Modelling the Exposure of Satellites in Medium Earth Orbit to Proton Belt Radiation

Alexander Richard Lozinski

Clare College

September 2021

This thesis is submitted for the degree of Doctor of Philosophy

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text. It is not substantially the same as any work that has already been submitted before for any degree or other qualification except as declared in the preface and specified in the text. It does not exceed the prescribed word limit for the Mathematics Degree Committee.

Alexander Richard Lozinski
September 2021

Abstract

Modelling the Exposure of Satellites in Medium Earth Orbit to Proton Belt Radiation

Alexander Richard Lozinski

Geomagnetically trapped protons forming Earth's proton radiation belt pose a hazard to orbiting spacecraft. In particular, solar cells are prone to degradation caused by non-ionising collisions with protons in the energy range of several megaelectron volts, which can ultimately shorten the lifespan of a mission. Dynamic enhancements in trapped proton flux following solar energetic particle events have been observed to last several months, and there is a strong need for physics-based modelling in order to predict the impact these changes may have on orbiting spacecraft. This thesis addresses the need for physics-based modelling by presenting an investigation into inner proton belt variability with a 3D numerical model created from scratch, and by quantifying the impact that variability has on the solar cell degradation of orbiting satellites.

After a review of background concepts in Chapter 1, Chapter 2 presents a case study on satellites undergoing electric orbit raising to geostationary orbit. The increase in solar cell degradation that can occur during a period of proton belt enhancement is calculated for three example orbits. It is found that a large enhancement can cause an additional degradation in solar cell output power by up to $\sim$5% over the course of orbit raising, and further changes of a few percent are shown to occur based on the choice of trajectory, or for a 50μm change in solar cell coverglass shielding thickness.

In Chapter 3 a physics-based numerical model is constructed, solving for proton

phase space density in terms of the first, second and third adiabatic invariants μ, K and L. This chapter demonstrates how key processes can be quantified, including transport via radial diffusion, the cosmic ray albedo neutron decay source and coulomb collisional loss. In Chapter 4, a 2D version of the model is applied to derive proton radial diffusion coefficients for a period of solar maximum. This is achieved by varying parameters controlling the rate of radial diffusion in order to optimise the fit between model and data from the CRRES satellite, under the assumption of steady state. Results are compared with diffusion coefficients derived in other literature, and the validity of the steady state assumption underlying this technique is discussed.

In Chapter 5, the 3D numerical model is applied to investigate time variability at energies of 1-10 megaelectron volts, a crucial energy range for solar cell degradation. Three sets of diffusion coefficients from previous literature are applied to model the time evolution of proton phase space density over the four year period 2014-2018. The sensitivity of modelling results to the choice of diffusion coefficients is discussed, including the effect on the anisotropy of proton pitch angle distributions. In the final research chapter of this thesis, Chapter 6, these modelling results are then applied to calculate solar cell degradation over the modelling period for an example satellite in 1200km inclined circular orbit. This demonstrates the final step in an end-to-end physics-based calculation of solar cell degradation.

Acknowledgements

Thank you to my three supervisors Richard, Sarah and Giulio, for your patience and mentorship. You have each inspired me in different ways and set me on a brighter course.

Thank you to my parents and brother, for your constant love, support and encouragement which has always enabled me to chase my dreams.

Contents

Chapter 1

Introduction

1.1 Earth's Proton Radiation Belt

The motion of charged particles in space is constrained in the vicinity of Earth by interactions with the geomagnetic field. The field somewhat resembles a dipole, with an equatorial intensity inversely proportional to distance cubed. The constraints on motion are a result of the Lorentz force, which causes some protons and electrons to undergo three periodic types of motion: gyration centred about geomagnetic field lines; bounce motion up and down field lines; and a longitudinal drift around Earth, with each type of motion occurring on a respectively longer timescale. Example proton and electron trajectories are illustrated in Figure 1.1. The three types of motion lead protons with energies ranging from several hundred kiloelectron volts (keV) up to hundreds of megaelectron volts (MeV) to occupy a toroidal region surrounding Earth called the proton radiation belt, which is the subject of this thesis. Protons are said to be "trapped" in this region, which is characterised by sustained orbits along closed drift shells.

The theory of charged particle motion in a dipole field was developed by Carl Störmer in 1903. Considering the analogy of Earth's geomagnetic field, he demonstrated the existence of a region within the field where particles could be trapped, but showed that access was blocked for a particle approaching from infinity. Decades later, Siegfried Fred Singer hypothesised the existence of charged particles of solar origin trapped within this region in an effort to explain observations of an

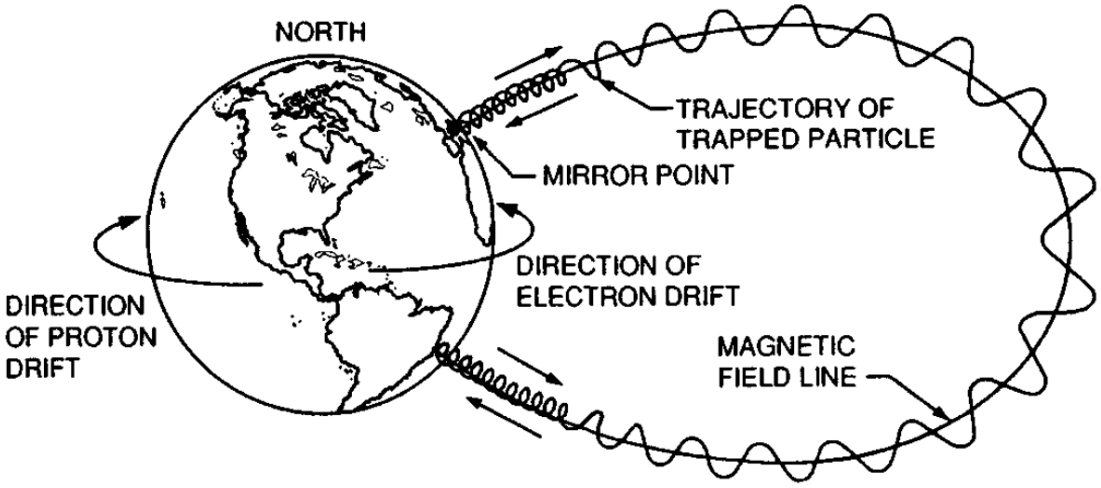


Figure 1.1: Trajectories of trapped protons and electrons under the influence of the geomagnetic field. Taken from Figure 2.7 of Walt (1994)

azimuthal ring current around Earth, thought to be responsible for the main phase of geomagnetic storms. Soon after, the Explorer-1 satellite (launched 31 January 1958) verified the existence of trapped particles using an onboard Geiger counter by recording an increase in counting rate high enough to saturate the instrument. James Van Allen provided the correct interpretation of these results, inferring the existence of Earth's proton and electron radiation belts (Van Allen et al., 1958; Van Allen, 1959).

