JGR | Machine Learning and Computation

RESEARCH ARTICLE
10.1029/2024JH000221

Clustering of Global Magnetospheric Observations

James Edmond[1], Joachim Raeder[1], Banafsheh Ferdousi[2], Matthew Argall[1], and Maria Elena Innocenti[3]

[1]University of New Hampshire, Durham, NH, USA, [2]Air Force Research Laboratory, Albuquerque, NM, USA, [3]Ruhr-Universität Bochum, Bochum, Germany

Key Points:

- MMS and THEMIS measurements are classified into magnetosphere, magnetosheath, and solar wind regions using unsupervised methods
- We created a data set of 5228 magnetopause and 3047 bow shock crossings inferred from the classifications
- The model is capable of detecting Hot Flow Anomalies, Foreshock Bubbles, and Bursty Bulk Flows

Correspondence to:

J. Edmond,
james.edmond@unh.edu

Citation:

Edmond, J., Raeder, J., Ferdousi, B., Argall, M., & Innocenti, M. E. (2024). Clustering of global magnetospheric observations. *Journal of Geophysical Research: Machine Learning and Computation*, *1*, e2024JH000221. https://doi.org/10.1029/2024JH000221

Received 16 APR 2024
Accepted 5 SEP 2024

Abstract The use of supervised methods in space science have demonstrated powerful capability in classification tasks, but purely unsupervised methods have been less utilized for the classification of spacecraft observations. We use a combination of unsupervised methods, being principal component analysis, Self-Organizing Maps, and hierarchical agglomerative clustering, to classify THEMIS and MMS observations as having occurred in the magnetosphere, magnetosheath, or the solar wind. The resulting classification are validated visually by analyzing the distribution of classifications and studying individual time series as well as by comparison to the labeled data set of a previous model, against which ours has an accuracy of 99.4%. The model has a variety of applications beyond region classification such as deeper hierarchical analysis, magnetopause and bow shock crossing identification, and identification of bursty bulk flows, hot flow anomalies, and foreshock bubbles.

Plain Language Summary Machine learning in space science often uses supervised methods for classification, but we explore using unsupervised methods for classifying spacecraft observations. We combine principal component analysis, self-organizing maps, and hierarchical clustering to classify whether observations occurred in the magnetosphere, magnetosheath, or solar wind for THEMIS and MMS. We verify classifications both visually and using a preexisting labeled data set, achieving 99.4% accuracy. Our model has additional applications such as the ability to analyze subgroups of clusters, identify boundary regions between the clusters, and flag important transient events related to the dynamics of the magnetosphere.

1. Introduction

The region of space where Earth is directly affected by solar activity can be divided into various regions, such as the solar wind, the magnetosheath, and the magnetosphere itself. Since the first measurements of the solar wind were made incidentally by Gringauz et al. (1962) and intentionally by Neugebauer and Snyder (1962), many missions have recorded measurements in these different regions to investigate various space plasma processes. The solar wind is a continual stream of plasma ejecta originating from the Sun that is accelerated in the solar corona, although the exact mechanisms through which it does this have not yet been confirmed (Cranmer & Winebarger, 2019). This incident plasma is slowed to sub-magnetosonic speeds in order to be diverted about the Earth's magnetosphere, resulting in a denser region of heated and slowed plasma called the magnetosheath (Lucek et al., 2005). The bow shock separates these two regions and causes these changes in plasma flow. It has been generally modeled as a hyperbola in the x-y plane in GSE coordinates where its flaring and position vary in response to solar wind parameters (Fairfield et al., 2001). The magnetosphere itself can be further compartmentalized into a number of sub-regions containing plasma with different properties, including the ring current, the radiation belts, the ionosphere, the tail, etc. The boundary layer separating this region from the magnetosheath is the magnetopause, which results from the pressure equilibrium between the magnetic pressure of the earth and the dynamic pressure of the shocked solar wind (Willis, 1971).

In analyzing the measurements of spacecraft that frequent these regions, it is generally not difficult to identify the location of these regions given a brief time history. The same can often be said for doing so with a joint set of measurements at a single moment in time. Drafting general mathematical relationships that can always correctly classify the regions is more challenging. Classification in the context of machine learning is ideal for this task as it involves determining what class a data point belongs to. This requires that one curates a data set and provides the class label for each measurement. There have already been many successful efforts in using this approach for different missions, such as those by Breuillard et al. (2020), M. R. Argall et al. (2020), and Olshevsky et al. (2021) using deep learning, or those of Nguyen et al. (2022), Smith et al. (2020), and Camporeale et al. (2017) for more

traditional nonlinear approaches. Clustering is an unsupervised method in which data are amalgamated into homogeneous groups and more recently, some have used related methods to classify data, like Amaya et al. (2020) with solar wind classifications from ACE data, Innocenti et al. (2021) with identifying different regions in magnetospheric simulation results, and Köhne et al. (2023) for classifying PIC simulations involving the tearing instability.

Our methodology is based on an unsupervised approach to separate the solar wind, magnetosheath, and magnetosphere measurements from spacecraft data. Our data are recorded at different time resolutions, so methods reliant upon a consistent time step cannot be utilized and must focus on the joint set of measurements alone. We use principal component analysis to reduce the dimensionality and correlations in our data set, along with a visualization technique to add greater interpretability to the dimensionality reduction. Self-Organizing Maps (SOMs) (Kohonen, 1982) are then used to effectively reduce the size of the training set so that a larger number of clustering algorithms can be considered. We finally use hierarchical agglomerative clustering to cluster the individual nodes of the SOM and propagate the cluster assignments of the nodes to the data they represent. The use of hierarchical clustering coupled with a SOM provides a unique advantage in that in addition to being able to separate the data into clusters, these clusters are composed of subclusters which can be further investigated. This combination distinguishes it from other more common clustering methods. The paper is outlined as follows: data sources, data preprocessing, dimensionality reduction, SOMs, clustering of SOM nodes, results, and derived boundary crossings.

2. Data Sources

We use data from two missions, Time History of Events and Macroscale Interactions during Substorms (Angelopoulos, 2008), or THEMIS, and the Magnetospheric Multiscale Mission (Burch et al., 2016), or MMS. These data sets include measurements of magnetic field **B**, the ion velocity **V**, the ion scalar temperature T, and the ion density n, a cumulative eight features. The vector data is in GSE coordinates. Below, we describe for each mission how the data is prepared.

2.1. THEMIS

THEMIS is a collection of five spacecraft (THEMIS-A, B, C, D, and E) with equatorial orbits with the purpose of observing different aspects of magnetic storms and substorms We used data from March 2007 to the end of December 2020. THEMIS-B and C were moved to lunar orbit in 2009 to become the Acceleration, Reconnection, Turbulence, and Electrodynamics of the Moon's Interaction with the Sun (Angelopoulos, 2014) (ARTEMIS) mission where they would make measurements departing from what would normally be seen by THEMIS-A D, and E. We only use THEMIS-B and C data up until end of year 2009.

The ion velocity, temperature, and density measurements of THEMIS are from the Electrostatic Analyzer instrument (McFadden et al., 2008) and are available at multiple time resolutions, such as "reduced" (ESAR) and "full" (ESAF) data packets. The ESAR offers higher time resolution at once per spin (∼3 s), but the cold temperatures of typical solar wind mean that their distributions are narrow and require sufficiently high angular resolution to resolve. The ESAF packets sacrifice time resolution for higher angular resolution and are available in two formats, 32-spin (96 s) in fast survey mode and 128-spin (∼6.5 min) in slow survey mode. Figure 5 of McFadden et al. (2008) illustrates the difference in angular resolution. The data are flagged for quality and we use quality zero data, indicating no issues. The magnetic field measurements are from the Flux Gate Magnetometer (FGM) (Auster et al., 2008) and are collected at spin resolution. This data is then averaged down to the resolution of the ESAF measurements to synchronize them.

2.2. MMS

MMS, the Magnetospheric Multiscale Mission (Burch et al., 2016), is a constellation of four spacecraft (MMS-1, 2, 3, and 4) flying in low-to mid-inclination orbits in tight formation to make electron-scale measurements. The ion measurements are taken from the Dual Ion Spectrometer as part of the Fast Plasma Investigation (Pollock et al., 2016) suite. Multiple ion spectrometers per spacecraft makes it possible to make measurements below spin resolution. The magnetic field measurements are taken from the FGM (Russell et al., 2016) and are available at 10 ms. These magnetic field measurements and ion measurements are averaged down together to 1 min

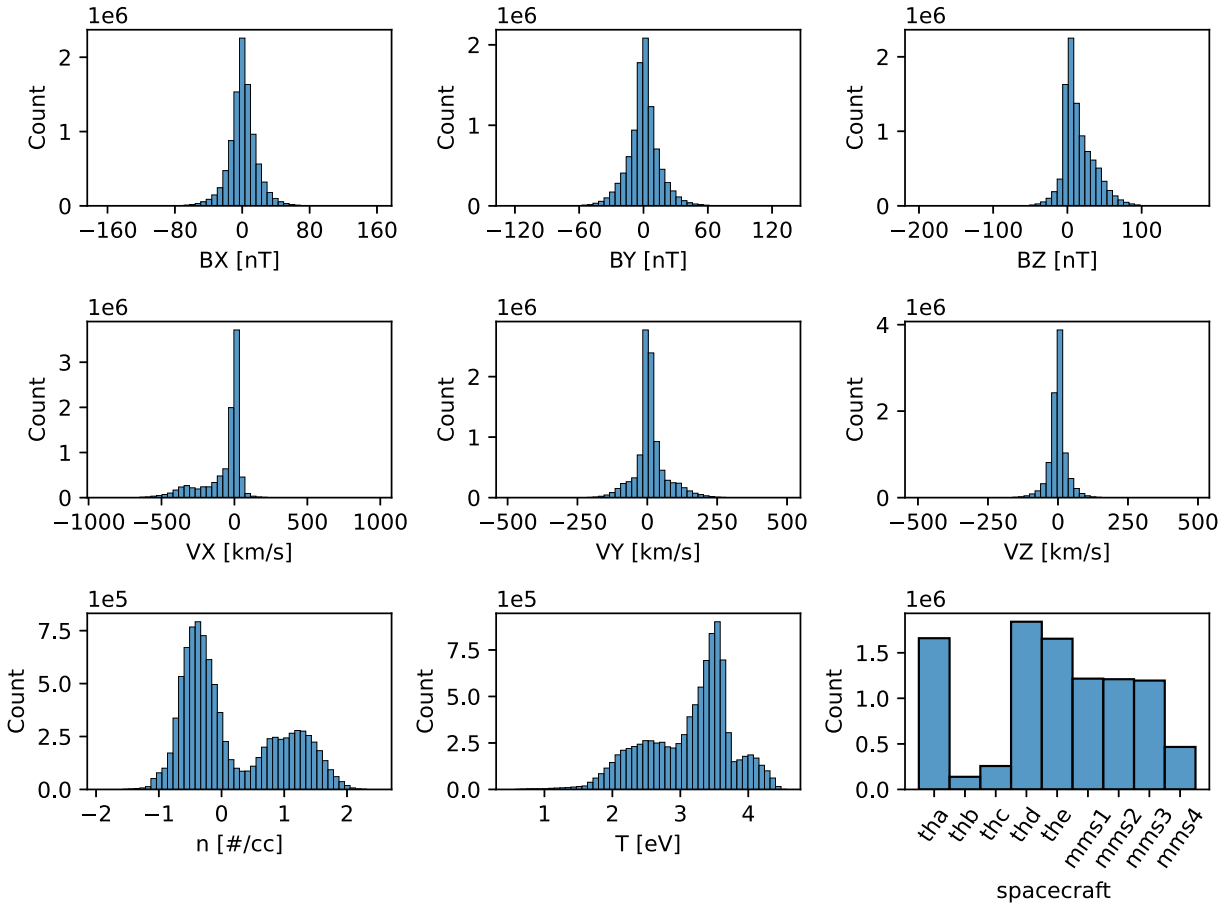

Figure 1. The distributions of all data collected, both training and testing. It is apparent from the density and temperature distributions (both in log10 scale) that multiple populations are present: Sparse (0.1 #/cc < n < 1 #/cc), moderate-density (1 #/cc < n < 30 #/cc), and dense (n > 30 #/cc) plasma and very cold (T < 10s eV), warm (10s eV < T < 1 keV) and hot (T > 1 keV) plasma. These different peaks in distributions are ideal for clustering.

resolution. Data from MMS 1, 2, and 3 span September 2015 to December 2021. Due to damage to the spectrometers of MMS 4, we only use data from September 2015 to 7 June 2018.

2.3. Data Cleaning

The THEMIS and MMS data sets possess 8.13 and 4.09 million points, respectively. The methods we apply to these data can be very sensitive to outliers and the size of magnetic field measurements closer to Earth could impact our ability to separate them in an unsupervised manner, so we constrain our data to be between 7 and 35 Earth radii. This final filtering leaves us with 9.64 million points, 4.09 million (42.4%) being MMS and 5.55 million (57.6%) being THEMIS. We separate our data with a test-train split of 95% and 5%, giving us a training size of ~482k points. The distributions of the magnetic field, ion velocity, ion density, and ion temperature measurements are shown in Figure 1.

3. Data Preprocessing

The eight variables, **V**, **B**, n, and T, hereafter referred to as features, of our data set do not possess enough variance for many unsupervised methods to sufficiently separate the regions. It is very common within machine learning to engineer derived features from the original in hopes of capturing non-linear relationships (Horn et al., 2020) because what is non-linearly separable in some space might become linearly separable in a higher dimensional space. To this end, we include the ion speed V, the magnetic field magnitude B, and the ion momentum density,

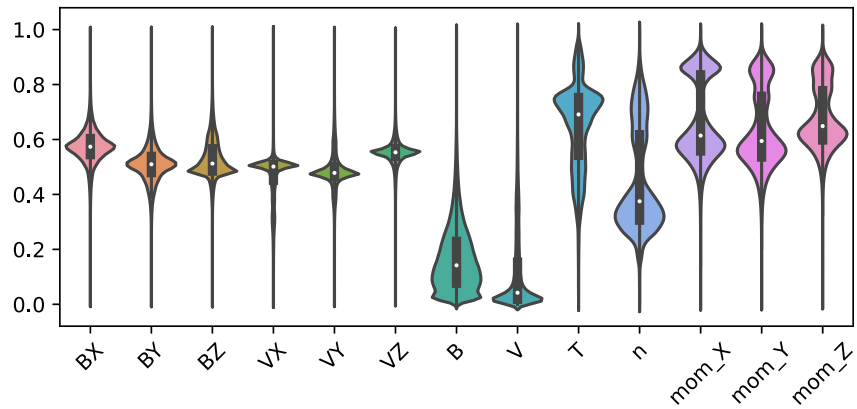

Figure 2. Violin plots representing the distributions of input features of our min-max scaled training set. The violin plots here show the kernel density estimate (KDE) as the width, the range of the estimate as a thin vertical gray bar, the interquartile range as a thick vertical black bar, and the median as a white dot. The KDE for each variable is scaled according to the width so that the distributions are more visible.

mom $= n$ **V** (with ion mass set to 1), as five additional features, giving us a total of 13 features. The addition of the ion momentum density vector is to help better separate the magnetosheath from the solar wind and magnetosphere as the magnetosheath acts as a transition region between them.

