Journal of Computational Science 97 (2026) 102855

Contents lists available at ScienceDirect

Journal of Computational Science

journal homepage: www.elsevier.com/locate/jocs

Time-invariant properties and principal components of in-situ measurements used for outlier detection in space missions☆

Jonah Ekelund [a,*], Savvas Raptis [b], Vicki Toy-Edens [b], Wenli Mo [b], Drew L. Turner [b], Ian J. Cohen [b], Stefano Markidis [a]

[a] Computer Science Dept. KTH Royal Institute of Technology, Stockholm, Sweden
[b] Johns Hopkins University Applied Physics Laboratory, Laurel, MD, USA

ARTICLE INFO

Keywords:
Outlier detection
Incremental PCA
In-site time series measurements
Online learning

ABSTRACT

Understanding Earth's magnetosphere requires analyzing complex, multiscale interactions captured through diverse in-situ measurements such as particle distributions (e.g., ion spectra and flow velocities), and electromagnetic fields. Modern space missions, including NASA's MMS and THEMIS, produce large volumes of such time-series data, where scientifically relevant events often appear as short-lived, context-dependent outliers. In this work, we investigate the structure of time-windowed, multi-feature datasets from Earth's magnetosphere and show that their information content can be effectively represented using a reduced set of largely time-invariant principal components. Using Principal Component Analysis (PCA) for static datasets and Incremental PCA for streaming observations, we characterize these components and leverage the associated reconstruction error to develop an unsupervised outlier detection method suitable for evolving data distributions. Building on our earlier approach, this method robustly identifies significant plasma phenomena in both structured and streaming-mode measurements. Applied to MMS and THEMIS observations, it successfully recovers known events, including bow shock and magnetopause crossings, foreshock transients, and plasma bubbles, while also highlighting additional scientifically relevant structures. This demonstrates that dimensionality-reduction-based outlier detection provides an effective pathway for automated event identification in complex space plasma datasets. This is an extension of previous work presented in Ekelund et al. (2025).

1. Introduction

Earth's magnetosphere is a complex and dynamic environment, shaped by multiscale interactions and strong coupling, both internally and with solar activity via the interplanetary magnetic field and the solar wind [1]. This complexity necessitates the collection of diverse, multi-featured data, including particle distributions (e.g., ion spectra and flow velocities), and electromagnetic fields [2]. Space physics missions such as NASA's Magnetospheric Multiscale (MMS) Mission [1], the Time History of Events and Macroscale Interactions during Substorms (THEMIS) [3], and ESA's Cluster Mission [4] generate large volumes of such time-series data through in-situ observations. The multiscale nature of magnetospheric processes, ranging from large-scale solar wind dynamics to small-scale events such as magnetic reconnection, further complicates analysis. As spacecraft traverse this ever-changing environment, even major interactions may comprise only a small subset of the recorded measurements, making the identification of scientifically relevant events particularly challenging.

A central difficulty lies in recognizing scientifically significant events [5]. These include well-established phenomena such as Magnetopause and Bow Shock crossings [6], magnetic reconnection [1], as well as transient or not yet fully characterized structures [7]. A common feature of such events is their manifestation as short-term changes in the plasma environment, observed as spacecraft move between regions. Consequently, these signatures often appear as statistical outliers within the data [8], motivating the development of detection methods that incorporate temporal context. Simple approaches, such as computing a mean over preceding samples [9], provide a baseline, while more advanced techniques treat each window of data as a multi-feature sample [8].

Historically, the identification and prioritization of such events have relied on manual searches [5], rule-based algorithms [2,10], and supervised machine learning methods relying on human-labeled data [6,11]. Recently, research has shifted toward unsupervised machine learning [12], automating discovery and reducing reliance on labeled datasets.

☆ This article is part of a Special issue entitled: 'ICCS 2025' published in Journal of Computational Science.

* Corresponding author.
E-mail address: jonahek@kth.se (J. Ekelund).

https://doi.org/10.1016/j.jocs.2026.102855
Received 14 November 2025; Received in revised form 6 March 2026; Accepted 3 April 2026
Available online 4 April 2026

In this study, we investigate the structure of time-windowed in-situ space plasma measurements collected from Earth's dayside. Our analysis includes both structured datasets with minimal event activity and streaming-mode data, where ion spectra, magnetic field measurements, and ion velocity observations from the MMS mission are divided into non-overlapping windows. To characterize the underlying structure, we employ principal component analysis (PCA) for static datasets and Incremental PCA for streaming contexts, with emphasis on interpreting the first principal components.

Building on these analyses, we design an unsupervised approach to detect scientifically significant events in multi-feature in-situ datasets. Our method extends the outlier detection algorithm introduced in our previous work [13] and is well-suited for streaming applications. By leveraging the PCA reconstruction error arising from dimensionality reduction and adapting to evolving data distributions via Incremental PCA, the method robustly identifies relevant plasma phenomena. We demonstrate its effectiveness in detecting events such as foreshock transients and Bow Shock crossings in MMS observations, as well as foreshock bubbles in THEMIS data.

2. Background

This work focuses on multi-dimensional space-plasma time-series data, utilizing data from the MMS [1] and THEMIS satellites, launched in 2015 and 2007, respectively, as case studies. NASA launched MMS to explore magnetic reconnection in Earth's magnetosphere with exceptional temporal resolution. It comprises four spacecraft that navigate in formation through both the dayside magnetopause and nightside magnetotail regions [1]. NASA launched the THEMIS mission to explore the onset and progression of substorms. This mission includes a fleet of five satellites positioned to track particles along the magnetotail. Although the primary objective was to conduct measurements in the magnetotail, specifically in the nightside region of Earth's magnetosphere, the spacecraft also gathered data from the dayside region [3]. In this study, we use a segment of THEMIS dayside data.

As spacecraft orbit Earth, they pass through various regions, such as the magnetosphere, where Earth's magnetic field shapes the plasma environment, and regions beyond its protective zone, where the solar wind and the Sun's magnetic fields become predominant. These spacecraft also encounter critical boundaries, including the bow shock, where the solar wind decelerates and redirects, and the magnetopause, which delineates Earth's magnetic influence from the solar wind.

Downstream of Earth's bow shock, the heated and compressed solar wind forms the magnetosheath region, where plasma is slowed down until it reaches the magnetopause. This region is the primary interface through which energy from the solar wind is transferred into Earth's magnetosphere. When the upstream magnetic field is almost parallel to the bow shock, it forms a more turbulent quasi-parallel magnetosheath region, while when the magnetic field is almost perpendicular, the region is called quasi-perpendicular

The Earth's nightside plasma sheet is a region of hot plasma located within the magnetotail, between the low-density regions called lobes. It acts as the primary region where particle transport takes place and where magnetic reconnection and the formation of fast earthward flows can be found.

Classifying the regions where measurements are taken enables scientists to focus on the most significant geospace phenomena [12]. To support such analyses, Grison et al. [5] introduced the GRMB dataset, a manually labeled dataset from the Cluster satellites. In this dataset, each 20-minute interval is assigned to one of 15 geospace regions or transition categories. This structured labeling facilitates targeted investigations of region-specific phenomena in geospace.

Numerous studies have developed techniques for automatically identifying plasma regions, such as the solar wind and magnetosheath, and their transitions, like the bow shock. Olshevsky et al. [11] and Breuillard et al. [6] employed neural networks to classify various plasma regions. Bow shock transitions can then be determined either by regions with a high probability of neighboring area occurrence [11] or by specific classification labels [6]. These methods demonstrate high accuracy but depend on training datasets, limiting adaptability to new data.

Recent research by Toy-Edens et al. [12] employed a Gaussian Mixture Model to effectively classify the MMS dayside region, offering high precision with less complexity than a complete neural network. Bakrania et al. [14] implemented a classification pipeline involving an autoencoder, PCA, and Agglomerative Clustering on data from ESA's Cluster spacecraft [15]. Innocenti et al. [16] applied PCA, self-organizing maps, and K-means clustering to analyze various parameters, like B-field, particle velocity, and density, for plasma region classification.

2.1. Outlier detection

Outlier detection focuses on identifying individual or clustered samples that deviate from the predominant pattern observed in the rest of the dataset. Conventionally, most outlier detection methods are applied to static datasets, in which the entire data spectrum is preestablished [17].

In scenarios involving high-dimensional datasets, dimensionality reduction strategies such as Principal Component Analysis (PCA) or Autoencoders [14] are often employed to reduce the dimensionality of the search space [18]. PCA constitutes a linear dimensionality reduction technique that extracts the leading N features, known as Principal Components (PCs), which encapsulate the greatest variance, hence, the most information, from the initial feature space [19]. After reduction to this feature space, various clustering methods [14,16] can be employed to categorize the samples into distinct groups. Outliers can subsequently be identified by assessing their distances to the cluster centroids or by examining density metrics.

The methodology outlined by Finley et al. [8] involves a dual-phase process to identify noteworthy occurrences in MMS data. This is achieved by partitioning the data region into smaller segments and applying Principal Component Analysis (PCA) to the resulting subregion matrix. Subsequently, a One-Class Support Vector Machine (OCSVN) is employed to identify anomalous windows. Despite its effectiveness, this method is computationally intensive, as it requires recalculating PCA and OCSVN for each region assessed. Similarly, Zamry et al. [20] employs PCA for feature reduction alongside a One-Class Support Vector Machine to detect irregularities in sensor data from a Wireless Sensor Network.

Alternatively, to identify outliers, one might revert the samples to their original feature space and measure the deviation of the reconstructed features (R_f) from the initial features (F_f). This deviation is the reconstruction error (E_f):

$$E_f = R_f - F_f \tag{1}$$

Samples exhibiting substantial reconstruction errors in their features tend to diverge from the PCA model and can consequently be classified as outliers relative to the remaining samples [21].

