Generalized Time-Series Analysis for In-Situ Spacecraft Observations: Anomaly Detection and Data Prioritization using Principal Components Analysis and Unsupervised Clustering

Matthew G. Finley[1], Miguel Martínez Ledesma[2], William Paterson[3], Matthew R Argall[4], David Michael Miles[5], John C. Dorelli[6], and Eftyhia Zesta[7]

[1]University of Maryland, College Park
[2]Catholic University of America
[3]NASA Goddard Spaceflight Center
[4]University of New Hampshire
[5]University of Iowa
[6]NASA-GSFC
[7]NASA Goddard

April 26, 2024

Abstract

In-situ spacecraft observations are critical to our study and understanding of the various phenomena that couple mass, momentum, and energy throughout near-Earth space and beyond. However, on-orbit telemetry constraints can severely limit the capability of spacecraft to transmit high-cadence data, and missions are often only able to telemeter a small percentage of their captured data at full rate. This presents a programmatic need to prioritize intervals with the highest probability of enabling the mission's science goals. Larger missions such as the Magnetospheric Multiscale mission (MMS) aim to solve this problem with a Scientist-In-The-Loop (SITL), where a domain expert flags intervals of time with potentially interesting data for high-cadence data downlink and subsequent study. Although suitable for some missions, the SITL solution is not always feasible, especially for low-cost missions such as CubeSats and NanoSats. This manuscript presents a generalizable method for the detection of anomalous data points in spacecraft observations, enabling rapid data prioritization without substantial computational overhead or the need for additional infrastructure on the ground. Specifically, Principal Components Analysis and One-Class Support Vector Machines are used to generate an alternative representation of the data and provide an indication, for each point, of the data's potential for scientific utility. The technique's performance and generalizability is demonstrated through application to intervals of observations, including magnetic field data and plasma moments, from the CASSIOPE e-POP/Swarm-Echo and MMS missions.

Generalized Time-Series Analysis for In-Situ Spacecraft Observations: Anomaly Detection and Data Prioritization using Principal Components Analysis and Unsupervised Clustering

Matthew G. Finley[1,2], Miguel Martinez-Ledesma[2,3], William R. Paterson[2], Matthew R. Argall[4], David M. Miles[5], John. C. Dorelli[2], and Eftyhia Zesta[2]

[1]Department of Astronomy, University of Maryland, College Park, MD, USA.

[2]Geospace Physics Laboratory, NASA Goddard Space Flight Center, Greenbelt, MD, USA.

[3]Department of Physics, Catholic University of America, District of Columbia, USA.

[4]Space Science Center, University of New Hampshire, Durham, NH, USA.

[5]Department of Physics and Astronomy, University of Iowa, Iowa City, IA, USA.

Corresponding author: Matthew G. Finley (matthew.g.finley@nasa.gov)

Key Points:

- Spacecraft often cannot transmit all measurements to Earth at full cadence due to telemetry bandwidth limitations.
- Many missions must implement complex data prioritization schemes to ensure only the highest-priority data is transmitted at high cadence.
- The proposed data prioritization technique is highly generic, compatible with inexpensive hardware, and suitable for low-cost missions.

Abstract

In-situ spacecraft observations are critical to our study and understanding of the various phenomena that couple mass, momentum, and energy throughout near-Earth space and beyond. However, on-orbit telemetry constraints can severely limit the capability of spacecraft to transmit high-cadence data, and missions are often only able to telemeter a small percentage of their captured data at full rate. This presents a programmatic need to prioritize intervals with the highest probability of enabling the mission's science goals. Larger missions such as the Magnetospheric Multiscale mission (MMS) aim to solve this problem with a Scientist-In-The-Loop (SITL), where a domain expert flags intervals of time with potentially interesting data for high-cadence data downlink and subsequent study. Although suitable for some missions, the SITL solution is not always feasible, especially for low-cost missions such as CubeSats and NanoSats. This manuscript presents a generalizable method for the detection of anomalous data points in spacecraft observations, enabling rapid data prioritization without substantial computational overhead or the need for additional infrastructure on the ground. Specifically, Principal Components Analysis and One-Class Support Vector Machines are used to generate an alternative representation of the data and provide an indication, for each point, of the data's potential for scientific utility. The technique's performance and generalizability is demonstrated through application to intervals of observations, including magnetic field data and plasma moments, from the CASSIOPE e-POP/Swarm-Echo and MMS missions.

Plain Language Summary

Measurements captured by spacecraft are necessary to our understanding the space environment near Earth and throughout our solar system. However, spacecraft can often only transmit a small portion of the data they capture back to Earth. This means that many spacecraft must prioritize intervals of data that have the highest probability of helping to further our understanding of these environments. Some missions utilize humans, on Earth, to help select these scientifically important intervals. This solution, called the Scientist-In-The-Loop, can be too expensive or programmatically complex for many small missions to implement. This manuscript presents a technique for the detection of anomalous events in spaceflight measurements using statistical analysis and machine learning. These detected anomalies can be used to prioritize data that has a high probability of scientific relevance. Further, the proposed technique is highly generalizable and computationally lightweight, making it suitable for a variety of missions. Several case studies from multiple existing missions will be analyzed throughout this paper.

1 Introduction

Magnetic field sensors are one of the many science instruments that have been a fundamental part of space exploration since its inception. Some of the first satellites, such as the late-1950's Sputnik 3 and Explorer 6, carried fluxgate magnetometers to collect scientific data (Gordon & Brown, 1972). Since then, the science of magnetometry and spaceflight has been advanced such that we can make magnetic field measurements of far-flung bodies such as asteroids (Weiss et al., 2023), Mars (Connerney et al., 2015), Jupiter (Connerney et al., 2017), and the Sun (Bale et al., 2016). The need to understand fundamental physical processes in space, such as magnetic reconnection, has driven requirements for the telemetry of measurements at higher and higher cadences (Phan et al., 2016). Additionally, our desire to enable comprehensive understanding, nowcasting, and forecasting of Earth's near-space environment has led to the development of the inexpensive CubeSat form factor and a greater number of multi-spacecraft missions. CubeSats allow cost effective proliferation of measurements that can vastly improve our models through data assimilation and machine learning techniques. The last several years have seen the emergence of large constellations, leading to a multitude of challenges, particularly in the handling and telemetry of the massive quantities of raw data available onboard each spacecraft (Liddle et al., 2020; Zhan et al., 2020).

Many missions, due to programmatic constraints on telemetry rates, are unfortunately unable to downlink all of their captured data to Earth for analysis. Instead, mission operators and science teams must make decisions about which intervals of time to transmit high-cadence data (i.e., *burst data*). Some missions will transmit a lower-cadence data product (i.e., *survey data*) during intervals deemed less important (Lepping et al., 1995), and some missions will simply not telemeter these intervals (Yau & James, 2015). Intervals of burst data to be telemetered are typically determined by estimated spacecraft position, by humans on the ground diligently monitoring low-cadence data, or by carefully calibrated onboard algorithms which search the high-cadence data for mission-specific triggers.

The Magnetospheric Multiscale (MMS) mission (Burch et al., 2016) utilizes a particularly thorough approach to the identification and prioritization of burst data to telemeter to the ground (Baker et al., 2016). During spacecraft traversal through predetermined regions of interest, the MMS instruments always capture data at their high cadence burst rates. This high-cadence data is stored onboard while the lower-cadence survey data is telemetered to the ground and analyzed. If the survey data shows potentially interesting phenomena, short intervals of burst data can be downlinked from the spacecraft. The MMS mission prioritizes burst data using two techniques: the Scientist in the Loop (SITL) and the Automated Burst System (ABS).

The ABS, as its name indicates, runs automatically onboard each spacecraft and provides a data ranking metric to be downlinked alongside the survey data (Baker et al., 2016). This system uses data quality indicators calculated by each instrument to rank the available burst data in a downlink prioritization queue. The last item in the queue will be the first to be overwritten should a higher-ranking interval be identified. Although 34 data quality indicators are available for burst triggering, the early mission used only large gradients in the Z_{GSM}-component of the measured magnetic field to prioritize burst intervals while the data quality indicators were characterized. After two years of careful parameterization, ~6 data quality indicators are now used by the ABS for prioritization of burst data containing mission-specific phenomena of interest (Argall et al., 2020).

The SITL is a manual option that can validate or override the selections made by the ABS. A domain expert – with access to MMS survey data, spacecraft-calculated trigger metadata, and data from other satellites or ground systems – determines the priority of data for downlink using specialized software, ensuring that data with high scientific significance is telemetered (Baker et al., 2016).

Both the ABS and SITL burst selection schemes for data prioritization require substantial scientific infrastructure and potentially costly overhead in their implementation. Extensive onboard triggering logic, years of parameter characterization, dedicated interval labeling time from experts during in-situ mission operations, and bespoke burst prioritization software are almost certainly infeasible for low-budget missions.

One method has been recently proposed which aims to reduce the need for infrastructure-intensive SITL activities on MMS using supervised machine learning (Argall et al., 2020). Although their technique shows excellent results and enables a great reduction in the reliance on manual labeling tasks, it requires a large set of expert-labeled data during the training of their segmentation network and therefore does not eliminate the need for the SITL infrastructure entirely.

This manuscript proposes the use of a common dimensionality reduction technique, coupled with unsupervised clustering, to provide a robust and generalizable method for detecting anomalous intervals of time series spacecraft observations. This method is intended for use as a component of a drop-in burst data prioritization system for missions where the infrastructure and cost associated with more sophisticated and mission-specific solutions are not feasible. Although the SpaceX Starlink constellation currently dominates the Low Earth Orbit environment with over 2,000 satellites currently in orbit and approval granted for 12,000 total satellites, it is unlikely to remain the only major constellation in orbit (Ma et al., 2023; McDowell, 2020). These mega-constellations are a heavy burden on ground systems, requiring complex protocols for dealing with telemetry, command, and tracking (Berner, 2019). The proposed tool would be an invaluable asset for such constellation missions, enabling a higher degree of distributed autonomy in their space operations.

The following sections of this manuscript describe the proposed technique and demonstrate its performance with several case studies on observational measurements obtained from the CASSIOPE e-POP/Swarm-Echo spacecraft and one of the Magnetospheric Multiscale mission satellites.

2 Methodology

2.1 *Dimensionality Reduction via Principal Components Analysis*

Principal Components Analysis (PCA) is one of the oldest and most popular multivariate statistical analysis techniques used to reduce the dimensionality of large datasets (Jolliffe & Cadima, 2016). Mathematically, PCA is performed by identifying the eigenvectors of the covariance matrix associated with the data matrix under observation ($\boldsymbol{X}$) via

$$\boldsymbol{X}^T\boldsymbol{X} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{\boldsymbol{T}} \tag{1}$$

where the columns of $\boldsymbol{V}$ correspond to the eigenvectors and the diagonal elements of $\boldsymbol{\Lambda}$ are the associated eigenvalues. For convenience, let each eigenvector V_i be ordered by the magnitude of its associated eigenvalue.

The projection and subsequent dimensionality reduction can be realized through

$$\boldsymbol{P} = \boldsymbol{X}\mathbf{V}_{\boldsymbol{q}} \tag{2}$$

where $\boldsymbol{V_q}$ is a matrix whose columns are only the first q eigenvectors from $\boldsymbol{V}$. Throughout this manuscript, the dimensionality of the output projection is fixed to two (i.e., $q = 2$) in order to reduce the computational complexity associated with the analysis of the projected data.

The specific data matrix being analyzed in this manuscript is generated by concatenating R consecutive time intervals, of length L, from the original time series (x, with length N) into a single matrix via

$$\boldsymbol{X} = \begin{bmatrix} x(1) & x(L+1) & \dots & x(RL-L+1) \\ \dots & \dots & \dots & \dots \\ x(L) & x(2L) & \dots & x(RL) \end{bmatrix}. \quad (3)$$

It should be noted that $\boldsymbol{X}$ is a reduced version of the trajectory matrix passed into PCA by the Singular Spectrum Analysis (SSA) technique (Finley et al., 2023). Both SSA's trajectory matrix and the data matrix used throughout this paper are constructed to enable information about the temporal variation of a signal to be utilized. The reduced trajectory matrix (i.e., no overlapping samples in each consecutive column) was used here to reduce the computational complexity associated with the overall algorithm.

An example of this variation on PCA, applied to one axis of the vector magnetometer data captured by the CASSIOPE e-POP/Swarm-Echo magnetic field instrument (Wallis et al., 2015; Yau & James, 2015), is shown in Fig. 1. Figure 1(a) illustrates the 2.5-minute interval of data, at a sampling rate of 160 Hz, to be processed. Figure 2(b) shows the two-dimensional representation, given by $\boldsymbol{P}$, of the data following the application of PCA (with $q = 2$) on a data matrix constructed from 0.5-second consecutive intervals taken from the signal in Fig. 1(a).

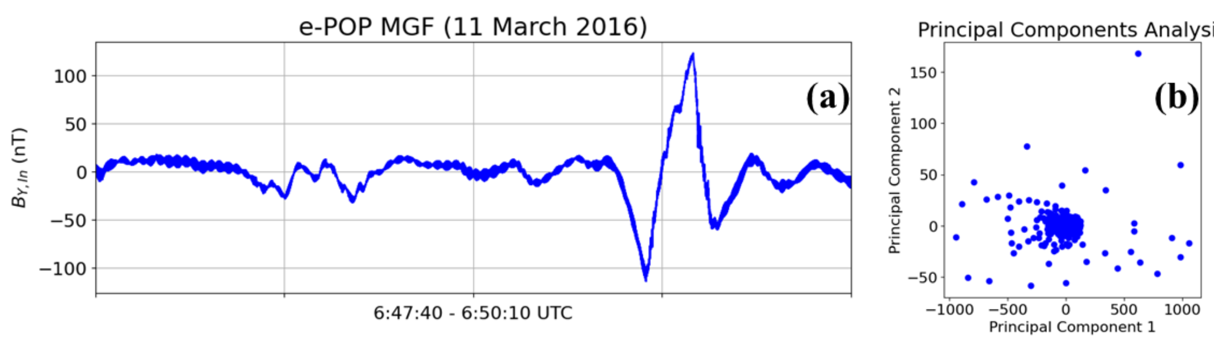

Figure 1: Dimensionality reduction via application of PCA on concatenated 0.5-second intervals of magnetic field data. (a) Inboard magnetometer data from CASSIOPE e-POP/Swarm-Echo MGF; (b) Two-dimensional representation of the 0.5-second intervals of (a) following PCA while retaining only two principal components.

Each point in Fig. 1(b) is a projection of one of the 0.5-second intervals of Fig. 1(a). It can be seen that many of these points are clustered tightly near the origin, whereas some of the points are outlying near the periphery. This implies that the majority of the time intervals exhibit similar behavior when represented using only the first two principal components (i.e., those that describe the largest variance in the original data matrix). However, some intervals show very different behavior in terms of these maximum-variance components. Automatic clustering of this two-dimensional representation should reveal anomalous time intervals in the original signal and is discussed in detail in Sec. 2.2.

2.2 *Clustering with One-Class Support Vector Machine*

Machine learning techniques have become increasingly popular in the various space physics research domains. Successful application of these techniques has been seen in methods for auroral image classification (Clausen & Nickisch, 2018), recreating magnetohydrodynamic environments from sparse sample spaces (Bard & Dorelli, 2021), space weather forecasting (Camporeale, 2019), and many others. However, machine learning models can often be complex, requiring large quantities of training data and computational resources. Once trained, these large and complex models are often treated as 'black boxes,' and can lack interpretability (Angelov et al., 2021). To increase potential applicability to low-cost and in-situ spaceflight hardware, a machine learning-based clustering solution that is computationally efficient and easily understood must be utilized instead of a more complex model.

Support Vector Machines (SVMs) are a popular means of performing classification tasks throughout a variety of fields including the biomedical sciences (Zhou et al., 2005) and industrial engineering (Shin et al., 2005). This data labeling technique has seen widespread adoption due to its high degree of robustness and interpretability (Hearst et al., 1998). Traditional SVMs are trained by first projecting the labeled training data to a higher dimension feature space using a user-selected *kernel*. Next, a hyperplane that best separates the classes is calculated, although a slack parameter is considered in this optimization. This slack parameter enables the trained SVM to handle a small number of data points that cannot be separated using a hyperplane in the higher-dimension feature space, which is a common situation in real-world datasets (Noble, 2006). This trained SVM can now be used to classify new data not seen during the training process.

Slight modification of the traditional SVM framework leads to a technique known as the One-Class Support Vector Machine (OC-SVM), a common unsupervised approach to data classification and anomaly detection (Yin et al., 2014). These OC-SVMs operate in a similar fashion to the traditional SVM but calculate a hyperplane that optimally separates the data from the origin, not by separating pre-labeled classes (Amer et al., 2013). Here, the primary user-defined control parameter is ν, which lies in the range (0,1] and determines the upper bound on the number of allowed errors and a lower bound on the number of data points used when calculating the separating hyperplane (Chang & Lin, 2001).

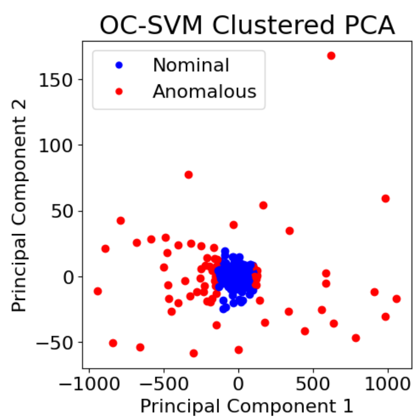

Figure 2: One-Class Support Vector Machine clustering applied to the reduced-dimension data illustrated in Fig. 1.

Figure 2 illustrates the result of passing the two-dimensional output of PCA, shown in Fig. 1(b), through an OC-SVM. This OC-SVM was trained with a Gaussian kernel and a ν-value of 0.3. Points shown in red were those that were considered anomalous, whereas the points shown in blue were considered nominal. Since each point in Fig. 2 represents a 0.5-second interval of the original input shown in Fig. 1(a), the associated labels can be directly applied to each interval in the original input time series. The result of this inversion procedure, and additional examples, will be discussed in detail in Sec. 4.

3 Data and Preprocessing

3.1 *CASSIOPE e-POP/Swarm-Echo MGF*

One primary source of the data analyzed in previous sections and throughout the remainder of this manuscript is the CASSIOPE/Swarm-Echo Magnetic Field instrument (Wallis et al., 2015; Yau & James, 2015). The Magnetic Field instrument (MGF) consists of a pair of identical fluxgate magnetometers mounted inline on a single boom at approximately 0.6 m and 0.9 m from the body of the spacecraft. Both magnetometers capture the local magnetic field at a rate of 160 Hz and downlink the data when telemetry constraints allow. Although the magnetometer pair (i.e., gradiometer) could be used to mitigate local interference from the host spacecraft and improve data fidelity (Finley et al., 2023), the measurements used in this manuscript were taken from only the inboard magnetometer mounted closer to the spacecraft. This provides some insight into how the proposed anomaly detection technique handles data contaminated by local interference, which is almost constantly observed at the magnetometers as high-frequency oscillations caused primarily by the spacecraft's attitude control systems (i.e., reaction wheels). The magnetometer measurements used in this manuscript have had a near-DC baseline removed using a 20-s moving average prior to analysis and visualization.

