JGR Space Physics

DATA ARTICLE

10.1029/2024JA032431

Classifying 8 Years of MMS Dayside Plasma Regions via Unsupervised Machine Learning

Vicki Toy-Edens[1], Wenli Mo[1], Savvas Raptis[1], and Drew L. Turner[1]

[1]Johns Hopkins University Applied Physics Laboratory, Laurel, MD, USA

Key Points:

- We use an extensible unsupervised Gaussian mixture model (GMM) algorithm to automatically classify 8 years of dayside magnetospheric multiscale (MMS) plasma data
- Our model distinguishes between solar wind, ion foreshock, magnetosheath, and magnetosphere with 97.8% accuracy compared to manual labels
- We provide classified labels and transitions at 1-min resolution and stable region specific lists for all 8 years of dayside MMS data

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

V. Toy-Edens,
vicki.toy-edens@jhuapl.edu

Citation:

Toy-Edens, V., Mo, W., Raptis, S., & Turner, D. L. (2024). Classifying 8 years of MMS dayside plasma regions via unsupervised machine learning. *Journal of Geophysical Research: Space Physics*, *129*, e2024JA032431. https://doi.org/10.1029/2024JA032431

Received 11 JAN 2024
Accepted 6 JUN 2024

Abstract The Magnetospheric Multiscale (MMS) mission has probed Earth's magnetosphere, magnetosheath, and near-Earth solar wind for over 8 years. We utilize an unsupervised learning algorithm, Gaussian mixture model clustering, along with feature generation and simple post-cleaning methods to automatically classify 8 years of MMS dayside observations into four plasma regions (magnetosphere, magnetosheath, solar wind, and ion foreshock) at 1-min resolution. With these plasma regions distinguished, we have also identified boundary surfaces (e.g., magnetopause, bow shock). We validate our results on manually generated and rule based region labels described in the literature. We report overlap rates in our cluster determined magnetopauses and bow shocks against Scientist-in-the Loop (SITL) identified transitions and published databases. Our features are general and our model is extensible, potentially making it applicable to observational data from multiple other missions.

1. Introduction

In recent decades, large-scale heliophysics missions such as Magnetospheric Multiscale (MMS, Burch et al., 2015), Cluster (Escoubet et al., 2001), and Time History of Events and Macroscale Interactions during Substorms (THEMIS, Angelopoulos, 2008) have collected copious amounts of data from Earth's magnetosphere, magnetosheath, and near-Earth solar wind. Processing the data for scientific purposes requires identifying when the spacecraft is within a certain plasma region (e.g., magnetosphere, magnetosheath, solar wind). Though distinct regions are often identifiable by eye, there is no automated way to systematically identify plasma regions for in situ data with high certainty. While physical models have been developed to predict when transitional periods (e.g., bow shocks and magnetopauses) may occur, the actual transition depends on the measured values of the magnetic field vector, plasma moments (e.g., velocity, density, and temperature), and particles' energy distribution functions. Furthermore, it is challenging to distinguish these regions with a universal approach since these properties can vary based on the upstream solar wind conditions and the actual values change in time.

Thus far, the majority of efforts have been to identify and assign regions through empirical thresholds based on visual identification of different plasma regions (Chaston et al., 2013; Karlsson et al., 2021; Raptis et al., 2020), by using models of bow shock and magnetopause locations (Chapman & Cairns, 2003; Dimmock et al., 2017; Dimmock & Nykyri, 2013; Lin et al., 2010), or based on models and in situ measurements (Jelínek et al., 2012; Plaschke et al., 2013).

However, machine learning has caught on within the scientific community as a way to process and analyze the data in ways that are more nuanced than traditional methods. Many authors have employed machine learning methods to automate the identification of plasma regions. For example, Olshevsky et al. (2021) used a supervised neural network trained on 2 months's worth of MMS ion spectra data to predict the region identification of 2 years of MMS data. Lalti et al. (2022) then extended classifications with these models to 5 years of data. While Argall et al. (2020) and Breuillard et al. (2020) also employed a supervised neural network, they trained on Scientist-in-the-Loop (SITL) report labels instead.

Though supervised methods have yielded promising results, creating the training data to train these models is a non-trivial process. Generating enough data to properly train a supervised model is tedious; it typically requires manually labeling thousands of regions. Additionally, since data can change as a function of mission time (e.g., MMS changing orbits, seasons) and model performance may degrade if the model is only trained on narrow subsets of the data set, one must carefully select times that fully represent the entire data set. Improper or underrepresentative training can lead to unstable models that cannot be applied to future data or even prior data that is too different from the training data set. Furthermore, solar cycle and solar activity can drastically change the

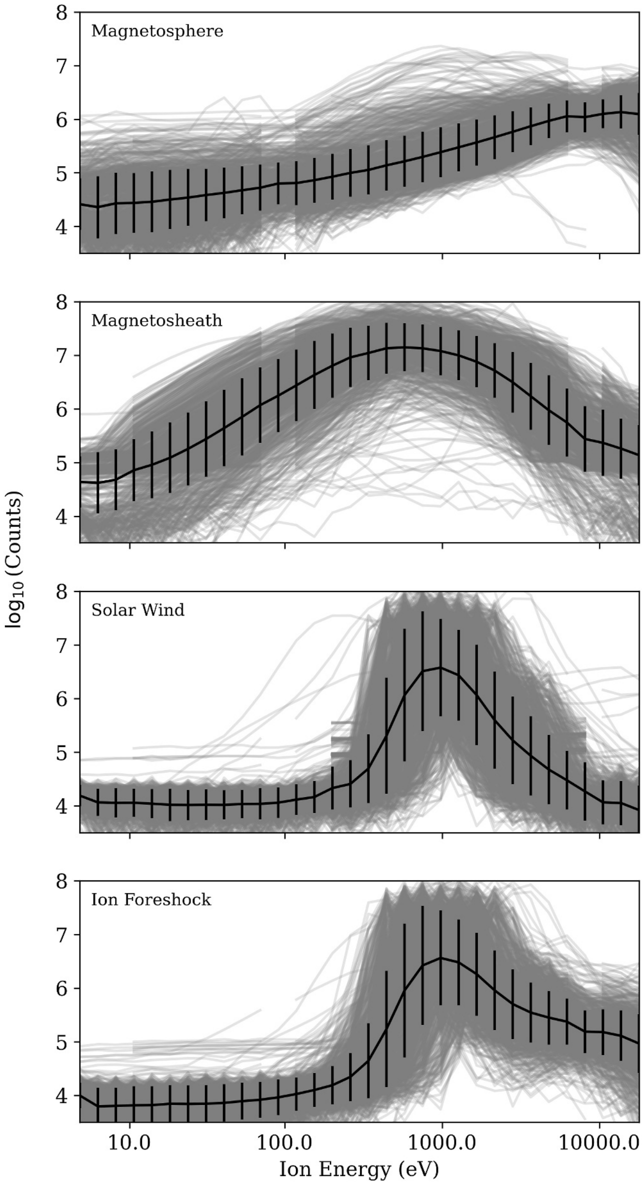

Figure 1. From top to bottom, the ion counts versus energy for different plasma regions (magnetosphere, magnetosheath, solar wind, and ion foreshock) are shown for a random sample of 1,000 min per plasma region in MMS3. Overlaid in black is the sample mean with 1 sigma errorbars. As shown, the plasma regions have distinct structures in a time-slice of the ion spectra. The magnetosphere has a linearly increasing signal, the magnetosheath has a single broadened intensity peak, and the solar wind and ion foreshock have a narrow dominant peak. Note that while the solar wind and ion foreshock look similar, there is an extended high intensity tail of high energy (>3,000 eV) ions in the ion foreshock whereas the solar wind follows a more symmetric distribution.

statistical properties of solar wind which affect the plasma environments close to Earth (Lepri et al., 2013; Richardson & Kasper, 2008; Xu & Borovsky, 2015). Koller et al. (2024) specifically mentions that machine learning methods trained exclusively during a specific solar cycle can have intrinsic biases that lead to unreliable results for other solar cycle phases. Some machine learning methods do not scale well with large data volumes. For example, MMS collects ~100 gigabits of mission data daily amongst its four probes combined. Despite this, there have been limited attempts to utilize unsupervised learning for plasma region labeling, focusing primarily on computer simulations (Innocenti et al., 2021).

In this work, we present an automated method for plasma region identification in MMS using an unsupervised clustering model. While we apply additional post-cleaning methods following the clustering, we refer to this method as unsupervised as the majority of labels are assigned by clustering. The model labels four distinct regions: magnetosphere, magnetosheath, solar wind, and ion foreshock. This method is computationally lightweight and does not require extensive hand-labeling of data for training. Features are generated using a series of ratios and sums from a combination of measurements that include the magnetic field, the total ion temperature, and the ion energy spectrogram, which can be applied easily to data from other missions with similar data products. We apply this model to identify plasma regions at a 1-min resolution for MMS dayside observations from September 2015 through 31 January 2024. Identified regions and raw transitions at 1-min resolution are provided as data products accessible by the reader. Additionally, we produce specific lists where the spacecraft is well within each of the four plasma regions to aid in region specific studies. In Section 2, our unsupervised clustering model is described in detail, including the data and engineered features used within the model. The results and validation of our identification method are presented in Section 3. Finally, we summarize and discuss our findings in Section 4.