Data streams complicate outlier identification due to their inherent temporal variability. Changes in temporal data can alter the classification of outliers. Additionally, both the baseline feature values and their significance can fluctuate over time [18]. As a result, this can invalidate an initial data model, necessitating that outlier detection methods adapt to these data shifts. Incremental PCA represents a PCA variation that facilitates the progressive construction of the PCA model [19]. This approach becomes essential when the dataset is too large to be entirely loaded into memory or, as discussed in the paper, when the complete dataset is not accessible during the initial model's creation. Bhushan et al. [21] utilized an Incremental PCA on streaming data with spatial distribution, focusing only on outliers in a single feature type from one sensor, with outliers artificially injected into the dataset.

Table 1
MMS intervals containing foreshock transient events from the Earth dayside [7] and fast plasma flows in the Earth nightside data [22].

Nr	Data interval	Transient/Flow location	Day or nightside
1	2017-12-17 16:00 → 2017-12-17 22:00	17:52 → 17:54	Dayside
2	2018-01-12 00:50 → 2018-01-12 06:00	01:50 → 01:52	Dayside
3	2018-12-14 02:00 → 2018-12-14 07:20	04:21 → 04:22	Dayside
		04:40 → 04:42	
4	2018-12-10 04:00 → 2018-12-10 11:00	05:12 → 05:25	Dayside
		06:27 → 06:31	
5	2019-01-05 16:00 → 2019-01-05 19:00	17:38 → 17:41	Dayside
6	2021-01-12 00:00 → 2021-01-12 06:00	01:18 → 01:21	Dayside
7	2021-02-13 10:00 → 2021-02-13 18:00	11:05 → 11:06	Dayside
8	2022-05-02 18:00 → 2022-05-02 22:00	18:23 → 18:25	Dayside
9	2022-11-24 02:00 → 2022-11-24 09:00	04:16 → 04:18	Dayside
10	2023-01-16 06:00 → 2023-01-16 11:00	08:21 → 08:24	Dayside
11	2017-07-23 12:00 → 2017-07-23 18:00	16:55 → 16:56	Nightside
12	2021-08-14 16:00 → 2021-08-15 06:00	01:23 → 01:25	Nightside

3. Methodology

3.1. Data and pre-processing

MMS Data: The principal data utilized in this study is sourced from the MMS mission. Specifically, it comprises the omni-directional ion spectra and ion velocities in Geocentric Solar Ecliptic (GSE) coordinates obtained from the Fast Plasma Investigation (FPI) instrument [23], as well as the magnetic field data (hereafter referred to as the B-field) acquired from fluxgate magnetometers (FGM), which are components of the FIELDS instrument suite [24], onboard the MMS-1 spacecraft. Table 1 enumerates ten dayside intervals of MMS data, along with specified regions of interest. These intervals represent instances when MMS transitions into or out of Earth's magnetosphere from the solar wind. Transitions from the solar wind to the magnetosheath are identified as *bow shock* crossings, while those from the magnetosheath to the magnetosphere are termed *magnetopause* crossings. The FGM data is acquired at a notably higher frequency, approximately 8–16 Hz, compared to the FPI data, which is around 0.2 Hz; hence, the FGM data was downsampled to match the FPI sampling frequency. The dataset employed is level-2 data, post-processed on Earth, and is not accessible onboard the spacecraft.

In this study, we used two MMS datasets: the first is based on the labeled dataset compiled by Olshevsky et al. [11], which we used to evaluate the principal components. The second dataset comprises 10 dayside and 2 nightside intervals known to contain scientifically relevant events (Table 1) and is used to evaluate the algorithm's ability to detect these and other events.

Olshevsky et al. [11] labeled the MMS data from November and December 2017 into the regions: Solar Wind, Ion Foreshock, Magnetosheath, Magnetosphere, or Unknown. Using the SpacePhyML dataset creation tool [25] and the labels from Olshevsky et al. [11], we construct a structured dataset for November 2017 comprising Ion Spectrum, B-field, and Ion Velocity data. The dataset is filtered to include only samples with ion-spectrum channels ranging from 1 to 10^4 eV and is split into homogeneous time windows of size N for each label, excluding samples with the Unknown label. Fig. 1 depicts two examples of each region with a time window of $N = 133$ samples, approximately 10 min. Here we see that the Solar Wind and Ion Foreshock regions consist of a narrow central spectral line, with most of the flux concentrated at the 10^3 energy level. The ion foreshock has an additional component with slightly lower flux above this line. In the magnetosheath region, the peak flux is spread across more energy levels, with a relatively high flux down to approximately 10^2 eV and, in some cases, a lower flux throughout the measured spectrum. Lastly, the magnetosphere has a lower peak flux than the other regions; however, at higher energies, around 10^4 eV.

We also use the same data interval, November 2017, to evaluate the principal components computed with Incremental PCA. In this case, the data are unstructured, meaning it is partitioned into non-overlapping homogeneous data windows without considering specific regions. The only filtering performed is to include samples with ion-spectrum channels ranging from 1 to 10^4 eV and to exclude any windows with a time jump between samples exceeding the sampling time (~4.5s).

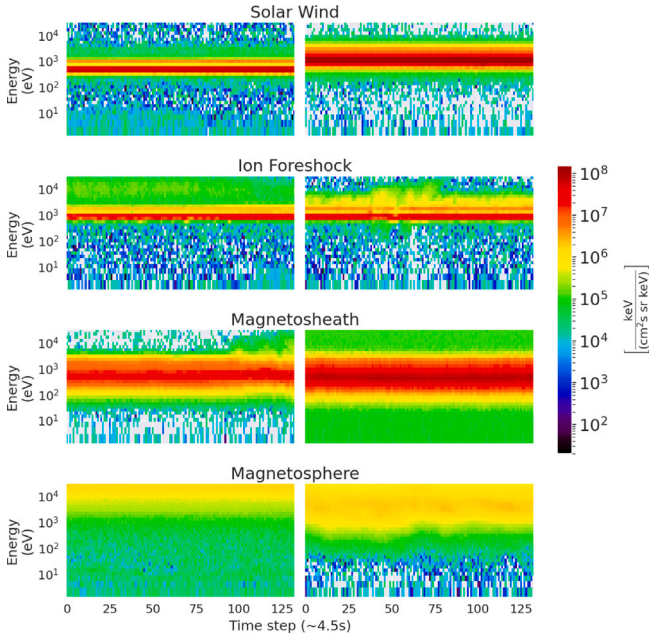


Fig. 1. Examples of 10-minute (133 time steps) spectra for the four different regions in the Earth magnetospheric dayside labeled by Olshevsky et al. [11], each time step is approximately 4.5 s.

The second dataset presented in Table 1 comprises events from Raptis et al. [7] for the dayside and from Richard et al. [22] for the nightside. The events presented by Raptis et al. [7] are transient phenomena occurring upstream of the bow shock in the ion foreshock region. These phenomena are recognized for their role in particle energization and the induction of space weather effects [26]. Consequently, detecting these events, even within level-2 data, is crucial for advancing scientific studies. Additionally, these data intervals also include other phenomena, such as bow shock and magnetopause crossings, which are essential for automatic operational identification.

The two nightside data intervals from MMS-1, as outlined in Table 1, were utilized to assess the method's ability to generalize across different regions. These intervals originate from Earth's nightside magnetosphere and include two fast plasma flows examined by Richard et al. [22]. Such fast plasma flows are linked to geomagnetic disturbances and are therefore capable of inducing space weather effects [27,28].

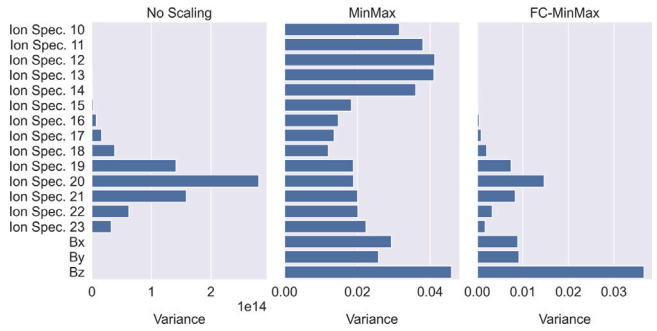


Fig. 2. Comparison of the variance of the different features for no scaling (Left), MinMax scaling (Middle), and FC-MinMax scaling (Right). The features are MMS omnidirectional ion spectrum, channel 10 to 23, and B-field from 2017-12-17, 20:00 to 21:50, while MMS is traversing the magnetosheath region.

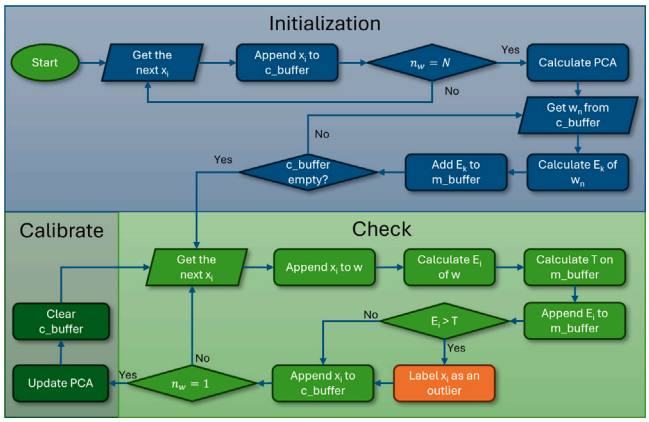


Fig. 3. Flowchart showing the conceptual flow of the outlier detection algorithm: samples are received, checked for outliers, and added to the calibration buffer (c_buffer). If the c_buffer contains one full data window, it is used to update the PCA model..