3.2 *Magnetospheric Multiscale Mission FGM*

Another source of magnetic field data used in the remaining sections of this manuscript is the Magnetospheric Multiscale mission (MMS) Magnetometers (Burch et al., 2016; Russell et al., 2016). Although data is available from all four MMS satellites, only data from MMS1 was utilized throughout this manuscript. The MMS Magnetometers consist of a near-identical pair of fluxgate sensors, with each sensor mounted at the end of two separate 5-meter booms. The measured data is reported as a high-fidelity triaxial vector data product (called FGM) with three possible sampling rates: slow survey at 8 Hz; fast survey at 16 Hz; and, burst data at 128 Hz. In this manuscript, only fast survey and slow survey data were utilized. If both fast and slow survey data were present in the time interval to be analyzed, the data was resampled to match the slow survey data rate for consistency of analysis. As with the CASSIOPE data discussed in Sec. 3.1., a 20-s moving average has been removed from the MMS data prior to analysis and visualization.

3.3 *Magnetospheric Multiscale Mission FPI*

An additional set of data used in the remaining sections of this manuscript is the MMS Fast Plasma Investigation (Burch et al., 2016; Pollock et al., 2016). As with the magnetic field data, only data from the MMS1 satellite was used in this manuscript. The Fast Plasma Investigation (FPI) for MMS comprises multiple top-hat electrostatic analyzers (Carlson et al., 1982) to determine in situ the fluxes of electrons and ions as functions of energy and direction. The FPI, its measurements, and methods of computation are described in detail in (Pollock et al., 2016). FPI acquires a full 3D set of electron samples (32 energies × 32 azimuths × 16 polar sections) every 30 ms, and an equivalent set of ion samples every 150 ms. When telemetered to the ground, these samples are employed to compute electron and ion fluid parameters such as number density, bulk-flow velocity, pressure, and others. Each parameter is computed as a summation of fluxes weighted by appropriate physical factors. These parameters are essential quantities required for many scientific analyses of space plasmas, for example in comparisons of observations with the output of magnetohydrodynamic (MHD) simulations that predict the overall behavior of plasma as a fluid.

The ion parameters (i.e., number density and velocity) shown in this manuscript are reported at the Fast Survey rate of ~0.22 Hz (i.e., captured at full cadence and averaged once every 4.5 s).

For the FPI, approximate fluid-like parameters are also computed in a simplified fashion by the instrument processor onboard the spacecraft. The computation is approximate because the summations are not weighted by the proper factors necessary to obtain true physical quantities. However, it has been shown that these onboard quantities, known as pseudo-moments, can be rescaled to serve as proxies for the true physical parameters (Barrie et al., 2018). As a result, analysis of the derived plasma moments using the proposed anomaly detection technique is indicative of the technique's performance when applied to the onboard pseudo-moments.

3.3 *Magnetospheric Multiscale Mission SITL*

To verify the capability of the proposed technique to identify intervals of time relevant to the MMS mission's science objectives, this manuscript also includes data from the MMS Scientist-in-the-Loop (SITL) report. MMS is the first mission with both the spatial and temporal resolution to resolve electron-scale dynamics. This requires MMS to capture much more data than it can telemeter to the ground. As discussed in Sec. 1, this has led to the development of MMS' burst management system, consisting of the Automated Burst System (ABS), Scientist-in-the-Loop (SITL), and the Ground Loop System (GLS). The SITL, specifically, is a role passed among MMS team member volunteers who search through the survey data to identify and select time intervals that may contain relevant events for burst-mode downlink. Survey data has insufficient resolution to capture electron-scale dynamics, so the SITL must over-select events to ensure that mission-critical science data is captured. The SITL is guided by mission-level science objectives that enable assignment of a figure of merit (FOM) value to each interval examined (Argall et al., 2020; Hasegawa et al., 2023; Phan et al., 2016). These FOMs subsequently inform the mission of the highest-priority data to be selected and downlinked at burst rate. Each SITL selection is accompanied by a short description that is searchable and parsable. These descriptions have previously been used to train a supervised machine learning model that uses the SITL report to make future predictions about which data should be selected for downlink (Argall et al., 2020). This model is installed in the near real-time data processing GLS so that predictions can be made as soon as the preliminary low-cadence data is downlinked in order to guide the SITL (along with the ABS selections). The SITL data itself is not publicly available, as it is not considered science-quality, but the reports describing the selected events and their time range can be searched through the MMS Mission Events webpage (https://lasp.colorado.edu/mms/sdc/public/about/events/#/). Additional tools, such as PyMMS, have been developed that enable rapid searching of these reports (Argall et al., 2022).

4 Results

4.1 *Observations of Current Sheets with Embedded Alfvén Waves (CASSIOPE)*

Section 2 illustrated (through Fig. 1 and Fig. 2, respectively) the dimensionality reduction and unsupervised clustering techniques utilized in the proposed method of automated anomaly detection by applying them to a short interval of magnetic field data captured by the CASSIOPE e-POP/Swarm-Echo inboard magnetometer. Figure 3(a) shows the proposed method's output after inverting the labels given to each point shown in Fig. 2 back onto the original magnetometer data shown in Fig. 1(a), as described in Sec. 2.2. Points in blue are those considered nominal by the technique, whereas points labeled red have been flagged as anomalous or potentially scientifically relevant. Figure 3(b) provides the spectral content for the interval shown in Fig. 3(a). The spectral content and time series for this 2.5-minute interval shows previously identified Alfvénic activity embedded in a large current sheet at roughly 6:48:55 UTC (Miles et al., 2018), with a smaller current sheet observed near 6:48:10 UTC. The current sheet containing embedded Alfvénic activity is highlighted in gray in Fig. 3(a). It can be seen that the proposed method of anomaly detection was able to accurately identify these interesting intervals containing broadband magnetic activity, even in the presence of the potentially obfuscating reaction wheel interference observed at ~15 Hz.

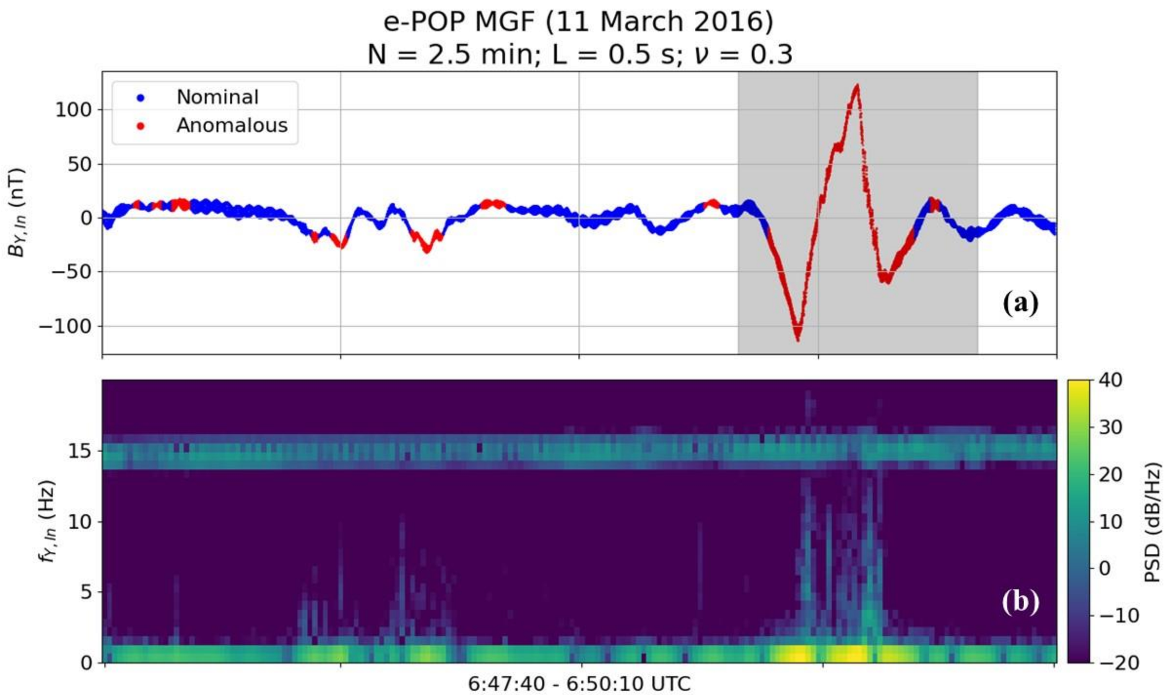

Figure 3: Proposed method of anomaly detection applied to CASSIOPE e-POP/Swarm-Echo MGF data containing Alfvénic activity. (a) Time series with anomalous intervals plotted in red, nominal regions plotted in blue; (b) Spectral content of the interval shown in (a).

4.2 *Observations of Spacecraft Maneuvers (CASSIOPE)*

The next interval of data analyzed in this section is a 20-minute interval of CASSIOPE e-POP/Swarm-Echo magnetometer data captured when the spacecraft was performing maneuvers. As a result, the spacecraft's four reaction wheels were changing their spin frequency rapidly during this period. Figure 4(a) shows the labeled time series output by the method with a window length of five seconds. Anomalous intervals are shown in red, normal points are shown in blue. Figure 4(b) shows the spectral content of the interval shown in Fig. 4(a). Highly dynamic high-frequency activity resulting from the reaction wheels' diverging operational rates can be seen for a five-minute interval starting at approximately 4:55:00 UTC, with other low-frequency perturbations occurring throughout the total 20-minute interval. It can be seen that the detected anomalies directly correspond to changes in behavior in the measured magnetic field due to the spacecraft maneuvers. This example helps to both illustrate the ability of the proposed technique to simultaneously identify dynamic behavior in both high- and low-frequency bands, as well as identify intervals critical to spacecraft operations (i.e., spacecraft maneuvers).

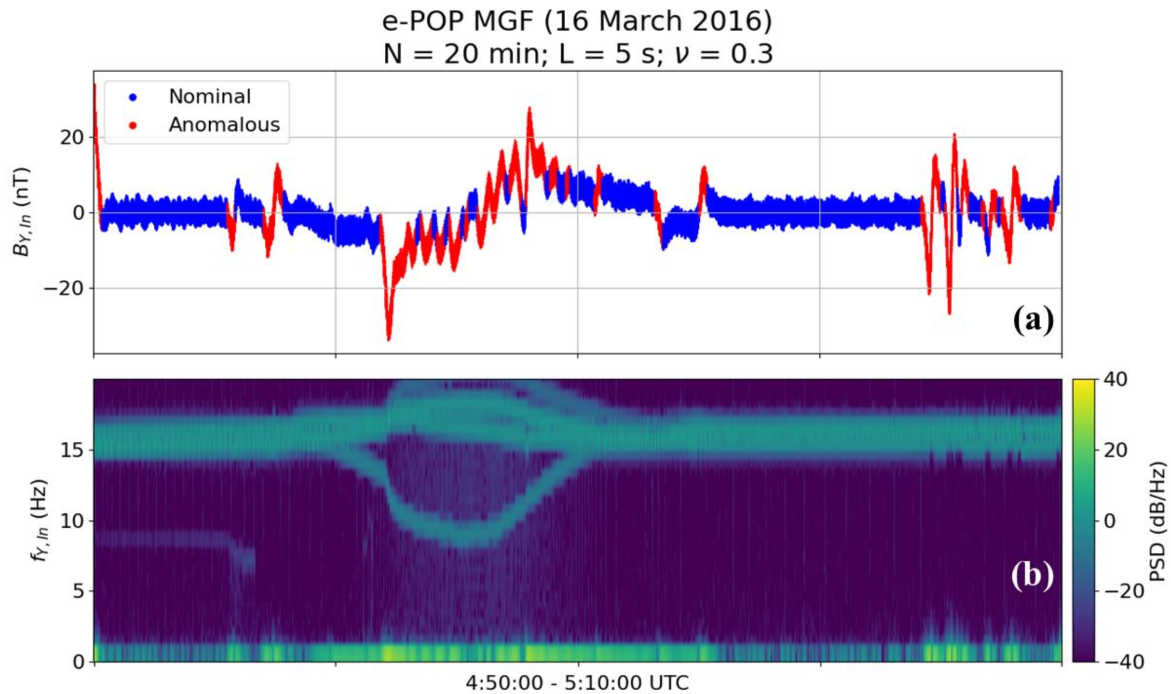

Figure 4: Proposed method of anomaly detection applied to interval of CASSIOPE e-POP/Swarm Echo MGF data containing highly dynamic high-frequency signatures caused by spacecraft reaction wheels. (a) Time series with anomalous intervals plotted in red, nominal in blue. (b) Spectrogram associated with the data in (a).

4.3 *Observations of Interplanetary Shocks and EMIC Waves (MMS)*

The next period of data analyzed in this manuscript is a full 24 hours of magnetic field data captured by the MMS magnetometer suite in December 2015. Figure 5(a) shows the result of the proposed method of anomaly detection applied to the full day of data with a window length (L) of 5 minutes. As in the previous examples, points shown in red correspond to detected anomalies and points shown in blue have been identified as nominal. Figure 5(b) shows the spectrogram associated with the magnetic field measurements in Fig. 5(a). Note that this data has been resampled to match the slowest sampling rate present in the interval (i.e., 8 Hz), as described in Sec. 3.2. The frequent broadband magnetic phenomena seen throughout the first ~40% of this 24-hour interval is attributed, by the MMS SITL, to observations of the bow shock and magnetopause. The broadband magnetic activity seen in the last ~15% of the day was similarly reported by SITL to correspond to observations of the magnetopause. The large-amplitude (i.e., > 200 nT) phenomena at approximately 17:00 UTC corresponds to the perigee of the MMS1 spacecraft. This day of data, which was analyzed extensively by Engebretson et al. (2018), also contains MMS' observations prior to, during, and after an interplanetary shock. The resultant compression of the magnetosheath was observed by MMS at approximately 13:24 UTC, with structured EMIC wave activity occurring before and after. The gray highlighted region in Fig. 5(a) corresponds to the region of interest (from 13:00-14:00 UTC) containing the majority of this activity (Engebretson et al., 2018). It can be seen that the proposed method of anomaly detection is able to accurately label a significant portion of the interesting magnetic field data occurring at bow shock and magnetopause crossings before noon and near midnight. In addition, the technique also labels several of the large-amplitude perturbations in the magnetic field measurement near and during the 13:00-14:00 UTC region of interest. Figures 6(a) and 6(b) provide a zoomed view of this region's time series and spectral content, respectively. No change has been made to the anomaly labeling, which still corresponds to the result when the proposed method is applied to the full 24-hour period. It is important to note that although the proposed technique has successfully identified the broadband signal corresponding to the magnetosheath compression, much of the structured EMIC wave activity occurring during this region of interest has not been labeled as anomalous. This result is obtained due to the input parameters selected for this example. The next example, shown in Fig. 7, will illustrate the effect of varying the input parameters when the anomaly detection technique is applied.

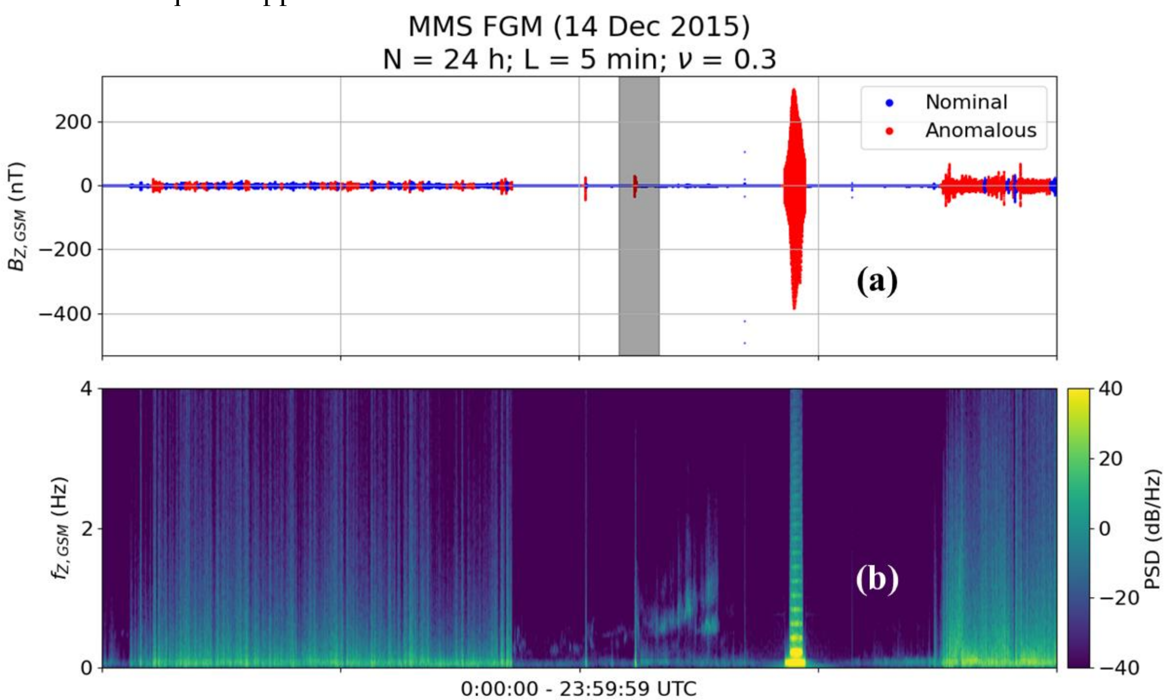

Figure 5: Anomaly detection technique applied to a 24-hour interval of MMS FGM data containing a variety of scientifically interesting phenomena including observations of the bow shock, magnetopause, and compressions of the magnetopause due to an interplanetary shock. (a) Time series FGM measurements with anomalous intervals identified by the proposed technique plotted in red, nominal intervals plotted in blue. (b) Spectrogram associated with the data shown in (a).

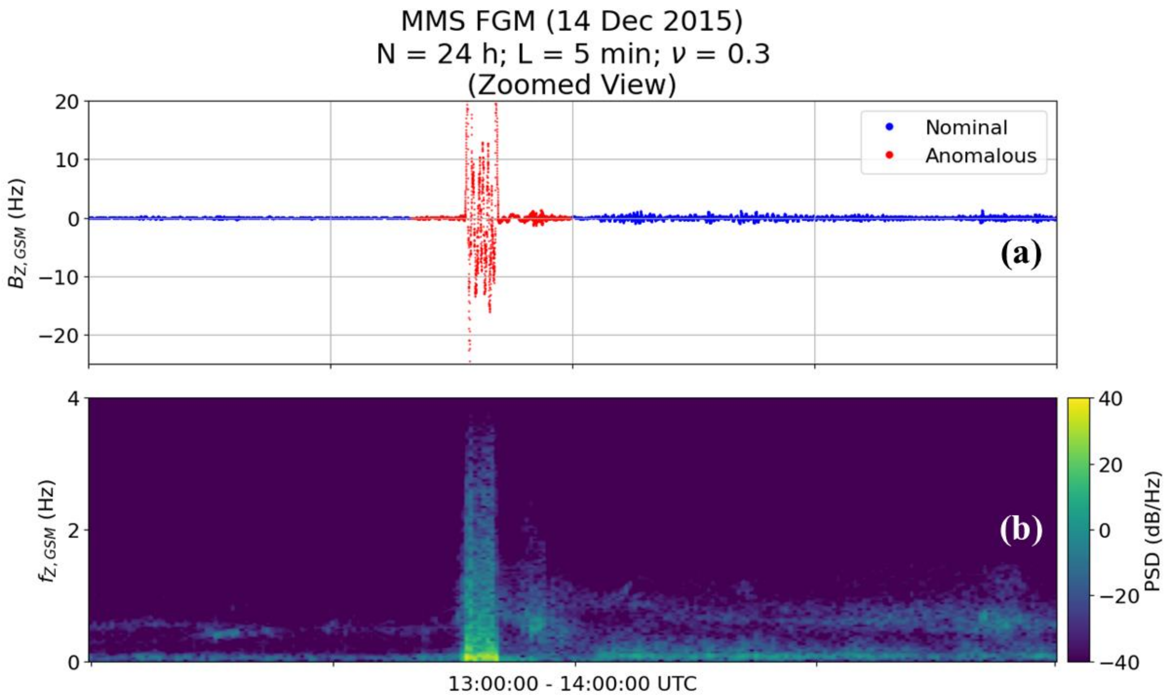

Figure 6: Zoomed view of the gray-highlighted region shown in Fig. 5, with previously identified EMIC activity. (a) Time series FGM measurement with anomalous points shown in red and nominal points shown in blue. (b) Spectral content corresponding to the interval shown in (a).

