Development of a Layered Unsupervised Classifier of Plasma Regions and a Bootstrap-Ensemble Neural Network Bow Shock Model

BY

JAMES EDMOND

B.Sc. in Physics, Auburn University, 2017

DISSERTATION

Submitted to the University of New Hampshire
in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

in

Physics

September, 2024

This dissertation has been examined and approved in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Physics by:

Dissertation Director, Joachim Raeder
Professor of Physics

Matthew Argall
Research Assistant Professor of Physics

Banafsheh Ferdousi
Senior Research Physicist

Amy Keesee
Associate Professor of Physics

Marek Petrik
Associate Professor of Computer Science

On September, 2024

Original approval signatures are on file with the University of New Hampshire Graduate School.

Dedication

Life is a disco inferno, but it was made slightly easier thanks to the work of Alexandra Elbakyan.

Acknowledgments

I came, I saw, I was compelled to submit

Table of Contents

List of Tables

List of Figures

ABSTRACT

Development of a Layered Unsupervised Classifier of Plasma Regions and a Bootstrap-Ensemble Neural Network Bow Shock Model

by

James Edmond

University of New Hampshire, September, 2024

The use of supervised methods in space science have demonstrated powerful capability in classification tasks, but purely unsupervised methods have been less utilized for the classification of spacecraft observations. We use a combination of unsupervised methods, being principal component analysis, self-organizing maps, and hierarchical agglomerative clustering, to classify THEMIS and MMS observations as having occurred in the magnetosphere, magnetosheath, or the solar wind. The resulting classification are validated visually by analyzing the distribution of classifications and studying individual time series as well as by comparison to the labeled dataset of a previous model, against which ours has an accuracy of 99.4%. The model has a variety of applications beyond region classification such as deeper hierarchical analysis, magnetopause and bow shock crossing identification, and identification of bursty bulk flows, hot flow anomalies, and foreshock bubbles. Then using the bow shock crossings inferred from the previous model as well as the results of another machine learning model and an online bow shock crossing catalogue, we create a bow shock dataset of almost 16k crossings. An ensemble of neural networks are trained to predict the coefficients of a bow shock model function using traditional bow shock parameters as well as magnetic clock and cone angles. The small size of the dataset means that typical partitioning of the training set cannot reasonably be done, so the ensemble members are

bootstrap-trained. We show the ensemble to perform better than a single model trained on the full dataset. The model results show mixed agreement with previous observations and performs better than the Chao model for varying clock and cone angles and for nightside bow shock crossings, but slightly underperforms for dayside crossings.

Chapter 1

Introduction

1.1 General Overview of the Relevant Regions of the Magnetosphere

The Sun is a large celestial body at the center of the solar system that continually ejects some of its plasma as it rotates. Thermal pressure gradients continually accelerate this outgoing plasma (*Parker* [1958]) where it transitions from subsonic to supersonic speeds. It is in the latter regime thereafter that this plasma is referred to as the solar wind. This solar wind is generally bimodal in its speed and density distributions in that repeated measurements show a "slow" solar wind, with average speed and density of 400 km/s and 10 $\#/cc^3$, and a "fast" solar wind, with average speed and density 700 km/s and 5 $\#/cc^3$ (see chapter 5 of *C. T. Russell* [2016]).

The solar wind later arrives at Earth and is diverted due to its electromagnetic environment acting as a fluidic obstacle. In the GSE (Geocentric Solar Ecliptic) coordinate system, where the x-axis points from the Earth to the sun and the z axis is perpendicular to the ecliptic plane in which the Earth orbits the Sun, this diversion begins around $X =\sim 13.5$ Earth radii, or R_E (1 R_E = 6,371 km). This diverting causes the solar wind to pile up and become denser as the rate at which more plasma piles on exceeds the rate at which it can be diverted; this condensing both slows down and heats up the incident plasma, converting

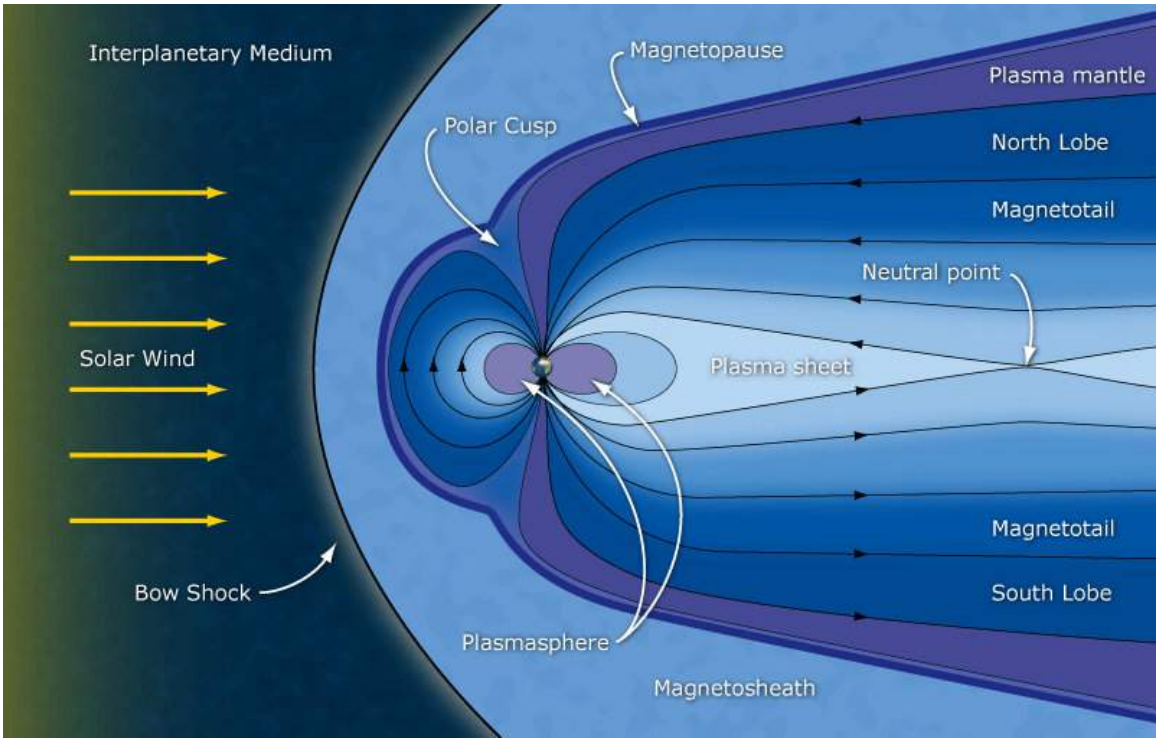

Figure 1-1: Depiction of a vertical cross section of the magnetosphere and its outer environs where the Earth is in the center of the image. Towards the left is the sun with the flow of solar wind from the Sun to Earth represented with yellow arrows. For the context of this thesis, all the distinct regions can be represented as one of three groups: the solar wind (the region of space to the left of the bow shock boundary), the magnetosphere (cumulatively representing all regions encompassed within the magnetopause boundary), and the magnetosheath as the light blue region in between. Image credit to ESA / C. T. Russell.

the supersonic flow to subsonic flow and much incident kinetic energy into thermal energy (see chapter 10 of *Parks* [2004]). The boundary that separates the denser, hotter plasma from the incident solar wind is called the bow shock, and the area that encompasses this "shocked" plasma is called the magnetosheath. Further in towards Earth is the magnetosphere and the boundary it shares with the magnetosheath is called the magnetopause, which acts as an equilibrium between the pressure of the incoming solar wind and the magnetic pressure of the magnetosphere. The magnetosphere encompasses multiple populations of varying density and energy, but is often low in density ($0.1\ \#/cc^3 \leq n_{ion} \leq 1\ \#/cc^3$) and high in energy ($T_{ion} \geq 1$ keV). These regions are depicted in figure 1-1.

The phrase "bow shock" stems from comparisons with a ship (specifically the front, or

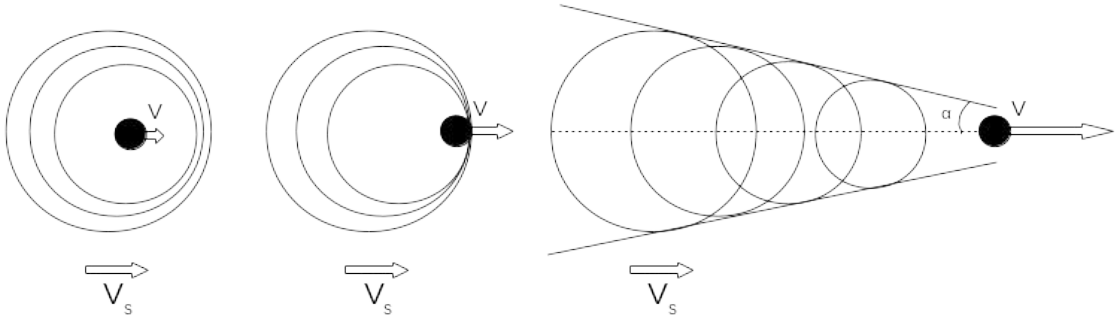

Figure 1-2: Shock steepening depicted with diagrams. The solid black circle is the source of waves moving at speed V and the radially propagating waves are represented with the outer circles which move at speed V_S. The left image shows the wave source moving at V $< V_S$. The middle image shows the source moving at V $= V_S$, which causes the waves to steepen along the direction of source movement. The right image shows the source moving at speed V $> V_S$, causing the waves to steepen instead at an angle α and forming a Mach cone of the same angle.

"bow" of the ship) crossing water. In general, a "shock" is a separation between two fluidic regions: an upstream (fluid which is flowing towards the shock and has yet to experience its effects) and a downstream (fluid which has already passed the shock). Classically, quantities of this downstream fluid can be predicted relative to parameters of the upstream fluid and are described by the Rankine-Hugoniot (RH) relations by assuming a conservation of mass, momentum, and energy. A relatively simple way to describe movement through a fluid progressing to that of a shock can be seen in the form of a wave source and its movement through a fluid, as seen in Figure 1-2. A physical example of the contours of a shock created around a hemispherical object can be seen in Figure 1-3. The direction that is perpendicular and outward-facing from the shock is termed the bow shock normal $\hat{n}$.

Classically, the motion of a fluid can be described by the Navier-Stokes equations. Magnetohydrodynamics (MHD) describes the motion of plasma as a two-fluid representation, one of electrons and one of ions, by marrying Navier-Stokes with Maxwell's equations. And just as there are classical shocks, so also exists the MHD equivalent (see chapter 5 of *Priest* [2014] as well as *Kennel* [1994]). Also classically, shocks convert kinetic energy to thermal

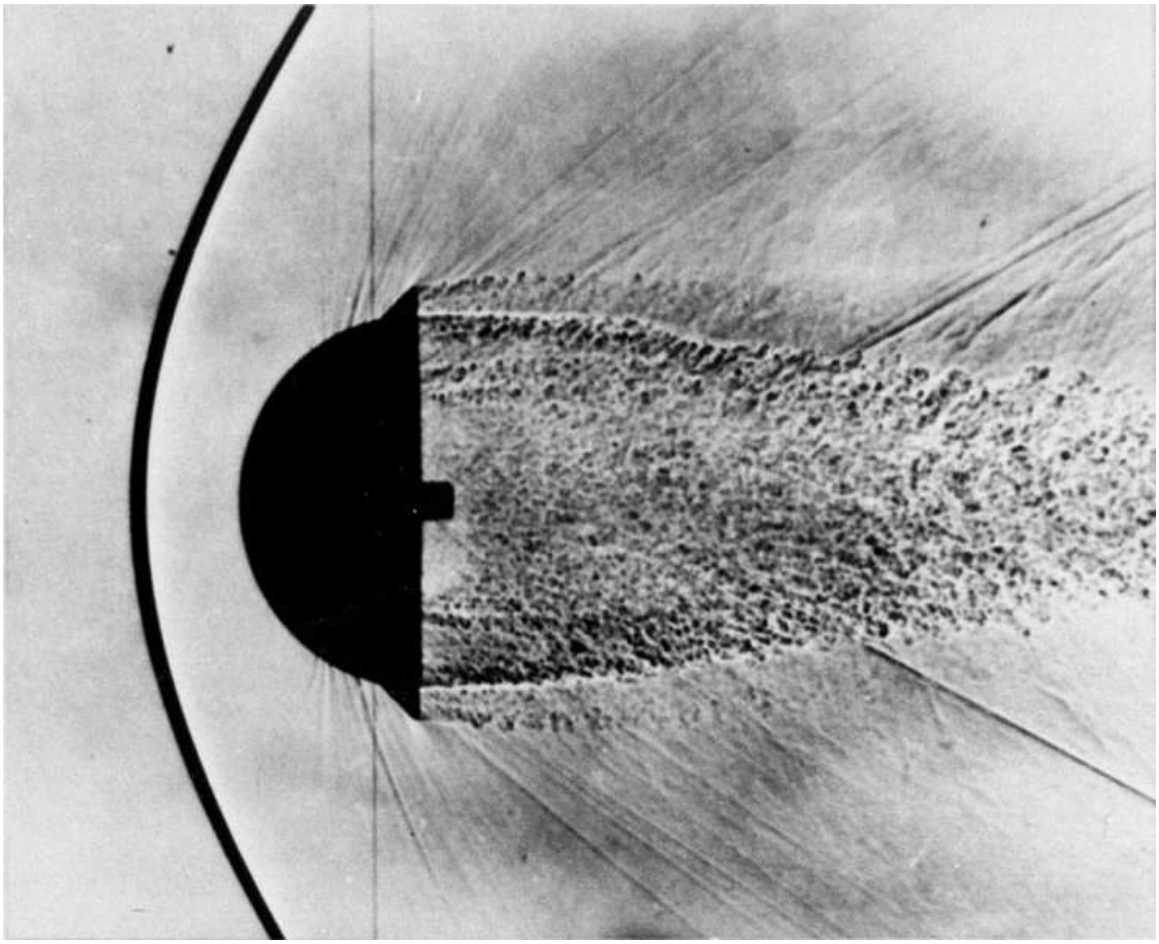

Figure 1-3: A shadowgraph of a high speed fluid being diverted about a rounded object. A loose comparison between the regions in this image and the magnetospheric environment can be made by analogizing the solar wind to the incident fluid flow on the left, the magnetosheath as the shocked fluid just downstream of the shock, and the magnetosphere as the object and its resulting trail. Image credit at NASA.

via collisions, resulting in an increase in entropy downstream of the shock. The presence of collisions is inferred from the mean-free-path of the particles constituting the fluid, which is very small in many cases and often less than the thickness of the shock. In plasmas, however, electromagnetic fields can guide the motion of particles in addition to collisions. This can generate shocks wherein the thickness is much smaller than the mean-free-path, with such shocks being termed "collisionless."

Incorporating magnetic fields into the equations of motion for a two-species fluid introduces a large variety of complexities and instabilities (such as the Alfvén wave [*Nariyuki* [2022]], tearing [*Galeev and Zelenyi* [1976]]], and Kelvin-Helmholtz [*Masson and Nykyri* [2018]] instabilities to name a small handful), but one most relevant to the distinguishing of plasma regions is the angle θ_{Bn}, which is the angle between the bow shock normal and the magnetic field vector (or interplanetary magnetic field [IMF] in the context of the solar wind). This is important because, while in classical fluid dynamics, fluid elements travelling downstream cannot go back upwards, this angle can produce such an effect in collisionless shocks separately for ion and electron populations. Upstream particles with a combination of sufficiently high energy and pitch angle can enter the shock, gain energy, and be reflected back upstream (see section 8.4 of *Gurnett and Bhattacharjee* [2017]). Since the solar wind is ever-flowing, there is a continuous presence of these energetic backstreaming particles and is referred to as the foreshock.

Due to the mass difference between ions and electrons, there are different populations of foreshock. Distinguishing the range of angles in θ_{Bn} as either quasi-perpendicular ($\theta_{Bn} \geq 45°$) or quasi-parallel ($\theta_{Bn} \leq 45°$), it is known that the electron foreshock can be seen in both regimes (*Fitzenreiter* [1995]) while the ion foreshock is restricted to quasi-parallel angles as reflected ions often get turned around within an ion gyroradius (see chapter 5 of *Balogh and Treumann* [2013]). To distinguish between foreshock-contaminated solar wind

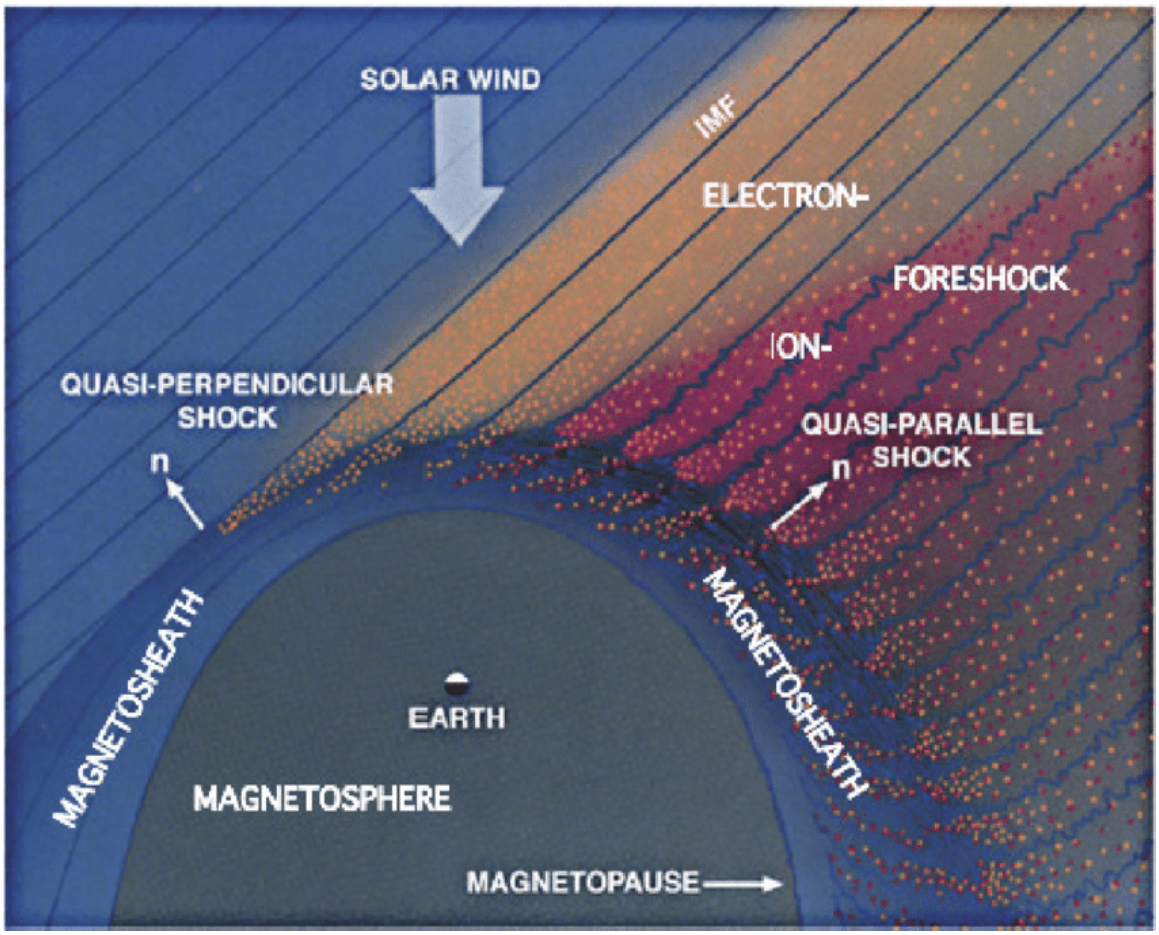

Figure 1-4: Depiction of both ion and electron foreshocks in the solar wind — Credit to *Oliveira* [2015] for captioned image and *Kennel et al.* [1985] for original image.

and otherwise, solar wind that is far enough upstream such that it hasn't encountered any reflected populations or the instabilities such can generate is referred to as pristine solar wind. Measurements of this plasma are made by spacecraft orbiting the L1 Lagrange point, such as ACE or Wind, at about 230 R_E upstream.

1.2 Bow Shock Modeling

The first empirical model of the bow shock was created by *Fairfield* [1971] by fitting a conic section to the bow shock crossings of multiple Imp spacecraft. Although their data included crossings at non-equatorial altitudes, they rotate their crossings such that all points are in

the X-Y plane, and then rotate these 2D points by a consistent 4 degrees to account for angle that solar wind arrives at Earth. Being a 2d problem, fitting a conic section involves finding the optimal constants to the equation

$$y^2 + Axy + Bx^2 + Cy + Dx + E = 0 \tag{1.1}$$

that best describe the average bow shock position. Note that the recovered constants are not parameterized relative to changing pristine solar wind effects and that this model describes the average boundaries of a 2d asymmetric bow shock.

The first 3d empirical model was constructed by *Formisano* [1979] and used bow shock crossings from both HEOS and Imp observations. These crossings were not rotated or aberrated but the radii were modified by normalizing them relative to the average pressure according to $R_{norm} = R_{obs}\frac{n_{obs}V_{obs}^2}{n_0 V_0}$ where n_{obs} and V_{obs} are the solar wind number density and speed of the observations and n_0 and V_0 were the averages of those quantities for years 1972-3 of their dataset. Additionally, they weight their crossings against those of HEOS-2 in inverse proportion to the amount of time each spacecraft spends in each 1 R_E x 1 R_E x 1 R_E box across their dataset to discourage fitting solutions from focusing too much on crossings located in similar small regions of space. They also fit a conic section, but a simplified 3d one of the form

$$a_{11}x^2 + a_{22}y^2 + a_{33}z^2 + a_{12}xy + a_{14}x + a_{24}y + a_{44} = 0. \tag{1.2}$$

They assumed symmetries relative to the equatorial plane (i.e. azimuthal [Y-Z], X-Z, and Z symmetries), hence there are no yz, xz, or z terms. They binned their data across three intervals of Alfveń Mach number M_A 1-6, 6-10, and 10-20, showing increasing bow shock compression for higher values. They also binned according to magnetosonic Mach

number M_{MS} for the intervals 1-5, 5-7, and 7-20 and estimated similar constants as in the M_A binning.

Slavin and Holzer [1981] created a 2d empirical model using data from Imp 3 and 4, HEOS 1, and Prognoz 1 and 2 in which they modified their coordinates to account for the V_Y aberration of the solar wind. This was not done like with *Fairfield* [1971], where a blanket 4 degree aberration correction was incorporated, but rather the aberration correction was done for each individual crossing (for crossings in which they lacked upstream solar wind measurements, they assumed a mean solar wind speed of 430 km/s). They start from a second order surface like that described by Equation 1.1, but remove the xy term. They do this because the xy interaction term represents an aberration in the coordinates. Both *Fairfield* [1971] and *Formisano* [1979] compared the aberrations of their models to the expected aberration due to the orbital motion of the Earth, but since *Slavin and Holzer* [1981] have preemptively removed the aberration from their data, there is no need for this term in the fitting. They show that by algberaic manipulation, this surface can instead be expressed in polar form as

$$r = \frac{L}{1 + \epsilon \cos\theta} \tag{1.3}$$

where L is the semi-latus rectum and ϵ is the eccentricity. Note however that r is not the radius to the bow shock from Earth, but is instead the distance from the foci to the bow shock.

In gas dynamical analysis, an expression relating the bow shock standoff distance Δ and magnetopause standoff distance D to the sonic Mach number M and ratio of specific heats γ of the upstream solar wind based on numerical simulations is

$$\frac{\Delta}{D} = q\frac{(\gamma+1)M^2+2}{(\gamma-1)M^2}, \tag{1.4}$$

which comes from combining Equations 29 and 30 of *Spreiter et al.* [1966] (also see sources therein) where $q = 1.1$. This relation is the firmament of many bow shock models, although modifications are usually incorporated, such as substituting different Mach numbers (magnetosonic, Alfveń, etc), changing the constant q, and using different values for γ. *Němeček and Šafránková* [1991] used bow shock crossings derived from Imp 8 and Prognoz 10 observations to take such an approach. They used the magnetosonic Mach number, made q linearly dependent on the ratio of the upstream magnetic field relative to its average, and used a γ of 1.8 to improve upon their best-fit second order bow shock conic surface. Their model was updated by *Jeřáb et al.* [2005] using bow shock crossings from Prognoz, Magion-4, Geotail, Imp, and Cluster by re-expressing their Mach number term as

$$\frac{(\gamma+1)M^2+2}{(\gamma-1)M^2} \rightarrow \frac{(\gamma-1)M^2+2}{(\gamma+1)(M^2-1)}, \tag{1.5}$$

using instead the Alfveń Mach number and an improved linear dependence on the magnetic field ratio. Their Mach number correction was inspired by the success of the bow shock model created by *Farris and Russell* [1994], who cite *Landau and Lifshitz* [1987] in relating the upstream sonic Mach number and downstream sonic Mach number via γ according to gas dynamics theory and the RH relations. The bow shock model of *Farris and Russell* [1994] estimate the standoff distance of the bow shock, that is, the position of the nose of the bow shock along the X axis in GSE coordinates, and its radius of curvature to describe the bow shock shape. Their usage of Expression 1.5 is to satisfy the condition that the bow shock move out to infinity as the Mach number approaches 1.

Cairns and Lyon [1995] used 3d ideal MHD simulations to simulate the bow shock for low Alfveń Mach number. *Verigin et al.* [2001] uses a similar approach as *Farris and Russell* [1994] in modeling the bow shock shape by calculating the bow shock standoff distance and curvature. *Fairfield et al.* [2001] notes in the rare cases of low Alfveń Mach number in which

the standoff distance moves out to 40+ R_E that *Cairns and Lyon* [1995], a modified *Farris and Russell* [1994], and *Verigin et al.* [1997] predict this quite well.

Peredo et al. [1995] constructed a 3d model using bow shock crossing positions across a variety of spacecraft coupled with hourly-averaged upstream solar wind estimates collected using *King* [1979] after both aberrating and pressure-normalizing. They then binned their data according sonic, Alfvénic and magnetosonic Mach numbers and derived the best-fit 3d conic sections using the simplified form seen in Equation 1.2. They also made a further attempt in their fitting to account for the variability in B_Z by rotating from aberrated GSE into geocentric interplanetary medium (GIPM, *Bieber and Stone* [1979]) coordinates such that B_Z in the resulting coordinate system is zero. Finally, they use their best-fit parameters for the pressure-normalized GIPM-rotated data to express them as explicit parameterizations of M_A by fitting them to a second-order polynomial, producing a final model that predicts the bow shock location as a function of solar wind dynamic pressure, IMF, and Alfvén Mach number. After reports of biases in the model predictions (*Safránková et al.* [1999]; *Merka et al.* [2003]), *Merka et al.* [2005] improved their model where fitting was done using the Levenburg-Marquardt algorithm (*Marquardt* [1963]) and the errors were calculated using a superior error estimation technique (*Efron* [1979]) which has shown good results in space physics (*Kawano and Higuchi* [1995]).

Many of the previous mentioned bow shock models have been fitted for close-to-Earth bow shock positions (often ≤ 50 R_E), but at least one model (*Bennett et al.* [1997]) was made to account for downtail crossings 100+ R_E away, which were observed by Galileo and Pioneer 7. *Greenstadt et al.* [1990] notes that ISEE 3 bow shock crossings made 110 R_E downtail were consistent with the predictions of the tail-symmetric version of the *Fairfield* [1971] model, inspiring *Bennett et al.* [1997] to use a similar modeling approach. They pressure-normalize their crossings and construct a modified cylindrical model with the expressions

$$\rho_{shock} = \rho_1 + \Delta\rho, \tag{1.6}$$

where

$$\Delta\rho = (x_{n3} - x)(\tan(\psi_p) - \tan(\psi_b)) \tag{1.7}$$

describes the deviation from the base model given by

$$\rho_1 = \sqrt{\frac{L_1}{1 + \cos\theta} - (x - x_{03})^2}. \tag{1.8}$$

ψ_b and ψ_b are the Mach cone angles for the prevailing and average magnetosonic Mach numbers, x_n is the bow shock nose distance from Earth, x_{n3} is a modified calculation of x_n involving the MHD and gas dynamical assumptions of *Cairns and Lyon* [1995] and *Farris and Russell* [1994] respectively, and L_1 is their updated semi-latus rectum (and not the L_1 Lagrange point).

Chapman and Cairns [2003] created a paraboloidal bow shock based on the simulated bow shock locations of *Cairns and Lyon* [1995] for Parker spiral angles of $\theta_{IMF} = 45°$ and 90° according to the expression

$$x = a_s - b_s(y^2 + z^2) \tag{1.9}$$

where (x,y,z) are in GSE coordinates, a_s is the standoff distance of the bow shock, and b_s is a flaring parameter that causes the bow shock to expand or contract. a_s is parameterized by a $1/M_A^2$ and pressure-normalized dependence and b_s is fitted for different azimuthal angles in ϕ.

Arguably the most widely used bow shock model is that of *Chao et al.* [2002] owing to its

extensive parameterization, careful selection of bow shock crossings, and easily interpretable coordinates and implementability. It models the bow shock using the radially symmetric functional form

$$R(\theta) = R_0(\frac{1+\epsilon}{1+\epsilon\cos\theta})^{\alpha}, \tag{1.10}$$

which was inspired from a similar form (with $\epsilon = 1$) used to model the magnetopause in *Shue et al.* [1998]. Here, R_0 is the standoff distance of the bow shock, α is the flaring angle, ϵ is a parameter similar to the eccentricity (but not exactly due to the exponentiation by α), and θ is the cone angle from the x axis. Note that $\alpha \geq 0.5$ creates an "open" bow shock (corresponding to a modified parabola) while $\alpha < 0.5$ results in a "closed" bow shock (one corresponding to a modified ellipse such that the contours on the nightside eventually converge). ϵ is used here in order that it can be made compatible with distant bow shock crossings (*Bennett et al.* [1997]). Many of the symmetric models mentioned thus far have utilized parabolic coordinates in which the origin is at the focus of the modeled conic, but this model can easily be used and interpreted where the origin is simply the origin in GSE (the center of the Earth). In the preparation of their dataset, they took care to only select bow shock crossings that corresponded to a gradual crossing, and in the case of multiple bow shock crossings, only the middle one was selected. The first point is subtle, but it can be noted that just because a bow shock crossing is observed in spacecraft data does not mean that the upstream solar wind conditions at time of measurement cause the equilibrium bow shock boundary to be at the spacecraft position; rather, in cases of sharp immediate crossings, it can be the case that the bow shock boundary moved past the spacecraft to a new equilibrium position beyond it, giving the erroneous impression that the given upstream solar wind conditions at that time caused the bow shock boundary to be at the spacecraft position. Applying these filters to their dataset reduced their data size, but resulted in

a high quality bow shock crossing catalogue. Their fitting was done on crossings that accounted for the V_Y aberration but were not pressure-normalized. The $D_p^{-1/6}$ relationship coming from theoretical calculations of the magnetopause pressure balance and the bow shock standoff distance being made a function of Mach number as per Equation 1.4 are both incorporated in the parameterization of R_0. *Dmitriev et al.* [2003] compared their model along with *Peredo et al.* [1995], *Verigin et al.* [2001], and *Farris and Russell* [1994] and found that the model of *Chao et al.* [2002] was the best for bow shock prediction.

Wang et al. [2015] used Space Weather Modeling Framework global MHD simulations (*Tóth et al.* [2005]) to simulate the bow shock under a spectrum of dipole tilt angles to investigate the dipole tilt dependence. They use a model function comparable to that of *Shue et al.* [1998] in which

$$R = R_0(\frac{2}{1+\cos\theta})^{(\alpha+\beta\cos^2\phi)} \tag{1.11}$$

and α and β are z-axis dependent parameters such that $\alpha = \alpha_n(\alpha_s)$ for Z ≥ 0 (< 0) and $\beta = \beta_n(\beta_s)$ for Z ≥ 0 (< 0). Note that the $\cos^2\phi$ term alone (without Z-dependent α and β) can describe azimuthal (that is, Y-Z) asymmetry but the use of different parameters for each hemisphere is required to incorporate North-South (i.e. Z-axis) asymmetry. They then proceed to make each of their five parameters functions of the dipole tilt, showing a continual increase in the bow shock standoff distance with increasing dipole tilt and strong North-South asymmetry.

Lu et al. [2019] built off of the model of *Wang et al.* [2015] and created a bow shock model using Imp 8, Geotail, Magion-4, and Cluster bow shock crossings (with upstream solar wind measurements provided by ACE and Wind). They aberrate their crossings into corrected GSM (cGSM) where x points along the solar wind flow (accounting for both V_Y and V_Z aberrations), z points along the component of the geocentric dipole moment that

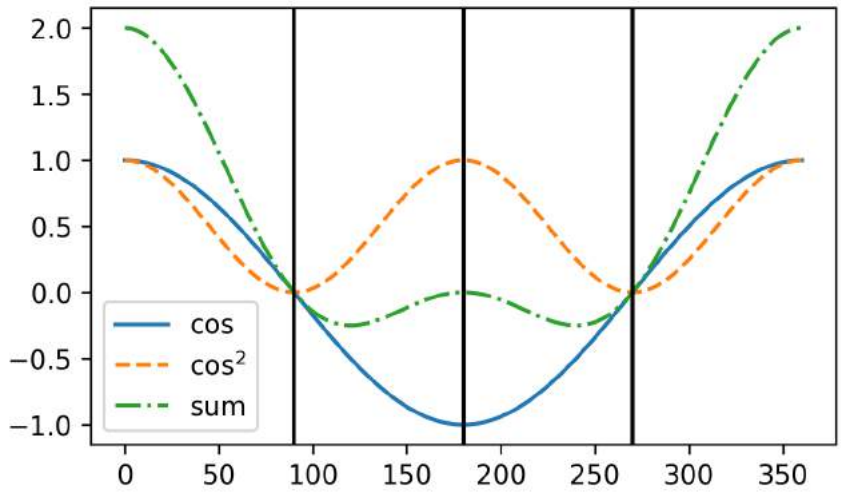

Figure 1-5: Plots of $\cos\phi$, $\cos^2\phi$, and their sum from 0 to 360^{circ}. cos is anti-symmetric for the ranges (0,90) and (90,180) whereas $\cos^2\phi$ is symmetric, and their sum breaks the symmetry for the interval (0,180). However, this function is still symmetric between the (0,180) and (180,360) intervals (meaning the model is inherently dawn-dusk symmetric).

is perpendicular to x, and y is defined such that $\hat{y} = \hat{z} \times \hat{x}$, and then pressure-normalize their bow shock position and IMF vectors. They used a model function that expresses the North-South and azimuthal asymmetry using a single set of parameters, which is given by

$$R = R_0\left(\frac{2}{1+\cos\theta}\right)^{\alpha_0+\alpha_1\cos\phi+\alpha_2\cos^2\phi} \tag{1.12}$$

where the total flaring angle α is broken into three terms: α_0 representing the rotationally symmetric flaring angle, α_2 describing the North-South asymmetry similar to Equation 1.11, and α_1 capturing the azimuthal asymmetry. Note that any dawn-dusk asymmetry cannot be captured by this model (see Figure 1-5 for a graphical explanation). They then parameterize the four parameters as functions of upstream solar wind B_Z, dynamic pressure, magnetosonic Mach number, plasma beta, and dipole tilt. Using a test set of Cluster 3 bow shock crossings, they show that overall their model is comparable to *Chao et al.* [2002], but that theirs performs noticeably better for dipole tilts with magnitude $> 30°$.

1.3 The Influence of B_Y on the Bow Shock

The IMF clock angle ϕ_B represents both IMF B_Y and B_Z. An analysis of the influence of B_Y on the bow shock was made with the clock angle using global MHD simulations by *Wang et al.* [2016]. They observed that the tail bow shock cross section can be viewed as an ellipse in which the direction of the major axis is perpendicular to the IMF direction. They note that for northward IMF, the eccentricity of the tail bow shock cross section increases with the IMF clock angle, and for southward IMF, the eccentricity decreases with the clock angle

Studying B_Y outright on observed bow shock crossings was done by *Wang et al.* [2018]. They noted several properties, including: (1) The bow shock standoff distance increasing and flaring angle decreasing as eastward B_Y (that is, positive B_Y) increases, (2) the standoff distance not changing significantly and flaring angle decreasing less as westward B_Y (that is, negative B_Y) increases, (3) the magnitude of eastward B_Y influences the location of the bow shock nose comparably to that of B_Z. They attribute the third point to the fact that B_Z has more influence on the shape of the magnetopause, compensating for some of its effects on the bow shock.

1.4 The Novel Contributions of This Thesis

In this thesis, we have created a layered unsupervised classifier that can classify THEMIS and MMS observations into belonging to the solar wind, the magnetosheath, or the magnetosphere. This is done using Self-Organizing Maps to get a greatly discretized interface of the data and hierarchical clustering to classify the resulting nodes in an unsupervised manner. This approach allows for unique identification of data via feature maps in which events corresponding to multiple particular features can be visually identified as belonging

to particular nodes. The hierarchical organization of the nodes also means that the constituent clusters can be "unpacked" and investigated to reveal sub-clusters. We will cover the source data, data preparation, development, validation, and applications of the classifier in chapters 3 through 7.

Bow shock crossing times and positions are inferred from the classifications of the model with upstream solar wind estimates provided by OMNI. Additional crossings are taken from other sources, including Cluster, Imp 8, Magion-4, and Geotail. These crossings are then used to create a bow shock model with a neural network implementation using traditional bow shock parameters as well as the magnetic clock and cone angles. Instead of directly predicting the bow shock radius for these inputs, we predict coefficients which we supply to a bow shock model function with known interpretations of the coefficients. Due to the small size of the training set, we create bootstrap samples of the training data and train an ensemble of neural networks on these samples. We prune the ensemble and choose a member size of 12, showing that it outperforms the single fully trained model on the validation set. The model results show mixed agreement with previous observations and performs better than the Chao model for varying clock and cone angles and for nightside bow shock crossings, but slightly underperforms for dayside crossings. The collection of the source data, data preparation, model development, and the model results are covered in chapters 3, 8, 9, and 10.

Chapter 2

An Overview of Statistical Concepts and Machine Learning Methods Used

A number of statistical concepts and machine learning methods were used to generate results, and we cover introductions to them separately here.

2.1 Principal Component Analysis

Principal Component Analysis (PCA, *Jolliffe* [2011]), provides a matrix Q with shape D x K to reduce the dimensionality D of a dataset to a reduced dimension K via a linear transform where K is specified and can vary between one and D. The goal of PCA is to find a new set of uncorrelated variables, called the principal components, that capture the maximum variance in the data. These principal components are ordered by the amount of variance they explain, with the first component explaining the most variance and subsequent components explaining less. A common way this is done is by computing the eigenvalues and eigenvectors of the covariance matrix of the data. The eigenvalues quantify the proportions of variance captured by the eigenvectors and these eigenvectors are the principal components.

It is analogous to a hyper-rotation of the D-dimensional space in which the cardinal axes, or principal components, are oriented along directions of decreasing variance. If a variance threshold is chosen, then a number of the principal components can be selected that cumulatively represent that variance. This method has limitations in that it is a obviously a linear method of dimensionality reduction. When data are characterized by non-linear correlations, this complicated structure can be destroyed in the transformation and can cause misinterpretations of the resulting components. However, this linearity can also make it readily interpretable. Once Q is known, its elements, or "loadings", can be inspected to ascertain the influence of each feature along any principal component. PCA uses linear combinations of features to ascertain directions of maximal variance with projections of the form $PCA_i = \sum_a^D z_{ia} F_a$ where PCA_i indicates the i^{th} principal component, $\{F_a\}$ is the set of D features, and $\{z_{ia}\}$ are the loadings in the linear combination. Using just the first two principal components, we can visualize these loadings as vectors that can visually communicate the importance of each feature in the projection. Plotting these vectors on top of the first two components of the projection is called a biplot. Using biplots to infer information from PCA results has a rich history and an introduction to the concept is covered in *Kohler and Luniak* [2005]. An example showing how PCA is related to a hyper rotation and preserves cumulative variance in shown in Figure 2-1.

2.2 K-Means

K-Means (*Lloyd* [1982]) is one of the most popular clustering algorithms. It partitions data into k Voronoi-separated clusters where k is pre-specified. To accomplish this, k random points from the data are selected to act as initial cluster centroids. The distances between points in the data and the centroids are computed and points are assigned to the cluster whose centroid they are closest do. The centroid positions are re-computed as the average

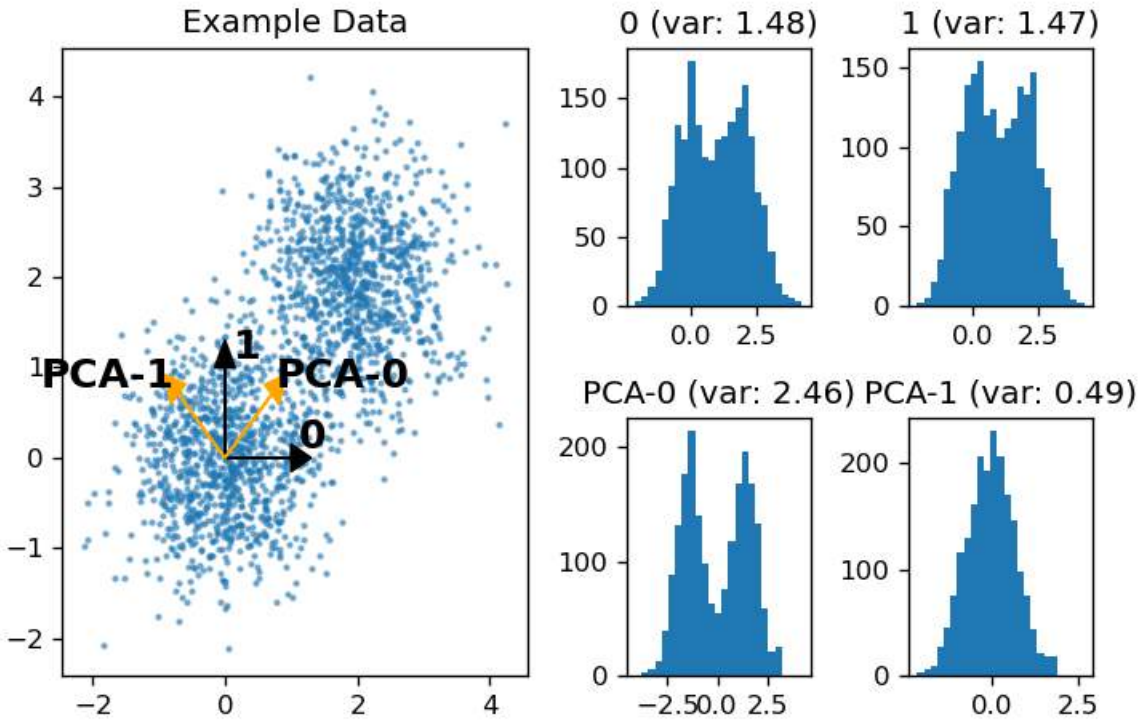

Figure 2-1: Two isotropic 2d gaussians comprised of several hundred points are plotted in the left plot under the title "Example Data". The black arrows indicate the positive direction along each axis with **0** representing x and **1** as y. Histograms along these axes are shown along the top row of plots on the right. The variance along each dimension, computed directly as the population-normalized squared deviation from the mean, i.e. $1/n \sum_i (q_i - \mu)^2$, is shown in the title of each plot. Using PCA, new axes are derived that indicate directions of decreasing variance in their cardinal ordering, shown in the left plot as **PCA-0** and **PCA-1**. Histograms of the data along these new axes are shown on the bottom row of plots as well as the variances along these axes. Note that (1) these two gaussians are largely (but not completely) separable along the **PCA-0** axis and (2) the sum of variances across all axes is preserved (e.g. $1.48 + 1.47 = 2.46 + 0.49$) as rotations will preserve sum-of-square calculations. (1) implies that the dimensionality inherent to separating these two gaussians can be reduced from two dimensions, **0** and **1**, to one, just **PCA-0**.

of all the points in their respective clusters. The inertia, or the sum of square distances of each point from their closest representative, is re-calculated after each iteration, and the procedure continues until its consecutive changes are small or until a max iteration number is reached. A common initialization method is K-Means++ (*Arthur and Vassilvitskii* [2007]), which makes better choices for cluster centroids by weighting data in proportion to their square distance from the previously created centroid.

This type of learning is purely competitive in that centroid updates are only affected by the data in their own clusters. It will have limited success with data that do not contain spherically separable clusters, particularly those non-convex in shape. A work-around to the non-convexity difficulty has been to use KMeans on data to resolve many clusters (often several hundred or more) and then apply a more resilient clustering method to the cluster centroids and propagate the predictions of this second stage clustering to the data represented by the centroids. However, this method will still be subject to the competitive learning biases inherent in K-Means solutions.

2.3 Self-Organizing Maps

In a higher-dimensional space of N points, finding m "prototype" points where m << N while also minimizing some predefined distortion criteria is the main goal of vector quantization (*de Bodt et al.* [2004]; *Gray* [1984]). The distortion criteria changes for different methods, but the most common one involves computing the inertia.

A more robust method for clustering is the Self-Organizing Map (*Kohonen* [1982]), or SOM, which uses a combination of competitive and cooperative updates. A SOM is a method of clustering that is meant to resemble the structure of a neural network. The neurons are referred to as nodes and they are usually arranged in a square 2d grid (i.e. the "node-space"). Each node has a weight vector $\mathbf{w}$ which is the position of the node

relative to the data (i.e. in "feature-space"). The relationship between higher-dimensional data in the feature-space and the 2d grid of the node-space allows for 2d visualizations of higher dimensional data. To train a SOM, a data point $\mathbf{q}$ is presented to the network and the closest node in feature-space, called the best-matching-unit or BMU, is identified. The BMU will then be moved closer to $\mathbf{q}$. If we interpret similarity between two points as being related to their proximity, then moving the BMU closer to $\mathbf{q}$ can be described as making the BMU more similar to $\mathbf{q}$, or more representative of it. An example of the convergence of a simple SOM on 2d data is shown in Figure 2-2.