The **THEMIS spacecraft** serves as an alternative source for space plasma data, utilizing different instrumentation compared to the MMS. Specifically, we employ the THEMIS-C probe's data on ion energy flux and ion velocity obtained from the Electrostatic Analyzer (ESA). These measurements undergo minimal processing and are readily available for onboard use [3,29]. The analysis focuses on a time interval beginning on 2008-07-15, which includes a foreshock transient [30].

3.2. Multiple features

The PCA technique identifies the direction with the greatest variance; thus, for PCA to be effective, all features must be normalized to a common scale. When PCA is applied to datasets with a single feature type, such as the omnidirectional ion spectrum discussed in this study, it is expected that all feature values will exhibit similar magnitudes. However, when datasets include multiple feature types, these additional features may exhibit significantly varying magnitudes. In such cases, the feature types with higher magnitudes tend to dominate the PCA. This is illustrated in the leftmost plot of Fig. 2, where Ion spectrum channels 19 to 21 display the greatest variance, whereas the B-field variance is negligible. A standard approach to address this is feature scaling, such as MinMax scaling, which adjusts each feature to lie between zero and one. However, this method disrupts the relative variance within a feature type. This effect is visible in the middle plot of Fig. 2, where the channels with the highest variance in the ion spectrum data have shifted to channels 12 and 13. To address this challenge, we propose introducing a coupling mechanism among features of the same type during scaling. Our Feature Coupled MinMax (FC-MinMax) scaling method adjusts the coupled features to the same minimum and maximum values, determined over the group of features of the same type.

The rightmost plot of Fig. 2 illustrates the variance following the application of the FC-MinMax scaling. This method maintains the relative variance within each specified group, namely the ion spectrum and B-field, while normalizing the features to achieve uniform magnitudes. It becomes evident that both the B-field and the ion spectrum will affect the PCA. Our experiments utilized the data interval one from Table 1 to determine the scaling for each feature group when analyzing the MMS dayside interval data. For other datasets, the scaling was computed based on the specific data characteristics.

3.3. Outlier detection

In previous work [13], we introduced an outlier-detection algorithm that operates in three modes: Initialization, Check, and Calibration. The system builds an initial PCA model using early samples, then enters Check mode, in which each new sample is evaluated against a dynamic threshold derived from the reconstruction errors of recent non-outlier samples. Samples exceeding this threshold are flagged as outliers and added to a calibration buffer. If a non-outlier sample is received, the calibration buffer is cleared. Once the calibration buffer is filled with a number of consecutive outlier samples (defined by a control parameter), the algorithm enters Calibration mode, updates the PCA model using the samples in the calibration buffer, and then returns to Check mode. This logic for when to add samples to the calibration buffer made the algorithm sensitive to parameter selection. In this work, we simplify the algorithm by removing the logic for deciding when to add samples to the calibration buffer and instead adding all samples to the calibration buffer. This has removed the sensitivity to parameter selection. We also extend the algorithm to operate on a time window of data rather than on individual samples. By operating on time windows, we leverage the inherently stable properties of the different plasma regions and the *time-invariant* nature of the resulting PCA components, as will be presented in Section 4.1.1, to identify the abrupt changes associated with outlier events.

Similar to the previous algorithm, this updated outlier detection algorithm operates in three different modes: Initialization, Check, and Calibrate. In the initialization mode, the initial model is built based on the first samples retrieved by the algorithm. Following the initialization mode is the check mode, which is the main operating mode. Here, samples are retrieved added to a data buffer (where the earliest sample is removed), and this buffer is then evaluated to determine if it contains outliers. When enough samples are collected to fill a data window, a calibration update is initiated. In calibration mode, the PCA model is updated based on the latest window of samples. Fig. 3 shows a flowchart of the algorithm execution flow.

The new algorithm has five control parameters: the data window size (S_d), the number of components (N) used in the PCA, the mean buffer size (S_m), the threshold (λ), and the threshold maximum (T_{max}). The parameter S_d controls the size of the data window to evaluate. This firstly controls the size of the moving window over which the algorithm evaluates the reconstruction error, and secondly sets how many samples are collected in a calibration buffer before the PCA can be updated. The mean buffer size (S_m) determines the size of the circular buffer that stores the latest reconstruction errors (E_i). This is used to calculate the mean μ and standard deviation σ and together with λ, determine the error threshold according to the equation

$$T_i = \mu + \lambda\sigma \tag{2}$$

Lastly, the threshold maximum (T_{max}) is used to limit the calculated threshold, preventing the extreme values stemming from abrupt large

Table 2
Summary of computational and memory complexities for the three algorithm modes.

Mode	Memory	Computational
Initialization	$O(NS_dF)$	$O(F^2S_d^2N)$
Check	$O(S_dF(2+N))$	$O(FS_d(2N+1)+S_m)$
Calibration	$O(S_dF(2+N))$	$O(FS_d)$

changes in the reconstruction error [13].

$$T_i = \begin{cases} \mu + T_{max} & \text{if } \lambda\sigma > T_{max} \\ \mu + \lambda\sigma & \text{otherwise} \end{cases} \quad (3)$$

Initialization mode: represented by the blue region of Fig. 3, establishes the initial PCA model. Establishing this initial model requires N samples, where N is the number of components used by the model. In this case, we are calculating PCA on a window of S_d data samples and therefore require $S_d \cdot N$ samples. The algorithm starts by retrieving the first $S_d \cdot N$ samples and adding them to the calibration buffer (c_buffer). When the number of data windows in the calibration buffer n_w is equal to the number of components N, the initial PCA model is computed. Following this, the reconstruction errors on a moving window of size S_d are then calculated on the samples used for calibration, and these are then added to the mean buffer for use in calculating the error threshold T_i.

Check mode: represented by the light green region of Fig. 3, here, the reconstruction error (E_i) is calculated as the Euclidean norm on a moving window containing the S_d latest data samples. The reconstruction error is then compared to a threshold calculated using Equation (3), where μ and σ are calculated on the latest S_m reconstruction errors stored in the mean buffer. If E_i exceeds T_i, the latest sample x_i is marked as an outlier. E_i is then added to the mean buffer and x_i is added to the calibration buffer (c_buffer). In this mode, the calibration buffer only needs to contain one data window of size S_d before the calibration can be triggered and PCA model updated.

The calibration mode: dark green region in Fig. 3, is where the PCA is updated using the samples from the calibration buffer. After updating, the calibration buffer is cleared, the algorithm reenters check mode, and the next sample is processed.

3.4. Time windowing

While useful information can be extracted from single samples, such as plasma regions [9], many of the events of interest, such as the bow shock crossings or foreshock transients, have a time component to them. You can, for example, identify the bowshock by looking for the transition from the solar wind or ion foreshock region to the magnetosheath region [11]. It can therefore be beneficial to evaluate the samples not individually, but together with other samples within a time window.

To apply PCA, each spectral time window is converted into a one-dimensional vector by flattening all feature values across the temporal dimension. For a 10-minute window containing 133 samples and 32 features (ion spectral energy channels), this results in a vector of length $133 \cdot 32$. Multiple such vectors, each representing a non-overlapping spectral window, are then concatenated to form a two-dimensional data matrix suitable for PCA.

3.5. Statistical change detection

As an alternative to simple thresholding, we explore statistical change detection approaches, focusing on the Adaptive Windowing (ADWIN) algorithm. ADWIN is an online method for detecting changes in data streams with potentially non-stationary distributions. The algorithm maintains a dynamically sized sliding window that adapts to the incoming data: with each new data point, ADWIN evaluates whether the window can be split into two segments with significantly different mean values. If a statistically significant difference is detected, ADWIN infers a change in the underlying distribution and removes the older segment of the window. Otherwise, the window expands. This adaptive process ensures the window always contains the most recent and relevant data [31].

In our work, we have used the ADWIN implementation found in River [32] version 0.23.0 and have applied ADWIN in two ways. First, it serves as an alternative to thresholding the reconstruction error, allowing us to detect statistically significant changes in the signal. Second, we use it as a standalone detector operating directly on normalized data. When ADWIN identifies a change, the most recent samples are marked as outliers, which may introduce a slight delay in detection. However, by considering the size of the ADWIN window at the time of detection, we can label a broader region as potentially containing the outlier event.

Since the base ADWIN operates on a single feature, we extend it to multivariate data by deploying a separate detector for each feature. Detected changes from each individual detector are then aggregated through a voting scheme, in which at least k detectors must signal a change for a sample to be classified as an outlier [33]. In this work, we have used $k = 19$, representing half of the features for the multi-feature case.

3.6. Computational complexity

The three operating modes, initialization, check, and calibration, each have distinct computational and memory requirements, which are summarized in Table 2. Although the initialization mode is used only briefly at startup and accounts for a minority of the total runtime, it is the most computationally and memory-intensive stage. To construct the initial PCA model, the number of samples must be at least N, the number of retained components. The memory complexity of this stage, therefore, scales as $O(NS_dF)$, where S_d denotes the size of the time window and F is the number of extracted features (either 32 for the ion-spectrum-only configuration or 38 for the multi-feature configuration). Direct PCA is applied to N samples, each of dimension $F \times S_d$ with $FS_d > N$. The resulting computational complexity is $O(F^2S_d^2N)$ corresponding to the construction of the covariance matrix and its eigendecomposition.

