

# Northumbria Research Link

Citation: Bakrania, Mayur R., Rae, Jonathan, Walsh, Andrew P., Verscharen, Daniel and Smith, Andy W. (2020) Using Dimensionality Reduction and Clustering Techniques to Classify Space Plasma Regimes. *Frontiers in Astronomy and Space Sciences*, 7. p. 593516. ISSN 2296-987X

Published by: Frontiers

URL: <https://doi.org/10.3389/fspas.2020.593516>  
<<https://doi.org/10.3389/fspas.2020.593516>>

This version was downloaded from Northumbria Research Link:  
<http://nrl.northumbria.ac.uk/id/eprint/49238/>

Northumbria University has developed Northumbria Research Link (NRL) to enable users to access the University's research output. Copyright © and moral rights for items on NRL are retained by the individual author(s) and/or other copyright owners. Single copies of full items can be reproduced, displayed or performed, and given to third parties in any format or medium for personal research or study, educational, or not-for-profit purposes without prior permission or charge, provided the authors, title and full bibliographic details are given, as well as a hyperlink and/or URL to the original metadata page. The content must not be changed in any way. Full items must not be sold commercially in any format or medium without formal permission of the copyright holder. The full policy is available online: <http://nrl.northumbria.ac.uk/policies.html>

This document may differ from the final, published version of the research and has been made available online in accordance with publisher policies. To read and/or cite from the published version of the research, please visit the publisher's website (a subscription may be required.)



# Using Dimensionality Reduction and Clustering Techniques to Classify Space Plasma Regimes

Mayur R. Bakrania<sup>1\*</sup>, I. Jonathan Rae<sup>1,2</sup>, Andrew P. Walsh<sup>3</sup>, Daniel Verscharen<sup>1,4</sup> and Andy W. Smith<sup>1</sup>

<sup>1</sup>Department of Space and Climate Physics, Mullard Space Science Laboratory, University College London, Dorking, United Kingdom, <sup>2</sup>Department of Mathematics, Physics and Electrical Engineering, University of Northumbria, Newcastle, United Kingdom, <sup>3</sup>European Space Astronomy Centre, ESA, Madrid, Spain, <sup>4</sup>Space Science Center, University of New Hampshire, Durham, NH, United States

## OPEN ACCESS

### Edited by:

Vladislav Izmodenov,  
Space Research Institute (RAS),  
Russia

### Reviewed by:

Anton Artemyev,  
Space Research Institute (RAS),  
Russia  
Olga Khabarova,  
Institute of Terrestrial Magnetism  
Ionosphere and Radio Wave  
Propagation (RAS), Russia

### \*Correspondence:

Mayur R. Bakrania  
mayur.bakrania@ucl.ac.uk

### Specialty section:

This article was submitted  
to Space Physics,  
a section of the journal  
Frontiers in Astronomy and Space  
Sciences

**Received:** 10 August 2020

**Accepted:** 22 September 2020

**Published:** 21 October 2020

### Citation:

Bakrania MR, Rae IJ, Walsh AP,  
Verscharen D and Smith AW (2020)  
Using Dimensionality Reduction and  
Clustering Techniques to Classify  
Space Plasma Regimes.  
Front. Astron. Space Sci. 7:593516.  
doi: 10.3389/fspas.2020.593516

Collisionless space plasma environments are typically characterized by distinct particle populations. Although moments of their velocity distribution functions help in distinguishing different plasma regimes, the distribution functions themselves provide more comprehensive information about the plasma state, especially at times when the distribution function includes non-thermal effects. Unlike moments, however, distribution functions are not easily characterized by a small number of parameters, making their classification more difficult to achieve. In order to perform this classification, we propose to distinguish between the different plasma regions by applying dimensionality reduction and clustering methods to electron distributions in pitch angle and energy space. We utilize four separate algorithms to achieve our plasma classifications: autoencoders, principal component analysis, mean shift, and agglomerative clustering. We test our classification algorithms by applying our scheme to data from the Cluster-Plasma Electron and Current Experiment instrument measured in the Earth's magnetotail. Traditionally, it is thought that the Earth's magnetotail is split into three different regions (the plasma sheet, the plasma sheet boundary layer, and the lobes), that are primarily defined by their plasma characteristics. Starting with the ECLAT database with associated classifications based on the plasma parameters, we identify eight distinct groups of distributions, that are dependent upon significantly more complex plasma and field dynamics. By comparing the average distributions as well as the plasma and magnetic field parameters for each region, we relate several of the groups to different plasma sheet populations, and the rest we attribute to the plasma sheet boundary layer and the lobes. We find clear distinctions between each of our classified regions and the ECLAT results. The automated classification of different regions in space plasma environments provides a useful tool to identify the physical processes governing particle populations in near-Earth space. These tools are model independent, providing reproducible results without requiring the placement of arbitrary thresholds, limits or expert judgment. Similar methods could be used onboard spacecraft to reduce the dimensionality of distributions in order to optimize data collection and downlink resources in future missions.

**Keywords:** space plasma environments, particle populations, distribution functions, dimensionality reduction, clustering techniques

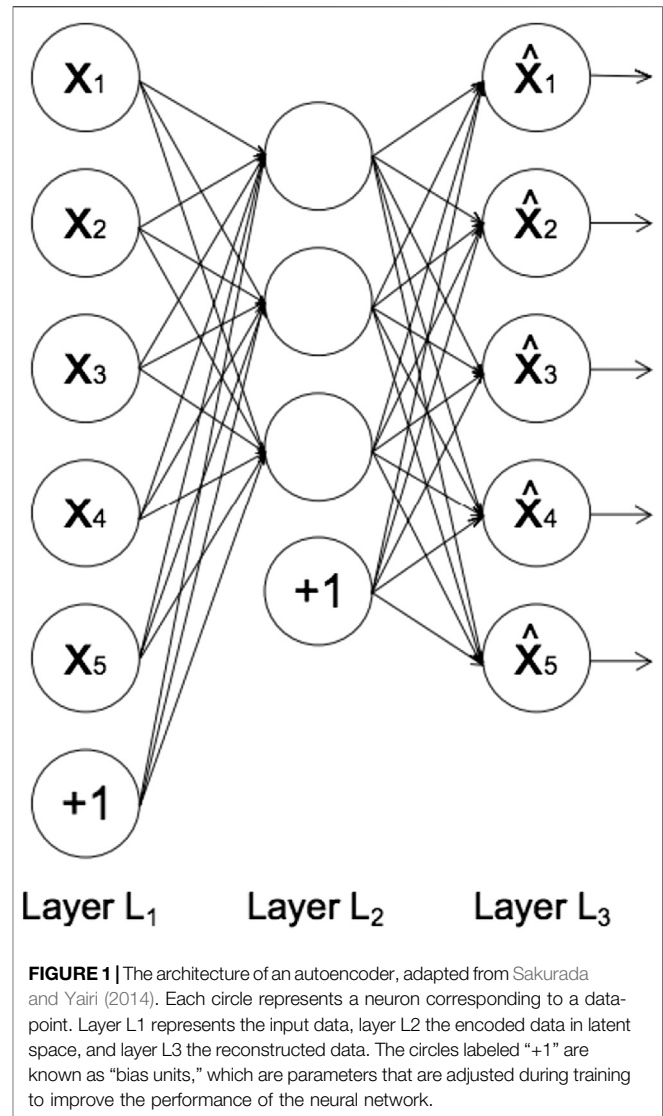
## 1 INTRODUCTION

Particle populations in collisionless space plasma environments, such as the Earth's magnetotail, are traditionally characterized by the moments of their distribution functions. 2D distribution functions in pitch angle and energy, however, provide the full picture of the state of each plasma environment, especially when non-thermal particle populations are present that are less easily characterized by a Maxwellian fit. These non-thermal plasma populations are ubiquitous across the solar system. They make crucial contributions to the bulk properties of a plasma, such as the temperature and collisionality (Hapgood et al., 2011). Magnetic reconnection, for example, heats non-thermal seed populations in both the diffusion and outflow regions, making them an important component of the overall energization process (Øieroset et al., 2002). High-quality measurements and analysis of collisionless plasmas are consequently of key importance when attempting to understand these non-thermal populations.

Distribution functions, unlike moments, are not easily classified by a small number of parameters. We therefore propose to apply dimensionality reduction and clustering methods to particle distributions in pitch angle and energy space as a new method to distinguish between the different plasma regions. 2D distributions functions in pitch angle and energy are derived from full 3D distributions in velocity space based on the magnetic field direction and the assumption of gyrotropy of electrons. With these novel methods, we robustly classify variations in particle populations to a high temporal and spatial resolution, allowing us to better identify the physical processes governing particle populations in near-Earth space. Our method also has the advantage of being independent of the model applied, as these methods do not require prior assumptions of the distributions of each population.