Advancements in theory and observation have since lead to a more detailed picture. The proton belt has an energy-dependent radial profile that extends up to to $r \sim 3.5R_E$ (distance in Earth radii from the centre of Earth) at tens of MeV (Kovtyukh, 2020). At $r \lesssim 1.5R_E$, the belt tends to be shielded from time-variation of the geomagnetic field and so exhibits variability over long timescales of years to decades (Selesnick et al., 2016), but requires centuries to form (Selesnick et al., 2007). At $r \sim 2R_E$, variation in MeV intensity nominally occurs on timescales of a year or so (Albert and Ginet, 1998), and at $r \gtrsim 2R_E$ the proton belt may exhibit rapid variability due to geomagnetic disturbances. Observations since the early space age have also recorded transient enhancements in intensity at $r > 2R_E$, associated with the spontaneous injection of high energy (tens of MeV) protons originating from the Sun. Such enhancements have been observed to occur more

frequently during periods of maximum sunspot number, and may last several months (Sawyer and Vette, 1976; Lorentzen et al., 2002).

The proton radiation belt poses a hazard to orbiting spacecraft due to the abundance of high energy particles. In particular, solar cells are prone to degradation caused by non-ionising collisions with protons in the energy range of several MeV, which can ultimately shorten the lifespan of a mission. In light of the observed enhancements, predicting variability is therefore of practical importance for spacecraft mission planning and operations. Earth's electron belts also contribute towards this degradation, but for orbits that pass through the proton belt the proton contribution is much greater (Lejosne, 2019a). Currently, radiation belt models based on previously-collected data are used by mission designers to prescribe suitable spacecraft shielding. In the case of the NASA AE-8 (electron belt) and AP-8 (proton belt) models (Vette, 1991), modelled intensity is based on mission-averaged conditions with variation attributed to solar cycle; results do not take into account dynamic changes. The more recent NASA AE-9 and AP-9 models have been developed by incorporating more recent observational data, and allow the user to address the problem of variability by selecting a confidence level that the flux will not exceed a given value (Ginet et al., 2014). For different confidence levels, spatial variations in intensity can be estimated based on templates of previously-observed or modelled states of the radiation belts (Ginet et al., 2014). However, it has also been shown that statistical models such as AP8 and AP9 are liable to under-predict exposure. In the case of the Tacsat-4 satellite, launched in September 2011, solar cell degradation due to energetic protons was 5 - 15% greater than model predictions after two years in orbit (Jenkins et al., 2014).

The hazard to spacecraft, potential shortcomings of statistical models, and need to better understand variability have created a strong impetus for physics-based modelling of the proton belt, which is the subject of this thesis. This work begins with an overview of charged particle motion within the radiation belts, followed by a summary of some key physical processes, then an overview of how to calculate the solar cell degradation of orbiting satellites due to trapped protons. Following chapters describe the construction of a physics-based proton belt model, which is then put to use to investigate variability and make calculations of solar cell degradation for an example satellite in low Earth orbit.

1.1.1 Motion of a Geomagnetically Trapped Particle

In the presence of an electric field $\mathbf{E}$ and magnetic field $\mathbf{B}$, a particle with charge q and velocity $\mathbf{v}$ is subject to the Lorentz force given by

$$\mathbf{F} = \frac{\mathrm{d}\boldsymbol{\rho}}{\mathrm{d}t} = q\mathbf{E} + q\mathbf{v} \times \mathbf{B} \tag{1.1}$$

where $\boldsymbol{\rho}$ is the particle momentum given by $\boldsymbol{\rho} = m\mathbf{v}$. The quantity m is relativistic mass given by $m = \gamma m_0$, where $\gamma = 1/\sqrt{1 - v^2/c^2}$ is the Lorentz factor. Equation 1.1 is the fundamental equation of motion for radiation belt particles, which are subject to forces arising from both static and time-varying magnetic and electric fields. However, the influence of the static geomagnetic field predominates, and leads to "trapping" of radiation belt particles around Earth in the sense that their resultant motion is periodic and unobstructed over timescales much longer than one orbit.

Figure 1.1 shows each of the three types of periodic motion exhibited by electrons and protons in the radiation belts, which are referred to as gyration, bounce and drift. Throughout this chapter, each type of particle motion is derived by considering Equation 1.1 in several simplified cases with idealised static electric and magnetic fields.

1.1.1.1 Gyromotion

In a static magnetic field with direction vector $\hat{\mathbf{b}}$, it is convenient to separate the Lorentz force on a particle into components parallel and perpendicular to the magnetic field. When $\mathbf{E} = 0$, the force on the particle parallel to the magnetic field can be extracted from Equation 1.1 like so:

$$\frac{\mathrm{d}\boldsymbol{\rho}}{\mathrm{d}t}_{\parallel} = \frac{\mathrm{d}\boldsymbol{\rho}}{\mathrm{d}t} \cdot \hat{\mathbf{b}} = \frac{\mathrm{d}\boldsymbol{\rho} \cdot \hat{\mathbf{b}}}{\mathrm{d}t} = \frac{\mathrm{d}\boldsymbol{\rho}_{\parallel}}{\mathrm{d}t} = 0 \tag{1.2}$$

since $\hat{\mathbf{b}}$ is not time varying. Equation 1.2 has the solution $\boldsymbol{\rho}_{\parallel} = m\mathbf{v}_{\parallel} = \text{constant}$, meaning that particle velocity does not change in the direction of the magnetic