In the check mode, the memory requirements are more involved. Two buffers must be maintained: a time-window buffer and a calibration buffer, each requiring storage for an $F \times S_d$ matrix. In addition, the PCA projection matrix of size $N \times (FS_d)$ must be stored. The total memory complexity is therefore $O(S_dF(2+N))$. Four computations dominate this mode: (i) projection of the windowed data into and out of the PCA reduced feature space (two matrix–vector multiplications), (ii) computation of the reconstruction-error norm, (iii) computation of the mean, and (iv) computation of the standard deviation on the mean buffer. The projection operations dominate and incur a cost of $O(2NFS_d)$. The reconstruction-error norm contributes an additional $O(FS_d)$, and computing the mean and standard deviation together contributes $O(S_m)$. The total computational complexity of the check mode is thus $O(FS_d(2N+1)+S_m)$.

In calibration mode, the PCA projection matrix and the calibration buffer are reused, so the memory complexity remains unchanged from the check mode. Updating the PCA components via Incremental PCA has complexity $O(dm^2)$, where d is the dimensionality of the data matrix and m is the number of new samples incorporated per update [19]. In our case, the calibration buffer has dimension FS_d, and only a single new window is incorporated at each update; the resulting computational complexity is therefore $O(FS_d)$. Calibration is triggered only when the calibration buffer becomes full, i.e., once every S_d samples.

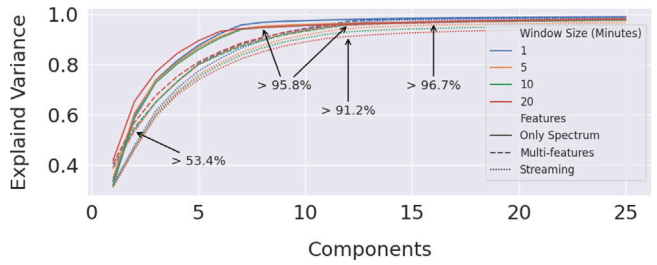


Fig. 4. The cumulative sum of the explained variance for an increasing number of principal components when the PCA is performed on the omnidirectional ion spectrum (full lines), multi-featured, B-field and Ion Velocities included (Dashed lines), and multi-featured streaming, Incremental PCA (dotted lines).

4. Results

4.1. Information content in magnetospheric data

To understand how well different time windows of the magnetospheric dayside regions can be represented by PCA, we perform PCA on time windows of different sizes and evaluate how much information is retained as the number of components increases. Fig. 4 depicts the cumulative sum of the explained variance for an increasing number of principal components when only using the ion Spectrum (full line) and when using extending the feature space to include the B-field and ion velocities (dashed line), with each feature scaled using the previously mentioned FC-MinMax scaling, on the structured dataset with only data from the regions: 'Solar Wind', 'Ion Foreshock', 'Magnetosheath' and 'Magnetosphere'. We also evaluate the explained variance when Incremental PCA is applied to the full November 2017 data, where the PCA model is built incrementally, as in the algorithm presented in Section 3.3.

For the structured data, applying a static PCA requires only two components to retain more than half of the information. However, when only using the ion spectrum, the explained variance increases more rapidly with the number of components, and only 8 components are needed to retain more than 95% of the information, whereas 12 components are needed when using multiple features. Furthermore, Fig. 4 reveals that, for structured data, the window size has only a small effect on the explained variance. This can be explained by the relative stability of the plasma regions, as seen in Fig. 1. There are temporal variations within the time windows of plasma regions, but across the full dataset, they are a minority of the data.

In the multi-feature streaming case (dotted line), there may be greater variation within the time windows. Because the windows are constructed in a streaming fashion without regard for specific regions, each window may contain multiple regions, including transitions between them, or even unknown regions, i.e., transient events or similar. This can be seen in the lower explained variance as more components are used, for example, just above 91% for 12 components and a time window of 20 min. For the streaming data, there is also a clear decrease in explained variance for a fixed number of components as the time window increases. This is expected because, as the time window increases, there is more room for transition regions and transient phenomena within it, thereby requiring more time-dependent information to explain the data variance.

4.1.1. Eigenspectrum properties

The components the PCA gives us are the eigenvectors of the data covariance matrix. If we take these flattened vectors and reshape them to the shape of the original spectrum, we get what we can call the eigenspectrum. Fig. 5 depicts the first 16 PCA eigenspectra for a window of 10 min. In the PCA, each pairing of an energy level and a time step within the data window is treated as an independent feature. The

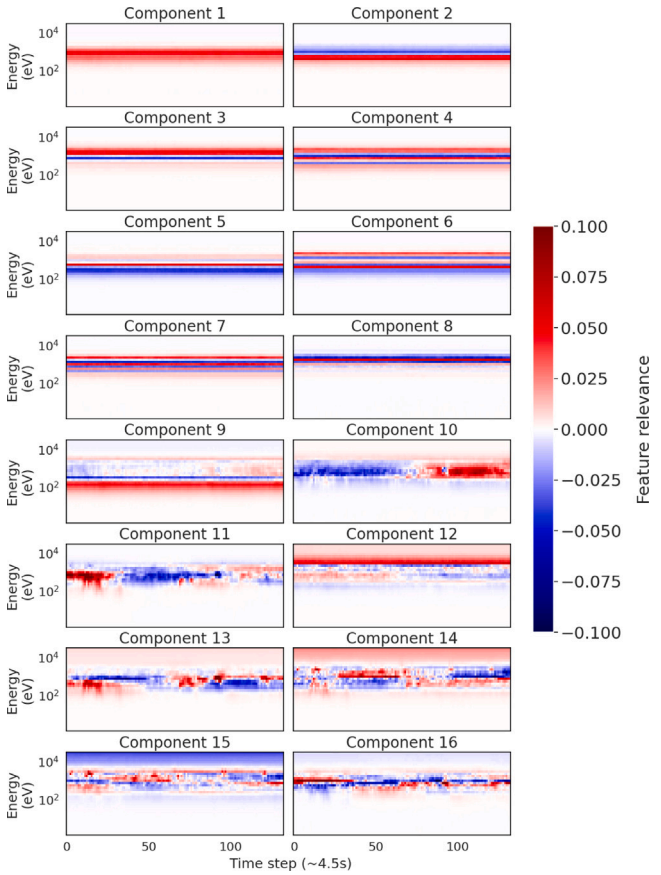


Fig. 5. The first 16 principal component eigenspectra for a 10-minute (133 samples) window. The PCA treats each combination of energy level and time step within the data window as a separate feature. The color spectrum represents how much each feature affects the corresponding principal component (feature relevance).

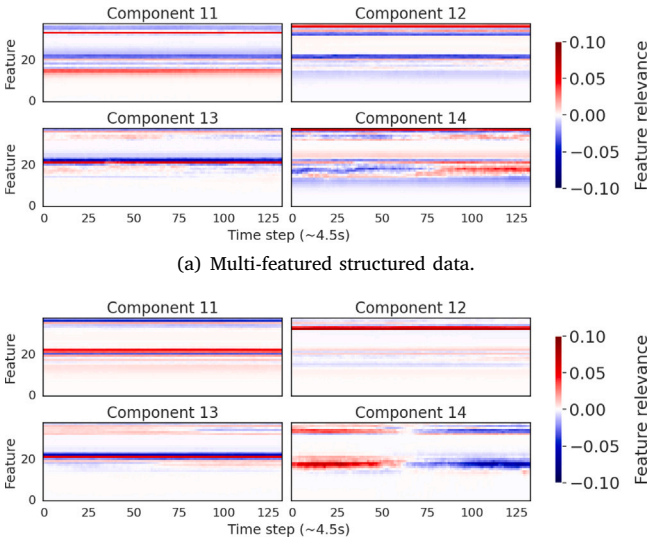
color spectrum illustrates the extent to which each feature influences the principal components, i.e., the feature relevance.

For all component eigenspectra, a central spectral line is the most prominent feature. The first component contains the central spectrum line that is present in the solar wind, ion foreshock, and, to some extent, in the magnetosheath spectra. As the number of components increases, more variation is added to the information content for the different energy levels (y-direction) by adding higher frequency components. The first eight components ar mostly *time-invariant*, i.e. the variation is mostly in the different energy levels (y-axis). However, after eight components, we start to see variations along the time axis (x-axis) also. First, low frequency changes, then higher for the later components.

The seemingly under-representation of the 10^4 region in the energy spectrum in the first eight components can partly be explained by the fact that the magnetosphere accounts for only one-fourth of the data used for the PCA, so it has less impact on the PCA components.

4.1.2. Multi-featured eigenspectrum

When extending the feature space to include the B-field and ion velocities, Fig. 4 showed that more components were required to achieve the same explained variance level as when only the ion spectrum was used. This increase corresponds well to the number of components required to express the *time-invariant* feature variance in the data window, increasing from 8 to 11 or 12. As shown in Fig. 6(a), at 11 and 12 components, slight variation along the time axis begins to appear; At 13 and 14 components, there are large variations, going from negative to positive feature relevance along the time axis for the same feature.



(a) Multi-featured structured data.

(b) Multi-featured streaming data.

Fig. 6. PCA Components 11–14 for structured multi-featured data (a) calculated on 10-minute time windowed data where each window only contains one of the plasma regions in Fig. 1, and for streaming data (b) applied to the full November 2017 dataset.

When using Incremental PCA on streaming data (Fig. 6(b)), we see a similar pattern: *time-invariant* components up to 12, with the introduction of pronounced time variance in the 13th component. Interestingly, in component 14, there is a very pronounced edge across the features in the middle of the eigenspectrum. This edge is likely due to boundary crossings in the data, e.g., bowshock crossings, as these events exhibit the largest difference between the two sides of the window. Because the data windows used to update the PCA are constructed as samples arrive, these events are not necessarily at the center of the data window. However, a central divide, as seen in component 14, likely provides the best general representation for further refinement, with higher frequencies in the higher components.

