Applying Machine Learning to Heliophysics Problems to Broaden Space-Weather Understanding

Téo Bloch

A thesis presented for the degree of
Doctor of Philosophy

University of Reading
School of Mathematical, Physical and Computational Sciences
United Kingdom
August 2021

Declaration

I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Téo Bloch

Publications

Chapters 4, 5 and 6 are reproduced from the following publications:

1. Bloch, T., Watt, C. E. J., Owens, M. J., McInnes, L., & Macneil, A. R. Data-Driven Classification of Coronal Hole and Streamer Belt Solar Wind. Sol Phys 295, 41 (2020).

   https://doi.org/10.1007/s11207-020-01609-z.

2. Bloch, T., Watt, C. E. J., Owens, M. J., Thompson, R. L., & Agiwal, O. (2021). Constraining the location of the outer boundary of Earth's outer radiation belt. Earth and Space Science, 8, e2020EA001610.

   https://doi.org/10.1029/2020EA001610.

3. Bloch, T., Tigas, P., Watt, C. E. J., & Owens, M. J. Modelling of Flux at the Radiation Belt's Outer Boundary with Ensembles of Probabilistic Deep Neural Networks. Under review by AGU's Space Weather.

All work undertaken in these publications was carried out by Téo Bloch, with co-authors providing guidance and review.

Abstract

Understanding space-weather phenomena is a growing requisite given our day-to-day reliance upon space-based infrastructure. This entails identifying the causal factors of space-weather phenomena, quantifying the magnitude of response of space-weather events, and jointly using this information for forecasting. Machine learning (ML), as a set of mathematical and statistical tools, has been successfully used across many fields of research, demonstrating vast potential to improve our understanding of space-weather phenomena.

We apply unsupervised ML (dimension-reduction and clustering) to derive robust solar wind classifications – providing further insight into space-weather driving. Our unsupervised techniques are applied to a theoretically-motivated set of extant composition variables - which are non-evolving with solar wind propagation. We demonstrate that solar-wind-speed-based classifications lose latent information regarding solar source regions. Our dimension-reduction suggests a more informative latent-space to represent streamer-belt-origin solar wind.

Subsequently, we investigate the outer boundary of the outer radiation belt (OBORB). Modelling of the energetic-electrons in the outer radiation belt is crucial to the effective operation of many Earth-orbiting satellites, and the outer boundary conditions for such models are critical to accurate simulation. We applied simple ML models to a dataset of electron distribution functions, testing a range of potential boundary locations – yielding an empirical identification of the quiet-time boundary location. Next, we employed Bayesian neural networks to construct parameterised, probabilistic models providing synthetic nowcasts of the electron fluxes at the boundary. These models bridge the gap between the empirically identified OBORB location and the information required by modellers to construct the outer boundary conditions.

This work showcases how a broad spectrum of ML techniques can be applied to a variety of space-weather related problems. We present novel scientific results with significant implications for future studies into the solar wind and radiation belts, and ultimately, space-weather forecasting.

Acknowledgements

My foremost thanks go to my supervisors, Clare and Matt. Your guidance and support has shaped the researcher I like to think I've become. Clare, to me, you're the epitome of what a researcher should be. Genuinely interested in and curious about everything, and always knowing the right questions to ask. Your understanding of plasma physics has taught me so much and your insight was completely integral to interpreting so many of my results. I think your constant enthusiasm and excitement for space-physics is really inspirational, and was probably one of the biggest temptations for me to continue with research. Matt, trying to learn your preternatural ability with both scoping a project and knowing what figures to make really helped me understand what I wanted to achieve through research projects and kept me sane and safe from endless rabbit-holes. I think we all took a chance with the whole ML thing, and I really appreciate that you were both intrigued and patient enough to let me make it foundational to my PhD.

The group formerly known as ORB@R, Clare, Oliver, Sarah and Rhys, thank you for letting me pick the name. Even if we all did somewhat different things, our journal clubs and general conversations were so helpful to broadening my own understanding of space-physics. But more importantly to me, it felt like we had our own little community, and I would never have guessed it would have been such a boost to my morale on the slower and more frustrating weeks. Sarah, you especially made me feel like my pushing for ML was having a real impact, and that really helped as a contrast to the inertia that I could see elsewhere. You also helped/encouraged/forced me to really dig into the theory and get a proper understanding of the how and why of algorithms etc., which I might not have done as well otherwise, and my work was the better off for it.

Greenham Gang! Zak, Allan, Sarah, Omakshi, Mollie (was there anyone else? We should vote on it) it was great living with you guys. Zak, where would we be without all of our conversations about our favourite department? Just like the BBC, we all benefited from having your opinions to bring 'balance' to our conversations. Going to the cinema and sampling your baking were always highlights! Allan, mi familia, just like you said, FF seemingly has ended up dominating phase two of Greenham for me. It was great having a climbing buddy for so long! I want to say thanks for putting aside all the time that you did to help me talk through work and also for providing so many interesting kitchen conversations! You've definitely helped change a lot of my opinions for the better. Sarah, your unique perspectives and understanding of people were always really cool to chat about. I loved that you would understand those weird and indescribable thoughts/sensations that we both knew, even if Allan and Zak would just be looking at us like we were crazy! I wanted to get some reference to the 4th dimension here, but that seems to have been hoping for too much... Omakshi, you get your own paragraph below, so don't be too offended that I'm skipping you here. Mollie, even if it wasn't for that long, you made for a fun housemate. Between you and Allan I think you defined most of the Greenham memes! Memes that I'm still just-about-ironically using and can't shake (damn you!).

To all of my friends, thank you guys for all the hanging out! I can't imagine the last four years without all of you and your support.

Omakshi. Where do I even begin? You shaped so much of me getting to this point. From getting through uni with your revision help, to the convoluted process of my PhD offer which you were integral to. I am way more grateful than I can really express with these words. This journey would have been a lot more of a mess without all your love and support! I've also really liked and thoroughly valued that we've been able to teach and help each other throughout. I know that I, at least, have massively benefited from how smart and insightful you are, and that extends well beyond just science.

Finally, I have my Grandma and Jaygo to thank. Grandma, thank you for always being interested in my work and forcing me to figure out the answer to "but what's the point?" haha. Jaygo, you didn't do anything, but stay cool bro, get that Michelin Star! Ultimately, I have my parents, Manni and Stu, to thank. You guys always made me feel like curiosity was a good thing, and helped give me the self-confidence and self-reliance that has been crucial to getting to where I am now.

Contents

Chapter 1

Introduction

This thesis explores how machine learning can be applied in the heliospheric domain to better understand and characterise phenomena related to Space Weather.

This chapter serves to set the scene for the thesis, covering the background physics underpinning the research that has been undertaken. We explore the dynamics of charged particles and space-plasmas, and then move to the wider context for how these physics fit into the Sun-Earth system. Chapter 2 will provide an overview of machine learning as well as expounding the specific areas which are most pertinent to the work subsequently presented. Chapters 4 - 6 present the scientific analyses undertaken for this thesis. Chapters 4 and 5 are published works, and Chapter 6 is a submitted manuscript (*verbatim*) completed over the course of this PhD. Finally, in Chapter 7 we will summarise, conclude and discuss the future directions that work based on this thesis might take.

1.1 Basic Plasma Physics

Plasma - ionised gas - is one of the four fundamental states of matter [Langmuir, 1928]. It is the most abundant type of discovered matter in the universe [versus dark matter and energy, Chu and Lu, 2013]. Plasma is not commonly encountered in day-to-day life on Earth, since it typically requires extremely high temperatures or intense radiation to ionise enough of a gas for it to be classed as a plasma - though phenomena such as lightning can generate it in small, short-lived quantities.

An ideal plasma is typically described by three criteria:

- Quasi-neutrality: the typical length scales, L, of the plasma (as a whole) must be large enough, that any local deviations in charge density can be effectively shielded from the rest of the plasma. This is codified with a useful, local plasma length scale λ_D, known as the Debye length, given by:

$$\lambda_D = \sqrt{\frac{\epsilon_0 k_B T_e}{n_e q^2}} \tag{1.1}$$

where ϵ_0 is the permittivity of free space, k_B is the Boltzmann constant, and T_e, n_e and q are the temperature, density and elemental charge of electrons. Hence, quasi-neutrality is maintained while $L \gg \lambda_D$.

- Collective behaviour: quasi-neutrality is theoretically maintained when $L \gg \lambda_D$, though it may not be practically maintained if there are not enough charged particles to effectively shield a local variation. Thus, to ensure that collective shielding occurs, the number of charged particles in a sphere of radius λ_D must be much larger than 1.
- Collisionless: despite a plasma being able to co-exist with neutral atoms or molecules, the neutral particles are required to be sparse enough that their collisions with charged particles do not affect the bulk ability of the plasma to collectively shield. In general an ideal plasma contains only charged particles and is collisionless.

Given that plasma comprises charged particles (ions and electrons), the dynamics of and within a plasma are primarily controlled by electric and magnetic fields - $\underline{\mathbf{E}}$ and $\underline{\mathbf{B}}$, respectively. The properties and interactions of these fields are described through Maxwell's equations:

$$\nabla \cdot \underline{\mathbf{E}} = \frac{\rho_q}{\epsilon_0} \tag{1.2}$$

$$\nabla \cdot \underline{\mathbf{B}} = 0 \tag{1.3}$$

$$\nabla \times \underline{\mathbf{E}} = -\frac{\partial \underline{\mathbf{B}}}{\partial t} \tag{1.4}$$

$$\nabla \times \underline{\mathbf{B}} = \mu_0 \left(\underline{\mathbf{j}} + \epsilon_0 \frac{\partial \underline{\mathbf{E}}}{\partial t} \right) \tag{1.5}$$

where ρ_q is the charge density, μ_0 is the permeability of free space, $\underline{\mathbf{j}}$ is the current density and t is time. Equation 1.2 is known as Gauss's law, and is interpreted as specifying that the net flux of an electric field passing through a closed surface is proportional to the total charge within the surface. Similarly, Equation 1.3 (Gauss's law for magnetism) specifies that there is no net magnetic flux through a closed surface (and equivalently that there are no magnetic monopoles). Equation 1.4, the Maxwell-Faraday equation, expresses the generation of an electric field due to a time-varying magnetic field. Finally, Equation 1.5, Ampère's circuital law, describes the ways in which a magnetic field can be generated - either by simple electric current ($\mu_0 \underline{\mathbf{j}}$) or time-varying electric field ($\mu_0 \epsilon_0 \frac{\partial \underline{\mathbf{E}}}{\partial t}$).

1.2 Particle Motion

In general, the force, $\underline{\mathbf{F}}$, exerted on a charged particle by electric and magnetic fields is described by the Lorentz equation:

$$\underline{\mathbf{F}} = q(\underline{\mathbf{E}} + \underline{\mathbf{v}} \times \underline{\mathbf{B}}) \tag{1.6}$$

where $\underline{\mathbf{v}}$ is the velocity of the charged particle (in the reference frame of the magnetic field). From this equation, we can determine the direction of the force originating from the electric and magnetic components. The force due to the electric field is a scalar multiple of the electric field, $\underline{\mathbf{E}}$, itself. Hence, the force is directed parallel to the electric field. In contrast, the force due to the magnetic field is the cross-product between the particle velocity and the magnetic field, and so the force acts perpendicularly to both $\underline{\mathbf{v}}$ and $\underline{\mathbf{B}}$.

1.2.1 Gyro-motion

In a uniform magnetic field in the Z-direction ($\underline{\mathbf{B}} = (0, 0, B_z) = const.$) absent of an electric field, one may simplify Equation 1.6 into the Lorentz force:

$$\underline{\mathbf{F}} = q(\underline{\mathbf{v}} \times \underline{\mathbf{B}}) \tag{1.7}$$

The three components of the Lorentz force (substituting $\underline{\mathbf{F}} = m\frac{d\underline{\mathbf{v}}}{dt}$, where m is mass) obtained by taking the cross product of the velocity with a magnetic field are:

$$m\frac{dv_x}{dt} = qv_yB_z; \qquad m\frac{dv_y}{dt} = -qv_xB_z; \qquad m\frac{dv_z}{dt} = 0,$$

Thus, the resulting force due to the magnetic field acts perpendicularly to both the velocity and the magnetic field. If we differentiate both non-zero forces with respect to time, we get:

$$\frac{d^2v_x}{dt^2} = \left(\frac{qBz}{m}\right)\frac{dv_y}{dt}; \qquad \frac{d^2v_y}{dt^2} = -\left(\frac{qBz}{m}\right)\frac{dv_x}{dt}$$

We can then substitute the single time derivatives to get:

$$\frac{d^2v_x}{dt^2} = -\left(\frac{qBz}{m}\right)^2 v_x; \qquad \frac{d^2v_y}{dt^2} = -\left(\frac{qBz}{m}\right)^2 v_y$$

Which are typical equations of motion for a simple harmonic oscillator. Knowing this, we can infer without further derivation that the angular frequency of the oscillation is:

$$\omega_c = \frac{qB}{m} \tag{1.8}$$

Solving these equations:

$$x = \frac{v_\perp}{\omega_c}\sin(\omega_c t) + x_0; \qquad y = \frac{v_\perp}{\omega_c}\cos(\omega_c t) + y0$$

we see that the gyration of charged particles in a uniform magnetic field is circular. Additionally, since v_z remains unchanged, the complete motion of the particle when $v_z \neq 0$ is a helix. Note that depending on the charge of a particle, the circular motion will be in different directions - electrons travelling parallel to the field gyrate right-handedly, whilst ions gyrate left-handedly (such particle motion is foundational to the radiation belts, for example).

It is useful to consider the particle velocity by its parallel and perpendicular components ($v_\parallel$ and $v_\perp$, respectively), and define the concept of pitch-angle, α:

$$\alpha = \arctan\left(\frac{v_\perp}{v_\parallel}\right) \tag{1.9}$$

which gives the angle between the velocity of a particle and the magnetic field, as shown in Figure 1.1. A pitch-angle of $90°$ corresponds to particle motion entirely perpendicular to the magnetic field, and $0°$ or $180°$ represents field-aligned motion.

It can be shown that the kinetic energy is conserved, despite the centripetal

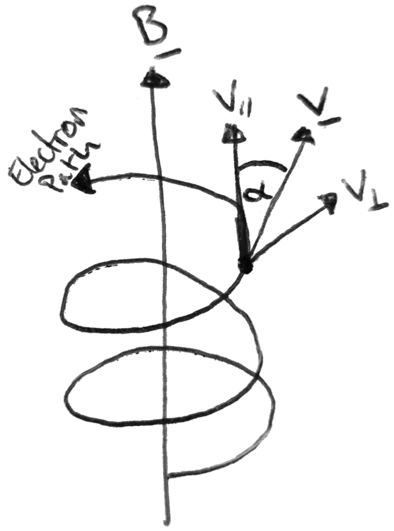


Figure 1.1: *A schematic diagram of an electron gyrating about a magnetic field line. Directions are signified with arrows. The electron's velocity is presented and decomposed into parallel and perpendicular components. The pitch angle is indicated.*

acceleration, by taking the dot product of the velocity with-respect-to both sides of Equation 1.7 (substituting $\underline{\mathbf{F}} = m\frac{d\underline{\mathbf{v}}}{dt}$), giving:

$$m\underline{\mathbf{v}} \cdot \frac{d\underline{\mathbf{v}}}{dt} = q\underline{\mathbf{v}} \cdot (\underline{\mathbf{v}} \times \underline{\mathbf{B}}) \tag{1.10}$$

This in turn can be simplified with the knowledge that $\underline{\mathbf{v}} \cdot (\underline{\mathbf{v}} \times \underline{\mathbf{B}}) = 0$ (since a dot product between orthogonal vectors is 0 and we know that $\underline{\mathbf{v}} \times \underline{\mathbf{B}}$ is orthogonal to $\underline{\mathbf{v}}$, because of the cross product) to:

$$m\underline{\mathbf{v}} \cdot \frac{d\underline{\mathbf{v}}}{dt} = \frac{d}{dt}\left(\frac{1}{2}m\underline{\mathbf{v}}^2\right) = 0 \tag{1.11}$$

hence, there is no change in the kinetic energy. This highlights that magnetic fields do not do work on charged particles, despite forces acting between them.

Due to the circular motion traced out by particles and the conservation of kinetic energy, we deduce that the centripetal motion is produced by the Lorentz force:

$$\frac{mv_{\perp}^2}{r} = |q|v_{\perp}B \tag{1.12}$$

where $v_\perp$ and B are now the magnitudes of the velocity and magnetic field strength. This allows us to define another characteristic length scale for a particle's motion - the gyro-radius:

$$r_{gyro} = \frac{mv_\perp}{|q|B} \tag{1.13}$$

where we take the absolute value of the charge to ensure that the radius is physical. Unsurprisingly, the radius is directly proportional the perpendicular speed. Though, we additionally see that the stronger the field is, the smaller the gyro-radius is.

We can also derive from Equation 1.13 an associated gyro-period:

$$T_{gyro} = 2\pi \frac{m}{|q|B} \tag{1.14}$$

which is independent of the energy, instead only depending on the magnetic field strength and particle charge. Though, once energies become relativistic, and the mass exceeds the rest mass, the gyro-period becomes energy-dependent.

1.2.2 Bounce Motion

If the magnetic field experienced by a particle is approximately static (i.e., the field does not change on length- or time-scales comparable to the gyro-radius or gyro-frequency, respectively) it implies - through Faraday's law - that the magnetic flux, Φ, through the surface, S, traced by a particle's gyro-orbit is constant:

$$\Phi = \iint_S \underline{\mathbf{B}} \cdot d\underline{\mathbf{s}} = B\pi r_{gyro}^2 = constant \tag{1.15}$$

On top of this, the magnetic moment, μ:

$$\mu = \frac{mv_\perp^2}{2B} \tag{1.16}$$

must also be conserved. μ is also referred to as the first adiabatic invariant.

When there are spatial gradients in the magnetic field, such that the field converges along the parallel direction, magnetic mirroring occurs. As a particle 'slowly' experiences the convergence of the magnetic field, the perpendicular velocity must increase to conserve μ (its gyro-radius must also become smaller to conserve Φ). However, as we have shown in Equation 1.11, the total kinetic energy of the particle is constant. Therefore, we deduce that the energy must be converted from the parallel kinetic energy (as the magnetic field does no work). This means that the parallel velocity must decrease - eventually reaching zero at the mirror point, and then reversing to travel anti-parallel along the field.

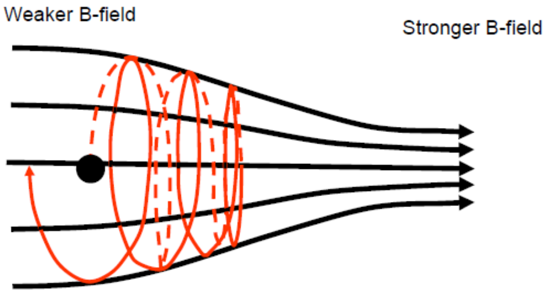


Figure 1.2: *A schematic diagram of magnetic mirroring. The dashed line represents the motion towards the stronger field, and the solid line represents the motion subsequent to being mirrored. Original image taken from lecture notes for Space Plasma and Magnetospheric Physics, UCL, I. J. Rae and C. J. Owen, 2017.*

The magnetic field strength at the mirror point, B_M, is given by:

$$B_M = \frac{B_0}{\sin^2(\alpha)} \tag{1.17}$$

where B_0 is the initial field strength and α is the pitch-angle. The minimum pitch-angle for which mirroring can occur (in ideal conditions) is given by:

$$\alpha_{min} = \arcsin\left(\sqrt{\frac{B_0}{B_{max}}}\right) \tag{1.18}$$

Where Equation 1.17 has been rearranged for α, and B_M has been substituted for the maximum field strength (giving us the lower bound).

In addition to the first adiabatic invariant associated with gyro-motion, there is an adiabatic invariant associated with the bounce-motion:

$$\underline{\mathsf{J}}_{bounce} = m \oint v_{\|} dl \tag{1.19}$$

which can be interpreted as the total parallel momentum over the full bounce motion being conserved (provided that the field does not vary on time-scales smaller than the bounce period). More comprehensive treatment of the adiabatic invariants is provided by Roederer and Zhang [2014] and Ukhorskiy and Sitnov

[2013], and relativistically in Öztürk [2012].

1.2.3 Drift-motion

In general, when a force acts upon a charged particle in a magnetic field, the resultant velocity due to that force is given by:

$$\underline{\mathbf{V}}_{drift} = \frac{\underline{\mathbf{F}} \times \underline{\mathbf{B}}}{qB^2} \tag{1.20}$$

which signifies a resultant perpendicular motion. Evidently, it is only the perpendicular (to $\underline{\mathbf{B}}$) components of the force which yield a drift. Figure 1.3 represents some of the drifts we will discuss below.

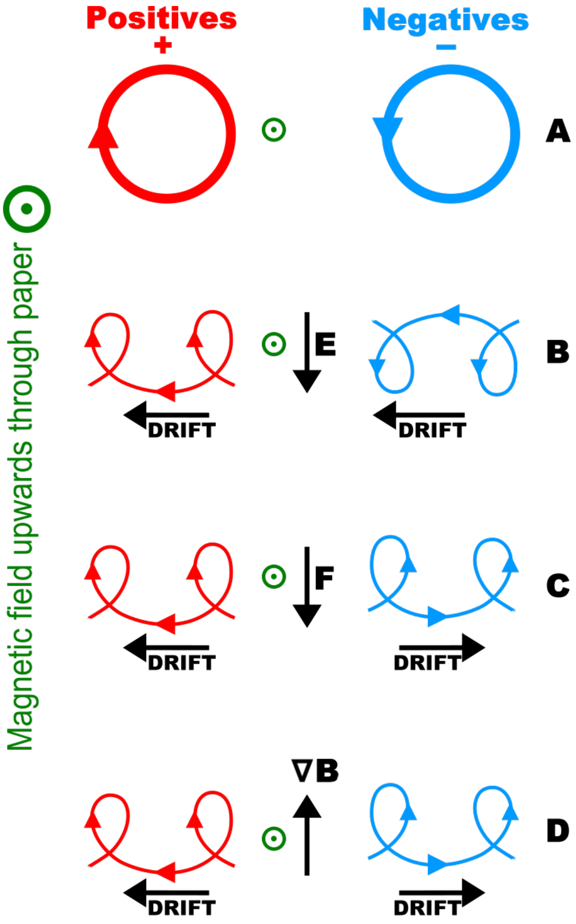


Figure 1.3: *A schematic diagram of the direction of charged particle motion with respect to a simple uniform magnetic field directed outwards of the page (panel a) and: an electric field directed downwards (panel b); a general force acting downwards (panel c), and a gradient in the magnetic field present in the upwards direction (panel d). Original image taken from Wikipedia, originally produced by user Stannerd, and improved by user Maschen. Used under Creative Commons Attribution 2.5 Generic license.*

Gradient drift occurs when there is a gradient in the magnetic field perpendicular to the direction of the magnetic field. In this case the drift velocity is given by:

$$\underline{\mathbf{V}}_{drift} = -\frac{\mu \underline{\nabla \mathbf{B}} \times \underline{\mathbf{B}}}{qB^2} \tag{1.21}$$

where the force due to the gradient is given by $\underline{\mathbf{F}} = -\mu \underline{\nabla \mathbf{B}}$. This drift can be understood through the equation of the gyro-radius (Equation 1.13), which describes how the gyro radius must decrease when the field increases. The process is shown visually in the bottom panel of Figure 1.3. We also note, that given the dependence on q, positively and negatively charged particles will drift in opposing directions, creating an electric field. This process is negligible on gyro-period time-scales, and is only evident when a particle experiences the magnetic gradient over many gyrations - thus μ is conserved.

In the presence of an electric field, particles are acted upon by a force $\underline{\mathbf{F}} = q\underline{\mathbf{E}}$, and so experience a drift given by:

$$\underline{\mathbf{V}}_{drift} = \frac{\underline{\mathbf{E}} \times \underline{\mathbf{B}}}{B^2} \tag{1.22}$$

This drift (the $\underline{\mathbf{E}} \times \underline{\mathbf{B}}$ drift) originates due to the electric field perpendicularly (to $\underline{\mathbf{B}}$) accelerating and decelerating the particle over many drift periods. As the gyro-radius (Equation 1.13) is dependent on the perpendicular velocity, it becomes smaller when $v_{\perp}$ is minimised (similarly to $\underline{\nabla \mathbf{B}} \times \underline{\mathbf{B}}$ drift).

In more complex magnetic field geometries, field lines may be curved. Particles undergoing bounce-motion along a curved field experience a centrifugal force, $\underline{\mathbf{F}}_c = \frac{mv_{\parallel}^2}{R_c^2}\underline{\mathbf{R}}_c$, acting perpendicularly to $\underline{\mathbf{B}}$ along the direction of the the radius of curvature, $\underline{\mathbf{R}}_c$. Substituting $\underline{\mathbf{F}}_c$ into Equation 1.20, we get the curvature drift:

$$\underline{\mathbf{V}}_{drift} = \frac{mv_{\parallel}^2 \underline{\mathbf{R}}_c \times \underline{\mathbf{B}}}{qR_c^2 B^2} \tag{1.23}$$

where we see again that the drift velocity is different depending on the charge of the particle (meaning there is a net electric field). As the curvature drift depends on $v_{\parallel}$, it is effectively dependent on the energy and pitch-angle of the particle (it also acts more strongly on more massive particles).

In a system comprising all of these forces, the total drift velocity is the sum of each of these individual drifts (gradient drift, $\underline{\mathbf{E}} \times \underline{\mathbf{B}}$ drift and curvature drift).

Earth's magnetosphere is often approximated to be dipolar. In a dipole field for a 90° pitch-angle particle, there is no bounce motion ($v_\parallel = 0$) and so there is no curvature drift. But, due to the radial gradient in a dipole field, a particle still experience gradient drift in the azimuthal direction. If the pitch-angle is not 90°, then due to bounce motion, a particle will experience a curvature force (maximally at the equator where the pitch-angle is furthest from 90°, and 0 at the mirror point). In concert, both of these effects cause particles to drift azimuthally about the dipole magnet. As with gyro-motion and bounce-motion, there is an invariant quantity associated with the drift-motion:

$$\underline{\mathbf{J}}_{drift} = \oint (\underline{\mathbf{p}} + q\underline{\mathbf{A}}) \cdot d\underline{\mathbf{l}} \tag{1.24}$$

where $\underline{\mathbf{p}}$ is the momentum, and $\underline{\mathbf{A}}$ is the magnetic vector potential:

$$\underline{\mathbf{B}} = \underline{\nabla} \times \underline{\mathbf{A}} \tag{1.25}$$

Using Stokes theorem ($\oint(\underline{\nabla} \times \underline{\Psi}) \cdot d\underline{\mathbf{s}} = \oint \underline{\Psi} \cdot d\underline{\mathbf{l}}$, for some vector field $\underline{\Psi}$), we can substitute $\underline{\mathbf{A}}$ in Equation 1.24, and if we also neglect $\underline{\mathbf{p}}$ due to the drift momentum being negligibly small we end up with:

$$\underline{\mathbf{J}}_{drift} = q \oint (\underline{\mathbf{B}}) \cdot d\underline{\mathbf{s}} = q\underline{\Phi} \tag{1.26}$$

where we have substituted the integral for the magnetic flux, as per Equation 1.15. This invariant is semantically and mathematically similar to the first invariant - the amount of flux contained within a drift-path is constant (provided temporal and spatial changes to the field are small). This means that we consider the third invariant in a bounce-averaged sense. The drift path is mapped along field lines to the surface of a sphere about the magnet and the flux through the enclosed surface is calculated. This is equivalent to calculating the flux outside of the drift path [see, e.g., Roederer and Zhang, 2014].

1.3 Magneto-hydrodynamics

To better describe the collective behaviour of a plasma Alfvén [1942] introduced a hydro-dynamic formalism, which is now referred to as magneto-hydrodynamics (or MHD). An ideal hydro-magnetic fluid conserves mass, momentum and energy

- allowing for the following general equation of motion:

$$\rho_m \frac{d\underline{\mathbf{V}}}{dt} = \rho_m \underline{\mathbf{g}} + \underline{\mathbf{j}} \times \underline{\mathbf{B}} - \underline{\nabla} \cdot \underline{\mathbf{P}} + \rho_q \underline{\mathbf{E}} \approx \underline{\mathbf{j}} \times \underline{\mathbf{B}} - \underline{\nabla} \cdot \underline{\mathbf{P}} \tag{1.27}$$

where ρ_m is the mass density, $\underline{\mathbf{V}}$ is the bulk-velocity of the fluid, $\underline{\mathbf{P}}$ is the pressure, and $\underline{\mathbf{g}}$ is the gravitational acceleration. In a space plasma, the gravitational and electric field terms are generally neglected (since $\underline{\mathbf{g}}$ is weak and quasi-neutrality prevents $\underline{\mathbf{E}}$). $\underline{\mathbf{j}} \times \underline{\mathbf{B}}$ is regarded as the magnetic force, and can be rewritten as:

$$\underline{\mathbf{j}} \times \underline{\mathbf{B}} = \frac{1}{\mu_0}(\underline{\nabla} \times \underline{\mathbf{B}}) \times \underline{\mathbf{B}} = \frac{(\underline{\mathbf{B}} \cdot \underline{\nabla})\underline{\mathbf{B}}}{\mu_0} - \frac{\underline{\nabla}(\underline{\mathbf{B}} \cdot \underline{\mathbf{B}})}{2\mu_0} \tag{1.28}$$

by substituting $\underline{\mathbf{j}}$ using Ampère's law (Equation 1.5) with no electric field. The penultimate term relates to the restoring force of a curved magnetic field line (commonly, the magnetic tension force) and the ultimate term is the magnetic pressure gradient force, for a magnetic pressure $P_B = \frac{B^2}{2\mu_0}$. Note that the vector calculus identity used to expand $(\underline{\nabla} \times \underline{\mathbf{B}}) \times \underline{\mathbf{B}}$ (which is not a triple product) is the following:

$$\underline{\nabla}(\underline{\mathbf{a}} \cdot \underline{\mathbf{b}}) = (\underline{\mathbf{a}} \cdot \underline{\nabla})\underline{\mathbf{b}} + (\underline{\mathbf{b}} \cdot \underline{\nabla})\underline{\mathbf{a}} + \underline{\mathbf{a}} \times (\underline{\nabla} \times \underline{\mathbf{b}}) + \underline{\mathbf{b}} \times (\underline{\nabla} \times \underline{\mathbf{a}})$$

rather than the typically used expansion of the triple product.

It is also possible to derive a generalised form of Ohms law [Somov, 2006]:

$$\underline{\mathbf{j}} = \sigma(\underline{\mathbf{E}} + \underline{\mathbf{V}} \times \underline{\mathbf{B}}) \tag{1.29}$$

which can be combined with Equations 1.4 and 1.5 to derive the induction equation:

$$\frac{\partial \underline{\mathbf{B}}}{dt} = \underline{\nabla} \times (\underline{\mathbf{V}} \times \underline{\mathbf{B}}) + \frac{1}{\mu_0 \sigma} \underline{\nabla}^2 \underline{\mathbf{B}} \tag{1.30}$$

which describes the time evolution of the magnetic field. The fist term on the right-hand side is known as the convective term, while the second is known as the diffusive term. By calculating the ratio of these terms, one can determine which is dominating the field evolution. This ratio is know, short-handedly as the Reynolds number, R_m:

$$R_m = \frac{|\underline{\nabla} \times (\underline{\mathbf{V}} \times \underline{\mathbf{B}})|}{|\frac{1}{\mu_0 \sigma} \underline{\nabla}^2 \underline{\mathbf{B}}|} \tag{1.31}$$

Practically, R_m is large when the conductivity (σ) or length-scales are large (i.e., the distances over which the derivatives, $\underline{\nabla}$, are taken). This implies that convective term dominates, and so the field and plasma are coupled (frozen-in). In contrast, if the plasma is weakly conductive (say, because there are many neutrals and collisions) or spatial gradients occur over small length scales, then the plasma and field are not coupled together (not frozen-in).

To determine, in general, whether plasma is frozen into the field or *vice versa* one calculates the plasma beta, β:

$$\beta = nk_BT\left(\frac{B^2}{2\mu_0}\right)^{-1} \tag{1.32}$$

which designates the ratio between the thermal pressure and the magnetic pressure (where pressure and energy density are interchangeable). If $\beta \gg 1$ then the thermal pressure dominates the motion of the plasma, and hence the field must be frozen into the plasma. Conceptually, a high plasma beta describes the situation when the pressure gradient force in the plasma dominates the $\underline{\mathbf{j}}\times\underline{\mathbf{B}}$ force in the equation of motion (Equation 1.27), and so the dynamics of the hydro-magnetic fluid are dominated by the plasma, not the field.

1.4 The Sun-Earth System

Space is uniquely well-suited for the production of plasma due to the extreme heat (i.e., collisional ionisation) and/or radiation (photo-ionisation) provided by stars, and the low chance of recombination (ions and electrons rejoining to form un-charged atoms or molecules) due to the sparsity of interstellar matter. Stars naturally provide an extreme environment which satisfies the requirements, and the interstellar and interplanetary mediums are so diffuse that the recombination rate is effectively zero. As such, once the atoms or molecules in space are ionised, they remain ionised effectively indefinitely - until they encounter a different environment such as a planetary atmosphere.

This section will serve to describe the various domains investigated in this thesis, and link them in to the descriptions and formalisms of plasma physics from sections 1.1 and 1.2. We'll begin with the solar wind, as this corresponds to the work in Chapter 4, then move onto Earth's magnetosphere which we investigate in the second and third research chapters, and finally we'll briefly explore the ways in which the two systems are coupled (as this also links into the third research chapter).

1.4.1 The Solar Wind

Solar wind originates in the solar corona [see Figure 1.4, Parker, 1958], and permeates the entire heliosphere - the interstellar region within which the Sun's magnetic field dominates [Parker, 1961; Dialynas et al., 2017; Burlaga et al., 2019; Rankin et al., 2019]. The first documented conjecture of a stream of ions from the Sun was published by Eddington [1910] in his investigation of the envelope of comet Morehouse. This was further developed by Biermann [1951], estimating solar wind speeds between $500 - 1500\,\mathrm{km\,s^{-1}}$ and number densities between $500 - 10^5\,\mathrm{cm^{-3}}$. However these ideas weren't well-understood until Parker [1958] demonstrated that, due to its temperature, the solar corona could not be in hydrostatic equilibrium - and so a radial pressure-gradient force acts on the coronal plasma, overcoming the gravitational force and causing a continuous stream of plasma to leave the Sun.

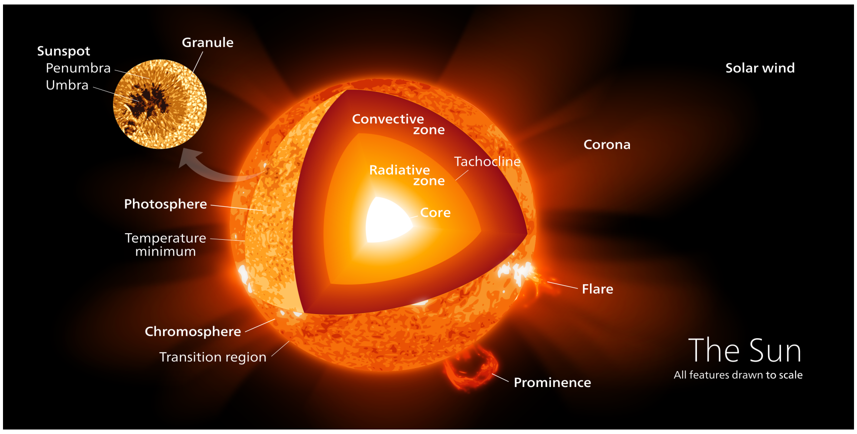


Figure 1.4: *A schematic diagram of the Sun's internal and external structure. Original image taken from Wikipedia, originally produced by user Kelvin13. Used under Creative Commons Attribution-Share Alike 3.0 Unported license.*

Describing Parker's model with MHD, the equation of motion simplifies to:

$$\rho_m \underline{\mathbf{V}} \cdot \underline{\nabla \mathbf{V}} = \rho_m \underline{\mathbf{g}} - \nabla \cdot \underline{\mathbf{P}} \tag{1.33}$$

by assuming an isotropic pressure and negligible electromagnetic effects (i.e., no electric field, or $\underline{\mathbf{j}} \times \underline{\mathbf{B}}$). We have reintroduced the $\underline{\mathbf{V}} \cdot \underline{\nabla \mathbf{V}}$ term (which is often neglected in ideal MHD) and removed the $\frac{d\mathbf{V}}{dt}$ term, due to searching for a static solution (i.e., no temporal variations). Additionally, a radially symmetric solution is assumed, and the pressure at infinity must be zero.

These conditions produce a radial solar wind flow, wherein the solar wind continues to accelerate due to the finite pressure gradient, while density, temperature and magnetic field strength all decrease. One of the key predictions from Parker's model is that the solar wind becomes supersonic - since the local sound-speed is density dependent, and the density drops off radially. The flow begins travelling below the sound-speed in the low corona, but exceeds it at a height known as the critical point. Eventually the solar wind also exceeds the Alfvén speed [Alfvén, 1942; Barnes and Hollweg, 1974], given by:

$$v_a = \frac{B}{\sqrt{\mu_0 \rho_m}} \tag{1.34}$$

The Alfvén speed is conceptually similar to the sound-speed in that it is the maximum speed of information transfer, but different in that the information is transmitted via electromagnetic rather than kinetic interactions. The Alfvén critical point is the radial distance where the solar wind flow becomes super-Alfvénic, typically lying at around $15-50$ solar radii [where 1 solar radius is 6.95700×10^8 metres, Weber and Davis, 1967].

At low radial distances from the Sun (less than 50 solar radii), the plasma co-rotates because it is 'frozen-in' to the solar magnetic field as the magnetic field pressure greatly exceeds the thermal and dynamic flow pressures [Alfvén, 1942; Roberts, 2007]. In contrast, at large radial distances the plasma flows approximately radially, and the field is frozen-in to the plasma (with the thermal flow pressure dominating). Recent work using Parker Solar Probe data [Kasper et al., 2019] has show that even at distances as great as $35\,\mathrm{R}_\odot$ there can be significant azimuthal components to the solar wind velocity[1].

Due to the Sun's rotation, the field being frozen into the plasma, and the radial flow direction of the solar wind, it can be shown that the interplanetary magnetic field forms a spiral - the Parker spiral [Parker, 1958], see Figure 1.5.

---

[1] Though, these results may be due to a calibration issue, rather than being a feature in-and-of themselves [see Appendix B in Finley et al., 2020].

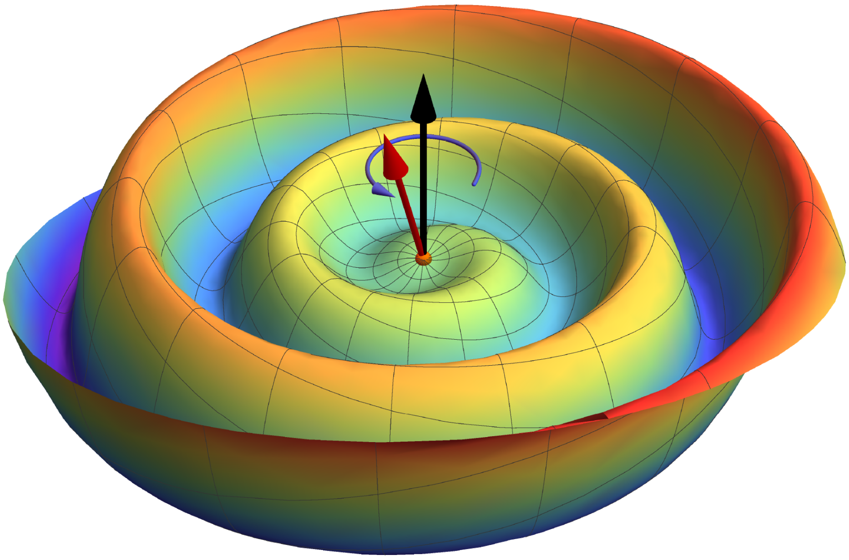


Figure 1.5: *A schematic diagram of the heliospheric current sheet in a Parker solar wind model. The black arrow designates the axis of rotation (rotating as per the purple arrow), and the red arrow designates the magnetic dipole direction. The Parker spiral structure is clearly visible. Original image taken from Orcinha et al. [2019].*

1.4.1.1 Solar Wind Structures and Transients

From the earliest *in situ* observations, it was clear that the solar wind could be broadly classified into two types: fast and slow [Neugebauer and Snyder, 1966; Stakhiv et al., 2015], see Figure 1.6. This duality was found to extend beyond the local solar wind speed, but is present in the elemental composition and ion charge states of the solar wind, indicating distinct coronal source properties of fast and slow wind [von Steiger et al., 2000; Geiss et al., 1995]. Fast wind is found to originate from coronal holes [Sheeley et al., 1976]. These are magnetically open regions of the corona where the plasma can freely escape, meaning that coronal holes appear dark in EUV emission (since there is less plasma and less time for it to be heated). The formation and release of the slow wind is a current area of research, but it originates from the vicinity of closed coronal magnetic structures such as the streamer belt [Antiochos et al., 2011; Ko et al., 2006; Xu and Borovsky, 2015; Brooks et al., 2015]. Arguments have been made that the slow solar wind may be due to small-scale reconnection [as per Sheeley et al., 1997], and thus CMEs (coronal mass ejections, see below) and the slow wind may exist on a spectrum. At solar minimum, coronal holes cover the polar regions, with the streamer belt confined close to the solar equator. At solar maximum, the coronal field is far less ordered. The resulting variation of solar wind speed can be seen in Figure 1 of McComas et al. [2013]. Despite the breakdown of the latitudinal dependence at solar maximum, there is still a good separation between streams of different speeds. This suggests that despite the activity, the source regions remain isolated from one another, and there is not significant mixing of the streams.

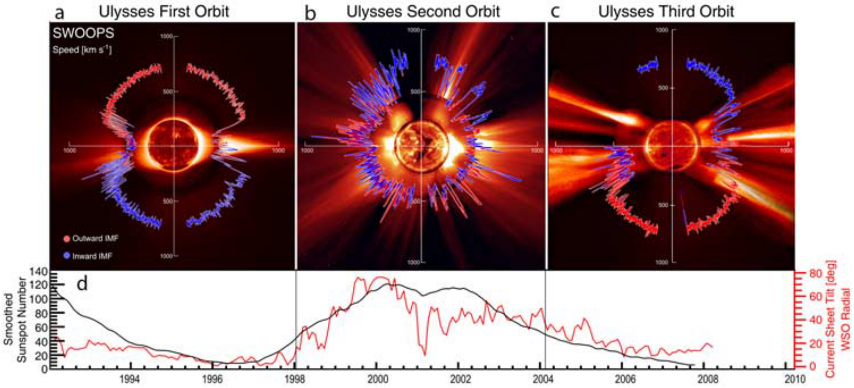


Figure 1.6: *Polar plots of solar wind speed measured by the Ulysses spacecraft as a function of heliospheric latitude. Panels (a) and (c) represent solar minimum, panel (b) represents solar maximum, and panel (d) shows the sunspot number (left axis, proxy for the solar cycle) and the current sheet tilt angle. The blue/red colours in panels (a-c) represent the directions of the Sun's magnetic field. The fast/slow solar wind duality is clearly present, especially at solar minimum when the Sun's magnetic field is more ordered. Original image taken from McComas et al. [2008].*

Having solar wind streams with different velocities leads to additional heliospheric structures. When faster (supersonic) solar wind catches up with slower (subsonic) wind, a shock region forms. These the shocks are typically embedded within regions of compressed solar wind known as co-rotating interaction regions (CIRs) [Wilcox and Ness, 1965; Neugebauer and Snyder, 1966; Smith and Wolfe, 1976; Gosling, 1996; Richardson, 2018], because they recur on the time-scale of solar rotation (≈ 27 days). Figure 1.7 shows a schematic diagram of a CIR. If coronal holes exist for less than one solar rotation - which most do [Hewins et al., 2020] - then the CIR will not be a recurrent feature, but given that low-latitude coronal holes may exist for up to three years [Hewins et al., 2020], recurrence is frequently observed.

Through a super-posed epoch analysis of 25 CIRs, Gosling [1996] showed what a typical CIR might look like - as measured *in situ* (see Figure 1.8 below):

- There is a moderate positive gradient in solar wind speed between the preceding slow wind and anteceding fast wind.
- There is a very sharp gradient in the proton thermal pressure - $P_T = nk_BT_p$ [Burlaga and Ogilvie, 1970].

- The proton density spikes at the interface between the two streams.
- The flow is deflected. The preceding slow wind is deflected in the positive azimuthal direction (i.e., in the sense of solar rotation), while the anteceding fast wind is deflected in the negative azimuthal direction.

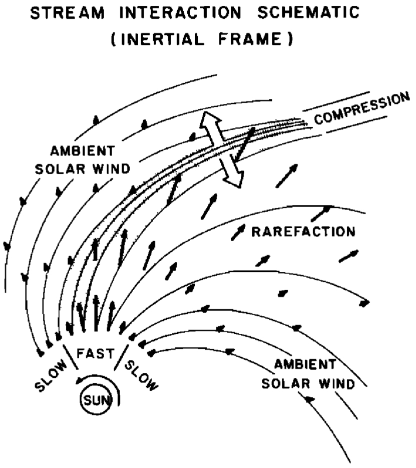


Figure 1.7: *A schematic diagram of the corotating interaction region formed when fast solar wind catches up with slower solar wind. Diagram taken from Pizzo [1978].*

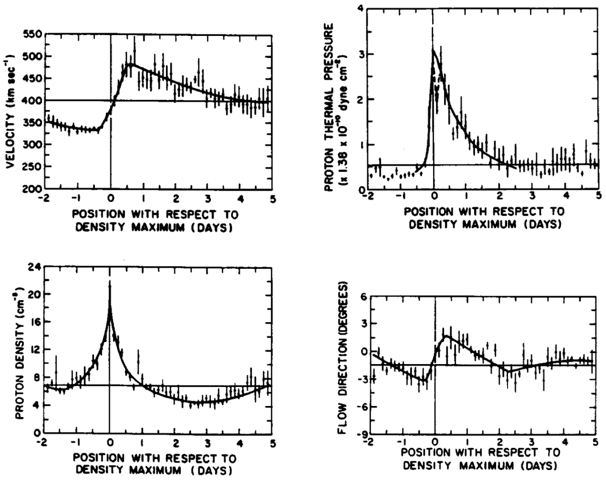


Figure 1.8: *Super-posed epoch analysis of 25 CIRs as presented in Gosling [1996]. The panels present the solar wind velocity, thermal pressure, density and flow directions, centred on the time of maximum density. Each panel also includes error-bars corresponding to the standard deviation at each point.*

In relation to solar wind classification, it is the first point which makes solar wind speed a bad descriptor for solar wind type. We know observationally and theoretically that the solar wind speed changes as CIRs develop. This effect also increases as the solar wind propagates through the heliosphere. So even if one considers the effect at 1 AU (AU - astronomical unit - is the distance between the Sun and the Earth) negligible, it means that the same classification scheme cannot be used on *in situ* data from more distant locations [Gosling and Pizzo, 1999; Hanlon et al., 2004].

CIRs exist, almost exclusively, in the low-latitude regions of the heliosphere (i.e., the ecliptic plane). The primary reason for this is that the slower solar wind originates from these regions. Secondarily, coronal holes typically exist at higher latitudes - where lower latitude coronal holes are often equator-ward distensions of the polar coronal holes. Because the solar magnetic axis is offset from its rotational axis, over the course of a solar rotation, faster solar wind is granted access to lower heliospheric latitudes, meaning it might catch up with slower wind.

In contrast to these broad structures, the solar wind also has transient structures. Coronal mass ejections (CMEs) are the most notable source of transient solar wind events. CMEs can release $10^{11} - 10^{13}$ kg of plasma into the interplanetary medium [Kahler, 1992; Chen, 2011]. Once the CME plasmoid (the magnetically enclosed plasma ejected from the Sun) enters interplanetary space, it is referred to as an ICME (interplanetary coronal mass ejection). Figure 1.9 presents a schematic of a CME.

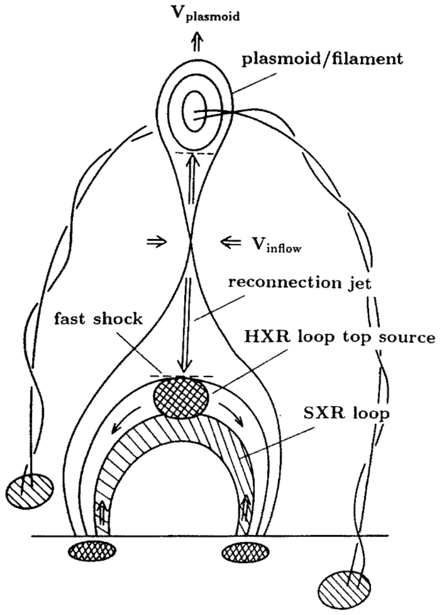


Figure 1.9: *A schematic diagram of a CME occurring on the Sun. The inflow velocity relates to the speed of plasma streaming into the reconnection region, with the fast shock being due to the plasma being ejected from the reconnection region. The HXR and SXR refer to the regions from which hard and soft x-rays are emitted. The plasmoid is the closed loop of magnetic field lines which are no longer topologically connected to the the Sun. Original figure presented in Shibata [1996].*

I/CMEs and their structure can be identified using telescopes to observe for activity on the Sun or Thomson scattering within the solar wind - e.g., much CME work uses the two-spacecraft STEREO mission [Solar TErrestrial RElations Observatory, Kaiser, 2005] to understand their 3D structure and evolution [Eyles et al., 2009; Scott et al., 2019; Jones et al., 2020]. ICMEs can also be identified by *in situ* measurements [see the super-posed epoch analysis below, Cane and Richardson, 2003].

As with CIRs, we can understand what a ‘typical’ ICME structure is via a super-posed epoch analysis of observed ICMEs. Masiás-Meza et al. [2016] present such a study, and we surmise the following properties:

1. The velocity of ICMEs is much faster than solar wind within which they are embedded - giving rise to a shock front.
2. The magnetic field is two (or more) times greater than the interplanetary magnetic field.
3. The number density spikes - though this occurs in relation to the shock front, with the plasma inside of the ICME remaining similar to that of the solar wind due to the isothermal expansion of the ICME.
4. As with the number density, the proton temperature also increases within the shock region, but is typically lower than the solar wind inside of the ICME.

1.4.1.2 The Solar Cycle

The Sun exhibits large-scale reconfiguration of its magnetic field on 11 (22) year time-scales - this is known as the Schwabe (Hales) solar cycle. Over the course of this cycle, the Sun’s magnetic field reverses polarity (twice for the Hales cycle). Morphologically, the polarity of the Sun’s magnetic field has very little effect on the processes discussed in context of the solar cycle. Instead the solar cycle is split into four main categories, related to the number of sunspots or sunspot groups visible in the corona. Sunspot maximum, minimum, the declining phase, and the enhancement phase.

The first evidence of the '10 year' solar cycle was presented by Schwabe [1843], who identified a periodic enhancement in the number of sunspot groups on the Sun. Subsequently, most characterisations have relied on sunspot (group) numbers to quantify solar cycles - though more recent work uses the 10.7 cm radio flux as a proxy [Tapping and Charrois, 1994].

A prototypical solar cycle is as follows (shown in Figure 1.10):

1. Solar minimum: coronal holes are limited to the polar region, with a well-defined low-latitude streamer belt (Figure 1.10a).
2. The internal magnetic field becomes increasingly toroidal due to the differential rotation of the Sun, amplifying the mid-to-low-latitude magnetic field [Figure 1.10b, Babcock, 1961].
3. Due to the increased field strength sunspot groups begin to appear in the mid-latitude regions, around $20 - 50^{\circ}$ [Figure 1.10c-e, Takalo, 2020]. The sunspot themselves come in pairs of opposing polarity. With the leading sunspot having the same polarity to the polarity of the respective hemispheric pole (Figure 1.10f). This is known as the Hale-Nicholson law of sunspot polarity [Hale, 1924]. Joy's law explains that the leading sunspot tends to lie closer to the equator (Figure 1.10f), and the angle between them increases with latitude [Hale et al., 1919].
4. During the cycle, sunspot groups begin to appear at lower and lower latitudes. This effect is known as Spörer's law [Maunder, 1903].
5. As the sunspots form closer and closer to the equator, the leading and opposite polarity sunspots from each hemisphere neutralise their respective magnetic fields.
6. Due to meridional flows, the remaining sunspots are preferentially transported towards the poles. This results in a build-up of sunspots with polarity opposing the hemispheric polarity (Figure 1.10g).
7. Once enough opposite polarity sunspots exist in a given hemisphere, their magnetic flux dominates the polarity of the hemisphere and the solar magnetic polarity is effectively reversed (Figure 1.10h-i).
8. This cycle then recurs (with opposite polarity) over the next 11 years, completing the 22 year cycle.

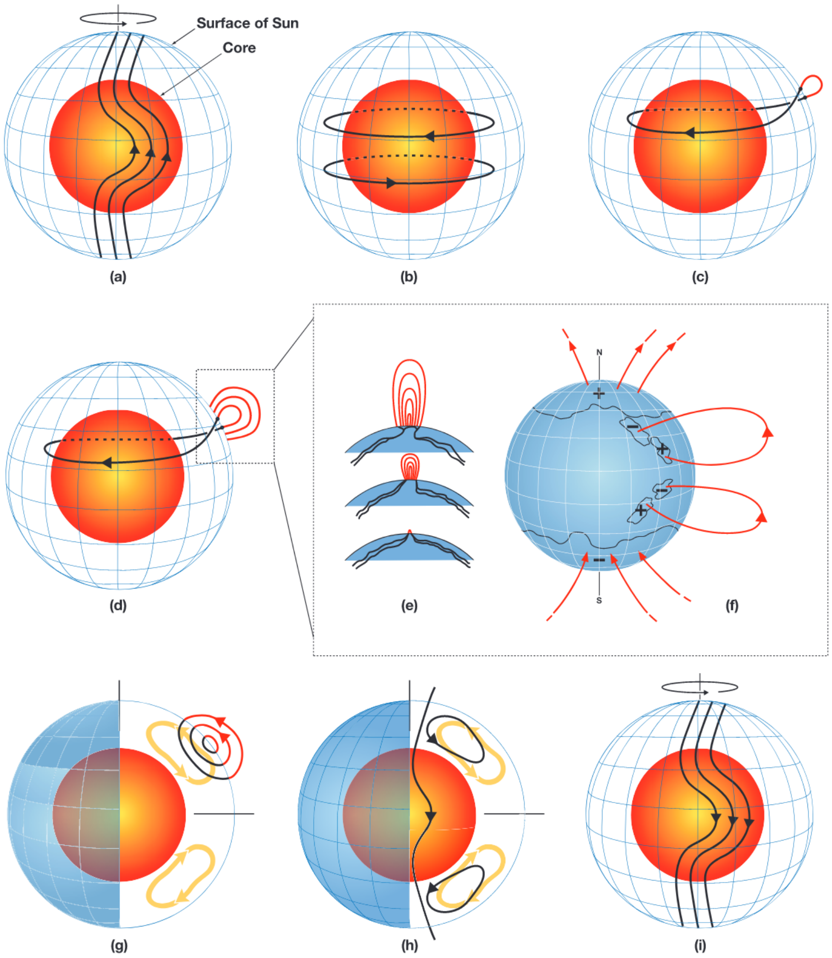


Figure 1.10: *A schematic diagram of the Schwabe solar cycle, as present by Dikpati and Gilman [2007]. Panel (a) shows the Sun's poloidal field at the 'beginning' of the cycle. Panel (b) shows the result of the Sun's transition to a toroidal field. Panel (c) shows the emergence of sunspots, and panels (d-f) provide greater detail, with (f) specifically visualises both Hales-Nicholson's and Joy's law (see text). Panel (g) shows the meridional flows which transport the remaining flux pole-wards. Panel (h) serves to show that the flux is transported downwards into the Sun, resulting in the polarity inversion which can be seen in panel (i).*

The solar cycle is important, from a solar wind point-of-view, because it has a profound effect on how well-ordered the streamer belt and coronal hole magnetic fields are. During solar maximum, where the dipole components of the field are weaker compared to the multi-polar components, large-scale solar wind structures are much less frequent, as well-defined coronal hole and streamer belt regions are not present, but CMEs are approximately ten times more common [Lamy et al., 2019]. In contrast, during the declining phase of the solar cycle, when the system is well-ordered, there are more CIRs [Legrand, 1985; Mursula and Zieger, 1996; Mursula et al., 2015] and flares[2] [Wagner, 1988].