Figure 7 illustrates the impact of changing the signal length (N) and window length (L) on the performance of this technique. Specifically, Figure 7(a) shows the labeling output by the proposed method of anomaly detection when applied to only the 13:00-14:00 UTC region of interest highlighted in Fig. 5(a) and shown in Fig. 6(a). Note that this impacts the signal length, N, without changing the window length, L. Although more of the structured wave activity in this region is correctly labeled as anomalous, this result may still not be sufficient for some applications. Reducing the window length (L) from five minutes to one second, as seen in Fig. 7(b), provides a more detailed labeling of the anomalous samples within the total interval. As a result, much more of the structured EMIC wave activity occurring after the magnetosheath incursion is identified as anomalous. Although the parameter space inherent to the proposed method of anomaly detection is small, these results clearly show the relevance of the input parameters to the method's result. Selection of these parameters, and their effect on the method's computational complexity, will be discussed in greater detail in Sec. 5.1.

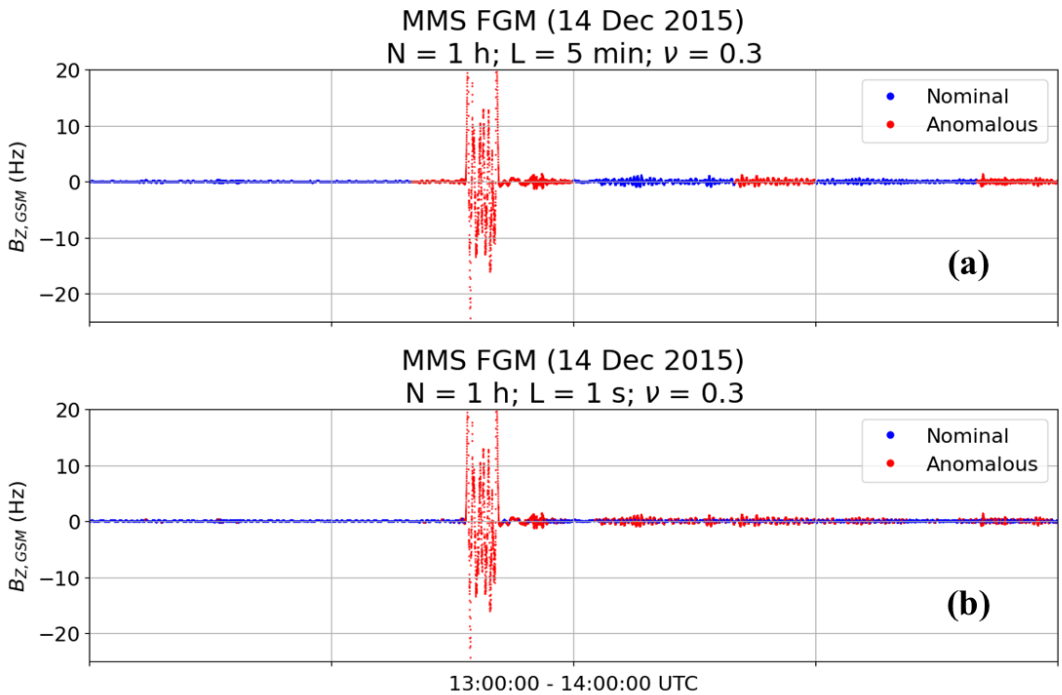

Figure 7: Illustration of the impact of the user-defined input parameters on the output of the proposed anomaly detection, shown for the same data seen in Fig. 6. (a) Time series output by the anomaly detection technique for a signal length of one hour and a window length of five minutes. (b) Time series output by the anomaly detection technique for a signal length of one hour and a window length of one second.

4.4 *Observations of Magnetopause Crossings and the Bow Shock (MMS)*

The next experiment in this manuscript demonstrates the applicability of the proposed anomaly detection technique to spacecraft observations other than magnetic field measurements. Specifically, the technique is applied to a large interval of magnetic field data, as well as ion number density and ion velocity, as measured by the MMS spacecraft on 15 May 2023. Figure 8(a) shows a 24-hour period of magnetic field measurements, Fig. 8(b) and Fig. 8(c) show the ion number density and ion velocity corresponding to the same period. The points illustrated in red are those considered anomalous by the proposed technique when it was applied to the total interval, whereas points shown in blue are considered nominal. It should be noted that the last 3.75 hours of ion measurements were not available because the spacecraft transitioned from Fast Survey to Slow Survey rate, and the FPI does not operate during Slow Survey mode. This period of missing data has been padded with nominal-labeled zeroes, after the anomaly detection algorithm was applied, for consistency with the magnetic field data. The gray region highlighted in Fig. 8 corresponds to a region of data containing various physical phenomena identified by the MMS SITL, such as observations of the bow shock, magnetopause, and boundary layer crossings. These SITL identifications, including the time range and the associated Figure of Merit (FOM) are detailed in Tab. 1. A higher FOM is assigned to a selection that should be downlinked at higher priority, with a maximum value of 199 being assigned to typical events, while higher FOMs are reserved for calibration intervals or other special events (Argall et al., 2020).

Table 1: MMS SITL report for 15 May 2023. Columns 1-2 provide the start and stop time for each interval of interest. Column 3 shows the Figure of Merit (FOM) associated with each interval. Column 4 shows SITL remarks for each interval.

Start Time (UTC)	Stop Time (UTC)	Figure of Merit (FOM)	Discussion
12:49:43	13:17:33	100	Partial Bow Shock
13:25:33	13:32:43	75	Partial Bow Shock
13:37:33	13:46:43	75	String of Partial Bow Shock Crossings
15:34:03	15:58:13	90	Partial Magnetopause
16:24:53	16:49:43	150	Full Medium Shear Magnetopause
17:51:43	17:56:13	55	Boundary Layer Traversal

In addition to these SITL-identified phenomena, the magnetic field data shows frequent broadband perturbations throughout the first half of the day, as well as a large data spike occurring near midnight that corresponds to a perigee pass of the MMS satellite. It can be seen that, for all three sets of input data, the anomaly detection technique successfully identifies much of the gray-highlighted region selected by the SITL as anomalous, but with a significant number of false positives shown for the magnetic field data when compared to only the SITL selections. These false positives in relationship to the SITL selection can be attributed to successful identification of the broadband perturbations of the magnetic field during the first half of the day. This number of false positives is greatly reduced in the labels associated with both the ion density and velocity, implying that a combination of parameters classified by the proposed technique may be used in the identification of only phenomena similar to those prioritized by the MMS SITL.

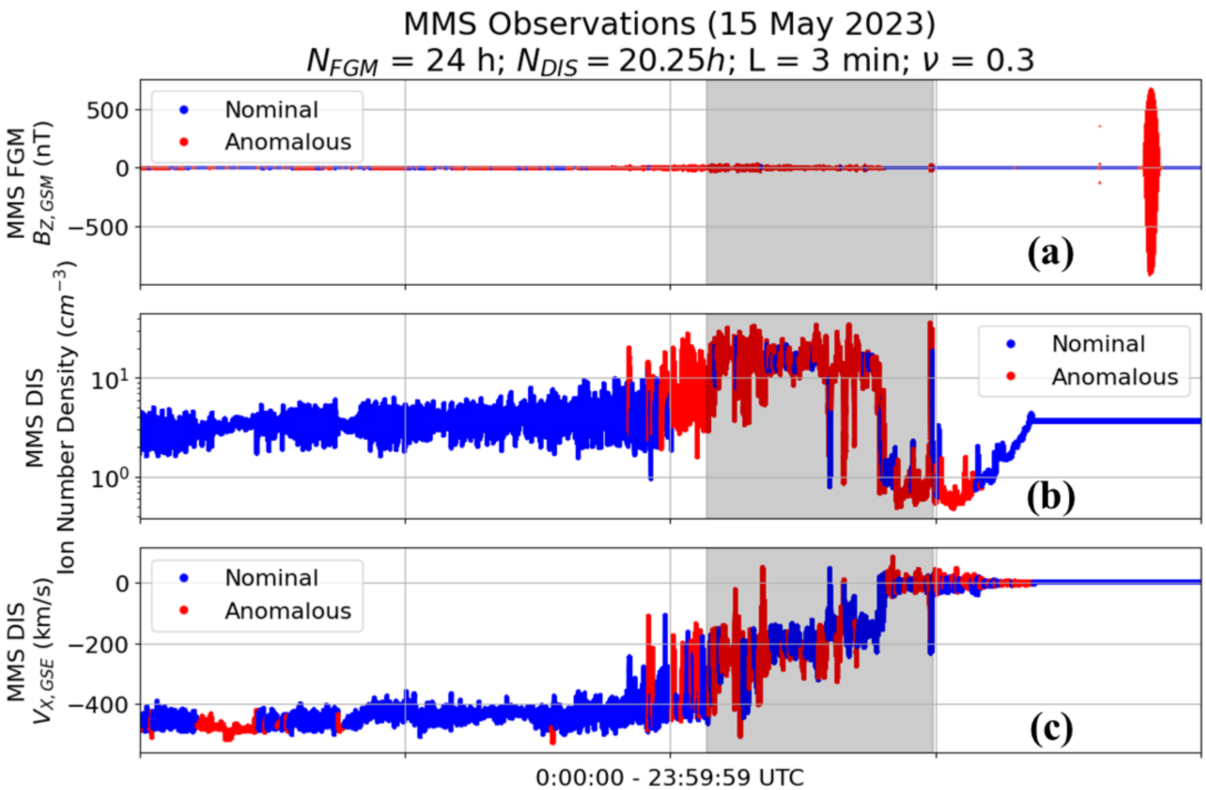

Figure 8: Anomaly detection technique applied to a long interval of MMS FGM and FPI measurements. This interval contains SITL-identified observations of the bow shock, magnetopause, and boundary layer crossings. Points highlighted in red are considered anomalous, points highlighted in blue are considered nominal. (a) FGM magnetic field measurements. (b) FPI ion number density measurements. (c) FPI ion velocity measurements.

Figure 9 provides a zoomed view of only the gray-highlighted region shown in Fig. 8. As before, Fig. 9(a) – Fig. 9(c) show the magnetic field, the ion number density, and the ion velocity measured by MMS. The nominal and anomalous labels assigned by the proposed anomaly detection technique correspond to the blue and red points, respectively. These labels are identical to the labels seen in Fig. 8; only the time scale of the plot has been changed to visualize the technique's output more clearly during the intervals selected by SITL. The regions highlighted in gray in Fig. 9 correspond to the specific time periods identified by SITL (described in Tab. 1) as containing observations of the bow shock, magnetopause, and boundary layer crossings. In a similar trend to the larger period shown in Fig. 8, the magnetic field data contains a greater number of false positive identifications of anomalous events when compared to only the SITL selections, and the ion moments, or combinations of the ion moments and magnetic field, may provide a greater degree of accuracy for this case study.

The example shown in Fig. 8 and Fig. 9 clearly demonstrates the generic applicability of the anomaly detection algorithm to a variety of spaceflight data products. Additionally, it shows that greater utility may be leveraged from the proposed method by analyzing its output when applied to several data products. For example, a weighting scheme generated by some logical operation of the binary labels (nominal or anomalous) associated with the method's output, when applied to multiple data products, may provide a reduced number of potential false positives depending on the target application. This is left as a potential topic for future study, and Section 5 will discuss several other avenues for future work utilizing this technique.

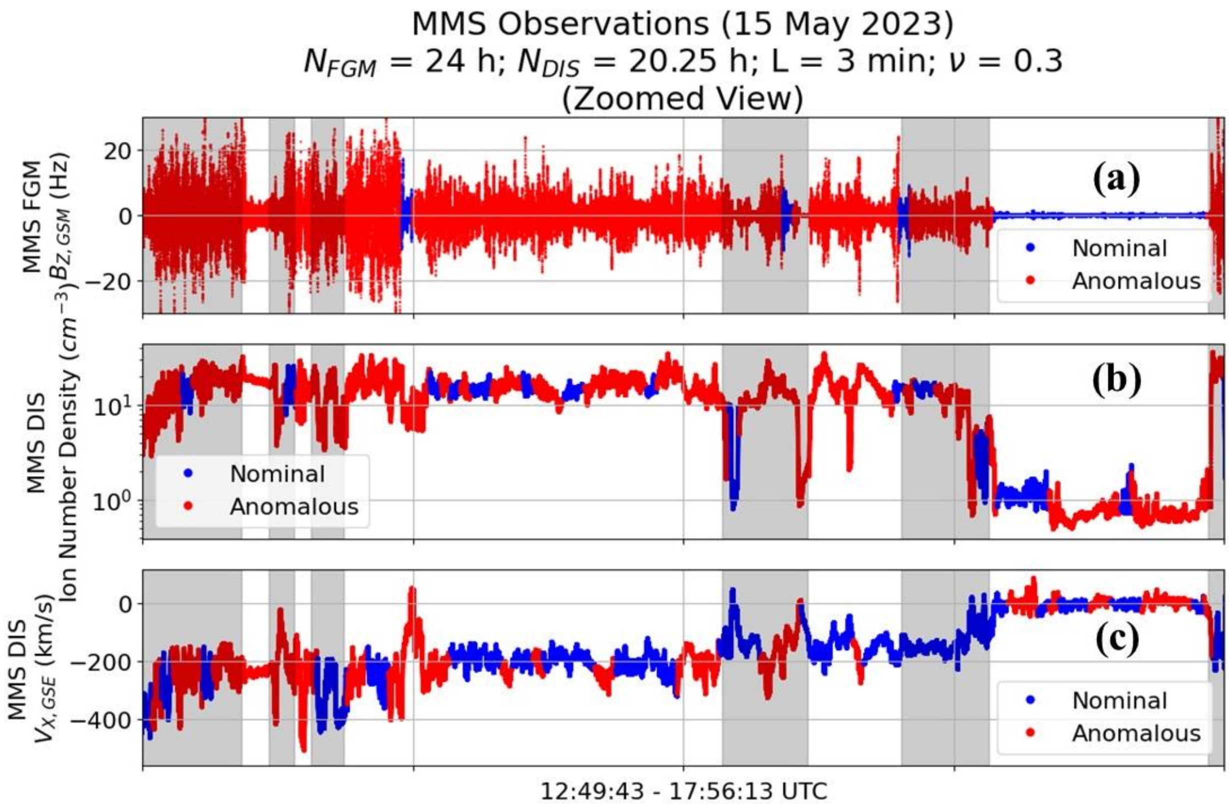

Figure 9: Zoomed view of the region of interest shown in Fig. 8., with SITL-identified observations of the bow shock, magnetopause, and boundary layer crossings. (a) FGM magnetic field data. (b) FPI ion number density. (c) FPI ion velocity.

5 Discussion & Future Work

5.1 *Parameter Selection & Computational Complexity*

This manuscript has demonstrated the capabilities of the proposed method of anomaly detection to analyze data from a variety of instruments and successfully identify various interesting phenomena; however, the example shown in Fig. 5 through Fig. 7 has illustrated the need for appropriate parameter selection to enable the highest possible scientific return. The proposed technique has a small parameter space, with only three variables impacting the output result. Thus, each of these parameters plays a critical role in the outcome of the analysis, as well is its computational complexity.

The input signal length (N) determines the total length of data to be analyzed when determining anomalous activity. The window length, L, determines the length of the consecutive observations used as features when performing PCA. Finally, the tuning parameter ν used in the OC-SVM is, practically, the fraction of the consecutive observations that can be considered anomalous when the data is clustered. In practice, this means that the input signal length (N) must be large enough that the majority of the data (i.e., at least the fraction given by $1 - \nu$) should be considered nominal, based on the target application. The window length must also be carefully considered when applying the technique as it corresponds to the scale on which anomalies are detected. If large, slowly varying changes are to be identified, a longer window length might be utilized; if small, rapid changes must be flagged, a shorter window length may be more suitable.

Changes in these parameters also have a significant impact on the computational complexity of the previously described technique. PCA, which is one of the fundamental building blocks for this algorithm, has two basic steps: computation of the covariance matrix, which has a complexity of $\mathrm{O}(L^2 R)$ for a given window length of L and a given number of total consecutive observations, R; and, eigendecomposition of the covariance matrix, which has a complexity of up to $\mathrm{O}(R^3)$ in the worst-case scenario (Zhang et al., 2015). For the OC-SVM, the most computationally expensive component is the model training, with a complexity of up to $\mathrm{O}(R^3)$. However, this complexity has been proven to be reduced by approximately one order of magnitude using more complex techniques such as sequential minimal optimization (Kang et al., 2019).

It can be seen that the window length, the number of consecutive observation intervals, and the OC-SVM tuning parameter have a direct impact on the computational complexity, or the detected anomalies associated with the proposed algorithm. As a result, users must tune these parameters as appropriate to suite the capabilities of their hardware and the data being processed.

5.2 *Potential for Implementation on Low-Cost Hardware*

Although PCA has historically been one of the most popular statistical analysis techniques used to reduce the dimensionality of large datasets (Jolliffe & Cadima, 2016), until recently relatively few studies have evaluated the potential for PCA's implementation on embedded hardware (Korat & Alimohammad, 2019). Early efforts to perform PCA on embedded hardware relied on bespoke Very Large Scale Integration (VLSI) integrated circuits with complex architecture (Tung-Chien Chen et al., 2008), but more recent works have leveraged modern and relatively inexpensive FPGA technologies to perform either portions of the PCA computations (Ali et al., 2013) or complete implementations of the PCA algorithm (Bravo et al., 2010; Korat & Alimohammad, 2019). Additional research has proven that embedded implementation of PCA is generic and highly scalable, enabling substantial improvements in the computational speed of the technique across a range of applications (Shahrouzi & Perera, 2019).

Support Vector Machines (SVMs), which fall into the broad category of machine learning algorithms, have historically provided excellent performance when classifying complex and continuous features (Saidi et al., 2021). Generally, machine learning techniques are considered computationally expensive and challenging to implement on embedded hardware (Sze et al., 2017). However, several recent studies have shown the potential for the SVM algorithm to be implemented on a variety of embedded devices including VLSI integrated circuits and FPGAs (Amezzane et al., 2020; Loukrakpam & Choudhury, 2020).

In addition to being suitable for implementation on embedded hardware, both Principal Components Analysis and/or Support Vector Machines have been previously utilized for spacecraft fault detection and diagnosis (Yu Gao et al., 2012), the onboard detection of anomalous behavior in CubeSat solar panels (Cespedes et al., 2022), and other intelligent decision making applications onboard spacecraft (Jallad & Mohammed, 2014). This illustrates that not only can PCA and SVM be implemented in hardware, but that they have history in successful implementation for spaceflight applications.