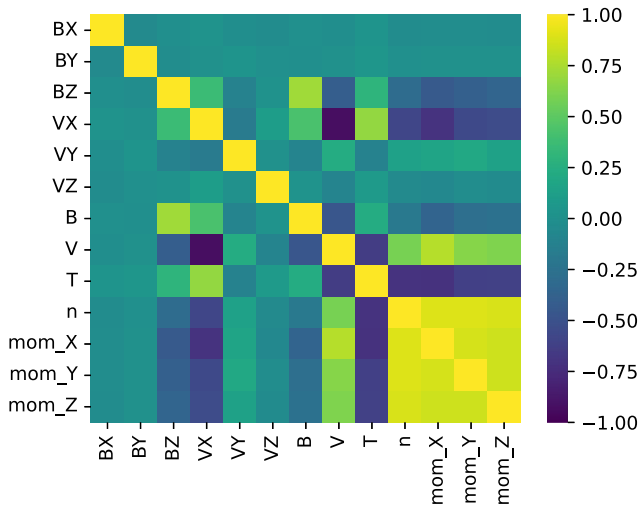

Figure 3. A heatmap of the correlations between variables in the min-max rescaled training set. The plot is symmetric across the diagonal. It is to be interpreted as showing the correlation of each feature with every other feature in the training set, for example, correlation (log10(n), log10(T))~ −0.6, or the log10 of the density is moderately negatively correlated with log10 of the temperature. There is a visible number of variable pairs with large magnitude in correlation (the bright or dark colored boxes in the off-diagonal). Also apparent is the absence of correlation of *BX* and *BY* with all other variables–even with (b) This is because the distributions of *BX* and *BY* are symmetric around 0, which is visible in Figure 1. The *VX* and *VY* components have correlation with V because large speeds (>350 km/s) are often going to be associated with solar wind, which generally possess large negative magnitudes in VX and slightly positive VY, on average (30 km/s), due to the angle that the solar wind arrives at the Earth. Lastly, note the positive correlation between VX and T. This is legitimate as the lowest values of VX occur in the solar wind, which is characterized by the lowest temperatures; more moderate values of VX and T occur in the magnetosheath; the largest (read, most positive) values of VX occur in the magnetosphere, which also possesses the largest temperatures.

Most of the features have ranges over a few orders of magnitude whereas the density, temperature, and momentum density components cover more than several. We convert the density and temperature to log10 scale, but the same cannot be done for the momentum density due to the negative values. This is circumvented by transforming the momentum density using the log10 of the absolute values of their components instead. After, these data still possess uneven ranges that can impact the performance of the dimensionality reduction and clustering methods we will use. To avoid feature bias, we rescale our training data using min-max normalization such that the new minimum and maximum of each feature is 0 and 1, respectively. The distributions of this rescaled training data is shown in Figure 2.

Non-negligible feature correlation is certain given our choice of features and this is evident in the correlation heatmap of Figure 3. The high number of correlated features means that direct clustering methods would be biased in the favor of these correlated components. Further still, the dimensionality can make some methods computationally expensive or cause them to find poor solutions due to the curse of dimensionality. The implication of the latter here is that distances between points will become smaller as the dimensionality increases, reducing the quality of clustering solutions. For data that does not possess significant outliers or that has been meticulously cleaned, the loss in quality of these solutions may be small, but it can become an issue for noisy data, especially data that are observations. We address both the correlation of features and dimensionality in the dimensionality reduction method to follow.

4. Dimensionality Reduction

Our training set prohibits using many clustering methods due to a combination of the training size, the dimensionality, and the presence of correlated features. We can simultaneously reduce the number of dimensions and the number of correlated features using one of the most prominent dimensionality reduction techniques, principal component analysis (Jolliffe, 2011), or PCA. This method provides a matrix Q with shape D × K to reduce the dimensionality D of a data set to a reduced dimension K via a linear transform where

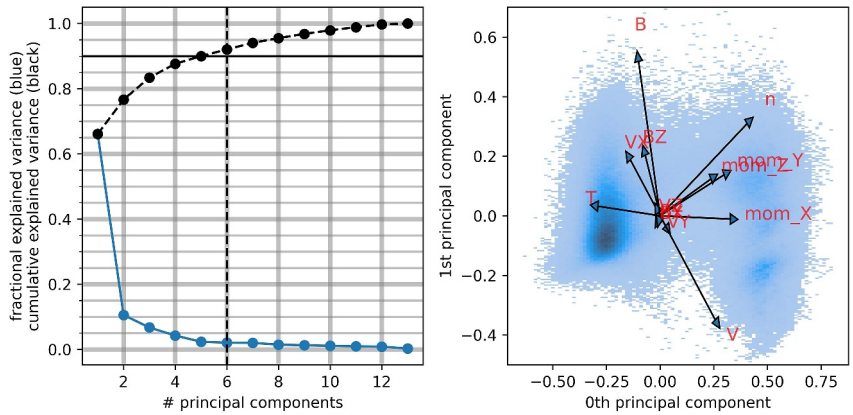

Figure 4. Left: The normalized eigenvalues from the PCA decomposition are plotted in descending order as the solid blue line. The cumulative sum of these normalized eigenvalues is plotted as the dashed black line. We choose to select a number of components representing at least 90% of the variance (the horizontal black line), so 6 components are chosen that represent 93% (the vertical dashed black line). Right: A bivariate histogram of the training data projected onto the first two principal components, representing 76% variance. It is evident from the first two components that several clusters are present in the data. The arrows plotted here are the loadings for our features across the first two principal components. The length of an arrow represents the influence that feature had for the PCA projection along that direction. All arrow lengths are normalized to the longest arrow, that of the B feature. From the plot, the temperature feature, T, significantly influenced the 0^{th} component but barely for the 1^{st} and points to the cluster on the left. This means that cluster is likely to correspond to higher temperatures than the data on the right. The density, n, roughly equally contributed to both components and indicates that the top right region is related to higher densities and by its antiparallel direction, the cluster on the left is largely associated with lower densities. Since VX points to the top left and V to the bottom right, the bottom right region is related to data with high speeds and large negative values of VX. The BX, BY, VY, and VZ features are clustered at the origin, indicating that they did not influence the first two components (although they may have impacted the higher order components). Overall, we can surmise from this plot alone that the left, top right, and bottom right areas are associated with higher temperature, higher density, and higher speeds, respectively. Thus, it is likely that these clusters are the magnetosphere, magnetosheath, and solar wind populations.

K is specified and can vary between one and D. It is analogous to a hyper-rotation of the D-dimensional space in which the cardinal axes, or principal components, are oriented along directions of decreasing variance. If a variance threshold is chosen, then a number of the principal components can be selected that cumulatively represent that variance. This method has limitations in that it is a obviously a linear method of dimensionality reduction. When data are characterized by non-linear correlations, this complicated structure can be destroyed in the transformation and can cause misinterpretations of the resulting components.

However, since PCA is linear, it can also be interpretable. Once Q is known, its elements, or "loadings," can be inspected to ascertain the influence of each feature along any principal component. Using just the first two principal components, we can visualize these loadings as vectors that can visually communicate the importance of each feature in the projection. Plotting these vectors on top of the first two components of the projection is called a biplot and is shown in Figure 4. Using biplots to infer information from PCA results has a rich history and an introduction to the concept is covered in Kohler and Luniak (2005).

Although feature correlations and dimensionality are simultaneously addressed using PCA, there is still the matter of a large training size after the PCA transform. The size can be reduced by simply randomly selecting fewer points, but this will only trade variance for sample size. Choosing enough points to represent a similar amount of variance will still require a large population size. In the next section, we use a method in which distinct points act as "representative" of their local distribution such that their amalgamation reflects the distribution of the training set.

5. Self Organizing Maps

K-Means (Lloyd, 1982) is the most popular clustering method and creates clustering solutions that separate data into k Voronoi-separated clusters where k is specified in advance. The most common convergence criteria used for this is the sum of square distances of all points from their cluster centroids, also called the inertia or quantization error, which is common among vector quantization methods (de Bodt et al., 2004; Gray, 1984). Self

Organizing Maps (Kohonen, 1982, 2014) can be viewed as a more powerful alternative to K-Means because they utilize a combination of competitive and cooperative updates during training. Individual cluster centroids, or nodes, of the map are updated to represent data in such a way that the topological relationships between the nodes are maintained throughout training. Preserving this relationship means that the average inter-node distance can be used to effectively create 2D visualizations of the data. Any data point will always be closest to some node (referred to as the Best-Matching-Unit, or BMU), and that node can represent the local distribution of data. Once a map is trained, this property can be utilized such that the nodes are used as input for other clustering algorithms. This greatly reduces the data size for clustering and expands the types of clustering methods we can use.

5.1. Implementation

There are several open-source python packages implementing SOMs available. The most common is minisom (Vettigli, 2018), which uses a vectorized design to speed up computations. For large data sets or network sizes, the time to completion may still be quite long. Traditionally, training a SOM has been a computationally expensive process for two reasons: The network adapts to one point at a time, and it is fairly common that multiple trainings are done. The latter occurs because SOM initialization and training are done stochastically and there is a large number of hyperparameter choices available (the number of iterations, the network size, the decay function, the neighborhood function, the initial and final learning rate and neighborhood size, etc). Since the network with the lowest quantization error is usually selected as the best fitting, this significantly increases the total amount of time needed to get a complete and robust model.

The one-at-a-time training constraint is resolved using SOMs that train over batch-updates. These usually involve computing weighted averages of the neighborhood values across a batch of samples. This approach is taken by two popular python packages Somoclu (Wittek et al., 2017) and XPySom (Mancini et al., 2020) and speed-up on CPU resources alone can be close to a factor of 100, sometimes greater. We have used the XPySom package for our results.

5.2. Hyperparameter Optimization and Training

To expedite the process of finding the best fitting SOM with the most appropriate set of hyperparameters, we create a micro training set. First, we min-max normalize the PCA-projected training data in order to avoid bias to any particular feature. Next, we run K-Means to resolve 10,000 clusters with a K-Means++ initialization method for 100 runs and select the optimal run based on minimal inertia. This initialization method makes better choices for cluster centroids by weighting data in proportion to their square distance from the previously created centroid. Then for each centroid, the closest point in the training data is extracted, and the resulting 10,000 points form the micro training set. The remaining points in the training data set are referred to as the macro training set with a size of 470k.

We consider a number of different SOM hyperparameters and that each SOM will be trained on the micro training set and validated on the macro training set. The maps are validated in this way because the macro set will contain a larger number of outliers, and given the noise evident in Figure 4, resolving these outliers correctly will be critical. The hyperparameters of the map with the lowest value for our loss function will be retained and a final SOM will be trained using these hyperparameters on the macro training set. We define our loss function to be

$$L = Q * \left(\frac{n_x n_y}{(n_x)_{\max}(n_y)_{\max}} + \frac{\max\{n_x, n_y\}}{\min\{n_x, n_y\}} \right). \quad (1)$$

where Q is the quantization error of the SOM, n_x and n_y are the dimensions of the 2D node grid, and $(n_x)_{\max}$ and $(n_y)_{\max}$ are the maximum values permitted for the x and y dimensions. The $\max\{n_x, n_y\}/\min\{n_x, n_y\}$ term penalizes non-square networks and will only allow for non-square maps should they provide a sizably lower quantization error.

It should be noted that the use of a custom loss function for SOM validation is critical for our purposes. With the number of training iterations and training data set held constant, increasing the map size will generally reduce the quantization error for many choices of hyperparameters. A larger map size may better represent the training data, and in many cases even the test data, than a smaller map, but a larger number of nodes and their distributions may

be suboptimal for clustering methods that will fit to these nodes. This can be loosely seen as a form of overfitting, but not in the sense of a model not generalizing well to unseen data. To illustrate this concept by example, consider training a "small" map on a large data set containing heterogeneous groups whose distributions are somewhat (but not extremely) non-convex. One might find that the distributions of the nodes mapping to these different groups are approximately spherically separable because there are few to no nodes mapping to outliers. This would be a good motivation to use K-Means to cluster the nodes for such maps. However, as the map size is increased, the node distributions will begin to better resemble the more complicated, original distribution of the data, which contains harder-to-resolve non-convex distributions that clustering algorithms like K-Means or Gaussian Mixture Models (GMM) may struggle to resolve.

The python-based optimization library Optuna (Akiba et al., 2019) is used to choose hyperparameter values. The training of each SOM on the micro training set is referred to as a trial. Optuna offers a variety of samplers to generate hyperparameter choices, and we use the Tree-structured Parzen Estimator with independent sampling as the sampler. It generates hyperparameter choices by fitting two sets of GMM per trial, one set for the better performing trials, $l(x)$, and another for the remaining, $g(x)$. Each set involves fitting a GMM for each hyperparameter x and the hyperparameter value selected is that which maximizes the ratio of density estimates $l(x)/g(x)$. Maximizing this ratio is consistent with choosing a hyperparameter that is simultaneously most likely to be generated by $l(x)$ (the "good" models) and least so by $g(x)$ (the "poor" models).

For our optimization, we considered the following hyperparameters. The number of nodes for the SOM grid n_x and n_y, the initial learning rate α, the initial neighborhood size σ, the neighborhood function H, and the decay function D. We have fixed the number of training epochs to be 50, the final learning rate and neighborhood size to be 0.01, and the maximum n_x and n_y dimensions to be 30. The values the hyperparameters are permitted to take are in the following list:

1. $5 \leq n_x, n_y \leq 30$
2. $1 \leq \sigma \leq \sqrt{n_x n_y}$
3. $0.1 \leq \alpha \leq 1$
4. D: {linear, exponential}
5. H: {Gaussian, Ricker}

5.3. SOM Results

After 500 trials, the best hyperparameter options are $(n_x, n_y) = (14, 14)$, $\sigma = 5.518$, $\alpha = 0.843$, $D =$ exponential, and $H =$ Ricker. We train a SOM with these hyperparameters on the macro training set which completes in 7 min. The resulting SOM has a quantization error of 0.0702 and 0.0703 on the macro training and test sets. The loss function rounds to 0.0855 and 0.0856. With a Intel Xeon 2.90 GHz E5-2690 (32 cores, 64 threads) CPU and 64 GB of RAM available, the entire process of hyperparameter optimization and final model training takes approximately 3 hr.

While the SOM we have trained has a good quantization error, there are visualization techniques we can use to further assess how well it represents the data. Since the goal of a SOM is to give a vector-quantized representation of the data, one simple approach is to create plots of the data itself with the SOM node positions overlaid. If it is an effective representation, it should roughly map to positions of high data density, both in scatter plots and histogram marginals. We show pairplots over the first three min-max scaled principal components of the test set in Figure 5. When scaling up the marginal histograms of the node positions to that of the marginal histograms of the test set, there is good agreement over the 0^{th} and 2^{nd} components. The 1^{st} component shows partial agreement with the node histogram, only somewhat capturing the peak in density between 0.3 and 0.4.

Another method uses the ordered nature of the SOM to create a heatmap of distances between the nodes. Since the nodes of a SOM have an ordered topological relationship, we can compute the average distance between a node and its immediate neighbors and create a heatmap of these average neighbor distances. The 2D matrix of these values is referred to as the U-Matrix. The U-Matrix for the test data is shown in the top left of Figure 6. Moreover, since each data point can be uniquely associated with its corresponding BMU in the SOM, we can then compute the average of all data per node. This average value per node can be used to create heatmaps of the SOM for any feature from the data, as seen in the remaining plots of Figure 6.