---

[2] Though, other works suggest that the frequency of flares may not have a one-to-one correlation

1.4.2 Earth's Magnetosphere

Earth has an internally generated magnetic field [Gillbert, 1600; Larmor, 1919a,b; Bullard, 1949]. In the absence of any external forces, this magnetic field (to a first-order approximation) would present itself as a dipole. Earth's field points south to north, with the 'northern' magnetic pole located in the southern hemisphere and *vice versa*. The dipole axis is not perfectly aligned with the spin axis[3], instead being offset by roughly $11.5°$ [Chapman and Bartels, 1941]. Additionally, the longitude of the poles varies between the hemispheres with the northern and southern poles longitude being approximately $73°$ west and $107°$ east, respectively.

As a consequence of Alfvén's frozen-in theorem, the magnetic field and plasma contained within the solar wind cannot - under the conditions of ideal MHD - mix with those of the Earth. This yields a heliospheric cavity within which Earth's magnetic field dominates - 'The Magnetosphere'. The plasma within the magnetosphere is frozen into the field with $\beta \ll 1$.

In reality, due to the interactions between the solar wind and Earth's magnetosphere, the magnetospheric field is compressed on the Sun-facing side (the day-side), and elongated into a tail-like structure on the anti-Sunward side (the night-side). When the super-sonic solar wind reaches the effectively (radially) stationary magnetosphere, it transitions to sub-sonic speeds causing the formation of a standing bow shock up-stream of the planet. The sub-sonic wind continues towards the magnetosphere, compressing it until pressure balance (Equation 1.35, below) between the solar wind and the magnetosphere is established.

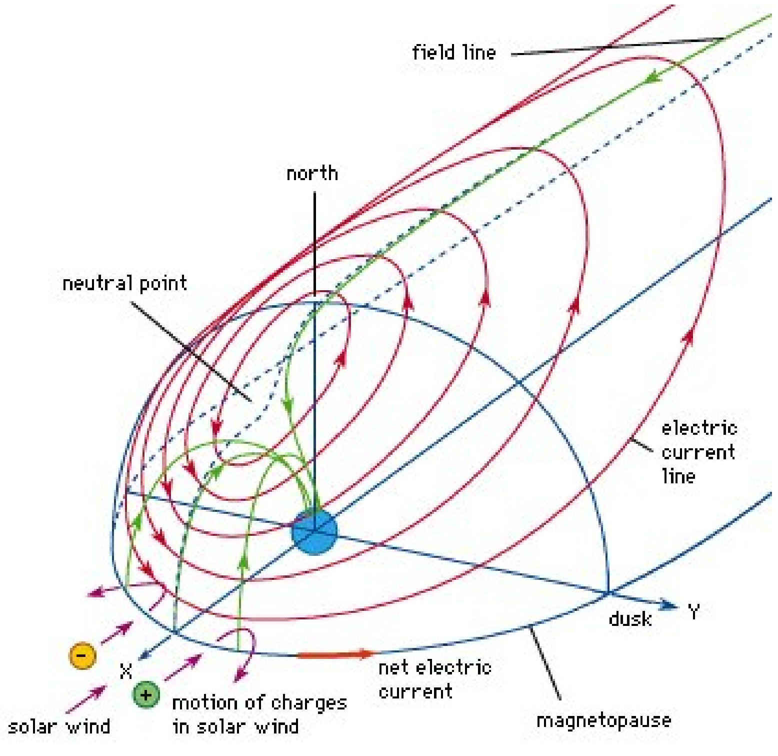
As the solar wind cannot go through the magnetosphere, it must flow around. In general, this flow is equivalent to a a typical fluid flowing around an obstacle. However, because we're dealing with plasma and magnetic fields there is an additional feature - a current is formed in the region between the two different plasmas. In the context of MHD, we have a current set up due to the spatial gradients in the magnetic field (from Equation 1.5, assuming no electric field). One may, alternatively, use kinetic plasma physics to explain the phenomena: as ions and electrons encounter the magnetosphere, they perform half a gyra-

---

to the solar cycle [Temmer, 2010].

[3] Meaning that Cowling's anti-dynamo theorem does not pose a problem Cowling [1933]. Though others contest that the offset may be an artefact the of spherical harmonic representation [Akasofu, 2002].

tion about the geomagnetic field lines, and are subsequently expelled from the magnetosphere [Chapman and Ferraro, 1931; Ganushkina et al., 2018]. Since electrons and ions are oppositely charged, they gyrate in opposing directions, and so a current is formed (electrons preferentially travelling eastwards and ions preferentially travelling westwards) - this is the Chapman-Ferraro Current [Chapman and Ferraro, 1931]. The current system can be seen in Figure 1.11. Whilst the currents are westwards on the dayside of the magnetosphere, as the solar wind flow continues past the Earth and encounters the night-side field lines, the sense of current flow is reversed, allowing the currents to close eastwards across the surface of the magnetosphere.



©1994 Encyclopaedia Britannica, Inc.

Figure 1.11: *A schematic diagram of the Chapman-Ferraro currents in the northern hemisphere. The blue circle in the middle represents Earth, with the Sun in the positive x-directions. The solar wind direction is in the negative x-direction, as shown by the purple arrow. The purple arrows further demonstrates the particle kinetics of the electrons and ions as they encounter the magnetosphere (whose shape is outlined by the blue curves). The directions of the Chapman-Ferraro currents are shown by the red curves - exemplifying the westwards direction in the equatorial plane, and eastwards direction above the magnetotail. This figure is taken from the Encyclopædia Britannica.*

The region within which the Chapman-Ferraro current flows is known as the magnetopause. The magnetopause is co-incident with where pressure balance is reached between the solar wind and the magnetosphere. In this case the balance is given by:

$$P_{SW} + \frac{B_{SW}}{2\mu_0} + n_{SW}k_BT_{SW} = \frac{B_{MS}}{2\mu_0} + n_{MS}k_BT_{MS} \tag{1.35}$$

where P_{SW} is the ram pressure (ρv^2), and the subscripts 'SW' and 'MS' refer to the solar wind and magnetosphere, respectively. Since the solar wind is a high-beta plasma, and the magnetosphere is a low-beta plasma, we can remove the solar wind magnetic pressure term and the magnetospheric thermal pressure term. The solar wind thermal pressure is very low compared to the ram pressure, so we neglect that too [Kivelson and Russell, 1995]. This leaves us with:

$$P_{SW} = \frac{B_{MS}}{2\mu_0} \tag{1.36}$$

If we assume a dipole field, the field strength at a particular radial distance, r, is given by:

$$B(r) = B_E\left(\frac{R_E}{r}\right)^3 \tag{1.37}$$

where B_E is the magnetic field strength at the equator of the Earth's surface, and R_E is the radius of the Earth. At the magnetopause there is an additional contribution from the Chapman-Ferraro currents and general compression state of the magnetosphere on the day-side. Hence, we more accurately employ $B(r) \approx 2B_E\left(\frac{R_E}{r}\right)^3$ [Ganushkina et al., 2018; Kivelson and Russell, 1995]. With this we can re-write Equation 1.36 in terms of the radial distance, r, giving a simplified and idealised model of the magnetopause location:

$$r = R_E\left(\frac{2B_E^2}{\mu_0\rho_{SW}v_{SW}^2}\right)^{1/6} \tag{1.38}$$

In practice, a model of the magnetopause location - e.g., the Shue model [Shue et al., 1998] - is often used instead, as such models provide less idealised and more representative values of the magnetopause location.

1.4.2.1 Magnetospheric Reconnection

Magnetic reconnection is one of the primary drivers of magnetospheric dynamics [Eastwood et al., 2015]. For magnetic reconnection to happen, the topology of the magnetic field has to change. This occurs when magnetic fields have exceedingly

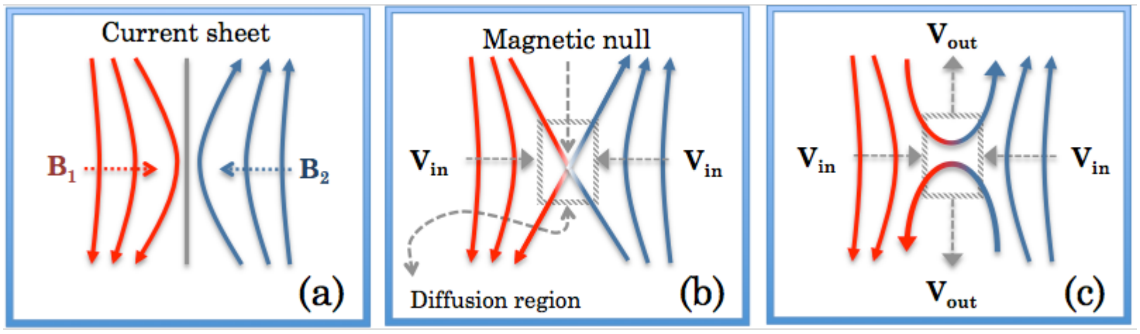


Figure 1.12: *A schematic diagram of reconnection between anti-parallel magnetic field lines (red and blue). Panel a is pre-reconnection, as pressures force the field lines together - enhancing the current sheet. In panel c reconnection has occurred and the previously-anti-parallel field has begun to dipolarise - with the magnetic tension acting on the field such that it rapidly advects in the directions of V_{out}. This figure is taken from Genestreti et al. [2012].*

strong spatial or temporal gradients. In such magnetic topography the electrons and ions may decouple from the field, at which point their motion is no longer governed as described in Section 1.2. This is mirrored in the MHD description as the magnetic Reynolds number, R_m, becomes small ($\lesssim 1$) due to sharp spatial gradients in the magnetic field.

Specifically in the case of spatial gradients in the field, if the changes are smaller than the characteristic length-scales of the plasma, then a phenomenon known as magnetic reconnection may occur (see Figure 1.12). Aside from the gyro-radius as one of these length-scales, there is also the inertial length $\lambda_{ף}$ for a species ף given by:

$$\lambda_{ף} = \frac{c}{\omega_{pף}} \tag{1.39}$$

where c is the speed of light, and $\omega_{pף}$ is the species plasma frequency. The plasma frequency corresponds to the frequency of oscillation of a species of particle in the plasma, and is given by:

$$\omega_{pף} = \left(\frac{4\pi n_{ף} Z_{ף}^2 e^2}{m_{ף}} \right) \tag{1.40}$$

where $n_{ף}$, $Z_{ף}$ and $m_{ף}$ are the species number density, charge state (e.g., 1 for protons and electrons) and mass, respectively. Reconnection occurs when the spatial gradients of the field are small enough to cause the electrons to decouple from the field.

The cycle of reconnective behaviour at Earth (locations shown in Figure 1.13) is known as the Dungey cycle [Dungey, 1961; Milan et al., 2007]. The Dungey cycle describes the process by which energy is transferred between the solar wind and the magnetosphere. The Dungey cycle is as follows:

1. Magnetic reconnection occurs around the nose (the point facing the Sun) of the magnetopause, leading to coupling between the magnetic fields of the solar wind and magnetosphere.
2. Due to the interplanetary magnetic field being frozen into the plasma, it continues streaming past the Earth. Since the fields are now coupled, field lines connected to the Earth are pulled along with the solar wind, forming an extended magnetotail.
3. The solar wind pressure and build-up of magnetic flux in the magnetotail lead to conditions favourable for reconnection. The reconnection allows for the interplanetary magnetic field to decouple from the magnetosphere and continue along with the solar wind.
4. Once the magnetospheric field lines are reconnected, the magnetic tension causes the field lines to travel back towards the planet, where they subsequently convect back to the day-side, setting up for continuation of the cycle.

Dungey [1961] initially proposed equivalent rates for the reconnection at the nose and in the tail, however, this steady-state condition is not a valid approximation for the behaviour observed [Milan et al., 2007].

Day-side reconnection provides the primary source of plasma flow into the magnetosphere. When the field between the solar wind and the Earth is coupled, it allows for solar wind plasma to flow along the field lines into the magnetosphere. Additionally, the convection of the field (both after day-side reconnection and after night-side reconnection) transfers momentum from the solar wind to the magnetosphere [Nakamura et al., 2017; Borovsky and Valdivia, 2018].

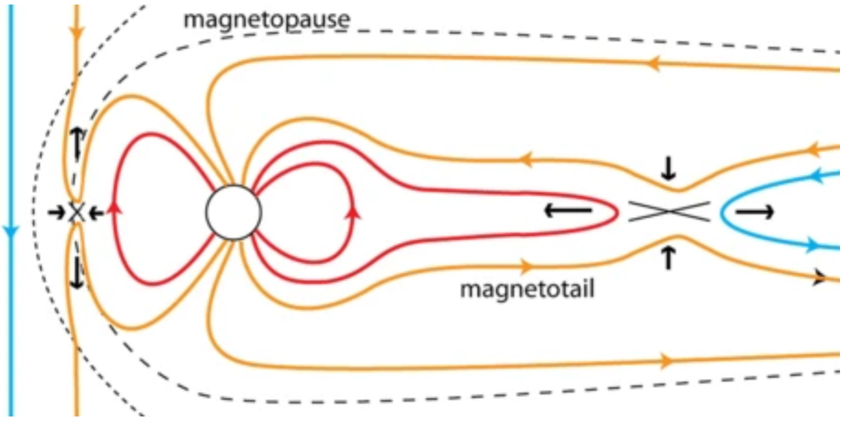


Figure 1.13: *A schematic diagram of the main reconnection site in Earth's magnetosphere - denoted by black crosses, with arrows representing the direction of plasma flows. The blue lines represent the interplanetary magnetic field; the orange lines are open magnetic field lines, where Earth's magnetic field has reconnected to the interplanetary magnetic field; and the red lines represent Earth's closed magnetic field lines (those whose foot-points are both connected to the Earth). The black dashed line is the magnetopause and the black dotted line is the bow shock. This figure is originally produced in Eastwood et al. [2017b].*

1.4.2.2 The Plasma Sheet and Ring Current

There are many different magnetospheric regions, current systems and structures. We will discuss some in this section, and cover others - such as the radiation belts - in subsequent sections.

In the night-side magnetosphere there is a dense region of relatively hot plasma known as the plasma sheet [with temperatures greater than 1 keV, Bame et al., 1967]. It can exist as close to the Earth as $7\,\mathrm{R_E}$ [Stawarz et al., 2017] and beyond $60\,\mathrm{R_E}$ [Meng and Mihalov, 1972]. within the plasma sheet is an embedded current sheet. The current is driven by the spatial gradients in the magnetotail fields. During tail reconnection plasma sheet electrons are transported to the inner magnetosphere due to the convection of the field lines, where they form the source population for the relativistic radiation belt (see the following section) population [Thorne et al., 2013; Jaynes et al., 2015].

Closer to the planet there is a ring current, shown in Figure 1.14. Though, it is more easily discussed as three separate ring currents. The ring currents are predominantly driven by a pressure gradient force acting on the plasma. The pressure peaks at roughly $3\,\mathrm{R_E}$ [Lui et al., 1987; Lui, 2003], and hence drives the first, eastwards ring current beyond $3\,\mathrm{R_E}$, and the second, westwards ring current within. The drift velocity is given by:

$$V_{drift} = -\frac{\nabla \mathbf{P} \times \mathbf{B}}{\rho q B^2} \tag{1.41}$$

which can be re-arranged to give the current, $\underline{\mathbf{j}}$ (where, in general, $\underline{\mathbf{j}} = \rho q \underline{\mathbf{v}}$):

$$\underline{\mathbf{j}} = -\frac{\nabla \mathbf{P} \times \mathbf{B}}{B^2} \tag{1.42}$$

The third ring current arises due to Earth's day-side magnetosphere having two magnetic field minima (due to solar wind compression), one above, the other below the magnetic equator [Antonova and Ganushkina, 2000; Ganushkina et al., 2018]. The bifurcation of the magnetic minima contour on the day-side leads to a bifurcation in the outer ranges of the ring current, leading to the cut-ring current.

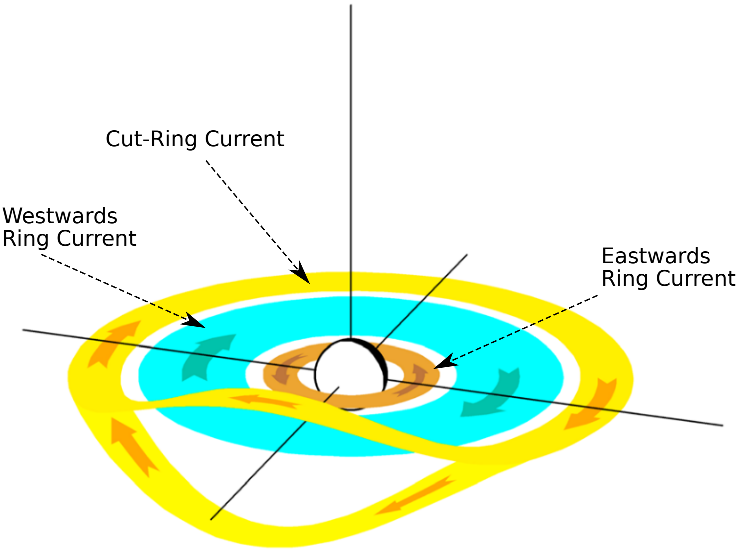


Figure 1.14: *A schematic diagram of Earth's symmetric ring current, taken from Ganushkina et al. [2018]. The diagram has Earth in the middle with the night-side mostly hidden behind the representation of the planet. In cyan and brown are the westwards and eastwards currents, respectively, as well as the bifurcating cut-ring current in yellow. In this diagram, the density gradient is implicitly maximised in the region between the cyan and brown ring currents, leading to their generation and opposite directions.*

The westwards ring current is enhanced subsequent to magnetic reconnection in the tail. As previously mentioned, the convection of newly reconnected field lines brings plasma from the plasma sheet into the inner magnetosphere. This plasma is comprised of ions and electrons each having a range of pitch-angles. This plasma experiences a magnetic gradient force due to interacting with the stronger field closer to the planet. If these particles have a non-90° pitch-angle, then they also experience a curvature force. In concert, these forces both act to drive a westwards flow of ions and eastwards flow of electrons. Thus a current is produced, adding to the effect of the ring current.

The ring current produces a measurable magnetic disturbance [Ganushkina et al., 2018] on the surface of the Earth (right hand grip rule). If the westwards current dominates, then the generated magnetic field opposes that of the Earth, and so measurements at the surface are depressed and *vice versa* [Ganushkina et al., 2018]. Equivalently, at positions beyond the ring currents, there is a measurable effect on the magnetic field which acts to enhance and depress the field depending on which flow direction dominates (enhancement occurs when the westwards flow dominates).

The ring current itself can be enhanced by various mechanisms - most significantly solar wind driving, ionospheric outflow and plasma-sheet dynamics [e.g., storms and substorms, see Section 1.4.2.4, Daglis, 2007; Sandhu et al., 2018].

1.4.2.3 The Radiation Belts

Earth's radiation belts typically manifest in the magnetosphere as two toroidal regions of magnetically confined, energetic plasma. They were discovered in 1958 when Explorer 1 [Van Allen et al., 1958] data was understood to be showing the results of a saturated Geiger counter. As the name suggests, the radiation belts are a radioactive environment. Despite Explorer 1 measuring only the low-altitude inner belt, it was quickly realised that there were in fact two radiation belts [Van Allen and Frank, 1959; Lyons and Thorne, 1973] - one closer to the Earth, on further away - with a slot region, typically devoid of higher energy particles, in between (see Figure 1.15). Both radiation belts contain trapped ions and electrons, though they are typically discussed as the proton (inner) and electron (outer) radiation belts due to these species dominating the respective dynamics [Li and Hudson, 2019]. The relativistic electrons commonly observed in the outer radiation belt pose a threat to spacecraft via surface charging and electrostatic discharges between internal components [Frederickson et al., 1991; Baker, 2001; Eastwood et al., 2017a]. Many spacecraft orbits overlap with the outer radiation belt (ORB), meaning that a large portion of space-based infrastructure exists in the radiation belts. Hence, there is significant interest in being able to accurately model and forecast its electron properties - thus motivating the work presented in Chapters 5 and 6.

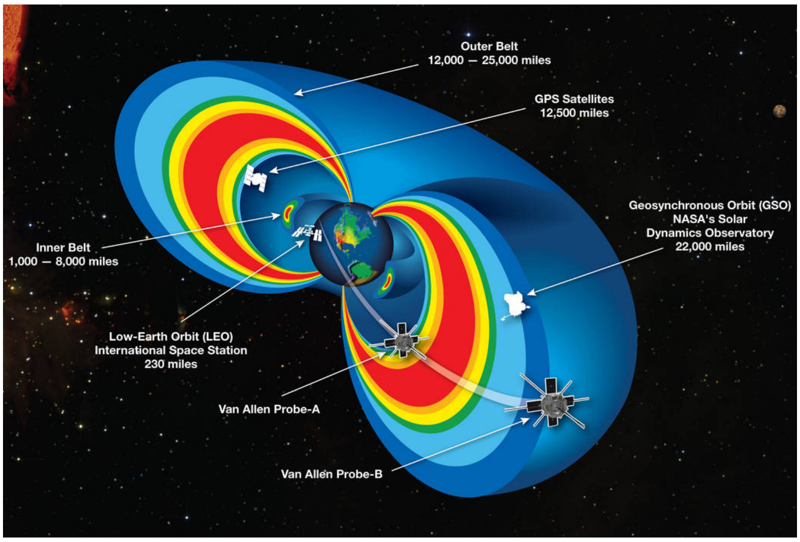


Figure 1.15: *A schematic diagram of Earth's radiation belts. The colours represent the relative intensity of radiation at a given location, with red being the highest intensity. The diagram shows that the inner belt is typically much smaller and more confined. Superimposed are some spacecraft and orbits for reference (e.g., the International Space Station orbiting within the inner belt, the Solar Dynamics Observatory orbiting at the outer boundary of the outer radiation belt, and the Van Allen Probes cutting through the radiation belts). Image credit: NASA.*

The inner belt, approximately located at $2\,\mathrm{R_E}$, is generated by a combination of CRAND [Cosmic Ray Albedo Neutron Decay, Singer, 1958] and SEP events [solar energetic protons, Blake et al., 1992; Selesnick et al., 2007].

The ORB (outer radiation belt) comprises a highly dynamic electron population, where fluxes can change by orders of magnitudes on minute timescales [Blake et al., 1992]. It has a peak ($2\,\mathrm{MeV}$) flux at approximately $4\,\mathrm{R_E}$ and frequently extends beyond geosynchronous orbit - roughly $6.6\,\mathrm{R_E}$ [Li et al., 2013]. Typically, the ORB is considered as the relativistic[4] tail of the electron energy distribution, as these particles have very weak interactions with the electric fields in the magnetosphere. Though, the 10s to 100s of keV source and seed electrons are also often considered as they ultimately become radiation belt electrons [Jaynes et al., 2015].

The motion of the higher energy electrons are governed by the particle motion described in Section 1.2 - gyro-motion about field lines, bounce-motion along field lines, and drift-motion about the Earth. With these motions, and knowledge

---

[4] Relativistic effects (loosely) become important when the particle energy approaches the rest mass of the particle. For electrons this is 511 keV.

of the geography, geometry and topology of the system, we can consider the boundaries of the ORB.

The inner boundary is two-fold. Firstly, there is the loss cone - this is the range of pitch-angles for which the mirror point (bounce-motion - Section 1.2.2) is within the atmosphere. If an electron enters the atmosphere, it is considered lost from the system due to the large increase in collisional frequency. Particles are scattered, in pitch-angle, into the loss cone by wave-particle interactions [WPIs, Thorne and Kennel, 1971; Gamble et al., 2008; Miyoshi et al., 2008; Ukhorskiy et al., 2010; Rodger et al., 2015]. Secondly, there is the 'impenetrable barrier' - a region close to the Earth which relativistic electrons cannot breach. Early work [Lyons and Thorne, 1973] suggested that this boundary was governed by Coulomb scattering and wave-driven pitch-angle scattering. More recently, it has been suggested that singularly wave-driven radial diffusion can explain the phenomenon [Ozeke et al., 2018], or, alternatively, it may be due to the suppression of chorus wave growth due to Earth-based electro-magnetic transmitter signals [Foster et al., 2020].

The outer, spatial boundary occurs at the boundary between trapped and untrapped electrons. This is commonly referred to as the last closed drift-shell for an electron which conserves all three of the adiabatic invariants (Section 1.2), where a closed drift-shell is the surface traced out by the combination of electron bounce-motion and drift-motion (both gradient and curvature). Practically, it is easier to think only of the third invariant [Xiang et al., 2017; Albert et al., 2018; Olifer et al., 2018], since the drift motion has the lowest frequency and largest spatial extent and so the first two invariants are conserved by definition[5] [though some works do consider the first invariant explicitly, e.g., Sivadas et al., 2019].

The outer boundary is dependent on energy and pitch-angle due to the gradient drift being related to $v_\perp$ through μ (and the gradient itself being dependent on the gyro-radius[6]), and the curvature drift being related to $v_\parallel$. Technically, electric fields also act, but due to the energy dependence of the aforementioned drifts the $\underline{\mathbf{E}} \times \underline{\mathbf{B}}$ drift is comparatively negligible [Buck, 1987; Öztürk, 2012].

The outer boundary is further dependent on the the solar wind. If the solar wind compresses the magnetosphere, then it is possible that the electron drift paths

---

[5] Recall that an adiabatic invariant is conserved if the field varies on spatial scales smaller than the associated motion, or timescales shorter that the periodic motion.

[6] A larger gyro-orbit would sample a greater extent of magnetic field dropping off with $\frac{1}{r^3}$, and the computed average gradient increases as a larger spatial extent of the field is considered.

will encounter the magnetopause and so be lost from the magnetosphere - this process is known as magnetopause shadowing, and is one of the dominant loss processes in the outer radiation belt [Turner et al., 2012b; Xiang et al., 2018; Staples et al., 2020].

Other significant processes for electron acceleration and transport are typically related to resonant WPIs. WPIs between seed population electrons and very low frequency (VLF) whistler mode waves, act to energise radiation belt electrons to relativistic energies [Baker et al., 1998a; Horne et al., 1998; Summers et al., 1998; Meredith, 2002; Horne et al., 2005]. Ultra-low frequency (ULF) waves transport electrons through radial diffusion [Fälthammar, 1965; Jaynes et al., 2015; Thompson et al., 2020] and can play a role in electron acceleration through WPIs [Elkington et al., 1999; Mann et al., 2013]. Radial electron transport via ULF wave activity results in betatron acceleration of electrons as electrons are transported radially inward (and *vice versa*). A transport and acceleration method, not linked with WPIs, occurs when the ring current is enhanced, causing electrons to be adiabatically transported outwards and thus decelerate [Dessler and Karplus, 1961; McIlwain, 1966].

1.4.2.4 Storms and Substorms

Geomagnetic storms are periods of intense solar-wind-magnetosphere coupling, during which large-scale reconfiguration of the magnetospheric geometry occurs, driving significant intensification of the ring current [Chapman and Bartels, 1941; Kamide et al., 1998]. Storms are characterised by horizontal reductions in the geomagnetic field, measured by equatorial magnetometer stations, and represented via geomagnetic indices such as Dst [Gonzalez et al., 1994] or SYM-H [Iyemori, 1990; Wanliss and Showalter, 2006]. Moderate storms are characterised by a ring-current-induced reduction in the surface geomagnetic field of $50\,\mathrm{n}T$ [c.f., Earth's typical surface field strength of $25-65\,\mu$T Finlay et al., 2010], which occurs ≈ 8 % of the time [Gonzalez et al., 1994]. Primarily, geomagnetic storms are associated with transient structures in the solar wind such as CMEs or CIRs [Hutchinson et al., 2011].

A prototypical storm consists of three phases, an initial, main and recovery phase. During the initial phase the SYM-H/Dst index increases, due to CME/CIR driven magnetospheric compression increasing the geomagnetic field strength. Subsequently, the main phase is characterised by a large reduction in the geomagnetic field due to enhancement of the ring current [through processes currently unknown, Hutchinson et al., 2011]; this is controlled by a strong southwards interplanetary magnetic field causing increased day-side reconnection, and depositing large amount of solar wind energy into the magnetosphere. The start of the recovery phase is typically linked to the solar wind driving conditions returning to quiescent conditions (reduced solar wind speed and ram pressure, less southwards interplanetary magnetic field etc.) and occurs on timescales of hours to days [Gonzalez et al., 1994; Hutchinson et al., 2011]. The recovery phase is associated with a decay in ring current intensity [Yermolaev et al., 2012; Daglis et al., 1999], possibly via resonant WPIs between bouncing particles and ULF waves [Hutchinson et al., 2011].

Storms are an important aspect of Space Weather research due to the effects they have on the geo-space environment. Whilst many storm-related effects are well understood, the ways they affect radiation belt electron fluxes are not. Reeves et al. [2003] conducted a statistical study of storms, finding that $\approx 50\%$ of storms caused an enhancement of relativistic electrons at geosynchronous orbit, $\approx 20\%$ caused a depletion, and $\approx 30\%$ caused no significant change (within a factor of 2). Anderson et al. [2015] conducted a similar study, but instead examined less geo-effective storms (i.e., a smaller Dst response). Compared to large storms, these smaller storms were found to be 10% less likely to result in a flux enhancement and 10% more likely to result in a flux depletion at geosynchronous orbit. In a similar fashion to Reeves et al. [2003], Turner et al. [2015], Moya et al. [2017] and Turner et al. [2019] used data from NASA's Van Allen Probes mission to investigate the radiation belt electron flux response to storms. Turner et al. [2015] and Turner et al. [2019] found that lower energy electrons (100s of keV) closer to the Earth ($3-4\,\mathrm{R_E}$) were enhanced in almost all ($> 90\%$) of the storms studied. Though both Turner et al. [2015] and Moya et al. [2017] found similar results to Reeves et al. [2003] for the MeV-energy electrons.

In contrast to these results, Murphy et al. [2018] and Murphy et al. [2020] use a proxy for the number of electrons in the outer radiation belt over ranges of different energies and find that the storm main-phase is accompanied by a consistent decrease in electrons, followed by an increase over the recovery phase.

Reeves et al. [2016] generalise the morphological response of radiation belt electrons to storms, later validated by Turner et al. [2019]: in the main phase, the slot region is filled with a new population of source and seed electrons [Jaynes et al., 2015], while the higher energy electrons are lost throughout most if not all of the outer radiation belt; during the recovery phase, the source and seed electrons at lower L-shells are lost, reforming the slot region, whilst typically, the relativistic electrons become enhanced within the outer radiation belt.

Associated with storms are substorms. During a storm there are always substorms, though substorms may also occur outside of storm-times [Gonzalez et al., 1994]. Substorms are a primary, if indirect, source of the energetic electrons comprising the outer radiation belt. Through magnetotail reconnection, substorms inject the low-energy source and seed electron populations into the inner magnetosphere [Akasofu, 1968; Reeves et al., 1990; Borovsky et al., 1993; Forsyth et al., 2016]. The injected seed population is not isotropic and so supports the generation of whistler mode chorus waves [Li et al., 2010b]. The seed population is subsequently locally accelerated through wave-particle interactions with these whistler mode chorus waves up to relativistic energies [Horne et al., 1998; Summers et al., 1998; Horne et al., 2005; Jaynes et al., 2015].

Much like the Reeves et al. [2003] work, investigations into the radiation belt electron response to substorms provide somewhat inconclusive results. Forsyth et al. [2016] found that the chances of an increase or decrease in the content of the radiation belts in response to substorms was 50/50. This tells us that, despite the current paradigm being to discuss Space Weather in terms of storms and substorms, these are unfortunately not the relevant phenomena for the radiation belts. Discovering the next paradigm is clearly necessary, but appears to be remaining elusively out of reach for the time being.

1.5 Space Weather

Space Weather is the umbrella term used to describe the dynamic behaviour in the geo-spatial environment. According to COST (European Cooperation in Science and Technology) Action 724, Space Weather is defined as: "*the physical and phenomenological state of natural space environments. The associated discipline aims, through observation, monitoring, analysis and modelling, at un-*

derstanding and predicting the state of the Sun, the interplanetary and planetary environments, and the solar and non-solar driven perturbations that affect them, and also at forecasting and nowcasting the potential impacts on biological and technological systems". As with terrestrial weather, Space Weather is always occurring, though the focus of Space Weather research typically falls on the more extreme behaviour which can have negative impacts on human life [Pulkkinen, 2007].

Space Weather events can lead to the loss of satellite service [Garrett, 1981; Baker et al., 1987; Frederickson et al., 1991; Baker et al., 1998b; Baker, 2001; Sreeja, 2016], increased radiation danger to astronauts [Cucinotta et al., 2010], plane diversions from polar routes [Eastwood et al., 2017a] and power outages [Pirjola, 2007; Piccinelli and Krausmann, 2014], among others.

Mitigating for these effects requires comprehensive understanding of causal factors, quantitative characterisation of the various levels of effects that may be induced, and ideally, the ability to forecast the effects.

Many impactful Space Weather events are directly linked to the solar wind, specifically, CMEs, flares, solar energetic proton (SEP) events and solar wind stream interactions. Predicting when the Sun might produces flares [Leka et al., 2019], CMEs and SEPs [Klein and Dalla, 2017] is difficult and a current area of research. However, the fact the solar eruptions that produce these phenomena can be observed on the Sun allows forecasting work to focus on the specific arrival times [e.g., for CMEs Riley et al., 2018; Owens et al., 2020; Riley and Ben-Nun, 2021] - an arguably more straightforward task. The maximum lead times for predicting effects due to these processes differs due to the different characteristic speeds they have. The effects of flares are due to the x-rays they produce - these travel at the speed of light and reach Earth in a matter of minutes. SEPs are particles that typically travel at relativistic speeds - these can reach Earth in an hour. CMEs are, by comparison, much slower, travelling at speeds more similar to the solar wind - CMEs typically reach Earth a few days a few days.

However, since the impacts of these events are present within the magnetosphere, it is necessary to be able to map from solar wind conditions to the magnetospheric response. Given the many spacecraft which exist in the radiation belts, it is crucial to understand the plasma environment within which they exist, as such understanding helps to prolong operational lifetimes. Much work has been done to model the radiation belt response to the solar wind [Rigler et al., 2005; Hudson et al., 2008; Turner et al., 2012a; Li et al., 2015; Lugaz et al., 2016; Baker et al., 2019; Kalliokoski et al., 2020; Smirnov et al., 2020; Xiang et al., 2021]. Equivalently, work has also been done to model the auroral oval size [Milan, 2009; Xiong et al., 2014; Xiong and Lühr, 2014; Hu et al., 2017; Han et al., 2020] and ground-induced currents or their related local magnetic perturbations [Wintoft, 2005; Pulkkinen et al., 2013; Camporeale et al., 2020; Keesee et al., 2020] with-respect-to the solar wind.

The work presented in Chapter 4 aims to aid Space Weather research by providing objective, data-driven classifications of solar wind type, providing additional information for future statistical studies in analysing the Space Weather impacts that different solar wind types have.

In Chapter 5 we explore whether we can extract an empirical location for the outer radiation belt boundary. This boundary is crucial in radiation belt modelling, and so an improved identification may lead to improvements in the characterisation of the radiation belt environment.

In Chapter 6, we take similar ideas to those used in solar wind parameterisations of the radiation belts to construct our own parameterisation of the electron flux at the boundary. Whilst the work presented in Chapter 5 is useful in-and-of itself, radiation belt modellers use the boundary electron flux specifically in their boundary condition. Therefore, this chapter serves to demonstrate how a synthetic spacecraft dataset of the electron fluxes can be created using machine learning. Results from this work should make it much easier for radiation belt modellers to include information regarding the new boundary location into their models.

1.6 Summary

This Chapter set out to introducing the relevant background physics required to understand the research projects presented in Chapters 4, 5 and 6.

In Section 1.1, the fundamentals of plasma physics were laid out - from what a plasma is to the equations governing the electromagnetic interactions which dominate the motion. Subsequently, in Section 1.2, the motion of charged particles in electric and magnetic fields was expounded. This motion is crucial to understanding the dynamics of radiation belt particles (e.g., in Chapters 5 and 6). Moving on from a particle description of plasma motions, the fluid theory of magneto-hydrodynamics was explored in Section 1.3. MHD provides a natural framework for exploring the bulk motions of charged fluids (especially useful when considering the solar wind and interactions between the solar wind and a planetary magnetosphere).

Section 1.4 saw a shift from more broad physics concepts to the specific phenomena and morphologies present within the heliospheric and magnetospheric systems. We initially described the solar wind (Section 1.4.1), its generation, basic properties, large scale structures and the solar cycle. Many of these concepts are foundational to the topic presented in Chapter 4, but additionally, there are profound links between the solar wind and Earth's magnetosphere - essentially manifesting the field of Space Weather research. After the solar wind, the narrative became more Earth focused (Section 1.4.2). Initially, the pathway for energy transfer between the solar wind and the magnetosphere was described (Section 1.4.2.1), followed by some broad geography of the relevant (to Chapters 5 and 6, at least) structures in Sections 1.4.2.2 and 1.4.2.3. We finished-off Section 1.4.2 with broad descriptions of the paradigm phenomena of Space Weather - geomagnetic storms and substorms (Section 1.4.2.4), with a focus on their effect on the outer radiation belt electron population (the focus of Chapters 5 and 6).

Chapter 2

Machine Learning

In the context of this thesis, machine learning (ML) is the umbrella term used to describe computational methods of (semi-)automated pattern recognition.

Throughout this chapter we will give a broad overview of supervised and unsupervised machine learning[1] - with further details regarding the algorithms and methods specific to the work presented in Chapters 4, 5 and 6.

2.1 Terminology

Before beginning the more detailed discussion of the algorithms, it is useful to set out the terminology that will be used.

Inputs, **input feature**, and **X** all refer to the data from the parameter space that you want to map *from*. In our specific uses, each individual input is a 1D vector of all of the pertinent data from the input parameter space. Each individual parameter in the input vector, denoted X_i, represents one of measured bits of data.

Outputs, **target feature**, **targets** and **y** all refer to the data from the parameter space that you want to map *to*. Typically, in space-physics and our use-cases, this is a single value or, as seen in Chapter 6, it may also be a vector - these values may be an integer for classification problems or a continuous variable for regression problems.

An **ML model** represents a trained version of an ML algorithm, which is ready to

---

[1] Note that whilst supervised and unsupervised ML are, in my experience, the most frequently discussed, there are other branches of ML too: semi-supervised and reinforcement learning. The latter is what is used to teach a robot to walk, or train a computer to play a video-game etc. The former is a half-way between supervised and unsupervised, wherein the training dataset has a small number of labelled cases and a large number of unlabelled cases - learning may focus on labelling the whole dataset, or using the unlabelled data to bolster the learning of the mapping from inputs to outputs.

be used as an input-output mapping tool.

Model architecture refers to the specific overall structure of a given ML model. In the case of decision trees and random forests, this is the depth of the tree and the rules for creating leaf nodes. In neural networks, it is the specific number of layers, nodes in each layer and method by which information is passed through the network.

Neural network or **NN** serves as a shorthand for artificial neural network (since we aren't literally training other brains to do this work...). It further serves as a shorthand for feed-forward artificial neural networks, as this is the specific architecture we use in Chapter 6.

Hyper-parameters are the parameters which govern the inner workings and architecture of a machine-learning model. These might be the maximum depth of a decision tree, or the number of nodes in a neural network layer. They also include the loss function, optimisation algorithm and learning rates used to train a model.

2.2 Supervised Machine Learning

Supervised learning refers to the type of ML where a chosen algorithm is provided with input-output pairs of data [Russell and Norvig, 2009]. Essentially, supervised ML algorithms are function approximators, trying to optimise the ability to correctly map from the inputs to the outputs. The inputs are samples of the parameter space from which one seeks to model the outputs. For example, as an input to a model you might include the weather (temperature, humidity etc.) and time-of-day, with the aim of predicting the number of ice-creams that will be sold (the output).

There exist too many supervised algorithms to feasibly discuss here, so we will focus this section on decision trees (and their ensemble counterparts) and neural networks, as these are the specific techniques used in Chapters 5 and 6.

2.2.1 Decision Trees

A basic decision tree is a relatively simple algorithm [Hastie et al., 2008]. This class of algorithm is so-named due to the dendrograms used to represent the architecture appearing vaguely similar to a tree[2].

Practically, a decision tree is just a set of nested if/else conditions. These conditions create linear partitions in input parameter space that best separate the data. Typically simple decision trees are no longer used, with researchers favouring more sophisticated methods, however, they have been used applied to identifying noise types in Hubble Space Telescope data [Salzberg et al., 1995]. Note that, despite the linear partitioning, the recursive nature of the algorithm allows decision trees to approximate non-linear functions. The algorithm can be considered in the following way:

1. The model finds the optimum partition value (over all of the input features) to make a single split in the data. The partition is selected to optimise for some criterion[3] in the post-partition regions - frequently, this is related to a loss function such as mean-squared error in regression tasks or the entropy/information gain (related to the distributions of the class labels) in classification problems.
2. After a partition, the data are split between two child nodes (children of the root node). The representative target value for each node is given as either the average (typical in regression) or the majority (in classification) of the target value of all of the data it contains.
3. This process is repeated until either the number of decision ‘layers’ (i.e., the depth of the tree) has reached the specified maximum and/or every node has the minimum allowed number of data samples (1 is the fundamental limit). When a node can no longer be split due to either condition being met, it is considered a ‘leaf’ node; further promoting the tree analogy.

Due to the linear partitions, when decision trees are used for regression, the

---

[2] Though, somewhat amusingly, the dendrograms are almost always shown with the root node - the initial part of the tree - at the top, with the dendrogram spreading downwards.

[3] There are various algorithms for this optimisation, though the most commonly used one is likely CART [classification and regression tree, Breiman et al., 1984], as this is the default in the Scikit-learn python package. Others, such as ID3 [Iterative Dichotomiser 3, Quinlan, 1986] or C4.5 [Quinlan, 1993] are also widely used (though mostly the latter, as it is an update to the former).

regression surface is not smooth and continuous, it is step-like, with discrete changes in the regression values.

In decision-tree learning, the depth and the number of samples per leaf are the most important hyper-parameters. They effectively control whether a model will over- or under-fit the data. Taking some extreme examples of these we find:

- If the tree depth is fixed at 1, so there is a single split in the data, then the model can only generalise a very simple binary classification, otherwise it will heavily under-fit the data. On its own, a decision tree stump model is rarely useful or informative[4].
- If, in contrast, the tree is allowed to grow such that each leaf node contains a single data-sample, then the model will typically be over-fitting (see Section 2.2.4), as it won't have learnt to generalise the data well.

These issues are typically overcome by optimising for these hyper-parameters, allowing for the optimum balance between over- and under-fitting to be determined by empirical testing of different values.

2.2.2 Ensemble Decision Trees

Trees are very fast to train due to the algorithms used to make the splits, and typically easy to explain and understand. However, individual trees do not perform machine-learning tasks as well as other algorithms, and they aren't robust to changes in training data [James et al., 2013]. To this end, there has been a lot of work on ensemble tree methods. These make good use of the speed of decision-tree training, to create large, robust ensembles which have significantly better performance than single trees and other machine-learning algorithms [Engelhardt et al., 2014; Gredell et al., 2019; Uddin et al., 2019]. Note, we do not comment on over-fitting here, as we will make arguments in Section 2.2.4 for it not being an algorithm problem, but instead being a failing of the scientist.

The most straightforward ensemble method is known as 'bagging,' which we will

---

[4] Though, when used as part of more complex ensemble methods, they can be put to good use [e.g., these examples from other fields: Kawaguchi and Nishii, 2007; Chen et al., 2019].

describe first, before covering random forests[5] and boosting. These topics are covered in detail in the following textbooks, from which we draw much of the material: James et al. [2013]; Hastie et al. [2008]

'Bagging' is a portmanteau of bootstrap aggregating, and is a term used when training an ensemble using subsets of the training dataset rather than the full training dataset in one big chunk. The motivation for bagging originates from calculating sample statistics: given a set of n independent observations of some population statistic, y, which have variance σ^2, the variance associated with the mean value, $\bar{y}$ is given by σ^2/n. This means that by averaging our observations, we can reduce the variance. Equivalently, if n independent decision trees are averaged, the variance in the predicted value is reduced. The immediate problem with this is that decision trees are typically not stochastic, and so the trees will not be independent of one another. Hence the use of bagging[6] - by training many trees with subsets of the training data - each tree is effectively de-correlated from the others and can be considered independent. In the context of classification, one would take a majority classification from the ensemble and use that as the predicted class for a given data-sample.

Random forests build from the concept of bagging, making a small adjustment to the algorithm to further de-correlate the trees: for each node, the algorithm determining the optimum partition is given a random sub-sample of the input features. By making this small change, the propensity of trees to focus on the same input features (i.e., the most informative ones) is removed, forcing more variability in the decision trees. Random forests have been used to optimise the multi-dimensional partitioning of solar wind data to better predict the power of certain magnetospheric waves [Bentley et al., 2020].

Unlike bagging and random forests, boosting does not use subsets of the data for training ensemble members. Instead, each tree is 'grown' sequentially and applied additively, focusing on reducing the errors from the previous trees [Friedman, 2001, 2002; James et al., 2013]. As an additive model, the ensemble is not averaged like in bagging or random forests. The general algorithm, MART [Multiple Additive Regression Trees, Friedman, 2001, 2002], proceeds as follows:

1. An initial prediction, y', of 0 is given for all training samples, X: $y'_0 = f_0(X) = 0$.

---

[5] Very committed to the tree analogy now.

[6] Ideally one would use n independent training datasets, one for each tree. But alas, we very rarely have enough data to construct our models this way. It's likely that bagging the full dataset would still be more effective anyway, since there would be more training data.

And, a model weighting, λ, is specified (typically 0.01-0.00001).

2. The residuals, r, between the true values, y, minus the predictions are calculated: $r = y - y'$.
3. A decision tree model, $f_n^{DT}(X)$, is fit to predict the residuals: $r_n \approx f_n^{DT}(X)$.
4. The initial predictions are added with the weighted output of the decision tree to give a new y': $y'_{n+1} = y'_0 + \lambda f_n^{DT}(X)$
5. The residuals are then recalculated: $r_{n+1} = y - y'_{n+1}$.
6. Steps 3 - 5 are repeated until the specified number of trees have been trained, or the model performs sufficiently well. The final boosted model, f^B, is then given by: $f^B(X) = \sum_{n=0}^{n=N} \lambda f_n^{DT}(X) \approx f(X)$. Where N is the total number of trees to be trained, and $f(X)$ is the underlying mapping that the model is being taught to approximate.

In this way, an ensemble grown through gradient boosting is specialised to reduce the residuals between predictions - providing a model with good performance, even with very small trees [James et al., 2013]. The downside is that when the ensemble becomes very large, the trees added later in the process don't have an impact beyond the residuals relating to a few data-samples [Freund, 2001; Rashmi and Gilad-Bachrach, 2015].

To address this downside, alterations can be made to the general algorithm. Specifically, DART [Dropouts meet MART, Rashmi and Gilad-Bachrach, 2015] has been proposed and shown to significantly improve performance. The primary difference is that DART calculates the updated residual based on a probabilistic subset of the previous trees. Then, to compensate for training trees using a subset residual, the new tree is normalised, and the dropped trees are scaled such the the overall effect of the dropped and new trees is the same as the dropped trees' initial effect. This method is implemented in Chapter 5 and Bloch et al. [2021].

As explained in Rashmi and Gilad-Bachrach [2015], DART is effectively a balance between random forests (all trees being dropped at each iteration) and MART (no tree being dropped). This allows, with an appropriately selected dropout probability, each tree to have optimum significance in the final model - allowing the final model to perform significantly better than either MART or random forests.

2.2.3 Neural Networks

Here we will give a high level introduction to feed-forward neural networks (FFNNs), though we recommend Gurney [1997] for a comprehensive introduction to the subject. We won't discuss other network architectures, such as convolutional (CNN) or recurrent (RNN) neural networks, as they do not contribute to the work presented in Chapter 6.

Simple neural networks are trained to approximate a single function. To do so, a network architecture is chosen, which can be represented as a computational graph. Where a computation graph is constructed out of two types of building blocks: nodes and edges (edges connect nodes). Nodes can be considered in two ways: those that represent input values, and those that represent a function for combining input values. The input values can either be a vector of raw data passed to the model, or the values from previous nodes in the network being propagated through the network along edges (with associated weightings). Figure 2.1 presents a schematic of a neural network. The nodes typically perform the following to map between its input to its output (show visually in the blown up node in Figure 2.1):

1. The inputs (previous node outputs which have been weighted) and bias term (typically equal to 1) are summed together.
2. The summed inputs are passed to a non-linear activation function (e.g., a sigmoid or hyperbolic tangent function).
3. The output from the activation function is provided as the node output.

For linearly separable problems, only a single node is required to map from the inputs to the outputs. That is to say, if the function being approximated is composed of a linear superposition of the inputs, then the network only needs learn the weightings (no activation function is required) such that the summation of weighted inputs gives the output.

However, as problems become more complicated, a single node becomes insufficient to separate the data. To address this, more nodes can be added to the computational graph, allowing the model to approximate non-linear functions more accurately. When additional nodes are added between the input and output layers, they are referred to as 'hidden layers'. Theoretically, a network with a

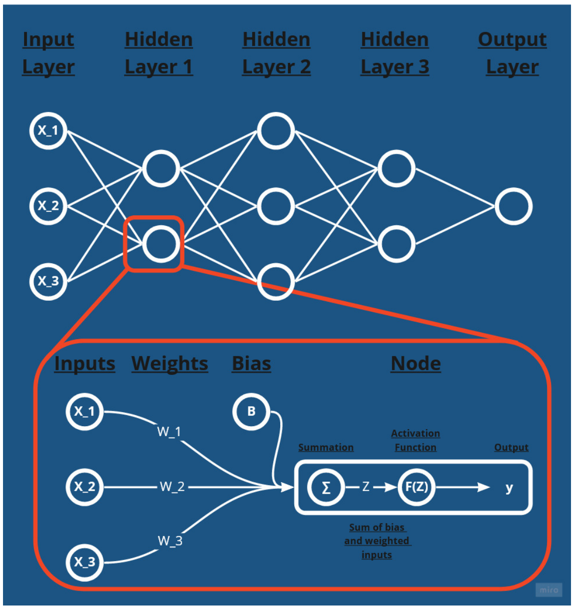


Figure 2.1: *A schematic diagram of a simple neural network, with an expanded look at processing through a network node.*

single hidden layer is all that is required to approximate any non-linear function [Cybenko, 1989; Hornik, 1991] - begging the question "why does anyone use deep learning?", where 'deep-learning' refers to networks with multiple hidden layers. Whilst, theoretically, a single layer is enough, in practice this layer may require up-to an infinite number of nodes to approximate the function. Deep learning has developed as a computationally efficient method of approximating functions that may be intractable with only a single hidden layer. So, the reason that deep-learning is so widely used is that, in many situations, it can better approximate functions which are computationally intractable to approximate with a single hidden layer.

So far, our description covers only the graph representation of a network and how data are propagated across. This gives us the framework to explain how one trains a NN.

The first step is to choose a network architecture. In theory, this should be a 'simple' case of making assumptions about how complicated the mapping between inputs and outputs is, the format of the inputs and outputs, and which non-linear activation function to use. But in practice, it is most straightforward to consider the architecture as a set of hyper-parameters to be optimised for.

One may make certain inferences about the architecture based on the type of mapping being undertaken. Some examples of this include using CNNs for image-type data, RNNs for time-series data, and FFNNs for mapping independent input-output samples. Though, in many cases, CNN and RNN frameworks often also include FFNNs, and 1D CNNs can also be used for time-series data.

As mentioned previously, there are weights associated with each of the edges in the network graph. In a NN model, these weights are free parameters for the model to optimise during training. The weights are initialised randomly; typically sampled from the uniform distribution with a standard deviation of $\sqrt{1/\mathrm{N_{input}}}$ [Montavon et al., 2012] or bounds [Glorot and Bengio, 2010] given by $\pm\sqrt{1/\mathrm{N_{input}}}$, where $\mathrm{N_{input}}$ is the number of inputs to the layer. The weight distribution can also be optimised, though this is less common.

Training is typically performed in epochs, where one epoch represents a single pass of the whole training dataset through the network. After each epoch, the model performance is quantified by a loss function (commonly the mean-square error in regression tasks). Then an optimisation algorithm is implemented to minimise the loss with-respect-to each edge weight in the network. This process is repeated, updating the weights until the loss has been reduced to the requisite amount or until a sufficient number of training epochs have been performed (‘sufficient’ is commonly taken to be the number of epochs required for the loss to asymptote).

2.2.3.1 Optimisation Algorithms

Optimisation algorithms utilise gradient descent to tweak the network weights and biases. Gradient descent is a method used to find the minima of a differentiable function [Cauchy, 1847; Nelder and Mead, 1965; Lemaréchal, 2012]. Simplistically, gradient descent can be described as follows for the 1D case:

1. Start with a function to minimise. The function should continuously and differentiably map from some input space, X to the output space, y.
2. Choose an initial random point, X_n, in the input space.
3. Randomly select two points in the input space with some distance λ between them and X_n: $X_a = X_n + \lambda$ and $X_b = X_n - \lambda$.