5.3 *Semantic Labeling of Prioritized Data*

This manuscript has proposed a technique for the automated binary classification of time-series data as either anomalous (i.e., potentially of high scientific importance) or nominal, one useful avenue of future work would be to provide additional semantic labels for the high-priority data. These labels could indicate whether an identified event falls into a particular class of geophysical event, such as shocks, magnetopause crossings, or whistler-mode waves. Several machine learning techniques have been previously developed for the identification of events in spaceflight data archives (Fordin et al., 2023; Vech & Malaspina, 2021), although the computational intensity or large dimensionality of some of these algorithms make them potentially unsuitable for deployment on spaceflight hardware. The proposed anomaly detection technique provides utility to the pursuit of semantic labeling in two ways: firstly, as a low-cost data reduction tool it can reduce the number of samples that must be processed by more complex algorithms, decreasing the overall time complexity of the problem; secondly, the binary-labeled data can serve as a powerful input feature vector if data reduction is not desired or required, potentially increasing model performance.

5.4 Generalizability

This manuscript has illustrated the applicability of the proposed method of anomaly detection to magnetic field data from MMS and CASSIOPE/Swarm-Echo, as well as the ion density and velocity moments from MMS. Although this shows promise for the generalizability of the technique to other platforms and other physical observations, thorough exploration of this generalizability to other spacecraft observations (such as ion and electron pressures and temperatures, as well as the housekeeping measurements critical to spacecraft operations) remains an avenue for future work.

Testing the generalizability of the clustering model trained on a specific interval of time against different intervals is also an interesting future project. If the clustering learned by an OC-SVM trained on one interval of time (e.g., one of the 24-hour periods of MMS data) is applied to subsequent intervals (e.g., the next several days of MMS data) with meaningful results, the computational complexity of the proposed technique would decrease by eliminating the need to retrain a model for every interval under observation, enabling more rapid in-situ application.

6 Conclusions

The scientific measurements captured by in-situ spacecraft are critical to our study of the physical phenomena that control the flow of mass, momentum, and energy throughout our solar system. However, due to burdens on ground systems as a result of a rapid increase in the number of active spaceflight missions and the ever-growing need for higher-cadence data, spacecraft are often unable to transmit all of their data to Earth at full rate. As a result, missions must develop techniques that enable prioritization of specific intervals with a high probability of importance to their science goals. The techniques that have historically been used by large missions have provided excellent results but come with high design and implementation costs, leaving them potentially unsuitable for application on low-cost missions. This manuscript has proposed a generic technique for the prioritization of data through the detection of anomalous data points in spaceflight observations. The technique’s utility has been demonstrated through the successful identification of various physical phenomena in a variety of data products from several missions, and its potential applicability to low-cost spaceflight hardware has been discussed. Additionally, several avenues for potential future research utilizing the proposed technique have been highlighted.

Acknowledgments

This work was supported in part by the National Aeronautics and Space Administration HERMES mission and grants 80NSSC21M0180, 80NSSC23K1295. Additional support was provided by the Center for HelioAnalytics, funded through the NASA ISFM program. The computing resources utilized in the generation of the figures and analysis in this manuscript were in part provided by HelioCloud, a service managed by NASA's Heliophysics Digital Resource Library. The authors thank Barbara J. Thompson for her thoughtful insights towards the potential applications of the proposed method.

Open Research

The CASSIOPE/Swarm-Echo MGF data used in this manuscript is publicly available at https://epop-data.phys.ucalgary.ca/. The MMS FGM, FPI, and SITL data used in this manuscript is publicly available through https://lasp.colorado.edu/mms/sdc/public/. Code used to implement the algorithm described in this manuscript, along with sample data used to illustrate the technique's performance, are currently stored at https://drive.google.com/drive/folders/1151OzBKW9Mgf6xVQ8P_bGLMmxTfBjHyG?usp=sharing. Upon acceptance of this manuscript, the code and sample data will be moved to a University of Maryland institutional repository, or similar digital repository, for long-term storage and reuse.

References

Ali, A. A. S., Amira, A., Bensaali, F., & Benammar, M. (2013). Hardware PCA for gas identification systems using high level synthesis on the Zynq SoC. *2013 IEEE 20th International Conference on Electronics, Circuits, and Systems (ICECS)*, 707–710. https://doi.org/10.1109/ICECS.2013.6815512

Amer, M., Goldstein, M., & Abdennadher, S. (2013). Enhancing one-class support vector machines for unsupervised anomaly detection. *Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description*, 8–15. https://doi.org/10.1145/2500853.2500857

Amezzane, I., Fakhri, Y., El Aroussi, M., & Bakhouya, M. (2020). Hardware Acceleration of SVM Training for Real-Time Embedded Systems: Overview. In S. Dos Santos, M. Maslouhi, & K. A. Okoudjou (Eds.), *Recent Advances in Mathematics and Technology: Proceedings of the First International Conference on Technology, Engineering, and Mathematics, Kenitra, Morocco, March 26-27, 2018* (pp. 131–139). Springer International Publishing. https://doi.org/10.1007/978-3-030-35202-8_7

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: An analytical review. *WIREs Data Mining and Knowledge Discovery*, *11*(5), e1424. https://doi.org/10.1002/widm.1424

Argall, M. R., Small, C. R., & Petrik, M. (2022). *argallmr/pymms: V0.4.6 (2022-05-19)* (0.4.6) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.6564714

Argall, M. R., Small, C. R., Piatt, S., Breen, L., Petrik, M., Kokkonen, K., Barnum, J., Larsen, K., Wilder, F. D., Oka, M., Paterson, W. R., Torbert, R. B., Ergun, R. E., Phan, T., Giles, B. L., & Burch, J. L. (2020). MMS SITL Ground Loop: Automating the Burst Data Selection Process. *Frontiers in Astronomy and Space Sciences*, *7*. https://www.frontiersin.org/articles/10.3389/fspas.2020.00054

Baker, D. N., Riesberg, L., Pankratz, C. K., Panneton, R. S., Giles, B. L., Wilder, F. D., & Ergun, R. E. (2016). Magnetospheric Multiscale Instrument Suite Operations and Data System. *Space Science Reviews*, *199*(1), 545–575. https://doi.org/10.1007/s11214-014-0128-5

Bale, S. D., Goetz, K., Harvey, P. R., Turin, P., Bonnell, J. W., Dudok de Wit, T., Ergun, R. E., MacDowall, R. J., Pulupa, M., Andre, M., Bolton, M., Bougeret, J.-L., Bowen, T. A., Burgess, D., Cattell, C. A., Chandran, B. D. G., Chaston, C. C., Chen, C. H. K., Choi, M. K., … Wygant, J. R. (2016). The FIELDS Instrument Suite for Solar Probe Plus. *Space Science Reviews*, *204*(1), Article 1. https://doi.org/10.1007/s11214-016-0244-5

Bard, C., & Dorelli, J. C. (2021). Neural Network Reconstruction of Plasma Space-Time. *Frontiers in Astronomy and Space Sciences*, *8*. https://www.frontiersin.org/articles/10.3389/fspas.2021.732275

Barrie, A. C., da Silva, D., Elkington, S., Sternovsky, Z., Rager, A. C., Gershman, D. J., Paterson, W. R., Dorelli, J. C., & Giles, B. (2018). Physically Accurate Large Dynamic Range Pseudo Moments for the MMS Fast Plasma Investigation. *Earth and Space Science*, *5*(9), 503–515. https://doi.org/10.1029/2018EA000407

Berner, J. B. (2019). Deep Space Network in the CubeSat Era. *IEEE Aerospace and Electronic Systems Magazine*, *34*(4), 46–54. https://doi.org/10.1109/MAES.2019.2913266

Bravo, I., Mazo, M., Lázaro, J. L., Gardel, A., Jiménez, P., & Pizarro, D. (2010). An Intelligent Architecture Based on Field Programmable Gate Arrays Designed to Detect Moving Objects by Using Principal Component Analysis. *Sensors*, *10*(10), Article 10. https://doi.org/10.3390/s101009232

Burch, J. L., Moore, T. E., Torbert, R. B., & Giles, B. L. (2016). Magnetospheric Multiscale Overview and Science Objectives. *Space Science Reviews*, *199*(1), Article 1. https://doi.org/10.1007/s11214-015-0164-9

Camporeale, E. (2019). The Challenge of Machine Learning in Space Weather: Nowcasting and Forecasting. *Space Weather*, *17*(8), Article 8. https://doi.org/10.1029/2018SW002061

Carlson, C. W., Curtis, D. W., Paschmann, G., & Michel, W. (1982). An instrument for rapidly measuring plasma distribution functions with high resolution. *Advances in Space Research*, *2*(7), 67–70. https://doi.org/10.1016/0273-1177(82)90151-X

Cespedes, A. J. J., Pangestu, B. H. B., Hanazawa, A., & Cho, M. (2022). Performance Evaluation of Machine Learning Methods for Anomaly Detection in CubeSat Solar Panels. *Applied Sciences*, *12*(17), Article 17. https://doi.org/10.3390/app12178634

Chang, C.-C., & Lin, C.-J. (2001). Training v-Support Vector Classifiers: Theory and Algorithms. *Neural Computation*, *13*(9), 2119–2147. https://doi.org/10.1162/089976601750399335

Clausen, L. B. N., & Nickisch, H. (2018). Automatic Classification of Auroral Images From the Oslo Auroral THEMIS (OATH) Data Set Using Machine Learning. *Journal of Geophysical Research: Space Physics*, *123*(7), Article 7. https://doi.org/10.1029/2018JA025274

Connerney, J. E. P., Benn, M., Bjarno, J. B., Denver, T., Espley, J., Jorgensen, J. L., Jorgensen, P. S., Lawton, P., Malinnikova, A., Merayo, J. M., Murphy, S., Odom, J., Oliversen, R., Schnurr, R., Sheppard, D., & Smith, E. J. (2017). The Juno Magnetic Field Investigation. *Space Science Reviews*, *213*(1), Article 1. https://doi.org/10.1007/s11214-017-0334-z

Connerney, J. E. P., Espley, J., Lawton, P., Murphy, S., Odom, J., Oliversen, R., & Sheppard, D. (2015). The MAVEN Magnetic Field Investigation. *Space Science Reviews*, *195*(1), Article 1. https://doi.org/10.1007/s11214-015-0169-4

Engebretson, M. J., Posch, J. L., Capman, N. S. S., Campuzano, N. G., Bělik, P., Allen, R. C., Vines, S. K., Anderson, B. J., Tian, S., Cattell, C. A., Wygant, J. R., Fuselier, S. A., Argall, M. R., Lessard, M. R., Torbert, R. B., Moldwin, M. B., Hartinger, M. D., Kim, H., Russell, C. T., … Singer, H. J. (2018). MMS, Van Allen Probes, GOES 13, and Ground-Based Magnetometer Observations of EMIC Wave Events Before, During, and After a Modest Interplanetary Shock. *Journal of Geophysical Research: Space Physics*, *123*(10), 8331–8357. https://doi.org/10.1029/2018JA025984

Finley, M. G., Broadfoot, R. M., Shekhar, S., & Miles, D. M. (2023). Identification and Removal of Reaction Wheel Interference From In-Situ Magnetic Field Data Using Multichannel Singular Spectrum Analysis. *Journal of Geophysical Research: Space Physics*, *128*(2), Article 2. https://doi.org/10.1029/2022JA031020

Fordin, S., Shay, M., III, L. B. W., Maruca, B., & Thompson, B. J. (2023). A Machine Learning–Based Approach to Time-series Wave Identification in the Solar Wind. *The Astrophysical Journal*, *949*(2), 40. https://doi.org/10.3847/1538-4357/acc8d5

Gordon, D., & Brown, R. (1972). Recent advances in fluxgate magnetometry. *IEEE Transactions on Magnetics*, *8*(1), Article 1. https://doi.org/10.1109/TMAG.1972.1067268

Hasegawa, H., Argall, M. R., Aunai, N., Bandyopadhyay, R., Bessho, N., Cohen, I. J., Denton, R. E., Dorelli, J. C., Egedal, J., Fuselier, S. A., Garnier, P., Genot, V., Graham, D. B., Hwang, K. J., Khotyaintsev, Y. V., Korovinskiy, D. B., Lavraud, B., Lenouvel, Q., Li, T. C., … Zenitani, S. (2023). *Advanced methods for analyzing in-situ observations of magnetic reconnection* (arXiv:2307.05867). arXiv. https://doi.org/10.48550/arXiv.2307.05867

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and Their Applications*, *13*(4), 18–28. https://doi.org/10.1109/5254.708428

Jallad, A.-H. M., & Mohammed, L. B. (2014). Hardware Support Vector Machine (SVM) for satellite on-board applications. *2014 NASA/ESA Conference on Adaptive Hardware and Systems (AHS)*, 256–261. https://doi.org/10.1109/AHS.2014.6880185

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Kang, S., Kim, D., & Cho, S. (2019). Approximate training of one-class support vector machines using expected margin. *Computers & Industrial Engineering*, *130*, 772–778. https://doi.org/10.1016/j.cie.2019.03.029

Korat, U. A., & Alimohammad, A. (2019). A Reconfigurable Hardware Architecture for Principal Component Analysis. *Circuits, Systems, and Signal Processing*, *38*(5), 2097–2113. https://doi.org/10.1007/s00034-018-0953-y

Lepping, R. P., Acũna, M. H., Burlaga, L. F., Farrell, W. M., Slavin, J. A., Schatten, K. H., Mariani, F., Ness, N. F., Neubauer, F. M., Whang, Y. C., Byrnes, J. B., Kennon, R. S., Panetta, P. V., Scheifele, J., & Worley, E. M. (1995). The WIND magnetic field investigation. *Space Science Reviews*, *71*(1), 207–229. https://doi.org/10.1007/BF00751330

Liddle, J. D., Holt, A. P., Jason, S. J., O'Donnell, K. A., & Stevens, E. J. (2020). Space science with CubeSats and nanosatellites. *Nature Astronomy*, *4*(11), Article 11. https://doi.org/10.1038/s41550-020-01247-2

Loukrakpam, M., & Choudhury, M. (2020). Hardware-Efficient VLSI Design for Cascade Support Vector Machine with On-Chip Training and Classification Capability. *Circuits, Systems, and Signal Processing*, *39*(10), 5272–5297. https://doi.org/10.1007/s00034-020-01415-9

Ma, S., Chou, Y. C., Zhao, H., Chen, L., Ma, X., & Liu, J. (2023). Network Characteristics of LEO Satellite Constellations: A Starlink-Based Measurement from End Users. *IEEE INFOCOM 2023 - IEEE Conference on Computer Communications*, 1–10. https://doi.org/10.1109/INFOCOM53939.2023.10228912

McDowell, J. C. (2020). The Low Earth Orbit Satellite Population and Impacts of the SpaceX Starlink Constellation. *The Astrophysical Journal Letters*, *892*(2), L36. https://doi.org/10.3847/2041-8213/ab8016

Miles, D. M., Mann, I. R., Pakhotin, I. P., Burchill, J. K., Howarth, A. D., Knudsen, D. J., Lysak, R. L., Wallis, D. D., Cogger, L. L., & Yau, A. W. (2018). Alfvénic Dynamics and Fine Structuring of Discrete Auroral Arcs: Swarm and e-POP Observations. *Geophysical Research Letters*, *45*(2), Article 2. https://doi.org/10.1002/2017GL076051

Noble, W. S. (2006). What is a support vector machine? *Nature Biotechnology*, *24*(12), Article 12. https://doi.org/10.1038/nbt1206-1565

Phan, T. D., Shay, M. A., Eastwood, J. P., Angelopoulos, V., Oieroset, M., Oka, M., & Fujimoto, M. (2016). Establishing the Context for Reconnection Diffusion Region Encounters and Strategies for the Capture and Transmission of Diffusion Region Burst Data by MMS. *Space Science Reviews*, *199*(1), 631–650. https://doi.org/10.1007/s11214-015-0150-2

Pollock, C., Moore, T., Jacques, A., Burch, J., Gliese, U., Saito, Y., Omoto, T., Avanov, L., Barrie, A., Coffey, V., Dorelli, J., Gershman, D., Giles, B., Rosnack, T., Salo, C., Yokota, S., Adrian, M., Aoustin, C., Auletti, C., … Zeuch, M. (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., Le, G., Leinweber, H. K., Leneman, D., Magnes, W., Means, J. D., Moldwin, M. B., Nakamura, R., Pierce, D., Plaschke, F., Rowe, K. M., Slavin, J. A., Strangeway, R. J., Torbert, R., … Richter, I. (2016). The Magnetospheric

Multiscale Magnetometers. *Space Science Reviews*, *199*(1), Article 1. https://doi.org/10.1007/s11214-014-0057-3

Saidi, A., Ben Othman, S., Dhouibi, M., & Ben Saoud, S. (2021). FPGA-based implementation of classification techniques: A survey. *Integration*, *81*, 280–299. https://doi.org/10.1016/j.vlsi.2021.08.004

Shahrouzi, S. N., & Perera, D. G. (2019). Optimized hardware accelerators for data mining applications on embedded platforms: Case study principal component analysis. *Microprocessors and Microsystems*, *65*, 79–96. https://doi.org/10.1016/j.micpro.2019.01.001

Shin, H. J., Eom, D.-H., & Kim, S.-S. (2005). One-class support vector machines—An application in machine fault detection and classification. *Computers & Industrial Engineering*, *48*(2), 395–408. https://doi.org/10.1016/j.cie.2005.01.009

Sze, V., Chen, Y.-H., Emer, J., Suleiman, A., & Zhang, Z. (2017). Hardware for machine learning: Challenges and opportunities. *2017 IEEE Custom Integrated Circuits Conference (CICC)*, 1–8. https://doi.org/10.1109/CICC.2017.7993626

Tung-Chien Chen, Wentai Liu, & Liang-Gee Chen. (2008). VLSI architecture of leading eigenvector generation for on-chip principal component analysis spike sorting system. *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 3192–3195. https://doi.org/10.1109/IEMBS.2008.4649882

Vech, D., & Malaspina, D. M. (2021). A Novel Machine Learning Technique to Identify and Categorize Plasma Waves in Spacecraft Measurements. *Journal of Geophysical Research: Space Physics*, *126*(9), e2021JA029567. https://doi.org/10.1029/2021JA029567

Wallis, D. D., Miles, D. M., Narod, B. B., Bennest, J. R., Murphy, K. R., Mann, I. R., & Yau, A. W. (2015). The CASSIOPE/e-POP Magnetic Field Instrument (MGF). *Space Science Reviews*, *189*(1), Article 1. https://doi.org/10.1007/s11214-014-0105-z

Weiss, B. P., Merayo, J. M. G., Ream, J. B., Oran, R., Brauer, P., Cochrane, C. J., Cloutier, K., Elkins-Tanton, L. T., Jørgensen, J. L., Maurel, C., Park, R. S., Polanskey, C. A., de Soria Santacruz-Pich, M., Raymond, C. A., Russell, C. T., Wenkert, D., Wieczorek, M. A., & Zuber, M. T. (2023). The Psyche Magnetometry Investigation. *Space Science Reviews*, *219*(3), 22. https://doi.org/10.1007/s11214-023-00965-z

Yau, A. W., & James, H. G. (2015). CASSIOPE Enhanced Polar Outflow Probe (e-POP) Mission Overview. *Space Science Reviews*, *189*(1), Article 1. https://doi.org/10.1007/s11214-015-0135-1

Yin, S., Zhu, X., & Jing, C. (2014). Fault detection based on a robust one class support vector machine. *Neurocomputing*, *145*, 263–268. https://doi.org/10.1016/j.neucom.2014.05.035

Yu Gao, Tianshe Yang, Nan Xing, & Minqiang Xu. (2012). Fault detection and diagnosis for spacecraft using principal component analysis and support vector machines. *2012 7th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, 1984–1988. https://doi.org/10.1109/ICIEA.2012.6361054

Zhan, Y., Wan, P., Jiang, C., Pan, X., Chen, X., & Guo, S. (2020). Challenges and Solutions for the Satellite Tracking, Telemetry, and Command System. *IEEE Wireless Communications*, *27*(6), 12–18. https://doi.org/10.1109/MWC.001.2000089

Zhang, Q., Tian, Y., Wang, T., Yuan, F., & Xu, Q. (2015). ApproxEigen: An approximate computing technique for large-scale eigen-decomposition. *2015 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 824–830. https://doi.org/10.1109/ICCAD.2015.7372656

Zhou, J., Chan, K. L., Chong, V. F. H., & Krishnan, S. M. (2005). Extraction of Brain Tumor from MR Images Using One-Class Support Vector Machine. *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, 6411–6414. https://doi.org/10.1109/IEMBS.2005.1615965

Generalized Time-Series Analysis for In-Situ Spacecraft Observations: Anomaly Detection and Data Prioritization using Principal Components Analysis and Unsupervised Clustering

Matthew G. Finley[1,2], Miguel Martinez-Ledesma[2,3], William R. Paterson[2], Matthew R. Argall[4], David M. Miles[5], John. C. Dorelli[2], and Eftyhia Zesta[2]

[1]Department of Astronomy, University of Maryland, College Park, MD, USA.