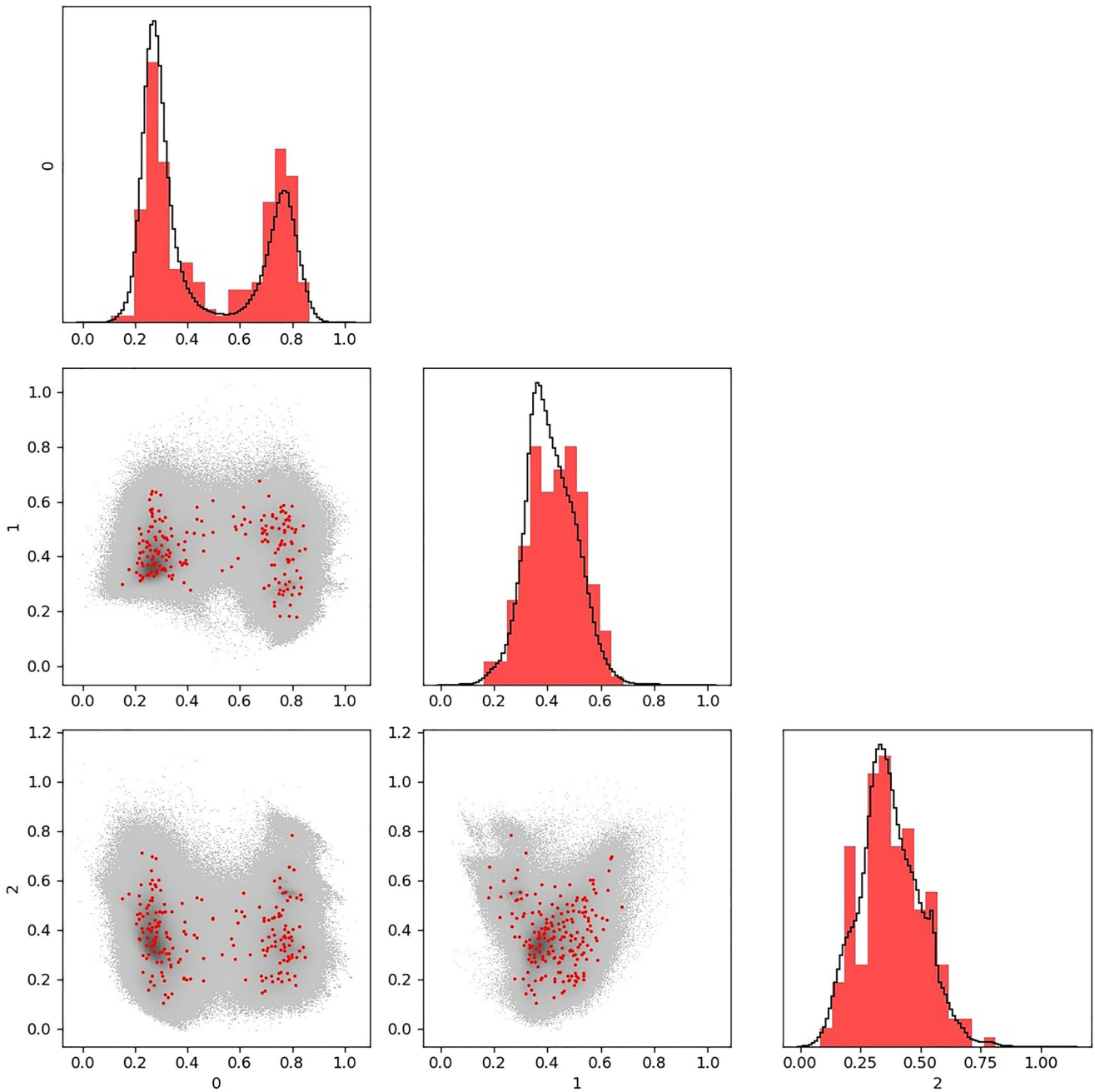

Figure 5. Pairplots over the first three min-max normalized principal components (83% variance) of the test set. The off diagonal plots are bivariate histograms for the test data in greyscale. Scatter plots of the Self-Organizing Map (SOM) node position are plotted in red on top of the bivariate histograms. The diagonal plots are the marginal distributions where the black line is the test data distributed over 100 bins. The SOM node positions are simultaneously binned but at a smaller resolution of 25 bins. The nodes generally match the histograms of the 0^{th} and 2^{nd} components with a dip noticeable in the nodes histogram of the 1^{st} component.

6. Clustering of SOM Nodes

Applying direct clustering methods caused difficulties involving size, dimensionality, and multicollinearity. We resolved the latter two using PCA and have addressed the first by training a SOM to act as a further discretized representation of the data. With a SOM representation, we now can consider a much wider choice of methods to cluster the data as training size is no longer a constraining factor. Once a clustering method is trained, it can separate the SOM nodes automatically, classifying which nodes belong to which cluster. These node classifications can then be propagated to the data that the nodes represent, that is, if a node A is assigned to cluster 1, then

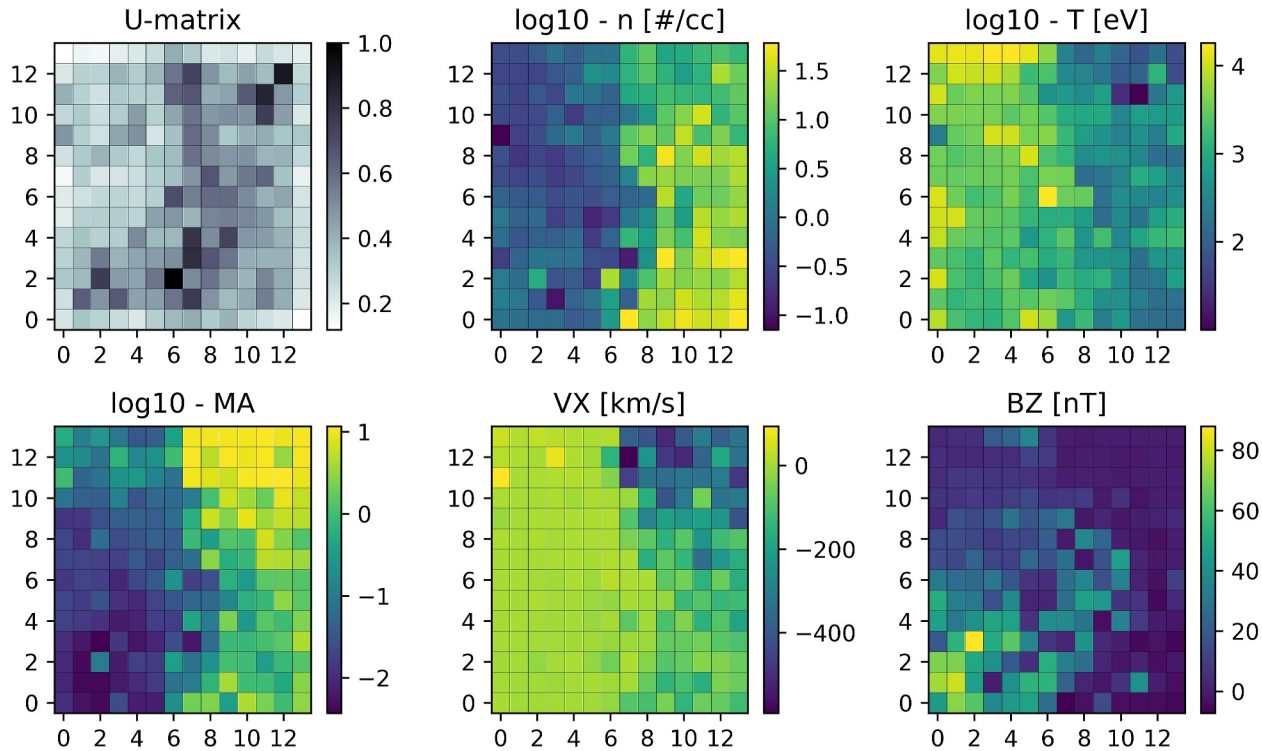

Figure 6. 2D heatmaps of the test data as seen through the Self-Organizing Map. In the U-matrix, plotted in the top left, nodes are colored according to their distance to the nearest neighbors: the lighter nodes are more similar to the neighbors than darker nodes. Note that neighbors here is defined in the square topological sense; nodes in the corners only have two neighbors, nodes along the rest of the perimeter have three neighbors, and all other nodes have four neighbors. The fewer neighbors among those on the perimeter means that there will usually be less variance among them such that the perimeter nodes have a lower (lighter) U-matrix value. A region of dark gray nodes partitions the U-Matrix into two areas of lighter color in the top left and bottom right. This means that there are two relatively homogeneous groups of nodes. To interpret what groups of data these nodes represent, we can look at the feature maps in the remaining plots. In these plots, the average feature value per node is depicted as a heatmap. It is apparent from the feature maps that the group of nodes on the left side of the U-Matrix correspond to regions of low density and high temperature. The nodes to the right correspond to moderate-to-high densities, low-to-moderate temperatures and negative values of VX.

all data for which node A is the BMU will be assigned to cluster 1. We used an agglomerative, or "bottom-up," form of hierarchical clustering as implemented in the scikit-learn package (an overview of various hierarchical clustering methods is covered in Nielsen (2016)). In order to focus on separating clusters based on homogeneity, we used a Ward linkage to determine the merge order of clusters. In hierarchical agglomerative clustering, if there are N clusters, then all N-choose-2 cluster pairings are considered for possible merging. The optimal merger is determined using a linkage function, which produces a number representing the similarity of the clustering where smaller numbers indicate more similar clusters, and the pair with the smallest linkage function value are merged. In some linkage functions, this can be interpreted as a distance, such as with the single, complete, average, and centroid linkages. The linkage we used, Ward's linkage, is instead concerned with identifying the cluster pair that minimizes the in-cluster variance. The entire model pipeline, including the approach used for hyperparameter optimization of the SOM, is shown in Figure 7.

The dendrogram of the clustered SOM nodes and their cluster assignments are shown in Figure 8. From the dendrogram, we make cluster classifications using a distance threshold of 1.65 and propagate the cluster assignments of the SOM nodes to the test data. The number of data points in the test set mapped per node is also shown in the same figure. Histograms of the classifications for each cluster are shown in Figure 9. These clusters were obtained in an unsupervised manner and an a posteriori analysis shows that they are in line with expert understanding of the solar wind, magnetosheath, and magnetosphere. In Figure 9, the solar wind corresponds to moderate density and supersonic Alfvén Mach number ($\log10\ M_A > 0$), the magnetosheath has the largest densities and shocked Alfvén Mach number ($\log10\ M_A$ 0), and the magnetosphere has the lowest densities and subsonic Alfvén Mach number ($\log10\ M_A < 0$). The Alfvén Mach number is used as a loose metric of success in that these regions can largely be distinguished with it.

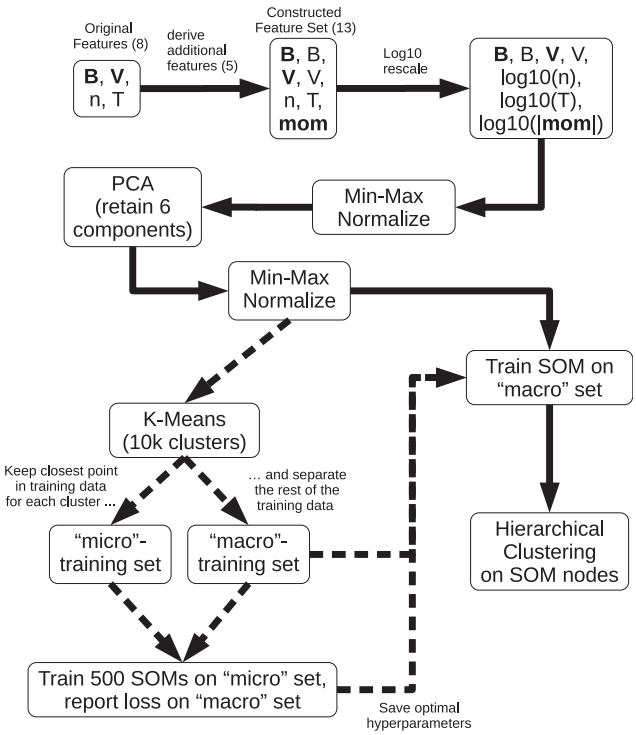

Figure 7. The pipeline of methods in our model. Solid arrows indicate a component of the model and dashed lines show how the optimal hyperparameters were learned using a micro and macro training set.

The clustering of the SOM nodes in PCA space is shown in Figure 10. We previously made conjectures as to what portions of the biplot from Figure 4 are associated with the solar wind, magnetosheath, and magnetosphere, and they are confirmed with the clustering depicted. In both the (0,1) and (0,2) plots of Figure 10, the magnetosheath cluster has overlap with both the magnetosphere and the solar wind clusters but the magnetosphere and solar wind clusters have little overlap with each other, as one can expect from the physics of the magnetospheric system. Higher order components possess less variance and show considerable overlap as seen in the (1,2) plot. This is a consequence of using PCA for dimensionality reduction: The first PCA components will generally capture the majority of the variance and subsequent components will be less significant.

In GSE coordinates, the solar wind tends to be in the sunward (here, rightward) direction, the magnetosphere in the tailward (leftward) direction, and the magnetosheath is a curved transition region between the two. The histograms of log10 density and log10 Alfvén Mach number of Figure 9 reflect this and show the clustering is very effective in separating supersonic, moderate density plasma (solar wind) from shocked, dense plasma (magnetosheath) and very subsonic, thin plasma (magnetosphere). Note that since the Alfvén Mach number is plotted in log10 scale, the supersonic to subsonic transition occurs as a change in sign. Overlap between these distributions can certainly occur and this is reflected in their histograms. Incorrect classifications are also visible in Figure 9, such as scattered magnetosheath and solar wind classifications occurring in the nightside at $-20\ R_E \leq Y_{GSE} \leq 20\ R_E$, a swath of magnetosheath classifications at $-10\ R_E \leq X_{GSE} \leq -5\ R_E$, and magnetosphere classifications well out into the dayside. In analyzing time series, these are generally spurious in that misclassifications occur but are relatively infrequent (such as in Figure 11 where the magnetosheath misclassifications are correlated with jumps in VX) and rarely part of consecutive misclassifications. We show two sample classifications of time series, one for THEMIS-C where the classification is exactly correct (Figure 11) and one where the majority of classifications are correct but suffer from spurious misclassifications (Figure 12). Analyzing when MMS 1 is in the solar wind in Figure 12, it's apparent that the magnetosheath-misclassifications correspond to higher temperature and lower absolute value of the velocity, as in the magnetosheath. When MMS 1 is in the magnetosheath, the solar wind-misclassifications correspond to higher absolute value in velocity and the magnetosphere-misclassifications correspond to lower density, again consistent with the characteristics of the region to which the measurements are incorrectly assigned.

In Figure 8, it is evident that the different clusters are largely segregated spatially in the node grid but exceptions are present. There are multiple nodes that are at best somewhat adjacent to the remainder of their cluster. Notably, the magnetosheath cluster has nodes at grid positions (12,12) and (8,12) that are surrounded by the solar wind cluster. The magnetosheath cluster also has a node that is surrounded by the magnetosphere cluster at (6,2) and a vertical streak of magnetosphere-classified nodes starting at (10,4). Results like this are not entirely unexpected as we are analyzing observations and the magnetosheath acts as a transition region between the magnetosphere and solar wind. We analyze the data that map to these nodes in detail in the appendix.

Lastly, we comment on our choice of SSD cutoff shown in Figure 8. In hierarchical agglomerative clustering, spaces on the dendrogram that show long vertical drops before another cluster bifurcation indicate that the clusters before the bifurcation are largely heterogeneous. Looking to Figure 8, this means that the two clusters that would be formed using an SSD = 2 cutoff would be quite distinct from each other. When we analyzed the four clusters resulting from an SSD cutoff of 1.3, inspection of this revealed that these four clusters corresponded to a solar wind cluster, a magnetosheath cluster, and a split magnetosphere cluster into two pieces. Since an SSD = 1.65 cutoff cleanly yielded a solar wind, magnetosheath, and single magnetosphere cluster, it was decided to use that cutoff instead. The two resulting magnetosphere sub-clusters are shown in Section 7.1. An interpretation of why the two magnetosphere sub-clusters possess such a high SSD (relative to the merged magnetosheath and solar wind) is that the variety of magnetospheric observations is comparable to the mutual variety of the solar wind and

