The impact of dimensionality reduction of ion counts distributions on preserving moments, with applications to data compression

D. da Silva[1,2,3]*, C. Bard[1], J. Dorelli[1], M. Kirk[1], B. Thompson[1] and J. Shuster[1,4]

[1]Heliophysics Sciences Division, NASA Goddard Spaceflight Center, Greenbelt, MD, United States, [2]Goddard Planetary Heliophysics Institute, University of Maryland Baltimore County, Baltimore, MD, United States, [3]Laboratory for Atmospheric and Space Physics, University of Colorado Boulder, Boulder, CO, United States, [4]Department of Astronomy, University of Maryland College Park, College Park, MD, United States

The field of space physics has a long history of utilizing dimensionality reduction methods to distill data, including but not limited to spherical harmonics, the Fourier Transform, and the wavelet transform. Here, we present a technique for performing dimensionality reduction on ion counts distributions from the Multiscale Mission/Fast Plasma Investigation (MMS/FPI) instrument using a data-adaptive method powered by neural networks. This has applications to both feeding low-dimensional parameterizations of the counts distributions into other machine learning algorithms, and the problem of data compression to reduce transmission volume for space missions. The algorithm presented here is lossy, and in this work, we present the technique of validating the reconstruction performance with calculated plasma moments under the argument that preserving the moments also preserves fluid-level physics, and in turn a degree of scientific validity. The method presented here is an improvement over other lossy compressions in loss-tolerant scenarios like the Multiscale Mission/Fast Plasma Investigation Fast Survey or in non-research space weather applications.

KEYWORDS

machine learning, dimensionality reduction, compression, ions, plasma instrumentation, fluid theory

Highlights

- Algorithm is developed for reducing the dimensionality of ion counts distributions from MMS/FPI
- Algorithm performance is described and analyzed in terms of its effect on calculated plasma moments
- Usage of the method to perform data compression problems is discussed and demonstrated

Introduction

Dimensionality reduction is the process of reducing complex high-dimensional data to lower-dimensional representations wherein the number of parameters is reduced, at minimal cost to a certain perspective of information contained within the data. The concepts of dimensionality reduction have a rich history in mathematical physics (Brunton and Nathan Kutz 2022). They undeniably play a role in the field of space plasma physics, with examples ranging from reduced-order spherical harmonic modeling of the external geomagnetic field (Chulliat et al., 2015), truncated Fourier Series representations of plasma waves, and wavelet decomposition of solar imagery (Stenborg and Cobelli 2003; Stenborg et al., 2008).

Dimensionality reduction is typically performed by series decompositions designed from idealized geometries with attractive mathematical properties. Among the most common of these is an ordering for sets of basis functions wherein the leading functions capture the broadest structure, and the following functions capture increasingly fine detail. A further attractive property of such series decompositions is that when the number of basis functions approaches infinity, the error between the reconstructed model and the original data converges to zero within some appropriate error metric. This is the case for both Fourier series and spherical harmonics. In some cases, transformed representations serve as convenient coordinate systems for discussing high-level physical features of the untransformed structure. For example, the time-rate-of-change of specific spherical harmonic terms is a convenient language to discuss movement of a major deviation of Earth's magnetic field from the dipole model known as the South Atlantic Anomaly (SAA) (Heynderickx 1996; Finlay et al., 2020). Additionally, previous work has analyzed the use of spherical harmonics for modeling velocity distribution functions with Cluster data (Viñas and Gurgiolo, 2009), which showed that relatively few spherical harmonic basis functions were necessary to reproduce moments in that dataset.

A recent prospect within the field of data-driven science and machine learning is learning dimensionality reduction for given sets of data, including the derivation of custom latent representations. This is commonly done with neural network autoencoders (Goodfellow et al., 2016; Bank et al., 2020) and Principal Component Analysis (PCA) (Ringnér, 2008; Hastie et al., 2009). With autoencoders, the parameters of a highly dynamic functional space are optimized to produce a copy of its input data as output whilst passing its intermediary representation through a low-dimensional bottleneck. By pushing the information through such a bottleneck, both a transform and inverse transform are learned simultaneously. The portion of the network up to and including the bottleneck becomes a transform method (sometimes called encoder), the portion including and following the bottleneck becomes the inverse transform method (sometimes called

decoder), and the data at the stage of the bottleneck is a latent, low-dimensional representation. The advantage of neural network auto-encoders is that they specialize their representation capability to the specific training data used, avoiding unnecessary generalizability in the process.

In this work, we look at using a neural network autoencoder to reduce the dimensionality of raw ion counts distributions from different plasma regions sampled by the Fast Plasma Investigation (FPI) Dual Ion Spectrometers (DIS) on the Magnetospheric Multiscale Mission (MMS) (Burch et al., 2016; Pollock et al., 2016). The FPI instrument currently uses a discrete wavelet transform method (Yeh et al., 2005) with a high compression level for fast survey quality data and a lower, generally lossless, compression level for burst quality data (Barrie et al., 2017; Barrie et al., 2019).

In general computing, different types of data have their own respective compression algorithms adapted to the needs of the data. For example, the problem of image compression has been approached with JPEG or PNG; audio with MP3; and video with MPEG-4 AVC/H.264. Each approach adapts to its domain by identifying the highest priority information to be retained for that type of data and designing itself to be effective for the intended application. This paper moves towards the first compression algorithm designed originally for plasma distribution data, made with an eye towards physical validity and applicability to spaceflight implementation.