To incorporate a form of Hebbian learning, or "neurons that fire together wire together," nodes near the BMU are moved closer to $\mathbf{q}$ as well (strictly speaking, SOMs are classified as competitive Hebbian learners). The amount they are moved is proportional both to their feature-space distance to $\mathbf{q}$ (the distance from the node to $\mathbf{q}$ as seen in the left plots of Figure 2-2) and their node-space distance to the BMU (i.e. the distance from the node to the BMU as seen in the right plots of Figure 2-2). This node-space distance is supplied to the neighborhood function and usually involves exponentially diminishing distances. Common neighborhood functions are the gaussian and Ricker wavelet functions, both of which are reliant upon the neighborhood distance hyperparameter σ to determine the sharpness of the distribution. The update to the weights of node a per iteration can be expressed with

$$\mathbf{w}_a(i+1) = \mathbf{w}_a(i) + \alpha(i)h(i, a, BMU)(\mathbf{q} - \mathbf{w}_a(i)) \tag{2.1}$$

where i is the iteration number, $\mathbf{w}_a(i)$ is the weight vector for node a at iteration i, $\alpha(i)$ is the learning rate at iteration i, and $h(i, a, BMU)$ is the neighborhood value between nodes a and the BMU at iteration i. The $\mathbf{q} - \mathbf{w}_a(i)$ term represents the feature-space distance of node a from $\mathbf{q}$ and the $h(i, a, BMU)$ term is the neighborhood distance of node a from the BMU. Convergence of the SOM is guaranteed by specifying a finite number of iterations.

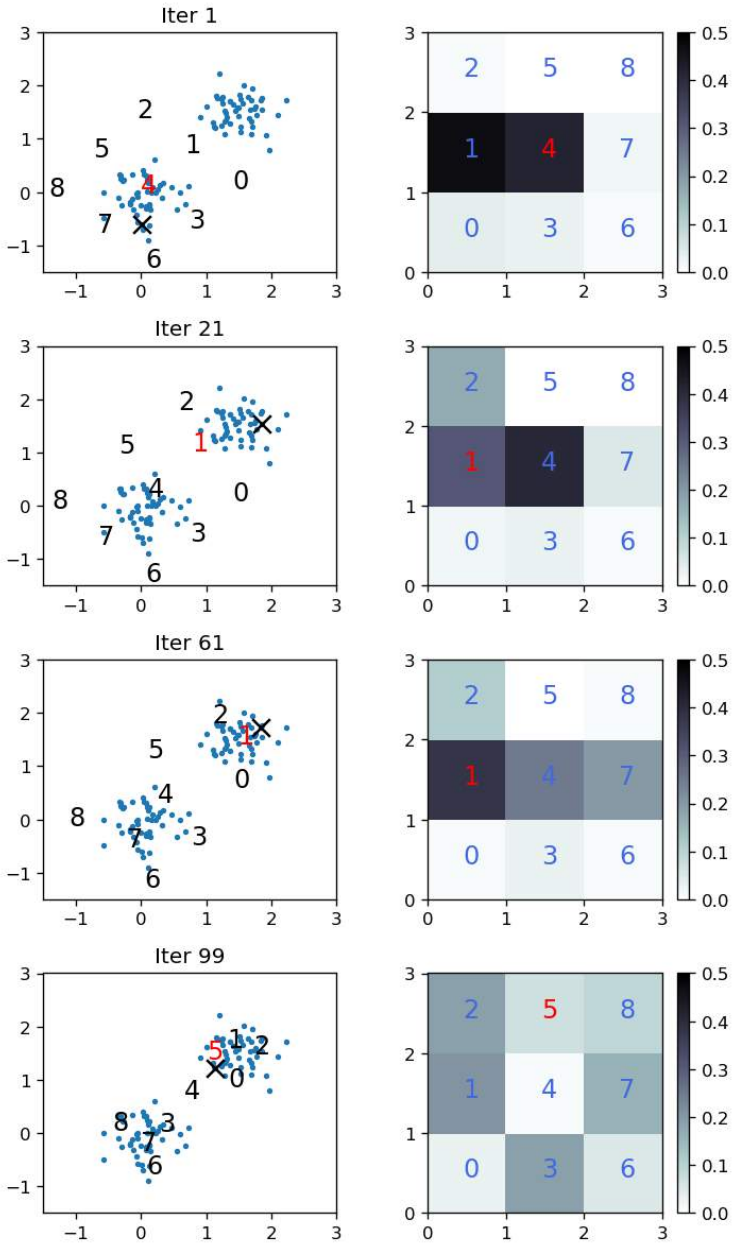

Figure 2-2: These plots show the convergence of a SOM over iterations on a simple 2d dataset consisting of two isotropic gaussians. **Left**: The node positions per iteration in the data are shown. The positions are initialized over the first two principal components of the data. For clarity, the positions are plotted as their number in the SOM grid (node 3 is depicted as a '3'). At each iteration, one data point is selected to train the network against (plotted as X) and the BMU for that point is identified and shown in red. **Right**: The 2d SOM grid is shown per iteration. Heatmaps show the fraction of data points that any node is closest to. At iteration 1, nodes 1 and 4 are the closest nodes to (or, "represent") about 90% of the data. By the final iteration, the data representation is more equidistributed across the network. The distribution could be further improved with a better choice of hyperparameters for this SOM.

Defining an initial and final α and σ, a decay function determines the learning rate and neighborhood distance at each iteration i. Commonly used decay functions include a linear or exponential decay. The weight update for each neuron as seen in Equation 2.1 can be directly contrasted with the K-Means centroid update, which is given by

$$\mathbf{w}_a(i+1) = \mathbf{w}_a(i) + \frac{1}{|C_a(i)|} \sum_{\mathbf{x}_j \in C_a(i)} (\mathbf{x}_j - \mathbf{w}_a(i)) \tag{2.2}$$

where $C_a(i)$ is the set of points belonging to cluster a at iteration number i, $|C_a(i)|$ is the cardinality of the set, and the sum is only over points that belong to cluster a at iteration i. The position for K-Means centroid a at iteration number *i+1* is also proportional to the difference between a point $\mathbf{x_j}$ and the original centroid position $\mathbf{w}_a(i)$ and is basically the same as the $(\mathbf{q} - \mathbf{w}_a)$ term in Equation 2.1. However, there is no learning rate (or more technically, there is a constant learning rate of magnitude 1) and no neighborhood contributions. This is because there is no defined topological structure organizing the K-Means cluster centers.

Another aspect of SOMs is the way the nodes are structured relative to each other. In the example SOM used in Figure 2-2, a node's immediate neighbors are those vertically or horizontally adjacent to it such that node 4 has nodes 1, 5, 3, and 7 as immediate neighbors, node 5 has nodes 2, 4, and 8 as immediate neighbors, and node 0 only has nodes 1 and 3 as immediate neighbors. This structured relationship is referred to as the topology of the network and the type used here is a square topology. Topologies involving any convex shape are technically possible, but only a handful are really utilized, such as the hexagonal topology. If a map with a square topology represents a grid-like structure, then one with a hexagonal topology resembles a style of honeycomb structure. Maps with square topologies are simpler and easier to visualize whereas hexagonal ones can be more difficult to visualize. However, this more complex topology can often more compactly represent data

than a square topology and can have less edge effects at the borders of the map. We only consider a square topology for our model.

The quantification of similarity between a SOM and the data is the quantization error Q, given as

$$Q = \frac{1}{N} \sum_{i=1}^{N} |x_i - BMU(x_i)| \tag{2.3}$$

where x_i is the i^{th} point of a dataset of size N and $BMU(x_i)$ is the BMU for point x_i. This is simply the concept of inertia described previously, but normalized to the size of the data. A survey of SOM applications and metrics used to verify their accuracy can be found in *Kohonen* [2014].

2.4 Hierarchical Clustering

2.4.1 The Method

Hierarchical clustering involves using one of two approaches. Agglomerative clustering assumes that all data points are individual clusters and that they can be iteratively merged based on the clusters' similarity. This is called the "bottom-up" approach to hierarchical clustering. The complement to this is divisive clustering, which assumes all data initially belongs to a single cluster and iteratively separates data into heterogeneous subclusters, called the "top-down" approach. We use the hierarchical agglomerative method as implemented in the scikit-learn package (an overview of various hierarchical clustering methods is covered in *Nielsen* [2016]).

The quantification of similarity between a cluster A and B is computed using a linkage function, for which there are several common types: Maximum (or complete) linkage will define the distance from A to B to be the largest pairwise distance between a point in

A and another in B. Minimum (or single) linkage defines the distance as the minimum pairwise distance. Average linkage will define the distance to be the average of the data of A to the average of the data of B. Ward's linkage does not use distance in the previous senses. Rather, it defines the distance between two clusters A and B as the variance of a new cluster obtained combining A and B. Since the variance is computed as the sum of square deviations from the mean (SSD), the clusters that will be merged are the ones which minimise this sum.

Because this method is hierarchical, one needs to define stopping criteria for cluster merging. This is done by visualizing the order of merging using a dendrogram where clusters are shown on the x axis as individual vertical lines and their merge order can be inferred from when their lines are horizontally merged. The position on the y axis of the merging is the SSD of the merged clusters. The dendrogram of the entire agglomerative merging process is visualized first and then a threshold distance is chosen so that only clusters with SSD below this cutoff will be considered. A small example showcasing agglomerative hierarchical clustering with a Ward linkage on eight data points is shown in Figure 2-3.

In hierarchical agglomerative clustering, if there are N clusters, then all N-choose-2 cluster pairings are considered for possible merging. The optimal merger is determined using a linkage function, which produces a number representing the similarity of the clustering where smaller numbers indicate more similar clusters, and the pair with the smallest linkage function value are merged. In some linkage functions, this can be interpreted as a distance, such as with the single, complete, average, and centroid linkages. The linkage we used, Ward's linkage, is instead concerned with identifying the cluster pair that minimizes the in-cluster variance.

A disadvantage of using hierarchical clustering is that getting predictions on data not seen during training can be difficult as the method is inherently transductive, i.e. it is

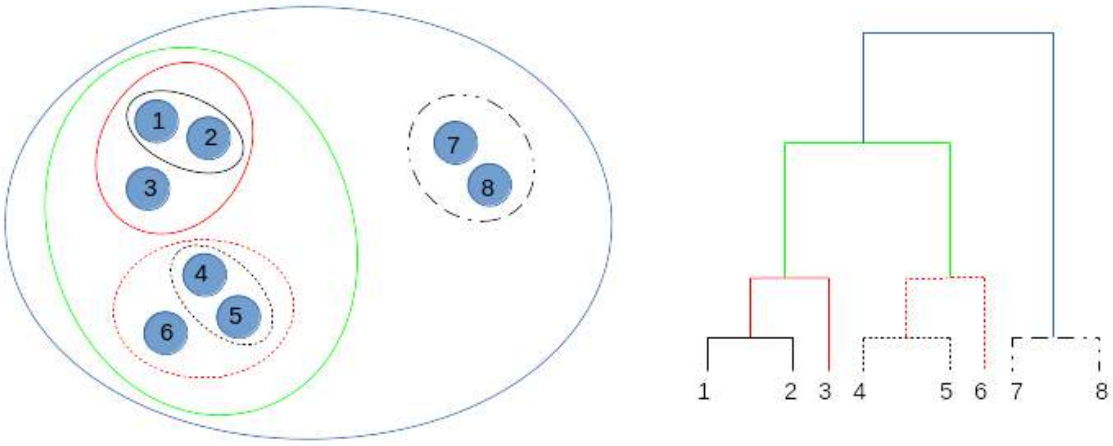

Figure 2-3: A basic example demonstrating how clusters are merged with agglomerative hierarchical clustering using a Ward linkage on a handful of labeled data points. The 2d distribution of the data are shown on the left and their cluster mergings are depicted using different colors and linestyles of ovals. These data have three "natural" clusters given by (1,2,3), (4,5,6), and (7,8). The merge order of the hierarchical clustering is also shown on the right in the form of a dendrogram. Some of the first clusters to be merged are the (1,2), (4,5), and (7,8) clusters due to the proximity / inertia of the constituent points. The next clusters to form are the (1,2,3) and (4,5,6) clusters because the points 3 and 6 are in close proximity to the (1,2) and (4,5) clusters. Then the (1,2,3,4,5,6) cluster is created, followed by the merging of all data into a single cluster. The "compactness" of clusters is also evident by the heights of the mergings in the dendrogram. The mergings in the tree of the clusters (1,2), (1,2,3), (4,5), (4,5,6), and (7,8) are quite short, indicating that the SSD of those clusters is small (e.g. possessing little variance / being quite similar). However, the SSD of the (1,2,3,4,5,6) and (1,2,3,4,5,6,7,8) clusters is very tall, representing how there is a large jump in variance once these clusters are formed.

trained on a specific dataset and does not generalize to unseen data. This limitation can sometimes be circumvented depending on the linkage used but certainly not in general. For example, in using a centroid linkage, one could simply assign new data to clusters whose centroids are closest, but this concept of closeness or similarity becomes vague in the context of other linkages, as some linkage types are capable of uncovering non-convex cluster distributions.

2.4.2 Hierarchical Clustering + SOMs

Hierarchical clustering being a transductive method is an enormous limitation for many clustering purposes as well as the resources required to compute the solution (in terms of memory) and thoroughly analyze said solution (user-hours spent analyzing). This issue can be addressed by creating an interface for the data, something that will represent the distribution of the data but is static and not modified. A trained SOM can act as such an interface. Hierarchical agglomerative clustering can then be used to organize the nodes of the SOM into clusters. The cluster assignments of the nodes are then propagated to the data according to their BMUs. In this way, data not seen during training is always assignable to some node of the SOM and all nodes belong to some cluster. By using a SOM to represent the data, we are able to use what is traditionally a transductive method in an inductive manner. We show an example of this in Figure 2-4.

2.5 Artificial Neural Networks

Artificial neural networks are a subset of machine learning algorithms that can act as advanced functional estimators (*Funahashi* [1989]). Their architecture is inspired from the structure of the human brain in which neurons process, receive, and transmit signals to other neurons (*McCulloch and Pitts* [1943]; *Rosenblatt* [1958]; *Hebb* [1949] and see a basic

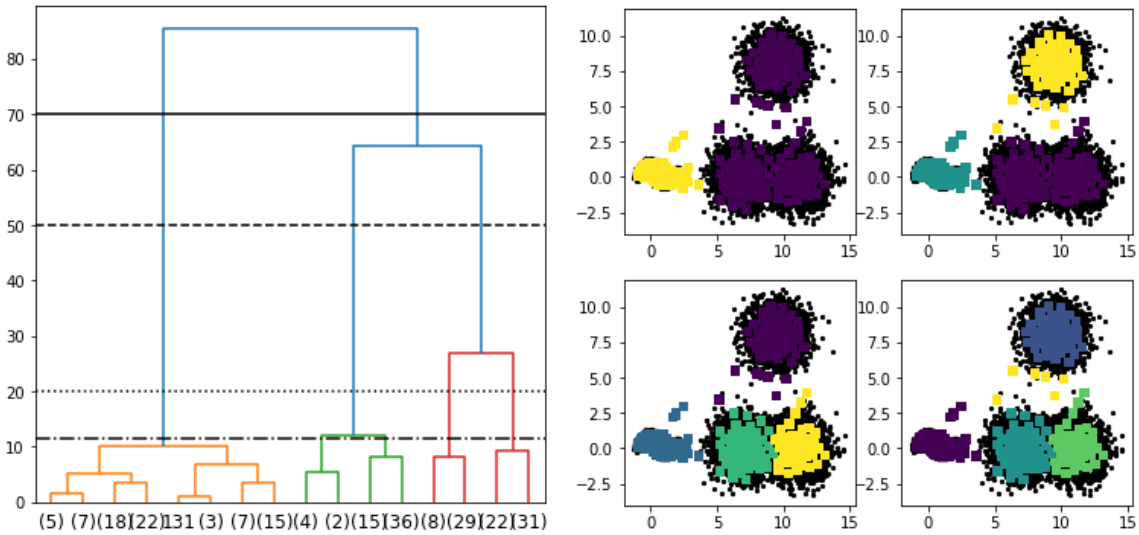

Figure 2-4: An example showing the combined usage of hierarchical agglomerative clustering with a Ward linkage on a SOM consisting of $\sim$ 230 nodes. The 2d data is composed of several 10,000's of points across five distributions, an isotropic gaussian at the top, two conjoined isotropic gaussians in the bottom, and two interlocking semicircles on the bottom left, all of which are seen in black on the four plots on the right. The nodes of the SOM are seen in different colors across each of the four plots where the different colors represent different clustering solutions. The dendrogram at left shows the merge order up to the last three mergings and the horizontal lines indicated different cutoff SSD values we choose to analyze different clustering solutions. Note that the number of times the horizontal lines intersects with the vertical lines of the dendrogram indicate the number of clusters. The clustering solutions found by using a cutoff SSD of 70, 50, 20, and 11 are shown in the plots on the right in top-left, top-right, bottom-left, and bottom-right order. The relative heights of the dendrogram indicate that the "best" clustering solution is likely one with a 50 SSD cutoff, corresponding to the top right plot. However, notice that finer solutions can largely distinguish between the three isotropic gaussians, the semicircles, and the nodes caught between the gaussians!

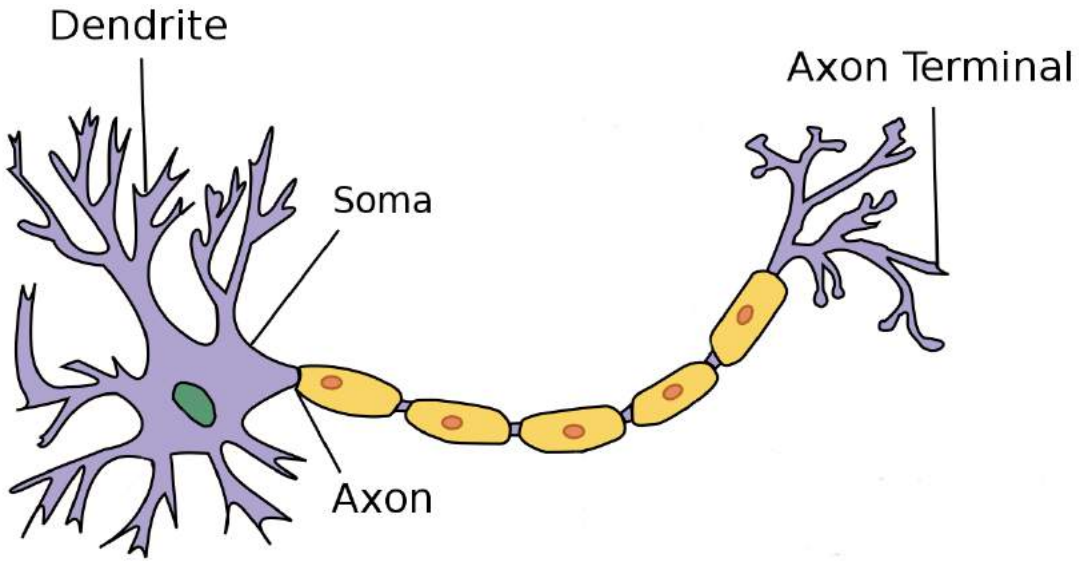

Figure 2-5: A simple depiction of a neuron that illustrates the dendrites (signal receivers), soma (the body of the neuron), axon (signal transmitters), and axon terminals (connections to other neurons) - Original image in U.S. public domain, has been modified to remove additional labels - Retrieved from https://commons.wikimedia.org/wiki/File:Neuron.jpg

diagram depicted in Figure 2-5). Their ability to approximate functions stems from multiple fundamental aspects, some of which will be briefly elaborated on.

One component is the modification (or, "activation") of signals from neuron-to-neuron via activation functions, analogous to the soma in the neural structure. Expressing the concept in terms of linear algebra, a sequence of linear transformations can itself be represented as a single linear transformation. However, a nonlinear modification of the signal can enable modeling of nonlinear behavior.

Another aspect is the degree of connections between neurons. Organizing the neurons into layers where adjacent layers are fully connected is a standard "feed-forward" architecture and consists of an input layer (with the number of neurons in this layer matching the dimensionality of the input data), an output layer (with size equal to the dimensionality of the desired output), and in-between layers called hidden layers. This type of structure is referred to as a Multilayer Perceptron (MLP). An example is shown in Figure 2-6. An

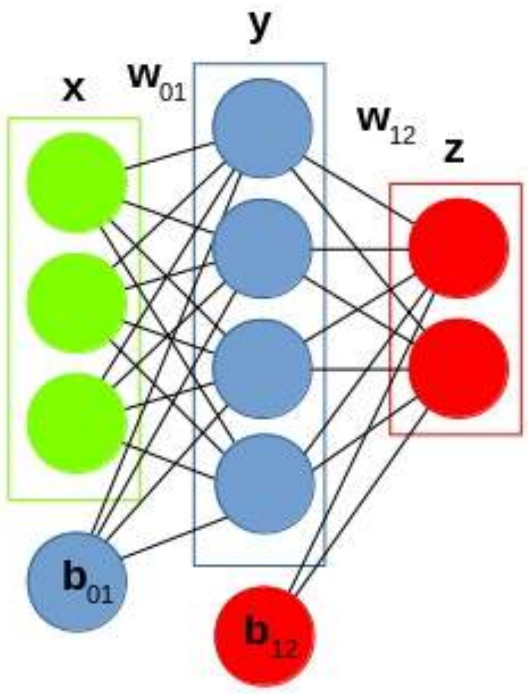

Figure 2-6: A simple diagram of a single-hidden layer multilayer perceptron (MLP). $\mathbf{x}$ is the 3-dimensional input data colored in green on the left and corresponds to the input layer. $\mathbf{Z}$ is the 2-dimensional output colored in red on the right and is the output layer. The layer in blue in between is called the hidden layer and is represented with the 4-dimensional variable $\mathbf{y}$. The neural connections between the input and hidden layer consists of scalar values that are collectively represented by the matrix $\mathbf{W_{01}}$. The values for the connections between the hidden and output layers are $\mathbf{W_{12}}$. The $\mathbf{W_{01}}$ and $\mathbf{W_{12}}$ terms are collectively referred to as "weights." The weights can only perform multiplicative changes (e.g. hyper-stretching) to the data and the bias weights, $\mathbf{b_{01}}$ for the hidden layer and $\mathbf{b_{12}}$ for the output layer, supply the additive changes (e.g. hyper-shift). If the activation function is σ, then expressions for $\mathbf{y}$ and $\mathbf{z}$ are $\mathbf{y} = \sigma(\mathbf{W_{01}x} + \mathbf{b_{01}})$ and $\mathbf{z} = \sigma(\mathbf{W_{12}y} + \mathbf{b_{12}})$.

MLP with a single hidden layer has been proven to be a universal approximator (*Cybenko* [1989]; *Hornik et al.* [1989]) and the same is true for multiple hidden layers (*Hornik* [1991]). Increasing the number of neurons in a network can increase its adaptive capability. However, it is usually better to add connections by adding more hidden layers instead of just increasing the size of the layers (*Schölkopf et al.* [2007]).

Last is the appropriate modification of the weights. This requires quantifying how "wrong" the neural network is given a prediction and how to update the weights accordingly. The inaccuracy of the neural network in estimating a function is expressed using a loss

function, the most popular for regression being the mean squared error (MSE). If N is the size of the dataset, and y_i and $\hat{y}_i$ were the observation and prediction for point i, then the MSE would is written as $\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$.

Updating the weights of an MLP can be expressed as a hyperparameter optimization problem, e.g. what weights minimize the loss function? Many classical optimization methods use the derivative (either in the form of the gradient or the hessian) of the loss function to find optima, such as Newton-Raphson or Gradient Descent (*Cauchy* [2009]), but derivative-free methods exist, such as Nelder-Mead (*Nelder and Mead* [1965]) and Powell's method (*Powell* [1964]). A common optimizer used to update the weights in neural networks is Stochastic Gradient Descent (SGD, *Robbins and Monro* [1951]) in which the full gradient (that is, the calculation of the gradient over the entire dataset) is replaced with an estimation of the gradient over a fraction of the dataset. The weight update, for an arbitrary weight w at iteration number t, is simply $w_t = w_{t-1} - \alpha\frac{\partial L}{\partial w_{t-1}}$ where α is the learning rate and L is the loss function. A variety of optimizers for neural network training exist, and one of the most powerful is Adam (*Kingma and Ba* [2014]). It uses the first ($\hat{m}_t$) and second ($\hat{v}_t$) moments of the gradients to estimate the new weight as $w_t = w_{t-1} - \alpha\frac{\hat{m}_t}{\sqrt{\hat{v}_t}+\epsilon}$.

However, across many of these optimizers, the derivatives of the loss function are still necessary. The most popular method used to supply these gradient calculations is backpropagation (*Rumelhart et al.* [1986]). In essence, it applies the chain rule to the loss function with respect to the weights. The differentiability of the loss function and having neurons output over continuous intervals are improvements over original applications.

2.6 Ensemble Models and Bootstrap Aggregation

It is a common thought that when presented with a distribution of predictions, using the average prediction is a good choice. In 1906, Sir Francis Galton observed amongst partici-

pants at the West of England Fat Stock and Poultry Exhibition in Plymouth, England that the average guess (1207 lbs) of an ox's weight (1197 lbs) across 800 participants was correct to within 1% error and better than any individual guess (*Galton* [1907]). In this historical example, each participant is a predictor and the crowd is the ensemble of predictors.

In general, any collection of models can together create an ensemble model. The models can be of the same architecture or different types. Ensemble modeling is used in many different forms in machine learning, and *Zhou* [2012] covers a number of them. Famous examples of ensemble models include Random Forests (*Breiman* [2001]) and Bagging Trees (*Breiman* [1996]), using bootstrap aggregation, and XGBoost (*Chen and Guestrin* [2016]), using boosting.

The form of ensemble learning that will be used in this thesis is bootstrap aggregation, or bagging. The concept of bootstrapping was introduced by *Efron* [1979] and showed an alternative way to calculate parameter estimates by sampling from the data with replacement. To explain by example, consider estimating the variance of some data. Traditionally, the variance would just be expressed as the difference of the expected value of the square from the square of the expected value, e.g. $\sigma^2 = E[X^2] - E[X]^2$. Using bootstrapping, one would create multiple bootstrap samples of length N from the data and calculate the variance for each bootstrap sample. This yields a distribution of estimates for the variance and a good estimate would be to simply take then mean.

Bootstrap aggregation, or "bagging" involves creating multiple training datasets by sampling from the original training set with replacement (bootstrapping) and training a model on each dataset so that all the model's results can collected (aggregation). The aggregation of model results can be done differently, either explicitly keeping the distribution of predictions or simply taking the average. Ensemble methods using bagging may utilize this concept in different ways. While Bagging Trees explicitly use bagging in this way,

Random Forests are distinct in that they randomly select a subset of features for subsets of each tree to learn from so that the generated trees possess less correlation with each other (making them better predictors).

For some ensemble methods, different metrics may be used for constructing the optimal ensemble. One such method is the Out-of-Bag (OOB) error for a single trained model of the ensemble. The training data that was not included in the bootstrap sample is referred to as the OOB sample, and the model's error on it is computed as the OOB error. The full OOB error (that is, the OOB error across all the models) may be computed in different ways, such as a simple or weighted average of the predictions vs the true answer using MSE. Another method is to use a validation set that is distinct from the training set and to calculate the errors on the validation set.

Determining the proper number of members of an ensemble is also done in different ways, dependent upon the models or methods being used. One is ensemble growing in which the ensemble, starting from a single model, has more models added to it until the ensemble error begins to worsen. An opposite approach is ensemble pruning where a suite of models are trained and models are removed until the ensemble error begins to increase.

Chapter 3

Source Data

With respect to the final results, there are two forms of source data used. One is the data used to build an unsupervised classification model of the plasma regions in and around the magnetosphere and will be covered first in Section 3.1. Next are the data used to create a new bow shock model covered in Section 3.2. The latter are taken from a variety of spacecraft as well as the results of another spacecraft region classifier.

3.1 Unsupervised Classifier Source Data

We use data from two missions, Time History of Events and Macroscale Interactions during Substorms (THEMIS, *Angelopoulos* [2008]) and the Magnetospheric Multiscale Mission (MMS, *Burch et al.* [2016]). These datasets includes measurements of magnetic field $\mathbf{B}$, the ion velocity $\mathbf{V}$, the ion scalar temperature T, and the ion density n, a cumulative eight features. We show how the data is prepared for both missions.

3.1.1 THEMIS

THEMIS is a collection of five spacecraft (THEMIS-A, B, C, D, and E) with equatorial orbits with the purpose of observing different aspects of magnetic storms and substorms. We used data from March 2007 to the end of December 2020. THEMIS-B and C were moved to lunar orbit in 2009 to become the Acceleration, Reconnection, Turbulence, and

Electrodynamics of the Moon's Interaction with the Sun (ARTEMIS, *Angelopoulos* [2014]) mission where they would make measurements departing from what would normally be seen by THEMIS-A D, and E. We only use THEMIS-B and C data up until end of year 2009.

The ion velocity, temperature, and density measurements of THEMIS are from the Electrostatic Analyzer instrument (ESA, *McFadden et al.* [2008]) and are available at multiple time resolutions, such as "reduced" (ESAR) and "full" (ESAF) data packets. The ESAR offers higher time resolution at once per spin (∼3 secs), but the cold temperatures of typical solar wind mean that their distributions are narrow and require sufficiently high angular resolution to resolve. The ESAF packets sacrifice time resolution for higher angular resolution and are available in two formats, 32-spin (96 sec) in fast survey mode and 128-spin (∼6.5 minutes) in slow survey mode. Figure 5 of *McFadden et al.* [2008] illustrates the difference in angular resolution. The data are flagged for quality and we use quality zero data, indicating no issues. The magnetic field measurements are from the Flux Gate Magnetometer (FGM, *Auster et al.* [2008]) and are collected at spin resolution. This data is then averaged down to the resolution of the ESAF measurements to synchronize them, as illustrated in Figure 3-1.

3.1.2 MMS

MMS is a constellation of four spacecraft (MMS-1, 2, 3, and 4) flying in equatorial orbits in mutual close proximity to make electron-scale measurements. The ion measurements are taken from the Dual Ion Spectrometer (DIS) as part of the Fast Plasma Investigation (FPI, *Pollock et al.* [2016]) suite. Multiple ion spectrometers per spacecraft makes it possible to make measurements below spin resolution. The magnetic field measurements are taken from the Flux Gate Magnetometer (FGM, *Russell et al.* [2016]) and are available at 10 ms. These magnetic field measurements and ion measurements are averaged down together to

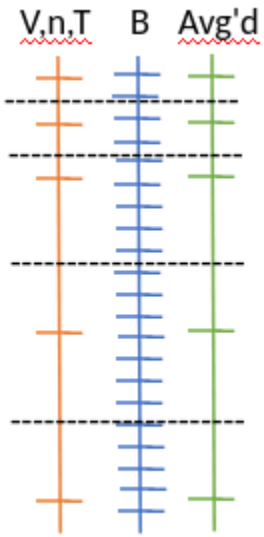

Figure 3-1: Basic graphic visualizing how the THEMIS data is averaged down across different time resolutions. The ESAF data products (**V**, n, and T) have two different time resolutions but the FGM products (**B**) used have constant resolution. The middle time is selected between adjacent points and the new data point is created by averaging over all data in that interval. The joined data (Avg'd) will then have similar resolution as the ESAF products.

1 minute resolution. Data from MMS 1, 2, and 3 span September 2015 to December 2021. Due to damage to the spectrometers of MMS 4, we only use data from September 2015 to 7 June 2018.

3.1.3 Data Cleaning

The THEMIS and MMS datasets possess 8.13 and 4.09 million points, respectively. The methods we apply to these data can be very sensitive to outliers and the size of magnetic field measurements closer to Earth could impact our ability to separate them in an unsupervised manner, so we constrain our data to be between 7 and 35 Earth radii. This final filtering leaves us with 9.64 million points, 4.09 million (42.4%) being MMS and 5.55 million (57.6%) being THEMIS. We separate our data with a test-train split of 95% and 5%, giving us a training size of ∼482k points. The distributions of the magnetic field, ion velocity, ion

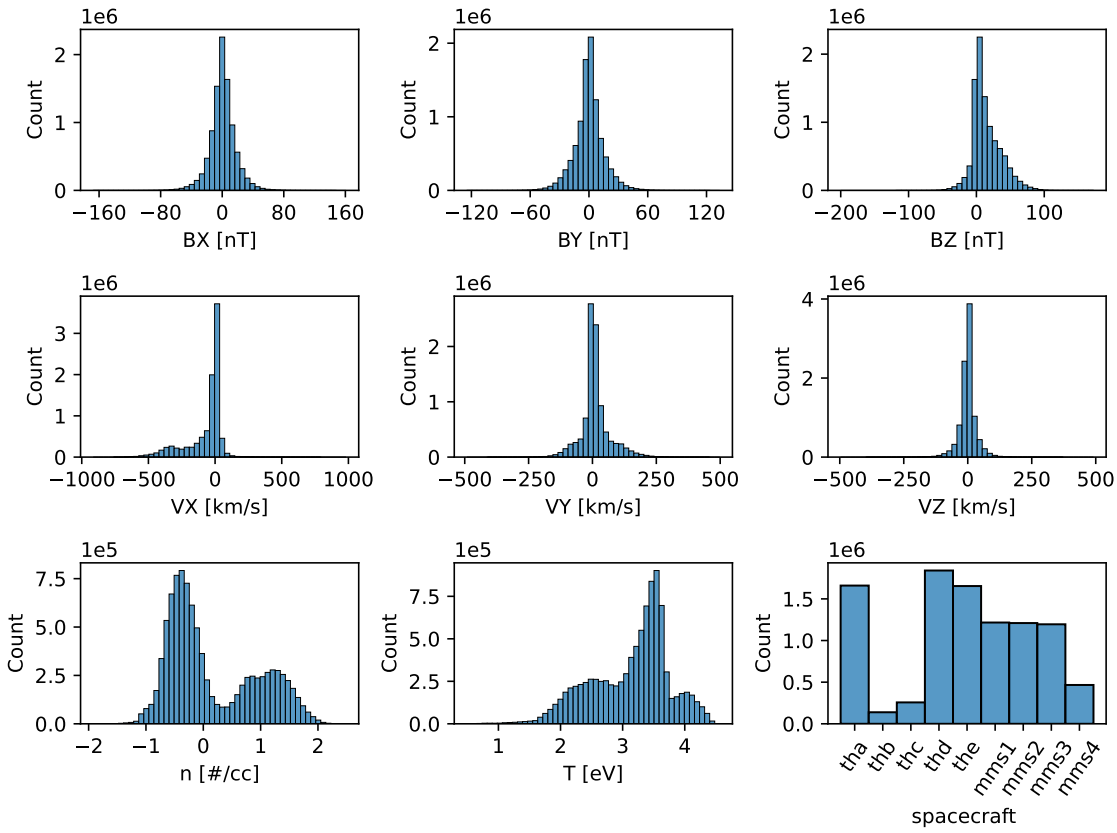

Figure 3-2: The distributions of all data collected, both training and testing. It is apparent from the density and temperature distributions (both in log10 scale) that multiple populations are present: Sparse (0.1 #/cc < n < 1 #/cc), moderate-density (1 #/cc < n < 30 #/cc), and dense (n > 30 #/cc) plasma and very cold (T < 10's eV), warm (10's eV < T < 1 keV) and hot (T > 1 keV) plasma. These different peaks in distributions are ideal for clustering.

density, and ion temperature measurements are shown in Figure 3-2. The data described thus far will be used to build a classifier to classify THEMIS and MMS observations into solar wind, magnetosheath, or magnetosphere.

3.2 Bow Shock Model Source Data

To estimate the position of the bow shock as a function of upstream solar wind parameters, we need a way of getting reliable upstream solar wind estimates. For THEMIS and MMS, we cannot simply take the solar wind-classified data in proximity to a bow shock crossing as

such observations are contaminated by the ion foreshock in quasi-parallel regions, meaning that such measurements will be hotter and more turbulent, affecting quantities such as the plasma beta and magnetic clock and cone angles. OMNI data will be used to provide more accurate estimates of these quantities for some spacecraft missions.

3.2.1 Cluster

Cluster (*Escoubet et al.* [2001]) is a constellation of four spacecraft (termed Cluster-1, 2, 3, and 4) orbiting in highly non-equatorial orbits so as to make measurements in the polar cusp. The magnetic field measurements are taken from the Flux Gate Magnetometer (FMG, *Balogh et al.* [1997]) at spin-averaged resolution (∼4.5s) and then averaged down to 1 minute resolution. The ion measurements are taken from the Hot Ion Analyzer (HIA) which is part of the Cluster Ion Spectrometry experiment (CIS, *Rème et al.* [1997]) and are spin-averaged to be available at spin resolution. Due to technical failures and anomalies, the data used here are only for Cluster 1 (up to 1 April 2011) and Cluster 3 (up to 11 November 2009). The quality of the measurements are represented with integers in which a three denotes a measurement with no known errors, a two indicates that minor issues are present, and a four represents specially cleaned data. We take all data with quality flags $\geq$ two and average them down to 1 minute resolution. These data are conjoined with the prepared magnetic field data to yield 7.83 million points.

3.2.2 OMNI

OMNI (*King and Papitashvili* [2005]) is a database created from 20 spacecraft, including Imp 8, Wind, ACE, and Geotail, containing estimates of upstream solar wind that impacts the Earth's magnetosphere. These estimates are determined from the planar propagation method, the key assumption being that the magnetic field and plasma observations measured upstream correspond to a point on a plane that propagates at the speed and direction

of the velocity vector V. Taking these observations as points of phase fronts measured at time t and position Ro, determining when this phase front meets a point downstream at time t' and position Rd is simply a ballistics problem. Rd is taken to be the bow shock nose, which is calculated using both the Shue magnetopause (*Shue et al.* [1998]) and Farris and Russell bow shock (*Farris and Russell* [1994]) models (accounting for the +30 km/s aberration in the Y direction in GSE). n is the phase front normal (PFN) and is determined using Minimum Variance Analysis (MVA, *Sonnerup and Cahill Jr.* [1967]). All these parameters known, the time shift of these phase fronts is then given by $\Delta_t = t' - t = n(Rd - Ro)/nV$.

Note that two immediate shortcomings of this approach is (1) the planar propagation technique can't account for curvature and (2) estimates for data that are far away from the bow shock nose (e.g. nightside flank measurements of Imp 8) may be inaccurate, although the time delay between these positions may be slight. There have been attempts to provide better estimates than the phase-front propagation technique (*O'Brien et al.* [2023]).

These data offer magnetic field and plasma observations as well as geomagnetic indices. They are made available at multiple resolutions, including 1-hour, 5-minute and 1-minute. We use the 5-minute resolution dataset. In addition to the primary publication, a detailed description of how the 1-min and 5-min resolution OMNI datasets are prepared can be found at https://omniweb.gsfc.nasa.gov/html/omni_min_data.html

3.2.3 Imp 8

Imp 8, also known as Explorer 50, was a single spacecraft containing instrumentation to measure magnetic fields, plasma flow, and energetic particles. Intended to make observations in both the solar wind and magnetotail, its average orbital radius was about 35 R_E. Its exact orbital patterns varied throughout the mission with largest apogee and smallest perigee being 45 and 22 R_E, respectively. An overview of history and instrumentation of

the spacecraft can be found in *Paularena and King* [1999].

3.2.4 Geotail

Geotail was a spacecraft whose primary science goal was to investigate the structure and dynamics of the geomagnetic tail. It possessed instrumentation to measure magnetic and electric fields, plasma flows, energetic particles, and plasma waves. An introductory overview and brief description of its instrumentation and orbits can be found in *Nishida* [1994].

3.2.5 Magion-4

Magion-4 was a subsatellite of the Interball project, a four spacecraft mission. Interball was comprised of two pairs of spacecraft: Interball-1 and Magion-4 as the "tail" pair and Interball-2 and Magion-5 as the "auroral" pair. All spacecraft had highly elliptical, polar orbits. The scientific intent of the mission was to investigate the global dynamic characteristics of magnetospheric processes as well as small scale features of these processes at the key plasma regions of the Earth's magnetosphere and its neighborhoods. It had instrumentation to measure plasma flows, electric and magnetic fields, plasma waves, solar radio and X-ray emissions, and UV auroral emissions. A description of the instruments, mission objectives, and early observations can be found in *Zelenyi et al.* [1997].

3.2.6 Wind

Wind is a spacecraft orbiting the L1 Lagrange point (with an X_{GSE} / Y_{GSE} displacement of $\sim$ 35 and 100 R_E from the Sun-Earth line). It, along with ACE, acts as an upstream monitor of the solar wind and is used to provide the upstream data for OMNI calculations. It has instrumentation to measure magnetic fields, radio and plasma waves, solar wind plasma composition, 3d distributions of ions and electrons, energetic particles, and gamma rays. An brief description of the spacecraft, its instruments, and its early results can be

	THEMIS-A	B	C	D	E	Total
Original	222	166	264	363	294	1,309
OMNI-aligned	191	155	243	324	260	1,173

Table 3.1: THEMIS Crossings

	MMS-1	2	3	4	Total
Original	535	534	535	134	1,738
OMNI-aligned	306	307	310	131	1,054

Table 3.2: MMS Crossings

found in *Ogilvie and Desch* [1997], and a reviews of its contributions to the space physics community can be found in *Wilson III et al.* [2021].

3.2.7 Bow Shock Crossing Preparation

Our determination of a bow shock crossing from data and the resulting collection of upstream solar wind parameters changes across the different datasets. We outline the three methods below. Note that for methods where we align crossing times with OMNI, we remove crossings containing any amount of NaNs among the parameters of interest (magnetic field vector, plasma beta, magnetosonic mach number, and dynamic pressure).

THEMIS + MMS

Using our unsupervised classifier, we infer the times of 3,047 bow shock crossings from classifications of THEMIS and MMS data. The times are then aligned with 5-minute resolution OMNI data to have estimates of upstream solar wind conditions during the times these spacecraft cross the bow shock. The positions of the spacecraft are retained and conjoined with the OMNI estimations of pristine solar wind magnetic field components, magnetosonic mach number, plasma beta, and dynamic pressure. The exact procedure for constructing these crossings are given in section 7.2. The crossing counts per spacecraft for both THEMIS and MMS missions are seen in Tables 3.1 and 3.2.

Cluster

The bow shock crossing times from Cluster 1 and 3 are taken from *Nguyen et al.* [2022] wherein they used their classifier to classify a variety of spacecraft observations into solar wind, magnetosheath, and magnetosphere. They reported 3,225 and 2,005 (5,229 in total) from Cluster 1 and 3, respectively. Specifically, they are retrieved from https://zenodo.org/records/5668298#.ZFedt0jMJYh. To determine these crossings from their classifications (as well as bow shock crossings from other spacecraft in their dataset), they defined a bow shock crossing event as a 10 min interval that contains as much magnetosheath points as solar wind points. Their Cluster bow shock catalogue records the crossings according to crossing number (for the particular spacecraft), the start and end of the 10 minute interval, and the probability they estimate as a bow shock crossing occurring within the interval. Upon manual inspection of the Cluster 1 catalogue, it is found to contain 567 redundant entries. These are redundant in that the same time intervals are repeated from earlier crossing events. No redundancies are present in the Cluster 3 catalogue. Thus, the number of unique bow shock crossings for Cluster 1 is 2,658 with a revised total of 4,663.

To select a single time to represent the crossing, we select the middle time of each 10-minute interval. Then using the Cluster dataset we originally prepared for our classifier (but ultimately did not use), we extract the positions of Cluster 1 and 3 according to these middle times. Our Cluster dataset was created by preserving data with quality flags of 2 (indicating minor problems) and above (3 for no issues, 4 for particularly cleaned). When we align the middle times of these bow shock crossing intervals with our Cluster dataset, we recover 4,543 data points, indicating that the remaining 120 points were inferred from poor-quality Cluster data.

With these 4,543 crossings for which we have the Cluster spacecraft positions, we now

	Cluster 1	Cluster 3	Total
Nguyen et al. [2022]	2,658	2,005	4,663
Cluster Quality ≥ 2	2,610	1,933	4,543
OMNI-aligned	2,476	1,848	4,324

Table 3.3: Cluster Crossings - Note that the crossings reported here for *Nguyen et al.* [2022] do not include the repeated entries.

align these with 5-minute resolution OMNI data. We find that the number of crossings for which we have available OMNI data is 2,476 and 1,848 crossings for Cluster 1 and 3, respectively (with 4,324 in total). A table showing the reduction in crossing counts across these filters is shown in Table 3.3.

Imp 8 + Geotail + Magion-4

The crossings from Imp 8, Geotail, and Magion-4 were pre-prepared and hosted in the Space Physics Data Facility database. They additionally contained some crossings for Cluster 1, but we did not include these since we have already prepared crossings from Cluster. The spacecraft data cover from 1973 - 2000 for Imp 8, 1995 - 1997 for Geotail, and 1995 - 1996 for Magion-4. A statement indicating those responsible for the creation of the dataset is taken directly from the webpage and included here exactly:

> "The basic IMP 8 observed parameters were provided by IMP magnetometer and plasma teams at GSFC (A. Szabo, J. Merka) and MIT (J. Richardson, K. Paularena). The database will grow to encompass all 1973-2001 IMP 8 bow shock crossings. Geotail bow shock crossings currently include 1995, 1996 and 1997 at the bow shock flanks but will grow to include all crossings from 1995 onward. The Geotail crossing were identified by R. Kessel (NSSDC). Two years worth - 1995, 1996 - of Magion-4 (Interball-Tail's sub satellite) bow shock crossings were supplied by J. Safrankova and Z.Nemecek. Cluster bow shock crossings were determined from the prime parameters by R. Kessel and students.

	Imp 8	Geotail	Magion-4	Total
Original	11,455	813	834	13,102
Time span $\leq$ 10 min	10,036	n/a	n/a	n/a
NaNs removed	6,937	810	825	8,572

Table 3.4: Crossings for Imp 8, Geotail, and Magion-4. Note that the time span variable is only used for Imp 8 and measures the time gap between adjacent observations in which a bow shock crossing was inferred to have occurred between.

> Cluster crossings are currently supplied for the first two years of operation - 2001 and 2002 - from one satellite, but will grow to include subsequent years."

These data include variables that are only defined for particular spacecraft. The Imp 8 data include the time gap (called "Timspan" in the data itself) between consecutive observations in which a bow shock crossing was inferred. In processing this data, we remove all crossings for Magion-4 and Geotail that contain NaNs. We repeat this for Imp 8 crossings, but also require that the time span be $\leq$ 10 minutes (to avoid crossings inferred from large gaps). This reduces the number of crossings from 13,102 to 11,683. Removing the NaNs from the dataset reduces this further to 8,572.

The way these crossings are connected to upstream solar wind conditions is done differently per spacecraft. Imp 8 serves as its own upstream solar wind monitor such that the reported upstream conditions are taken directly from Imp 8 when it was in the solar wind. Geotail and Magion-4 use 10-minute averaged upstream Wind data that are time-shifted to the approximate bow shock nose location at 14 R_E according to the solar wind V_X component and the X_{GSE} position of Wind using to $t = (X_{WIND} - 14R_E)/V_X$, a cruder form of how OMNI time-shifts L1 observations. Note that, just as with OMNI estimates, time-shifts for crossings occurring far out along the flanks can involve greater errors due to the greater distance. A table of the crossing counts for the original dataset, the time span filtered dataset, and for NaNs being removed is given by Table 3.4.