### 1.1 Machine Learning Models

In this section, we give a detailed account of the internal operations of each of the unsupervised machine learning algorithms used in our method. In unsupervised learning, algorithms discover the internal representations of the input data without requiring training on example output data. Dimensionality reduction is a specific type of unsupervised learning in which data in high-dimensional space is transformed to a meaningful representation in lower dimensional space. This transformation allows complex datasets, such as 2D pitch angle and energy distributions, to be characterized by analysis techniques (e.g., clustering algorithms) with much more computational efficiency. Our machine learning method utilizes four separate algorithms: autoencoders (Hinton and Salakhutdinov, 2006), principal component analysis (PCA, Abdi and Williams, 2010), mean shift (Fukunaga and Hostetler, 1975), and agglomerative clustering (AC) (Lukasová, 1979). We obtain the autoencoder algorithm from the Keras library (Chollet et al., 2015), and the PCA, mean shift, and AC algorithms from the scikit-learn library (Pedregosa et al., 2011).



**FIGURE 1** | The architecture of an autoencoder, adapted from Sakurada and Yairi (2014). Each circle represents a neuron corresponding to a data-point. Layer L<sub>1</sub> represents the input data, layer L<sub>2</sub> the encoded data in latent space, and layer L<sub>3</sub> the reconstructed data. The circles labeled “+1” are known as “bias units,” which are parameters that are adjusted during training to improve the performance of the neural network.

We use the autoencoder to compress the data by a factor of 10 from a high-dimensional representation. We subsequently apply the PCA algorithm to further compress the data to a three-dimensional representation. The PCA algorithm has the advantage of being a lot cheaper computationally than an autoencoder, however the algorithm only captures variations that emerge from linear relationships in the data, while autoencoders also account for non-linear relationships in the dimensionality reduction process (Bishop, 1998). For this reason, we only utilize the PCA algorithm after the data have been compressed *via* an autoencoder. After compressing the data, we use the mean shift algorithm to inform us of how many populations are present in the data using this three-dimensional representation. While the mean shift algorithm provides us with this estimate of the requisite number of clusters, the algorithm is ineffective in constraining the shapes of the clusters to determine which population each data-point belongs to. Therefore, we use an AC algorithm to assign each data-point to one of the populations.

### 1.1.1 Autoencoders

Autoencoders are a particular class of unsupervised neural networks. They are trained to learn compressed representations of data by using a bottleneck layer which maps the input data to a lower dimensional space, and then subsequently reconstructing the original input. By minimizing the “reconstruction error,” or “loss,” the autoencoder is able to retain the most important information in a representative compression and reconstruction of the data. As a result, autoencoders have applications in dimensionality reduction (e.g., Hinton and Salakhutdinov, 2006), anomaly detection (e.g., Kube et al., 2019) and noise filtering (e.g., Chandra and Sharma, 2014).

During training, an autoencoder runs two functions simultaneously. The first, called an “encoder,” maps the input data,  $x$ , to the coded representation in latent space,  $z$ . The second function, called a “decoder,” maps the compressed data,  $z$ , to a reconstruction of the input data,  $\hat{x}$ . The encoder,  $E(x)$ , and decoder,  $D(z)$ , are defined by the following deterministic posteriors:

$$\begin{aligned} E(x) &= p(z|x; \theta_E), \\ D(z) &= p(\hat{x}|z; \theta_D), \end{aligned} \quad (1)$$

where  $\theta_E$  and  $\theta_D$  are the trainable parameters of the encoder and decoder respectively. **Figure 1** illustrates the standard architecture of an autoencoder.

In feed-forward neural networks, such as autoencoders, each neuron computes the following sum:

$$y = \sum_i w_i x_i + b, \quad (2)$$

where  $x_i$  represents the input from the previous layer,  $w_i$  denotes the weights associated with the connections between neurons in different layers, and  $b$  denotes the bias term associated with each layer (represented by the circles labeled “+1” in **Figure 1**). The number of neurons in each layer defines the dimension of the data representation in that layer. The output of each neuron,  $f(y)$ , is called the activation function. Rectified Linear Unit, (Hahnioser et al., 2000) is the most commonly used activation function due to its low computational cost (Agarap, 2018). The function is described as:

$$f(y) = \max(0, y). \quad (3)$$

The sigmoid activation function (Chandra and Singh, 2004) is also commonly used. It is defined by:

$$f(y) = \frac{1}{1 + e^{-y}}, \quad (4)$$

where  $y$  is defined in **Eq. 2**. Analysis of the use of various activation functions in the remit of plasma physics are given by Kube et al. (2019).

In order to improve the representation of the compressed data in layer L2 and minimize the discrepancy between the input and reconstruction layer, the autoencoder adjusts the weights and biases by minimising a loss function through an optimiser (described below). The binary cross-entropy loss function

(de Boer et al., 2005) is typically used when the input data,  $x$ , are normalized to values between 0 and 1. The loss value,  $c$ , increases as the reconstruction data,  $\hat{x}$ , diverge from the input data. The loss function is defined as:

$$c = -[x \ln(\hat{x}) + (1 - x) \ln(1 - \hat{x})]. \quad (5)$$

An overview of various loss functions is provided by Janocha and Czarnecki (2017). Optimisers are used to ensure the autoencoder converges quickly to a minimum loss value by finding the optimum value of the weight,  $w_i$ , of each neuron. This is achieved by running multiple iterations with different weight values, known as gradient descent (Ruder, 2016). The weights are adjusted in each iteration,  $t$ , according to:

$$w_t = w_{t-1} - \alpha \frac{\partial c}{\partial w}, \quad (6)$$

where  $\partial c / \partial w$  is the gradient, which is a partial derivative of the loss value with respect to the weight. The learning rate,  $\alpha$ , updates all the weights simultaneously with respect to the gradient descent. This learning rate is randomly initialized between 0 and 1 by the Algorithm. A low learning rate results in a slower convergence to the global minimum loss value. However a too high value for the learning rate impedes the gradient descent (**Eq. 6**) from converging on the optimum weights. The Adadelta optimiser (Zeiler, 2012) is commonly used due to its rapid convergence to the minimum loss value and its ability to adapt the learning rate depending on each parameter. The optimiser updates each parameter,  $\theta$ , according to:

$$\Delta \theta_t = -\frac{\text{RMS}[\Delta \theta]_{t-1}}{\text{RMS}[g]_t} g_t, \quad (7)$$

where  $\Delta \theta_t$  is the parameter update at the  $t$ -th iteration,  $g_t$  is the gradient of the parameters at the  $t$ -th iteration, and RMS is the root mean square. An overview of the various optimisers is provided by Khandelwal (2019).

### 1.1.2 Principal Component Analysis

PCA is a statistical procedure that, as well as autoencoders, also reduces the dimensionality of input data. The algorithm achieves this by transforming the input data from a large number of correlated variables to a smaller number of uncorrelated variables, known as principal components. These principal components account for most of the variation in the original input data, making them a useful tool in feature extraction.

Before the procedure, the original data,  $X_0$ , are represented by a  $(n \times Q)$  matrix, where  $n$  is the number of observations and  $Q$  is the number of variables (also called dimensions). In the first step, the algorithm scales and centers the data:

$$X = (X_0 - \bar{X}_0) D^{-1}, \quad (8)$$

where  $\bar{X}_0$  contains the means of each of the variables, and  $D$  is a diagonal matrix that contains the scaling coefficient of each variable. Typically,  $D_{ii} = \sigma_i$  where  $\sigma_i$  is the standard deviation of variable with index  $i$  (Peerenboom et al., 2015). The algorithm then uses  $X$  to calculate the covariance matrix:

$$C_X = \frac{1}{n-1} X^T X, \quad (9)$$

which measures the correlation between the different variables. The principal components are calculated as the eigenvectors,  $A$ , of the covariance matrix:

$$C_X = ALA^T, \quad (10)$$

where  $L$  is a diagonal matrix containing the eigenvalues associated with  $A$ . These principal components are ordered in decreasing order, whereby the first principal components account for most of the variation in the input data. These input data are finally projected into the principal component space according to:

$$Z = XA, \quad (11)$$

where  $Z$  represents the output data containing the principal component scores. The dimensionality of these output data are determined by the number of principal components used.

### 1.1.3 Mean Shift

The mean shift algorithm is a non-parametric clustering technique that is used for locating the maxima of a density function in a sample space. The algorithm aims to discover the number of clusters within a dataset, meaning no prior knowledge of the number of clusters is necessary.