[2]Geospace Physics Laboratory, NASA Goddard Space Flight Center, Greenbelt, MD, USA.

[3]Department of Physics, Catholic University of America, District of Columbia, USA.

[4]Space Science Center, University of New Hampshire, Durham, NH, USA.

[5]Department of Physics and Astronomy, University of Iowa, Iowa City, IA, USA.

Corresponding author: Matthew G. Finley (matthew.g.finley@nasa.gov)

Key Points:

- Spacecraft often cannot transmit all measurements to Earth at full cadence due to telemetry bandwidth limitations.
- Many missions must implement complex data prioritization schemes to ensure only the highest-priority data is transmitted at high cadence.
- The proposed data prioritization technique is highly generic, compatible with inexpensive hardware, and suitable for low-cost missions.

Abstract

In-situ spacecraft observations are critical to our study and understanding of the various phenomena that couple mass, momentum, and energy throughout near-Earth space and beyond. However, on-orbit telemetry constraints can severely limit the capability of spacecraft to transmit high-cadence data, and missions are often only able to telemeter a small percentage of their captured data at full rate. This presents a programmatic need to prioritize intervals with the highest probability of enabling the mission's science goals. Larger missions such as the Magnetospheric Multiscale mission (MMS) aim to solve this problem with a Scientist-In-The-Loop (SITL), where a domain expert flags intervals of time with potentially interesting data for high-cadence data downlink and subsequent study. Although suitable for some missions, the SITL solution is not always feasible, especially for low-cost missions such as CubeSats and NanoSats. This manuscript presents a generalizable method for the detection of anomalous data points in spacecraft observations, enabling rapid data prioritization without substantial computational overhead or the need for additional infrastructure on the ground. Specifically, Principal Components Analysis and One-Class Support Vector Machines are used to generate an alternative representation of the data and provide an indication, for each point, of the data's potential for scientific utility. The technique's performance and generalizability is demonstrated through application to intervals of observations, including magnetic field data and plasma moments, from the CASSIOPE e-POP/Swarm-Echo and MMS missions.

Plain Language Summary

Measurements captured by spacecraft are necessary to our understanding the space environment near Earth and throughout our solar system. However, spacecraft can often only transmit a small portion of the data they capture back to Earth. This means that many spacecraft must prioritize intervals of data that have the highest probability of helping to further our understanding of these environments. Some missions utilize humans, on Earth, to help select these scientifically important intervals. This solution, called the Scientist-In-The-Loop, can be too expensive or programmatically complex for many small missions to implement. This manuscript presents a technique for the detection of anomalous events in spaceflight measurements using statistical analysis and machine learning. These detected anomalies can be used to prioritize data that has a high probability of scientific relevance. Further, the proposed technique is highly generalizable and computationally lightweight, making it suitable for a variety of missions. Several case studies from multiple existing missions will be analyzed throughout this paper.

1 Introduction

Magnetic field sensors are one of the many science instruments that have been a fundamental part of space exploration since its inception. Some of the first satellites, such as the late-1950's Sputnik 3 and Explorer 6, carried fluxgate magnetometers to collect scientific data (Gordon & Brown, 1972). Since then, the science of magnetometry and spaceflight has been advanced such that we can make magnetic field measurements of far-flung bodies such as asteroids (Weiss et al., 2023), Mars (Connerney et al., 2015), Jupiter (Connerney et al., 2017), and the Sun (Bale et al., 2016). The need to understand fundamental physical processes in space, such as magnetic reconnection, has driven requirements for the telemetry of measurements at higher and higher cadences (Phan et al., 2016). Additionally, our desire to enable comprehensive understanding, nowcasting, and forecasting of Earth's near-space environment has led to the development of the inexpensive CubeSat form factor and a greater number of multi-spacecraft missions. CubeSats allow cost effective proliferation of measurements that can vastly improve our models through data assimilation and machine learning techniques. The last several years have seen the emergence of large constellations, leading to a multitude of challenges, particularly in the handling and telemetry of the massive quantities of raw data available onboard each spacecraft (Liddle et al., 2020; Zhan et al., 2020).

Many missions, due to programmatic constraints on telemetry rates, are unfortunately unable to downlink all of their captured data to Earth for analysis. Instead, mission operators and science teams must make decisions about which intervals of time to transmit high-cadence data (i.e., *burst data*). Some missions will transmit a lower-cadence data product (i.e., *survey data*) during intervals deemed less important (Lepping et al., 1995), and some missions will simply not telemeter these intervals (Yau & James, 2015). Intervals of burst data to be telemetered are typically determined by estimated spacecraft position, by humans on the ground diligently monitoring low-cadence data, or by carefully calibrated onboard algorithms which search the high-cadence data for mission-specific triggers.

The Magnetospheric Multiscale (MMS) mission (Burch et al., 2016) utilizes a particularly thorough approach to the identification and prioritization of burst data to telemeter to the ground (Baker et al., 2016). During spacecraft traversal through predetermined regions of interest, the MMS instruments always capture data at their high cadence burst rates. This high-cadence data is stored onboard while the lower-cadence survey data is telemetered to the ground and analyzed. If the survey data shows potentially interesting phenomena, short intervals of burst data can be downlinked from the spacecraft. The MMS mission prioritizes burst data using two techniques: the Scientist in the Loop (SITL) and the Automated Burst System (ABS).

The ABS, as its name indicates, runs automatically onboard each spacecraft and provides a data ranking metric to be downlinked alongside the survey data (Baker et al., 2016). This system uses data quality indicators calculated by each instrument to rank the available burst data in a downlink prioritization queue. The last item in the queue will be the first to be overwritten should a higher-ranking interval be identified. Although 34 data quality indicators are available for burst triggering, the early mission used only large gradients in the Z_{GSM}-component of the measured magnetic field to prioritize burst intervals while the data quality indicators were characterized. After two years of careful parameterization, ~6 data quality indicators are now used by the ABS for prioritization of burst data containing mission-specific phenomena of interest (Argall et al., 2020).

The SITL is a manual option that can validate or override the selections made by the ABS. A domain expert – with access to MMS survey data, spacecraft-calculated trigger metadata, and data from other satellites or ground systems – determines the priority of data for downlink using specialized software, ensuring that data with high scientific significance is telemetered (Baker et al., 2016).

Both the ABS and SITL burst selection schemes for data prioritization require substantial scientific infrastructure and potentially costly overhead in their implementation. Extensive onboard triggering logic, years of parameter characterization, dedicated interval labeling time from experts during in-situ mission operations, and bespoke burst prioritization software are almost certainly infeasible for low-budget missions.

One method has been recently proposed which aims to reduce the need for infrastructure-intensive SITL activities on MMS using supervised machine learning (Argall et al., 2020). Although their technique shows excellent results and enables a great reduction in the reliance on manual labeling tasks, it requires a large set of expert-labeled data during the training of their segmentation network and therefore does not eliminate the need for the SITL infrastructure entirely.

This manuscript proposes the use of a common dimensionality reduction technique, coupled with unsupervised clustering, to provide a robust and generalizable method for detecting anomalous intervals of time series spacecraft observations. This method is intended for use as a component of a drop-in burst data prioritization system for missions where the infrastructure and cost associated with more sophisticated and mission-specific solutions are not feasible. Although the SpaceX Starlink constellation currently dominates the Low Earth Orbit environment with over 2,000 satellites currently in orbit and approval granted for 12,000 total satellites, it is unlikely to remain the only major constellation in orbit (Ma et al., 2023; McDowell, 2020). These mega-constellations are a heavy burden on ground systems, requiring complex protocols for dealing with telemetry, command, and tracking (Berner, 2019). The proposed tool would be an invaluable asset for such constellation missions, enabling a higher degree of distributed autonomy in their space operations.

The following sections of this manuscript describe the proposed technique and demonstrate its performance with several case studies on observational measurements obtained from the CASSIOPE e-POP/Swarm-Echo spacecraft and one of the Magnetospheric Multiscale mission satellites.

2 Methodology

2.1 *Dimensionality Reduction via Principal Components Analysis*

Principal Components Analysis (PCA) is one of the oldest and most popular multivariate statistical analysis techniques used to reduce the dimensionality of large datasets (Jolliffe & Cadima, 2016). Mathematically, PCA is performed by identifying the eigenvectors of the covariance matrix associated with the data matrix under observation ($\boldsymbol{X}$) via

$$\boldsymbol{X}^T\boldsymbol{X} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{\boldsymbol{T}} \tag{1}$$

where the columns of $\boldsymbol{V}$ correspond to the eigenvectors and the diagonal elements of $\boldsymbol{\Lambda}$ are the associated eigenvalues. For convenience, let each eigenvector V_i be ordered by the magnitude of its associated eigenvalue.

The projection and subsequent dimensionality reduction can be realized through

$$\boldsymbol{P} = \boldsymbol{X}\mathbf{V}_{\boldsymbol{q}} \tag{2}$$

where $\boldsymbol{V_q}$ is a matrix whose columns are only the first q eigenvectors from $\boldsymbol{V}$. Throughout this manuscript, the dimensionality of the output projection is fixed to two (i.e., $q = 2$) in order to reduce the computational complexity associated with the analysis of the projected data.

The specific data matrix being analyzed in this manuscript is generated by concatenating R consecutive time intervals, of length L, from the original time series (x, with length N) into a single matrix via

$$\boldsymbol{X} = \begin{bmatrix} x(1) & x(L+1) & \dots & x(RL-L+1) \\ \dots & \dots & \dots & \dots \\ x(L) & x(2L) & \dots & x(RL) \end{bmatrix}. \tag{3}$$

It should be noted that $\boldsymbol{X}$ is a reduced version of the trajectory matrix passed into PCA by the Singular Spectrum Analysis (SSA) technique (Finley et al., 2023). Both SSA's trajectory matrix and the data matrix used throughout this paper are constructed to enable information about the temporal variation of a signal to be utilized. The reduced trajectory matrix (i.e., no overlapping samples in each consecutive column) was used here to reduce the computational complexity associated with the overall algorithm.

An example of this variation on PCA, applied to one axis of the vector magnetometer data captured by the CASSIOPE e-POP/Swarm-Echo magnetic field instrument (Wallis et al., 2015; Yau & James, 2015), is shown in Fig. 1. Figure 1(a) illustrates the 2.5-minute interval of data, at a sampling rate of 160 Hz, to be processed. Figure 2(b) shows the two-dimensional representation, given by $\boldsymbol{P}$, of the data following the application of PCA (with q = 2) on a data matrix constructed from 0.5-second consecutive intervals taken from the signal in Fig. 1(a).

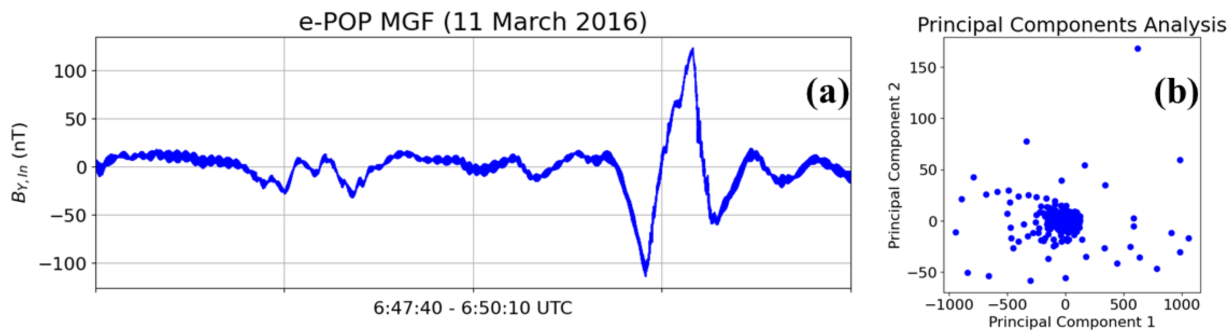

Figure 1: Dimensionality reduction via application of PCA on concatenated 0.5-second intervals of magnetic field data. (a) Inboard magnetometer data from CASSIOPE e-POP/Swarm-Echo MGF; (b) Two-dimensional representation of the 0.5-second intervals of (a) following PCA while retaining only two principal components.

Each point in Fig. 1(b) is a projection of one of the 0.5-second intervals of Fig. 1(a). It can be seen that many of these points are clustered tightly near the origin, whereas some of the points are outlying near the periphery. This implies that the majority of the time intervals exhibit similar behavior when represented using only the first two principal components (i.e., those that describe the largest variance in the original data matrix). However, some intervals show very different behavior in terms of these maximum-variance components. Automatic clustering of this two-dimensional representation should reveal anomalous time intervals in the original signal and is discussed in detail in Sec. 2.2.

2.2 *Clustering with One-Class Support Vector Machine*

Machine learning techniques have become increasingly popular in the various space physics research domains. Successful application of these techniques has been seen in methods for auroral image classification (Clausen & Nickisch, 2018), recreating magnetohydrodynamic environments from sparse sample spaces (Bard & Dorelli, 2021), space weather forecasting (Camporeale, 2019), and many others. However, machine learning models can often be complex, requiring large quantities of training data and computational resources. Once trained, these large and complex models are often treated as 'black boxes,' and can lack interpretability (Angelov et al., 2021). To increase potential applicability to low-cost and in-situ spaceflight hardware, a machine learning-based clustering solution that is computationally efficient and easily understood must be utilized instead of a more complex model.

Support Vector Machines (SVMs) are a popular means of performing classification tasks throughout a variety of fields including the biomedical sciences (Zhou et al., 2005) and industrial engineering (Shin et al., 2005). This data labeling technique has seen widespread adoption due to its high degree of robustness and interpretability (Hearst et al., 1998). Traditional SVMs are trained by first projecting the labeled training data to a higher dimension feature space using a user-selected *kernel*. Next, a hyperplane that best separates the classes is calculated, although a slack parameter is considered in this optimization. This slack parameter enables the trained SVM to handle a small number of data points that cannot be separated using a hyperplane in the higher-dimension feature space, which is a common situation in real-world datasets (Noble, 2006). This trained SVM can now be used to classify new data not seen during the training process.

Slight modification of the traditional SVM framework leads to a technique known as the One-Class Support Vector Machine (OC-SVM), a common unsupervised approach to data classification and anomaly detection (Yin et al., 2014). These OC-SVMs operate in a similar fashion to the traditional SVM but calculate a hyperplane that optimally separates the data from the origin, not by separating pre-labeled classes (Amer et al., 2013). Here, the primary user-defined control parameter is ν, which lies in the range (0,1] and determines the upper bound on the number of allowed errors and a lower bound on the number of data points used when calculating the separating hyperplane (Chang & Lin, 2001).

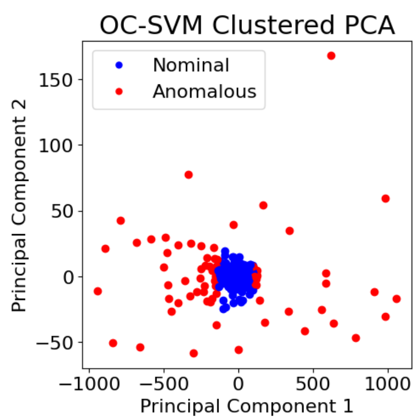

Figure 2: One-Class Support Vector Machine clustering applied to the reduced-dimension data illustrated in Fig. 1.

Figure 2 illustrates the result of passing the two-dimensional output of PCA, shown in Fig. 1(b), through an OC-SVM. This OC-SVM was trained with a Gaussian kernel and a ν-value of 0.3. Points shown in red were those that were considered anomalous, whereas the points shown in blue were considered nominal. Since each point in Fig. 2 represents a 0.5-second interval of the original input shown in Fig. 1(a), the associated labels can be directly applied to each interval in the original input time series. The result of this inversion procedure, and additional examples, will be discussed in detail in Sec. 4.

3 Data and Preprocessing

3.1 *CASSIOPE e-POP/Swarm-Echo MGF*

One primary source of the data analyzed in previous sections and throughout the remainder of this manuscript is the CASSIOPE/Swarm-Echo Magnetic Field instrument (Wallis et al., 2015; Yau & James, 2015). The Magnetic Field instrument (MGF) consists of a pair of identical fluxgate magnetometers mounted inline on a single boom at approximately 0.6 m and 0.9 m from the body of the spacecraft. Both magnetometers capture the local magnetic field at a rate of 160 Hz and downlink the data when telemetry constraints allow. Although the magnetometer pair (i.e., gradiometer) could be used to mitigate local interference from the host spacecraft and improve data fidelity (Finley et al., 2023), the measurements used in this manuscript were taken from only the inboard magnetometer mounted closer to the spacecraft. This provides some insight into how the proposed anomaly detection technique handles data contaminated by local interference, which is almost constantly observed at the magnetometers as high-frequency oscillations caused primarily by the spacecraft's attitude control systems (i.e., reaction wheels). The magnetometer measurements used in this manuscript have had a near-DC baseline removed using a 20-s moving average prior to analysis and visualization.

3.2 *Magnetospheric Multiscale Mission FGM*

Another source of magnetic field data used in the remaining sections of this manuscript is the Magnetospheric Multiscale mission (MMS) Magnetometers (Burch et al., 2016; Russell et al., 2016). Although data is available from all four MMS satellites, only data from MMS1 was utilized throughout this manuscript. The MMS Magnetometers consist of a near-identical pair of fluxgate sensors, with each sensor mounted at the end of two separate 5-meter booms. The measured data is reported as a high-fidelity triaxial vector data product (called FGM) with three possible sampling rates: slow survey at 8 Hz; fast survey at 16 Hz; and, burst data at 128 Hz. In this manuscript, only fast survey and slow survey data were utilized. If both fast and slow survey data were present in the time interval to be analyzed, the data was resampled to match the slow survey data rate for consistency of analysis. As with the CASSIOPE data discussed in Sec. 3.1., a 20-s moving average has been removed from the MMS data prior to analysis and visualization.