4. Calculate the gradient between the two points to approximate the gradient at X_n: $\nabla f(X_n) = \frac{f(X_b)-f(X_a)}{X_b-X_a}$.
5. Choose a new point X_{n+1} in the space some distance from X_n in the direction of most-decreasing gradient: $X_{n+1} = X_n + \eta \nabla f(X_n)$, where η is the step size.
6. Repeat steps 3-5 until the (local) minimum is reached, to some tolerance level.

This process can be extended to higher dimensions by creating a local hypersurface about X_n, and calculating the direction of most-decreasing gradient.

A downside to this approach is that by using an inappropriate step size the algorithm may not converge or may be inefficient. If the step size is too small, then convergence will take too many iterations. If it is too large, then the algorithm may diverge from or oscillate about the true minimum. Early ideas suggested that a solution to this would be to choose a sub-optimal direction (i.e., add some perturbation angle, θ) and allow for better convergence over a larger number of iterations [Wolfe, 1969].

More recently, stochastic gradient descent has come to the forefront of machine learning as a more efficient approach to gradient descent [Bottou, 1991; Bach and Moulines, 2011]. This stochastic approximation method [Robbins and Monro, 1951], is effectively the same as gradient descent, but uses singular random samples of the data each iteration to calculate the losses and gradients. In this way, optimisation requires more iterations to converge, but each iteration is both significantly faster, and also feasible for very large datasets which can't be held in memory. Between the extremes of using all of the data and a single sample is mini-batch gradient descent. This utilises random subsets of the data to calculate the loss and gradients. By using mini-batches of the data, the variance of the estimated gradient is reduced, leading to smoother convergence to the optimal solution [Khirirat et al., 2017].

There are many further adaptions to these algorithms, as well as other implementations [Nwankpa, 2020]. In the following we will briefly summarise those used in Chapter 6. ASDG [averaged stochastic gradient descent, Polyak and Juditsky, 1992] updates the weights using the average of the weights over previous iterations of the algorithm, making the optimisation smoother. Adam [adaptive moment estimation, Kingma and Ba, 2014] is an update to RMSprop. Where the RMSprop [Graves, 2013] performs stochastic gradient descent, but weights the learning rate based on the running average of the gradients from previous iterations - meaning the weights are changed less when the gradients are strong. Adam builds on this by also weighting the learning rate by the running variance of the gradients. AdamW [Adam with weight decay, Loshchilov and Hutter, 2017] builds yet further, changing the algorithm so that the updates to the weights become smaller over iterations.

Neural networks are likely the most abundantly used supervised ML technique in Space Physics, hence there are simply too many papers to reference them all, though non-exhaustive lists can be found within: Camporeale et al. [2018]; Camporeale [2019]; Smirnov et al. [2020]; Azari et al. [2020].

2.2.4 Over-fitting

Over-fitting is a common issue researchers face when creating machine-learning models and it is frequently discussed as a model failing.

What is over-fitting? In general terms, over-fitting describes the situation when a model is fit too closely to some data such that it loses generalisability. In practical terms, over-fitting arises when either the training set is poorly chosen, or when there is too much noise in the data. In either case, fitting too closely to poor data naturally leads to a poor model - i.e., ‘rubbish in, rubbish out’ [The Hammond Indiana Times, 1957].

Perhaps pedantically, it should be clarified that in many cases, over-fitting is precisely what algorithms are designed to do. The problem arises instead, when researchers try to balance accurate fitting of the model to data with fitting a model that can generalise well to out-of-sample data.

For example, a decision tree can be trained such that every training sample terminates in its own unique leaf node. Such a model is exceptionally well-fit to the data. In practice, many of the subsequent uses of such a decision tree revolve around applying it to other data. When the perfectly fit tree is applied to unseen data it will typically perform poorly, with high variance. So despite the tree perfectly performing what it was programmatically designed for, its subsequent application results in a ‘failure’.

To counter the fact that the algorithm, by definition, frequently performs its intended purpose ‘too well’, new adaptions to the model must be included. In the case of trees, these adaptions are primarily limits to the maximum depth of the tree and the minimum number of samples a node can partition. This raises questions about how to choose such parameters.

These choices are made through hyper-parameter optimisation - the process in which many possible hyper-parameter combinations are trialed. By tracking the performance of the model on an independent validation dataset, the approximately[7] optimum parameters can be selected to be used in the final model.

In neural networks, which are iteratively trained (unlike e.g., decision trees), another compounding factor is that the number of training epochs must be limited to prevent over-fitting. To account for this, a concept known as ‘early stopping’ was introduced. This boils down to tracking the model performance on the training dataset and an independent validation dataset for each training epoch. When the loss functions begin to diverge, as shown in Figure 2.2, it is indicative of the model is beginning to over-fit and lose generalisability.

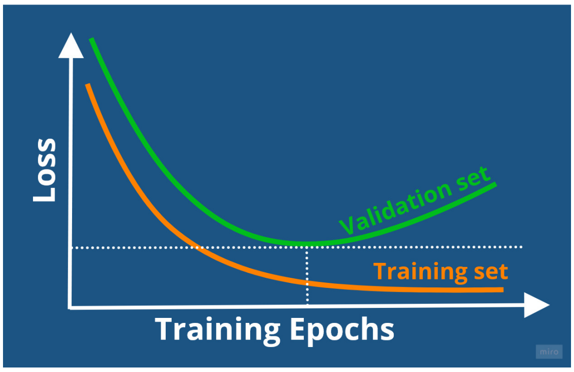


Figure 2.2: *An example of the training and validation loss curves showing the optimum training epoch to halt neural network training, so as to avoid over-fitting the model.*

In general, there are many techniques to check for and reduce over-fitting. As such, when over-fitting becomes problematic in a machine learning experiment, it is invariably due to poor experiment design.

---

[7] It’s often intractable to try every combination of hyper-parameters. It isn’t uncommon to be optimising for tens of parameters, each with five or more possible values, meaning that to fully sample the hyper-parameter space 5^{10} (ten million) models would have to be trained

2.3 Unsupervised Machine learning

In contrast to supervised learning, which requires labelled datasets to learn a functional mapping, unsupervised learning does not require labelled data, instead being designed to learn latent spaces and their structure. Broadly, unsupervised learning falls into two categories: clustering and dimension reduction.

Dimension reduction seeks to take a set of extant input parameters and map them to a space comprised of fewer, latent variables which can be used to adequately represent the original data.

Clustering refers to the techniques which can be used to identify groupings of data in either an extant or latent space. In this way, clustering algorithms can identify similarities between groups of data-samples, and label the data according to the groupings.

2.3.1 Dimension Reduction

Dimension reduction has a diverse set of uses:

- In research, it is often useful to identify which subset of measured or derived variables are most important to use in an empirical model. Often, these variables may be selected using theoretical deduction or a scientist making a (black-box) decision. But neither of these methods necessarily provide the optimal set of descriptive parameters. Instead, using dimension reduction techniques such as PCA (principle component analysis) the subset of parameters that best describe the data can be identified.
- When dealing with very high dimensional datasets, dimension reduction can be used to reduce the dataset down to a more manageable size - both in terms of physical storage size of the data, and for human interpretation (e.g., for visualisation).
- The multi-faceted 'curse of dimensionality' also becomes a problem in high dimensional spaces [Aggarwal et al., 2001; Domingos, 2012]. Some of the issues include: for a fixed dataset size, the parameter space is increasingly

sparsely sampled as the dimensions increase, making empirical learning difficult; the relative difference in euclidean distance between closest and furthest points tends towards 1 as the dimensionality increases; too many noisy features can swamp the signal from a few good features. Given these issues, the possible benefits of dimension reduction become self-evident.

Dimension reduction algorithms typically work in one of a few ways: by making assumptions about what a reduced dimensional space should 'look' like (e.g., PCA, Figure 2.3); by trying to maximise the distances between known clusters of data (e.g., linear discriminant analysis, LDA, Figure 2.4); retaining distance-based structures in the data (e.g., t-distributed stochastic neighbour embedding, t-SNE); or by making assumptions about the high-dimensional manifold (e.g., UMAP).

2.3.1.1 Principle Component Analysis

PCA (principle component analysis) is one of the most widely-used methods of dimension reduction. In a magnetospheric context, PCA has been used to study the ionospheric current systems [Milan et al., 2015; Alken et al., 2017]. Technically, PCA is actually a transformation algorithm. It maps from m-dimensional input space to an orthogonal m-dimensional output space with uncorrelated dimensions. PCA sequentially calculates the new dimensions' directions by finding the directions of maximum variance (see Figure 2.3). The final output of PCA is then a set of mappings from the input space ordered such that each subsequent mapping maps to a direction which contains less variance than those previous.

Note that if a direction contains more variance, then it means that there is the most co-variation in the original input parameters in that direction. This is useful, because it allows many correlated parameters to be broadly captured in a single new parameter (i.e., the PCA directions).

PCA becomes dimension reduction when only a subset of the directions (with most variance) are selected, rather than the full decomposition. This can either by an *a priori* choice of a fixed number of the most important directions, or *a posteriori* by calculating the variance in each direction and selecting some threshold of acceptable captured variance.

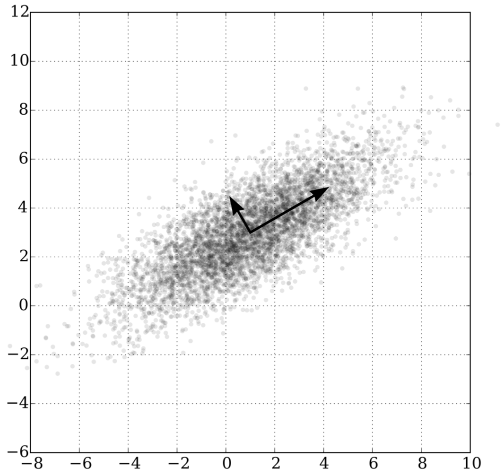


Figure 2.3: *An example of applying PCA to a multivariate Gaussian distribution. Initially the data are distributed in x and y, but the direction of largest variance is the long black arrow, followed by the smaller black arrow. PCA learns these directions and the linear transform required to map between the two coordinate systems. Image credit: Wikipedia user Nicoguaro.*

2.3.1.2 Linear Discriminant Analysis

An example of applying LDA to two clusters of data are presented in Figure 2.4. Initially the data are distributed in x and y (axes not shown here), but the direction LD1 allows for the maximum separation of the two Gaussian distributions (one for each cluster). LD2 is an example of another direction LDA might test, but ultimately discard in favour of LD1. For cluster analysis, the original 2D dataset can be represented singularly by distances along LD1, thus reducing the dimensions required to characterise the clusters. LDA is less widely used than PCA in Space Physics research, but it has been implemented to better understand the dynamic location of dayside reconnection [Hoshi et al., 2018].

2.3.1.3 UMAP and t-SNE

A comparative example of t-SNE [van der Maaten and Hinton, 2008] and UMAP [McInnes et al., 2018] as applied to the MNIST dataset is presented in Figure 2.5. On the left-hand side of Figure 2.5 is an example of the dataset. The MNIST dataset is comprised of 28×28 pixel images of hand-written digits, which can be

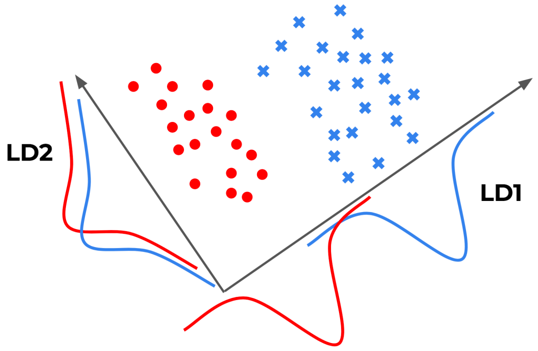
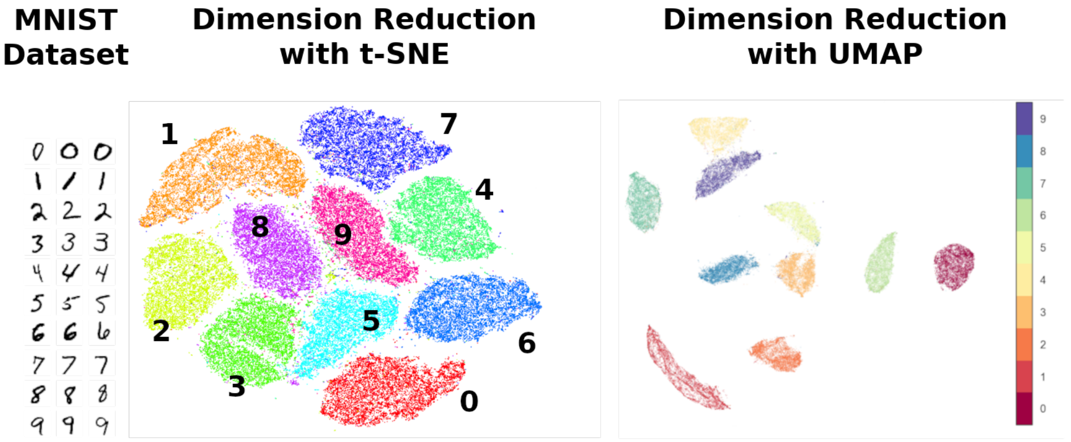


Figure 2.4: *A diagram depicting the axis (LD1) LDA finds to best separate the two classes (red dots, and blue crosses), as well as a secondary example direction the algorithm may consider (LD2). The Gaussian distributions are shown as the separation is maximised based on an assumed Gaussian distribution of data. Image credit: George Ho.*

decomposed as a 784 dimensional vector. To the right of the example dataset are two representations of the MNIST dataset in 2D - one by UMAP and the other by t-SNE. As can be seen in the figure, both UMAP and t-SNE can reduce this 784 dimensional vector down to 2 dimensions whilst still capturing the discrete digits. UMAP performs a comparatively better reduction, as inferred from the more discrete clusters (the clusters are more well-separated from each other). UMAP also retains more of the structure within each of the clusters than t-SNE does [as discussed in McInnes et al., 2018]. UMAP is described in more detail in Chapter 4.


MNIST Dataset

Dimension Reduction with t-SNE

Dimension Reduction with UMAP

Figure 2.5: *An example of applying t-SNE and UMAP to the MNIST dataset of hand written digits. MNIST image credit: Josef Steppan. T-SNE image credit: Van Der Maaten [2015]. UMAP image credit: Leland McInnes.*

2.3.2 Clustering

Clustering, as the name suggests, is primarily used to identify groupings in data. The main tenet of clustering is that within a cluster the samples of data should be more similar to each other, than to samples from other clusters.

Clustering is typically applied to the following scenarios:

- When groupings want to be found in high-dimensional space, which can't easily be visualised by a researcher.
- When groupings want to be found mathematically or free of researcher biases[8].
- When groupings in the data need to be identified automatically - e.g., in image segmentation, where manual processing is intractable.

The nature of clusters and clustering algorithms is diverse, and thus difficult to categorise [Estivill-Castro, 2002]. For the sake of discussion, we will focus on three types of clustering: distribution models; density models; and hierarchical models.

2.3.2.1 Distribution Models

Distribution models assume that a particular type of distribution can be used to generalise each of the clusters in the data. Gaussian mixture models are one of the foremost of this model type. Gaussian mixtures are typically fit using expectation maximisation, shown in Figure 2.6 and proceeding algorithmically as follows, for some dataset X [Bishop, 2006]:

1. K Gaussians are initialised, with random mean positions, $\mu^{(k)}$ within the space, and unitary covariance and weights - $\sigma^{(k)} = 1 = w^{(k)}$. Note, k refers to one of the Gaussians in the mixture - $k \in \{0, ..., K\}$.

---

[8] Not that it isn't possible to still cherry-pick the data and the algorithm.

2. The posterior probability of each point having been generated by each Gaussian is calculated: $p_{i,j}^{(k)} = p(x_j|\mu_i^{(k)}, \sigma_i^{(k)}, w_i^{(k)}) = w_i^{(k)}\mathcal{N}(x_j|\mu_i^{(k)}, \sigma_i^{(k)})$, where $\mathcal{N}$ represents a normal distribution, subscript i refers to the distribution iteration number, and subscript j refers to the j^{th} sample of data.
3. The normalised probability of each point belonging to each Gaussian is calculated: $\rho_{i,j}^{(k)} = \frac{p_{i,j}^{(k)}}{\sum_{k=0}^{K} p_{i,j}^{(k)}}$.
4. Each mean position is updated as the probability-weighted mean of the dataset: $\mu_{i+1}^{(k)} = \frac{1}{N_i^{(k)}} \sum_{j=0}^{N} \rho_{i,j}^{(k)} \cdot x_j$. Where N is the number of samples in X and $N_i^{(k)}$ is the effective number of samples associated with the k^{th} Gaussian, given by: $N_i^{(k)} = \sum_{j=0}^{N} \rho_{i,j}^{(k)}$.
5. Each covariance is updated as the probability-weighted covariance of the dataset: $\sigma_{i+1}^{(k)} = \frac{1}{N_i^{(k)}} \sum_{j=0}^{N} \rho_{i,j}^{(k)} \cdot (x_j - \mu_{i+1}^{(k)})(x_j - \mu_{i+1}^{(k)})^T$.
6. Each weight is updated as the relative contribution of each Gaussian to the mixture: $w_{i+1}^{(k)} = \frac{N_i^{(k)}}{N}$.
7. Steps 2-6 are repeated until some convergence criteria is reached - e.g., the mean position changes by less than ϵ etc.
8. Classifications are then either given probabilistically (one probability for each Gaussian) or deterministically to the Gaussian with the largest probability of generating the point.

This process is somewhat similar to the k-means clustering algorithm [though slower, Bishop, 2006]. K-means is just a special case of a Gaussian mixture model, but with a different objective function - as explained in Chapter 4.5.

Mixture models are useful when the underlying distributions can be assumed *a priori*. They provide a statistical description of the data, meaning that higher-order moments can be calculated etc. If one is searching for an efficient way to summarise a large dataset, then a Gaussian mixture model is very effective as it can represent a large dataset with only $3 \times K$ variables (one mean, covariance and weighting for each component), allowing an arbitrarily large dataset to be stored with fixed size.

The down-side to mixture models is that often the distribution cannot just be assumed, meaning that the mixture may inherently be flawed from the outset. The process of fitting a mixture can be more costly than other methods (compare

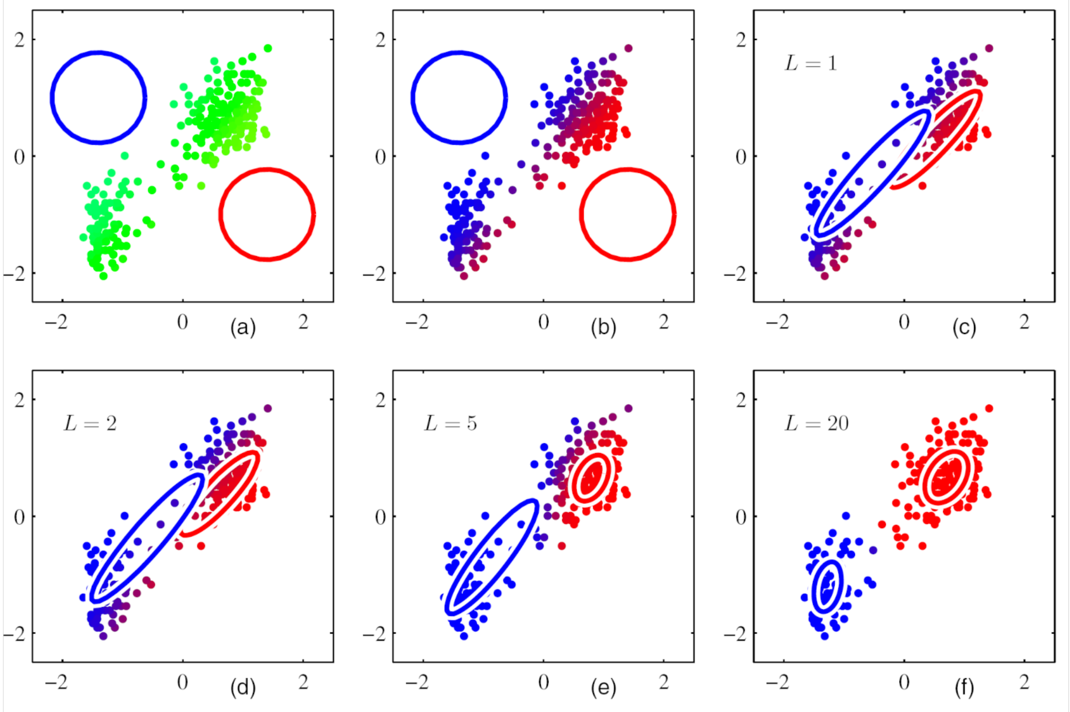


Figure 2.6: *An example of fitting a Gaussian mixture to some data. Panel (a) presents the setup, the dataset is represented by green data-points, and two component Gaussians are represented by the red and blue circles (the circles are the unitary covariance). In panel (b) steps 2 and 3 have been performed - the colour represents the normalised probability of each point belonging to each Gaussian. In panel (c) steps 4, 5 and 6 have been performed, yielding new mean positions, covariances and weights (though the latter is not a visible change). Then in panels (d)-(f) the iterative process has been repeated L times, showing how the mixture ends up fitting the data. Image credit: Bishop [2006].*

k-means to a Gaussian mixture, for example). For many distributions (e.g., Gaussian) the mixture can inherently only find convex (circular or elliptical) clusters, which causes issues when clusters have more exotic shapes. These methods also require the user to specify the number of clusters to find *a priori*, which in many cases precludes a completely data-driven clustering approach.

If the number of clusters is not known, one method is to fit a Gaussian mixture with significantly more components than the expected number of clusters. Then, evaluate the weightings of each Gaussian and drop those below some threshold, leaving only the most significant components, and thus, clusters.

2.3.2.2 Density Models

Density-based models work with the assumption that clusters in the parameter space can be represented by contiguous regions of high density, where density refers to the number of data-samples in some defined region.

The benefit to a density-based model is that it isn't limited to convex clusters, instead being able to form complex cluster shapes (even clusters within clusters if hyper-parameters are appropriately tuned). Additionally, the number of clusters does not have to be specified.

DBSCAN [Density-based spatial clustering of applications with noise, Ester et al., 1996] is an award winning algorithm, receiving the test of time award (an award given to algorithms which have received substantial attention in theory and practice) at the data mining conference, ACM SIGKDD. Additional benefits of this algorithm are: it only has two primary hyper-parameters (meaning that a parameter search is tractable), and it can identify outlier points, rather than incorporating them into cluster descriptions. There are draw-backs too: the algorithm is not deterministic [though there are variations that are, Campello et al., 2013; McInnes and Healy, 2017]; due to relying on distance metrics, it can suffer from the 'curse of dimensionality'; and, variable density clusters often fail to be appropriately captured due to fixed region sizes.

DBSCAN (Figure 2.7) works as follows:

1. Choose a neighbourhood size, ϵ, and the minimum number of points, m.
2. Pick a random point, x_i.
3. Count the number of other points, N, within a distance of ϵ from x_i.
4. If the the number of points is greater than the minimum number of points, $N \geq m$, label the point a 'core' point.
5. Repeat steps 2 and 3 for the remaining unconnected points.
6. After all the core points are identified, all of the 'reachable' points are identified (this can happen during step 5 if a core point is being examined). Where reachable points are those whose ϵ neighbourhood has a non-zero population less than the minimum, $m > N > 0$, but still contain another core point.
7. If any point is not reachable from a core point, label the point as noise.

8. Clusters are identified by reachable sets of core points and reachable points (there can be a size threshold).

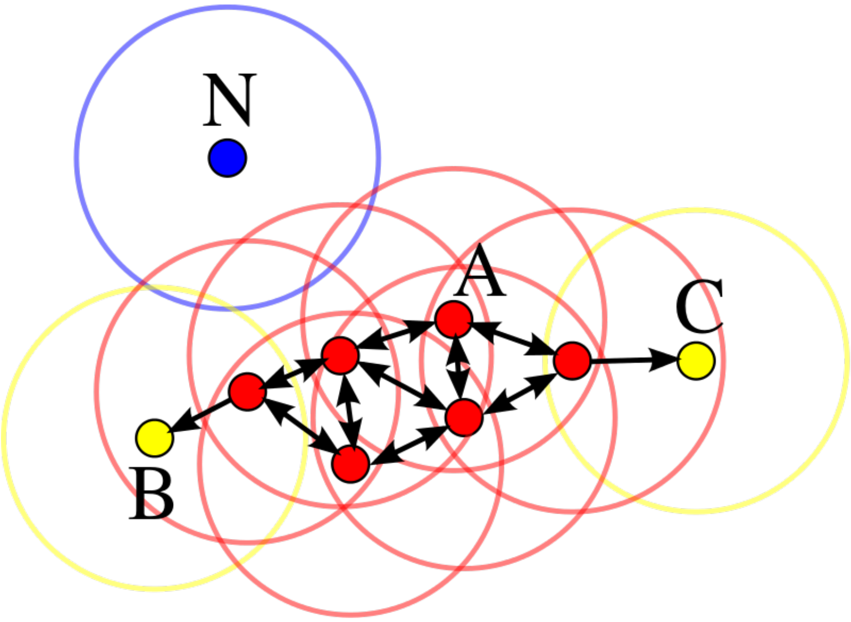


Figure 2.7: *An example of applying DBSCAN to some data (the coloured dots). The neighbourhood size, ϵ, is represented by the circles, and the minimum size, m, in this example is 3. The red points each have at least 3 other points within their respective circles, making them all core points. All of the core points are reachable from each other, as represented by the double-headed arrows. The yellow points (B) and (C) are only reachable from the set of core points, but do not have enough neighbouring points to be labelled as core points. The blue point (N) has no neighbouring points (reachable or core) and so is labelled as noise. Image credit: Wikipedia user Chire.*

The reason that DBSCAN performs less well when clusters have different densities is due to the fixed neighbourhood size. To be able to accurately capture small, dense clusters, a small neighbourhood is required - otherwise the cluster may include points which should be classified as noise. However, if there is a sparse cluster, then too small a neighbourhood will yield an incomplete cluster, or the cluster will be labelled as noise. These requirements are naturally at odds when a dataset has variable density clusters.

2.3.2.3 Hierarchical Models

Hierarchical models aren't so different to density models, though they typically focus on distance metrics rather than density metrics. There are two approaches to building hierarchical models: bottom up (agglomerative), and top down (divisive). In agglomerative models, each data-sample starts in its own cluster, subsequently being merged up the hierarchy until all the data are in a single cluster. Contrastingly, divisive models begin with all the data in a single cluster, subsequently being split down the hierarchy until all the data-samples are in their own cluster.

Figure 2.8 presents an example dataset, distributed in 2 dimensions, that one may want to cluster.

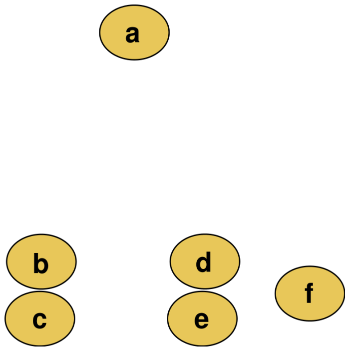


Figure 2.8: *An example distribution of data in a 2-dimensional space. Image credit: Wikipedia user Snubcube.*

Applying agglomerative clustering[9] to the data in Figure 2.8 would proceed as follows:

1. Assign each sample (ellipse with a letter) to its own cluster.
2. Calculate the distances between the points (we will assume Euclidean distance as our metric for this example).
3. Join pairs of samples which are closest. In this case *b* and *c* are equally close to each other as *d* and *e* are - so at this stage both these pairs are merged. By only combining the closest samples each iteration, the hierarchy is stratified by the cluster distances.
4. Re-define our points to reflect the clusters - e.g., our example data would now be {*a*, *bc*, *de*, *f*}.
5. Calculate the distances between the new points. Now, however, there is an added complication - 'what distance should one use for the already-clustered data?'. This gives rise to different linkage criteria (which we will expand upon below). For this example, we will consider single-linkage, which takes the

---

[9] Divisive clustering is performed in an inverse way.

minimum inter-cluster distance, and uses that to determine what should be combined.

6. Repeat steps 3-5 until all the data are in a single cluster or some other specified lower limit on the number of clusters, or until the clusters are a sufficient distance apart,

Following this procedure would yield the dendrogram presented in Figure 2.9.

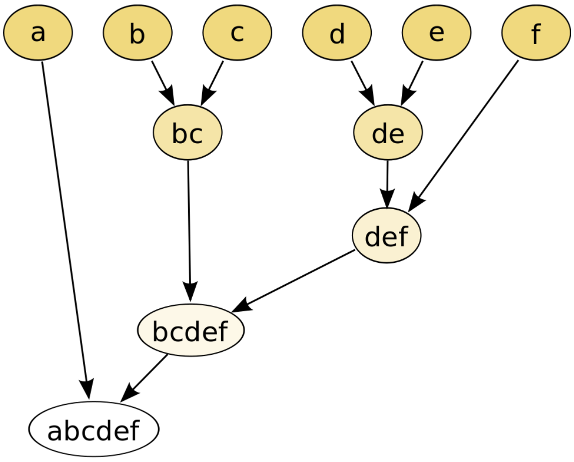


Figure 2.9: *An example of a dendrogram created by applying agglomerative clustering to the data presented in Figure 2.8. Image credit: Wikipedia user Mhbrugman.*

Because the constructed tree is stratified by the minimum distances at each level, it allows the user to examine the result of different minimum cluster distances on the overall clustering. That is to say, one can take a horizontal slice though one of the levels of the tree shown in Figure 2.9, and examine the clusters present directly above. Say for example, we took a slice about the level of the *def* cluster. We would find for that particular distance that the clusters a, bc and *def* were pertinent.

Being able to take slices of the dendrogram and view the clusters present at that level allows for clusters of various scale sizes to be identified (so long as the inter-cluster distances are great enough.

As mentioned above, there are various possible linkage criteria:

- Single linkage: the minimum inter-cluster distance (between all of the points in each cluster) is used as the distance between the clusters.
- Complete linkage: the maximum inter-cluster distance (between all of the points in each cluster) is used as the distance between the clusters.

- Average linkage: the mean inter-cluster distance (between all of the points in each cluster) is used as the distance between the clusters.
- Ward's method [Ward, 1963]: the increase in variance from merging two clusters is used as the distance between the clusters.

One can think of single linkage as the least conservative way of clustering; as long as two points (one from each cluster) are close together, the two clusters will be merged - regardless of the rest of the shapes of the two clusters. Complete linkage is then much more conservative in requiring that all points be closer than some maximum distance. Average linkage gives some balance between the two. Ward's method is a bit of a step in a different direction. It requires *a priori* knowledge of the distributions of the clusters being sought to determine if it may be the most appropriate method to use. In cases where the clusters are normally distributed (or even just convex), it is likely applicable to use Ward's method. But, in situations with more complex cluster shapes, it's likely that the variance would yield incorrect clusters.

Possibly the most useful application of hierarchical clustering algorithms is to combine them with density-based models - as has been done with the OPTICS [Kriegel et al., 2011] and HDBSCAN [Campello et al., 2013; McInnes and Healy, 2017] algorithms. These algorithms don't use a fixed neighbourhood distance (ϵ), instead forming the hierarchy over varying neighbourhood sizes. This significantly improves the ability of these algorithms to handle clusters with different densities. However, OPTICS only allows for a horizontal slice through the dendrogram, making the resultant clusters less robust if there are differences in the densities [Campello et al., 2013]. In contrast, HDBSCAN allows for variable level slicing, based on the robustness of the present clusters. It is this robustness in variable-density clustering which motivated its use in Chapter 4.

2.4 Summary

This section set out to introduce the relevant background machine learning required to understand the research projects presented in Chapters 4, 5 and 6.

In Section 2.2, we described supervised machine learning, and how it is used to approximate the latent mapping between examples of input-output pairs of data. We subsequently focused on the specific methods pertinent to Chapters 5 and 6 - namely, decision trees (and their ensemble methods) and neural networks. As well as a short discussion on over-fitting as a researcher issue rather than an algorithmic issue.

In Section 2.3 we covered unsupervised machine learning - the set of techniques used to identify latent structure in complex datasets. We briefly covered the more esoteric field of dimension reduction, which covers a range of techniques used to reduce high-dimensional data into a more efficient form (efficient in terms of utility and information density). We followed this by exploring clustering, and the relevant methods that are employed in Chapter 4.

Chapter 3

Datasets and Instrumentation

Throughout the work presented in Chapters 4, 5 and 6, a variety of different datasets are used: Chapter 4 makes use of data from the Ulysses and ACE missions; Chapter 5 makes use of data from the THEMIS mission; and Chapter 6 makes use of data from the OMNI dataset, the THEMIS mission and GOES data. This chapter will serve to provide a brief description of the source of each dataset.

3.1 Ulysses

The Ulysses mission [Wenzel et al., 1992] was a joint ESA-NASA mission to study the solar wind outside of the ecliptic plane. One of its primary science objectives was "*to study the origin of the solar wind by measuring the composition of solar wind plasma at different latitudes*", which is why we use this mission's data in Chapter 4's solar wind classification work.

The mission lasted approximately 19.5 years, running from 1990 to 2009. Its solar-polar orbit had a perihelion and aphelion at ≈ 1.3 AU and ≈ 5.4 AU, respectively. The inclination and ellipticity of the $\approx 6-$year orbit allowed for the spacecraft to quickly (≈ 1 year) sample the full range of heliospheric latitudes during the 'fast latitude-scan' at perihelion.

Due to the 6-year length of the orbit, there are significant solar cycle variations over an orbit. This adds an extra layer of complication when trying to analyse the solar wind over a given orbit. In contrast, the quick perihelion passes are fast enough that the latitude sampling can effectively be considered static with respect to the solar cycle.

Ulysses data has been extensively used to classify the solar wind based on speed [e.g., McComas et al., 2008; Ebert et al., 2009; McComas et al., 2013] and composition [e.g., Geiss et al., 1995; Gloeckler et al., 2003; Zurbuchen, 2007; Stakhiv et al., 2015].

The *SWOOPS* [*Solar Wind Observations Over the Poles of the Sun*, Bame et al., 1992] instrument measures the abundance of electrons and ions (H^+ to Fe^{16+}) using electrostatic analysers. Due to only measuring the mass-charge ratio, there is ambiguity in the species identification - hence the inclusion of the *SWICS* instrument (see below). In general, SWOOPS' electrostatic analyser works as follows:

- Electrons/ions enter an instrument aperture.
- Once inside the instrument, particles are subject to a potential difference between plates. The potential difference causes particles to deflect based on their mass and charge-state.
- Different voltages are applied to deflect different energy and charge-state particles to the detector, where the number of particles is recorded.
- The energy-charge-state data can then be compared with the solar wind kinetic temperature, providing information about the species and charge states.

Whilst it's possible for SWOOPS to measure ions, it can only resolve them reliably when the solar wind kinetic temperature is low enough that the species mass contributes significantly to the kinetic energy.

The *SWICS* [*Solar Wind Ion Composition*, Gloeckler et al., 1992] instrument also uses an electrostatic analyser. However, by also employing a time-of-flight calculation and recording the energy specifically, the mass can be deduced. In this way, the detector can reliably measure the requisite parameters to resolve the different species regardless of the solar wind kinetic temperature.

The magnetometer [Balogh et al., 1992] instrument on Ulysses uses a combination of two different magnetometers - a vector Helium magnetometer (VHM) and a flux-gate magnetometer (FGM). Both magnetometers are comprised of three orthogonal detectors, which can be used to identify the magnetic field in 3D with-respect-to the detectors. Where Ulysses magnetometer data has been used, it is only the from the FGM, not the VHM.

The VHM leverages the fact that a Helium gas responds to incident radiation differently depending on the magnitude and direction of a permeating magnetic field. To maintain a constant infrared output from the Helium gas, the instrument generates its own magnetic field to counter the external, permeating magnetic field. This generated magnetic field is thus known, and the external field can be deduced.

The FGM on the other hand works by leveraging the saturation of highly magnetically permeable materials. An FGM is comprised (in each orthogonal direction) of a magnetically permeable ring-shaped core which is toroidally wrapped in a wire coil (the drive-winding), and wrapped again by another wire coil (the sense-winding). When an AC current is passed through the drive-winding coil, it magnetically saturates the core in alternating directions. In the absence of an external magnetic field, the core is saturated, de-saturated and re-saturated evenly. However, in the presence of an external magnetic field, the saturation becomes asymmetric due to the external magnetic field adding to the current-driven magnetic field. The sector of the coil with an induced field in the direction of the external field remains saturated for longer than the opposite sector. This sector would also saturate faster when the induced field alternates back. This asymmetry yields a measurable voltage, from which the external field strength can be inferred.

3.2 ACE

The ACE (Advanced Composition Explorer) mission [Stone et al., 1998] is a NASA mission to study and monitor the composition of the heliospheric medium. One of its primary science objectives was to study the acceleration of the solar wind and to "*study plasma conditions at the source of solar wind and solar energetic particles by measuring and comparing the charge states of these two populations.*", which is why we use this mission's data in Chapter 4's solar wind classification work.

The mission has been running since 1998 (23 years), though due to degradation of one of the instruments, composition data post-2012 is less reliable (and hence, not used in the presented work). The ACE spacecraft orbits the first lagrange point (L1) between the Sun and the Earth, $\approx 240\,\mathrm{R_E}$ from the Earth. The positioning of the spacecraft upstream of Earth (with-respect-to the solar wind), makes it exceptionally useful for Space Weather forecasting. ACE (and other spacecraft also orbiting L1) can provide information about the solar wind with a lead time of between $\approx 8 - 90$ minutes depending on the velocity of the solar wind.

ACE data has been extensively used to classify the solar wind [e.g., Zhao et al., 2009; Xu and Borovsky, 2015; Camporeale et al., 2017; Zhao et al., 2017; Heidrich-Meisner and Wimmer-Schweingruber, 2018; Amaya et al., 2020; Li et al., 2020]

The SWICS [*Solar Wind Ion Composition Spectrometer*, Gloeckler et al., 1998] instrument on ACE is almost identical to that used for the Ulysses mission - it is the Ulysses flight spare.

The SWEPAM [*Solar Wind Electron Proton Alpha Monitor*, McComas et al., 1998] instrument on ACE is also almost identical to the SWOOPS instrument on Ulysses - using the refurbished flight spare as well.

The magnetometer [Smith et al., 1998] instrument comprises two fluxgate magnetometers (as described previously) each mounted on its own boom.

3.3 OMNI

The OMNI dataset is a multi-mission dataset available through OMNIweb. For the time-periods examined in Chapter 6 (post-2006), the spacecraft making up OMNI's solar wind measurements were ACE and Wind. In addition to these data, OMNI also contains geomagnetic indices, derived by the university of Kyoto's World Data Center for Geomagnetism.

OMNI's solar wind data is automatically selected from the most appropriate spacecraft at L1 providing better coverage and measured data. This data is subsequently propagated to the nose of the magnetosphere. These factors make the OMNI dataset extremely usable for Space Weather related research - especially removing the need for the complex propagation.

The solar wind propagation is carried out in two ways, the first uses minimum variance analysis to identify the inclination of solar wind fronts, and is extensively discussed in the back-and-forth papers Weimer [2003, 2004]; Bargatze [2005]; Haaland et al. [2006] and Bargatze et al. [2006]. The second uses a cross-product method [Knetter, 2004], which specifically aims to address shortfalls in using minimum variance analysis. Once the inclination of the fronts is determined, they are ballistically propagated to the nose of the magnetosphere.

The geomagnetic index data are calculated using ground magnetometers. Because the magnetometers measure the current state of the magnetosphere (as opposed to the solar wind data which corresponds to the driving), the derived indices are effectively state variables used to summarise the current response of the magnetosphere to previous solar wind driving. In Chapter 6 we use the AL, AU, AE, SYM-H, SYM-D, ASY-H and ASY-D indices. The A-indices represent auroral behaviour, with AL and AU giving the upper and lower bound of the geomagnetic variations measured by the contributing stations in the auroral region, and AE gives the difference between AL and AU. The SYM-indices represent the disturbance field acting on the geomagnetic field (H for the magnetic North direction and D for the East-West direction). The ASY-indices are the range between the maximum and minimum deviations of the SYM-indices subtracted from the disturbance field. The SYM and ASY indices are derived from equatorial magnetometer stations. During quiet times, the SYM-H index reflects ring current activity.

The OMNI dataset has been used in numerous studies of radiation belt phenomena [e.g., Morley et al., 2010a; Turner et al., 2012b; Borovsky and Denton, 2016; Antonova et al., 2018; Wing and Johnson, 2019; Turner et al., 2019; Smirnov et al., 2020; Pires de Lima et al., 2020; Borovsky, 2020].

3.4 THEMIS

The THEMIS (Time History of Events and Macroscale Interactions during Substorms[1]) mission [Angelopoulos, 2008] was a NASA mission to primarily study the cause and evolution of substorm processes. The secondary mission goal was to understand the “*production of storm-time MeV electrons*”. Both the goals required a comprehensive set of electron measuring instruments, which is why we use this mission’s data in Chapters 5 and 6.

---

[1] The must have had to work for a while to get a backronym that worked for this...

The THEMIS mission comprised 5 individual spacecraft[2] - A, B, C, D and E. Magnetospheric data is available from THEMIS probes A, D and E between 2007 till the present (a 14 year mission duration so far), whilst data from probes B and C is available up till 2010, at which point these probes were moved to a lunar orbit [Russell and Angelopoulos, 2014] as part of the ARTEMIS mission. The five spacecraft were in elliptical orbits - see Figure 3.1. Initially, B had an apogee of $\approx 30\,\mathrm{R_E}$, C had an apogee of $\approx 19\,\mathrm{R_E}$, D and E had apogees of $\approx 12\,\mathrm{R_E}$ and A had an apogee of $\approx 10\,\mathrm{R_E}$.

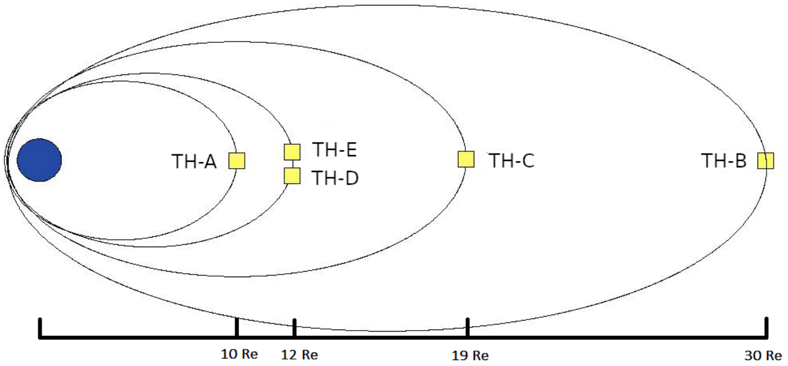


Figure 3.1: *A schematic diagram of the orbits of the 5 THEMIS spacecraft. This diagram highlights the elliptical nature of the orbits, allowing the spacecraft to sample a wide range of radial distances. The orbits precess around the Earth on approximately year-long timescales. Original image credit: University College Berkley [Kramer, 2021], with minor edits to reflect the spacecraft apogee distances in Earth radii, and to give the alphabetic designation of the spacecraft.*

THEMIS data has been widely used in radiation belt research [e.g., Ni et al., 2011; Turner et al., 2012a,b; Baker et al., 2013; Shin and Lee, 2013; Turner et al., 2013; Shin et al., 2014; Maget et al., 2015; Turner et al., 2015; Boyd et al., 2018; Liu et al., 2018; Turner et al., 2019].

The THEMIS ESA [electrostatic analyser, McFadden et al., 2008] works the same way as the Ulysses and ACE electrostatic analysers, though the THEMIS ESA instrument comprises two (‘top-hat’) detectors, giving it a full 4π steradian field of view, rather than the more limited aperture on the aforementioned detectors.

The THEMIS SST (Solid State Telescope) does not have an instrument paper (*very helpfully*). Fortunately, the instrument is based upon the similar instrument flown on the WIND spacecraft, for which there is an instrument paper [Lin et al., 1995]. The SST instrument comprises a set of two double-sided detectors. On one side of each detector, a thin foil is added to prevent proton entry, and on the other

---

[2] The THEMIS mission was the first NASA mission to include an extra spacecraft for redundancy - nominally probe A.

side, strong magnets deflect incoming electrons. This effectively allows one direction to act as a proton detector, and the other as an electron detector. The SST works the same was as a semiconductor detector. Incident ions or electrons hit the silicon detectors and induce ionisation currents which can be measured. These currents correspond to the intensity of the incident particle beam, allowing the energy spectra to be deduced based on which detectors the currents are measured in. This means that the instrument measures the differential number flux (allowing the kinetic energy flux to be determined too[3]).

Later in the THEMIS mission, it was realised that cross-species contamination of the electron flux by protons was present in the SST data. This is briefly mentioned by Turner et al. [2012a, 2013], though the precise amount of contamination or exactly where it occurs seems to remain elusive to published work. In a private communication D. Turner explained that the contamination is most prevalent inside of geosynchronous orbit altitudes, and practically negligible outside of $8\,\mathrm{R_E}$. We speculate that the contamination may increase electron fluxes measured in the SSTs lower energy channel.

3.5 GOES

The GOES (geostationary operational environmental satellites) are a series of spacecraft operated by NOAA (National Oceanic and Atmospheric Administration), designed to provide data for operational weather and Space Weather monitoring and forecasting. In Chapter 6 we use data from GOES 13-15 (N-P) [Kramer, 2021, and references therein].

GOES data is currently used in the construction of the outer boundary condition for the BAS-RBM [British Antarctic survey's radiation belt model Glauert et al., 2018], which is why we incorporate it into our outer boundary electron flux parameterisations in Chapter 6.

The various GOES are placed into two geostationary orbits (meaning that they are located above a fixed latitude and longitude), east and west. The eastern orbit is located at longitude 75° west and a latitude of $\approx 11^\circ$ north of the magnetic

---

[3] Relativistic energies, though present in the radiation belts, are not accounted for in the instrument design or subsequent data processing.

equator, and the western orbit is located at $135°$ west and a latitude of $\approx 4°$ north of the magnetic equator [Meredith et al., 2015].

These data were obtained from CDAWeb, and are specifically the background, contamination and dead time corrected data from the Energetic Proton, Electron and Alpha Detector (EPEAD) [Onsager et al., 1996]. The detectors are dome detectors, but work in much same was as the solid state telescope described for THEMIS. One difference is that GOES measures integral fluxes rather than differential fluxes. It measures these in 3 bands > 0.6 MeV, > 2 MeV and > 4 MeV[4], by placing modulating plates of different thicknesses in front of each detector (effectively preventing lower energy electrons from entering through the thicker plates).

Due to the different latitudes, the eastern and western orbits sample different magnetic field lines, with the eastern orbit typically sampling magnetic field lines which map further outwards in the equatorial plane - thus the eastern spacecraft often measures lower electron fluxes [Meredith et al., 2015].

---

[4] CDAweb does not provide the > 4 MeV electron fluxes for GOES 13-15.

Chapter 4

Data-Driven Classification of Coronal Hole and Streamer Belt Solar Wind

Abstract

We present two new solar wind origin classification schemes developed independently using unsupervised machine learning. The first scheme aims to classify solar wind into three types: coronal hole wind, streamer belt wind, and 'unclassified' which does not fit into either of the previous two categories. The second scheme independently derives three clusters from the data; the coronal hole and streamer belt winds, and a differing unclassified cluster. The classification schemes are created using non-evolving solar wind parameters, such as ion charge states and composition, measured during the three *Ulysses* fast latitude-scans. The schemes are subsequently applied to the *Ulysses* and the *Advanced Composition Explorer* (*ACE*) datasets. The first scheme is based on oxygen charge state ratio and proton specific entropy. The second uses these data, as well as the carbon charge state ratio, the alpha-to-proton ratio, the iron-to-oxygen ratio, and the mean iron charge state. Thus, the classification schemes are grounded in the properties of the solar source regions. Furthermore, the techniques used are selected specifically to reduce the introduction of subjective biases into the schemes. We demonstrate significant best case disparities (minimum $\approx$8%, maximum $\approx$22%) with the traditional fast and slow solar wind determined using speed thresholds. By comparing the results between the in- (*ACE*) and out-of-ecliptic (*Ulysses*) data, we find morphological differences in the structure of coronal hole wind. Our results show how a data-driven approach to the classification of solar wind origins can yield results which differ from those obtained using other methods. As such, the results form an important part of the information required to validate how well current understanding of solar origins and the solar wind match with the data we have.

4.1 Introduction

The solar wind comprises streams of ionised particles which travel nearly radially from the Sun through the heliosphere. From the earliest *in situ* observations, it was clear that the solar wind could be broadly classified into two types, fast and slow [Neugebauer and Snyder, 1966; Stakhiv et al., 2015]. This duality was found to extend beyond the local solar wind speed, but is present in the elemental composition and ion charge states of the solar wind, indicating very different coronal source properties of fast and slow wind [von Steiger et al., 2000; Geiss et al., 1995]. Fast wind is found to originate from coronal holes [Sheeley et al., 1976]. These are magnetically open regions of the corona where the plasma can freely escape, meaning that coronal holes appear dark in EUV emission. The formation and release of the slow wind is a current area of research, but it originates from the vicinity of closed coronal magnetic structures such as the streamer belt [Antiochos et al., 2011; Ko et al., 2006; Xu and Borovsky, 2015; Brooks et al., 2015]. At solar minimum, coronal holes cover the polar regions, with the streamer belt confined close to the solar equator. At solar maximum, the coronal field is far less ordered. The resulting variation of solar wind speed can be seen in Figure 1 of McComas et al. [2013]. Despite the breakdown of the latitudinal dependence at solar maximum, there is still a good separation between streams of different speeds. This suggests that despite the activity, the source regions remain isolated from one another, and there is not significant mixing of the streams.

While appealing, the traditional two-type solar wind paradigm is not unique, with a number of different observationally-determined solar wind types proposed. A two-type scheme has been proposed by Zhao et al. [2009], a three-type scheme has been proposed by Stakhiv et al. [2015], a four-type scheme has been proposed by Xu and Borovsky [2015] and been built upon using machine learning by Camporeale et al. [2017], and even a six-type scheme has been proposed by Zhao et al. [2017]. Furthermore, Heidrich-Meisner and Wimmer-Schweingruber [2018] have proposed a two-type classification scheme, and a two – seven type scheme (depending on interpretation) using the k-means clustering algorithm [MacQueen, 1967] implemented using the C++ library, Shark [Igel et al., 2008]. In each of these categorisation schemes the properties of each solar wind type are quantitatively different from one another, an essential factor when performing statistical studies of heliospheric phenomena driven by the solar wind.

The Zhao et al. [2009] scheme sought to classify the solar wind into coronal hole wind or non-coronal hole wind. Stakhiv et al. [2015] classified the solar wind into coronal hole wind, wind due to magnetic reconnection at the boundary of large scale streamers, and a boundary wind which originates from the edges of coronal holes. Xu and Borovsky [2015] described a scheme which encompasses coronal hole wind, sector reversal region wind emitted from the top of helmet streamers, and streamer belt wind. The streamer belt wind is comprised of two types: pseudostreamers and helmet streamers. These occur when two loop arcades separate a pair of like-signed coronal holes, and when a single loop arcade separates two coronal holes of opposite polarity, respectively [Panasenco and Velli, 2013; Owens et al., 2014]. Zhao et al. [2017] split the solar wind into six types: coronal hole, active region, quiet Sun, active-region boundary, coronal-hole boundary and helmet streamer. The regions in this case are not determined by coronal signatures in the solar wind, but instead by direct mapping to the Sun. A ballistic method is used to map to the solar source-surface, and then an extrapolation is made using the potential-field source-surface model [Altschuler and Newkirk, 1969; Schatten et al., 1969] to map to the photosphere.

As part of the Machine Learning Techniques for Space Weather book [Camporeale et al., 2018], Heidrich-Meisner and Wimmer-Schweingruber [2018] present a systematic analysis of applying a simple unsupervised machine learning algorithm, k-means, to the classification of solar wind types. A variety of parameter spaces are investigated (13 different sets are used), as is the choice of the number of clusters for which the algorithm should search. The first k-means scheme proposed is a coronal hole *versus* slow wind scheme, whilst the second uses k-means to find seven clusters (where the number of clusters to find was a data-driven choice). The latter scheme provides results which are significantly more open to interpretations. The authors state that they find: two coronal hole wind classes, though one may comprise interplanetary coronal mass ejection (ICME) plasma; one primary slow solar wind class; and four potential sub-classes of slow solar wind, where two are compressional/rarefaction regions surrounding a stream interaction, another is very slow, dense and cool wind, and the final is even more dense, has high charge states and is cool (though again, this may represent undetected ICMEs).

Aside from the growing evidence that the simplistic solar wind speed categorisation scheme is not adequate for distinguishing solar source regions, there is a more direct reason that such a scheme is not appropriate for many of the datasets that exist: co-rotating interaction regions (CIRs). CIRs are the compression regions that form when high-speed solar wind streams catch up to low-speed streams as they travel through the heliosphere. Since coronal hole wind (CHW) and streamer belt wind (SBW) typically show latitudinal dependence, CIRs do not tend to form everywhere. Instead, there is a tendency towards the ecliptic plane due to the inclination of the solar rotation axis [Crooker et al., 1999; Borovsky and Denton, 2016]. The result is that much of the solar wind in the ecliptic plane undergoes interaction of high- and low-speed streams. Such mixing causes high-speed streams to slow down and slow-speed streams to speed up. Thus, speed is not the most reliable means to distinguish different coronal sources.

The scientific motivation behind the current work is to provide two new solar wind classification schemes. They will be developed using unsupervised machine learning techniques so as to reduce scientific subjectivity. By using novel techniques which have their own unique biases, this work will provide new information towards validation or benchmarking of existing solar wind classification models. As with any scientific work, the total removal of any subjective influence is near impossible. Our methods hope to address the scientific subjectivity in the determination of classification boundaries, and number of solar wind types. The Bayesian Gaussian Mixture (BGM) scheme addresses the former point, whilst the Uniform Manifold and Projection (UMAP) scheme addresses both.