For a dataset containing  $n$  data-points  $x_i$ , the algorithm starts finding each maximum of the dataset's density function by randomly choosing a data-point to be the mean of the distribution,  $x$ . The algorithm then uses a kernel function,  $K$ , to determine the weights of the nearby data-points for re-estimating the mean. The variable  $h$  is the width of the kernel window. Typically, a Gaussian kernel,  $k$ , is used:

$$K\left(\frac{x-x_i}{h}\right) = c_k k\left(\left\|\frac{x-x_i}{h}\right\|^2\right) = \exp\left(-c_k \left\|\frac{x-x_i}{h}\right\|^2\right), \quad (12)$$

where  $c_k$  is the normalising constant. With the kernel function, the multivariate kernel density estimator is obtained:

$$f(x) = \frac{1}{nh^d} \sum_{i=1}^n K\left(\frac{x-x_i}{h}\right), \quad (13)$$

where  $d$  is the dimensionality of the dataset. The gradient of the density estimator is then:

$$\begin{aligned} \nabla f(x) &= \frac{2c_k}{nh^{d+2}} \sum_{i=1}^n (x_i - x) g\left(\left\|\frac{x-x_i}{h}\right\|^2\right) \\ &= \frac{2c_k}{nh^{d+2}} \left[ \sum_{i=1}^n g\left(\left\|\frac{x-x_i}{h}\right\|^2\right) \right] m_h(x), \end{aligned} \quad (14)$$

where  $g(x) = -k'(x)$ . The first term is proportional to the density estimate at  $x$ , and the second term,  $m_h(x)$ , is:

$$m_h(x) = \frac{\sum_{i=1}^n x_i g\left(\left\|\frac{x-x_i}{h}\right\|^2\right)}{\sum_{i=1}^n g\left(\left\|\frac{x-x_i}{h}\right\|^2\right)} - x, \quad (15)$$

which is the mean shift vector and points toward the direction of the maximum increase in density. The mean shift algorithm

therefore iterates between calculating the mean shift vector,  $m_h(x^t)$ , and translating the kernel window:

$$x^{t+1} = x^t + m_h(x^t), \quad (16)$$

where  $t$  is the iteration step. Once the window has converged to a point in feature space where the density function gradient is zero, the algorithm carries out the same procedure with a new window until all data-points have been assigned to a maximum in the density function.

### 1.1.4 Agglomerative Clustering

AC is a type of hierarchical clustering that uses a "bottom-up" approach, whereby each data-point is first assigned a different cluster. Then pairs of similar clusters are merged until the specified number of clusters has been reached. During each recursive step, the AC algorithm combines clusters typically using Ward's criterion (Ward, 1963), which finds pairs of clusters that lead to the smallest increase in the total intra-cluster variance after merging. The increase is measured by a squared Euclidean distance metric:

$$d_{ij} = d(C_i, C_j) = \|C_i - C_j\|^2, \quad (17)$$

where  $C_i$  represents a cluster with index  $i$ . The algorithm implements Ward's criterion using the Lance-Williams formula (Lance and Williams, 1967):

$$\begin{aligned} d(C_i \cup C_j, C_k) &= \frac{n_i + n_k}{n_i + n_j + n_k} d(C_i, C_k) \\ &\quad + \frac{n_j + n_k}{n_i + n_j + n_k} d(C_j, C_k) - \frac{n_k}{n_i + n_j + n_k} d(C_i, C_j), \end{aligned} \quad (18)$$

where  $C_i$ ,  $C_j$ , and  $C_k$  are disjoint clusters with sizes  $n_i$ ,  $n_j$ , and  $n_k$ , and  $d(C_i \cup C_j, C_k)$  is the squared Euclidean distance between the new cluster  $C_i \cup C_j$  and  $C_k$ . The clustering algorithm uses Eq. 18 to find the optimal pair of clusters to merge.

## 1.2 The Magnetotail

We use electron data from the magnetotail in order to test the effectiveness of our method. The magnetotail is traditionally divided into three different regions: the plasma sheet, the plasma sheet boundary layer, and the lobes (Hughes, 1995). These regions are defined by their plasma and magnetic field characteristics. The low temperature ( $\sim 85$  eV) outermost northern and southern lobes are on open magnetic field lines which results in a much lower plasma density of  $\sim 0.01$  cm (Lui, 1987). The plasma sheet boundary layer exists on the reconnected magnetic field lines. This region forms the transition region in between the plasma sheet and the lobes, and is characterized by a population of field-aligned particles and a plasma  $\beta$ , which is the ratio of the plasma pressure to the magnetic pressure, of  $\sim 0.1$  (Lui, 1987).

The innermost plasma sheet typically contains a comparatively hot ( $\sim 4,250$  eV) and isotropic plasma with a relatively high particle density of  $\sim 0.01$  cm $^{-3}$ . At the center of the plasma sheet is the thin neutral current sheet, which is



characterized by a relatively high plasma  $\beta$  of  $\sim 0$ , and a magnetic field strength of near zero (Lui, 1987). Although isotropic electron pitch angle distributions (PADs) are the most dominant in the plasma sheet, many cases of pitch angle anisotropy have also been found (e.g., Walsh et al., 2013; Artemyev et al., 2014; Liu et al., 2020). These intervals correspond to a colder and denser electron population and are linked to: cold anisotropic ionospheric outflows (Walsh et al., 2013), and a penetration of cold electrons from the magnetosheath near the flanks (Artemyev et al., 2014).

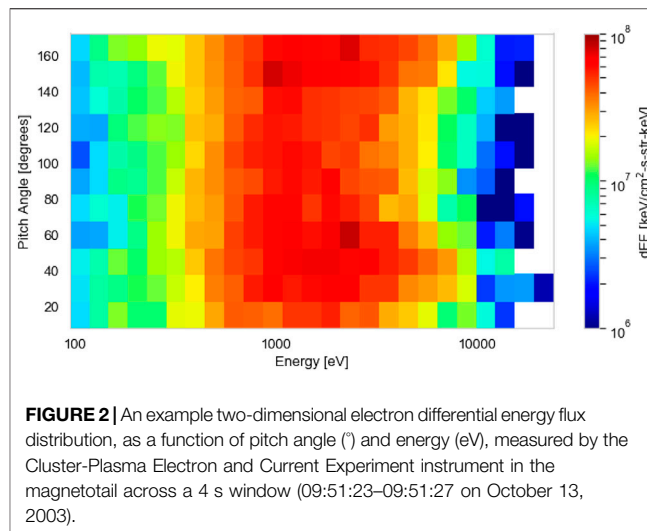
## 2 METHOD AND APPLICATION

In this section, we detail the steps required to classify different regions within a space plasma environment using machine learning techniques. As an example, we classify Cluster-Plasma Electron and Current Experiment (PEACE, Johnstone et al., 1997; Fazakerley et al., 2010) data (Laakso et al., 2010) from the Earth's magnetotail to showcase our method, as this allows us to compare to the Cluster-ECLAT (European Cluster Assimilation Technology) (Boakes et al., 2014) database for evaluation. The same method, however, can be applied to any plasma regime where energy and pitch angle measurements are available. Our steps are as follows:

- (1) Data preparation: We obtain the Cluster-PEACE data from different magnetotail regions based on the Cluster-ECLAT database, and prepare the data for testing.
- (2) Reducing dimensionality: We build our autoencoder and use the encoder part to reduce the dimensionality of each pitch angle and energy distribution by a factor of 10. We use a PCA algorithm to further compress each distribution to a set of coordinates in 3D space.
- (3) Clustering: We apply the mean-shift algorithm to determine how many clusters exist within the compressed magnetotail electron data, and use an AC algorithm to separate the compressed dataset into this number of clusters. This allows us to determine how many plasma regimes exist within the overall dataset.
- (4) Evaluation: We estimate the probabilities of the AC labels and compare our clustering results to the original ECLAT labels in order to evaluate our method.

### 2.1 Data Preparation

We prepare PEACE instrument data from the Cluster mission's C4 spacecraft (Escoubet et al., 2001) to test and present our method. The Cluster mission comprises of four spacecraft, each spinning at a rate of  $4 \text{ s}^{-1}$ . The PEACE data have a 4 s time resolution and are constructed from two instantaneous PAD measurements per spin. Each of our distributions is a two-dimensional differential energy flux product containing twelve  $15^\circ$  wide pitch angle bins and 26 energy bins, spaced logarithmically between 93 eV and 24 keV. The dimensionality of each distribution is 312 ( $12 \times 26$ ). We normalize the differential energy flux linearly between 0 and 1 based on the maximum flux value in the dataset. An example of an individual differential



energy flux distribution used in our analysis is shown in **Figure 2**. We correct for spacecraft potential with measurements from the Cluster-EFW (Electric Field and Wave Experiment) instrument (Gustafsson et al., 2001) and corrections (19% increase) according to the results of Cully et al. (2007).