2. Methods

We perform unsupervised learning via clustering to classify different plasma regimes in Earth's magntosphere region. Clustering methods are extensible and do not require intensive manually generated labels that are difficult to collect. In particular, clustering is well-suited for this type of multi-label classification because each plasma regime has a very distinct ion spectral signal (Figure 1) that SITL utilize to make decisions on what burst data to select. In order to generate distinct clusters that each represent a plasma regime, we engineer features that enhance characteristics between different plasma regimes. We utilize additional data (e.g., total magnitude of the magnetic field) besides ion spectra to help distinguish less clear cut plasma region observations discussed in depth in Sections 2.2 and 2.3.

2.1. Data

The MMS mission, a constellation of four identical satellites, launched on 13 March 2015 and started collecting scientific data in September 2015. For this study, we required Level 2 science data in fast survey mode from the Fast Plasma Investigation (FPI, Pollock et al., 2016) instrument, Fluxgate Magnetometers (FGM, Torbert et al., 2014), and the Magnetic Ephemeris and Coordinates (MEC, Henderson et al., 2022) data product.

The FPI instrument consisted of four dual electron (DES) and ion (DIS) spectrometers for each MMS probe. DIS consisted of 32 channels ranging from 10 eV to 30 keV with a temporal resolution of 150 ms. We used the ion omni-directional flux and temperature from the DIS fast survey moments data set which was reported on an integrated 4.5 s time cadence. FGM consists of analog and digital flux-gate magnetometers as part of the FIELDS instrument suite. DC measurements of the magnetic field are reported at a temporal resolution of 10 ms and range of 500 nT to 8,200–10,500 nT (for details, see Russell et al., 2014). We use the component magnetic field values from the FGM data set. The MEC data product is a part of the ephemerides released by the mission. We use the `epht89d` version of the ephemeris which calculates the ephemeris using the Tsyganenko (1989) magnetic field model for disturbed magnetospheric conditions. The MEC data in survey mode has temporal cadence of 30 s.

To join together the data from FPI, FGM, and MEC, we standardize the temporal cadence of the three data sets to 1-min resolution. For each 1-min time increment, we take the mean of all the values within that time window in order to find the typical value during this time period.

Mission data is publicly accessible through the MMS Science Data Center. We consider all data from September 2015 to 31 January 2024 for all four MMS probes. For this study, we only consider data on Earth's dayside. Therefore, we only include data where the spacecraft's ephemeris met the dayside criteria, $X > 0$, in the Geocentric Solar Ecliptic (GSE) coordinate system at 1-min resolution, and where FPI and FGM data were available. In total, we have 1,622 unique days with at least one dayside datapoint in at least one MMS probe and a total of 66,498 hr of dayside data across all four MMS spacecraft.

2.2. Feature Engineering

We generate three features (`ratio_max_width`, `ratio_high_low`, and `norm_Bt`) from ion spectra data and the total magnitude of the magnetic field, B_{tot}, to enhance distinguishable differences between the four plasma regions (magnetosphere, magnetosheath, solar wind, and ion foreshock) seen in Figure 1. We note that in addition to these three features, we include an indirect feature (pseudofeature) that alters other features based on its value (e.g. when the pseudofeature is >0.5 then set the other three features to 0), but is not explicitly included in an algorithm as a feature. We describe each feature and pseudofeature in depth in this section.

Our features rely heavily on peak fitting. We use the standard `scipy find_peak` algorithm to identify and fit all peaks in each 1-min averaged ion spectra log counts data. We call `scipy find_peak` with the following parameters - width = 1, height = 1, and prominence = 0.2 - which specify the minimum required values to identify a peak. For instance, the log counts peak must be at least 0.2 above the surrounding baseline to be identified as a peak. These parameters were selected based on manual inspection of ion spectra such as those seen in Figure 1 and low thresholds are specified for parameters like height in order to trigger the function to calculate and output these measured quantities. `scipy find_peak` returns the peak position, the height, and the interpolated positions of left and right intersection points of each peak. We define the width of these peaks as twice the distance between the left intersection point (low energy) and the peak position. We use this definition of width because the right intersection point (high energy) can be miscalculated due to large plateaus in energy like those seen in ion foreshock signatures in Figure 1. Our ion spectra features utilize the peak with the global maximum intensity.

The `ratio_max_width` feature is the ratio between the width of the most prominent peak in energy channel bins to the number of energy channels. In our case, this refers to the peak width of the 32 FPI DIS energy channels. This feature is similar to the thermal velocity or temperature, however, because we only use twice the left intersection point to the peak as the width we can eliminate high energy fluctuations that may cause confusion between ion foreshock and magnetosheath. The `ratio_high_low` feature is the ratio of the mean of the log intensity of energies >4,000 eV (`large_energy_mean`) to the mean of the log intensity of energies <100 eV (`small_energy_mean`). The `norm_Bt` feature is B_{tot} normalized to 50 nT, where anything above 50 nT is set to 1. We additionally create a pseudofeature to characterize magnetospheric plasmas by fitting a linear regression to the higher energy half of all ion energy spectra channels (>300 eV). If the Pearson correlation coefficient of the linear fit is > 0.7 or a peak is not detected, we set the other three features to zero to place this type of signature in a distinct location in feature space.

These three features, after they have been altered by the pseudofeature, are input into a clustering algorithm where each distinct cluster is assigned to a different plasma regime. Figure 2 shows an example ion spectra that passes through all four plasma regimes and displays how the features change in time in panels C–E.

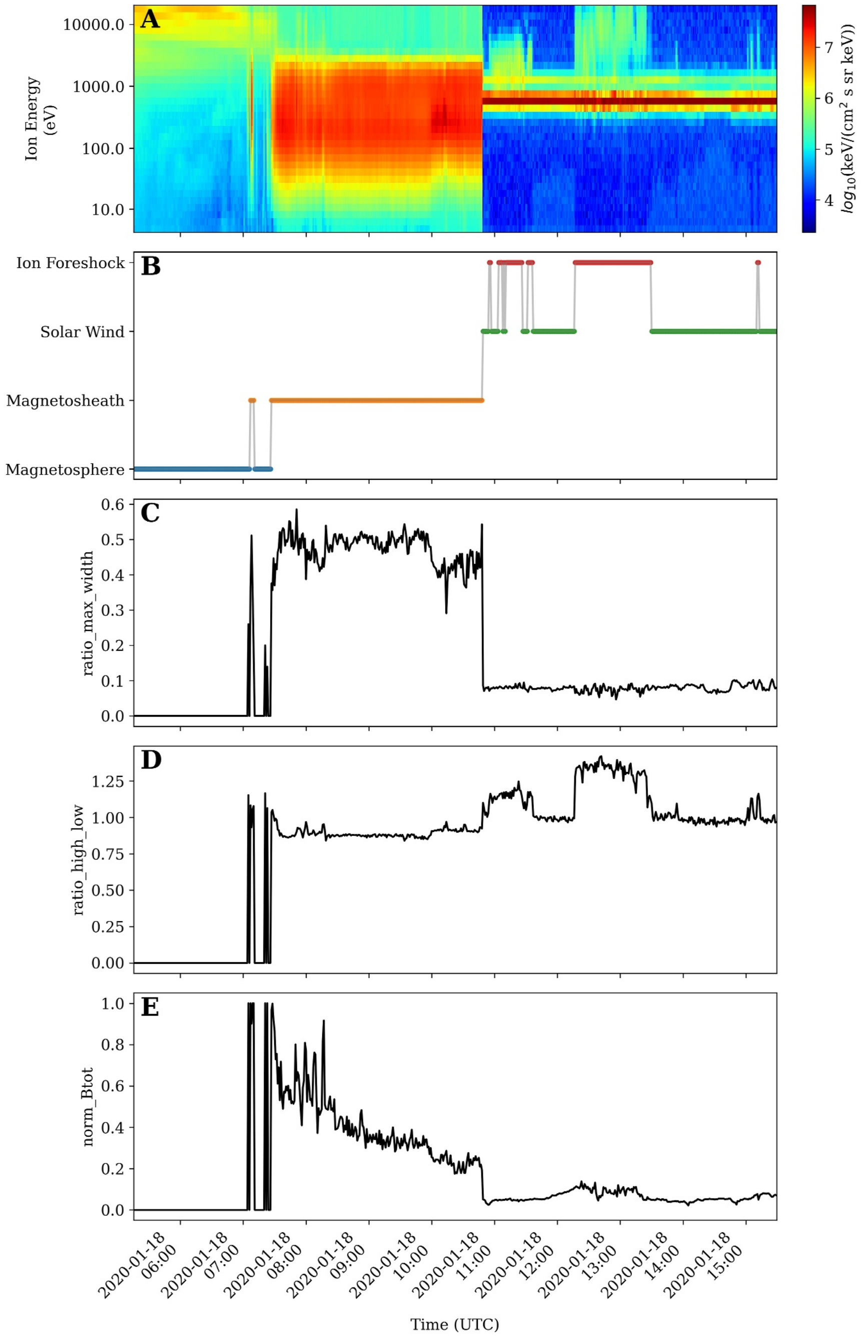

Figure 2. Example of clustering for 18 January 2020, chosen because it has clean and distinct coverage of the four plasma regimes: (a) Ion spectra rolled up to 1-min cadence, (b) Gaussian Mixture Model (GMM) labeled clusters, and (c–e) the features that go into the GMM cluster model. Note that for early times all the features are set to 0 due to a pseudofeature that corresponds to a linear regression r-value thresholding described in the text.