All Crossings Together

Merging all previously mentioned cleaned crossing datasets results in 15,123 bow shock crossings. All but one crossing has a radius $\leq 50R_E$ with this one crossing belonging to Magion-4 and possessing a very large radius of 106.8 R_E. We remove this single crossing, reducing the dataset to 15,122 points. Henceforth, when there is reference to the "original" crossing dataset, we shall be referring to this one.

Last, our intent is to use a physics-informed neural network approach to the construction of a bow shock model. Neural networks can struggle with performing regression on data that correspond to extrema for one of the training features or ranges not seen during training. To deal with this, we take a rather blunt approach of removing the first and last 0.5 percentiles of data (inclusive) for the plasma beta, magnetosonic mach number, dynamic pressure, and BZ (that is, the bottom 0.5% and top 0.5% of each of these features is removed) of the upstream estimates. This is not applied to the magnetic cone and clock angles as they are periodic features. This removal leaves 14,664 crossings. A table of the resulting crossings per spacecraft is seen in Table 3.5 and histograms of the modified features are shown in Figure 3-3. Any remaining modifications of the data and how it is prepared for the development of a bow shock model will be covered in Chapter 8, after the development and validation of the unsupervised classifier.

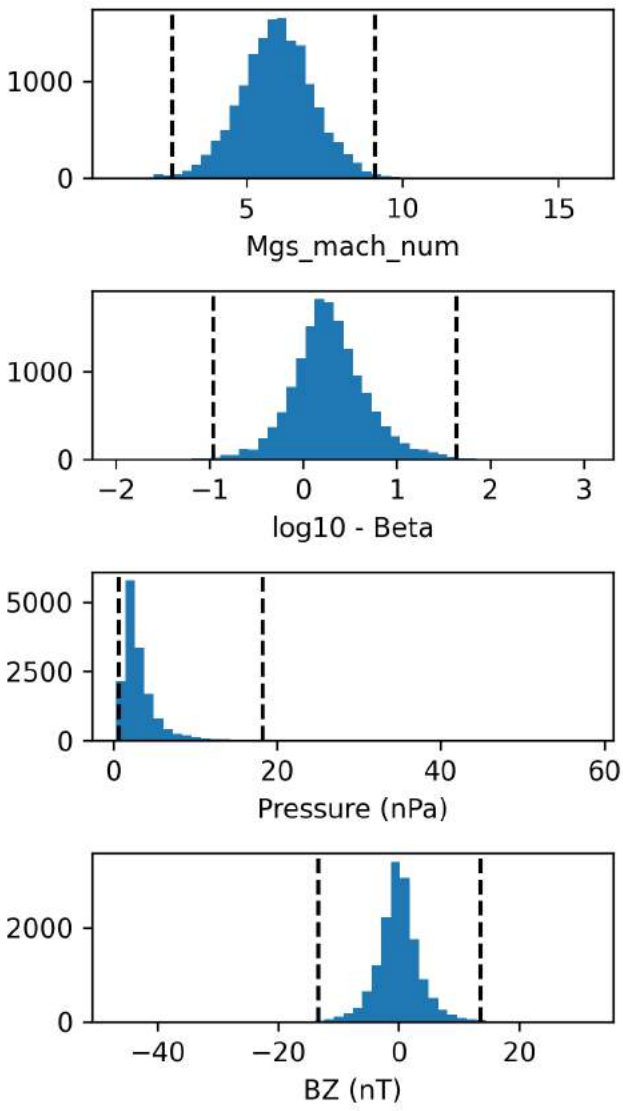

Figure 3-3: Histograms of the magnetosonic mach number, plasma beta (in log10 scale), dynamic pressure (in nPa), and BZ (in nT). The min and max values along the x axis for each figure indicate that data exist with those extreme values but are very rare in the original dataset. Note that for the plasma beta, this implies points with $\beta = 0.01$ and 1000 in linear scaling. Applying the 1% extrema removal results in truncating the data along the dashed black lines. These limits are [2.60, 9.10] for the magnetosonic mach number, [0.11, 43.38] for the plasma beta in linear scale ([-0.96, 1.64] in log10 scale), [0.64, 18.24] for the dynamic pressure, and [-13.33, 13.60] for B_Z.

	Original*	1% Extrema Removed
THEMIS-A	191	186
B	155	150
C	243	240
D	324	319
E	260	253
MMS-1	306	303
2	307	303
3	310	304
4	131	129
Cluster-1	2,476	2,425
3	1,848	1,819
Imp 8	6,937	6,623
Geotail	810	799
Magion-4	824	811
Total	15,122	14,664

Table 3.5: * Note that we have removed the single 100+ R_E crossing due to Magion-4 here.

Chapter 4

Unsupervised Classifier Data Preparation

The eight variables $\mathbf{V}$, $\mathbf{B}$, n, and T, hereafter referred to as features, of our dataset do not possess enough variance for many unsupervised methods to sufficiently separate the regions. It is very common within machine learning to engineer derived features from the original in hopes of capturing non-linear relationships (*Horn et al.* [2020]) because what is non-linearly separable in some space might become linearly separable in a higher dimensional space (see the example in Figure 4-1). To this end, we include the ion speed V, the magnetic field magnitude B, and the ion momentum density, $\mathbf{mom} = n\ \mathbf{V}$ (with ion mass set to 1), as five additional features, giving us a total of 13 features. The addition of the ion momentum density vector is to help better separate the magnetosheath from the solar wind and magnetosphere as it acts as a transition region between them.

Most of the features have ranges over a few orders of magnitude whereas the density, temperature, and momentum density components cover more than several. We convert the density and temperature to log10 scale, but the same cannot be done for the momentum density due to the negative values. This is circumvented by transforming the momentum density using the log10 of the absolute values of their components instead. After, these data still possess uneven ranges that can impact the performance of the dimensionality reduction

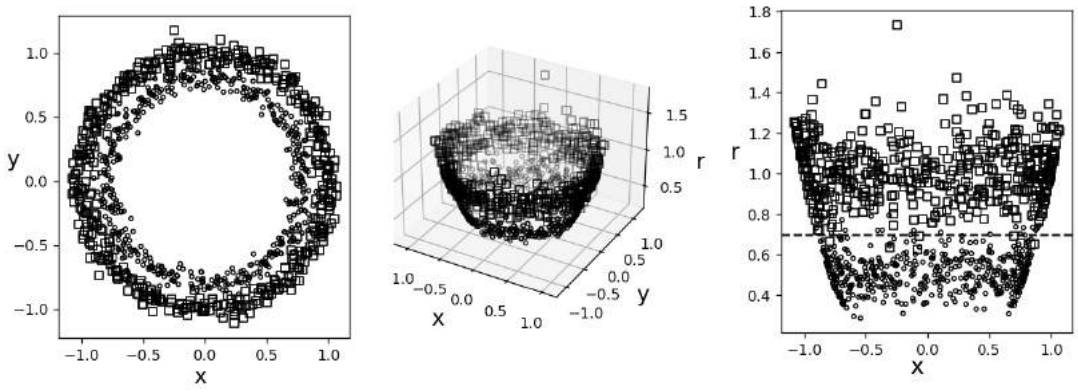

Figure 4-1: A basic example demonstrating the importance of feature engineering. Two noisy concentric circles (plotted in the left figure with different markers) cannot be separated from each other linearly according to their x and y positions alone. Creating the feature $r = \sqrt{x^2 + y^2}$ and plotting (x,y,r) shows a correlation of one circle with higher values of r. This is seen more clearly in a plot of (x,r) in the right figure where a line can separate the circles to high accuracy.

and clustering methods we will use. To avoid feature bias, we rescale our training data using min-max normalization such that the new minimum and maximum of each feature is 0 and 1, respectively. The distributions of this rescaled training data is shown in Figure 4-2.

Non-negligible feature correlation is certain given our choice of features and this is evident in the correlation heatmap of Figure 4-3. The high number of correlated features means that direct clustering methods would be biased in the favour of these correlated components. Further still, the dimensionality can make some methods computationally expensive or cause them to find poor solutions due to the curse of dimensionality. The implication of the latter here is that distances between points will become smaller as the dimensionality increases, reducing the quality of clustering solutions. For data that does not possess significant outliers or that has been meticulously cleaned, the loss in quality of these solutions may be small, but it can become an issue for noisy data, especially data that are observations. These issues will be addressed in the following chapter so that an accurate unsupervised classifier can be constructed.

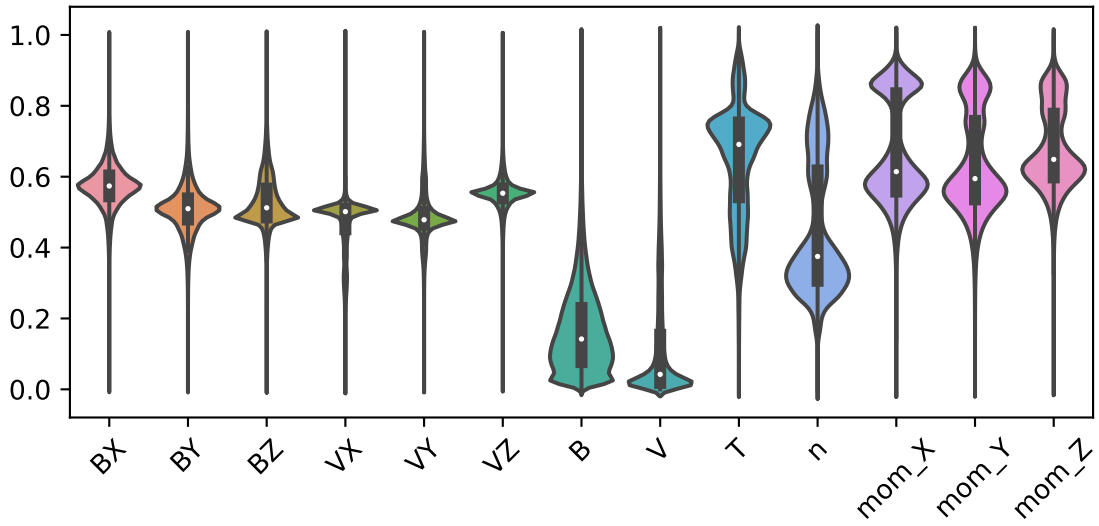

Figure 4-2: Violin plots representing the distributions of input features of our min-max scaled training set. The violin plots here show the kernel density estimate (KDE) as the width, the range of the estimate as a thin vertical grey bar (for example, all features here have this thin vertical grey bar reaching from 0 to 1), the interquartile range as a thick vertical black bar, and the median as a white dot. The KDE for each variable is scaled according to the width so that the distributions are more visible.

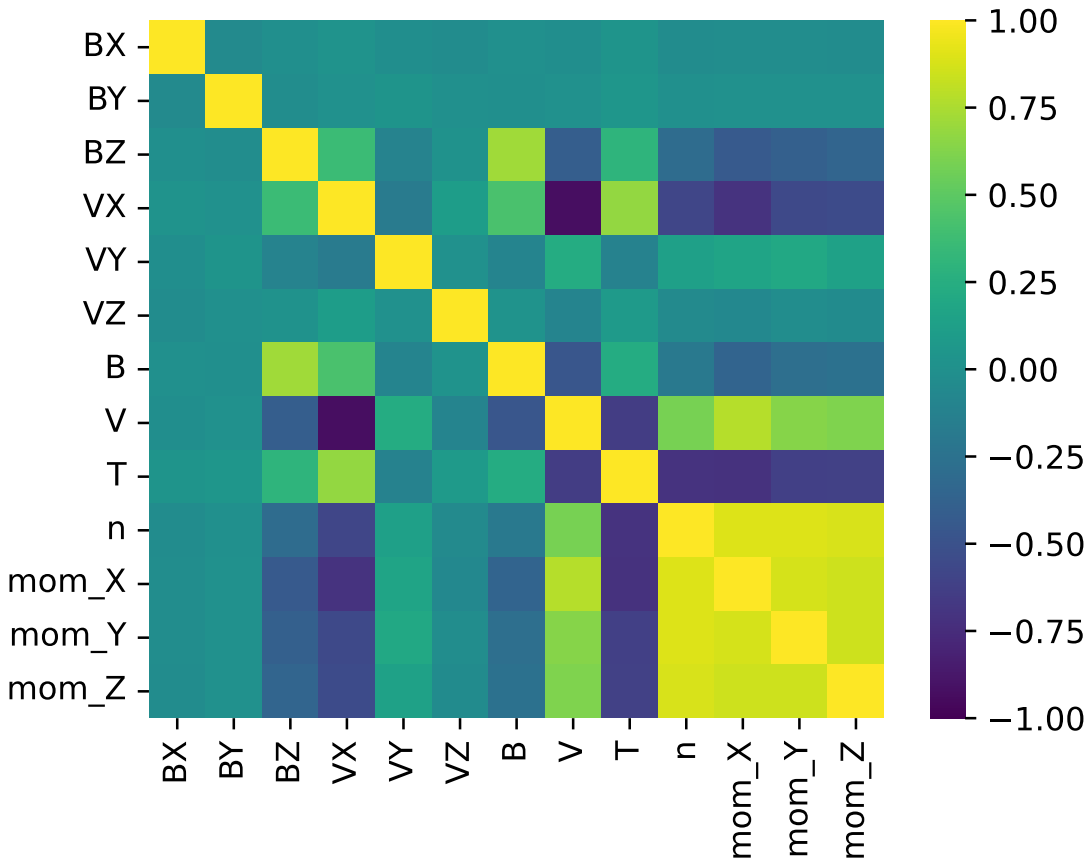

Figure 4-3: A heatmap of the correlations between variables in the min-max rescaled training set. The plot is symmetric across the diagonal. It is to be interpreted as showing the correlation of each feature with every other feature in the training set, e.g. correlation(log10(n), log10(T)) $\sim$ -0.6, or the log10 of the density is moderately negatively correlated with log10 of the temperature. There is a visible number of variable pairs with large magnitude in correlation (the bright or dark colored boxes in the off-diagonal). Also apparent is the absence of correlation of BX and BY with all other variables - even with B. This is because the distributions of BX and BY are symmetric around 0, which is visible in Figure 3-2. The VX and VY components have correlation with V because large speeds (>350 km/s) are often going to be associated with solar wind, which generally possess large negative magnitudes in VX and slightly positive VY, on average ($\sim$30 km/s), due to the angle that the solar wind arrives at the Earth. Lastly, note the positive correlation between VX and T. This is legitimate as the lowest values of VX occur in the solar wind, which is characterized by the lowest temperatures; more moderate values of VX and T occur in the magnetosheath; the largest (read, most positive) values of VX occur in the magnetosphere, which also possesses the largest temperatures.

Chapter 5

Building the Unsupervised Classifier Pipeline

The unsupervised classifier we create requires multiple methods to function. We outline here the full pipeline of methods that allow us to take our 13-dimensional data and classify it into magnetosheath, magnetosphere, and solar wind. Analyses and validation of the model results are carried out in the subsequent chapter. The methods covered here include PCA, SOMs, K-Means, and Hierarchical Agglomerative Clustering.

5.1 Reducing Dimensionality with PCA

We possess a dataset of thirteen features, some of which have correlations with each other as seen in Figure 4-3. Applying clustering techniques to this data outright could create biased solutions due to both the presence of these correlations and the curse of dimensionality making it hard to resolve outliers correctly. We rectify this by applying PCA to the data to recover uncorrelated components that represent the majority of the variance of the data. The decomposition of the data is seen in Figure 5-1 and shows two subplots, one describing the proportions of variance associated with each component (left) and the other showing the resulting biplot of the data (right). We choose a 90% variance cutoff as our threshold, meaning that we extract six components from the decomposition. The biplot shows that

the left, top right, and bottom right areas are associated with higher temperature, higher density, and higher speeds, respectively, thus making it likely that these clusters are the magnetosphere, magnetosheath, and solar wind populations.

While this decomposition addresses both the correlations and dimensionality of the data, there is still the matter of a large training size after the PCA transform. This size can be reduced by simply randomly selecting fewer points, but this will only trade variance for sample size. Choosing enough points to represent a similar amount of variance will still require a large population size. In the next section, we will use a Self-Organizing Map (SOM) to create distinct points that can act as "representatives" of their local distributions such that their amalgamation reflects the distribution of the training set.

5.2 Vector Quantization via Self Organizing Maps

5.2.1 Implementation

There are several open-source python packages implementing SOMs available. The most common is minisom (*Vettigli* [2018]), which uses a vectorized design to speed up computations. For large datasets or network sizes, the time to completion may still be quite long. Traditionally, training an SOM has been a computationally expensive process for two reasons: The network adapts to one point at a time, and it is fairly common that multiple trainings are done. The latter occurs because SOM initialization and training are done stochastically and there is a large number of hyperparameter choices available (the number of iterations, the network size, the decay function, the neighborhood function, the initial and final learning rate and neighborhood size, etc). Since the network with the lowest quantization error is usually selected as the best fitting, this significantly increases the total amount of time needed to get a complete and robust model.

The one-at-a-time training constraint is resolved by using SOMs that train over batch-

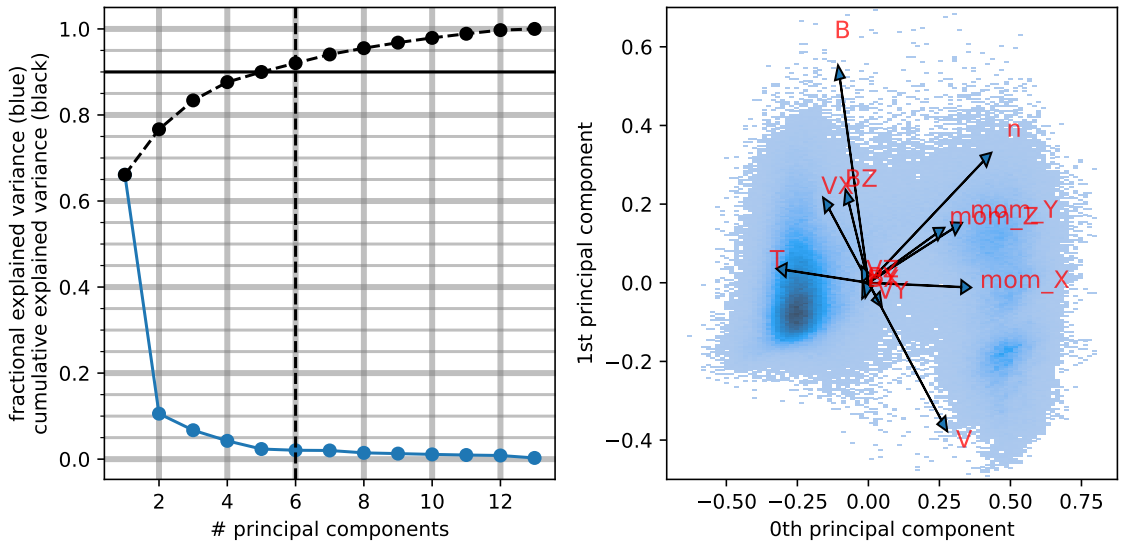

Figure 5-1: **Left**: The normalized eigenvalues from the PCA decomposition are plotted in descending order as the solid blue line. The cumulative sum of these normalized eigenvalues is plotted as the dashed black line. We choose to select a number of components representing at least 90% of the variance (the horizontal black line), so 6 components are chosen that represent 93% (the vertical dashed black line). **Right**: A bivariate histogram of the training data projected onto the first two principal components, representing 76% variance. It is evident from the first two components that several clusters are present in the data. The arrows plotted here are the loadings for our features across the first two principal components. The length of an arrow represents the influence that feature had for the PCA projection along that direction. All arrow lengths are normalized to the longest arrow, that of the B feature. From the plot, the temperature feature, T, significantly influenced the 0^{th} component but barely for the 1^{st} and points to the cluster on the left. This means that that cluster is likely to correspond to higher temperatures than the data on the right. The density, n, roughly equally contributed to both components and indicates that the top right region is related to higher densities and by its antiparallel direction, the cluster on the left is largely associated with lower densities. Since VX points to the top left and V to the bottom right, the bottom right region is related to data with high speeds and large negative values of VX. The BX, BY, VY, and VZ features are clustered at the origin, indicating that they did not influence the first two components (although they may have impacted the higher order components). Overall, we can surmise from this plot alone that the left, top right, and bottom right areas are associated with higher temperature, higher density, and higher speeds, respectively. Thus, it is likely that these clusters are the magnetosphere, magnetosheath, and solar wind populations.

updates. These usually involve computing weighted averages of the neighborhood values across a batch of samples. This approach is taken by two popular python packages Somoclu (*Wittek et al.* [2017]) and XPySom (*Mancini et al.* [2020]) and speed-up on CPU resources alone can be close to a factor of 100, sometimes greater. We have used the XPySom package for our results.

5.2.2 Hyperparameter Optimization and Training

To expedite the process of finding the best fitting SOM with the most appropriate set of hyperparameters, we create a micro training set. First, we again min-max normalize the PCA-projected training data in order to avoid bias to any particular feature. Next, we run K-Means 100 times to resolve 10,000 clusters with a K-Means++ initialization method and select the optimal run based on minimal inertia. This initialization method makes better choices for cluster centroids by weighting data in proportion to their square distance from the previously created centroid. Then for each centroid, the closest point in the training data is extracted, and the resulting 10,000 points form the micro training set. The remaining points in the training dataset are referred to as the macro training set with a size of 472k.

We consider a number of different SOM hyperparameters and that each SOM will be trained on the micro training set and validated on the macro training set. The maps are validated in this way because the macro set will contain a larger number of outliers, and given the noise evident in the biplot of Figure 5-1, resolving these outliers correctly will be crucial. The hyperparameters of the map with the lowest value for our loss function will be retained and a final SOM will be trained using these hyperparameters on the macro training set. We define our loss function to be

$$L = Q * \left(\frac{n_x n_y}{(n_x)_{max}(n_y)_{max}} + \frac{max\{n_x, n_y\}}{min\{n_x, n_y\}} \right). \tag{5.1}$$

where Q is the quantization error of the SOM, n_x and n_y are the dimensions of the 2D node grid, and $(n_x)_{max}$ and $(n_y)_{max}$ are the maximum values permitted for the x and y dimensions. The $max\{n_x, n_y\}/min\{n_x, n_y\}$ term penalizes non-square networks and will only allow for non-square maps should they provide a sizably lower quantization error.

It should be noted that the use of a custom loss function for SOM validation is critical for our purposes. With the number of training iterations and training data set held constant, increasing the map size will generally reduce the quantization error for many choices of hyperparameters. A larger map size may better represent the training data, and in many cases even the test data, than a smaller map, but a larger number of nodes and their distributions may be suboptimal for clustering methods that will fit to these nodes. This can be loosely seen as a form of overfitting, but not in the sense of a model not generalizing well to unseen data. To illustrate this concept by example, consider training a "small" map on a large dataset containing heterogeneous groups whose distributions are somewhat (but not extremely) non-convex. One might find that the distributions of the nodes mapping to these different groups are approximately spherically separable because there are few to no nodes mapping to outliers. This would be a good motivation to use K-Means to cluster the nodes for such maps. However, as the map size is increased, the node distributions will begin to better resemble the more complicated, original distribution of the data, which contains harder-to-resolve non-convex distributions that clustering algorithms like K-Means or Gaussian Mixture Models may struggle to resolve. For another example of defining a custom loss function for SOMs in space physics applications, see the loss function defined in *Amaya et al.* [2020].

The python-based optimization library Optuna (*Akiba et al.* [2019]) is used to choose hyperparameter values. The training of each SOM on the micro training set is referred to as a trial. Optuna offers a variety of samplers to generate hyperparameters choices, and we use

the Tree-structured Parzen Estimator (TPE) with independent sampling as the sampler. It generates hyperparameter choices by fitting two sets of Gaussian Mixture Models (GMM) per trial, one set for the better performing trials, $l(x)$, and another for the remaining, $g(x)$. Each set involves fitting a GMM for each hyperparameter x and the hyperparameter value selected is that which maximizes the ratio of density estimates $l(x)/g(x)$. Maximizing this ratio is consistent with choosing a hyperparameter that is simultaneously most likely to be generated by $l(x)$ (the "good" models) and least so by $g(x)$ (the "poor" models).

For our optimization, we considered the following hyperparameters. The number of nodes for the SOM grid n_x and n_y, the initial learning rate α, the initial neighborhood size σ, the neighborhood function H, and the decay function D. We have fixed the number of training epochs to be 50, the final learning rate and neighborhood size to be 0.01, and the maximum n_x and n_y dimensions to be 30. The values the hyperparameters are permitted to take are enumerated below:

1. $5 \leq n_x, n_y \leq 30$
2. $1 \leq \sigma \leq \sqrt{n_x n_y}$
3. $0.1 \leq \alpha \leq 1$
4. D: {linear, exponential}
5. H: {Gaussian, Ricker}

5.2.3 SOM Results

After 500 trials, the best hyperparameter options are $(n_x, n_y) = (14, 14)$, $\sigma = 5.518$, $\alpha = 0.843$, $D =$ exponential, and $H =$ Ricker. We train a SOM with these hyperparameters on the macro training set which completes in 7 minutes. The resulting SOM has a quantization error of 0.0702 and 0.0703 on the macro training and test sets. The loss function rounds to

0.0855 and 0.0856. With a Intel Xeon 2.90GHz E5-2690 (32 cores, 64 threads) CPU and 64 GB of RAM available, the entire process of hyperparameter optimization and final model training takes approximately 3 hours.

While the SOM we have trained has a good quantization error, there are visualization techniques we can use to further assess how well it represents the data. Since the goal of a SOM is to give a vector-quantized representation of the data, one simple approach is to create plots of the data itself with the SOM node positions overlaid. If it is an effective representation, it should roughly map to positions of high data density, both in scatter plots and histogram marginals. We show pairplots over the first three min-max scaled principal components of the test set in Figure 5-2. When scaling up the marginal histograms of the node positions to that of the marginal histograms of the test set, there is good agreement over the 0^{th} and 2^{nd} components. The 1^{st} component shows partial agreement with the node histogram, only somewhat capturing the peak in density between 0.3 and 0.4

Another method uses the ordered nature of the SOM to create a heatmap of distances between the nodes. Since the nodes of a SOM have an ordered topological relationship, we can compute the average distance between a node and its immediate neighbors and create a heatmap of these average neighbor distances. The 2D matrix of these values is referred to as the U-Matrix. The U-Matrix for the test data is shown in the top left of Figure 5-3. Moreover, since each data point can be uniquely associated with its corresponding BMU in the SOM, we can then compute the average of all data per node. This average value per node can be used to create heatmaps of the SOM for any feature from the data, as seen in the remaining plots of Figure 5-3.

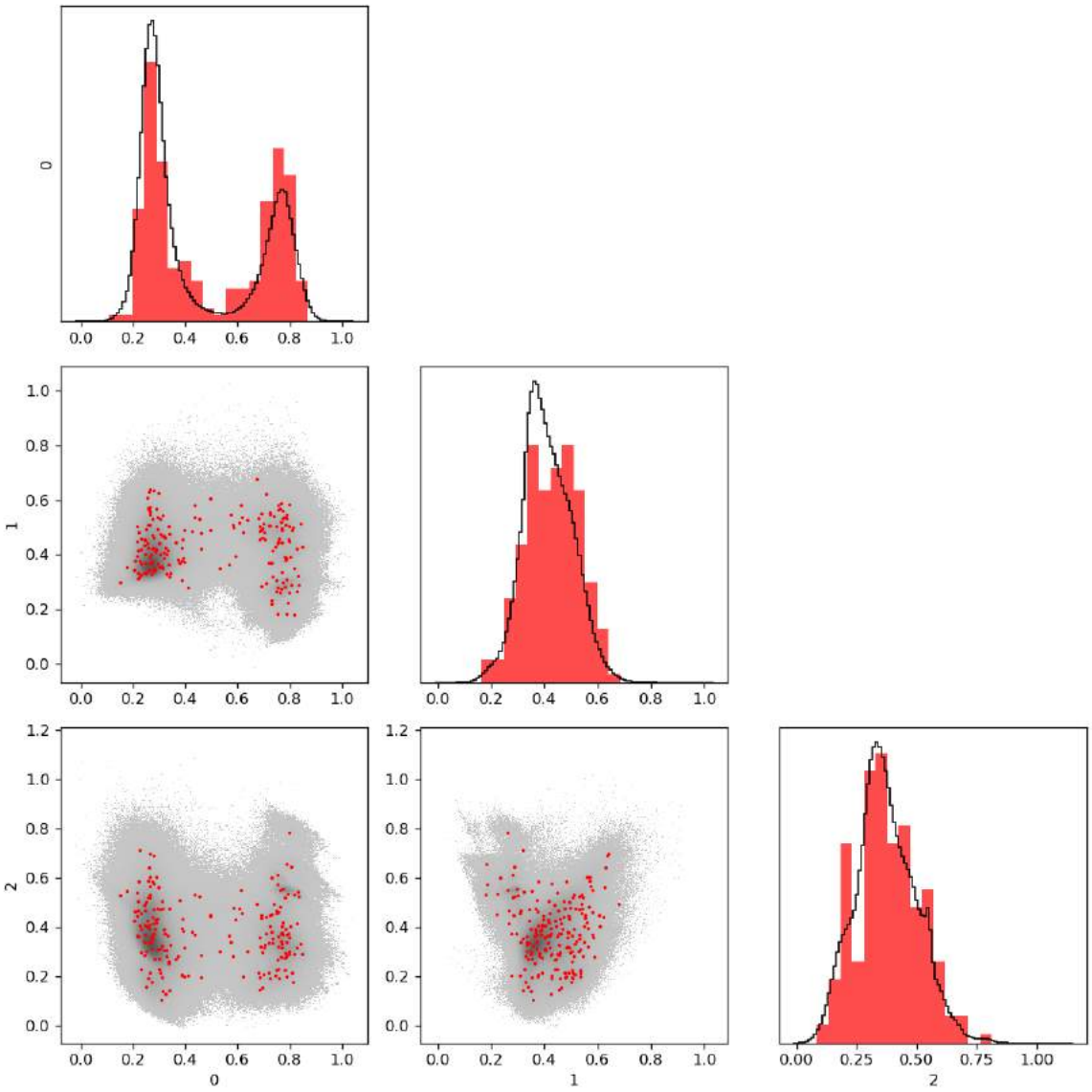

Figure 5-2: Pairplots over the first three min-max normalized principal components (83% variance) of the test set. The off diagonal plots are bivariate histograms for the test data in greyscale. Scatter plots of the SOM node position are plotted in red on top of the bivariate histograms. The diagonal plots are the marginal distributions where the black line is the test data distributed over 100 bins. The SOM node positions are simultaneously binned but at a smaller resolution of 25 bins. The nodes generally match the histograms of the 0^{th} and 2^{nd} components with a dip noticeable in the nodes histogram of the 1^{st} component.

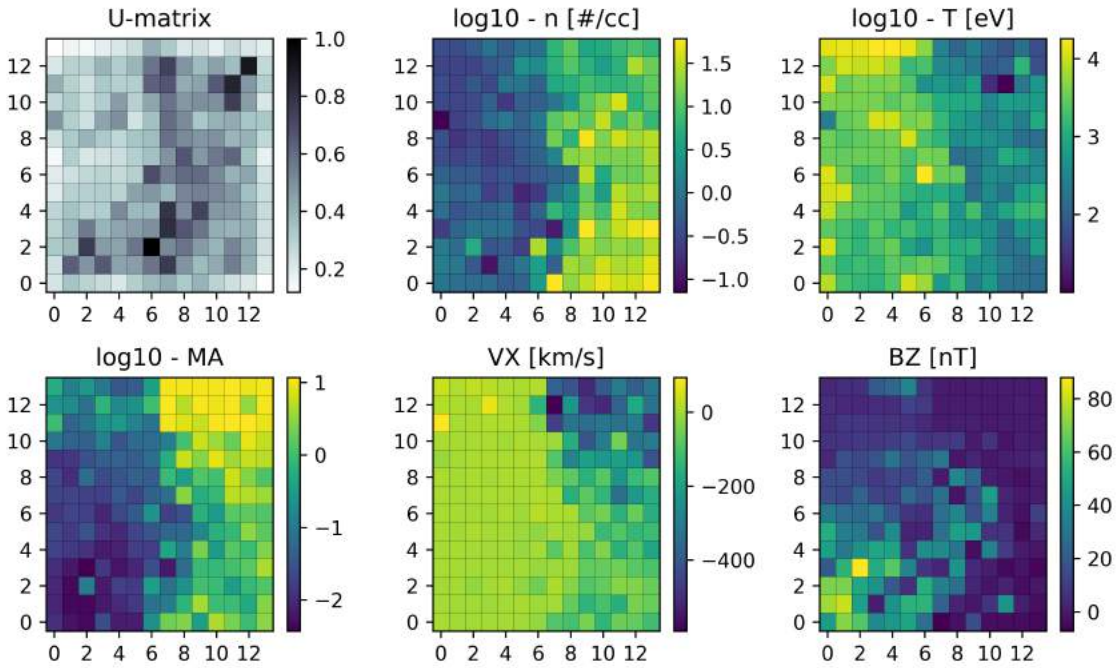

Figure 5-3: 2D heatmaps of the test data as seen through the SOM. In the U-matrix, plotted in the top left, nodes are coloured according to their distance to the nearest neighbours: the lighter nodes are more similar to the neighbours than darker nodes. Note that neighbors here is defined in the square topological sense; nodes in the corners only have two neighbors, nodes along the rest of the perimeter have three neighbors, and all other nodes have four neighbors. The fewer neighbors among those on the perimeter means that there will usually be less variance among them such that the perimeter nodes have a lower (lighter) U-matrix value. A region of dark grey nodes partitions the U-Matrix into two areas of lighter color in the top left and bottom right. This means that there are two relatively homogeneous groups of nodes. To interpret what groups of data these nodes represent, we can look at the feature maps in the remaining plots. In these plots, the average feature value per node is depicted as a heatmap. It is apparent from the feature maps that the group of nodes on the left side of the U-Matrix correspond to regions of low density and high temperature. The nodes to the right correspond to moderate-to-high densities, low-to-moderate temperatures and negative values of VX.

5.3 Hierarchical Clustering of the SOM

Applying direct clustering methods caused difficulties involving size, dimensionality, and multicollinearity. We resolved the latter two using PCA and have addressed the first by training a SOM to act as a further discretized representation of the data. With a SOM representation, we now can consider a much wider choice of methods to cluster the data as training size is no longer a constraining factor. Once a clustering method is trained, it can separate the SOM nodes automatically, classifying which nodes belong to which cluster. These node classifications can then be propagated to the data that the nodes represent, i.e. if a node A is assigned to cluster 1, then all data for which node A is the BMU will be assigned to cluster 1. We use an agglomerative, or "bottom-up," form of hierarchical clustering as implemented in the scikit-learn package with a Ward linkage in order to focus on separating clusters based on homogeneity. The entire model pipeline, including the approach used for hyperparameter optimization of the SOM, is shown in Figure 5-4.

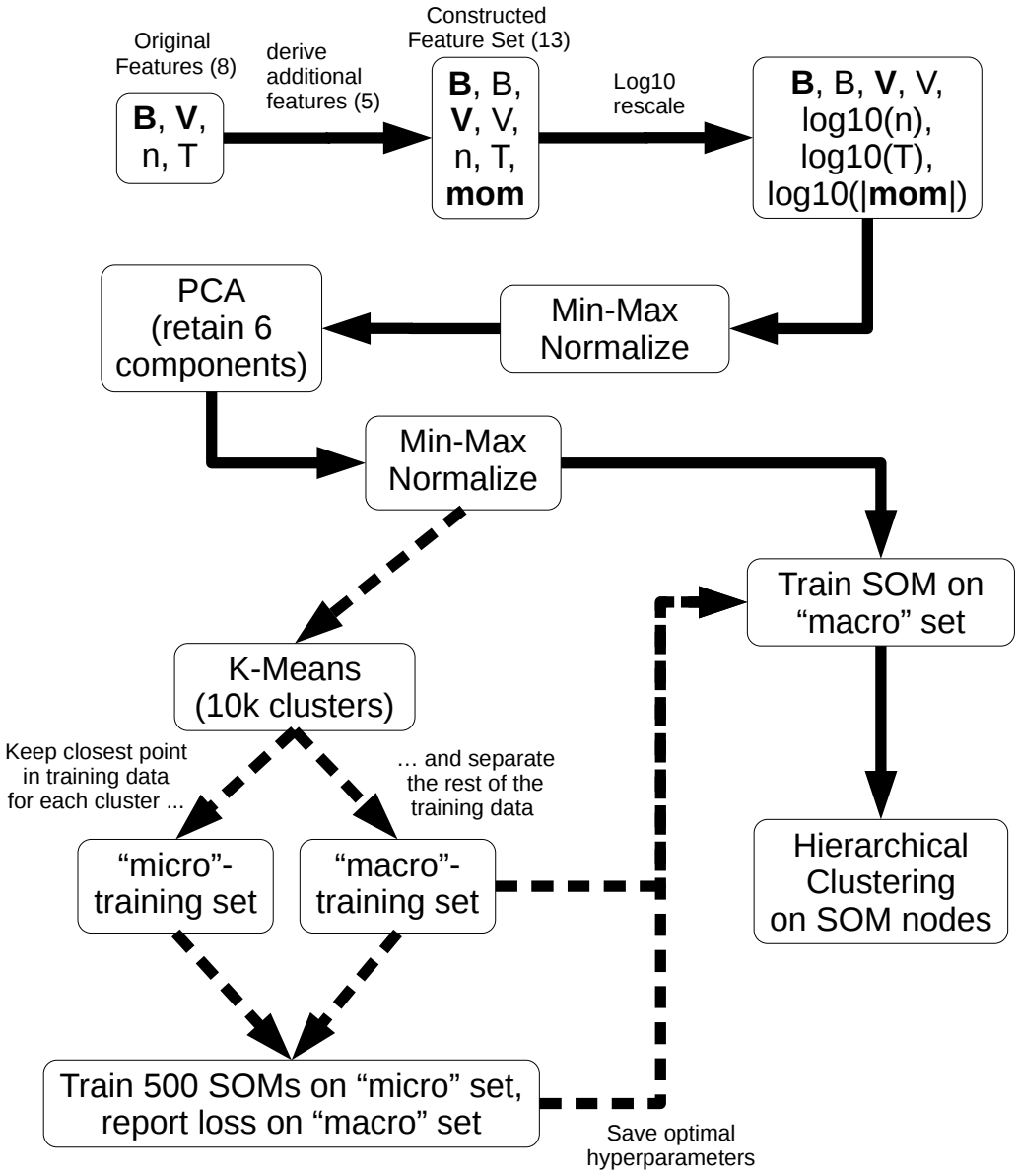

Figure 5-4: The pipeline of methods in our model. Solid arrows indicate a component of the model and dashed lines show how the optimal hyperparameters were learned using a micro and macro training set.

Chapter 6

GMClustering

In this chapter, we validate the unsupervised classifier, called Global-Magnetosphere-Clustering, or GMClustering, that we have constructed.

6.1 Model Results

The dendrogram of the hierarchical clustering of the SOM nodes and the resulting cluster assignments are shown in Figure 6-1. From the dendrogram, we make cluster classifications using a distance threshold of 1.65 and propagate the cluster assignments of the SOM nodes to the test data. The number of data points in the test set mapped per node is also shown in the same figure. Histograms of the classifications for each cluster are shown in Figure 6-2. These clusters were obtained in an unsupervised manner and a posteriori analysis shows that they correspond with specific regions, those being the magnetosphere, magnetosheath, and solar wind. The clustering of the SOM nodes in PCA space is shown in Figure 6-3. We previously made conjectures as to what portions of the biplot from Figure 5-1 are associated with the solar wind, magnetosheath, and magnetosphere, and they are confirmed with the clustering depicted. In both the (0,1) and (0,2) plots of Figure 6-3, the magnetosheath cluster has overlap with both the magnetosphere and the solar wind clusters but the magnetosphere and solar wind clusters have little overlap with each other, as one can expect from the physics of the magnetospheric system. Higher order components possess

less variance and show considerable overlap as seen in the (1,2) plot. This is a consequence of using PCA for dimensionality reduction: The first PCA components will generally capture the majority of the variance and subsequent components will be less significant.

In GSE coordinates, the solar wind tends to be in the sunward (here, rightward) direction, the magnetosphere in the tailward (leftward) direction, and the magnetosheath is a curved transition region between the two. The histograms of log10 density and log10 Alfvén Mach number of Figure 6-2 reflect this and show the clustering is very effective in separating supersonic, moderate density plasma (solar wind) from shocked, dense plasma (magnetosheath) and very subsonic, thin plasma (magnetosphere). Note that since the Alfvén Mach number is plotted in log10 scale, the supersonic to subsonic transition occurs as a change in sign. Overlap between these distributions can certainly occur and this is reflected in their histograms. Incorrect classifications are also visible in Figure 6-2, such as scattered magnetosheath and solar wind classifications occurring in the nightside at -20 $R_E \leq Y_{GSE} \leq 20\ R_E$, a swath of magnetosheath classifications at -10 $R_E \leq X_{GSE} \leq$ -5 R_E, and magnetosphere classifications well out into the dayside. In analyzing time series, these are generally spurious and rarely part of consecutive misclassifications. We show two sample classifications of time series, one for THEMIS-C where the classification is exactly correct (Figure 6-4) and one where the majority of classifications are correct but suffer from spurious misclassifications (Figure 6-5). Analyzing when MMS 1 is in the solar wind in Figure 6-5, it's apparent that the magnetosheath-misclassifications correspond to higher temperature and lower absolute value of the velocity, as in the magnetosheath. When MMS 1 is in the magnetosheath, the solar wind-misclassifications correspond to higher absolute value in velocity and the magnetosphere-misclassifications correspond to lower density, again consistent with the characteristics of the region to which the measurements are incorrectly assigned.

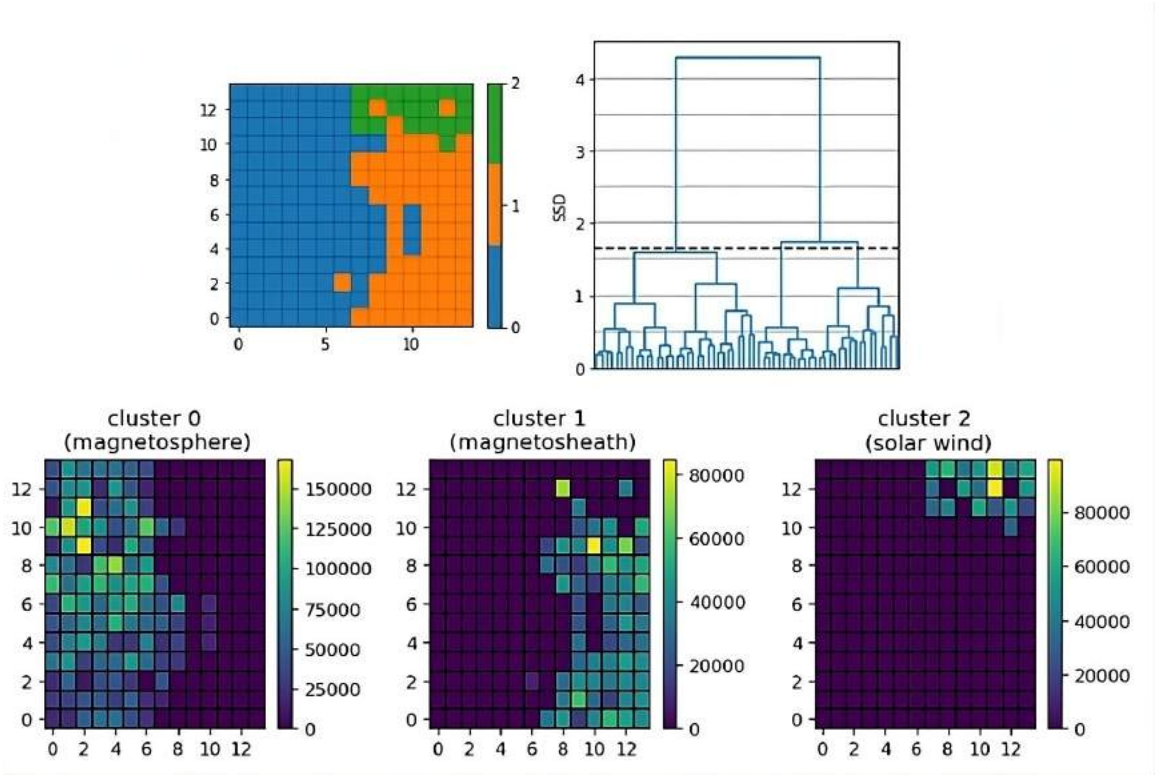

Figure 6-1: **Top Right**: A dendrogram of the clustered nodes using a Ward linkage. Separate clusters only up to the five most recent mergings are shown. We chose a cutoff sum of square deviations from the mean (SSD) of 1.65 to extract three clusters, as shown by the horizontal dashed black line. The number of times the line intersects with the vertical lines of clusters is the number of clusters recovered. The cluster assignments are visualized in the top left image. **Top Left**: Cluster assignments of the SOM nodes shown on the 2D node grid. The region of low density and high temperature observed in Figure 5-3 has been assigned to cluster 0 (blue), the region of low VX is largely cluster 2 (green) and the region of high density is largely cluster 1 (orange). The color scheme used to represent the different clusters will remain the same. **Bottom Row**: For each cluster, the number of test points per node is shown. Note that the magnetosphere-classified nodes (10,6), (10,5), and (10,4) within the magnetosheath cluster contain few hits and the magnetosheath-classified node (6,2) within the magnetosphere cluster also contains few hits. However, the magnetosheath nodes (12,12) and (8,12) within the solar wind cluster are responsible for a sizable number of hits.