The paper is structured as follows. In the Applications of Compression section, the application to data compression in the context of space physics missions is reviewed and discussed. In the Dimensionality Reduction Method section we present the method and discuss the architectural choices and data preparation. In the Impact on Moments section the performance is analyzed per orbit configuration, and the differences between orbital configurations are discussed. This lays the foundation for the Demonstration of Compression Algorithm section, which describes a fully functional compression algorithm utilizing the trained neural network model capable of being run on board a spacecraft or on the ground. Finally, in the Conclusion and Outlook section, we review the results of the paper, describe future work, and look forward to improvements to theoretical understanding from the study.

Applications of compression

At the most basic level, space science missions in the heliosphere measure particles and fields *in-situ*. Field data are 3-coordinate vectors which are collected through magnetometer and electric field instruments, generated once per collection cycle. In contrast, instruments measuring particle distributions produce much larger arrays per collection cycle—up to four orders of magnitude higher. Data transmitted for plasma

TABLE 1 Number of elements per velocity distribution for various plasma instruments.

Mission	Launch year	Instrument	Number of elements per velocity distribution
Magnetospheric Multiscale Mission (MMS)	2015	Fast Plasma Instrument/Dual Ion Spectrometers (FPI/DIS)	**16,384** (=32 azimuth x 16 elevation x 32 energy)
		Fast Plasma Instrument/Dual Electron Spectrometers (FPI/DES)	**16,384** (=32 azimuth x 16 elevation x 32 energy)
THEMIS	2007	Ion Electrostatic Analyzer (iESA)	**15,872** (= 32 azimuth x 16 elevation x 31 energy)
Cluster	2000	Cluster Ion Spectroscopy Experiment (CIS)	**3,968** (=16 azimuth x 8 elevation x 31 energy)

Number of elements per collection timestep for various plasma instruments. For the MMS mission, see Pollock et al., 2016. For THEMIS mission, see McFadden et al., 2008. For the Cluster mission, see Rème et al., 1997. The number of elements per velocity distribution here is purely the size of the counts array—it does not include any type of supplementary information that may accompany the downlink such as headers, calibration parameters, *etc.*

measurements is dominated by the plasma velocity distributions (Table 1). Therefore, to efficiently manage the telemetry reserve for Heliophysics science goals, one should first look at the plasma measurements.

To date, methods for managing the overwhelming amount of plasma data include using selective downlink *via* scientist-in-the-loop systems and various levels of data compression. In a selective downlink solution, the mission will downlink a lower quality version of the data that is less than ideal for the targeted research application ("review quality") and allow human reviewers to pick a subset for transmission at full quality ("full quality") (Fuselier et al., 2016; Baker et al., 2016; Argall et al., 2020). The idea is that review quality data is made much smaller—through lossy compression and/or time averaging, with the quality sufficient to just give a broad picture of whether the collection warrants a full quality retrieval.

The use of data compression, which can be used to minimize the size of both review quality and full quality data, is another method. In the MMS mission, the image compression algorithm set by the Consultive Committee for Space Data Systems (CCSDS) standard was utilized (Yeh et al., 2005). However, while utilizing a standard was convenient for implementation, the plasma-agnostic nature of the algorithm led to small errors appearing in the worst areas for maintaining physics integrity. In the MMS mission, the full quality data utilized lossy compression during the first phase of the mission, but this decision was reversed in favor of lossless compression for the following phases (Barrie et al., 2017; Barrie et al., 2019; da Silva et al., 2020).

The use of lossy compression for the measurement of physical variables and scientific inquiry is a delicate subject. Research missions, such as MMS, demand near-lossless versions of the full quality data accompanied by a high guarantee that all sources of noise and artifacts are well understood. Such research-grade science quality data is not the target for the compression presented here. Instead, review quality data (which is already lossy in the case of MMS/FPI) provides a much safer opportunity to better apply novel lossy compression algorithms. Furthermore, there exists another class

of non-research operational space weather missions which, on a case-to-case basis, may be judged more tolerant of lossy compression.

Dimensionality reduction method

The neural network dimensionality reduction method is based on an autoencoder (as illustrated in Figure 1). The autoencoder simultaneously maps the full-dimensional data to a lower dimensionality and from the low-dimensional representation back to the original representation. Based on previous work in da Silva et al. (2020) to remove compression artifacts from satellite data, we use a multi-layer perceptron network operating on patches of the raw ion counts distributions from MMS/FPI DIS (Pollock et al., 2016). We use rectified linear unit (ReLU) activations on the hidden and final layers. The use of the ReLu instead of linear activation in the final layer guarantees the output counts are never negative. As da Silva et al. (2020) discuss, the multi-layer-perception neural network is chosen over more complex architectures (e.g. convolution neural networks) because of an interpretability aspect. Specifically, the weights of the trained network can be visualized as basis functions. Another result from da Silva et al. (2020) is that the emergent basis functions exhibit non-random spatial structures, comparable to spherical harmonics, despite having no spatial relationships integrated into the network. This is in contrast to, e.g., Convolutional Neural Networks (CNNs) that use spatial locations of input pixels relative to one another. We also experimented with multi-layer networks, adding additional "transition" layers before/after the low-dimensional representation. However, this did not improve performance and made the resulting network more volatile and dependent on the random weight initialization.

