Data-driven classification of coronal hole and streamer belt solar wind

Solar Phys (2020) 295:41
https://doi.org/10.1007/s11207-020-01609-z

Data-Driven Classification of Coronal Hole and Streamer Belt Solar Wind

Téo Bloch[1] · **Clare Watt**[1] · **Mathew Owens**[1] · **Leland McInnes**[2] · **Allan R. Macneil**[1]

Received: 28 January 2019 / Accepted: 2 March 2020

Abstract We present two new solar wind origin classification schemes developed independently using unsupervised machine learning. The first scheme aims to classify solar wind into three types: coronal-hole wind, streamer-belt wind, and 'unclassified' which does not fit into either of the previous two categories. The second scheme independently derives three clusters from the data; the coronal-hole and streamer-belt winds, and a differing unclassified cluster. The classification schemes are created using non-evolving solar wind parameters, such as ion charge states and composition, measured during the three *Ulysses* fast latitude scans. The schemes are subsequently applied to the *Ulysses* and the *Advanced Compositional Explorer* (ACE) datasets. The first scheme is based on oxygen charge state ratio and proton specific entropy. The second uses these data, as well as the carbon charge state ratio, the alpha-to-proton ratio, the iron-to-oxygen ratio, and the mean iron charge state. Thus, the classification schemes are grounded in the properties of the solar source regions. Furthermore, the techniques used are selected specifically to reduce the introduction of subjective biases into the schemes. We demonstrate significant best case disparities (minimum ≈8%, maximum ≈22%) with the traditional fast and slow solar wind determined using speed thresholds. By comparing the results between the in- (ACE) and out-of-ecliptic (*Ulysses*) data, we find morphological differences in the structure of coronal-hole wind. Our results show how a data-driven approach to the classification of solar wind origins can yield results which differ from those obtained using other methods. As such, the results form an important part of the information required to validate how well current understanding of solar origins and the solar wind match with the data we have.

Keywords Solar wind · Classification · Machine learning

✉ T. Bloch
t.bloch@pgr.reading.ac.uk

[1] Department of Meteorology, University of Reading, Reading, UK

[2] Tutte Institute for Mathematics and Computing, Ottawa, Canada

Published online: 17 March 2020

1. Introduction

The solar wind comprises streams of ionised particles which travel nearly radially from the Sun through the heliosphere. From the earliest *in situ* observations, it was clear that the solar wind could be broadly classified into two types, fast and slow (Neugebauer and Snyder, 1966; Stakhiv *et al.*, 2015). This duality was found to extend beyond the local solar wind speed, but is present in the elemental composition and ion charge states of the solar wind, indicating very different coronal source properties of fast and slow wind (von Steiger *et al.*, 2000; Geiss, Gloeckler, and Von Steiger, 1995). Fast wind is found to originate from coronal holes (Sheeley, Harvey, and Feldman, 1976). These are magnetically open regions of the corona where the plasma can freely escape, meaning that coronal holes appear dark in EUV emission. The formation and release of the slow wind is a current area of research, but it originates from the vicinity of closed coronal magnetic structures such as the streamer belt (Antiochos *et al.*, 2011; Ko *et al.*, 2006; Xu and Borovsky, 2015; Brooks, Ugarte-Urra, and Warren, 2015). At solar minimum, coronal holes cover the polar regions, with the streamer belt confined close to the solar equator. At solar maximum, the coronal field is far less ordered. The resulting variation of the solar wind speed can be seen in Figure 1 of McComas *et al.* (2013). Despite the breakdown of the latitudinal dependence at solar maximum, there is still a good separation between streams of different speeds. This suggests that despite the activity, the source regions remain isolated from one another, and there is not significant mixing of the streams.

While appealing, the traditional two-type solar wind paradigm is not unique, with a number of different observationally-determined solar wind types proposed. A two-type scheme has been proposed by Zhao, Zurbuchen, and Fisk (2009), a three-type scheme has been proposed by Stakhiv *et al.* (2015), a four-type scheme has been proposed by Xu and Borovsky (2015) and been built upon using machine learning by Camporeale, Carè, and Borovsky (2017), and even a six-type scheme has been proposed by Zhao *et al.* (2017). Furthermore, Heidrich-Meisner and Wimmer-Schweingruber (2018) have proposed a two-type classification scheme, and a two–seven type scheme (depending on interpretation) using the k-means clustering algorithm (MacQueen, 1967) implemented using the C++ library, Shark (Igel, Heidrich-Meisner, and Glasmachers, 2008). In each of these categorisation schemes the properties of each solar wind type are quantitatively different from one another, an essential factor when performing statistical studies of heliospheric phenomena driven by the solar wind.

The Zhao, Zurbuchen, and Fisk (2009) scheme sought to classify the solar wind into coronal-hole wind or non-coronal-hole wind. Stakhiv *et al.* (2015) classified the solar wind into coronal-hole wind, wind due to magnetic reconnection at the boundary of large scale streamers, and a boundary wind which originates from the edges of coronal holes. Xu and Borovsky (2015) described a scheme which encompasses coronal-hole wind, sector-reversal region wind emitted from the top of helmet streamers, and streamer-belt wind. The streamer-belt wind is comprised of two types: pseudostreamers and helmet streamers. These occur when two loop arcades separate a pair of like-signed coronal holes, and when a single loop arcade separates two coronal holes of opposite polarity, respectively (Panasenco and Velli, 2013; Owens, Crooker, and Lockwood, 2014). Zhao *et al.* (2017) split the solar wind into six types: coronal hole, active region, quiet Sun, active-region boundary, coronal-hole boundary, and helmet streamer. The regions in this case are not determined by coronal signatures in the solar wind, but instead by direct mapping to the Sun. A ballistic method is used to map to the solar source-surface, and then an extrapolation is made using the potential-field source-surface model (Altschuler and Newkirk, 1969; Schatten, Wilcox, and Ness, 1969) to map to the photosphere.

As part of the Machine Learning Techniques for Space Weather book (Camporeale, Wing, and Johnson, 2018), Heidrich-Meisner and Wimmer-Schweingruber (2018) present a systematic analysis of applying a simple unsupervised machine learning algorithm, k-means, to the classification of solar wind types. A variety of parameter spaces are investigated (13 different sets are used), as is the choice of the number of clusters for which the algorithm should search. The first k-means scheme proposed is a coronal hole *versus* slow wind scheme, whilst the second uses k-means to find seven clusters (where the number of clusters to find was a data-driven choice). The latter scheme provides results which are significantly more open to interpretations. The authors state that they find: two coronal-hole wind classes, though one may comprise interplanetary coronal mass ejection (ICME) plasma; one primary slow solar wind class; and four potential sub-classes of slow solar wind, where two are compressional/rarefaction regions surrounding a stream interaction, another is very slow, dense and cool wind, and the final is even more dense, has high charge states and is cool (though again, this may represent undetected ICMEs).

Aside from the growing evidence that the simplistic solar wind speed categorisation scheme is not adequate for distinguishing solar source regions, there is a more direct reason that such a scheme is not appropriate for many of the datasets that exist: co-rotating interaction regions (CIRs). CIRs are the compression regions that form when high-speed solar wind streams catch up to low-speed streams as they travel through the heliosphere. Since coronal-hole wind (CHW) and streamer-belt wind (SBW) typically show latitudinal dependence, CIRs do not tend to form everywhere. Instead, there is a tendency towards the ecliptic plane due to the inclination of the solar rotation axis (Crooker *et al.*, 1999; Borovsky and Denton, 2016). The result is that much of the solar wind in the ecliptic plane undergoes interaction of high- and low-speed streams. Such mixing causes high-speed streams to slow down and slow-speed streams to speed up. Thus, speed is not the most reliable means to distinguish different coronal sources.

The scientific motivation behind the current work is to provide two new solar wind classification schemes. They will be developed using unsupervised machine learning techniques so as to reduce scientific subjectivity. By using novel techniques which have their own unique biases, this work will provide new information towards validation or benchmarking of existing solar wind classification models. As with any scientific work, the total removal of any subjective influence is near impossible. Our methods hope to address the scientific subjectivity in the determination of classification boundaries, and number of solar wind types. The Bayesian Gaussian Mixture (BGM) scheme addresses the former point, whilst the Uniform Manifold and Projection (UMAP) scheme addresses both.

2. Data

For this analysis, data from the *Ulysses* spacecraft's (Wenzel *et al.*, 1992) *Solar Wind Observations Over the Poles of the Sun* (SWOOPS, Bame *et al.*, 1992) and magnetometer (Balogh *et al.*, 1992) instruments have been primarily used. The motivation for this usage is that, unlike in-ecliptic spacecraft such as the *Advanced Composition Explorer* (ACE), *Ulysses* has a polar solar orbit and enables sampling of the pure CHW during the high-latitude phase of the mission at solar minimum. The *Ulysses* mission contains three 'fast latitude scans', which are periods at perihelion when the spacecraft covers almost the full solar latitude range in a relatively short amount of time (approximate dates: 15/08/94 – 20/08/95, 01/11/2000 – 01/11/2000, and 01/02/07 – 01/02/08.). Particularly for the two solar minimum fast latitude scans, solar wind types can be well separated by their latitudinal dependence. The latitude scans comprise ≈3 years worth of total data, whilst the whole dataset is ≈19.5 years

(1990 – 2009). The data are mean-resampled into three-hourly steps to match the cadence of the compositional data. In practice, this yields 8227 (8139) latitude-scan data points, 46893 (45463) total *Ulysses* data points and 38108 (23665) ACE data points for the BGM (UMAP) scheme. The UMAP scheme has fewer points due to a larger parameter space; the whole data point must be discarded if any one parameter is bad.

Solar Wind Ion Composition Spectrometer (SWICS, Gloeckler *et al.*, 1998), *Solar Wind Electron Proton Alpha Monitor* (SWEPAM, McComas *et al.*, 1998) and magnetometer (Smith, Heureux, and Ness, 1998) data from ACE (Gloeckler *et al.*, 1998) are also used. As ACE is confined to the ecliptic plane (at the first Lagrange point, L1, just upstream of Earth), it rarely samples the CHW without it first interacting with SBW.

The classification scheme developed from the *Ulysses* data will be applied to ≈13 years of ACE data (1998-2011). This will allow more statistical insight into the link between solar wind source regions and space-weather events. All the *Ulysses* and ACE data are mean-resampled into three-hourly data.

Since solar wind speed is a poor parameter choice for classification, other parameters must be used. In order to relate properties to the coronal source conditions, parameters should ideally remain constant as the solar wind flows from the Sun. For this task, ion charge state ratios are the obvious choice, since they are well known to be non-evolving parameters after a few solar radii (Pagel, 2004; Geiss, Gloeckler, and Von Steiger, 1995). The reason for this is that the electron mean free path becomes so large that interactions are negligible (Owocki, Holzer, and Hundhausen, 1983). Between the *Ulysses* and ACE spacecraft, the common charge state and composition measurements are: O^{7+}/O^{6+}, C^{6+}/C^{5+}, Fe/O, $< q_{Fe} >$, and He^{2+}/H^{1+} (where fractions signify the relative density ratio). Further to these, Burlaga, Mish, and Whang (1990) describe how the proton specific entropy, S_p, is a good stream signature since it only diverges 10% between 1 – 5 AU ($S_p = \frac{T_p}{\sqrt{n_p}}$).

3. An Intuitive Classification Scheme

Many studies of the classification of the solar wind often rely on scientists' intuition. Xu and Borovsky (2015) (and by proxy Camporeale, Carè, and Borovsky, 2017) state that their method of determining classification boundaries is that they are "chosen by eye". Zhao, Zurbuchen, and Fisk (2009) base their identification of ICMEs (Cane and Richardson, 2003) on the work of Richardson (2004), who state their choice of parameter boundaries are "somewhat arbitrary". Such expert intuition is undoubtedly valuable, but extending this intuition from the abstract to the mathematical is a necessary progression.

In order to enable comparison with the machine learning approaches introduced below, we essentially reproduce the threshold approach in two-parameter space. The chosen parameters are O^{7+}/O^{6+} and S_p. These parameters are chosen based on the work of Zhao, Zurbuchen, and Fisk (2009) and Xu and Borovsky (2015), respectively. The methodology is as follows. Firstly, we take the log of our data and then plot the occurrence density. By visually inspecting the result, we see groupings within the data. These groupings are subsequently separated by placing a line (linear in log-space) to divide them. This dividing line forms the classification boundary between the two groupings. Not only does such a model allow for investigating the physical premise of the classification schemes introduced subsequently, but it will also be used as a benchmark to show that further results are not wholly unique to more complicated methods.