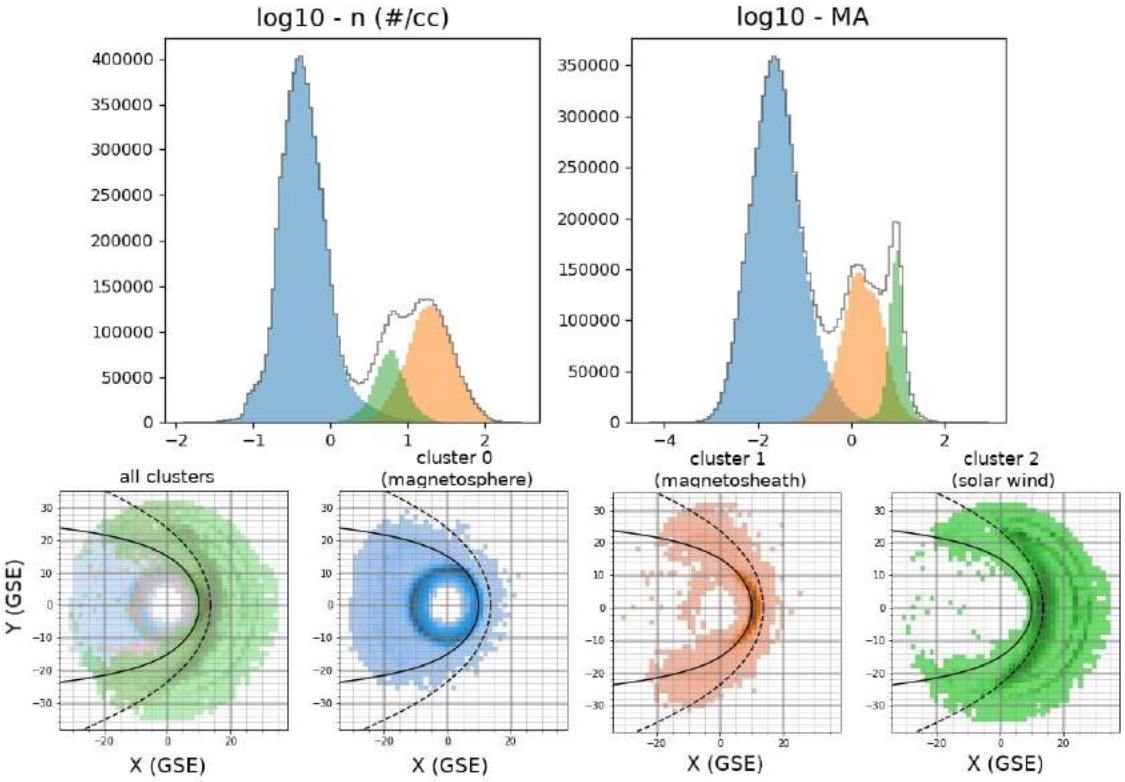

Figure 6-2: **Top / univariate histograms**: Histograms of the log10 density and log10 Alfvén Mach number. The histogram over the entire test set is in black and the histograms of the three clusters of the test set are represented in color. The magnetosphere is in blue (cluster 0), the magnetosheath is in orange (cluster 1) and the solar wind is in green (cluster 2). **Bottom / bivariate histograms**: ($X_{GSE}[R_E]$, $Y_{GSE}[R_E]$) bivariate histograms of cluster occupancy where the sun is on the right. The leftmost plot shows the histogram over the entire test set and each other plot shows an occupancy histogram for a particular cluster of the test set. The cluster color scheme used is the same as in Figure 6-1. A darker shade of color indicates a higher count in the bivariate bin. The solid line is a Shue magnetopause and the dashed line is a Chao bow shock. The parameters for these models are BZ = 0.15 nT, $D_p = 2$ nPa, $M_{MS} = 6$, and $\beta = 2$.

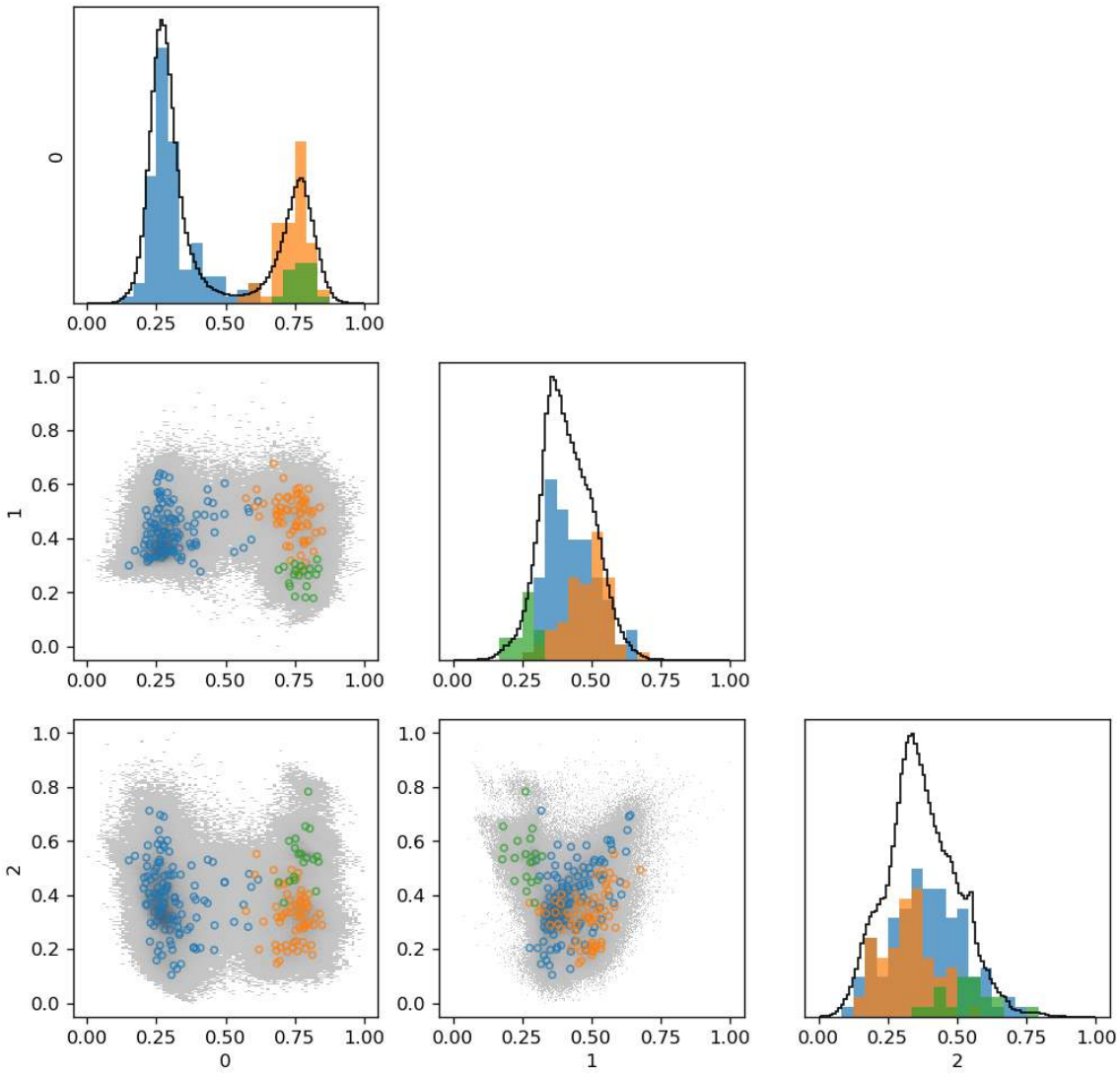

Figure 6-3: Cluster assignments of SOM nodes over the first three min-max normalized principal components of the test set. Comparing the plot of the (0,1) component-transformed data (center-left plot) to the biplot over the first two principal components in Figure 5-1, we observe that the region on the left is the magnetosphere, the upper right is the magnetosheath, and the lower right is the solar wind. The marginal histograms of all clusters are shown along the diagonal using the same bin ratio (100 bins for data and 25 for SOM nodes) as in Figure 5-2.

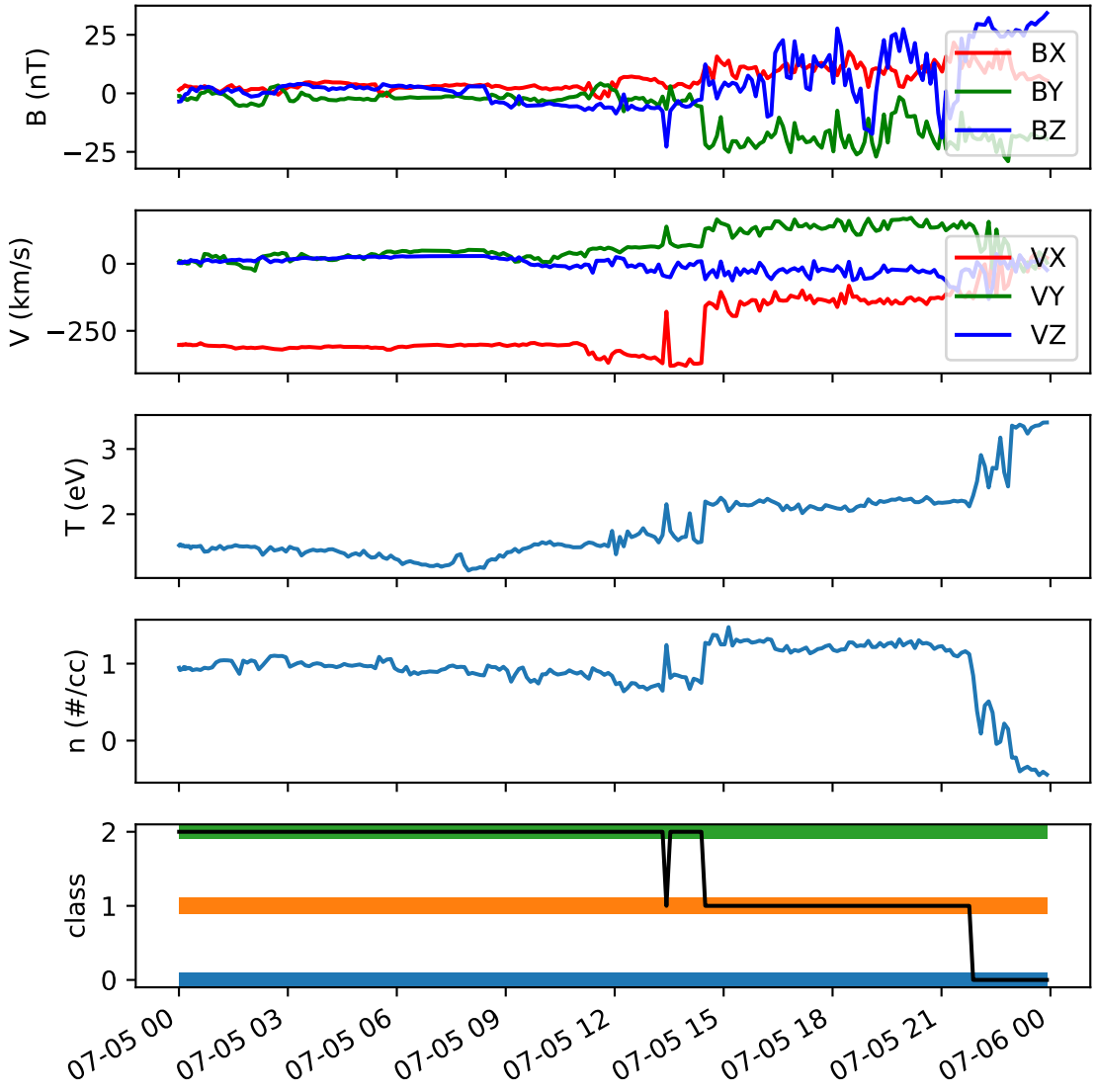

Figure 6-4: THEMIS-C measurements from 2008-07-05 to 2008-07-06. The temperature and density are in log10 scale. The classifications are shown in the bottom plot with the same cluster color scheme as Figure 6-1. The model successfully classifies the solar wind, magnetosphere, and magnetosheath measurements according to our visual verification. Noticeably, it also catches the "blip" when THEMIS-C is briefly in the magnetosheath before again crossing the bow shock and going back into the magnetosheath at 14:00 UT.

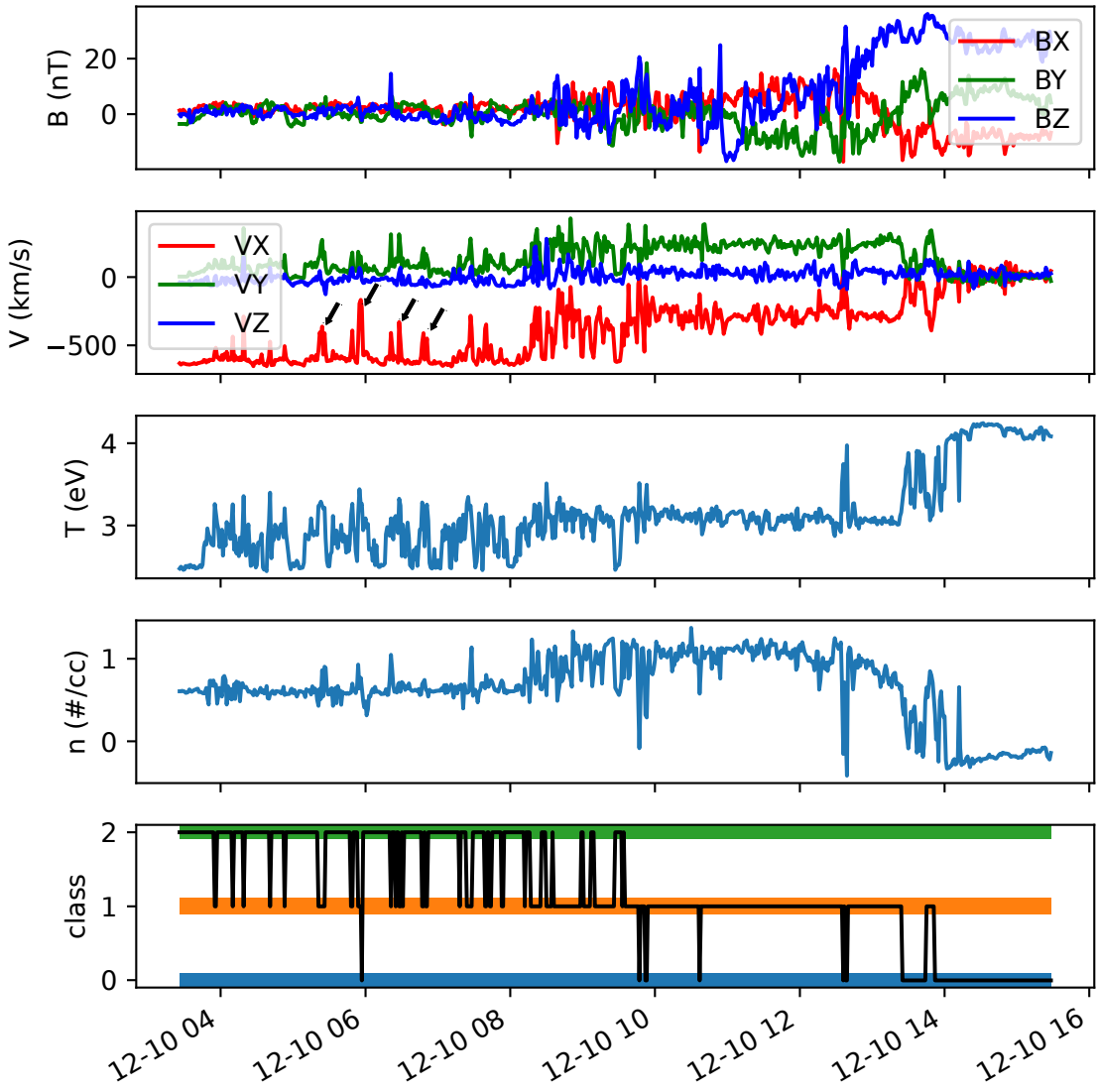

Figure 6-5: MMS 1 measurements from 2018-12-10. The plot structure is the same as Figure 6-4. MMS 1 crosses the bow shock at about 8:00 UT and the magnetopause shortly after 13:00. The majority of the classifications prior to crossing the bow shock are solar wind, but there are a number of incorrect and spurious magnetosheath classifications that occur with sharp increases in VX (as indicated by the black arrows) as well as one magnetosphere classification around 10:30 UT. After 8:00 UT, the majority of classifications changes to magnetosheath with rarer solar wind and magnetosphere classifications occurring. In the interval when MMS 1 is in the magnetosheath, magnetosphere misclassifications correspond with sudden drops in density measurements.

In Figure 6-1, it is evident that the different clusters are largely segregated spatially in the node grid but exceptions are present. There are multiple nodes that are at best somewhat adjacent to the remainder of their cluster. Notably, the magnetosheath cluster has nodes at grid positions (12,12) and (8,12) that are surrounded by the solar wind cluster. The magnetosheath cluster also has a node that is surrounded by the magnetosphere cluster at (6,2) and a vertical streak of magnetosphere-classified nodes starting at (10,4). Results like this are not entirely unexpected as we are analyzing observations and the magnetosheath acts as a transition region between the magnetosphere and solar wind. We analyze the data that map to these nodes in detail in Section 6.3.

6.2 Comparison with *Olshevsky et al.* [2021]

Ours is not the only model that has attempted to classify spacecraft observations into different plasma regions. *Olshevsky et al.* [2021] used a convolutional neural network trained on the ion energy distributions of MMS to classify them as magnetosphere, magnetosheath, pristine solar wind, or ion foreshock and *Nguyen et al.* [2022] used a gradient-boosted decision tree trained on magnetic field and ion moments of a variety of spacecraft to classify them as magnetosphere, magnetosheath, and solar wind classes. *Breuillard et al.* [2020] also used a convolutional neural network on MMS measurements of the magnetic field components $\mathbf{B}$ and magnitude B, the ion velocity components $\mathbf{V}$ and magnitude V, the ion density, and the parallel, perpendicular, and total ion temperatures to classify them into pristine solar wind, ion foreshock, bow shock, magnetosheath, magnetopause, boundary layer, magnetosphere, plasma sheet, plasma sheet boundary layer, and lobe.

Olshevsky et al. [2021] created a labeled dataset and has comparable classes to our model, so we have made comparisons with their model and data. They curated two month's worth of MMS1 data, covering November and December 2017 to the total of 469k points and

created two models. One of their models was trained on the November 2017 data and tested against the December 2017 data and the training and testing were reversed for the other. They did not use the full datasets for training and instead used about ∼25k points each for November and December, making sure to evenly sample from the four classes to avoid class imbalances. We use their better performing model, which was trained on December 2017 and tested against November 2017, as a comparison. We prepared both magnetic field and ion observations (averaging the magnetic field measurements to the latency of the ion observations at 4.5 sec resolution) and assigned their labels to our prepared dataset of 467k points, discarding the 2k unrecognized points. Since their model relied on correctly classifying the ion sky maps, they anticipated that complex mixing of distributions could occur at the magnetopause and bow shock, and so any data that indicates distribution mixing was assigned to the class "Unknown," comprising about 15% of their dataset. We mask these points out when comparing the accuracy of these models.

As explored in a previous section, the hierarchical capability of our model means that we can further derive sub-classes from our original classification. To directly compare against the model of *Olshevsky et al.* [2021], we will unpack our solar wind cluster into two sub-clusters and regard one as the pristine solar wind and the other as the ion foreshock. To compare model performance in our 3-class classification, we fold together the ion foreshock and pristine solar wind labels collectively as solar wind. Confusion matrices of the classifications of both models in both cases and their overall accuracy for each are shown in Figure 6-6. For magnetosphere / magnetosheath / solar wind classification, our model's overall accuracy (99.41%) is approximately equal to theirs (99.39%) but the per-class accuracy varies. Our model's accuracy for magnetosphere and magnetosheath predictions is quite high at 100% and 99.8% but our solar wind classification is only 99.1%. The false negatives of the solar wind and magnetosheath classes are almost entirely magnetosheath and mag-

netosphere labeled data at 0.9% and 0.02%, respectively. Their model's most accurately classified category is solar wind at 99.8% followed by magnetosphere and magnetosheath at 99.1% and 98.6% and the amount of solar wind false positives is 1.4%. For magnetosphere / magnetosheath / ion foreshock / pristine solar wind classification, our model's accuracy is only 86.7% with a per-class accuracy of 83.0% and 76.4% for the ion foreshock and pristine solar wind. It can also be seen that the solar wind-labeled data that our model misclassified as magnetosheath almost always corresponded to ion foreshock labels. Their model certainly outperforms here, correctly classifying the ion foreshock and pristine solar wind classes at 92.4% and 98.2% accuracy. This is not surprising as they used a supervised 3D convolutional neural network with a much more diverse dataset of 32x16x32 features and our model is an unsupervised neural network using data with only 13 features.

Rather, it should be expressed that our model is able to achieve a similar 3-class accuracy compared to a much more robust model. Moreover, the most significant advantage of this model is that it utilizes a SOM's ability to analyze data using feature maps in which unique data can be uncovered using either a "node-to-data" or "data-to-node" approach. In the next chapter, we will outline the applications of the model.

6.3 Investigating Topologically Distinct Nodes

The clustering of the nodes in the SOM is largely separated, but the topological overlap of classified nodes merits further investigation to reveal if the classification is correct or improper. We analyze the anomalous node positions in the SOM, namely the separated magnetosheath nodes at positions (12,12), (8,12), and (6,2) as well as the vertical streak of magnetosphere nodes at positions (10,6), (10,5), and (10,4).

The node at (12,12) has the largest U-Matrix value seen in Figure 5-3, indicating that it is farther from its neighbors than all other nodes in the SOM. This is not surprising since

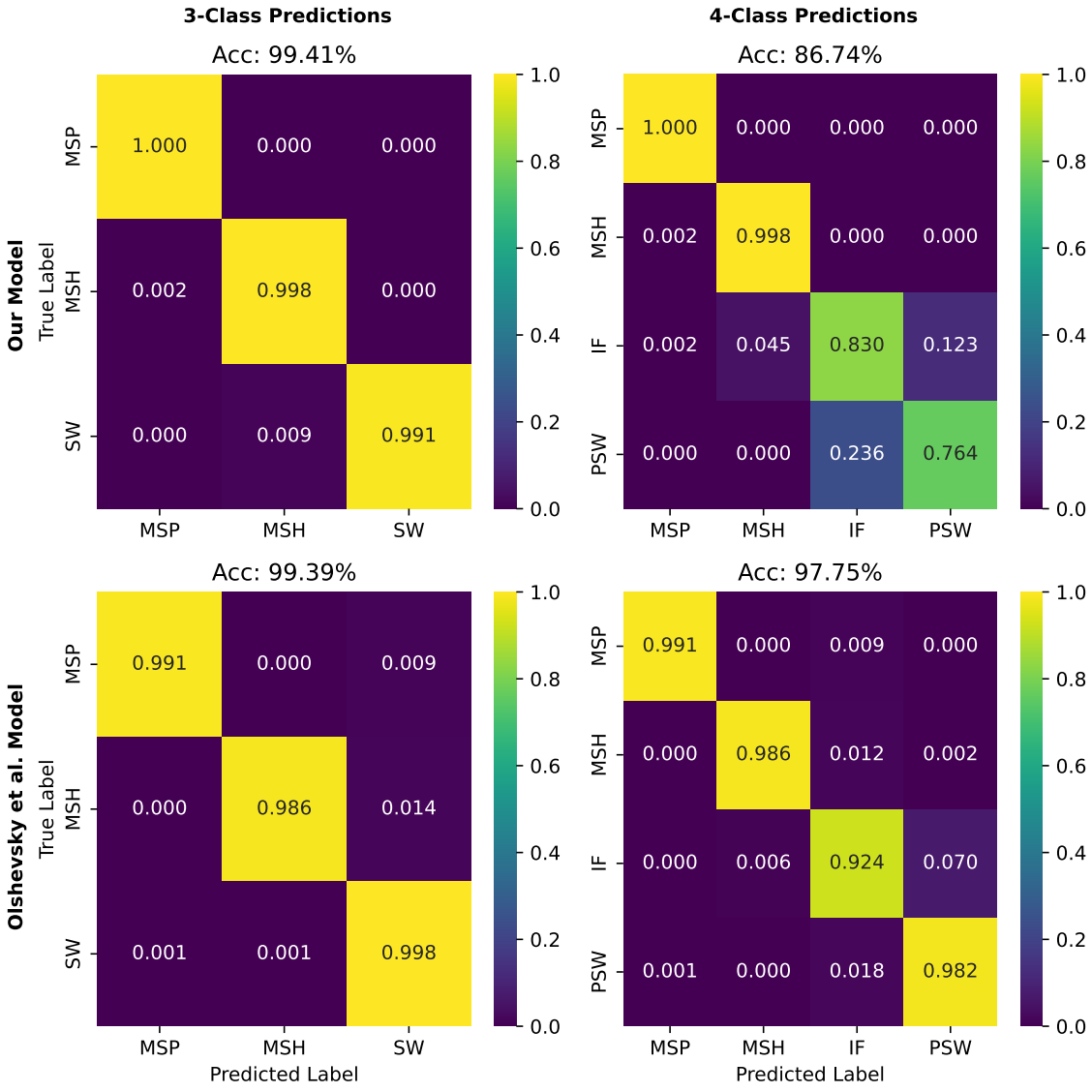

Figure 6-6: Confusion matrices for our model (top row) and the model of *Olshevsky et al.* [2021] (bottom row) against the labeled dataset of *Olshevsky et al.* [2021] for magnetosphere (MSP), magnetosheath (MSH), and solar wind (SW) classifications (left column) and for magnetosphere, magnetosheath, ion foreshock (IF), and pristine solar wind (PSW) (right column). Note that about 15% of their dataset was labeled as being "Unknown" and these comparisons are done using only the remaining 85%.

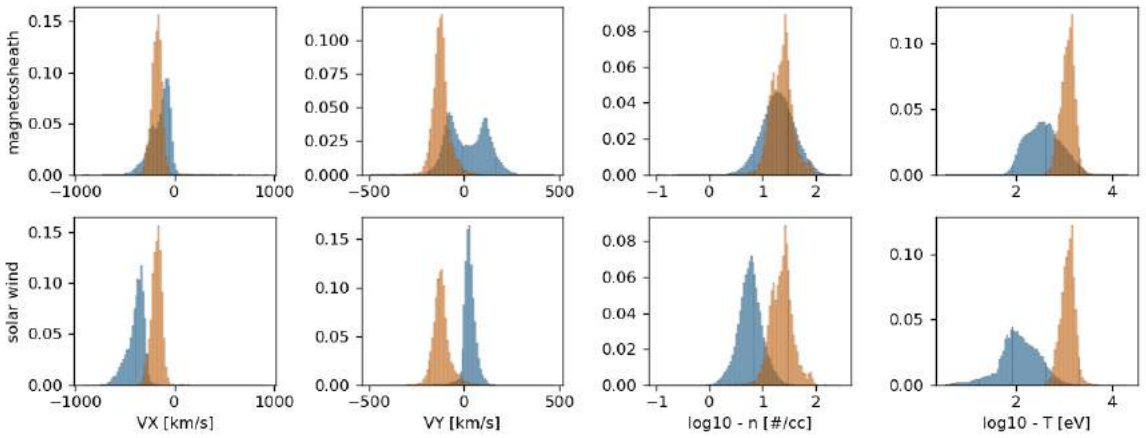

Figure 6-7: The VX, VY, log10 density, and log10 temperature empirical probability distributions of all magnetosheath-classified test data are plotted along the top row in blue. Similar features but for all solar wind-classified test data are plotted along the bottom row, also in blue. The empirical probability distribution of all test data that maps to node (12,12) is plotted in all plots as the orange distribution. The probability distributions are plotted here because of the large size differences between the number of magnetosheath observations (2.17 million) and solar wind observations (883k) of the test set and number of data mapping to node (12,12) (33k).

it is classified as a magnetosheath node and is surrounded by solar wind-classified nodes. There are about 2.18 million magnetosheath points in the test set and 34k (1.5%) of them map to this node. Categorizing this node's data by spacecraft, we find that almost all are MMS observations with only about 100 belonging to THEMIS. We plot the empirical probability distributions of all magnetosheath and solar wind measurements in the test set in Figure 6-7 as well as the data belonging to this node for comparison. From the figure, we can see that there is much more overlap with the distributions of node (12,12) with the magnetosheath observations than that of solar wind, indicating that although the node's position in the grid is unusual, it corresponds well with magnetosheath observations.

Node (6,2) is another topologically isolated magnetosheath node that also possesses a very high U-Matrix value, except that this one is surrounded by magnetosphere-classified nodes. It is responsible for only about 7.7k (0.35%) points of the magnetosheath-classified data of the test set and is almost evenly split by spacecraft with 56% points belonging to

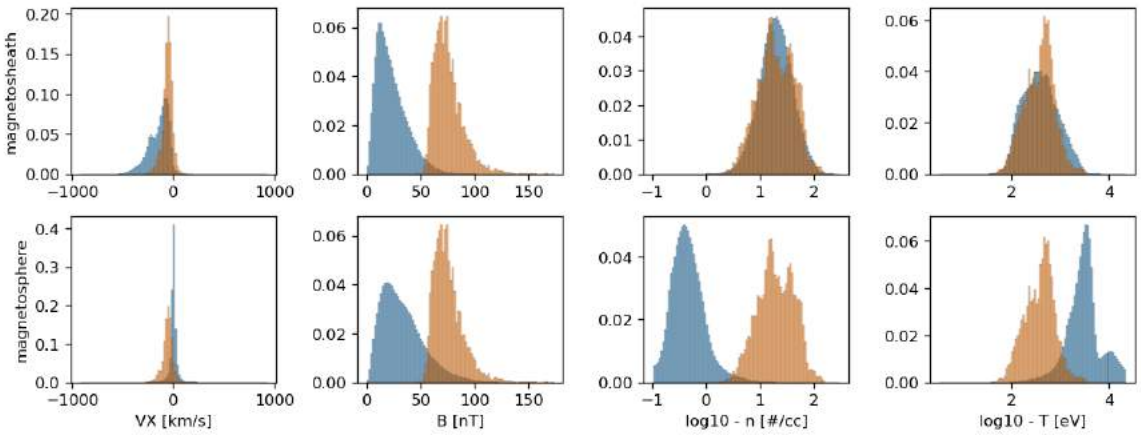

Figure 6-8: The VX, B, log10 density, and log10 temperature empirical probability distributions of all magnetosheath-classified test data are plotted along the top row in blue. Similar features but for all magnetosphere-classified test data are plotted along the bottom row, also in blue. The empirical probability distribution of the 7.7k magnetosheath-classified observations of node (6,2) are plotted in orange for each feature. The VX, log10 density, and log10 temperature distributions for this node all align more with the magnetosheath data than that classified as magnetosphere whereas the B distribution reflects high magnitude observations. Overall, this node has captured data with magnetosheath characteristics in velocity, density, and temperature, but also possessing high field magnitudes.

THEMIS and 44% to MMS. The empirical probability distributions of all magnetosheath-classified and magnetosphere-classified data in the test set are plotted alongside the observations mapped to this node in Figure 6-8 and multiple distinctions can immediately be made: data mapping to this node exhibit more magnetosheath characteristics in velocity, density, and temperature and also possess high magnetic field magnitudes. It seems correct that this node is classified as magnetosheath and the sparsity of points mapping to this node is understood in the context that magnetosheath observations possessing such large magnetic field magnitudes is relatively rare. The large U-Matrix value is justified with these observations.

Node (8,12) is diagonally topologically adjacent to the magnetosheath cluster but otherwise surround by solar wind nodes. This SOM uses a square topology, so this diagonal proximity does not factor into its U-Matrix value. It maps 68k (3.1%) points from the magnetosheath-classified data of the test set with 11% being THEMIS observations and

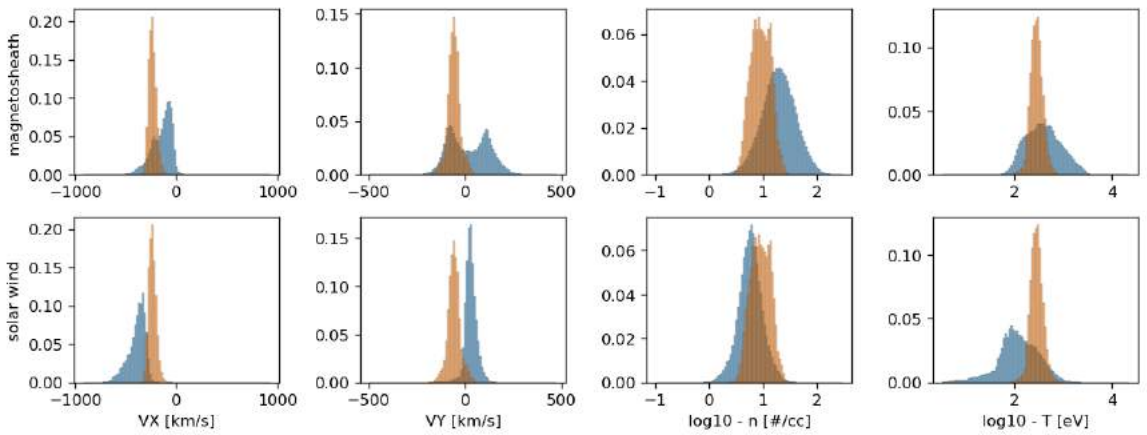

Figure 6-9: The VX, VY, log10 density and log10 temperature empirical probability distributions of all magnetosheath-classified data from the test set are plotted in blue along the top row. The solar wind-classified test data are plotted in blue along the bottom. The empirical probability distribution of the 68k magnetosheath-classified observations of node (8,12) are plotted in orange for each feature. The log10 density and log10 temperature distributions of the data from this node have sizeable mixing between both magnetosheath and solar wind observations whereas the VX and VY distributions are more distinctly magnetosheath than solar wind.

89% being MMS. The VX, VY, log10 temperature, and log10 density empirical probability distributions of the data mapping to this node are shown in Figure 6-9 alongside all magnetosheath-classified and solar wind-classified test data. They indicate magnetosheath observations with respect to the VX and VY distributions, but the log10 temperature and log10 density distributions somewhat resemble a blend of solar wind and magnetosheath. This lack of uniform agreement across these features can explain why node (8,12) is adjacent to solar wind-classified nodes but the VX and VY distributions in particular indicate that it is correct to classify it as a magnetosheath node.

Lastly, we analyze the magnetosphere-classified nodes at positions (10,6), (10,5) and (10,4) that occur topologically within the magnetosheath cluster. Together, these nodes account for 46k (0.75%) of the 6.1 million magnetosphere-classified points of the test set with 76% being THEMIS observations and 24% belonging to MMS. Their VX, VY, log10 density and log10 temperature empirical probability distributions are plotted in Figure 6-

10 along with the distributions of all three clusters in the test set. The data that map to these nodes are unusual in that the node distributions do not fully overlap with all of the distributions for any cluster. These data are classified as magnetosphere, but exist along the extrema of all the magnetosphere distributions shown. They resemble the VY, log10 density, and log10 temperature distributions of the solar wind, but the VX would be quite low for solar wind. The VX, VY, and log10 temperature distributions match up well with the magnetosheath distributions, but the log10 density is conspicuously low. Across all of the clusters, the measurements have much more in common with magnetosheath observations than magnetosphere or solar wind and are likely misclassifications. A time series of MMS1 observations containing many points that map to one of these nodes is shown in Figure 6-11. The magnetosheath plasma is of relatively low density, reflective of how these nodes are misclassified as magnetosphere. These nodes are responsible for 0.50% of the total test set.

Overall, the magnetosheath cluster has nodes in several aberrant positions in the SOM grid in which they were surrounded by nodes belonging to other clusters. Investigating these nodes in detail, however, has shown that the data correspond well with magnetosheath observations and are deserving of being classified as such. It was also seen that three magnetosphere-classified nodes are likely misclassified and should be recognized as magnetosheath. These three nodes contain few points (46k points, or 0.50% of the test set), together containing slightly less than the average number of test points per node (47k), and so do not significantly impact the strength of the results. Furthermore, it should be noted that such a misclassification occurred between the magnetosheath and the magnetosphere and that the separation between solar wind and magnetosphere plasma is quite distinct in the map.

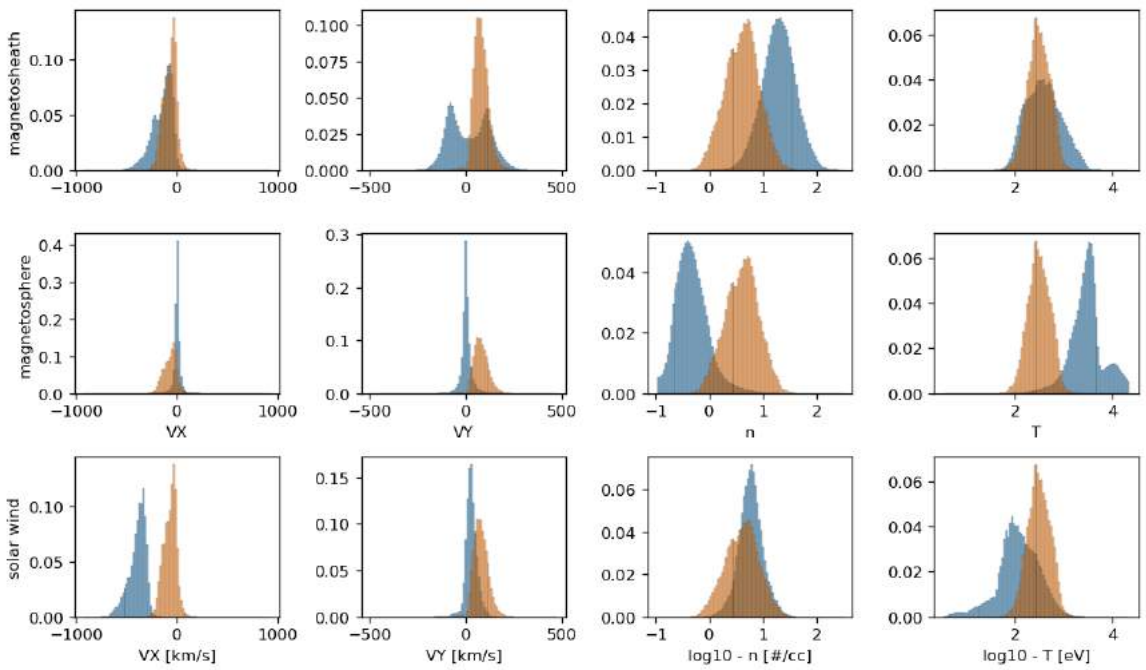

Figure 6-10: The VX, VY, log10 density and log10 temperature empirical probability distributions of all magnetosheath-, magnetosphere-, and solar wind-classified test data are plotted in blue along the top, middle, and bottom rows, respectively. All test data that map to nodes (10,6), (10,5) and (10,4) are collectively plotted here as the orange empirical probability distributions. These data are anomalous and exhibit characteristics found in all magnetosheath, magnetosphere, and solar wind observations. The VY, log10 density, and log10 temperature align well with the solar wind distributions, but the VX distribution is far too low. The VX, VY, and log10 temperature distributions correspond with magnetosheath observations, but there are very low densities. All of these distributions seem to have the least in common with the magnetosphere cluster, being along the extrema in all cases.

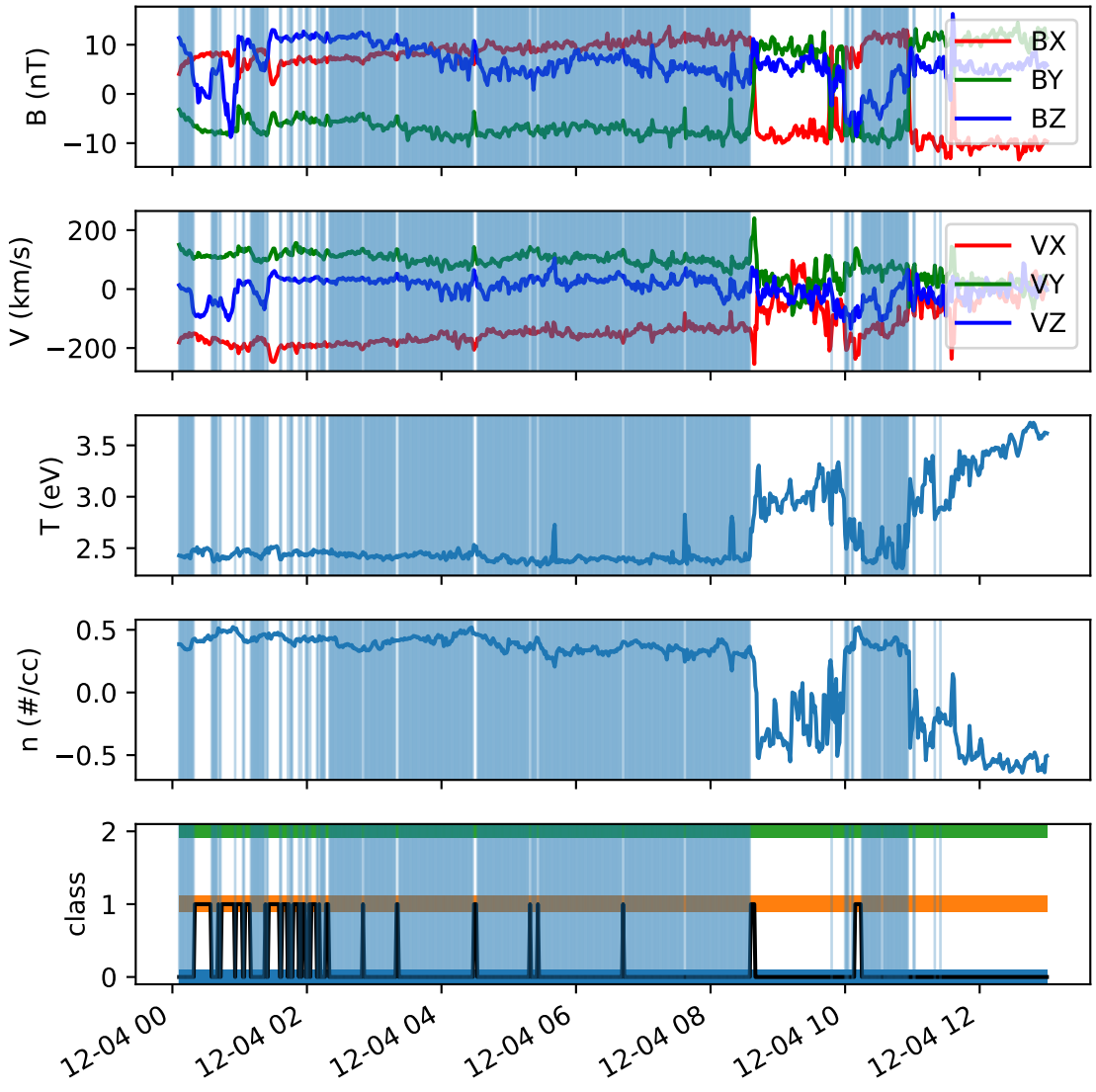

Figure 6-11: MMS 1 measurements from midnight to 13:00 UT on 2020-12-04. The plot structure is the same as Figure 6-4. The transparent vertical blue lines indicate that the measurement at that time maps to the (10,6), (10,5), or (10,4) node. MMS1 is measuring low-density magnetosheath plasma from midnight to 8:30 UT and from 10:00 to 11:00 UT. 473 points (84.3%) of the magnetosphere-classified data in the midnight to 11:00 UT interval map to one of these nodes. These 473 points are also almost 1% of all data that map to these nodes.

Chapter 7

GMClustering Applications

7.1 Subpopulation Analysis

We show in brief the capability of subpopulation analysis with this clustering method. Since we have used a hierarchical method to cluster the SOM nodes, we can pick any cluster and investigate the previously merged clusters that compose it. We "unpack" the magnetosphere cluster in Figures 7-1 and 7-2 to show how distinct magnetospheric populations were collectively recognized as the magnetosphere. From the histograms, we see that the feature that changes most clearly between the two clusters is the Alfvénic Mach number. Note that the subclusters of the magnetosphere in Figure 7-1 are not as evenly topologically separated like the original clustering solution seen in Figure 6-1. This is not surprising given the large overlap in features between these subclusters seen in the univariate histograms of Figure 7-2 and indicates that the variance between these two subclusters is less than the variance between the magnetosphere, magnetosheath, and solar wind clusters, hence these two subclusters appearing earlier in the merge order with a Ward linkage. In simpler terms, it is easier to distinguish solar wind measurements from those of the magnetosheath or magnetosphere than it is to separate magnetospheric populations by Alfvén Mach number.

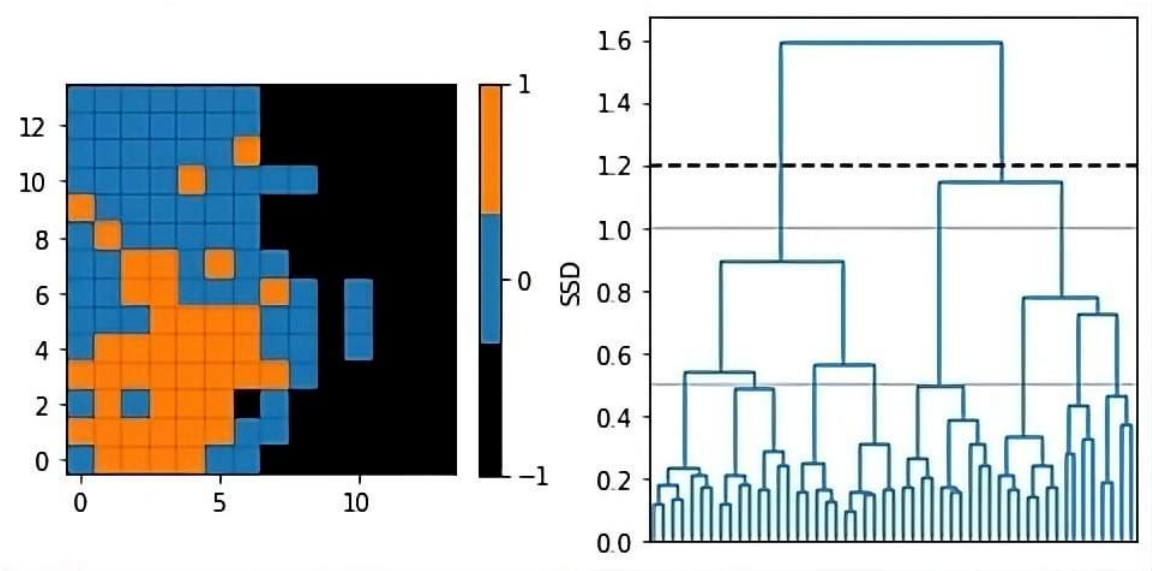

Figure 7-1: Like Figure 6-1 but only focusing on the magnetosphere cluster. **Right**: A dendrogram showing the merge order of the magnetosphere cluster. This tree is a subset of the dendrogram in Figure 6-1. We use a cutoff SSD of 1.2 and extract two clusters from the magnetosphere cluster. **Left**: Subcluster assignments of the SOM nodes based on the distance chosen in the dendrogram. The nodes that did not belong to the magnetosphere cluster are masked out in black and assigned a label of -1. Looking back to the feature maps in Figure 5-3, we can see that the blue cluster (0) is related to higher subsonic Alfvén Mach number and the orange cluster (1) is related to lower subsonic Alfvén Mach number.