During Phase 4 of the MMS mission (9/28/2018–9/30/2019), four orbital configurations were executed to sample different

FIGURE 1
The neural network encoding and decoding steps of the dimensionality reduction method. A normalization to guarantee the mean number of counts is preserved is not shown here. In the encoding step, a patch is retrieved from the full counts distribution structure, flattened, and processed through the transform encoder network. The encoded form of this patch is stored as a column in a matrix, where the number of columns is the number of patches. For the decoding step, an encoded column is retrieved from the table, passed through the decoder, and placed back in its original location in the full counts distribution.

magnetosphere populations: the day-side magnetopause region (4B), the tail region (4D), and the dawn/dusk flanks (4C/4A). In this paper, we looked at the neural network dimensionality reduction for a per-orbital-configuration training set. Data was used exclusively from the MMS1 spacecraft to prevent any spacecraft-specific issues from complicating the experiment. All bursts were downloaded from each of the orbital configurations, and the number of timesteps present in each burst was recorded. From each orbital configuration, a total of 50,000 timesteps were selected randomly for training and 5,000 timesteps for testing/validation (with no overlap between the two sets). We excluded the portion of the data recorded for a solar wind study; this used a non-standard instrument configuration which scanned different energies from the normal operating mode. This data was identified by the string "12–14" in the "Energy_table_name" header field. Finally, we trained individual sets of networks for each orbital configuration (4A-4D).

The raw ion counts distribution is the number of ion detections by the instrument within a collection cycle, over a look direction solid angle and energy range. It is related to the ion velocity distribution in units of phase space density $s^3 cm^{-6}$ through an instrument response variable known as the geometric factor (Collinson et al., 2012). The raw ions count distributions are of shape (N_{az}, N_{el}, N_{en}) where N_{az} is the number of azimuth pixels (32), N_{el} is the number of elevation pixels (16), and N_{en} is the number of energy channels (32). We split each raw ion count distribution into smaller patches, which cover the full range of azimuth and elevation but are restricted to two energy channels. This results in the shape $(N_{az}, N_{el}, 2)$. The network architecture is duplicated between each patch location in the image. This means that one network is trained for the first two energies, while another network is independently trained for the next two energies. This was chosen to enable each network to tailor itself to a more specific portion of data. Experimentation with patches that spanned three energy shells was attempted, but

it was found that performance was worse. Experimentation with one energy shell was also attempted, but it was found that this lead to worse overall dimensionality reduction when the entire counts distribution was considered.

Figure 1 illustrates the dimensionality reduction method applied on a per-patch basis. Starting with a patch selected from the training set, the transform encoder encodes the patch into a latent representation with dimensional size N (where N is a user-selected parameter). The fully encoded counts distribution from the autoencoder is thus $16 \times N$, where 16 is the number of patch slots for the DIS data. The process for decoding is the reverse of encoding: the latent representation is run through the middle and final layers of the autoencoder network. The decoded representation of each patch is stored in an array holding the full reconstructed count distribution.

A critical question is the dimensionality of the latent representation (as given by the parameter N), which controls the amount of dimensionality reduction. This is left as a tunable parameter for the algorithm, with investigations into the ideal parameter discussed in the Impact on Moments section.

To process the data more efficiently, we structured the problem with as much physical knowledge as possible. This is done by providing the neural network input data pre-processed in a form that is intended to simplify the task the network performs.

First, we align the data as best as possible in the GSE frame. In the native data layout, the pixels are ordered in the azimuth of the spacecraft frame. We reorder the data across the azimuth dimension so that counts at the same azimuth index, elevation, and energy correspond to the same look direction (within the size of azimuth and elevation bins). This allows the network to learn representations of the training set in a non-spinning frame. This is preferred because the network can, for instance, learn correlations between count rates in X and Y GSE directions.

We also define a patch size of two energy channels. This allows the network to take advantage of information redundancy between energy channels and learn both the broad structure common across energy channels and a concise relative structure that captures how it changes between energy channels.

Finally, we adjust the data reconstructed by the network so that the mean number of counts in each energy shell prior to encoding is preserved after the reconstruction. By doing this we require only N_{en} additional parameters for the entire (N_{az}, N_{el}, N_{en}) counts distribution, and in turn achieve increased performance in the moments (particularly the number density). As the network itself decodes two energy shells at a time (patch size of 2), we force the mean number of counts to be equal between the original and reconstructed energy shells, each of shape (N_{az}, N_{el}). This is described in Eq. 1, where $\vec{C}^E_{recon,adjusted}$ is the reconstructed vector of counts for energy shell E, $\vec{C}^E_{recon}$ is the reconstructed counts for energy shell

E, $\bar{C}^E_{orig}$ is the mean number of counts in the original energy shell E, and $\bar{C}^E_{recon}$ is the mean number of counts in the reconstructed energy shell E. We note that merely preserving the mean number of counts per patch (instead of per energy shell) leads to artifacts and discontinuities along the patch boundaries in the final reconstruction.

Eq. 1 – Adjustment to Preserve Mean Number of Counts per Energy Shell.

$$\vec{C}^E_{recon,adjusted} = \vec{C}^E_{recon} \left(\frac{\bar{C}^E_{orig}}{\bar{C}^E_{recon}} \right) \qquad (1)$$

The networks are trained using the ADAM optimizer with a learning rate of .001 (Kingma and Jimmy, 2014). The loss function uses the squared residual error of the counts. We also experimented with custom loss functions to compare moments integrated over the subset of velocity space associated with a patch (i.e. partial moments). In a partial moment, the triple integral over all of velocity space is rewritten to integrate only over the subset of velocity space where the energies are within the range associated with the patch. Geometrically, this corresponds to restricting the integration domain to the volume between a sphere at constant energy E_1 (lowest energy of the patch), and a larger sphere at constant energy E_2 (highest energy of the patch).

The moments are computed directly after converting the counts to phase space density. The performance of a dimensionality reduction is evaluated using the moments derived from the reconstructed ion velocity distribution. This is done under the perspective of preserving conservation laws in the reduced model data. As an example, a dimensionality reduction which performs well to preserve n and $\vec{v}$ on data can be argued that to also preserve any equation which uses only those variables, such as $\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \vec{v}) = 0$ (with the understanding that $\rho = \rho(n)$). Therefore, the performance in accurately reconstructing the moments can be said to generate confidence in the ability to preserve continuum-level physics. It is understood that this does not necessarily speak to the ability to preserve kinetic-level physics, and such an analysis is left for future work.

