A Generative Atlas of the Earth's Magnetosphere

Stefano Markidis, Jonah Ekelund, and Luca Pennati

KTH Royal Institute of Technology, Stockholm, Sweden

Abstract. We propose a methodology to identify and characterize different plasma environments observed by magnetospheric mission spacecraft instruments, creating a data-driven magnetosphere atlas. To enable an accurate analysis under highly non-uniform sampling, we design a 3D adaptive octree data structure that refines magnetospheric regions with high sampling counts. A variational autoencoder that leverages the octree structure allows us to automatically identify different regions of the Earth's magnetosphere. The model also generates per-region feature distributions, improving interpretability by summarizing both typical conditions and variability. We apply this methodology to 10 years of plasma observations from NASA's Magnetospheric Multiscale (MMS) mission, identify known magnetospheric regions, and create an atlas of plasma environments in space. The resulting atlas can be used to better understand magnetospheric plasma regions, detect anomalies, automatically generate labels for space applications, and study the magnetospheric response across geomagnetic activities.

Keywords: Earth magnetosphere · atlas · unsupervised learning · VQ-VAE · generative models · adaptive grids

1 Introduction

The Earth's magnetosphere is a prime example of a complex system. Driven by the interaction between the supersonic solar wind and Earth's dipole magnetic field, it contains a wide range of plasma environments with distinct physical properties [12]. As the solar wind encounters the geomagnetic field, it is decelerated and heated at the bow shock, forming the magnetosheath downstream. Closer to Earth, the magnetopause separates shocked solar-wind plasma from magnetospheric plasma and hosts boundary layers where the two populations mix. On the nightside, stretched field lines form the magnetotail, with a hot, relatively dense plasma sheet embedded between low-density lobes. All these magnetospheric regions are separated by transitions and boundary layers and exhibit characteristic differences in density, flow speed, temperature, and magnetic field. A magnetospheric atlas aims to identifying these regions, determines their spatial extent and shapes, and maps them to their typical plasma signatures [12].

Most previous characterizations of magnetospheric plasma regions have relied on global MHD simulations and theory. These studies provide complete spatial coverage and controlled conditions [28,20,15]. In contrast, we adopt a purely

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_31

data-driven approach that leverages the large in situ dataset from NASA's Magnetospheric Multiscale (MMS) mission [2]. We define a methodology that makes data collected by space missions amenable to learning by introducing a high-dimensional state vector that summarizes average values and variability of key plasma quantities within a time window. Using this representation, we assemble a decade of MMS observations into a dataset for unsupervised regime discovery and atlas construction.

A critical challenge in using spacecraft measurements for atlas building is the strong non-uniformity of sampling. In fact, observations are concentrated along orbital trajectories and shaped by mission science priorities, e.g., a mission can be designed to study only a particular region of space. For example, during MMS's Phase 1 [2], the orbit strategy and operations focused on the dayside and magnetopause crossings, resulting in dense sampling near expected magnetopause locations and sparse coverage elsewhere. To address this imbalance, we introduce an adaptive 3D octree data structure [16] that refines the spatial discretization where measurement counts are high while keeping sparsely sampled regions coarse.

To identify distinct plasma environments, we use an autoencoder-based approach that captures the data's probabilistic structure and partitions observations into a fixed number of classes [11,18]. We further adapt the autoencoder to exploit the octree layout through an octree-aware sampling strategy that reduces spatial sampling bias during training. Each identified plasma environment is characterized by a distribution of underlying plasma features. The resulting atlas is generative in the sense that, for a given location (or set of conditioning variables), it can produce a representative distribution for comparison with observations. This enables tasks such as anomaly detection and automated labeling. The main contributions of this work are the following:

- We establish a methodology for analyzing data from space missions and apply it to approximately 660k MMS state windows spanning 10 years of observations.
- To represent spatial context under highly non-uniform orbital sampling, we introduce a 3D adaptive octree that refines regions of dense coverage while keeping sparsely sampled regions coarse.
- We learn a discrete, unsupervised atlas of magnetospheric regimes with a variational autoencoder, without expert thresholds or hand-labeled regions.
- We provide summaries of the distribution functions that characterize each identified magnetospheric region to support physical interpretation.
- We quantify how geomagnetic activity shifts the different magnetospheric regions by comparing quiet and storm times.

2 Methodology

Our methodology has four phases: (1) construction of fixed-duration MMS1 plasma states, (2) adaptive spatial discretization with a 3D octree, (3) unsupervised regime learning with an octree-balanced VQ-VAE, and (4) robustness and geomagnetic-activity analysis.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_31

Table 1. Summary of the MMS1 state dataset and adaptive octree support used for atlas construction. Each state is represented by the mean ($\bar{\cdot}$) and standard deviation (σ) of the listed variables within a 2-min window.