4.2 Data

For this analysis, data from the *Ulysses* spacecraft's [Wenzel et al., 1992] *Solar Wind Observations Over the Poles of the Sun* [*SWOOPS*, Bame et al., 1992], *Solar Wind Ion Composition* [*SWICS*, Gloeckler et al., 1992] and magnetometer [Balogh et al., 1992] instruments have been primarily used. The motivation for this usage is that, unlike in-ecliptic spacecraft such as the *Advanced Composition Explorer (ACE)*, *Ulysses* has a polar solar orbit and enables sampling of the pure CHW during the high-latitude phase of the mission at solar minimum. The *Ulysses* mission contains three 'fast latitude-scans', which are periods at perihelion when the spacecraft covers almost the full solar latitude range in a relatively short amount of time (approximate dates: 15/08/94 – 20/08/95; 01/11/2000 – 01/11/2000, and 01/02/07 – 01/02/08.). Particularly for the two solar minimum fast latitude-scans, solar wind types can be well separated by their latitudinal dependence. The latitude-scans comprise ≈3 years worth of total data, whilst the whole dataset is ≈19.5 years (1990 – 2009). The data are mean-resampled into three-hourly steps to match the cadence of the compositional data. In practice, this yields 8227 (8139) latitude-scan data points, 46893 (45463) total *Ulysses* data points and 38108 (23665) *ACE* data points for the BGM (UMAP) scheme. The UMAP scheme has fewer points due to a larger parameter space; the whole data point must be discarded if any one parameter is bad.

Solar Wind Ion Composition Spectrometer [*SWICS*, Gloeckler et al., 1998], Solar Wind Electron Proton Alpha Monitor [*SWEPAM*, McComas et al., 1998] and magnetometer [Smith et al., 1998] data from *ACE* [Stone et al., 1998] are also used. As *ACE* is confined to the ecliptic plane (at the first Lagrange point, L1, just upstream of Earth), it rarely samples the CHW without it first interacting with SBW.

The classification scheme developed from the *Ulysses* data will be applied to ≈13 years of *ACE* data (1998-2011). This will allow more statistical insight into the link between solar wind source-regions and Space Weather events. All the *Ulysses* and *ACE* data are mean-resampled into three-hourly data.

Since solar wind speed is a poor parameter choice for classification, other parameters must be used. In order to relate properties to the coronal source conditions, parameters should ideally remain constant as the solar wind flows from the Sun. For this task, ion charge state ratios are the obvious choice, since they are well known to be non-evolving parameters after a few solar radii [Pagel, 2004; Geiss et al., 1995]. The reason for this is that the electron mean free path becomes so large that interactions are negligible [Owocki et al., 1983]. Between the *Ulysses* and *ACE* spacecraft, the common charge state and composition measurements are: O^{7+}/O^{6+}, C^{6+}/C^{5+}, Fe/O, $< q_{Fe} >$, and He^{2+}/H^{1+} (where fractions signify the relative density ratio). Further to these, Burlaga et al. [1990] describe how the proton specific entropy, S_p, is a good stream signature since it only diverges 10% between 1 – 5 AU ($S_p = \frac{T_p}{\sqrt{n_p}}$).

4.3 An Intuitive Classification Scheme

Many studies of the classification of the solar wind often rely on scientists' intuition. Xu and Borovsky [2015] [and by proxy Camporeale et al., 2017] state that their method of determining classification boundaries is that they are "chosen by eye". Zhao et al. [2009] base their identification of ICMEs [Cane and Richardson, 2003] on the work of Richardson [2004], who state their choice of parameter boundaries are "somewhat arbitrary". Such expert intuition is undoubtedly valuable, but extending this intuition from the abstract to the mathematical is a necessary progression.

In order to enable comparison with the machine learning approaches introduced below, we essentially reproduce the threshold approach in two-parameter space. The chosen parameters are O^{7+}/O^{6+} and S_p. These parameters are chosen based on the work of Zhao et al. [2009] and Xu and Borovsky [2015], respectively. The methodology is as follows. Firstly, we take the log of our data and then plot the occurrence density. By visually inspecting the result, we see groupings within the data. These groupings are subsequently separated by placing a line (linear in log-space) to divide them. This dividing line forms the classification boundary between the two groupings. Not only does such a model allow for investigating the physical premise of the classification schemes introduced subsequently, but it will also be used as a benchmark to show that further results are not wholly unique to more complicated methods.

Figures 4.1a and 4.1b present two identical occurrence density plots of the whole *Ulysses* dataset, wherein two populations are clearly visible (the colourbar is log scaled). The panels include a different choice of threshold line which plausibly separates the populations in the data. Thresholding has been performed twice to highlight the fact that the result is not specific to a given threshold. Figures 4.1c and 4.1d show the results of the classification in linear space. Despite being a simplistic classification scheme, there are benefits to its use; it is transparent and based on parameters which are known indicators of solar source regions, thus reducing the impact of solar wind stream interaction. From panels a, b, c and d it is inferred that the data found to the left (right) of the classification thresholds are CHW (SBW) due to their lower (higher) charge state ratios and higher (lower) proton specific entropy.

We make a simple comparison to demonstrate how the classical speed-threshold scheme does not well-capture the origins of the solar wind, as compared to the intuitive scheme. The *Ulysses* data are divided into fast- and slow-stream wind according to various speed thresholds. We calculate the proportion of points in each case where the results of the speed-threshold scheme do not agree with the results of the intuitive scheme. The total number of discrepant points is divided by the total number of points in the data used, giving a fraction describing the relative difference between the speed-threshold and intuitive schemes. This is shown in Figures 4.1e and 4.1f.

Whilst differences can be seen due to individual thresholds used in the two intuitive schemes, it is their similarity which is of most import. Specifically, both show comparative inaccuracies of the speed-threshold scheme greater than 10% for all speed thresholds. Already, this simple scheme highlights the potential shortfalls of the speed-threshold scheme.

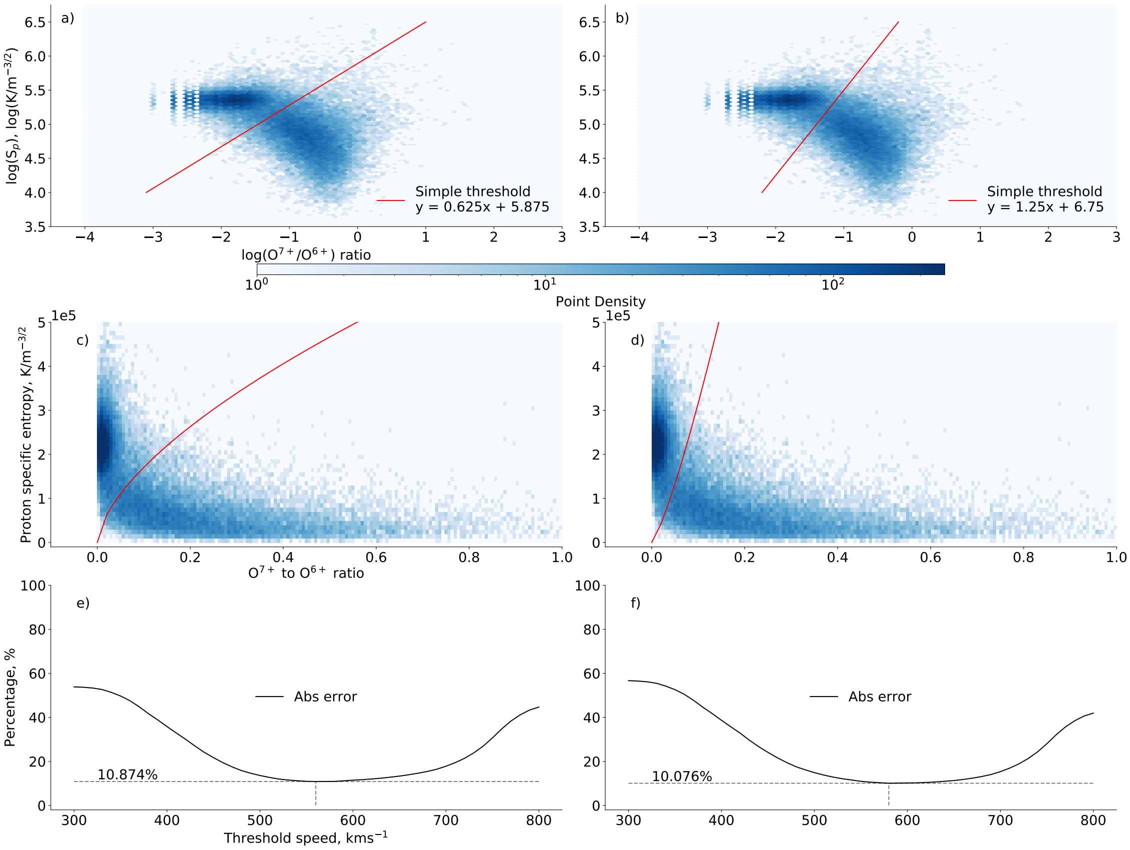


Figure 4.1: *The intuitive classification scheme with comparison to classical speed-threshold methods. Panels a and b show occurrence density plots of ≈19.5 years of Ulysses data as a function of O^{7+}/O^{6+} and S_p. Note the logarithmic colourbar scale. Each plot shows a different threshold chosen to separate the two main solar wind populations, assumed to represent coronal hole and streamer belt winds. Panels c and d present this classification in linear space. The coronal hole wind is represented by low O^{7+}/O^{6+} and high S_p, and vice versa for the streamer belt wind. Finally, panels e and f present the resulting mis-classification of solar wind using the speed-threshold method, in terms of the absolute error.*

4.4 Machine Learning Schemes

Whilst the intuitive scheme is undoubtedly useful, it still contains subjective decisions about which parameters to use, the number of solar wind types to identify, and the decision boundaries. Here, more objective (data-driven) and mathematical methods are presented. Unsupervised machine learning will be used to create two new classification schemes; with reduced subjectivity and more algorithmic reproducibility. The latter point specifically contrasting the selection of decision boundaries 'by eye'.

Machine learning (ML, see Chapter 2) can be split into two main categories; supervised and unsupervised. Supervised ML describes the techniques which produce and optimise a function to map from an input (data) to an output (class label), given a set of example (training) input-output pairs [Russell and Norvig, 2009]. By contrast, unsupervised ML describes the subset of techniques which are used to determine effective ways of mathematically separating data with no predetermined class labels. Instead of a boundary function being optimised by a predictive performance metric, the optimisation is often focused on improving the separation of data clusters. In this way, unsupervised ML can be applied to data with less bias, allowing for groupings in the data to be found mathematically rather than being influenced by what one may expect to find *a priori*. Unsupervised ML is a data-driven approach to classification. Its purpose is to determine an underlying structure in the data and find quantitative separations between discrete regions. As such, the algorithms find that which is already present in the data (subject to algorithm specific limitations).

The first new scheme will allow for the determination of a third solar wind category. This category represents data which is difficult to assign to either CHW or SBW, and hence be referred to as unclassified data. The second proposed scheme will independently determine the number of solar wind categories.

To cluster the whole *Ulysses* dataset is a bad idea for several reasons: as previously mentioned, there is limited pristine data; clustering is computationally expensive; and, it is inefficient for classifying new data (since the clustering would have to be re-performed). To address these issues, the clustering is performed on the three latitude-scans. This allows the clustering to be performed only on the more pristine data, with higher latitudinal dependence; provides a more manageable dataset for clustering, reducing computational complexity; and ensures the ability to classify any new data efficiently.

Subsequently, the results of the clustering are applied to classify $\approx$19.5 years of *Ulysses* data and $\approx$13 years of L1 *ACE* data. The independence of the classifications from solar wind speed allow them to be applied to the *ACE* dataset despite significant solar wind stream interactions.

4.5 The Bayesian Gaussian Mixture Scheme

Some of the literature regarding solar wind classification is built upon classification boundaries which are chosen subjectively (*e.g.* "arbitrarily" or "by eye"). We present a Bayesian Gaussian Mixture (BGM) classification scheme which uses unsupervised machine learning to mathematically determine the optimum data-driven decision boundary between solar wind types (subject to the suitability of the Gaussian assumption).

The BGM algorithm iteratively fits a Gaussian mixture [McLachlan and Peel, 2000] to the data. During each iteration, variational inference is implemented to do two things: firstly, find the probability of each point being generated by the mixture; and secondly, refit the mixture to the points using information from the prior distributions [Attias, 2000; Bishop, 2006] (for further information regarding variational inference, see the Appendix A). Once convergence has been reached, the algorithm outputs the cluster label for each point (*i.e.* the label of the Gaussian in the mixture to which it belongs), and the information describing the distributions (*e.g.* mean and variance). The latter information is extremely useful, as it allows for the Gaussian mixture to be stored, removing the need to run the algorithm every time. With the Gaussian mixture stored, application to data classification is straight-forward: firstly, each new data point is mapped into the pre-established normalised space; then, the posterior probability of each component Gaussian given the data point is calculated [further detail given in Appendix A, or see *e.g.* Gelman et al., 2013]; and finally the point is assigned to the component with the highest probability of generating it [as per Camporeale et al., 2017]. The BGM is here applied using the algorithm from the scikit-learn package available for python [Pedregosa et al., 2011].

We do not use k-means as Heidrich-Meisner and Wimmer-Schweingruber [2018] have done. From the standpoint of the objective functions being optimised by k-means and the BGM (as opposed to the algorithms used to attempt the optimisation), k-means is strictly a special case of a Gaussian mixture model. That is, if you choose a Gaussian mixture with K components and fix the Gaussians to be spherical (scalar multiple of the identity for covariance) then the means of the maximum-likelihood estimate for the mixture are the centroids that minimise the distance from the data to the centroids. Algorithmically, k-means and the BGM method use different optimisation techniques, but philosophically k-means is a subset of a Gaussian mixture model. Using a BGM rather than k-means allows for non-circular clusters to be appropriately described using ellipses, by relaxing the restriction that the Gaussians must be spherical.

To test the validity of the above arguments against using k-means, we have investigated how the results differ from the BGM scheme. Overall, the results from k-means are qualitatively the same as those from the BGM (*i.e.* the majority of data are assigned the same class in both schemes), but with drawbacks. Such drawbacks include an apparent increase in the mis-classification of *Ulysses* CHW data, and incongruent speed distributions for the unclassified data between *Ulysses* and *ACE*. These differences are due to the comparatively poor way of determining classification boundaries, and the changes in the objective functions being optimised. These differences both highlight that k-means is less suited to classification in the way we have applied the BGM.

The BGM approach allows probabilistic classifications, antithetical to the intuitive scheme. Whilst fitting Gaussians to data is a common practice, there is the inherent shortfall of the approximation becoming less valid as a dataset diverges from being normally distributed. As such, whilst we may be more objective in the fitting procedure and gain information (*e.g.* probabilities), the results must always be considered carefully in terms of the validity of the Gaussian assumption.

Despite the BGM producing probabilistic results, this study will use hard decision boundaries. Points will be assigned to the Gaussian which most likely generated it. Such an approach is entirely adequate for comparing between different solar wind classification schemes (since most others use hard boundaries). In theory, problems may arise if there were many data points yielding comparable (60%:40% say) probabilities of belonging to multiple classes, but in our case fewer than 1-in-10 data points have probabilities below 90%. Hence, minimal data are affected by our use of hard boundaries.

Each parameter in the dataset is normalised to a zero mean and unit standard deviation to reduce any bias that the heteroscedasticity of the variables could introduce to the algorithm. The method of normalisation is through the standard score:

$$x' = \frac{x - \mu}{\sigma}. \tag{4.1}$$

Where x' is the normalised value, x is the initial value, μ is the mean of the population, and σ is the standard deviation of the population.

The BGM algorithm does require user-specified parameters. 1) The number of components in the mixture. Since this study focuses on classifying the solar wind into coronal hole and streamer belt wind whilst accounting for data which is difficult to classify, the algorithm is set to fit a three-component Gaussian mixture to the data. 2) The precision prior on the mean distribution. The motivation of this research is to avoid incorporating bias where possible. Therefore, the prior was set to be flat, allowing all possible mean positions to be equally weighted. 3) The number of initialisations. The algorithm was set to perform clustering with 30 random initialisations of the means to ensure that the convergence was not on a local maximum/minimum. The result with the largest value of the lower bound of the likelihood is kept. Convergence is reached when the change in likelihood is less than 10^{-5} between iterations. Higher values of the likelihood correspond to higher degrees of confidence that the model could produce the data (see e.g. Gelman et al. [2013] for a detailed description of likelihood). As such, by choosing the model with the highest lower bound, the baseline degree of confidence is highest.

The Gaussian Mixture best describing the data can be described by the three component means $\mu_{1-3}(O^{7+}/O^{6+}, S_p)$, covariances $cov_{1-3}(O^{7+}/O^{6+}, S_p)$ and their respective weightings. In the normalised space the means and covariances are as follows:

$$\mu_1 = (-0.3779, 0.6252),\ cov_1 = \begin{pmatrix} 0.0019 & -0.0012 \\ -0.0012 & 0.1833 \end{pmatrix}$$

$$\mu_2 = (0.4235, -1.1145),\ cov_2 = \begin{pmatrix} 0.3759 & -0.1009 \\ -0.1009 & 0.1067 \end{pmatrix}$$

$$\mu_3 = (3.2836, -0.0057),\ cov_3 = \begin{pmatrix} 15.9839 & -3.0918 \\ -3.0918 & 6.3546 \end{pmatrix}$$

and the weights are 0.6238, 0.3497, and 0.0265, respectively.

Figure 4.2a presents the results of applying the BGM clustering algorithm to the latitude-scan data, and 4.2b shows how the clusters map to the solar wind speed and proton temperature. The combination of the two parameters used allow the clustering to map the solar wind well to either coronal holes or the streamer belt. Cluster one, with low average O^{7+}/O^{6+} and high average S_p, represents CHW. Cluster two, with higher average O^{7+}/O^{6+} and lower average S_p, represents SBW. Cluster three is thus the unclassified data. The projection of the clustering into the solar wind proton speed and temperature shows clearly that the clustering is capturing distinct populations. The interaction of CHW and SBW can be seen by the overlapping of the two groups along the speed axis.

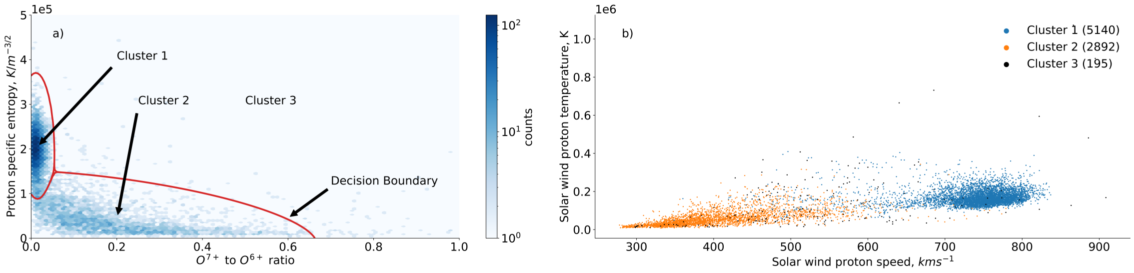


Figure 4.2: *Classification with the BGM scheme. Panel a presents the results of Bayesian Gaussian Mixture algorithm clustering on the normalised* O^{7+}/O^{6+} *and* S_p *data. The plot has been trimmed in both x and y to better display clusters one and two, as such a small number of data points from clusters two and three are not shown. Panel b presents the projection of the clustered data into into solar wind proton speed and temperature.*

To investigate the stability of the clustering, the procedure was performed a further 300 times using random sub-samples of 90% of the data. Upon completion of each iteration, the mean value of each component Gaussian was recorded. Once completed, the standard deviations and inter-quartile ranges of the distributions of means are calculated. If the clustering had found a local maximum/minimum in the data, we would expect there to be significant differences between the results of a single run compared with the statistical results of many runs. The results of the analysis are presented in Table 4.1. The BGM algorithm does not systematically label the Gaussians and so eight of the recorded mean values were incorporated into the incorrect group, and thus removed.

The proportionally small standard deviations and inter-quartile ranges signify that the clustering is stable, and that the individual runs do not deviate greatly from the average values. The means of the component Gaussians used to classify the data in the normalised space are [-0.3779, 0.6252], [0.4235, -1.115], and [3.284, -0.0057] for clusters one, two, and three respectively. Comparing these values to those presented in Table 4.1, we see that they are very much in-line with the standard behaviour.

4.5.1 BGM Scheme: Application

The clustering described above has been used to develop a solar wind classification scheme, based on the fast latitude-scan subset of the *Ulysses* data. Here we will apply the classification scheme to the whole *Ulysses* and *ACE* datasets.

		Sub-sampled BGM Model Component		
		One (296)	Two (296)	Three (300)
O^{7+}/O^{6+}	μ	-0.3778	0.4235	3.281
	σ	0.0002 (0.0003)	0.0052 (0.0064)	0.128 (0.165)
S_p	μ	0.6251	-1.115	-0.0093
	σ	0.0022 (0.0028)	0.003 (0.005)	0.0704 (0.0984)

Table 4.1: The results of three-hundred 90% sub-sample runs of the Bayesian Gaussian Mixture algorithm on the fast latitude-scan *Ulysses* data. The values in brackets after the component number indicate the number of data points being included (some are excluded due inconsistent labelling). μ is the mean of the individual component means. σ is the standard deviation of the means, the value in brackets represents the inter-quartile range. All values of the mean, standard deviation, and inter-quartile range are given in the normalised space.

4.5.1.1 *Ulysses*

Figure 4.3 presents the results of the classification of the whole *Ulysses* dataset. In Figure 4.3a one can see how the SBW (cluster two) appears to deviate from a Gaussian. On the contrary (though less obvious from the plot) is that the CHW (cluster one) is well approximated by a Gaussian, especially in comparison to the SBW. However, a significant portion of the data that we might consider to be difficult-to-classify has been captured as such by the algorithm. When compared with taking simple speed thresholds, Figure 4.3b, there is a minimum of ≈22% disparity in the results. Again, this is highly suggestive that the traditional method falls short of adequate for many applications. Also shown in Figure 4.3b is the disparity when the unclassified data are ignored. This has been included to allow a more like-for-like comparison (since both schemes can be considered two-type schemes). In this way the disparity is reduced to ≈6%, suggesting that the speed threshold captures the cores of the clusters. Nonetheless, the speed-threshold scheme oversimplifies the classification of solar wind data, and importantly gives too much confidence to the classification of borderline data.

4.5.1.2 *ACE*

Figure 4.4 presents the results of the BGM classification of the whole *ACE* dataset. In Figure 4.4a there are considerable difference as compared with the *Ulysses* data; as expected due to Earth's orbital position, there is significantly less CHW (cluster one) and more SBW (cluster two). When compared with simple speed-

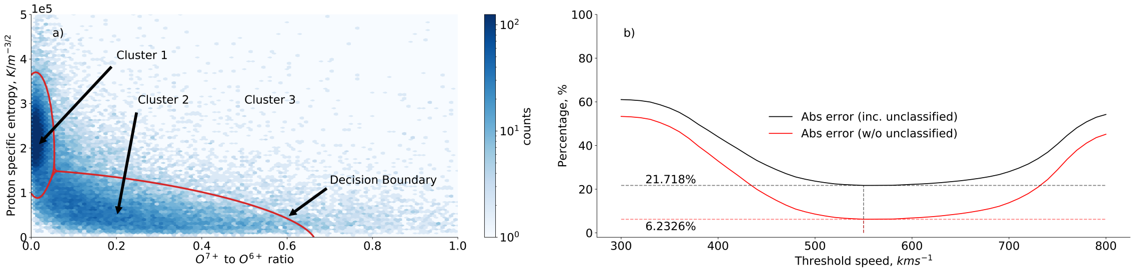


Figure 4.3: *Panel a presents the results of applying the BGM fast-latitude scan classification to the whole Ulysses dataset, shown in the* O^{7+}/O^{6+} *and* S_p *space. Cluster one, two, and three represent CHW, SBW, and unclassified data, respectively. Panel b shows the mis-classification of the data using a simple speed threshold, both including and excluding unclassified data (as determined by the BGM technique).*

threshold classification, Figure 4.4b shows that there is a minimum ≈18% disparity in the results. Again, ignoring the unclassified data, the disparity is reduced; though, the same issues persist.

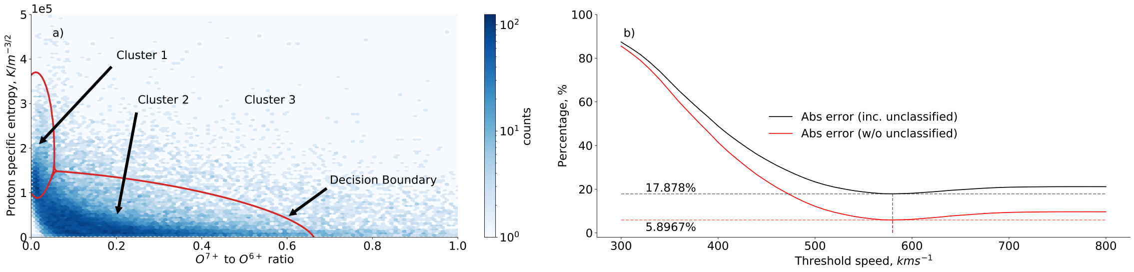


Figure 4.4: *The results of applying the BGM scheme to the whole ACE dataset, and the subsequent comparison to taking speed thresholds. Panels a and b are presented in the same format to Figure 4.3. Cluster one, two, and three represent CHW, SBW, and unclassified data, respectively.*

4.5.2 BGM Scheme: Analysis

To view the way in which the classification of the *Ulysses* data maps to velocity and solar latitude, a McComas et al. [2013] style visualisation is presented in Figure 4.5a. Both the latitudinal and speed dependent nature are clearly present. These dependent variables have not been used in the classification scheme, but the correlation is expected. Capturing the predicted behaviour shows that the initial choice of parameters is well-informed.

Figure 4.5b shows that the unclassified data is skewed towards the aphelion of the orbit. It is worth noting that, due to the slower motion of the spacecraft, there is considerably more data per latitudinal increment at aphelion than other portions

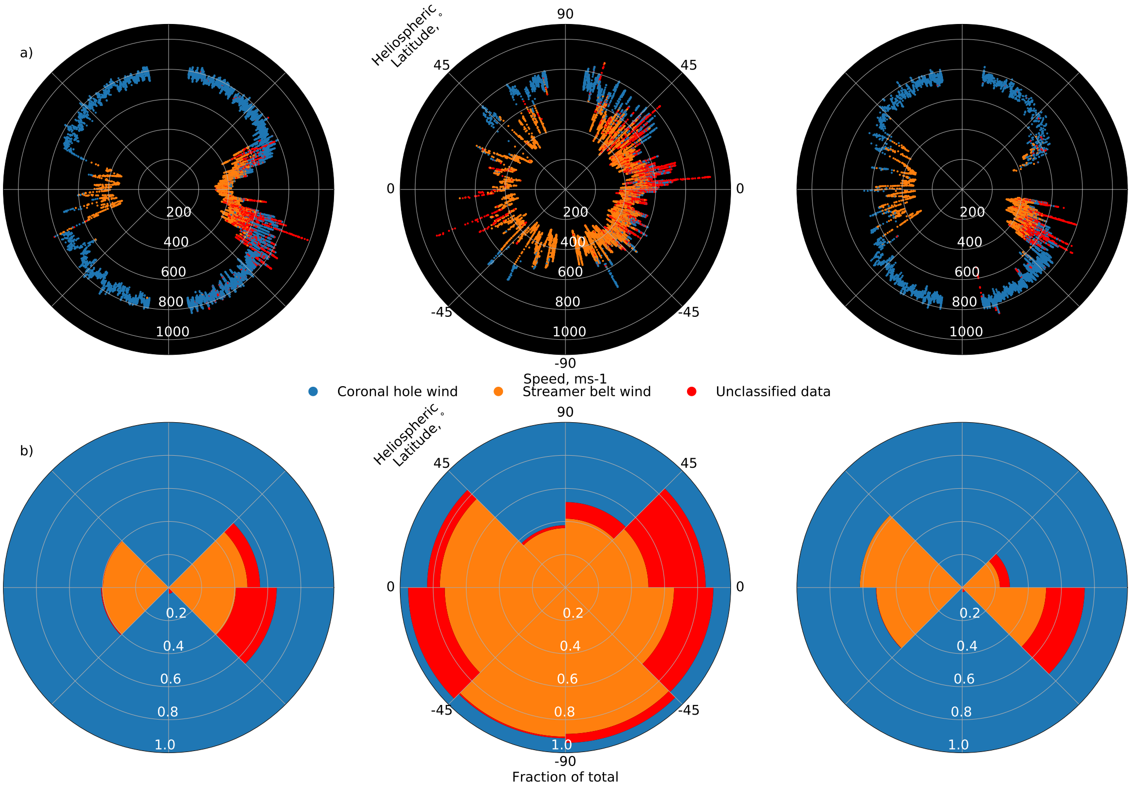


Figure 4.5: *The BGM classification of the whole Ulysses dataset. Panel a presents the projection of the BGM classification scheme onto radial plots of solar wind speed and Ulysses' heliospheric latitude. Panel b presents the fraction of each classification type in each of the octet segments of the plot. Each plot represents an orbit of Ulysses around the Sun, with the ecliptic plane in the east-west direction. In both panels, time increments clockwise, starting from aphelion at 8.6 degrees below east. The first and third plots of panels a and b are the orbits where perihelion occurred at solar minimum, whilst in the middle plots, perihelion occurred at solar maximum.*

of the orbit. After accounting for the expected increase in unclassified data, there remains a significant disparity in the distribution of unclassified data. The aphelion regions of the orbit present more unclassified data than the perihelion regions.

Figure 4.6a presents the speed distributions of the three BGM classifications from *Ulysses* data. Note that these speeds were not used in the classification in any way. The SBW shows no significant bi-modality, and appears to follow a Maxwellian distribution. Both the CHW and the unclassified data show some suggestion of being bi-modal, each with their secondary peak aligning close to the primary peak of the other (whilst subtle, the secondary peaks are present, *viz.* ≈ 775 kms^{-1} for the unclassified, and ≈ 500 kms^{-1} for the CHW). This suggests that the classification scheme may be having trouble differentiating between the two types (see discussion of Figure 4.7). Figure 4.6b presents the same distributions, but obtained from the classification of the *ACE* dataset. We see the significant drop in CHW, but observe that the distributions of both CHW and SBW are not double-peaked. The unclassified data in the *ACE* classification is considerably flattened, suggesting that the difficulty in classification may be ubiquitous in the ecliptic plane.

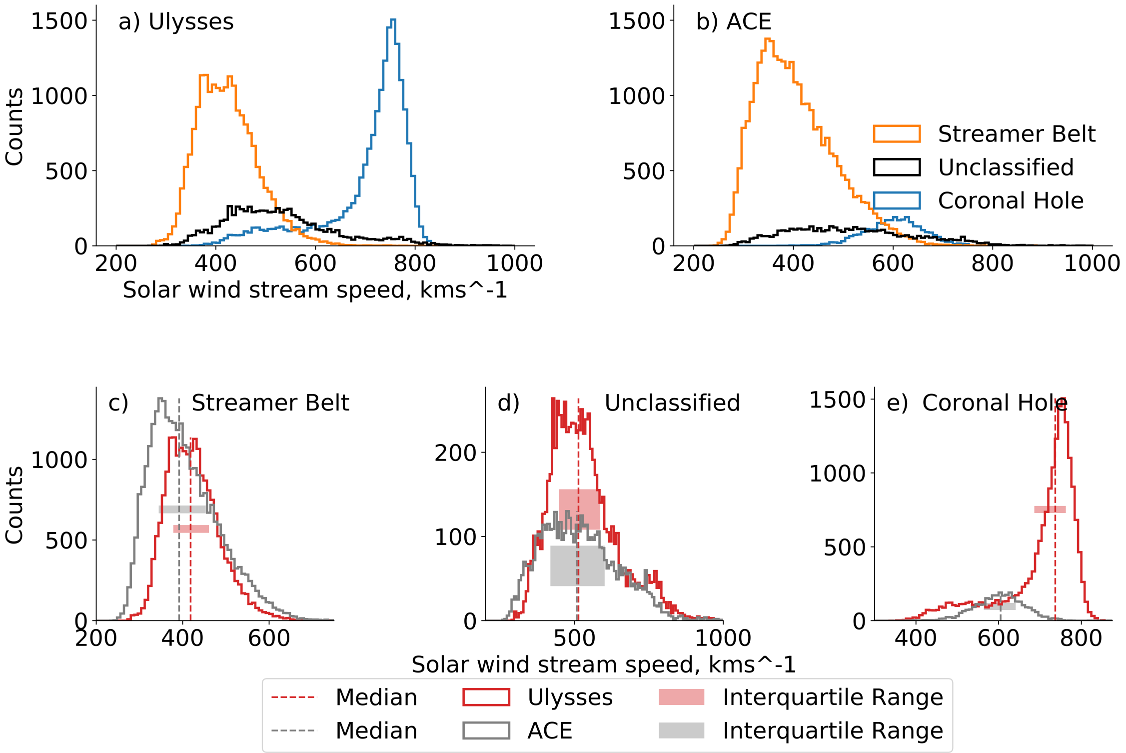


Figure 4.6: *Distributions of solar wind speeds within each cluster found by applying the BGM classification scheme. Panels a and b present the comparisons for the distributions in the Ulysses and ACE datasets respectively. Panels c, d and e present the comparisons of like-clusters between the datasets. The latter panels also include the mean and inter-quartile range of each distribution.*

Figures 4.6c, 4.6d and 4.6e show direct comparisons of the distributions of solar wind classifications in the *ACE* and *Ulysses* datasets. There is good qualitative agreement in the distributions of the SBW and unclassified data, suggesting that the scheme well-captures streamer-belt solar wind structures, as well as consistently identifying the unclassified data. However, the CHW distributions for *ACE* and *Ulysses* are very different. Given the general trend of coronal holes towards higher latitudes, observing significantly less CHW in the ecliptic plane is not unexpected. Furthermore, seeing that the CHW in the ecliptic plane is generally slower is in line with the idea that the fast wind is slowed down due to stream-stream interactions in the solar wind. The difference in the unclassified data is almost exclusively related to the amplitude of the peak. The means and inter-quartile ranges show good agreement.

To better understand the double peak in the *Ulysses* CHW speed distribution, the data are further split by spacecraft location. We use a threshold of 600 kms^{-1} to separate the two CHW peaks. Figure 4.7 shows the occurrence of the two CHW distributions as a function of radial distance from the Sun. Given the long orbital duration of *Ulysses* and its associated latitudinal variation, the data the spacecraft obtains is convolved with the solar cycle and latitude. Thus, the average latitude and sunspot number (and the respective standard deviations) are calculated for each histogram bin. The sunspot number shows very little structure, and no overall trends matching the distributions of radial distance shown. In contrast, there is a clear trend between the absolute heliospheric latitude and the radial distances contained within the secondary peak (lower speed) of the CHW speed distribution. The trend suggests that the majority of the secondary peak data is obtained both far from the Sun and closer to the ecliptic plane.

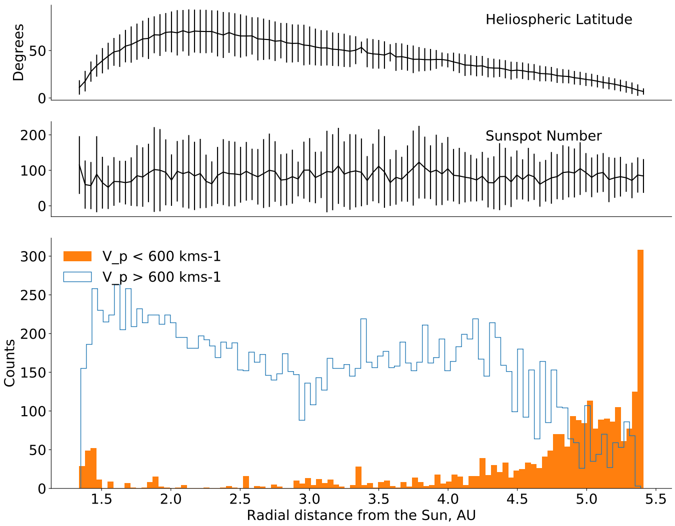


Figure 4.7: *Occurrence of high- and low-speed CHW (using a 600 kms^{-1} threshold) as a function of radial distance. This speed threshold is chosen to isolate the two peaks observed in the distribution of speed in the CHW classification of Ulysses data. The top two panels present the average absolute heliospheric latitude of the Ulysses spacecraft, and the average sunspot number within each bin of the histogram. The error-bars are the standard deviation of the data contained in each bin.*

4.6 The UMAP Scheme

The BGM scheme presents a step forward in creating a classification scheme which is more objective and physically motivated. However, there remain some drawbacks: only a subset of the possible parameters are used; there is an inherent assumption that the data are normally distributed; the number of component Gaussians must be specified in advance (reducing the objectivity of the scheme); moving to higher dimensions reduces the interpretability of the results (what does a six-dimensional (6D) Gaussian look like or mean?); and, simply having three Gaussians does not provide much information about the substructure within each cluster. These issues are addressed by creating a further classification scheme using dimensional reduction and clustering. This scheme will specifically address the subjectivity introduced when designating decision boundaries by-eye. Further, it will remove the subjectivity in determining the number of types of solar wind by deriving the number of clusters from the latent structures in the data itself. We will apply the UMAP algorithm for dimension reduction, and the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm to subsequently cluster the low-dimensional representation of the data.

Datasets are often expressed in terms of a large number of measurements or features. This means that each sample in the dataset is expressed as a vector, or point, in a high dimensional space. It is often the case that the underlying structure of the dataset as a whole can be described in terms of a much smaller number of latent features, which dimensional reduction seeks to determine. More formally, while the ambient space in which a dataset lives may be high dimensional there often exists a much lower dimensional manifold from which the data samples are (noisily) drawn. The UMAP algorithm [McInnes et al., 2018] seeks to learn the topological structure of this manifold, and then find a low dimensional representation of the data that has an equivalent topological structure. In this way UMAP can transform highly complex datasets into much simpler representations that still capture meaningful structural features of the original dataset. Due to the algorithm using stochastic gradient descent [Kusher and Yin, 2003], there are minor variations in results produced by UMAP each time it is performed. For further technical information, see Appendix B.

The HDBSCAN algorithm [Campello et al., 2013] seeks to find dense regions (clusters) of a dataset that are otherwise separated from the rest of the data by regions where data are sparse. In particular it seeks to do this even when the dataset contains background noise. To achieve this, HDBSCAN makes use of a density threshold (expressed as a minimum number of data samples required before a region can be considered "dense") and constructs a hierarchical tree of contiguous regions of density. Given a minimum size for a cluster, this tree can then be simplified resulting in a nested hierarchy of clusters. By selecting out the most persistent such clusters (over ranges of distance scales) a single flat clustering may be extracted. This results in an output of cluster labels where each point is either labelled with a cluster identity, or as noise. For further technical information, see the Appendix C.

It is worth noting that the nature of the way UMAP works is almost guaranteed to result in non-convex clusters, and hence a clustering technique that is robust to this is required. By necessity this essentially means either a hierarchical method such as single linkage [Florek et al., 1951] or average linkage [Sokal and Michener, 1958], or a density based technique such as Density-based spatial clustering of applications with noise (DBSCAN) [Ester et al., 1996] or Mean Shift [Fukunaga and Hostetler, 1975] is required. That one should therefore consider the hybrid hierarchical density based approach of HDBSCAN, over more simplistic methods such as k-means, is entirely natural.

UMAP does not limit the number of parameters that can be used. We apply UMAP to all of the non-evolving parameters (O^{7+}/O^{6+}, C^{6+}/C^{5+}, Fe/O, $< q_{Fe} >$, He^{2+}/H^{1+} and S_p). This 6D data-structure is projected into 2D, allowing subsequent clustering to be independent of any potential user-biases, since there is not a physical interpretation of the reduced dimension axes [McInnes et al., 2018]. Whilst the axes are some non-linear function of the input dimensions, it is not possible to derive this function from the mapping. HDBSCAN clustering does not require any user specified number of clusters, instead finding groupings by the intrinsic density structures present in the data.

As with the BGM scheme, the UMAP classifications are determined using the *Ulysses* fast latitude-scan data. The data is normalised using the MinMaxScaler function available in scikit-learn, as shown in the documentation for UMAP. This method individually normalises all of the parameters to be in the 0 – 1 range:

$$X_{scaled} = \frac{X_i - min\{X\}}{max\{X\} - min\{X\}} \tag{4.2}$$

where X_{scaled} is the scaled value, X_i is the un-scaled data value, and $min\{X\}$

and $max\{X\}$ are the corresponding minimum and maximum values of that parameter. It is of no consequence that different normalisation schemes are used between the UMAP and BGM schemes, since comparison is only made in real space. Normalisation is simply a tool to facilitate unbiased dimension reduction and classification on a per-scheme basis. The normalised dataset is then reduced and clustered. The function mapping from 6D to 2D is stored, as are the classification parameters obtained by HDBSCAN.

Figure 4.8a presents the results of reducing and clustering the latitude scan data. Figures 4.8b and 4.8c present the clustering projected into the O^{7+}/O^{6+} and S_p, and proton speed and temperature spaces, respectively.

To stress a point, the dimension reduction is simply one of the steps required to build the classification scheme. The lack of physical interpretation of the 2D space axes adds to the validity of the results, rather than facilitating the influence of current scientific ideas on the classification. The classification becomes entirely based on the latent structure in the data and as such, independent of biases or expectations we may hold.

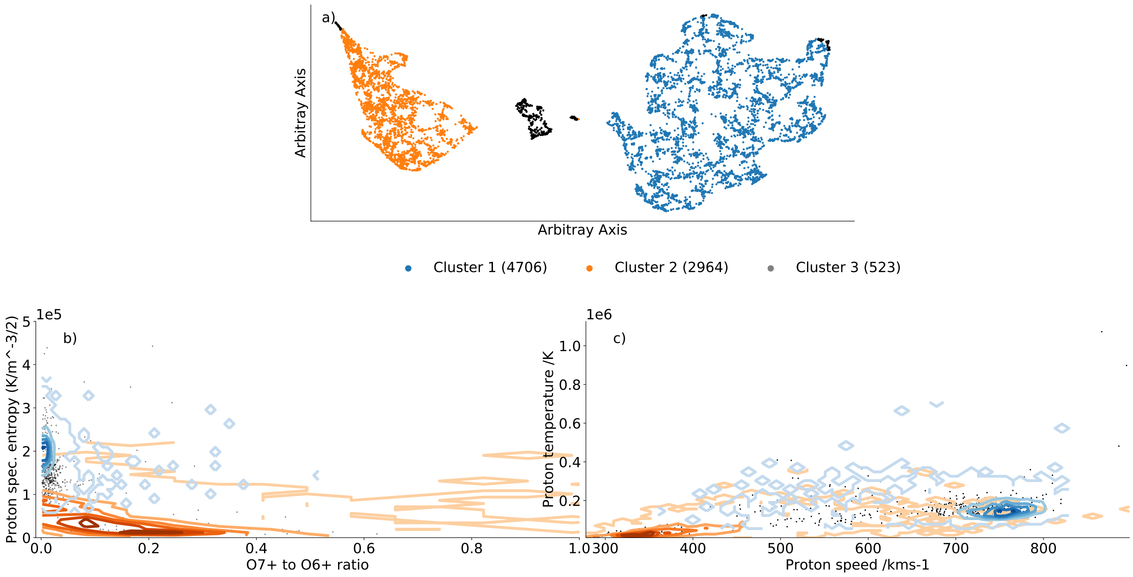


Figure 4.8: *The results of performing the UMAP dimension reduction and HDBSCAN clustering on the Ulysses fast latitude-scan data. Panel a presents the dimensional reduction and subsequent clustering of the non-time-evolving solar wind parameters. Panel b presents the clustering of the reduced data projected onto the* O^{7+}/O^{6+} *and* S_p *space used in the previous two classification schemes. Panel c presents the clustering projected onto solar wind speed and proton temperature. From the latter panels it is inferred that clusters one, two, and three represent CHW, SBW and unclassified data, respectively. The contours are representative of the point-density of data. Given the two different contours and that the data remains largely the same as in Figure 4.2, colourbars are not included. The noisy, final contour line is at the one-point level.*

The data in Figure 4.8a shows distinct groupings in the data. This implies that there are fundamental differences between the three groups in the 6D space. By inspecting panels b and c it is apparent that the distinction is the type of solar wind present. From panel b we infer that cluster one is CHW, cluster two is SBW, and cluster three is the unclassified data. However, the unclassified data is more complicated than with the BGM scheme. Here, there are four distinct regions. Because the UMAP reduction preserves the latent structure of the higher dimensional space, the isolation of the groupings provides information. The spatial separation between the unclassified data associated with the CHW, SBW or middle cluster suggests that these are fundamentally different from one another. However, within-cluster separation of the unclassified data does not necessarily imply fundamental differences, since this could just be an artefact of imperfect projection of the 6D structure onto the 2D plane.

The distribution of the unclassified data is different in each of panels b and c. In panel b much of the unclassified data is grouped in the region where the CHW and SBW signatures overlap. This is expected due to the region being where the parameter values transition between the two types of solar origin, and as such classification uncertainty should exist. The remaining unclassified data which is spread throughout the CHW and SBW is due to the small pockets of unclassified data connected to each of the clusters in panel a. In panel c the unclassified data is more evenly spread around the core regions of each group. Had the unclassified data been grouped where the faster and slower regions overlapped, it would have suggested that the speed could be providing useful information about the types of solar wind present. Since it was not, the stream speed of the solar wind appears to be a less-informative parameter for classification schemes such as this one.

Since the UMAP reduction aims to maintain the structures in the 6D space, one can extract information based on the structures present in the data. The CHW group shows more spread in its internal structure, despite being understood to be less variable than SBW. This could suggest that there is some underlying variability in the CHW's parameter space, or that the manifold covering the CHW is a shape that does not lend itself to 2D reduction (*e.g.* a spherical manifold is difficult to project into 2D whilst retaining the topological structure).

4.6.1 UMAP Scheme: Application

The BGM scheme allows a simple way of classifying new data (the probabilities of each Gaussian giving the data). Using UMAP and HDBSCAN is not quite as straightforward. Fortunately, both techniques (as well as the scikit-learn MinMaxScaler) allow for the created mappings to be stored and applied to new data. In this way, reducing and clustering the whole *Ulysses* and *ACE* datasets is both straightforward and consistent.

4.6.1.1 *Ulysses*

Figure 4.9a presents the results of projecting the entire *Ulysses* dataset to the reduced-dimension space and the subsequent clustering. The data maintains the structure found in the reduction of the fast latitude-scans. However, there are larger pockets of unclassified data, as well as significant linkage between the central unclassified data and clusters one and two. These features are likely to be the cause of the increased amount of unclassified data dispersed throughout the CHW in Figure 4.9b. There is proportionally more unclassified data present than there was in the original fast latitude-scan result. However, as shown in previous sections, the data being used is expected to be more variable and thus, may exhibit more unclassified data. Again, this scheme yields large disparities with the speed-threshold scheme; ≈20% (≈10% excluding the unclassified data).

4.6.1.2 *ACE*

The UMAP classification scheme is now applied to data from the *ACE* spacecraft; Figure 4.10 shows the results. Whereas the classification of the *Ulysses* data resulted in an almost even split between CHW and SBW with some unclassified data interspersed throughout, the *ACE* results shown in Figure 4.10a are dominated by SBW, with only a small fraction being CHW or unclassified data. Comparing with the BGM scheme, the UMAP scheme identifies approximately half (proportionally) as much CHW.

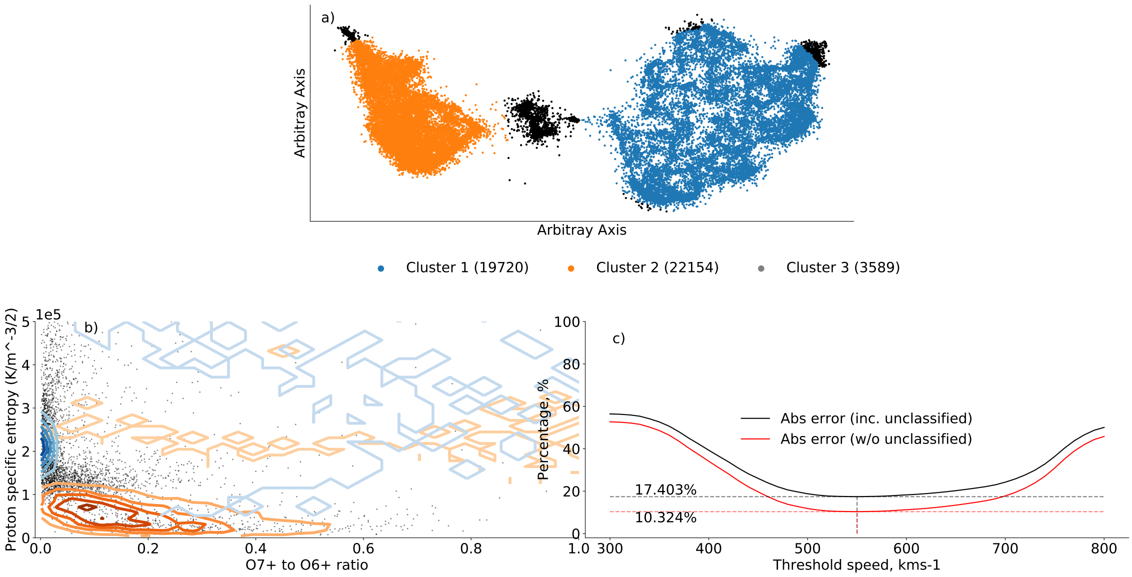


Figure 4.9: *The results of classifying the full Ulysses dataset using the determined UMAP classification scheme. Panel a presents the whole Ulysses dataset reduced to 2D, and the results of the subsequent mapping to the clustering model created using the latitude-scan data. Panel b presents the clustering of the reduced data projected into O^{7+}/O^{6+} and S_p space. The contours are representative of the point density of data as in Figure 4.8, showing similar data to Figure 4.3. Panel c presents the comparison between the UMAP classification scheme and the traditional speed-threshold scheme. Clusters one, two, and three represent CHW, SBW and unclassified data, respectively.*

Comparing to the speed-threshold method, we see that the UMAP classification of ACE data has a disparity of ≈8% (≈4% excluding unclassified data). This suggests closer agreement of UMAP with the traditional method than any of the other classifications. Such a low disparity is promising for the speed-threshold method. However, taking a threshold for skewed data may not be a fair way to split the data. Taking a speed threshold above any value found in the data gives a prediction error rate of ≈9%, simply due to small ratio of CHW and unclassified data to SBW.

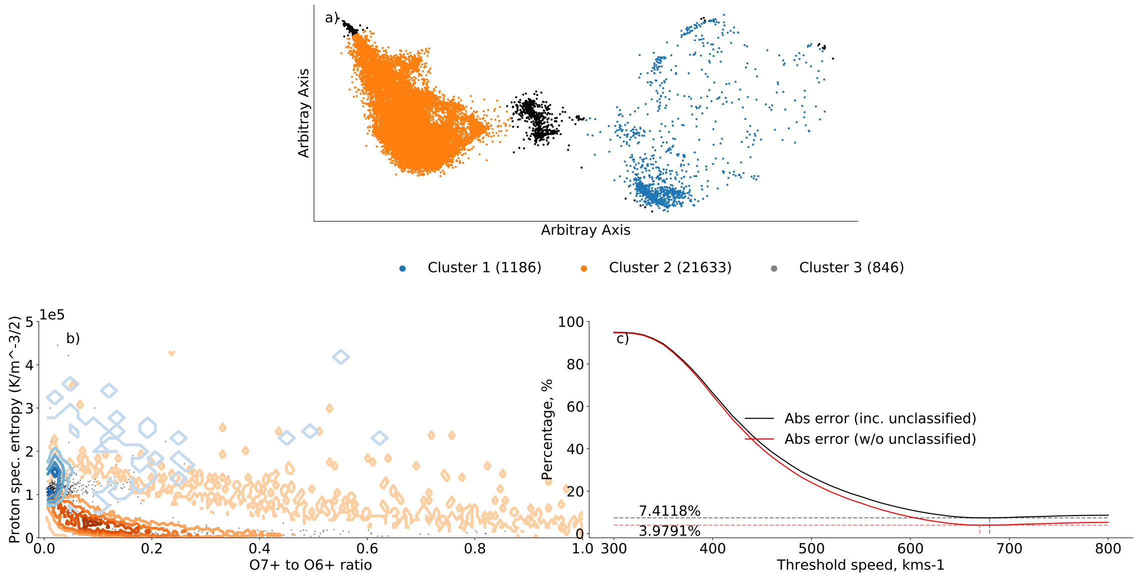


Figure 4.10: *The results of classifying the ACE dataset using the determined UMAP classification scheme. Panel a presents the whole ACE dataset reduced to 2D, and the results of the subsequent mapping to the clustering model created using the fast latitude-scan data. Panel b presents the clustering of the reduced data projected into* O^{7+}/O^{6+} *and* S_p *space. The contours are representative of the point density of data as in Figure 4.8, showing similar data to Figure 4.4. Panel c presents the comparison between the UMAP classification scheme and the traditional speed-threshold scheme. Clusters one, two, and three represent CHW, SBW and unclassified data, respectively.*

4.6.2 UMAP Scheme: Analysis

To further validate the UMAP scheme, radial plots are shown in Figure 4.11. As before, the plots show that the classification scheme captures the overall speed and latitudinal dependence in the data (despite neither being used in the classification scheme itself). Whilst there is good agreement between the UMAP and BGM radial plots (Figures 4.11a and 4.5a, respectively), it is the differences which are interesting. The UMAP scheme unclassified data is more uniformly distributed and there is an increased amount of CHW at lower speeds as compared with the BGM scheme results. The first point is further evidenced in Figure 4.11b where the fraction of unclassified data is clearly more evenly spread throughout the octet of bins. This implies that the UMAP scheme is more able to classify points at the aphelion of the orbit in the ecliptic plane.