The ECLAT dataset consists of a detailed list of plasma regions encountered by each of the four Cluster spacecraft in the nightside magnetosphere. The dataset is available from July to October during the years 2001–2009. Using plasma and magnetic field moments from the PEACE, Cluster Ion Spectrometry (CIS) (Rème et al., 2001), and Fluxgate Magnetometer (FGM) (Balogh et al., 1997) instruments, the dataset provides a list of (inner and outer) plasma sheet, boundary layer, and lobe times. These regions are identified based on the plasma  $\beta$ , the magnetic field measurements, and the current density vectors. A comprehensive account of the ECLAT identification routine for each plasma region is provided by Boakes et al. (2014). To ensure that we test our method on a large number of data from each of the magnetotail regions ( $>50,000$  samples), we obtain PEACE data from times when the C4 spacecraft has spent at least 1 h in each region, according to ECLAT.

### 2.2 Reducing Dimensionality

After preparing the dataset to include a series of  $>50,000$  time intervals, each with its associated 2D pitch angle and energy distributions (e.g., **Figure 2**), the first step toward reducing the dataset's dimensionality is to build a suitable autoencoder (described in **Section 1.1.1**). We construct our autoencoder using the Keras library. This step requires defining the number of neurons in each layer. The input and reconstruction layer should have the same number, which is equal to the dimensionality of the original dataset (312 for each time interval in this example). The middle encoded layer typically contains a compressed representation of the data that is by a factor of 10 smaller than the input data (Hinton and Salakhutdinov, 2006). We therefore specify our encoded layer to contain 32 neurons. The next step involves specifying the activation function for the neurons in the first and middle layers.

We use the standard Rectified Linear Unit activation function (Hahnloser et al., 2000) in the encoder part of our autoencoder and the sigmoid activation function (Chandra and Singh, 2004) in the decoder part, as this function is used to normalize the output between 0 and 1.

The next step defines which loss function and optimiser the autoencoder uses in order to representatively compress and reconstruct the input data. As we use normalized output data, we choose the standard binary cross-entropy loss function (de Boer et al., 2005). In terms of the optimiser, we utilize the Adadelta optimiser (Zeiler, 2012) due to its speed and versatility. All of the activation functions, loss functions, and optimisers are available in the Keras library.

In the next step, we set the hyperparameters used for training the autoencoder. These hyperparameters include: the number of epochs, the batch size, and the validation split ratio. The number of epochs represents the number of training iterations undergone by the autoencoder, with the weights and biases updated at each iteration. The batch size defines the number of samples that are propagated through the network at each iteration. It is equal to  $2^n$ , where  $n$  is a positive integer. The batch size (256 in our case) is ideally set as close to the dimensionality of the input data as possible. The validation split ratio determines the percentage of the input data that should remain “unseen” by the autoencoder in order to verify that the algorithm is not overfitting the remaining training data. We set the validation split ratio to 1/12, which is commonly used for large datasets (Guyon, 1997). At each iteration, a training loss value and a validation loss value are produced, which are determined by the binary cross-entropy loss function. Both of these values converge to their minima after a certain number of iterations, at which point the autoencoder cannot be optimized to the input data any further. Loss values  $<0.01$  are typically considered ideal (Le et al., 2018).

After retrieving the compressed representation of the input data from the encoding layer (with a dimensionality of 32 in our case), we apply a PCA algorithm (see Section 1.1.2) to the compressed data to reduce the dimensionality to 3. We obtain the PCA algorithm from the scikit-learn library. We set the output dimensionality of the PCA algorithm to 3 as the following clustering algorithms used in this method are computationally expensive and their performance scales poorly with increasing dimensionality (Lukasova, 1979; Comaniciu and Meer, 2002). Setting the dimensionality to 3 has the added benefit that the clusters can be visualised.

## 2.3 Clustering

Once the dimensionality reduction stage has taken place and each pitch angle and energy distribution is represented by three PCA values, we use clustering algorithms to separate the dataset into the different particle populations. To first determine how many populations exist within the dataset (8 in our case), we apply a mean shift clustering algorithm (see Section 1.1.3) to the data to find the number of maxima,  $n_c$ , in the distribution of data-points. We obtain the mean shift algorithm from the scikit-learn library. We set the bandwidth, represented by  $h$  in Eq. 15 to 1, which we

find optimizes the time taken for the algorithm to converge on the maxima in the density distribution.

After determining the number of clusters in the dataset, we use an AC algorithm (see Section 1.1.4) to assign each data-point to one of the clusters. We obtain the AC algorithm from the scikit-learn library and instantiate the algorithm by specifying the number of clusters,  $n_c$ , before applying it to the compressed dataset. Assigning several clusters to a large dataset with three dimensions is a computationally expensive task, however we find the AC algorithm converges relatively quickly in comparison to other clustering algorithms. A further advantage of the hierarchical clustering procedure, used in the AC algorithm, is that data-points belonging to a single non-spherical structure in the 3-dimensional parameter space are not incorrectly separated into different clusters, unlike the more widely used K-means algorithm (Arthur, 2007).

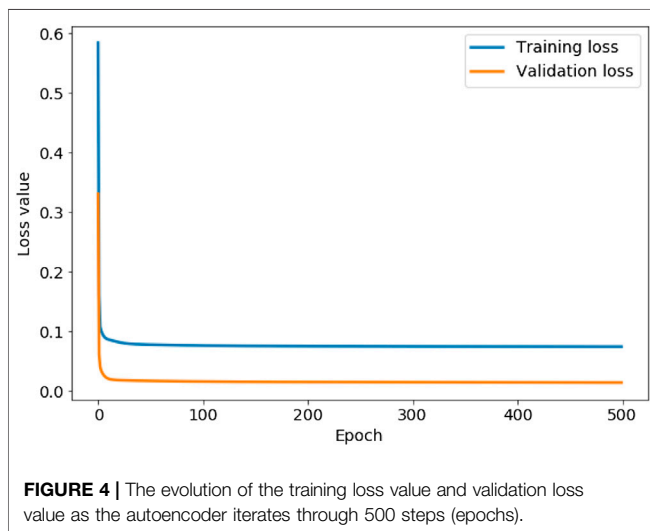
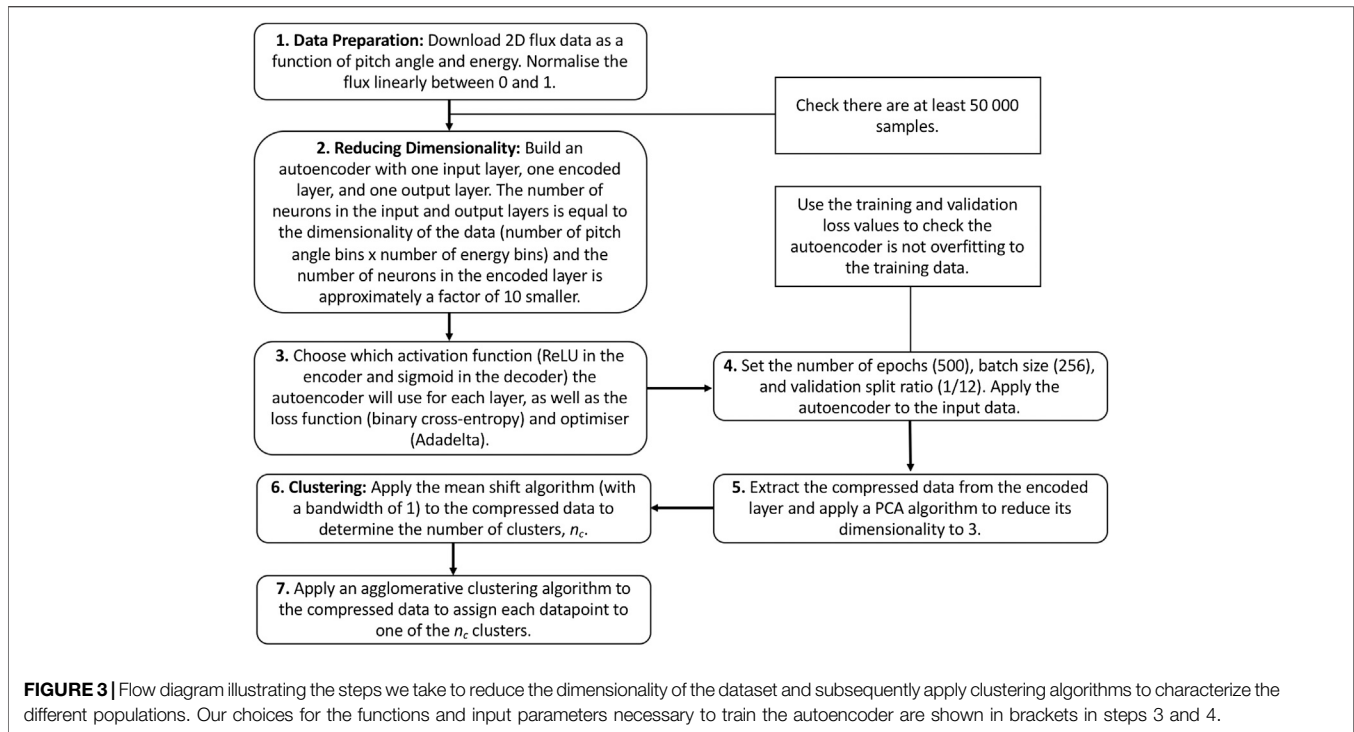
Figure 3 contains a flow diagram detailing our method.