Item	Value
Total states	660,340
Time span (UTC)	2015-09-20 to 2025-10-30
Unique days	2,614
State windowing	$\Delta t = 2$ min, cadence 10 s, stride 1 min (50% overlap)
Spatial extent (R_E)	$x \in [-28.37, 28.48]$, $y \in [-28.46, 28.70]$, $z \in [-20.38, 10.38]$
Quiet-time fraction	86.2% with $\lvert Dst\rvert \leq 20$ nT
Features ($d = 14$)	$\bar{x}$ and $\sigma(x)$ for $x \in \{\lvert B\rvert, B_z, n_i, T_i, \lvert V_i\rvert, V_{i,x}, \beta_i^*\}$
Octree support	Adaptive 3D octree on median window position; base cell size $2\,R_E$, refine if > 25 states/cell; 65,352 leaf cells

We construct an Earth magnetospheric atlas from observations by NASA's Magnetospheric Multiscale (MMS) mission, a four-spacecraft constellation designed for high-resolution plasma and field measurements [2]. We use survey-resolution magnetic-field vectors from the Fluxgate Magnetometer (FGM) [25], ion density, bulk velocity, and temperature moments from FPI/DIS [21], and spacecraft position in geocentric solar magnetospheric (GSM) coordinates when available. Our analysis targets ion-scale plasma conditions that are broadly comparable across magnetospheric environments while remaining computationally tractable at mission scale. Because MMS spacecraft separations are typically smaller than ion kinetic scales for much of the mission [2], the four spacecraft do not provide independent ion-scale samples. We therefore use MMS1 data throughout this study.

Phase 1: Data preparation and state construction. Our atlas is built from fixed-duration *states* that summarize local plasma conditions over short windows. Starting from time-tagged MMS1 measurements of magnetic field $\mathbf{B}(t)$, ion number density $n_i(t)$, ion temperature $T_i(t)$, and ion bulk velocity $\mathbf{V}_i(t)$, we first resample the native Level-2 survey products onto a common 10 s grid using the median within each bin. We then segment the resampled multivariate time series into windows of duration $\Delta t = 2$ min with a stride of 1 min, leading to $N = 12$ samples per window and 50% overlap between consecutive windows. This choice captures mesoscale variability and common magnetospheric transitions while keeping the plasma conditions within each window approximately stationary for regime identification.

Each window is represented by a 14-dimensional feature vector containing the mean and standard deviation of seven physically motivated variables: $|B|$, B_z, n_i, T_i, $|V_i|$, $V_{i,x}$, and β_i^*. Here, β_i^* is an ion plasma-β proxy computed from the ion moments and magnetic-field magnitude; it helps distinguish high-β regions such as the magnetosheath and plasma sheet from low-β regions such as the lobes. We also store the median spacecraft position (x, y, z) within each window, in units of R_E, for spatial binning and bias mitigation, but we do not use position as a model input. Table 1 summarizes the resulting dataset. Before training, we remove fill values, discard windows with insufficient valid samples, and standardize features

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_31

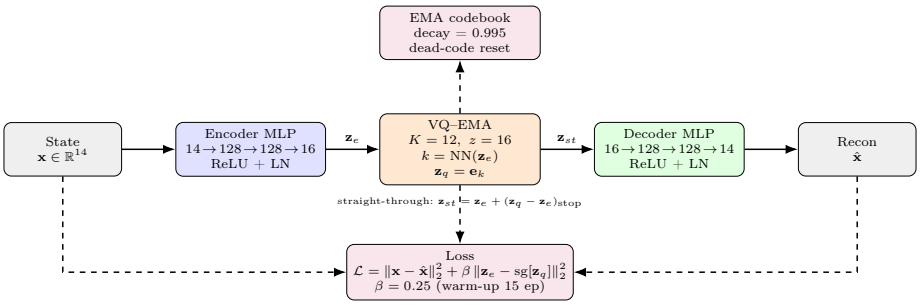

Fig. 1. Compact VQ-VAE used for the atlas (hidden=128, LayerNorm, z=16, K=12 with EMA updates).

using statistics from the training split. A fixed training/validation split is used to monitor generalization and select the atlas checkpoint.

Phase 2: Adaptive octree spatial support. To represent regime occurrence without imposing an arbitrary uniform Cartesian mesh, and to mitigate strong spatial sampling bias, we build an adaptive 3D octree over the state positions [16]. We initialize a base grid with cubic cells of side length $2\,R_E$ and recursively subdivide any cell containing more than 25 states into eight children. Refinement continues until all leaf cells satisfy the occupancy threshold or a minimum cell size is reached. This produces a data-adaptive spatial discretization that is fine where MMS sampling is dense and coarse where it is sparse, improving the robustness of cell-level summaries. In our dataset, the octree contains 65,352 leaf cells.

We also use the octree during training to reduce the influence of densely sampled orbital tracks. Let n_c denote the number of states in leaf cell c, and let $c(i)$ be the cell containing state i. We define per-sample weights $w_i \propto n_{c(i)}^{-\alpha}, \alpha \in [0, 1]$, so that densely sampled cells are downweighted and sparsely sampled cells are upweighted. Mini-batches are drawn with a weighted sampler proportional to w_i. In this work, we set $\alpha = 0.5$ and normalize the weights so that $w_{\max} = 1$.

Phase 3: VQ-VAE regime learning. We learn discrete plasma regimes with a vector-quantized variational autoencoder (VQ-VAE) [18]. The architecture is summarized in Figure 1. A key hyperparameter is the codebook size K. We set $K = 12$ to obtain an atlas that remains interpretable and comparable across runs: smaller values merge physically distinct environments into overly broad categories, whereas larger values fragment the atlas into rare codes that are difficult to interpret.The codebook is updated using exponential moving averages (EMAs) with decay 0.995, and we enable a dead-code reset mechanism to reduce the risk of unused codes. Gradients are passed through the quantization step with the straight-through estimator.