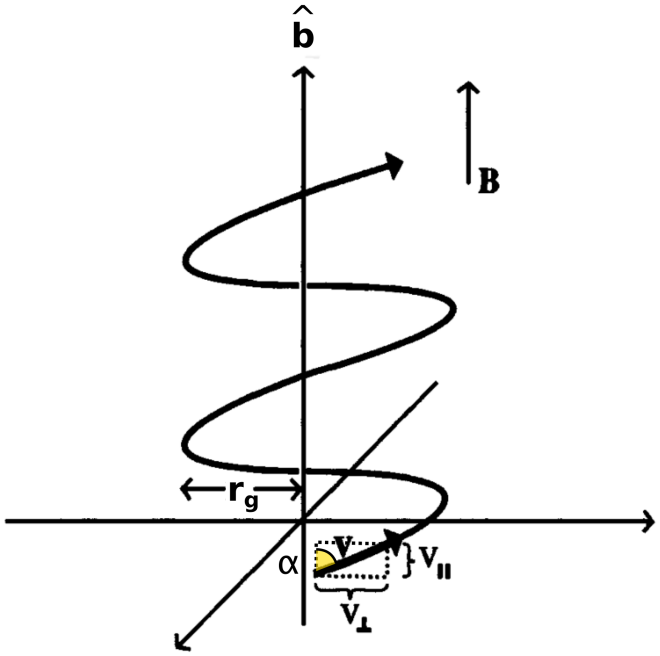


Figure 1.2: Helical trajectory of a negatively charged particle in a region of uniform $\mathbf{B}$ aligned with the vertical axis. Pitch angle α of the particle is highlighted in yellow, with $\tan(\alpha) = |\mathbf{v}_\perp| \,/\, \left|\mathbf{v}_\parallel\right|$. This figure has been adapted and modified from Figure 3.2 of Cravens (1997).

field. Likewise, the perpendicular force is given by

$$\frac{\mathrm{d}\boldsymbol{\rho}}{\mathrm{d}t}_{\perp} = \frac{\mathrm{d}\boldsymbol{\rho}_\perp}{\mathrm{d}t} = \frac{\mathrm{d}\boldsymbol{\rho}}{\mathrm{d}t} - \frac{\mathrm{d}\boldsymbol{\rho}}{\mathrm{d}t}_{\parallel} = q\mathbf{v} \times \mathbf{B} \tag{1.3}$$

Since the Lorentz force is perpendicular to $\mathbf{B}$ and $\mathbf{v}$ when $\mathbf{E} = 0$, no work is done on the particle and velocity remains constant in magnitude. Furthermore, when $\mathbf{B}$ is uniform, the Lorentz force is constant in magnitude and motion of the particle is circular when projected in 2D across a plane perpendicular to $\mathbf{B}$. Because $\mathbf{v}_\parallel$ remains constant, the overall motion is helical if $\mathbf{v}_\parallel$ is initially non-zero. An example trajectory is illustrated in Figure 1.2 for a negatively charged particle, showing helical motion in the presently discussed field configuration.

Figure 1.2 highlights (in yellow) the angle α between $\mathbf{v}_\parallel$ and $\mathbf{v}_\perp$, which remains constant along the trajectory in this case. This angle is called the pitch angle, and its size controls the shape of the helix; motion is entirely circular when $\alpha = 90°$.

Pitch angle is a key parameter characterising the trajectory of charged particles even in more complex field configurations, and is defined like so:

$$\tan(\alpha) = \frac{\left|\mathbf{v}_\perp\right|}{\left|\mathbf{v}_\parallel\right|} = \frac{v_\perp}{v_\parallel} \tag{1.4}$$

Figure 1.2 also labels the radius of the circular motion r_g. This radius, known as the gyroradius, is another key parameter to characterise the motion of individual particles. It can be derived by balancing the magnitude of the centrifugal force by the Lorentz force:

$$\frac{m{v_\perp}^2}{r_g} = q v_\perp B \tag{1.5}$$

$$r_g = \frac{m v_\perp}{Bq} \tag{1.6}$$

where $|\mathbf{B}| = B$. The angular frequency of gyration $\boldsymbol{\Omega_1}$, also called the gyrofrequency or Larmor frequency, can then be found by rewriting Equation 1.3 as

$$\frac{\mathrm{d}\boldsymbol{\rho}_\perp}{\mathrm{d}t} = q\mathbf{v}_\perp \times \mathbf{B} = \boldsymbol{\rho}_\perp \times \boldsymbol{\Omega_1} \tag{1.7}$$

where $\boldsymbol{\Omega_1} = \Omega_1 \hat{\mathbf{b}}$ with units rad s^{-1} and magnitude $\Omega_1 = Bq/m$. The corresponding gyroperiod, or time taken to complete one rotation of 2π radians, is given by

$$T_g = \frac{2\pi}{|\Omega_1|} = 2\pi \frac{m}{|q|\,B} \tag{1.8}$$

For radiation belt particles, r_g is small compared to the scale length of inhomogeneities in the geomagnetic field, such that the magnetic field experienced by a radiation belt particle over a gyration is approximately uniform. Other types of periodic motion arise due to the deviation from helical motion caused by $\nabla\mathbf{B}$ over many gyrations, and therefore occur over far greater distances than r_g and over longer periods than the gyroperiod. In this sense, gyromotion is the most fundamental type of periodic motion exhibited by radiation belt particles.

1.1.1.2 Bounce Motion

The motion of a charged particle is now considered in the case of a non-uniform static magnetic field where $\mathbf{E} = 0$. In any non-uniform field, the Lorentz force leads to more complicated trajectories, as the particle is impelled by any gradients $\nabla\mathbf{B}$ such that motion is not strictly helical or circular. Therefore, concepts invoked in Section 1.1.1.1, such as a gyroperiod, are approximations, but still useful for characterising the gyrational component of motion. To derive expressions for the forces that cause other types of periodic motion, it is convenient to invoke the additional approximation of a "guiding centre". The particle position can then be written as $\mathbf{r} = \mathbf{R} + \mathbf{r_g}$, where $\mathbf{R}$ is the location of the guiding centre and $\mathbf{r_g}$ is the instantaneous gyroradius.

The second type of periodic motion arises due to convergence of magnetic field lines, resulting in $\nabla\mathbf{B}$ along a field line. An example field line configuration is shown in Figure 1.3 using cylindrical coordinates, where $\mathbf{B}$ is axisymmetric around a central axis $\hat{\mathbf{z}}$ (Chen, 1984). In Figure 1.3, convergence results in a radial magnetic field component B_r, with $B_\theta = 0$ due to axisymmetry. Using $\boldsymbol{\nabla} \cdot \mathbf{B} = 0$ (Gauss' law for magnetism), one can therefore write

$$\begin{aligned} \boldsymbol{\nabla} \cdot \mathbf{B} &= \frac{1}{r}\frac{\partial}{\partial r}(rB_r) + \frac{1}{r}\frac{\partial B_\theta}{\partial \theta} + \frac{\partial B_z}{\partial z} \\ &= \frac{1}{r}\frac{\partial}{\partial r}(rB_r) + \frac{\partial B_z}{\partial z} = 0 \end{aligned} \tag{1.9}$$