Figures 1a and 1b present two identical occurrence density plots of the whole *Ulysses* dataset, wherein two populations are clearly visible (the colourbar is log scaled). The panels

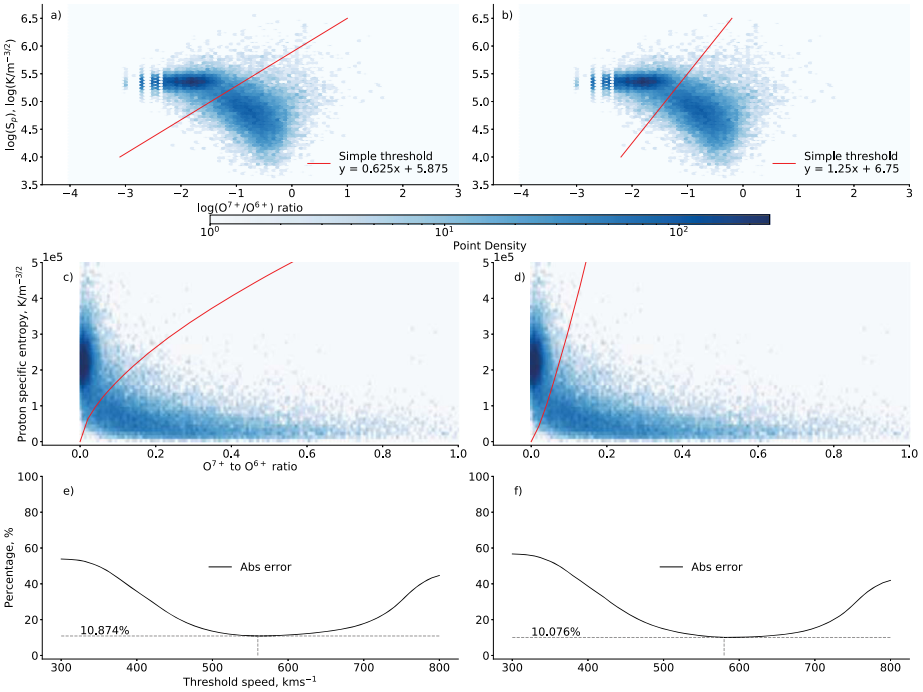

Figure 1 The intuitive classification scheme with comparison to classical speed-threshold methods. Panels a and b show occurrence density plots of ≈19.5 years of *Ulysses* data as a function of O^{7+}/O^{6+} and S_p. Note the logarithmic colourbar scale. Each plot shows a different threshold chosen to separate the two main solar wind populations, assumed to represent coronal-hole and streamer-belt winds. Panels c and d present this classification in linear space. The coronal-hole wind is represented by low O^{7+}/O^{6+} and high S_p, and vice versa for the streamer-belt wind. Finally, panels e and f present the resulting mis-classification of solar wind using the speed-threshold method, in terms of the absolute error.

include a different choice of threshold line which plausibly separates the populations in the data. Thresholding has been performed twice to highlight the fact that the result is not specific to a given threshold. Figures 1c and 1d show the results of the classification in linear space. Despite being a simplistic classification scheme, there are benefits to its use; it is transparent and based on parameters which are known indicators of solar source regions, thus reducing the impact of solar wind stream interaction. From panels a, b, c and d it is inferred that the data found to the left (right) of the classification thresholds are CHW (SBW) due to their lower (higher) charge state ratios and higher (lower) proton specific entropy.

We make a simple comparison to demonstrate how the classical speed-threshold scheme does not well-capture the origins of the solar wind, as compared to the intuitive scheme. The *Ulysses* data are divided into fast- and slow-stream wind according to various speed thresholds. We calculate the proportion of points in each case where the results of the speed-threshold scheme do not agree with the results of the intuitive scheme. The total number of discrepant points is divided by the total number of points in the data used, giving a fraction describing the relative difference between the speed-threshold and intuitive schemes. This is shown in Figures 1e and 1f.

Whilst differences can be seen due to individual thresholds used in the two intuitive schemes, it is their similarity which is of most import. Specifically, both show comparative

inaccuracies of the speed-threshold scheme greater than 10% for all speed thresholds. Already, this simple scheme highlights the potential shortfalls of the speed-threshold scheme.

4. Machine Learning Schemes

Whilst the intuitive scheme is undoubtedly useful, it still contains subjective decisions about which parameters to use, the number of solar wind types to identify, and the decision boundaries. Here, more objective (data-driven) and mathematical methods are presented. Unsupervised machine learning will be used to create two new classification schemes; with reduced subjectivity and more algorithmic reproducibility. The latter point specifically contrasting the choice of decision boundaries by eye.

Machine learning (ML) can be split into two main categories; supervised and unsupervised. Supervised ML describes the techniques which produce and optimise a function to map from an input (data) to an output (class label), given a set of example (training) input-output pairs (Russell and Norvig, 2009). By contrast, unsupervised ML describes the subset of techniques which are used to determine effective ways of mathematically separating data with no predetermined class labels. Instead of a boundary function being optimised by a predictive performance metric, the optimisation is often focussed on improving the separation of data clusters. In this way, unsupervised ML can be applied to data with less bias, allowing for groupings in the data to be found mathematically rather than being influenced by what one may expect to find *a priori*. Unsupervised ML is a data-driven approach to classification. Its purpose is to determine an underlying structure in the data and find quantitative separations between discrete regions. As such, the algorithms find that which is already present in the data (subject to algorithm specific limitations).

The first new scheme will allow for the determination of a third solar wind category. This category represents data which is difficult to assign to either CHW or SBW, and hence be referred to as unclassified data. The second proposed scheme will independently determine the number of solar wind categories.

To cluster the whole *Ulysses* dataset is a bad idea for several reasons: as previously mentioned, there is limited pristine data; clustering is computationally expensive; and, it is inefficient for classifying new data (since the clustering would have to be re-performed). To address these issues, the clustering is performed on the three latitude scans. This allows the clustering to be performed only on the more pristine data, with higher latitudinal dependence; provides a more manageable dataset for clustering, reducing computational complexity; and ensures the ability to classify any new data efficiently.

Subsequently, the results of the clustering are applied to classify $\approx$19.5 years of *Ulysses* data and $\approx$13 years of L1 ACE data. The independence of the classifications from solar wind speed allow them to be applied to the ACE dataset despite significant solar wind stream interactions.

5. The Bayesian Gaussian Mixture Scheme

Some of the literature regarding solar wind classification is built upon classification boundaries which are chosen subjectively (*e.g.* "arbitrarily" or "by eye"). We present a Bayesian Gaussian Mixture (BGM) classification scheme which uses unsupervised machine learning to mathematically determine the optimum data-driven decision boundary between solar wind types (subject to the suitability of the Gaussian assumption).

The BGM algorithm iteratively fits a Gaussian mixture (McLachlan and Peel, 2000) to the data. During each iteration, variational inference is implemented to do two things: firstly, find the probability of each point being generated by the mixture; and secondly, refit the mixture to the points using information from the prior distributions (Attias, 2000; Bishop, 2006) (for further information regarding variational inference, see Appendix A). Once convergence has been reached, the algorithm outputs the cluster label for each point (*i.e.* the label of the Gaussian in the mixture to which it belongs), and the information describing the distributions (*e.g.* mean and variance). The latter information is extremely useful, as it allows for the Gaussian mixture to be stored, removing the need to run the algorithm every time. With the Gaussian mixture stored, application to data classification is straightforward: firstly, each new data point is mapped into the pre-established normalised space; then, the posterior probability of each component Gaussian given the data point is calculated (further detail given in Appendix A, or see *e.g.* Gelman *et al.*, 2013); and finally the point is assigned to the component with the highest probability of generating it (as per Camporeale, Carè, and Borovsky, 2017). The BGM is here applied using the algorithm from the scikit-learn package available for python (Pedregosa *et al.*, 2011).

We do not use k-means as Heidrich-Meisner and Wimmer-Schweingruber (2018) have done. From the standpoint of the objective functions being optimised by k-means and the BGM (as opposed to the algorithms used to attempt the optimisation), k-means is strictly a special case of a Gaussian mixture model. That is, if you choose a Gaussian mixture with K components and fix the Gaussians to be spherical (scalar multiple of the identity for covariance) then the means of the maximum-likelihood estimate for the mixture are the centroids that minimise the distance from the data to the centroids. Algorithmically k-means and the BGM method use different optimisation techniques, but philosophically k-means is a subset of a Gaussian mixture model. Using a BGM rather than k-means allows for non-circular clusters to be appropriately described using ellipses, by relaxing the restriction that the Gaussians must be spherical.

To test the validity of the above arguments against using k-means, we have investigated how the results differ from the BGM scheme. Overall, the results from k-means are qualitatively the same as those from the BGM (*i.e.* the majority of data are assigned the same class in both schemes), but with drawbacks. Such drawbacks include an apparent increase in the mis-classification of *Ulysses* CHW data, and incongruent speed distributions for the unclassified data between *Ulysses* and *ACE*. These differences are due to the comparatively poor way of determining classification boundaries, and the changes in the objective functions being optimised. These differences both highlight that k-means is less suited to classification in the way we have applied the BGM.

The BGM approach allows probabilistic classifications, antithetical to the intuitive scheme. Whilst fitting Gaussians to data is a common practice, there is the inherent shortfall of the approximation becoming less valid as a dataset diverges from being normally distributed. As such, whilst we may be more objective in the fitting procedure and gain information (*e.g.* probabilities), the results must always be considered carefully in terms of the validity of the Gaussian assumption.

Despite the BGM producing probabilistic results, this study will use hard decision boundaries. Points will be assigned to the Gaussian which most likely generated it. Such an approach is entirely adequate for comparing between different solar wind classification schemes (since most others use hard boundaries). In theory, problems may arise if there were many data points yielding comparable (60% or 40%) probabilities of belonging to multiple classes, but in our case fewer than 1-in-10 data points have probabilities below 90%. Hence, minimal data are affected by our use of hard boundaries.

Each parameter in the dataset is normalised to a zero mean and unit standard deviation to reduce any bias that the heteroscedasticity of the variables could introduce to the algorithm. The method of normalisation is through the standard score:

$$x' = \frac{x - \mu}{\sigma}. \tag{1}$$

Here x' is the normalised value, x is the initial value, μ is the mean of the population, and σ is the standard deviation of the population.

The BGM algorithm does require user-specified parameters. i) The number of components in the mixture. Since this study focuses on classifying the solar wind into coronal-hole and streamer-belt wind whilst accounting for data which is difficult to classify, the algorithm is set to fit a three-component Gaussian mixture to the data. ii) The precision prior on the mean distribution. The motivation of this research is to avoid incorporating bias where possible. Therefore, the prior was set to be flat, allowing all possible mean positions to be equally weighted. iii) The number of initialisations. The algorithm was set to perform clustering with 30 random initialisations of the means to ensure that the convergence was not on a local maximum/minimum. The result with the largest value of the lower bound of the likelihood is kept. Convergence is reached when the change in likelihood is less than 10^{-5} between iterations. Higher values of the likelihood correspond to higher degrees of confidence that the model could produce the data (see *e.g.* Gelman *et al.* (2013) for a detailed description of likelihood). As such, by choosing the model with the highest lower bound, the baseline degree of confidence is highest.

The Gaussian mixture best describing the data can be described by the three-component means $\mu_{1-3}(\mathrm{O}^{7+}/\mathrm{O}^{6+}, S_p)$, covariances $cov_{1-3}(\mathrm{O}^{7+}/\mathrm{O}^{6+}, S_p)$ and their respective weightings. In the normalised space the means and covariances are as follows:

$$\mu_1 = (-0.3779, 0.6252),\ cov_1 = \begin{pmatrix} 0.0019 & -0.0012 \\ -0.0012 & 0.1833 \end{pmatrix},$$

$$\mu_2 = (0.4235, -1.1145),\ cov_2 = \begin{pmatrix} 0.3759 & -0.1009 \\ -0.1009 & 0.1067 \end{pmatrix},$$

$$\mu_3 = (3.2836, -0.0057),\ cov_3 = \begin{pmatrix} 15.9839 & -3.0918 \\ -3.0918 & 6.3546 \end{pmatrix},$$

and the weights are 0.6238, 0.3497, and 0.0265, respectively.