However, there is mathematical support that strong agreement in moments leads to agreement in the distribution function. The result of the Hausdorff moment problem states that if two distributions $f_1(\vec{v})$ and $f_2(\vec{v})$ have the same moments $M_1^{(k)} = M_2^{(k)}$ for all moment order $k = 0, 1, 2, \ldots \infty$, then it is necessarily true that $f_1 = f_2$ (Hausdorff 1921a; Hausdorff 1921b; Shohat and David Tamarkin 1950). This is only guaranteed because f_1 and f_2 are defined on a bounded space (which can be taken here as the velocity space subset given by $||\vec{v}|| < c$) and is not necessarily true for distributions defined on unbounded space. We note a major limitation of this theorem is that it is stated in terms of absolute equality for an infinite sequence, and similarly concludes that $f_1 = f_2$ exactly.

In practice absolute equality is not achievable and therefore the literal interpretation is limited. However, we believe the intuitive principle of the theorem is still useful and provides a foundation of mathematical support for the methodology.

The final loss function is given in Eq. 2, where $\vec{C}_{recon}$ is the reconstructed counts patch, $\vec{C}_{orig}$ is the original counts patch, $N_{moments}$ spans each scalar moment $(n, nV_x, nV_y, nV_z, nk_bT_{xx}, \ldots nk_bT_{zz})$, N_{pixels} is the number of pixels in the patch, the variable λ_m spans the penalty weights for moments, λ_{SC} is the penalty weight for squared counts residuals, $M_m(\vec{C})$ is the m'th moment of a counts array per the above ordering, and $\vec{w}_m$ holds information from both the conversion from counts to phase space as well as the numerical quadrature weights for trapezoidal integration for the moments integral in a single set of weights. That is, calculating $\vec{w}_m \cdot \vec{C}$ applies numerical integration. We note that $N_{moments}$ uses number density times the velocity and temperature moments instead of the direct velocity and temperature moments because division by zero occurs for some empty counts distributions where $n = 0$.

Eq. 2 – Loss Function including Moments Comparison

$$L\left(\vec{C}_{recon}, \vec{C}_{orig}\right) = \lambda_{SC} \sum_{i=0}^{N_{pixels}-1} \left(C_{orig}^i - C_{recon}^i\right)^2$$
$$+ \sum_{m=0}^{N_{moms}-1} \lambda_m \left[M_m\left(\vec{C}_{recon}\right) - M_m\left(\vec{C}_{orig}\right)\right]^2 \quad (2)$$

$$M_m(\vec{C}) = \vec{w}_m \cdot \vec{C} \quad (3)$$

Different values of the penalties λ_m were experimented with include a weighted combination of the squared residual of counts and the squared error of moments, including where the weight on the square residual of counts was zero. However, it found was found that while conceptually elegant, these methods did not lead to better performing networks. It is hypothesized that the reasons these networks trained with such loss functions did not perform well was because the additional penalty term produces an especially non-convex, difficult to train optimization space.

Impact on moments

We must tune the dimensionality reduction parameters to maximize the preservation of the count distribution moments after encoding and decoding. Specifically, we focus on the number density n, the bulk velocity $\vec{v}$, and the temperature T. Moment preservation is a proxy for how well one can trust that fundamental continuum conservation laws are being followed.

An independent variable in the study is the dimensionality of the latent representation (N), which determines the number of parameters each patch (and the cumulation of all patches) is (are) condensed into. The performance of the moment

reconstruction is evaluated in terms of the correlation coefficient r^2, with the correlation done between the original and reconstructed data. The performance is also demonstrated through comparison of the original and reconstructed moments in the form of moments time series located later in this manuscript.

The correlation of the number density is strong for all orbital configurations, with the r^2 peaking between .998 and .999 for each. Figure 2 shows the per-orbit correlation coefficient for the bulk velocity as a function of the dimensionality reduction fraction (DRF). The DRF represents the total dimensionality of each latent patch representation relative the original dimensionality of the patch, where 1 corresponds to no dimensionality reduction. Specifically, it is given by $DRF = N/1024$, where $1024 = 32 \times 16 \times 2$ is the original size of each patch.

In the V_x direction, the r^2 falls between .97 and .98; in V_y, it falls between .93 and .98; and in V_z, it falls between .92 and .97. First, we note that a knee in the performance curve as a function of the reduction fraction occurs around .05 and .10 for each curve, corresponding to total dimensionality reductions of a factor between 10 and 20. It is noted here that we consider one DRF for all patches; but with a deeper investigation one could set the DRF per patch location to account for varying levels of condensability between patches. Readers eager to see a time-series comparison of original and reconstructed data using $DRF = .098$ (closest discrete point to .1) may jump ahead to Figure 4 before coming back to read this section.

We observe that the performance of V_x, V_y, and V_z is not only dependent on the DRF and the orbit, but also the bulk velocity component (GSE X, Y, Z). In some cases, the performance counter-intuitively drops once the DRF is increased past a certain point. Simply put, this is due to overfitting. To explain these phenomena, we note that each autoencoder is trained to just a locally optimal solution susceptible to imperfection from the training process: it is not a global minimum. The decrease in performance is thus a result of the network becoming more difficult to train as the number of parameters becomes too large. This is evident because the larger network contains the smaller network, so the additional parameters must over-complicate the reconstruction. The fact that this effect appears for some orbital configurations and not others is understood to be a dependence on the data it is trying to fit to.