Solving for B_r using the assumption that $\frac{\partial B_z}{\partial z}$ is constant and known at $r = 0$ (Chen, 1984), Equation 1.9 leads to:

$$\begin{aligned} rB_r &= -\int_0^r r\frac{\partial B_z}{\partial z}\mathrm{d}r \approx -\frac{1}{2}r^2\left[\frac{\partial B_z}{\partial z}\right]_{r=0} \\ B_r &= -\frac{1}{2}r\left[\frac{\partial B_z}{\partial z}\right]_{r=0} \end{aligned} \tag{1.10}$$

Evaluating the component of the Lorentz force from Equation 1.1 in the $\hat{\mathbf{z}}$ direction (anywhere in this system) gives

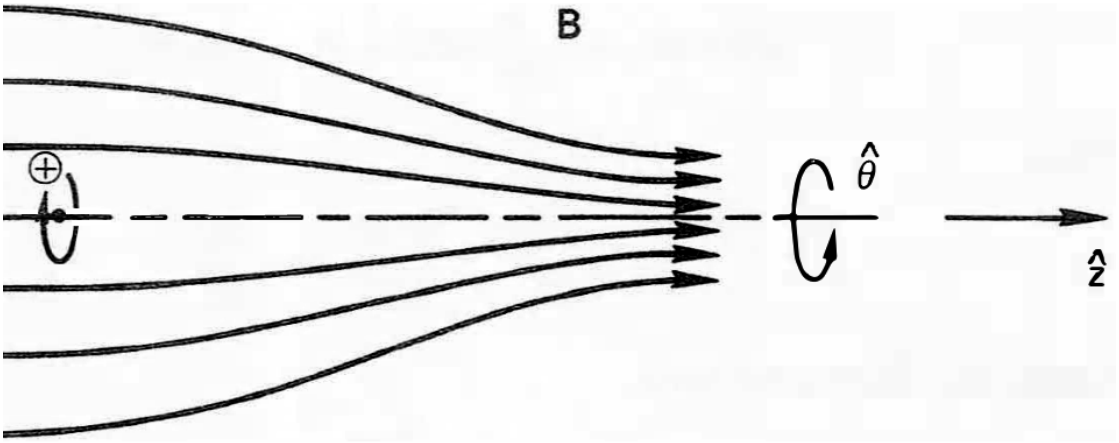


Figure 1.3: Orientation of field lines in a magnetic field that causes mirroring. Gyration of a positively charged particle is shown on the left side. Orientations of the cylindrical coordinate system axis vectors $\hat{\boldsymbol{\theta}}$ and $\hat{\mathbf{z}}$ are shown on the right side. This figure has been adapted and modified from Figure 2.7 of Chen (1984).

$$\begin{aligned} \frac{\mathrm{d}\boldsymbol{\rho}}{\mathrm{d}t}_z &= q(v_r B_\theta - v_\theta B_r) = -q v_\theta B_r \\ &= \frac{1}{2} q v_\theta r \frac{\partial B_z}{\partial z} \end{aligned} \tag{1.11}$$

Figure 1.3 shows an example charged particle with a guiding centre $\mathbf{R}$ that lies on the $\hat{\mathbf{z}}$ axis. For positive charges, gyration occurs such that $v_\theta = -v_\perp$ in the cylindrical coordinate system (the $\hat{\boldsymbol{\theta}}$ direction is anticlockwise around $\hat{\mathbf{z}}$). Therefore, Equation 1.11 can be re-written in terms of the quantities r_g and $v_\perp$, to consider the average force on the guiding centre over one gyration (also making use of Equation 1.8):

$$\begin{aligned} \frac{\mathrm{d}\boldsymbol{\rho}}{\mathrm{d}t}_z &= -\frac{1}{2} |q| \, v_\perp r_g \frac{\partial B_z}{\partial z} \\ &= -\frac{1}{2} \frac{m {v_\perp}^2}{B} \frac{\partial B_z}{\partial z} \end{aligned} \tag{1.12}$$

Equation 1.12 is for a force which opposes the motion of a particle into a region

where field lines converge. This force acts on radiation belt particles which travel along converging geomagnetic field lines into Earth's polar regions. Eventually the particle may be decelerated in the parallel direction to the point at which $v_{\parallel} = 0$, which leads in the next instant to a reversal of the guiding centre direction along the field line, and the particle is said to have "mirrored". On approach to the mirror point, the local pitch angle of the particle increases to 90° as the ratio $v_{\perp}/v_{\parallel}$ increases so as to conserve kinetic energy, then after mirroring pitch angle decreases. This phenomenon leads to periodic reversal of the guiding centre direction as particles bounce from pole to pole, and pitch angle varies from a minimum on the magnetic equator up to a maximum of 90°.

The local pitch angle of a radiation belt particle at the magnetic equator (equatorial pitch angle, α_{eq}) is a significant quantity since it controls the latitude at which mirroring takes place. For example, a particle with $\alpha_{\mathrm{eq}} = 90°$ will be contained at the magnetic equator by the mirror force. On the other hand, a particle with $\alpha_{\mathrm{eq}} \sim 0°$ will travel with a total velocity closely aligned with the magnetic field, and the mirror point may be below the height of the atmosphere such that the particle is lost via atmospheric collision before it can bounce back. In this case, the particle is said to be in the loss cone, which is defined by a critical equatorial pitch angle α_{lc}. A particle with this pitch angle will mirror just above the top of the atmosphere such that it remains trapped in the radiation belts, whilst particles with $\alpha_{\mathrm{eq}} < \alpha_{\mathrm{lc}}$ will be lost to the atmosphere.

1.1.1.3 Drift Motion

The third type of periodic motion exhibited by radiation belt particles involves drift of the guiding centre across the magnetic field due to a force $\mathbf{F}_{\perp}$ applied to the particle in the plane of gyromotion. To facilitate a general description, the Lorentz force perpendicular to the magnetic field can be written

$$\frac{\mathrm{d}\boldsymbol{\rho}_{\perp}}{\mathrm{d}t} = \boldsymbol{\rho}_{\perp} \times \boldsymbol{\Omega}_{\mathbf{1}} + \mathbf{F}_{\perp} \tag{1.13}$$

By denoting particle position $\mathbf{r} = \mathbf{R} + \mathbf{r_g}$ and velocity $\mathbf{v} = \mathbf{V} + \mathbf{v}_{\perp}$, velocity of the guiding centre $\mathbf{V}$ can be solved for. Two assumptions are made before proceeding which can later be verified: firstly, that $\mathbf{V}$ is perpendicular to the magnetic field;

and secondly, that $\mathbf{V}$ is constant in time. Substituting the overall velocity $\mathbf{v}$ into Equation 1.13 gives

$$\frac{\mathrm{d}\boldsymbol{\rho}_\perp}{\mathrm{d}t} = m\left[\frac{\mathrm{d}\mathbf{V}}{\mathrm{d}t} + \frac{\mathrm{d}\mathbf{v}_\perp}{\mathrm{d}t}\right] = m\left[\mathbf{V} + \mathbf{v}_\perp\right] \times \boldsymbol{\Omega_1} + \mathbf{F}_\perp \tag{1.14}$$