Figure 2a presents the results of applying the BGM clustering algorithm to the latitude-scan data, and 2b shows how the clusters map to the solar wind speed and proton temperature. The combination of the two parameters used allow the clustering to map the solar wind well to either coronal holes or the streamer belt. Cluster one, with low average $\mathrm{O}^{7+}/\mathrm{O}^{6+}$ and high average S_p, represents CHW. Cluster two, with higher average $\mathrm{O}^{7+}/\mathrm{O}^{6+}$ and lower average S_p, represents SBW. Cluster three is thus the unclassified data. The projection of the clustering into the solar wind proton speed and temperature shows clearly that the clustering is capturing distinct populations. The interaction of CHW and SBW can be seen by the overlapping of the two groups along the speed axis.

To investigate the stability of the clustering, the procedure was performed a further 300 times using random sub-samples of 90% of the data. Upon completion of each iteration, the mean value of each component Gaussian was recorded. Once completed, the standard deviations and inter-quartile ranges of the distributions of means are calculated. If the clustering had found a local maximum/minimum in the data, we would expect there to be significant differences between the results of a single run compared with the statistical results of many

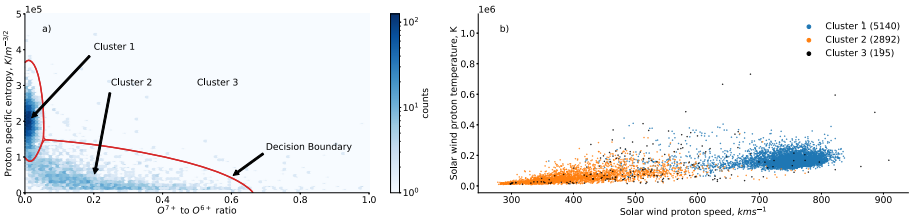

Figure 2 Classification with the BGM scheme. Panel a presents the results of Bayesian Gaussian Mixture algorithm clustering on the normalised O^{7+}/O^{6+} and S_p data. The plot has been trimmed in both x and y to better display clusters one and two, as such a small number of data points from clusters two and three are not shown. Panel b presents the projection of the clustered data into solar wind proton speed and temperature.

Table 1 The results of three-hundred 90% sub-sample runs of the Bayesian Gaussian Mixture algorithm on the fast latitude-scan *Ulysses* data. The values in brackets after the component number indicate the number of data points being included (some are excluded due inconsistent labelling). μ is the mean of the individual component means. σ is the standard deviation of the means, the value in brackets represents the inter-quartile range. All values of the mean, standard deviation, and inter-quartile range are given in the normalised space.

		Sub-sampled Bayesian Gaussian Mixture Model		
		Component one (296)	Component two (296)	Component three (300)
O^{7+}/O^{6+}	μ	−0.3778	0.4235	3.281
	σ	0.0002(0.0003)	0.0052(0.0064)	0.128(0.165)
S_p	μ	0.6251	−1.115	−0.0093
	σ	0.0022(0.0028)	0.003(0.005)	0.0704(0.0984)

runs. The results of the analysis are presented in Table 1. The BGM algorithm does not systematically label the Gaussians and so eight of the recorded mean values were incorporated into the incorrect group, and thus removed.

The proportionally small standard deviations and inter-quartile ranges signify that the clustering is stable, and that the individual runs do not deviate greatly from the average values. The means of the component Gaussians used to classify the data in the normalised space are [−0.3779, 0.6252], [0.4235, −1.115], and [3.284, −0.0057] for clusters one, two, and three, respectively. Comparing these values to those presented in Table 1, we see that they are very much in-line with the standard behaviour.

5.1. BGM Scheme: Application

The clustering described above has been used to develop a solar wind classification scheme, based on the fast latitude-scan subset of the *Ulysses* data. In this section we will apply the classification scheme to the whole *Ulysses* and ACE datasets.

5.1.1. Ulysses

Figure 3 presents the results of the classification of the whole *Ulysses* dataset. In Figure 3a one can see how the SBW (cluster two) appears to deviate from a Gaussian. On the contrary (though less obvious from the plot) is that the CHW (cluster one) is well approximated by a Gaussian, especially in comparison to the SBW. However, a significant portion of the

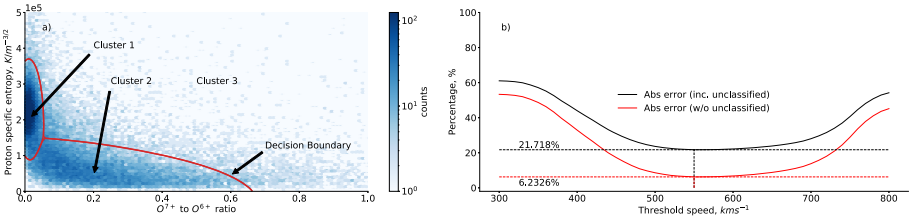

Figure 3 Panel a presents the results of applying the BGM fast-latitude-scan classification to the whole *Ulysses* dataset, shown in the O^{7+}/O^{6+} and S_p space. Cluster one, two, and three represent CHW, SBW, and unclassified data, respectively. Panel b shows the mis-classification of the data using a simple speed threshold, both including and excluding unclassified data (as determined by the BGM technique).

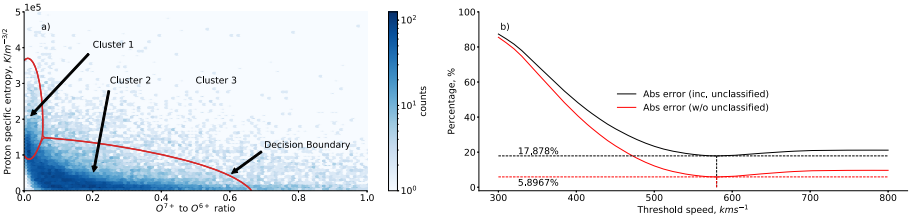

Figure 4 The results of applying the BGM scheme to the whole *ACE* dataset, and the subsequent comparison to taking speed thresholds. Panels a and b are presented in the same format to Figure 3. Cluster one, two, and three represent CHW, SBW, and unclassified data, respectively.

data that we might consider to be difficult-to-classify has been captured as such by the algorithm. When compared with taking simple speed thresholds, Figure 3b, there is a minimum of ≈22% disparity in the results. Again, this is highly suggestive that the traditional method falls short of adequate for many applications. Also shown in Figure 3b is the disparity when the unclassified data are ignored. This has been included to allow a more like-for-like comparison (since both schemes can be considered two-type schemes). In this way the disparity is reduced to ≈6%, suggesting that the speed threshold captures the cores of the clusters. Nonetheless, the speed-threshold scheme oversimplifies the classification of solar wind data, and importantly gives too much confidence to the classification of borderline data.

5.1.2. ACE

Figure 4 presents the results of the BGM classification of the whole ACE dataset. In Figure 4a there are considerable difference as compared with the *Ulysses* data; as expected, there is significantly less CHW (cluster one) and more SBW (cluster two). When compared with simple speed-threshold classification, Figure 4b shows that there is a minimum ≈18% disparity in the results. Again, ignoring the unclassified data, the disparity is reduced; though, the same issues persist.

5.2. BGM Scheme: Analysis

To view the way in which the classification of the *Ulysses* data maps to velocity and solar latitude, a McComas *et al.* (2013) style visualisation is presented in Figure 5a. Both the latitudinal and speed dependent nature are clearly present. These dependent variables have

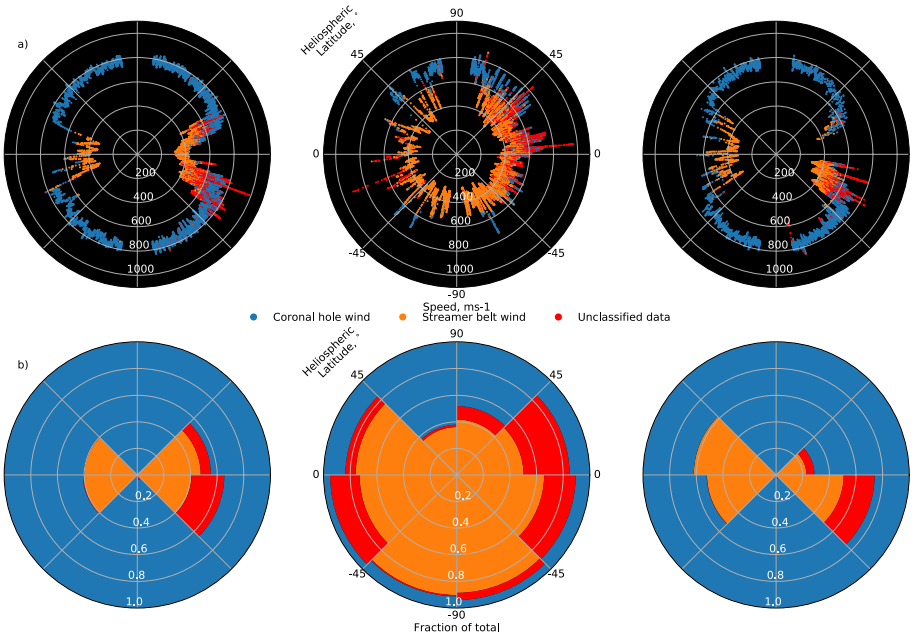

Figure 5 The BGM classification of the whole *Ulysses* dataset. Panel a presents the projection of the BGM classification scheme onto radial plots of solar wind speed and *Ulysses* heliospheric latitude. Panel b presents the fraction of each classification type in each of the octet segments of the plot. Each plot represents an orbit of *Ulysses* around the Sun, with the ecliptic plane in the east-west direction. In both panels, time increments clockwise, starting from aphelion at 8.6 degrees below east. The first and third plots of panels a and b are the orbits where perihelion occurred at solar minimum, whilst in the middle plots, perihelion occurred at solar maximum.

not been used in the classification scheme, but the correlation is expected. Capturing the predicted behaviour shows that the initial choice of parameters is well-informed.

Figure 5b shows that the unclassified data is skewed towards the aphelion of the orbit. It is worth noting that, due to the slower motion of the spacecraft, there is considerably more data per latitudinal increment at aphelion than other portions of the orbit. After accounting for the expected increase in unclassified data, there remains a significant disparity in the distribution of unclassified data. The aphelion regions of the orbit present more unclassified data than the perihelion regions.

Figure 6a presents the speed distributions of the three BGM classifications from *Ulysses* data. Note that these speeds were not used in the classification in any way. The SBW shows no significant bi-modality, and appears to follow a Maxwellian distribution. Both the CHW and the unclassified data show some suggestion of being bi-modal, each with their secondary peak aligning close to the primary peak of the other (whilst subtle, the secondary peaks are present, *viz.* ≈ 775 km s^{-1} for the unclassified, and ≈ 500 km s^{-1} for the CHW). This suggests that the classification scheme may be having trouble differentiating between the two types (see discussion of Figure 7). Figure 6b presents the same distributions, but obtained from the classification of the ACE dataset. We see the significant drop in CHW, but observe that the distributions of both CHW and SBW are not double-peaked. The unclassified data in the ACE classification is considerably flattened, suggesting that the difficulty in classification may be ubiquitous in the ecliptic plane.

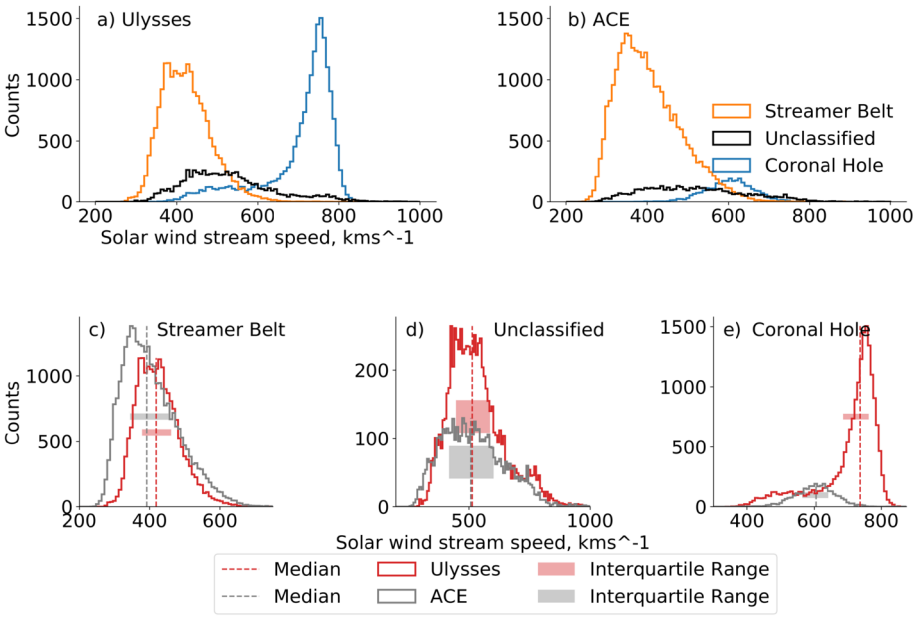

Figure 6 Distributions of solar wind speeds within each cluster found by applying the BGM classification scheme. Panels a and b present the comparisons for the distributions in the *Ulysses* and ACE datasets, respectively. Panels c, d and e present the comparisons of like-clusters between the datasets. The latter panels also include the mean and inter-quartile range of each distribution.