## 2.4 Evaluation

Figure 4 shows the training and validation loss values associated with each iteration during the training of our autoencoder. We use this graph to check if the autoencoder is overfitting to the training data, which is evident if the training loss starts to decrease more rapidly than the validation loss. In this case, our autoencoder is not overfitting at any iteration during training. Figure 4 shows that both the loss values start to rapidly level off in less than 100 epochs. Both loss values, however, continue to decrease, with the training loss value converging to 0.0743 after 444 iterations, and the validation loss value converging to 0.0140 after 485 iterations. We therefore set the number of epochs to 500. As both loss values are lower than 0.01, we conclude the autoencoder is accurately reconstructing both sets of input data, assuring us that the encoded data with a lower dimensionality is representative of the original 2D distribution functions. The lower validation loss than training loss in Figure 4 indicates the presence of anomalous data in the training set that is not represented in the validation set. We discuss this anomalous data later in this section.

Figure 5 shows the result of applying the AC algorithm to the compressed magnetotail electron data after the implementation of the autoencoder and PCA algorithms. The 3-dimensional representation shows that the clustering algorithm is able to assign data-points of varying PCA values to the same cluster if they belong to the same complex non-spherical structure, e.g., clusters 0, 4, and 6. The clustering algorithm is able to form clear boundaries between clusters with adjacent PCA values, e.g., between clusters 0, 1, and 7, with no mixing of cluster labels on either side of the boundaries. The clustering algorithm locates the boundaries by finding areas with a low density of data-points in comparison to the centers of the clusters.

Figure 6 shows the results of averaging the 2D differential energy flux distributions in pitch angle and energy space for each of the eight clusters. Using moments data collected by the PEACE, FGM, and CIS instruments, we compare the proton plasma  $\beta_s$ , electron densities and temperatures, and magnetic field strengths to the average 2D distribution of each cluster. This process allows us to verify the consistency of the clustering



method and provide general region classifications in order to make comparisons with the ECLAT labels. Our classifications (shown in the captions below each sub-figure) are produced with the aid of previous analyses of electron PADs (e.g., Walsh et al., 2011; Artemyev et al., 2014) and the plasma and magnetic field parameters (e.g., Lui, 1987; Artemyev et al., 2014) in the magnetotail.

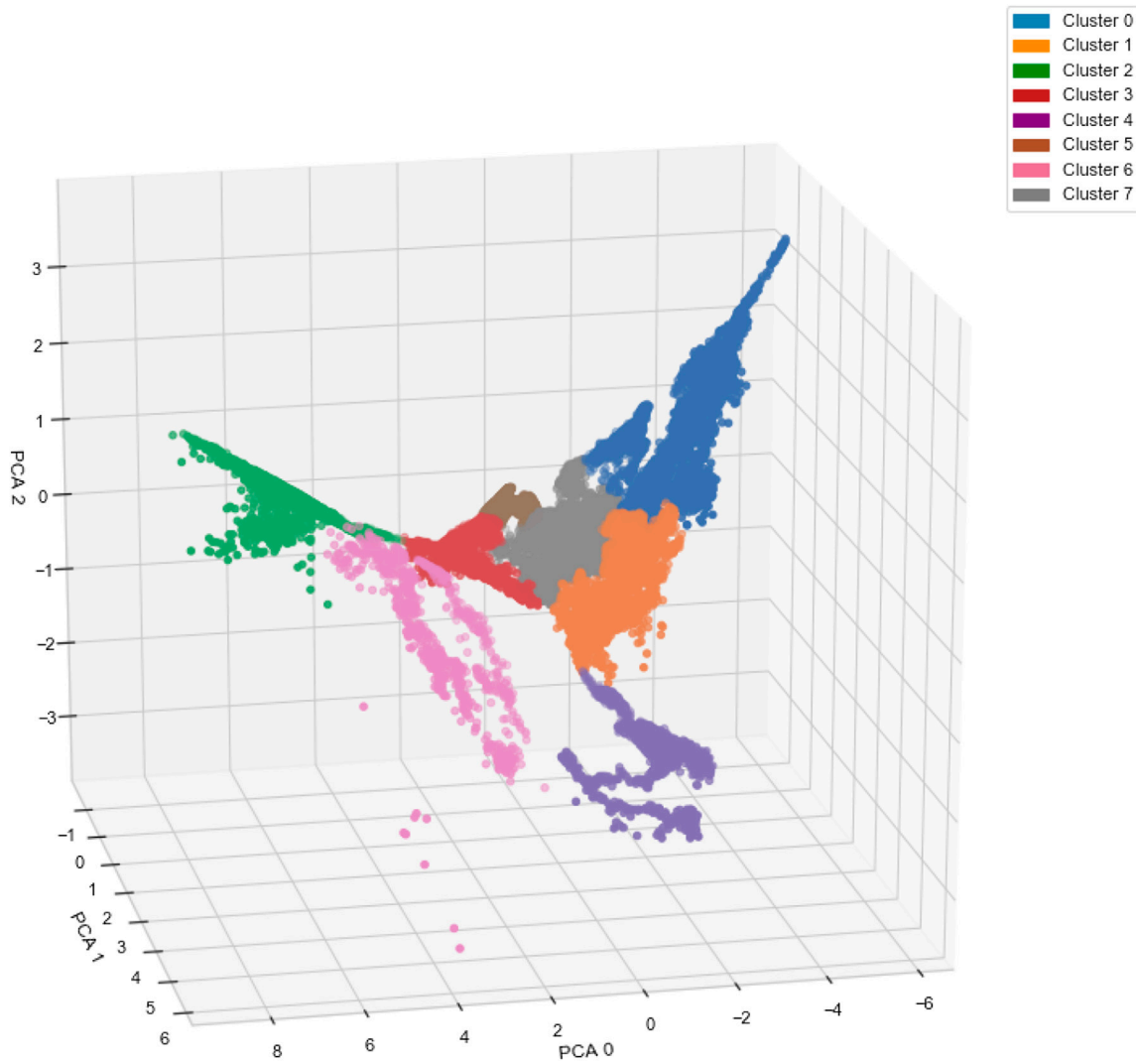
The individual sub-figures in **Figure 6** display large differences in the average electron 2D pitch angle and energy distributions. Each average distribution differs by either: the energy of the peak flux, the peak value of the flux, or the amount of pitch angle anisotropy, i.e., the difference in flux between the parallel and

perpendicular magnetic field direction. The lack of identical average distributions amongst the clusters shows the mean shift algorithm has not overestimated the number of clusters. By observing the individual 2D distributions within each cluster, we see a distinct lack of intra-cluster variance, showing the mean shift algorithm does not underestimate the number of clusters.

A limitation of using the AC algorithm is that outliers or anomalous data are not differentiated from the main clusters. Clustering a sizable number of outliers with the main populations can lead to ambiguity in the defining characteristics of each population, reducing the robustness of our method. In our case, **Figure 5** shows only nine data points, within cluster 6, that are disconnected from the main structure of cluster 6 due to their distinct PCA values. We observe similar phenomenon to a lesser extent with cluster 2. To counteract this issue, we perform an outlier detection procedure using the reconstructed output of the autoencoder. By calculating the mean square error (MSE) between each input data-point and its reconstructed output, we isolate outliers in the dataset from the AC analysis based on their large MSE values, in comparison to 99.95% of the data-points. During training, the autoencoder learns the latent space representation that defines the key characteristics of the bulk populations present in the dataset. The most relevant features of an anomalous particle distribution are not present in this subspace, resulting in a large MSE between the reconstructed data, which lacks these important features, and the original data. This technique effectively identifies the nine obvious outliers observable by eye in **Figure 5**, along with six from cluster 2 and 5 from cluster 1.