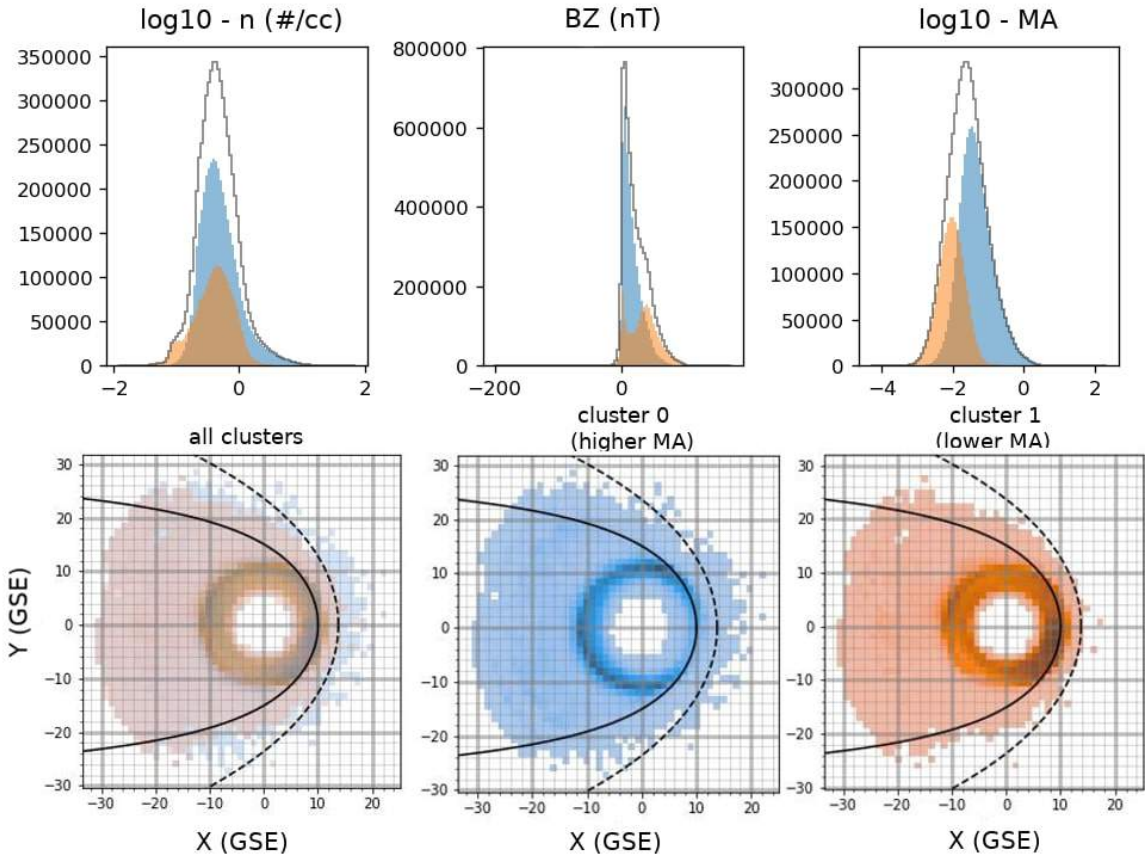

Figure 7-2: Like Figure 6-2, but analyzing only the magnetosphere cluster of the test set. **Bottom / bivariate histograms**: The occupancy of cluster 0 (blue) and 1 (orange) are plotted as bivariate histograms in ($X_{GSE}[R_E]$, $Y_{GSE}[R_E]$). They cover a similar region, but cluster 1 is much less pronounced on the dayside. **Top / univariate histograms**: The histograms of log10 density, BZ, and log10 Alfvén Mach number are plotted in black and the cluster populations are plotted in their respective colors. As could be inferred from Figure 5-3, cluster 0 is related to higher subsonic Alfvén Mach number and cluster 1 to lower subsonic values.

7.2 Derived Boundary Crossings

With a model that can classify when a measurement occurs in the magnetosphere, magnetosheath, or solar wind, we can study the time series of these classifications and infer when a spacecraft has crossed the magnetopause or bow shock. To select crossings, we used a moving window over the time series of classifications and find where the classification changes from magnetosheath to solar wind or vice-versa. We considered such a change in classification to be a crossing if all points half a window length before belong to one cluster and all points half a window length ahead belong to the other. The changing time resolution in the THEMIS data means that we need to consider different window lengths between MMS and THEMIS observations. A window length of 20 minutes was used for MMS to give up to 10 points per half window length and a window length of 40 minutes for THEMIS to give up to 13 points per half window length when the ESA is in Fast-Survey Mode (32 spins, 96 sec, going from the magnetosheath to the solar wind) or up to 3 points per window when it is in Slow-Survey Mode (128 spins, 6.4 minutes, going from the solar wind to the magnetosheath). A total of 3047 bow shock crossings and 5228 magnetopause crossings are extracted using these parameters. Bivariate histograms of the (X_{GSE}, Y_{GSE}) positions of these crossings is depicted in Figure 7-3 alongside a Shue magnetopause (*Shue et al.* [1998]) and Chao bow shock model (*Chao et al.* [2002]) and show good agreement with respect to both.

For the bow shock crossings, we select the most recent solar wind point relative to the time of crossing and see how they're distributed in the SOM grid in Figure 7-4. When cross-comparing these with the number of counts in the test set from Figure 6-1, we see that the two most activated nodes of bow shock crossings are nodes (10,11) and (12,11). These nodes are responsible for 21.7% of the crossings but only 11.5% (training + testing) of the solar wind classifications. In the case of operational use of this model, a solar wind measurement

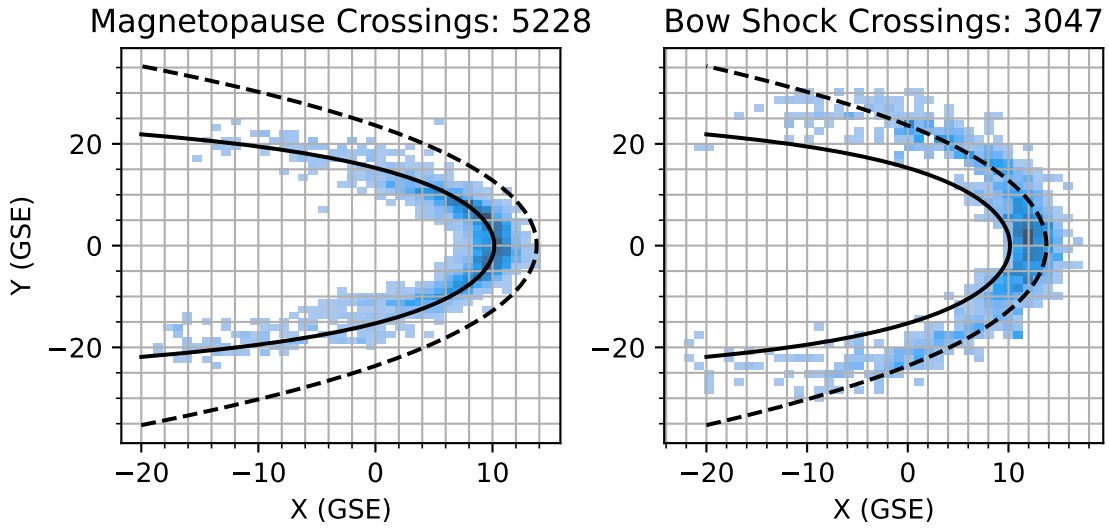

Figure 7-3: Bivariate histograms of the magnetopause (left) and bow shock (right) crossings in ($X_{GSE}[R_E]$, $Y_{GSE}[R_E]$). In both figures, the solid line is a Shue magnetopause with parameters n = 8 #/cc, V = 400 km/s, and BZ = 0.15 nT and the dashed line is a Chao bow shock with parameters BZ = 0.15 nT, D_p = 2 nPa, M_{MS} = 6, and β = 2. Many of the crossings are in line with expectations of magnetopause and bow shock positions although a handful of errant crossings are evident, such as the magnetopause crossings at (X=-4, Y=7) and (X=5, Y=25).

assigned to one of these nodes could be flagged as having an increased probability of being a solar wind point adjacent to a bow shock crossing. Additionally, the node with the highest count in the test set for solar wind points, node (11,12), has only a small number of bow shock crossing points (6.2%) relative to the previous nodes.

We perform a similar analysis for the magnetosheath points relative to the magnetopause crossings. The nodes with the highest number of counts of magnetosheath points associated with magnetopause crossings are the nodes (9,5), (8,8), and (8,2). These are responsible for 18.2% of the magnetopause crossings but only 3.0% of the magnetosheath classifications (training + testing). The node with the largest number of magnetosheath points in the test set, node (10,9) at 3.6%, only contains 15 magnetosheath points of the crossings, or 0.29% of the magnetopause crossings. These three nodes could be used to flag possible magnetopause crossings.

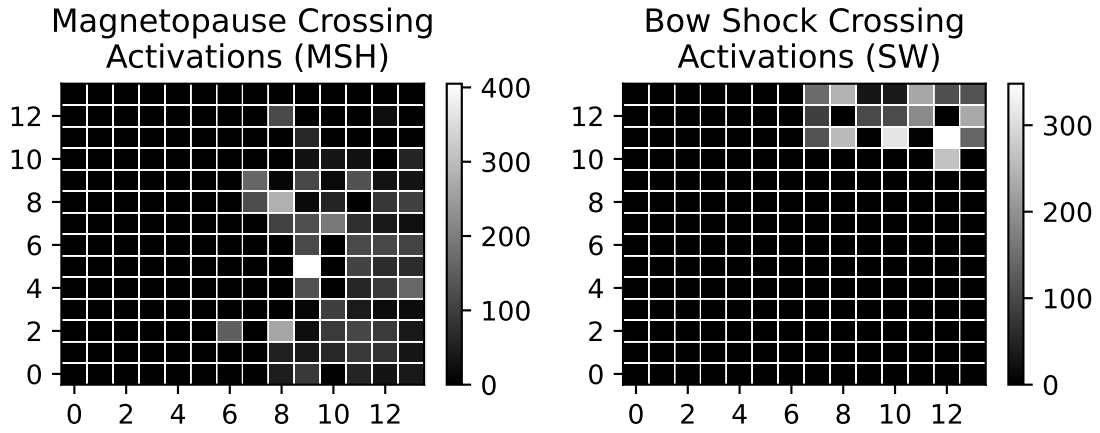

Figure 7-4: For each magnetopause (bow shock) crossing, we select the most recent magnetosheath (solar wind) point. Each point maps to, or "activates", some node in the SOM. The distribution of these counts is shown for the magnetosheath points for the magnetopause on the left and the solar wind points for the bow shock on the right. For the magnetosheath points, the most activated nodes are at positions (9,5), (8,8), and (8,2) and are together responsible for 949 crossings. For the solar wind points, the most activated nodes are at positions (10,11) and (12,11) and are responsible for 660 crossings.

7.3 Identifying Bursty Bulk Flows

Bursty Bulk Flows (BBF) are earthward-moving plasma flows in the magnetotail that are often characterized by large speeds towards Earth (hence a large, positive VX component), dipolarizations, depletions in density, and increases in temperature and are an important process in the earthward transport of mass, energy, and magnetic flux in the magnetosphere (*Angelopoulos et al.* [1994]). Detecting a dipolarization in magnetic field data alone is inherently a time-dependent comparison, but detecting large VX components can be done in a time-independent manner. Using the feature maps from Figure 5-3, we see that nodes (0,11) and (3,12) are magnetosphere-classified nodes that have large average VX values of almost 100 km/s. Thus we can use these nodes to identify possible BBFs. A dataset of BBFs as observed by MMS from 2017 to 2021 was created by *Pitkänen et al.* [2023], and they show two examples in their paper. We show that the BBF of their first example corresponds to many activations of the (0,11) node in Figure 7-5. Not every activation

corresponds to a BBF, but a rolling window method counting the number of activations could be used to flag possible BBFs.

7.4 Identifying Hot Flow Anomalies and Foreshock Bubbles

Hot Flow Anomalies (HFA) and Foreshock Bubbles (FB) are transient phenomena that are often observed in the ion foreshock. HFAs form from the interaction of a tangential discontinuity with the bow shock and can result in particle energization, diminished density and magnetic field, and flow turning sunward (*Schwartz et al.* [1985]; *Omidi and Sibeck* [2007]). FBs are instead formed prior to this interaction but can possess similar characteristics of low density and field strength and reduced VX / sunward flows (*Omidi et al.* [2010, 2020]). These properties mean that these observations could be classified as magnetosheath or magnetosphere. Thus a simple way to identify possible HFAs and FBs using this model is to track sequential solar wind-classified data and find gaps in the classifications. *Liu et al.* [2022] compiled a list of observations of HFAs and FBs from MMS1 and THEMIS-A, 47 of which are from November and December 2017 of MMS1. Using the same 4.5 sec resolution dataset we previously prepared, we extract solar wind classification gaps of up to 2 minute duration. Allowing an observation to be within up to 30 seconds of an identified gap, we find that we can identify 39 of the 47 observations. An example interval of MMS1 data containing seven HFA / FB observations is shown in Figure 7-6.

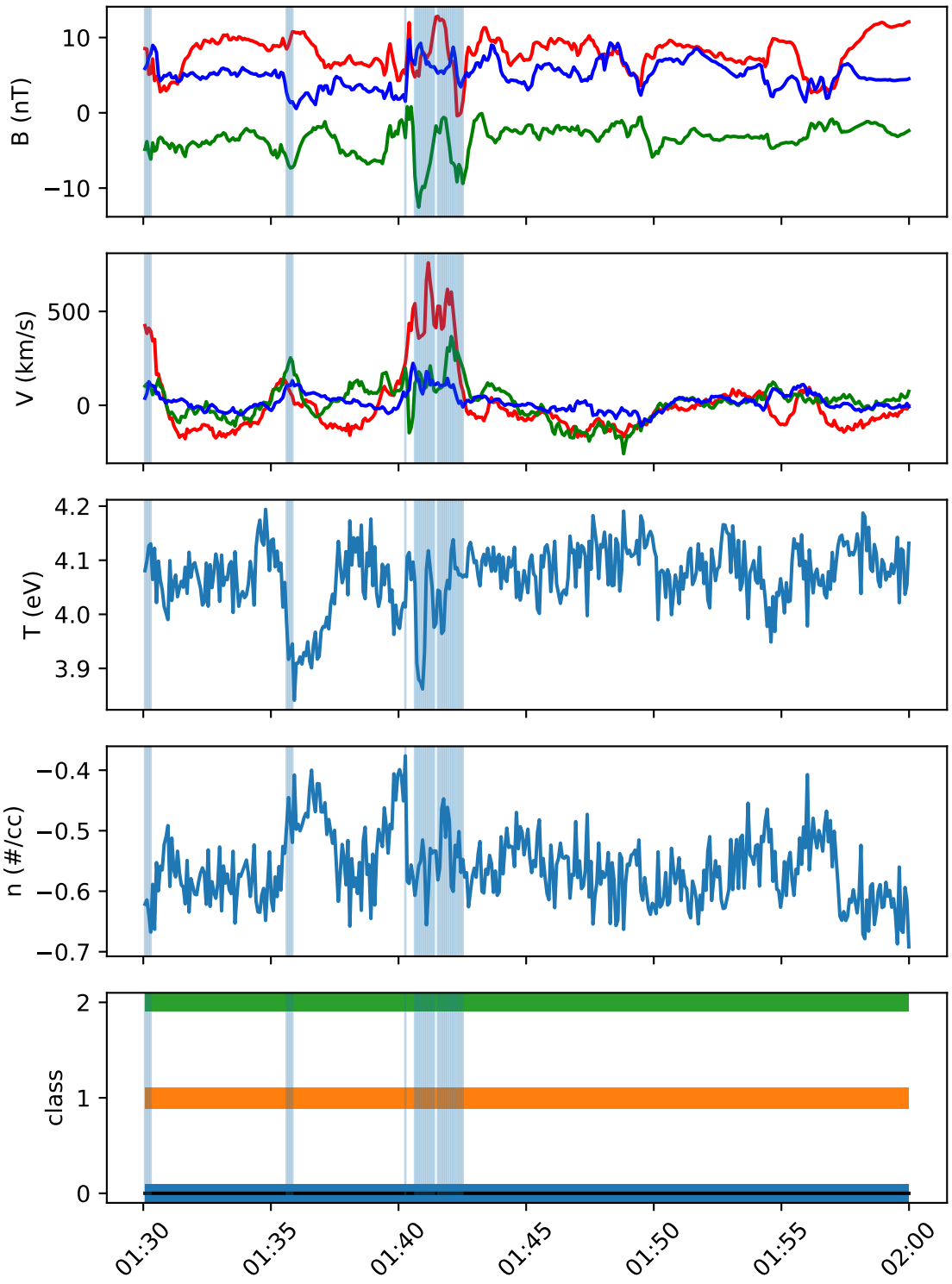

Figure 7-5: MMS 1 measurements from 01:30 to 02:00 UT on 2021-08-15 at 4.5 sec resolution. The plot structure is the same as Figure 6-4. The vertical blue lines here denote the magnetosphere-classified points that mapped to node (0,11). *Pitkänen et al.* [2023] identified the BBF interval as lasting from 01:39:56 to 01:42:38 UT, and 26 of the 36 points in that 2 minutes 42 seconds interval map to node (0,11).

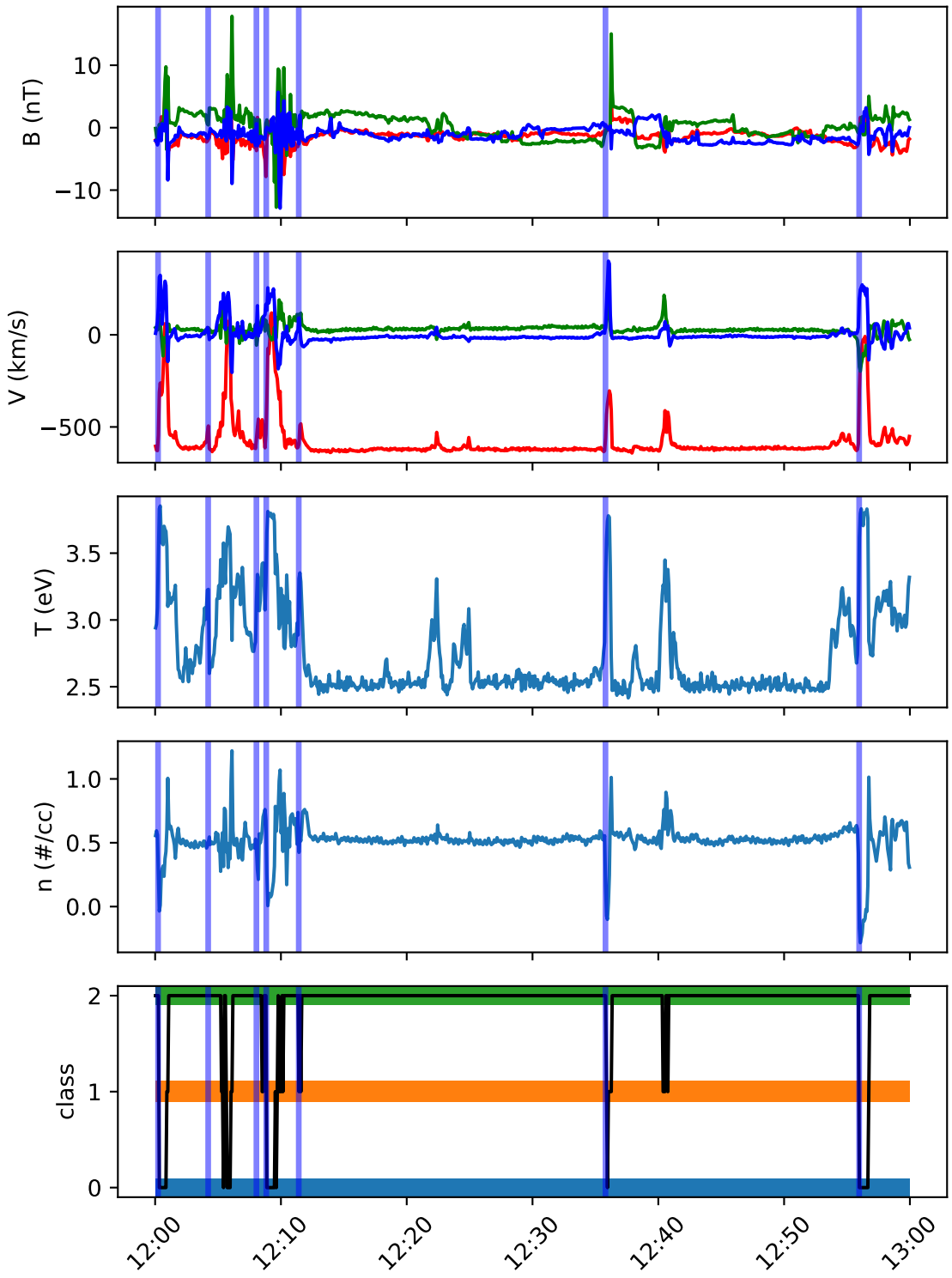

Figure 7-6: MMS 1 measurements from 12:00 to 13:00 UT on 2017-12-18. The plot structure is the same as Figure 6-4. The vertical purple lines here denote a HFA / FB time as recorded by *Liu et al.* [2022]. Six of the seven observations were recognized with our method, the exemption being the observation at 12:04:13 UT. This missed observation is still reflected in the sequence of magnetosheath / magnetosphere classifications occurring near 12:05 UT, but is beyond our 30 second window.

Chapter 8

Bow Shock Model Data Preparation

The bow shock crossing times derived from Section 7.2 are aligned with upstream solar wind estimates of OMNI. Additional crossings along with upstream solar wind parameters are prepared using Cluster, Imp 8, Geotail, and Magion-4. The creation of this cumulative dataset is outlined in Section 3.2. Here we will show how we modify our crossing dataset into an aberrated coordinate system to account for offsets in VY and VZ as well as the transformations and rescalings required for training a neural network on the data.

8.1 Aberrated Coordinates

The orbital motion of the Earth around the Sun causes a +30 km/s offset in V_Y observations in GSE. There is no such offset along Z, but large V_Y and V_Z upstream solar wind values for bow shock crossings can affect our ability to effectively parameterize the bow shock relative to upstream solar wind effects. We will transform from GSE coordinates to an aberrated coordinate system. Note that this aberration correction will not simply be a single 3d rotation matrix to be applied to all data, but a unique 3d rotation matrix will be learned for each point such that aberrated V_Z will be 0, aberrated V_Y will be -30 km/s (such that aberrated $V_Y + V_{Earth} = 0$), and aberrated V_X will be close to the original speed (as the

velocity components due to V_Y and V_Z have been collectively rotated into just V_X). The angles to aberrate the coordinates for V_Y and V_Z are

$$\alpha_{abr} = \arctan(\frac{V_Y + 30}{|V_X|}) \tag{8.1}$$

and

$$\beta_{abr} = \arctan(\frac{V_Z}{\sqrt{V_X^2 + (V_Y + 30)^2}}). \tag{8.2}$$

For each point, the full aberrated rotational matrix is then given by the matrix product $\mathbf{R}(\beta;\mathrm{Y}^*)$ $\mathbf{R}(\alpha;\mathrm{Z})$ where $\mathbf{R}(\alpha;\mathrm{Z})$ means the 3d rotation matrix for rotating about the Z axis by an angle α. The full aberrated rotation matrix is then given by

$$\begin{aligned}
\vec{X}' &= \mathbf{R}(\beta; Y*)\mathbf{R}(\alpha; Z)\vec{X} \\
&= \begin{pmatrix} \cos\beta & 0 & \sin\beta \\ 0 & 1 & 0 \\ -\sin\beta & 0 & \cos\beta \end{pmatrix} \begin{pmatrix} \cos\alpha & -\sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}, \\
&= \begin{pmatrix} \cos\alpha\cos\beta & -\sin\alpha\cos\beta & -\sin\beta \\ \sin\alpha & \cos\alpha & 0 \\ \cos\alpha\sin\beta & -\sin\alpha\sin\beta & \cos\beta \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix},
\end{aligned}$$

where $\vec{X}'$ and $\vec{X}$ are the position vectors for the aberrated GSE and GSE coordinates respectively. These equations can also be seen as part of Equations 1 and 2 of *Gencturk Akay et al.* [2019]. Plots of the crossings in X-Y, X-Z, and Y-Z positions in GSE and aberrated GSE can be seen in Figure 8-1. Histograms of α_{abr} and β_{abr} can be seen in Figure 8-2. It should be carefully noted in this aberrated system that the rotation accounting for the aberration due to V_Z is not a rotation about the original Y axis but is instead about

the Y* axis (the new Y axis that results from the Z axis rotation).

8.2 Data Rescaling

Neural networks generally work better with normally distributed data such that the means and variances of each input feature are zero and one, respectively (*LeCun et al.* [2012]). Magnetosonic mach number, plasma beta, dynamic pressure, and BZ are classic features used to parameterize many different bow shock models as covered in Chapter 1. We incorporate these features, as well as the magnetic cone and clock angles. The dynamic pressure being proportional to $n_{ion}V_{ion}^2$ and plasma beta involving this term means that both of these variables are typically skewed distributions (see the log10 beta and linear dynamic pressure histograms in Figure 3-3). To reign this in, we apply a log10 transform to the pressure and plasma beta, then standardize each training feature such that the resulting mean and variance of each feature is 0 and 1, respectively. Histograms of the distributions of the training data are seen in Figure 8-3. We note that while the histograms of the clock and cone angles are shown in degrees for easier visual interpretation, we will use them in units of radians for any model calculations. After, we do a 70%-15%-15% training-validation-test split. This means our training, validation, and test set sizes are 10,264, 2,199, and 2,201 points, respectively. Violin plots of the standardized training data are shown in Figure 8-4.

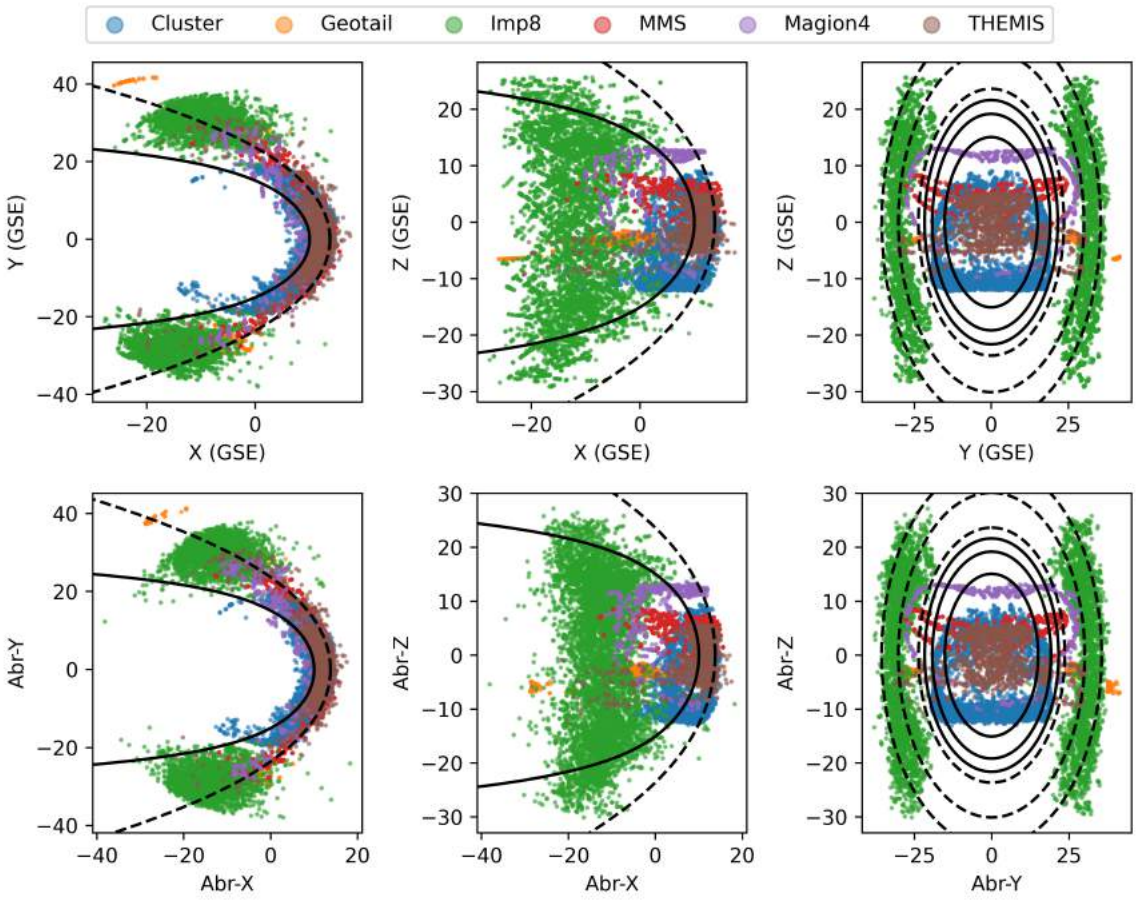

Figure 8-1: Plots of the X-Y, X-Z, and Y-Z positions in GSE for the unaberrated 14,664 bow shock crossings are along plotted along the top row, respectively. The aberrated coordinates are shown along the bottom row. Both plots are in units of R_E and show each spacecraft mission using a different color. Solid and dashed lines correspond to a Shue magnetopause and a Chao bow shock using the same parameters as in Figure 6-2 (being $B_Z = 0.15$ nT, $\beta = 2$, $M_{MS} = 6$, and $D_p = 2$ nPa). Both models are symmetric with respect to the Y and Z axes so there is no difference between the model estimates in the X-Y and X-Z plots. Both Y-Z plots show three Shue and Chao models which correspond to Y-Z slices at X = 0, -10, and -20 R_E moving from the center of the plot outward.

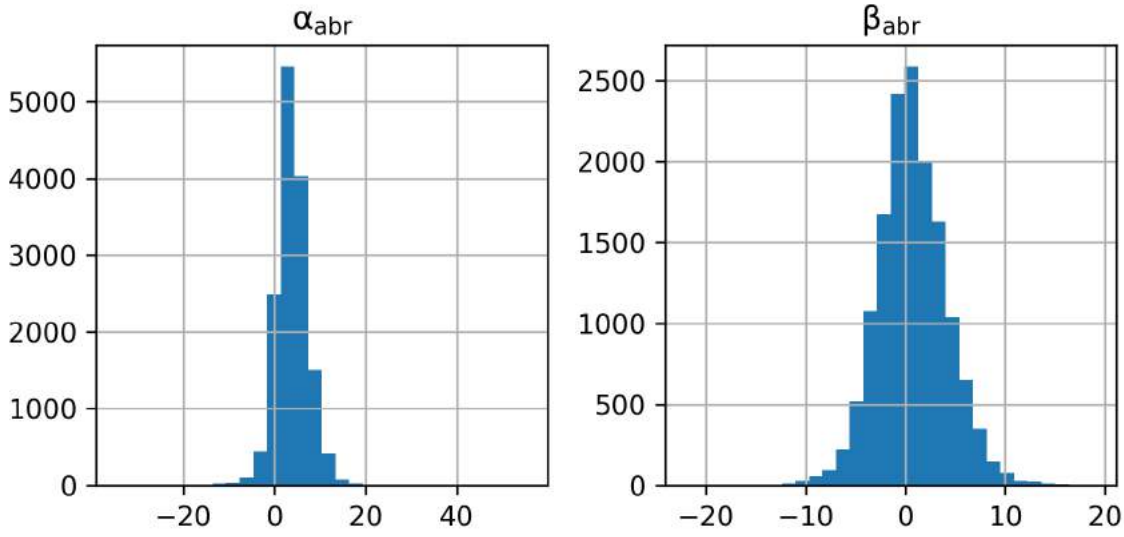

Figure 8-2: Histograms of the aberration angles α_{abr} and β_{abr}. Many of the angles are small with the means close to 4° and 0° as expected, but there are crossings present with quite large angles.

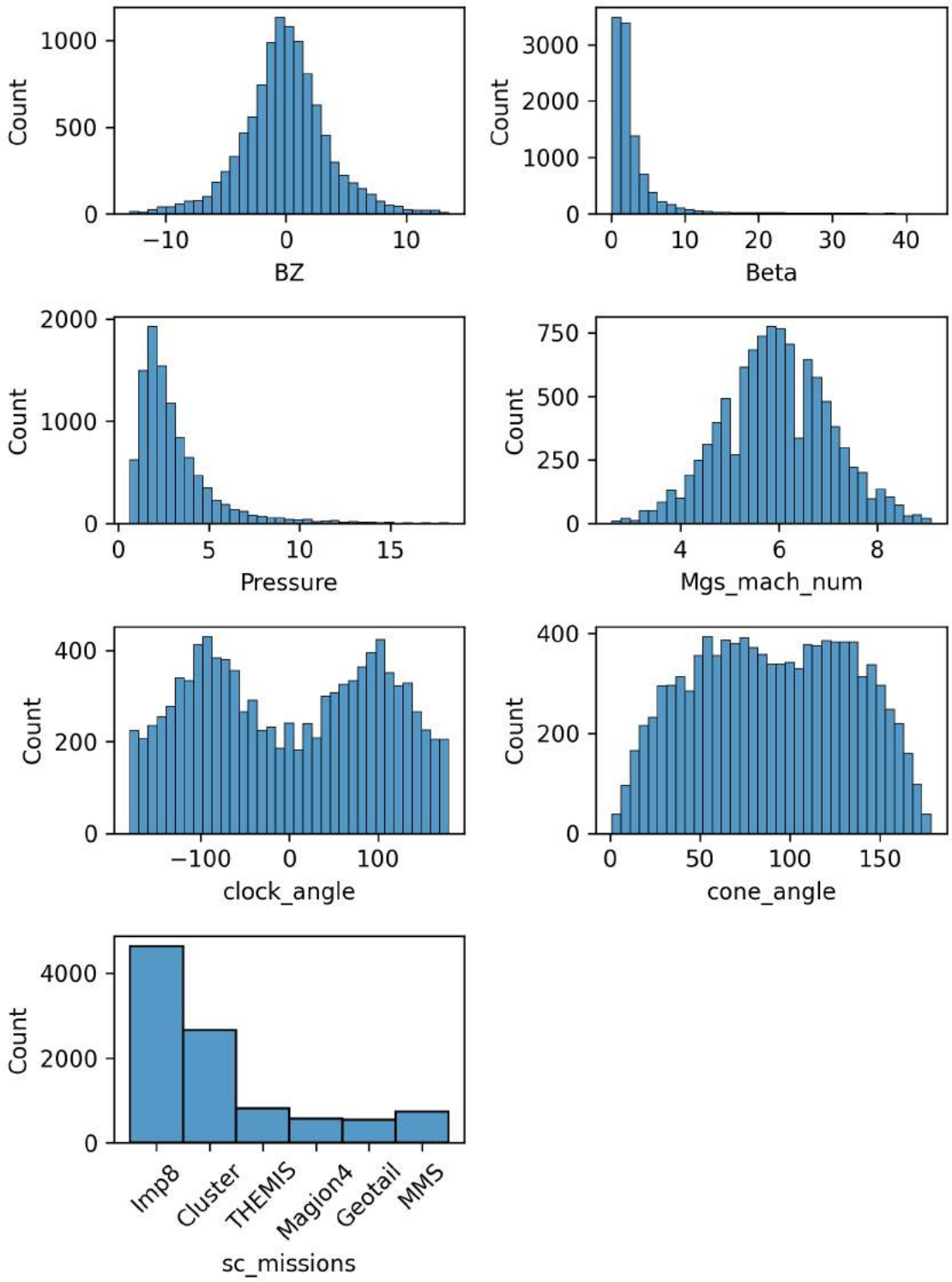

Figure 8-3: Histograms of the training data for each of the training features as well as the spacecraft mission. Note that dips present in the M_{MS} (labeled as "Mgs_mach_num") are due to the single decimal precision of Magnetosonic Mach number calculations provided by OMNI.

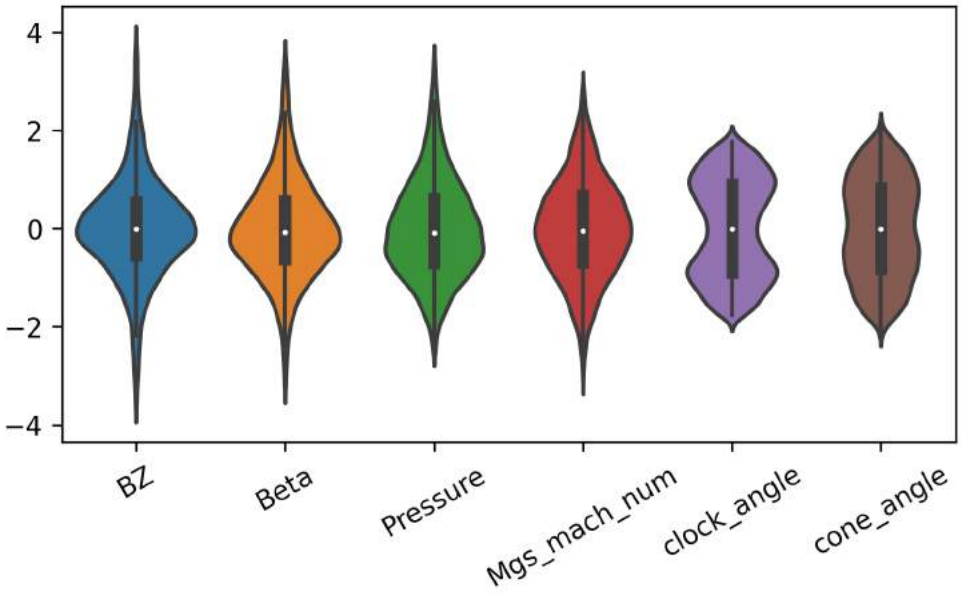

Figure 8-4: Violin plots of the features of the training data.

Chapter 9

Bow Shock Model Development

In this chapter, we explain how the bow shock model is developed using neural networks. We first discuss how we make a prediction in contrast to standard regression tasks, followed by discovering the optimal hyperparameters for our architecture, then exploring the use of an ensemble model in comparison to a single standard-trained model.

9.1 Prediction Design

In many regression tasks, the ultimate prediction of a neural network is directly compared against the observed value in order to calculate the residual between them. In the context of our bow shock modeling, this would lend the interpretation that a neural network needs to predict a radius so that a residual can be computed. This is true, but the predicted radius need not be the final output of a neural network designed for this task.

An important component of multiple previously developed bow shock models is their parameterization with respect to upstream solar wind conditions. Training a neural network to take input features and predict outright a corresponding bow shock shape can work, but can be difficult to interpret. We take a different approach by assuming a functional form for the radius that is reliant upon coefficients and use a neural network to predict them. The bow shock radius function we will use has already been introduced as Equation 1.12, and we repeat it here:

$$R(\theta, \phi; R_0, \alpha_0, \alpha_1, \alpha_2) = R_0(\frac{2}{1+\cos\theta})^{\alpha_0+\alpha_1\cos\phi+\alpha_2\cos^2\phi}. \tag{9.1}$$

The expression $R(\theta, \phi; R_0, \alpha_0, \alpha_1, \alpha_2)$ means that the bow shock radius R is a function θ and ϕ assuming the coefficients R_0, α_0, α_1, α_2 are provided. The interpretations of the coefficients and the symmetries and asymmetries they represent were previously covered in Section 1.2 when discussing the bow shock model of *Lu et al.* [2019]. Worth noting is that traditionally, functional characteristics of R_0 and α_0 were made by analyzing dayside ($X > 0$) fits for the former and nightside ($X < 0$) fits for the latter (again, see *Lu et al.* [2019]). This approach is not taken here and crossing fits via neural network estimation are done without compartmentalization.

Since three of the four coefficients are part of an exponential, care must be taken in handling the predictions of a neural network because it is common that sensible predictions only begin to arise after some training has undergone. To limit them, we truncate the coefficients for each α_i using a scaled and shifted sigmoid function of the form

$$\alpha_i \leftarrow -2 + 4(\frac{1}{1+e^{-a_i}}). \tag{9.2}$$

The * 4 multiplier stretches the limits of the sigmoid from (0,1) to (0,4) and the -2 shift centers the output at 0 such that the new scale is (-2,2).

9.2 Neural Network Architecture and Hyperparameter Optimization

The architecture we will use to predict these coefficients from the input features is that of a standard feed-forward neural network, or Multilayer Perceptron (MLP). We use the Pytorch (*Ansel et al.* [2024]) python library to implement it, and the optimal hyperparameters we

will determined using Optuna (*Akiba et al.* [2019]). To this end, we train the neural network on the training data and report the loss on the validation set. We use a learning rate of 10^{-3}, batch size of 32, L2 regularization constant of 10^{-4}, and 100 epochs of training with an early-stopping patience counter of 5. We vary the dropout probability, the number of hidden layers, and the number of neurons for each hidden layer according to:

1. $0.01 \leq dr \leq 0.5$,
2. $10 \leq h_i \leq 100$ for each hidden layer h_i,
3. $2 \leq n_h \leq 4$.

It may come across as unusual that, for all the hyperparameters associated with the training of neural networks, that the learning rate was not a varied hyperparameter. This is because varying this while also incorporating early stopping on a validation set can cause certain trials to be seen as best when realistically the validation loss only briefly dropped to a small value by chance. There are more advanced methods to avoid this during hyperparameter optimization, but nevertheless this is how it was done.

We perform three separate analyses which test for the optimal network hyperparameters using only two, three, or four hidden layers. We separate the trials in this way so that the sampler does not produce biased results (e.g. if the sampler if asked to generate a guess for how many neurons should be in the third hidden layer, but only two hidden layers are used). Each analysis consists of three hundred trials. The model parameters are saved at the end of each epoch in order that they can be retrieved according to the optimal validation loss. Parallel coordinate plots for each of the three analyses can be seen in Figures 9-1, 9-2, and 9-3. The best trial among all three analyses comes from the optimization done for four layers and we next train a single model using similar hyperparameters.

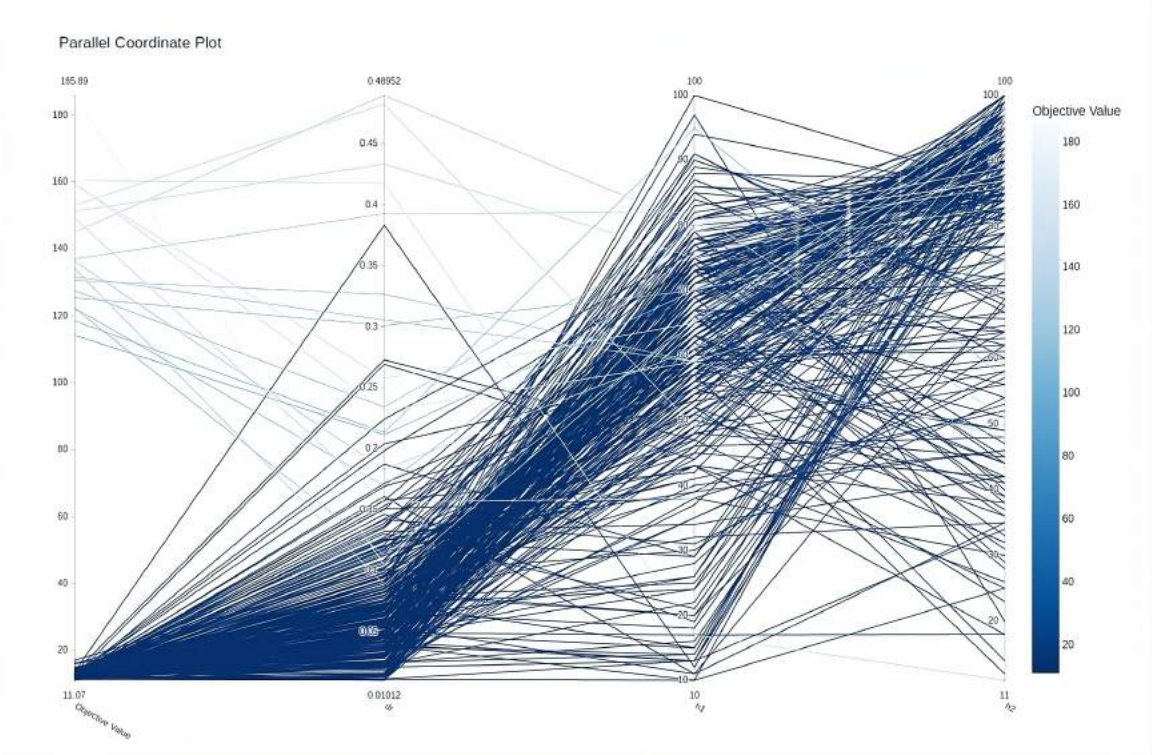

Figure 9-1: A parallel-coordinates plot of all trials done for the hyperparameter optimization for two layers. The leftmost axis, labeled "Objective Value," shows the range of losses on the validation set where it's clear that many of the trials result in comparable losses $\leq$ 20. The axis to its right, "dr" for dropout, shows the dropout value that was used for each trial. The subsequent axes, "h1" and "h2," represent the number of neurons sizes chosen for hidden layers one and two. The lines that connect across these axes represent the values used for each trial. The validation loss is also illustrated with a color bar to the far right, with larger losses in lighter blue and smaller losses in darker blue. The best trial had validation loss 11.06 with parameters $dr = 0.0440$, $h_1 = 78$, and $h_2 = 86$.

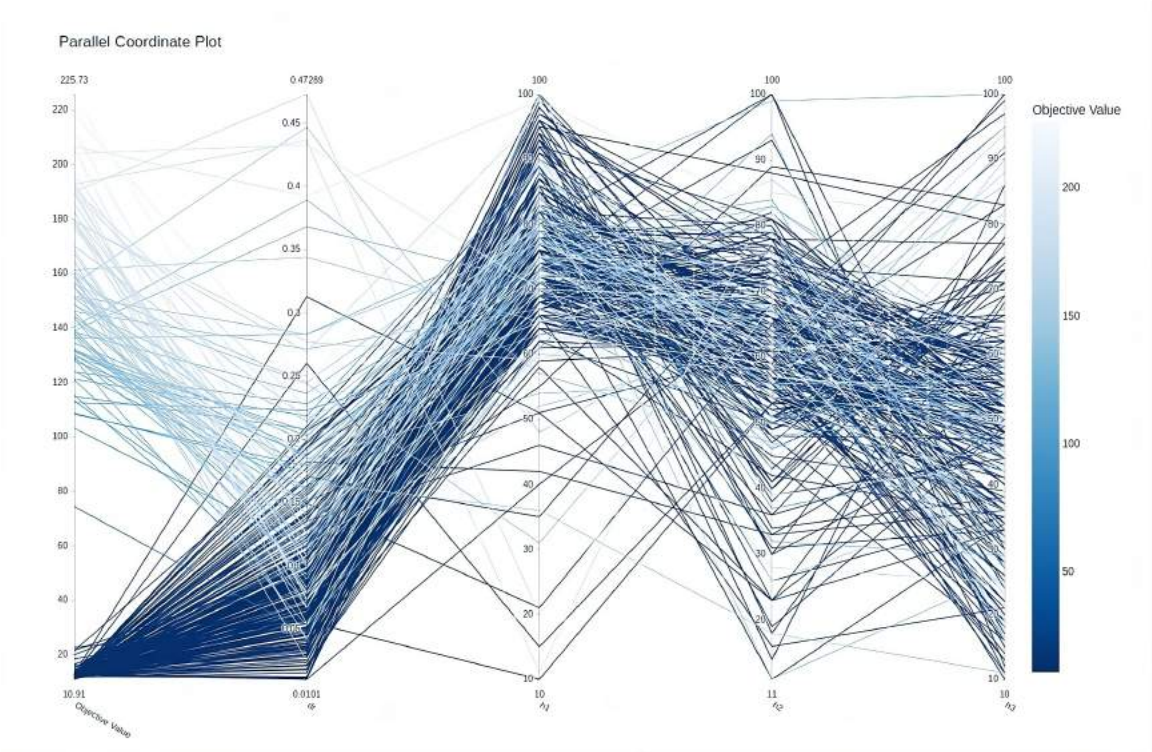

Figure 9-2: A parallel-coordinates plot like Figure 9-1, but for training neural networks with three hidden layers. The best trial had a validation loss of 10.91 with parameters $dr = 0.0862$, and $\{h_i\} = [70, 74, 59]$.