Figures 6c, 6d and 6e show direct comparisons of the distributions of solar wind classifications in the ACE and *Ulysses* datasets. There is good qualitative agreement in the distributions of the SBW and unclassified data, suggesting that the scheme well-captures streamer-belt solar wind structures, as well as consistently identifying the unclassified data. However, the CHW distributions for ACE and *Ulysses* are very different. Given the general trend of coronal holes towards higher latitudes, observing significantly less CHW in the ecliptic plane is not unexpected. Furthermore, seeing that the CHW in the ecliptic plane is generally slower is in line with the idea that the fast wind is slowed down due to stream-stream interactions in the solar wind. The difference in the unclassified data is almost exclusively related to the amplitude of the peak. The means and inter-quartile ranges show good agreement.

To better understand the double peak in the *Ulysses* CHW speed distribution, the data are further split by spacecraft location. We use a threshold of 600 km s^{-1} to separate the two CHW peaks. Figure 7 shows the occurrence of the two CHW distributions as a function of radial distance from the Sun. Given the long orbital duration of *Ulysses* and its associated latitudinal variation, the data the spacecraft obtains is convolved with the solar cycle and latitude. Thus, the average latitude and sunspot number (and the respective standard deviations) are calculated for each histogram bin. The sunspot number shows very little structure, and no overall trends matching the distributions of radial distance shown. In contrast, there is a clear trend between the absolute heliospheric latitude and the radial distances contained within the secondary peak (lower speed) of the CHW speed distribution. The trend suggests that the majority of the secondary peak data is obtained both far from the Sun and closer to the ecliptic plane.

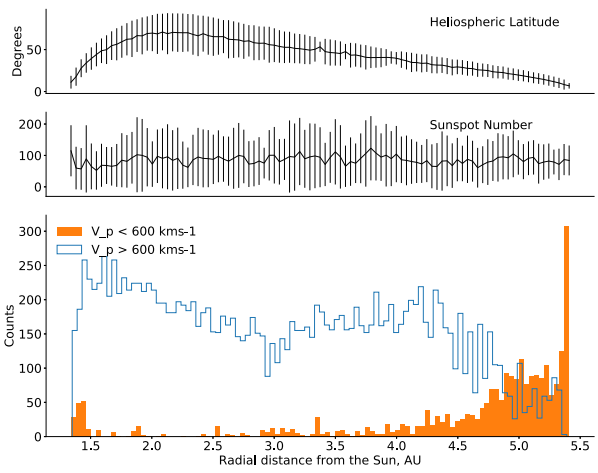

Figure 7 Occurrence of high- and low-speed CHW (using a 600 km s^{-1} threshold) as a function of radial distance. This speed threshold is chosen to isolate the two peaks observed in the distribution of speed in the CHW classification of *Ulysses* data. The top two panels present the average absolute heliospheric latitude of the *Ulysses* spacecraft, and the average sunspot number within each bin of the histogram. The error-bars are the standard deviation of the data contained in each bin.

6. The UMAP Scheme

The BGM scheme presents a step forward in creating a classification scheme which is more objective and physically motivated. However, there remain some drawbacks: only a subset of the possible parameters are used; there is an inherent assumption that the data are normally distributed; the number of component Gaussians must be specified in advance (reducing the objectivity of the scheme); moving to higher dimensions reduces the interpretability of the results (what does a six-dimensional (6D) Gaussian look like or mean?); and, simply having three Gaussians does not provide much information about the substructure within each cluster. These issues are addressed by creating a further classification scheme using dimensional reduction and clustering. This scheme will specifically address the subjectivity introduced when designating decision boundaries by eye. Further, it will remove the subjectivity in determining the number of types of solar wind by deriving the number of clusters from the latent structures in the data itself. We will apply the UMAP algorithm for dimension reduction, and the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm to subsequently cluster the low-dimensional representation of the data.

Datasets are often expressed in terms of a large number of measurements or features. This means that each sample in the dataset is expressed as a vector, or point, in a high-dimensional space. It is often the case that the underlying structure of the dataset as a whole can be described in terms of a much smaller number of latent features, which dimensional reduction seeks to determine. More formally, while the ambient space in which a dataset lives may be high dimensional there often exists a much lower-dimensional manifold from which the data samples are (noisily) drawn. The UMAP algorithm (McInnes, Healy, and Melville, 2018) seeks to learn the topological structure of this manifold, and then find a low-dimensional representation of the data that has an equivalent topological structure. In this way UMAP can transform highly complex datasets into much simpler representations that still capture meaningful structural features of the original dataset. Due to the algorithm using stochastic gradient descent (Kusher and Yin, 2003), there are minor variations in results produced by UMAP each time it is performed. For further technical information, see Appendix B.

The HDBSCAN algorithm (Campello, Moulavi, and Sander, 2013) seeks to find dense regions (clusters) of a dataset that are otherwise separated from the rest of the data by regions where data are sparse. In particular it seeks to do this even when the dataset contains

background noise. To achieve this, HDBSCAN makes use of a density threshold (expressed as a minimum number of data samples required before a region can be considered "dense") and constructs a hierarchical tree of contiguous regions of density. Given a minimum size for a cluster, this tree can then be simplified resulting in a nested hierarchy of clusters. By selecting out the most persistent such clusters (over ranges of distance scales) a single flat clustering may be extracted. This results in an output of cluster labels where each point is either labelled with a cluster identity, or as noise. For further technical information, see Appendix C.

It is worth noting that the nature of the way UMAP works is almost guaranteed to result in non-convex clusters, and hence a clustering technique that is robust to this is required. By necessity this essentially means either a hierarchical method such as single linkage (Florek *et al.*, 1951) or average linkage (Sokal and Michener, 1958), or a density-based technique such as density-based spatial clustering of applications with noise (DBSCAN) (Ester *et al.*, 1996) or mean shift (Fukunaga and Hostetler, 1975) is required. That one should therefore consider the hybrid hierarchical density-based approach of HDBSCAN, over more simplistic methods such as k-means, is entirely natural.

UMAP does not limit the number of parameters that can be used. We apply UMAP to all of the non-evolving parameters (O^{7+}/O^{6+}, C^{6+}/C^{5+}, Fe/O, $< q_{Fe} >$, He^{2+}/H^{1+} and S_p). This 6D data-structure is projected into 2D, allowing subsequent clustering to be independent of any potential user-biases, since there is not a physical interpretation of the reduced-dimension axes (McInnes, Healy, and Melville, 2018). Whilst the axes are some non-linear function of the input dimensions, it is not possible to derive this function from the mapping. HDBSCAN clustering does not require any user-specified number of clusters, instead finding groupings by the intrinsic density structures present in the data.

As with the BGM scheme, the UMAP classifications are determined using the *Ulysses* fast latitude-scan data. The data is normalised using the MinMaxScaler function available in scikit-learn, as shown in the documentation for UMAP. This method individually normalises all of the parameters to be in the 0 – 1 range:

$$X_{scaled} = \frac{X_i - \min\{X\}}{\max\{X\} - \min\{X\}} \tag{2}$$

where X_{scaled} is the scaled value, X_i is the un-scaled data value, and $\min\{X\}$ and $\max\{X\}$ are the corresponding minimum and maximum values of that parameter. It is of no consequence that different normalisation schemes are used between the UMAP and BGM schemes, since comparison is only made in real space. Normalisation is simply a tool to facilitate unbiased dimension reduction and classification on a per-scheme basis. The normalised dataset is then reduced and clustered. The function mapping from 6D to 2D is stored, as are the classification parameters obtained by HDBSCAN.

Figure 8a presents the results of reducing and clustering the latitude-scan data. Figures 8b and 8c present the clustering projected into the O^{7+}/O^{6+} and S_p, and proton speed and temperature spaces, respectively.

To stress a point, the dimension reduction is simply one of the steps required to build the classification scheme. The lack of physical interpretation of the 2D space axes adds to the validity of the results, rather than facilitating the influence of current scientific ideas on the classification. The classification becomes entirely based on the latent structure in the data and as such, independent of biases or expectations we may hold.

The data in Figure 8a shows distinct groupings in the data. This implies that there are fundamental differences between the three groups in the 6D space. By inspecting panels b and c it is apparent that the distinction is the type of solar wind present. From panel b we

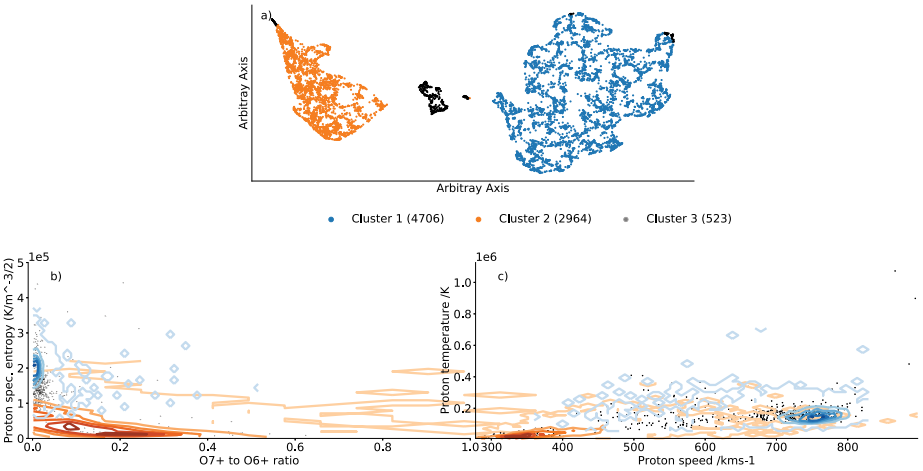

Figure 8 The results of performing the UMAP dimension reduction and HDBSCAN clustering on the *Ulysses* fast latitude-scan data. Panel a presents the dimensional reduction and subsequent clustering of the non-time-evolving solar wind parameters. Panel b presents the clustering of the reduced data projected onto the O^{7+}/O^{6+} and S_p space used in the previous two classification schemes. Panel c presents the clustering projected onto solar wind speed and proton temperature. From the latter panels it is inferred that clusters one, two, and three represent CHW, SBW and unclassified data, respectively. The contours are representative of the point density of data. Given the two different contours and that the data remains largely the same as in Figure 2, colourbars are not included. The noisy, final contour line is at the one-point level.

infer that cluster one is CHW, cluster two is SBW, and cluster three is the unclassified data. However, the unclassified data is more complicated than with the BGM scheme. Here, there are four distinct regions. Because the UMAP reduction preserves the latent structure of the higher-dimensional space, the isolation of the groupings provides information. The spatial separation between the unclassified data associated with the CHW, SBW or middle cluster suggests that in these are fundamentally different from one another. However, within-cluster separation of the unclassified data does not necessarily imply fundamental differences, since this could just be an artefact of imperfect projection of the 6D structure onto the 2D plane.

The distribution of the unclassified data is different in each of panels b and c. In panel b much of the unclassified data is grouped in the region where the CHW and SBW signatures overlap. This is expected due to the region being where the parameter values transition between the two types of solar origin, and as such classification uncertainty should exist. The remaining unclassified data which is spread throughout the CHW and SBW is due to the small pockets of unclassified data connected to each of the clusters in panel a. In panel c the unclassified data is more evenly spread around the core regions of each group. Had the unclassified data been grouped where the faster and slower regions overlapped, it would have suggested that the speed could be providing useful information about the types of solar wind present. Since it was not, the stream speed of the solar wind appears to be a less-informative parameter for classification schemes such as this one.

Since the UMAP reduction aims to maintain the structures in the 6D space, one can extract information based on the structures present in the data. The CHW group shows more spread in its internal structure, despite being understood to be less variable than SBW. This could suggest that there is some underlying variability in the CHW's parameter space, or that the manifold covering the CHW is a shape that does not lend itself to 2D reduction (*e.g.* a spherical manifold is difficult to project into 2D whilst retaining the topological structure).