The influences of orbital configuration and bulk velocity on the reconstruction performance is linked to the magnetospheric populations represented in each configuration. While it is difficult to give definitive conclusions, it is hypothesized that for some orbital configurations the relevant information in each population is more difficult to condense into a low-dimensional representation. The portions of the counts distribution relevant to the bulk velocity are possibly less redundant with the rest of the

FIGURE 2
The correlation coefficient r^2 performance of the bulk velocity components V_x, V_y, V_z (GSE) on a per-orbital-configuration basis. The independent variable on the y-axis in the dimensionality reduction fraction (DRF) corresponds to the size of the latent representations for each patch. In this plot, we observe a knee in the performance curve occurring at DRF's between .05 and .10, corresponding to a dimensionality reduction between 10 and 20. Other characteristics of the performance, such as the dependence on orbital configuration and bulk velocity component, as well as the low performance as the DRF before the knee, are discussed in the text. DRFs above one corresponds to networks that do not actually perform dimensionality reduction and instead demonstrates information loss not associated with a reduced number of parameters.

distribution and require more information for accurate reconstruction.

Demonstration of Compression Algorithm

The effect of the compression algorithm can be seen in Figure 3 with before and after spectrograms for a scientifically relevant event, and in Figure 4 with before and after time series for the same event. This event originated in the 4B dayside orbital configuration. This complements the full end-to-end compression algorithm based on what is described in Dimensionality Reduction Method, parameterized by a latent patch size of $N = 100$ corresponding to $DRF = .098$. This example resulted in a compression ratio of 30.2 times, which is almost double the average ratio of 16.67 times for the lossy review quality (fast survey) data from MMS/FPI DIS instrument over Phases 1A and 1B of the mission, as reported in Table 1 of Barrie et al. (2019).

Because the means are preserved between the original and reconstructed data, the error in the reconstructed counts spectrogram (Figure 3, bottom) is very low. This result is taken with a grain of salt because the spectrogram perspective ignores important errors in the directional distribution of counts. In a sense, the error in the spectrogram is inevitable because we (effectively) transmit the entire spectrogram as part of the compressed data *via* the individual means of each energy shell. Therefore, in the spectrogram, error is only caused primarily by quantization of each mean. This is particularly apparent in area of high counts and areas of very low counts where a single count difference is significant. In the time series perspective (Figure 4), the apparent relative errors are, as expected, on a scale higher than the errors in the spectrogram.

The compressed data in Figure 3 includes additional processing of quantization and entropy coding. For compression algorithms, the quantization step compliments the dimensionality reduction to further reduce the number of bits by truncating the precision of the latent representation coefficients. The entropy coding step finalizes the compression through a lossless process where the final compressed bytes are reduced by alternating symbols to reduce information entropy. In this demonstration, the entropy coding was performed using the GZIP software which implements the DEFLATE algorithm (Deutsch 1996a; 1996b). Throughout the intervals experimented with using these settings, the dimensionality reduction reduced the size by a factor 10.24, the quantization by a factor of 1.6, and entropy coding by a varying amount between a factor of 1.8–2.2.

The quantization approach used is to reduce the number of bits allocated to each latent representation coefficient. First, each floating-point coefficient is converted to a 16-bit floating point number, which itself is composed of a single bit dedicated to the sign of the number and bits corresponding to the fractional/exponent parts of a scientific notation representation of the number. In the traditional IEEE specification of a 16-bit floating point number, the sign bit is 1 bit, the fractional part is 10 bits, and the exponent is 5 bits. In our quantization scheme, we reduce the 16-bit floating point number to a 10-bit floating point number by reducing the fractional part to 4 bits.

In Figure 3 the compression demonstrates strong ability to capture the energy spread of the distribution as well as transitions between cold and hot plasma. Features of the spectrogram such as bimodal populations (first 10 s) and populations skewed with tails extending to lower energies (about 10–25 s in) are also well preserved. The background flux of the spectrogram (dark purple) is similar between the original and reconstructed data, with some cells where zero counts occur matching in a way that is consistent

FIGURE 3
Demonstration of the compression algorithm in its end-to-end form displaying the original ion data, the compressed (reconstructed) ion data, and the relative error between the two given by (original—reconstructed)/original. This uses a version of the model with a latent patch size of $n = 100$ corresponding to $DRF = .098$. In addition to the dimensionality reduction, quantization is used to trim 6 bits off the fractional parts of IEEE 16-bit floating point latent representation coefficients, and the GZIP software implementing DEFLATE is used for lossless entropy coding. This interval comes from data measured during the 4B dayside orbital configuration.

with the underlying random detection process. Between 17:41:00 and 17:41:30 the dip in the peak of the main population is also well preserved. The relative error is lowest in areas where the flux is the highest, and more apparent in regions of low flux.

We analyze the effect of the reconstruction on the fluid moments. The relationship between original and reconstructed moments shows large agreement in qualitative physics arising from analysis of the data (Figure 4). This is using the full end-to-end compression algorithm outlined in the Demonstration of Compression Algorithm section, including quantization and entropy coding, and is the same interval displayed in the spectrogram from that section (Figure 3). This comparison shows good agreement in the moments of the data, wherein an analysis looking at only the reconstructed moments would come to similar conclusions about the plasma environment as if the original moments were used.

In the beginning of the interval around 17:39:05 is a transition from a cold, lower-density ion population to a much hotter and

higher-density population. This transition is captured well in the original and reconstructed data. Following the fly-through of this hot and higher-density population is an abrupt change in flow velocity. Until the end of the interval, the flow velocity shifts in the v_y and v_z directions and more slightly in the v_x direction. This is generally tracked well between the original and reconstructed data, with differences existing but generally not impacting an understanding of the environment. Towards the end of the interval, starting around 17:42:20, is the appearance of oscillations in the number density possibly indicating plasma wave behavior. This is captured very well in the number densities obtained from the reconstructed data. Similarly, the other variables during the time of oscillation are in good agreement to support understanding of the plasma wave phenomenon.