Equation 1.14 can be simplified using $\mathrm{d}\mathbf{V}/\mathrm{d}t = 0$, leaving:

$$\begin{aligned} m\left[\frac{\mathrm{d}\mathbf{v}_\perp}{\mathrm{d}t}\right] &= m\left[\mathbf{V} + \mathbf{v}_\perp\right] \times \boldsymbol{\Omega_1} + \mathbf{F}_\perp \\ &= m\left[\mathbf{V} \times \boldsymbol{\Omega_1}\right] + m\left[\mathbf{v}_\perp \times \boldsymbol{\Omega_1}\right] + \mathbf{F}_\perp \end{aligned} \tag{1.15}$$

From Equation 1.7, $\mathrm{d}\mathbf{v}_\perp/\mathrm{d}t = \mathbf{v}_\perp \times \boldsymbol{\Omega_1}$. Cancelling terms in Equation 1.15 leads to

$$0 = m\left[\mathbf{V} \times \boldsymbol{\Omega_1}\right] + \mathbf{F}_\perp \tag{1.16}$$

Velocity $\mathbf{V}$ can then be isolated by taking the cross product of both sides with $\boldsymbol{\Omega_1}$ like so:

$$\begin{aligned} 0 &= m\left[\mathbf{V} \times \boldsymbol{\Omega_1}\right] \times \boldsymbol{\Omega_1} + \mathbf{F}_\perp \times \boldsymbol{\Omega_1} \\ &= -m\left[\boldsymbol{\Omega_1} \cdot \boldsymbol{\Omega_1}\right]\mathbf{V} + \left[\mathbf{V} \cdot \boldsymbol{\Omega_1}\right]\boldsymbol{\Omega_1} + \mathbf{F}_\perp \times \boldsymbol{\Omega_1} \\ &= -m\Omega_1^2\mathbf{V} + \mathbf{F}_\perp \times \boldsymbol{\Omega_1} \end{aligned} \tag{1.17}$$

Drift velocity of the guiding centre due to a general force $\mathbf{F}_\perp$ is finally given by

$$\mathbf{V} = \frac{\mathbf{F}_\perp \times \boldsymbol{\Omega_1}}{m\Omega_1^2} = \frac{\mathbf{F}_\perp \times \mathbf{B}}{qB^2} \tag{1.18}$$

Equation 1.18 shows that a constant force applied to the particle over a gyration results in a guiding centre drift perpendicular to the force, with a direction that depends on charge. For radiation belt particles, a force $\mathbf{F}_\perp$ strong enough to result in drift arises from $\nabla\mathbf{B}$ as the magnetic field strength increases closer to Earth, as well as from the centrifugal force exerted by the guiding centre as it travels up and down curved field lines. The resulting components of velocity $\mathbf{V_G}$ and $\mathbf{V_C}$ are called gradient and curvature drifts respectively, and cause the drift of radiation

belt protons from East to West, whilst electrons drift from West to East. This gives rise to a westward azimuthal current around Earth at $\sim$3 to $5R_E$ called the ring current (Walt, 1994). Electrostatic fields also cause drift motion in the case that $\mathbf{F}_\perp = \mathbf{E}_\perp q$, but the q term cancels, leading to drift in the same direction.

Drift due to electrostatic fields drives convection of magnetospheric plasma at energies $\lesssim$ 1keV because the gradient and curvature drift becomes negligible at this energy (Schulz and Lanzerotti, 1974). On the contrary, the electrostatic field exerts little influence on particles with energies $\gtrsim$ 200keV, leading to a drift path that follows a contour of approximately constant magnetic field intensity on the magnetic equator. By this criteria, particles are broadly characterised as belonging to either the ring current or radiation belts depending on their energy being below or above $\sim$ 200keV.

Gradient Drift

Force on the guiding centre due to a gradient in the magnetic field is derived with respect to a Cartesian frame using the example field configuration shown in Figure 1.4 from Walt (1994), where $\mathbf{B} = B_z(y)\hat{\boldsymbol{z}}$ and $\mathbf{E} = 0$. The magnetic field at the particle position $\mathbf{r} = \mathbf{R} + \mathbf{r_g}$ can be approximated by considering a Taylor series expansion about the point $\mathbf{R}$. When $\mathbf{R}$ coincides with the origin, the Lorentz force on the particle in the x-y plane can be written

$$\begin{aligned} F_x(y) &= qv_y B_z(y) = qv_y \left[B_{z0} + y \frac{\mathrm{d}B_z}{\mathrm{d}y}_{\,0} + ... \right] \\ F_y(y) &= -qv_x B_z(y) = -qv_x \left[B_{z0} + y \frac{\mathrm{d}B_z}{\mathrm{d}y}_{\,0} + ... \right] \end{aligned} \tag{1.19}$$

When $\mathrm{d}B_z/\mathrm{d}y$ is small, $\mathbf{F}_\perp \approx 0$ in Equation 1.13 and the position of the particle (x, y) at the moment $\mathbf{R}$ coincides with the origin can be described by purely helical motion. The solution to the Lorentz force in this case (from Cravens (1997)) is:

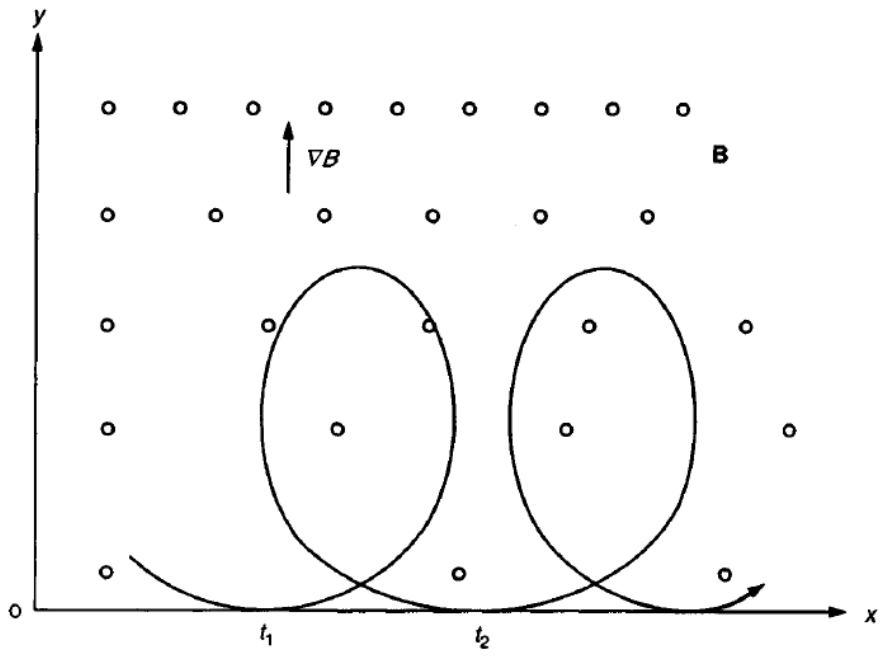


Figure 1.4: Trajectory of a negatively charged particle due to the Lorentz force in a field where $\mathbf{B} = B_z(y)\hat{\boldsymbol{z}}$ such that $\nabla\mathbf{B}$ points in the positive $\hat{\boldsymbol{y}}$ direction. This figure has been taken from Figure 2.4 of Walt (1994).

$$
\begin{aligned}
x &= r_g \sin(\Omega_1 t + \delta) \\
y &= \pm r_g \cos(\Omega_1 t + \delta) \\
v_x &= v_\perp \cos(\Omega_1 t + \delta) \\
v_y &= \mp v_\perp \sin(\Omega_1 t + \delta)
\end{aligned}
\tag{1.20}
$$

where δ is initial phase of the gyration orbit and the plus/minus symbol indicates dependence on particle charge. Substituting this into Equation 1.19 gives

$$
\begin{aligned}
F_x &= \mp q v_\perp \sin(\Omega_1 t + \delta)(B_{z0} \pm r_g \cos(\Omega_1 t + \delta) \frac{\mathrm{d}B_z}{\mathrm{d}y}_{\,0} + ...) \\
F_y &= -q v_\perp \cos(\Omega_1 t + \delta)(B_{z0} \pm r_g \cos(\Omega_1 t + \delta) \frac{\mathrm{d}B_z}{\mathrm{d}y}_{\,0} + ...)
\end{aligned}
\tag{1.21}
$$

The average of forces F_x and F_y over one gyroperiod are therefore given by

$$
\begin{aligned}
\langle F_x \rangle &\approx \mp q v_\perp \left[B_0 \langle \sin(\Omega_1 t) \rangle \pm r_g \langle \sin(\Omega_1 t) \cos(\Omega_1 t) \rangle \frac{\mathrm{d}B_z}{\mathrm{d}y} \right] \\
\langle F_y \rangle &\approx -q v_\perp \left[B_0 \langle \cos(\Omega_1 t) \rangle \pm r_g \left\langle \cos^2(\Omega_1 t) \right\rangle \frac{\mathrm{d}B_z}{\mathrm{d}y} \right]
\end{aligned}
\tag{1.22}
$$

where dependence on δ has been eliminated due to averaging. Furthermore, the following equalities can be made use of to simply Equation 1.22: $\langle \sin(\Omega_1 t) \rangle = 0$; $\langle \sin(\Omega_1 t) \cos(\Omega_1 t) \rangle = 0$; $\langle \cos(\Omega_1 t) \rangle = 0$; and $\langle \cos^2(\Omega_1 t) \rangle = 1/2$. This yields

$$
\begin{aligned}
\langle F_x \rangle &= 0 \\
\langle F_y \rangle &\approx -\frac{q v_\perp r_g}{2} \frac{\mathrm{d}B_z}{\mathrm{d}y}
\end{aligned}
\tag{1.23}
$$

indicating a force on the particle in the direction of $-\nabla \mathbf{B}$. Over half a gyration the particle is therefore accelerated, before being decelerated over the returning half. When the particle is accelerated, the gyroradius increases as per Equation 1.6. The resulting motion is illustrated in Figure 1.4 for the trajectory of a negatively

charged particle. Substituting $\langle F_y \rangle$ into Equation 1.18 to find the drift velocity of the guiding centre gives

$$\begin{aligned} \mathbf{V_G} &= -\frac{q v_\perp r_g}{2} \frac{\frac{\mathrm{d}B_z}{\mathrm{d}y}\hat{\mathbf{y}} \times \mathbf{B}}{qB^2} \\ &= \frac{m{v_\perp}^2}{2qB^3}\mathbf{B} \times \frac{\mathrm{d}B_z}{\mathrm{d}y}\hat{\mathbf{y}} \\ &= \frac{m{v_\perp}^2}{2qB^3}\mathbf{B} \times \nabla\mathbf{B} \end{aligned} \tag{1.24}$$

Curvature Drift

Particles with $\alpha < 90°$ have a component of velocity along the magnetic field line, leading to traversal of the guiding centre approximately along a field line. Geomagnetic field lines curve towards the polar regions of Earth, and therefore exert a centrifugal force on the guiding centre of such particles. This force can be written in terms of the local field line radius of curvature, R_c, as

$$\mathbf{F_c} = \frac{m{v_\parallel}^2}{R_c}\hat{\mathbf{n}} \tag{1.25}$$

where $\hat{\mathbf{n}}$ represents the unit vector in the direction of the field line radius of curvature. Substituting this into Equation 1.18 to find the drift velocity of the guiding centre gives

$$\mathbf{V_C} = -\frac{m{v_\parallel}^2}{qR_cB^2}\hat{\mathbf{n}} \times \mathbf{B} \tag{1.26}$$

The $\hat{\mathbf{n}} \times \mathbf{B}$ term in Equation 1.26 and $\mathbf{B} \times \nabla\mathbf{B}$ term in Equation 1.24 point in opposite directions since $\hat{\mathbf{n}}$ points outwards, thus gradient and curvature drift cause drift in the same direction due to the minus sign in Equation 1.26. However, the relative importance of each term varies with pitch angle, with curvature drift being more important than gradient drift for particles with high $v_\parallel$, and vice versa for particles with high $v_\perp$.