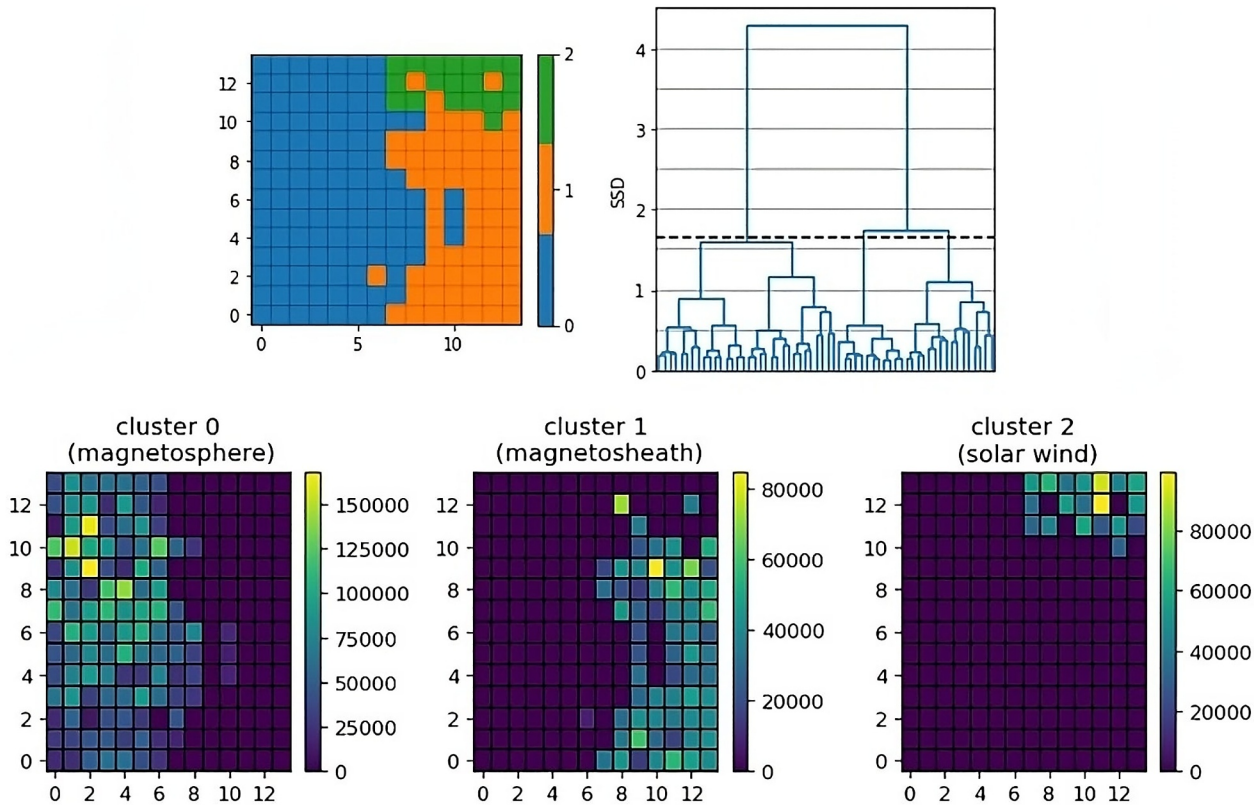

Figure 8. Top Right: A dendrogram of the clustered nodes using a Ward linkage. Separate clusters only up to the five most recent mergings are shown. We chose a cutoff sum of square deviations from the mean (SSD) of 1.65 to extract three clusters, as shown by the horizontal dashed black line. The number of times the line intersects with the vertical lines of clusters is the number of clusters recovered. The cluster assignments are visualized in the left image. Top Left: Cluster assignments of the Self-Organizing Map nodes shown on the 2D node grid. The region of low density and high temperature observed in Figure 6 has been assigned to cluster 0 (blue), the region of low VX is largely cluster 2 (green) and the region of high density is largely cluster 1 (orange). The color scheme used to represent the different clusters will remain the same. Bottom Row: For each cluster, the number of test points per node is shown. Note that the magnetosphere-classified nodes (10,6), (10,5), and (10,4) within the magnetosheath cluster contain few hits and the magnetosheath-classified node (6,2) within the magnetosphere cluster also contains few hits. However, the magnetosheath nodes (12,12) and (8,12) within the solar wind cluster are responsible for a sizable number of hits.

magnetosheath (or more simply, the magnetosphere, as observed by THEMIS and MMS, has almost as much "variance" as the magnetosheath and solar wind combined).

7. Applications

7.1. Subpopulation Analysis

We show in brief the capability of subpopulation analysis with this clustering method. Since we have used a hierarchical method to cluster the SOM nodes, we can pick any cluster and investigate the previously merged clusters that compose it. We "unpack" the magnetosphere cluster in Figures 13 and 14 to show how distinct magnetospheric populations were collectively recognized as the magnetosphere. From the histograms, we see that the feature that changes most clearly between the two clusters is the Alfvénic Mach number. Note that the subclusters of the magnetosphere in Figure 13 are not as evenly topologically separated like the original clustering solution seen in Figure 8. This is not surprising given the large overlap in features between these subclusters seen in the univariate histograms of Figure 14 and indicates that the variance between these two subclusters is less than the variance between the magnetosphere, magnetosheath, and solar wind clusters, hence these two subclusters appearing earlier in the merge order with a Ward linkage. In simpler terms, it is easier to distinguish solar wind measurements from those of the magnetosheath or magnetosphere than it is to separate magnetospheric populations by Alfvén Mach number.

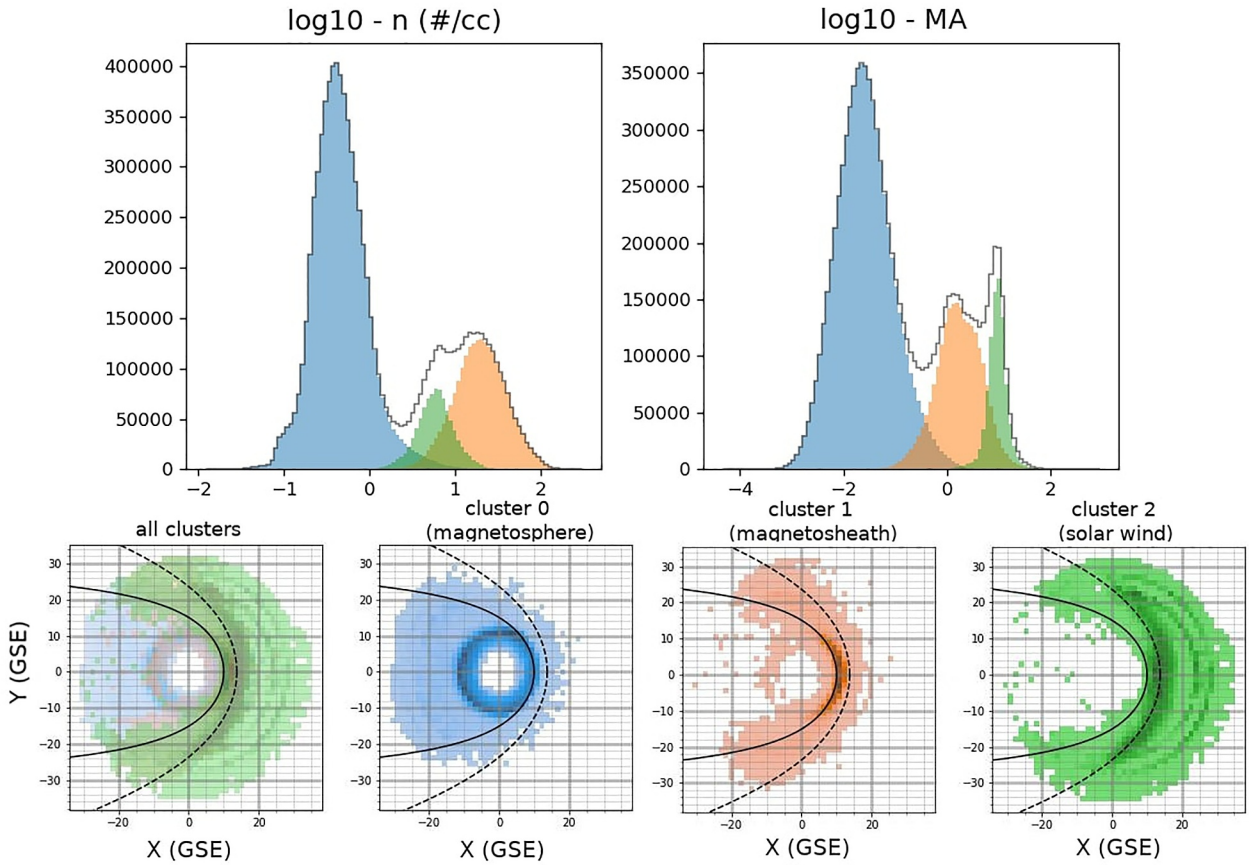

Figure 9. Top/univariate histograms: Histograms of the log10 density and log10 Alfvén Mach number. The histogram over the entire test set is in black and the histograms of the three clusters of the test set are represented in color. The magnetosphere is in blue (cluster 0), the magnetosheath is in orange (cluster 1) and the solar wind is in green (cluster 2). Bottom/bivariate histograms: ($X_{GSE}[R_E]$, $Y_{GSE}[R_E]$) bivariate histograms of cluster occupancy where the sun is on the right. The leftmost plot shows the histogram over the entire test set and each other plot shows an occupancy histogram for a particular cluster of the test set. The cluster color scheme used is the same as in Figure 8. A darker shade of color indicates a higher count in the bivariate bin. The solid line is a Shue magnetopause and the dashed line is a Chao bow shock. The parameters for these models are BZ = 0.15 nT, $D_p = 2$ nPa, $M_{MS} = 6$, and $\beta = 2$.

7.2. Derived Boundary Crossings

With a model that can classify when a measurement occurs in the magnetosphere, magnetosheath, or solar wind, we can study the time series of these classifications and infer when a spacecraft has crossed the magnetopause or bow shock. To select crossings, we used a moving window over the time series of classifications and find where the classification changes from magnetosheath to solar wind or vice-versa. We considered such a change in classification to be a crossing if all points half a window length before belong to one cluster and all points half a window length ahead belong to the other. The changing time resolution in the THEMIS data means that we need to consider different window lengths between MMS and THEMIS observations. A window length of 20 min was used for MMS to give up to 10 points per half window length and a window length of 40 min for THEMIS to give up to 13 points per half window length when the ESA is in Fast-Survey Mode (32 spins, 96 s, going from the magnetosheath to the solar wind) or up to 3 points per window when it is in Slow-Survey Mode (128 spins, 6.4 min, going from the solar wind to the magnetosheath). A total of 3,047 bow shock crossings and 5,228 magnetopause crossings are extracted using these parameters. Bivariate histograms of the (X_{GSE}, Y_{GSE}) positions of these crossings is depicted in Figure 15 alongside a Shue magnetopause (Shue et al., 1998) and Chao bow shock model (Chao et al., 2002) and show good agreement with respect to both.

For the bow shock crossings, we select the most recent solar wind point relative to the time of crossing and see how they're distributed in the SOM grid in Figure 16. When cross-comparing these with the number of counts in the test set from Figure 8, we see that the two most activated nodes of bow shock crossings are nodes (10,11) and (12,11). These nodes are responsible for 21.7% of the crossings but only 11.5% (training + testing) of the solar

10.1029/2024JH000221

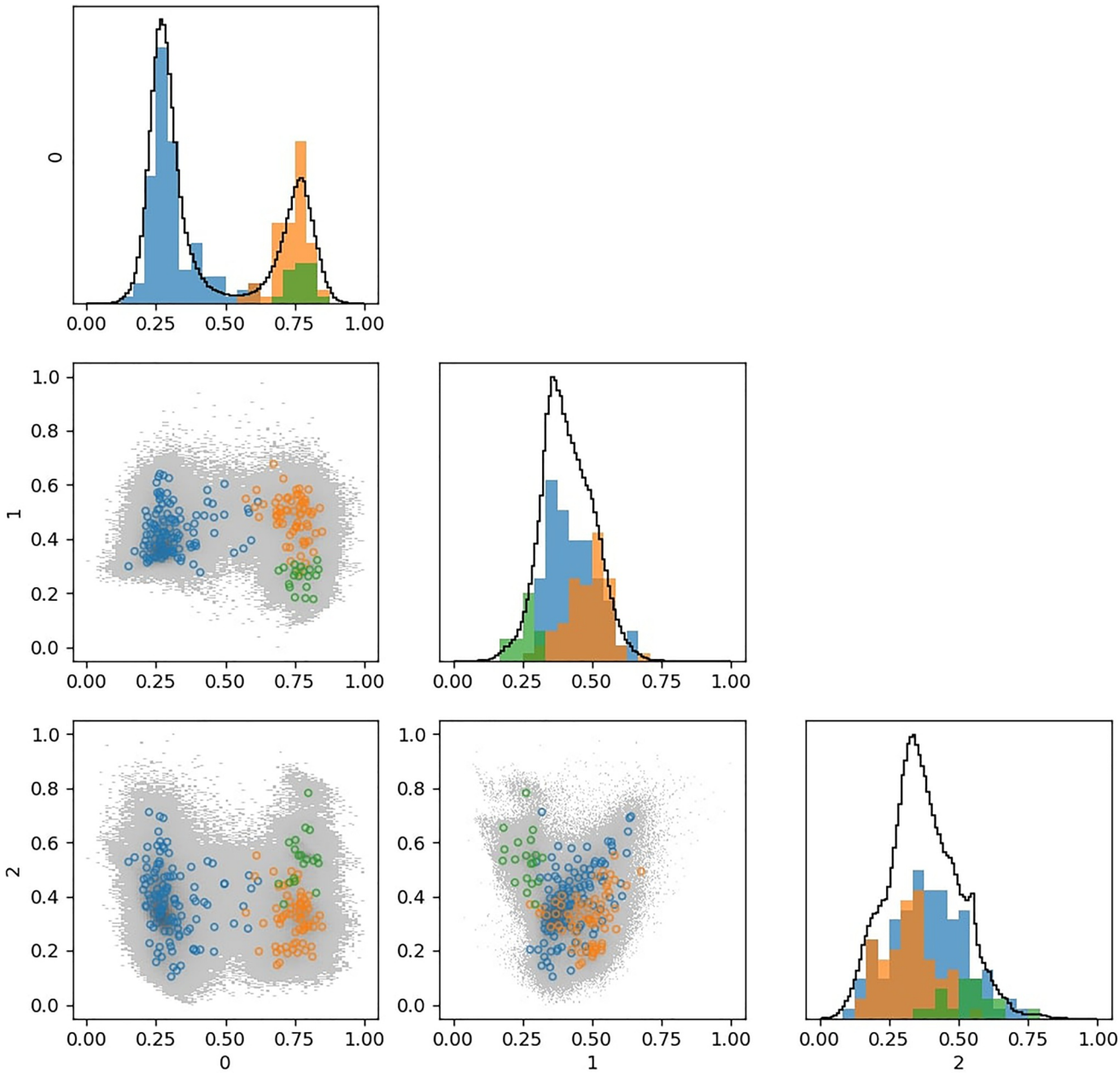

Figure 10. Cluster assignments of Self-Organizing Map (SOM) nodes over the first three min-max normalized principal components of the test set. Comparing the plot of the (0,1) component-transformed data (center-left plot) to the biplot over the first two principal components in Figure 4, we observe that the region on the left is the magnetosphere, the upper right is the magnetosheath, and the lower right is the solar wind. The marginal histograms of all clusters are shown along the diagonal using the same bin ratio (100 bins for data and 25 for SOM nodes) as in Figure 5.

wind classifications. This means that the model could be used such that a solar wind measurement assigned to one of these nodes could be flagged as having an increased probability of being a solar wind point adjacent to a bow shock crossing. Additionally, the node with the highest count in the test set for solar wind points, node (11,12), has only a small number of bow shock crossing points (6.2%) relative to the previous nodes.

We perform a similar analysis for the magnetosheath points relative to the magnetopause crossings. The nodes with the highest number of counts of magnetosheath points associated with magnetopause crossings are the nodes (9,5), (8,8), and (8,2). These are responsible for 18.2% of the magnetopause crossings but only 3.0% of the magnetosheath classifications (training + testing). The node with the largest number of magnetosheath points in

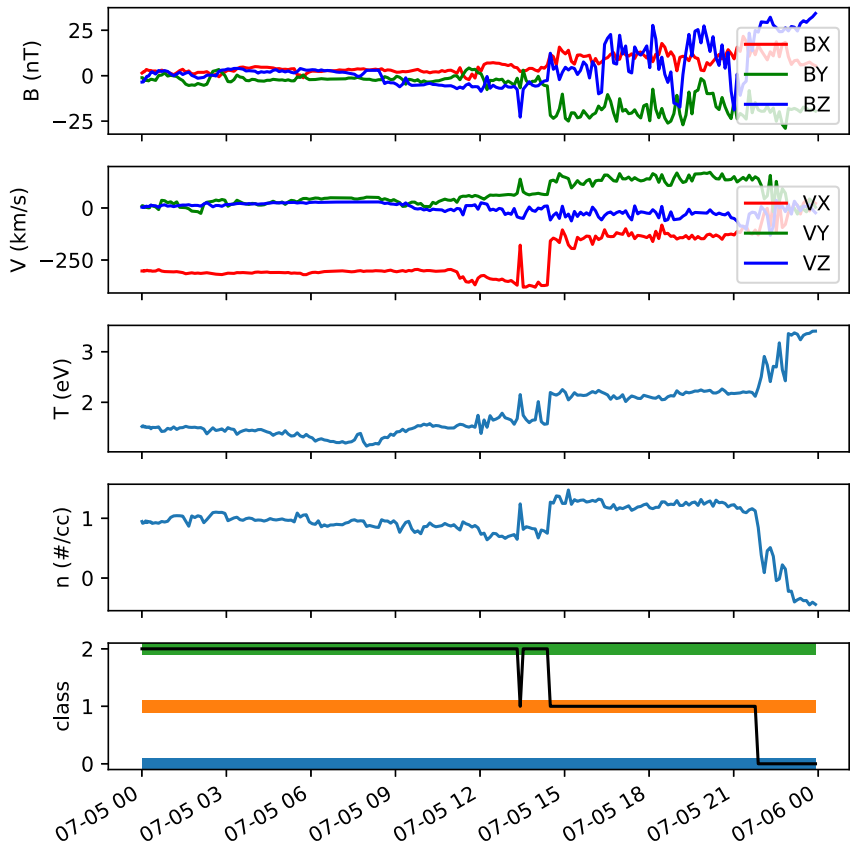

Figure 11. THEMIS-C measurements from 2008 to 07-05 to 2008-07-06. The temperature and density are in log10 scale. The classifications are shown in the bottom plot with the same cluster color scheme as Figure 8. The model successfully classifies the solar wind, magnetosphere, and magnetosheath measurements according to our visual verification. Noticeably, it also catches the "blip" when THEMIS-C is briefly in the magnetosheath before again crossing the bow shock and going back into the magnetosheath at 14:00 UT.

the test set, node (10,9) at 3.6%, only contains 15 magnetosheath points of the crossings, or 0.29% of the magnetopause crossings. These three nodes could be used to flag possible magnetopause crossings.