2.3. Clustering

We initially evaluated four different clustering methods: KMeans (Lloyd, 1982), Density Based Scan (DBScan; Ester et al., 1996), Hierarchical Density Based Scan (HDBScan; Campello et al., 2013), and Gaussian Mixture Model (GMM). All of these methods are unsupervised models that do not require labels. The common terminology for segmenting the data into distinct clusters is “training” as the model is learning how to separate the

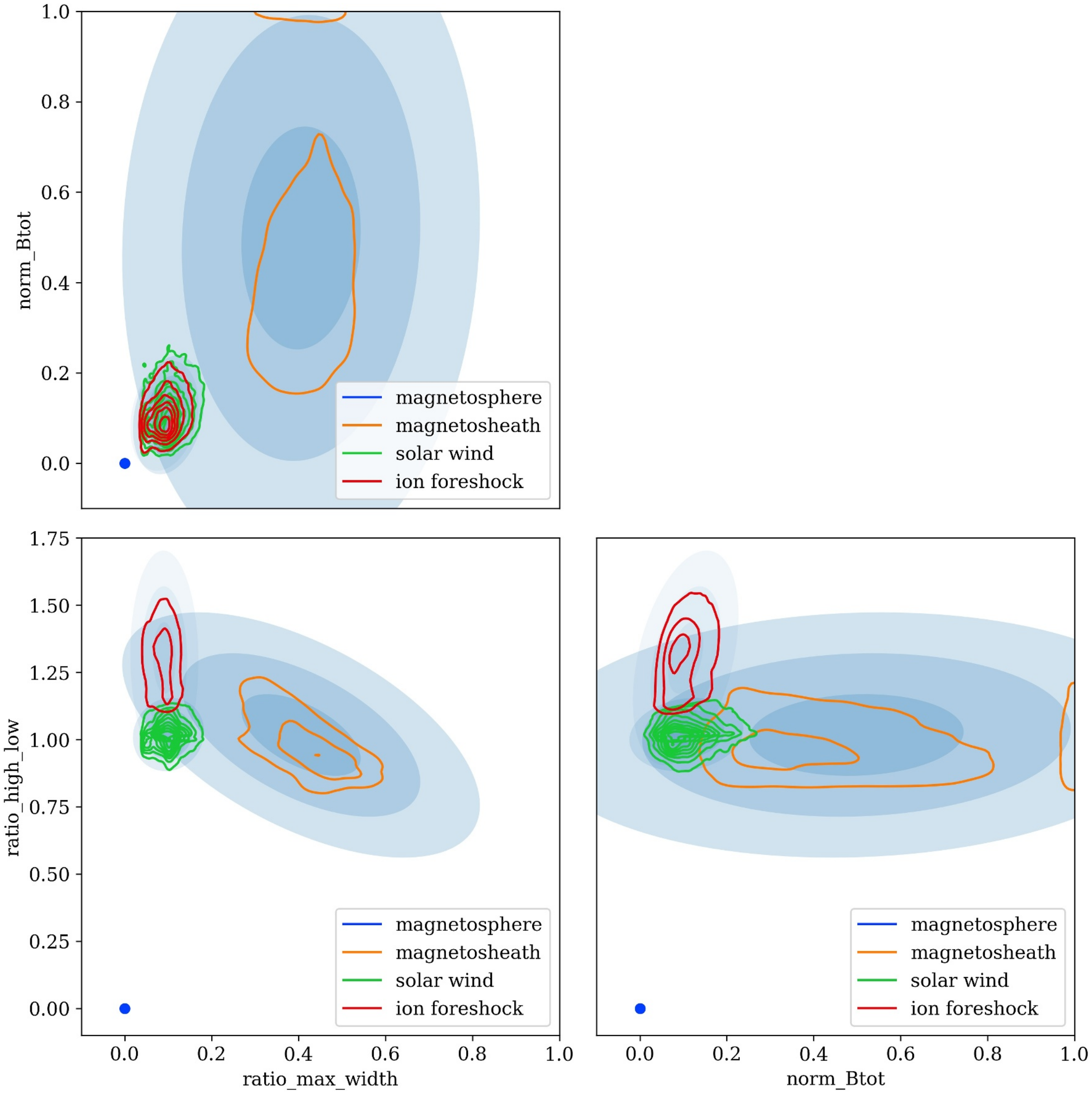

Figure 3. Gaussian mixture model clustering on all 1-min data dayside data. Contours are shown for the three features and 1-, 2-, and three-σ shown for each cluster. Note that there is a dense magnetosphere point for each feature because these are artificially set to 0 with a pseudofeature (see Section 2.2).

clusters and we use this terminology throughout this paper. We note that this is distinct from the use of training data in a supervised model where the training data require manual labels.

While DBScan was appealing because it is well-suited for scaling with larger data sets, it cannot be pre-trained on a previous data set and applied to new data; it must instead be applied to the full data set in order to generate clusters. For any future training, this requires having access to the full data set in order to get any cluster predictions. This is unsustainable for large data sets like MMS and the cluster order would change every time data was added; making predictions on new data cumbersome and intensive. HDBScan is a variant of DBScan that avoids this constraint and allows for an approximate prediction. However, it did not perform consistently on the same data set it was trained on, often giving entirely different plasma labels during the train stage compared to the prediction stage. Finally, KMeans assumes that clusters are spherical in nature whereas GMM assumes the clusters follow Gaussian distributions and, in general, GMM is computationally slower than KMeans. When we evaluated our generated feature space, we found that the features are more elliptical in nature in our feature space demonstrated in Figure 3 and we therefore select GMM over the faster KMeans.

Table 1
Relabeling Rules

Original label	Relabel rule	Resultant label
Magnetosheath	ratio_max_width ≤ 0.15	Solar Wind
Magnetosheath	(t_{tot} < 100 eV) and (ratio_max_width ≤ 0.25)	Next most probable label
Magnetosheath	$t_{\text{tot}} \geq$ 1,000 eV	Magnetosphere
Ion Foreshock	large_energy_mean < 4.6	Solar Wind

One consideration in our clustering model selection was the ability to set the number of clusters. In the majority of cases it is favorable to allow the algorithm to select the number of clusters, but we have a very clear scientific use case for manually selecting four clusters. Instead of utilizing clustering to find how many unique groups or clusters exist in a data set, we generate features that accentuate differences between plasma regimes so they are separated farther apart and can more easily be identified as distinct groups or clusters. We utilize a metric to select models that weighs model performance against model complexity (Bayesian Information Criteria - BIC) to verify that our resultant clustering model is optimized for four clusters. A plot displaying the BIC values for different number of clusters can be found in Figure S1.

In order to train the model, we first generate the features described in Section 2.2 on the data set. We utilize all 1-min data for each MMS probe that is in dayside (i.e., $X_{\text{GSE}} > 0$ during fast survey periods), a total of 66,498 hr of data. We perform this analysis on the Amazon Web Services (AWS) cloud using the Heliocloud framework (Thomas et al., 2022), which provided computational tools and cloud object storage of MMS data. Then, we train a GMM model (sklearn GaussianMixture) with this data and are left with a lightweight clustering model that can be applied to new data.

Figure 3 shows the distribution of clusters in feature space after performing the clustering. There is some overlap between the solar wind cluster and the ion foreshock cluster, which is expected since they are not two entirely unique regimes isolated by a well-defined and narrow boundary layer. There is also some less significant overlap between the solar wind and magnetosheath. This is to be somewhat expected again since the magnetosheath consists of shocked solar wind plasma.

We evaluated the initial GMM quality through manual review of plasma labels of hundreds of randomly selected dayside days and finally with validation data described in Section 3.2. Figure 2 shows an example of the cluster labels (panel B) for a time period that displays all four plasma regimes.

2.4. Post-Cleaning

We modify the cluster labels to correct for spurious cluster labels. In a sequence of five adjacent classifications (5-min time window), if the central datapoint is classified differently than the other four datapoints, we consider the central datapoint spurious and clean it to match the classification of the other datapoints. However, we do not modify the central datapoint if it is classified as magnetosheath while the remainder are magnetosphere, or vice versa, because this is consistent with the expected physical behavior of boundary layers. In this case, the central datapoint retains its original label and we tag the transition as "boundary layer."