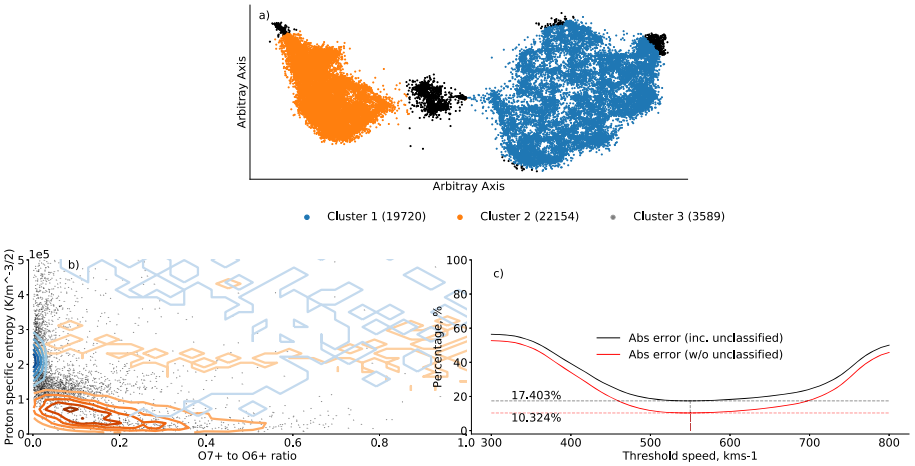

Figure 9 The results of classifying the full *Ulysses* dataset using the determined UMAP classification scheme. Panel a presents the whole *Ulysses* dataset reduced to 2D, and the results of the subsequent mapping to the clustering model created using the latitude-scan data. Panel b presents the clustering of the reduced data projected into O^{7+}/O^{6+} and S_p space. The contours are representative of the point density of data as in Figure 8, showing similar data to Figure 3. Panel c presents the comparison between the UMAP classification scheme and the traditional speed-threshold scheme. Clusters one, two, and three represent CHW, SBW and unclassified data, respectively.

6.1. UMAP Scheme: Application

The BGM scheme allows a simple way of classifying new data (the probabilities of each Gaussian giving the data). Using UMAP and HDBSCAN is not quite as straightforward. Fortunately, both techniques (as well as the scikit-learn MinMaxScaler) allow for the created mappings to be stored and applied to new data. In this way, reducing and clustering the whole *Ulysses* and ACE datasets is both straightforward and consistent.

6.1.1. Ulysses

Figure 9a presents the results of projecting the entire *Ulysses* dataset to the reduced-dimension space and the subsequent clustering. The data maintains the structure found in the reduction of the fast latitude scans. However, there are larger pockets of unclassified data, as well as significant linkage between the central unclassified data and clusters one and two. These features are likely to be the cause of the increased amount of unclassified data dispersed throughout the CHW in Figure 9b. There is proportionally more unclassified data present than there was in the original fast latitude-scan result. However, as shown in previous sections, the data being used is expected to be more variable and thus, may exhibit more unclassified data. Again, this scheme yields large disparities with the speed-threshold scheme; ≈20% (≈10% excluding the unclassified data).

6.1.2. ACE

The UMAP classification scheme is now applied to data from the ACE spacecraft; Figure 10 shows the results. Whereas the classification of the *Ulysses* data resulted in an almost even split between CHW and SBW with some unclassified data interspersed throughout, the

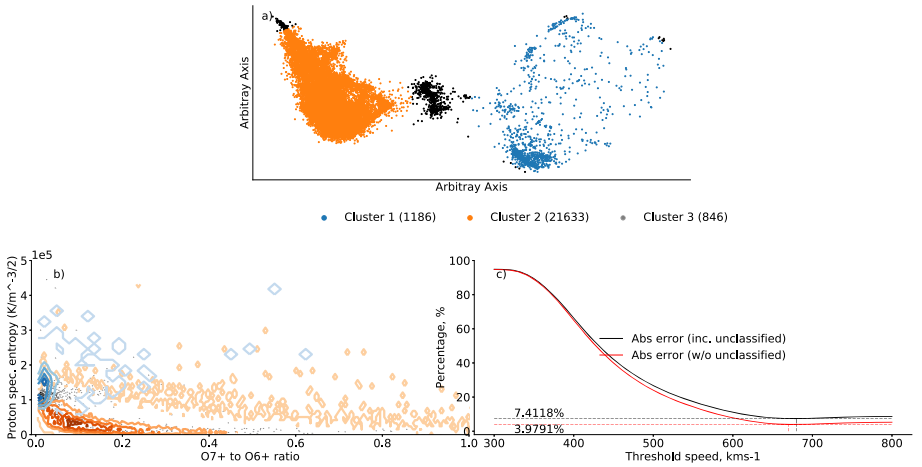

Figure 10 The results of classifying the ACE dataset using the determined UMAP classification scheme. Panel a presents the whole ACE dataset reduced to 2D, and the results of the subsequent mapping to the clustering model created using the fast latitude-scan data. Panel b presents the clustering of the reduced data projected into O^{7+}/O^{6+} and S_p space. The contours are representative of the point density of data as in Figure 8, showing similar data to Figure 4. Panel c presents the comparison between the UMAP classification scheme and the traditional speed-threshold scheme. Clusters one, two, and three represent CHW, SBW and unclassified data, respectively.

ACE results shown in Figure 10a are dominated by SBW, with only a small fraction being CHW or unclassified data. Comparing with the BGM scheme, the UMAP scheme identifies approximately half (proportionally) as much CHW.

Comparing to the speed-threshold method, we see that the UMAP classification of *ACE* data has a disparity of ≈8% (≈4% excluding unclassified data). This suggests closer agreement of UMAP with the traditional method than any of the other classifications. Such a low disparity is promising for the speed-threshold method. However, taking a threshold for skewed data may not be a fair way to split the data. Taking a speed threshold above any value found in the data gives a prediction error rate of ≈9%, simply due to small ratio of CHW and unclassified data to SBW.

6.2. UMAP Scheme: Analysis

To further validate the UMAP scheme, radial plots are shown in Figure 11. As before, the plots show that the classification scheme captures the overall speed and latitudinal dependence in the data (despite neither being used in the classification scheme itself). Whilst there is good agreement between the UMAP and BGM radial plots (Figures 11a and 5a, respectively), it is the differences which are interesting. The UMAP scheme unclassified data is more uniformly distributed and there is an increased amount of CHW at lower speeds as compared with the BGM scheme results. The first point is further evidenced in Figure 11b where the fraction of unclassified data is clearly more evenly spread throughout the octet of bins. This implies that the UMAP scheme is more able to classify points at the aphelion of the orbit in the ecliptic plane.

Figure 12 shows the distributions of the different solar wind classifications as a function of solar wind speed. The CHW and SBW distributions of the *Ulysses* data in panel a match

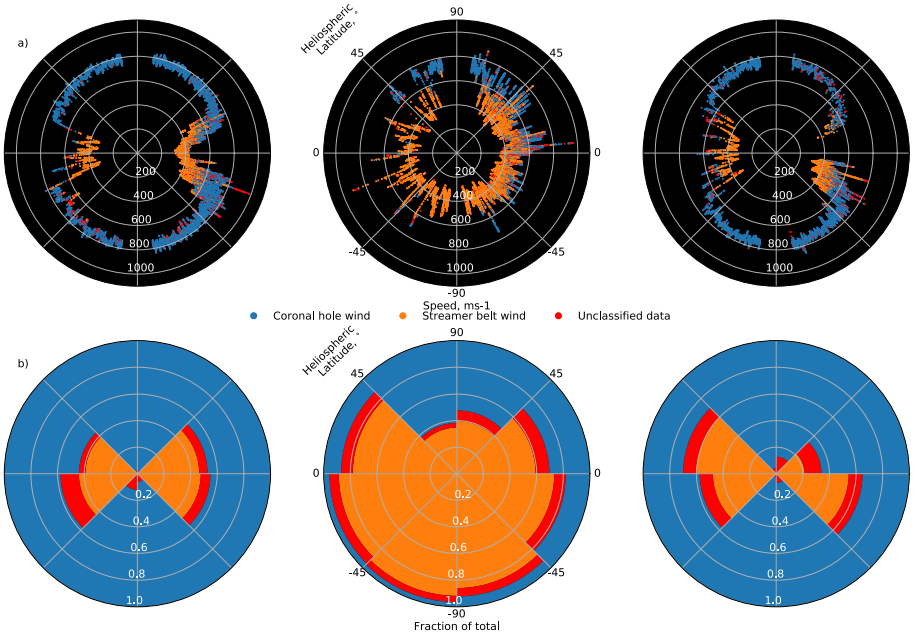

Figure 11 Plots showing how the UMAP classification of the whole *Ulysses* dataset maps to solar wind speed and solar latitude across its three orbits, as well as how the distribution of unclassified data changes over these orbits. Panels a and b are presented identically to those for the BGM scheme in Figure 5.

well with the distributions found from the BGM classification, including the secondary peak in the CHW. However, the unclassified data exhibits a more significantly bi-modal distribution. The peaks of the bi-modality align with the peak of the SBW and primary peak of the CHW. This may suggest that the unclassified data is comprised of points which lie in the tails of the CHW and SBW 6D distributions. Furthermore, the ACE data distribution matches the equivalent BGM result, despite displaying a heavy-tail distribution. The unclassified data, again, displays a bi-modal distribution, and the CHW is too sparse to make a fair comparison of anything but the predicted occurrence rate.

The comparative plots in Figures 12c, 12d and 12e present the differences between the distributions more clearly. In panel c the overall similarity is clear, though the ACE distribution shows the heavy tail. In panel d the peaks of the unclassified ACE data are shifted towards lower speeds. Finally, panel e shows further expected behaviour; the CHW from ACE is slower and a relatively minor contribution.

As with the BGM scheme, the double peak in CHW has been investigated, yielding results which are qualitatively equivalent to the results presented in Figure 7.

7. Discussion

The intuitive classification scheme, wherein an arbitrary threshold is applied to non-evolving solar wind parameters (such as ion charge states ratios and proton entropy), is a bridge between simple solar wind speed-threshold classification and the machine learning methods presented herein. It shows significant differences to the speed-threshold method. The latter method is not without its merits, for many applications the speed of the solar wind is

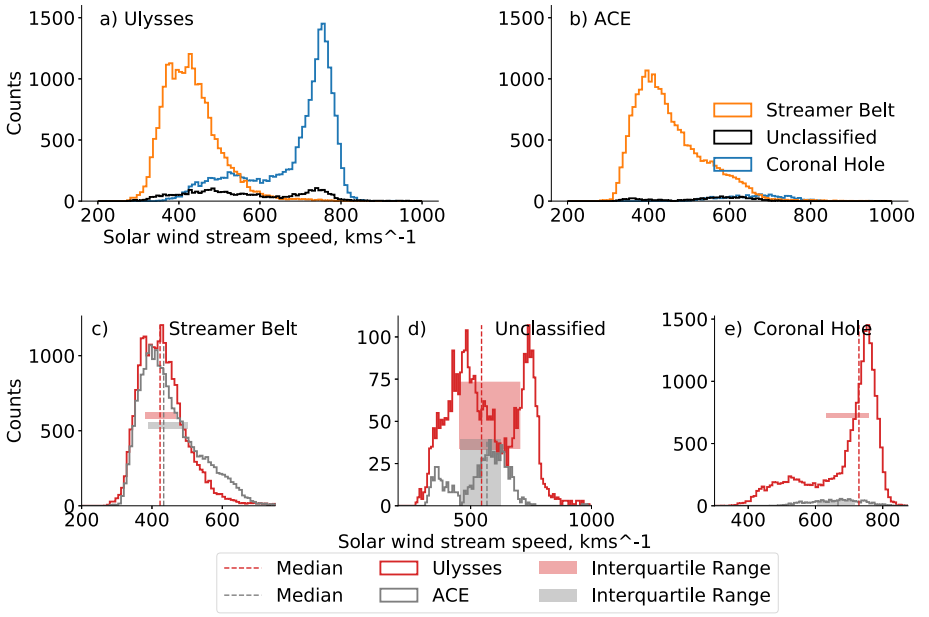

Figure 12 The distributions of solar wind speeds within each UMAP classification. Panels a and b present the comparisons for the distributions in the *Ulysses* and *ACE* datasets, respectively. Panels c, d and e present the comparisons of like-clusters between datasets. The latter panels also include the mean and inter-quartile ranges of each distribution.

the driving factor. However, in situations where the solar source is important, simply splitting the solar wind up using arbitrary speeds may be misleading. The disparity of ≈10% to ≈11% between the intuitive scheme and the speed-threshold scheme highlights the potential flaws in statistical analyses performed using solar wind data. The two intuitive scheme classification boundaries in Figures 1a and 1b are quite different. One has twice the gradient of the other in linear space, yet they produce similar results in terms of coronal-hole and streamer-belt winds, suggesting a degree of robustness in the classifications. However, this approach is entirely deterministic and there is no means to assess uncertain or difficult-to-classify solar wind intervals.