Training minimizes a reconstruction-plus-commitment objective,

$$\mathcal{L} = \mathcal{L}_{\mathrm{rec}} + \beta\,\mathcal{L}_{\mathrm{com}}, \qquad \mathcal{L}_{\mathrm{rec}} = \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2, \qquad \mathcal{L}_{\mathrm{com}} = \|\mathbf{z}_e - \mathrm{sg}[\mathbf{z}_q]\|_2^2, \tag{1}$$

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_31

where sg[·] denotes stop-gradient. The reconstruction term encourages the discrete latent code to preserve the information needed to reproduce the observed plasma features, while the commitment term stabilizes the encoder-codebook assignments. We train for 25 epochs with the Adam optimizer (learning rate 10^{-3}), a batch size of 1024, and a linear warm-up of β to 0.25 over the first 15 epochs. During training, we monitor reconstruction and commitment losses on a held-out validation set and track the effective perplexity $\exp(H(p(k)))$ of the code-assignment distribution to detect codebook collapse.

Phase 4: Robustness and geomagnetic-activity analysis. Because unsupervised discrete partitions can vary with random initialization, we assess robustness across multiple seeds. Regime identities are aligned across runs using a Hungarian assignment on a validation subset, and agreement is quantified with the adjusted Rand index (ARI) [9].

To quantify how geomagnetic activity redistributes the learned regimes, each window i is assigned a regime $k \in \{0, \ldots, 11\}$. We define quiet samples $\mathcal{Q} = \{i : |Dst_i| \leq 20 \text{ nT}\}$ and storm samples $\mathcal{S} = \{i : \ Dst_i \leq -50 \text{ nT}\}$, where Dst is the disturbance storm time index [14]. Because MMS sampling is highly nonuniform in (x, y, z), we estimate occupancies with a spatially stratified octree scheme based on the window-centered spacecraft positions $\mathbf{r}_i = (x_i, y_i, z_i)$. Starting from a root bounding box covering the sampled domain, we recursively subdivide any occupied cell into eight equal subcells until either *(i)* the cell edge length falls below a prescribed minimum resolution $\ell_{\min}$ or *(ii)* the cell contains fewer than n_{split} samples. The resulting leaf cells form a partition $\mathcal{C}_{\text{leaf}}$.

For each leaf cell $c \in \mathcal{C}_{\text{leaf}}$, we compute regime counts separately for $\mathcal{Q}$ and $\mathcal{S}$ and form within-cell regime frequencies. We then average these frequencies across cells that are sufficiently supported in both subsets. If

$$C = \{c \in \mathcal{C}_{\text{leaf}} : \ n^{\mathcal{Q}}(c) \geq n_{\min}, \ n^{\mathcal{S}}(c) \geq n_{\min}\},$$

then

$$\Delta p_k = 100 \left(\frac{1}{|C|} \sum_{c \in C} \frac{n_k^{\mathcal{S}}(c)}{n^{\mathcal{S}}(c)} - \frac{1}{|C|} \sum_{c \in C} \frac{n_k^{\mathcal{Q}}(c)}{n^{\mathcal{Q}}(c)} \right). \tag{2}$$

Here, $n_k^X(c)$ is the number of samples assigned to regime k in leaf cell c for subset $X \in \{\mathcal{Q}, \mathcal{S}\}$, and $n^X(c)$ is the total number of samples from subset X in that cell. We report Δp_k in percentage points and, when needed, separately for dayside ($x_c \geq 0$) and nightside ($x_c < 0$) cells using the GSM x coordinate of the leaf-cell center $\mathbf{r}_c$.

3 Results

We begin by assessing training dynamics and robustness to random initialization, then present the learned spatial atlas and per-regime feature-distribution summaries, and finally quantify storm-time redistribution of regime occupancy. Figure 2 shows a rapid reduction in both training and validation loss during the first few epochs, followed by a stable plateau, indicating effective learning with

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_31

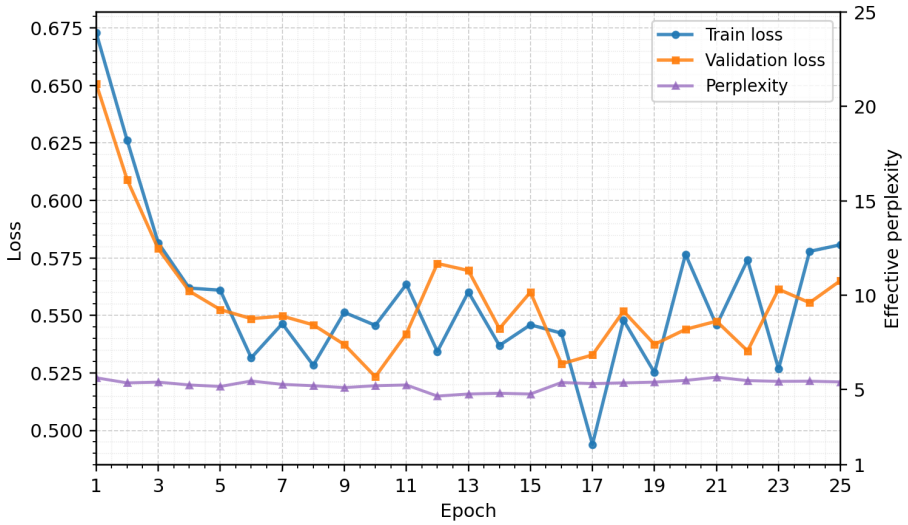

Fig. 2. Training and validation loss (left axis) and effective code-assignment perplexity (right axis) versus epoch for the octree-balanced VQ-VAE (25 epochs).

a limited generalization gap. The lowest validation loss is reached around epoch ~ 10, after which the validation curve varies only modestly, consistent with convergence rather than progressive overfitting. A key failure mode in discrete-latent models is codebook collapse, where only a few regimes are used. We therefore track code utilization using the effective perplexity of the assignment distribution, see purple line in Figure 2. Perplexity remains near 5-6 throughout training, indicating effective use of multiple codes and avoiding degeneration into a trivial single-regime solution.