We use Gaussian mixture models (GMMs, McLachlan and Peel, 2000) to establish the probabilities of each of the data-points





**FIGURE 5 |** Three-dimensional representation of the magnetotail data after undergoing dimensionality reduction via an autoencoder and principal component analysis (PCA) algorithm. The colors represent the clustering results from the agglomerative clustering algorithm.

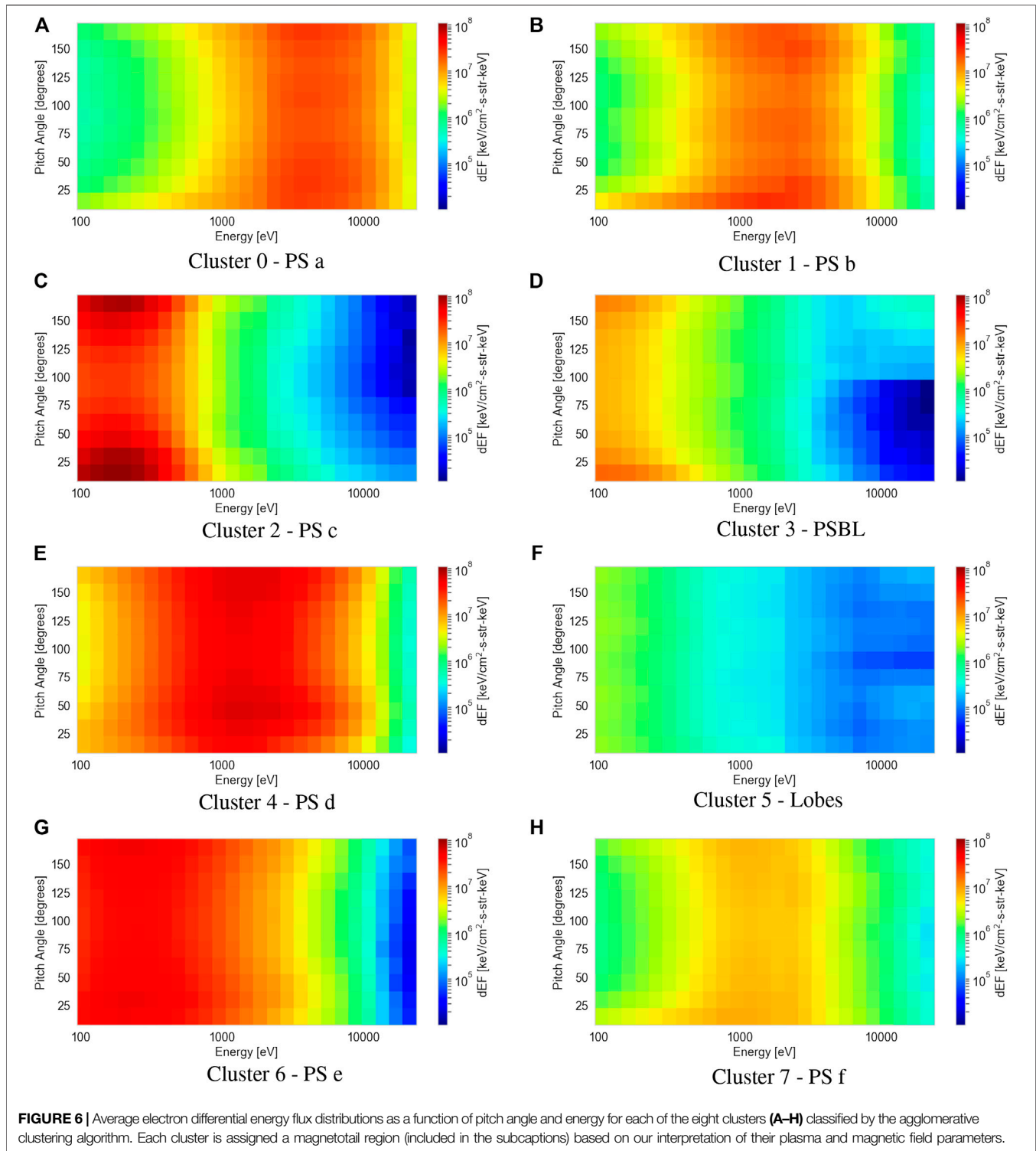
belonging to the clusters they have been assigned to by the AC algorithm, providing useful information on the uncertainty associated with our region classification method. We obtain the GMM from the scikit-learn library. For each data-point,  $x_i$ , a GMM fits a normal distribution,  $\mathcal{N}$ , to each cluster and computes the sum of probabilities as:

$$p(x_i) = \sum_{j=1}^k \phi_j \mathcal{N}(x_i; \mu_j, \tau_j) = 1, \quad (19)$$

where  $\mu_j$  and  $\tau_j$  are the mean and covariance of the normal distribution belonging to cluster  $j$ , and  $\phi_j$  is the mixing coefficient which represents the weight of Gaussian  $j$  and is calculated by the Expectation–Maximisation algorithm (Dempster et al., 1977). A complete description of GMMs and the Expectation–Maximisation algorithm is provided by Dupuis et al. (2020).

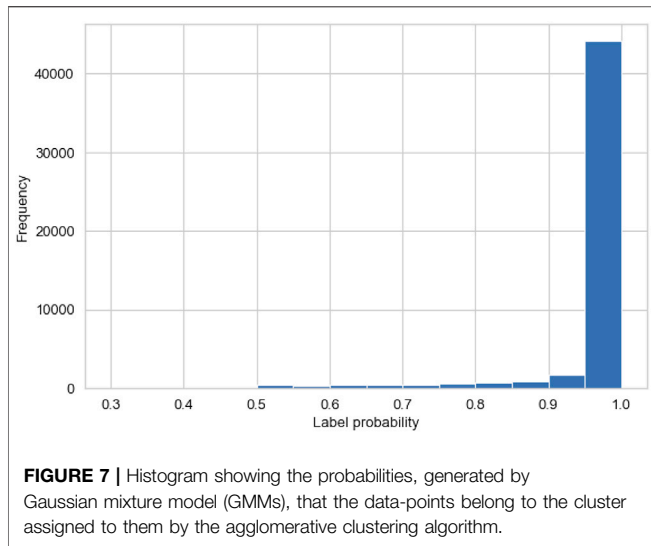
**Figure 7** shows a histogram of the probabilities, calculated by the GMM, associated with each data-point belonging to the cluster it is assigned to by the AC algorithm. More than 92% of the data-points have a probability of over 0.9, and <1% of the data-points have a probability of <0.5. This indicates a high certainty in our clustering method and validates the high precision in our region classifications. Further investigations of the data-points with associated probabilities of <0.5 show that these data-points exist on the boundary between clusters 0 and 1, i.e., two plasma sheet populations that differ by temperature. This illustrates a small limitation in the AC method when distinguishing between relatively similar plasma regimes.

**Table 1** shows the median and upper and lower quartiles of the electron density, electron temperature, magnetic field, and ion plasma  $\beta$  for each of the eight clusters designated by our AC algorithm. None of the eight clusters have comparable median and quartile values across all four of the chosen parameters.



Certain pairs of clusters exhibit similarities in the median and quartile values for one or two of the four parameters, e.g., clusters 0 and 4 exhibit similar electron densities and magnetic field strengths, and clusters 3 and 6 exhibit similar magnetic field strengths. However there are large differences in the values of the remaining parameters for these pairs of clusters. These results

show that clear differences in the 2D pitch angle and energy distributions (see **Figure 6**) can translate into distinctions between certain but not all plasma parameter measurements, providing a strong indicator that full 2D distributions can effectively be used to distinguish between similar particle populations. Regarding the ECLAT classifications, which are



based on magnetic field and plasma  $\beta$  measurements, certain pairs of clusters exhibit a similar range of values in both of these measurements, e.g., clusters 0 and 4 and clusters 1 and 7. As the majority of data-points for all of these clusters are considered the same plasma sheet population by ECLAT (see **Table 2**), we conclude that using a limited number of parameters to provide classifications overlooks distinctions between different populations and incorrectly groups them into the same category.

**Table 2** shows our comparison between the eight AC labels and the region names given in the ECLAT database, for the magnetotail data used in our example.