The BGM scheme mathematically extends the intuitive scheme using the same parameters, O^{7+}/O^{6+} and S_p. Instead of using visual inspection, classification boundaries are derived by optimising the fit of a Gaussian mixture to the *Ulysses* fast latitude-scan data. This method also allows for the inclusion of a third category: unclassified data. The stability of the classification is assessed through repeated trials on sub-samples of the data, and found to be robust. Applying the classification scheme to the whole of the ACE and *Ulysses* datasets shows, again, significant disparities with the speed-threshold method: ≈18% and ≈22%, respectively. As expected, much less CHW is found in the ACE dataset than the full *Ulysses* dataset.

The unclassified *Ulysses* data was found to be skewed towards the aphelion of the orbit. This could be indicative of the increased time that turbulence or solar wind stream interactions have to develop before reaching *Ulysses*. It may be that the assumption of no plasma mixing breaks down on these long timescales, either as a result of differential streaming of ions (Marsch, 2006; Schwadron *et al.*, 2005) or magnetic reconnection (Gosling, 2012).

The similarity of the speed distributions, Figure 6, in the *Ulysses* and ACE SBW suggests repeatable classification despite the different occurrence density in the two datasets. The slightly higher mean speed for SBW at *Ulysses* compared to ACE is consistent with the increased radial distance and hence continued acceleration and/or interaction time with faster CHW. The CHW distributions, however, show little similarity. This shows there are quantitative differences in the speed of CHW streams in and out of the ecliptic plane. A low-speed CHW population is found primarily at the aphelion of the orbit and perihelion at low latitudes (the aphelion data is also generally closer to the ecliptic plane). This could be a further result of the factors causing the unclassified data to be skewed towards the aphelion of the orbit (*e.g.* turbulence, stream interactions, differential streaming and magnetic reconnection).

The UMAP scheme builds on the principles of the previous schemes: choosing non-evolving parameters for classification, and using the fast latitude-scan data to establish the classifications. Unlike the BGM scheme, there is no user-specified number of categories to discover, nor is the distribution of data assumed to approximate a multivariate Gaussian. The UMAP algorithm takes six non-evolving parameters (O^{7+}/O^{6+}, C^{6+}/C^{5+}, Fe/O, $< q_{Fe} >$, He^{2+}/H^{1+}, and S_p) and approximates the latent structure of the space into a 2D representation. The reduced data presented in Figure 8a shows that there are two primary groupings in the data, as well as another small grouping. The clusters in the data are extracted using HDBSCAN, a density-based clustering algorithm. By mapping the clusters from the arbitrary 2D space into O^{7+}/O^{6+} and S_p, Figure 8b, it is clear that the groupings in the data match well with the expected properties of CHW and SBW. Such information is determined using accepted domain-specific knowledge about the solar wind (*e.g.* CHW is generally cooler and has higher S_p). Furthermore, the majority of the unclassified data is found in the boundary region between the CHW and SBW, supporting the idea that it is difficult to definitively classify, especially in a lower-dimensional representation.

When applying the UMAP classification scheme to the whole *Ulysses* data, Figure 9a, the two primary clusters become saturated with data points. There is a proportional increase in the amount of both CHW and unclassified data. As with the BGM scheme, there is a large disparity in the comparison with taking speed thresholds: ≈20%. The application of the UMAP scheme to the ACE dataset, Figure 10a, shows a lack of CHW. Interestingly, in Figure 10c we see much better agreement between the speed-threshold classification and the results of the UMAP scheme on the ACE data, ≈8%. However, the results are not much better than just classifying everything as SBW, as one may expect with such skewed data. Despite the link between the CHW and unclassified data, the latter shows very little dependence on the orbital position in Figure 11b.

Comparing the speed distributions of each class, we see qualitative similarities with the results of the BGM scheme: the SBW speed distributions match well, though the ACE distribution displays a heavy tail, and the CHW is bi-modal. Different however, are the unclassified data distributions, which are also bi-modal. This highlights the different ways in which the unclassified data is characterised in the two schemes. The UMAP results show double-peaks in speed close to the peaks of the CHW and SBW. This suggests that the unclassified data may be data which belongs to one or other of the distributions, whose parameters deviate from their respective norm.

In Figure 12d, the peaks of the unclassified ACE data are shifted towards lower speeds. This is contrary to the expectation of the slower stream to be sped up and the faster stream slowed down, due to stream interactions. As such, the data may signify that the unclassified data comprises solar wind transients (all slowed due to the increase in SBW in the ecliptic plane). Else, there may be a process acting to generally slow the unclassified data found in the ecliptic plane.

Table 2 The proportions of each solar wind type found when classifying *Ulysses* data. S15 refers to the results of Stakhiv *et al.* (2015). BGM and UMAP refer to the results of the presented classification schemes.

SW Type	S15	BGM	UMAP
CHW		≈38.8%	≈40.3%
SBW		≈44.7%	≈48.3%
Unclassified		≈16.5%	≈11.3%
Fast	≈20%		
Slow	≈65%		
Intermediate	≈15%		

Table 3 The proportions of each solar wind type found when classifying ACE or OMNI data. The column labels refer to Zhao, Zurbuchen, and Fisk (2009), Zhao *et al.* (2017), Xu and Borovsky (2015), and Camporeale, Carè, and Borovsky (2017), respectively. BGM and UMAP refer to the results of the presented classification schemes. AR refers to active-regions and SR refers to sector-reversal regions.

SW Type	Z09	Z17	X15	C17	BGM	UMAP
CHW	≈58%	≈8.2%	≈25.2%	≈20.4%	≈8.4%	≈4.8%
SBW			≈41.7%	≈27.6%	≈78.9%	≈90.9%
Unclassified					≈12.7%	≈4.4%
Non-CHW	≈37%					
ICME	≈5%					
CH-boundary		≈10.2%				
Quiet Sun		≈25.6%				
AR		≈31.1%				
AR boundary		≈10.8%				
Helmet streamer		≈13%				
Ejecta			≈12.9%	≈13.9%		
SR region			≈20.2%	≈38.0%		

The heavy-tail distribution of the ACE SBW (classified by UMAP) may suggest that a more complex model is required to characterise all of the structure in the data. Alternatively, it may highlight the presence of another process which accelerates SBW in the ecliptic plane, such as interactions with faster CHW or the inclusion of more solar wind transients (*e.g.* ICMEs). Investigating the slower CHW peak in the *Ulysses* data produces the same results as for the BGM scheme. The data within the peak is largely from the aphelion of the orbit.

Tables 2 and 3 show the distributions of the classification results from some of the papers discussed in the introduction. This, again, draws attention to the lack of consensus on how the solar wind should be classified. Note that these comparisons are relate only to proportions, since the results are not all obtained from the same data. Direct comparison of classifications for the same data are given in Table 4.

Of the classification schemes mentioned, only Stakhiv *et al.* (2015) (S15) have results for classifying *Ulysses* data. The results in Table 2 have been estimated from their Figure 5. These results show less fast and more slow wind than we find of our comparable CHW and SBW, respectively. However, these differences are reduced if we account for the errors we predict for taking such speed thresholds.

Both of Zhao, Zurbuchen, and Fisk (2009) and Zhao *et al.* (2017) (Z09 and Z17) have results from classifying ACE data, though the latter paper is more difficult to compare given its six-type classification scheme. The results in Table 3 are estimated from Figure 1 of

Table 4 Confusion matrices (contingency tables) inter-comparing the classification results of various schemes on ACE or OMNI data. The column labels refer to Zhao, Zurbuchen, and Fisk (2009) and Camporeale, Carè, and Borovsky (2017). BGM and UMAP refer to the presented classification schemes. C17's results are obtained by matching timestamps between the data provided in their paper with those in either the BGM or UMAP scheme. Z09's results are obtained by applying the classification criteria from their paper to the data used in the BGM or UMAP scheme. The green highlighting serves to draw attention to the diagonals, which represent the number of samples of agreed classifications between any two schemes.

		UMAP				C17				Z09		
BGM		CHW	SBW	Ej/Unc		CHW	SBW	Ej/Unc		CHW	SBW	Ej/Unc
	CHW	587	892	392	CHW	2936	181	71	CHW	3194	0	0
	SBW	71	17668	302	SBW	3585	22677	3496	SBW	15425	13620	1017
	Ej/Unc	527	3023	152	Ej/Unc	1856	1522	1421	Ej/Unc	626	514	3712
						CHW	SBW	Ej/Unc		CHW	SBW	Ej/Unc
UMAP					CHW	1129	18	38	CHW	1092	17	15
					SBW	4549	13774	3286	SBW	11725	8106	1678
					Ej/Unc	560	162	124	Ej/Unc	772	220	40
										CHW	SBW	Ej/Unc
					C17				CHW	6295	390	1692
									SBW	11450	11507	1423
									Ej/Unc	1396	2019	1573

Z09 and Figure 6 in Z17. The Z09 results do not match well with the CHW or SBW results of either the BGM or UMAP scheme. However, the ICME value is not dissimilar to UMAP's unclassified data, possibly supporting the idea that the unclassified data (especially in the UMAP scheme) could be composed of ICMEs and other transients. Z17's pure CHW shows good agreement with our results (especially from the BGM classification). If the CH-boundary class is taken as a part of the CHW, then the agreement diminishes. The rest of the classifications are not usefully comparable due to the differences to our scheme.

Both Xu and Borovsky (2015) (X15) and Camporeale, Carè, and Borovsky (2017) (C17) apply their classifications to *OMNI* data (King and Papitashvili, 2005). This incorporates ACE data as well as other data from L1, allowing comparison. The results in Table 3 are taken from Table 3 in X15, and the results of the C17 classification (see their acknowledgements for data location). The X15 results differ from those found with our schemes. However, if we consider that the sector-reversal region solar wind is a part of our SBW, then there is some agreement between these results and those from the BGM classification (the UMAP classification still differs significantly). The C17 results differ slightly from those of X15, and present more agreement with our results.

To compare some of these results in a more rigorous way, Table 4 presents confusion matrices (contingency tables) comparing the results of two of the schemes on the same data. The Z09 and C17 results have been simplified by assuming the non-CHW is equivalent to our SBW, and combining the SBW and sector-reversal region wind, respectively.

Most noteworthy of these results is the agreement between what the BGM and UMAP schemes classify as CHW compared with the Z09 and C17 schemes. This is exemplified by the horizontal rows of CHW from the BGM and UMAP. In these rows the proportion of what our schemes classify as CHW and the other schemes classify otherwise (reading along horizontally) is very low. However, reading vertically along the CHW columns, we see there are many samples in other columns. This suggests that our schemes (trained on the *Ulysses* data) provide accurate-but-conservative classifications of CHW as compared to the other models.

Comparison between each of our (BGM and UMAP) SBW classifications to Z09's and C17's schemes present less consistent results. The BGM:C17 and UMAP:C17 results are broadly in agreement, with the majority of our SBW also being classed as SBW by the other schemes. In contrast, the Z09 scheme classifies the majority of solar wind as CHW. Hence,

the results of our classifications are in conflict (as are the Z09 in conflict with all of the others in Table 3).

Given the disparate methods of determining the ejecta/unclassified wind, it is unsurprising that there is little agreement between any of the schemes (save some between the BGM and Z09 schemes). Using a broadened feature space for the UMAP scheme and identifying an unclassified cluster not found in the BGM scheme highlights the importance of applying domain-specific knowledge, even in data-driven approaches.

The inter-comparison between the BGM and UMAP schemes quantifies the evident differences and similarities between the two methods. As one may expect from the comparisons of speed distributions, the SBW is in good agreement. However, the CHW is more diverse. Given the larger feature space, and less constraining method of clustering, we would posit that the UMAP CHW is a more accurate representation of the class. It is more difficult to comment on the accuracy of the unclassified wind from UMAP given that there are contributions from other areas of the feature space.