3.3 *Magnetospheric Multiscale Mission FPI*

An additional set of data used in the remaining sections of this manuscript is the MMS Fast Plasma Investigation (Burch et al., 2016; Pollock et al., 2016). As with the magnetic field data, only data from the MMS1 satellite was used in this manuscript. The Fast Plasma Investigation (FPI) for MMS comprises multiple top-hat electrostatic analyzers (Carlson et al., 1982) to determine in situ the fluxes of electrons and ions as functions of energy and direction. The FPI, its measurements, and methods of computation are described in detail in (Pollock et al., 2016). FPI acquires a full 3D set of electron samples (32 energies × 32 azimuths × 16 polar sections) every 30 ms, and an equivalent set of ion samples every 150 ms. When telemetered to the ground, these samples are employed to compute electron and ion fluid parameters such as number density, bulk-flow velocity, pressure, and others. Each parameter is computed as a summation of fluxes weighted by appropriate physical factors. These parameters are essential quantities required for many scientific analyses of space plasmas, for example in comparisons of observations with the output of magnetohydrodynamic (MHD) simulations that predict the overall behavior of plasma as a fluid.

The ion parameters (i.e., number density and velocity) shown in this manuscript are reported at the Fast Survey rate of ~0.22 Hz (i.e., captured at full cadence and averaged once every 4.5 s).

For the FPI, approximate fluid-like parameters are also computed in a simplified fashion by the instrument processor onboard the spacecraft. The computation is approximate because the summations are not weighted by the proper factors necessary to obtain true physical quantities. However, it has been shown that these onboard quantities, known as pseudo-moments, can be rescaled to serve as proxies for the true physical parameters (Barrie et al., 2018). As a result, analysis of the derived plasma moments using the proposed anomaly detection technique is indicative of the technique's performance when applied to the onboard pseudo-moments.

3.3 *Magnetospheric Multiscale Mission SITL*

To verify the capability of the proposed technique to identify intervals of time relevant to the MMS mission's science objectives, this manuscript also includes data from the MMS Scientist-in-the-Loop (SITL) report. MMS is the first mission with both the spatial and temporal resolution to resolve electron-scale dynamics. This requires MMS to capture much more data than it can telemeter to the ground. As discussed in Sec. 1, this has led to the development of MMS' burst management system, consisting of the Automated Burst System (ABS), Scientist-in-the-Loop (SITL), and the Ground Loop System (GLS). The SITL, specifically, is a role passed among MMS team member volunteers who search through the survey data to identify and select time intervals that may contain relevant events for burst-mode downlink. Survey data has insufficient resolution to capture electron-scale dynamics, so the SITL must over-select events to ensure that mission-critical science data is captured. The SITL is guided by mission-level science objectives that enable assignment of a figure of merit (FOM) value to each interval examined (Argall et al., 2020; Hasegawa et al., 2023; Phan et al., 2016). These FOMs subsequently inform the mission of the highest-priority data to be selected and downlinked at burst rate. Each SITL selection is accompanied by a short description that is searchable and parsable. These descriptions have previously been used to train a supervised machine learning model that uses the SITL report to make future predictions about which data should be selected for downlink (Argall et al., 2020). This model is installed in the near real-time data processing GLS so that predictions can be made as soon as the preliminary low-cadence data is downlinked in order to guide the SITL (along with the ABS selections). The SITL data itself is not publicly available, as it is not considered science-quality, but the reports describing the selected events and their time range can be searched through the MMS Mission Events webpage (https://lasp.colorado.edu/mms/sdc/public/about/events/#/). Additional tools, such as PyMMS, have been developed that enable rapid searching of these reports (Argall et al., 2022).

4 Results

4.1 *Observations of Current Sheets with Embedded Alfvén Waves (CASSIOPE)*

Section 2 illustrated (through Fig. 1 and Fig. 2, respectively) the dimensionality reduction and unsupervised clustering techniques utilized in the proposed method of automated anomaly detection by applying them to a short interval of magnetic field data captured by the CASSIOPE e-POP/Swarm-Echo inboard magnetometer. Figure 3(a) shows the proposed method's output after inverting the labels given to each point shown in Fig. 2 back onto the original magnetometer data shown in Fig. 1(a), as described in Sec. 2.2. Points in blue are those considered nominal by the technique, whereas points labeled red have been flagged as anomalous or potentially scientifically relevant. Figure 3(b) provides the spectral content for the interval shown in Fig. 3(a). The spectral content and time series for this 2.5-minute interval shows previously identified Alfvénic activity embedded in a large current sheet at roughly 6:48:55 UTC (Miles et al., 2018), with a smaller current sheet observed near 6:48:10 UTC. The current sheet containing embedded Alfvénic activity is highlighted in gray in Fig. 3(a). It can be seen that the proposed method of anomaly detection was able to accurately identify these interesting intervals containing broadband magnetic activity, even in the presence of the potentially obfuscating reaction wheel interference observed at ~15 Hz.

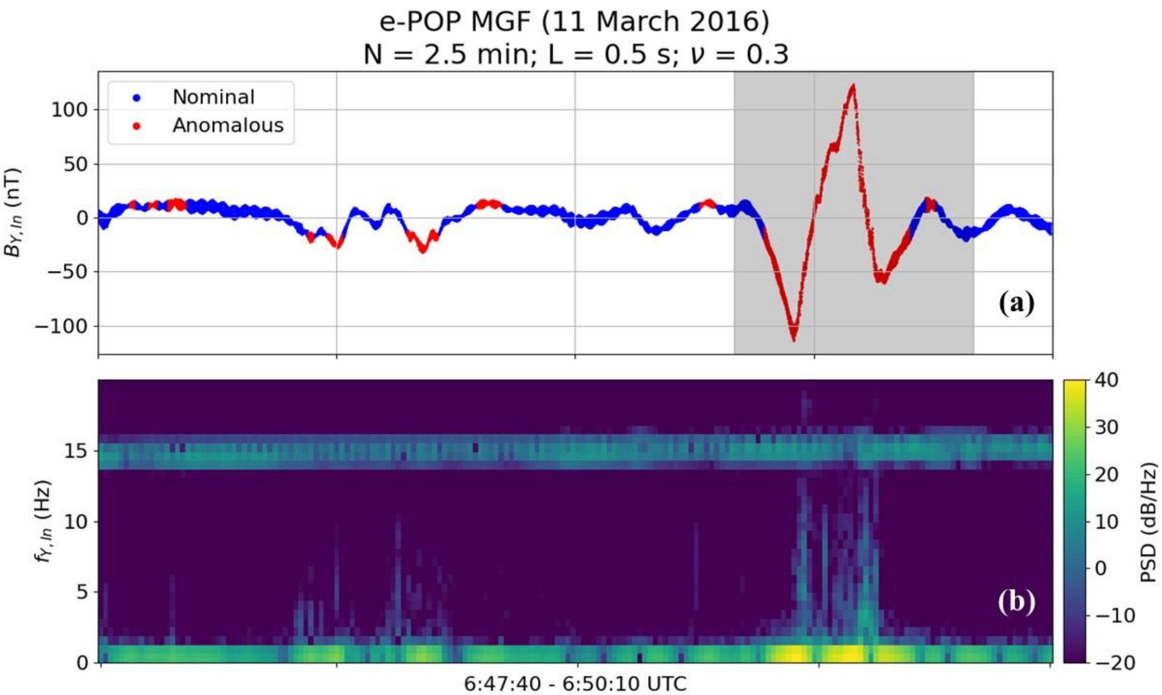

Figure 3: Proposed method of anomaly detection applied to CASSIOPE e-POP/Swarm-Echo MGF data containing Alfvénic activity. (a) Time series with anomalous intervals plotted in red, nominal regions plotted in blue; (b) Spectral content of the interval shown in (a).

4.2 *Observations of Spacecraft Maneuvers (CASSIOPE)*

The next interval of data analyzed in this section is a 20-minute interval of CASSIOPE e-POP/Swarm-Echo magnetometer data captured when the spacecraft was performing maneuvers. As a result, the spacecraft's four reaction wheels were changing their spin frequency rapidly during this period. Figure 4(a) shows the labeled time series output by the method with a window length of five seconds. Anomalous intervals are shown in red, normal points are shown in blue. Figure 4(b) shows the spectral content of the interval shown in Fig. 4(a). Highly dynamic high-frequency activity resulting from the reaction wheels' diverging operational rates can be seen for a five-minute interval starting at approximately 4:55:00 UTC, with other low-frequency perturbations occurring throughout the total 20-minute interval. It can be seen that the detected anomalies directly correspond to changes in behavior in the measured magnetic field due to the spacecraft maneuvers. This example helps to both illustrate the ability of the proposed technique to simultaneously identify dynamic behavior in both high- and low-frequency bands, as well as identify intervals critical to spacecraft operations (i.e., spacecraft maneuvers).

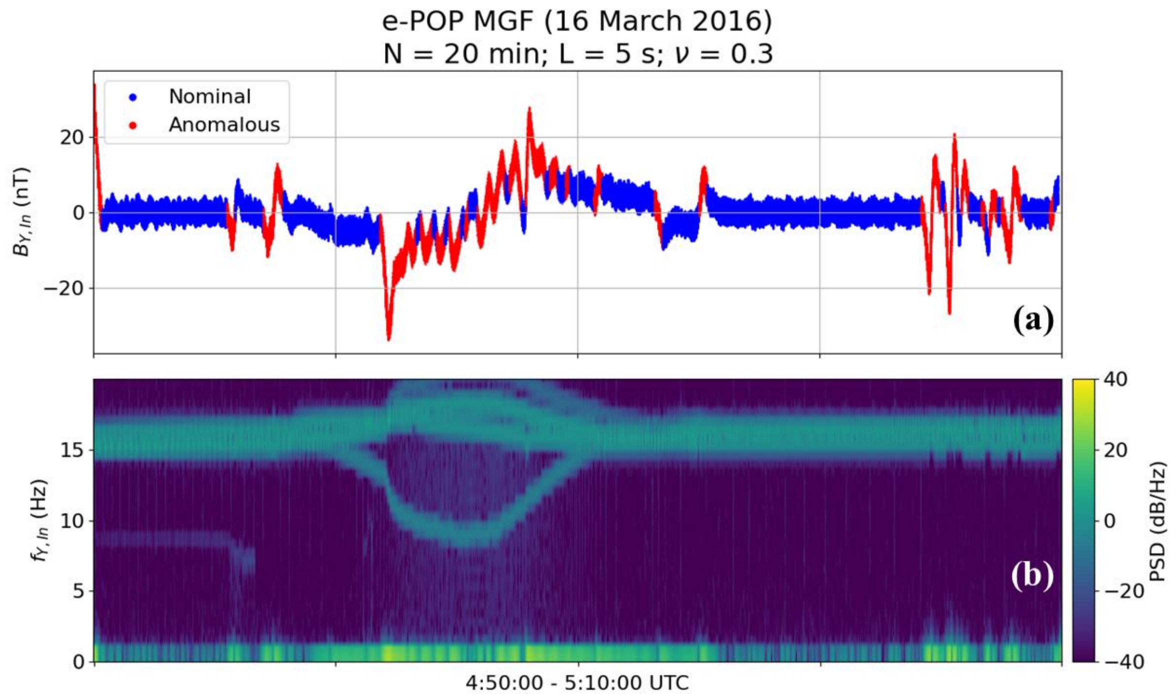

Figure 4: Proposed method of anomaly detection applied to interval of CASSIOPE e-POP/Swarm Echo MGF data containing highly dynamic high-frequency signatures caused by spacecraft reaction wheels. (a) Time series with anomalous intervals plotted in red, nominal in blue. (b) Spectrogram associated with the data in (a).

4.3 *Observations of Interplanetary Shocks and EMIC Waves (MMS)*

The next period of data analyzed in this manuscript is a full 24 hours of magnetic field data captured by the MMS magnetometer suite in December 2015. Figure 5(a) shows the result of the proposed method of anomaly detection applied to the full day of data with a window length (L) of 5 minutes. As in the previous examples, points shown in red correspond to detected anomalies and points shown in blue have been identified as nominal. Figure 5(b) shows the spectrogram associated with the magnetic field measurements in Fig. 5(a). Note that this data has been resampled to match the slowest sampling rate present in the interval (i.e., 8 Hz), as described in Sec. 3.2. The frequent broadband magnetic phenomena seen throughout the first ~40% of this 24-hour interval is attributed, by the MMS SITL, to observations of the bow shock and magnetopause. The broadband magnetic activity seen in the last ~15% of the day was similarly reported by SITL to correspond to observations of the magnetopause. The large-amplitude (i.e., > 200 nT) phenomena at approximately 17:00 UTC corresponds to the perigee of the MMS1 spacecraft. This day of data, which was analyzed extensively by Engebretson et al. (2018), also contains MMS' observations prior to, during, and after an interplanetary shock. The resultant compression of the magnetosheath was observed by MMS at approximately 13:24 UTC, with structured EMIC wave activity occurring before and after. The gray highlighted region in Fig. 5(a) corresponds to the region of interest (from 13:00-14:00 UTC) containing the majority of this activity (Engebretson et al., 2018). It can be seen that the proposed method of anomaly detection is able to accurately label a significant portion of the interesting magnetic field data occurring at bow shock and magnetopause crossings before noon and near midnight. In addition, the technique also labels several of the large-amplitude perturbations in the magnetic field measurement near and during the 13:00-14:00 UTC region of interest. Figures 6(a) and 6(b) provide a zoomed view of this region's time series and spectral content, respectively. No change has been made to the anomaly labeling, which still corresponds to the result when the proposed method is applied to the full 24-hour period. It is important to note that although the proposed technique has successfully identified the broadband signal corresponding to the magnetosheath compression, much of the structured EMIC wave activity occurring during this region of interest has not been labeled as anomalous. This result is obtained due to the input parameters selected for this example. The next example, shown in Fig. 7, will illustrate the effect of varying the input parameters when the anomaly detection technique is applied.

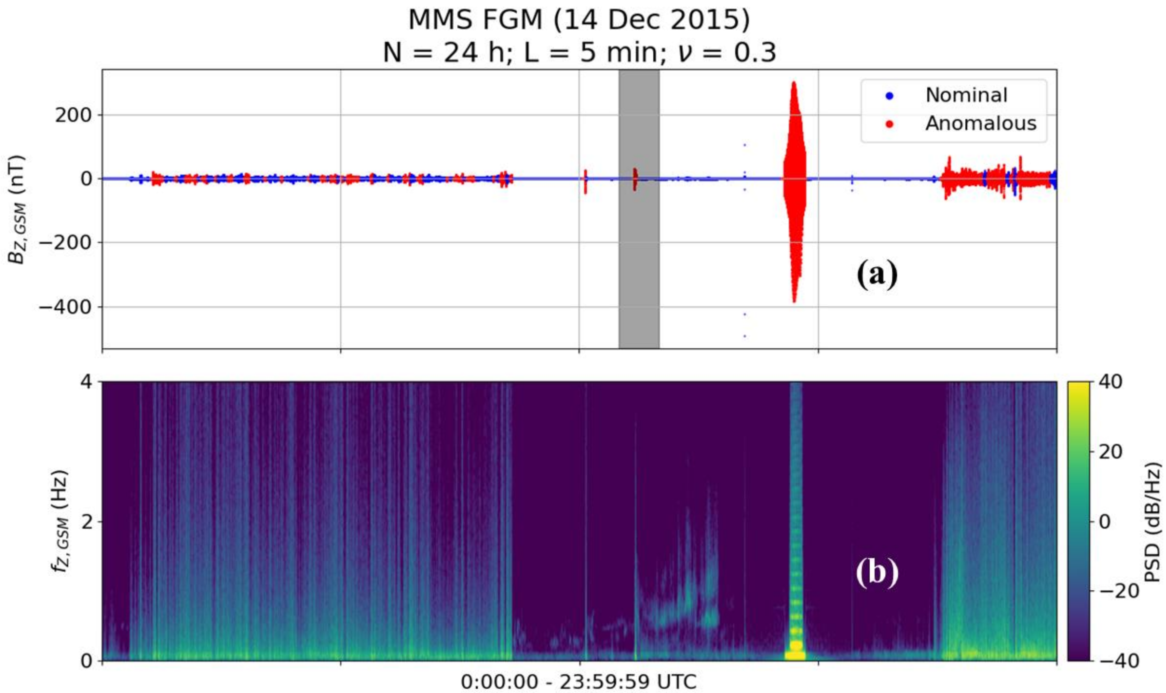

Figure 5: Anomaly detection technique applied to a 24-hour interval of MMS FGM data containing a variety of scientifically interesting phenomena including observations of the bow shock, magnetopause, and compressions of the magnetopause due to an interplanetary shock. (a) Time series FGM measurements with anomalous intervals identified by the proposed technique plotted in red, nominal intervals plotted in blue. (b) Spectrogram associated with the data shown in (a).

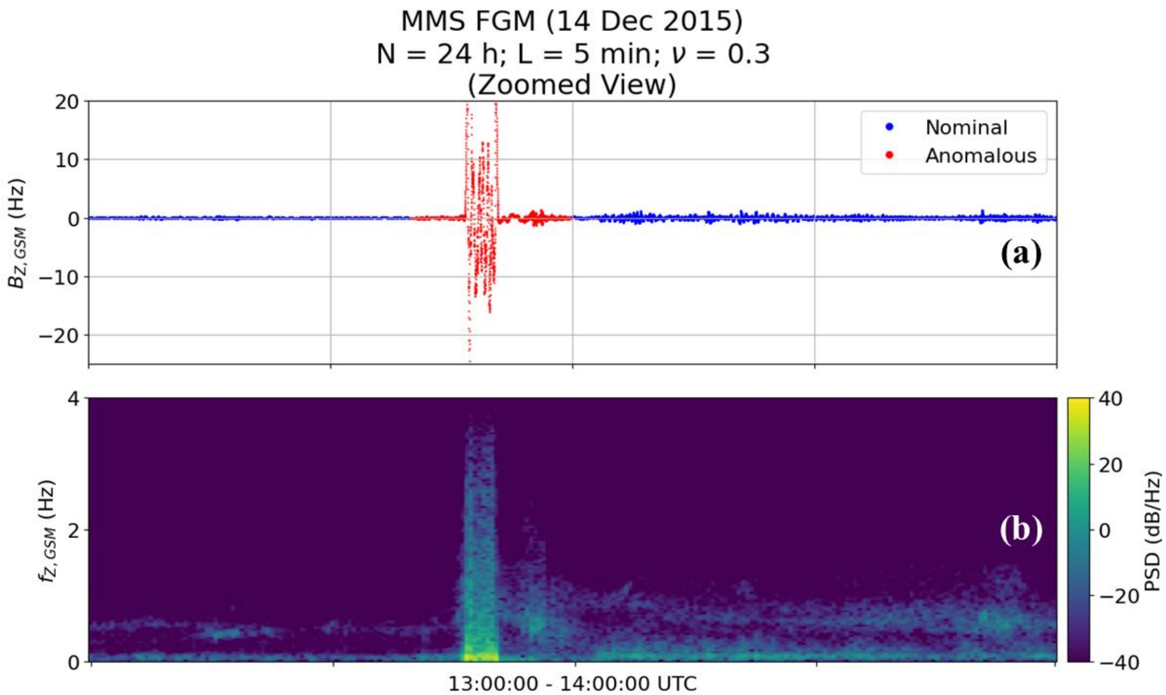

Figure 6: Zoomed view of the gray-highlighted region shown in Fig. 5, with previously identified EMIC activity. (a) Time series FGM measurement with anomalous points shown in red and nominal points shown in blue. (b) Spectral content corresponding to the interval shown in (a).