Utilizing 12 components in the multi-feature model would allow it to recreate the feature space structures but not the temporal structures, both in the structured and non-structured streaming case.

4.2. Outlier detection

To evaluate how the addition of the windowing mechanism affects the algorithm, and then how we apply it to the 2nd interval, January 12th 2018, in Fig. 7. The initial three panels in Fig. 7, labeled as (a–c), illustrate the physical characteristics. Here, several events of interest occur. Firstly, there is the foreshock transient from Table 1 at 01:50, then there is a bowshock crossing at 03:22 leading to a magnetosheath interval, a quick passage out to the solar wind region and back again at 05:15, followed by a quick entry and exit into the magnetosphere at about 05:45.

The following two panels (d and e) display the results of the algorithm when it is exclusively applied to the Ion Spectrum. The data is scaled using the FC-MinMax scaling with the scaling factors calculated on the 1st data interval from Table 1. The parameters related to the threshold calculation are set to $S_m = 170$, $\lambda = 4$, $T_{max} = 0.5$. The analysis of the principal components used for the ion spectrum in the previous section showed that eight components are enough to represent the time-invariant part of the spectrum. This means that the reconstruction error, when calculated over a time window, will be larger where the environment is more dynamic and there are sudden shifts in the data.

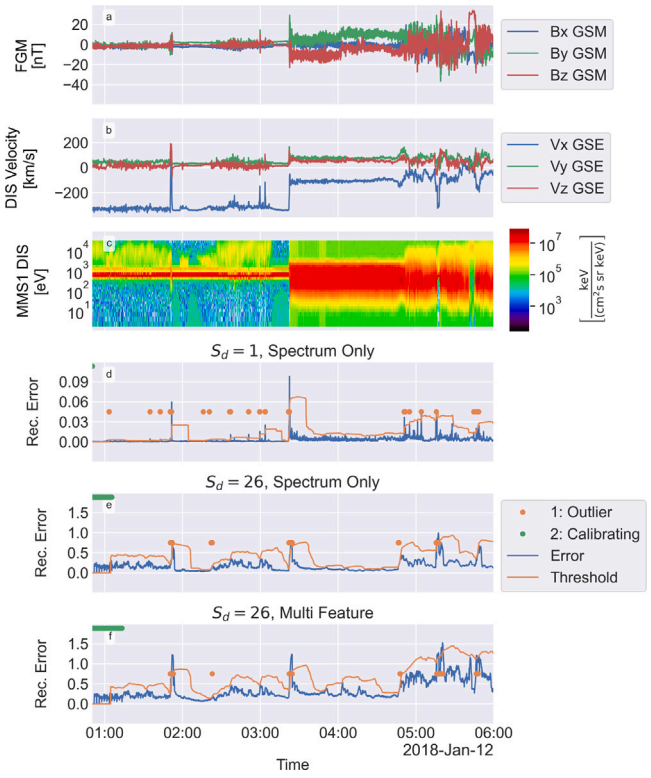


Fig. 7. The algorithm applied to the 2nd interval in Table 1. The top three panels show the physical features, B-field (a), Ion Velocities (b), and Omni-directional Ion Energy Spectrum (c). The next two panels show the algorithm's output when applied only to the Ion Spectrum, using eight principal components with window sizes $S_d = 1$ (d) and $S_d = 26$ (e), each approximately 2 min. In the bottom-most panel, the algorithm is applied to all the feature types in panels a–c, with twelve principal components and a window size $S_d = 26$.

Most of the intervals in Table 1 are around two minutes long. A natural time window for the algorithm to detect these events is 2 min, corresponding to 26 samples. Keeping the data window length short will require fewer samples for the initial model, because the number of data windows must be at least as large as the number of components to create the initial PCA model. For the spectrum-only case (d and e), this is eight to align with the *time-invariant* components, and with a window size of 26, the total number of samples needed is 208, or approximately 16 min with a sampling frequency of 4.5 s. In the top left of panels d–f of Fig. 7, the samples used for the initial calibration are marked with green dots. It becomes apparent that using long time windows is not feasible when applied to a single short interval of data.

Using eight components and a time window of $S_d = 1$, a single sample, and $S_d = 26$, approximately two minutes, panels d and e in Fig. 7, it can be seen that the reconstruction error, blue line, grows with an increase in window size; panels d and e have different scales on the y-axis to accommodate this. When the window size is $S_d = 1$, the model does not consider the time aspect of the spectrum, only evaluating the individual sample received at each time step. The reconstruction error is therefore generally low with sudden spikes when the sample does not fit the model. This is also reflected in the calculated error threshold: the low standard deviation in the reconstruction error yields a small threshold and multiple detections (orange dots), many of which can be considered erroneous.

When the reconstruction error is instead calculated on a moving window over the input samples, the dimensionality of the data increases, and the feature reduction is therefore more extreme, and the reconstruction error increases. We also see broader spikes in the reconstruction error as outlier samples transition through the moving

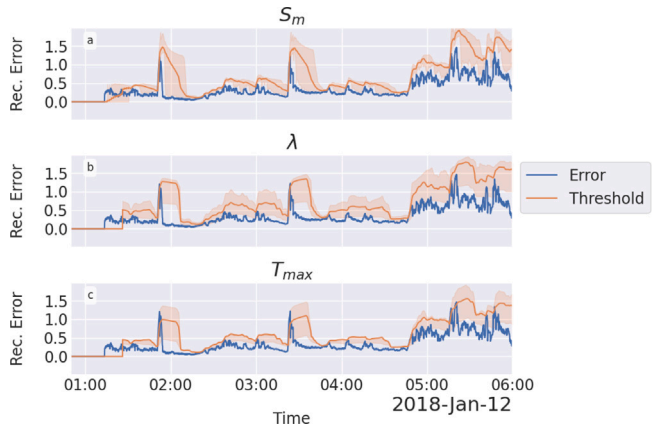
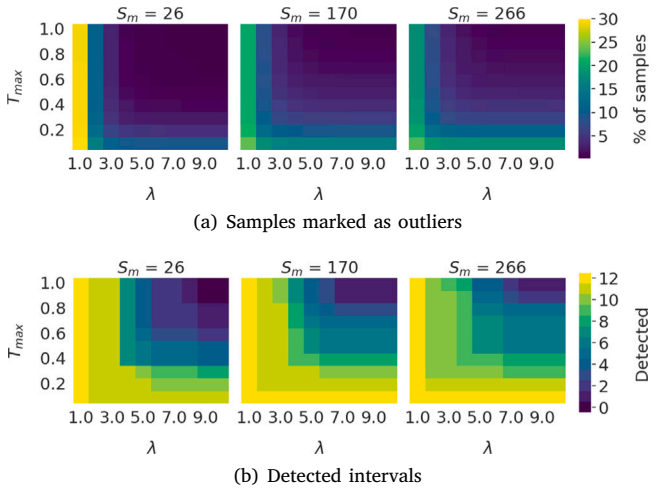


Fig. 8. Reconstruction error (blue) and threshold values (orange) for the 2nd interval in Table 1 when varying threshold parameters. The shaded orange region represents the range of maximum and minimum threshold values. Panel a show how the threshold varies when S_m is in range $[110, 210]$, with $\lambda = 5$ and $T_{max} = 0.5$. Panel b show how the threshold varies when λ is in range $[1, 10]$, with $S_m = 170$ and $T_{max} = 0.5$. Panel c show how the threshold varies when T_{max} is in range $[0.1, 1.]$, with $\lambda = 5$ and $S_m = 170$.



(a) Samples marked as outliers

(b) Detected intervals

Fig. 9. Heatmap for a grid search of the parameter space for the threshold parameters S_m, λ, and T_{Max}, where the color indicates (a) the percentage of samples marked as outliers or (b) the number of detected intervals from the dayside regions in Table 1.

window. Furthermore, the error signal in the transition regions becomes more distinct. This is especially clear in the transitions at around 05:20 and 05:50. While these transitions can be seen in the error signal for $S_d = 1$, there is no clear distinction between these and the error signals at 05:00, where the spectrum is more homogeneous. Similarly, the computed threshold for $S_d = 26$ (panel e) is larger due to the reconstruction error's higher standard deviation. The threshold remains slightly above the general data fluctuation, leading to fewer outlier detections.

However, the definition of the threshold as a function of the mean and standard deviation, see Equation (3), over a window of reconstruction error, leads to the threshold being affected by the increased error at the events themselves. This can, for example, be seen after the detection of the foreshock transient at 01:50; The sudden increase in the reconstruction error results in a corresponding increase in the threshold value, which persists until the peak has passed out of the mean buffer. The effect of this peak is limited by the T_{max} value, leading to a plateau in the threshold value. Similar plateaus can be seen after each of the peaks associated with the boundary crossings.

In the last panel (f) in Fig. 7, the algorithm is executed with all feature types presented in panels a–c as input. The data is scaled using the FC-MinMax technique, with scaling factors obtained from the 1st data interval. Here, the number of components used has been increased to twelve to account for the additional features, and a window size of $S_d = 26$ is used. The parameters related to the threshold calculation are set to $S_m = 150$, $\lambda = 4$, and $T_{max} = 0.5$.

By including the B-field and ion velocities in the feature space, the reconstruction error signal gets more pronounced for both the foreshock transient at 01:50 and the later region transitions. Furthermore, there is a clear peak in the reconstruction error at the short entry into the magnetosphere region at around 05:45, which was not present when only the ion spectrum was used. However, this transition is missed due to the threshold.