Analysis of the signed moment errors in the test set was performed and is presented in Figure 5. The line drawn is centered at the mean error for each moment and phase, and
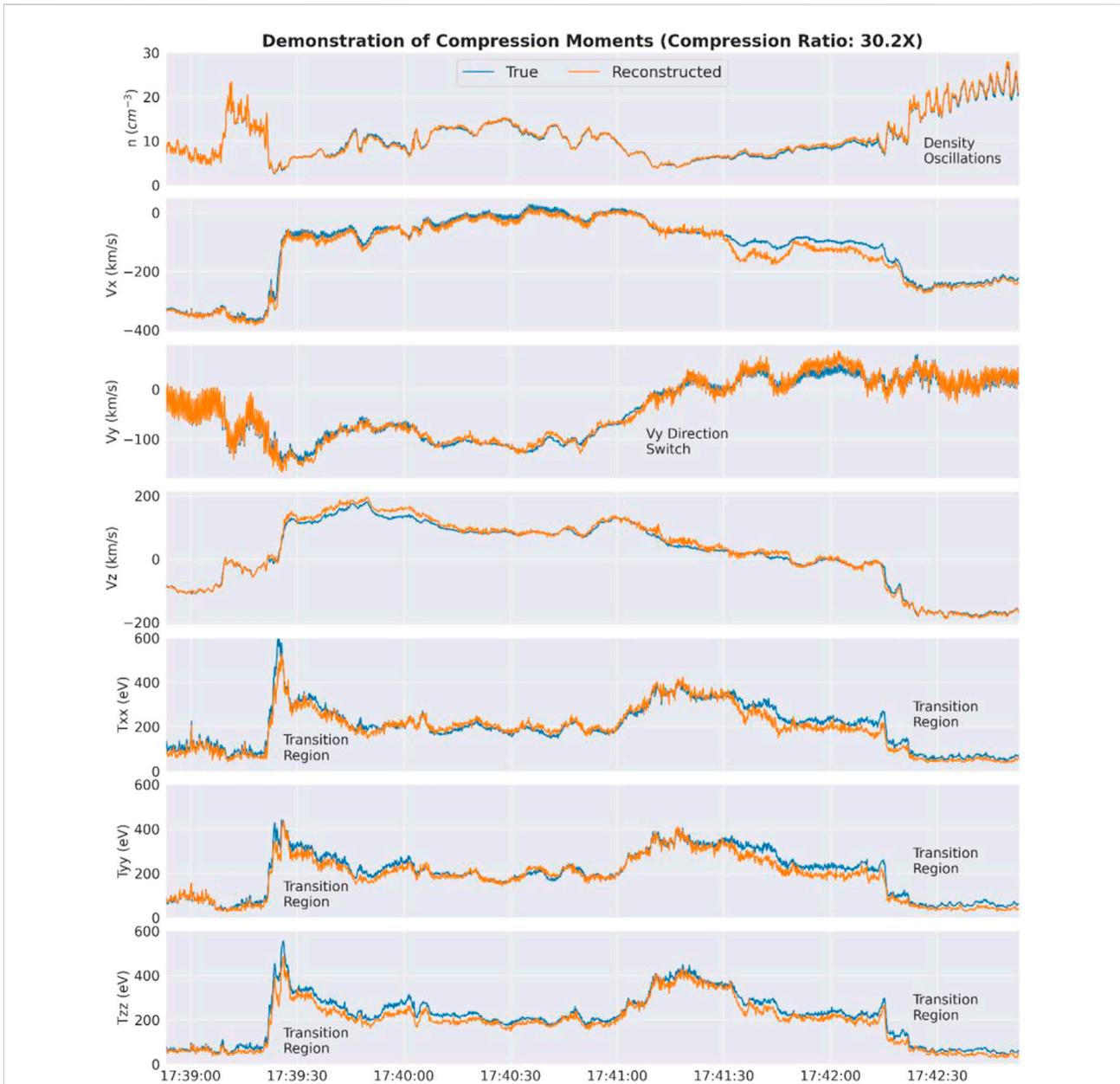
FIGURE 4
A time series comparison of moments from the original and reconstructed versions of the data. This is using the full end-to-end compression algorithm outlined in the Demonstration of Compression Algorithm section, including quantization and entropy coding. This is the same interval displayed in spectrogram from that section. This comparison shows good agreement in the moments of the data, wherein an analysis looking at only the reconstructed moments would come to similar conclusions about the plasma environment as if the original moments were used. The Txx, Tyy, and Tzz variables are the diagonal elements of the temperature tensor. All coordinates are in the GSE coordinate system.

the error bars correspond to the 95% confidence interval. To calculate the 95% confidence interval, the list of errors was sorted in ascending order and the 2.5% and 97.5% positions were selected. The confidence interval can be interpreted to indicate that 95% of the errors found in the test set fell between the drawn error bars. In the number density moment, the 95% confidence interval is low enough to distinguish sheath plasmas from

magnetosphere plasmas. Similarly, the temperature moments are sufficient to distinguish a cold plasma from a hot plasma. Asymmetries in the error distribution are observed where the radius of the confidence interval above and below are not equal, indicating that the error distribution is first and foremost not Gaussian, but also not symmetric in general. For all variables and orbital configurations, the mean signed error is close to zero

FIGURE 5
Analysis of Signed Errors for each of the moments (reconstructed—original), investigated through 95% confidence intervals calculated from the test set (sample size 5000). The primary line drawn corresponds to the mean signed error, and the error bars correspond to the confidence interval. The confidence interval can be interpreted to indicate that 95% of the signed errors found in the test set fell between the drawn error bars. The density and temperature moments appear suitable to distinguish dense plasma from sparse plasma and hot plasma from cold plasma, and the velocity components appear suitable to determine the general flow direction.