Figure 7 illustrates the impact of changing the signal length (N) and window length (L) on the performance of this technique. Specifically, Figure 7(a) shows the labeling output by the proposed method of anomaly detection when applied to only the 13:00-14:00 UTC region of interest highlighted in Fig. 5(a) and shown in Fig. 6(a). Note that this impacts the signal length, N, without changing the window length, L. Although more of the structured wave activity in this region is correctly labeled as anomalous, this result may still not be sufficient for some applications. Reducing the window length (L) from five minutes to one second, as seen in Fig. 7(b), provides a more detailed labeling of the anomalous samples within the total interval. As a result, much more of the structured EMIC wave activity occurring after the magnetosheath incursion is identified as anomalous. Although the parameter space inherent to the proposed method of anomaly detection is small, these results clearly show the relevance of the input parameters to the method's result. Selection of these parameters, and their effect on the method's computational complexity, will be discussed in greater detail in Sec. 5.1.

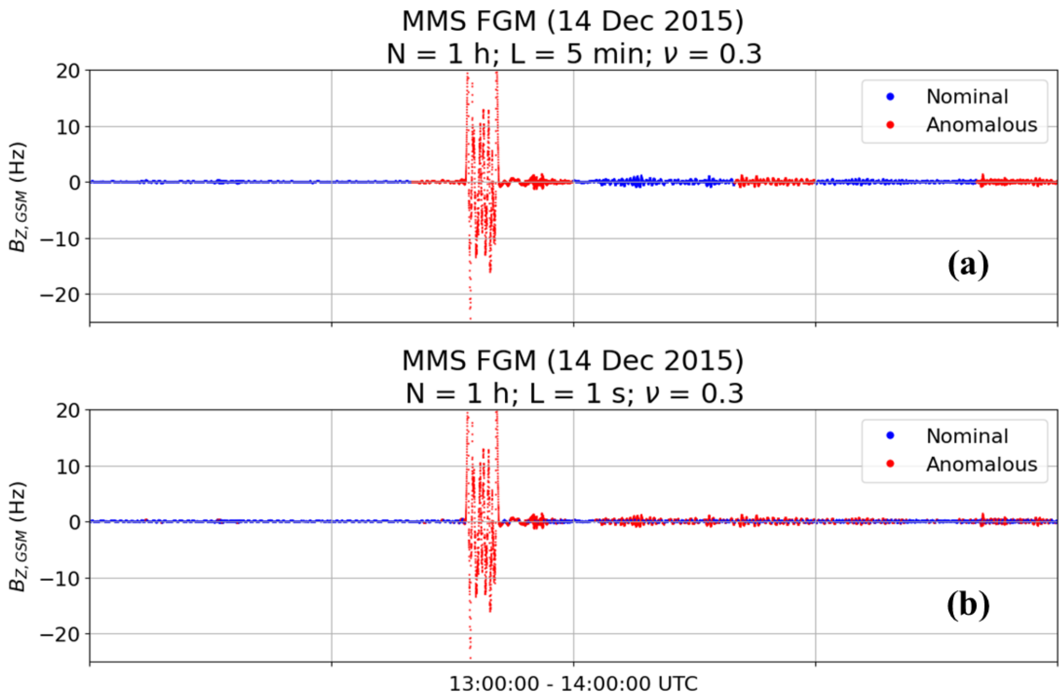

Figure 7: Illustration of the impact of the user-defined input parameters on the output of the proposed anomaly detection, shown for the same data seen in Fig. 6. (a) Time series output by the anomaly detection technique for a signal length of one hour and a window length of five minutes. (b) Time series output by the anomaly detection technique for a signal length of one hour and a window length of one second.

4.4 *Observations of Magnetopause Crossings and the Bow Shock (MMS)*

The next experiment in this manuscript demonstrates the applicability of the proposed anomaly detection technique to spacecraft observations other than magnetic field measurements. Specifically, the technique is applied to a large interval of magnetic field data, as well as ion number density and ion velocity, as measured by the MMS spacecraft on 15 May 2023. Figure 8(a) shows a 24-hour period of magnetic field measurements, Fig. 8(b) and Fig. 8(c) show the ion number density and ion velocity corresponding to the same period. The points illustrated in red are those considered anomalous by the proposed technique when it was applied to the total interval, whereas points shown in blue are considered nominal. It should be noted that the last 3.75 hours of ion measurements were not available because the spacecraft transitioned from Fast Survey to Slow Survey rate, and the FPI does not operate during Slow Survey mode. This period of missing data has been padded with nominal-labeled zeroes, after the anomaly detection algorithm was applied, for consistency with the magnetic field data. The gray region highlighted in Fig. 8 corresponds to a region of data containing various physical phenomena identified by the MMS SITL, such as observations of the bow shock, magnetopause, and boundary layer crossings. These SITL identifications, including the time range and the associated Figure of Merit (FOM) are detailed in Tab. 1. A higher FOM is assigned to a selection that should be downlinked at higher priority, with a maximum value of 199 being assigned to typical events, while higher FOMs are reserved for calibration intervals or other special events (Argall et al., 2020).

Table 1: MMS SITL report for 15 May 2023. Columns 1-2 provide the start and stop time for each interval of interest. Column 3 shows the Figure of Merit (FOM) associated with each interval. Column 4 shows SITL remarks for each interval.

Start Time (UTC)	Stop Time (UTC)	Figure of Merit (FOM)	Discussion
12:49:43	13:17:33	100	Partial Bow Shock
13:25:33	13:32:43	75	Partial Bow Shock
13:37:33	13:46:43	75	String of Partial Bow Shock Crossings
15:34:03	15:58:13	90	Partial Magnetopause
16:24:53	16:49:43	150	Full Medium Shear Magnetopause
17:51:43	17:56:13	55	Boundary Layer Traversal

In addition to these SITL-identified phenomena, the magnetic field data shows frequent broadband perturbations throughout the first half of the day, as well as a large data spike occurring near midnight that corresponds to a perigee pass of the MMS satellite. It can be seen that, for all three sets of input data, the anomaly detection technique successfully identifies much of the gray-highlighted region selected by the SITL as anomalous, but with a significant number of false positives shown for the magnetic field data when compared to only the SITL selections. These false positives in relationship to the SITL selection can be attributed to successful identification of the broadband perturbations of the magnetic field during the first half of the day. This number of false positives is greatly reduced in the labels associated with both the ion density and velocity, implying that a combination of parameters classified by the proposed technique may be used in the identification of only phenomena similar to those prioritized by the MMS SITL.

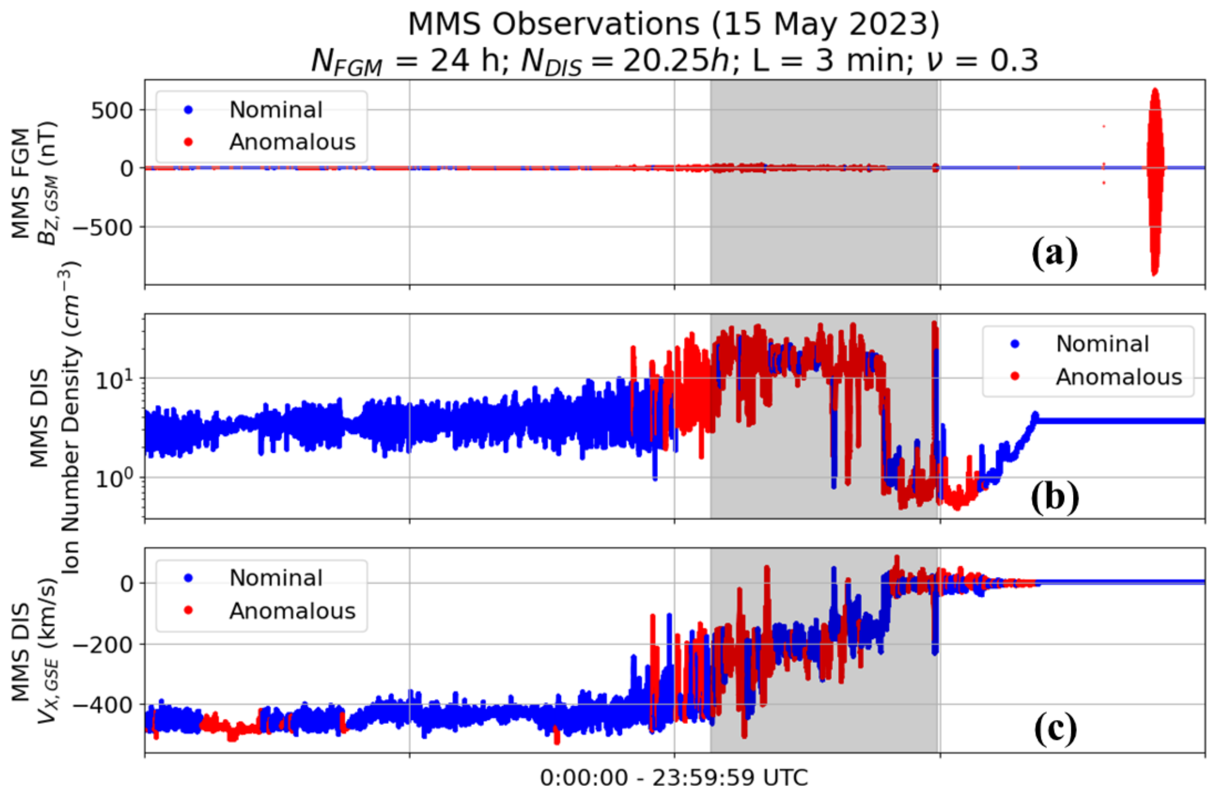

Figure 8: Anomaly detection technique applied to a long interval of MMS FGM and FPI measurements. This interval contains SITL-identified observations of the bow shock, magnetopause, and boundary layer crossings. Points highlighted in red are considered anomalous, points highlighted in blue are considered nominal. (a) FGM magnetic field measurements. (b) FPI ion number density measurements. (c) FPI ion velocity measurements.

Figure 9 provides a zoomed view of only the gray-highlighted region shown in Fig. 8. As before, Fig. 9(a) – Fig. 9(c) show the magnetic field, the ion number density, and the ion velocity measured by MMS. The nominal and anomalous labels assigned by the proposed anomaly detection technique correspond to the blue and red points, respectively. These labels are identical to the labels seen in Fig. 8; only the time scale of the plot has been changed to visualize the technique's output more clearly during the intervals selected by SITL. The regions highlighted in gray in Fig. 9 correspond to the specific time periods identified by SITL (described in Tab. 1) as containing observations of the bow shock, magnetopause, and boundary layer crossings. In a similar trend to the larger period shown in Fig. 8, the magnetic field data contains a greater number of false positive identifications of anomalous events when compared to only the SITL selections, and the ion moments, or combinations of the ion moments and magnetic field, may provide a greater degree of accuracy for this case study.

The example shown in Fig. 8 and Fig. 9 clearly demonstrates the generic applicability of the anomaly detection algorithm to a variety of spaceflight data products. Additionally, it shows that greater utility may be leveraged from the proposed method by analyzing its output when applied to several data products. For example, a weighting scheme generated by some logical operation of the binary labels (nominal or anomalous) associated with the method's output, when applied to multiple data products, may provide a reduced number of potential false positives depending on the target application. This is left as a potential topic for future study, and Section 5 will discuss several other avenues for future work utilizing this technique.

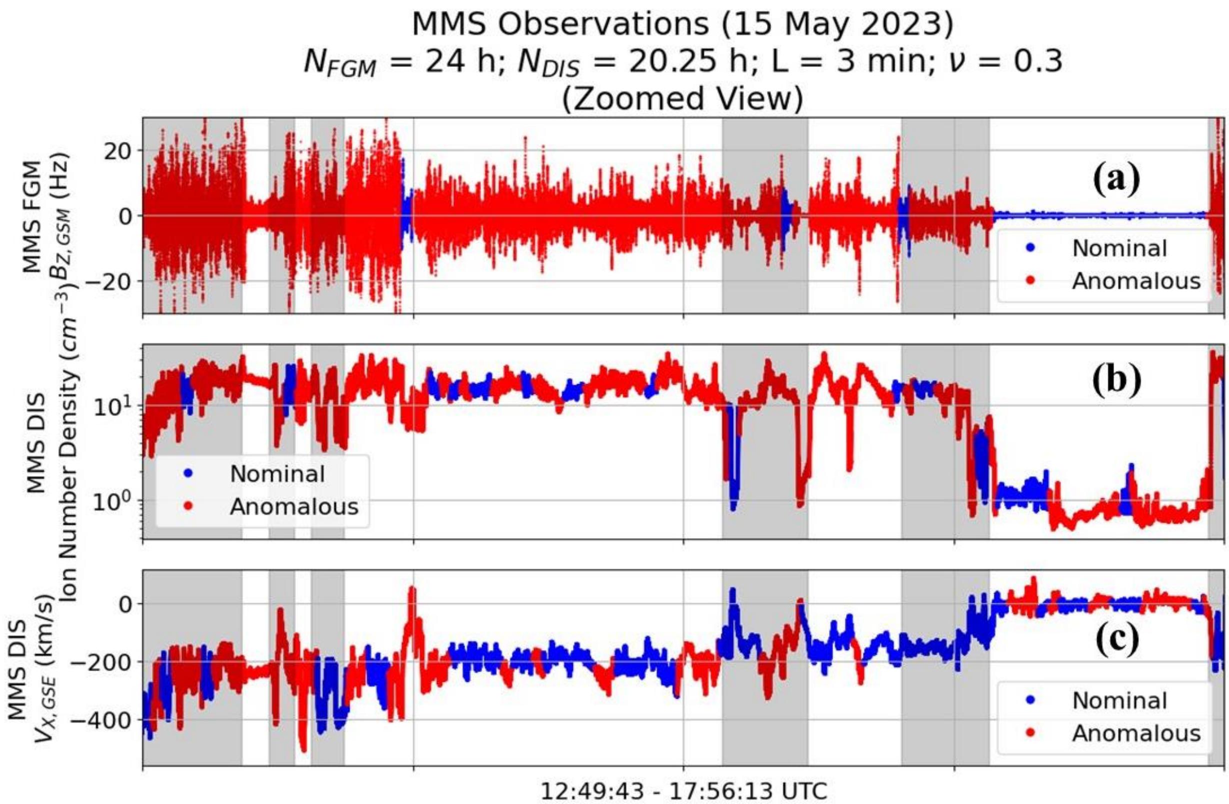

Figure 9: Zoomed view of the region of interest shown in Fig. 8., with SITL-identified observations of the bow shock, magnetopause, and boundary layer crossings. (a) FGM magnetic field data. (b) FPI ion number density. (c) FPI ion velocity.

5 Discussion & Future Work

5.1 *Parameter Selection & Computational Complexity*

This manuscript has demonstrated the capabilities of the proposed method of anomaly detection to analyze data from a variety of instruments and successfully identify various interesting phenomena; however, the example shown in Fig. 5 through Fig. 7 has illustrated the need for appropriate parameter selection to enable the highest possible scientific return. The proposed technique has a small parameter space, with only three variables impacting the output result. Thus, each of these parameters plays a critical role in the outcome of the analysis, as well is its computational complexity.

The input signal length (N) determines the total length of data to be analyzed when determining anomalous activity. The window length, L, determines the length of the consecutive observations used as features when performing PCA. Finally, the tuning parameter ν used in the OC-SVM is, practically, the fraction of the consecutive observations that can be considered anomalous when the data is clustered. In practice, this means that the input signal length (N) must be large enough that the majority of the data (i.e., at least the fraction given by $1 - \nu$) should be considered nominal, based on the target application. The window length must also be carefully considered when applying the technique as it corresponds to the scale on which anomalies are detected. If large, slowly varying changes are to be identified, a longer window length might be utilized; if small, rapid changes must be flagged, a shorter window length may be more suitable.

Changes in these parameters also have a significant impact on the computational complexity of the previously described technique. PCA, which is one of the fundamental building blocks for this algorithm, has two basic steps: computation of the covariance matrix, which has a complexity of $O(L^2R)$ for a given window length of L and a given number of total consecutive observations, R; and, eigendecomposition of the covariance matrix, which has a complexity of up to $O(R^3)$ in the worst-case scenario (Zhang et al., 2015). For the OC-SVM, the most computationally expensive component is the model training, with a complexity of up to $O(R^3)$. However, this complexity has been proven to be reduced by approximately one order of magnitude using more complex techniques such as sequential minimal optimization (Kang et al., 2019).

It can be seen that the window length, the number of consecutive observation intervals, and the OC-SVM tuning parameter have a direct impact on the computational complexity, or the detected anomalies associated with the proposed algorithm. As a result, users must tune these parameters as appropriate to suite the capabilities of their hardware and the data being processed.

5.2 *Potential for Implementation on Low-Cost Hardware*

Although PCA has historically been one of the most popular statistical analysis techniques used to reduce the dimensionality of large datasets (Jolliffe & Cadima, 2016), until recently relatively few studies have evaluated the potential for PCA’s implementation on embedded hardware (Korat & Alimohammad, 2019). Early efforts to perform PCA on embedded hardware relied on bespoke Very Large Scale Integration (VLSI) integrated circuits with complex architecture (Tung-Chien Chen et al., 2008), but more recent works have leveraged modern and relatively inexpensive FPGA technologies to perform either portions of the PCA computations (Ali et al., 2013) or complete implementations of the PCA algorithm (Bravo et al., 2010; Korat & Alimohammad, 2019). Additional research has proven that embedded implementation of PCA is generic and highly scalable, enabling substantial improvements in the computational speed of the technique across a range of applications (Shahrouzi & Perera, 2019).

Support Vector Machines (SVMs), which fall into the broad category of machine learning algorithms, have historically provided excellent performance when classifying complex and continuous features (Saidi et al., 2021). Generally, machine learning techniques are considered computationally expensive and challenging to implement on embedded hardware (Sze et al., 2017). However, several recent studies have shown the potential for the SVM algorithm to be implemented on a variety of embedded devices including VLSI integrated circuits and FPGAs (Amezzane et al., 2020; Loukrakpam & Choudhury, 2020).

In addition to being suitable for implementation on embedded hardware, both Principal Components Analysis and/or Support Vector Machines have been previously utilized for spacecraft fault detection and diagnosis (Yu Gao et al., 2012), the onboard detection of anomalous behavior in CubeSat solar panels (Cespedes et al., 2022), and other intelligent decision making applications onboard spacecraft (Jallad & Mohammed, 2014). This illustrates that not only can PCA and SVM be implemented in hardware, but that they have history in successful implementation for spaceflight applications.

5.3 *Semantic Labeling of Prioritized Data*

This manuscript has proposed a technique for the automated binary classification of time-series data as either anomalous (i.e., potentially of high scientific importance) or nominal, one useful avenue of future work would be to provide additional semantic labels for the high-priority data. These labels could indicate whether an identified event falls into a particular class of geophysical event, such as shocks, magnetopause crossings, or whistler-mode waves. Several machine learning techniques have been previously developed for the identification of events in spaceflight data archives (Fordin et al., 2023; Vech & Malaspina, 2021), although the computational intensity or large dimensionality of some of these algorithms make them potentially unsuitable for deployment on spaceflight hardware. The proposed anomaly detection technique provides utility to the pursuit of semantic labeling in two ways: firstly, as a low-cost data reduction tool it can reduce the number of samples that must be processed by more complex algorithms, decreasing the overall time complexity of the problem; secondly, the binary-labeled data can serve as a powerful input feature vector if data reduction is not desired or required, potentially increasing model performance.