In **Table 2**, there is some disagreement with three of our clusters, namely AC labels 3, 5, and 7, which correspond to the plasma sheet boundary layer, the lobes, and a plasma sheet population respectively. However for each of these clusters, the majority of labels are in agreement with the ECLAT regions (72.4, 86.8, and 86.9% for AC clusters 3, 5, and 7 respectively). For AC labels 0, 1, 2, 4, and 6, which represent various other populations within the plasma sheet, there is 100% agreement with the ECLAT label 0, which denotes the plasma sheet. By using this

method to characterize full electron pitch angle and energy distributions, instead of using the derived moments, we are successfully able to distinguish between multiple populations within what has historically been considered as one region, due to the lack of variation in the plasma moments (see **Table 1**) as well as the similarity in spatial location. Using 2D pitch angle and energy distributions also improves the time resolution of the plasma region classifications, due to a higher cadence in the spacecraft flux and counts data (e.g., 4 s resolution for the PEACE instrument) in comparison to the moments data (e.g., 8 s resolution for CIS moments and 16 s resolution for PEACE moments).

### 3 CONCLUSION

We present a novel machine learning method that characterizes full particle distributions in order to classify different space plasma regimes. Our method uses autoencoders and subsequently PCA to reduce the dimensionality of the 2D particle distributions to three dimensions. We then apply the mean shift algorithm to discover the number of populations in the dataset, followed by the AC algorithm to assign each data-point to a population.

To illustrate the effectiveness of our method, we apply it to magnetotail electron data and compare our results to previous classifications, i.e. the ECLAT database, that utilizes moments. With our method, we find multiple distinct electron populations within the plasma sheet, which previous studies have identified as one region (**Table 2**). These findings show that key features in particle distributions are not fully characterized by the plasma moments (e.g., **Table 1**), resulting in important distinctions between populations being overlooked. For example, we find two separate cold dense anisotropic populations in the plasma sheet (clusters 2 and 6), which are less abundant than the hotter and more isotropic plasma sheet populations. By using our clustering method to specify an exact list of times when populations like these are observed, we create a more comprehensive picture of their spatial distribution. Inherent time-dependencies may also contribute to our finding of

**TABLE 1 |** Comparisons of the median, Q2, and upper, Q3, and lower, Q1, quartile values of the electron density  $n_e$ , electron temperature  $T_e$ , magnetic field  $|B|$ , and plasma  $\beta$  associated with each of the eight clusters.

AC labels	$n_e$ ( $\text{cm}^{-3}$ )			$T_e$ (eV)			$ B $ (nT)			Plasma $\beta$		
	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3
0	0.21	0.22	0.23	2,057.26	2,515.28	2,913.33	32.45	35.09	37.86	3.54	4.33	5.92
1	0.12	0.14	0.19	1,487.67	1,838.97	2,258.43	10.56	13.65	19.42	15.31	24.54	39.75
2	1.08	1.18	1.30	106.44	114.19	123.33	20.87	22.67	24.25	0.80	0.99	1.17
3	0.21	0.25	0.28	79.21	83.63	93.00	16.64	35.07	39.76	0.28	0.39	0.60
4	0.22	0.28	0.82	783.33	879.77	997.85	24.58	39.10	41.93	1.23	1.52	6.29
5	0.01	0.02	0.03	116.63	170.57	252.85	32.97	41.34	49.00	0.00	0.06	0.27
6	1.29	1.49	1.65	164.41	214.64	269.16	41.65	44.74	46.56	0.30	0.43	0.60
7	0.08	0.10	0.13	669.64	882.30	1,217.77	5.42	17.16	25.95	3.69	9.05	128.11

AC, agglomerative clustering. The AC labels 0, 1, 2, 4, 6, and 7 belong to the plasma sheet, according to ECLAT, three belongs to the plasma sheet boundary layer, and five belongs to the lobes.

**TABLE 2 |** Contingency table comparing the agglomerative clustering (AC) labels of the magnetotail electron data to the original ECLAT labels (0 = PS, 1 = PSBL, and 2 = lobes).

AC labels	ECLAT labels		
	0	1	2
0	6,549	0	0
1	3,074	0	0
2	5,092	0	0
3	1,590	4,188	0
4	2,097	0	0
5	156	2,228	15,641
6	1,029	0	0
7	7,020	1,057	0

The AC labels are the same as in **Table 1**.

multiple plasma sheet populations. Even in this case, our method is effective in characterising the evolution of particle populations, made possible by the high time resolution of our region classifications. In a follow up study, we will use this information to link the occurrence of these populations to other high-resolution spacecraft measurements in different plasma regions, in order to understand the physical processes driving changes in the less abundant particle populations. As an example analysis, our high resolution classifications of the observed anisotropic plasma sheet populations could be combined with previous theories on the sources of these populations (e.g., Walsh et al., 2013; Artemyev et al., 2014), to understand the relative contributions of particle outflows from distinct magnetospheric regions, such as the magnetosheath or ionosphere.

Comparisons between this original method and the previous classifications from ECLAT also show specific periods of disagreement (e.g., we classify a small number of ECLAT periods of plasma sheet as the plasma sheet boundary layer). This discrepancy shows that using the full 2D pitch angle and energy distributions, without requiring prior assumptions about magnetospheric plasma regions, may redefine the classifications of electron populations, along with our understanding of their plasma properties. Our method, which uses open-source and easily accessible machine learning techniques, can be used to better characterize any space plasma regime with sufficient *in-situ* observations. By not being constrained to a small number of parameters, this method allows for a more complete understanding of the interactions between various thermal and non-thermal populations. With increasingly large datasets being collected by multi-spacecraft missions, such as Cluster (Escoubert et al., 2001) ( $>10^9$  full distributions in 20 years) and Magnetospheric Multiscale Mission (Sharma and Curtis, 2005), similar methods would provide a useful tool to reduce the dimensionality of distributions, thereby optimising data retrieval on Earth.

## REFERENCES

- Abdi, H., and Williams, L. J. (2010). Principal component analysis. *WIREs Comput. Stat.* 2, 433–459. doi:10.1002/wics.101
- Agarap, A. F. (2018). Deep learning using rectified linear units (ReLU). arXiv: 1803.08375.

Furthermore, combining this method with large-scale survey data, such as NASA/GSFC's OMNI database, would allow users to isolate a specific population or plasma region for analysis of its properties.

## DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: The datasets analysed in this study can be found in the Cluster Science Archive (<https://csa.esac.esa.int/csa-web/>).

## AUTHOR CONTRIBUTIONS

MRB developed the method described in the manuscript, tested it on the magnetotail data and wrote the manuscript. IJR was the lead supervisor who guided the direction of the project and provided insight at every stage. AW provided expertise on the magnetotail and the various populations that we observed, aiding the evaluation of our method. DV was also important in classifying the plasma regimes and provided insights into the physical processes governing electrons in space plasmas. AWS was key to the development of the method due to his expertise in machine learning. All co-authors made important contributions to the manuscript.

## FUNDING

MB is supported by a UCL Impact Studentship, joint funded by the ESA NPI program. IR the STFC Consolidated Grant ST/S000240/1 and the NERC grants NE/P017150/1, NE/P017185/1, NE/V002554/1, and NE/V002724/1. DV is supported by the STFC Consolidated Grant ST/S000240/1 and the STFC Ernest Rutherford Fellowship ST/P003826/1. AS is supported by the STFC Consolidated Grant ST/S000240/1 and by NERC Grants NE/P017150/1 and NE/V002724/1.

## ACKNOWLEDGMENTS

We thank the Cluster instrument teams (PEACE, FGM, CIS, EFW) for the data used in this study, in particular the PEACE operations team at the Mullard Space Science Laboratory. We also acknowledge the European Union Framework 7 Program, the ECLAT Project FP7 Grant No. 263325, and the ESA Cluster Science Archive.

- Artemyev, A. V., Walsh, A. P., Petrukovich, A. A., Baumjohann, W., Nakamura, R., and Fazakerley, A. N. (2014). Electron pitch angle/energy distribution in the magnetotail. *J. Geophys. Res. Space Phys.* 119, 7214–7227. doi:10.1002/2014JA020350
- Arthur, D. (2007). “k-means++: the advantages of careful seeding,” in Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms, New Orleans, United States, January 7, 2007. 1027–1035