4.3. Threshold stability

To evaluate the sensitivity of the algorithm to the threshold parameters, mean buffer size S_m, threshold value λ, and threshold maximum T_{max}, we begin by applying the algorithm to the 2nd interval in Table 1 using multiple features, varying each parameter individually, the number of components and the data window have been kept fixed to $N = 12$ and $S_d = 26$. In the top panel (a) of Fig. 8, S_m is varied from 26 to 226 while $\lambda = 4$ and $T_{max} = 1.0$. The orange line shows the mean threshold value, and the shaded region indicates the range from minimum to maximum thresholds. For smaller S_m values, the most recent reconstruction error comprises a larger portion of the buffer, exerting a greater influence on the calculated threshold. This has two effects: the buffer mean is closer to the current sample, and sudden changes in the reconstruction error lead to larger changes in the standard deviation (σ). This is reflected in spikes in the maximum threshold value that coincide with spikes in the reconstruction error, and in a rapid decrease in the minimum threshold value after such spikes. Increasing S_m results in much smoother threshold variations, while, as noted earlier, the threshold remains elevated after spikes in the reconstruction error.

In panel (b), λ is varied from 1 to 10 while $S_m = 170$ and $T_{max} = 1.0$. The minimum threshold value corresponds to $\lambda = 1$, and the maximum to $\lambda = 10$. When the mean and standard deviation of the reconstruction error change slowly, the value of λ greatly affects whether the threshold intersects the reconstruction error and whether outliers are detected. For $\lambda = 1$, the threshold closely follows the reconstruction error, resulting in a large proportion of samples being marked as outliers. Conversely, for larger λ values (tested up to $\lambda = 10$), most small-scale changes in the reconstruction error are not flagged as outliers. At points where there are spikes in the reconstruction error, such as at 01:50 in Fig. 8, the standard deviation of the buffer increases, causing the threshold to be limited primarily by T_{max}. This is evident as the mean (dark orange line) approaches the upper boundary of the shaded region, and is further illustrated in panel (c), where T_{max} is varied from 0.1 to 1.0 with $\lambda = 4$ and $S_m = 170$. In panel (c), the effect of T_{max} on the threshold is the opposite of λ's effect in panel (b): T_{max} has the largest impact in regions with sudden spikes (large standard deviation), but minimal impact where the standard deviation is small.

Fig. 8 shows that selecting T_{max} is most critical for detecting spikes in the reconstruction error, particularly when multiple spikes occur, as illustrated around 05:20. However, in regions with less distinct spikes, such as at 05:45, the λ value plays a more significant role. For S_m, the choice involves balancing a more stable threshold achieved with a larger S_m against the tradeoff of maintaining a high threshold for an extended period after spikes.

To further understand the combinatorial space of the threshold parameters, we performed a grid search using the same ranges as in Fig.

Table 3
Comparison of the method used in this study to ADWIN [31].

Method	Detected intervals	Outlier groups	% of total samples
Our Method	7	70	2.89
Our Method with ADWIN	2	136	0.32
ADWIN with vote	6	66	0.16
ADWIN with length estimation	6	66	7.27

8 and plotted this as heatmaps against percentage of samples marked as outliers, Fig. 9(a), and number of detected dayside intervals from Table 1, Fig. 9(b). It becomes apparent that the parameter selection becomes a trade-off between the number of detected intervals and the percentage of total samples marked as outliers, especially for selecting λ and T_{max}, while S_m is less relevant. To maintain a low percentage of samples marked as outliers, roughly 2%, while detecting as many of the dayside intervals in Table 1 as possible, we have elected to use $S_m = 170$, $\lambda = 5$, and $T_{max} = 0.5$.

4.4. Statistical change detection

A comparison between our approach, which uses *time-invariant* components and thresholded reconstruction error, and the ADWIN-based method shows that our approach performs comparably or better, as shown in Table 3. The ADWIN algorithm was implemented using the default River parameters, except for the clock parameter, which determines the frequency of change-point evaluations. The clock parameter was set to 26 to align with the S_d parameter.

When ADWIN is applied to the reconstruction error produced by Incremental PCA, the resulting performance is notably poor. In contrast, applying ADWIN directly to the normalized data, in conjunction with a voting scheme, achieves performance that approaches that of our proposed method. Nevertheless, designating a sample as an outlier only upon detecting drift results in the identification of single-sample outliers. Consequently, this approach does not convey information about the duration of the outlier event, and the proportion of samples marked as outliers is substantially lower than in our method. Alternatively, if the length of the outlier is inferred based on the size of the ADWIN window at the time of detection, as illustrated in the final row of Table 3, the proportion of samples marked as outliers increases excessively.

4.5. Application on MMS and THEMIS data

Selecting to use $N = 12$ due to the *time-invariant* nature of these first components, and the data window to $S_d = 26$ (~2 min), as most of the events in Table 1 are around 2 min long. The remaining parameter are set to $S_m = 170$, $\lambda = 5$ and $T_{max} = 0.5$. With these parameters, the algorithm is applied to each interval in Table 1, resetting only the mean buffer between intervals while keeping the PCA model. With this, seven of the twelve foreshock transients (dayside events) from the table are detected from Table 1. Of the five missed events, one, the 5th interval, has a distinct peak in the reconstruction error. However, the threshold's base level is sufficiently high to react quickly to the increase. For the remaining four missed detections, the events are in more dynamic environments, making their detection challenging. While many of them have visual peaks in the reconstruction error, it can be hard to distinguish these peaks from peaks surrounding them. If subsequent samples marked as outliers are grouped, seven outliers are connected to the foreshock transient events from Table 1, another 17 are other foreshock transient events, 24 are related to bow shock crossings, another nine are possible partial bow shock crossings in succession, and five are likely transitions between quasi-parallel and quasi-perpendicular regions of the magnetosheath. The last eight detections require more detailed analysis. Of all the samples, 0.7% are used for the initial calibration, and 2.9% are marked as outliers of potential interest.

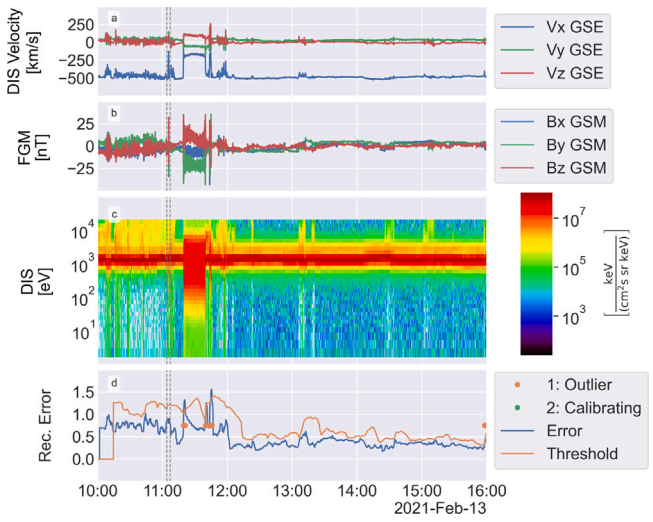


Fig. 10. The algorithm applied to the 7th interval in Table 1. The top three panels show the physical features, B-field (a), Ion Velocities (b), and Omni-directional Ion Energy Spectrum (c). Panel d shows the reconstruction error, error threshold, and samples classified as outliers. Here, the algorithm does not find the foreshock transient (dashed lines); however, the bowshock crossings are found.

Fig. 10 shows the event from the 7th interval, which contains the least clear peak. Here, the foreshock transient at 11:05 is not detected. There is a slight bump in the reconstruction error signal at the time of the event. However, even visually, it is challenging to pick out the event in the reconstruction error without prior knowledge. The algorithm, however, does find the bow shock crossing out of the magnetosheath at 11:19 and the subsequent quick crossing in and out of the magnetosheath, at 11:44 to 11:45.

MMS-1 Nightside Data: To evaluate the algorithm's generalization to a different region of the magnetosphere, we use it on the nightside data from Table 1 with the same features and settings as the dayside data, with the FC-MinMax scaling calculated on the corresponding nightside interval. In this case, both region transitions between the plasma sheet and the inner magnetosphere are found, along with the transient fast flows. However, the number of outlier detections appears to be very high. The T_{max} value that worked for the dayside is therefore too low for the nightside. Increasing this to $T_{max} = 0.8$ leads to fewer detections but also causes the event in interval 11 to be missed.

The nightside magnetosphere is, in general, a more dynamic environment than the dayside, with less well-defined spatial boundaries [28]. This is reflected in the reconstruction error as shown in Fig. 11, panel d. At the beginning of the Figure, from 16:00 to around 21:00, the spacecraft is in a low activity region of the plasma sheet. This is reflected in the reconstruction error, panel d, which has values similar to what could be seen in the dayside regions. There are some smaller peaks, most of which are not classified as outliers. Following this, between 21:00 to around midnight, is a very calm boundary layer toward the inner magnetosphere. Then, from midnight onward, the spacecraft enters a highly dynamic region of the plasma sheet with multiple fast plasma flows. Here, the variance in the reconstruction error is larger with multiple sharp peaks. At 01:23, marked with gray dashed lines, there is a peak corresponding to the fast plasma flow from Table 1 and has been discussed by Richard et al. [22]. In this case, most of the parameters for the algorithm were the same as for the dayside regions, $N = 12$, $S_d = 26$, $S_m = 170$, and $\lambda = 5$. However, the T_{max} was increased to 0.8 to reflect the more dynamic environment.