After visually inspecting samples of our labeled plasma regions, we additionally clean for systematic over-prediction of one label over another (e.g., magnetosheath over ion foreshock). We inspect the frequency distribution of a particular feature for two clusters in question, inspect where the distribution appears bimodal, and institute a threshold to correct for confusion between the two labels. The rules are based on a combination of total temperature, ratio_max_width, and large_energy_mean, defined as the mean of the log intensity for energies >4,000 eV used in ratio_high_low. The total temperature is defined as

$$t_{\text{tot}} = \frac{2t_{\text{perp}} + t_{\text{para}}}{3}, \tag{1}$$

where t_{perp} and t_{para} are the perpendicular and parallel temperature components, with respect to the local magnetic field, from DIS moments data. We also use the posterior probability, predict_proba, for each prediction provided by the GMM clustering algorithm. This probability can effectively be considered the probability that the datapoint should be classified as each of the four labels.

Table 2
Clustering Results: Cleaned Labels

Plasma regime	No. of labels[a]
Magnetosheath	1,499,968 (37.6%)
Solar Wind	1,285,244 (32.2%)
Magnetosphere	730,408 (18.3%)
Ion Foreshock	474,294 (11.9%)

[a] 1-minute cadence for all four MMS spacecraft.

We institute three rules to correct for magnetosheath misclassifications over other labels and one to correct for ion foreshock misclassifications over solar wind. These relabeling rules are enabled only for a particular initial cluster label and are shown in Table 1. The only use of posterior probability in post-cleaning is when a datapoint is labeled magnetosheath but has low total temperature and moderate `ratio_max_width`, we relabel this datapoint to the classification with the next highest posterior probability.

Overall, these modifications alter only 6.8% of the data. We briefly note the difference in accuracy between the raw clustering labels and the post-cleaned labels on the validation data in Section 3.2.1. This is an improvement of solely rule-based labeling because we only modify labels that have a particular starting label, reducing the number of affected label modifications that may be on the cusp of a threshold.

2.5. Transitions

When the plasma regime is different than the preceding minute in a continuous set of epochs, it is tagged as a transition. The transition names are the same regardless of the direction (i.e., magnetosheath to magnetosphere and magnetosphere to magnetosheath are both labeled as magnetopause). The six different types of transition layers employed here are:

1. Foreshock compressional boundary: between solar wind and ion foreshock (Omidi et al., 2009).
2. Quasi-perpendicular bow shock: between solar wind and magnetosheath.
3. Quasi-parallel bow shock: between ion foreshock and magnetosheath.
4. Magnetopause: between magnetosheath and magnetosphere.
5. Potential ion foreshock transient: single central point of magnetosheath in ion foreshock/solar wind or vice versa in 5 min time window.
6. Potential boundary layer: single central point of magnetosheath in magnetosphere or vice versa in 5 min time window.

One caveat to the clustering is that slow or complex quasi-perpendicular bow shocks have the potential to reflect ions and mimic ion foreshock which may potentially be miscategorized as a quasi-parallel bow shock. We expect this to be a rare occurrence and have not encountered an example through our manual review of the data.

The potential ion foreshock transient and potential boundary layer represent unresolvable/anomalous transitions and are a non-exhaustive list. We only note these regions of interest and leave it up to the reader to validate the nature of these flagged transitions. We also have an additional "unphysical" transition which represents non-physical transitions like solar wind or ion foreshock to magnetosphere. We first tag the transition as unphysical and then compare against the previous cluster label until a physical transition or the same plasma regime is found.

Table 3
Clustering Results: Transitions

Transition name	No. of transitions[a]
Foreshock Compressional Boundary	69,068
Magnetopause	49,580
Quasi-Parallel Bow Shock	13,491
Quasi-Perpendicular Bow Shock	12,730
Potential Ion Foreshock Transient	567
Potential Boundary Layer	45
Unphysical	7,940

[a] 1-minute cadence for all four MMS spacecraft.

3. Results

3.1. Data Set

We present 3,989,914 labeled plasma region 1-min epochs across all MMS spacecraft (∼1,000,000 labeled minutes per probe). Our publicly available data (see Open Research (Data Availability Statement)) includes both the raw GMM produced labels and the post-cleaned cluster labels (see 2.3 for details) as well as all labeled transitions. In addition, we include spacecraft ephemeris (in GSE and GSM coordinates) and the calculated clustering features. Tables 2 and 3 include the breakdown of plasma regimes and transitions for the 8 years data set. The majority of observations are of magnetosheath plasma, which is expected considering the orbital history and mechanics plus the prioritization for the science region-of-interest along the dayside magnetopause. Still, our

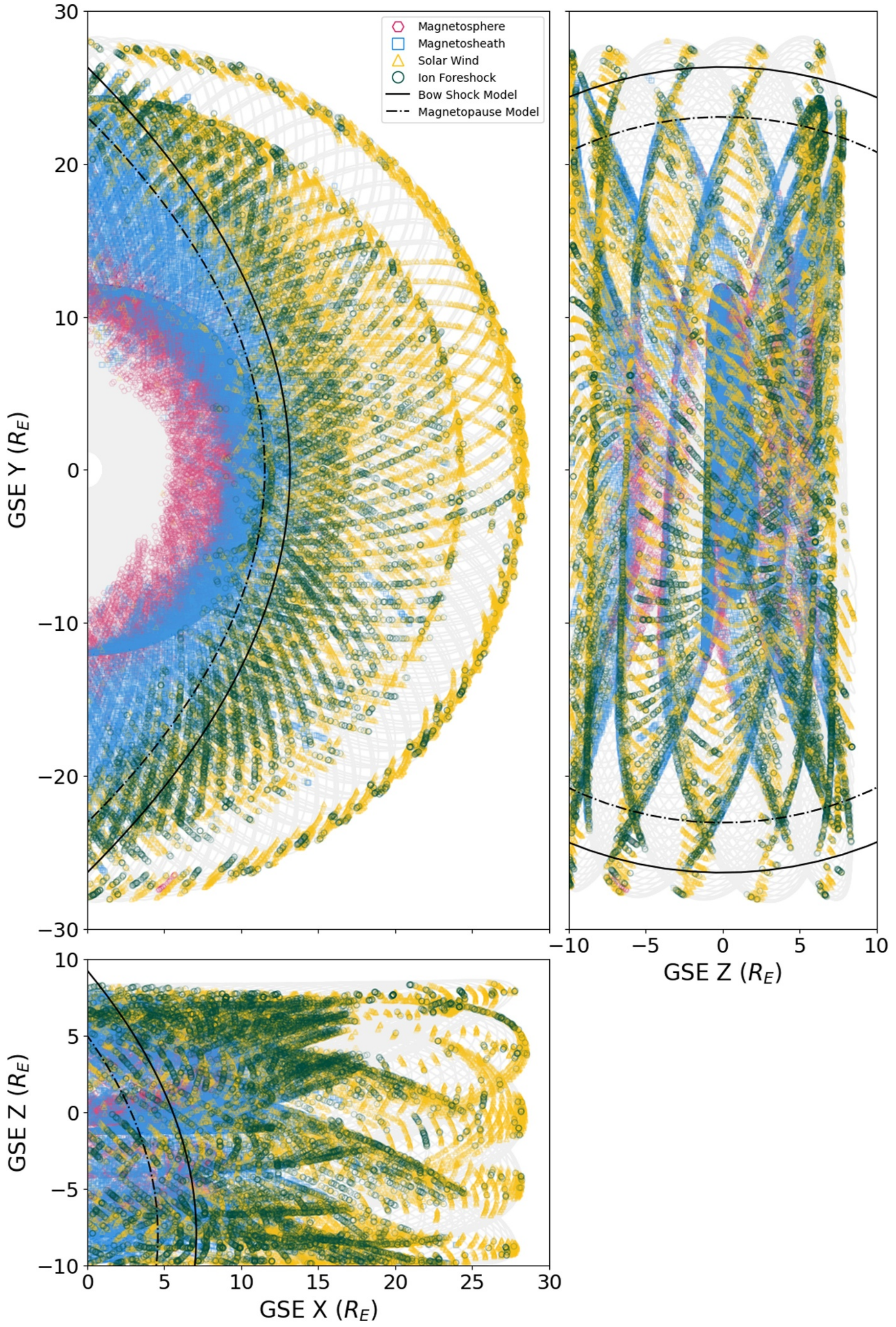

Figure 4. Plasma regions labeled by Gaussian mixture model clustering in the geocentric solar ecliptic coordinate plane. The gray lines represent the magnetospheric multiscale (MMS) orbit trajectories and the black lines are cross-sections of the bow shock and magnetopause models from Jelínek et al. (2012), assuming a typical solar wind pressure of 2 nPa, to help guide the eye. There is a gap between ∼23–28 R_E in the XY projections due to preferential orbit selection for MMS science regions-of-interest.

data set contains significant portions of each type of plasma regime. This data set can be used for large statistical studies of magnetopauses or bow shocks as well as single plasma region studies. To aid single plasma region studies, we also provide four plasma region lists (see Section 5) that specify times when the spacecraft is within each plasma region (magnetosphere, magnetosheath, solar wind, and ion foreshock) for 15 min or longer.