- Balogh, A., Dunlop, M. W., Cowley, S. W. H., Southwood, D. J., Thomlinson, J. G., Glassmeier, K. H., et al. (1997). The cluster magnetic field investigation. *Space Sci. Rev.* 79, 65–91. doi:10.1023/A:1004970907748
- Bishop, C. M. (1998). “Bayesian PCA,” in Proceedings of the 11th international conference on neural information processing systems, Denver, United States, December 1, 1997. (Cambridge, MA: MIT Press), 382–388.
- Boakes, P. D., Nakamura, R., Volwerk, M., and Milan, S. E. (2014). ECLAT cluster spacecraft magnetotail plasma region identifications (2001–2009). *Dataset Papers Sci.* 2014, 684305. doi:10.1155/2014/684305
- Chandra, B., and Sharma, R. K. (2014). “Adaptive noise schedule for denoising autoencoder,” in *Neural information processing*. Editors C. K. Loo, K. S. Yap, K. W. Wong, A. Teoh, and K. Huang (Berlin, Germany: Springer International Publishing), 535–542.
- Chandra, P., and Singh, Y. (2004). An activation function adapting training algorithm for sigmoidal feedforward networks. *Neurocomputing* 61, 429–437. doi:10.1016/j.neucom.2004.04.001
- Chollet, F., Allison, K., Wicke, M., Bileschi, S., Bailey, B., Gibson, A., et al. (2015). Keras. Available at: <https://keras.io> (Accessed February 15, 2020).
- Comaniciu, D., and Meer, P. (2002). Mean shift: a robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Machine Intell.* 24, 603–619. doi:10.1109/34.1000236
- Cully, C. M., Ergun, R. E., and Eriksson, A. I. (2007). Electrostatic structure around spacecraft in tenuous plasmas. *J. Geophys. Res. Atmos.* 112, A09211. doi:10.1029/2007JA012269
- de Boer, P.-T., Kroese, D. P., Mannor, S., and Rubinstein, R. Y. (2005). A tutorial on the cross-entropy method. *Ann. Oper. Res.* 134, 19–67. doi:10.1007/s10479-005-5724-z
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B Methodol.* 39, 1–22. doi:10.1111/j.2517-6161.1977.tb01600.x
- Dupuis, R., Goldman, M. V., Newman, D. L., Amaya, J., and Lapenta, G. (2020). Characterizing magnetic reconnection regions using Gaussian mixture models on particle velocity distributions. *Astrophys. J.* 889, 15. doi:10.3847/1538-4357/ab5524
- Escoubet, C. P., Fehringer, M., and Goldstein, M. (2001). Introduction: the cluster mission. *Ann. Geophys.* 19, 1197–1200. doi:10.5194/angeo-19-1197-2001
- Fazakerley, A. N., Lahiff, A. D., Wilson, R. J., Rozum, I., Anekallu, C., West, M., et al. (2010). “Peace data in the cluster active archive,” in *The cluster active archive*. Editors H. Laakso, M. Taylor, and C. P. Escoubet (Dordrecht, Netherlands: Springer Netherlands), 129–144.
- Fukunaga, K., and Hostetler, L. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans. Inf. Theory.* 21, 32–40. doi:10.1109/tit.1975.1055330
- Gustafsson, G., André, M., Carozzi, T., Eriksson, A. I., Fälthammar, C.-G., Gard, R., et al. (2001). First results of electric field and density observations by Cluster EFW based on initial months of operation. *Ann. Geophys.* 19, 1219–1240. doi:10.5194/angeo-19-1219-2001
- Guyon, I. (1997). *A scaling law for the validation-set training-set size ratio*. Murray Hill, NJ: AT & T Bell Laboratories.
- Hahnloser, R., Sarpeshkar, R., Mahowald, M., Douglas, R., and Seung, H. (2000). Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature* 405, 947–951. doi:10.1038/35016072
- Hapgood, M., Perry, C., Davies, J., and Denton, M. (2011). The role of suprathermal particle measurements in crossscale studies of collisionless plasma processes. *Planet. Space Sci.* 59, 618–629. doi:10.1016/j.pss.2010.06.002
- Hinton, G. E., and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science* 313, 504–507. doi:10.1126/science.1127647
- Hughes, W. J. (1995). “The magnetopause, magnetotail, and magnetic reconnection,” in *Introduction to space physics*. Editors M. G. Kivelson and C. T. Russell (Cambridge, UK: Cambridge University Press), Chap. 9, 227–288.
- Janocha, K., and Czarnecki, W. M. (2017). On loss functions for deep neural networks in classification. arXiv.
- Johnstone, A. D., Alsop, C., Burge, S., Carter, P. J., Coates, A. J., Coker, A. J., et al. (1997). *Peace: a plasma electron and current experiment*. Dordrecht, Netherlands: Springer Netherlands, 351–398
- Khandelwal, R. (2019). Overview of different optimizers for neural networks. Available at: <https://medium.com/datadriveninvestor/overview-of-different-optimizers-for-neural-networks-e0ed119440c3>. (Accessed July 7, 2020).
- Kube, R., Bianchi, F. M., Brunner, D., and LaBombard, B. (2019). Outlier classification using autoencoders: application for fluctuation driven flows in fusion plasmas. *Rev. Sci. Instrum.* 90, 013505. doi:10.1063/1.5049519
- Laakso, H., Perry, C., McCaffrey, S., Herment, D., Allen, A. J., Harvey, C. C., et al. (2010). “Cluster active archive: Overview,” in *The cluster active archive*. Editors H. Laakso, M. Taylor, and C. P. Escoubet (Dordrecht, Netherlands: Springer Netherlands), 3–37
- Lance, G. N., and Williams, W. T. (1967). A general theory of classificatory sorting strategies: 1. Hierarchical systems. *Comput. J.* 9, 373–380. doi:10.1093/comjnl/9.4.373
- Le, L., Patterson, A., and White, M. (2018). “Supervised autoencoders: improving generalization performance with unsupervised regularizers,” in *Advances in neural information processing systems* (Red Hook, NY: Curran Associates, Inc.), Vol. 31, 107–117.
- Liu, C. M., Fu, H. S., Liu, Y. Y., Wang, Z., Chen, G., Xu, Y., et al. (2020). Electron pitch-angle distribution in earth’s magnetotail: pancake, cigar, isotropy, butterfly, and rolling-pin. *J. Geophys. Res. Space Phys.* 125, e2020JA027777. doi:10.1029/2020JA027777
- Lui, A. T. Y. (1987). *Road map to magnetotail domains*. Baltimore, MD: John Hopkins University Press, 3–5.
- Lukasová, A. (1979). Hierarchical agglomerative clustering procedure. *Pattern Recogn.* 11, 365–381. doi:10.1016/0031-3203(79)90049-9
- McLachlan, G., and Peel, D. (2000). *ML fitting of mixture models*. Hoboken, NJ: John Wiley & Sons, Ltd., Chap. 2, 40–80. doi:10.1002/0471721182.ch2
- Øieroset, M., Lin, R. P., Phan, T. D., Larson, D. E., and Bale, S. D. (2002). Evidence for electron acceleration up to ~300 keV in the magnetic reconnection diffusion region of earth’s magnetotail. *Phys. Rev. Lett.* 89, 195001. doi:10.1103/PhysRevLett.89.195001
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830. doi:10.5555/1953048.2078195
- Peerenboom, K., Parente, A., Kozák, T., Bogaerts, A., and Degrez, G. (2015). Dimension reduction of non-equilibrium plasma kinetic models using principal component analysis. *Plasma Sources Sci. Technol.* 24, 025004. doi:10.1088/0963-0252/24/2/025004
- Rème, H., Aoustin, C., Bosqued, J. M., Dandouras, I., Lavraud, B., Sauvaud, J. A., et al. (2001). First multispacecraft ion measurements in and near the Earth’s magnetosphere with the identical Cluster ion spectrometry (CIS) experiment. *Ann. Geophys.* 19, 1303–1354. doi:10.5194/angeo-19-1303-2001
- Ruder, S. (2016). *The overview of gradient descent optimization algorithms*. arXiv e-prints.
- Sakurada, M., and Yairi, T. (2014). “Anomaly detection using autoencoders with nonlinear dimensionality reduction,” in Proceedings of the MLSDA 2014 2nd workshop on machine learning for sensory data analysis, Gold Coast, Australia, December 2, 2014. (New York, NY: Association for Computing Machinery), 4–11. doi:10.1145/2689746.2689747
- Sharma, A. S., and Curtis, S. A. (2005). *Magnetospheric multiscale mission*. Dordrecht, Netherlands: Springer Netherlands, 179–195. doi:10.1007/1-4020-3109-28
- Walsh, A. P., Fazakerley, A. N., Forsyth, C., Owen, C. J., Taylor, M. G. G. T., and Rae, I. J. (2013). Sources of electron pitch angle anisotropy in the magnetotail plasma sheet. *J. Geophys. Res. Space Phys.* 118, 6042–6054. doi:10.1002/jgra.50553
- Walsh, A. P., Owen, C. J., Fazakerley, A. N., Forsyth, C., and Dandouras, I. (2011). Average magnetotail electron and proton pitch angle distributions from cluster PEACE and CIS observations. *Geophys. Res. Lett.* 38, 1029. doi:10.1029/2011GL046770
- Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* 58, 236–244. doi:10.1080/01621459.1963.10500845
- Zeiler, M. D. (2012). Adadelta: an adaptive learning rate method. arXiv e-prints.

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Bakrania, Rae, Walsh, Verscharen and Smith. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.