TABLE 2 95% range (2.5% and 97.5% percentiles) of test set values to better interpret Figure 5.

	4A Dusk Flank	4B Dayside	4C Dawn Flank	4D Tail
n (cm^{-3})	.499–66.145	.822–60.546	.467–49.276	.146–4.887
v_x (km/s)	-476.115–26.707	-554.104–45.632	-414.732–63.921	-284.746–173.642
v_y (km/s)	-44.425–208.268	-167.317–183.669	-193.874–47.641	-83.245–100.394
v_z (km/s)	-70.394–125.645	-89.795–136.619	-113.052–99.698	-55.794–68.221
T_{xx} (eV)	31.983–2,847.648	3.710–2,631.109	14.099–2,953.167	83.653–3,120.441
T_{yy} (eV)	9.971–2,847.975	1.399–2,609.728	7.110–2,800.843	74.695–2,796.996
T_{zz} (eV)	19.018–2,565.776	2.152–2,300.594	8.079–2,682.937	82.152–2,883.957

within the scale of the data, a reassuring feature. For reference, a table of the 95% ranges for the test set data (2.5% and 97.5% percentiles for each variable) is included in Table 2. This is meant to provide context for those not familiar with the data to comprehend the signed errors reported in Figure 5.

Conclusion and outlook

In this paper we look at the topic of learning dimensionality reduction on counts distributions as they would be measured by a spacecraft. The capability of dimensionality reduction for counts distributions has a natural application to compression but is

more generally a continuation of traditional methods (spherical harmonics, Fourier series, Wavelet transforms) in the era of data-intensive modeling. In addition to compression, dimensionality reduction is also applicable to synthesizing spacecraft data into condensed features for other machine learning algorithms, saving them the work of learning to distill counts distributions themselves. The performance reported here is an initial attempt at using machine learning for compressing plasma data. In our demonstration, we showed that our initial attempt provided a compression ratio of 30.2 times, almost double the average ratio of 16.67 times for the lossy review quality (fast survey) data from MMS/FPI DIS instrument during mission Phases 1A and 1B (Barrie et al., 2019). We believe that there is

room for improvement; further experimentation with different kinds of network architectures is a promising avenue for investigation. We emphasize that the use of lossy compression should be aimed at review quality data or non-research space weather applications judged on a case-by-case basis.

In this paper, we contribute the perspective of validating the dimensionality reduction in terms of preserved physics. The approach utilized here is to consider its impact on moments and the fluid perspective—best applicable to Maxwell-Boltzmann-conforming distributions—as a metric for the ability to preserve scientific integrity. More advanced capabilities exist for analyzing the integrity of reconstructed counts distributions—including analysis of terms of the Vlasov equation measured directly from spacecraft data (Shuster et al., 2019; Shuster et al., 2021). Careful analysis of these Vlasov equation terms may offer increased confidence in the scientific validity of the counts distributions at the kinetic level for collisionless plasma environments. Finally, in this paper we only look at the ion counts distributions; a natural extension would be a similar study with electron counts distributions.

A future avenue for this work includes studying the practical details required to implement such a compression systems in flight. In the context of mission-development and systems engineering, developing a sample concept of operations and data format specification would increase the technical readiness level (TRL) and ease development burden on the benefactor. In the spirit of this, it would be of significant advantage to understand whether data from a past space mission could be used to train a compression algorithm for a future spaceflight mission, particularly before that future spaceflight mission launches. For instance, could THEMIS or CLUSTER be used to train a compression algorithm for MMS FPI/DIS, pretending that MMS had not yet launched? If such could be done successfully, a subsequent question would be how closely the orbits and instrumentation specifications need to match to retain strong performance.

For an *in-situ* plasma sensing mission seeking measurements of the distribution function, the data volume from the plasma instruments will overwhelm the transmission bandwidth compared to the magnetometer contribution. When there are compression artifacts, such artifacts should be quantified for both error analysis and development of artifact removal algorithms. The quantification of compression artifacts in terms of non-naïve error metrics is an active area of research in data compression, which we leave for future work. Finally, this paper aims to remind the machine learning community that contributions to the problem of data compression would be a directly measurable and practical problem in Heliophysics for us to undertake.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, The code for this paper is made open source at https://github.com/ddasilva/plasma-compression-neurips-2022 further inquiries can be directed to the corresponding author.

Author contributions

The primary research labor was done by DdS and CB. JD and BT provided senior/advisor-style feedback on the work throughout the project. MK and JS provided useful feedback and editing support during the preparation of the manuscript.

Funding

Acknowledgments

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Argall, M. R., Small, C. R., Piatt, S., Breen, L., Petrik, M., Kokkonen, K., et al. (2020). MMS SITL Ground Loop: Automating the burst data selection process. *Front. astronomy space Sci.* 7, 54. doi:10.3389/fspas.2020.00054

Baker, D. N., Riesberg, L., Pankratz, C. K., Panneton, R. S., Giles, B. L., Wilder, F. D., et al. (2016). Magnetospheric multiscale instrument suite operations and data system. *Space Sci. Rev.* 199, 545–575. doi:10.1007/s11214-014-0128-5

Bank, D., Koenigstein, N., and Giryes, R. (2020). Autoencoders. arXiv preprint arXiv:2003.05991.