7.3. Identifying Bursty Bulk Flows

Bursty Bulk Flows (BBF) are earthward-moving plasma flows in the magnetotail that are often characterized by large speeds toward Earth (hence a large, positive VX component), dipolarizations, depletions in density, and increases in temperature and are an important process in the earthward transport of mass, energy, and magnetic flux in the magnetosphere (Angelopoulos et al., 1994). Detecting a dipolarization in magnetic field data alone is inherently a time-dependent comparison, but detecting large VX components can be done in a time-independent manner. Using the feature maps from Figure 6, we see that nodes (0,11) and (3,12) are magnetosphere-classified nodes that have large average VX values of almost 100 km/s. Thus we can use these nodes to identify possible BBFs. A data set of BBFs as observed by MMS from 2017 to 2021 was created by Pitkänen et al. (2023), and they show two examples in their paper. We show that the BBF of their first example corresponds to many activations of the (0,11) node in Figure 17. Not every activation corresponds to a BBF, but a rolling window method counting the number of activations could be used to flag possible BBFs.

7.4. Identifying Hot Flow Anomalies and Foreshock Bubbles

Hot Flow Anomalies (HFA) and Foreshock Bubbles (FB) are transient phenomena that are often observed in the ion foreshock. HFAs form from the interaction of a tangential discontinuity with the bow shock and can result in particle energization, diminished density and magnetic field, and flow turning sunward (Omidi & Sibeck, 2007; Schwartz et al., 1985). FBs are instead formed prior to this interaction but can possess similar characteristics of

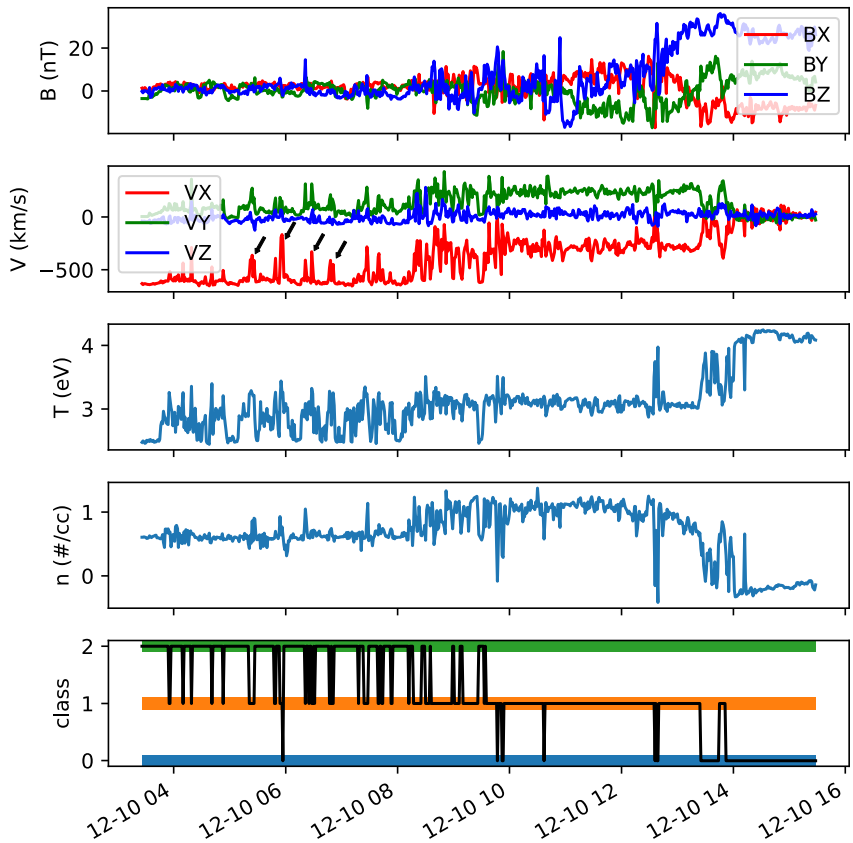

Figure 12. MMS 1 measurements from 2018 to 12-10. The plot structure is the same as Figure 11. MMS 1 crosses the bow shock at about 8:00 UT and the magnetopause shortly after 13:00. The majority of the classifications prior to crossing the bow shock are solar wind, but there are a number of incorrect and spurious magnetosheath classifications that occur with sharp increases in VX (as indicated by the black arrows) as well as one magnetosphere classification around 10:30 UT. After 8:00 UT, the majority of classifications changes to magnetosheath with rarer solar wind and magnetosphere classifications occurring. In the interval when MMS 1 is in the magnetosheath, magnetosphere misclassifications correspond with sudden drops in density measurements.

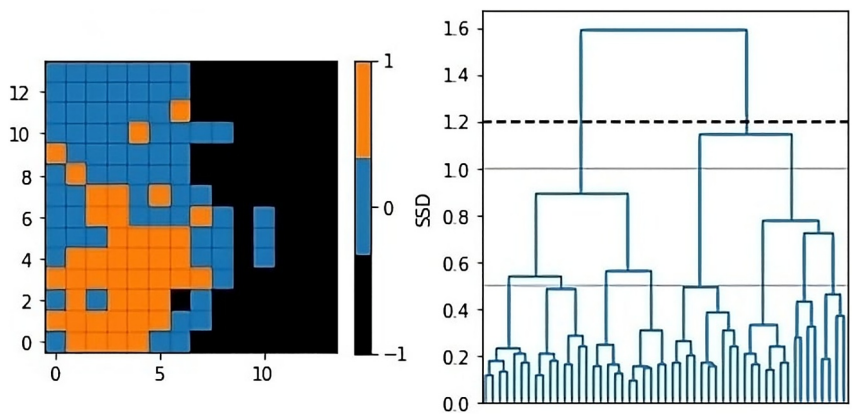

Figure 13. Like Figure 8 but only focusing on the magnetosphere cluster. Right: A dendrogram showing the merge order of the magnetosphere cluster. This tree is a subset of the dendrogram in Figure 8. We use a cutoff SSD of 1.2 and extract two clusters from the magnetosphere cluster. Left: Subcluster assignments of the Self-Organizing Map nodes based on the distance chosen in the dendrogram. The nodes that did not belong to the magnetosphere cluster are masked out in black and assigned a label of −1. Looking back to the feature maps in Figure 6, we can see that the blue cluster (0) is related to higher subsonic Alfvén Mach number and the orange cluster (1) is related to lower subsonic Alfvén Mach number.

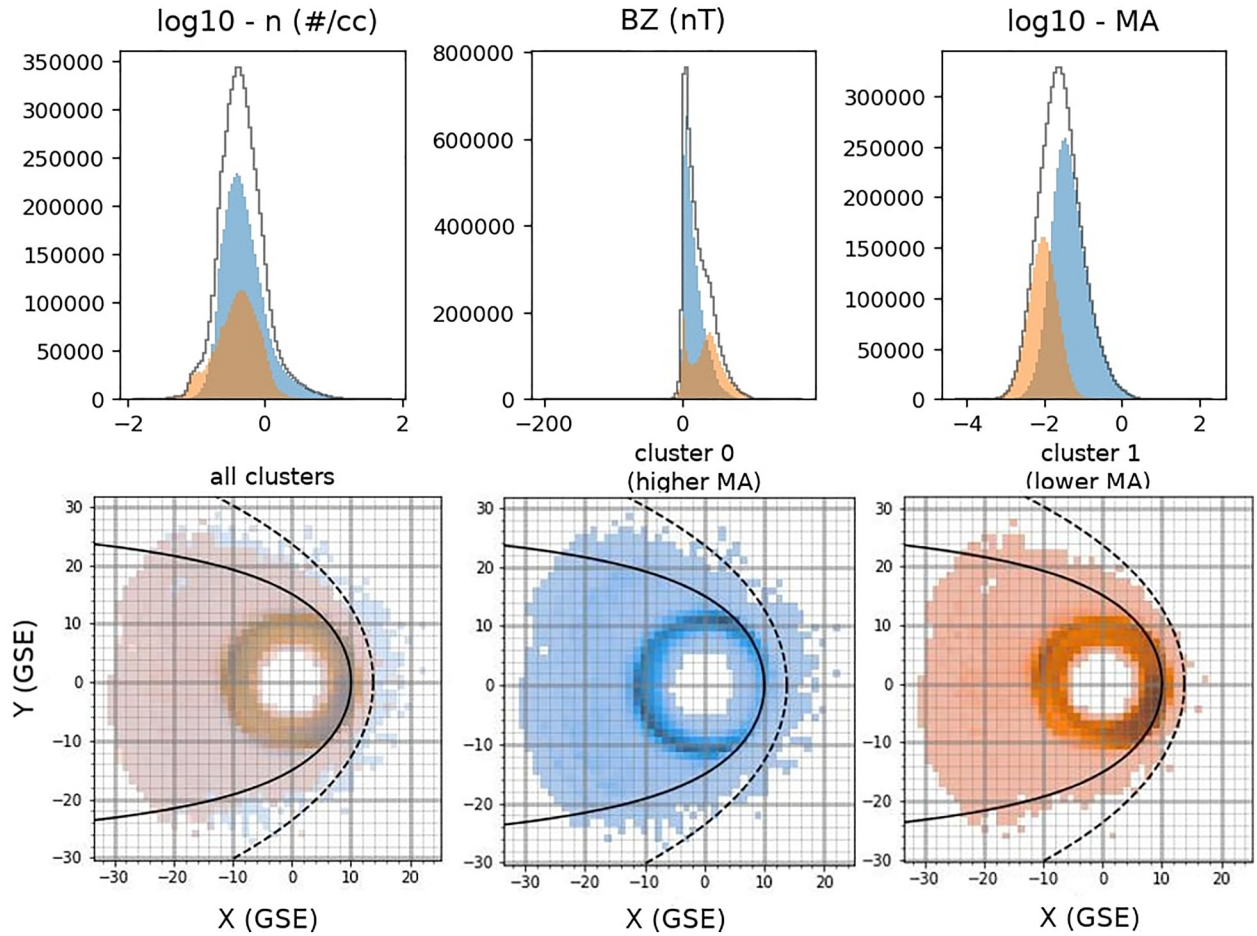

Figure 14. Like Figure 9, but analyzing only the magnetosphere cluster of the test set. Bottom/bivariate histograms: The occupancy of cluster 0 (blue) and 1 (orange) are plotted as bivariate histograms in ($X_{GSE}[R_E]$, $Y_{GSE}[R_E]$). They cover a similar region, but cluster 1 is much less pronounced on the dayside. The solid and dashed lines are again a Shue magnetopause and Chao bow shock using the same parameters described in Figure 9. Top/univariate histograms: The histograms of log10 density, BZ, and log10 Alfvén Mach number are plotted in black and the cluster populations are plotted in their respective colors. As could be inferred from Figure 6, cluster 0 is related to higher subsonic Alfvén Mach number and cluster 1 to lower subsonic values.

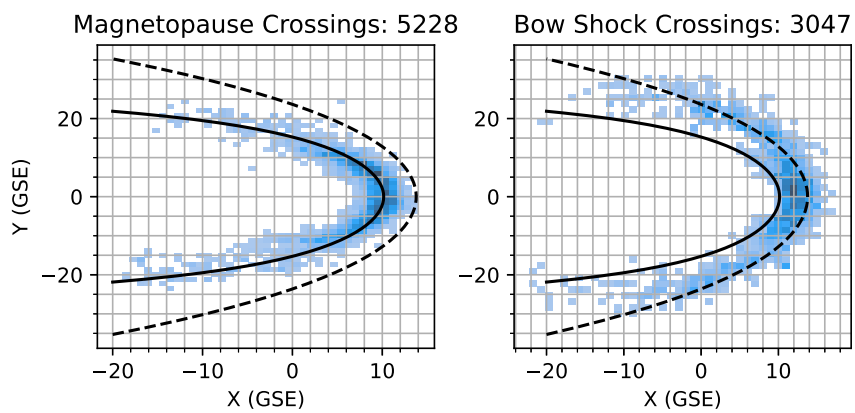

Figure 15. Bivariate histograms of the magnetopause (left) and bow shock (right) crossings in ($X_{GSE}[R_E]$, $Y_{GSE}[R_E]$). In both figures, the solid line is a Shue magnetopause and the dashed line is a Chao bow shock. The parameters for these models are the same as described in Figure 9. Many of the crossings are in line with expectations of magnetopause and bow shock positions although a handful of errant crossings are evident, such as the magnetopause crossings at ($X = -4$, $Y = 7$) and ($X = 5$, $Y = 25$). Nightside bow shock crossings at $X <= -10$ start to deviate from the Chao model due to the orbital bias of THEMIS and MMS wherein the bow shock is only crossed due to its compression from higher solar wind dynamic pressure.

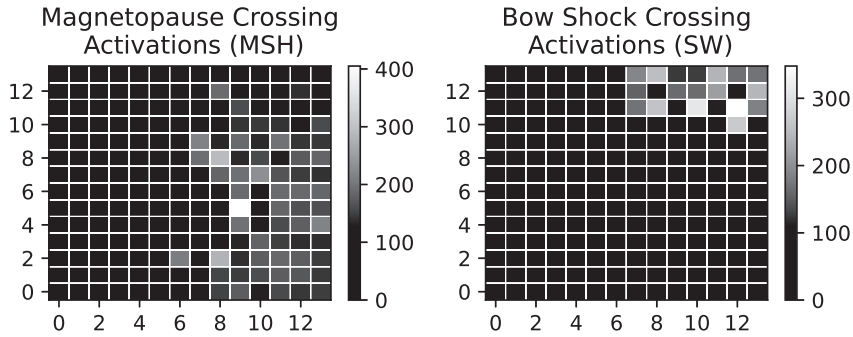

Figure 16. For each magnetopause (bow shock) crossing, we select the most recent magnetosheath (solar wind) point. Each point maps to, or "activates," some node in the Self-Organizing Map. The distribution of these counts is shown for the magnetosheath points for the magnetopause on the left and the solar wind points for the bow shock on the right. For the magnetosheath points, the most activated nodes are at positions (9,5), (8,8), and (8,2) and are together responsible for 949 crossings. For the solar wind points, the most activated nodes are at positions (10,11) and (12,11) and are responsible for 660 crossings.

low density and field strength and reduced VX/sunward flows (Omidi et al., 2010, 2020). These properties mean that these observations could be classified as magnetosheath or magnetosphere. Thus a simple way to identify possible HFAs and FBs using this model is to track sequential solar wind-classified data and find gaps in the classifications. Liu et al. (2022) compiled a list of observations of HFAs and FBs from MMS1 and THEMIS-A, 47 of which are from November and December 2017 of MMS1. Using the same 4.5 s resolution data set we previously prepared, we extract solar wind classification gaps of up to 2 min duration. Allowing an observation to be within up to 30 s of an identified gap, we find that we can identify 39 of the 47 observations. An example interval of MMS1 data containing seven HFA/FB observations is shown in Figure 18.