We acknowledge that there may be some systematic bias in the classifications of ACE data. It is possible that by limiting the training set to the *Ulysses* latitude scans, we created classification boundaries which generalise less well to the ACE data (despite our use of non-evolving parameters). One potential source of this may be that our training data heavily samples very large polar-coronal holes. As such, in the ecliptic plane where we see generally smaller coronal holes, and are more likely to sample boundary regions (see the percentages of CHW and CH-boundary wind in the Z17 column of Table 3), the algorithms may classify such winds as SBW mistakenly.

Whilst our choice of training data may bias the classification, the benefits of training on out-of-ecliptic data which samples almost the entire range of heliospheric latitudes are significant: a more complete range of solar wind is sampled, and that wind is less likely to be interfered with by processes relating to stream interaction. Furthermore, it should be noted that discovering results which differ from the norm when using novel techniques does not necessarily mean the results are wrong. It could very well be the case that there is less CHW in the ecliptic plane than current classifications recognise.

8. Conclusions

This work presents two novel, data-driven schemes to classify the solar origin of solar wind streams using unsupervised machine learning. The schemes are built using non-evolving parameters which retain information about the source regions. Each classification model is created using the *Ulysses* fast latitude-scan data, before being applied to the whole *Ulysses* and ACE datasets. The BGM scheme reduces the subjectivity in determining classification boundaries between solar wind types. It was specified to fit three clusters in the solar wind data. As expected, two of these are the coronal-hole and streamer-belt winds. The third remains unclassified. The UMAP scheme addresses subjectivity in the choices of decision boundaries and the number of clusters to find in the data; it independently derives three clusters in the latent topological structure of the solar wind data. These clusters correspond to coronal-hole and streamer-belt winds as before, but find a different type of unclassified solar wind. Application of the UMAP scheme to *Ulysses* and ACE shows morphological differences in the coronal-hole wind seen in and out of the ecliptic plane.

For both schemes, and both spacecraft datasets, the classification results are compared with the traditional approach of taking speed thresholds. In each case, there are significant best case disparities between the speed-threshold approach relative to the machine learning

classifications: The BGM scheme applied to *Ulysses*, ≈22% and ACE, ≈18%; the UMAP scheme applied to *Ulysses*, ≈20% and ACE, ≈8%.

Whilst our results differ from those of other works, our data-driven methods are designed to increase objectivity and reduce the introduction of scientifically subjective biases. Thus, the differences do not take away from the results presented. Instead, such differences should motivate further work investigating objective methods of solar wind classification, and their differences to current schemes.

Acknowledgements We acknowledge the NASA National Space Science Data Center and the Space Physics Data Facility for usage of *Ulysses* data. *Ulysses* data is publicly available at: ftp://spdf.gsfc.nasa.gov/pub/data/ulysses/. See https://wind.nasa.gov/data_sources.php for more information on different data products. We thank the ACE-SWEPAM, ACE-SWICS, ACE-MAG instrument teams and the ACE Science Center for providing the ACE data. Téo Bloch is funded by Science and Technology Facilities Council (STFC) training grant number ST/R505031/1. Clare Watt is part funded by STFC grant number ST/R000921/1 and Natural Environment Research Council (NERC) grant number NE/P017274/1. Mathew Owens is part-funded by STFC grant number ST/R000921/1 and NERC grant number NE/P016928/1. Allan Macneil is funded by STFC grant number ST/R000921/1.

Disclosure of Potential Conflicts of Interest The authors declare that they have no conflicts of interest.

Appendix A: Bayesian Statistics and Variational Inference

Bayes' theorem is the statistical description of the probability that an event happens, given some prior knowledge of the conditions of the event. Bayes' theorem is notated, for two events A and B, as

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)},$$

where $P(A \mid B)$ (the posterior probability) is the conditional probability that event A occurs given event B, $P(B \mid A)$ (the likelihood) is the conditional probability that event B occurs given event A, and $P(A)$ (the prior probability) and $P(B)$ (marginal likelihood) are the probabilities of events A and B happening independently. In Bayesian inference, the interpretation of the posterior probability is the degree of belief in a hypothesis. This can be envisioned as a situation where you have a number of Gaussians from which a point measurement may be sampled (*e.g.* a solar wind measurement and Gaussian mixture for its classification). To determine which Gaussian is most likely given the data point, you calculate the posterior using Bayes' theorem, taking each Gaussian, A, given the data point, B, and compare the probabilities for each Gaussian.

Variational inference (used in the BGM) is an extension to Bayesian inference developed for ML and is a current area of research in statistics. A brief description will be given here based on the work of Blei, Kucukelbir, and McAuliffe (2017) and Gelman *et al.* (2013). Variational inference is a method used to approximate probability densities through optimisation, rather than sampling techniques (*e.g.* Markov Chain Monte Carlo, MCMC). MCMC is used to create an empirical estimate of the posterior distribution based on collected samples, and is very effective on smaller or more simple models. However, when models are complex or datasets are large a different approach is needed for computational practicality. Variational inference chooses a family of probability density functions (PDFs) as an approximation to the true PDF. The member of the PDF family which minimises the Kullback–Leibler (KL) divergence to the exact posterior is sought (explained further, below). The member which minimises the KL divergence is then optimised and used as the approximate distribution for the posterior distribution. Variational inference is usually faster than MCMC methods and better suited to scaling for large datasets. The drawback is that while MCMC is known to converge asymptotically to the correct solution, variational inference is not. Despite this, Figure 3 and Table 1 in Blei and Jordan (2006), show how variational inference can be much faster, while also remaining competitive to MCMC methods.

The Kullback–Leibler divergence is a measure of how one probability distribution diverges from another (for the derivation and further information, see Kullback, 1978). For the probability distribution of a continuous random variable x the Kullback–Leibler divergence, $D_{KL}(P||Q)$ of distribution $Q(x)$ from a given distribution $P(x)$ is defined as:

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx,$$

where $p(x)$ and $q(x)$ are the probability densities of $P(x)$ and $Q(x)$, respectively.

Appendix B: Uniform Manifold Approximation and Projection

Expanding on the description given in Section 6, the following will discuss the methodology of the UMAP dimension reduction technique using the appropriate mathematical terminology. This description is quite involved, but should provide an interested reader with all of the vocabulary needed to further investigate the algorithm. Naturally, the full mathematical description is given in McInnes, Healy, and Melville (2018).

The field of topological data analysis (Carlsson, 2009) uses methods from topology to better understand complex datasets. One such technique is the construction of the Čech complex (Ghrist, 2014) which provides a combinatorial representation of a topological space inferred from a given dataset. To construct the Čech complex one forms a cover given by open balls of a fixed radius about each of the datapoints. The Čech complex is then the simplicial complex (Ghrist, 2014) given by the nerve of that open cover (see, *e.g.* Ghrist, 2014, for more detail)). Informally the process proceeds essentially as follows: to each open ball one assigns a point; whenever a pair of open balls have non-empty intersection one joins the corresponding points with line segment; whenever three open balls share a non-empty intersection one adds a filled triangle joining the points; and so on, adding higher dimension pieces for more complex intersections. By the nerve theorem (Borsuk, 1948), the resulting simplicial complex is homotopy equivalent (Ghrist, 2014) to the manifold formed by the union of the open cover. Informally, the topological space pieced together by points, lines, triangles, tetrahedrons, *etc.*, captures the same fundamental topological structure as

the space being covered by open balls. In this manner manifold structure latent in data can be discovered.

Unfortunately this will only successfully capture the underlying manifold from which the data was drawn when the data samples are uniformly distributed on the manifold. Since this is rarely the case for data under the ambient metric we must instead use the existing data distribution to infer the Riemannian metric on the manifold that would result in such a uniform distribution. This can be done by examining local data distributions and approximating a locally constant Riemannian metric at each point. While this recovers the uniform distribution assumption it introduces a new difficulty in that the metric spaces local to each point are mutually incompatible.

By translating the local metric spaces into fuzzy simplicial sets (see Spivak, 2012; Goerss and Jardine, 2009) the incompatibility can be overcome by taking the union of the entire family of fuzzy simplicial sets. The result is a single fuzzy simplicial set that provides a coherent view of the topological structure of the underlying manifold from which the data was sampled. UMAP then uses an optimisation process to find a low-dimensional representation of the data that has a fuzzy simplicial set representation that matches the topological representation of the source dataset as closely as possible.

Appendix C: Hierarchical Density Based Spatial Clustering for Applications with Noise

As with Appendix B, we present an extension to the description of HDBSCAN given in Section 6. Again, it is quite involved, but should be of interest to those familiar with ML or who wish to learn more.

A dataset of measurements can be assumed to have been (noisily) sampled from some probability density function. 'Noisily sampled' in this case refers to sampling a value when there is noise (*e.g.* the inherent uncertainty in spacecraft measurements). Given a probability density function f, where $f(x)$ is the likelihood of sampling a point x and $\int_{\mathbb{R}^n} f(x)dx = 1$, one can consider the level sets $\{x \in \mathbb{R}^n \mid f(x) \geq \lambda\}$. As $\lambda \geq 0$ varies these level sets will nest in such a way as to form an infinite tree, called the *cluster tree*. Each cluster is a branch of the tree, extending over the range of λ values for which it is distinct. The goal of hierarchical density-based clustering algorithms is to approximate this cluster tree given only a finite set of sampled data.

Classical hierarchical clustering techniques such as single linkage clustering (Everitt *et al.*, 2011) provide a partial solution. Results by Hartigan (1981) demonstrate consistency with the cluster tree for single linkage clustering in the case of 1D data. In higher dimensions, however, single linkage clustering becomes too sensitive to noise: it suffers from chaining effects where spurious points result in clusters merging prematurely. To remedy this we need to introduce a notion of density. Let $X = \{x_1, x_2, \ldots, x_N\} \subset \mathbb{R}^n$ be the dataset, and define the *core-distance* $\kappa(x_i)$ of a point x_i as the distance to the kth nearest neighbour of x_i. The core-distance can act as a proxy for density (since sparse areas of the sample space will have larger core-distances). We can then define a new metric, called mutual-reachability-distance, defined as

$$d(x_i, x_j) = \begin{cases} \max\{\kappa(x_i), \kappa(x_j), \|x_i - x_j\|_2\} & x_i \neq x_j, \\ 0 & x_i = x_j. \end{cases}$$

In effect the mutual-reachability distance between a pair of points is the smallest distance scale at which both points will be dense and considered to be neighbouring each other. Performing single linkage clustering under this new density-sensitive metric yields a more robust clustering algorithm that can be shown to converge to the cluster tree of the probability density function from which the data was drawn Eldridge, Belkin, and Wang (2015).

The resulting cluster hierarchy is often exceptionally complex. Much of the complexity is the result of single, or small numbers of, points separating off into new clusters. To simplify the resulting cluster hierarchy we can consider a minimum allowable cluster size m. We can then re-process the hierarchy considering any child cluster with fewer than m points to be spurious—we denote those points as "falling out of the parent cluster". The resulting simplified tree allows for better cluster analysis. A further step can then be taken by selecting those clusters within the tree that persist for the largest ranges of distance scales. This can be posed as a simple optimisation problem using the notion of relative-excess-of-mass from probability theory. This allows for the production of a flat clustering where each data point is either assigned a cluster label or, if it fell out of a cluster further up the hierarchy, is labelled as noise.

References

Altschuler, M.D., Newkirk, G.: 1969, Magnetic fields and the structure of the solar corona. *Solar Phys.* **9**(1), 131. DOI.

Antiochos, S.K., Mikić, Z., Titov, V.S., Lionello, R., Linker, J.A.: 2011, A model for the sources of the slow solar wind. *Astrophys. J.* **731**(2), 2 DOI.

Attias, H.: 2000, A variational Bayesian framework for graphical models. *Adv. Neural Inf.* **12**, 209. ISBN 0262194503.

Balogh, A., Beek, T.J., Forsyth, R.J., Hedgecock, P.C., Marquedant, R.J., Smith, E.J., Southwood, D.J., Tsurutani, B.T.: 1992, The magnetic field investigation on the Ulysses mission: instrumentation and preliminary scientific results. *Astron. Astrophys. Suppl. Ser.* **92**, 221.

Bame, S.J., McComas, D.J., Barraclough, B.L., Phillips, J.L., Sofaly, K.J., Chavez, J.C., Goldstein, B.E., Sakurai, R.K.: 1992, The ULYSSES solar wind plasma experiment. *Astron. Astrophys. Suppl.* **92**, 237.

Bishop, C.M.: 2006, *Pattern Recognition and Machine Learning*, 1st edn. Springer, Cambridge. ISBN 978-0-387-31073-2.

Blei, D.M., Jordan, M.I.: 2006, Variational inference for Dirichlet process mixtures. *Bayesian Anal.* **1**(1), 121. DOI. http://projecteuclid.org/euclid.ba/1340371077.