Figure 4.12 shows the distributions of the different solar wind classifications as a function of solar wind speed. The CHW and SBW distributions of the *Ulysses* data in panel a match well with the distributions found from the BGM classification, including the secondary peak in the CHW. However, the unclassified data exhibits

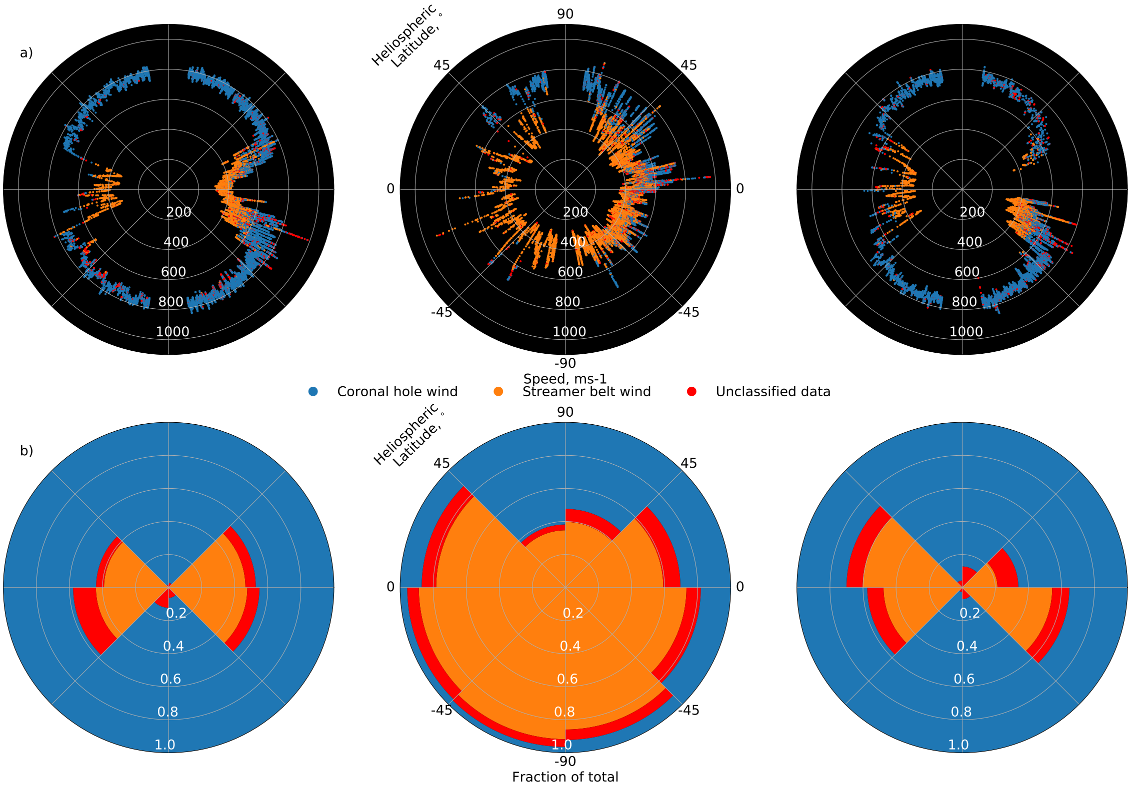


Figure 4.11: *Plots showing how the UMAP classification of the whole Ulysses dataset maps to solar wind speed and solar latitude across its three orbits, as well as how the distribution of unclassified data changes over these orbits. Panels a and b are presented identically to those for the BGM scheme in Figure 4.5.*

a more significantly bi-modal distribution. The peaks of the bi-modality align with the peak of the SBW and primary peak of the CHW. This may suggest that the unclassified data is comprised of points which lie in the tails of the CHW and SBW 6D distributions. Furthermore, the *ACE* data distribution matches the equivalent BGM result, despite displaying a heavy-tail distribution. The Unclassified data, again, displays a bi-modal distribution, and the CHW is too sparse to make a fair comparison of anything but the predicted occurrence rate.

The comparative plots in Figures 4.12c, 4.12d and 4.12e present the differences between the distributions more clearly. In panel c the overall similarity is clear, though the *ACE* distribution shows the heavy tail. In panel d the peaks of the unclassified *ACE* data are shifted towards lower speeds. Finally, panel e shows further expected behaviour; the CHW from *ACE* is slower and a relatively minor contribution.

As with the BGM scheme, the double peak in CHW has been investigated, yielding results which are qualitatively equivalent to the results presented in Figure 4.7.

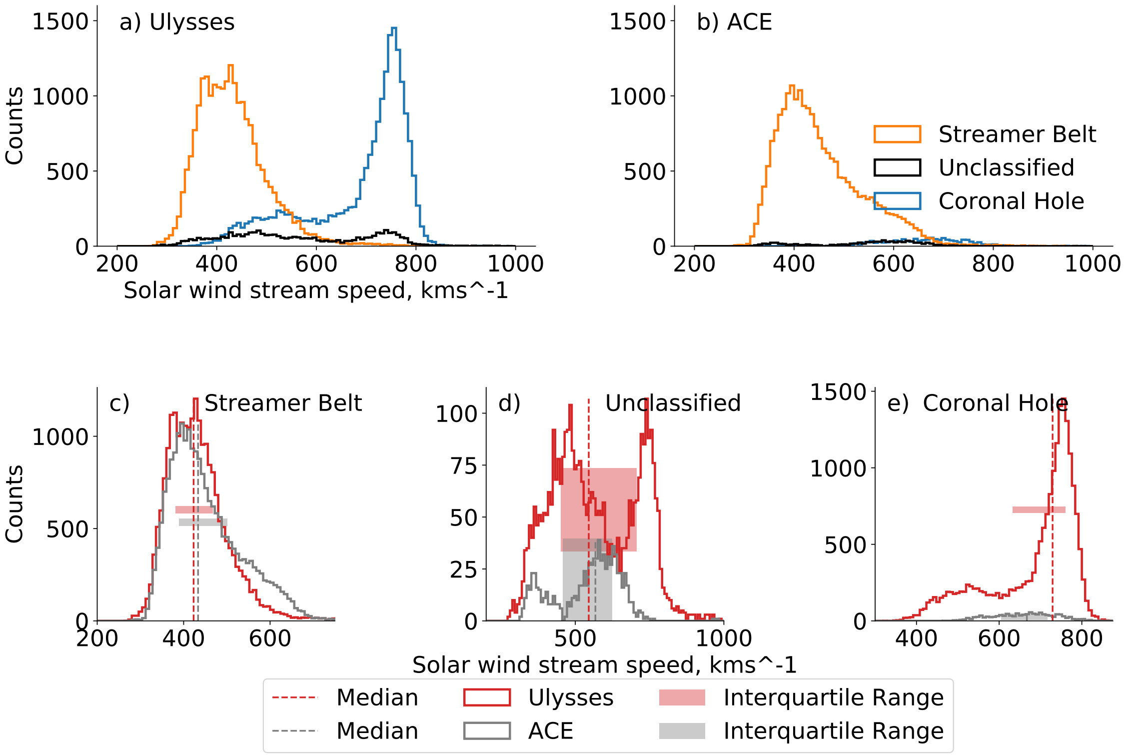


Figure 4.12: *The distributions of solar wind speeds within each UMAP classification. Panels a and b present the comparisons for the distributions in the Ulysses and ACE datasets respectively. Panels c, d and e present the comparisons of like-clusters between datasets. The latter panels also include the mean and inter-quartile ranges of each distribution.*

4.7 Discussion

The intuitive classification scheme, wherein an arbitrary threshold is applied to non-evolving solar wind parameters (such as ion charge states ratios and proton entropy), is a bridge between simple solar wind speed-threshold classification and the machine learning methods presented herein. It shows significant differences to the speed-threshold method. The latter method is not without its merits, for many applications the speed of the solar wind is the driving factor. However, in situations where the solar source is important, simply splitting the solar wind up using arbitrary speeds may be misleading. The disparity of ≈10% to ≈11% between the intuitive scheme and the speed-threshold scheme highlights the potential flaws in statistical analyses performed using solar wind data. The two intuitive scheme classification boundaries in Figures 4.1a and 4.1b are quite different. One has twice the gradient of the other in linear space, yet they produce similar results in terms of coronal hole and streamer belt winds, suggesting a degree of robustness in the classifications. However, this approach is entirely deterministic and there is no means to assess uncertain or difficult-to-classify solar wind intervals.

The BGM scheme mathematically extends the intuitive scheme using the same parameters, O^{7+}/O^{6+} and S_p. Instead of using visual inspection, classification boundaries are derived by optimising the fit of a Gaussian mixture to the *Ulysses* fast latitude-scan data. This method also allows for the inclusion of a third category: unclassified data. The stability of the classification is assessed through repeated trials on sub-samples of the data, and found to be robust. Applying the classification scheme to the whole of the *ACE* and *Ulysses* datasets shows, again, significant disparities with the speed-threshold method: ≈18% and ≈22%, respectively. As expected, much less CHW is found in the *ACE* dataset than the full *Ulysses* dataset.

The unclassified *Ulysses* data was found to be skewed towards the aphelion of the orbit. This could be indicative of the increased time that turbulence or solar wind stream interactions have to develop before reaching *Ulysses*. It may be that the assumption of no plasma mixing breaks down on these long timescales, either as a result of differential streaming of ions [Marsch, 2006; Schwadron et al., 2005] or magnetic reconnection [Gosling, 2012].

The similarity of the speed distributions, Figure 4.6, in the *Ulysses* and *ACE* SBW suggests repeatable classification despite the different occurrence density in the two datasets. The slightly higher mean speed for SBW at *Ulysses* compared to *ACE* is consistent with the increased radial distance and hence continued acceleration and/or interaction time with faster CHW. The CHW distributions, however, show little similarity. This shows there are quantitative differences in the speed of CHW streams in and out of the ecliptic plane. A low speed CHW population is found primarily at the aphelion of the orbit and perihelion at low latitudes (the aphelion data is also generally closer to the ecliptic plane). This could be a further result of the factors causing the unclassified data to be skewed towards the aphelion of the orbit (*e.g.* turbulence, stream interactions, differential streaming and magnetic reconnection).

The UMAP scheme builds on the principles of the previous schemes: choosing non-evolving parameters for classification, and using the fast latitude-scan data to establish the classifications. Unlike the BGM scheme, there is no user-specified number of categories to discover, nor is the distribution of data assumed to approximate a multivariate Gaussian. The UMAP algorithm takes six non-evolving parameters (O^{7+}/O^{6+}, C^{6+}/C^{5+}, Fe/O, $< q_{Fe} >$, He^{2+}/H^{1+}, and S_p) and approximates the latent structure of the space into a 2D representation. The reduced data presented in Figure 4.8a shows that there are two primary groupings in the data, as well as another small grouping. The clusters in the data are extracted using HDBSCAN, a density based clustering algorithm. By mapping the clusters from the arbitrary 2D space into O^{7+}/O^{6+} and S_p, Figure 4.8b, it is clear that the groupings in the data match well with the expected properties of CHW and SBW. Such information is determined using accepted domain-specific knowledge about the solar wind (*e.g.* CHW is generally cooler and has higher S_p). Furthermore, the majority of the unclassified data is found in the boundary region between the CHW and SBW, supporting the idea that it is difficult to definitively classify, especially in a lower dimensional representation.

When applying the UMAP classification scheme to the whole *Ulysses* data, Figure 4.9a, the two primary clusters become saturated with data points. There is a proportional increase in the amount of both CHW and unclassified data. As with the BGM scheme, there is a large disparity in the comparison with taking speed thresholds: ≈20%. The application of the UMAP scheme to the *ACE* dataset, Figure 4.10a, shows a lack of CHW. Interestingly, in Figure 4.10c we see much better agreement between the speed-threshold classification and the results of the UMAP scheme on the *ACE* data: ≈8%. However, the results are not much better than just classifying everything as SBW, as one may expect with such skewed data. Despite the link between the CHW and unclassified data, the latter shows very little dependence on the orbital position in Figure 4.11b.

Comparing the speed distributions of each class, we see qualitative similarities with the results of the BGM scheme: the SBW speed distributions match well, though the *ACE* distribution displays a heavy tail, and the CHW is bi-modal. Different however, are the unclassified data distributions, which are also bi-modal. This highlights the different ways in which the unclassified data is characterised in the two schemes. The UMAP results show double-peaks in speed close to the peaks of the CHW and SBW. This suggests that the unclassified data may be data which belongs to one or other of the distributions, whose parameters deviate from their respective norm.

In Figure 4.12d, the peaks of the unclassified *ACE* data are shifted towards lower speeds. This is contrary to the expectation of the slower stream to be sped up and the faster stream slowed down, due to stream interactions. As such, the data may signify that the unclassified data comprises solar wind transients (all slowed due to the increase in SBW in the ecliptic plane). Else, there may be a process acting to generally slow the unclassified data found in the ecliptic plane.

The heavy-tail distribution of the *ACE* SBW (classified by UMAP) may suggest that a more complex model is required to characterise all of the structure in the data. Alternatively, it may highlight the presence of another process which accelerates SBW in the ecliptic plane, such as interactions with faster CHW or the inclusion of more solar wind transients (*e.g.* ICMEs). Investigating the slower CHW peak in the *Ulysses* data produces the same results as for the BGM scheme. The data within the peak is largely from the aphelion of the orbit.

Tables 4.2 and 4.3 show the distributions of the classification results from some of the papers discussed in the introduction. This, again, draws attention to the lack of consensus on how the solar wind should be classified. Note that these comparisons relate only to proportions, since the results are not all obtained from the same data. Direct comparison of classifications for the same data are given in Table 4.4.

SW Type	S15	BGM	UMAP
CHW		≈38.8%	≈40.3%
SBW		≈44.7%	≈48.3%
Unclassified		≈16.5%	≈11.3%
Fast	≈20%		
Slow	≈65%		
Intermediate	≈15%		

Table 4.2: The proportions of each solar wind type found when classifying *Ulysses* data. S15 refers to the results of Stakhiv et al. [2015]. BGM and UMAP refer to the results of the presented classification schemes.

Of the classification schemes mentioned, only Stakhiv et al. [2015] (S15) have results for classifying *Ulysses* data. The results in Table 4.2 have been estimated from their Figure 5. These results show less fast and more slow wind than we find of our comparable CHW and SBW, respectively. However, these differences are reduced if we account for the errors we predict for taking such speed thresholds.

Both of Zhao et al. [2009] and Zhao et al. [2017] (Z09 and Z17) have results from classifying *ACE* data, though the latter paper is more difficult to compare given its six-type classification scheme. The results in Table 4.3 are estimated from Figure 1 of Z09 and Figure 6 in Z17. The Z09 results do not match well with the CHW or SBW results of either the BGM or UMAP scheme. However, the ICME value

SW Type	Z09	Z17	X15	C17	BGM	UMAP
CHW	58%	8.2%	25.2%	20.4%	8.4%	4.8%
SBW			41.7%	27.6%	78.9%	90.9%
Unclassified					12.7%	4.4%
Non-CHW	37%					
ICME	5%					
CH-boundary		10.2%				
Quiet Sun		25.6%				
AR		31.1%				
AR boundary		10.8%				
Helmet streamer		13%				
Ejecta			12.9%	13.9%		
SR region			20.2%	38.0%		

Table 4.3: The approximate proportions of each solar wind type found when classifying *ACE* or *OMNI* data. The column labels refer to Zhao et al. [2009], Zhao et al. [2017], Xu and Borovsky [2015], and Camporeale et al. [2017], respectively. BGM and UMAP refer to the results of the presented classification schemes. AR refers to active-regions and SR refers to Sector Reversal Regions.

is not dissimilar to UMAP's unclassified data, possibly supporting the idea that the unclassified data (especially in the UMAP scheme) could be composed of ICMEs and other transients. Z17's pure CHW shows good agreement with our results (especially from the BGM classification). If the CH-boundary class is taken as a part of the CHW, then the agreement diminishes. The rest of the classifications are not usefully comparable due to the differences to our scheme.

Both Xu and Borovsky [2015] (X15) and Camporeale et al. [2017] (C17) apply their classifications to *OMNI* data [King and Papitashvili, 2005]. This incorporates *ACE* data as well as other data from L1, allowing comparison. The results in Table 4.3 are taken from Table 3 in X15, and the results of the C17 classification (see their acknowledgements for data location). The X15 results differ from those found with our schemes. However, if we consider that the sector-reversal region solar wind is a part of our SBW, then there is some agreement between these results and those from the BGM classification (the UMAP classification still differs significantly). The C17 results differ slightly from those of X15, and present more agreement with our results.

To compare some of these results in a more rigorous way, Table 4.4 presents confusion matrices (contingency tables) comparing the results of two of the schemes on the same data. The Z09 and C17 results have been simplified by assuming the non-CHW is equivalent to our SBW, and combining the SBW and sector-reversal region wind, respectively.

Most noteworthy of these results is the agreement between what the BGM and UMAP schemes classify as CHW compared with the Z09 and C17 schemes. This is exemplified by the horizontal rows of CHW from the BGM and UMAP. In these rows the proportion of what our schemes classify as CHW and the other schemes classify otherwise (reading along horizontally) is very low. However, reading vertically along the CHW columns, we see there are many samples in other columns. This suggests that our schemes (trained on the *Ulysses* data) provide accurate-but-conservative classifications of CHW as compared to the other models.

Comparison between each of our (BGM and UMAP) SBW classifications to Z09's and C17's schemes present less consistent results. The BGM:C17 and UMAP:C17 results are broadly in agreement, with the majority of our SBW also being classed as SBW by the other schemes. In contrast, the Z09 scheme classifies the majority of solar wind as CHW. Hence, the results of our classifications are in conflict (as are the Z09 in conflict with all of the others in Table 4.3).

Given the disparate methods of determining the ejecta/unclassified wind, it is unsurprising that there is little agreement between any of the schemes (save some between the BGM and Z09 schemes). Using a broadened feature space for the UMAP scheme and identifying an unclassified cluster not found in the BGM scheme highlights the importance of applying domain-specific knowledge, even in data-driven approaches.

The inter-comparison between the BGM and UMAP schemes quantifies the evident differences and similarities between the two methods. As one may expect from the comparisons of speed distributions, the SBW is in good agreement. However, the CHW is more diverse. Given the larger feature-space, and less constraining method of clustering, we would posit that the UMAP CHW is a more accurate representation of the class. It is more difficult to comment on the accuracy of the unclassified wind from UMAP given that there are contributions from other areas of the feature space.

We acknowledge that there may be some systematic bias in the classifications of *ACE* data. It is possible that by limiting the training set to the *Ulysses* latitude-scans, we created classification boundaries which generalise less well to the *ACE* data (despite our use of non-evolving parameters). One potential source of this may be that our training data heavily samples very large polar-coronal holes. As such, in the ecliptic plane where we see generally smaller coronal holes, and are more likely to sample boundary regions (see the percentages of CHW and

Table 4.4: Confusion matrices (contingency tables) intercomparing the classification results of various schemes on *ACE* or *OMNI* data. The column labels refer to Zhao et al. [2009] and Camporeale et al. [2017]. BGM and UMAP refer to the presented classification schemes. C17's results are obtained by matching timestamps between the data provided in their paper with those in either the BGM or UMAP scheme. Z09's results are obtained by applying the classification criteria from their paper to the data used in the BGM or UMAP scheme. The green highlighting serves to draw attention to the diagonals, which represent the number of samples of agreed classifications between any two schemes.

	UMAP				C17				Z09			
		CHW	SBW	Ej/Unc		CHW	SBW	Ej/Unc		CHW	SBW	Ej/Unc
BGM	CHW	587	892	392	CHW	2936	181	71	CHW	3194	0	0
	SBW	71	17668	302	SBW	3585	22677	3496	SBW	15425	13620	1017
	Ej/Unc	527	3023	152	Ej/Unc	1856	1522	1421	Ej/Unc	626	514	3712
						CHW	SBW	Ej/Unc		CHW	SBW	Ej/Unc
				UMAP	CHW	1129	18	38	CHW	1092	17	15
					SBW	4549	13774	3286	SBW	11725	8106	1678
					Ej/Unc	560	162	124	Ej/Unc	772	220	40
										CHW	SBW	Ej/Unc
								C17	CHW	6295	390	1692
									SBW	11450	11507	1423
									Ej/Unc	1396	2019	1573

CH-boundary wind in the Z17 column of Table 4.3), the algorithms may classify such winds as SBW mistakenly.

Whilst our choice of training data may bias the classification, the benefits of training on out-of-ecliptic data which samples almost the entire range of heliospheric latitudes are significant: a more complete range of solar wind is sampled, and that wind is less likely to be interfered with by processes relating to stream interaction. Furthermore, it should be noted that discovering results which differ from the norm when using novel techniques does not necessarily mean the results are wrong. It could very well be the case that there is less CHW in the ecliptic plane than current classifications recognise.

4.8 Conclusions

This work presents two novel, data-driven schemes to classify the solar origin of solar wind streams using unsupervised machine learning. The schemes are built using non-evolving parameters which retain information about the source regions. Each classification model is created using the *Ulysses* fast latitude-scan data, before being applied to the whole *Ulysses* and *ACE* datasets. The BGM scheme reduces the subjectivity in determining classification boundaries between solar wind types. It was specified to fit three clusters in the solar wind data. As expected, two of these are the coronal hole and streamer belt winds. The third remains unclassified. The UMAP scheme addresses subjectivity in the choices of decision boundaries and the number of clusters to find in the data; it independently derives three clusters in the latent topological structure of the solar wind data. These clusters correspond to coronal hole and streamer belt winds as before, but find a different type of unclassified solar wind. Application of the UMAP scheme to *Ulysses* and *ACE* shows morphological differences in the coronal hole wind seen in and out of the ecliptic plane.

For both schemes, and both spacecraft datasets, the classification results are compared with the traditional approach of taking speed thresholds. In each case, there are significant best case disparities between the speed-threshold approach relative to the machine learning classifications: The BGM scheme applied to *Ulysses*, ≈22% and *ACE*, ≈18%; the UMAP scheme applied to *Ulysses*, ≈20% and *ACE*, ≈8%.

Whilst our results differ from those of other works, our data-driven methods are designed to increase objectivity and reduce the introduction of scientifically subjective biases. Thus, the differences do not take away from the results presented. Instead, such differences should motivate further work investigating objective methods of solar wind classification, and their differences to current schemes.

4.9 Appendices

4.9.1 Bayesian Statistics and Variational Inference

Bayes' theorem is the statistical description of the probability that an event happens, given some prior knowledge of the conditions of the event. Bayes' theorem is notated, for two events A and B, as:

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)},$$

where $P(A \mid B)$ (the posterior probability) is the conditional probability that event A occurs given event B, $P(B \mid A)$ (the likelihood) is the conditional probability that event B occurs given event A, and $P(A)$ (the prior probability) and $P(B)$ (marginal likelihood) are the probabilities of events A and B happening independently. In Bayesian inference, the interpretation of the posterior probability is the degree of belief in a hypothesis. This can be envisioned as a situation where you have a number of Gaussians from which a point measurement may be sampled (*e.g.* a solar wind measurement and Gaussian mixture for its classification). To determine which Gaussian is most likely given the data point, you calculate the posterior using Bayes' theorem, taking each Gaussian, A, given the data point, B, and compare the probabilities for each Gaussian.

Variational inference (used in the BGM) is an extension to Bayesian inference developed for ML and is a current area of research in statistics. A brief description will be given here based on the works of Blei et al. [2017] and Gelman et al. [2013]. Variational inference is a method used to approximate probability densities through optimisation, rather than sampling techniques (*e.g.* Markov Chain Monte Carlo, MCMC). MCMC is used to create an empirical estimate of the posterior distribution based on collected samples, and is very effective on smaller or more simple models. However, when models are complex or datasets are large a different approach is needed for computational practicality. Variational inference chooses a family of probability density functions (PDFs) as an approximation to the true PDF. The member of the PDF family which minimises the Kullback-Leibler (KL) divergence to the exact posterior is sought (explained further, below). The member which minimises the KL divergence is then optimised and used as the approximate distribution for the posterior distribution. Variational inference is usually faster than MCMC methods and better suited to scaling for large datasets. The drawback is that while MCMC is known to converge asymptotically to the correct solution, variational inference is not. Despite this, Figure 3 and Table 1 in Blei and Jordan [2006], show how variational inference can be much faster, while also remaining competitive to MCMC methods.

The Kullback-Leibler divergence is a measure of how one probability distribution diverges from another (for the derivation and further information, see Kullback and Smith [1978]). For the probability distribution of a continuous random variable

x the Kullback-Leibler divergence, $D_{KL}(P||Q)$ of distribution $Q(x)$ from a given distribution $P(x)$ is defined as:

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) log\left(\frac{p(x)}{q(x)}\right) dx,$$

where $p(x)$ and $q(x)$ are the probability densities of $P(x)$ and $Q(x)$, respectively.

4.9.2 Uniform Manifold Approximation and Projection

Expanding on the description given in Section 6, the following will discuss the methodology of the UMAP dimension reduction technique using the appropriate mathematical terminology. This description is quite involved, but should provide an interested reader with all of the vocabulary needed to further investigate the algorithm. Naturally, the full mathematical description is given in McInnes et al. [2018].

The field of topological data analysis [Carlsson, 2009] uses methods from topology to better understand complex datasets. One such technique is the construction of the Čech complex [Ghrist, 2014] which provides a combinatorial representation of a topological space inferred from a given dataset. To construct the Čech complex one forms a cover given by open balls of a fixed radius about each of the datapoints. The Čech complex is then the simplicial complex [Ghrist, 2014] given by the nerve of that open cover [see, e.g. Ghrist, 2014, for more detail)]. Informally the process proceeds essentially as follows: to each open ball one assigns a point; whenever a pair of open balls have non-empty intersection one joins the corresponding points with line segment; whenever three open balls share a non-empty intersection one adds a filled triangle joining the points; and so on, adding higher dimension pieces for more complex intersections. By the nerve theorem [Borsuk, 1948], the resulting simplicial complex is homotopy equivalent [Ghrist, 2014] to the manifold formed by the union of the open cover. Informally, the topological space pieced together by points, lines, triangles, tetrahedrons, etc., captures the same fundamental topological structure as the space being covered by open balls. In this manner manifold structure latent in data can be discovered.

Unfortunately this will only successfully capture the underlying manifold from which the data was drawn when the data samples are uniformly distributed on the manifold. Since this is rarely the case for data under the ambient metric we must instead use the existing data distribution to infer the Riemannian metric on the manifold that would result in such a uniform distribution. This can be done by examining local data distributions and approximating a locally constant Riemannian metric at each point. While this recovers the uniform distribution assumption it introduces a new difficulty in that the metric spaces local to each point are mutually incompatible.

By translating the local metric spaces into fuzzy simplicial sets [see Spivak, 2012; Goerss and Jardine, 2009] the incompatibility can be overcome by taking the union of the entire family of fuzzy simplicial sets. The result is a single fuzzy simplicial set that provides a coherent view of the topological structure of the underlying manifold from which the data was sampled. UMAP then uses an optimisation process to find a low dimensional representation of the data that has a fuzzy simplicial set representation that matches the topological representation of the source dataset as closely as possible.

4.9.3 Hierarchical Density Based Spatial Clustering for Applications with Noise

As with Appendix B, we present an extension to the description of HDBSCAN given in Section 6. Again, it is quite involved, but should be of interest to those familiar with ML or who wish to learn more.

A dataset of measurements can be assumed to have been (noisily) sampled from some probability density function. ‘Noisily sampled’ in this case refers to sampling a value when there is noise (e.g. the inherent uncertainty in spacecraft measurements). Given a probability density function f, where $f(x)$ is the likelihood of sampling a point x and $\int_{\mathbb{R}^n} f(x)dx = 1$, one can consider the level sets $\{x \in \mathbb{R}^n \mid f(x) \geq \lambda\}$. As $\lambda \geq 0$ varies these level sets will nest in such a way as to form an infinite tree, called the *cluster tree*. Each cluster is a branch of the tree, extending over the range of λ values for which it is distinct. The goal of hierarchical density based clustering algorithms is to approximate this cluster tree given only a finite set of sampled data.

Hierarchical clustering techniques such as single linkage clustering [Everitt et al., 2011] provide a partial solution. Results by Hartigan [1981] demonstrate consistency with the cluster tree for single linkage clustering in the case of 1D data. In higher dimensions, however, single linkage clustering becomes too sensitive to noise: it suffers from chaining effects where spurious points result in clusters merging prematurely. To remedy this we need to introduce a notion of density. Let $X = \{x_1, x_2, \ldots, x_N\} \subset \mathbb{R}^n$ be the dataset, and define the *core-distance* $\kappa(x_i)$ of a point x_i as the distance to the k^{th} nearest neighbour of x_i. The core-distance can act as a proxy for density (since sparse areas of the sample space will have larger core-distances). We can then define a new metric, called mutual-reachability-distance, defined as

$$d(x_i, x_j) = \begin{cases} \max\{\kappa(x_i), \kappa(x_j), \|x_i - x_j\|_2\} & x_i \neq x_j \\ 0 & x_i = x_j \end{cases}.$$

In effect the mutual-reachability distance between a pair of points is the smallest distance scale at which both points will be dense and considered to be neighbouring each other. Performing single linkage clustering under this new density-sensitive metric yields a more robust clustering algorithm that can be shown to converge to the cluster tree of the probability density function from which the data was drawn Eldridge et al. [2015].

The resulting cluster hierarchy is often exceptionally complex. Much of the complexity is the result of single, or small numbers of, points separating off into new clusters. To simplify the resulting cluster hierarchy we can consider a minimum allowable cluster size m. We can then re-process the hierarchy considering any child cluster with fewer than m points to be spurious – we denote those points as “falling out of the parent cluster”. The resulting simplified tree allows for better cluster analysis. A further step can then be taken by selecting those clusters within the tree that persist for the largest ranges of distance scales. This can be posed as a simple optimisation problem using the notion of relative-excess-of-mass from probability theory. This allows for the production of a flat clustering where each data point is either assigned a cluster label or, if it fell out of a cluster further up the hierarchy, is labelled as noise.

Chapter 5

Constraining the Location of the Outer Boundary of Earth's Outer Radiation Belt

---

Preface

In Chapter 4 we investigated whether *in situ* solar wind data encoded latent information about its solar source region. We found that it is possible to identify the coronal hole and streamer belt solar wind while minimising the impact of subjective biases.

Such results are useful to the Space Weather community as this latent information may also include proxies for other related quantities which are not measurable, but may help to improve modelling. Alternatively, such classifications may serve as a tool to reduce the dimensionality of down-stream modelling tasks, rather than being required to include parameters relating to the composition of the solar wind.

While Chapter 4 sought to gain new insight into the the solar wind, as the main driver of Space Weather, this chapter and Chapter 6 shift the focus more directly onto Space Weather phenomena. Both of these chapters provide novel insight into the the outer boundary of the outer radiation belt. As will be expanded upon in each chapter, understanding the outer radiation belt is crucial to much of humanity's space-based infrastructure. This chapter specifically targets the identification of the physical location of the outer boundary, while Chapter 6 builds on this results, creating a synthetic dataset of electron spectra at the boundary location - a necessary ingredient to radiation belt models.

Abstract

Characterising the location of the outer boundary of the outer radiation belt is a key aspect of improving radiation belt models and helps to constrain our understanding of the mechanisms by which the source and seed electron populations are transported into the radiation belts. In this paper, we hypothesise that there are statistical differences in the electron distribution function across the radiation belt outer boundary, and thus analyse electron flux data from the THEMIS (Time History of Events and Macroscale Interactions during Substorms) satellites to identify this location. We validate our hypothesis by using modelled electron L* values to approximately characterise the differences between electron distribution functions inside and outside of the radiation belts. Initially, we perform a simple statistical analysis by studying the radial evolution of the electron distribution functions. This approach does not yield a clear discontinuity, thus highlighting the need for more complex statistical treatment of the data. Subsequently, we employ machine learning (with no dependence on radial position or L*) to test a range of candidate outer boundary locations. By analysing the performance of the models at each candidate location, we identify a statistical boundary at ≈ 8 R_E, with results suggesting some variability. This statistical boundary is typically further out than those used in current radiation belt models.

Plain Language Summary

Earth's magnetic field traps highly-energetic particles in a doughnut shaped region, referred to as 'the radiation belts'. Our work focuses on the outer belt, comprised of electrons. Many spacecraft orbit within this region, exposing them to potential damage. To mitigate this, the radiation belts must be understood and modelled. The outer boundary is crucial to modelling, driving changes in radiation belt activity. The boundary is also important because its location helps us to understand which processes form the radiation belts.

In this paper, we analyse electron data measured by satellites to identify the

location of the radiation belt's outer boundary by using simple statistical methods and machine learning. Our results show that simple statistical methods cannot be used to deduce an outer boundary. Using machine learning, we test many candidate boundary locations and by quantifying the model performances at each of these locations, we are able to identify a statistical boundary location. This boundary is located at approximately 8 Earth radii away from the planet, which is typically further out than the boundaries currently used by radiation belt models, although our analysis suggests the boundary location may be variable.

5.1 Introduction

Earth's radiation belts typically manifest as two toroidal regions of magnetically confined, energetic plasma. The outer radiation belt (ORB) comprises a highly dynamic electron population, where fluxes can change by orders of magnitudes on minute timescales [Blake et al., 1992]. The relativistic electrons commonly observed in the ORB pose a threat to spacecraft via surface charging and electrostatic discharges between internal components [Frederickson et al., 1991; Baker, 2001; Eastwood et al., 2017a]. As the well-used geostationary and medium earth orbits overlap with the ORB, there is significant interest in being able to accurately model and forecast its electron properties.

There exist a number of radiation belt models, including: Salammbô [Beutier and Boscher, 1995; Boscher et al., 2000; Bourdarie et al., 2005]; VERB [Versatile Electron Radiation Belt, Subbotin and Shprits, 2009]; STEERB [Storm-Time Evolution of Electron Radiation Belt Su et al., 2010b,a, 2011]; DREAM [Dynamic Radiation Environment Assimilation Model Reeves et al., 2012], and the BAS-RBM [British Antarctic Survey's Radiation Belt Model, Glauert et al., 2014]. One of the critically important aspects of defining the boundary conditions for these models is the outer boundary of the ORB (OBORB), since this boundary acts as a time dependent source for the simulations.

There are two aspects of specifying this boundary condition. Firstly, the location must be specified either in physical or adiabatic invariant coordinates, and secondly the source distribution must be specified for the chosen boundary location. Typically, a boundary location is chosen around geosynchronous orbit or an equivalent position in adiabatic invariant coordinates, and the source distribution is taken from either a model output [e.g., Vette, 1991] or observational data. The model boundary locations used do not necessarily correspond to the physical outer boundary, but instead are chosen to maximise the amount of data available to construct the source distribution (more recently this has been data from geosynchronous orbit or the apogee of the Van Allen Probes mission). Importantly, there may be physical processes outside of the arbitrary, data-maximising boundary location which cannot be included through these modelling approaches. Until radiation belt models capture the entire physics of the radiation belts, they will have difficulty in predicting future behaviour, since they will be limited to using reanalysis of past behaviour rather than being able to fully model the dynamics into the future.

Determining the extent of the outer radiation belt relative to the location of the tail plasma sheet may help to identify mechanisms which may provide the crucial trapped seed population [Jaynes et al., 2015]. Since Earth's plasma sheet is known to be an important source of electrons that ultimately form the radiation belt, though the precise mechanism of transport is not well understood [e.g., Forsyth et al., 2014, 2016; Sergeev et al., 2015].

Given the importance of the OBORB, and the lack of empirical investigation into its location, we here attempt to identify a statistical boundary location. This investigation is built upon the following hypotheses about the ORB and its electron content:

1. The distribution function of the trapped radiation belt electron population differs from the distribution function of the untrapped electrons.
2. There exists statistically - or explicitly - a radial limit at which the distribution functions of trapped and untrapped electrons will diverge.

Here, trapped electrons refer to radiation belt electrons which exhibit closed drifting and bouncing trajectories, as opposed to the untrapped electrons, whose drift paths lead to them being lost to different magnetospheric regions. Distributions functions in this work are as a function of energy. A further point of note is that different distribution functions for the untrapped electrons have been observed between dawn and dusk, due to electrons injected in the midnight sector being lost to the magnetopause without reaching the dusk sector [Li et al., 2010a; Sorathia et al., 2017]. Thus, comparing the differences in the distribution functions between dawn and dusk should allow us to identify the radial extent of the bound electrons more easily.

In Section 5.2 the data and data processing will be discussed. In Section 5.3.1 the current definition of what constitutes the radiation belt (i.e., where a trajectory has a defined L*) is used to set a benchmark for the type of differences between the ORB and untrapped distribution functions. In Section 5.3.2 the statistical radial evolution of the distribution function is presented. In Section 5.3.3, machine learning (ML) is employed as a hypothesis testing tool and a statistical boundary location is found for both the dawn and dusk MLT sectors. Finally, we will summarise and make concluding remarks in Sections 5.4 and 5.5.

5.2 Data

Given that this investigation requires data over a large range of radial distances, we use data from the Time History of Events and Macroscale Interactions during Substorms (THEMIS) spacecraft (publicly available through NASA's CDAWeb archive). The distribution functions are derived from electron flux data from the electrostatic analyser (ESA) to give us the energy range 10 eV to 30 keV and the solid state telescope (SST) to give us the energy range 30 keV to 719 keV [Angelopoulos, 2008; McFadden et al., 2008]. Data is taken from THEMIS probes A, D and E between 2007/09/27 and 2019/09/29, whilst data from probes B and C is taken up till 2010, at which point they were moved to a lunar orbit [Russell and Angelopoulos, 2014]. Note that for the L* analysis in Section 5.3.1, data is only used up until 2017 due to the availability of OMNI data in the SpacePy L* calculator [Morley et al., 2010b]. Qualitatively, this limitation is very unlikely to affect the results.

This investigation will focus on identifying the equatorial boundary location, and will use data from the dawn and dusk MLT sectors. We use the spacecraft's position in GSM co-ordinates to specify dawn and dusk data (6 and 18 ± 3 MLT hours), and we use geomagnetically-aligned (GEOMAG) co-ordinates to specify data from the magnetic equatorial region ($Z = 0 \pm 0.5$ R_E). This latter step is done to ensure that the region we are sampling corresponds to the magnetic equator in the appropriate coordinate system.

To construct the distribution functions for the electrons we convert the direction-averaged differential electron (kinetic) energy flux (DEF, $eV/cm^2 \cdot s \cdot sr \cdot eV$) into phase-space density (PSD, s^3/m^6) as follows:

$$PSD = \frac{DEF \cdot 10^6 \cdot m_e^2}{2E^2} \tag{5.1}$$

where E is the measured energy of electrons (in Joules) and m_e is the rest mass of an electron.

Figure 5.1 presents the equatorial plane (left) and radial (right) distribution of the THEMIS data used. From this, we note that the data is not evenly distributed, but instead has a radial bias with a maximum ≈ 11.5 R$_\text{E}$. This distribution is expected given the orbital parameters of the various spacecraft. Two spacecraft (probes D and E) have their apogee at ≈ 11.5 R$_\text{E}$, meaning that they are travelling most slowly at this region and so the density of measurements is higher. Probes B and C have apogee at ≈ 30 and 19 R$_\text{E}$, and so their measurements of the inner magnetosphere are more spatially sparse. Probe A has an orbit with apogee at ≈ 10 R$_\text{E}$.

In the following analysis, it will be important to ensure that results are not biased by the radial sampling. To address this, we construct ensembles of randomly sub-sampled data. In each of dawn and dusk, we take n equally-spaced radial bins between $5 - 13.5$ R$_\text{E}$ (the amount of available data drops after this radial limit). We find the bin with the fewest samples, m (where $m \approx 3000$ if $n = 20$). We then construct a new dataset by randomly sub-sampling m points from every bin 100 times (with replacement). This new dataset is now uniformly populated in radial distance.

Such ensemble sampling addresses positional biases of the spacecraft measurements. Furthermore, we maintain the underlying statistical properties of the PSD distributions in each of the radial bins [Efron and Tibshirani, 1986]. There also exist biases in the MLT distribution of the data. However, these biases are much smaller than the radial biases (as can be seen in figure 1), and the distribution functions are expected to show less of a trend with MLT than radius, so we do not mitigate for them.

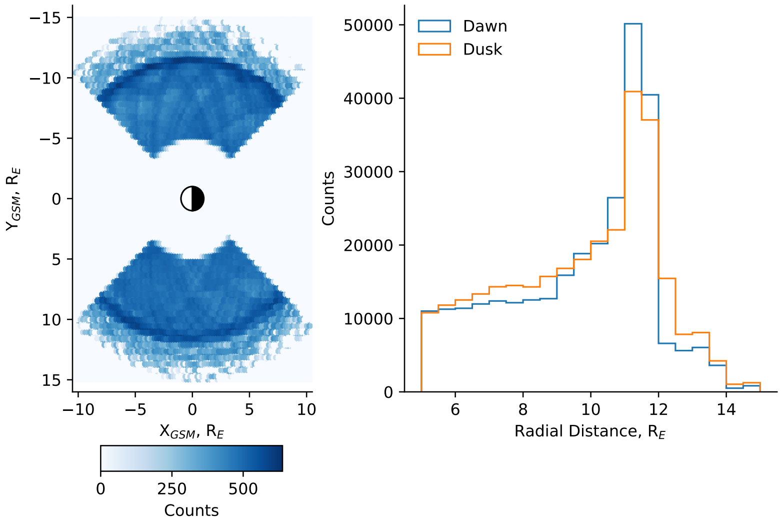


Figure 5.1: *The left plot presents the distribution of magnetically equatorial data samples in GSM co-ordinates, with a representation of Earth's day- (white) and night-side (black). The right plot presents the same data, but explicitly showing the radial distribution.*

5.3 Analysis

In this section we explore various methods which might be used to identify the location of the OBORB. Each method involves comparing the electron distribution function within various radial limits. We look at this through the lens of the hypotheses in Section 5.1. Initially, we use a non-empirical method based upon the evaluation of L* [Roederer, 1967] to investigate our hypotheses within the typical adiabatic invariant coordinate framework. Following this, we use radial binning to observe the radial evolution of the electron distribution function and look for discontinuous behaviour signifying the OBORB. Lastly, we employ machine learning methods as a tool for searching for the radial position of the OBORB through a hypothesis testing approach (though not the same hypotheses as in Section 5.1).

5.3.1 L* Analysis

Our study focuses on finding the radial extent of the ORB in real space (cf. adiabatic invariant space) by analysing positional differences in the electron distribution function. This naturally leads to using L* to classify whether data is inside or outside of the radiation belts. L* is a modelled property of magnetically trapped particles, which is used to define the extent of the radiation belts [Roederer, 1967; Roederer and Zhang, 2014; Roederer and Lejosne, 2018]. In a dipole field, the modelled L* corresponds to the radial distance of the point where the drift path of an electron intersects the magnetic equator. Employing L* as a definition of the radiation belts themselves allows us test our first hypothesis - that the electron distribution functions within and without the ORB differ. We stress that this is only an approximation for the radiation belts, since it relies on empirical field models (L* is not a measured quantity), which have significant disparities [see e.g., Thompson et al., 2020; Albert et al., 2018], and as such we do not use L* to try to quantify the location of the OBORB.

To incorporate the information L* provides (whether or not the electrons are on a closed drift-path), we employ seven magnetic field models to determine L* for a given datapoint - calculated using the SpacePy's wrapper of the IRBEM library [Morley et al., 2010b; Roederer and Zhang, 2014; Albert et al., 2018; Thompson et al., 2020] for 90° pitch-angle electrons (we comment on pitch angle in the discussion section). These models are: T89 [Tsyganenko, 1989]; OPQuiet [Olson and Pfitzer, 1974]; T96 [Tsyganenko, 1995]; OSTA [Ostapenko and Maltsev, 1997]; T01Quiet [Tsyganenko, 2002]; T01Storm [Tsyganenko et al., 2003], and T05 [Tsyganenko, 2005]. These models range from being analytic (OPQUIET) to quite heavily solar wind/geomagnetic index parameterised (T05). Given the seven models used, we specify that so long as at least four models returns a finite L* value, the datapoint corresponds to a trapped drift trajectory for at least some of the electrons measured, and is therefore within the radiation belts. This choice was informed by Thompson et al. [2020], who choose three models but suggest that using more models can reduce model-specific biases.

Figure 5.2 presents the results of the L* analysis. We have employed the subsampling method described in Section 5.2, with $n = 20$, to ensure that there is no sampling bias in the results. In panel a of Figure 5.2, the L* occurrence distribution and median L* values (based on the 4-model agreement criteria) are plotted over the range of radial distances. Below 8 R_E, $> 90\%$ of the data is located within the radiation belts (in that it has a valid L* value in 4 of the 7 magnetic field models). The occurrence fraction of L* values show a monotonically decreasing relationship with increasing radial distance (except > 12 R_E), in agreement with theory. We speculate that the increasing occurrence above 12 R_E and the decreasing median L* values above 11.5 R_E are spurious and represent some of the issues in trying to solely use modelling to define the OBORB [further issues with using current magnetic field models are highlighted in Albert et al., 2018].

Panels *b*-*e* in Figure 5.2 present comparisons between dawn and dusk, and inside and outside of the ORB (on the basis of L* being defined or not). Comparing vertically (i.e., panels *b* with *d*, and *c* with *e*) shows the difference between dawn (top) and dusk (bottom). There is a clear enhancement of the ≈ 10 keV seed population electrons [Jaynes et al., 2015] at dawn which is not present at dusk. There is also a depletion of the ≈ 1 keV source population electrons [Jaynes et al., 2015] which only appears outside of the radiation belts. The medians of the THEMIS SST data (> 30 keV), follow a power-law-type distribution as other works have found [e.g., Whittaker et al., 2013; Zhao et al., 2019]. Comparing between inside and outside of the radiation belts, the main differences (aside from the aforementioned depletion of source population electrons) are the typically more variable PSDs at energies $\lesssim 100$ keV outside the belt compared to inside. In contrast, the PSDs above this energy are much less variable outside the belt compared to inside. The distribution functions also have a shallower gradient and more variability inside of the radiation belt, highlighting a considerably more enhanced electron population.

5.3.2 Simple Radial Analysis

To investigate the OBORB, we calculate the median and interdecile (i.e., 10 to 90th percentile) range of data in nine radial bins between $5 - 13.5$ R_E. These results are presented in Figure 5.3. These distributions are calculated using the random sub-sampling technique described in section 5.2, with $n = 9$, to ensure comparable statistics between each of the bins.

We find significant radial evolution in both the dawn and dusk distribution functions. Both display flattening over the mid-range energies, suggesting either wave-particle interactions [Meredith et al., 2020], or the plasma sheet source [Kurita et al., 2011]. The notable difference between dawn and dusk is the pronounced bulge in the dawn distribution at ≈ 10 keV, mirrored in the interdecile ranges of the dawn data. We observe that the dawn and dusk distributions diverge with increasing radial distance up to $r \approx 9.7$ R_E, after which they converge to similar distributions. At low radial distances, the dawn and dusk data may

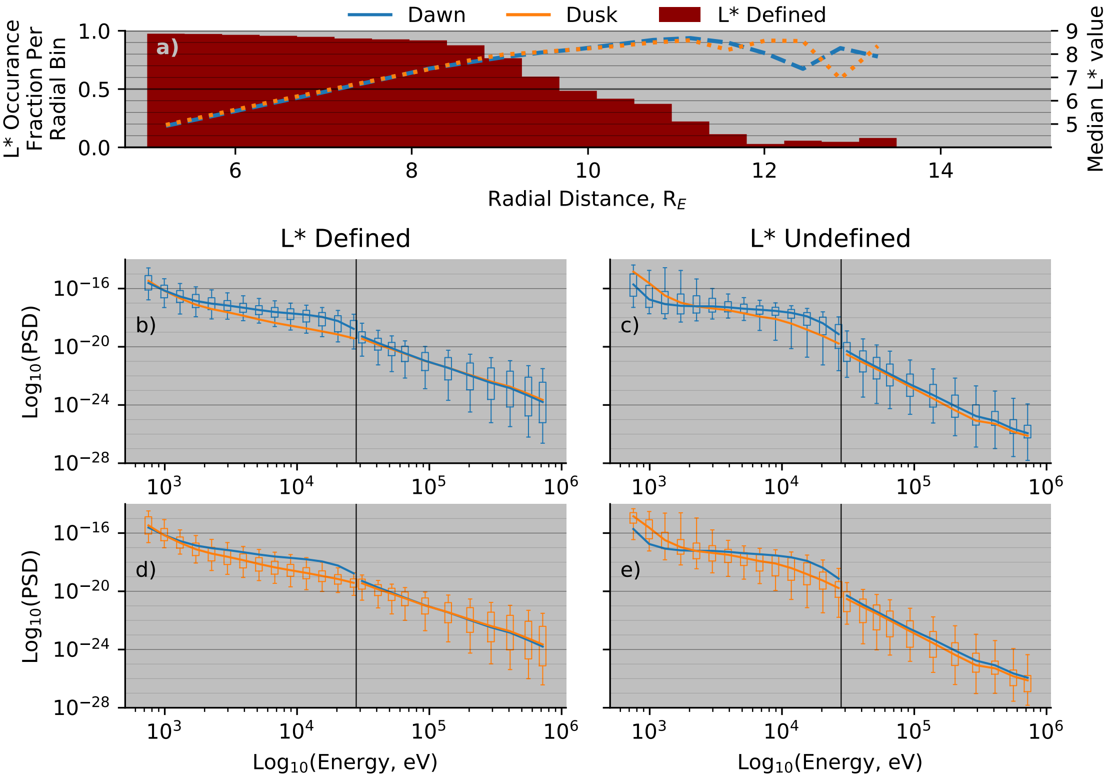


Figure 5.2: *Panel a presents the radial distribution of datapoints where L* is defined (i.e., the electrons are on closed field lines) normalised per radial bin, as well as the median L* value in each bin. The following box-plots present the per-energy-channel distribution and median trend-line of PSD at dawn and dusk, for data with and without a defined L*, respectively. These plots have a vertical line separating the ESA and SST instrument measurements. The box-plots represent dawn (b and c) and dusk (d and e), with the alternate line representing the median of the other for comparison.*

be more consistent because most of the data is inside the radiation belts, and equivalently at the higher radial distance most of the data is likely to be outside of the radiation belts. We observe that the dawn data exhibits the elbow at lower radial limits, and suggest that this may be the contribution of untrapped electrons. This is supported by the dusk distribution converging to the enhancement as the radial limit is increased beyond the expected limit of the OBORB and trapped electrons.

The distribution function at $5.0 - 5.9$ R$_E$ is very different in form from that at $12.6 - 13.5$ R$_E$, but the change in form occurs gradually, with no obvious discontinuity as a function of radial distance. This may imply that either there is not a hard boundary, or that the boundary location is highly variable. By not finding such a marker, we infer that this simplistic approach isn't best suited to locating the OBORB.

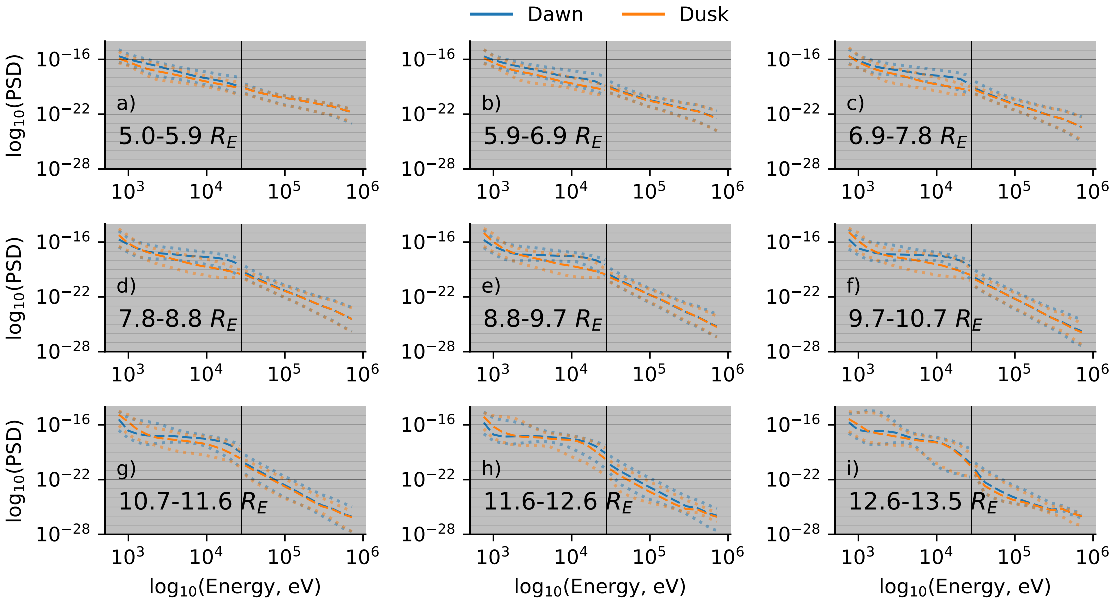


Figure 5.3: *The median and interdecile range of PSD in dawn and dusk, binned by radial distance. The black vertical line represents the break between data from the ESA and SST instruments.*

5.3.3 Machine Learning Analysis

With the previous method unable to find a clear radial distinction between electron populations, we now employ machine learning. We approach this much like hypothesis testing - a variety of radial limits are proposed as potential OBORBs (hypotheses) and empirically tested to determine which is most appropriate (the validity of an OBORB radial location, and how we might determine it, are discussed below). We constrain the data to the SST energy channels before applying machine learning, ensuring the results are not biased by lower energy particles, strongly affected by the $E \times B$ drift [Roederer and Zhang, 2014].

Our empirical analysis for a single set of proposed dawn and dusk radial limits is as follows:

1. Make a hypothesis by selecting a candidate radial limit for the OBORB (e.g., 7 R_E in the dusk or dawn sector).
2. Label each datapoint with a `0` if the measurement was made inside of the candidate radial limit, else label it with a 1. These class labels form the targets that a machine learning model (explained later in the text) will try to predict on the basis of the electron distributions.

3. Combine the dawn and dusk labelled data into a single dataset.
4. Provide a machine learning model each of the electron distribution functions as features (i.e., what the model will use to form a prediction). Each input is a 1 dimensional array of the values of PSD at each energy.
5. Train the machine learning model for the given set of input features (electron distribution functions) and targets (whether the data is inside or outside the chosen radial limit). The training set corresponds to 80% of the data, allowing for model performance to be quantified on an un-seen test set (the remaining 20% of the data).
6. Quantify the model performance of estimating whether a datapoint is inside or outside the chosen radial distance using un-seen electron distribution functions from the testing set. Metrics quantify the differences between the model-predicted class labels (0/1, inside/outside) with the class labels prescribed by the boundary location choices.

Note that neither the radial boundary locations, nor the radial locations of the measurements are provided to the machine learning model. Instead, the model tries to improve classification accuracy by inferring differences in the input features (PSD at each energy) between each set of class labels. By considering how well the model performs, we are assessing how much information is present in the electron distribution functions about the chosen radial distance. As electron distribution functions are expected to show the greatest difference either side of the OBORB, this in turn provides a measure for how good an approximation the chosen radial distance is for the OBORB. By, ‘greatest difference’ we are referring back to our initial hypothesis that the electron distribution functions of trapped electrons are different to those of untrapped electrons.

Each model used in the following analysis is a gradient-boosted [Friedman, 2001] ensemble of decision trees [Belson, 1959] implemented using the LightGBM framework for Python [Ke et al., 2017]. For each set of hypothetical boundary locations, a new model is trained, but the model architecture remains the same. Each ensemble is comprised of 256 decision trees [chosen to exceed suggestions from Oshiro et al., 2012, since LightGBM is cheap to run], which each contain 32 leaf nodes. Each model is gradient boosted using the dart algorithm [Rashmi and Gilad-Bachrach, 2015], where gradient boosting is a method of constructing the ensemble such that each subsequent decision tree in the ensemble is trained to correct for mis-classified predictions from the previous decision trees.