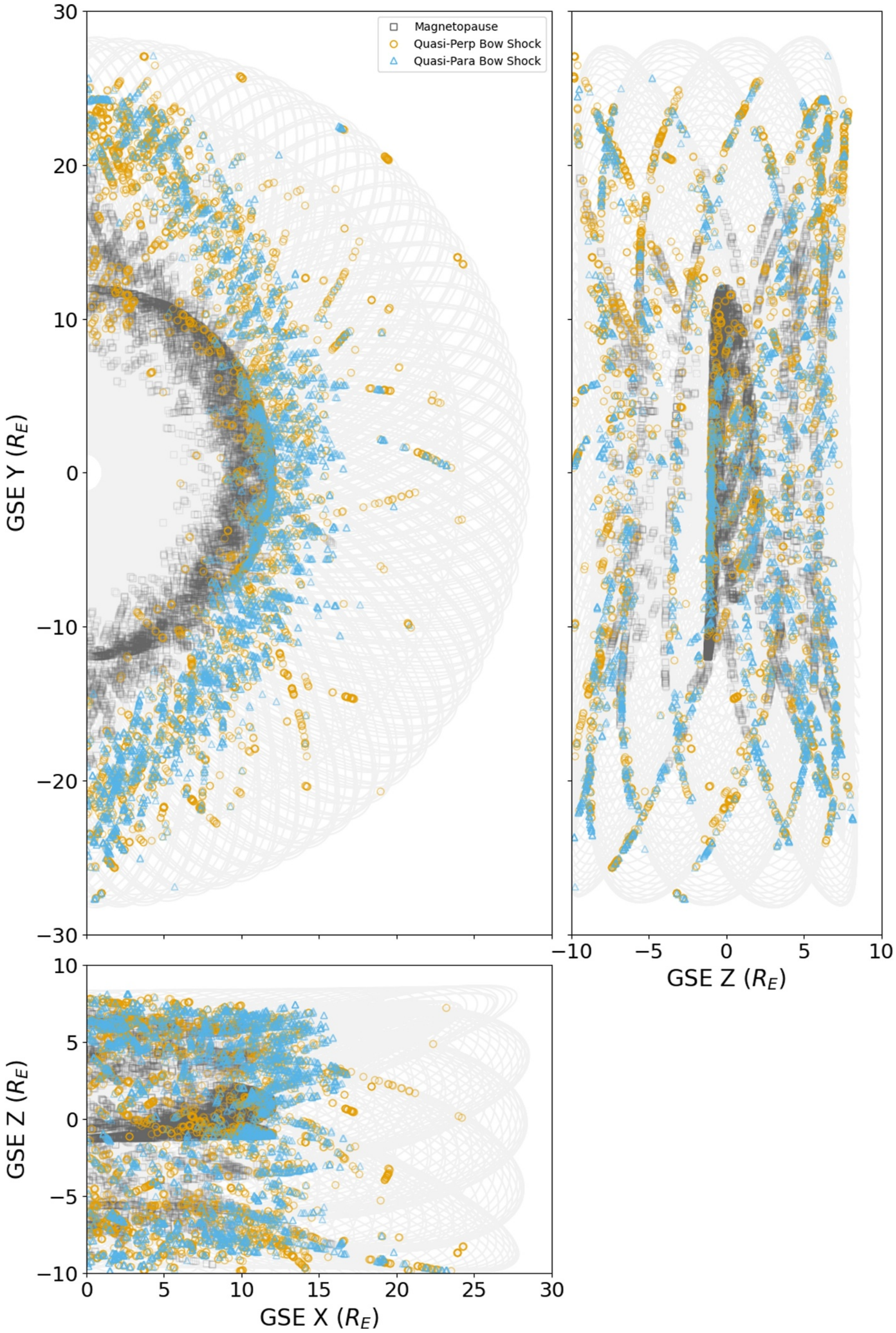

Figure 5. Bow shocks and magnetopauses labeled by Gaussian mixture model clustering in the geocentric solar ecliptic coordinate plane. The gray lines represent the magnetospheric multiscale (MMS) orbit trajectories. The dark band at ∼12 R_E occurs because there is an abundance of MMS orbits that only extend to ∼12 R_E during the first several years of the mission rather than there being a hard edge to the magnetopause transitions.

Figures 4 and 5 display the distribution of plasma regions and transitions in GSE coordinate space respectively. These align well with the expected plasma regime positions and bow shock/magnetopause transitions, which is the focus of a follow-on study. These figures capture some physically meaningful aspects such the statistical ranges of bow shock and magnetopause locations. The bow shock locations span across a large range of X_{GSE} in Earth radii (R_E) as expected due to the variable solar wind conditions that the MMS spacecraft encountered across

Table 4
Comparison of Region Label Results of Olshevsky et al. (2021) to Clustering

	GMM magnetosphere	GMM magnetosheath	GMM solar wind	GMM ion foreshock
Manual Magnetosphere	99.9% (2,801)	0.0% (1)	0.0% (0)	0.0% (0)
Manual Magnetosheath	0.1% (4)	99.5% (8,640)	0.9% (124)	0.3% (10)
Manual Solar Wind	0.0% (0)	0.0% (0)	97.0% (14,091)	5.0% (160)
Manual Ion Foreshock	0.0% (0)	0.5% (45)	2.1% (307)	94.7% (3,059)

a broad range of the solar cycle. We leave the investigation of how the GMM method performs in different solar conditions for future work. Note that these figures also display the orbital effects of the MMS mission and how it changed over the years. For example, there is boundary at ∼12 R_E in both the region and transition plot not because there is a hard edge to the magnetopause transitions, but instead because there is an abundance of MMS orbits that only extend to ∼12 R_E during the first several years of the mission. Additional patterns emerge due to the abundance or lack of availability of fast survey data from prescribed selection of the MMS science region-of-interest. A prime example of this is the gap of data points between approximately 23 and 28 R_E geocentric distances in the XY projections in Figure 4.

Though difficult to see, we find that ion foreshock observations are more prevalent closer to the dawn and dusk local time sectors (in Magnetic Local Time—MLT). In order to measure this we look at the ratio of the occurrences of ion foreshock to the all upstream occurrences (solar wind and ion foreshock). For dawn ($6 \geq \text{MLT} \leq 9$) and dusk ($15 \geq \text{MLT} \leq 18$) this ratio is 30.4% and for subsolar ($9 < \text{MLT} < 18$) this ratio is 24.3%. The ion foreshock increase towards dawn/dusk is consistent with statistics of interplanetary magnetic field (IMF) orientation and the nature of the Parker spiral (or ortho-spiral) solar wind at 1 AU (Borovsky, 2010).

3.2. Validation

In order to validate our unsupervised clustering model, we compare against manual labels, rule based labels, SITL reports of magnetopause and bow shock transitions, and databases of magnetopause and bow shock transitions.

3.2.1. Comparing to Manually Labeling

Olshevsky et al. (2021) hand labeled every datapoint taken from FPI in fast resolution (i.e., 4.5 s) for MMS1 in November and December 2017 into five categories: magnetosphere, magnetosheath, solar wind, ion foreshock, and unknown. We compare our clustering labels to their manual labels in Table 4 where we exclude data with the unknown label. We rebin the epochs in this validation data set to 1-min windows and assign the mode label within the window. We have 97.8% agreement with the non-unknown labels where our largest discrepancy coming from mislabeling solar wind as ion foreshock and vice versa. For reference, there is a 94.3% agreement with the manually labeled data if we instead use the raw clustering labels opposed to the post-cleaned labels.

3.2.2. Comparing to Rule Based Labeling

Raptis et al. (2020) identified magnetosheath and solar wind plasma regions using manually chosen thresholds for ion number density, velocity, temperature, and differential energy flux of high-energy ions for September 2015 through June 2020 at 4.5 s resolution. We rebin the epochs in this data set to 1-min windows and assign the mode label within the 1-min time interval. When comparing the labels, we use the GMM ion foreshock and solar wind labels interchangeably as a solar wind label because the authors made no distinction between ion foreshock and solar wind. We have 88.0% agreement (Table 5). From manual inspection it appears that the rule based identification often underperforms due to the lack of flexibility in rule based classification as plasma conditions change.