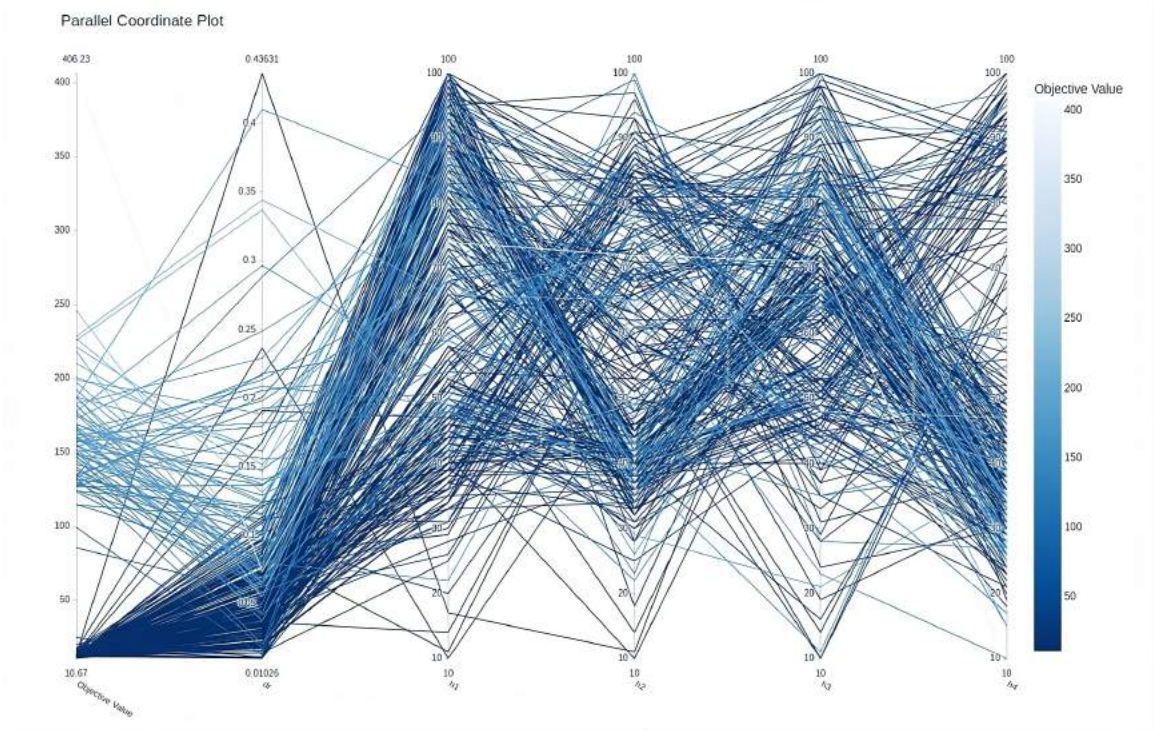

Figure 9-3: A parallel-coordinates plot like Figure 9-1, but for training neural networks with four hidden layers. The best trial had a validation loss of 10.67 with parameters $dr = 0.0253$, and $\{h_i\} = [90, 59, 43, 28]$.

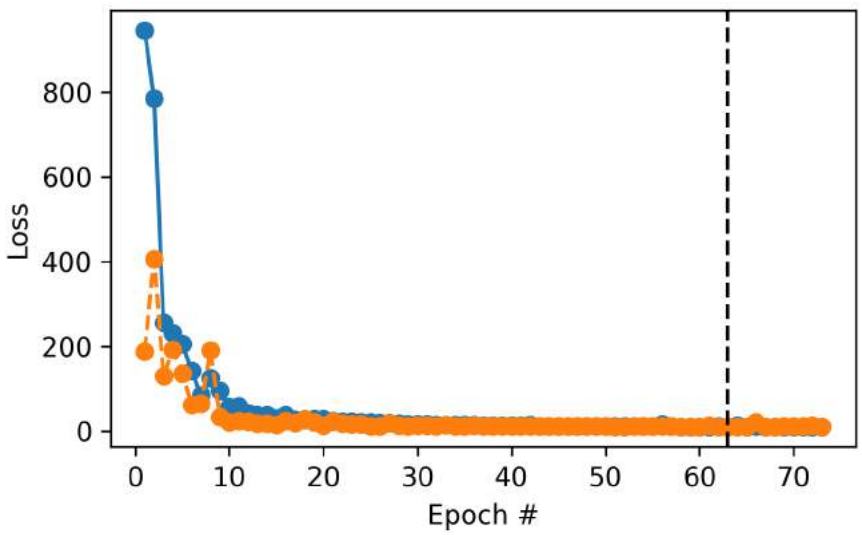

Figure 9-4: This plot shows the training (blue, solid lines) and validation (orange, dashed lines) losses as a function of epoch number (with counting starting from 1). The vertical dashed black line (epoch 63) indicates where the model weights were loaded from after the patience count had been reached due to early-stopping. The training and validation losses at that epoch are 10.64 and 10.69.

9.3 Training a Single Model

For training a single model, we will again use an L2 regularization of 10^{-4}, a learning rate of 10^{-3}, and 100 epochs of training. However, we increase the patience of early stopping to 10. We use the hyperparameters of the best trial which comes from the four-layer analysis and has hyperparameters $dr = 0.0253$ and $\{h_i\} = [90, 59, 43, 28]$. The training and validation losses of the resulting model can be seen as a function of epoch number in Figure 9-4.

9.4 Building the Ensemble

To build the ensemble, we will train multiple models using the same hyperparameters as the single model from the previous section. However, to introduce greater diversity into the model predictions, we will not simply train an ensemble on the same exact training dataset. Instead we will utilize bootstrap aggregation, or bagging, to train models on unique subsets of the training data, which will meaningfully diversify the cumulative predictions of the

ensemble. While each member of the ensemble is trained based on the residuals between its radius predictions for the bootstrap sample and the observations, it is the coefficients predicted by each network that will ultimately comprise the ensemble as seen in Figure 9-5.

We prepare 300 bootstrap training samples by sampling from the training set with replacement. Defining the ensemble to be the collection of all 300 models is too naive an approach as some of the models could be poorly trained or have been trained on insufficiently diverse training samples. We prune the ensemble by initially defining the ensemble to contain all 300 members. The ensemble validation loss is then computed based on the mean-coefficient predictions. The validation loss of each ensemble member is also computed, and the model with the worst validation loss is then removed. The ensemble is then reformed from the remaining members and this process is repeated until a single bootstrap-trained model, the one with the best validation loss, remains. The resulting validation loss as a function of number of models removed is shown in Figure 9-6. It is clear from the figure that the model trained using the entire training set performs better than the best bootstrap-trained model; it is also evident that the ensemble, even for just a meager three members, is clearly superior to both. The optimal value corresponds to a validation loss of 10.0703 for a 16-member ensemble. Since there is a comparable loss out to the third decimal place for a 12-member ensemble, we opt for this instead as it is more parsimonious. Comparisons of the parameterizations of this ensemble vs the single model trained in Section 9.3 will be made in the following section.

9.5 Ensemble and Single Model Comparisons

We define our ensemble to make radius predictions according to the mean of the coefficients instead of the mean of the radii. This does not afford any significant error as we find the mean-squared-error loss on the validation set using the mean-coefficient method to be

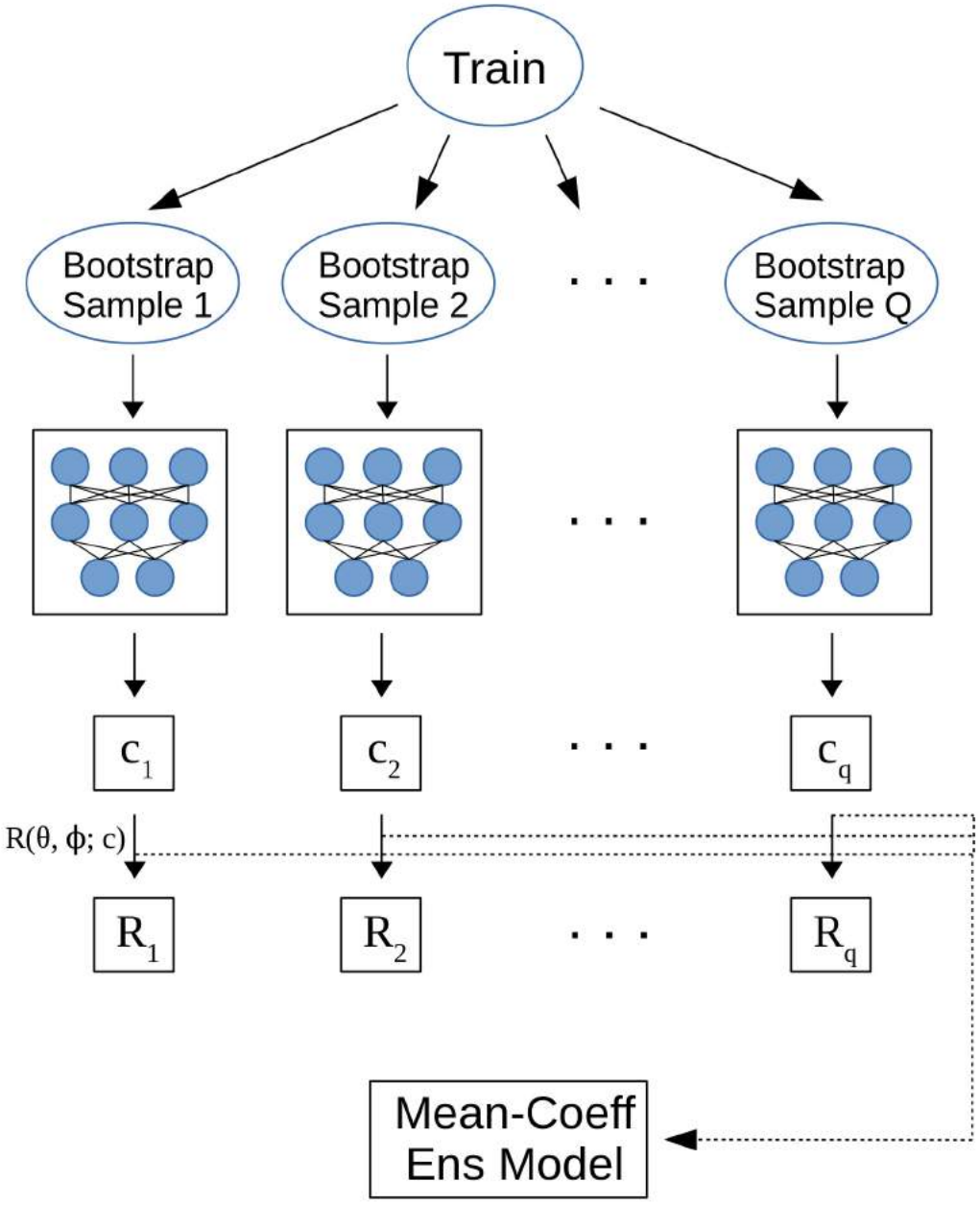

Figure 9-5: Diagram representing how the ensemble is formed. For an ensemble of size Q, Q bootstrap samples are created from the original training set and individual MLPs are trained on each bootstrap sample such that each MLP predicts its own coefficient vector. These coefficients, along with the θ and ϕ associated with each bow shock crossing, can be supplied to our model function (represented here as $R(\theta, \phi; c)$) to predict a radius. The solid lines descending from the Train set at the top correspond to the models and predictions for each bootstrap sample. The dashed lines branching off between the coefficient and radius predictions represent how the ensemble is formed from the mean of the coefficients instead of the radii, allowing parameterization of each bootstrap model to be analyzed individually and collectively as shown in Figure 9-7.

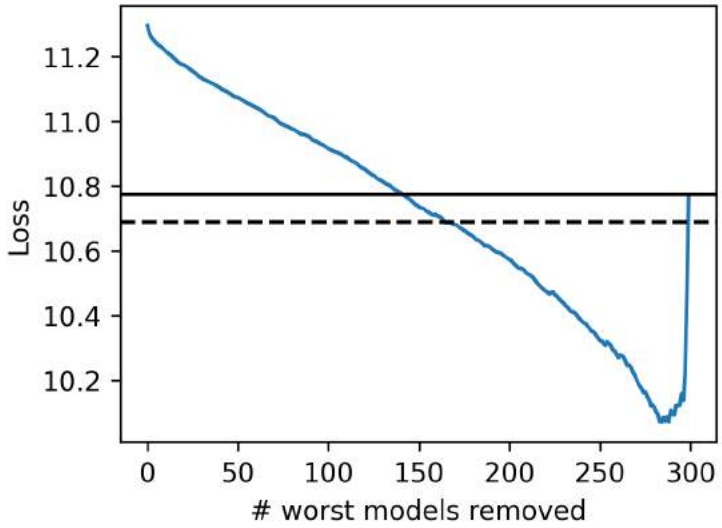

Figure 9-6: The ensemble validation loss as a function of number of worst models removed is shown above. The solid horizontal line is the validation loss of the best bootstrap-trained model, and the dashed horizontal line is the validation loss of the single model that was trained seeing the entire training set. Note that the true minimum occurs at x = 284 with loss = 10.0703 (or a 16-member ensemble). A loss of 10.0707 occurs for x = 288, a 12-member ensemble. In the interest of being more parsimonious, we will use the 12-member ensemble since it has a comparable loss out to three decimal places with a 25% reduced ensemble size.

slightly below that of the mean-radius method (10.041 vs 10.071). The mean-coefficient methods means that we can analyze the coefficient parameterizations for each model of the ensemble as well as the ensemble mean, which we show in Figure 9-7. Some coefficients have tight ensemble distributions such as the paramaterizations of R_0 and α_0 with respect to M_{MS} and dynamic pressure, which bear resemblances to their constrained functional form used in other bow shock models (recall the inspiration of the inverse Mach number relationship from *Spreiter et al.* [1966] or the $D_p^{-1/6}$ dynamic pressure relation mentioned in Section 1.2). Others have wider distributions, such as R_0 and α_0 as a function of plasma beta or R_0 as a function of cone angle. Also of note is the smaller ranges of α_1 and α_2 in the parameterizations. This is not that surprising as the terms represent the North-South and azimuthal asymmetries respectively and the bow shock can be modeled quite well as a radially symmetric paraboloid (a "rule-of-thumb" rough bow shock fitting can be approximated using our model function with $R_0 = 13.75\ R_E$, $\alpha_0 = 0.75$, and $\alpha_{1,2} = 0$).

We can also make comparisons between the ensemble mean and fully-trained single model, shown in Figure 9-8. Several ensemble mean profiles exhibit similar characteristics as the fully-trained counterpart, looking again at R_0 and α_0 as functions of M_{MS} and dynamic pressure. Others however show notable deviation, such as a number of the $\alpha_{1,2}$ plots and the continual decrease in the ensemble-mean of α_0 with respect to plasma beta. Still others exhibit almost static behavior for some thresholds, such as the parameterizations of α_2 for plasma beta and dynamic pressure. Some of these parameterizations may exhibit unusual curves, but we will ultimately gauge it on its resulting bow shock contours and it performance relative to another bow shock model in the next section.

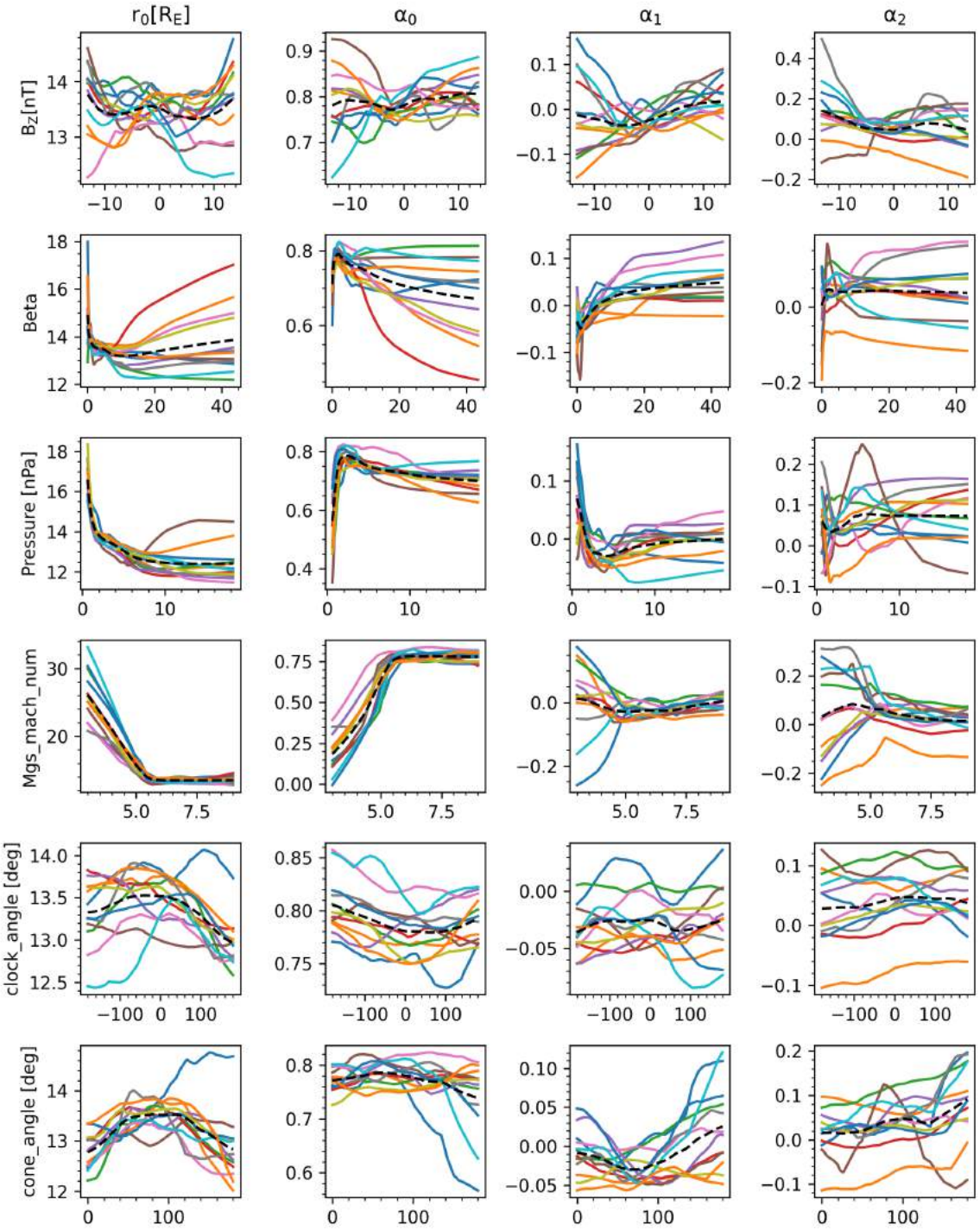

Figure 9-7: Each solid line is a parameterization of a coefficient by a single model of the ensemble. The dashed black line is the ensemble mean. Note the "inverted" nature of the x and y labels relative to the plots. The column labels denote the coefficient being parameterized (e.g. the left-most column shows the predictions of R_0 relative to the input features) and the row labels represent the input feature being varied for the coefficient predictions (e.g. the top row shows the parameterizations of the four coefficients as a function of B_Z). So in effect, the row labels correspond to the values on the x axis and the column labels to the values on the y axis.

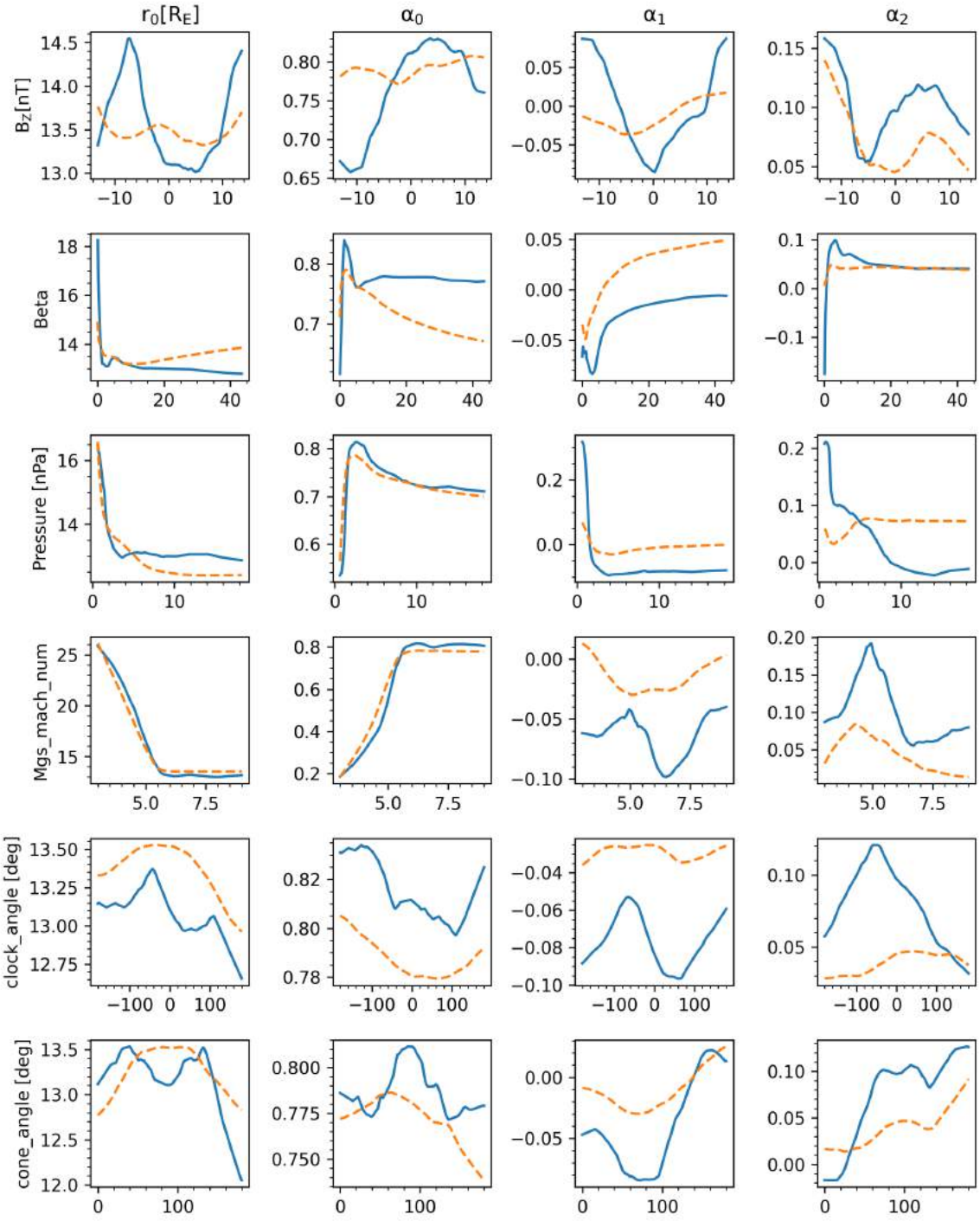

Figure 9-8: The x and y axes of the plots has similar structure to the plots of Figure 9-7. The difference here is to showcase the difference in predictions between the fully-trained single model coefficients (shown with a solid blue line) and the ensemble-mean coefficients (shown with a dashed orange line).

Chapter 10

Bow Shock Model Results

In this chapter, we will analyze the contours predicted by the constructed bow shock model and compare it with the model of *Chao et al.* [2002].

10.1 BS Shape Predictions

To analyze the contours of the model, we need a consistent set of coefficients to base our results on. Since our ensemble can produce a 3d array of coefficients (of shape [ensemble size, number test points, number coefficients]) from the test set, we will take the ensemble average and then the average over the test set (i.e. averaging over the 1st and 2nd dimensions in that order). This yields coefficient values of (R_0, α_0, α_1, α_2) = (13.8486 R_E, 0.7670, -0.0141, 0.0355). Cross sections in the x-y, x-z, and y-z planes are seen in Figure 10-1 which show both the Z and azimuthal asymmetry. The aspect ratio is the range over Z divided by the range over Y. To further illustrate the azimuthal asymmetry, the aspect ratio is computed as a function of X from 10 to -30 R_E in the bottom right of the same figure.

We next explore the changes in shape of our model for the parameters B_Z, plasma beta, dynamic pressure, and magnetosonic mach number in Figures 10-2, 10-3, 10-4, and 10-5, respectively. The Chao model for the corresponding parameters is shown as well. In these comparisons, we will use a consistent set of parameters by taking the average parameter values from the test set. These parameters (to three decimal places) are B_Z = -0.101 nT,

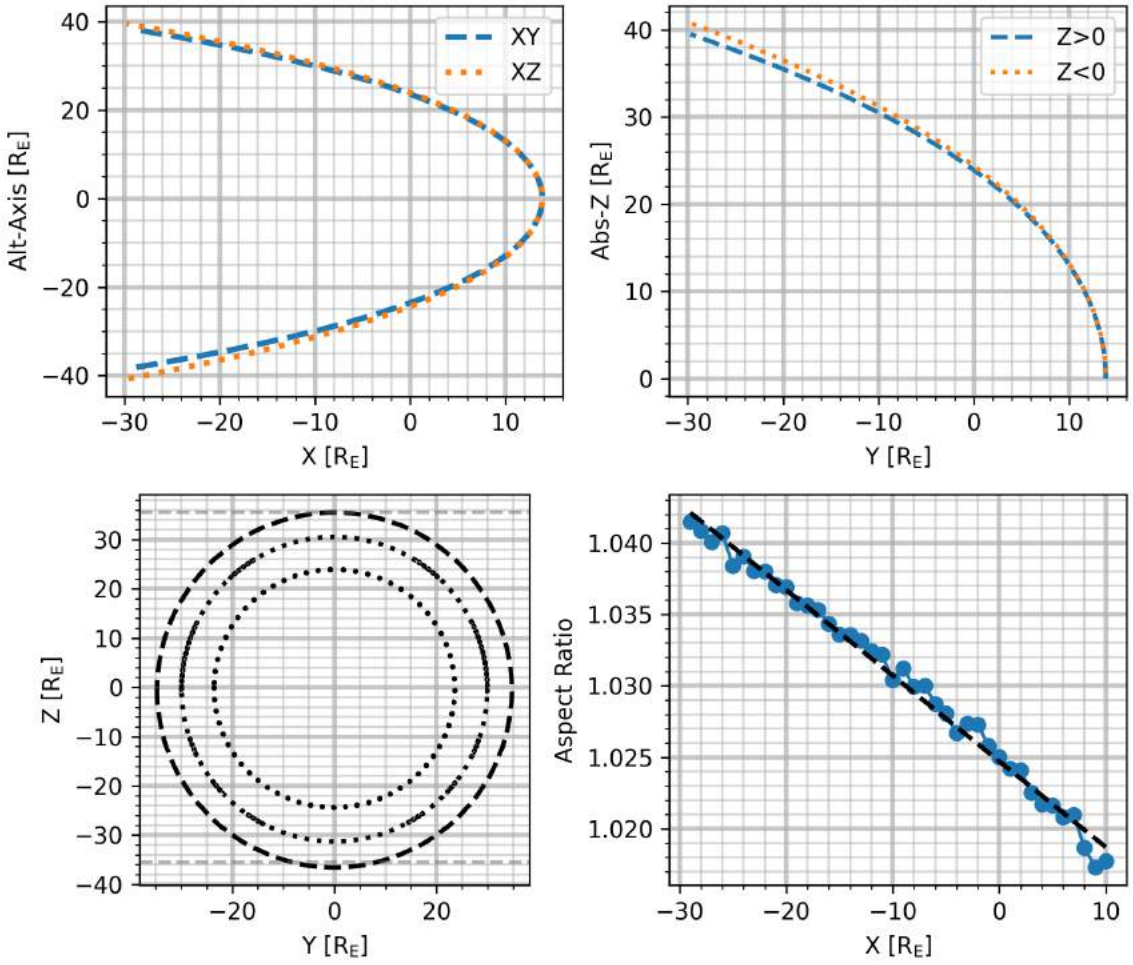

Figure 10-1: **Top Left:** The cross sections for x-y (along z=0) and x-z (for y=0) are plotted as the dashed inner blue and dotted outer orange lines. The difference between them grows with a decrease in X, illustrating the azimuthal asymmetry of the model. **Top Right:** The same x-z cross section as shown in the top left plot, but the y axis is the absolute value of Z. The dashed inner blue line is the x-z cross section for the positive Z axis and the dotted outer orange line is the same cross section but corresponding to the negative Z axis, showing the shows the Z asymmetry generated by the coefficients. **Bottom Left:** Three different y-z cross sections are shown for x=0 (inner-most, dotted), x=-10 (middle, dash-dotted), and x=-20 (outer-most, dashed). The top horizontal dashed grey line indicates the maximum z value reached by the x=-20 cross section. The same line is plotted symmetrically along negative Z to show the slight north-south asymmetry. **Bottom Right:** The aspect ratio (z-range along y=0 divided by y-range along z=0) for each y-z cross section is plotted as a function of X from 10 to -30 R_E.

$\beta = 2.954$, $D_p = 3.028$ nPa, $M_{MS} = 5.962$, ϕ_B = -2.641 degrees, and $\theta_B = 89.214$ degrees. These will not be modified except when varying a single parameter at a time.

First we explore B_Z in Figure 10-2. There is little change in the X-Y or X-negative Z cross sections for our model, but there is an increase along X-positive Z, resulting in an enhanced aspect ratio for positive B_Z. As seen in Figure 10-3, beta causes a continual contraction in X-Y due to the diminishing α_0. X-positive Z is practically static, but X-negative Z contracts with increasing beta, yielding an increasing aspect ratio.

The dynamic pressure in Figure 10-4 shows that for a relatively quiet solar wind pressure of 2 nPa, our model aligns with Chao in the X-Y plane. Increasing pressure causes the X-Y cross section to contract but not as much as that of Chao where the 10 nPa contour of our model aligns with the 5 nPa contour of Chao. R_0 decreases from almost 14 to 12 R_E. The aspect ratio for the for the low pressure case aligns almost exactly with the average and higher pressures exhibit a similarly large aspect ratio, reflecting that the Y-axis compression is larger than the Z-axis compression. This is a reflection of the α_1 and α_2 parameterizations wherein α_1 begins converging to 0 for $D_p > 10$ nPa (meaning a North-South symmetry) and α_2 converges to a constant ~ 0.075 for $D_p > 5$ nPa (meaning a constant azimuthal asymmetry).

The last parameter in common between our model and the Chao model is the magnetosonic mach number M_{MS}, which we show in Figure 10-5. The lower M_{MS} value of 5 shows a quite large R_0 close to 16 R_E with visually notably smaller X-Y flaring (here, just the α_0 term because $\cos(90) = \cos(270) = 0$) than Chao. The parameterization of R_0 with M_{MS} does technically show a monotonically decreasing R_0 for increasing values of M_{MS}, but with an almost "switch-on" occurring at $M_{MS} = 5$. The X-Z contour only slightly changes in response to higher M_{MS}, causing a continual decrease in the aspect ratio.

Since we have incorporated both B_Z and the clock angle as input features, we need

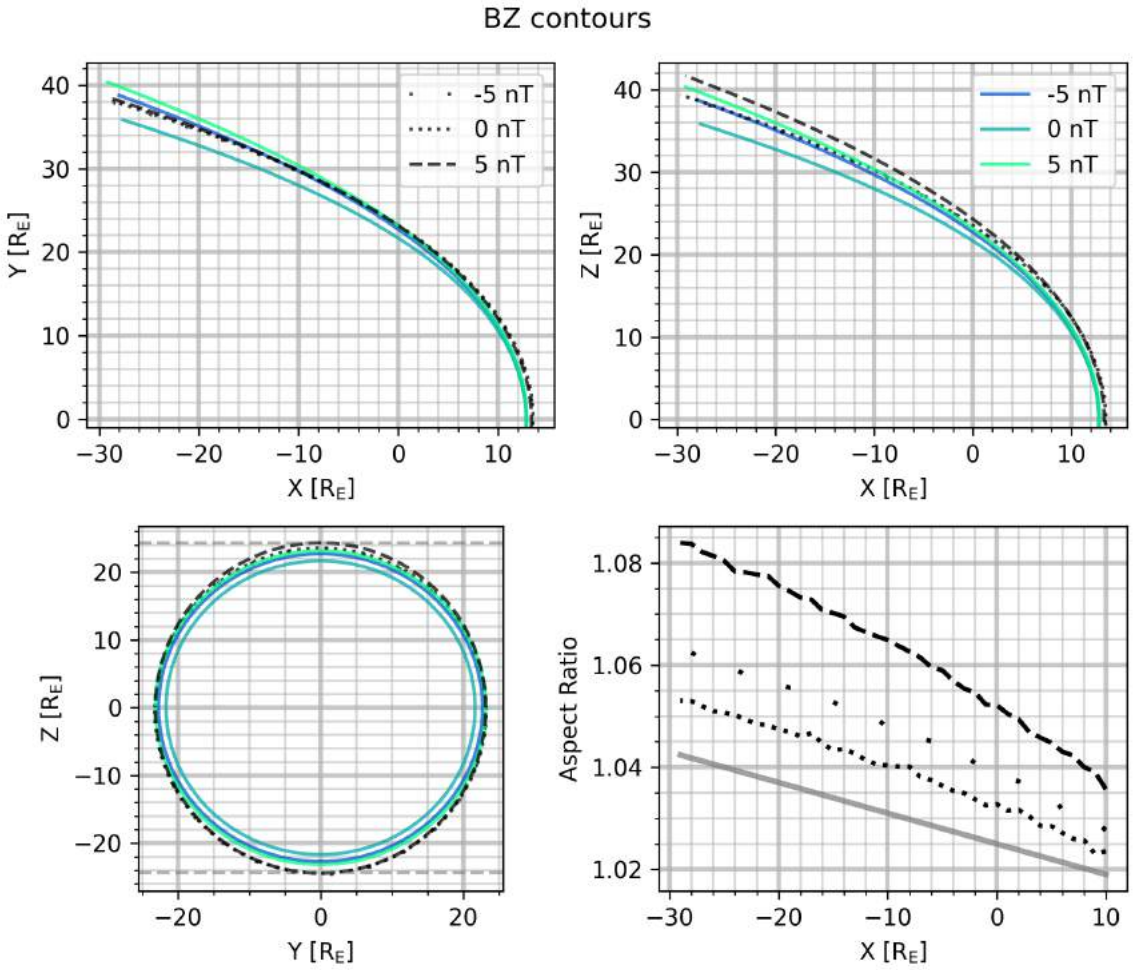

Figure 10-2: Three different model contours are shown for both our model and the Chao model according to B_Z = -5, 0, and 5 nT. Our model contours for these values correspond to sparsely dotted (top of the legend in the top left plot), normally dotted (middle of the legend), and dashed (bottom of the plot) lines. The Chao model contours for these values are colored according to blue (the darkest color, top of the legend in the top right plot), azure (the lighter blue color, the middle of the legend) and cyan (the brightest color, the bottom of the legend). **Top left:** The X-Y cross sections for the positive Y axis at Z=0. **Top right:** The X-Z cross sections for the positive Z axis at Y=0. **Bottom left:** The Y-Z cross sections at X=0. The horizontal dashed grey line at the top indicates the maximum z-axis value attained by our model. A symmetric horizontal dashed line is shown at the bottom to help visually accentuate our the Z-axis asymmetry of our model. **Bottom right:** The aspect ratio of our model as a function of X. The solid grey line corresponds to the linear fit of the aspect ratio as a function of X seen in the bottom right of Figure 10-1 (that is, the aspect ratio according to average coefficients). Note that our model has Y axis symmetry but not Z axis or azimuthal symmetry, meaning that the X-Y cross sections are symmetric for the negative Y axis but not for the negative Z axis. The Chao model has Y axis, Z axis, and Y-Z symmetry and so forms a radially symmetric parabola in the X-Y and X-Z planes and constitutes a perfect circle in Y-Z cross sections. The full B_Z parameterization can be seen in the top row of Figure 9-8.

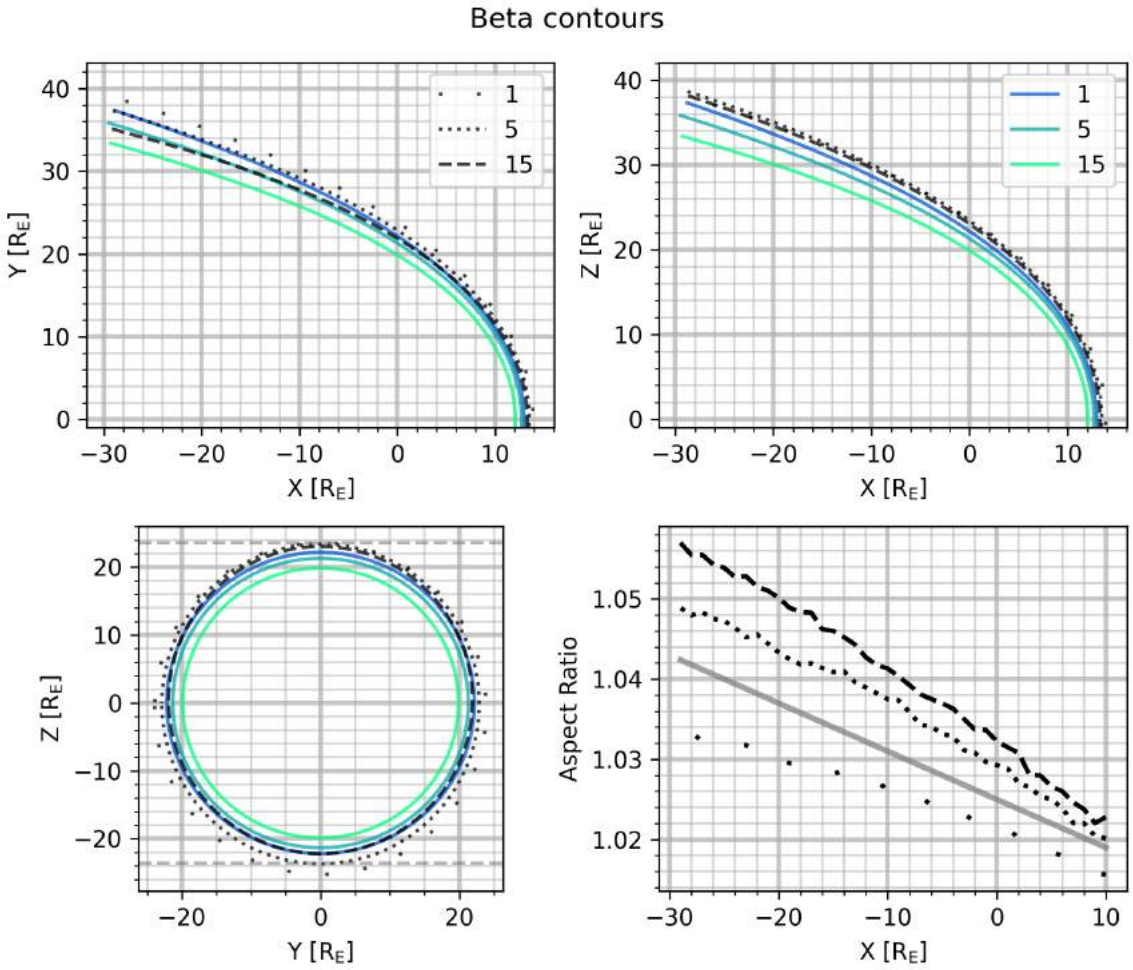

Figure 10-3: Similar to the plot and model structure of Figure 10-2, but considers the different model contours for $\beta = 1$, 5, and 15. The full β parameterization can be seen in the second row of Figure 9-8.

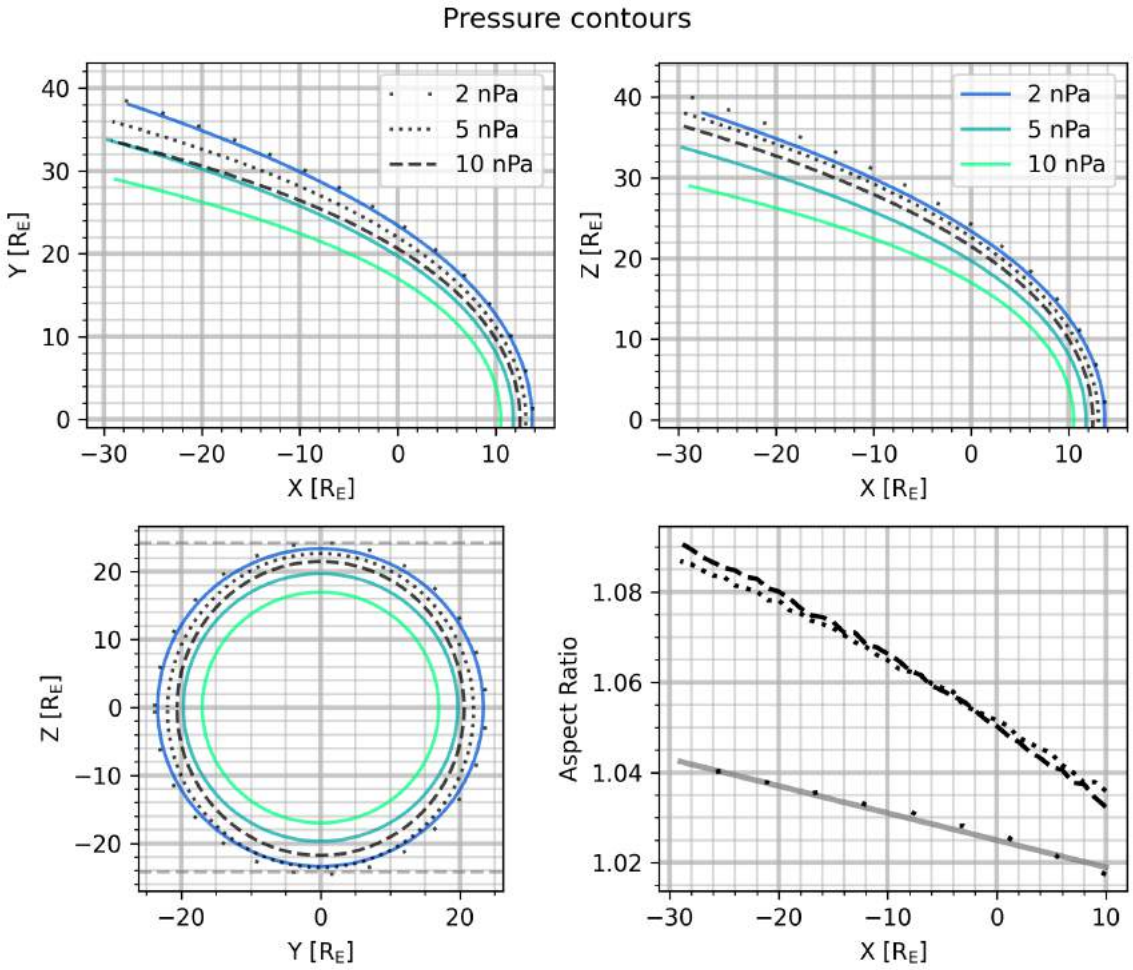

Figure 10-4: Similar to the plot and model structure of 10-2, but considers the different model contours for D_p = 2, 5, and 10. The full D_p parameterization can be seen in the third row of Figure 9-8.

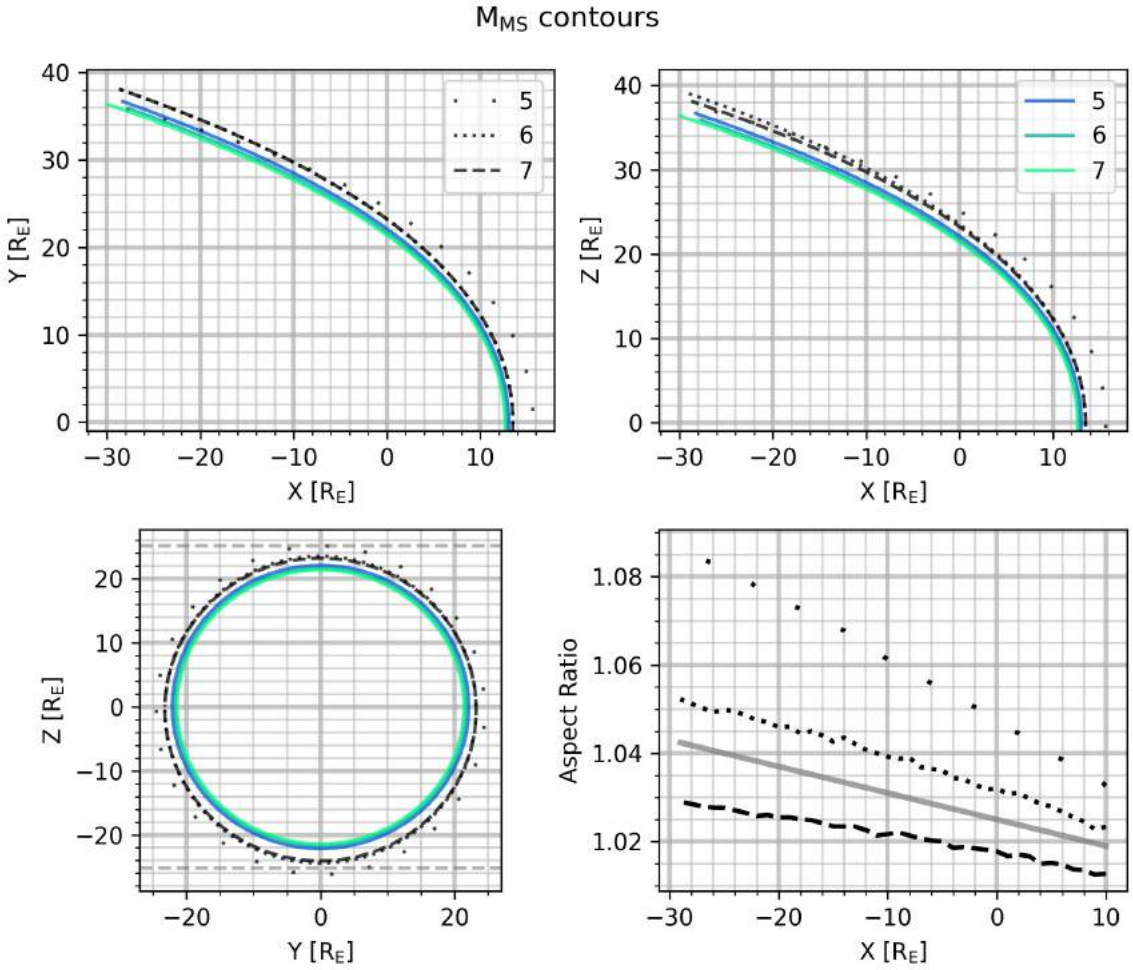

Figure 10-5: Similar to the plot and model structure of 10-2, but considers the different model contours for $M_{MS} = 5$, 6, and 7. The full M_{MS} parameterization can be seen in the fourth row of Figure 9-8.

to take care in the value supplied to B_Z when studying the changes in clock angle. The average B_Z we have been using, $\sim$ -0.1 nT, means that the varying clock angle will only reflect small values of B_Y or be outright incorrect if the clock angle indicates that the azimuthal IMF_{YZ} vector is in the wrong quadrant (i.e. if ϕ_B is -135°, then B_Z cannot be positive). This means that B_Z should be positive for $|\phi_B| \leq 90°$ and negative otherwise. Because of this, we will show two plots representing the change of the model in clock angle. One will use a B_Z magnitude corresponding to the average -0.1 nT and the other will use a B_Z with magnitude 1 nT (with sign correction accounted for in clock angle for both plots).