Because unsupervised discrete partitions can vary with random initialization, we explicitly assess robustness by retraining the model with five different random seeds and comparing the resulting assignments on the full dataset. Across the ten seed pairs, the ARI is 0.708 ± 0.043, indicating agreement in the recovered partition. After aligning code identities with a Hungarian assignment, the mean fraction of identically labeled states (relative to a reference run) is 0.729 ± 0.041. The learned regimes themselves are also consistent across runs. The cosine similarity between aligned prototype feature vectors is 0.956 ± 0.046, showing that both assignments and per-regime characteristics are stable.

As a second step, we project the learned discrete regimes back into physical space to form a data-driven magnetospheric atlas. We visualize this atlas as maps of the dominant regime within each adaptive octree cell. Figure 3 shows an equatorial (XY) slice. Each cell aggregates all windows whose median positions fall within that volume, and the color indicates the modal (most frequent) regime label. Black outlines mark octree leaf-cell boundaries, with finer cells appearing where MMS sampling is densest. Figure 4 complements the equatorial view with

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_31

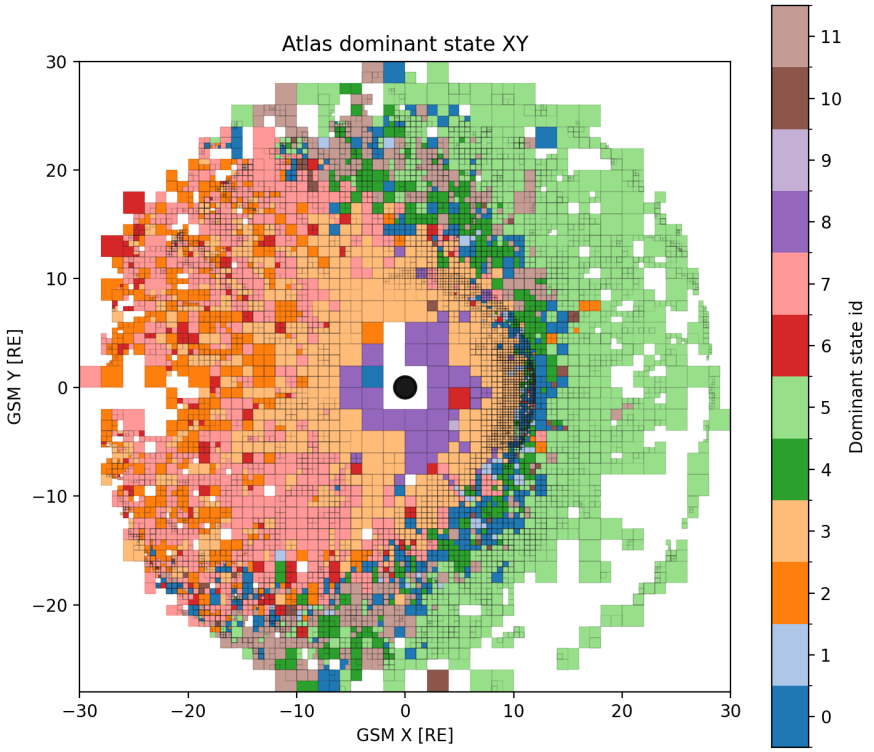

Fig. 3. Magnetosphere equatorial plane atlas map. Colors denote the most frequent discrete regime (state id 0-11). Black outlines indicate the octree cell boundaries; finer cells appear where MMS sampling is dense. Earth is shown at the origin.

meridional (XZ) and flank (YZ) slices, highlighting vertical structure and dawn-dusk asymmetries in the learned partition.

To interpret the learned atlas states beyond their spatial maps, we summarize each state by its feature distributions. Figure 5 shows per-state distribution functions for selected variables, where the violin shapes capture the full spread of conditions within each state (including possible multi-modality). The location panels indicate where each state is most frequently observed, while density and temperature separate low-density/hot populations from denser/cooler plasma, and $V_{i,x}$ highlights strongly convecting intervals.