Table 5
Comparison of Region Label Results of Raptis et al. (2020) to Clustering

	GMM magnetosheath	GMM solar wind	GMM ion foreshock
Threshold Magnetosheath	90.8% (205,723)	21.5% (21,183)	3.2% (1,151)
Threshold Solar Wind	9.2% (20,926)	78.5% (77,566)	96.8% (34,930)

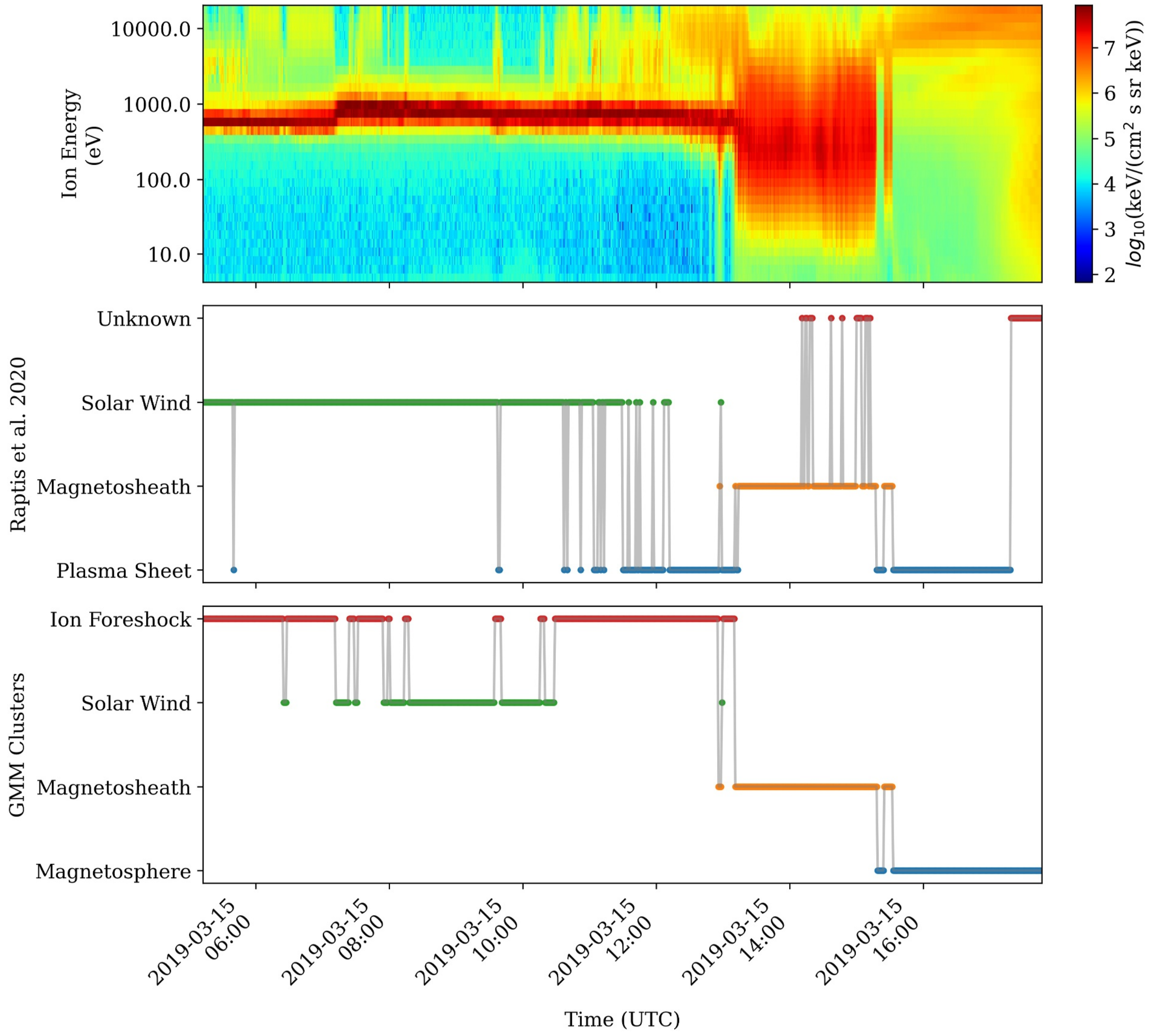

Figure 6. An example showing the ion spectra (top) of a disagreement between Raptis et al. (2020) rule based labeling (middle) and Gaussian mixture model clusters (bottom). The rule based approach identifies ion foreshock regions as magnetospheric around 10:00 UTC. This originates from a threshold criterion that magnetospheric plasma has a flux of high-energy ions (1–10 keV) significantly larger than the medium energy ranges (<1 keV). However, as we see in this example (between 12:00–13:00), the ion foreshock region can be populated with very high-energy ions causing the rule based approach to produce wrongly classified regions.

For instance, this rule based identification does not take into account varying velocity and density from solar wind variability and cannot adapt when there is high variation in high energy ions due to a significant foreshock population (e.g., 12:00 UT in Figure 6).

3.2.3. Comparing to SITL Reports

We extract SITL reports and identify freeform text descriptions relating to bow shocks or magnetopause with regular expressions. We get the time range of the SITL selected data and merge together any directly adjacent time regions: often SITLs label large chunks of time to identify similar behavior across a large time period. This means that we often include multiple bow shocks or magnetopauses within the same time period. For our comparison, we make no distinction based on how many bow shock crossings are specified in the description text.

We introduce a measure of how well the clustered plasma regions align with SITL reports based on the bow shock (or magnetopause) identification within the SITL time periods. A 100% match means the clustering model correctly identified every single SITL identified bow shock (or magnetopause). For example, a 60% match rate

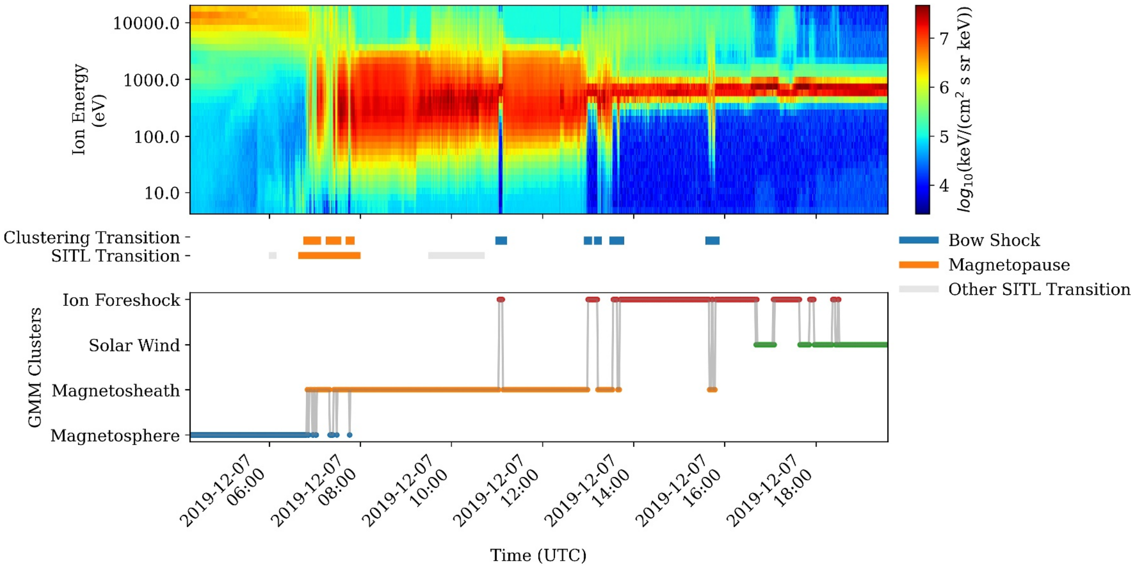

Figure 7. The ion spectra (top) and Gaussian mixture model (GMM) clusters (bottom) are shown for all the dayside observations of MMS1 on 7 December 2019. The GMM clustering transitions are overlaid with the SITL identified transitions (middle). The SITL transitions are found by identifying regex matches like shock, BS, MP, magnetopause, etc. and mapping them to bow shock (blue) and magnetopause (orange). We additionally display any other SITL identified epoch with gray in the event the SITL report did not match any of our regex searches. Note the region around 11:00 UT with two bow shock transitions identified by the GMM clusters but not noted in SITL reports and the additional bow shocks after 13:00 UT that all went unidentified by the SITL.

means that 40% of the SITL identified bow shocks (or magnetopauses) were not identified via clustering. We have a 75.1% match rate when we compare how well we identify any bow shock within the SITL time periods and a 76.0% match rate for magnetopause transitions.

We also include the complement of this measurement in order to indicate how many additional bow shocks (or magnetopauses) the clustering method identifies that are not included in SITL reports. A 100% match means that every single cluster identified bow shock is in SITL reports. For example, a 60% match means that 40% of the cluster identified bow shocks (or magnetopauses) were not identified in SITL reports. We have a 55.0% match rate when we compare how many of the SITL bow shocks are included in the GMM identified list of bow shock transitions and a 35.1% match rate for magnetopause transitions.

Manual review of the unmatched transitions indicate the issues can be attributed to human error, very unusual conditions (e.g., coronal mass ejections), or transient phenomena (e.g., ion foreshock transients) that can cause rapidly changing alternations between cluster plasma regions (Figure 7).

3.2.4. Comparing to Transitions in Literature

Magnetopause and bow shock transitions are decided by a slew of different factors that are often biased to the type of scientific problem they are aimed at studying. We present here an automated method for determining transitions and compare to manually curated transition lists from literature.