8. Discussion

Ours is not the only model that has attempted to classify spacecraft observations into different plasma regions. Olshevsky et al. (2021) used a convolutional neural network trained on the ion energy distributions of MMS to classify them as magnetosphere, magnetosheath, pristine solar wind (PSW), or ion foreshock and Nguyen et al. (2022) used a gradient-boosted decision tree trained on magnetic field and ion moments of a variety of spacecraft to classify them as magnetosphere, magnetosheath, and solar wind classes. Breuillard et al. (2020) also used a convolutional neural network on MMS measurements of the magnetic field components **B** and magnitude B, the ion velocity components **V** and magnitude V, the ion density, and the parallel, perpendicular, and total ion temperatures to classify them into PSW, ion foreshock, bow shock, magnetosheath, magnetopause, boundary layer, magnetosphere, plasma sheet, plasma sheet boundary layer, and lobe.

Olshevsky et al. (2021) created a labeled data set and has comparable classes to our model, so we have made comparisons with their model and data. They curated two month's worth of MMS1 data, covering November and December 2017 to the total of 469k points and created two models. One of their models was trained on the November 2017 data and tested against the December 2017 data and the training and testing were reversed for the other. They did not use the full data sets for training and instead used about ∼25k points each for November and December, making sure to evenly sample from the four classes to avoid class imbalances. We use their better performing model, which was trained on December 2017 and tested against November 2017, as a comparison. We prepared both magnetic field and ion observations (averaging the magnetic field measurements to the latency of the ion observations at 4.5 s resolution) and assigned their labels to our prepared data set of 467k points, discarding the 2k unrecognized points. Since their model relied on correctly classifying the ion sky maps, they anticipated that complex mixing of distributions could occur at the magnetopause and bow shock, and so any data that indicates distribution mixing was assigned to the class "Unknown," comprising about 15% of their data set. We mask these points out when comparing the accuracy of these models.

As explored in a previous section, the hierarchical capability of our model means that we can further derive subclasses from our original classification. To directly compare against the model of Olshevsky et al. (2021), we will unpack our solar wind cluster into two sub-clusters and regard one as the PSW and the other as the ion foreshock.

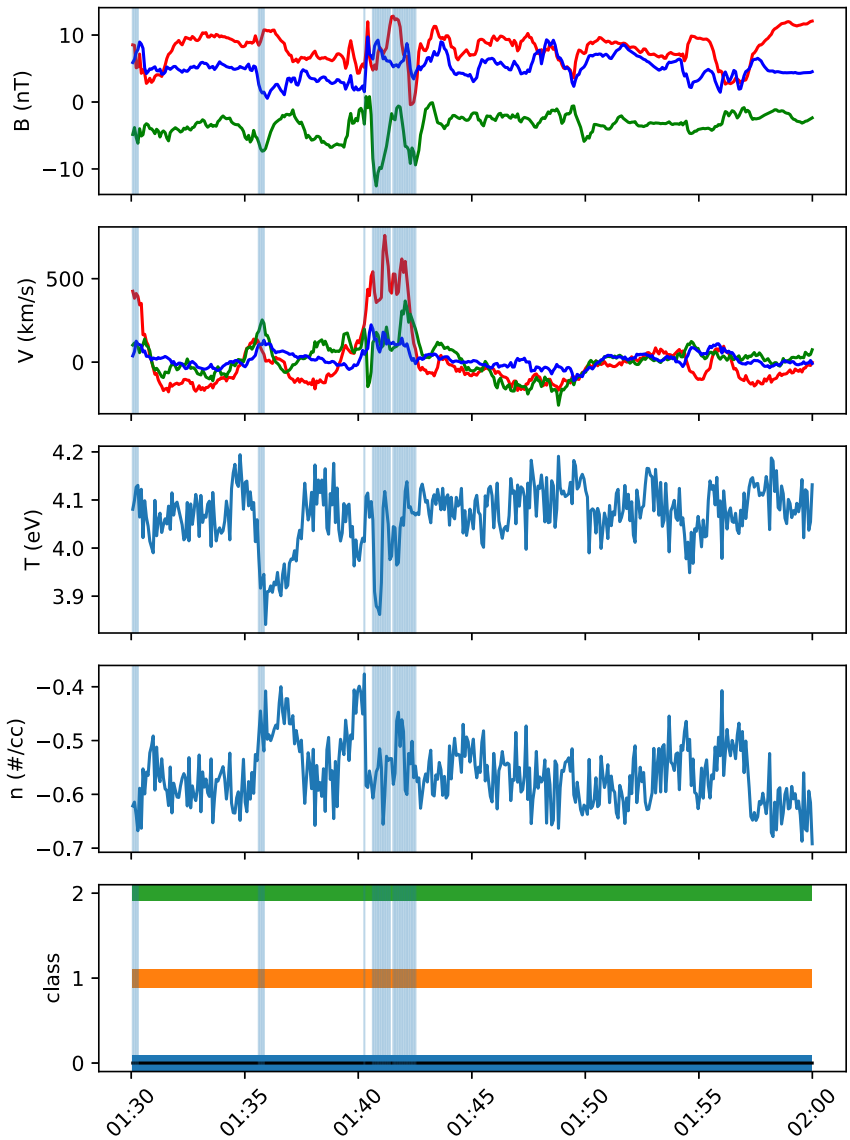

Figure 17. MMS 1 measurements from 01:30 to 02:00 UT on 2021-08-15 at 4.5 s resolution. The plot structure is the same as Figure 11. The vertical blue lines here denote the magnetosphere-classified points that mapped to node (0,11). Pitkänen et al. (2023) identified the Bursty Bulk Flows interval as lasting from 01:39:56 to 01:42:38 UT, and 26 of the 36 points in that 2 min 42 s interval map to node (0,11).

To compare model performance in our 3-class classification, we fold together the ion foreshock and PSW labels collectively as solar wind. Confusion matrices of the classifications of both models in both cases and their overall accuracy for each are shown in Figure 19. For magnetosphere/magnetosheath/solar wind classification, our model's overall accuracy (99.41%) is approximately equal to theirs (99.39%) but the per-class accuracy varies. Our model's accuracy for magnetosphere and magnetosheath predictions is quite high at 100% and 99.8% but our solar wind classification is only 99.1%. The false negatives of the solar wind and magnetosheath classes are almost entirely magnetosheath and magnetosphere labeled data at 0.9% and 0.02%, respectively. Their model's most accurately classified category is solar wind at 99.8% followed by magnetosphere and magnetosheath at 99.1% and 98.6% and the amount of solar wind false positives is 1.4%. For magnetosphere/magnetosheath/ion foreshock/PSW classification, our model's accuracy is only 86.7% with a per-class accuracy of 83.0% and 76.4% for the ion foreshock and PSW. It can also be seen that the solar wind-labeled data that our model misclassified as magnetosheath almost always corresponded to ion foreshock labels. Their model certainly outperforms here, correctly classifying the ion foreshock and PSW classes at 92.4% and 98.2% accuracy. This is not surprising as

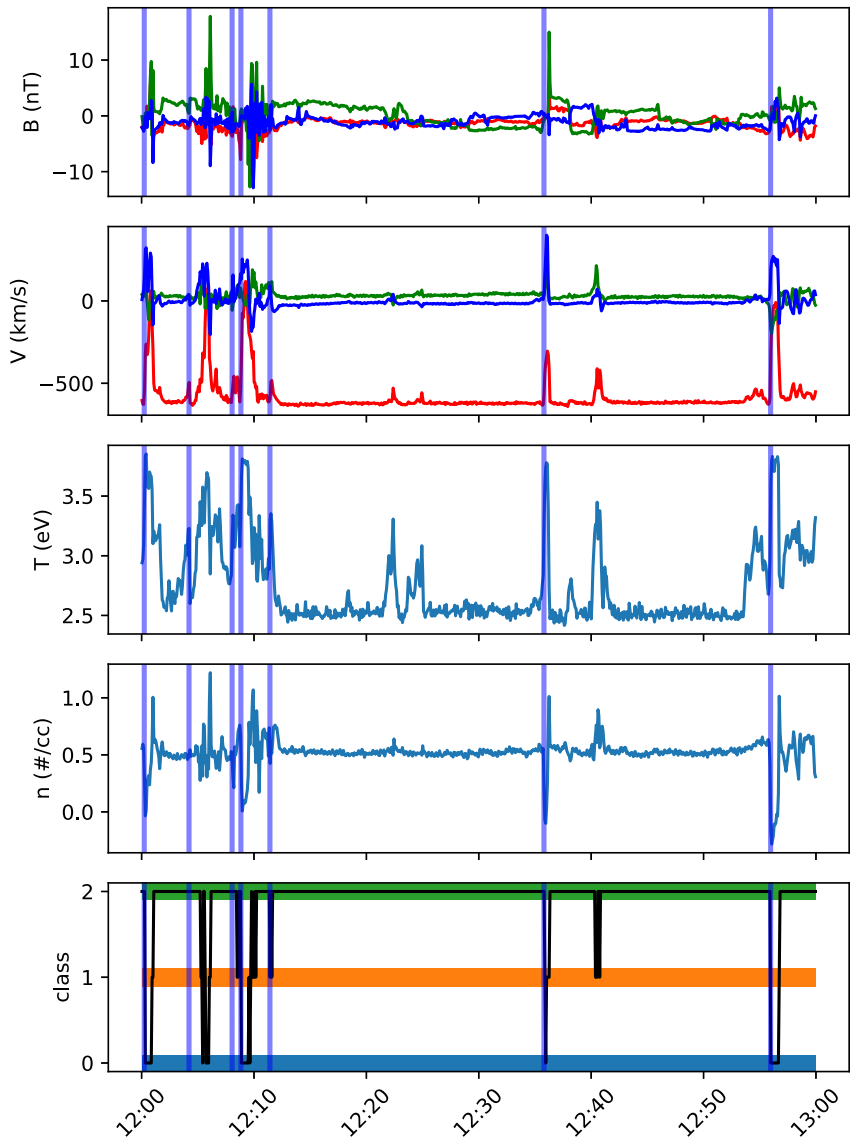

Figure 18. MMS 1 measurements from 12:00 to 13:00 UT on 2017-12-18. The plot structure is the same as Figure 11. The vertical purple lines here denote a HFA/FB time as recorded by Liu et al. (2022). Six of the seven observations were recognized with our method, the exception being the observation at 12:04:13 UT. This missed observation is still reflected in the sequence of magnetosheath/magnetosphere classifications occurring near 12:05 UT, but is beyond our 30 s window.

they used a supervised 3D convolutional neural network with a much more diverse data set of 32 × 16 × 32 features and our model is an unsupervised neural network using data with only 13 features.

Rather, it should be expressed that our model is able to achieve a similar 3-class accuracy compared to a much more robust model. Moreover, the most significant advantage of this model is that it utilizes a SOM's ability to analyze data using feature maps. One approach can be similar to how we used it to flag possible BBFs, in which there is a parameter of interest and one wants to identify other data points that have similar characteristics across multiple features (a "node-to-data" method). Or the reverse can be done, where one possesses unique data and wants to find other data like it or see if it corresponds to repeated activations of the map, like what was done in taking observations that were adjacent to boundary crossings and analyzing what nodes were activated in response (a "data-to-node" approach).

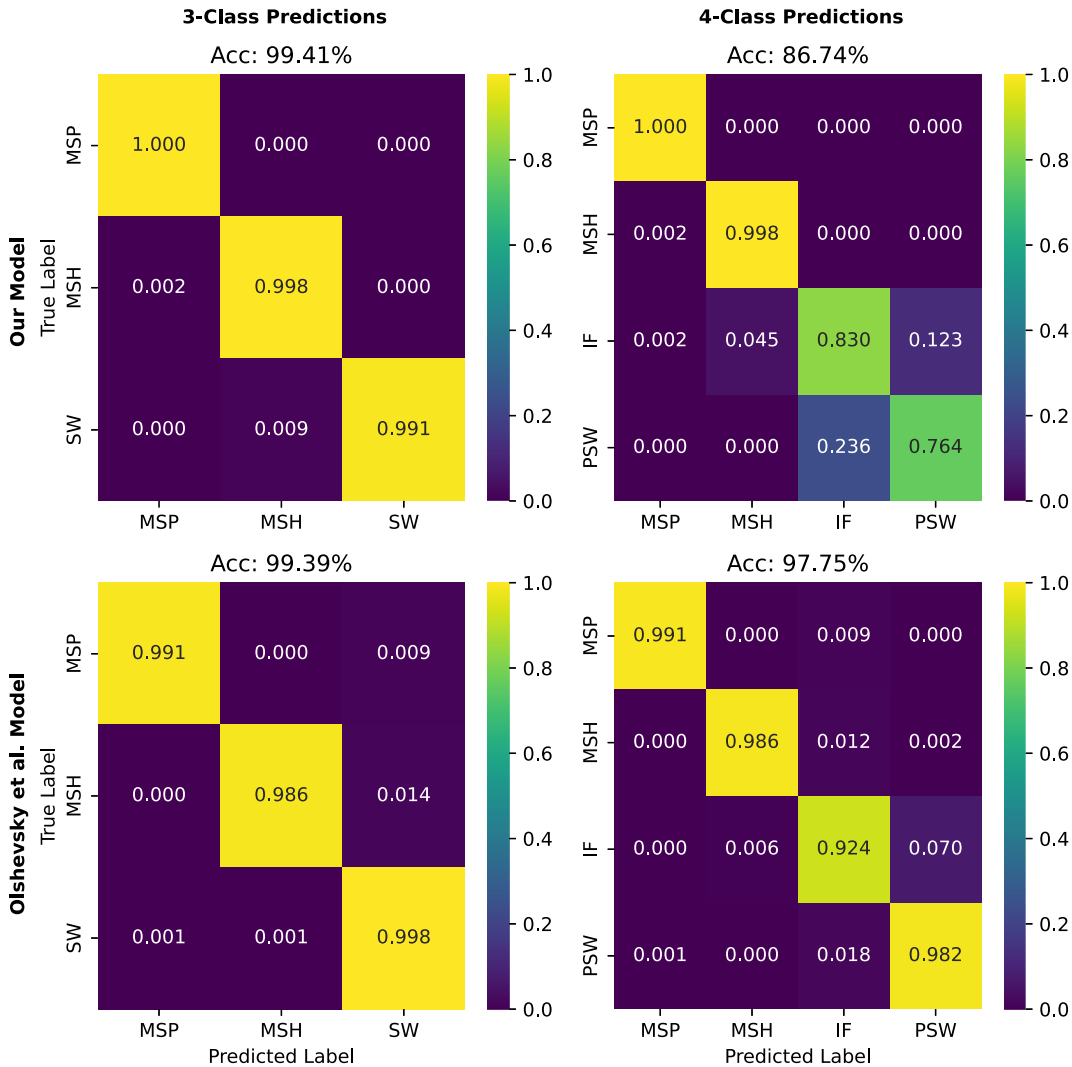

Figure 19. Confusion matrices for our model (top row) and the model of Olshevsky et al. (2021) (bottom row) against the labeled data set of Olshevsky et al. (2021) for magnetosphere (MSP), magnetosheath (MSH), and solar wind (SW) classifications (left column) and for magnetosphere, magnetosheath, ion foreshock (IF), and pristine solar wind (right column). Note that about 15% of their data set was labeled as being "Unknown" and these comparisons are done using only the remaining 85%.

Appendix A: Analysis of Topologically Distinct Nodes

The clustering of the nodes in the SOM as seen in the top left plot of Figure 8 is largely separated, but the topological overlap of classified nodes merits further investigation to reveal if the classification is correct or improper. We analyze the anomalous node positions, namely the separated magnetosheath nodes at positions (12,12), (8,12), and (6,2) as well as the vertical streak of magnetosphere nodes at positions (10,6), (10,5), and (10,4).