1.1.2 Adiabatic Invariants of Motion

Radiation belt particles are distinguished from surrounding populations by their three types of periodic motion. When this type of motion is sustained for many drift periods, a radiation belt particle is said to be trapped. The adiabatic theory of charged particle motion ascribes a quantity to each of the three periodic motions of a given particle. These three quantities characterise each orbit independently of one another and in doing so provide a convenient way to organise the population of trapped particles forming Earth's radiation belts, and a way to distinguish from particles outside the population such as cosmic rays or ring current particles. Each quantity is proportional to a canonical action integral J_i from Hamiltonian-Jacobi theory:

$$J_i = \oint_i [\boldsymbol{\rho} + q\mathbf{A}] \cdot \mathrm{d}\mathbf{l} \tag{1.27}$$

The subscript i represents the period motion: i =1, 2 and 3 for gyration, bounce and drift respectively, and $\mathbf{l}$ is along the corresponding orbit. The integrand of Equation 1.27 is the canonical momentum, with $\mathbf{A}$ the vector potential of the magnetic field. It can be shown that J_i is a conserved quantity when forces altering the particle trajectory act over a timescale much longer than that of the period of associated motion (see for example Walt, 1994), thus any proportional quantity is also conserved and may be considered an "adiabatic invariant".

1.1.2.1 The First Invariant

Using Equation 1.27 with i =1 for the quantity conserved over a gyration orbit leads to:

$$
\begin{aligned}
J_1 &= \oint_1 \boldsymbol{\rho} \cdot \mathrm{d}\mathbf{l} + q \oint_1 \mathbf{A} \cdot \mathrm{d}\mathbf{l} \\
&= \rho_\perp 2\pi r_g + q \oint_1 \nabla \times \mathbf{A} \cdot \mathrm{d}\mathbf{S} \\
&= \rho_\perp 2\pi r_g + q \oint_1 \mathbf{B} \cdot \mathrm{d}\mathbf{S} \\
&= \rho_\perp 2\pi r_g - q\pi {r_g}^2 B \\
&= \frac{{\rho_\perp}^2 \pi}{Bq}
\end{aligned}
\tag{1.28}
$$

where the element $\mathrm{d}\mathbf{l}$ is along the path taken by the particle over a gyration. Conservation of J_1 therefore implies that ${\rho_\perp}^2/B$ is conserved. However, ${\rho_\perp}^2 = \rho^2 \sin^2(\alpha)$, and momentum ρ is conserved, meaning that $\sin^2(\alpha)/B$ is also conserved. As $\alpha = 90^\circ$ at the magnetic mirror points where $B = B_m$, this implies that B_m is invariant along a particle drift.

The quantity $\sin^2(\alpha)/B$ can be evaluated at different points along a magnetic field line traversed by the guiding centre to relate magnetic field strength and local pitch angle. For example, equating ${\rho_\perp}^2/B$ at the mirror point, where $B = B_m$, and a point somewhere else along the field line, gives

$$
\begin{aligned}
\frac{\sin^2(90^\circ)}{B_m} &= \frac{\sin^2(\alpha)}{B} \\
\therefore B &= B_m \sin^2(\alpha)
\end{aligned}
\tag{1.29}
$$

Alternatively, when the magnetic field can be calculated using a model, conservation of ${\rho_\perp}^2/B$ is useful for mapping the local pitch angle of an observed particle to its equatorial pitch angle.

The first invariant is usually taken as the quantity proportional to J_1 given by

$$
\mu = \frac{{\rho_\perp}^2}{2m_0 B} = \frac{T(T + 2E_0)}{2E_0 B_m} \tag{1.30}
$$

where T is kinetic energy and E_0 is rest energy, so that total energy $E_{tot} = T + E_0$. The second equality in Equation 1.30 comes from the relationship between total energy and momentum ${E_{tot}}^2 = \rho^2 c^2 + {m_0}^2 c^4$, and using $B = B_m \sin^2(\alpha)$. Hereon,

the symbol E is also used for kinetic energy.

1.1.2.2 The Second Invariant

Using Equation 1.27 with i =2 for the quantity conserved over a bounce orbit leads to:

$$\begin{aligned} J_2 &= \oint_2 \boldsymbol{\rho} \cdot \mathrm{d}\mathbf{l} + q \oint_2 \mathbf{A} \cdot \mathrm{d}\mathbf{l} \\ &= \oint_2 \rho_{\parallel} \mathrm{d}l + q \oint_2 \nabla \times \mathbf{A} \cdot \mathrm{d}\mathbf{S} \\ &= \oint_2 \rho \cos(\alpha) \mathrm{d}l + q \oint_2 \mathbf{B} \cdot \mathrm{d}\mathbf{S} \\ &= \rho \oint_2 \cos(\alpha) \mathrm{d}l \end{aligned} \tag{1.31}$$

where $\oint_2 \mathbf{B} \cdot \mathrm{d}\mathbf{S}$ goes to zero because the bounce path traces a small area and is assumed to be parallel to the magnetic field, thus enclosing no magnetic flux. The element d$\mathbf{l}$ is along the bounce path taken by the guiding centre along a field line. The second invariant is $J = J_2$, and a proportional quantity I is also used that is independent of particle momentum like so:

$$\begin{aligned} I &= J/2\rho \\ &= \frac{1}{2} \oint_2 \cos(\alpha)\, dl \\ &= \int_{\mathrm{m}}^{\mathrm{m'}} \cos(\alpha)\, dl \end{aligned} \tag{1.32}$$

where m and m' are the conjugate mirror points in opposite hemispheres where $B = B_m$. The time period for one bounce can be written as

$$\frac{2\pi}{\Omega_2} = \oint_2 \frac{\mathrm{d}l}{v_{\parallel}} = \frac{m}{\rho} \oint_2 \frac{\mathrm{d}l}{\cos(\alpha)} \tag{1.33}$$

Equations 1.31, 1.32 and 1.33 contain an integral that cannot be solved analytically. However, the cosine term can be replaced using the relationship $B = B_m \sin^2(\alpha)$ which leads to:

$$\cos(\alpha) = \left[1 - \frac{B(l)}{B_m}\right]^{1/2} \tag{1.34}$$

By choosing a point in space to be the mirror point at which $\alpha = 90°$, one can evaluate I for any radiation belt particle by numerically solving particle or guiding centre motion until the conjugate mirror point. Invariance of I implies that the path taken by the guiding centre over one bounce does not depend on particle energy. The invariance of I also implies that even when the magnetic field is not axisymmetric, the particle must return to the same bounce path at a given longitude after one drift orbit. Therefore, when the assumptions of adiabatic theory hold, the whole drift path is independent of energy, and may be fully defined by a value of B_m and I. These quantities may be derived using only a point in space, the local pitch angle of the particle, and the magnetic field itself.