Blei, D.M., Kucukelbir, A., McAuliffe, J.D.: 2017, Variational inference: a review for statisticians. *J. Am. Stat. Assoc.* **112**(518), 859. DOI.

Borovsky, J.E., Denton, M.H.: 2016, The trailing edges of high-speed streams at 1 AU. *J. Geophys. Res.* **121**(7), 6107. DOI.

Borsuk, K.: 1948, On the imbedding of systems of compacta in simplicial complexes. *Fundam. Math.* **35**, 217. DOI.

Brooks, D.H., Ugarte-Urra, I., Warren, H.P.: 2015, Full-Sun observations for identifying the source of the slow solar wind. *Nat. Commun.* **6**(1), 5947. DOI.

Burlaga, L.F., Mish, W.H., Whang, Y.C.: 1990, Coalescence of recurrent streams of different sizes and amplitudes. *J. Geophys. Res.* **95**(A4), 4247. DOI.

Campello, R.J.G.B., Moulavi, D., Sander, J.: 2013, Density-based clustering based on hierarchical density estimates. *Lec. Notes Computer Sci.* **7819**, 160. DOI.

Camporeale, E., Carè, A., Borovsky, J.E.: 2017, Classification of solar wind with machine learning. *J. Geophys. Res.* **122**(11), 10,910. DOI.

Camporeale, E., Wing, S., Johnson, J.R.: 2018, *Machine Learning Techniques for Space Weather*, 1st edn. Elsevier, Amsterdam. DOI. ISBN 9780128117880.

Cane, H.V., Richardson, I.G.: 2003, Interplanetary coronal mass ejections in the near-Earth solar wind during 1996–2002. *J. Geophys. Res.* **108**, 1156. DOI.

Carlsson, G.: 2009, Topology and data. *Bull. Am. Math. Soc.* **46**(2), 255. DOI. http://www.ams.org/journal-getitem?pii=S0273-0979-09-01249-X.

Crooker, N.U., Gosling, J.T., Bothmer, V., Forsyth, R.J., Gazis, P.R., Hewish, A., Horbury, T.S., Intriligator, D.S., Jokipii, J.R., Kóta, J., Lazarus, A.J., Lee, M.A., Lucek, E., Marsch, E., Posner, A., Richardson, I.G., Roelof, E.C., Schmidt, J.M., Siscoe, G.L., Tsurutani, B.T., Wimmer Schweingruber, R.F., Wimmer-Schweingruber, R.F.: 1999, CIR morphology, turbulence, discontinuities, and energetic particles. *Space Sci. Rev.* **89**(1/2), 179. DOI. ISBN 0038-6308.

Eldridge, J., Belkin, M., Wang, Y.: 2015, Beyond Hartigan consistency: Merge distortion metric for hierarchical clustering. *Proc. Machine Learning Res.* **40**, 588. http://proceedings.mlr.press/v40/Eldridge15.html.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X.: 1996, A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Proc. KDD 96*, 226. ISBN 1-57735-004-9. https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf.

Everitt, B.S., Landau, S., Leese, M., Stahl, D.: 2011, *Cluster Analysis*, 5th edn., Wiley, Chichester. DOI. ISBN 9780470977811.

Florek, K., Łukaszewicz, J., Perkal, J., Steinhaus, H., Zubrzycki, S.: 1951, Sur la liaison et la division des points d'un ensemble fini. *Colloq. Math.* **2**(3-4), 282. DOI.

Fukunaga, K., Hostetler, L.: 1975, The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans. Inf. Theory* **21**(1), 32. DOI. http://ieeexplore.ieee.org/document/1055330/.

Geiss, J., Gloeckler, G., Von Steiger, R.: 1995, Origin of the solar wind from composition data. *Space Sci. Rev.* **72**(1-2), 49. DOI.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B.: 2013, *Bayesian Data Analysis*, 3rd edn. CRC Press/Taylor & Francis, Boca Raton/London. ISBN 978-1-4398-9820-8.

Ghrist, R.: 2014, *Elementary Applied Topology*, 1st edn. Createspace, Philadelphia. ISBN 978-1502880857.

Gloeckler, G., Cain, J., Ipavich, F.M., Tums, E.O., Bedini, P., Fisk, L.A., Zurbuchen, T.H., Bochsler, P., Fischer, J., Wimmer-Schweingruber, R.F., Geiss, J., Kallenbach, R.: 1998, Investigation of the composition of solar and interstellar matter using solar wind and pickup ion measurements with SWICS and SWIMS on the ACE spacecraft. *Space Sci. Rev.* **86**, 497. DOI.

Goerss, P.G., Jardine, J.F.: 2009, *Simplicial Homotopy Theory* **53**, Birkhäuser, Basel, 1689. DOI. ISBN 978-3-0346-0188-7.

Gosling, J.T.: 2012, Magnetic reconnection in the solar wind. *Space Sci. Rev.* **172**(1-4), 187. DOI.

Hartigan, J.A.: 1981, Consistency of single linkage for high-density clusters. *J. Am. Stat. Assoc.* **76**(374), 388. DOI. http://www.tandfonline.com/doi/abs/10.1080/01621459.1981.10477658.

Heidrich-Meisner, V., Wimmer-Schweingruber, R.F.: 2018, Solar wind classification via k-means clustering algorithm. In: *Machine Learning Techniques for Space Weather*, 1st edn., Elsevier, Amsterdam, 397. DOI. ISBN 9780128117880. https://linkinghub.elsevier.com/retrieve/pii/B9780128117880000160.

Igel, C., Heidrich-Meisner, V., Glasmachers, T.: 2008, Shark. *J. Mach. Learn. Res.* **9**, 993. http://www.jmlr.org/papers/volume9/igel08a/igel08a.pdf.

King, J.H., Papitashvili, N.E.: 2005, Solar wind spatial scales in and comparisons of hourly wind and ACE plasma and magnetic field data. *J. Geophys. Res.* **110**(A2), 1. DOI.

Ko, Y., Raymond, J.C., Zurbuchen, T.H., Riley, P., Raines, J.M., Strachan, L.: 2006, Abundance variation at the vicinity of an active region and the coronal origin of the slow solar wind. *Astrophys. J.* **646**(2), 1275. DOI.

Kullback, S.: 1978, *Information Theory and Statistics*, Dover Pub. Inc., Dover, Gloucester. ISBN 0844656259. http://index-of.co.uk/Information-Theory/Informationtheoryandstatistics-Solomon.pdf.

Kusher, H.J., Yin, G.G.: 2003, *Stochastic Approximation and Recursive Algorithms and Applications*, 2nd edn., *Stochastic Modelling and Applied Probability* **35**, Springer, New York. DOI. ISBN 0-387-00894-2.

MacQueen, J.: 1967, Some methods for classification and analysis of multivariate observations. In: *Proc. Fifth Berkeley Symp. Mathematical Statistics and Probability* **1**, 281. https://projecteuclid.org/euclid.bsmsp/1200512974.

Marsch, E.: 2006, Kinetic physics of the solar corona and solar wind. *Living Rev. Solar Phys.* **3**, 1 DOI.

McComas, D.J., Bame, S.J., Barker, P., Feldman, W.C., Phillips, J.L., Riley, P., Griffee, J.W.: 1998, Solar wind electron proton alpha monitor (SWEPAM) for the advanced composition explorer. *Space Sci. Rev.* **86**(1-4), 563. DOI.

McComas, D.J., Angold, N., Elliott, H.A., Livadiotis, G., Schwadron, N.A., Skoug, R.M., Smith, C.W.: 2013, Weakest solar wind of the space age and the current "mini" solar maximum. *Astrophys. J.* **779**(1), 2. DOI.

McInnes, L., Healy, J., Melville, J.: 2018, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv.

McLachlan, G., Peel, D.: 2000, *Finite Mixture Models*, *Wiley Series in Probability and Statistics*, Wiley, Hoboken. DOI. ISBN 9780471721185.

Neugebauer, M., Snyder, C.W.: 1966, Mariner 2 observations of the solar wind: 1. Average properties. *J. Geophys. Res.* **71**(19), 4469. DOI.

Owens, M.J., Crooker, N.U., Lockwood, M.: 2014, Solar cycle evolution of dipolar and pseudostreamer belts and their relation to the slow solar wind. *J. Geophys. Res.* **119**(1), 36. DOI.

Owocki, S.P., Holzer, T.E., Hundhausen, A.J.: 1983, The solar wind ionization state as a coronal temperature diagnostic. *Astrophys. J.* **275**, 354. DOI.

Pagel, A.C.: 2004, Correlation of solar wind entropy and oxygen ion charge state ratio. *J. Geophys. Res.* **109**(A1), A01113. DOI.

Panasenco, O., Velli, M.: 2013, Coronal pseudostreamers: source of fast or slow solar wind? *AIP Conference Proc.* **1539**, 50. DOI. ISBN 9780735411630.

Pedregosa, F., Varoquaux, G., Buitinck, L., Louppe, G., Grisel, O., Pedregosa, F., Mueller, A.: 2011, Scikit-learn. *J. Mach. Learn. Res.* **12**, 2825. DOI. http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html.

Richardson, I.G.: 2004, Identification of interplanetary coronal mass ejections at 1 AU using multiple solar wind plasma composition anomalies. *J. Geophys. Res.* **109**(A9), A09104. DOI.

Russell, S.J., Norvig, P.: 2009, *Artificial Intelligence a Modern Approach*, 3rd edn. Prentice Hall/Pearson Education, New York/Upper Saddle River. ISBN 978-0-13-604259-4.

Schatten, K.H., Wilcox, J.M., Ness, N.F.: 1969, A model of interplanetary and coronal magnetic fields. *Solar Phys.* **6**(3), 442. DOI.

Schwadron, N.A., McComas, D.J., Elliott, H.A., Gloeckler, G., Geiss, J., von Steiger, R.: 2005, Solar wind from the coronal hole boundaries. *J. Geophys. Res.* **110**(A4), 1. DOI.

Sheeley, N.R., Harvey, J.W., Feldman, W.C.: 1976, Coronal holes, solar wind streams, and recurrent geomagnetic disturbances: 1973–1976. *Solar Phys.* **49**(2), 271. DOI.

Smith, C.W., Heureux, J.L., Ness, N.F.: 1998, The ACE magnetic fields experiment. *Space Sci. Rev.* **86**(1-4), 613. DOI.

Sokal, R.R., Michener, C.: 1958, A statistical methods for evaluating relationships. *Univ. Kansas Sci. Bull.* **38**, 1409.

Spivak, D.I.: 2012, Metric realization of fuzzy simplicial sets. *Self published notes*. http://math.mit.edu/~dspivak/files/metric_realization.pdf.

Stakhiv, M., Landi, E., Lepri, S.T., Oran, R., Zurbuchen, T.H.: 2015, On the origin of mid-latitude fast wind: challenging the two-state solar wind paradigm. *Astrophys. J.* **801**(2), 100. DOI. http://stacks.iop.org/0004-637X/801/i=2/a=100?key=crossref.1a00bc7943b9d49d76751a8bf7f41587.

von Steiger, R., Schwadron, N.A., Fisk, L.A., Geiss, J., Gloeckler, G., Hefti, S., Wilken, B., Wimmer-Schweingruber, R.R., Zurbuchen, T.H.: 2000, Composition of quasi-stationary solar wind flows from Ulysses/Solar Wind Ion Composition Spectrometer. *J. Geophys. Res.* **105**(A12), 27217. DOI.

Wenzel, K.P., Marsden, R.G., Page, D.E., Smith, E.J.: 1992, The Ulysses mission. *Astron. Astrophys. Suppl.* **92**, 207. ADS.

Xu, F., Borovsky, J.E.: 2015, A new four-plasma categorization scheme for the solar wind. *J. Geophys. Res.* **120**(1), 70. DOI. ISBN 2169-9402.

Zhao, L., Zurbuchen, T.H., Fisk, L.A.: 2009, Global distribution of the solar wind during solar cycle 23: ACE observations. *Geophys. Res. Lett.* **36**(14), L14104. DOI.

Zhao, L., Landi, E., Lepri, S.T., Gilbert, J.A., Zurbuchen, T.H., Fisk, L.A., Raines, J.M.: 2017, On the relation between the in situ properties and the coronal sources of the solar wind. *Astrophys. J.* **846**(2), 135. DOI. http://stacks.iop.org/0004-637X/846/i=2/a=135?key=crossref.852c60d0fb4792c4e7d0b09e9fc9b323.