The node at (12,12) has the largest U-Matrix value seen in Figure 6, indicating that it is farther from its neighbors than all other nodes in the SOM. This is not surprising since it is classified as a magnetosheath node and is surrounded by solar wind-classified nodes. There are about 2.18 million magnetosheath points in the test set and 34k (1.5%) of them map to this node. Categorizing this node's data by spacecraft, we find that almost all are MMS observations with only about 100 belonging to THEMIS. We plot the empirical probability distributions of all magnetosheath and solar wind measurements in the test set in Figure A1 as well as the data belonging to this node for comparison. From the figure, we can see that there is much more overlap with the distributions of node (12,12) with the magnetosheath observations than that of solar wind, indicating that

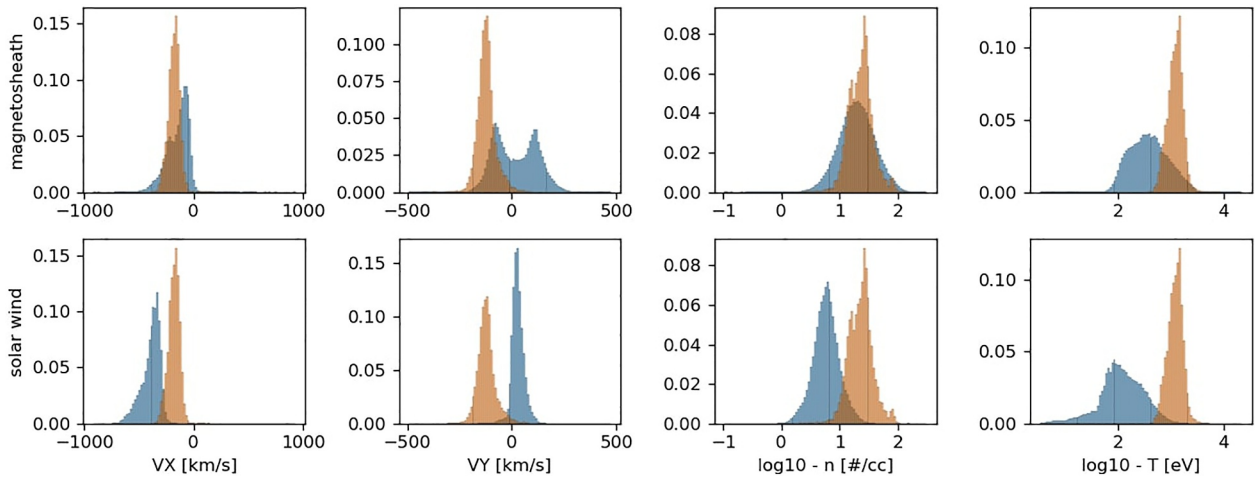

Figure A1. The VX, VY, log10 density, and log10 temperature empirical probability distributions of all magnetosheath-classified test data are plotted along the top row in blue. Similar features but for all solar wind-classified test data are plotted along the bottom row, also in blue. The empirical probability distribution of all test data that maps to node (12,12) is plotted in all plots as the orange distribution. The probability distributions are plotted here because of the large size differences between the number of magnetosheath observations (2.17 million) and solar wind observations (883k) of the test set and number of data mapping to node (12,12) (33k).

although the node's position in the grid is unusual, it corresponds well with magnetosheath observations. We believe this node's population being dominated by MMS observations is primarily due to the difference in time resolution between MMS and THEMIS observations. The MMS data set we prepared has a higher time resolution, 4.5 s resolution averaged down to 1 min, than the THEMIS data set, 1.5 min in higher temperature plasmas (magnetosphere, magnetosheath, and sometimes ion foreshock) and 6.5 min in colder plasmas (solar wind). The mode change is not done immediately upon crossing the bow shock but rather after consecutive observations showing higher/lower temperatures. Since these data are formed from either 32-spin/1.5 min (on the outbound passes where the spacecraft are going from the magnetosheath to the solar wind) or 128-spin/6.5 min (solar wind to magnetosheath) averages, the magnetosheath observations that correspond to this node could be more uncommon for THEMIS observations. These data are somewhat uncommon magnetosheath observations as seen in the node distributions relative to the distributions of the magnetosheath and solar wind classified data in Figure A1 in which they tend toward ends of the VY and temperature distributions. Overall, this uniqueness in MMS observations for this particular node could be due to the higher time resolution that is unavailable to THEMIS.

Node (6,2) is another topologically isolated magnetosheath node that also possesses a very high U-Matrix value, except that this one is surrounded by magnetosphere-classified nodes. It is responsible for only about 7.7k (0.35%) points of the magnetosheath-classified data of the test set and is almost evenly split by spacecraft with 56% points belonging to THEMIS and 44% to MMS. The empirical probability distributions of all magnetosheath-classified and magnetosphere-classified data in the test set are plotted alongside the observations mapped to this node in Figure A2 and multiple distinctions can immediately be made: data mapping to this node exhibit more magnetosheath characteristics in velocity, density, and temperature and also possess high magnetic field magnitudes. It seems correct that this node is classified as magnetosheath and the sparsity of points mapping to this node is understood in the context that magnetosheath observations possessing such large magnetic field magnitudes is relatively rare. The large U-Matrix value is justified with these observations.

Node (8,12) is diagonally topologically adjacent to the magnetosheath cluster but otherwise surround by solar wind nodes. This SOM uses a square topology, so this diagonal proximity does not factor into its U-Matrix value. It maps 68k (3.1%) points from the magnetosheath-classified data of the test set with 11% being THEMIS observations and 89% being MMS. The VX, VY, log10 temperature, and log10 density empirical probability distributions of the data mapping to this node are shown in Figure A3 alongside all magnetosheath-classified and solar wind-classified test data. They indicate magnetosheath observations with respect to the VX and VY distributions, but the log10 temperature and log10 density distributions somewhat resemble a blend of solar wind and

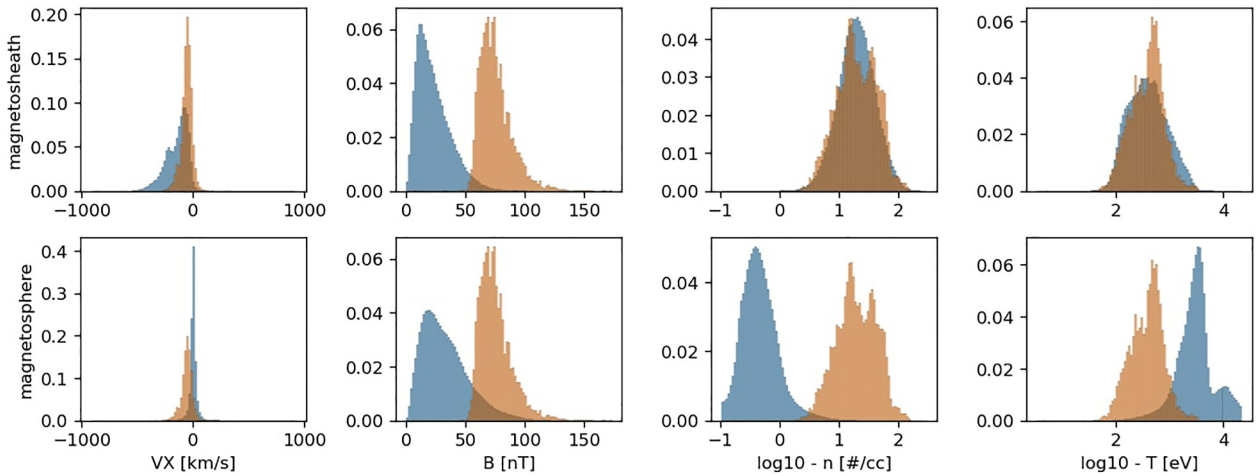

Figure A2. The VX, B, log10 density, and log10 temperature empirical probability distributions of all magnetosheath-classified test data are plotted along the top row in blue. Similar features but for all magnetosphere-classified test data are plotted along the bottom row, also in blue. The empirical probability distribution of the 7.7k magnetosheath-classified observations of node (6,2) are plotted in orange for each feature. The VX, log10 density, and log10 temperature distributions for this node all align more with the magnetosheath data than that classified as magnetosphere whereas the B distribution reflects high magnitude observations. Overall, this node has captured data with magnetosheath characteristics in velocity, density, and temperature, but also possessing high field magnitudes.

magnetosheath. This lack of uniform agreement across these features can explain why node (8,12) is adjacent to solar wind-classified nodes but the VX and VY distributions in particular indicate that it is correct to classify it as a magnetosheath node.

Lastly, we analyze the magnetosphere-classified nodes at positions (10,6), (10,5) and (10,4) that occur topologically within the magnetosheath cluster. Together, these nodes account for 46k (0.75%) of the 6.1 million magnetosphere-classified points of the test set with 76% being THEMIS observations and 24% belonging to MMS. Their VX, VY, log10 density and log10 temperature empirical probability distributions are plotted in Figure A4 along with the distributions of all three clusters in the test set. The data that map to these nodes are unusual in that the node distributions do not fully overlap with all of the distributions for any cluster. These data

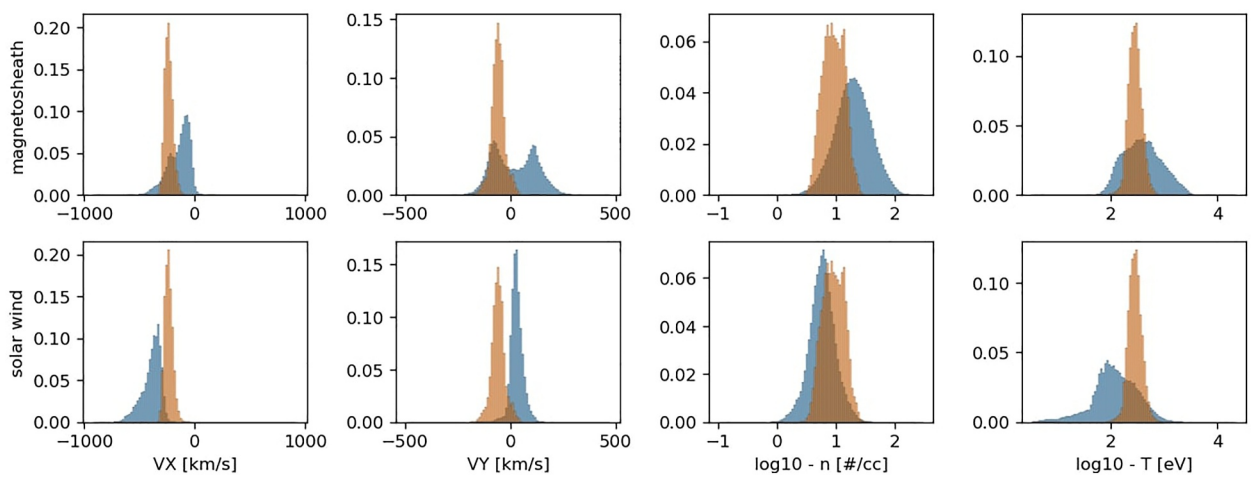

Figure A3. The VX, VY, log10 density and log10 temperature empirical probability distributions of all magnetosheath-classified data from the test set are plotted in blue along the top row. The solar wind-classified test data are plotted in blue along the bottom. The empirical probability distribution of the 68k magnetosheath-classified observations of node (8,12) are plotted in orange for each feature. The log10 density and log10 temperature distributions of the data from this node have sizable mixing between both magnetosheath and solar wind observations whereas the VX and VY distributions are more distinctly magnetosheath than solar wind.

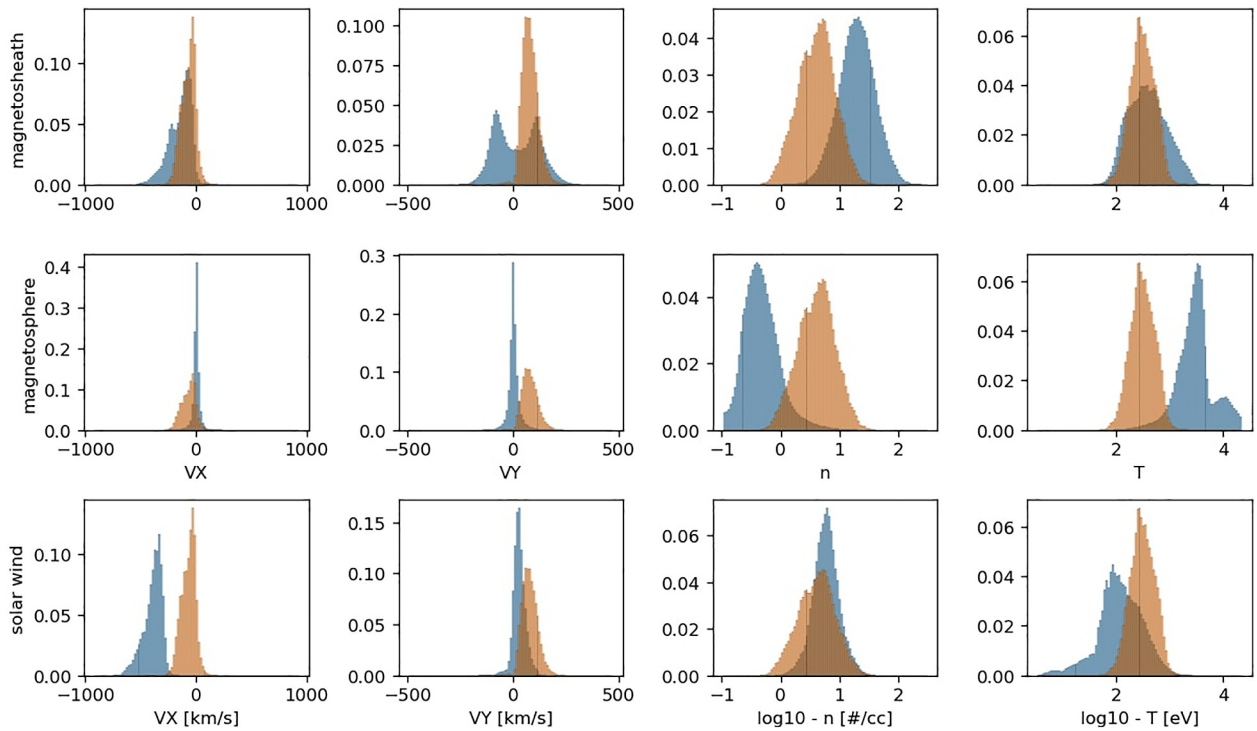

Figure A4. The VX, VY, log10 density and log10 temperature empirical probability distributions of all magnetosheath-, magnetosphere-, and solar wind-classified test data are plotted in blue along the top, middle, and bottom rows, respectively. All test data that map to nodes (10,6), (10,5) and (10,4) are collectively plotted here as the orange empirical probability distributions. These data are anomalous and exhibit characteristics found in all magnetosheath, magnetosphere, and solar wind observations. The VY, log10 density, and log10 temperature align well with the solar wind distributions, but the VX distribution is far too low. The VX, VY, and log10 temperature distributions correspond with magnetosheath observations, but there are very low densities. All of these distributions seem to have the least in common with the magnetosphere cluster, being along the extrema in all cases.

are classified as magnetosphere, but exist along the extrema of all the magnetosphere distributions shown. They resemble the VY, log10 density, and log10 temperature distributions of the solar wind, but the VX would be quite low for solar wind. The VX, VY, and log10 temperature distributions match up well with the magnetosheath distributions, but the log10 density is conspicuously low. Across all of the clusters, the measurements have much more in common with magnetosheath observations than magnetosphere or solar wind and are likely misclassifications. A time series of MMS1 observations containing many points that map to one of these nodes is shown in Figure A5. The magnetosheath plasma is of relatively low density, reflective of how these nodes are misclassified as magnetosphere. These nodes are responsible for 0.50% of the total test set.

Overall, the magnetosheath cluster has nodes in several aberrant positions in the SOM grid in which they were surrounded by nodes belonging to other clusters. Investigating these nodes in detail, however, has shown that the data correspond well with magnetosheath observations and are deserving of being classified as such. It was also seen that three magnetosphere-classified nodes are likely misclassified and should be recognized as magnetosheath. These three nodes contain few points (46k points, or 0.50% of the test set), together containing slightly less than the average number of test points per node (47k), and so do not significantly impact the strength of the results. Furthermore, it should be noted that such a misclassification occurred between the magnetosheath and the magnetosphere and that the separation between solar wind and magnetosphere plasma is quite distinct in the cluster solution seen in the top left of Figure 8.