Paschmann et al. (2018) presented a database of magnetopause transitions, 3,735 of which are flagged as complete or Harris-like (see Harris, 1962) magnetopause crossings within our dayside data epochs. We identify a match if our cluster magnetopause transition, with a 1 minute buffer to either side, is within the crossing time period in magnetopause database. We find that 67.4% of the Paschmann et al. (2018) magnetopause crossings are identified by our clustering method. Manual inspection of magnetopause disagreements shows that often mismatches occur because (a) our 1-min resolution is coarser than the transition or (b) it is either a boundary layer or region where the spacecraft is rapidly moving back and forth across the transition. In the second case, we typically label at least one magnetopause transition during the time period, but are unable to distinguish every single magnetopause (see Figure 8).

Lalti et al. (2022) present a database of bow shock transitions, 2,418 of which overlap with our dayside data epochs. We identify a match if our cluster bow shock transition, with a 1 minute buffer to either side, is within the

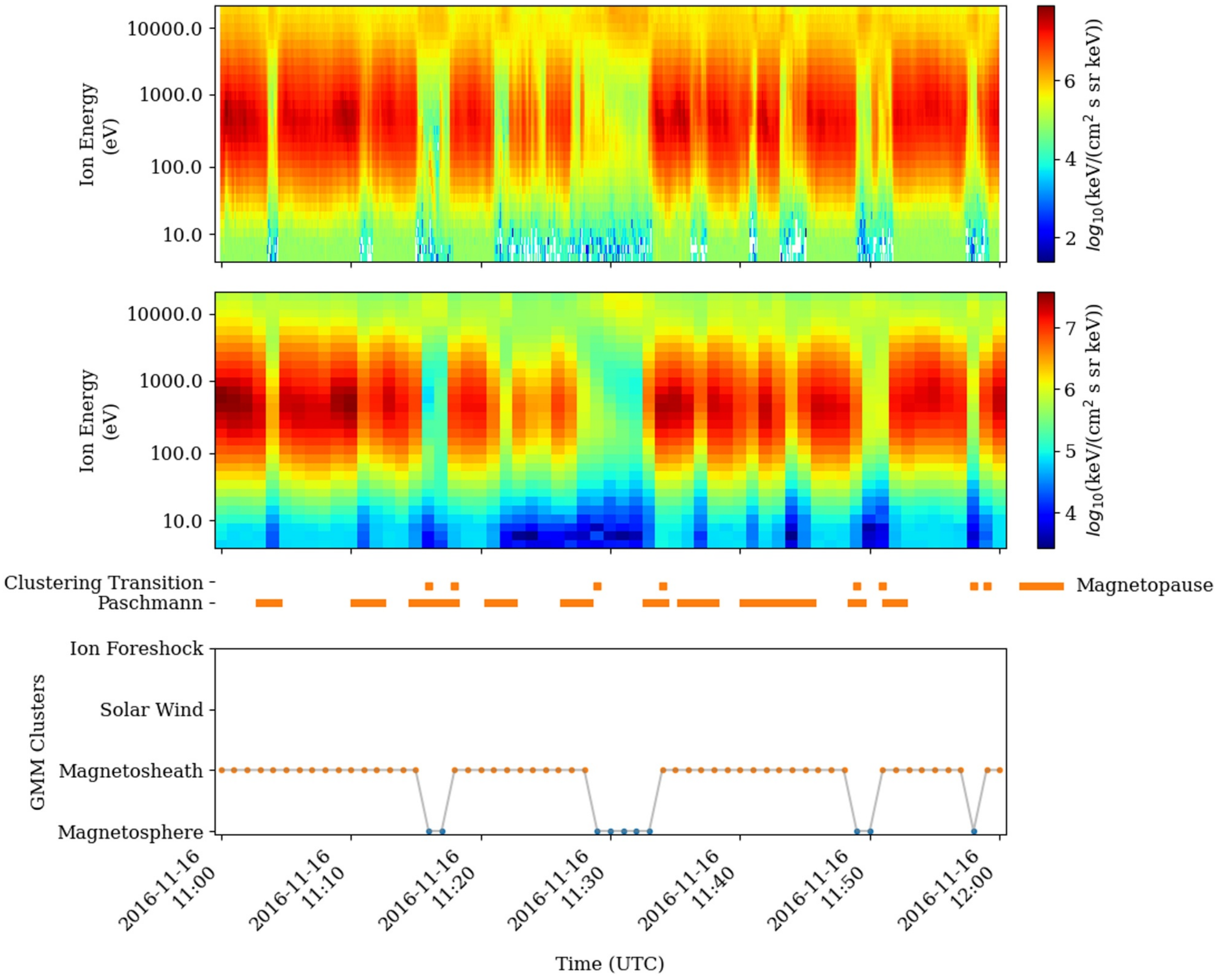

Figure 8. An example of multiple missed magnetopause transitions compared to Paschmann et al. (2018) on 16 November 2016 on MMS2. From top to bottom we have (a) the ion spectra at approximately the original resolution (rounded to 1-s), (b) the ion spectra at 1-min resolution, (c) the Paschmann et al. (2018) identified complete or Harris-like magnetopause transitions, and (d) the Gaussian mixture model cluster labels. Our clustering model misses the rapidly changing plasma regions because they are likely too short-lived and/or blended by boundary layer processes. However, our model can detect the longer-lasting magnetopause transitions where there are more clear differences between magnetosheath and magnetosphere.

crossing time period in bow shock database. 63.6% of the Lalti et al. (2022) bow shock crossings are identified by our GMM clustering method. Many disagreements come from spurious bow shocks that are removed during post-processing. This step greatly reduces the number of false rapidly changing plasma regions at the cost of removing real rapid bow shocks. We note that when our method misses classifying these types of bow shocks that we would under-count two bow shocks (the rapid change would create a bow shock in both directions of the plasma region change).

4. Summary and Discussion

We present a lightweight and extensible unsupervised method, GMM, to identify plasma regions on MMS's dayside observations at a 1-min cadence based on generated features from the ion spectra and B_{tot}. We initially trained the model on a sample across all the MMS data, but found that the more complex tailside plasma regimes (e.g., plasma sheet and lobes) blended across the dayside features and caused noise and misclassifications when included in the training data set. We therefore leave tailside plasma region classification for future work. Our dayside cleaned plasma region labels perform well against manually labeled data (97.8% accuracy) and rule based data (88.0% accuracy).

We have a reasonable match rate against SITL reported bow shocks (75.1%) and magnetopauses (76.0%). There is a significant over-contribution of bow shocks (45.0% are not seen in SITL reports) and magnetopause (64.9% are not seen in SITL reports) but these may be attributed to rapidly changing cluster classifications during unusual conditions and human error in SITL reports. Additionally, we report the percentage of overlap of magnetopause (67.4%) and bow shock (63.6%) databases in the literature. These types of transitions often have biases associated with them depending on what type of science each group is trying to achieve. For example, some groups are only interested in the most pristine data sets because it may affect their statistics and they do not include multiple rapid transitions or partial magnetopause crossings. Our data set includes every automatically identified transition and leaves it to the users to clean the data as they see fit.

We provide both initial and cleaned labeled plasma regions and transitions from September 2015 to 31 January 2024 for the four MMS spacecraft with $X_{\mathrm{GSE}} > 0$ at 1-min resolution. In addition, we provide four plasma region lists that specify times when the spacecraft is within each plasma region for 15 min or longer to aid in region specific studies. In future work, we plan to compare the magnetopause and bow shock transitions to models and refit 3D parameterized models based on this data set (Mo et al., *in prep*).

Due to the generalized nature of our features, which are only dependent on ratios and sums, this trained model could potentially be successfully applied to other missions like Cluster and THEMIS and even to other spacecraft orbiting other planetary bodies with magnetospheres (e.g., Cassini, Juno, and MESSENGER). A natural continuation of this work would be cross validating near Earth plasma properties obtained by traditional (King & Papitashvili, 2005) and/or machine learning approaches (e.g., O'Brien et al., 2023). Moreover, our labeled data set can be used as the basis of connecting plasma regions properties and upstream solar wind conditions. This could greatly contribute in advancing our knowledge on the solar wind - magnetosphere coupling in a statistical manner. Finally, our openly-available data set can aid in the statistical analysis of dayside transient phenomena and the study of their effects on the magnetosphere. It has been recently shown that such phenomena can have a large impact (Zhang et al., 2022), but statistical research has thus far been limited due to the difficulty of obtaining a large data set of these events.

Data Availability Statement

The data used in this work is publicly available through the MMS Science Data Center and is available at https://lasp.colorado.edu/mms/sdc/public/. We access finalized SITL reports from https://research.ssl.berkeley.edu/~moka/eva/bss_table_full.html.

The labeled clusters, transitions, and region lists produced in this work are available on Zenodo (Toy-Edens, 2024) at https://zenodo.org/records/10491878.