To connect the data-driven atlas states to familiar magnetospheric regions, we summarize each state with simple, physically interpretable statistics. Table 2 reports an overview of the 12 discrete atlas states learned by the VQ-VAE. Each state arises from the unsupervised assignment of two-minute windows in feature space. They do not include region labels, boundary definitions, or the expert thresholds used during training. For interpretation, we aggregate all win-

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_31

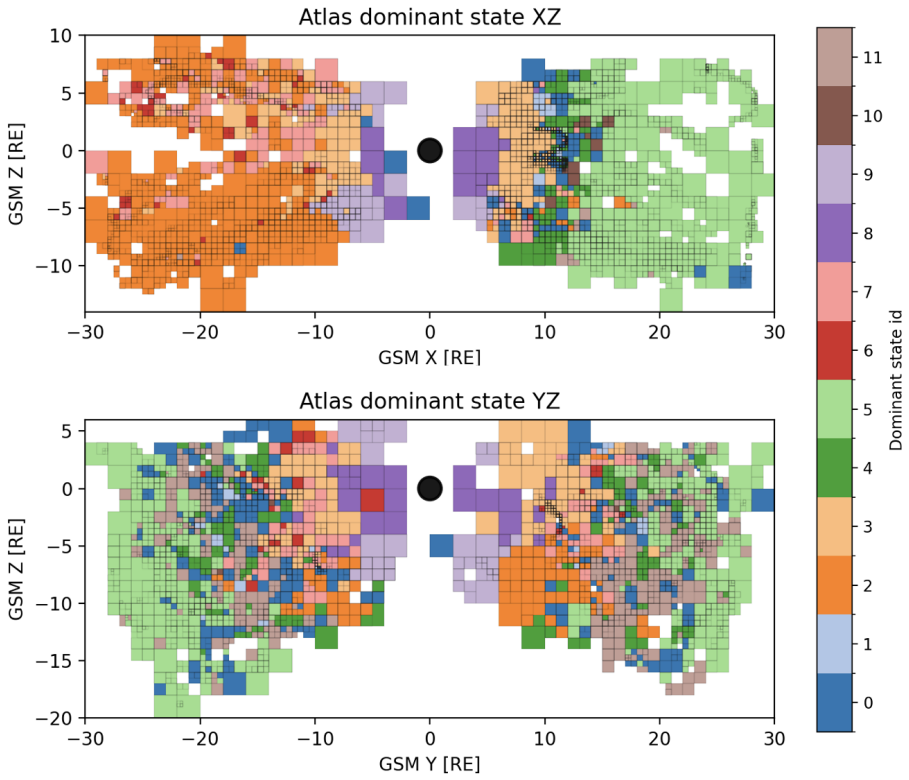

Fig. 4. Magnetosphere meridional (top) and flank (bottom) planes. Each cell is colored by the most frequent discrete regime (state id 0-11)

dows assigned to each state and report median values for key quantities that traditionally separate magnetospheric environments: plasma density and temperature, bulk flow along the Sun-Earth line, and magnetic-field magnitude. We then relate these summaries to established qualitative signatures of typical regions and to the spatial occurrence patterns visible in the atlas maps. States with higher n_i and comparatively lower T_i align with magnetosheath-like conditions, whereas low-n_i and multi-keV T_i states align with plasma-sheet-like populations. State 9 is distinctive, with very large $|\mathbf{B}|$ and very low density, consistent with a strong-field, low-β environment whose spatial context distinguishes lobe-like from near-Earth dipolar intervals.

Finally, we use the atlas to quantify how geomagnetic activity redistributes the prevalence of the learned states. Figure 6 shows the storm-quiet change in state occupancy, $\Delta p_k = p_k^{\text{storm}} - p_k^{\text{quiet}}$, computed with octree stratification to reduce spatial sampling bias and reported separately for dayside ($x_c \geq 0$) and nightside ($x_c < 0$) cells. Interpreting the states using Table 2, the storm-time response has a clear physical signature. It is dominated by nightside reconfig-