THEMIS C Dayside Data: The algorithm's generalization to a different feature set from a different spacecraft was assessed by applying the algorithm to a daytime interval of THEMIS C data. The

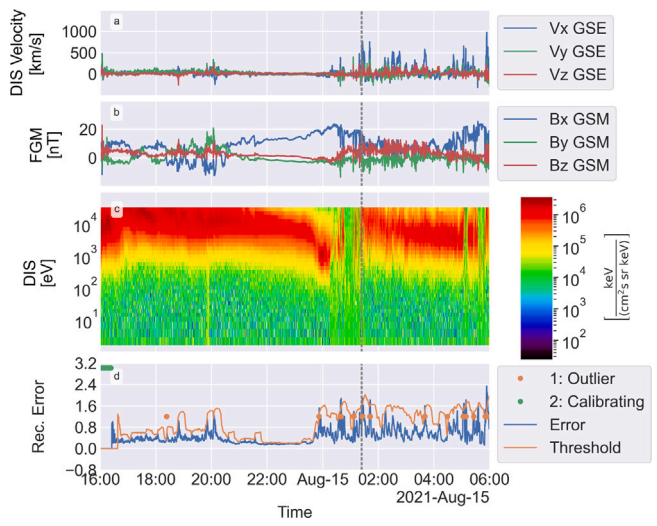


Fig. 11. The algorithm applied to MMS Nightside data, 12th interval from Table 1. The top three panels show the physical features, B-field (a), Ion Velocities (b), and Omni-directional Ion Energy Spectrum (c). The more dynamic nightside magnetosphere leads to greater variance in the reconstruction error (c).

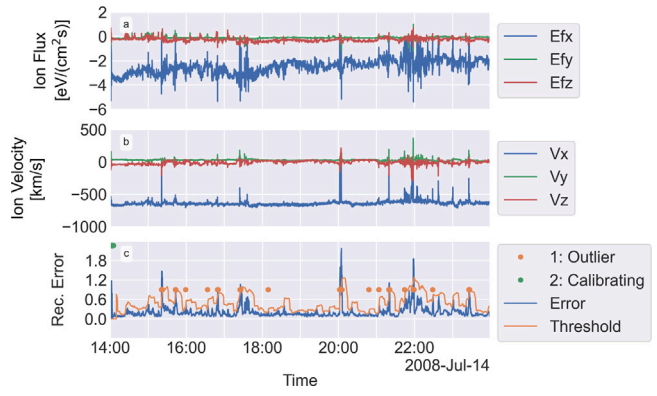


Fig. 12. The algorithm applied to an interval of dayside data from THEMIS C. The top two panels show the physical features, the ion energy flux (a), and the ion velocities (b). The peak in the reconstruction error in panel c at 21:57 corresponds to a foreshock bubble [30,34].

same parameters as for the MMS dayside were used, but the number of components was reduced to four due to only using six features from the THEMIS data, ion energy flux, and ion velocity in the x, y, and z GSE directions. Scaling was calculated on the corresponding THEMIS interval. The features, panel a–b, and the reconstruction error, panel c, are shown in Fig. 12. As can be seen by the marked outliers, several possible interesting events are found; The longest of these is between 20:01:51 and 20:05:42, and between 21:57:56 and 21:59:17. This second event, starting at 21:57:56, has been noted in previous works as a foreshock bubble [30,34]. The start and stop times indicated by our algorithm are similar to the start of expansion to the end of compression, shown by Liu et al. [30] in Fig. 2.

5. Discussion

In this study, we have proposed a method for finding scientifically important events in multi-featured space plasma measurements, based on time-invariant components from the PCA. The method provides a promising, interpretable, unsupervised approach for identifying scientifically relevant events in in-situ plasma measurements by leveraging structures in feature reduction. However, several challenges with the proposed method remain, especially with the adaptability of the method to new regions or when applying to different datasets; foremost of these are identifying the number of *time-invariant* PCA components, setting appropriate threshold parameters (S_m, λ and T_{max}), and the calibration of the FC-MinMax scaling required for multi-feature application.

Using a temporal data window, we have shown that a limited set of PCA components can account for over 92% of the variation in the data when applied to plasma-region data windows on Earth's dayside. We have also shown that these PCA components are essentially *time-invariant*, highlighting the short-term invariance of the data in these regions. Using these *time-invariant* components, we showed that scientifically relevant events can be located using a reconstruction-error-based approach with a dynamic threshold. However, identifying the number of *time-invariant* components requires prior knowledge of the data. An extended initialization phase, during which the algorithm autonomously calibrates the optimal number of components, could help reduce reliance on manual data evaluation. Furthermore, while the results presented in this study suggest that the algorithm can be applied to both dayside and nightside data, it is not yet fully evaluated how the number of *time-invariant* components changes when transitioning between distinct regions, such as from the dayside to the more dynamic nightside.

The selection of appropriate threshold parameters is the second challenge with the application of the method to new data. While an appropriately selected T_{max} can enable the identification of the most distinct outliers, and the selection of λ can enable the identification of events with less distinct error signals or events seen as slow shifts in the standard deviation of the reconstruction error. We have seen that at least T_{max} had to be adjusted when moving to the more dynamic nightside region. Nonetheless, using a threshold remains the most effective when compared to alternatives such as ADWIN. One solution for T_{max} is to define it in terms of the mean and std over a larger buffer than currently used to calculate the threshold. Thus, you would set the threshold to the minimum of the values from a fast-adapting and a slow-adapting component.

Lastly, the FC-MinMax scaling is used to normalize the feature values, enabling the integration of features with differing magnitudes. However, the reliance on this pre-scaling step limits the algorithm's applicability to regions where the features have magnitudes comparable to those used to compute the scaling values. Suppose a new region is encountered where one or more features have values with larger magnitudes. In that case, these will start to dominate the model, making the other features overrepresented in the reconstruction error. To enhance the method and expand its utility, future work will assess strategies to embed it directly into the algorithm. In such a scenario, the scaling would be set during the algorithm's initialization phase. Then, during recalibration, the new data could be compared with the existing scaler to determine whether updates are necessary.

In addition to guiding scientists toward relevant data, the presented method and algorithm could also be used to prioritize data for downlink onboard spacecraft. The THEMIS C dataset analyzed in this study reflects data that could be accessible on board, highlighting the algorithm's potential for onboard application in selecting which data to prioritize for downlink. In actual mission scenarios, the algorithm would be applied to data that has already passed through a robust onboard pre-processing step to detect sensor anomalies or failures. Nevertheless, further extensive testing using onboard data from multiple spacecraft is necessary to demonstrate the robustness of the approach for future space missions. Moreover, a key direction for future algorithm development is to implement an optimized version for embedded systems and evaluate its performance on representative hardware before deployment on space missions.

6. Conclusion

In this study, we have investigated in-situ space plasma measurements, primarily the omnidirectional ion spectrum, magnetic field, and ion velocities, from Earth's dayside magnetosphere. By using time windowing, we have shown that, in a dynamic environment, the information content can be expressed in a relatively small number of reduced features. Furthermore, the features that contain a majority (>95%) of the information are *time-invariant*, i.e., when the PCA component is projected onto a spectral image, there is almost no variation along the time axis.

By leveraging the *time-invariant* components and the reconstruction error from PCA feature reduction, we propose an interpretable, unsupervised method for identifying scientifically relevant events in the data using an updated version of the outlier detection algorithm for streaming data introduced in our previous work. We show that the proposed method can find a majority of events described by Raptis et al. [7] while keeping the percentage of samples marked as outliers low. In addition, the method can find events such as bowshock or magnetopause crossings.

CRediT authorship contribution statement

Jonah Ekelund: Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Conceptualization. **Savvas Raptis:** Writing – review & editing, Validation, Methodology, Data curation, Conceptualization. **Vicki Toy-Edens:** Writing – review & editing. **Wenli Mo:** Writing – review & editing. **Drew L. Turner:** Writing – review & editing. **Ian J. Cohen:** Writing – review & editing. **Stefano Markidis:** Writing – review & editing, Writing – original draft, Supervision, Resources, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by the European Union's HORIZON Research and Innovation action via ASAP - Automatics in Space (https://asap-space.eu) under the grant agreement no. 101082633.

SR acknowledges funding from the MMS Early Career Award 80NSSC25K7353 and from the Johns Hopkins University Applied Physics Laboratory independent R&D fund.

Data availability

Data will be made available on request.

References

[1] J.L. Burch, T.E. Moore, R.B. Torbert, B.L. Giles, Magnetospheric multiscale overview and science objectives, Space Sci. Rev. 199 (2016).

[2] D.N. Baker, L. Riesberg, C.K. Pankratz, R.S. Panneton, B.L. Giles, F.D. Wilder, R.E. Ergun, Magnetospheric multiscale instrument suite operations and data system, Space Sci. Rev. (ISSN: 1572-9672) 199 (1) (2016) 545–575, http://dx.doi.org/10.1007/s11214-014-0128-5.

[3] V. Angelopoulos, The THEMIS mission, Space Sci. Rev. 141 (1) (2008) 5–34.

[4] J. Credland, G. Mecke, J. Ellwood, The cluster mission: Esa's spacefleet to the magnetosphere, Space Sci. Rev. (ISSN: 1572-9672) 79 (1) (1997) 33–64, http://dx.doi.org/10.1023/A:1004914822769.

[5] B. Grison, F. Darrouzet, R. Maggiolo, M. Hajoš, M. Dvořák, M. Švanda, A. Jeřábková, M.G.G.T. Taylor, D. Herment, A. Masson, J. Souček, O. Santolík, J. De Keyser, Localization of the cluster satellites in the geospace environment, Sci. Data (ISSN: 2052-4463) 12 (1) (2025) 327, http://dx.doi.org/10.1038/s41597-025-04639-z.

[6] H. Breuillard, R. Dupuis, A. Retino, O. Le Contel, J. Amaya, G. Lapenta, Automatic classification of plasma regions in near-earth space with supervised machine learning: Application to magnetospheric multi scale 2016–2019 observations, Front. Astron. Space Sci. 7 (2020).