To test a large range of hypotheses we implement the above method in a training loop, stepping through each combination of dawn and dusk radial locations between 6 to 12 R_E (in increments of 0.2 R_E). By investigating the model performances over this range of plausible OBORB locations, we can assess the existence or otherwise of an OBORB, and whether the location can be constrained to a certain radial distance range. The existence of an OBORB can be judged by the magnitude of the quantified model performances; if models perform well, then it suggests that an OBORB or OBORB region exists. Once validated, the location of the OBORB can be constrained by comparing the relative skill of the different models and seeing if a particular set of boundary locations leads to models which perform better. Where we find radial limits with the best model performance, we know that these locations correspond to a split which maximises the differences in the distribution function data between the two classes (i.e., inside/outside, 0/1). In our context, this would represent the statistical OBORB.

Before detailing the results, we present the distribution of data obtained by our various radial limits. Figure 5.4 presents the proportion of data labelled as 'inside' at each dawn and dusk limit. There is a noticeable increase in the fraction of data within the radial limit at ≈ 11.5 R_E. This is due to the radial bias in the data distribution presented in 5.1. Generally the central regions of the plot have balanced data distributions. This distribution will be important in evaluating the performance metrics to ensure that they are not biased by having uneven class distributions.

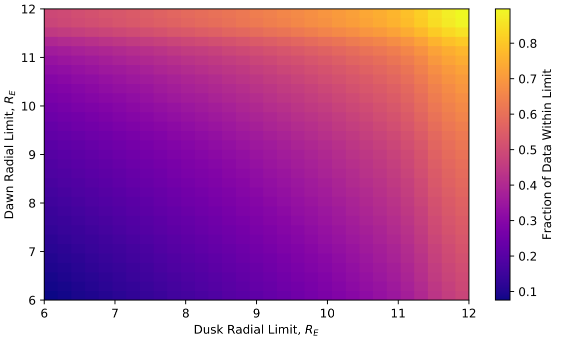


Figure 5.4: *A 2D histogram presenting the fraction of data classed as inside the radiation belts, as determined by various radial limits. The radial limits are independently chosen for dawn and dusk.*

To quantify our model performances, we employ a variety of binary classification metrics: Accuracy, Gilbert Skill Score (GSS), G-mean, F-measure and Critical Success Index (CSI) [Gilbert, 1884; Kubat et al., 1998; Lewis and Gale, 1994]. These metrics (aside from accuracy) have been chosen because they are designed to take into account class imbalances. Since different metrics focus on quantifying different aspects of predictive performance (see how the different metrics are constructed in 5.6.1), we present the results of multiple metrics to get a more complete view of the model performances. We also consider the inverted F-measure and CSI to account for the fact that they only consider one correct classification label (namely, the true positive predictions, ignoring the true negative predictions), and finally an aggregated metric comprised of the geometric mean of results from all metrics used. These metrics can all be derived from a confusion matrix of the results of our binary classification. See 5.6.1 for further details of the metrics and how they relate to confusion matrices.

Figure 5.5 presents the results of our machine learning analysis. Each panel presents a 2D histogram of the performance of a metric at each combination of dawn and dusk boundary conditions. Over-plotted are six contours evenly-spaced between the $70-100^{th}$ percentiles of the data. By all the metrics used, there are models which perform relatively high for at least a subset of the hypothesised boundary locations. The GSS has the lowest numeric model performance, but still has a constrained region of performance exceeding 0.7 (a score of 0 would represent no-skill and -1/3 is the lowest possible value). Aside from the GSS, each metric is constrained to the range 0-1. If our approach were flawed, and machine learning was not a suitable tool, we would expect to find that the models did not perform especially well at any location. Seeing as there are high-performing models (by each metric), we infer this as validation of our machine leaning approach. The contours of model performance presented allow us to constrain the locations of best performance, which we attribute to the OBORB location. However, before we analyse these contours we will discuss the issue of class imbalance.

Of the traditional metrics used, it appears that the GSS and G-mean metrics perform most robustly against the class imbalance, as can be seen by the lack of bias towards the upper right, or lower left areas (where the class imbalance is most pronounced). The average of the metrics also provides a class-balanced representation of the results. One thing to note from these results is the similarity between the accuracy, F-measure and CSI. This likely originates from the algebraic similarity between the definitions of these metrics (see 5.6.1). By using the inverted versions of these metrics we address the class imbalance when we take our average of the results, and observe how sensitive the results are to the class imbalance (the metric behaviour completely changes by focusing on a different true class prediction). Accuracy is inadequate as a metric when used on imbalanced data, since it is easily biased. This bias can be demonstrated in the following hypothetical case. If one has 100 data points, split into two classes (0 and 1), with 99 points falling in the 0 class. Then a model trained on this data may obtain a predictive accuracy of 99% by predicting everything to be in the 0 class. If it is important to be able to correctly predict the other classification, then this model will have no skill, despite the high accuracy.

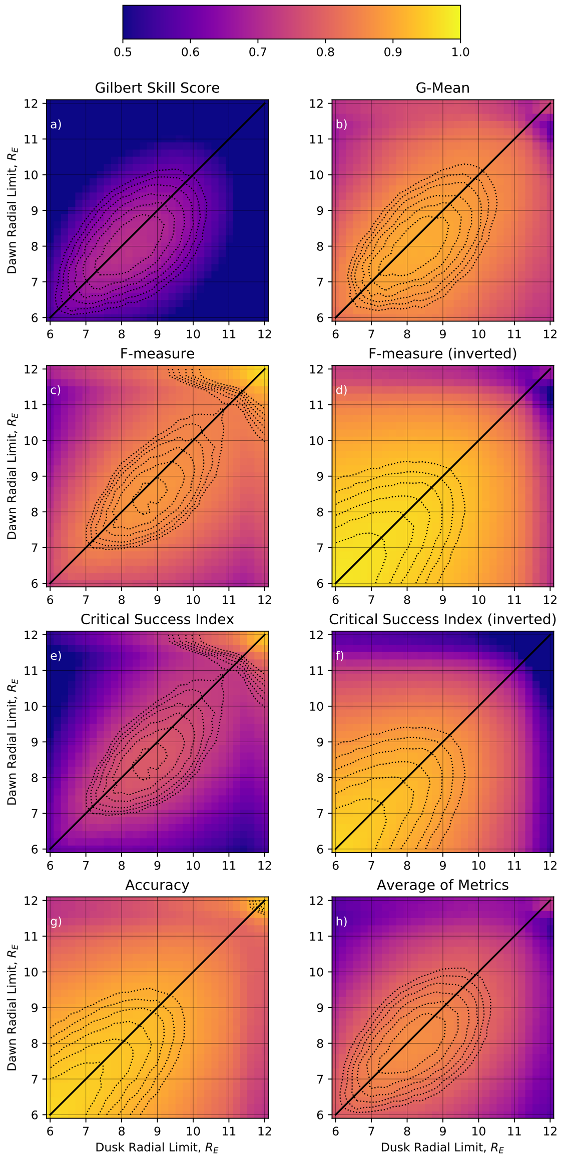


Figure 5.5: *2D histograms presenting the machine learning model performance, through various metrics. The average of the metrics presented in panel h represents the geometric mean of the metrics presented in panels a-g. Over-plotted are six contours between the $70 - 100^{th}$ percentiles of the data, used to draw attention to the regions of best performance.*

Whilst we present all of the metric results in Figure 5.5, for convenience we will focus the remaining discussion on the results of the average of the metrics, as this encapsulates the trends between all of the metrics. We observe a bounded region of best-performance between $\approx 6.9 - 9.1$ R_E in the dawn sector and $\approx 7.0 - 9.3$ R_E in the dusk sector. The contours show sharp decrease in the quantiles of performance outside of this area.

5.4 Discussion

For the sake of a clear methodology, we have generally made few comments on the results we've found. Here, we will start by discussing the machine learning aspect of this work, since it yields the most interesting results, and subsequently compare with the features found in our simple radial analysis.

In the machine learning analysis, we employed a fairly simple hypothesis testing approach to investigate various radial boundary locations for the OBORB. Our results suggest that a boundary exists, though its location may be highly variable. Variability in the boundary location may originate from myriad sources: pitch-angle dependence; energy dependence, and solar wind/geomagnetic activity. The pitch-angle of electrons is less likely to affect our results due to the focus on the dawn and dusk regions rather than day or night. For these latter MLT sectors, there is a strong pitch-angle dependence of the drift shell being observed at a given radial location [see figures in Roederer and Lejosne, 2018], which act in opposing directions between day and night. As such, this effect is much reduced in our data, though certainly some of the variability in the results is due to this. For electrons of different energies, there are magnetospheric processes that act preferentially. Thus, for each energy level there may be a different radial location corresponding to the last closed drift shell. By limiting our machine learning experiment to only the higher energies, we reduce the energy dependent effects. Though, as discussed below, we may still observe some of these effects even in our more limited energy range. Solar wind and geomagnetic activity are likely to have a significant effect on the OBORB, since such activity leads to large-scale reconfiguration of the magnetospheric topology and geometry. As this study represents (to our knowledge) the first empirical constraint on the OBORB location using *in situ* data, we do not account for activity, leaving such considerations for future work much in the same way as early research into the magnetopause location [Fairfield, 1971]. By doing so, our results likely represent a quiet-time or modal OBORB location. One final consideration is cross-species contamination of the electron flux by protons, which may increase the lower energy channel's PSDs and add additional variability to our results [Turner et al., 2012a, 2013].

From our results we infer that the lack of parameterising by solar wind or geomagnetic activity is the dominant factor for the variability in our boundary location, as activity will move the boundary physically, rather than softening it as the pitch-angle and energy dependence do. We infer this by the high skill scores (relative to the maximum value) which are distributed over a large range of potential radial limits. If instead, there was a softer boundary (i.e., a slow transition between the two characteristic distribution functions), we might still expect to see the smooth variation in the metric scores, but we would typically expect the quantitative values to be lower (e.g., all less than 0.5), as the models would find it more difficult to characterise the subtle differences in the slowly changing distribution functions.

Looking specifically at the average (geometric mean) of the metrics in Figure 5.5*h*, the distribution is shifted slightly in favour of a larger radial limit at dusk than dawn, but is otherwise quite a symmetric shape. The ovoid shape of the contours suggest a tendency for the boundary to favour similar values at dawn and dusk, though the implicit variability highlights that this may be only a weak tendency (taking the contours as the extrema of the variability, the dawn radial limit can be ± 2 R_E compared to dusk and *vice versa* the variability can be ± 2.5 R_E).

The dawn-dusk asymmetries observed might be explained by similar dawn-dusk asymmetries in the magnetosphere [Walsh et al., 2014; Haaland et al., 2017; Staples et al., 2020]. As we have excluded the lower energy particles from this portion of analysis, we do not expect this asymmetry to be primarily due to $E \times B$ drift, since the curvature and gradient drifts are energy dependent and hence will dominate over the electric field drift [though some recent works have shown that the electric field may still contribute: Sillanpää et al., 2017; Califf et al., 2017]. Instead, we speculate that this effect is more likely to be due to asymmetries in the (partial) ring current, whose effect is to increase the magnetic field strength at larger radial distances. This causes the electrons to follow the field and drift further out because of the gradient drift experienced. The sense of the dawn-dusk asymmetry suggests it is not simply the result of the algorithm identifying the magnetopause rather than the OBORB - the magnetopause can be compressed to below 8 R$_E$, but this happens much more frequently at dawn than dusk [Staples et al., 2020]. Whilst there may be some contamination of the data due to sampling the magnetopause or solar wind, we infer that this is negligible, since electron populations (and hence their distribution functions) are very different. It is expected that the difference between electron distribution functions inside of the magnetosphere and those in the magnetosheath or solar wind is much bigger than the differences between distribution functions inside and outside the radiation belt. A more easily identifiable dichotomy of distribution functions would be picked out more significantly by the algorithm and so we assert that the boundary identified by the algorithm is not the magnetopause, but the OBORB.

Our identification of the OBORB at ≈ 8 R$_E$ is typically larger than the values currently used in radiation belt modelling [e.g., Subbotin and Shprits, 2009; Shin and Lee, 2013; Glauert et al., 2014, 2018; Ozeke et al., 2014, 2018], suggesting that these modelling efforts are potentially missing radiation belt phenomena from the outer regions. Other empirical evidence, such as that in Sivadas et al. [2019], also support an OBORB location beyond the currently used limits ($9-12$ R$_E$ in their case). The OBORB being located further out opens up the possibility for smaller scale magnetotail behaviour (e.g., less severe substorms) to inject particles into the radiation belts, since they would not have to penetrate to such low L-shells. Such injections could lead to additional variability in the radiation belts [Turner et al., 2017; Jaynes et al., 2015] and to enhanced chorus wave activity in the outer regions [Meredith, 2002].

In Figure 5.3, we observed a flattening of the PSD at the mid-range energies and speculate that this is due to wave-particle interactions (WPIs). Given the energies of these electrons ($10-30$ keV) and their location (equatorial region, large radial distance) it is likely that whistler-mode chorus waves are the cause [Omura et al., 2008; Li et al., 2010a, 2011; Meredith et al., 2020]. The flattening occurs asymmetrically between dawn and dusk, with dawn being affected at lower radial distances. Meredith et al. [2020] present results showing that both lower- and upper-band chorus have a large dawn-dusk asymmetry. These results also show that specifically the lower-band chorus intensity is high at the large radial distances where we continue to observe the flattening of the distribution. Our presented results extend to larger radial distances than Meredith et al. [2020] or Li et al. [2010a], into regions close to the magnetopause. Due to the sparseness of data and research into WPIs in this region, we cannot speculate on whether or lower-band chorus remains the dominant wave affecting the electrons but these results suggest that more investigation may be required.

5.5 Conclusions

This study provides the first *in situ*, empirically-constrained location for the outer boundary of the outer radiation belt using THEMIS ESA and SST measurements. Characterising this boundary location accurately is an important aspect of radiation belt modelling, as it forms a time-varying source of electrons.

By applying simple statistical techniques, we observe significant radial evolution of the distribution functions, highlighting the intrinsic differences between the trapped (radiation belt) and untrapped electron populations. However, this approach did not yield a clear boundary location, instead showing a smooth transition between the two states. Such a transition signifies either a soft boundary, or a boundary with significant variability.

We employ machine learning (specifically, ensemble decision tree classification) in a hypothesis-testing framework, to assess whether there exists an identifiable change in electron distribution functions and hence outer boundary to the outer radiation belt, and where it may be located. The dataset was converted into 900 binary classification datasets, where data was labelled as either inside or outside of specified dawn and dusk radial limits (our hypothesised boundary locations). 900 machine learning models were then trained to learn this classification. Where the models perform better, we infer that our choices of boundary locations coincide more closely with identifiable changes in the electron distribution functions and hence the true statistical boundary location. By aggregating a series of metrics (many designed specifically for imbalanced datasets) we find a region of best performance between $\approx 6.9 - 9.1$ R$_E$ in the dawn sector and $\approx 7.0 - 9.3$ R$_E$ in the dusk sector.

This work presents a novel methodology for identifying the OBORB location, and opens up future research directions in parameterising the boundary location by solar wind and/or geomagnetic conditions. Our current results better constrain the statistical location of the OBORB and can be incorporated into the construction of radiation belt models, ensuring that they contain the physical processes of the radiation belts, and allowing future analyses to more appropriately capture the dynamics of injection events and how they influence the behaviour of the outer radiation belt.

5.6 Appendix

5.6.1 Metrics

All of the metrics used in this study can be derived from a confusion matrix. A confusion matrix is made up of True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN). How these correspond to model predictions can be seen in Table 5.1.

Table 5.1: A symbolic representation of a confusion matrix, with acronyms *TP*, *TN*, *FP*, *FN* referring to the different predictions True Positives, True Negatives, False Positives, and False Negatives, respectively.

		Model Prediction	
		0	1
True Value	0	TP	FN
	1	FP	TN

These relate to the following three commonly used, intermediary metrics and to H_R, which is used as a correction factor in the Gilbert Skill Score to account for the random chance of correctly categorising a sample.

$$precision = \frac{TP}{TP + FP} \tag{5.2}$$

$$recall = \frac{TP}{TP + FN} \tag{5.3}$$

$$specificity = \frac{TN}{TN + FP} \tag{5.4}$$

$$H_R = \frac{(TP + FP)(TP + FN)}{TP + TN + FP + FN} \tag{5.5}$$

We now define the metrics, and also present simplifications of the expansion into forms using only the four values from the confusion matrix.

$$
\begin{aligned}
Accuracy &= \frac{TP+TN}{TP+TN+FP+FN} && \qquad \text{(5.6)}\\
\text{G-mean} &= \sqrt{recall \cdot specificity} &&= \sqrt{\frac{TP \cdot TN}{(TP+FN)(TP+FP)}} && \qquad \text{(5.7)}\\
\text{F-measure} &= \frac{2 \cdot precision \cdot recall}{precision + recall} &&= \frac{2TP}{2TP+FP+FN} && \qquad \text{(5.8)}\\
Gilbert_{SS} &= \frac{TP-H_R}{TP+FN+FP-H_R} &&= \frac{TP \cdot TN - FP \cdot FN}{(TN+FN+FP)(TP+FN+FP)-FP \cdot FN} && \qquad \text{(5.9)}\\
CSI &= \frac{TP}{TP+FN+FP} && \qquad \text{(5.10)}
\end{aligned}
$$

The F-measure is the harmonic mean of the precision and recall and the G-mean is the geometric mean of the recall and specificity. On top of the proposed metrics, we also consider their values when the class labels are inverted, allowing us to investigate the robustness to the class imbalance (i.e., TP ↦ TN and FN ↦ FP and *vice versa*). Of metrics defined in Equations 6-10, we note that only the F-measure and CSI will be affected by this change, and so these are the only additional metric scores calculated.

Chapter 6

Deep-Ensemble Characterisation of the Flux at the Radiation Belt's Outer Boundary

---

Preface

In Chapter 5 we investigated whether *in situ* electron data could be used to empirically identify the outer boundary to the outer radiation belt. This outer boundary is crucial to radiation belt modelling, due to acting as a source of energy/electrons during simulation. The results from Chapter 5 suggest that there is a nominal quiet-time boundary located at ≈ 8.25 R_E.

Such a result is useful from a theoretical standpoint as it can be used in future work to further investigate the boundary location and its characteristics. Practically however, the electron distribution function at the boundary is a more useful result for ingestion into radiation belt models. This chapter explores how a probabilistic neural network can be employed the parameterise the outer boundary electron distribution function, enabling real-time translation of solar wind driving or geomagnetic indices into these distribution function.

Abstract

With ever-increasing space-based infrastructure, and day-to-day reliance upon it, it is more important than ever to be able to accurately model the environment within which spacecraft orbit the Earth. Modelling of the energetic electrons in the outer radiation belt is especially important, as these electrons interact with and can damage the large number of spacecraft in geosynchronous and medium-Earth orbits. Many different radiation belt models exist, sharing a common, critical dependence on accurate outer boundary conditions. We present two probabilistic models which can be used to create synthetic electron energy flux data in the region of the outer boundary. Both models are approximate Bayesian neural networks, constructed using ensembles of probabilistic neural networks. The synthetic data provided by the models can be used in the construction of the outer boundary condition for subsequent use in radiation belt models. The two models, nominally parameterised by either geomagnetic indices or the solar wind, show equivalent performance, predicting within a factor of 4.5 for the ≈ 720 keV flux (and better for the lower energy electrons). Both models generalise well with activity, showing relatively little variation in performance (errors remain within a factor of 6) over the deciles of B_z and SYM-H present in our test data.

Plain Language Summary

To ensure that the spacecraft orbiting the Earth (e.g., for GPS, or telecommunications etc.) don't get damaged during times of extreme space weather, we need to be able to simulate how the environment around the spacecraft will change over time. There are many different models which aim to do this. Between all of the models, one crucial aspect is to use an accurate outer boundary in the simulation - this outer boundary helps direct how the simulation will change over time.

We present two machine learning models which can be used to estimate the measurement a spacecraft might have made if it were located at the boundary

location. The estimates from our machine learning models can then be used by radiation belt modellers to more appropriately specify the properties of their outer boundary.

Both of our machine learning models, despite using different training data, perform as well as each other. Being able to estimate the spacecraft measurement within 4.5 times the true value, which is a small factor given that the measurements can increase by 100 times on minute timescales. We also find that our models still perform well when space weather is more severe, ensuring they are useful.

6.1 Introduction

Earth's inner magnetosphere contains regions of magnetically confined, energetic plasma. This plasma is often confined in two toroidal regions known as the radiation belts. The outer radiation belt (ORB) is predominantly populated with energetic electrons (in contrast to the inner belt which is predominantly protons) which can reach relativistic energies. The electron energy fluxes in the ORB are highly dynamic, changing by orders of magnitude on minute timescales [e.g., Blake et al., 1992]. Accurate modelling of the ORB is critically important to the continued operation of spacecraft in geosynchronous and medium earth orbits, as they are subject to hazardous surface charging and electrostatic discharges between internal components due to interactions with the energetic electrons present [Frederickson et al., 1991; Baker, 2001; Eastwood et al., 2017a].

A variety of radiation belt models have been developed to further our understanding of the ORB, including: Salammbô [Beutier and Boscher, 1995; Boscher et al., 2000; Bourdarie et al., 2005]; VERB (Versatile Electron Radiation Belt) [Subbotin and Shprits, 2009]; STEERB (Storm-Time Evolution of Electron Radiation Belt) [Su et al., 2010b,a, 2011]; DREAM (Dynamic Radiation Environment Assimilation Model) [Reeves et al., 2012], and BAS-RBM (British Antarctic Survey's Radiation Belt Model) [Glauert et al., 2014]. One of the critically important aspects of using these models is characterising the outer boundary conditions; the boundary acts as a time-dependent source during model runs.

The boundary condition is multifaceted. Initially, the location of the boundary must be specified, this is done in either physical or adiabatic invariant coordinates. Subsequently, the electron distribution function (DF) at the chosen location (and energy/pitch angle) must be determined. Often, the boundary location is selected to be close to geosynchronous orbit ($L \approx 6.6$ R$_\mathrm{E}$ or an equivalent position in adiabatic invariant space), and then the electron DF is reconstructed using spacecraft data in the vicinity (e.g., CRRESS, Polar, Van Allen probes or THEMIS), from model outputs [e.g., AE8 Vette, 1991], or even optimised for as a free parameter [e.g., Reeves et al., 2012].

When specifying a boundary location at which to construct the electron DF, there is currently a trade-off between the boundary being co-located with the *true* boundary location (dependent as it is on energy and pitch angle) and being chosen to maximise the available data. The latter choice reduces the processing necessary to map the flux from its measured location to the boundary location, and increases the volume of data that can be used. However, if the data is being measured at a location that is not co-located with the true boundary location, then it is plausible that modelling efforts may inadequately capture all of the relevant phenomena. Without fully capturing the physics, radiation belt models are essentially limited to reanalysis modelling or short-term forecasting, as forecasting errors will increase with lead-time due to inaccurate time-evolution of the system.

Other works have sought to characterise the boundary DF at radial distances outside of geosynchronous orbit [e.g., Shin and Lee, 2013; Maget et al., 2015]. Whilst these results are extremely useful, and present steps in the right direction towards appropriately capturing the outer boundary of the ORB (the OBORB), their application is limited. Neither Shin and Lee [2013] nor Maget et al. [2015] (hence S13 and M15, respectively) account for spatial variations in the construction of the boundary, being (essentially) drift-averaged or only focussing on a specific MLT sector, respectively. Whilst S13 parameterise their outer boundary with solar wind variables (V_{sw} and density), they do so deterministically, and so cannot provide the probabilistic functionality that M15 have. In contrast, M15 only use Kp to parameterise their results, and so likely miss lots of the time dynamics present in the boundary condition.

Our approach allows researchers to side-step the trade-off between boundary location choices - providing models (or a methodology) which can accurately map electron fluxes between their measured location and a boundary location.

Specifically, we create various models using, as inputs, GOES (Geostationary Operational Environmental Satellites) data, solar wind data and geomagnetic index data. Using these inputs, we employ a deep-ensemble of neural networks (NNs) to probabilistically model electron fluxes measured by THEMIS (Time History of Events and Macroscale Interactions during Substorms) spacecraft in the vicinity of our chosen boundary location. Our models can be considered to be synthetic spacecraft data in the vicinity of the the OBORB; allowing ingestion into radiation belt models agnostic of their specific data processing requirements.

Rather than choosing a boundary location arbitrarily, we use the nominal boundary location of $L \approx 8.25$ R$_\text{E}$ provided by Bloch et al. [2021]. The work presented in Bloch et al. [2021] represents one of the few empirically constrained boundary locations, and (to our knowledge) the only one constrained using *in situ* data or machine learning.

The deep-ensemble and data will be described further in Sections 6.2 and 6.3, respectively. We will explore the model performances in Section 6.4 and investigate how different modelling approaches affect the results in Section 6.5 before summarising our results in Section 6.6.

6.2 Methodology

Machine learning (ML) has already been widely used in many diverse areas of magnetospheric physics [non-exhaustive lists can be found within: Camporeale et al., 2018; Camporeale, 2019; Smirnov et al., 2020; Azari et al., 2020]. The subset of ML techniques which are relevant to this work are those known as 'supervised' techniques. Supervised learning refers to the type of ML where a chosen algorithm is provided with input-output pairs of data. The inputs are samples of the parameter space from which one seeks to model the outputs. For example, in our case one set of inputs are GOES and solar wind data, and the output is the flux at the OBORB. In this way, supervised ML algorithms are essentially function approximators, trying to optimise the ability to correctly map from the inputs to the outputs. There are a variety of different algorithms that might be employed for this task, but neural networks are ideally suited to complex non-linear mappings.

6.2.1 Neural Networks

Here we give a high level introduction to neural networks. See Gurney [1997] for a more comprehensive introduction to the subject.

Neural networks can be represented as computational graphs constructed out of two types of building blocks: nodes and edges (edges connect nodes). Nodes can be considered in two ways: those that represent input values, and those that represent a function for combining input values. The input values can either be a vector of raw data passed to the model, or the values from previous nodes in the network being propagated through the network along an edge (with an associated weighting). Figure 6.1 presents a schematic of a neural network. The nodes typically perform the following to map between its input and its output (shown visually in the blown up node in Figure 6.1):

1. The inputs (previous node outputs which have been weighted) and bias term (typically 1) are summed together.
2. The summed inputs are passed to an activation function (e.g., a sigmoid or hyperbolic tangent function).
3. The output from the activation function is provided as the node output.

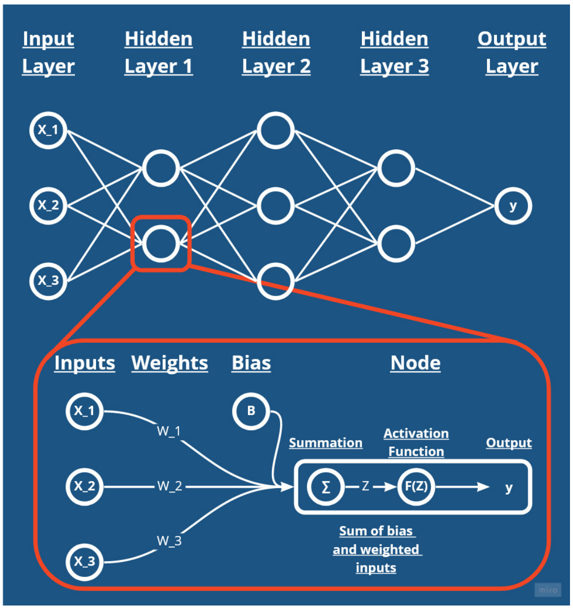


Figure 6.1: *A schematic diagram of a simple neural network, with an expanded look at processing which occurs through a network node.*

During training, the NN weights are optimised. The weights are initialised randomly; in our case they are sampled randomly from a uniform prior distribution between $\pm\sqrt{1/\mathrm{N}_{\mathrm{input}}}$, where $\mathrm{N}_{\mathrm{input}}$ is the number of input features given to the model. Training is typically performed in epochs, where one epoch represents passing the whole training data set through the network, once. After each epoch, the model performance is quantified by a loss function (e.g., mean-square error in many regression tasks). Then an optimisation algorithm is applied to minimise the loss with respect to the weightings of each edge in the network. This process is repeated until the loss has been reduced to the requisite amount or until a sufficient number of training epochs have been performed.

Abstracted from this description of training is the concept of hyper-parameter optimisation. This relates to how choices of architecture, optimisation algorithm and network features are made. Whilst there are more finessed options, the most simple and frequently used way of making these choices is to train many networks with different permutations of the hyper-parameters, and choose the hyper-parameters which yield the best performance.

6.2.2 Probabilistic Neural Networks and Bayesian Deep Learning

It is common in much Space Weather-related empirical modelling to create and optimise a deterministic model - one that only predicts a single value, without any associated uncertainty. Such approaches are indeed useful in many situations, but recent work highlights the need for, and benefits of, probabilistic forecasting/modelling [Bentley et al., 2018; Camporeale et al., 2019; Maget et al., 2015; Thompson et al., 2020; Owens and Riley, 2017; Owens et al., 2017; Watt et al., 2021]. Predicting a distribution provides more information than a single point, and can significantly improve down-stream usage of the model outputs [e.g., in data-assimilative models such as Maget et al., 2015]. Formally, this implies that for a given a dataset with training inputs X and outputs Y, we model the distribution $y \sim p(x;\theta)$, where x and y are individual samples from X and Y, and θ are the parameters of $p(\cdot\,;\theta)$. We then use probabilistic neural networks to parameterise such distributions. For example, for a Gaussian distribution, $\mathcal{N}$, parameterised by a neural network we would have $y \sim \mathcal{N}(y;\theta) = \mathcal{N}(y;\mu_\omega(x),\sigma^2_\omega(x))$, where $\mu_\omega(\cdot)$ and $\sigma^2_\omega(\cdot)$ are two neural networks with weights ω which, given an input x, output the statistics (mean and variance, respectively) of a Gaussian distribution. Training such a model requires the minimisation of the negative log likelihood of the data (equation 6.2). This is equivalent to maximum a posteriori (MAP) estimation of the network weights ω [Jospin et al., 2020].

However, so far we are assuming a distribution over the outcomes y. If, additionally, we assign a prior distribution $p(\omega)$ over the weights for each instantiation of a neural network, then we can consider training as a process of updating the weights of the network, Ω, after seeing the training dataset, D, or formally $p(\Omega|D) = \frac{p(D|\Omega)p(\Omega)}{p(D)}$. Such an update is called a 'posterior update' in Bayesian terminology and it is the basis of Bayesian neural networks [Mackay, 1995; Neal, 1996]. However, learning the posterior distribution is infeasible. Hence, in practice, we estimate the approximate posterior distribution with methods such as Monte Carlo Dropout [Gal, 2016] or Deep Ensembles [Lakshminarayanan et al., 2017]. In this work we used Deep Ensembles due to their effectiveness and simplicity to implement. Each ensemble member is trained on bootstrap samples of the training data, with randomised initial weights and stochastic optimisation - given us an ensemble where each member has different final weights (each a sample from the posterior distribution).

Note here that we have two distributions in place. One posterior distribution over the weights of the neural networks $p(\Omega|D)$ but also, for each sample from that posterior distribution, $\omega \sim p(\Omega|D)$, we have a distribution over the outcome $y \sim p(y|x, \omega)$. In order to incorporate both distributions, we marginalise over the distribution of the weights, resulting what is called the posterior predictive distribution $p(y|x, D) = \int p(y|x, \omega)p(\omega|D)d\omega$. By evaluating the variance of the posterior predictive distribution, a quantity called predictive variance, we can evaluate the total uncertainty of the fluxes our model predicts. This total uncertainty incorporates aleatoric (irreducible) uncertainty - the fundamental uncertainty due to the experiment design (i.e, limitations in the modelling approach; not measuring all the required state variables, and having imprecision in the measurements themselves), and epistemic uncertainty - the uncertainty which arises due to the lack of data. In practice, we expect the aleatoric uncertainty to be high in noisy regions of parameter space, and we expect the epistemic uncertainty to be high in areas of the parameter space which are sparsely sampled in the training dataset. By combining these uncertainties, we create a model that which can inform on its own confidence in the predicted flux. The total approximated uncertainty associated with a given prediction, y_i, is then given by the law of total variance:

$$Var(y_i|x_i) = \mathbb{E}_{\omega \sim p(\Omega|D)}[Var(y_i|x_i;\omega)] + Var_{\omega \sim p(\Omega|D)}(\mathbb{E}[y_i|x_i;\omega]) \tag{6.1}$$

Where x_i is the model input associated with the predictions, y_i is the mean flux predicted by the ensemble, and ω are the weights sampled from the posterior distribution of the weights. This equation can be interpreted as follows: the total variance associated with a prediction is given by the sum of the mean predicted variance of the ensemble and the variance of the ensemble flux predictions. We can then convert this variance into a standard deviation and use this to quantify the distribution associated with a particular prediction y_i.

6.2.3 Training Procedure

To train our networks, we split our data into three sets: training, validation and testing (see Section 6.2.3.1). During each training epoch, the full training dataset is passed through the model and weights are iteratively tuned to improve the model. The validation dataset is used during hyper-parameter optimisation and training to ensure model generalisability. In the former case, optimum hyper-parameters are chosen as those which minimise the validation loss. In the latter case, model training is halted when the validation loss begins to diverge from the training loss, as this signifies that the model has begun to over-fit to the training data and is losing generalisability. Since the validation dataset is used during the training procedure, it does not represent a dataset which can be used to independently quantify the model performance. Hence, we retain the test set, as our final unseen-by-the-model dataset, which is used to quantify the model performance without bias.

During training we use the negative log-likelihood (NLL) as our loss function. For a given pair of model outputs μ_i and σ_i^2 (the model architecture will be defined subsequently), the NLL of the true flux, y_i, is given by:

$$-log\left(\mathcal{L}(y_i|\mu_i,\sigma_i^2)\right) = \frac{1}{2}log(2\pi\sigma_i^2) + \frac{1}{2\sigma_i^2}(y_i - \mu_i)^2 \tag{6.2}$$

The NLL is calculated for each model output over the full training dataset, and averaged (mean) to get a single value of the loss for the optimiser to back-propagate through the network. This loss allows the model to condition the predicted variances - learning the aleatoric uncertainty.

To select hyper-parameters, many different iterations of training are performed with different hyper-parameters and the best model is selected for further analysis. In our case there are 4 hyper-parameters we optimise for:

1. The conjunction MLT: as will be described in Section 6.3, we optimise for the MLT distance between spacecraft. We vary the conjunction MLT between 1 and 12 hours of MLT in integer steps.
2. The network architecture: as will be described in Section 6.2.4, the width of our network is another aspect of our model that we optimise for. This allows us to empirically choose a suitable network size that gives our model good performance. We specify our network width's multiplicative factor as $N \in \{32, 64, 96, 128, 160, 192\}$.
3. The optimiser: different optimisers update weights in different ways, and this can have a significant impact on model performance, we investigate the following commonly-used optimisers - averaged stochastic gradient descent [Polyak and Juditsky, 1992], Adam [Kingma and Ba, 2014] and AdamW [Loshchilov and Hutter, 2017].
4. The learning rate: a part of the network optimisation process is characterised by how much to vary the weights based on the gradient of the loss function - this is controlled by the learning rate. Choosing a learning rate that is too small can lead to slow training and the algorithm getting stuck in a local minima, whereas too large a learning rate can lead to unstable training and a lack of convergence on the global minima. Here, we vary the learning rate between 0.01 and 0.00001.

For each of our input parameter sets we train approximately 250 models, with random permutations of the above hyper-parameters, and select the best hyper-parameters by the validation loss. Our hyper-parameter selection experiments are tracked and managed with Weights and Biases [Biewald, 2020], specifically using the sweep functionality. Training each model takes approximately 8 minutes on average, using a Telsa T4 GPU.

6.2.3.1 Train-Validation-Test Splitting

Two common methods of splitting a dataset up into a training, validation and testing dataset are: (1) to randomly split the data into the three sets ensuring some specified fraction of the data is in each set; or (2) to split the dataset up sequentially in time, such that the training dataset comprises the first fraction of the data, then the validation set, then the test set.

The first method is not suitable for time-series where there are auto-correlations of the parameters. If there is auto-correlation, then information from the larger testing dataset can leak into the validation and testing datasets, making them no longer independent of the training set, and giving unreliable performance metrics.

The second method is more commonly applied to time-series, as it reduces the information leak between the datasets. However, it can also introduce unreliability into the validation and test sets if there underlying drift in the dataset (i.e., if some process acts to change the distribution of the parameters). In our case, we expect that the solar cycle variation will impact the fluxes we aim to model. If our dataset was large enough to ensure that a whole solar cycle was represented in the training set, we might side-step this issue due to the model adequately learning the cyclic behaviour.

To reduce any possible bias in the performance metrics we evaluate, we use the following procedure to construct our training, testing and validation datasets:

1. Find the time range of the combined THEMIS-GOES-OMNI dataset.
2. Split the full time range into 50 bins.
3. If there are more than 100 samples of data in a bin, allocate the first 70% of the data to the training set, the next 20% to the validation set, and the final 10% to the testing set.
4. If there are fewer than 100 samples of data, then assign the whole bin to the training dataset.

This method ensures that each dataset samples approximately evenly from the solar cycle. But, by taking contiguous segments of data, reduces any biasing effects that auto-correlations may have.

6.2.4 Model Architecture

A schematic diagram of our general network architecture is presented in Figure 6.2. Our networks are multi-headed, meaning they predict multiple outputs for each set of input parameters. Specifically, the models predict 11 flux values (one corresponding to each THEMIS energy channel), and an associated variance. Semantically, our modelling is split into two parts. The first part (hidden layers 1-4) is a fully connected NN, whose purpose is to capture the general mapping between the inputs and the output fluxes. The second part passes the outputs from the fourth hidden layer to 11 smaller NNs, one for each of the energy channels. In this way, we hope to train each of the 11 smaller models to fine-tune the mapping for each energy channel. We use rectified linear units (ReLU) as our activation function between layers.

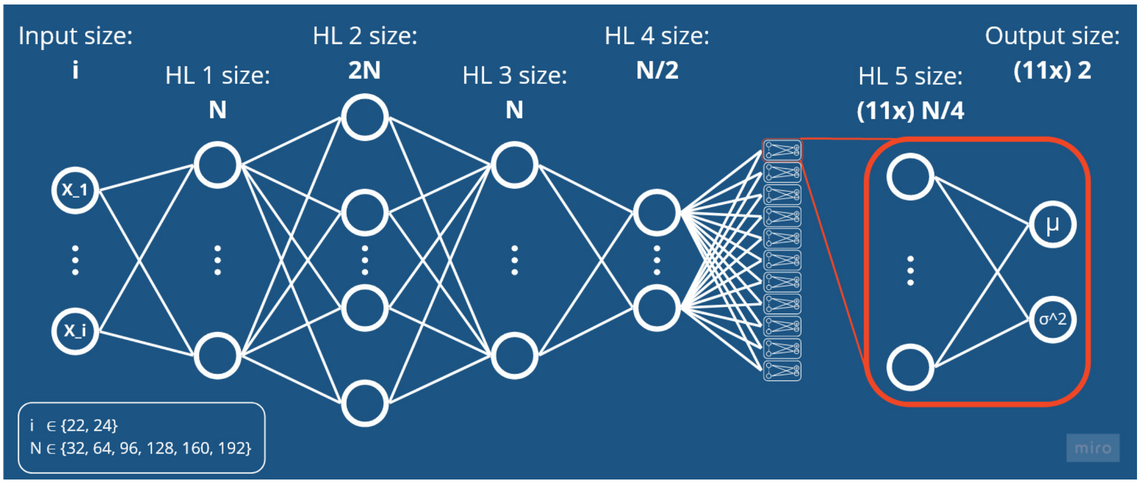


Figure 6.2: *A schematic diagram of the network architecture used in this study. The input size, i, represents the number of input features being passed to the network. N is a multiplicative factor that corresponding to the number of nodes in each hidden layer. μ and σ^2 represented the predicted flux and variance for a given energy channel.*

The specific architecture of our models is dictated by i and N. The input size, i, is the number of input features being used, and the number of nodes per hidden layer, N, is chosen through hyper-parameter optimisation. For the two models we train (whose names will be explained in Section 6.3), we find that the following hyper-parameters are optimal: for the index model we select ASGD as the optimiser, 0.005486846978788688 as the learning rate, 8 hours as the conjuction MLT and $N = 32$; for the solar wind model we again select ASGD as the optimser, 0.0015834503436594227 as the learning rate, 9 hours as the conjuction MLT and $N = 160$.

Our network width expands before tapering down. We use this model structure to ensure that an expansive set of combinations of the input parameters is captured, hopefully allowing the model to capture more of the appropriate non-linearities in the mapping.

6.3 Data

A variety of data is used in the construction of our models. To capture the driving conditions, we use 1-minute cadence upstream solar wind data from NASA/GSFC's OMNI dataset [through HelioPy, Stansby et al., 2021]. Specifically, we use: magnetic field measurements in Geocentric Solar Magnetospheric (GSM) coordinates; solar wind velocities in Geocentric Solar Ecliptic (GSE) coordinates; solar wind proton number density; solar wind proton temperature, and the solar wind flow pressure. All of these data have already been propagated to the nose of the magnetosphere. To capture the current state of the magnetosphere (this can be considered the integrated effect of past solar wind driving), we use the geomagnetic index data from the 1-minute cadence OMNI dataset, calculated using ground magnetometers. We select the AL, AU, AE, SYM-H, SYM-D, ASY-H and ASY-D indices. The A-indices represent auroral behaviour, with AL and AU giving the upper and lower bound of the geomagnetic variations measured by the contributing stations, and AE gives the difference between AL and AU. The SYM-indices represent the disturbance field acting on the geomagnetic field (H for the magnetic North direction and D for the East-West direction). The ASY-indices are the range between the maximum and minimum deviations of the SYM-indices subtracted from the disturbance field.

The GOES spacecraft orbit the Earth in a geosynchronous orbit ($L \approx 6.6$ R$_\text{E}$), providing measurements of the electron flux inside of the ORB. GOES data are used for radiation belt modelling due to the abundance of data and/or real-time availability [Glauert et al., 2018]. We use the 1-minute cadence > 800 keV and > 2 MeV integral electron energy fluxes, the standard deviations of those fluxes, the MLT (Magnetic Local Time) position of GOES and the respective spacecraft ID (reflecting whether data is from GOES 13, 14 or 15). These data were obtained from CDAWeb, and are specifically the background, contamination and dead time corrected data from the Energetic Proton, Electron and Alpha Detector (EPEAD)

[Onsager et al., 1996]. From GOES 13, 14 and 15, we use the flux measured by the westwards detector [as per Glauert et al., 2018; Meredith et al., 2015]. The GOES 13, 14 and 15 data between 2010/05/01 and 2017/12/14, 2012/10/01 and 2020/03/04, and 2010/03/26 and 2020/03/04 we used, respectively.

The choice to include the GOES spacecraft ID as an input parameter was made during our investigation and found to improve performance. It is likely that the performance improvement is due to the ID parameterising the different orbital positions of the GOES spacecraft, which are known to influence the measured fluxes [Meredith et al., 2015].

As our aim is to make a model that can map fluxes measured by the GOES spacecraft to an appropriate equatorial OBORB location, we use THEMIS data measured ± 0.5R$_E$ from the nominal boundary [identified by Bloch et al., 2021] as our target flux data (what our model will approximate). We take the flux measurements from the SST (Solid State Telescope) instrument [Angelopoulos, 2008], providing 11 differential electron energy flux measurements - approximately log-spaced, between $30-720$ keV - at each time-step. We also calculate the MLT position of the THEMIS spacecraft. Data is taken from THEMIS probes A, D and E between 2007/09/27 and 2019/09/29, whilst data from probes B and C is taken up till 2010, at which point they were moved to a lunar orbit [Russell and Angelopoulos, 2014].

Our dataset is comprised of GOES measurements which are contemporaneous with THEMIS crossing our equatorial boundary region. The THEMIS data are selected ± 0.5R$_E$ in X and z (in GSE coordinates). This greatly reduces the amount of data from the THEMIS mission which is applicable to be used in this study. A further reduction in the dataset size occurs as we constrain the relative positions of the THEMIS and GOES spacecraft in MLT. Ideally, we would fix our dataset to measurements where both spacecraft were located in the same MLT hour, as this limits the measurements to areas of most similar physics - The model would theoretically only have to learn the radial evolution, rather than both radial and azimuthal evolution. Unfortunately, there is a trade-off between the proximity (in MLT) of the spacecraft, and the amount of contemporaneous data we have. Hence, we add the conjunction MLT to our list of hyper-parameters to optimise (i.e., we empirically select the MLT proximity which maximises our model performance).

To account for the MLT between our spacecraft being different, we provide the

MLT of both spacecraft as model inputs (allowing the model to learn what effect different MLT distances have on the flux mapping). Since MLT is cyclically continuous, we provide the model with the sine and cosine of the MLTs. To help our models account for the cyclical temporal variations which occur on year-long (i.e., seasonal) and day-long (i.e., rotational) timescales we pass the sine and cosines of both the time-of-day and time-of-year of a given data-sample to our model as inputs. A complete list of our specific parameters is as follows:

- **Inputs**
  - Flux measurements
    - 800 keV flux, 2 MeV flux, standard deviation of 800 keV flux, standard deviation of 2 MeV flux, and GOES spacecraft ID.
  - Temporal information
    - Sin(time-of-year), Cos(time-of-year), Sin(time-of-day) and Cos(time-of-day).
  - Spatial information
    - Sin(GOES MLT) and Cos(GOES MLT)
    - Sin(THEMIS MLT) and Cos(THEMIS MLT)
  - Solar Wind variables
    - B_x, B_y, B_z, V_x, V_y, V_z, N_p, T_p and P.
  - Geomagnetic indices
    - AL, AU, AE, SYM-H, SYM-D, ASY-H and ASY-D.
- **Outputs**
  - THEMIS fluxes for energy channels:
    - 31 keV, 41 keV, 52 keV, 65.5 keV, 93 keV, 139 keV, 203.5 keV, 293 keV, 408 keV, 561.5 keV and 719.5 keV

To ensure that the heteroscedasticity (the differences in magnitudes of the variances) of the input parameters does not affect the training of our networks we standardise each of the input features, X_i, by their mean, μ_i, and standard deviations, σ_i:

$$X_i' = \frac{X_i - \mu_i}{\sigma_i} \tag{6.3}$$

We additionally scale our output features, y_i, by taking a log-10 transform and then standardising them, as above.

Of our selected feature set, it is likely that a subset of the parameters are performant, and hence can be used to more efficiently train adequate models. However, undertaking a comprehensive performance analysis of each parameter in beyond the scope of the present work. Instead, we take as base parameters all of the inputs above, except for the solar wind and geomagnetic indices. We then produce two models by choosing include either the solar wind data or the index data alongside the other base parameters. In this way we can determine whether the instantaneous solar wind is more important than its processed effect through the magnetosphere, and how much additional information is obtained using one set over the other. The solar wind model has 24 resulting input parameters, while the index model has 22 (this corresponds to i in Figure 6.2).

6.4 Model Performance

This section will serve to demonstrate the efficacy of the two ensemble models. We make two ensembles so as to investigate whether there is more pertinent information present in the solar wind or geomagnetic indices. The implications of the results presented here will be discussed in Section 6.5.

As the ultimate driver of radiation belt phenomena, one might theoretically expect the solar wind to encode all of the relevant information to predict the flux at the outer boundary. However, given that we only use instantaneous values as inputs to our model (vs. time-history) and don't account for atmospheric processes (e.g., ion outflow), we don't expect to be able to fully capture the effects related to the preconditioning of the magnetosphere (e.g., solar wind processing timescales, the geo-effectiveness of solar wind given the current state etc.).

By including indices - which act as state variables, summarising the integrated effect of past solar wind - in the modelling process, we hope that the model learns how the current state of the magnetosphere pertains to the mapping of the fluxes. Given that we only use instantaneous values as inputs to our model (vs. explicit time-history), one would reasonably expect the model to be unable to capture shock-like behaviour in real-time, instead requiring some processing time for the information to propagate through the magnetosphere.

Figure 6.3 presents a summary of the index model performance. The results from the solar wind model are qualitatively equivalent, and so we do not present them here. In Figure 6.3a-c, we observe how the model predictions vary with the true flux. In general, the predictions fall within a factor of five (dotted white lines) of the true value. With increasing energy (going from panel a to c) the predictions become distributed increasingly far from the solid white line (representing perfect predictions) - highlighting less accurate predictions.

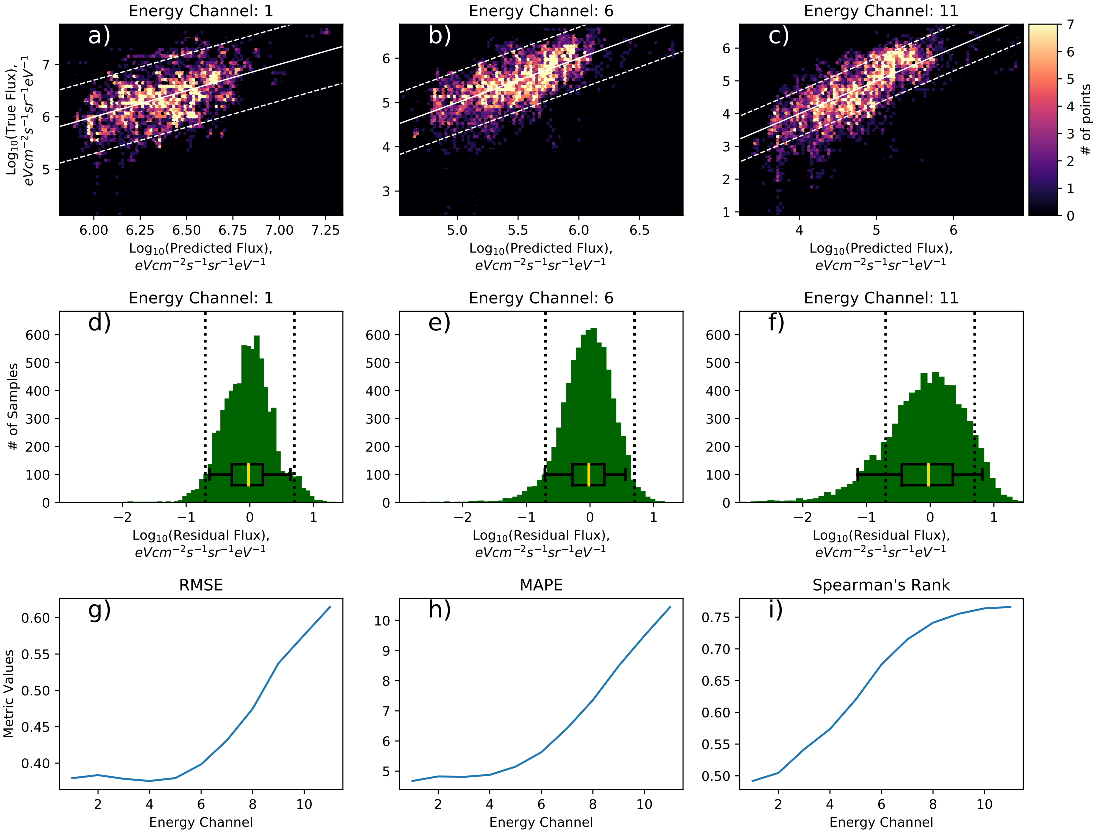


Figure 6.3: *Summary plots of the performance of the index model. Panels a-c present 2D histograms of the predicted values (x-axis) versus the true values (y-axis) for 3 representative energy channels (31, 139 and 720 keV). The colour corresponds to the number of values falling in each histogram bin. The solid white line represents y=x, where all the data would lie if the predictions were perfect. The dotted white lines represent a factor of five error in the prediction. Panels d-f present histograms of the residuals fluxes (true flux - predicted flux). The black dotted line represents a factor of 5 difference. Boxplots are also superimposed to signify the median (yellow line), interquartile range (the box itself), and the 5th and 95th quantiles of the distribution (the whiskers). Panels g-i present the RMSE, MAPE and Spearman's rank order correlation coefficient calculated for each of the energy channels.*

In Figure 6.3d-f we present the distributions of the residual fluxes (true flux - predicted flux) for the same energy channels. We see in the three examples that the interquartile range of the data is well within the factor of 5 error (dotted black lines). Additionally, for the lower energy channels, the 5th and 95th quantiles are within the a factor of 5 error as well. We see numerically the increased spread that comes with increasing energy too, with the 11th energy channel showing a significantly wider residual distribution. In each case, the tails of the residual fluxes are skewed below zero. So, despite the median (average) suggesting typically accurate predictions, when the model does over-predict , it does so by a larger amount than it under-predicts (this is seen by the larger negative values of the residuals).

In Figure 6.3g-i we present model performance with a range of metrics for each of the 11 energy channels our model predicts for. We see that for energy channels 1-6 the RMSE (root-mean-square error) is equivalent to less than a factor of 2.5 ($10^{0.4}$), and energy channels 7-11 fall within a factor of 4.5 for the higher energies ($10^{0.65}$). The MAPE (mean-absolute percentage error) shows a similar trend to the RMSE, increasing with energy channel. In contrast, the correlation (Spearman's rank order correlation coefficient) shows improvement with increasing energy channel - suggesting that even though the absolute values may be incorrect, the model is still providing useful discrimination between times of high and low flux in a relative sense.

In Figure 6.4 we can see how the true and predicted energy-flux distributions evolve with the energy channels. The lower energy channels have less variability than the higher energy channels. The predicted distributions do not capture the extremes of the distributions - this is common in empirical modelling, as there is often little training data for the extreme cases. The lower energy channels' true flux distributions appear to be distributed more normally, contrasting the multi-modal behaviour evident in the higher energy channels. This suggests two distinct states for the higher energy electrons. The modelled flux distributions do not seem to capture the multi-modal behaviour, instead predicting a distribution that looks closer to an 'average' distribution encompassing both states. The modelled fluxes have medians approximately in-line with the medians of the true distribution (signified by the white dot in the box-plot in the centre of each violin).

Comparing between the results of the two models in Figure 6.4, we see that generally they display a lot of similarities (neither capturing the full extent of the distribution or multi-modal behaviour). However, it appears that the solar wind model does manage to capture the multi-modal behaviour better than the index model.