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_31

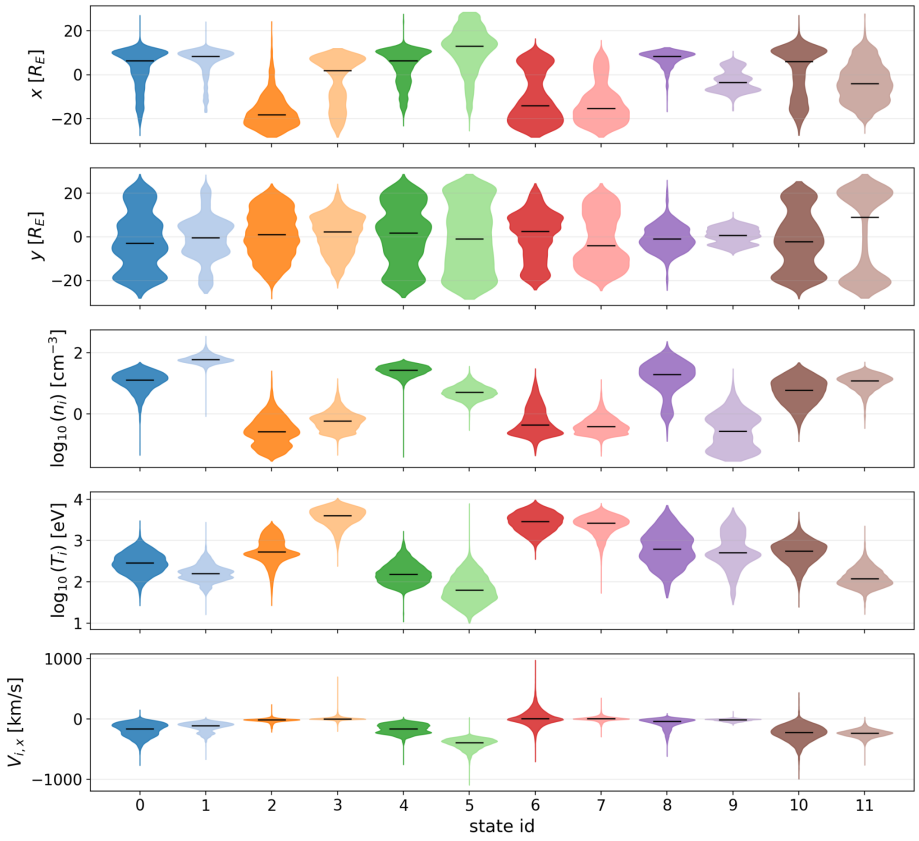

Fig. 5. Per-state feature distributions for spatial location and key plasma moments. Each violin shows the distribution of two-minute states assigned to each discrete regime (state ids 0-11). Rows show (top to bottom) GSM x and y position in R_E, $\log_{10}(n_i)$ in cm^{-3}, $\log_{10}(T_i)$ in eV, and ion bulk flow $V_{i,x}$ in km/s. Black horizontal markers indicate the median for each state.

uration, consistent with the magnetotail acting as the primary reservoir and release region for storm-time energy storage and unloading [24]. On the nightside, the occupancy of the *tail plasma sheet* category (state 7) drops strongly, while lobe-like and more strongly magnetized conditions become more common (state 9, characterized by very large $|\mathbf{B}|$ and low β). This pattern is consistent with plasma-sheet thinning and enhanced lobe magnetic pressure during disturbed intervals, together with increased intermittency and mixing between adjacent states. At the same time, the nightside distribution shifts toward hotter, more energetic plasma populations (e.g., state 3, interpreted as the hot plasma sheet/inner hot magnetosphere), compatible with storm-time injections and dipolarization-associated heating near Earth [7]. In contrast, the dayside changes are smaller and are more naturally interpreted in terms of large-scale

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_31

Table 2. Interpretability table for the learned atlas states. Tilde denotes the median value. Background colors provide an at-a-glance cue: $\tilde{x} > 0$ (dayside: orange) vs $\tilde{x} < 0$ (tailward: blue), and low-to-high gradients for $\tilde{n}_i$, $\tilde{T}_i$, $|\tilde{V}_{i,x}|$, and $|\tilde{\mathbf{B}}|$. The final column gives a physical interpretation by comparing these medians with typical magnetospheric condition ranges.

State	$\tilde{x}$ [R_E]	$\tilde{n}_i$ [cm^{-3}]	$\tilde{T}_i$ [eV]	$\lvert\tilde{V}_{i,x}\rvert$ [km s^{-1}]	$\lvert\tilde{\mathbf{B}}\rvert$ [nT]	Region Interpretation
0	6.4	13.0	286	163	19.7	MSH (hot/processed; boundary-adjacent)
1	8.3	59.9	160	109	24.5	Dense MSH (shocked solar wind)
2	-18.2	0.3	528	15	20.3	PSBL / lobe-adjacent tail transition
3	1.9	0.6	4056	5	34.9	Hot plasma sheet / inner hot magnetosphere
4	6.3	27.2	154	159	19.3	MSH
5	13.1	5.1	63	392	5.6	Solar wind
6	-14.0	0.4	2896	8	20.5	Plasma sheet (dynamic/bursty)
7	-15.3	0.4	2637	5	13.4	Tail plasma sheet
8	8.5	19.6	623	41	41.5	Compressed/stagnation sheath near dayside boundary (mixing)
9	-3.4	0.3	510	14	100.2	Strong-field, very low-β environment (lobe-like / near-Earth dipolar high-latitude)
10	6.0	6.0	561	220	15.9	Magnetopause boundary layer / hot sheath
11	-3.9	12.2	119	235	14.2	Flank MSH (downstream)

compression and boundary motion. The relative prevalence of *solar-wind-like* conditions (state 5) and downstream/flank magnetosheath (state 11) increases at fixed spatial bins. The boundary-adjacent magnetosheath/boundary-layer categories (e.g., states 0 and 10) decrease, consistent with an inward displacement of the magnetopause and bow shock under enhanced solar-wind driving [13]. An asymmetry is state 11, which increases on the dayside but decreases on the nightside, suggesting that storm forcing redistributes flank/sheath-like conditions across local time rather than producing a uniform global change. Overall, the Dst-conditioned occupancy shifts recovered by the atlas align with the expected storm-time picture. Compression on the dayside and a stronger, more reorganized tail system featuring reduced *quiet* plasma-sheet occupancy and increased strong-field/low-β and energized states.