5.4 Generalizability

This manuscript has illustrated the applicability of the proposed method of anomaly detection to magnetic field data from MMS and CASSIOPE/Swarm-Echo, as well as the ion density and velocity moments from MMS. Although this shows promise for the generalizability of the technique to other platforms and other physical observations, thorough exploration of this generalizability to other spacecraft observations (such as ion and electron pressures and temperatures, as well as the housekeeping measurements critical to spacecraft operations) remains an avenue for future work.

Testing the generalizability of the clustering model trained on a specific interval of time against different intervals is also an interesting future project. If the clustering learned by an OC-SVM trained on one interval of time (e.g., one of the 24-hour periods of MMS data) is applied to subsequent intervals (e.g., the next several days of MMS data) with meaningful results, the computational complexity of the proposed technique would decrease by eliminating the need to retrain a model for every interval under observation, enabling more rapid in-situ application.

6 Conclusions

The scientific measurements captured by in-situ spacecraft are critical to our study of the physical phenomena that control the flow of mass, momentum, and energy throughout our solar system. However, due to burdens on ground systems as a result of a rapid increase in the number of active spaceflight missions and the ever-growing need for higher-cadence data, spacecraft are often unable to transmit all of their data to Earth at full rate. As a result, missions must develop techniques that enable prioritization of specific intervals with a high probability of importance to their science goals. The techniques that have historically been used by large missions have provided excellent results but come with high design and implementation costs, leaving them potentially unsuitable for application on low-cost missions. This manuscript has proposed a generic technique for the prioritization of data through the detection of anomalous data points in spaceflight observations. The technique’s utility has been demonstrated through the successful identification of various physical phenomena in a variety of data products from several missions, and its potential applicability to low-cost spaceflight hardware has been discussed. Additionally, several avenues for potential future research utilizing the proposed technique have been highlighted.

Acknowledgments

This work was supported in part by the National Aeronautics and Space Administration HERMES mission and grants 80NSSC21M0180, 80NSSC23K1295. Additional support was provided by the Center for HelioAnalytics, funded through the NASA ISFM program. The computing resources utilized in the generation of the figures and analysis in this manuscript were in part provided by HelioCloud, a service managed by NASA's Heliophysics Digital Resource Library. The authors thank Barbara J. Thompson for her thoughtful insights towards the potential applications of the proposed method.

Open Research

The CASSIOPE/Swarm-Echo MGF data used in this manuscript is publicly available at https://epop-data.phys.ucalgary.ca/. The MMS FGM, FPI, and SITL data used in this manuscript is publicly available through https://lasp.colorado.edu/mms/sdc/public/. Code used to implement the algorithm described in this manuscript, along with sample data used to illustrate the technique's performance, are currently stored at https://drive.google.com/drive/folders/1151OzBKW9Mgf6xVQ8P_bGLMmxTfBjHyG?usp=sharing. Upon acceptance of this manuscript, the code and sample data will be moved to a University of Maryland institutional repository, or similar digital repository, for long-term storage and reuse.

References

Ali, A. A. S., Amira, A., Bensaali, F., & Benammar, M. (2013). Hardware PCA for gas identification systems using high level synthesis on the Zynq SoC. *2013 IEEE 20th International Conference on Electronics, Circuits, and Systems (ICECS)*, 707–710. https://doi.org/10.1109/ICECS.2013.6815512

Amer, M., Goldstein, M., & Abdennadher, S. (2013). Enhancing one-class support vector machines for unsupervised anomaly detection. *Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description*, 8–15. https://doi.org/10.1145/2500853.2500857

Amezzane, I., Fakhri, Y., El Aroussi, M., & Bakhouya, M. (2020). Hardware Acceleration of SVM Training for Real-Time Embedded Systems: Overview. In S. Dos Santos, M. Maslouhi, & K. A. Okoudjou (Eds.), *Recent Advances in Mathematics and Technology: Proceedings of the First International Conference on Technology, Engineering, and Mathematics, Kenitra, Morocco, March 26-27, 2018* (pp. 131–139). Springer International Publishing. https://doi.org/10.1007/978-3-030-35202-8_7

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: An analytical review. *WIREs Data Mining and Knowledge Discovery*, *11*(5), e1424. https://doi.org/10.1002/widm.1424

Argall, M. R., Small, C. R., & Petrik, M. (2022). *argallmr/pymms: V0.4.6 (2022-05-19)* (0.4.6) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.6564714

Argall, M. R., Small, C. R., Piatt, S., Breen, L., Petrik, M., Kokkonen, K., Barnum, J., Larsen, K., Wilder, F. D., Oka, M., Paterson, W. R., Torbert, R. B., Ergun, R. E., Phan, T., Giles, B. L., & Burch, J. L. (2020). MMS SITL Ground Loop: Automating the Burst Data Selection Process. *Frontiers in Astronomy and Space Sciences*, *7*. https://www.frontiersin.org/articles/10.3389/fspas.2020.00054

Baker, D. N., Riesberg, L., Pankratz, C. K., Panneton, R. S., Giles, B. L., Wilder, F. D., & Ergun, R. E. (2016). Magnetospheric Multiscale Instrument Suite Operations and Data System. *Space Science Reviews*, *199*(1), 545–575. https://doi.org/10.1007/s11214-014-0128-5

Bale, S. D., Goetz, K., Harvey, P. R., Turin, P., Bonnell, J. W., Dudok de Wit, T., Ergun, R. E., MacDowall, R. J., Pulupa, M., Andre, M., Bolton, M., Bougeret, J.-L., Bowen, T. A., Burgess, D., Cattell, C. A., Chandran, B. D. G., Chaston, C. C., Chen, C. H. K., Choi, M. K., … Wygant, J. R. (2016). The FIELDS Instrument Suite for Solar Probe Plus. *Space Science Reviews*, *204*(1), Article 1. https://doi.org/10.1007/s11214-016-0244-5

Bard, C., & Dorelli, J. C. (2021). Neural Network Reconstruction of Plasma Space-Time. *Frontiers in Astronomy and Space Sciences*, *8*. https://www.frontiersin.org/articles/10.3389/fspas.2021.732275

Barrie, A. C., da Silva, D., Elkington, S., Sternovsky, Z., Rager, A. C., Gershman, D. J., Paterson, W. R., Dorelli, J. C., & Giles, B. (2018). Physically Accurate Large Dynamic Range Pseudo Moments for the MMS Fast Plasma Investigation. *Earth and Space Science*, *5*(9), 503–515. https://doi.org/10.1029/2018EA000407

Berner, J. B. (2019). Deep Space Network in the CubeSat Era. *IEEE Aerospace and Electronic Systems Magazine*, *34*(4), 46–54. https://doi.org/10.1109/MAES.2019.2913266

Bravo, I., Mazo, M., Lázaro, J. L., Gardel, A., Jiménez, P., & Pizarro, D. (2010). An Intelligent Architecture Based on Field Programmable Gate Arrays Designed to Detect Moving Objects by Using Principal Component Analysis. *Sensors*, *10*(10), Article 10. https://doi.org/10.3390/s101009232

Burch, J. L., Moore, T. E., Torbert, R. B., & Giles, B. L. (2016). Magnetospheric Multiscale Overview and Science Objectives. *Space Science Reviews*, *199*(1), Article 1. https://doi.org/10.1007/s11214-015-0164-9

Camporeale, E. (2019). The Challenge of Machine Learning in Space Weather: Nowcasting and Forecasting. *Space Weather*, *17*(8), Article 8. https://doi.org/10.1029/2018SW002061

Carlson, C. W., Curtis, D. W., Paschmann, G., & Michel, W. (1982). An instrument for rapidly measuring plasma distribution functions with high resolution. *Advances in Space Research*, *2*(7), 67–70. https://doi.org/10.1016/0273-1177(82)90151-X

Cespedes, A. J. J., Pangestu, B. H. B., Hanazawa, A., & Cho, M. (2022). Performance Evaluation of Machine Learning Methods for Anomaly Detection in CubeSat Solar Panels. *Applied Sciences*, *12*(17), Article 17. https://doi.org/10.3390/app12178634

Chang, C.-C., & Lin, C.-J. (2001). Training v-Support Vector Classifiers: Theory and Algorithms. *Neural Computation*, *13*(9), 2119–2147. https://doi.org/10.1162/089976601750399335

Clausen, L. B. N., & Nickisch, H. (2018). Automatic Classification of Auroral Images From the Oslo Auroral THEMIS (OATH) Data Set Using Machine Learning. *Journal of Geophysical Research: Space Physics*, *123*(7), Article 7. https://doi.org/10.1029/2018JA025274

Connerney, J. E. P., Benn, M., Bjarno, J. B., Denver, T., Espley, J., Jorgensen, J. L., Jorgensen, P. S., Lawton, P., Malinnikova, A., Merayo, J. M., Murphy, S., Odom, J., Oliversen, R., Schnurr, R., Sheppard, D., & Smith, E. J. (2017). The Juno Magnetic Field Investigation. *Space Science Reviews*, *213*(1), Article 1. https://doi.org/10.1007/s11214-017-0334-z

Connerney, J. E. P., Espley, J., Lawton, P., Murphy, S., Odom, J., Oliversen, R., & Sheppard, D. (2015). The MAVEN Magnetic Field Investigation. *Space Science Reviews*, *195*(1), Article 1. https://doi.org/10.1007/s11214-015-0169-4

Engebretson, M. J., Posch, J. L., Capman, N. S. S., Campuzano, N. G., Bělik, P., Allen, R. C., Vines, S. K., Anderson, B. J., Tian, S., Cattell, C. A., Wygant, J. R., Fuselier, S. A., Argall, M. R., Lessard, M. R., Torbert, R. B., Moldwin, M. B., Hartinger, M. D., Kim, H., Russell, C. T., … Singer, H. J. (2018). MMS, Van Allen Probes, GOES 13, and Ground-Based Magnetometer Observations of EMIC Wave Events Before, During, and After a Modest Interplanetary Shock. *Journal of Geophysical Research: Space Physics*, *123*(10), 8331–8357. https://doi.org/10.1029/2018JA025984

Finley, M. G., Broadfoot, R. M., Shekhar, S., & Miles, D. M. (2023). Identification and Removal of Reaction Wheel Interference From In-Situ Magnetic Field Data Using Multichannel Singular Spectrum Analysis. *Journal of Geophysical Research: Space Physics*, *128*(2), Article 2. https://doi.org/10.1029/2022JA031020

Fordin, S., Shay, M., III, L. B. W., Maruca, B., & Thompson, B. J. (2023). A Machine Learning–Based Approach to Time-series Wave Identification in the Solar Wind. *The Astrophysical Journal*, *949*(2), 40. https://doi.org/10.3847/1538-4357/acc8d5

Gordon, D., & Brown, R. (1972). Recent advances in fluxgate magnetometry. *IEEE Transactions on Magnetics*, *8*(1), Article 1. https://doi.org/10.1109/TMAG.1972.1067268

Hasegawa, H., Argall, M. R., Aunai, N., Bandyopadhyay, R., Bessho, N., Cohen, I. J., Denton, R. E., Dorelli, J. C., Egedal, J., Fuselier, S. A., Garnier, P., Genot, V., Graham, D. B., Hwang, K. J., Khotyaintsev, Y. V., Korovinskiy, D. B., Lavraud, B., Lenouvel, Q., Li, T. C., … Zenitani, S. (2023). *Advanced methods for analyzing in-situ observations of magnetic reconnection* (arXiv:2307.05867). arXiv. https://doi.org/10.48550/arXiv.2307.05867

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and Their Applications*, *13*(4), 18–28. https://doi.org/10.1109/5254.708428

Jallad, A.-H. M., & Mohammed, L. B. (2014). Hardware Support Vector Machine (SVM) for satellite on-board applications. *2014 NASA/ESA Conference on Adaptive Hardware and Systems (AHS)*, 256–261. https://doi.org/10.1109/AHS.2014.6880185

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Kang, S., Kim, D., & Cho, S. (2019). Approximate training of one-class support vector machines using expected margin. *Computers & Industrial Engineering*, *130*, 772–778. https://doi.org/10.1016/j.cie.2019.03.029

Korat, U. A., & Alimohammad, A. (2019). A Reconfigurable Hardware Architecture for Principal Component Analysis. *Circuits, Systems, and Signal Processing*, *38*(5), 2097–2113. https://doi.org/10.1007/s00034-018-0953-y

Lepping, R. P., Acũna, M. H., Burlaga, L. F., Farrell, W. M., Slavin, J. A., Schatten, K. H., Mariani, F., Ness, N. F., Neubauer, F. M., Whang, Y. C., Byrnes, J. B., Kennon, R. S., Panetta, P. V., Scheifele, J., & Worley, E. M. (1995). The WIND magnetic field investigation. *Space Science Reviews*, *71*(1), 207–229. https://doi.org/10.1007/BF00751330

Liddle, J. D., Holt, A. P., Jason, S. J., O'Donnell, K. A., & Stevens, E. J. (2020). Space science with CubeSats and nanosatellites. *Nature Astronomy*, *4*(11), Article 11. https://doi.org/10.1038/s41550-020-01247-2

Loukrakpam, M., & Choudhury, M. (2020). Hardware-Efficient VLSI Design for Cascade Support Vector Machine with On-Chip Training and Classification Capability. *Circuits, Systems, and Signal Processing*, *39*(10), 5272–5297. https://doi.org/10.1007/s00034-020-01415-9

Ma, S., Chou, Y. C., Zhao, H., Chen, L., Ma, X., & Liu, J. (2023). Network Characteristics of LEO Satellite Constellations: A Starlink-Based Measurement from End Users. *IEEE INFOCOM 2023 - IEEE Conference on Computer Communications*, 1–10. https://doi.org/10.1109/INFOCOM53939.2023.10228912

McDowell, J. C. (2020). The Low Earth Orbit Satellite Population and Impacts of the SpaceX Starlink Constellation. *The Astrophysical Journal Letters*, *892*(2), L36. https://doi.org/10.3847/2041-8213/ab8016

Miles, D. M., Mann, I. R., Pakhotin, I. P., Burchill, J. K., Howarth, A. D., Knudsen, D. J., Lysak, R. L., Wallis, D. D., Cogger, L. L., & Yau, A. W. (2018). Alfvénic Dynamics and Fine Structuring of Discrete Auroral Arcs: Swarm and e-POP Observations. *Geophysical Research Letters*, *45*(2), Article 2. https://doi.org/10.1002/2017GL076051

Noble, W. S. (2006). What is a support vector machine? *Nature Biotechnology*, *24*(12), Article 12. https://doi.org/10.1038/nbt1206-1565

Phan, T. D., Shay, M. A., Eastwood, J. P., Angelopoulos, V., Oieroset, M., Oka, M., & Fujimoto, M. (2016). Establishing the Context for Reconnection Diffusion Region Encounters and Strategies for the Capture and Transmission of Diffusion Region Burst Data by MMS. *Space Science Reviews*, *199*(1), 631–650. https://doi.org/10.1007/s11214-015-0150-2

Pollock, C., Moore, T., Jacques, A., Burch, J., Gliese, U., Saito, Y., Omoto, T., Avanov, L., Barrie, A., Coffey, V., Dorelli, J., Gershman, D., Giles, B., Rosnack, T., Salo, C., Yokota, S., Adrian, M., Aoustin, C., Auletti, C., … Zeuch, M. (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., Le, G., Leinweber, H. K., Leneman, D., Magnes, W., Means, J. D., Moldwin, M. B., Nakamura, R., Pierce, D., Plaschke, F., Rowe, K. M., Slavin, J. A., Strangeway, R. J., Torbert, R., … Richter, I. (2016). The Magnetospheric

Multiscale Magnetometers. *Space Science Reviews*, *199*(1), Article 1. https://doi.org/10.1007/s11214-014-0057-3

Saidi, A., Ben Othman, S., Dhouibi, M., & Ben Saoud, S. (2021). FPGA-based implementation of classification techniques: A survey. *Integration*, *81*, 280–299. https://doi.org/10.1016/j.vlsi.2021.08.004

Shahrouzi, S. N., & Perera, D. G. (2019). Optimized hardware accelerators for data mining applications on embedded platforms: Case study principal component analysis. *Microprocessors and Microsystems*, *65*, 79–96. https://doi.org/10.1016/j.micpro.2019.01.001

Shin, H. J., Eom, D.-H., & Kim, S.-S. (2005). One-class support vector machines—An application in machine fault detection and classification. *Computers & Industrial Engineering*, *48*(2), 395–408. https://doi.org/10.1016/j.cie.2005.01.009

Sze, V., Chen, Y.-H., Emer, J., Suleiman, A., & Zhang, Z. (2017). Hardware for machine learning: Challenges and opportunities. *2017 IEEE Custom Integrated Circuits Conference (CICC)*, 1–8. https://doi.org/10.1109/CICC.2017.7993626

Tung-Chien Chen, Wentai Liu, & Liang-Gee Chen. (2008). VLSI architecture of leading eigenvector generation for on-chip principal component analysis spike sorting system. *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 3192–3195. https://doi.org/10.1109/IEMBS.2008.4649882

Vech, D., & Malaspina, D. M. (2021). A Novel Machine Learning Technique to Identify and Categorize Plasma Waves in Spacecraft Measurements. *Journal of Geophysical Research: Space Physics*, *126*(9), e2021JA029567. https://doi.org/10.1029/2021JA029567

Wallis, D. D., Miles, D. M., Narod, B. B., Bennest, J. R., Murphy, K. R., Mann, I. R., & Yau, A. W. (2015). The CASSIOPE/e-POP Magnetic Field Instrument (MGF). *Space Science Reviews*, *189*(1), Article 1. https://doi.org/10.1007/s11214-014-0105-z

Weiss, B. P., Merayo, J. M. G., Ream, J. B., Oran, R., Brauer, P., Cochrane, C. J., Cloutier, K., Elkins-Tanton, L. T., Jørgensen, J. L., Maurel, C., Park, R. S., Polanskey, C. A., de Soria Santacruz-Pich, M., Raymond, C. A., Russell, C. T., Wenkert, D., Wieczorek, M. A., & Zuber, M. T. (2023). The Psyche Magnetometry Investigation. *Space Science Reviews*, *219*(3), 22. https://doi.org/10.1007/s11214-023-00965-z

Yau, A. W., & James, H. G. (2015). CASSIOPE Enhanced Polar Outflow Probe (e-POP) Mission Overview. *Space Science Reviews*, *189*(1), Article 1. https://doi.org/10.1007/s11214-015-0135-1

Yin, S., Zhu, X., & Jing, C. (2014). Fault detection based on a robust one class support vector machine. *Neurocomputing*, *145*, 263–268. https://doi.org/10.1016/j.neucom.2014.05.035

Yu Gao, Tianshe Yang, Nan Xing, & Minqiang Xu. (2012). Fault detection and diagnosis for spacecraft using principal component analysis and support vector machines. *2012 7th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, 1984–1988. https://doi.org/10.1109/ICIEA.2012.6361054

Zhan, Y., Wan, P., Jiang, C., Pan, X., Chen, X., & Guo, S. (2020). Challenges and Solutions for the Satellite Tracking, Telemetry, and Command System. *IEEE Wireless Communications*, *27*(6), 12–18. https://doi.org/10.1109/MWC.001.2000089

Zhang, Q., Tian, Y., Wang, T., Yuan, F., & Xu, Q. (2015). ApproxEigen: An approximate computing technique for large-scale eigen-decomposition. *2015 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 824–830. https://doi.org/10.1109/ICCAD.2015.7372656

Zhou, J., Chan, K. L., Chong, V. F. H., & Krishnan, S. M. (2005). Extraction of Brain Tumor from MR Images Using One-Class Support Vector Machine. *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, 6411–6414. https://doi.org/10.1109/IEMBS.2005.1615965