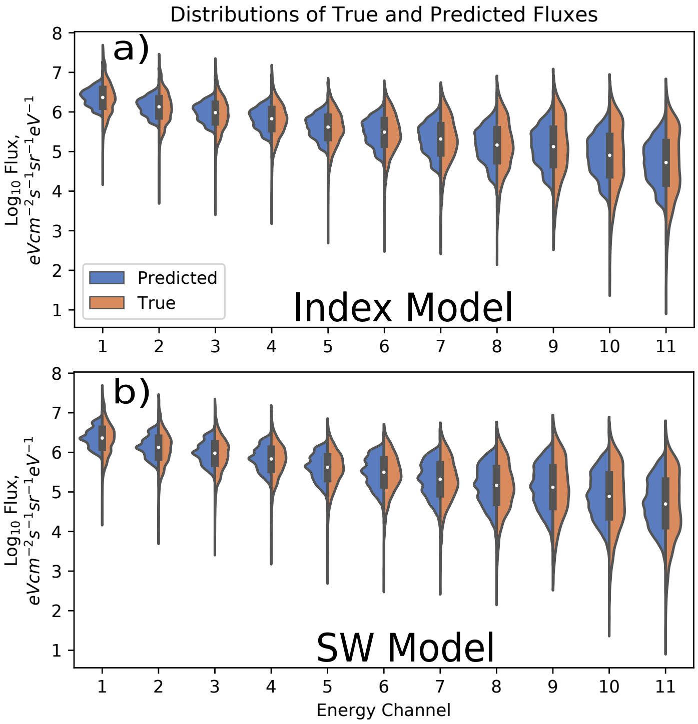


Figure 6.4: *Violin plots presenting the comparison between the true (orange) distribution of flux in the test set and the predicted (blue) distribution of the flux in the test set for each energy channel.*

6.5 Model Performance Analysis

Both of our models perform well and comparably to each other. Using the RMSE (which is in flux units) we calculate our models typically predict within a factor of 2.5 ($10^{0.4}$) for the lower energies and within a factor of 4.5 for the higher energies ($10^{0.65}$). This disparity is likely a result of the increased variability in the fluxes as we go up in energy channels (as can be seen in Figure 6.4). Additionally, we observe that the multi-modality is more prominent at higher energies. This suggests, that while both models use different parameters to construct the outer boundary flux, each set of parameters contains a significant portion of the required information to make an empirical model. Though, since we find variable performance over energy channels, it may be that neither model has the requisite parameters to completely capture the dynamics of the high energy electrons.

The index model has a smaller conjunction MLT limit (8 vs 9 for the solar wind model). Whilst we don't expect that this introduces more complex spatial mapping, it does change how much training data there is. The solar wind model has approximately 20% more training data (roughly 63000 vs 53000 samples). Given this difference, it is somewhat surprising that the model performance is quantitatively so similar. It may be the case that, even if the solar wind model has more data, the geomagnetic indices comprise better state variables to describe the response of the outer boundary fluxes. As such, the index model can learn the mapping more efficiently than the solar wind model can.

To explain why the correlation is better for the higher energy channels when the other metrics show the opposite, we suggest that it is due to the higher energy channels being more responsive to solar wind driving. If there are more significant and consistent changes to the flux at the higher energies (w.r.t. the solar wind/indices), then these features will be more straightforward for the model to learn. In contrast, if, at the lower energies, the response is subtle and more varied (due to the reduced relative contribution of the external driving), then we would expect the model to struggle to capture the behaviour. However, because the lower energy channels are less variable, they are easier to predict over (in terms of absolute flux, rather than relative variation).

It is not entirely clear why the true flux distributions are multi-modal. One contributing factor may be due to the fixed location of the outer boundary in real space sampling different areas of the radiation belts in L* space. On the day side, the fixed boundary may be sampling fluxes with an L* closer to the heart of the radiation belts, rather than the lower fluxes expected at the boundary. Depending on the temporal variability of the true boundary location relative to the static boundary we investigate, it may also be the case that the multi-modal behaviour is due to sampling a mixture of trapped and untrapped electron fluxes over time.

In a Space Weather context, it is important to be able to make accurate predictions during active times. Many metrics can be mis-representative of a model's ability to make good predictions during periods of enhanced geomagnetic activity (as these extremes are less well represented in the dataset). We know already from Figure 6.4 that both models struggle to capture the full variability of the data. However, to probe this more usefully, we present Figure 6.5. This figure presents the RMSE (colour) of each model's outputs per energy channel (x-axes) per decile of SYM-H and B_z (y-axes), respectively.

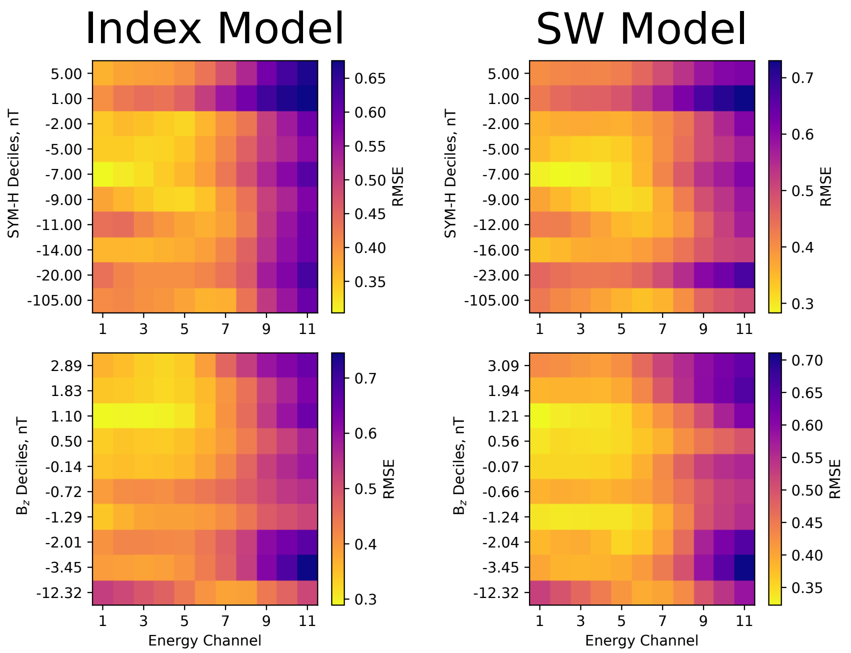


Figure 6.5: *The performance of both the index and solar wind model, split by energy channel and decile of either B_z or SYM-H. The colour represents the RMSE in the given bin.*

In Figure 6.5, the model performance remains significantly lower than the factor of 10 deemed acceptable for RB modelling [e.g., Glauert et al., 2014] for all deciles of B_z and SYM-H (an RMSE of 0.75 corresponds to a factor of 6 error in the prediction). We conclude from this that the model generalises well across activity levels. The main area where performance drops is in the higher deciles of SYM-H and the lower deciles of B_z. This likely corresponds to geomagnetic storm onsets. It is expected that the index model would not be able to capture such dynamic events, though as the solar wind model also fails to capture it suggests that it may be due to the sparsity of storm sudden commencements in the training data.

Finally, in Figure 6.6 we present an analysis of the probabilistic capability of each model. Each panel shows a histogram of the quantile of the predicted distribution into which each true values falls (for three representative energy channels). We use such a histogram to assess the reliability of the predicted distributions. Statistically, if our model reliably predicted the uncertainty, we would expect that the probability of each true value falling into a given quantile of the predicted distribution to be equal (i.e. the overall histogram to be completely flat). Conceptually, this is similar to ensemble verification using rank histograms [Hamill, 2001].

We observe in Figure 6.6a-c) that in general, the probabilities fall close to the 5% value which we would expect, though there is significant bias towards the first quantile bin (percentiles less than 5). With increasing energy a bias towards the 20th quantile (percentiles greater than 95).

In Figure 6.6d-f), we see broadly the same behaviour as in the previous panels, but the bias towards the first and last quantiles is more pronounced. In addition to this, the bias towards the first quantile remains stronger across the energy channels, in contrast to the levelling out at high energies observed for the index model.

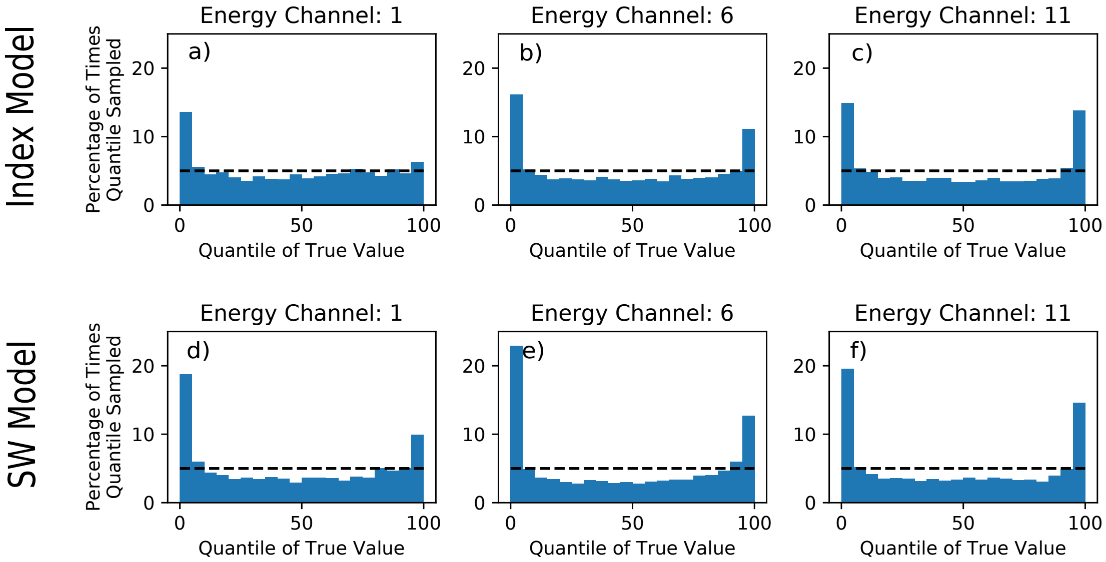


Figure 6.6: *The occurrence frequency of the true flux value lying in a given quantile of a predicted distribution from the index and solar wind models, respectively. The black lines represents the expected outcome from reliable predicted distributions.*

To understand the results presented in Figure 6.6, we take advantage of their similarity to rank histograms and use the discussion in Hamill [2001] (hence H01) to guide us. Figure 1 in H01 presents a range of different rank histograms constructed using a variety of different distributions (means and standard deviations). In general, we observe that if the standard deviation is under-estimated, then there is a bias towards the extremes (as we see in Figure 6.6). We also see that a bias towards the lower end is indicative of over-estimating the mean value. We know from Figure 6.3 that our models have a tendency to over-predict the flux. H01 suggest that the 'U-shape' distribution may be due conditional biases in the model, we believe that it is likely our ensemble models under-estimate the variances since each ensemble member we use (whilst only constituting one aspect of the variability) is fundamentally the same model with randomness injected through the random weights, the optimiser and the bootstrapped training data. This underestimation may originate from the ensemble only representing modes of the posterior distribution of the weights, rather than the full distribution [Izmailov et al., 2021]. It is also the case that compared to some datasets, ours is relatively small, and so the BNNs may not have sufficient information to inform the variances being output.

We do note though, that despite potentially under-estimating the variances, the model both performs well and is better than a simple deterministic model - even if the uncertainty is not perfectly captured. The probabilistic outputs can be used in data-assimilative radiation belt codes and hopefully allow for more accurate modelling.

6.6 Concluding Remarks

Modelling the radiation belts is increasingly important with the proliferation of space-based infrastructure. A crucial aspect of this modelling is the outer boundary condition. We build from the empirically identified boundary location found by Bloch et al. [2021], by here modelling the flux distribution at the boundary.

Our model outputs can be treated as synthetic, equatorial, omni-directional differential energy flux measurements about the radiation belt outer boundary. As such, any radiation belt model which uses 90° pitch-angle fluxes to construct their outer boundary condition can use our model outputs in place of their current flux data. For example, we understand that our model outputs could be used in place of the GOES-15 data used in Glauert et al. [2018], or in place of the CRRES (Combined Release and Radiation Effects Satellite) data used by Shprits et al. [2006] and subsequent works that use the same methodology [e.g., Subbotin and Shprits, 2009; Su et al., 2010a].

To provide both more informative and useful boundary fluxes (for e.g., data-assimilative radiation belt codes), we model the boundary flux probabilistically. To do so, we train deep-ensembles of neural networks (BNNs). Our BNNs capture the aleatoric (or irreducible) uncertainty present in our experiment design (input features and general architecture choice). We use our ensemble to estimate the epistemic (reducible) uncertainty by perturbing the training procedure (initial weights, stochastic optimisation and bootstrapping the training data), allowing us to sample the posterior distribution of the weights of the fully trained networks. We then combine these uncertainties to estimate the total uncertainty associated with each model prediction.

We find comparable deterministic performance between the two models (geomagnetic index and solar wind) - each predicting within a factor of 4.5 (and down to 2.5 for the lower energy channels). The solar wind model appears marginally better reconstructing the underlying distributions of the fluxes, and shows slightly less variability with activity (characterised by B_z and/or SYM-H). Analysing the probabilistic side of the models, we find evidence that both models under-predict the variances associated with predictions, but find that this effect is more pronounced for the solar wind model.

Most importantly, our methodology indicates how uncertainty may be captured and quantified in the models we use to describe the Earth's radiation belts. Though, further work is needed to better characterise the location of the outer boundary, so that methods such as ours can provide the most appropriate synthetic dataset. It seems that a better understanding of the multi-modal distributions we observe in the true flux would also be very useful, since it is appears that our analysis lacks the pertinent parameters needed to express it in the model outputs.

Chapter 7

Conclusions and Future Directions

Space Weather describes the variability of the near-Earth space environment, which can adversely affect the geo-spatial environment. Space Weather can therefore have a severe human impact via loss of satellite services such as GPS or telecommunications; by directly impact ground-based infrastructure such as electrical grids and long pipelines; by inducing magnetic fields which interfere with magnetic navigation systems; or even by increasing the radiation doses airline staff are subjected to due to the changing shape and size of the auroral oval.

Given our collective reliance on many of the systems which can be directly impacted by Space Weather events, the need to understand Space Weather is self-evident. Understanding comes in three main parts: being able to identify the causal factors of different Space Weather phenomena; being able to quantify the magnitude of Space Weather events based on their drivers; and, being able to jointly use this information to forecast Space Weather.

Space Weather research has much foundational work originating well before the spaceflight era, and has at least two dedicated journals - AGU's Space Weather and the Journal of Space Weather and Space Climate - as well as a significant amount of work published in many other journals.

Comparatively, machine learning is a much larger field, even if it is newer. Machine learning has been successfully used across various fields, from being employed for medical image recognition, to semantic language translation, to Space Weather forecasting models. In essence, machine learning is simply a set of tools for pattern recognition and identification. As such, there are clear uses for machine learning in Space Weather research to address the previously mentioned tenets of Space Weather understanding.

Complex machine learning models (e.g., neural networks) have been used in Space Weather research since the 1990s [Newell et al., 1991], though are only now becoming wide-spread as computational power and volumes of data continue to increase well-beyond the limits of less efficient (or manual) methods. More simple machine learning methods, such as linear regression, have certainly been in use for far longer.

This thesis is concerned with leveraging novel machine learning techniques to improve our collective understanding of Space Weather phenomena. The research chapters will subsequently be summarised in the context of space weather, followed by an exploration of their implications and limitations, concluding with thoughts and ideas on the directions that future work could take.

7.1 Solar Wind Classification

7.1.1 Summary of Results

In Chapter 4, we investigated the application of unsupervised machine learning to the problem of solar wind classification. Through both of our unsupervised approaches we found that there are two primary types of solar wind, originating from coronal holes and the streamer belt, respectively. Both of our machine learning classifications also identified a third category which we designated 'unclassified'. Speculation and subsequent work points to the unclassified data being related to ICMEs (see the HelioML textbook on Machine Learning, Statistics, and Data Mining for Heliophysics: https://helioml.org/10/notebook.html).

Our results show significant discrepancies between the two machine learning classifications and the classical speed-thresholding classification scheme. We conclude that using only the solar wind speed for the classification scheme is typically inappropriate, and may reduce the fidelity of subsequent statistical analyses based on such solar wind classifications. Our results in this regard can also be used to identify an appropriate speed threshold that minimises the discrepancy.

Our UMAP results indicate that the streamer belt wind may be more well-ordered than the coronal hole wind after dimension reduction. This contrasts with the typical view of the slow, streamer belt solar wind, which is typically considered to be less well-ordered. Such a result could represent the discovery of a more appropriate latent space to investigate streamer-belt solar wind, rather than the currently used methods which contain more variability.

The main limitation, which precluded the use of these classifications in Chapter 6, is the low temporal resolution of the compositional data. The multi-hour integration time reduces the amount of information that can be drawn from the classifications when they are being used with high-resolution (i.e., minutes) data to model swiftly evolving radiation belt phenomena. Higher temporal-resolution data can be obtained for the composition, but such data have larger uncertainties. Additionally, the composition data obtained from ACE is unreliable after 2012 due to instrument degradation, meaning the classifications can't be applied to more recent data (i.e., the data we use in Chapter 6).

7.1.2 Future Research Directions

Solar wind classification is an interesting problem, and there are myriad directions for future work. From a Space Weather perspective, the most import factors would be to create a classification scheme with higher temporal resolution, that specifically focuses on data from L1 (i.e., OMNI data). From a more esoteric view of solar wind classification, simply to better understand source regions and solar wind propagation, research directions would focus more specifically on using data from missions such as Parker Solar Probe and Solar Orbiter.

Ideally, a new solar wind mission with higher temporal resolution composition measurements would be launched and its data could be used for a new study, though that is an unlikely consequence of the work. Practically, future work should focus on creating classification schemes using data which is readily available at the desired resolution. To this end, more simplistic plasma measurements and magnetic field measurements are the best choice. Though, as most of these parameters vary with distance from the Sun, the classification likely requires a more nuanced approach. Fortunately, if the classification scheme is specifically for Space Weather research, one can simply use the abundant data from L1 - due to the effectively fixed distance from the Sun, the radial evolution is not problematic.

It may be possible to create a classification scheme which can inherently capture the radial evolution, though abundant data across radial distances would be required. If such a classification scheme was built using magnetic field and plasma measurements, then data from missions such as Helios, Ulysses, Parker Solar Probe and Solar Orbiter could all be used to gain a comprehensive understanding of the inner heliosphere.

Our classification schemes, in their current form, should be applicable to Solar Orbiter composition measurements, once they are processed sufficiently by the instrument teams. Doing so should allow for an interesting investigation into the distribution of classification at different radial distance - possibly elucidating the cause of the difference between the classification of Ulysses and ACE data that were observed in Chapter 4.

7.2 Radiation Belt Outer Boundary Location

7.2.1 Summary of Results

In Chapter 5, we used *in situ* data to identify the physical equatorial location of the outer boundary of the outer radiation belt (the OBORB). By using simple machine learning models, we were able to test a large variety of possible OBORB locations. We found that the typical boundary location was located at $8.25\,\mathrm{R_E}$, further out than geosynchronous orbit ($6.6\,\mathrm{R_E}$, where data is abundant) and many of the currently used boundary locations. Due to the nature of the analysis only being able to identify the statistical location, and because magnetospheric quiet times are more prevalent than disturbed times, we conclude that this boundary location corresponds to the nominal quiet-time OBORB.

Our empirical OBORB location serves as an important counter-point to the more subjective methods which are currently used to specify the boundary. Whilst not being fully representative of the likely-dynamic nature of the boundary, the quiet-time boundary is still important for the radiation belts, as the pre-conditioning of the system plays an important part in the subsequent dynamics.

The three main limitations of this result are that (1) it is a static, statistical representation of the boundary location, (2) it may only represent the quiet time boundary location, and (3) due to being at a location far from an abundance of data, it is more difficult to characterise the flux distribution at this location. Points (1) and (2) mean that during the inwards motion associated with disturbed periods of more significant Space Weather, our result no longer represents the true boundary, instead incorporating other magnetospheric (or even solar wind) plasma. It should be noted that the magnetopause does come inside of $6.6\,\mathrm{R_E}$ too, so radiation belt modellers likely already have a process set-up to deal with such situations. As for (3), this makes real-time nowcasting or forecasting of the outer radiation belt difficult, as modellers are left with either having to use a model of the electron fluxes or having to map data from geosynchronous orbit. Both options increase the potential for errors in the subsequent modelling, as there are uncertainties associated with each.

7.2.2 Future Research Directions

To address the static, quiet-time nature of the boundary location, using a similar methodology to that already employed, a larger dataset is required. By increasing the amount of data, it becomes easier (and more reliable/valid) to parameterise the data by either solar wind activity or some index of geomagnetic activity. The current THEMIS dataset could be augmented with data from the Cluster and MMS (magnetospheric multi-scale) missions.

Alternatively, if a method of empirically identifying boundary crossing from time-series of the phase-space density profiles could be developed [as has been done for other magnetospheric boundaries, e.g., Jelínek et al., 2012; Nguyen et al., 2019; Olshevsky et al., 2019; Argall et al., 2020], it may serve as a more appropriate and data-efficient method of identifying the dynamic boundary location. Such a method would require pre-labelled data if it employed supervised machine learning. Labels could be taken from (albeit non-empirical) works such as Ganushkina et al. [2011], and used to parameterise different input data (i.e., electron data rather than ion data). It may also be possible to address the problem in an unsupervised way, identifying time-series segments where the electron distribution functions change from representing radiation belt measurements to measurements of untrapped electrons - though, as is often the case with unsupervised learning, results may be more elusive and require a very nuanced approach to constructing a training dataset to work with (also, some method of disentangling the radial dependencies would be required).

As for the final limitation - it being more difficult to reconstruct the boundary fluxes - this is precisely what we demonstrate to be possible in Chapter 6.

7.3 Radiation Belt Outer Boundary Flux

7.3.1 Summary of Results

In Chapter 6, we created two Bayesian neural network models which can be used to construct a synthetic dataset of the electron flux distribution at the OBORB location. This work builds from the boundary location identified in Chapter 5, bridging the gap between the identification of the OBORB location and its actual use by radiation belt modellers. Aside from providing the OBORB fluxes at an often-more-appropriate location, our use of Bayesian methods gives probabilistic results which indicate the uncertainty in model predictions and thus provide more information than alternative, deterministic methods.

We find comparable deterministic performance between the two models - geomagnetic index and solar wind parameterised. Both models predict the electron fluxes within a factor of 4.5 of the measured fluxes (down to a factor of 2.5 for the lower energy channels).

By creating models which can adequately nowcast the fluxes at a prescribed OBORB location, we demonstrate the relative simplicity in reconstructing the OBORB fluxes at locations beyond those with readily available data. Such models have clear utility in radiation belt modelling, where the outer boundary conditions are derived from the electron flux distributions.

The main limitation of models stems directly from the limitations of our OBORB location - when the radiation belts are compressed inside of our boundary location, the flux distribution the models nowcast is not useful to radiation belt modellers. In the present manifestation there is not a direct solution, outside of the aforementioned ideas to better identify the dynamic boundary location. One alternative would be to create a model that captures the fluxes at all locations over the radiation belts. Some authors [e.g., Smirnov et al., 2020; Claudepierre and O'Brien, 2020] have attempted this, though these typically don't extend to the radial distances which we determine the OBORB to be at - primarily due to the choice of mission data (e.g., GPS satellites or the Van Allen Probes).

Aside from the baked-in limitation due to building from Chapter 5, the next most important missing piece is magnetospheric context. By that, we mean some sort of indication of the preconditioning of the magnetosphere. Geomagnetic indices provide information towards the current state of the magnetosphere, but a more holistic impression of the full state can be derived using time-histories of the solar wind or geomagnetic indices.

We also note in Chapter 6 that the predicted uncertainty is likely underestimated due to the relative homogeneity of the ensemble members' architectures. This means that the model produces over-confident predictions of the flux. This inappropriate uncertainty could lead to less reliable results in e.g., data-assimilative radiation belt models.

7.3.2 Future Research Directions

In terms of aiming to improve model performances, the next step for this project would be to include parameter time-histories as model inputs. However, this task is non-trivial at best, and considerably more tricky given the small dataset size. Initially, one might include instantaneous solar wind variables or geomagnetic indices from a range of previous time-steps, whilst maintaining the same model architecture. However, it is likely that significantly increasing the number of parameters would cause the model to under-fit, as it would not be complex enough to appropriately capture the non-linear parameter interactions. Hence, a larger model would be required to improve the modelling. But, a larger model typically requires more training data - something which is already in short-supply. So, to make the most effective use of time-history, the whole experiment would benefit from an augmented dataset - as with the OBORB location, data could be included from the Cluster and MMS missions. More data would also help to better condition the predicted uncertainties.

A more appropriate approach to including the time-history of the input variables would be to include a recurrent neural network (RNN) to encode the temporal information into a reduced latent-space. The RNN would take the same inputs as described above, but would effectively allow the most important temporal information to be summarised into a manageable set of output features, which could then be used to substitute the model inputs currently used and used with a model architecture similar to the current one.

Through the investigation of different time-histories and model architectures, it would also be possible to keep track of a more diverse range of performant model architectures. These varied models could then be used in ensemble format to produce better predictions and more robust uncertainties.

7.4 Final Remarks

This thesis has explored how machine learning can be applied in the heliospheric domain to better understand and characterise phenomena related to Space Weather. In Chapter 4 we focused on extracting latent information about the driver of Space Weather. This was followed in Chapters 5 and 6 by a closer examination of Space Weather.

Over these research projects, a broad range of machine learning models have been employed - some for the first time in the space-physics domain: from unsupervised to supervised machine learning; from deterministic to probabilistic models; from clustering to dimension reduction; and, from decision trees to Bayesian neural networks. Most of these tools have been very straightforward to implement - the art being in constructing an appropriate experiment and applying domain knowledge to understand the results and limitations.

As demonstrated, machine learning as an empirical tool has the ability to improve identification, theoretical understanding and the modelling of many Space Weather systems. The vast amounts of data that we continue to gather are ideally suited to being studied using the machine learning algorithms discussed in this thesis.

Bibliography

Aggarwal, C. C., Hinneburg, A., and Keim, D. A. (2001). On the Surprising Behavior of Distance Metrics in High Dimensional Space. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 1973, pages 420–434. Springer-Verlag.

Akasofu, S.-I. (2002). *Exploring the Secrets of the Aurora*, volume 273. Springer Netherlands, Dordrecht.

Akasofu, S.-T. (1968). The Growth of the Storm-Time Radiation Belt and the Magnetospheric Substorm. *Geophysical Journal of the Royal Astronomical Society*, 15:7–21.

Albert, J. M., Selesnick, R. S., Morley, S. K., Henderson, M. G., and Kellerman, A. C. (2018). Calculation of Last Closed Drift Shells for the 2013 GEM Radiation Belt Challenge Events. *Journal of Geophysical Research: Space Physics*, 123(11):9597–9611.

Alfvén, H. (1942). Existence of Electromagnetic-Hydrodynamic Waves. *Nature*, 150(3805):405–406.

Alken, P., Maute, A., Richmond, A. D., Vanhamäki, H., and Egbert, G. D. (2017). An application of principal component analysis to the interpretation of ionospheric current systems. *Journal of Geophysical Research: Space Physics*, 122(5):5687–5708.

Altschuler, M. D. and Newkirk, G. (1969). Magnetic fields and the structure of the solar corona. *Solar Physics*, 9(1):131–149.

Amaya, J., Dupuis, R., Innocenti, M. E., and Lapenta, G. (2020). Visualizing and Interpreting Unsupervised Solar Wind Classifications. *Frontiers in Astronomy and Space Sciences*, 7(September).

Anderson, B. R., Millan, R. M., Reeves, G. D., and Friedel, R. H. W. (2015). Acceleration and loss of relativistic electrons during small geomagnetic storms. *Geophysical Research Letters*, 42(23):10113–10119.

Angelopoulos, V. (2008). The THEMIS mission. *Space Science Reviews*, 141(1-4):5–34.

Antiochos, S. K., Mikić, Z., Titov, V. S., Lionello, R., and Linker, J. A. (2011). A model for the sources of the slow solar wind. *Astrophysical Journal*, 731(2).

Antonova, E. and Ganushkina, N. (2000). Inner magnetosphere currents and their role in the magnetosphere dynamics. *Physics and Chemistry of the Earth, Part C: Solar, Terrestrial & Planetary Science*, 25(1-2):23–26.

Antonova, E. E., Stepanova, M. V., Moya, P. S., Pinto, V. A., Vovchenko, V. V., Ovchinnikov, I. L., and Sotnikov, N. V. (2018). Processes in auroral oval and outer electron radiation belt. *Earth, Planets and Space*, 70(1):127.

Argall, M. R., Small, C. R., Piatt, S., Breen, L., Petrik, M., Kokkonen, K., Barnum, J., Larsen, K., Wilder, F. D., Oka, M., Paterson, W. R., Torbert, R. B., Ergun, R. E., Phan, T., Giles, B. L., and Burch, J. L. (2020). MMS SITL Ground Loop: Automating the Burst Data Selection Process. *Frontiers in Astronomy and Space Sciences*, 7(September):1–16.

Attias, H. (2000). A Variational Bayesian Framework for Graphical Models. *Adv. Neur. In.*, 12:209–215.

Azari, A. R., Biersteker, J. B., Dewey, R. M., Doran, G., Forsberg, E. J., Harris, C. D. K., Kerner, H. R., Skinner, K. A., Smith, A. W., Amini, R., Cambioni, S., Da Poian, V., Garton, T. M., Himes, M. D., Millholland, S., and Ruhunusiri, S. (2020). Integrating Machine Learning for Planetary Science: Perspectives for the Next Decade. *Bulletin of the AAS*, 53(4).

Babcock, H. W. (1961). The Topology of the Sun's Magnetic Field and the 22-YEAR Cycle. *The Astrophysical Journal*, 133(9):572.

Bach, F. and Moulines, E. (2011). Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011, NIPS 2011*.

Baker, D., Belian, R., Higbie, P., Klebesadel, R., and Blake, J. (1987). Deep dielectric charging effects due to high-energy electrons in earth's outer magnetosphere. *Journal of Electrostatics*, 20(1):3–19.

Baker, D., Li, X., Blake, J., and Kanekal, S. (1998a). Strong electron acceleration in the Earth's magnetosphere. *Advances in Space Research*, 21(4):609–613.

Baker, D. N. (2001). Satellite Anomalies due to Space Storms. In *Space Storms and Space Weather Hazards*, volume 38, pages 285–311. Springer Netherlands, Dordrecht.

Baker, D. N., Allen, J. H., Kanekal, S. G., and Reeves, G. D. (1998b). Disturbed space environment may have been related to pager satellite failure. *Eos, Transactions American Geophysical Union*, 79(40):477–477.

Baker, D. N., Hoxie, V., Zhao, H., Jaynes, A. N., Kanekal, S., Li, X., and Elkington, S. (2019). Multiyear Measurements of Radiation Belt Electrons: Acceleration, Transport, and Loss. *Journal of Geophysical Research: Space Physics*, 124(4):2588–2602.

Baker, D. N., Kanekal, S. G., Hoxie, V. C., Henderson, M. G., Li, X., Spence, H. E., Elkington, S. R., Friedel, R. H. W., Goldstein, J., Hudson, M. K., Reeves, G. D., Thorne, R. M., Kletzing, C. A., and Claudepierre, S. G. (2013). A Long-Lived Relativistic Electron Storage Ring Embedded in Earth's Outer Van Allen Belt. *Science*, 340(6129):186–190.

Balogh, A., Beek, T. J., Forsyth, R. J., Hedgecock, P. C., Marquedant, R. J., Smith, E. J., Southwood, D. J., and Tsurutani, B. T. (1992). The magnetic field investigation on the Ulysses mission: Instrumentation and preliminary scientific results. *Astron. Astrophys. Suppl. Ser.*, 92:221–236.

Bame, S. J., Asbridge, J. R., Felthauser, H. E., Hones, E. W., and Strong, I. B. (1967). Characteristics of the plasma sheet in the Earth's magnetotail. *Journal of Geophysical Research*, 72(1):113.

Bame, S. J., McComas, D. J., Barraclough, B. L., Phillips, J. L., Sofaly, K. J., Chavez, J. C., Goldstein, B. E., and Sakurai, R. K. (1992). The ULYSSES solar wind plasma experiment. *Astronomy and Astrophysics Supplement Series*, 92:237–265.

Bargatze, L. F. (2005). A new interpretation of Weimer et al.'s solar wind propagation delay technique. *Journal of Geophysical Research*, 110(A7):A07105.

Bargatze, L. F., McPherron, R. L., Minamora, J., and Weimer, D. (2006). Reply to comment by Haaland et al. on "A new interpretation of Weimer et al.'s solar wind propagation delay technique". *Journal of Geophysical Research*, 111(A6):A06103.

Barnes, A. and Hollweg, J. V. (1974). Large-amplitude hydromagnetic waves. *Journal of Geophysical Research*, 79(16):2302–2318.

Belson, W. A. (1959). Matching and Prediction on the Principle of Biological Classification. *Applied Statistics*, 8(2):65.

Bentley, S. N., Stout, J. R., Bloch, T. E., and Watt, C. E. (2020). Random Forest Model of Ultralow-Frequency Magnetospheric Wave Power. *Earth and Space Science*, 7(10):1–28.

Bentley, S. N., Watt, C. E., Owens, M. J., and Rae, I. J. (2018). ULF Wave Activity in the Magnetosphere: Resolving Solar Wind Interdependencies to Identify Driving Mechanisms. *Journal of Geophysical Research: Space Physics*, 123(4):2745–2771.

Beutier, T. and Boscher, D. (1995). A three-dimensional analysis of the electron radiation belt by the Salammbô code. *Journal of Geophysical Research*, 100(A8):14853.

Biermann, L. (1951). Kometenschweife und solare Korpuskularstrahlung. *Z Astrophys*, 29:274–286.

Biewald, L. (2020). Experiment Tracking with Weights & Biases.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer Science+Business Media, LLC, Cambridge, first edition.

Blake, J. B., Kolasinski, W. A., Fillius, R. W., and Mullen, E. G. (1992). Injection of electrons and protons with energies of tens of MeV into L < 3 on 24 March 1991. *Geophysical Research Letters*, 19(8):821–824.

Blei, D. M. and Jordan, M. I. (2006). Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–143.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Bloch, T., Watt, C. E. J., Owens, M. J., Thompson, R. L., and Agiwal, O. (2021). Constraining the Location of the Outer Boundary of Earth's Outer Radiation Belt. *Earth and Space Science*, pages 1–30.

Borovsky, J. E. (2020). A survey of geomagnetic and plasma time lags in the solar-wind-driven magnetosphere of earth. *Journal of Atmospheric and Solar-Terrestrial Physics*, 208:105376.

Borovsky, J. E. and Denton, M. H. (2016). The trailing edges of high-speed streams at 1 AU. *Journal of Geophysical Research: Space Physics*, 121(7):6107–6140.

Borovsky, J. E., Nemzek, R. J., and Belian, R. D. (1993). The occurrence rate of magnetospheric-substorm onsets: Random and periodic substorms. *Journal of Geophysical Research: Space Physics*, 98(A3):3807–3813.

Borovsky, J. E. and Valdivia, J. A. (2018). The Earth's Magnetosphere: A Systems Science Overview and Assessment. *Surveys in Geophysics*, 39(5):817–859.

Borsuk, K. (1948). On the imbedding of systems of compacta in simplicial complexes. *Fundamenta Mathematicae*, 35:217–234.

Boscher, D., Bourdarie, S., Thorne, R., and Abel, B. (2000). Influence of the wave characteristics on the electron radiation belt distribution. *Advances in Space Research*, 26(1):163–166.

Bottou, L. (1991). Stochastic Gradient Learning in Neural Networks. In *Proceedings of Neuro-Nîmes 91*, page 12, Nimes, France. EC2.

Bourdarie, S., Friedel, R. H. W., Fennell, J., Kanekal, S., and Cayton, T. E. (2005). Radiation belt representation of the energetic electron environment: Model and data synthesis using the Salammbô radiation belt transport code and Los Alamos geosynchronous and GPS energetic particle data. *Space Weather*, 3(4).

Boyd, A. J., Turner, D. L., Reeves, G. D., Spence, H. E., Baker, D. N., and Blake, J. B. (2018). What Causes Radiation Belt Enhancements: A Survey of the Van Allen Probes Era. *Geophysical Research Letters*, 45(11):5253–5259.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and Regression Trees*. Taylor & Francis, London, 1 edition.

Brooks, D. H., Ugarte-Urra, I., and Warren, H. P. (2015). Full-Sun observations for identifying the source of the slow solar wind. *Nature Communications*, 6(1):5947.

Buck, R. (1987). Energetic particle drift motions in the outer dayside magnetosphere. Technical Report December, Lawrence Livermore National Laboratory (LLNL), Livermore, CA (United States).

Bullard, E. C. (1949). The magnetic field within the earth. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 197(1051):433–453.

Burlaga, L. F., Mish, W. H., and Whang, Y. C. (1990). Coalescence of recurrent streams of different sizes and amplitudes. *J. Geophys. Res.*, 95(A4):4247.

Burlaga, L. F., Ness, N. F., Berdichevsky, D. B., Park, J., Jian, L. K., Szabo, A., Stone, E. C., and Richardson, J. D. (2019). Magnetic field and particle measurements made by Voyager 2 at and near the heliopause. *Nature Astronomy*, 3(11):1007–1012.

Burlaga, L. F. and Ogilvie, K. W. (1970). MAGNETIC AND THERMAL PRESSURES IN THESOLARWIND L. F. BURLAGA and K. W. OGILVIE. *Solar Physics*, 15(1969):61–71.

Califf, S., Li, X., Zhao, H., Kellerman, A., Sarris, T. E., Jaynes, A., and Malaspina, D. M. (2017). The role of the convection electric field in filling the slot region between the inner and outer radiation belts. *Journal of Geophysical Research: Space Physics*, 122(2):2051–2068.

Campello, R. J. G. B., Moulavi, D., and Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates. In *Lecture Notes in Computer Science*, volume 7819 LNAI, pages 160–172. Springer Berlin Heidelberg.

Camporeale, E. (2019). The Challenge of Machine Learning in Space Weather: Nowcasting and Forecasting. *Space Weather*, 17(8):1166–1207.

Camporeale, E., Carè, A., and Borovsky, J. E. (2017). Classification of Solar Wind with Machine Learning. *J. Geophys. Res. Space*, 122(11):910–10.

Camporeale, E., Cash, M. D., Singer, H. J., Balch, C. C., Huang, Z., and Toth, G. (2020). A Gray-Box Model for a Probabilistic Estimate of Regional Ground Magnetic Perturbations: Enhancing the NOAA Operational Geospace Model With Machine Learning. *Journal of Geophysical Research: Space Physics*, 125(11).

Camporeale, E., Chu, X., Agapitov, O. V., and Bortnik, J. (2019). On the Generation of Probabilistic Forecasts From Deterministic Models. *Space Weather*, 17(3):455–475.

Camporeale, E., Wing, S., and Johnson, J. R. (2018). *Machine Learning Techniques for Space Weather*. Elsevier, Amsterdam, 1 edition.

Cane, H. V. and Richardson, I. G. (2003). Interplanetary coronal mass ejections in the near-Earth solar wind during 1996–2002. *Journal of Geophysical Research*, 108(A4):1156.

Carlsson, G. (2009). Topology and data. *B. Am. Math. Soc.*, 46(2):255–308.

Cauchy, M. A. (1847). Méthode générale pour la résolution des systèmes d'équations simultanées. *Comptes rendus hebdomadaires des séances de l'Académie des Sciences*, 25(2):536–538.

Chapman, S. and Bartels, J. (1941). Geomagnetism. *The Mathematical Gazette*, 25(263):62.

Chapman, S. and Ferraro, V. C. A. (1931). A new theory of magnetic storms. *Journal of Geophysical Research*, 36(2):77.

Chen, P. F. (2011). Coronal mass ejections: Models and their observational basis. *Living Reviews in Solar Physics*, 8.

Chen, S., Shen, B., Wang, X., and Yoo, S.-J. (2019). A Strong Machine Learning Classifier and Decision Stumps Based Hybrid AdaBoost Classification Algorithm for Cognitive Radios. *Sensors*, 19(23):5077.

Chu, P. K. and Lu, X., editors (2013). *Low Temperature Plasma Technology*. CRC Press.

Claudepierre, S. G. and O'Brien, T. P. (2020). Specifying High-Altitude Electrons Using Low-Altitude LEO Systems: The SHELLS Model. *Space Weather*, 18(3).

Cowling, T. G. (1933). The Magnetic Field of Sunspots. *Monthly Notices of the Royal Astronomical Society*, 94(1):39–48.

Crooker, N. U., Gosling, J. T., Bothmer, V., Forsyth, R. J., Gazis, P. R., Hewish, A., Horbury, T. S., Intriligator, D. S., Jokipii, J. R., Kóta, J., Lazarus, A. J., Lee, M. A., Lucek, E., Marsch, E., Posner, A., Richardson, I. G., Roelof, E. C., Schmidt, J. M., Siscoe, G. L., Tsurutani, B. T., Wimmer Schweingruber, R. F., and Wimmer-Schweingruber, R. F. (1999). CIR Morphology, Turbulence, Discontinuities, and Energetic Particles. *Space Science Reviews, v. 89, Issue 1/2, p. 179 220 (1999).*, 89(1):179–220.

Cucinotta, F. A., Hu, S., Schwadron, N. A., Kozarev, K., Townsend, L. W., and Kim, M.-H. Y. (2010). Space radiation risk limits and Earth-Moon-Mars environmental models. *Space Weather*, 8(12):n/a–n/a.

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

Daglis, I. A. (2007). Ring Current Dynamics. *Space Science Reviews*, 124(1-4):183–202.

Daglis, I. A., Thorne, R. M., Baumjohann, W., and Orsini, S. (1999). The terrestrial ring current: Origin, formation, and decay. *Reviews of Geophysics*, 37(4):407–438.

Dessler, A. J. and Karplus, R. (1961). Some effects of diamagnetic ring currents on Van Allen radiation. *Journal of Geophysical Research*, 66(8):2289–2295.

Dialynas, K., Krimigis, S. M., Mitchell, D. G., Decker, R. B., and Roelof, E. C. (2017). The bubble-like shape of the heliosphere observed by Voyager and Cassini. *Nature Astronomy*, 1(5):0115.

Dikpati, M. and Gilman, P. A. (2007). Global solar dynamo models: simulations and predictions of cyclic photospheric fields and long-term non-reversing interior fields. *New Journal of Physics*, 9(8):297–297.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87.

Dungey, J. W. (1961). Interplanetary Magnetic Field and the Auroral Zones. *Physical Review Letters*, 6(2):47–48.

Eastwood, J. P., Biffis, E., Hapgood, M. A., Green, L., Bisi, M. M., Bentley, R. D., Wicks, R., McKinnell, L.-A., Gibbs, M., and Burnett, C. (2017a). The Economic Impact of Space Weather: Where Do We Stand? *Risk Analysis*, 37(2):206–218.

Eastwood, J. P., Hietala, H., Toth, G., Phan, T. D., and Fujimoto, M. (2015). What Controls the Structure and Dynamics of Earth's Magnetosphere? *Space Science Reviews*, 188(1-4):251–286.

Eastwood, J. P., Nakamura, R., Turc, L., Mejnertsen, L., and Hesse, M. (2017b). The Scientific Foundations of Forecasting Magnetospheric Space Weather. *Space Science Reviews*, 212(3-4):1221–1252.

Ebert, R. W., McComas, D. J., Elliott, H. A., Forsyth, R. J., and Gosling, J. T. (2009). Bulk properties of the slow and fast solar wind and interplanetary coronal mass ejections measured by Ulysses: Three polar orbits of observations. *Journal of Geophysical Research: Space Physics*, 114(1).

Eddington, A. S. (1910). The Envelopes of Comet Morehouse (1908 c). *Monthly Notices of the Royal Astronomical Society*, 70(5):442–458.

Efron, B. and Tibshirani, R. (1986). Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statistical Science*, 1(1):54–75.

Eldridge, J., Belkin, M., and Wang, Y. (2015). Beyond hartigan consistency: Merge distortion metric for hierarchical clustering. *Journal of Machine Learning Research*, 40(2015):1–19.

Elkington, S. R., Hudson, M. K., and Chan, A. A. (1999). Acceleration of relativistic electrons via drift-resonant interaction with toroidal-mode Pc-5 ULF oscillations. *Geophysical Research Letters*, 26(21):3273–3276.

Engelhardt, A., Kanawade, R., Knipfer, C., Schmid, M., Stelzle, F., and Adler, W. (2014). Comparing classification methods for diffuse reflectance spectra to improve tissue specific laser surgery. *BMC Medical Research Methodology*, 14(1):91.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, volume 1, pages 226–231. Elsevier.

Estivill-Castro, V. (2002). Why so many clustering algorithms. *ACM SIGKDD Explorations Newsletter*, 4(1):65–75.

Everitt, B. S., Landau, S., Leese, M., and Stahl, D. (2011). *Cluster Analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd, Chichester, UK, 5 edition.

Eyles, C. J., Harrison, R. A., Davis, C. J., Waltham, N. R., Shaughnessy, B. M., Mapson-Menard, H. C., Bewsher, D., Crothers, S. R., Davies, J. A., Simnett, G. M., Howard, R. A., Moses, J. D., Newmark, J. S., Socker, D. G., Halain, J. P., Defise, J. M., Mazy, E., and Rochus, P. (2009). The heliospheric imagers onboard the STEREO mission. *Solar Physics*, 254(2):387–445.

Fairfield, D. H. (1971). Average and unusual locations of the Earth's magnetopause and bow shock. *Journal of Geophysical Research*, 76(28):6700–6716.

Fälthammar, C.-G. (1965). Effects of time-dependent electric fields on geomagnetically trapped radiation. *Journal of Geophysical Research*, 70(11):2503–2516.

Finlay, C. C., Maus, S., Beggan, C. D., Bondar, T. N., Chambodut, A., Chernova, T. A., Chulliat, A., Golovkov, V. P., Hamilton, B., Hamoudi, M., Holme, R., Hulot, G., Kuang, W., Langlais, B., Lesur, V., Lowes, F. J., Lühr, H., Macmillan, S., Mandea, M., McLean, S., Manoj, C., Menvielle, M., Michaelis, I., Olsen, N., Rauberg, J., Rother, M., Sabaka, T. J., Tangborn, A., Tøffner-Clausen, L., Thébault, E., Thomson, A.

W. P., Wardinski, I., Wei, Z., and Zvereva, T. I. (2010). International Geomagnetic Reference Field: the eleventh generation. *Geophysical Journal International*, 183(3):1216–1230.

Finley, A. J., McManus, M. D., Matt, S. P., Kasper, J. C., Korreck, K. E., Case, A. W., Stevens, M. L., Whittlesey, P., Larson, D., Livi, R., Bale, S. D., Dudok de Wit, T., Goetz, K., Harvey, P. R., MacDowall, R. J., Malaspina, D. M., and Pulupa, M. (2020). The contribution of alpha particles to the solar wind angular momentum flux in the inner heliosphere. *Astronomy & Astrophysics*, pages 1–12.

Florek, K., Łukaszewicz, J., Perkal, J., Steinhaus, H., and Zubrzycki, S. (1951). Sur la liaison et la division des points d'un ensemble fini. *Colloquium Mathematicum*, 2(3-4):282–285.

Forsyth, C., Rae, I. J., Murphy, K. R., Freeman, M. P., Huang, C.-L., Spence, H. E., Boyd, A. J., Coxon, J. C., Jackman, C. M., Kalmoni, N. M. E., and Watt, C. E. J. (2016). What effect do substorms have on the content of the radiation belts? *Journal of Geophysical Research: Space Physics*, 121(7):6292–6306.

Forsyth, C., Watt, C. E. J., Rae, I. J., Fazakerley, A. N., Kalmoni, N. M. E., Freeman, M. P., Boakes, P. D., Nakamura, R., Dandouras, I., Kistler, L. M., Jackman, C. M., Coxon, J. C., and Carr, C. M. (2014). Increases in plasma sheet temperature with solar wind driving during substorm growth phases. *Geophysical Research Letters*, 41(24):8713–8721.

Foster, J. C., Erickson, P. J., Omura, Y., and Baker, D. N. (2020). The Impenetrable Barrier: Suppression of Chorus Wave Growth by VLF Transmitters. *Journal of Geophysical Research: Space Physics*, 125(9):1–11.

Frederickson, A., Mullen, E., Brautigam, D., Kerns, K., Robinson, P., and Holman, E. (1991). Radiation-induced insulator discharge pulses in the CRRES internal discharge monitor satellite experiment. *IEEE Transactions on Nuclear Science*, 38(6):1614–1621.

Freund, Y. (2001). An adaptive version of the boost by majority algorithm. *Machine Learning*, 43(3):293–318.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378.

Fukunaga, K. and Hostetler, L. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1):32–40.

Gal, Y. (2016). *Uncertainty in Deep Learning*. PhD thesis, University of Cambridge.

Gamble, R. J., Rodger, C. J., Clilverd, M. A., Sauvaud, J.-A., Thomson, N. R., Stewart, S. L., McCormick, R. J., Parrot, M., and Berthelier, J.-J. (2008). Radiation belt electron precipitation by man-made VLF transmissions. *Journal of Geophysical Research: Space Physics*, 113(A10):1–13.

Ganushkina, N. Y., Dandouras, I., Shprits, Y. Y., and Cao, J. (2011). Locations of boundaries of outer and inner radiation belts as observed by Cluster and Double Star. *Journal of Geophysical Research: Space Physics*, 116(A9).

Ganushkina, N. Y., Liemohn, M. W., and Dubyagin, S. (2018). Current Systems in the Earth's Magnetosphere. *Reviews of Geophysics*, 56(2):309–332.

Garrett, H. B. (1981). The charging of spacecraft surfaces. *Reviews of Geophysics*, 19(4):577.

Geiss, J., Gloeckler, G., and Von Steiger, R. (1995). Origin of the solar wind from composition data. *Space Science Reviews*, 72(1-2):49–60.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2013). *Bayesian Data Analysis*. CRC Press, Taylor and Francis Group, Boca Raton, third edition.

Genestreti, K. J., Kistler, L. M., and Mouikis, C. G. (2012). *The Role and Dynamics of Oxygen Of Ionospheric Origin in Magnetoapuse Reconnection*. PhD thesis, University of New Hampshire.

Ghrist, R. (2014). *Elementary Applied Topology*. Createspace, Philadelphia, 1 edition.

Gilbert, G. (1884). Finley's Tornado Predictions. *American Meteorological Journal*, 1(5):166–172.

Gillbert, W. (1600). *De Magnete*. Peter Short, London, 1 edition.

Glauert, S. A., Horne, R. B., and Meredith, N. P. (2014). Three-dimensional electron radiation belt simulations using the BAS Radiation Belt Model with new diffusion models for chorus, plasmaspheric hiss, and lightning-generated whistlers. *Journal of Geophysical Research: Space Physics*, 119(1):268–289.

Glauert, S. A., Horne, R. B., and Meredith, N. P. (2018). A 30-Year Simulation of the Outer Electron Radiation Belt. *Space Weather*, 16(10):1498–1522.

Gloeckler, G., Cain, J., Ipavich, F. M., Tums, E. O., Bedini, P., Fisk, L. A., Zurbuchen, T. H., Bochsler, P., Fischer, J., Wimmer-Schweingruber, R. F., Geiss, J., and Kallenbach, R. (1998). Investigation of the Composition of Solar and Interstellar Matter Using Solar Wind and Pickup Ion Measurements with SWICS and SWIMS on the Ace Spacecraft. In *The Advanced Composition Explorer Mission*, volume 86, pages 497–539. Springer Netherlands, Dordrecht.

Gloeckler, G., Geiss, J., Balsiger, H., Bedini, F., Cain, J., Fisher, J., Fisk, L., Galvin, A., Gliem, F., Hamilton, D., Hollweg, J., Ipavich, F., Joos, R., Livi, S., Lundgren, R., Mall, U., McKenzie, J. F., Ogilvie, K., Ottens, F., Rieck, W., Tums, E., Steiger, R., Weiss, W., and Wilken, B. (1992). The solar wind ion composition spectrometer. *Astron. Astrophys. Suppl. Ser.*, 92:267–289.

Gloeckler, G., Zurbuchen, T. H., and Geiss, J. (2003). Implications of the observed anticorrelation between solar wind speed and coronal electron temperature. *Journal of Geophysical Research: Space Physics*, 108(A4):1–11.

Glorot, G. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *Proceedings of Machine Learning Research*, 9(1):249–256.

Goerss, P. G. and Jardine, J. F. (2009). *Simplicial Homotopy Theory*, volume 53. Birkhäuser Basel, Basel.

Gonzalez, W. D., Joselyn, J. A., Kamide, Y., Kroehl, H. W., Rostoker, G., Tsurutani, B. T., and Vasyliunas, V. M. (1994). What is a geomagnetic storm? *Journal of Geophysical Research*, 99(A4):5771.

Gosling, J. T. (1996). COROTATING AND TRANSIENT SOLAR WIND FLOWS IN THREE DIMENSIONS. *Annual Review of Astronomy and Astrophysics*, 34(1):35–73.

Gosling, J. T. (2012). Magnetic Reconnection in the Solar Wind. *Space Science Reviews*, 172(1-4):187–200.

Gosling, J. T. and Pizzo, V. J. (1999). Formation and evolution of corotating interaction regions and their three dimensional structure. *Space Science Reviews*, 89(1-2):21–52.

Graves, A. (2013). Generating Sequences With Recurrent Neural Networks. *arXiv*, pages 1–43.

Gredell, D. A., Schroeder, A. R., Belk, K. E., Broeckling, C. D., Heuberger, A. L., Kim, S.-Y., King, D. A., Shackelford, S. D., Sharp, J. L., Wheeler, T. L., Woerner, D. R., and Prenni, J. E. (2019). Comparison of Machine Learning Algorithms for Predictive Modeling of Beef Attributes Using Rapid Evaporative Ionization Mass Spectrometry (REIMS) Data. *Scientific Reports*, 9(1):5721.

Gurney, K. (1997). *An Introduction to Neural Networks*. CRC Press, London, 1 edition.

Haaland, S., Hasegawa, H., De Keyser, J., and Maes, L. (2017). Dawn-Dusk Asymmetries at the Terrestrial Magnetopause. In *Dawn-Dusk Asymmetries in Planetary Plasma Environments*, pages 73–84. American Geophysical Union and John Wiley & Sons, Inc.

Haaland, S., Paschmann, G., and Sonnerup, B. U. □. (2006). Comment on “A new interpretation of Weimer et al.’s solar wind propagation delay technique” by Bargatze et al. *Journal of Geophysical Research*, 111(A6):A06102.

Hale, G. E. (1924). The Law of Sun-Spot Polarity. *Proceedings of the National Academy of Sciences*, 10(1):53–55.

Hale, G. E., Ellerman, F., Nicholson, S. B., and Joy, A. H. (1919). The Magnetic Polarity of Sun-Spots. *The Astrophysical Journal*, 49(26):153.

Hamill, T. M. (2001). Interpretation of Rank Histograms for Verifying Ensemble Forecasts. *Monthly Weather Review*, 129(3):550–560.

Han, Y., Han, B., Hu, Z., Gao, X., Zhang, L., Yang, H., and Li, B. (2020). Prediction and variation of the auroral oval boundary based on a deep learning model and space physical parameters. *Nonlinear Processes in Geophysics*, 27(1):11–22.