4 Related Work

Earth Magnetosphere Region Identification. Region identification is foundational in magnetospheric science and operations. Magnetospheric regions and boundaries are motivated by first-principles reasoning and supported by empirical boundary models for the magnetopause and bow shock [26,6], as well as data-based magnetic field models that provide large-scale context and mapping [30]. Recent work has increasingly explored automated and machine-learning-based region identification. Supervised deep learning has been used to classify MMS dayside plasma regions directly from particle distributions [17,4,5], and recurrent models have been applied to boundary-crossing detection [1]. Unsupervised

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_31

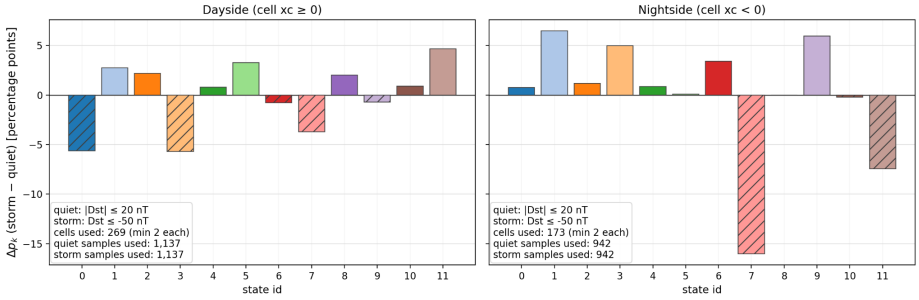

Fig. 6. Storm-quiet change in state occupancy Δp_k estimated with octree stratification, shown separately for dayside ($x_c \geq 0$) and nightside ($x_c < 0$) cells. Positive values indicate increased storm-time prevalence.

methods have also been developed to cluster spacecraft observations into broad regions using combinations of dimensionality reduction, self-organizing maps, and hierarchical clustering [10,3,29].

Generative Methods for Atlas Creation. Generative representation learning provides a way to compress high-dimensional data into reusable latent variables. Variational autoencoders (VAEs) learn probabilistic latent spaces that balance reconstruction fidelity with regularization, enabling sampling and uncertainty-aware downstream use [11]. Discrete latent models such as VQ-VAE replace continuous latents with a learned codebook of prototypes, producing a finite vocabulary of learned symbols that can be mapped, counted, and compared [18]. This discrete, generative structure is particularly well suited to atlas construction because it couples regime discovery with a compact generative description of each regime, supporting both classification-like indexing and distributional modeling.

5 Discussion and Conclusions

In this work, we presented a methodology for creating an atlas of magnetospheric regions characterized by different plasma parameters. Given the spatial imbalance, we introduce an adaptive strategy based on a 3D octree to divide the magnetosphere into regions, ensuring that each numerical cell approximately contains the same number of samples. We used this approach in training a VQ-VAE to classify different regions of the magnetosphere. We inspected the distribution function to interpret and map these regions. We used MMS1 data, and the dataset can be further augmented by also using data from the other spacecrafts, MMS2-MMS4. This would add samples only to the same point in space since the distance between the spacecrafts is relatively small. To expand the dataset for the magnetosphere atlas, one must rely on additional magnetospheric missions, such as CLUSTER and THEMIS. However, one challenge is to convert all measurements to a consistent reference system, e.g., GSM vs. GSE.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_31

The atlas is *generative* in the sense that it defines conditional probability distributions over plasma parameters. Practical applications of such a generative atlas include:

- Automatic labeling of MMS data. The VQ-VA can assign probabilities and discrete labels to observations via inference. These labels can be ingested into a knowledge-graph database to support scientific discovery, querying, and hypothesis testing. The automatic labeling can also support scientist-in-the-loop analysis.
- Anomaly detection. It is possible to identify unusual observation intervals by flagging observations with low likelihood under the atlas (so-called *out-of-distribution samples*). This can lead to the discovery of rare events or potential data-quality issues.
- Data-driven initialization of MHD simulations. The atlas can provide ensemble initial conditions by sampling from distributions. For example, conditioned on an upstream solar-wind and magnetic field states, the atlas can generate physically consistent distributions for $(n, \mathbf{v}, T)$ to initialize an MHD run.
- Gap filling/virtual diagnostics. The atlas allows us to infer distributions for missing parameters or unavailable instruments, conditioned on the available measurements. For example, when ion moments are missing or unreliable, the atlas can provide them as samples from probabilistic distributions.
- Data assimilation priors and regularization. We can use atlas-conditioned distributions as informative priors or regularizers in inversion and assimilation tasks, constraining solutions to physically consistent regions of parameter space. For example, in reconstructing local current density or pressure from multi-point magnetic field measurements, one can penalize solutions that fall outside the atlas-supported distribution for the inferred plasma environment.

A few directions for future work follow naturally from this study. First, the atlas can be extended from static, windowed states to an explicitly time-aware formulation by modeling sequences of discrete VQ codes with a probabilistic sequence model (e.g., a Hidden Markov Model) to estimate transition probabilities and identify typical evolution paths between regions, such as repeated crossings of the bow shock or magnetopause [22]. Second, incorporating additional context variables, such as upstream solar wind drivers, IMF orientation, and geomagnetic indices, would allow for richer conditioning and enable controlled comparisons. Third, cross-mission and cross-planet domain adaptation is an important next step. Learning shared discrete vocabularies and calibration-invariant features would enable the transfer of regime definitions between missions (MMS, Cluster, THEMIS) and other planetary magnetospheres (MAVEN, Juno, Cassini), while accounting for differences in instrumentation and sampling biases.

An additional direction for future work is to investigate whether alternative generative models can improve fidelity, calibration, and interpretability relative to the current VQ-VAE formulation. While VQ-VAE provides a codebook of

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_31

plasma regions that is convenient for labeling and atlas construction, other families of generative models may capture sharper distributions or more complex, multi-modal structures. For example, normalizing flows could provide exact likelihoods and flexible conditional densities for continuous parameters [23,19]. Additional diffusion models could be explored for high-dimensional representations (e.g., joint distributions over multi-instrument features) where accurate sample quality is critical [27,8].