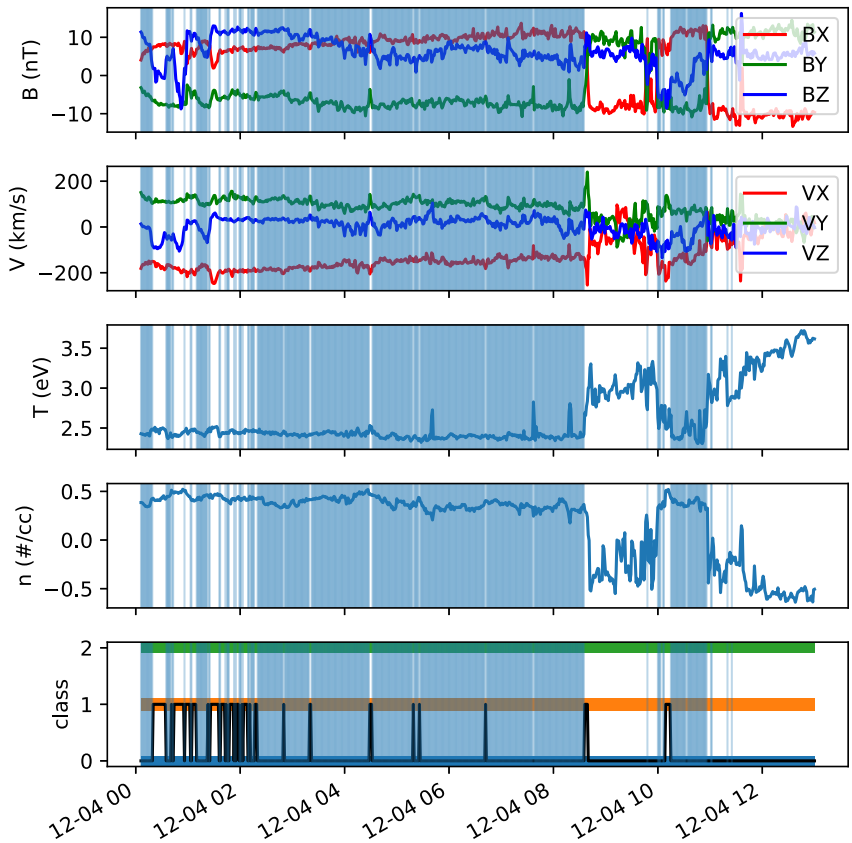

Figure A5. MMS 1 measurements from midnight to 13:00 UT on 2020-12-04. The plot structure is the same as Figure 11. The transparent vertical blue lines indicate that the measurement at that time maps to the (10,6), (10,5), or (10,4) node. MMS1 is measuring low-density magnetosheath plasma from midnight to 8:30 UT and from 10:00 to 11:00 UT. 473 points (84.3%) of the magnetosphere-classified data in the midnight to 11:00 UT interval map to one of these nodes. These 473 points are also almost 1% of all data that map to these nodes.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Data Availability Statement

The data set used for our model, the resulting crossings, and the MMS1 data set that we joined with the labels of Olshevsky et al. (2021) can be found in a Zenodo repository at Edmond et al. (2024a). The models, which have been serialized using Python's pickle module to allow them to be saved to and loaded from a hard disk, can be found in a separate repository at Edmond et al. (2024b). We have made a python package, GMClustering, that will easily make classifications and is pip-installable directly from its github repository at https://github.com/jae1018/GMClustering. It includes both an example python driver file and a small Jupyter notebook to showcase its use. Our modeling used various numerically-oriented python packages and we include the versions of those most relevant below.

- Numpy (Harris et al., 2020): 1.24.3
- Scikit-Learn (Pedregosa et al., 2011): 1.3.0
- XPySom (Mancini et al., 2020): 1.0.7
- Pandas (McKinney, 2010): 2.0.3
- SciPy (Virtanen et al., 2020): 1.11.1

Acknowledgments
The THEMIS data was downloaded and processed using the PySPEDAS python library (Grimes et al., 2022). The MMS data was downloaded and processed using the PyMMS python package (M. Argall et al., 2022). Maria Elena Innocenti acknowledges support from the Deutsche Forschungsgemeinschaft (German Science Foundation, DFG) within the Collaborative Research Center SFB1491 and within the DFG project 497938371. Work at UNH was supported through AFOSR Grant FA9550-18-1-0483 and from the NASA/THEMIS mission through subcontract SA405826326 from UC Berkeley.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining*.

Amaya, J., Dupuis, R., Innocenti, M. E., & Lapenta, G. (2020). Visualizing and interpreting unsupervised solar wind classifications. *Frontiers in Astronomy and Space Sciences*, *7*. https://doi.org/10.3389/fspas.2020.553207

Angelopoulos, V. (2008). The themis mission. *Space Science Reviews*, *141*(1), 5–34. https://doi.org/10.1007/s11214-008-9336-1

Angelopoulos, V. (2014). The Artemis mission. In C. Russell & V. Angelopoulos (Eds.), *The Artemis mission* (pp. 3–25). Springer New York. https://doi.org/10.1007/978-1-4614-9554-3_2

Angelopoulos, V., Kennel, C. F., Coroniti, F. V., Pellat, R., Kivelson, M. G., Walker, R. J., et al. (1994). Statistical characteristics of bursty bulk flow events. *Journal of Geophysical Research*, *99*(A11), 21257–21280. https://doi.org/10.1029/94JA01263

Argall, M., colinrsmall, & Petrik, M. (2022). argallmr/pymms: v0.4.6 (2022-05-19). *Zenodo*. https://doi.org/10.5281/zenodo.6564714

Argall, M. R., Small, C. R., Piatt, S., Breen, L., Petrik, M., Kokkonen, K., et al. (2020). Mms sitl ground loop: Automating the burst data selection process. *Frontiers in Astronomy and Space Sciences*, *7*. https://doi.org/10.3389/fspas.2020.00054

Auster, H. U., Glassmeier, K. H., Magnes, W., Aydogar, O., Baumjohann, W., Constantinescu, D., et al. (2008). The themis fluxgate magnetometer. *Space Science Reviews*, *141*(1), 235–264. https://doi.org/10.1007/s11214-008-9365-9

Breuillard, H., Dupuis, R., Retino, A., Le Contel, O., Amaya, J., & Lapenta, G. (2020). Automatic classification of plasma regions in near-earth space with supervised machine learning: Application to magnetospheric multi scale 2016–2019 observations. *Frontiers in Astronomy and Space Sciences*, *7*. https://doi.org/10.3389/fspas.2020.00055

Burch, J. L., Moore, T. E., Torbert, R. B., & Giles, B. L. (2016). Magnetospheric multiscale overview and science objectives. *Space Science Reviews*, *199*(1), 5–21. https://doi.org/10.1007/s11214-015-0164-9

Camporeale, E., Carè, A., & Borovsky, J. E. (2017). Classification of solar wind with machine learning. *Journal of Geophysical Research: Space Physics*, *122*(11), 10910–10920. https://doi.org/10.1002/2017JA024383

Chao, J., Wu, D., Lin, C.-H., Yang, Y.-H., Wang, X., Kessel, M., et al. (2002). Models for the size and shape of the earth's magnetopause and bow shock. In L.-H. Lyu (Ed.), *Space weather study using multipoint techniques* (Vol. 12, pp. 127–135). Pergamon. https://doi.org/10.1016/S0964-2749(02)80212-8

Cranmer, S. R., & Winebarger, A. R. (2019). The properties of the solar corona and its connection to the solar wind. *Annual Review of Astronomy and Astrophysics*, *57*(1), 157–187. https://doi.org/10.1146/annurev-astro-091918-104416

de Bodt, E., Cottrell, M., Letremy, P., & Verleysen, M. (2004). On the use of self-organizing maps to accelerate vector quantization. *Neurocomputing*, *56*, 187–203. https://doi.org/10.1016/j.neucom.2003.09.009

Edmond, J., Raeder, J., Ferdousi, B., Argall, M., & Innocenti, M. E. (2024a). Clustering of global magnetospheric observations. *Zenodo*. https://doi.org/10.5281/zenodo.10651397

Edmond, J., Raeder, J., Ferdousi, B., Argall, M., & Innocenti, M. E. (2024b). Clustering of global magnetospheric observations. *Zenodo*. https://doi.org/10.5281/zenodo.10651702

Fairfield, D. H., Iver, H. C., Desch, M. D., Szabo, A., Lazarus, A. J., & Aellig, M. R. (2001). The location of low mach number bow shocks at earth. *Journal of Geophysical Research*, *106*(A11), 25361–25376. https://doi.org/10.1029/2000JA000252

Gray, R. (1984). Vector quantization. *IEEE ASSP Magazine*, *1*(2), 4–29. https://doi.org/10.1109/MASSP.1984.1162229

Grimes, E. W., Harter, B., Hatzigeorgiu, N., Drozdov, A., Lewis, J. W., Angelopoulos, V., et al. (2022). The space physics environment data analysis system in python. *Frontiers in Astronomy and Space Sciences*, *9*. https://doi.org/10.3389/fspas.2022.1020815

Gringauz, K., Bezrukikh, V., Ozerov, V., & Rybchinskii, R. (1962). The study of interplanetary ionized gas, high-energy electrons and corpuscular radiation of the sun, employing three-electrode charged particle traps on the second soviet space rocket. *Planetary and Space Science*, *9*(3), 103–107. https://doi.org/10.1016/0032-0633(62)90180-0

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., et al. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Horn, F., Pack, R., & Rieger, M. (2020). The autofeat python library for automated feature engineering and selection. In P. Cellier & K. Driessens (Eds.), *Machine learning and knowledge discovery in databases* (pp. 111–120). Springer International Publishing.

Innocenti, M. E., Amaya, J., Raeder, J., Dupuis, R., Ferdousi, B., & Lapenta, G. (2021). Unsupervised classification of simulated magnetospheric regions. *Annales Geophysicae*, *39*(5), 861–881. https://doi.org/10.5194/angeo-39-861-2021

Jolliffe, I. (2011). Principal component analysis. In M. Lovric (Ed.), *International encyclopedia of statistical science* (pp. 1094–1096). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-04898-2_455

Kohler, U., & Luniak, M. (2005). Data inspection using biplots. *STATA Journal*, *5*(2), 208–223. https://doi.org/10.1177/1536867X0500500206

Köhne, S., Boella, E., & Innocenti, M. E. (2023). Unsupervised classification of fully kinetic simulations of plasmoid instability using self-organizing maps (SOMs). *Journal of Plasma Physics*, *89*(3), 895890301. https://doi.org/10.1017/S0022377823000454

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, *43*(1), 59–69. https://doi.org/10.1007/BF00337288

Kohonen, T. (2014). *Matlab implementations and applications of the self-organizing map*. Unigrafia Oy.

Liu, T. Z., Zhang, H., Turner, D., Vu, A., & Angelopoulos, V. (2022). Statistical study of favorable foreshock ion properties for the formation of hot flow anomalies and foreshock bubbles. *Journal of Geophysical Research: Space Physics*, *127*(8), e2022JA030273. https://doi.org/10.1029/2022JA030273

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, *28*(2), 129–137. https://doi.org/10.1109/TIT.1982.1056489

Lucek, E. A., Constantinescu, D., Goldstein, M. L., Pickett, J., Pinçon, J. L., Sahraoui, F., et al. (2005). The magnetosheath. *Space Science Reviews*, *118*(1), 95–152. https://doi.org/10.1007/s11214-005-3825-2

Mancini, R., Ritacco, A., Lanciano, G., & Cucinotta, T. (2020). Xpysom: High-performance self-organizing maps. In *2020 IEEE 32nd international symposium on computer architecture and high performance computing (SBAC-pad)* (pp. 209–216). https://doi.org/10.1109/SBAC-PAD49847.2020.00037

McFadden, J. P., Carlson, C. W., Larson, D., Ludlam, M., Abiad, R., Elliott, B., et al. (2008). The THEMIS ESA plasma instrument and in-flight calibration. *Space Science Reviews*, *141*(1), 277–302. https://doi.org/10.1007/s11214-008-9440-2

McKinney, W. (2010). Data structures for statistical computing in python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th python in science conference* (pp. 56–61). SciPy. https://doi.org/10.25080/Majora-92bf1922-00a

Neugebauer, M., & Snyder, C. W. (1962). Solar plasma experiment. *Science*, *138*(3545), 1095–1097. https://doi.org/10.1126/science.138.3545.1095.b

Nguyen, G., Aunai, N., Michotte de Welle, B., Jeandet, A., Lavraud, B., & Fontaine, D. (2022). Massive multi-mission statistical study and analytical modeling of the earth's magnetopause: 1. A gradient boosting based automatic detection of near-earth regions. *Journal of Geophysical Research: Space Physics*, *127*(1), e2021JA029773. https://doi.org/10.1029/2021JA029773

Nielsen, F. (2016). Hierarchical clustering. In *Introduction to HPC with MPI for data science* (pp. 195–211). Springer International Publishing. https://doi.org/10.1007/978-3-319-21903-5_8

Olshevsky, V., Khotyaintsev, Y. V., Lalti, A., Divin, A., Delzanno, G. L., Anderzén, S., et al. (2021). Automated classification of plasma regions using 3D particle energy distributions. *Journal of Geophysical Research: Space Physics*, *126*(10), e2021JA029620. https://doi.org/10.1029/2021JA029620

Omidi, N., Eastwood, J. P., & Sibeck, D. G. (2010). Foreshock bubbles and their global magnetospheric impacts. *Journal of Geophysical Research*, *115*(A6). https://doi.org/10.1029/2009JA014828

Omidi, N., Lee, S. H., Sibeck, D. G., Turner, D. L., Liu, T. Z., & Angelopoulos, V. (2020). Formation and topology of foreshock bubbles. *Journal of Geophysical Research: Space Physics*, *125*(9), e2020JA028058. https://doi.org/10.1029/2020JA028058

Omidi, N., & Sibeck, D. G. (2007). Formation of hot flow anomalies and solitary shocks. *Journal of Geophysical Research*, *112*(A1). https://doi.org/10.1029/2006JA011663

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.

Pitkänen, T., Chong, G. S., Hamrin, M., Kullen, A., Karlsson, T., Park, J.-S., et al. (2023). Statistical survey of magnetic forces associated with earthward bursty bulk flows measured by mms 2017–2021. *Journal of Geophysical Research: Space Physics*, *128*(5), e2022JA031094. https://doi.org/10.1029/2022JA031094

Pollock, C., Moore, T., Jacques, A., Burch, J., Gliese, U., Saito, Y., et al. (2016). Fast plasma investigation for magnetospheric multiscale. *Space Science Reviews*, *199*(1), 331–406. https://doi.org/10.1007/s11214-016-0245-4

Russell, C. T., Anderson, B. J., Baumjohann, W., Bromund, K. R., Dearborn, D., Fischer, D., et al. (2016). The magnetospheric multiscale magnetometers. *Space Science Reviews*, *199*(1), 189–256. https://doi.org/10.1007/s11214-014-0057-3

Schwartz, S. J., Chaloner, C. P., Christiansen, P. J., Coates, A. J., Hall, D. S., Johnstone, A. D., et al. (1985). An active current sheet in the solar wind. *Nature*, *318*(6043), 269–271. https://doi.org/10.1038/318269a0

Shue, J.-H., Song, P., Russell, C. T., Steinberg, J. T., Chao, J. K., Zastenker, G., et al. (1998). Magnetopause location under extreme solar wind conditions. *Journal of Geophysical Research*, *103*(A8), 17691–17700. https://doi.org/10.1029/98JA01103

Smith, A. W., Rae, I. J., Forsyth, C., Oliveira, D. M., Freeman, M. P., & Jackson, D. R. (2020). Probabilistic forecasts of storm sudden commencements from interplanetary shocks using machine learning. *Space Weather*, *18*(11), e2020SW002603. https://doi.org/10.1029/2020SW002603

Vettigli, G. (2018). Minisom: Minimalistic and numpy-based implementation of the self organizing map. Retrieved from https://github.com/JustGlowing/minisom/

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., et al., SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Willis, D. M. (1971). Structure of the magnetopause. *Reviews of Geophysics*, *9*(4), 953–985. https://doi.org/10.1029/RG009i004p00953

Wittek, P., Gao, S. C., Lim, I. S., & Zhao, L. (2017). somoclu: An efficient parallel library for self-organizing maps. *Journal of Statistical Software*, *78*(9), 1–21. https://doi.org/10.18637/jss.v078.i09