[7] S. Raptis, A. Lalti, M. Lindberg, D.L. Turner, D. Caprioli, J.L. Burch, Revealing an unexpectedly low electron injection threshold via reinforced shock acceleration, Nat. Commun. 16 (2025).

[8] M.G. Finley, M. Martinez-Ledesma, W.R. Paterson, M.R. Argall, D.M. Miles, J.C. Dorelli, E. Zesta, Generalized Time-Series Analysis for In Situ Spacecraft Observations: Anomaly Detection and Data Prioritization Using Principal Components Analysis and Unsupervised Clustering, Earth Space Sci. 11 (2024).

[9] J. Ekelund, R. Vinuesa, Y. Khotyaintsev, P. Henri, G.L. Delzanno, S. Markidis, AI in space for scientific missions: Strategies for minimizing neural-network model upload, in: 2024 IEEE 20th International Conference on E-Science (E-Science), 2024, pp. 1–10.

[10] S. Raptis, T. Karlsson, F. Plaschke, A. Kullen, P.-A. Lindqvist, Classifying magnetosheath jets using MMS: Statistical properties, J. Geophys. Res.: Space Phys. 125 (11) (2020).

[11] V. Olshevsky, Y.V. Khotyaintsev, A. Lalti, A. Divin, G.L. Delzanno, S. Anderzén, P. Herman, S.W.D. Chien, L. Avanov, A.P. Dimmock, S. Markidis, Automated classification of plasma regions using 3D particle energy distributions, J. Geophys. Res.: Space Phys. 126 (2021).

[12] V. Toy-Edens, W. Mo, S. Raptis, D.L. Turner, Classifying 8 Years of MMS Dayside Plasma Regions via Unsupervised Machine Learning, J. Geophys. Res.: Space Phys. 129 (2024).

[13] J. Ekelund, S. Raptis, V. Toy-Edens, W. Mo, D.L. Turner, I.J. Cohen, S. Markidis, Adaptive PCA-based outlier detection for multi-feature time series in space missions, in: M.H. Lees, W. Cai, S.A. Cheong, Y. Su, D. Abramson, J.J. Dongarra, P.M.A. Sloot (Eds.), Computational Science – ICCS 2025, Springer Nature Switzerland, ISBN: 978-3-031-97626-1, 2025, pp. 253–267.

[14] M.R. Bakrania, I.J. Rae, A.P. Walsh, D. Verscharen, A.W. Smith, Using dimensionality reduction and clustering techniques to classify space plasma regimes, Front. Astron. Space Sci. 7 (2020).

[15] C. Escoubet, M. Fehringer, M. Goldstein, Introduction the cluster mission, in: Annales Geophysicae, vol. 19, Copernicus GmbH, 2001, pp. 1197–1200.

[16] M.E. Innocenti, J. Amaya, J. Raeder, R. Dupuis, B. Ferdousi, G. Lapenta, Unsupervised classification of simulated magnetospheric regions, Ann. Geophys. 39 (2021).

[17] C.-C.M. Yeh, Y. Zhu, L. Ulanova, N. Begum, Y. Ding, H.A. Dau, D.F. Silva, A. Mueen, E. Keogh, Matrix profile I: All pairs similarity joins for time series: A unifying view that includes motifs, discords and shapelets, in: 2016 IEEE 16th International Conference on Data Mining, ICDM, 2016, pp. 1317–1322.

[18] I. Souiden, M.N. Omri, Z. Brahmi, A survey of outlier detection in high dimensional data streams, Comput. Sci. Rev. 44 (2022).

[19] D.A. Ross, J. Lim, R.-S. Lin, M.-H. Yang, Incremental learning for robust visual tracking, Int. J. Comput. Vis. 77 (2008).

[20] N.M. Zamry, A. Zainal, M.A. Rassam, E.H. Alkhammash, F.A. Ghaleb, F. Saeed, Lightweight Anomaly Detection Scheme Using Incremental Principal Component Analysis and Support Vector Machine, Sensors 21 (2021).

[21] A. Bhushan, M.H. Sharker, H.A. Karimi, Incremental principal component analysis based outlier detection methods for spatiotemporal data streams, in: ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, University of Pittsburgh, 2015, pp. 67–71.

[22] L. Richard, Y.V. Khotyaintsev, D.B. Graham, C.T. Russell, Are dipolarization fronts a typical feature of magnetotail plasma jets fronts? Geophys. Res. Lett. 49 (2022).

[23] C. Pollock, T. Moore, A. Jacques, J. Burch, U. Gliese, Y. Saito, T. Omoto, L. Avanov, A. Barrie, V. Coffey, J. Dorelli, D. Gershman, B. Giles, T. Rosnack, C. Salo, S. Yokota, M. Adrian, C. Aoustin, C. Auletti, S. Aung, V. Bigio, N. Cao, M. Chandler, D. Chornay, K. Christian, G. Clark, G. Collinson, T. Corris, A. De Los Santos, R. Devlin, T. Diaz, T. Dickerson, C. Dickson, A. Diekmann, F. Diggs, C. Duncan, A. Figueroa-Vinas, C. Firman, M. Freeman, N. Galassi, K. Garcia, G. Goodhart, D. Guererro, J. Hageman, J. Hanley, E. Hemminger, M. Holland, M. Hutchins, T. James, W. Jones, S. Kreisler, J. Kujawski, V. Lavu, J. Lobell, E. LeCompte, A. Lukemire, E. MacDonald, A. Mariano, T. Mukai, K. Narayanan, Q. Nguyan, M. Onizuka, W. Paterson, S. Persyn, B. Piepgrass, F. Cheney, A. Rager, T. Raghuram, A. Ramil, L. Reichenthal, H. Rodriguez, J. Rouzaud, A. Rucker, Y. Saito, M. Samara, J.-A. Sauvaud, D. Schuster, M. Shappirio, K. Shelton, D. Sher, D. Smith, K. Smith, S. Smith, D. Steinfeld, R. Szymkiewicz, K. Tanimoto, J. Taylor, C. Tucker, K. Tull, A. Uhl, J. Vloet, P. Walpole, S. Weidner, D. White, G. Winkert, P.-S. Yeh, M. Zeuch, Fast Plasma Investigation for Magnetospheric Multiscale, Space Sci. Rev. 199 (2016).

[24] R.B. Torbert, C.T. Russell, W. Magnes, R.E. Ergun, P.-A. Lindqvist, O. LeContel, H. Vaith, J. Macri, S. Myers, D. Rau, J. Needell, B. King, M. Granoff, M. Chutter, I. Dors, G. Olsson, Y.V. Khotyaintsev, A. Eriksson, C.A. Kletzing, S. Bounds, B. Anderson, W. Baumjohann, M. Steller, K. Bromund, G. Le, R. Nakamura, R.J. Strangeway, H.K. Leinweber, S. Tucker, J. Westfall, D. Fischer, F. Plaschke, J. Porter, K. Lappalainen, The FIELDS instrument suite on MMS: Scientific objectives, measurements, and data products, Space Sci. Rev. 199 (2016).

[25] J. Ekelund, S. Raptis, S. Markidis, SpacePhyML: Enabling access to space physics data for machine learning applications, 2025, http://dx.doi.org/10.5281/zenodo.17152371.
[26] P. Kajdič, X. Blanco-Cano, L. Turc, M. Archer, S. Raptis, T.Z. Liu, Y. Pfau-Kempf, A.T. LaMoury, Y. Hao, P.C. Escoubet, N. Omidi, D.G. Sibeck, B. Wang, H. Zhang, Y. Lin, Transient upstream mesoscale structures: drivers of solar-quiet space weather, Front. Astron. Space Sci. 11 (2024).
[27] V. Angelopoulos, W. Baumjohann, C. Kennel, F.V. Coroniti, M. Kivelson, R. Pellat, R. Walker, H. Lühr, G. Paschmann, Bursty bulk flows in the inner central plasma sheet, J. Geophys. Res.: Space Phys. 97 (A4) (1992) 4027–4039.
[28] M. Sitnov, J. Birn, B. Ferdousi, E. Gordeev, Y. Khotyaintsev, V. Merkin, T. Motoba, A. Otto, E. Panov, P. Pritchett, F. Pucci, J. Raeder, A. Runov, V. Sergeev, M. Velli, X. Zhou, Explosive magnetotail activity, Space Sci. Rev. 215 (2019).
[29] Themis, Variable descriptions, 2025, URL https://themis.igpp.ucla.edu/var_desc.shtml. (Accessed 23 February 2025).
[30] T.Z. Liu, V. Angelopoulos, S. Lu, Relativistic electrons generated at earth's quasi-parallel bow shock, Sci. Adv. 5 (2019).
[31] A. Bifet, R. Gavalda, Learning from time-changing data with adaptive windowing, in: Proceedings of the 2007 SIAM International Conference on Data Mining, SIAM, 2007, pp. 443–448.
[32] J. Montiel, M. Halford, S.M. Mastelini, G. Bolmier, R. Sourty, R. Vaysse, A. Zouitine, H.M. Gomes, J. Read, T. Abdessalem, et al., River: machine learning for streaming data in python, 2021.
[33] W.J. Faithfull, J.J. Rodríguez, L.I. Kuncheva, Combining univariate approaches for ensemble change detection in multivariate data, Inf. Fusion (ISSN: 1566-2535) 45 (2019) 202–214.
[34] L.B. Wilson, D.G. Sibeck, D.L. Turner, A. Osmane, D. Caprioli, V. Angelopoulos, Relativistic electrons produced by foreshock disturbances observed upstream of earth's bow shock, Phys. Rev. Lett. 117 (2016).