Barrie, A. C., Smith, S. E., Dorelli, J. C., Gershman, D. J., Yeh, P., Schiff, C., et al. (2017). Performance of a space-based wavelet compressor for plasma count data on the MMS Fast Plasma Investigation. *J. Geophys. Res. Space Phys.* 122, 765–779. doi:10.1002/2016ja022645

Barrie, A. C., Smith, D. L., Elkington, S. R., Sternovsky, Z., Silva, D., Giles, B. L., et al. (2019). Wavelet compression performance of MMS/FPI plasma count data with plasma environment. *Earth Space Sci.* 6, 116–135. doi:10.1029/2018ea000430

Brunton, S. L., and Nathan Kutz, J. (2022). *Data-driven science and engineering: Machine learning, dynamical systems, and control.* Cambridge: Cambridge University Press.

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

Chulliat, A., Macmillan, S., Alken, P., Beggan, C., Nair, M., Hamilton, B., et al. (2015). *NCEI Geomagnetic Modeling Team. 2015: World Magnetic Model 2015 Technical Report. 1.* NOAA National Centers for Environmental Information. doi:10.7289/V5TB14V7

Collinson, G. A., Dorelli, J. C., Avanov, L. A., Lewis, G. R., Moore, T. E., Pollock, C., et al. (2012). The geometric factor of electrostatic plasma analyzers: A case study from the fast plasma investigation for the magnetospheric multiscale mission. *Rev. Sci. Instrum.* 83, 033303. doi:10.1063/1.3687021

da Silva, D., Barrie, A., Gershman, D., Elkington, S., Dorelli, J., Giles, B., et al. (2020). Neural network repair of Lossy compression Artifacts in the September 2015 to March 2016 duration of the MMS/FPI data set. *J. Geophys. Res. Space Phys.* 125, e2019JA027181. doi:10.1029/2019ja027181

Deutsch, P. (1996a). DEFLATE compressed data format specification version 1.3. *No. rfc1951,* 1996.

Deutsch, P (1996b). *GZIP file format specification* version 4.3. *No. rfc1952.* 1996. United States: RFC Editor

Finlay, C. C., Kloss, C., Olsen, N., Hammer, M. D., Toffner-Clausen, L., Grayver, A., et al. (2020). The CHAOS-7 geomagnetic field model and observed changes in the South Atlantic Anomaly. *Earth, Planets Space* 72, 156–231. doi:10.1186/s40623-020-01252-9

Fuselier, S. A., Lewis, W. S., Schiff, C., Ergun, R., Burch, J. L., Petrinec, S. M., et al. (2016). Magnetospheric multiscale science mission profile and operations. *Space Sci. Rev.* 199, 77–103. doi:10.1007/s11214-014-0087-x

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning.* Cambridge: MIT press.

Hastie, T., Friedman, J., and Tibshirani, R. (2009). *The elements of statistical learning: data mining, inference, and prediction,* 2. New York: Springer.

Hausdorff, F. (1921a). Summationsmethoden und momentfolgen. I. *Math. Z.* 9, 74–109. doi:10.1007/bf01378337

Hausdorff, F. (1921b). Summationsmethoden und momentfolgen. II. *Math. Z.* 9, 280–299. doi:10.1007/bf01279032

Heynderickx, D. (1996). Comparison between methods to compensate for the secular motion of the South Atlantic Anomaly. *Radiat. Meas.* 263, 369–373. doi:10.1016/1350-4487(96)00056-x

Kingma, D. P., and Jimmy, B. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980.*

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

Pollock, C., Moore, T., Saito, Y., Aung, S., Giles, B., Omoto, T., et al. (2016). Fast plasma investigation for magnetospheric multiscale. *Space Sci. Rev.* 199, 331–406. doi:10.1007/s11214-016-0245-4

Reme, H., Cros, A., Martz, C., Romefort, D., Perrier, H., Scholer, M., et al. (1997). "The Cluster ion spectrometry (CIS) experiment," in *The cluster and phoenix missions* (Germany: Springer), 303–350.

Ringnér, M. (2008). What is principal component analysis? *Nat. Biotechnol.* 26, 303–304. doi:10.1038/nbt0308-303

Shohat, J. A., and David Tamarkin, J. (1950). *The problem of moments,* 1. United States: American Mathematical Society RI.

Shuster, J. R., Gershman, D. J., Chen, L., Wang, S., Bessho, N., Dorelli, J. C., et al. (2019). MMS measurements of the Vlasov equation: Probing the electron pressure divergence within thin current sheets. *Geophys. Res. Lett.* 46, 7862–7872. doi:10.1029/2019gl083549

Shuster, J. R., Gershman, D. J., Dorelli, J. C., Giles, B. L., Wang, S., Bessho, N., et al. (2021). Structures in the terms of the Vlasov equation observed at Earth's magnetopause. *Nat. Phys.* 17, 1056–1065. doi:10.1038/s41567-021-01280-6

Stenborg, G., and Cobelli, P. J. (2003). A wavelet packets equalization technique to reveal the multiple spatial-scale nature of coronal structures. *Astronomy Astrophysics* 398, 1185–1193. doi:10.1051/0004-6361:20021687

Stenborg, G., Vourlidas, A., and Howard, R. A. (2008). A fresh view of the extreme-ultraviolet corona from the application of a new image-processing technique. *Astrophysical J.* 674 (2), 1201–1206. doi:10.1086/525556

Viñas, A. F., and Gurgiolo, C. (2009). Spherical harmonic analysis of particle velocity distribution function: comparison of moments and anisotropies using Cluster data. *J. Geophys. Res. Space Phys.* 114, A1. doi:10.1029/2008ja013633

Yeh, P-S., Armbruster, P., Kiely, P., Moury, G., Thiebaut, C., Masschelein, B., et al. (2005). "The new CCSDS image compression recommendation," in 2005 IEEE Aerospace Conference, Big Sky, MT, USA, 05-12 March 2005 (IEEE).