The bow shock contours for varying clock angle with a magnitude $B_Z = 0.1$ nT are shown in Figure 10-6. It exhibits slight oscillation in the X-Y plane and even less so in the X-positive Z cross section. This results in two predominant clusters of lines in the aspect ratio: A largely eastward IMF ($45 \leq \phi_B \leq 135$ degrees) has a larger aspect ratio than typical before diminishing back to the average value for increasing (southward-turning) ϕ_B. And the other is the largely westward IMF (-45 $\leq \phi_B \leq$ -135 degrees) which exhibits smaller aspect ratios than average. Also present is the divergence of predictions as ϕ_B converges southward. Turning southward from east, the aspect ratio clearly decreases, but turning southward from west, the aspect ration remains clustered and does not increase.

Next we analyze the bow shock contours for varying clock angle with magnitude 1 nT in Figure 10-7. Like in the previous Figure, slight oscillations are again present in the X-Y cross sections but are a little more pronounced in the X-positive Z plane with visible contraction between eastward and westward ϕ_B. Correspondingly, a maximal aspect ratio occurs for eastward ϕ_B and minimal aspect ratios for north, south, and westward ϕ_B.

Overall, these results show mixed agreement with *Wang et al.* [2016] where they observed in global MHD simulation results that the tail cross sections are stretched in directions perpendicular to the $IMFYZ$ direction. In both bow shock contours for $|B_Z| = 0.1$ and

1 nT, it is clear that eastward ϕ_B results in an expansion along X-positive Z and the maximized aspect ratios reaffirm this. Northward and southward IMF_{YZ} both produce minimized aspect ratios, indicative of a relative stretching along X-Y (although a divergence in the periodicity of ϕ_B is visible in the X-Y cross sections). However, there is also a matter of explicit disagreement in the clock angle. *Wang et al.* [2018] analyzed spacecraft observations and observed an increase in R_0 with westward to eastward ϕ_B (or more exactly, going from B_Y = -10 to 10 nT). From the clock angle parameterizations of Figure 9-7, our model predicts R_0 to only slightly increase from south-to-west-to-north IMF_{YZ} and then to decrease from north-to-east-to-south. Our flaring angle α_0 also behaves oppositely where *Wang et al.* [2018] observed a slight increase going from west to north (barring a single sharp) before decreasing from north to east.

The cone angle θ_B dependence is seen in figure 10-8. For increasing cone angle, the X-Y cross section first expands then compresses, the X-positive Z cross section strictly increases, and R_0 only decreases slightly. Like with the clock angle, we again see two clusters of lines based on the cone angle. For quasi-sunward IMF ($\theta_B \leq 45$ degrees), the aspect ratio only increases from 1.01 to 1.02 for X = 10 to -30 R_E, indicating almost azimuthally symmetric Y-Z cross sections. The aspect ratio then continually increases to just above the average over a cone angle of 45 to 90 degrees (an azimuthal or radial IMF in which B_X reduces to $\sim$ 0 nT). The other cluster of lines then occurs for quasi-azimuthal IMF where $90 \leq \theta_B \leq 135$ degrees. This corresponds to the maximally expanded X-Y cross section, causing the aspect ratio to briefly slightly decrease. At largely anti-sunward cone angles ($\theta_B \geq 135$ degrees), the X-Y cross section begins to contract while the X-positive Z cross section continues expanding. It can be noted from the Y-Z cross sections that the contour continually contracts along the negative Z axis and that contour only slightly contracts along the positive Z axis. This is because these cross sections are taken as X = 0 cuts.

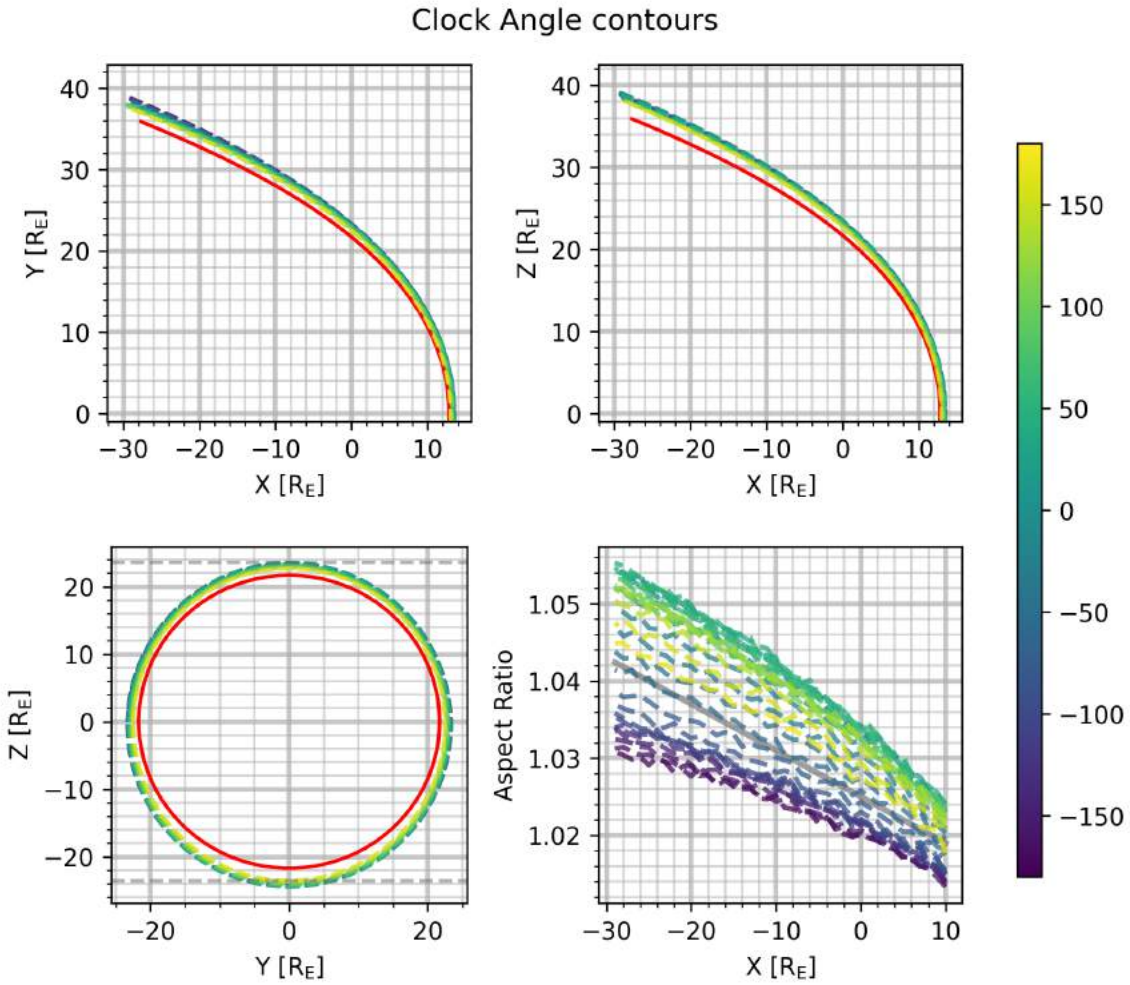

Figure 10-6: The plot structure is similar to that of 10-2 but now analyzes how our model behaves with respect to the clock angle ϕ_B for a B_Z magnitude of 0.1 nT. The Chao model is not parameterized for clock angle and so a single contour is shown using the average test set parameters. 30 different contours are plotted according to the colorbar at right ranging from -180 to 180 degrees. Recall that ϕ_B is defined as zero along the positive Z axis and increases along the direction of the positive y axis (i.e. clock-wise when looking straight along the negative X axis) such that ϕ_B = 0 is northward positive Z axis), ϕ_B = -180 = 180 is southward (negative Z axis), ϕ_B = 90 is eastward (towards dawn, positive Y axis), and ϕ_B = -90 is westward (towards dusk, negative Y axis). Eastward ϕ_B reduces the bow shock aspect ratio and westward ϕ_B causes it to increase. From the X-Y and X-Z plots, this change is moreso due to contraction and expansion along the Y axis, respectively. The lack of periodicity is apparent with ϕ_B = 180 exhibiting a comparable aspect ratio to the average and ϕ_B = -180 possessing the smallest aspect ratio. The full ϕ_B parameterization can be seen in the top fifth of Figure 9-8.

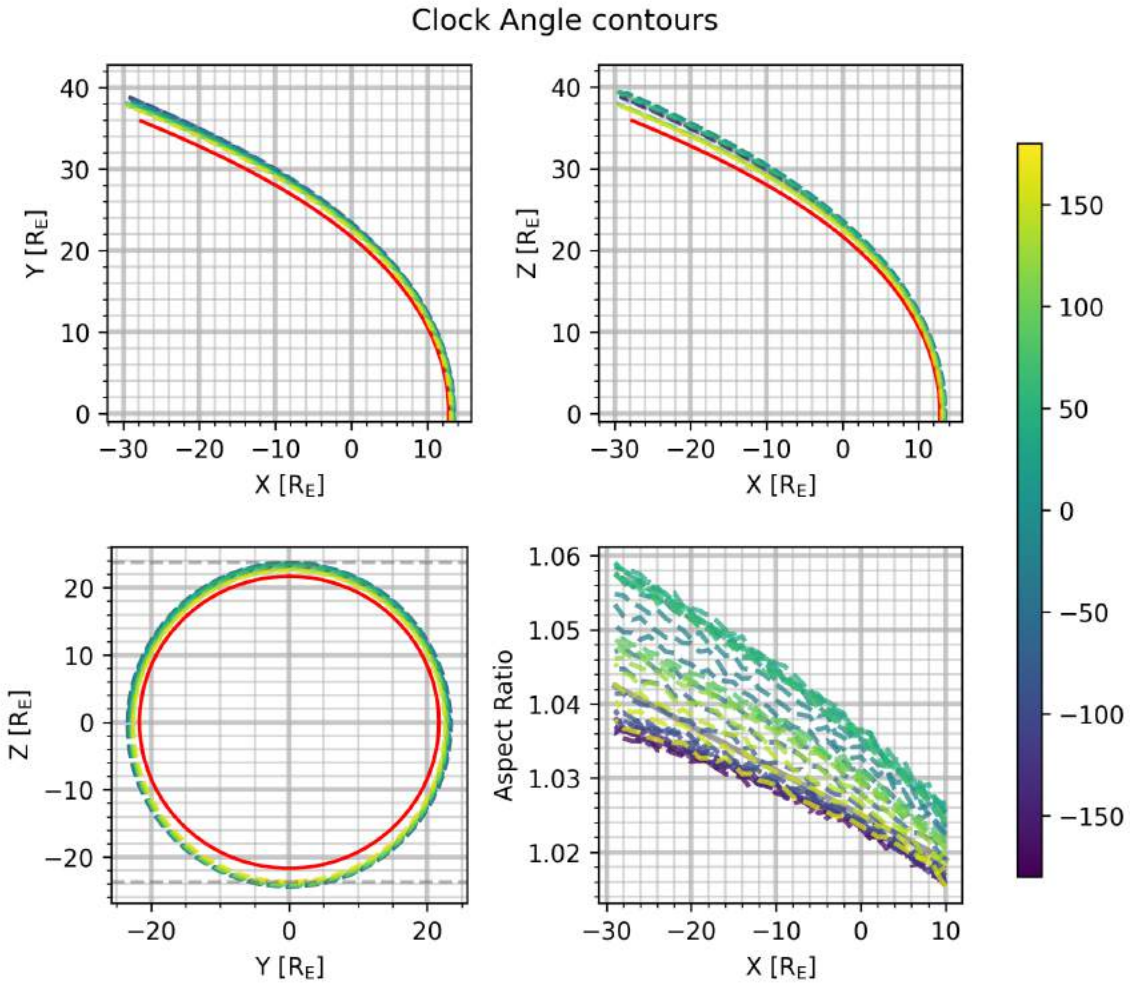

Figure 10-7: Same as Figure 10-6, but with a B_Z magnitude of 1 nT. A divergence from the periodicity of ϕ_B is again apparent in the X-Y and X-Z cross sections but is not present in the aspect ratios. The aspect ratio heatmap indicates that there is a maximal aspect ratio for roughly eastward ϕ_B. Conversely, a westward ϕ_B has comparable minimal aspect ratio to a northward or southward ϕ_B.

R_0 decrease throughout all of this. This X-Y contraction conjoined with an X-Z expansion causes the aspect ratio to climb.

10.2 Test Set Comparison with *Chao et al.* [2002]

Having covered the parameterizations and visual characteristics of the bow shock shape, we now proceed to show numerical comparisons with Chao on the test set. When interpreting the results, two points should be stressed: (1) It should be recalled from the discussion of *Chao et al.* [2002] in Section 1.2 that their fitting was done for crossings that were only aberrated according to V_Y and not V_Z whereas ours was aberrated for both, and (2) our test set contains a small number of large outliers that were not removed by our thresholding done in Section 3.2.7. To address the latter point, two residual plots will be shown for each comparison. One showing the residuals of both models on the original test set, and another showing the residuals with the poorest 5% of predictions (according to *Chao et al.* [2002]) removed, giving a small handicap in favor of the model of *Chao et al.* [2002].

The residuals for the test set are shown in Figure 10-9 and show largely good agreement for the original test set (barring the exceptional predicted outlier and a small handful of moderate outliers). Removing 5% of the poorest Chao predictions reveals that our model still has about one point less in loss, although the some of the largest errors in prediction are made by our model. This makes sense as the Chao model was designed to best fit the average bow shock positions as parameterized by B_Z, dynamic pressure, M_{MS}, and plasma beta whereas our coefficients are nonlinear functions of these inputs (as well as clock and cone angles). We also show how our model performs in the dayside and nightside regions in Figures 10-10 and 10-11. The dayside plot shows the Chao model to have a 7% lower error than ours (3.21 for Chao vs 3.46) and indicates that both models slightly overpredict the radii with ours overpredicting a bit further. The nightside plot shows our

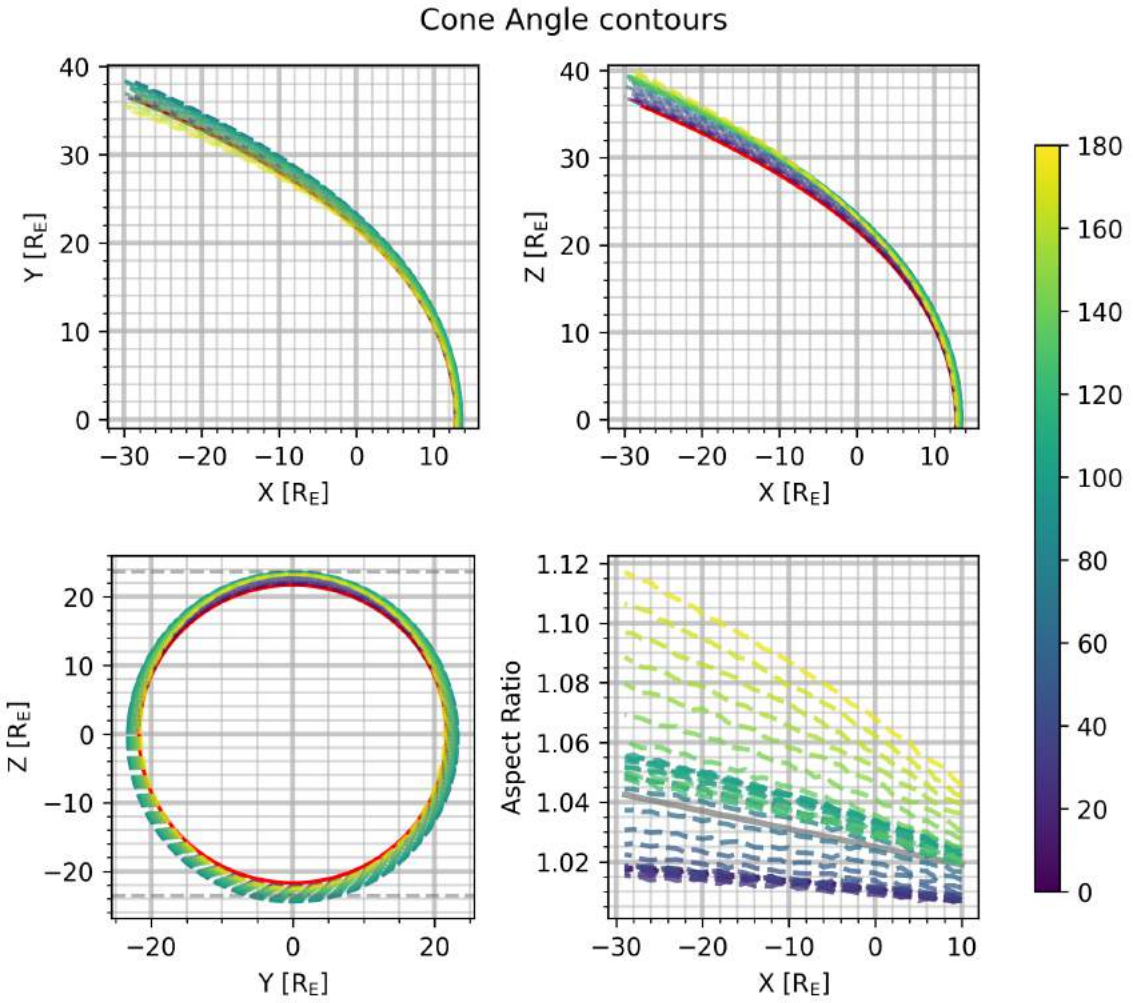

Figure 10-8: The plot structure is similar to that of 10-2 but now analyzes how our model behaves with respect to the cone angle θ_B. The Chao model is not parameterized for cone angle and so a single contour is shown using the average test set parameters. 30 different contours are plotted according to the colorbar at right ranging from 0 to 180 degrees. A θ_B between 0 and 72 degrees shows little change in X-Y, and increase in X-positive Z and X-negative Z, and aspect ratios below average. θ_B between 72 and 144 show slight increase in X-Y, a continuing upward shift of X-positive Z and X-negative Z, and comparable aspect ratios that are above the average. $\theta_B \geq 156$ degrees exhibit sharp contraction along X-Y, a further continuance of the upward shift in X-positive Z and X-negative Z and ever increasing aspect ratios. Note that the first and last values of cone angle are physically rare as they imply an IMF almost completely characterized by sunward / anti-sunward B_X. The full θ_B parameterization can be seen in the bottom row of Figure 9-8.

model outperforming the Chao model by about 25% less error.

We next analyze our model performance for clock and cone angles. Quasi-eastward and quasi-westward clock angles of the test set are shown in Figures 10-12 and 10-13, and both indicate our model to outperform in these respects. It should be noted after the analysis of the dayside region that both of these subsets include a number of dayside points and our model still manages to perform better despite this. The model performance on radial and non-radial cone angles is shown in Figures 10-14 and 10-15, showing that our model outperforms by about 1 point in loss in both regimes. And last, regarding the abnormally large bow shock contour predicted by our model for a M_{MS} value of 5 in Figure 10-5, we also show the model performance on test set points with $M_{MS} < 5$ in Figure 10-16. It shows that, even with this abnormally large R_0, our model shows a slightly lower loss in comparison to Chao.

10.3 Conclusions and Discussion

Machine learning approaches to regression have often been done by predicting a value outright wherein the network performs nonlinear transformations on the input to improve its predictions. Two unique contributions in our method are (1) the incorporation of the bow shock model function into our loss function and (2) the formation of an ensemble via bagging. These points could appear to run into conflict as coefficient prediction and loss calculations of the resulting radii would not necessarily imply that the ensemble could be well formed at the coefficient level, but we showed comparable error in doing so. This allows for visualization of the parameterization of each model in the ensemble as well as the ensemble-mean.

Our model has been shown to provide more accurate predictions than *Chao et al.* [2002] for nightside crossings and with respect to clock and cone angles. However, theirs slightly

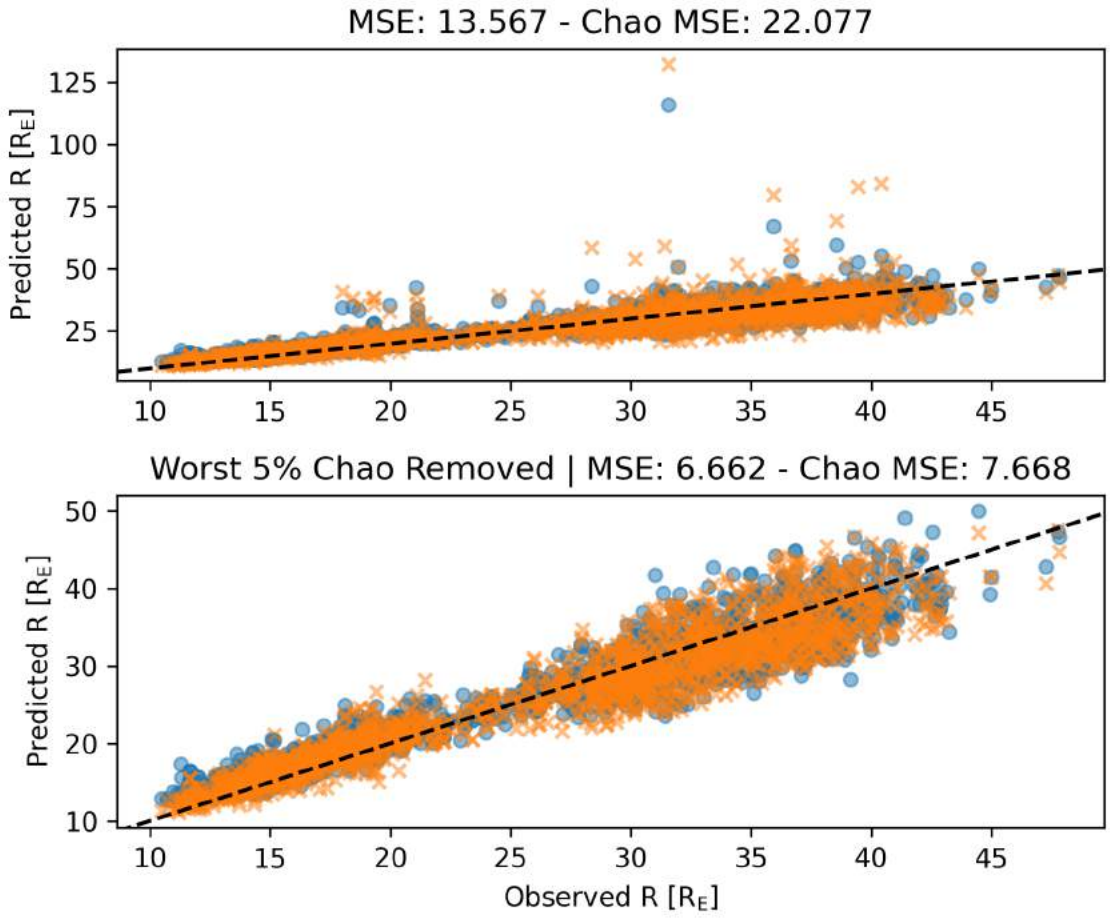

Figure 10-9: The residuals (that is, the difference between the observations and the model predictions) for both models on the original test set are shown in the top plot. The observations are plotted along the x axis, the model predictions along the y axis, and the line of perfect prediction (i.e. prediction = observed) is plotted along the diagonal as a dashed black line. To make a cleaner comparison without outliers, the residuals in which the poorest 5% of predictions of Chao are removed is shown in the bottom plot. Our model predictions are shown as a blue circle and Chao predictions are shown with an orange x. With 5% removed, the test set is reduced from 2,201 points to 2,090.

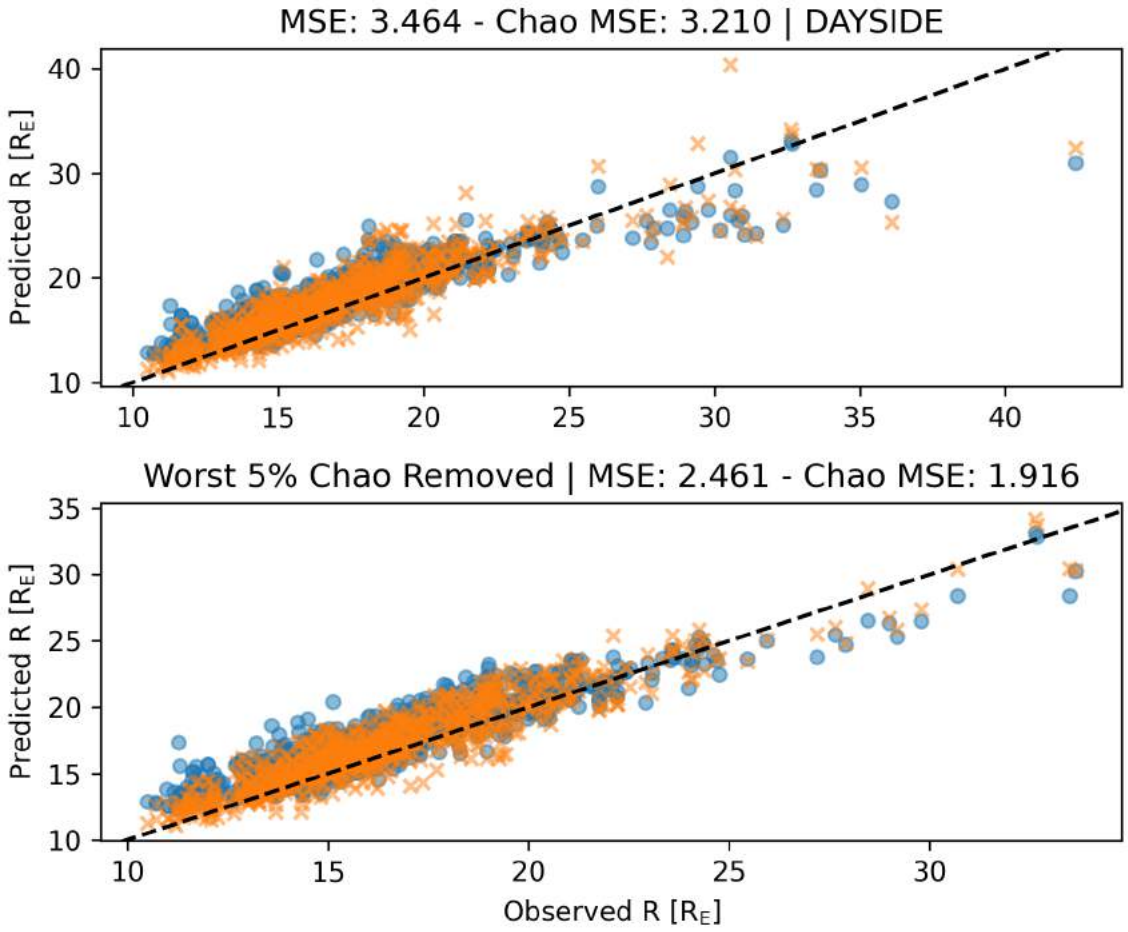

Figure 10-10: Like Figure 10-9, but only for dayside (X > 0) observations. The 5% removal decreases the test set size from 973 points to 924.

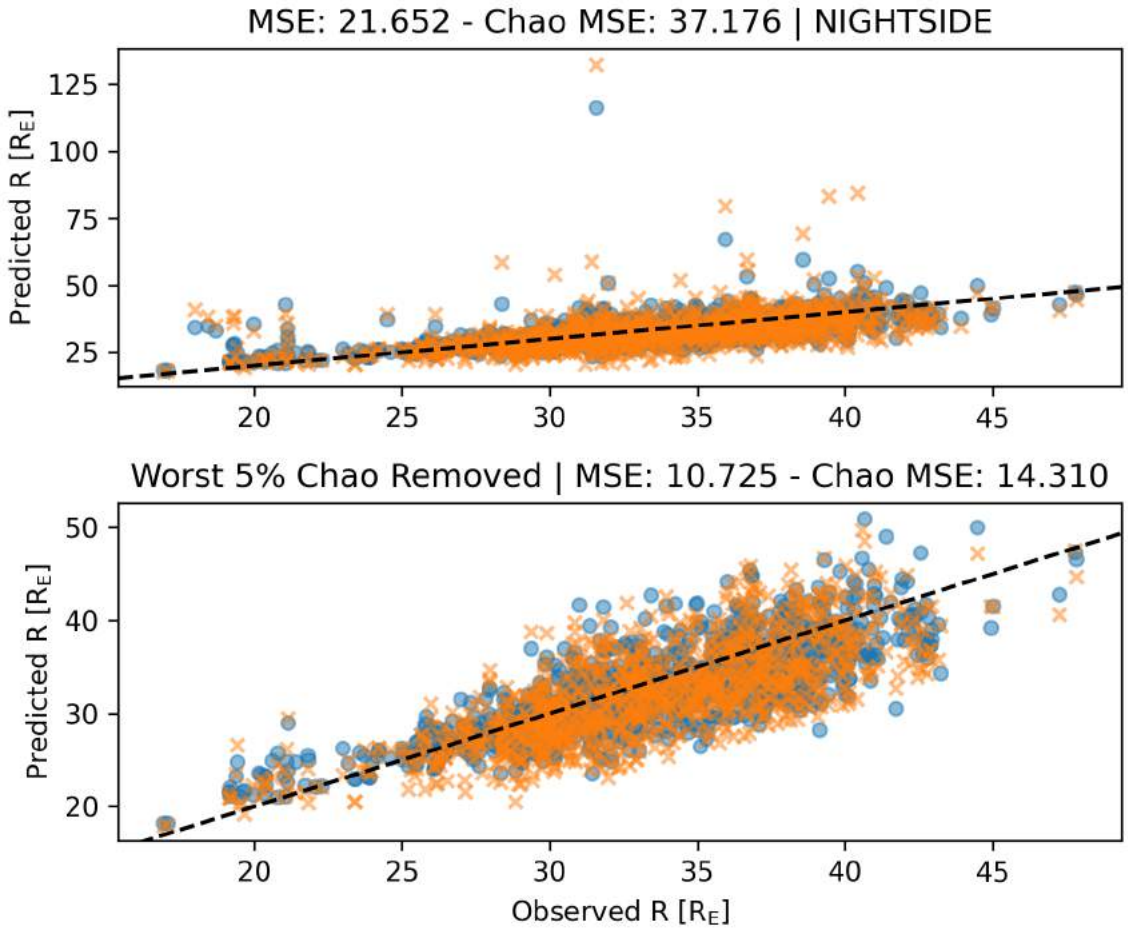

Figure 10-11: Like Figure 10-9, but only for nightside (X < 0) observations. The 5% removal decreases the test set size from 1,223 points to 1,161.

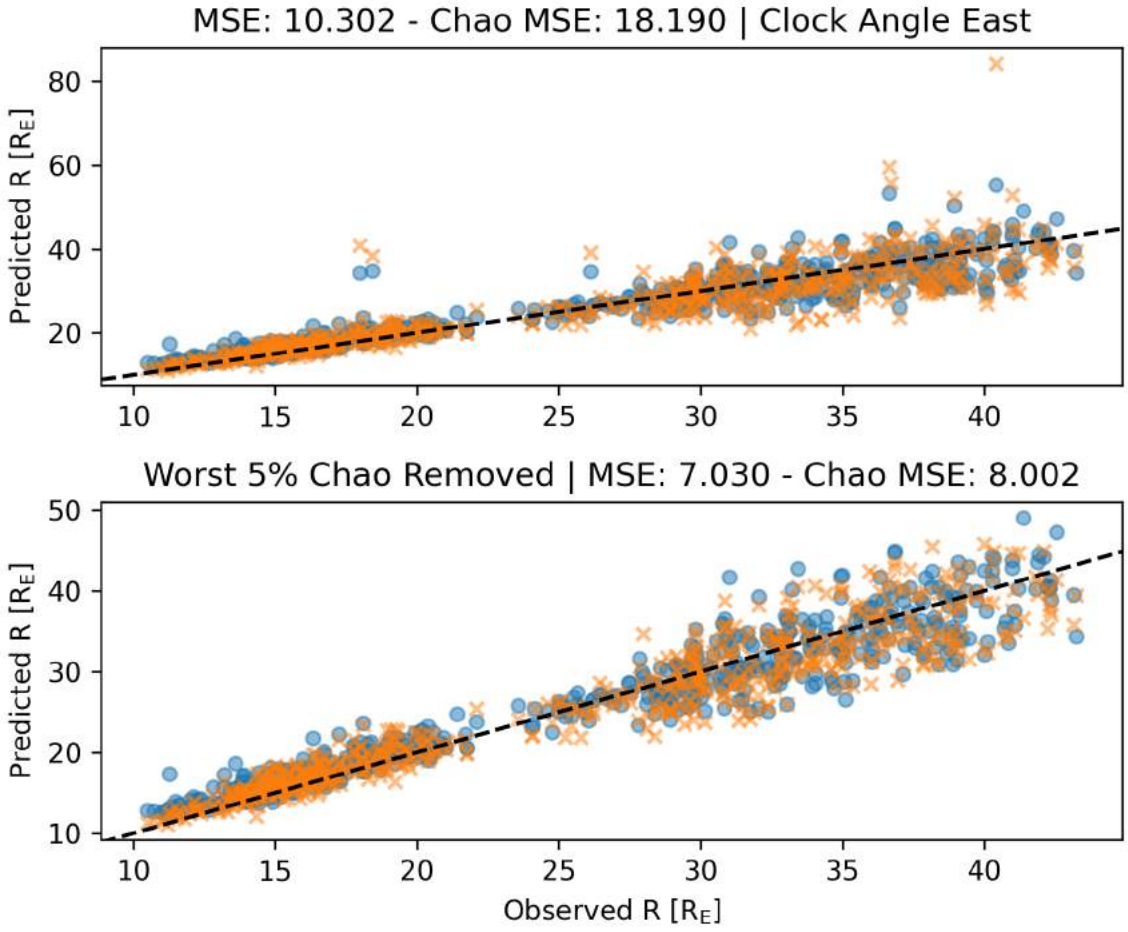

Figure 10-12: Like Figure 10-9, but only for quasi-eastward clock angles (i.e. $45° < \phi_B < 135°$ such that $\mathbf{B_{YZ}}$ largely points towards the positive Y axis). The 5% removal reduces the test set from 624 to 609 points.

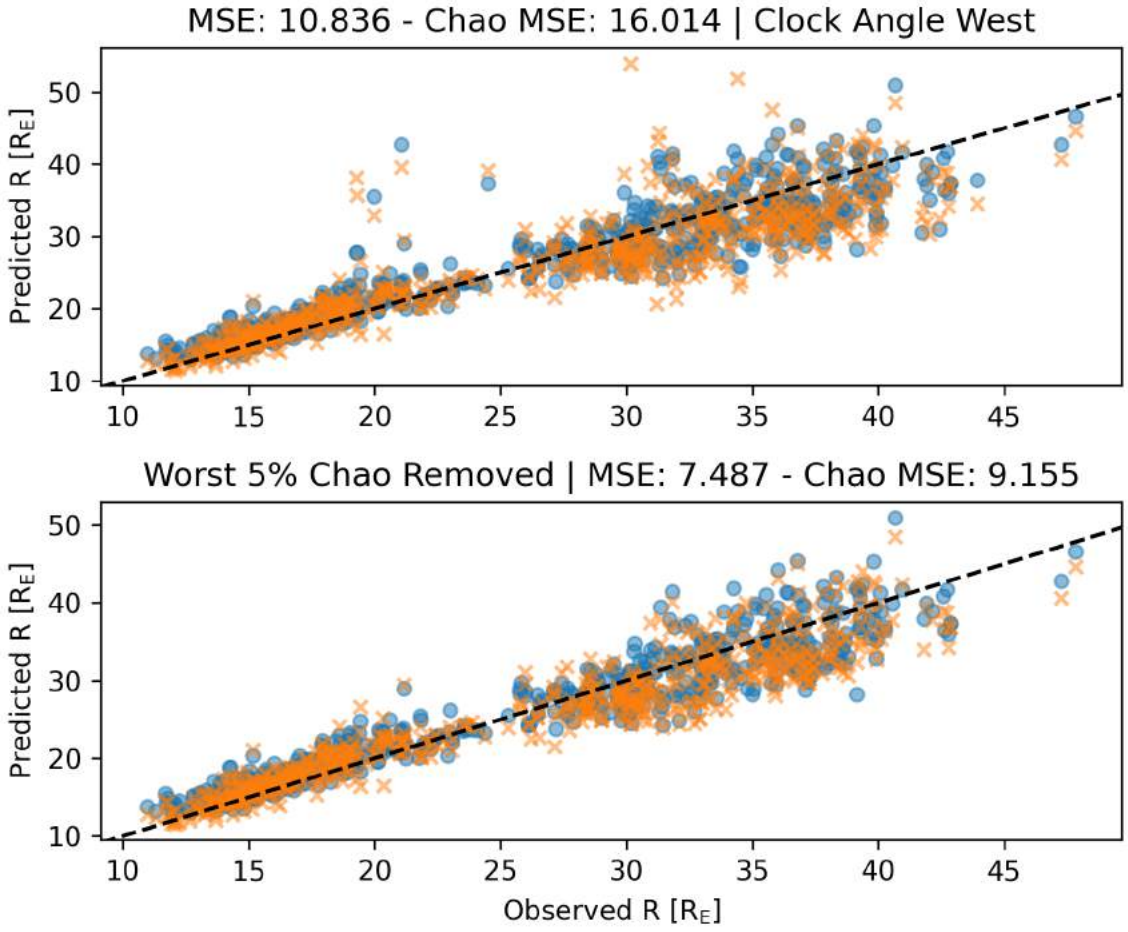

Figure 10-13: Like Figure 10-9, but only for quasi-westward clock angles (i.e. $-45° > \phi_B > -135°$ such that $\mathbf{B_{YZ}}$ largely points towards the negative Y axis). The 5% removal reduces the test set from 694 to 659 points.

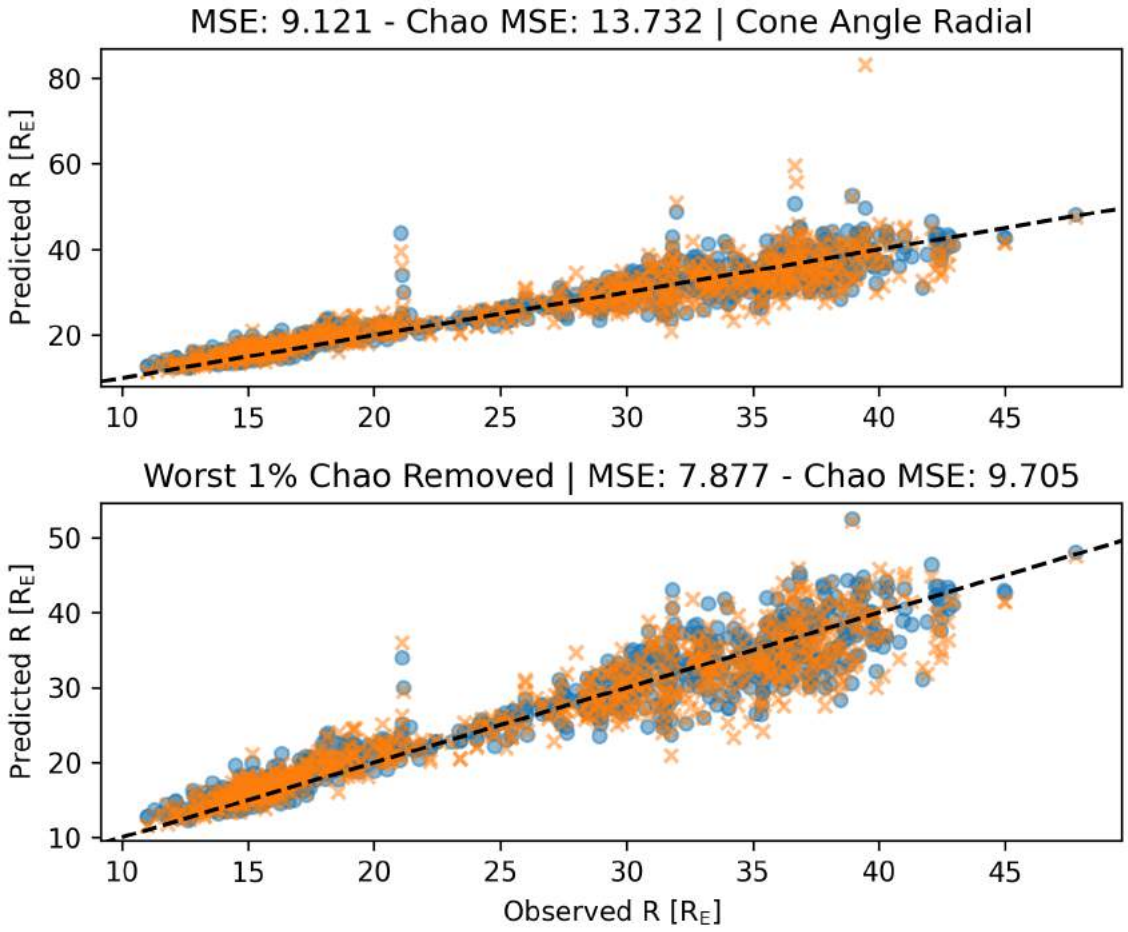

Figure 10-14: Like Figure 10-9, but only for radial cone angles (i.e. $\theta_B < 45°$ or $\theta_B > 135°$ such that it is B_X dominated). The 5% removal reduces the test set from 857 to 814 points.

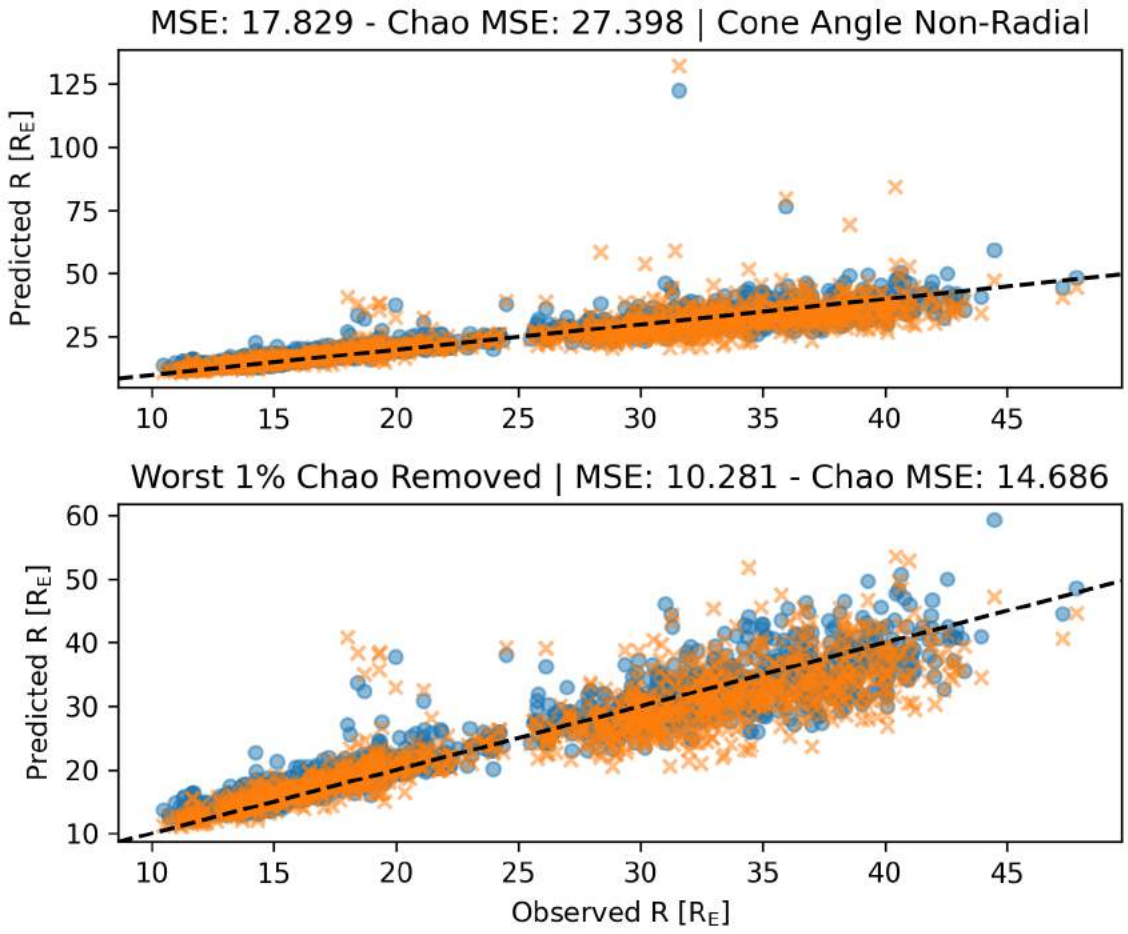

Figure 10-15: Like Figure 10-9, but only for non-radial cone angles (i.e. $45° < \theta_B < 135°$ such that B_{YZ} dominates such that it is B_X dominated). The 5% removal reduces the test set from 1,344 to 1,276 points.