Hanlon, P. G., Dougherty, M. K., Forsyth, R. J., Owens, M. J., Hansen, K. C., Tóth, G., Crary, F. J., and Young, D. T. (2004). On the evolution of the solar wind between 1 and 5 AU at the time of the Cassini Jupiter flyby: Multispacecraft observations of interplanetary coronal mass ejections including the formation of a merged interaction region. *Journal of Geophysical Research*, 109(A9):A09S03.

Hartigan, J. A. (1981). Consistency of Single Linkage for High-Density Clusters. *Journal of the American Statistical Association*, 76(374):388–394.

Hastie, T., Tibshirani, R., and Friedman, J. (2008). *The Elements of Statistical Learning*. Springer.

Heidrich-Meisner, V. and Wimmer-Schweingruber, R. F. (2018). Solar Wind Classification Via k -Means Clustering Algorithm. In *Machine Learning Techniques for Space Weather*, chapter 16, pages 397–424. Elsevier, Amsterdam, 1 edition.

Hewins, I. M., Gibson, S. E., Webb, D. F., McFadden, R. H., Kuchar, T. A., Emery, B. A., and McIntosh, S. W. (2020). The Evolution of Coronal Holes over Three Solar Cycles Using the McIntosh Archive. *Solar Physics*, 295(11):161.

Horne, R. B., Thorne, R. M., Horne, B., and Thorne, M. (1998). Potential waves for relativistic electron scattering and stochastic acceleration during magnetic storms. *Geophysical Research Letters*, 25(15):3011–3014.

Horne, R. B., Thorne, R. M., Shprits, Y. Y., Meredith, N. P., Glauert, S. A., Smith, A. J., Kanekal, S. G., Baker, D. N., Engebretson, M. J., Posch, J. L., Spasojevic, M., Inan, U. S., Pickett, J. S., and Decreau, P. M. (2005). Wave acceleration of electrons in the Van Allen radiation belts. *Nature*, 437(7056):227–230.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257.

Hoshi, Y., Hasegawa, H., Kitamura, N., Saito, Y., and Angelopoulos, V. (2018). Seasonal and Solar Wind Control of the Reconnection Line Location on the Earth's Dayside Magnetopause. *Journal of Geophysical Research: Space Physics*, 123(9):7498–7512.

Hu, Z. J., Yang, Q. J., Liang, J. M., Hu, H. Q., Zhang, B. C., and Yang, H. G. (2017). Variation and modeling of ultraviolet auroral oval boundaries associated with interplanetary and geomagnetic parameters. *Space Weather*, 15(4):606–622.

Hudson, M. K., Kress, B. T., Mueller, H. R., Zastrow, J. A., and Bernard Blake, J. (2008). Relationship of the Van Allen radiation belts to solar wind drivers. *Journal of Atmospheric and Solar-Terrestrial Physics*, 70(5):708–729.

Hutchinson, J. A., Wright, D. M., and Milan, S. E. (2011). Geomagnetic storms over the last solar cycle: A superposed epoch analysis. *Journal of Geophysical Research: Space Physics*, 116(9):1–16.

Igel, C., Heidrich-Meisner, V., and Glasmachers, T. (2008). Shark. *Journal of Machine Learning Research*, 9:993–996.

Iyemori, T. (1990). Storm-time magnetospheric currents inferred from mid-latitude geomagnetic field variations. *Journal of geomagnetism and geoelectricity*, 42(11):1249–1265.

Izmailov, P., Vikram, S., Hoffman, M. D., and Wilson, A. G. (2021). What Are Bayesian Neural Network Posteriors Really Like? *arXiv*.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning*, volume 103 of *Springer Texts in Statistics*. Springer New York, New York, NY.

Jaynes, A. N., Baker, D. N., Singer, H. J., Rodriguez, J. V., Loto'aniu, T. M., Ali, A. F., Elkington, S. R., Li, X., Kanekal, S. G., Claudepierre, S. G., Fennell, J. F., Li, W., Thorne, R. M., Kletzing, C. A., Spence, H. E., and Reeves, G. D. (2015). Source and seed populations for relativistic electrons: Their roles in radiation belt changes. *Journal of Geophysical Research: Space Physics*, 120(9):7240–7254.

Jelínek, K., Němeček, Z., and Šafránkov□, J. (2012). A new approach to magnetopause and bow shock modeling based on automated region identification. *Journal of Geophysical Research: Space Physics*, 117(5):1–8.

Jones, S. R., Scott, C. J., Barnard, L. A., Highfield, R., Lintott, C. J., and Baeten, E. (2020). The Visual Complexity of Coronal Mass Ejections Follows the Solar Cycle. *Space Weather*, 18(10).

Jospin, L. V., Buntine, W., Boussaid, F., Laga, H., and Bennamoun, M. (2020). Hands-on Bayesian Neural Networks - a Tutorial for Deep Learning Users. *arXiv*, 1(1):1–35.

Kahler, S. W. (1992). Solar flares and coronal mass ejections. *Annual Review of Astronomy and Astrophysics*, 30(1):113–141.

Kaiser, M. (2005). The STEREO mission: an overview. *Advances in Space Research*, 36(8):1483–1488.

Kalliokoski, M. M., Kilpua, E. K., Osmane, A., Turner, D., Jaynes, A., Turc, L., George, H., and Palmroth, M. (2020). Outer radiation belt and inner magnetospheric response to sheath regions of coronal mass ejections: a statistical analysis. *Annales Geophysicae*, 38(3):683–701.

Kamide, Y., Baumjohann, W., Daglis, I. A., Gonzalez, W. D., Grande, M., Joselyn, J. A., McPherron, R. L., Phillips, J. L., Reeves, E. G. D., Rostoker, G., Sharma, A. S., Singer, H. J., Tsurutani, B. T., and Vasyliunas, V. M. (1998). Current understanding of magnetic storms: Storm-substorm relationships. *Journal of Geophysical Research: Space Physics*, 103(A8):17705–17728.

Kasper, J. C., Bale, S. D., Belcher, J. W., Berthomier, M., Case, A. W., Chandran, B. D., Curtis, D. W., Gallagher, D., Gary, S. P., Golub, L., Halekas, J. S., Ho, G. C., Horbury, T. S., Hu, Q., Huang, J., Klein, K. G., Korreck, K. E., Larson, D. E., Livi, R., Maruca, B., Lavraud, B., Louarn, P., Maksimovic, M., Martinovic, M., McGinnis, D., Pogorelov, N. V., Richardson, J. D., Skoug, R. M., Steinberg, J. T., Stevens, M. L., Szabo, A., Velli, M., Whittlesey, P. L., Wright, K. H., Zank, G. P., MacDowall, R. J., McComas, D. J., McNutt, R. L., Pulupa, M., Raouafi, N. E., and Schwadron, N. A. (2019). Alfvénic velocity spikes and rotational flows in the near-Sun solar wind. *Nature*, 576(7786):228–231.

Kawaguchi, S. and Nishii, R. (2007). Hyperspectral Image Classification by Bootstrap AdaBoost With Random Decision Stumps. *IEEE Transactions on Geoscience and Remote Sensing*, 45(11):3845–3851.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 2017-Decem(Nips):3147–3155.

Keesee, A. M., Pinto, V., Coughlan, M., Lennox, C., Mahmud, M. S., and Connor, H. K. (2020). Comparison of Deep Learning Techniques to Model Connections Between Solar Wind and Ground Magnetic Perturbations. *Frontiers in Astronomy and Space Sciences*, 7(October):1–8.

Khirirat, S., Feyzmahdavian, H. R., and Johansson, M. (2017). Mini-batch gradient descent: Faster convergence under data sparsity. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, volume 56, pages 2880–2887. IEEE.

King, J. H. and Papitashvili, N. E. (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):1–9.

Kingma, D. P. and Ba, J. (2014). Adam: A Method for Stochastic Optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pages 1–15.

Kivelson, M. G. and Russell, C. T., editors (1995). *Introduction to Space Physics*. Cambridge University Press.

Klein, K. L. and Dalla, S. (2017). Acceleration and Propagation of Solar Energetic Particles. *Space Science Reviews*, 212(3-4):1107–1136.

Knetter, T. (2004). Four-point discontinuity observations using Cluster magnetic field data: A statistical survey. *Journal of Geophysical Research*, 109(A6):A06102.

Ko, Y., Raymond, J. C., Zurbuchen, T. H., Riley, P., Raines, J. M., and Strachan, L. (2006). Abundance Variation at the Vicinity of an Active Region and the Coronal Origin of the Slow Solar Wind. *The Astrophysical Journal*, 646(2):1275–1287.

Kramer, H. (2021). GOES-N , O , P Satellites.

Kriegel, H., Kröger, P., Sander, J., and Zimek, A. (2011). Density□based clustering. *WIREs Data Mining and Knowledge Discovery*, 1(3):231–240.

Kubat, M., Holte, R. C., and Matwin, S. (1998). Machine learning for the detection of oil spills in satellite radar images. *Machine Learning*, 30(2-3):195–215.

Kullback, S. and Smith, P. (1978). *Information Theory and Statistics*. Dover Publications, Inc., Gloucester, dover edit edition.

Kurita, S., Miyoshi, Y., Tsuchiya, F., Nishimura, Y., Hori, T., Miyashita, Y., Takada, T., Morioka, A., Angelopoulos, V., McFadden, J. P., Auster, H. U., Albert, J. M., Jordanova, V., and Misawa, H. (2011). Transport and loss of the inner plasma sheet electrons: THEMIS observations. *Journal of Geophysical Research: Space Physics*, 116(A3):1–11.

Kusher, H. J. and Yin, G. G. (2003). *Stochastic Approximation and Recursive Algorithms and Applications*, volume 35 of *Stochastic Modelling and Applied Probability*. Springer-Verlag, New York, second edition.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 2017-Decem(Nips):6403–6414.

Lamy, P. L., Floyd, O., Boclet, B., Wojak, J., Gilardy, H., and Barlyaeva, T. (2019). Coronal Mass Ejections over Solar Cycles 23 and 24. *Space Science Reviews*, 215(5):39.

Langmuir, I. (1928). Oscillations in Ionized Gases. *Proceedings of the National Academy of Sciences*, 14(8):627–637.

Larmor, J. (1919a). How could a Rotating Body such as the Sun Become a Magnet?, Concrete Mine Timbers. *Scientific American*, 88(2286supp):287–287.

Larmor, J. (1919b). Possible rotational origin of magnetic fields of sun and earth. *Electrical Review*, 85:412.

Legrand, J. P. ; Simon, P. A. (1985). Some solar cycle phenomena related to the geomagnetic activity from 1868 to 1980. I - The shock events, or the interplanetary expansion of the toroidal field. *Astronomy and Astrophysics*, 152(2):119–204.

Leka, K. D., Park, S.-H., Kusano, K., Andries, J., Barnes, G., Bingham, S., Bloomfield, D. S., McCloskey, A. E., Delouille, V., Falconer, D., Gallagher, P. T., Georgoulis, M. K., Kubo, Y., Lee, K., Lee, S., Lobzin, V., Mun, J., Murray, S. A., Hamad Nageem, T. A. M., Qahwaji, R., Sharpe, M., Steenburgh, R. A., Steward, G., and Terkildsen, M. (2019). A Comparison of Flare Forecasting Methods. II. Benchmarks, Metrics, and Performance Results for Operational Solar Flare Forecasting Systems. *The Astrophysical Journal Supplement Series*, 243(2):36.

Lemaréchal, C. (2012). Cauchy and the Gradient Method. *Documenta Mathematica*, ISMP:251–254.

Lewis, D. D. and Gale, W. A. (1994). A Sequential Algorithm for Training Text Classifiers. *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1994*, pages 3–12.

Li, H., Wang, C., Tu, C., and Xu, F. (2020). Machine Learning Approach for Solar Wind Categorization. *Earth and Space Science*, 7(5).

Li, W. and Hudson, M. K. (2019). Earth's Van Allen Radiation Belts: From Discovery to the Van Allen Probes Era. *Journal of Geophysical Research: Space Physics*, 124(11):8319–8351.

Li, W., Thorne, R. M., Bortnik, J., Baker, D. N., Reeves, G. D., Kanekal, S. G., Spence, H. E., and Green, J. C. (2015). Solar wind conditions leading to efficient radiation belt electron acceleration: A superposed epoch analysis. *Geophysical Research Letters*, 42(17):6906–6915.

Li, W., Thorne, R. M., Bortnik, J., Nishimura, Y., and Angelopoulos, V. (2011). Modulation of whistler mode chorus waves: 1. Role of compressional Pc4-5 pulsations. *Journal of Geophysical Research: Space Physics*, 116(A6):n/a–n/a.

Li, W., Thorne, R. M., Bortnik, J., Nishimura, Y., Angelopoulos, V., Chen, L., McFadden, J. P., and Bonnell, J. W. (2010a). Global distributions of suprathermal electrons observed on THEMIS and potential mechanisms for access into the plasmasphere. *Journal of Geophysical Research: Space Physics*, 115(12):1–14.

Li, W., Thorne, R. M., Nishimura, Y., Bortnik, J., Angelopoulos, V., McFadden, J. P., Larson, D. E., Bonnell, J. W., Le Contel, O., Roux, A., and Auster, U. (2010b). THEMIS analysis of observed equatorial electron distributions responsible for the chorus excitation. *Journal of Geophysical Research: Space Physics*, 115(A6):n/a–n/a.

Li, X., Schiller, Q., Blum, L., Califf, S., Zhao, H., Tu, W., Turner, D. L., Gerhardt, D., Palo, S., Kanekal, S., Baker, D. N., Fennell, J., Blake, J. B., Looper, M., Reeves, G. D., and Spence, H. (2013). First results from CSSWE CubeSat: Characteristics of relativistic electrons in the near-Earth environment during the October 2012 magnetic storms. *Journal of Geophysical Research: Space Physics*, 118(10):6489–6499.

Lin, R. P., Anderson, K. A., Ashford, S., Carlson, C., Curtis, D., Ergun, R., Larson, D., McFadden, J., McCarthy, M., Parks, G. K., Rème, H., Bosqued, J. M., Coutelier, J., Cotin, F., D'Uston, C., Wenzel, K. P., Sanderson, T. R., Henrion, J., Ronnet, J. C., and Paschmann, G. (1995). A three-dimensional plasma and energetic particle investigation for the wind spacecraft. *Space Science Reviews*, 71(1-4):125–153.

Liu, S., Yan, Q., Yang, C., Zhou, Q., He, Z., He, Y., Gao, Z., and Xiao, F. (2018). Quantifying Extremely Rapid Flux Enhancements of Radiation Belt Relativistic Electrons Associated With Radial Diffusion. *Geophysical Research Letters*, 45(3):1262–1270.

Loshchilov, I. and Hutter, F. (2017). Decoupled Weight Decay Regularization. *arXiv*.

Lugaz, N., Farrugia, C. J., Huang, C. L., Winslow, R. M., Spence, H. E., and Schwadron, N. A. (2016). Earth's magnetosphere and outer radiation belt under sub-Alfvénic solar wind. *Nature Communications*, 7:1–7.

Lui, A. T. Y. (2003). Inner magnetospheric plasma pressure distribution and its local time asymmetry. *Geophysical Research Letters*, 30(16):7–10.

Lui, A. T. Y., McEntire, R. W., and Krimigis, S. M. (1987). Evolution of the ring current during two geomagnetic storms. *Journal of Geophysical Research*, 92(A7):7459.

Lyons, L. R. and Thorne, R. M. (1973). Equilibrium structure of radiation belt electrons. *Journal of Geophysical Research*, 78(13):2142–2149.

Mackay, D. J. C. (1995). Probable networks and plausible predictions — a review of practical Bayesian methods for supervised neural networks. *Network: Computation in Neural Systems*, 6(3):469–505.

MacQueen, J. (1967). Some Methods For Classification and Analysis of Multivariate Observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297.

Maget, V., Sicard-Piet, A., Bourdarie, S., Lazaro, D., Turner, D. L., Daglis, I. A., and Sandberg, I. (2015). Improved outer boundary conditions for outer radiation belt data assimilation using THEMIS-SST data and the Salammbo-EnKF code. *Journal of Geophysical Research: Space Physics*, 120(7):5608–5622.

Mann, I. R., Lee, E. A., Claudepierre, S. G., Fennell, J. F., Degeling, A., Rae, I. J., Baker, D. N., Reeves, G. D., Spence, H. E., Ozeke, L. G., Rankin, R., Milling, D. K., Kale, A., Friedel, R. H. W., and Honary, F. (2013). Discovery of the action of a geophysical synchrotron in the Earth's Van Allen radiation belts. *Nature Communications*, 4(1):2795.

Marsch, E. (2006). Kinetic Physics of the Solar Corona and Solar Wind. *Living Reviews in Solar Physics*, 3.

Masiás-Meza, J. J., Dasso, S., Démoulin, P., Rodriguez, L., and Janvier, M. (2016). Superposed epoch study of ICME sub-structures near Earth and their effects on Galactic cosmic rays. *Astronomy and Astrophysics*, 592:1–13.

Maunder, E. W. (1903). Spöerer's Law of Zones. *Observatory*, 26:329–330.

McComas, D. J., Angold, N., Elliott, H. A., Livadiotis, G., Schwadron, N. A., Skoug, R. M., and Smith, C. W. (2013). WEAKEST SOLAR WIND OF THE SPACE AGE AND THE CURRENT "MINI" SOLAR MAXIMUM. *The Astrophysical Journal*, 779(1):2.

McComas, D. J., Bame, S. J., Barker, P., Feldman, W. C., Phillips, J. L., Riley, P., and Griffee, J. W. (1998). Solar wind electron proton alpha monitor (SWEPAM) for the Advanced Composition Explorer. *Space Science Reviews*, 86(1-4):563–612.

McComas, D. J., Ebert, R. W., Elliott, H. A., Goldstein, B. E., Gosling, J. T., Schwadron, N. A., and Skoug, R. M. (2008). Weaker solar wind from the polar coronal holes and the whole Sun. *Geophysical Research Letters*, 35(18):1–5.

McFadden, J. P., Carlson, C. W., Larson, D., Ludlam, M., Abiad, R., Elliott, B., Turin, P., Marckwordt, M., and Angelopoulos, V. (2008). The THEMIS ESA plasma instrument and in-flight calibration. *Space Science Reviews*, 141(1-4):277–302.

McIlwain, C. E. (1966). Ring current effects on trapped particles. *Journal of Geophysical Research*, 71(15):3623–3628.

McInnes, L. and Healy, J. (2017). Accelerated Hierarchical Density Based Clustering. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, volume 2017-Novem, pages 33–42. IEEE.

McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv*.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Meng, C.-I. and Mihalov, J. D. (1972). Average plasma-sheet configuration near 60 Earth radii. *Journal of Geophysical Research*, 77(10):1739–1755.

Meredith, N. P. (2002). Outer zone relativistic electron acceleration associated with substorm-enhanced whistler mode chorus. *Journal of Geophysical Research*, 107(A7):1144.

Meredith, N. P., Horne, R. B., Isles, J. D., and Rodriguez, J. V. (2015). Extreme relativistic electron fluxes at geosynchronous orbit: Analysis of GOES E > 2 MeV electrons. *Space Weather*, 13(3):170–184.

Meredith, N. P., Horne, R. B., Shen, X., Li, W., and Bortnik, J. (2020). Global Model of Whistler Mode Chorus in the Near□Equatorial Region (| λ m |< 18°). *Geophysical Research Letters*, 47(11).

Milan, S. E. (2009). Both solar wind-magnetosphere coupling and ring current intensity control of the size of the auroral oval. *Geophysical Research Letters*, 36(18):2–5.

Milan, S. E., Carter, J. A., Korth, H., and Anderson, B. J. (2015). Principal component analysis of Birkeland currents determined by the Active Magnetosphere and Planetary Electrodynamics Response Experiment. *Journal of Geophysical Research: Space Physics*, 120(12):10415–10424.

Milan, S. E., Provan, G., and Hubert, B. (2007). Magnetic flux transport in the Dungey cycle: A survey of dayside and nightside reconnection rates. *Journal of Geophysical Research: Space Physics*, 112(A1):n/a–n/a.

Miyoshi, Y., Sakaguchi, K., Shiokawa, K., Evans, D., Albert, J., Connors, M., and Jordanova, V. (2008). Precipitation of radiation belt electrons by EMIC waves, observed from ground and space. *Geophysical Research Letters*, 35(23):L23101.

Montavon, G., Orr, G. B., and Müller, K.-R., editors (2012). *Neural Networks: Tricks of the Trade*. Number MAY 2000 in Lecture Notes in Computer Science. Springer Berlin Heidelberg, Berlin, Heidelberg, 2 edition.

Morley, S. K., Friedel, R. H. W., Spanswick, E. L., Reeves, G. D., Steinberg, J. T., Koller, J., Cayton, T., and Noveroske, E. (2010a). Dropouts of the outer electron radiation belt in response to solar wind stream interfaces: global positioning system observations. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 466(2123):3329–3350.

Morley, S. K., Welling, D. T., Koller, J., Larsen, B. A., Henderson, M. G., and Niehof, J. (2010b). SpacePy-A Python-based library of tools for the space sciences. In *Proceeding of the 9th Python in Science Conference*, pages 67–72, Austin.

Moya, P. S., Pinto, V. A., Sibeck, D. G., Kanekal, S. G., and Baker, D. N. (2017). On the Effect of Geomagnetic Storms on Relativistic Electrons in the Outer Radiation Belt: Van Allen Probes Observations. *Journal of Geophysical Research: Space Physics*, 122(11):100–11.

Murphy, K. R., Mann, I. R., Sibeck, D. G., Rae, I. J., Watt, C., Ozeke, L. G., Kanekal, S. G., and Baker, D. N. (2020). A Framework for Understanding and Quantifying the Loss and Acceleration of Relativistic Electrons in the Outer Radiation Belt During Geomagnetic Storms. *Space Weather*, 18(5):1–15.

Murphy, K. R., Watt, C. E. J., Mann, I. R., Jonathan Rae, I., Sibeck, D. G., Boyd, A. J., Forsyth, C. F., Turner, D. L., Claudepierre, S. G., Baker, D. N., Spence, H. E., Reeves, G. D., Blake, J. B., and Fennell, J. (2018). The Global Statistical Response of the Outer Radiation Belt During Geomagnetic Storms. *Geophysical Research Letters*, 45(9):3783–3792.

Mursula, K., Lukianova, R., and Holappa, L. (2015). Occurence of High-Speed Solar Wind Streams Over The Grand Modern Maximum. *The Astrophysical Journal*, 801(1):30.

Mursula, K. and Zieger, B. (1996). The 13.5-day periodicity in the Sun, solar wind, and geomagnetic activity: The last three solar cycles. *Journal of Geophysical Research: Space Physics*, 101(A12):27077–27090.

Nakamura, T. K., Eriksson, S., Hasegawa, H., Zenitani, S., Li, W. Y., Genestreti, K. J., Nakamura, R., and Daughton, W. (2017). Mass and Energy Transfer Across the Earth's Magnetopause Caused by Vortex-Induced Reconnection. *Journal of Geophysical Research: Space Physics*, 122(11):505–11.

Neal, R. M. (1996). *Bayesian Learning for Neural Networks*, volume 118 of *Lecture Notes in Statistics*. Springer New York, New York, NY.

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

Neugebauer, M. and Snyder, C. W. (1966). Mariner 2 observations of the solar wind: 1. Average properties. *Journal of Geophysical Research*, 71(19):4469–4484.

Newell, P. T., Wing, S., Meng, C.-I., and Sigillito, V. (1991). The auroral oval position, structure, and intensity of precipitation from 1984 onward: An automated online data base. *Journal of Geophysical Research*, 96(A4):5877.

Nguyen, G., Aunai, N., Michotte de Welle, B., Jeandet, A., and Fontaine, D. (2019). Automatic detection of the Earth Bow Shock and Magnetopause from in-situ data with machine learning. *Annales Geophysicae*, 20(November):1–22.

Ni, B., Shprits, Y., Hartinger, M., Angelopoulos, V., Gu, X., and Larson, D. (2011). Analysis of radiation belt energetic electron phase space density using THEMIS SST measurements: Cross-satellite calibration and a case study. *Journal of Geophysical Research: Space Physics*, 116(3):1–14.

Nwankpa, C. E. (2020). Advances in optimisation algorithms and techniques for deep learning. *Advances in Science, Technology and Engineering Systems*, 5(5):563–577.

Olifer, L., Mann, I. R., Morley, S. K., Ozeke, L. G., and Choi, D. (2018). On the Role of Last Closed Drift Shell Dynamics in Driving Fast Losses and Van Allen Radiation Belt Extinction. *Journal of Geophysical Research: Space Physics*, 123(5):3692–3703.

Olshevsky, V., Khotyaintsev, Y. V., Divin, A., Delzanno, G. L., Anderzen, S., Herman, P., Chien, S. W. D., Avanov, L., and Markidis, S. (2019). Automated classification of plasma regions using 3D particle energy distribution. *J. Geophys. Res. Space*, pages 1–16.

Olson, W. P. and Pfitzer, K. A. (1974). A quantitative model of the magnetospheric magnetic field. *Journal of Geophysical Research*, 79(25):3739–3748.

Omura, Y., Katoh, Y., and Summers, D. (2008). Theory and simulation of the generation of whistler-mode chorus. *Journal of Geophysical Research: Space Physics*, 113(A4):n/a–n/a.

Onsager, T., Grubb, R., Kunches, J., Matheson, L., Speich, D., Zwickl, R. W., and Sauer, H. (1996). Operational uses of the GOES energetic particle detectors. In Washwell, E. R., editor, *GOES-8 and Beyond*, volume 2812, pages 281–290.

Orcinha, M., Tomassetti, N., Barão, F., and Bertucci, B. (2019). Observation of a time lag in solar modulation of cosmic rays in the heliosphere. *Journal of Physics: Conference Series*, 1181(1):012013.

Oshiro, T. M., Perez, P. S., and Baranauskas, J. A. (2012). How Many Trees in a Random Forest? In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 7376 LNAI, pages 154–168. Springer Berlin Heidelberg.

Ostapenko, A. A. and Maltsev, Y. P. (1997). Relation of the magnetic field in the magnetosphere to the geomagnetic and solar wind activity. *Journal of Geophysical Research: Space Physics*, 102(A8):17467–17473.

Owens, M. J., Crooker, N. U., and Lockwood, M. (2014). Solar cycle evolution of dipolar and pseudostreamer belts and their relation to the slow solar wind. *Journal of Geophysical Research: Space Physics*, 119(1):36–46.

Owens, M. J., Lockwood, M., and Barnard, L. A. (2020). The Value of CME Arrival Time Forecasts for Space Weather Mitigation. *Space Weather*, 18(9).

Owens, M. J. and Riley, P. (2017). Probabilistic Solar Wind Forecasting Using Large Ensembles of Near-Sun Conditions With a Simple One-Dimensional “Upwind” Scheme. *Space Weather*, 15(11):1461–1474.

Owens, M. J., Riley, P., and Horbury, T. S. (2017). Probabilistic Solar Wind and Geomagnetic Forecasting Using an Analogue Ensemble or “Similar Day” Approach. *Solar Physics*, 292(5):69.

Owocki, S. P., Holzer, T. E., and Hundhausen, A. J. (1983). The solar wind ionization state as a coronal temperature diagnostic. *The Astrophysical Journal*, 275:354.

Ozeke, L. G., Mann, I. R., Murphy, K. R., Degeling, A. W., Claudepierre, S. G., and Spence, H. E. (2018). Explaining the apparent impenetrable barrier to ultra-relativistic electrons in the outer Van Allen belt. *Nature Communications*, 9(1):1844.

Ozeke, L. G., Mann, I. R., Murphy, K. R., Jonathan Rae, I., and Milling, D. K. (2014). Analytic expressions for ULF wave radiation belt radial diffusion coefficients. *Journal of Geophysical Research: Space Physics*, 119(3):1587–1605.

Öztürk, M. K. (2012). Trajectories of charged particles trapped in Earth's magnetic field. *American Journal of Physics*, 80(5):420–428.

Pagel, A. C. (2004). Correlation of solar wind entropy and oxygen ion charge state ratio. *Journal of Geophysical Research*, 109(A1):A01113.

Panasenco, O. and Velli, M. (2013). Coronal pseudostreamers: Source of fast or slow solar wind? In *AIP Conference Proceedings*, volume 1539, pages 50–53.

Parker, E. N. (1958). Dynamics of the Interplanetary Gas and Magnetic Fields. *The Astrophysical Journal*, 128(1905):664.

Parker, E. N. (1961). The Stellar-Wind Regions. *The Astrophysical Journal*, 134(600):20.

Pedregosa, F., Varoquaux, G., Buitinck, L., Louppe, G., Grisel, O., Pedregosa, F., and Mueller, A. (2011). Scikit-learn. *Journal of Machine Learning Research*, 12:2825–2830.

Piccinelli, R. and Krausmann, E. (2014). Space Weather and Power Grids – A Vulnerability Assessment 2 0 1 4. Technical report, European Commission.

Pires de Lima, R., Chen, Y., and Lin, Y. (2020). Forecasting Megaelectron□Volt Electrons Inside Earth's Outer Radiation Belt: PreMevE 2.0 Based on Supervised Machine Learning Algorithms. *Space Weather*, 18(2):1–23.

Pirjola, R. (2007). Space weather effects on power grids. In *Space Weather-Physics and Effects*, chapter 10, pages 269–288. Springer Berlin Heidelberg, Berlin, Heidelberg.

Pizzo, V. (1978). A three-dimensional model of corotating streams in the solar wind, 1. Theoretical foundations. *Journal of Geophysical Research*, 83(A12):5563.

Polyak, B. T. and Juditsky, A. B. (1992). Acceleration of Stochastic Approximation by Averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855.

Pulkkinen, A., Rastätter, L., Kuznetsova, M., Singer, H., Balch, C., Weimer, D., Toth, G., Ridley, A., Gombosi, T., Wiltberger, M., Raeder, J., and Weigel, R. (2013). Community-wide validation of geospace model ground magnetic field perturbation predictions to support model transition to operations. *Space Weather*, 11(6):369–385.

Pulkkinen, T. (2007). Space Weather: Terrestrial Perspective. *Living Reviews in Solar Physics*, 4:1.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1):81–106.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., 1 edition.

Rankin, J. S., McComas, D. J., Richardson, J. D., and Schwadron, N. A. (2019). Heliosheath Properties Measured from a Voyager 2 to Voyager 1 Transient. *The Astrophysical Journal*, 883(1):101.

Rashmi, K. V. and Gilad-Bachrach, R. (2015). DART: Dropouts meet Multiple Additive Regression Trees. *Journal of Machine Learning Research*, 38:489–497.

Reeves, G. D., Chen, Y., Cunningham, G. S., Friedel, R. W. H., Henderson, M. G., Jordanova, V. K., Koller, J., Morley, S. K., Thomsen, M. F., and Zaharia, S. (2012). Dynamic Radiation Environment Assimilation Model: DREAM. *Space Weather*, 10(3).

Reeves, G. D., Friedel, R. H. W., Larsen, B. A., Skoug, R. M., Funsten, H. O., Claudepierre, S. G., Fennell, J. F., Turner, D. L., Denton, M. H., Spence, H. E., Blake, J. B., and Baker, D. N. (2016). Energy-dependent dynamics of keV to MeV electrons in the inner zone, outer zone, and slot regions. *Journal of Geophysical Research: Space Physics*, 121(1):397–412.

Reeves, G. D., Fritz, T. A., Cayton, T. E., and Belian, R. D. (1990). Multi-satellite measurements of the substorm injection region. *Geophysical Research Letters*, 17(11):2015–2018.

Reeves, G. D., McAdams, K. L., Friedel, R. H. W., and O’Brien, T. P. (2003). Acceleration and loss of relativistic electrons during geomagnetic storms. *Geophysical Research Letters*, 30(10):n/a–n/a.

Richardson, I. G. (2004). Identification of interplanetary coronal mass ejections at 1 AU using multiple solar wind plasma composition anomalies. *Journal of Geophysical Research*, 109(A9):A09104.

Richardson, I. G. (2018). Solar wind stream interaction regions throughout the heliosphere. *Living Reviews in Solar Physics*, 15(1):1.

Rigler, E. J., Baker, D. N., Weigel, R. S., and Vassiliadis, D. (2005). Solar wind-driven electron radiation belt response functions at 100-min time scales. *Advances in Space Research*, 36(12):2401–2406.

Riley, P. and Ben-Nun, M. (2021). On the Sources and Sizes of Uncertainty in Predicting the Arrival Time of Interplanetary Coronal Mass Ejections Using Global MHD Models. *Space Weather*, 19(6):1–15.

Riley, P., Mays, M. L., Andries, J., Amerstorfer, T., Biesecker, D., Delouille, V., Dumbović, M., Feng, X., Henley, E., Linker, J. A., Möstl, C., Nuñez, M., Pizzo, V., Temmer, M., Tobiska, W. K., Verbeke, C., West, M. J., and Zhao, X. (2018). Forecasting the Arrival Time of Coronal Mass Ejections: Analysis of the CCMC CME Scoreboard. *Space Weather*, 16(9):1245–1260.

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

Roberts, P. H. (2007). Alfvén's Theorem and the Frozen Flux Approximation. *Encyclopedia of Geomagnetism and Paleomagnetism*, pages 7–11.

Rodger, C. J., Hendry, A. T., Clilverd, M. A., Kletzing, C. A., Brundell, J. B., and Reeves, G. D. (2015). High-resolution in situ observations of electron precipitation-causing EMIC waves. *Geophysical Research Letters*, 42(22):9633–9641.

Roederer, J. G. (1967). On the adiabatic motion of energetic particles in a model magnetosphere. *Journal of Geophysical Research*, 72(3):981–992.

Roederer, J. G. and Lejosne, S. (2018). Coordinates for Representing Radiation Belt Particle Flux. *Journal of Geophysical Research: Space Physics*, 123(2):1381–1387.

Roederer, J. G. and Zhang, H. (2014). *Dynamics of Magnetically Trapped Particles*. Astrophysics and Space Science Library. Springer Berlin Heidelberg, Berlin, Heidelberg, 2 edition.

Russell, C. and Angelopoulos, V. (2014). *The ARTEMIS Mission*, volume 9781461495. Springer New York, New York, NY.

Russell, S. J. and Norvig, P. (2009). *Artificial Intelligence A Modern Approach*. Prentice Hall, Pearson., Upper Saddle River, third edition.

Salzberg, S., Chandar, R., Ford, H., Murthy, S. K., and White, R. (1995). Decision trees for automated identification of cosmic-ray hits in Hubble Space Telescope images. *Publications of the Astronomical Society of the Pacific*, 107(709):279.

Sandhu, J. K., Rae, I. J., Freeman, M. P., Forsyth, C., Gkioulidou, M., Reeves, G. D., Spence, H. E., Jackman, C. M., and Lam, M. M. (2018). Energization of the

Ring Current by Substorms. *Journal of Geophysical Research: Space Physics*, 123(10):8131–8148.

Schatten, K. H., Wilcox, J. M., and Ness, N. F. (1969). A model of interplanetary and coronal magnetic fields. *Solar Physics*, 6(3):442–455.

Schwabe, H. (1843). Sonnen — Beobachtungen im Jahre 1843. *Astronomische Nachrichten*, 21(15):234–235.

Schwadron, N. A., McComas, D. J., Elliott, H. A., Gloeckler, G., Geiss, J., and von Steiger, R. (2005). Solar wind from the coronal hole boundaries. *Journal of Geophysical Research: Space Physics*, 110(A4):1–8.

Scott, C. J., Owens, M. J., de Koning, C. A., Barnard, L. A., Jones, S. R., and Wilkinson, J. (2019). Using Ghost Fronts Within STEREO Heliospheric Imager Data to Infer the Evolution in Longitudinal Structure of a Coronal Mass Ejection. *Space Weather*, 17(4):539–552.

Selesnick, R. S., Looper, M. D., and Mewaldt, R. A. (2007). A theoretical model of the inner proton radiation belt. *Space Weather*, 5(4):n/a–n/a.

Sergeev, V. A., Dmitrieva, N. P., Stepanov, N. A., Sormakov, D. A., Angelopoulos, V., and Runov, A. V. (2015). On the plasma sheet dependence on solar wind and substorms and its role in magnetosphere-ionosphere coupling. *Earth, Planets and Space*, 67(1):133.

Sheeley, N. R., Harvey, J. W., and Feldman, W. C. (1976). Coronal holes, solar wind streams, and recurrent geomagnetic disturbances: 1973–1976. *Solar Physics*, 49(2):271–278.

Sheeley, Jr., N. R., Wang, Y., Hawley, S. H., Brueckner, G. E., Dere, K. P., Howard, R. A., Koomen, M. J., Korendyke, C. M., Michels, D. J., Paswaters, S. E., Socker, D. G., St. Cyr, O. C., Wang, D., Lamy, P. L., Llebaria, A., Schwenn, R., Simnett, G. M., Plunkett, S., and Biesecker, D. A. (1997). Measurements of Flow Speeds in the Corona Between 2 and 30 R ▯. *The Astrophysical Journal*, 484(1):472–478.

Shibata, K. (1996). New observational facts about solar flares from Yohkoh studies — Evidence of magnetic reconnection and a unified model of flares. *Advances in Space Research*, 17(4-5):9–18.

Shin, D.-K. and Lee, D.-Y. (2013). Determining radial boundary conditions of outer radiation belt electrons using THEMIS observations. *Journal of Geophysical Research: Space Physics*, 118(6):2888–2896.

Shin, D.-K., Lee, D.-Y., Kim, J.-H., and Cho, J.-H. (2014). Prediction Model of the Outer Radiation Belt Developed by Chungbuk National University. *Journal of Astronomy and Space Sciences*, 31(4):303–309.

Shprits, Y. Y., Thorne, R. M., Friedel, R., Reeves, G. D., Fennell, J., Baker, D. N., and Kanekal, S. G. (2006). Outward radial diffusion driven by losses at magnetopause. *Journal of Geophysical Research*, 111(A11):A11214.

Shue, J.-H., Song, P., Russell, C. T., Steinberg, J. T., Chao, J. K., Zastenker, G., Vaisberg, O. L., Kokubun, S., Singer, H. J., Detman, T. R., and Kawano, H. (1998). Magnetopause location under extreme solar wind conditions. *Journal of Geophysical Research: Space Physics*, 103(A8):17691–17700.

Sillanpää, I., Ganushkina, N. Y., Dubyagin, S., and Rodriguez, J. V. (2017). Electron Fluxes at Geostationary Orbit From GOES MAGED Data. *Space Weather*, 15(12):1602–1614.

Singer, S. F. (1958). Trapped Albedo Theory of the Radiation Belt. *Physical Review Letters*, 1(8):300–300.

Sivadas, N., Semeter, J., Nishimura, Y., and Mrak, S. (2019). Optical Signatures of the Outer Radiation Belt Boundary. *Geophysical Research Letters*, 46(15):8588–8596.

Smirnov, A. G., Berrendorf, M., Shprits, Y. Y., Kronberg, E. A., Allison, H. J., Aseev, N. A., Zhelavskaya, I. S., Morley, S. K., Reeves, G. D., Carver, M. R., and Effenberger, F. (2020). Medium Energy Electron Flux in Earth's Outer Radiation Belt (MERLIN): A Machine Learning Model. *Space Weather*, 18(11):1–20.

Smith, C., Heureux, J. L., and Ness, N. (1998). The Ace Magnetic Fields Experiment. *Space Science Reviews*, 86(1-4):613–632.

Smith, E. J. and Wolfe, J. H. (1976). Observations of interaction regions and corotating shocks between one and five AU: Pioneers 10 and 11. *Geophysical Research Letters*, 3(3):137–140.

Sokal, R. R. and Michener, C. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38:1409–1438.

Somov, B. V. (2006). The Generalized Ohm's Law in Plasma. In *Plasma Astrophysics. Astrophysics and Space Science Library, vol 341*, pages 193–204. Springer New York.

Sorathia, K. A., Merkin, V. G., Ukhorskiy, A. Y., Mauk, B. H., and Sibeck, D. G. (2017). Energetic particle loss through the magnetopause: A combined global MHD and test-particle study. *Journal of Geophysical Research: Space Physics*, 122(9):9329–9343.

Spivak, D. I. (2012). Metric realization of fuzzy simplicial sets.

Sreeja, V. (2016). Impact and mitigation of space weather effects on GNSS receiver performance. *Geoscience Letters*, 3(1):24.

Stakhiv, M., Landi, E., Lepri, S. T., Oran, R., and Zurbuchen, T. H. (2015). On the origin of mid-latitude fast wind: Challenging the two-state solar wind paradigm. *The Astrophysical Journal*, 801(2):100.

Stansby, D., Rai, Y., Argall, M., JeffreyBroll, Haythornthwaite, R., Erwin, N., Teunissen, J., Shaw, S., Aditya, Saha, R., Ireland, J., Lim, P. L., Badman, S., Mishra, S., Badger, T. G., dupuisIRT, and tlml (2021). heliopython/heliopy: HelioPy 0.15.3.

Staples, F. A., Rae, I. J., Forsyth, C., Smith, A. R. A., Murphy, K. R., Raymer, K. M., Plaschke, F., Case, N. A., Rodger, C. J., Wild, J. A., Milan, S. E., and Imber, S. M. (2020). Do Statistical Models Capture the Dynamics of the Magnetopause During Sudden Magnetospheric Compressions? *Journal of Geophysical Research: Space Physics*, 125(4).

Stawarz, J. E., Eastwood, J. P., Varsani, A., Ergun, R. E., Shay, M. A., Nakamura, R., Phan, T. D., Burch, J. L., Gershman, D. J., Giles, B. L., Goodrich, K. A., Khotyaintsev, Y. V., Lindqvist, P. A., Russell, C. T., Strangeway, R. J., and Torbert, R. B. (2017). Magnetospheric Multiscale analysis of intense field-aligned Poynting flux near the Earth's plasma sheet boundary. *Geophysical Research Letters*, 44(14):7106–7113.

Stone, E. C., Frandsen, A. M., and Mewaldt, R. A. (1998). The advanced composition explorer e. c. stone, a. m. frandsen, r. a. mewaldt. *Sensors (Peterborough, NH)*, pages 1–22.

Su, Z., Xiao, F., Zheng, H., and Wang, S. (2010a). Combined radial diffusion and adiabatic transport of radiation belt electrons with arbitrary pitch angles. *Journal of Geophysical Research: Space Physics*, 115(A10).

Su, Z., Xiao, F., Zheng, H., and Wang, S. (2010b). STEERB: A three-dimensional code for storm-time evolution of electron radiation belt. *Journal of Geophysical Research: Space Physics*, 115(A9).

Su, Z., Xiao, F., Zheng, H., and Wang, S. (2011). Radiation belt electron dynamics driven by adiabatic transport, radial diffusion, and wave-particle interactions. *Journal of Geophysical Research: Space Physics*, 116(A4).

Subbotin, D. A. and Shprits, Y. Y. (2009). Three-dimensional modeling of the radiation belts using the Versatile Electron Radiation Belt (VERB) code. *Space Weather*, 7(10).

Summers, D., Thorne, R. M., and Xiao, F. (1998). Relativistic theory of wave-particle resonant diffusion with application to electron acceleration in the magnetosphere. *Journal of Geophysical Research: Space Physics*, 103(A9):20487–20500.

Takalo, J. (2020). Comparison of Latitude Distribution and Evolution of Even and Odd Sunspot Cycles. *Solar Physics*, 295(3):49.

Tapping, K. F. and Charrois, D. P. (1994). Limits to the accuracy of the 10.7 cm flux. *Solar Physics*, 150(1-2):305–315.

Temmer, M. (2010). Statistical properties of flares and sunspots over the solar cycle. *Arxiv preprint arXiv10020413*, 428:1–8.

The Hammond Indiana Times (1957). Work With New Electronic 'Brains' Opens Field for Army Math Experts.

Thompson, R. L., Morley, S. K., Watt, C. E. J., Bentley, S. N., and Williams, P. D. (2020). Pro□ L * □ A probabilistic L * mapping tool for ground observations. *Space Weather*.

Thorne, R. M. and Kennel, C. F. (1971). Relativistic electron precipitation during magnetic storm main phase. *Journal of Geophysical Research*, 76(19):4446–4453.

Thorne, R. M., Li, W., Ni, B., Ma, Q., Bortnik, J., Chen, L., Baker, D. N., Spence, H. E., Reeves, G. D., Henderson, M. G., Kletzing, C. A., Kurth, W. S., Hospodarsky, G. B., Blake, J. B., Fennell, J. F., Claudepierre, S. G., and Kanekal, S. G. (2013). Rapid local acceleration of relativistic radiation-belt electrons by magnetospheric chorus. *Nature*, 504(7480):411–414.

Tsyganenko, N. (1989). A magnetospheric magnetic field model with a warped tail current sheet. *Planetary and Space Science*, 37(1):5–20.

Tsyganenko, N. A. (1995). Modeling the Earth's magnetospheric magnetic field confined within a realistic magnetopause. *Journal of Geophysical Research*, 100(A4):5599.

Tsyganenko, N. A. (2002). A model of the near magnetosphere with a dawn-dusk asymmetry 2. Parameterization and fitting to observations. *Journal of Geophysical Research: Space Physics*, 107(A8):10–1.

Tsyganenko, N. A. (2005). Modeling the dynamics of the inner magnetosphere during strong geomagnetic storms. *Journal of Geophysical Research*, 110(A3):A03208.

Tsyganenko, N. A., Singer, H. J., and Kasper, J. C. (2003). Storm-time distortion of the inner magnetosphere: How severe can it get? *Journal of Geophysical Research: Space Physics*, 108(A5).

Turner, D. L., Angelopoulos, V., Li, W., Hartinger, M. D., Usanova, M., Mann, I. R., Bortnik, J., and Shprits, Y. (2013). On the storm-time evolution of relativistic electron phase space density in Earth's outer radiation belt. *Journal of Geophysical Research: Space Physics*, 118(5):2196–2212.

Turner, D. L., Angelopoulos, V., Shprits, Y., Kellerman, A., Cruce, P., and Larson, D. (2012a). Radial distributions of equatorial phase space density for outer radiation belt electrons. *Geophysical Research Letters*, 39(9):n/a–n/a.

Turner, D. L., Fennell, J. F., Blake, J. B., Claudepierre, S. G., Clemmons, J. H., Jaynes, A. N., Leonard, T., Baker, D. N., Cohen, I. J., Gkioulidou, M., Ukhorskiy, A. Y., Mauk, B. H., Gabrielse, C., Angelopoulos, V., Strangeway, R. J., Kletzing, C. A., Le Contel, O., Spence, H. E., Torbert, R. B., Burch, J. L., and Reeves, G. D. (2017). Multipoint Observations of Energetic Particle Injections and Substorm Activity During a Conjunction Between Magnetospheric Multiscale (MMS) and Van Allen Probes. *Journal of Geophysical Research: Space Physics*, 122(11):481–11.

Turner, D. L., Kilpua, E. K. J., Hietala, H., Claudepierre, S. G., O'Brien, T. P., Fennell, J. F., Blake, J. B., Jaynes, A. N., Kanekal, S., Baker, D. N., Spence, H. E., Ripoll, J., and Reeves, G. D. (2019). The Response of Earth's Electron Radiation Belts to Geomagnetic Storms: Statistics From the Van Allen Probes Era Including Effects From Different Storm Drivers. *Journal of Geophysical Research: Space Physics*, 124(2):1013–1034.

Turner, D. L., O'Brien, T. P., Fennell, J. F., Claudepierre, S. G., Blake, J. B., Kilpua, E.

K. J., and Hietala, H. (2015). The effects of geomagnetic storms on electrons in Earth's radiation belts. *Geophysical Research Letters*, 42(21):9176–9184.

Turner, D. L., Shprits, Y., Hartinger, M., and Angelopoulos, V. (2012b). Explaining sudden losses of outer radiation belt electrons during geomagnetic storms. *Nature Physics*, 8(3):208–212.

Uddin, S., Khan, A., Hossain, M. E., and Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making*, 19(1):281.

Ukhorskiy, A. Y., Shprits, Y. Y., Anderson, B. J., Takahashi, K., and Thorne, R. M. (2010). Rapid scattering of radiation belt electrons by storm-time EMIC waves. *Geophysical Research Letters*, 37(9):n/a–n/a.

Ukhorskiy, A. Y. and Sitnov, M. I. (2013). Dynamics of Radiation Belt Particles. *Space Science Reviews*, 179(1-4):545–578.

Van Allen, J. A. and Frank, L. A. (1959). Radiation Around the Earth to a Radial Distance of 107,400 km. *Nature*, 183(4659):430–434.

Van Allen, J. A., Ludwig, G. H., Ray, E. C., and McIlwain, C. E. (1958). Observation of High Intensity Radiation by Satellites 1958 Alpha and Gamma. *Journal of Jet Propulsion*, 28(9):588–592.

Van Der Maaten, L. (2015). Accelerating t-SNE using tree-based algorithms. *Journal of Machine Learning Research*, 15:3221–3245.

van der Maaten, L. and Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

Vette, I. (1991). The AE-8 trapped electron model environment. Technical report, NASA, Greenbelt.

von Steiger, R., Schwadron, N. A., Fisk, L. A., Geiss, J., Gloeckler, G., Hefti, S., Wilken, B., Wimmer-Schweingruber, R. R., and Zurbuchen, T. H. (2000). Composition of quasi-stationary solar wind flows from Ulysses/Solar Wind Ion Composition Spectrometer. *Journal of Geophysical Research: Space Physics*, 105(A12):27217–27238.

Wagner, W. (1988). Observations of 1–8 Å solar X-ray variability during solar cycle 21. *Advances in Space Research*, 8(7):67–76.

Walsh, A. P., Haaland, S., Forsyth, C., Keesee, A. M., Kissinger, J., Li, K., Runov, A., Soucek, J., Walsh, B. M., Wing, S., and Taylor, M. G. G. T. (2014). Dawn–dusk asymmetries in the coupled solar wind–magnetosphere–ionosphere system: a review. *Annales Geophysicae*, 32(7):705–737.

Wanliss, J. A. and Showalter, K. M. (2006). High-resolution global storm index: Dst versus SYM-H. *Journal of Geophysical Research: Space Physics*, 111(2):1–10.

Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58(301):236–244.

Watt, C. E. J., Allison, H. J., Thompson, R. L., Bentley, S. N., Meredith, N. P., Glauert, S. A., Horne, R. B., and Rae, I. J. (2021). The Implications of Temporal Variability in Wave￿Particle Interactions in Earth's Radiation Belts. *Geophysical Research Letters*, 48(1):1–10.

Weber, E. J. and Davis, L. (1967). the Average Momentum of the Solar Wind. *ApJ*, 148(April):217–227.

Weimer, D. R. (2003). Predicting interplanetary magnetic field (IMF) propagation delay times using the minimum variance technique. *Journal of Geophysical Research*, 108(A1):1026.

Weimer, D. R. (2004). Correction to "Predicting interplanetary magnetic field (IMF) propagation delay times using the minimum variance technique". *Journal of Geophysical Research*, 109(A12):A12104.

Wenzel, K. P., Marsden, R. G., Page, D. E., and Smith, E. J. (1992). The Ulysses Mission. *Astronomy and Astrophysics Supplement*, 92:207.

Whittaker, I. C., Gamble, R. J., Rodger, C. J., Clilverd, M. A., and Sauvaud, J.-A. (2013). Determining the spectra of radiation belt electron losses: Fitting DEMETER electron flux observations for typical and storm times. *Journal of Geophysical Research: Space Physics*, 118(12):7611–7623.

Wilcox, J. M. and Ness, N. F. (1965). Quasi-stationary corotating structure in the interplanetary medium. *Journal of Geophysical Research*, 70(23):5793–5805.

Wing, S. and Johnson, J. R. (2019). Applications of Information Theory in Solar and Space Physics. *Entropy*, 21(2):140.

Wintoft, P. (2005). Study of the solar wind coupling to the time difference horizontal geomagnetic field. *Annales Geophysicae*, 23(5):1949–1957.

Wolfe, P. (1969). Convergence Conditions for Ascent Methods. *SIAM Review*, 11(2):226–235.

Xiang, Z., Li, X., Kapali, S., Gannon, J., Ni, B., Zhao, H., Zhang, K., and Khoo, L. Y. (2021). Modeling the Dynamics of Radiation Belt Electrons With Source and Loss Driven by the Solar Wind. *Journal of Geophysical Research: Space Physics*, 126(6):1–14.

Xiang, Z., Tu, W., Li, X., Ni, B., Morley, S. K., and Baker, D. N. (2017). Understanding the Mechanisms of Radiation Belt Dropouts Observed by Van Allen Probes. *Journal of Geophysical Research: Space Physics*, 122(10):9858–9879.

Xiang, Z., Tu, W., Ni, B., Henderson, M. G., and Cao, X. (2018). A Statistical Survey of Radiation Belt Dropouts Observed by Van Allen Probes. *Geophysical Research Letters*, 45(16):8035–8043.

Xiong, C. and Lühr, H. (2014). An empirical model of the auroral oval derived from CHAMP field-aligned current signatures - Part 2. *Annales Geophysicae*, 32(6):623–631.

Xiong, C., Lühr, H., Wang, H., and Johnsen, M. G. (2014). Determining the boundaries of the auroral oval from CHAMP field-aligned current signatures - Part 1. *Annales Geophysicae*, 32(6):609–622.

Xu, F. and Borovsky, J. E. (2015). A new four-plasma categorization scheme for the solar wind. *Journal of Geophysical Research: Space Physics*, 120(1):70–100.

Yermolaev, Y. I., Lodkina, I. G., Nikolaeva, N. S., and Yermolaev, M. Y. (2012). Recovery phase of magnetic storms induced by different interplanetary drivers. *Journal of Geophysical Research: Space Physics*, 117(8):1–6.

Zhao, H., Johnston, W. R., Baker, D. N., Li, X., Ni, B., Jaynes, A. N., Kanekal, S. G., Blake, J. B., Claudepierre, S. G., Reeves, G. D., and Boyd, A. J. (2019). Characterization and Evolution of Radiation Belt Electron Energy Spectra Based on the Van Allen Probes Measurements. *Journal of Geophysical Research: Space Physics*, 124(6):4217–4232.

Zhao, L., Landi, E., Lepri, S. T., Gilbert, J. A., Zurbuchen, T. H., Fisk, L. A., and Raines, J. M. (2017). On the Relation between the In Situ Properties and the Coronal Sources of the Solar Wind. *The Astrophysical Journal*, 846(2):135.

Zhao, L., Zurbuchen, T. H., and Fisk, L. A. (2009). Global distribution of the solar wind during solar cycle 23: ACE observations. *Geophysical Research Letters*, 36(14):L14104.

Zurbuchen, T. H. (2007). A new view of the coupling of the sun and the heliosphere. *Annual Review of Astronomy and Astrophysics*, 45:297–338.