Acknowledgments

This work was funded by Grant 2225463 from the NSF GEM program. SR acknowledges the support by John Hopkins University Applied Physics Laboratory independent R&D fund. We used tools created by the HelioCloud Project. The authors thank the numerous individuals who supported to the MMS mission and made the data available. The authors gratefully acknowledge Mitsuo Oka for making the SITL reports accessible in a machine readable format and Sarah Vines for providing the Paschmann et al. (2018) magnetopause list. We thank the reviewers for their thorough comments which improved the quality of the analysis.

References

Angelopoulos, V. (2008). The THEMIS mission. *Space Science Reviews*, *141*(1–4), 5–34. https://doi.org/10.1007/s11214-008-9336-1

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*, 54. https://doi.org/10.3389/fspas.2020.00054

Borovsky, J. E. (2010). On the variations of the solar wind magnetic field about the parker spiral direction. *Journal of Geophysical Research*, *115*(A9), A09101. https://doi.org/10.1029/2009JA015040

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 observation. *Frontiers in Astronomy and Space Sciences*, *7*, 55. https://doi.org/10.3389/fspas.2020.00055

Burch, J. L., Moore, T. E., Torbert, R. B., & Giles, B. L. (2015). Magnetospheric multiscale overview and science objectives. *Space Science Reviews*, *199*(1–4), 5–21. https://doi.org/10.1007/s11214-015-0164-9

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In *Advances in knowledge discovery and data mining* (pp. 160–172). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-37456-2_14

Chapman, J. F., & Cairns, I. H. (2003). Three-dimensional modeling of earth's bow shock: Shock shape as a function of Alfvén Mach number. *Journal of Geophysical Research: Space Physics*, *108*(A5), 1174. https://doi.org/10.1029/2002JA009569

Chaston, C. C., Yao, Y., Lin, N., Salem, C., & Ueno, G. (2013). Ion heating by broadband electromagnetic waves in the magnetosheath and across the magnetopause. *Journal of Geophysical Research: Space Physics*, *118*(9), 5579–5591. https://doi.org/10.1002/jgra.50506

Dimmock, A. P., & Nykyri, K. (2013). The statistical mapping of magnetosheath plasma properties based on THEMIS measurements in the magnetosheath interplanetary medium reference frame. *Journal of Geophysical Research: Space Physics*, *118*(8), 4963–4976. https://doi.org/10.1002/jgra.50465

Dimmock, A. P., Nykyri, K., Osmane, A., Karimabadi, H., & Pulkkinen, T. I. (2017). *Dawn-dusk asymmetries of the Earth's dayside magnetosheath in the magnetosheath interplanetary medium reference frame*. Wiley. https://doi.org/10.1002/9781119216346.ch5

Escoubet, C. P., Fehringer, M., & Goldstein, M. (2001). Introduction: The cluster mission. *Annales Geophysicae*, *19*(10/12), 1197–1200. https://doi.org/10.5194/angeo-19-1197-2001

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the second international conference on knowledge discovery and data mining* (pp. 226–231). AAAI Press.

Harris, E. G. (1962). On a plasma sheath separating regions of oppositely directed magnetic field. *Il Nuovo Cimento - B*, *23*(1), 115–121. https://doi.org/10.1007/bf02733547

Henderson, M. G., Morley, S. K., & Burch, J. L. (2022). MMS 4 magnetic ephemeris and coordinates (MEC) and support (TSYGANENKO 1989 model, dynamic conditions), level 2 (l2), survey mode, 30 s data [Dataset]. *NASA Space Physics Data Facility*. https://doi.org/10.48322/552N-ER81

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

Jelínek, K., Němeček, Z., & Šafránková, J. (2012). A new approach to magnetopause and bow shock modeling based on automated region identification. *Journal of Geophysical Research: Space Physics*, *117*(A5), A05208. https://doi.org/10.1029/2011ja017252

Karlsson, T., Raptis, S., Trollvik, H., & Nilsson, H. (2021). Classifying the magnetosheath behind the quasi-parallel and quasi-perpendicular bow shock by local measurements. *Journal of Geophysical Research: Space Physics*, *126*(9), e29269. https://doi.org/10.1029/2021ja029269

King, J. H., & Papitashvili, N. E. (2005). Solar wind spatial scales in and comparisons of hourly wind and ace plasma and magnetic field data. *Journal of Geophysical Research: Space Physics*, *110*(A2), A02104. https://doi.org/10.1029/2004ja010649

Koller, F., Raptis, S., Temmer, M., & Karlsson, T. (2024). The effect of fast solar wind on ion distribution downstream of earth's bow shock. *The Astrophysical Journal Letters*, *964*(1), L5. https://doi.org/10.3847/2041-8213/ad2ddf

Lalti, A., Khotyaintsev, Y. V., Dimmock, A. P., Johlander, A., Graham, D. B., & Olshevsky, V. (2022). A database of MMS bow shock crossings compiled using machine learning. *Journal of Geophysical Research: Space Physics*, *127*(8), e30454. https://doi.org/10.1029/2022ja030454

Lepri, S. T., Landi, E., & Zurbuchen, T. H. (2013). Solar wind heavy ions over solar cycle 23: ACE/SWICS measurements. *The Astrophysical Journal*, *768*(1), 94. https://doi.org/10.1088/0004-637X/768/1/94

Lin, R. L., Zhang, X. X., Liu, S. Q., Wang, Y. L., & Gong, J. C. (2010). A three-dimensional asymmetric magnetopause model. *Journal of Geophysical Research: Space Physics*, *115*(A4), A04207. https://doi.org/10.1029/2009JA014235

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

O'Brien, C., Walsh, B. M., Zou, Y., Tasnim, S., Zhang, H., & Sibeck, D. G. (2023). Prime: A probabilistic neural network approach to solar wind propagation from l1. *Frontiers in Astronomy and Space Sciences*, *10*, 1250779. https://doi.org/10.3389/fspas.2023.1250779

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), e29620. https://doi.org/10.1029/2021ja029620

Omidi, N., Sibeck, D. G., & Blanco-Cano, X. (2009). Foreshock compressional boundary. *Journal of Geophysical Research: Space Physics*, *114*(A8), A08205. https://doi.org/10.1029/2008JA013950

Paschmann, G., Haaland, S. E., Phan, T. D., Sonnerup, B. U. O., Burch, J. L., Torbert, R. B., et al. (2018). Large-scale survey of the structure of the dayside magnetopause by MMS. *Journal of Geophysical Research: Space Physics*, *123*(3), 2018–2033. https://doi.org/10.1002/2017ja025121

Plaschke, F., Hietala, H., & Angelopoulos, V. (2013). Anti-sunward high-speed jets in the subsolar magnetosheath. *Annales Geophysicae*, *31*(10), 1877–1889. https://doi.org/10.5194/angeo-31-1877-2013

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

Raptis, S., Karlsson, T., Plaschke, F., Kullen, A., & Lindqvist, P.-A. (2020). Classifying magnetosheath jets using MMS: Statistical properties. *Journal of Geophysical Research: Space Physics*, *125*(11), e27754. https://doi.org/10.1029/2019ja027754

Richardson, J., & Kasper, J. (2008). Solar cycle variations of solar wind dynamics and structures. *Journal of Atmospheric and Solar-Terrestrial Physics*, *70*(2–4), 219–225. https://doi.org/10.1016/j.jastp.2007.08.039

Russell, C. T., Anderson, B. J., Baumjohann, W., Bromund, K. R., Dearborn, D., Fischer, D., et al. (2014). The magnetospheric multiscale magnetometers. *Space Science Reviews*, *199*(1–4), 189–256. https://doi.org/10.1007/s11214-014-0057-3

Thomas, B. A., Vandegriff, J. D., Antunes, A. K., Bradford, J. W., Yeakel, K., Mo, W., et al. (2022). HelioCloud: A community cloud-based approach to heliophysics analytics and software development. In *Agu fall meeting abstracts* (Vol. 2022, p. SH45B-01).

Torbert, R. B., Russell, C. T., Magnes, W., Ergun, R. E., Lindqvist, P.-A., LeContel, O., et al. (2014). The FIELDS instrument suite on MMS: Scientific objectives, measurements, and data products. *Space Science Reviews*, *199*(1–4), 105–135. https://doi.org/10.1007/s11214-014-0109-8

Toy-Edens, V. (2024). 8 years of dayside magnetospheric multiscale (MMS) unsupervised clustering plasma regions classifications [Dataset]. *Zenodo*. https://doi.org/10.5281/ZENODO.10491878

Tsyganenko, N. (1989). A magnetospheric magnetic field model with a warped tail current sheet. *Planetary and Space Science*, *37*(1), 5–20. https://doi.org/10.1016/0032-0633(89)90066-4

Xu, F., & Borovsky, J. E. (2015). A new four-plasma categorization scheme for the solar wind. *Journal of Geophysical Research: Space Physics*, *120*(1), 70–100. https://doi.org/10.1002/2014ja020412

Zhang, H., Zong, Q., Connor, H., Delamere, P., Facskó, G., Han, D., et al. (2022). Dayside transient phenomena and their impact on the magnetosphere and ionosphere. *Space Science Reviews*, *218*(5), 40. https://doi.org/10.1007/s11214-021-00865-0