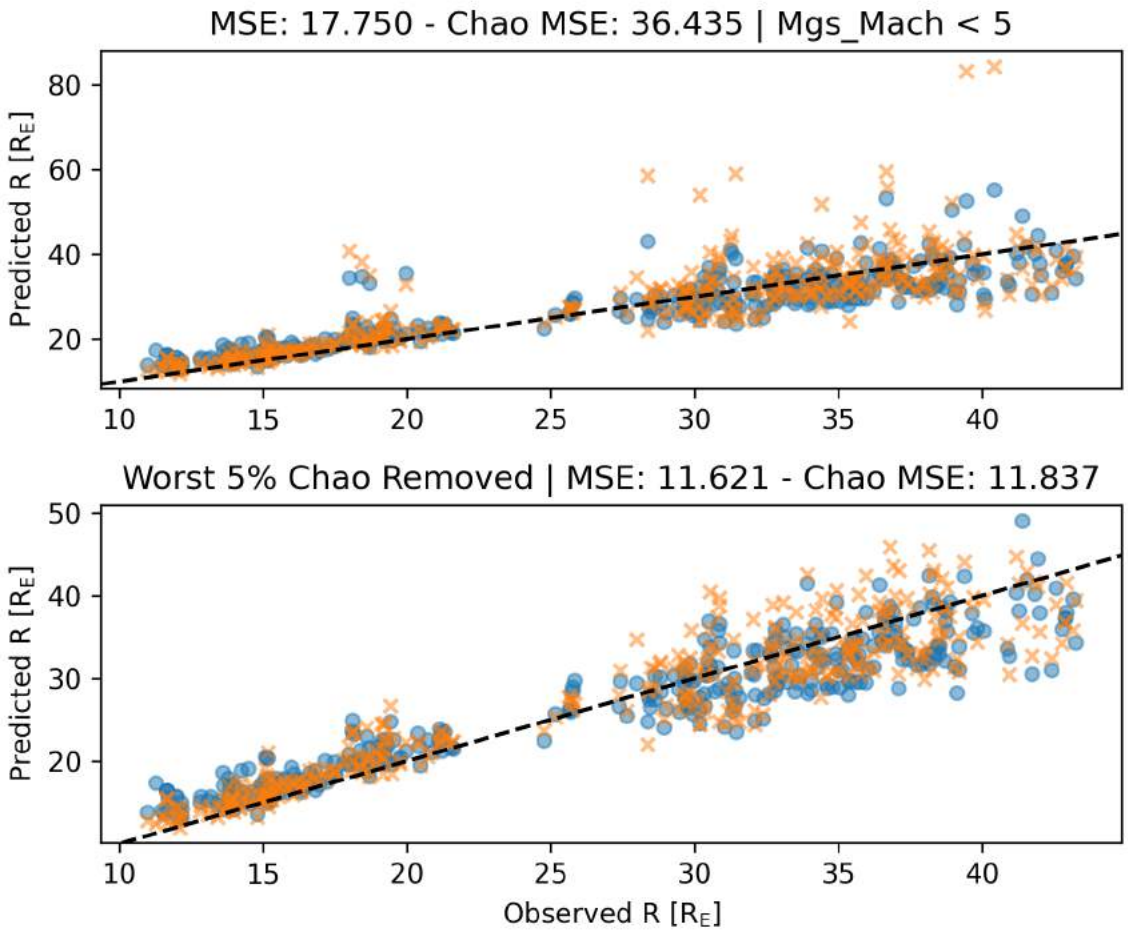

Figure 10-16: Like Figure 10-9, but only test data with $M_{MS} < 5$. The 5% removal reduces the test set from 390 to 370 points.

better predicts dayside crossings. This could be for a variety of reasons, possibly including the time range and spatial distributions of the dataset in relation to solar cycle. Many nightside crossings are taken from spacecraft with observations occurring from the 1970's to the 90's whereas the dayside crossings largely come from THEMIS, MMS, and Cluster, most of which were observed within the last 20 years.

Another caveat is the out-of-training or extrema predictions for the coefficients when comparing the training distributions of Figure 8-3 and the coefficient parameterizations in Figure 9-7. This is especially visible in the predictions for R_0 for $|B_Z| \geq 10$ nT, beta ≥ 10, low M_{MS}, which could have contributed to the overprediction in dayside crossings.

We have two ideas for improvement: One is inspired by the success of Physics-Informed Neural Networks (PINNs, *Raissi et al.* [2019]) and involves the further incorporation of coefficient parameterizations that are already widely known. For example, Equation 1.4 describes the theoretical gas dynamical relationship between R_0 and Mach number (as shown in *Spreiter et al.* [1966]) or its modification by *Farris and Russell* [1994] could be used. Another is to better represent the periodicity of the clock angle, which was not done. This could be implemented by either only considering the IMF Cartesian components alone (i.e. using the sin and cos components of the angle) or adding a custom penalty term to the loss function for a 2π shifted input (e.g a penalty $\propto$ —M(ψ) - M($\psi + 2\pi$)— for some angular input ψ to the model M).

Chapter 11

Summary and Future Work

11.1 Clustering Model Summary

We have taken magnetic field, ion velocity, ion density, and ion temperature measurements from THEMIS and MMS observations at different time resolutions in order to predict whether they occurred in the magnetosphere, magnetosheath, or the solar wind. The changing time step of THEMIS observations precludes the use of methods that need consistent timing and requires that we utilize tools that consider only the joint set of measurements. We have used a combination of unsupervised methods to cluster these data in a time independent way, and it works remarkably well with measurements from different spacecraft.

Seeking to express some of the nonlinear variances of the data linearly, we constructed additional features of magnetic field strength, ion speed, and the components of ion momentum density. Due to a range of orders of magnitude, features related to the ion density or temperature were converted to log10 scale, or the log10-absolute scale in the case of the ion momentum density. Partitioning the data into training and testing sets, these data were then min-max normalized based on the training set. PCA was used to yield a smaller set of uncorrelated features, both reducing the dimensionality and multicollinearity of the training data. Translating the loadings of a PCA transformation into a plot, we correctly predicted which regions of the PCA-transformed data would approximately correspond to certain an-

ticipated clusters: higher temperatures (magnetosphere), higher densities (magnetosheath), and higher speeds (solar wind).

To find a SOM that best represents outliers in the training data, we used KMeans to build a micro training set of 10k points from the training data and made the remainder of the training set a macro training set, or validation set. 500 maps were trained on this micro training set and the hyperparameters of the SOM that returned an optimal value on the macro training set were retained. A SOM with these hyperparameters was then trained on the macro training set. Using the nodes of this map as representative samples of our training set, the number of clustering options available for use was expanded such that even transductive clustering methods could be utilized. The bulk of SOM nodes mapped to the locations of expected clusters from the PCA transform along the 0^{th} and 1^{st} components with the remainder of nodes distributed between them.

Next, we used hierarchical agglomerative clustering to cluster the nodes and make predictions on test data by propagating node cluster assignments to the data that the nodes represent. We used this method with a Ward linkage, which focuses on building clusters based on minimization of intra-cluster variance. The solar wind and magnetosphere clusters were well separated from each other whereas the magnetosheath cluster had multiple nodes that overlapped into both the solar wind and magnetosphere clusters, which is not surprising as the magnetosheath acts as a transition region. We investigated the magnetosheath nodes that were surrounded by either solar wind or magnetosphere nodes to ascertain if there were possible misclassifications, but detailed analysis showed that it was indeed correct that these nodes were classified as magnetosheath. The data mapping to these nodes sometimes showed a mix of characteristics between magnetosheath and the surrounding cluster, reaffirming that these magnetosheath nodes appear adjacent to the nodes of other clusters. We also investigated where three magnetosphere-classified nodes were situated

within the magnetosheath cluster and inspection of these nodes' data revealed this classification to likely be incorrect. This data only represented 0.50% of the test set and did not have sufficient alignment across all distributions for magnetosphere, magnetosheath, or solar wind, so any classification of this data was done without strong confidence. Being a hierarchical method, we also demonstrated how subpopulation analysis is possible, an advantage not available to most other clustering methods. However, this does not mean that the uncovered subclusters will be as topologically "smooth" as parent clusters. We identified two magnetosphere-classified nodes in the map that indicate high VX-values, one of which we confirmed activated during an MMS 1 observation of a BBF. We also showed how our model can use gaps occurring in sequences of solar wind classifications to flag possible HFAs and FBs.

The validation of the model was done both by visual inspection of both time series and histograms as well as with comparison of the labeled dataset used in the training of a prexisting model (*Olshevsky et al.* [2021]). It shows comparable accuracy with respect to separating magnetosphere, magnetosheath, and solar wind at 99.4% but does not distinguish between pristine solar wind and ion foreshock as well as their model. On the whole, the model seems to be generally accurate but is capable of spurious and largely non-consecutive misclassifications.

Having separated data into magnetosphere, magnetosheath, or solar wind regions, we extract magnetopause and bow shock boundary crossings from the predictions on the full dataset. We accounted for both misclassifications and changing time resolution by using a 20 minute window for MMS and 40 minute window for THEMIS. We extracted 5,228 magnetopause crossings and 3,047 bow shock crossings. Analyzing the most recent solar wind points from the bow shock crossings in the SOM, we found that these points were distributed across most of the solar wind-classified SOM nodes with the two largest containing

almost 22% of the crossings. Performing the same analysis on the magnetosheath points in the context of magnetopause crossings, we found them less evenly distributed than what we saw for the solar wind nodes and noted that the three nodes with the highest number of counts for these points accounted for 18% of the crossings but only 3% of the magnetosheath predictions. In both cases, the nodes mapping the highest fraction of crossings can be useful in that data mapping to these nodes can be flagged as most likely to be related to a boundary crossing.

The dataset used for our model, the resulting crossings, and the MMS1 dataset that we joined with the labels of *Olshevsky et al.* [2021] can be found in a Zenodo repository at *Edmond et al.* [2024a]. The pickled models used can be found in a separate repository at *Edmond et al.* [2024b]. We have made a python package, GMClustering, that will easily make classifications and is pip-installable directly from its github repository at https://github.com/jae1018/GMClustering. It includes both an example python driver file and a small Jupyter notebook to showcase its use. Our modeling used various numerically-oriented python packages and we include the versions of those most relevant below.

- Numpy *Harris et al.* [2020] : 1.24.3
- Scikit-Learn *Pedregosa et al.* [2011] : 1.3.0
- XPySom *Mancini et al.* [2020] : 1.0.7
- Pandas *McKinney* [2010] : 2.0.3
- SciPy *Virtanen et al.* [2020] : 1.11.1

11.2 Bow Shock Model Summary

We have created a bagged ensemble of neural networks that predict the coefficients used for our model of the bow shock. The crossings used to train this model come from a variety of

sources: the THEMIS and MMS bow shock crossing time and positions are derived from the unsupervised classifier we created where upstream solar wind estimates provided by OMNI, the Cluster bow shock crossings are taken from the predictions of *Nguyen et al.* [2022] with upstream data also provided by OMNI, and with many of the nightside crossings taken from a pre-existing database of IMP 8, Geotail, and Magion-4 observations with upstream solar wind estimates either taken directly from the upstream measurements of the spacecraft (IMP 8) or ballistically-propagated from Wind data (Geotail and Magion-4). The resulting crossings are then filtered by removing the 1% extrema for B_Z, dynamic pressure, plasma beta, and M_{MS}, transformed from GSE coordinates to aberrated GSE by aberrating for both V_Y and V_Z, and split into train, validation, and test sets with proportions 70%, 15%, and 15%.

The input features for the model include common features of bow shock modeling such as B_Z, dynamic pressure, plasma beta, and magnetosonic Mach number. It is also the first model to incorporate both magnetic clock and cone angles. The bow shock model assumes a function to describe the radius as a function of θ, ϕ, and coefficients that govern the bow shock shape. The model takes these inputs features and predicts the coefficients for this function. This is advantageous to predicting the bow shock shape as these coefficients have known physical interpretation (i.e. R_0 being the subsolar point of the bow shock).

Since there are so few data points in the training set, splitting the training set into unique subsets without replacement and training individual models on them cannot reasonably produce good results. We address this by bootstrap-aggregating ("bagging") the trained networks on 300 bootstrap samples of the training set. For comparison, we also train a single model on the full training set. We define the ensemble prediction to correspond to the mean-coefficient prediction instead of the mean-radius and initially allow the ensemble to contain all 300 of these bootstrap-trained networks. The ensemble is pruned by ranking

each model according to its validation loss and removing the worst performing model. This process is repeated until only one model remains, and we find a close-to-optimal member size of 12 of these networks. The ensemble validation loss is shown to be lower than that of any single bootstrap-trained network or the model trained on the full training set. Moreover, we also find the the validation loss to be comparable when using either the mean-coefficient or mean-radius approach.

The parameterizations of the coefficients as functions of the inputs features are then analyzed in the context of the ensemble and in comparison to the model trained on the full training set in Figures 9-7 and 9-8. We find some similarities between the ensemble-mean parameterizations and known coefficient relationships, such as the decrease in R_0 with dynamic pressure and magnetosonic Mach number. Others deviate from expectation relative stability of α_0 with respect to B_Z or the gradual increase in R_0 for beta > 10. There is also mixed agreement with *Wang et al.* [2016] in that Y-Z cross sections are stretched perpendicular to the IMF direction for the east, north, and south clock angle cases, although westward does not produce a North-South stretching, and disagreement with *Wang et al.* [2018] in the behavior of R_0 and α with respect to the clock angle.

We compare our model to *Chao et al.* [2002] both on the original test set and a reduced test set in which 5% of the worst predictions (in favor of Chao) are removed for clearer comparison and find generally improved prediction with respect to clock and cone angles. It is also found to be more accurate in nightside bow shock crossing predictions, but slightly poorer on the dayside owing to overprediction of R_0.

11.3 Future Work

There is future work that can be done for both models. With regards to the classifier, we showed that we could potentially flag HFAs / FBs due to gaps in the solar wind classification,

finding agreement with some of the HFAs / FBs reported by *Liu et al.* [2022], and BBFs by checking for distinct node activations. Their could be a follow-up publication to show the full efficacy of such an approach or investigate if any other unique magnetospheric phenomena correlate with certain node activations. Additionally, the hierarchical organization of any of the classified nodes (with respect to magnetosphere, magnetosheath, or solar wind) could be investigated to see if sub-regions correspond to particular nodes.

As for the bow shock model, the results show an improvement over *Chao et al.* [2002] with respect to clock and cone angles but some mixed agreement with *Wang et al.* [2016] and disagreement with *Wang et al.* [2018]. An ensemble approach indisputably improved the performance overall but could also lead to abnormal coefficient prediction on out-of-training-bounds data. This approach could be improved by incorporating known relationships of some of the coefficients into the loss function as well as use of the Cartesian components of IMF as opposed to the clock angle.

Bibliography

Akiba, T., S. Sano, T. Yanase, T. Ohta, and M. Koyama (2019), Optuna: A next-generation hyperparameter optimization framework, in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Amaya, J., R. Dupuis, M. Innocenti, and G. Lapenta (2020), Visualizing and interpreting unsupervised solar wind classifications, *Frontiers in Astronomy and Space Sciences*, *7*, 553,207, doi:10.3389/fspas.2020.553207.

Angelopoulos, V. (2008), The themis mission, *Space Science Reviews*, *141*(1), 5, doi:10.1007/s11214-008-9336-1.

Angelopoulos, V. (2014), *The ARTEMIS Mission*, pp. 3–25, Springer New York, New York, NY, doi:10.1007/978-1-4614-9554-3_2.

Angelopoulos, V., C. F. Kennel, F. V. Coroniti, R. Pellat, M. G. Kivelson, R. J. Walker, C. T. Russell, W. Baumjohann, W. C. Feldman, and J. T. Gosling (1994), Statistical characteristics of bursty bulk flow events, *Journal of Geophysical Research: Space Physics*, *99*(A11), 21,257–21,280, doi:https://doi.org/10.1029/94JA01263.

Ansel, J., E. Yang, H. He, N. Gimelshein, A. Jain, M. Voznesensky, B. Bao, P. Bell, D. Berard, E. Burovski, G. Chauhan, A. Chourdia, W. Constable, A. Desmaison, Z. DeVito, E. Ellison, W. Feng, J. Gong, M. Gschwind, B. Hirsh, S. Huang, K. Kalambarkar, L. Kirsch, M. Lazos, M. Lezcano, Y. Liang, J. Liang, Y. Lu, C. Luk, B. Maher, Y. Pan, C. Puhrsch, M. Reso, M. Saroufim, M. Y. Siraichi, H. Suk, M. Suo, P. Tillet, E. Wang, X. Wang, W. Wen, S. Zhang, X. Zhao, K. Zhou, R. Zou, A. Mathews, G. Chanan, P. Wu, and S. Chintala (2024), Pytorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation, in *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*, ACM, doi:10.1145/3620665.3640366.

Arthur, D., and S. Vassilvitskii (2007), k-means++: the advantages of careful seeding, in *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, p. 1027–1035, Society for Industrial and Applied Mathematics, USA.

Auster, H. U., K. H. Glassmeier, W. Magnes, O. Aydogar, W. Baumjohann, D. Constantinescu, D. Fischer, K. H. Fornacon, E. Georgescu, P. Harvey, O. Hillenmaier, R. Kroth, M. Ludlam, Y. Narita, R. Nakamura, K. Okrafka, F. Plaschke, I. Richter, H. Schwarzl, B. Stoll, A. Valavanoglou, and M. Wiedemann (2008), The themis fluxgate magnetometer, *Space Science Reviews*, *141*(1), 235–264, doi:10.1007/s11214-008-9365-9.

Balogh, A., and R. A. Treumann (2013), *Physics of Collisionless Shocks, Space Plasma Shock Waves*, Springer.

Balogh, A., M. W. Dunlop, S. W. H. Cowley, D. J. Southwood, J. G. Thomlinson, K. H. Glassmeier, G. Musmann, H. LÜHR, S. Buchert, M. H. AcuÑA, D. H. Fairfield, J. A. Slavin, W. Riedler, K. Schwingenschuh, and M. G. Kivelson (1997), The cluster magnetic field investigation, *Space Science Reviews*, *79*(1), 65–91, doi:10.1023/A:1004970907748.

Bennett, L., M. G. Kivelson, K. K. Khurana, L. A. Frank, and W. R. Paterson (1997), A model of the earth's distant bow shock, *Journal of Geophysical Research: Space Physics*, *102*(A12), 26,927–26,941, doi:https://doi.org/10.1029/97JA01906.

Bieber, J., and E. Stone (1979), Energetic electron bursts in the magnetopause electron layer and in interplanetary space, *Magnetospheric Boundary Layers, edited by: Battrick, B., Mort, J., Haerendel, G., and Ortner, J*, *148*, 131–135.

Breiman, L. (1996), Bagging predictors, *Machine Learning*, *24*(2), 123–140, doi:10.1007/BF00058655.

Breiman, L. (2001), Random forests, *Machine Learning*, *45*, 5–32.

Breuillard, H., R. Dupuis, A. Retino, O. Le Contel, J. Amaya, and G. Lapenta (2020), Automatic classification of plasma regions in near-earth space with supervised machine learning: Application to magnetospheric multi scale 2016–2019 observations, *Frontiers in Astronomy and Space Sciences*, *7*, doi:10.3389/fspas.2020.00055.

Burch, J. L., T. E. Moore, R. B. Torbert, and B. L. Giles (2016), Magnetospheric multiscale overview and science objectives, *Space Science Reviews*, *199*(1), 5–21, doi:10.1007/s11214-015-0164-9.

C. T. Russell, R. J. S., J. G. Luhmann (2016), *Space Physics: An Introduction*, Cambridge University Press.

Cairns, I. H., and J. G. Lyon (1995), Mhd simulations of earth's bow shock at low mach numbers: Standoff distances, *Journal of Geophysical Research: Space Physics*, *100*(A9), 17,173–17,180, doi:https://doi.org/10.1029/95JA00993.

Cauchy, A.-L. (2009), *ANALYSE MATHÉMATIQUE. – Méthodc générale pour la résolution des systèmes d'équations simultanées*, p. 399–402, Cambridge Library Collection - Mathematics, Cambridge University Press.

Chao, J., D. Wu, C.-H. Lin, Y.-H. Yang, X. Wang, M. Kessel, S. Chen, and R. Lepping (2002), Models for the size and shape of the earth's magnetopause and bow shock, in *Space Weather Study Using Multipoint Techniques*, *COSPAR Colloquia Series*, vol. 12, edited by L.-H. Lyu, pp. 127–135, Pergamon, doi:https://doi.org/10.1016/S0964-2749(02)80212-8.

Chapman, J. F., and I. H. Cairns (2003), Three-dimensional modeling of earth's bow shock: Shock shape as a function of alfvén mach number, *Journal of Geophysical Research: Space Physics*, *108*(A5), doi:https://doi.org/10.1029/2002JA009569.

Chen, T., and C. Guestrin (2016), Xgboost: A scalable tree boosting system, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Cybenko, G. (1989), Approximation by superpositions of a sigmoidal function, *Mathematics of Control, Signals and Systems*, *2*(4), 303–314, doi:10.1007/BF02551274.

de Bodt, E., M. Cottrell, P. Letremy, and M. Verleysen (2004), On the use of self-organizing maps to accelerate vector quantization, *Neurocomputing*, *56*, 187–203, doi:https://doi.org/10.1016/j.neucom.2003.09.009.

Dmitriev, A. V., J. K. Chao, and D. J. Wu (2003), Comparative study of bow shock models using wind and geotail observations, *Journal of Geophysical Research: Space Physics*, *108*(A12), doi:https://doi.org/10.1029/2003JA010027.

Edmond, J., J. Raeder, B. Ferdousi, M. Argall, and M. E. Innocenti (2024a), Clustering of global magnetospheric observations, doi:10.5281/zenodo.10651397.

Edmond, J., J. Raeder, B. Ferdousi, M. Argall, and M. E. Innocenti (2024b), Clustering of global magnetospheric observations, `https://doi.org/10.5281/zenodo.10651702`.

Efron, B. (1979), Bootstrap methods: Another look at the jackknife, *Annals of Statistics*, *7*, 1–26.

Escoubet, C. P., M. Fehringer, and M. Goldstein (2001), Introduction the cluster mission, *Annales Geophysicae*, *19*(10/12), 1197–1200, doi:10.5194/angeo-19-1197-2001.

Fairfield, D. H. (1971), Average and unusual locations of the earth's magnetopause and bow shock, *Journal of Geophysical Research (1896-1977)*, *76*(28), 6700–6716, doi:https://doi.org/10.1029/JA076i028p06700.

Fairfield, D. H., H. C. Iver, M. D. Desch, A. Szabo, A. J. Lazarus, and M. R. Aellig (2001), The location of low mach number bow shocks at earth, *Journal of Geophysical Research: Space Physics*, *106*(A11), 25,361–25,376, doi:https://doi.org/10.1029/2000JA000252.

Farris, M. H., and C. T. Russell (1994), Determining the standoff distance of the bow shock: Mach number dependence and use of models, *Journal of Geophysical Research: Space Physics*, *99*(A9), 17,681–17,689, doi:https://doi.org/10.1029/94JA01020.

Fitzenreiter, R. (1995), The electron foreshock, *Advances in Space Research*, *15*(8), 9–27, doi:https://doi.org/10.1016/0273-1177(94)00081-B, proceedings of the D2.1 Symposium of COSPAR Scientific Commission D.

Formisano, V. (1979), Orientation and shape of the earth's bow shock in three dimensions, *Planetary and Space Science*, *27*(9), 1151–1161, doi:https://doi.org/10.1016/0032-0633(79)90135-1.

Funahashi, K.-I. (1989), On the approximate realization of continuous mappings by neural networks, *Neural Networks*, *2*(3), 183–192, doi:https://doi.org/10.1016/0893-6080(89)90003-8.

Galeev, A. A., and L. M. Zelenyi (1976), Tearing instability in plasma configurations, *Journal of Experimental and Theoretical Physics*, *43*, 1113.

Galton, F. (1907), Vox populi, *Nature*, *75*(1949), 450–451, doi:10.1038/075450a0.

Gencturk Akay, I., Z. Kaymaz, and D. G. Sibeck (2019), Magnetotail boundary crossings at lunar distances: Artemis observations, *Journal of Atmospheric and Solar-Terrestrial Physics*, *182*, 45–60, doi:https://doi.org/10.1016/j.jastp.2018.11.002.

Gray, R. (1984), Vector quantization, *IEEE ASSP Magazine*, *1*(2), 4–29, doi:10.1109/MASSP.1984.1162229.

Greenstadt, E. W., D. P. Traver, F. V. Coroniti, E. J. Smith, and J. A. Slavin (1990), Observations of the flank of earth's bow shock to 110 re by isee 3/ice, *Geophysical Research Letters*, *17*(6), 753–756, doi:https://doi.org/10.1029/GL017i006p00753.

Gurnett, D., and A. Bhattacharjee (2017), *Introduction to Plasma Physics, with Space, Laboratory, and Astrophysical Applications, 2nd Edition*, Cambridge University Press.

Harris, C. R., K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant (2020), Array programming with NumPy, *Nature*, *585*(7825), 357–362, doi:10.1038/s41586-020-2649-2.

Hebb, D. O. (1949), *The organization of behavior; a neuropsychological theory.*, xix, 335 pp., Wiley, Oxford, England.

Horn, F., R. Pack, and M. Rieger (2020), The autofeat python library for automated feature engineering and selection, in *Machine Learning and Knowledge Discovery in Databases*, edited by P. Cellier and K. Driessens, pp. 111–120, Springer International Publishing, Cham.

Hornik, K. (1991), Approximation capabilities of multilayer feedforward networks, *Neural Networks*, *4*(2), 251–257, doi:https://doi.org/10.1016/0893-6080(91)90009-T.

Hornik, K., M. Stinchcombe, and H. White (1989), Multilayer feedforward networks are universal approximators, *Neural Networks*, *2*(5), 359–366, doi:https://doi.org/10.1016/0893-6080(89)90020-8.

Jeřáb, M., Z. Němeček, J. Šafránková, K. Jelínek, and J. Měrka (2005), Improved bow shock model with dependence on the imf strength, *Planetary and Space Science*, *53*(1), 85–93, doi:https://doi.org/10.1016/j.pss.2004.09.032, dynamics of the Solar Wind - Magnetosphere Interaction.

Jolliffe, I. (2011), *Principal Component Analysis*, pp. 1094–1096, Springer Berlin Heidelberg, Berlin, Heidelberg, doi:10.1007/978-3-642-04898-2_455.

Kawano, H., and T. Higuchi (1995), The bootstrap method in space physics: Error estimation for the minimum variance analysis, *Geophysical Research Letters*, *22*(3), 307–310, doi:https://doi.org/10.1029/94GL02969.

Kennel, C. F. (1994), The magnetohydrodynamic Rankine-Hugoniot relations, *AIP Conference Proceedings*, *314*(1), 180–227, doi:10.1063/1.46750.

Kennel, C. F., J. P. Edmiston, and T. Hada (1985), *A Quarter Century of Collisionless Shock Research*, pp. 1–36, American Geophysical Union (AGU), doi:https://doi.org/10.1029/GM034p0001.

King, J. H. (1979), *Interplanetary Medium Data Book*, National Space Science Data Center, National Aeronautics and Space Administration, Goddard Space Flight Center.

King, J. H., and N. E. Papitashvili (2005), Solar wind spatial scales in and comparisons of hourly wind and ace plasma and magnetic field data, *Journal of Geophysical Research: Space Physics*, *110*(A2), doi:https://doi.org/10.1029/2004JA010649.

Kingma, D. P., and J. Ba (2014), Adam: A method for stochastic optimization, *CoRR*, *abs/1412.6980*.

Kohler, U., and M. Luniak (2005), Data inspection using biplots, *The Stata Journal*, *5*(2), 208–223, doi:10.1177/1536867X0500500206.

Kohonen, T. (1982), Self-organized formation of topologically correct feature maps, *Biological Cybernetics*, *43*(1), 59–69, doi:10.1007/BF00337288.

Kohonen, T. (2014), *MATLAB Implementations and Applications of the Self-Organizing Map*, Unigrafia Oy, Helsinki, Finland.

Landau, L., and E. Lifshitz (1987), *Fluid Mechanics*, *Course of Theoretical Physics*, vol. 6, 2nd ed., Pergamon Press.

LeCun, Y. A., L. Bottou, G. B. Orr, and K.-R. Müller (2012), *Efficient BackProp*, pp. 9–48, Springer Berlin Heidelberg, Berlin, Heidelberg, doi:10.1007/978-3-642-35289-8_3.

Liu, T. Z., H. Zhang, D. Turner, A. Vu, and V. Angelopoulos (2022), Statistical study of favorable foreshock ion properties for the formation of hot flow anomalies and foreshock bubbles, *Journal of Geophysical Research: Space Physics*, *127*(8), e2022JA030,273, doi:https://doi.org/10.1029/2022JA030273, e2022JA030273 2022JA030273.

Lloyd, S. (1982), Least squares quantization in pcm, *IEEE Transactions on Information Theory*, *28*(2), 129–137, doi:10.1109/TIT.1982.1056489.

Lu, J. Y., Y. Zhou, X. Ma, M. Wang, K. Kabin, and H. Z. Yuan (2019), Earth's bow shock: A new three-dimensional asymmetric model with dipole tilt effects, *Journal of Geophysical Research: Space Physics*, *124*(7), 5396–5407, doi:https://doi.org/10.1029/2018JA026144.

Mancini, R., A. Ritacco, G. Lanciano, and T. Cucinotta (2020), Xpysom: High-performance self-organizing maps, in *2020 IEEE 32nd International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD)*, pp. 209–216, doi:10.1109/SBAC-PAD49847.2020.00037.

Marquardt, D. W. (1963), An algorithm for least-squares estimation of nonlinear parameters, *Journal of the society for Industrial and Applied Mathematics*, *11*(2), 431–441.

Masson, A., and K. Nykyri (2018), Kelvin–helmholtz instability: Lessons learned and ways forward, *Space Science Reviews*, *214*(4), 71, doi:10.1007/s11214-018-0505-6.

McCulloch, W. S., and W. Pitts (1943), A logical calculus of the ideas immanent in nervous activity, *The bulletin of mathematical biophysics*, *5*(4), 115–133, doi:10.1007/BF02478259.

McFadden, J. P., C. W. Carlson, D. Larson, M. Ludlam, R. Abiad, B. Elliott, P. Turin, M. Marckwordt, and V. Angelopoulos (2008), The themis esa plasma instrument and in-flight calibration, *Space Science Reviews*, *141*(1), 277–302, doi:10.1007/s11214-008-9440-2.

McKinney, W. (2010), Data structures for statistical computing in python, in *Proceedings of the 9th Python in Science Conference*, edited by S. van der Walt and J. Millman, Proceedings of the Python in Science Conference, pp. 56–61, SciPy, doi: 10.25080/Majora-92bf1922-00a.

Merka, J., A. Szabo, T. W. Narock, J. H. King, K. I. Paularena, and J. D. Richardson (2003), A comparison of imp 8 observed bow shock positions with model predictions, *Journal of Geophysical Research: Space Physics*, *108*(A2), doi:https://doi.org/10.1029/2002JA009384.

Merka, J., A. Szabo, J. A. Slavin, and M. Peredo (2005), Three-dimensional position and shape of the bow shock and their variation with upstream mach numbers and interplanetary magnetic field orientation, *Journal of Geophysical Research: Space Physics*, *110*(A4), doi:https://doi.org/10.1029/2004JA010944.

Nariyuki, Y. (2022), Low-frequency alfvén waves and parametric instabilities in fluid and kinetic plasmas, *Reviews of Modern Plasma Physics*, *6*(1), 22, doi:10.1007/s41614-022-00085-1.

Nelder, J. A., and R. Mead (1965), A Simplex Method for Function Minimization, *The Computer Journal*, *7*(4), 308–313, doi:10.1093/comjnl/7.4.308.

Nguyen, G., N. Aunai, B. Michotte de Welle, A. Jeandet, B. Lavraud, and D. Fontaine (2022), Massive multi-mission statistical study and analytical modeling of the earth's magnetopause: 1. a gradient boosting based automatic detection of near-earth regions, *Journal of Geophysical Research: Space Physics*, *127*(1), e2021JA029,773, doi:https://doi.org/10.1029/2021JA029773, e2021JA029773 2021JA029773.

Nielsen, F. (2016), *Hierarchical Clustering*, pp. 195–211, Springer International Publishing, Cham, doi:10.1007/978-3-319-21903-5_8.

Nishida, A. (1994), The geotail mission, *Geophysical Research Letters*, *21*(25), 2871–2873, doi:https://doi.org/10.1029/94GL01223.

Němeček, Z., and J. Šafránková (1991), The earth's bow shock and magnetopause position as a result of the solar wind-magnetosphere interaction, *Journal of Atmospheric and Terrestrial Physics*, *53*(11), 1049–1054, doi:https://doi.org/10.1016/0021-9169(91)90051-8, the 7th International Scostep symposium on Solar-Terrestrial Physics.

Ogilvie, K., and M. Desch (1997), The wind spacecraft and its early scientific results, *Advances in Space Research*, *20*(4), 559–568, doi:https://doi.org/10.1016/S0273-1177(97)00439-0, results of the IASTP Program.

Oliveira, D. M. (2015), A study of interplanetary shock geoeffectiveness controlled by impact angles using simulations and observations, Ph.D. thesis, University of New Hampshire.

Olshevsky, V., Y. V. Khotyaintsev, A. Lalti, A. Divin, G. L. Delzanno, S. Anderzén, P. Herman, S. W. D. Chien, L. Avanov, A. P. Dimmock, and S. Markidis (2021), Automated classification of plasma regions using 3d particle energy distributions, *Journal of Geophysical Research: Space Physics*, *126*(10), e2021JA029,620, doi:https://doi.org/10.1029/2021JA029620, e2021JA029620 2021JA029620.

Omidi, N., and D. G. Sibeck (2007), Formation of hot flow anomalies and solitary shocks, *Journal of Geophysical Research: Space Physics*, *112*(A1), doi:https://doi.org/10.1029/2006JA011663.

Omidi, N., J. P. Eastwood, and D. G. Sibeck (2010), Foreshock bubbles and their global magnetospheric impacts, *Journal of Geophysical Research: Space Physics*, *115*(A6), doi:https://doi.org/10.1029/2009JA014828.

Omidi, N., S. H. Lee, D. G. Sibeck, D. L. Turner, T. Z. Liu, and V. Angelopoulos (2020), Formation and topology of foreshock bubbles, *Journal of Geophysical Research: Space Physics*, *125*(9), e2020JA028,058, doi:https://doi.org/10.1029/2020JA028058, e2020JA028058 2020JA028058.

O'Brien, C., B. M. Walsh, Y. Zou, S. Tasnim, H. Zhang, and D. G. Sibeck (2023), Prime: a probabilistic neural network approach to solar wind propagation from l1, *Frontiers in Astronomy and Space Sciences*, *10*, doi:10.3389/fspas.2023.1250779.

Parker, E. N. (1958), Dynamics of the Interplanetary Gas and Magnetic Fields., , *128*, 664, doi:10.1086/146579.

Parks, G. K. (2004), *Physics of Space Plasmas, 2nd Edition*, Westview Press.

Paularena, K. I., and J. H. King (1999), *NASA's IMP 8 Spacecraft*, pp. 145–154, Springer Netherlands, Dordrecht, doi:10.1007/978-94-011-4487-2_11.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. (2011), Scikit-learn: Machine learning in python, *Journal of machine learning research*, *12*(Oct), 2825–2830.

Peredo, M., J. A. Slavin, E. Mazur, and S. A. Curtis (1995), Three-dimensional position and shape of the bow shock and their variation with alfvénic, sonic and magnetosonic mach numbers and interplanetary magnetic field orientation, *Journal of Geophysical Research: Space Physics*, *100*(A5), 7907–7916, doi:https://doi.org/10.1029/94JA02545.

Pitkänen, T., G. S. Chong, M. Hamrin, A. Kullen, T. Karlsson, J.-S. Park, M. Nowada, S. T. Yao, A. W. Degeling, A. M. Tian, and Q. Q. Shi (2023), Statistical survey of magnetic forces associated with earthward bursty bulk flows measured by mms 2017–2021, *Journal of Geophysical Research: Space Physics*, *128*(5), e2022JA031,094, doi:https://doi.org/10.1029/2022JA031094, e2022JA031094 2022JA031094.

Pollock, C., T. Moore, A. Jacques, J. Burch, U. Gliese, Y. Saito, T. Omoto, L. Avanov, A. Barrie, V. Coffey, J. Dorelli, D. Gershman, B. Giles, T. Rosnack, C. Salo, S. Yokota, M. Adrian, C. Aoustin, C. Auletti, S. Aung, V. Bigio, N. Cao, M. Chandler, D. Chornay, K. Christian, G. Clark, G. Collinson, T. Corris, A. De Los Santos, R. Devlin, T. Diaz, T. Dickerson, C. Dickson, A. Diekmann, F. Diggs, C. Duncan, A. Figueroa-Vinas, C. Firman, M. Freeman, N. Galassi, K. Garcia, G. Goodhart, D. Guererro, J. Hageman, J. Hanley, E. Hemminger, M. Holland, M. Hutchins, T. James, W. Jones, S. Kreisler, J. Kujawski, V. Lavu, J. Lobell, E. LeCompte, A. Lukemire, E. MacDonald, A. Mariano, T. Mukai, K. Narayanan, Q. Nguyan, M. Onizuka, W. Paterson, S. Persyn, B. Piepgrass, F. Cheney, A. Rager, T. Raghuram, A. Ramil, L. Reichenthal, H. Rodriguez, J. Rouzaud, A. Rucker, M. Samara, J.-A. Sauvaud, D. Schuster, M. Shappirio, K. Shelton, D. Sher,

D. Smith, K. Smith, S. Smith, D. Steinfeld, R. Szymkiewicz, K. Tanimoto, J. Taylor, C. Tucker, K. Tull, A. Uhl, J. Vloet, P. Walpole, S. Weidner, D. White, G. Winkert, P.-S. Yeh, and M. Zeuch (2016), Fast plasma investigation for magnetospheric multiscale, *Space Science Reviews*, *199*(1), 331–406, doi:10.1007/s11214-016-0245-4.

Powell, M. J. D. (1964), An efficient method for finding the minimum of a function of several variables without calculating derivatives, *The Computer Journal*, *7*(2), 155–162, doi:10.1093/comjnl/7.2.155.

Priest, E. (2014), *Magnetohydrodynamics of the Sun*, Cambridge University Press.

Raissi, M., P. Perdikaris, and G. Karniadakis (2019), Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations, *Journal of Computational Physics*, *378*, 686–707, doi:https://doi.org/10.1016/j.jcp.2018.10.045.

Rème, H., J. M. Bosqued, J. A. Sauvaud, A. Cros, J. Dandouras, C. Aoustin, J. Bouyssou, T. Camus, J. Cuvilo, C. Martz, J. L. Médale, H. Perrier, D. Romefort, J. Rouzaud, C. d'Uston, E. Möbius, K. Crocker, M. Granoff, L. M. Kistler, M. Popecki, D. Hovestadt, B. Klecker, G. Paschmann, M. Scholer, C. W. Carlson, D. W. Curtis, R. P. Lin, J. P. McFadden, V. Formisano, E. Amata, M. B. Bavassano-Cattaneo, P. Baldetti, G. Belluci, R. Bruno, G. Chionchio, A. Di Lellis, E. G. Shelley, A. G. Ghielmetti, W. Lennartsson, A. Korth, H. Rosenbauer, R. Lundin, S. Olsen, G. K. Parks, M. McCarthy, and H. Balsiger (1997), *The Cluster Ion Spectrometry (CIS) Experiment*, pp. 303–350, Springer Netherlands, Dordrecht, doi:10.1007/978-94-011-5666-0_12.

Robbins, H., and S. Monro (1951), A Stochastic Approximation Method, *The Annals of Mathematical Statistics*, *22*(3), 400 – 407, doi:10.1214/aoms/1177729586.

Rosenblatt, F. (1958), The perceptron: a probabilistic model for information storage and organization in the brain., *Psychological review*, *65 6*, 386–408.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams (1986), Learning representations by back-propagating errors, *Nature*, *323*(6088), 533–536, doi:10.1038/323533a0.

Russell, C. T., B. J. Anderson, W. Baumjohann, K. R. Bromund, D. Dearborn, D. Fischer, G. Le, H. K. Leinweber, D. Leneman, W. Magnes, J. D. Means, M. B. Moldwin, R. Nakamura, D. Pierce, F. Plaschke, K. M. Rowe, J. A. Slavin, R. J. Strangeway, R. Torbert, C. Hagen, I. Jernej, A. Valavanoglou, and I. Richter (2016), The magnetospheric multiscale magnetometers, *Space Science Reviews*, *199*(1), 189–256, doi:10.1007/s11214-014-0057-3.

Safránková, J., Z. Němeek, and M. Borák (1999), Bow shock position: Observations and models.

Schwartz, S. J., C. P. Chaloner, P. J. Christiansen, A. J. Coates, D. S. Hall, A. D. Johnstone, M. P. Gough, A. J. Norris, R. P. Rijnbeek, D. J. Southwood, and L. J. C. Woolliscroft (1985), An active current sheet in the solar wind, *Nature*, *318*(6043), 269–271, doi:10.1038/318269a0.

Schölkopf, B., J. Platt, and T. Hofmann (2007), *Greedy Layer-Wise Training of Deep Networks*, pp. 153–160.

Shue, J.-H., P. Song, C. T. Russell, J. T. Steinberg, J. K. Chao, G. Zastenker, O. L. Vaisberg, S. Kokubun, H. J. Singer, T. R. Detman, and H. Kawano (1998), Magnetopause location under extreme solar wind conditions, *Journal of Geophysical Research: Space Physics*, *103*(A8), 17,691–17,700, doi:https://doi.org/10.1029/98JA01103.

Slavin, J. A., and R. E. Holzer (1981), Solar wind flow about the terrestrial planets 1. modeling bow shock position and shape, *Journal of Geophysical Research: Space Physics*, *86*(A13), 11,401–11,418, doi:https://doi.org/10.1029/JA086iA13p11401.

Sonnerup, B. U. , and L. J. Cahill Jr. (1967), Magnetopause structure and attitude from explorer 12 observations, *Journal of Geophysical Research (1896-1977)*, *72*(1), 171–183, doi:https://doi.org/10.1029/JZ072i001p00171.

Spreiter, J. R., A. L. Summers, and A. Y. Alksne (1966), Hydromagnetic flow around the magnetosphere, *Planetary and Space Science*, *14*(3), 223–253, doi:https://doi.org/10.1016/0032-0633(66)90124-3.

Tóth, G., I. V. Sokolov, T. I. Gombosi, D. R. Chesney, C. R. Clauer, D. L. De Zeeuw, K. C. Hansen, K. J. Kane, W. B. Manchester, R. C. Oehmke, K. G. Powell, A. J. Ridley, I. I. Roussev, Q. F. Stout, O. Volberg, R. A. Wolf, S. Sazykin, A. Chan, B. Yu, and J. Kóta (2005), Space weather modeling framework: A new tool for the space science community, *Journal of Geophysical Research: Space Physics*, *110*(A12), doi:https://doi.org/10.1029/2005JA011126.

Verigin, M., G. Kotova, A. Remizov, N. Shutte, K. Schwingenschuh, W. Riedler, T.-L. Zhang, H. Rosenbauer, K. Szego, M. Tatrallyay, and V. Styazhkin (1997), Studies of the martian bow shock response to the variation of the magnetosphere dimensions according to taus and magma measurements aboard the phobos 2 orbiter, *Advances in Space Research*, *20*(2), 155–158, doi:https://doi.org/10.1016/S0273-1177(97)00526-7, planetary Ionospheres and Magnetospheres.

Verigin, M., G. Kotova, A. Szabo, J. Slavin, T. Gombosi, K. Kabin, F. Shugaev, and A. Kalinchenko (2001), Wind observations of the terrestrial bow shock: 3-d shape and motion, *Earth, Planets and Space*, *53*(10), 1001–1009, doi:10.1186/BF03351697.

Vettigli, G. (2018), Minisom: minimalistic and numpy-based implementation of the self organizing map.

Virtanen, P., R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors (2020), SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python, *Nature Methods*, *17*, 261–272, doi:10.1038/s41592-019-0686-2.

Wang, M., J. Lu, H. Yuan, K. Kabin, Z.-Q. Liu, M. Zhao, and G. Li (2015), The dipole tilt angle dependence of the bow shock for southward imf: Mhd results, *Planetary and Space Science*, *106*, 99–107, doi:https://doi.org/10.1016/j.pss.2014.12.002.

Wang, M., J. Y. Lu, K. Kabin, H. Z. Yuan, X. Ma, Z.-Q. Liu, Y. F. Yang, J. S. Zhao, and G. Li (2016), The influence of imf clock angle on the cross section of the tail bow shock, *Journal of Geophysical Research: Space Physics*, *121*(11), 11,077–11,085, doi:https://doi.org/10.1002/2016JA022830.

Wang, M., J. Y. Lu, K. Kabin, H. Z. Yuan, Z.-Q. Liu, J. S. Zhao, and G. Li (2018), The influence of imf by on the bow shock: Observation result, *Journal of Geophysical Research: Space Physics*, *123*(3), 1915–1926, doi:https://doi.org/10.1002/2017JA024750.

Wilson III, L. B., A. L. Brosius, N. Gopalswamy, T. Nieves-Chinchilla, A. Szabo, K. Hurley, T. Phan, J. C. Kasper, N. Lugaz, I. G. Richardson, C. H. K. Chen, D. Verscharen, R. T. Wicks, and J. M. TenBarge (2021), A quarter century of wind spacecraft discoveries, *Reviews of Geophysics*, *59*(2), e2020RG000,714, doi:https://doi.org/10.1029/2020RG000714, e2020RG000714 2020RG000714.

Wittek, P., S. C. Gao, I. S. Lim, and L. Zhao (2017), somoclu: An efficient parallel library for self-organizing maps, *Journal of Statistical Software*, *78*(9), 1–21, doi:10.18637/jss.v078.i09.

Zelenyi, L., P. Triska, and A. Petrukovich (1997), Interball—dual probe and dual mission, *Advances in Space Research*, *20*(4), 549–557, doi:https://doi.org/10.1016/S0273-1177(97)00438-9, results of the IASTP Program.

Zhou, Z.-H. (2012), *Ensemble Methods: Foundations and Algorithms*, 1st ed., Chapman and Hall/CRC, doi:10.1201/b12207.