Acknowledgments

We acknowledge the MMS mission team and the instrument teams for producing the MMS data products used in this study [2,25,21]. MMS data are publicly available via NASA's MMS Science Data Center. This work is supported by the European Union's HORIZON RIA via ASAP - Automatics in Space (`asap-space.eu`) under the grant agreement no. 101082633.

References

1. Argall, M.R., et al.: MMS SITL ground loop: Automating the burst data selection process. Frontiers in astronomy and space sciences **7**, 54 (2020)
2. Burch, J.L., Moore, T.E., Torbert, R.B., Giles, B.L.: Magnetospheric multiscale overview and science objectives. Space Science Reviews **199**(1–4), 5–21 (2016)
3. Edmond, T., et al.: Clustering of global magnetospheric observations. Journal of Geophysical Research: Machine Learning and Computation (2024)
4. Ekelund, J., et al.: AI in space for scientific missions: Strategies for minimizing neural-network model upload. In: 2024 IEEE 20th International Conference on e-Science (e-Science). pp. 1–10. IEEE Computer Society (2024)
5. Ekelund, J., et al.: Adaptive PCA-based outlier detection for multi-feature time series in space missions. In: International Conference on Computational Science. pp. 253–267. Springer (2025)
6. Farris, M.H., Russell, C.T.: Determining the standoff distance of the bow shock: Mach number dependence and use of models. Journal of Geophysical Research: Space Physics **99**(A9), 17681–17689 (1994)
7. Forsyth, C., et al.: Increases in plasma sheet temperature with solar wind driving during substorm growth phases. Geophysical Research Letters **41**, 8713–8721 (2014)
8. Ho, J., et al.: Denoising diffusion probabilistic models. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) NeurIPS 2020. pp. 6840–6851 (2020)
9. Hubert, L., Arabie, P.: Comparing partitions. Journal of Classification **2**(1), 193–218 (1985)
10. Innocenti, M.E., et al.: Unsupervised classification of simulated magnetospheric regions. Annales Geophysicae **39**, 861–878 (2021)
11. Kingma, D.P., Welling, M.: Auto-encoding variational Bayes. In: ICLR (2014)
12. Kivelson, M.G., Russell, C.T. (eds.): Introduction to Space Physics. Cambridge University Press, Cambridge, UK (1995)
13. Kumar, S., Pulkkinen, T.I.: Statistical analysis of magnetopause response during substorm phases. Annales Geophysicae **43**, 137–149 (2025)

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_31

14. Markidis, S., et al.: Discovering governing equations of geomagnetic storm dynamics with symbolic regression. In: International Conference on Computational Science. pp. 122–136. Springer (2025)
15. Markidis, S., et al.: Exascale implicit kinetic plasma simulations on el capitan for solving the micro-macro coupling in magnetospheric physics. arXiv preprint arXiv:2507.20719 (2025)
16. Meagher, D.: Geometric modeling using octree encoding. Computer Graphics and Image Processing **19**(2), 129–147 (1982)
17. Olshevsky, V., et al.: Automated classification of plasma regions using 3D particle energy distributions. Journal of Geophysical Research: Space Physics **126**(10), e2021JA029620 (2021)
18. van den Oord, A., Vinyals, O., Kavukcuoglu, K.: Neural discrete representation learning. In: NeurIPS 2017. pp. 6306–6315 (2017)
19. Papamakarios, G., et al.: Normalizing flows for probabilistic modeling and inference. Journal of Machine Learning Research **22**(57), 1–64 (2021)
20. Peng, I.B., et al.: The formation of a magnetosphere with implicit particle-in-cell simulations. Procedia Computer Science **51**, 1178–1187 (2015)
21. Pollock, C., et al.: Fast plasma investigation for magnetospheric multiscale. Space Science Reviews **199**(1–4), 331–406 (2016)
22. Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE **77**(2), 257–286 (1989)
23. Rezende, D., Mohamed, S.: Variational inference with normalizing flows. In: Proceedings of the 32nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 37, pp. 1530–1538 (2015)
24. Rostoker, G., et al.: The roles of direct input of energy from the solar wind and unloading of stored magnetotail energy in driving magnetospheric substorms. Space Science Reviews **46**, 93–111 (3 1988)
25. Russell, C.T., et al.: The magnetospheric multiscale magnetometers. Space Science Reviews **199**(1–4), 189–256 (2016)
26. Shue, J.H., et al.: Magnetopause location under extreme solar wind conditions. Journal of Geophysical Research: Space Physics **103**(A8), 17691–17700 (1998)
27. Sohl-Dickstein, J., et al.: Deep unsupervised learning using nonequilibrium thermodynamics. In: Proceedings of the 32nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 37, pp. 2256–2265. PMLR (2015)
28. Tóth, G., et al.: Space weather modeling framework: A new tool for the space science community. Journal of Geophysical Research: Space Physics **110**(A12), A12226 (2005)
29. Toy-Edens, V., et al.: Classifying 8 years of MMS dayside plasma regions via unsupervised machine learning. Journal of Geophysical Research: Space Physics **129**(6), e2024JA032431 (2024)
30. Tsyganenko, N.A.: A model of the near magnetosphere with a dawn-dusk asymmetry 1. Mathematical structure. Journal of Geophysical Research: Space Physics **107**(A8), 1179 (2002)

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_31