As an alternative to numerical integration, Equations 1.31, 1.32 and 1.33 have been approximated in previous literature as simple functions of $y \equiv \sin(\alpha_{\mathrm{eq}})$. To derive these expressions, it is first necessary to assume a dipolar magnetic field centred and axisymmetric about Earth's rotational axis, defined as

$$\mathbf{B} = -\frac{B_0 a^3}{r^3}(2\cos(\theta)\hat{\mathbf{r}} + \sin(\theta)\hat{\boldsymbol{\theta}}) \tag{1.35}$$

where a is the radius of Earth and B_0 is the field strength measured at colatitude $\theta = \pi/2$ and radial distance $r = a$ (Walt, 1994, p. 30). By this definition, the field points downwards towards magnetic South (geographic North) at $\theta = \pi/2$. The differential equation for a field line is $\mathrm{d}r/\mathrm{d}\theta = rB_r/B_\theta = 2r\cos(\theta)/\sin(\theta)$, which can take the substitution $r = La$ to describe a field line that crosses the magnetic equator at L Earth radii (Schulz and Lanzerotti, 1974, Section I.4). The quantity L is known as the McIlwain L parameter (McIlwain, 1961), and is useful because it can describe the set of field lines along which a radiation belt particle executes a drift orbit when the magnetic field is dipolar. Integrating with respect to θ leads to the field line equation

$$r = La\sin^2(\theta) \tag{1.36}$$

and substituting this into Equation 1.35 thus leads to the following expression for

dipole magnetic field strength:

$$B = \frac{B_0}{L^3 \sin^6(\theta)}(1 + 3\cos^2(\theta))^{1/2} \tag{1.37}$$

The element of length along the field line $\mathrm{d}l = [(\mathrm{d}r)^2 + (r\mathrm{d}\theta)^2]^{1/2}$ can also be re-written using Equation 1.36 to give

$$\mathrm{d}l = La(1 + 3\cos^3(\theta))^{1/2} sin(\theta)\mathrm{d}\theta \tag{1.38}$$

Next, Equation 1.29 is used to derive $B_m = B_e/\sin^2(\alpha_{eq}) = B_e/y^2$, where B_e is magnetic field strength at the Equator. However, Equation 1.37 shows that $B_e = B_0/L^3$, and therefore $B_m = B_0 L^{-3} y^{-2}$ for a dipole. Finally, substituting this expression, along with Equations 1.34, 1.37 and 1.38 above into Equation 1.33 leads to the following expression for the bounce time:

$$\frac{2\pi}{\Omega_2} = \frac{4mLa}{\rho} T(y) \tag{1.39}$$

where

$$T(y) = \int_{\theta_m}^{\pi/2} \frac{[1 + 3\cos^2(\theta)]^{1/2} \sin(\theta)\mathrm{d}\theta}{[1 - y^2 \sin^{-6}(\theta)[1 + 3\cos^2(\theta)]^{1/2}]^{1/2}} \tag{1.40}$$

Davidson (1976) shows that $T(y)$ can be approximated within 0.57% error as

$$T(y) \approx 1.380173 - 0.639693 y^{3/4} \tag{1.41}$$

Similarly, using $B_m = B_0 L^{-3} y^{-2}$ and substituting Equations 1.34, 1.37 and 1.38 into Equation 1.31 leads to the following expression for J:

$$J = 2\rho LaY(y) \tag{1.42}$$

where

$$Y(y) = 2\int_{\theta_m}^{\pi/2} \left[1 + 3\cos^2(\theta) - y^2(1 + 3\cos^2(\theta)^{3/2} \sin^{-6}(\theta)\right]^{1/2} \sin(\theta)\mathrm{d}\theta \tag{1.43}$$

From the definitions of $T(y)$ and $Y(y)$ in Equations 1.40 and 1.43, it follows that

$$\frac{\mathrm{d}}{\mathrm{d}y}\left(\frac{Y}{y}\right) = -\frac{2T(y)}{y^2} \tag{1.44}$$

and therefore the approximation in Equation 1.41 can also be used to solve for an approximate $Y(y)$, leading to the following:

$$Y(y) \approx 2.760346 + 2.357194y - 5.117540y^{3/4} \tag{1.45}$$

which approximates $Y(y)$ within 0.51 % error (Davidson, 1976).

Accurate approximations of the bounce period and second invariant in a dipole magnetic field can therefore be calculated quickly using Equations 1.39 and 1.42 above.

Lastly, the second invariant can also be taken as K, given by

$$K = \sqrt{B_m} I \tag{1.46}$$

This quantity is useful for modelling purposes, due to the property that as K increases at a constant rate (or by a fixed interval), the increase in α_{eq} becomes less and less until the loss cone is reached.

1.1.2.3 The Third Invariant

Using Equation 1.27 with i =3 for the quantity conserved over a drift orbit leads to:

$$\begin{aligned} J_3 &= \oint_3 \boldsymbol{\rho} \cdot \mathrm{d}\mathbf{l} + q \oint_3 \mathbf{A} \cdot \mathrm{d}\mathbf{l} \\ &= q \oint_3 \nabla \times \mathbf{A} \cdot \mathrm{d}\mathbf{S} \\ &= q \oint_3 \mathbf{B} \cdot \mathrm{d}\mathbf{S} \\ &= q\Phi \end{aligned} \tag{1.47}$$

where $\boldsymbol{\rho} \cdot \mathrm{d}\mathbf{l}$ goes to zero because the velocity of the particle is small in the direction of drift, and Φ is the magnetic flux enclosed by the drift path. The

element d$\mathbf{l}$ is along the path taken by the guiding centre during one drift around Earth. The magnetic flux Φ can be calculated by considering the net flux outside the drift path, which is equal to the net flux enclosed by the drift path. For a dipole field specified by Equation 1.35, considering the drift path of a particle at radial distance r_0 in the magnetic equatorial plane thus leads to:

$$\begin{aligned}\Phi &= \int_{r_0}^{\infty} \frac{B_0 a^3}{r^3} 2\pi r \ \mathrm{d}r \\ &= 2\pi B_0 \frac{a^3}{r_0}\end{aligned} \tag{1.48}$$

Compression of the geomagnetic field increases the net magnetic flux enclosed within a given radius and therefore, to keep Φ constant, conservation of the third invariant requires that a drift path will shrink towards Earth. On the other hand, expansion of the geomagnetic field requires a drift path to expand away from Earth. Therefore, slow compressions/expansions of the field, such as that caused by secular variation in Earth's dynamo, cause the inward/outward motion of drift orbits.

Equation 1.48 can be rearranged for $L = r_0/a$, the McIlwain L parameter, which identifies the drift shell on which the particle at r_0 lies. Φ is the same for any particle on the drift shell L because the path element d$\mathbf{l}$ in Equation 1.47 lies within the magnetic shell surface (and therefore the surface S outlined by l is bounded by the same field lines regardless of latitude). Therefore the L parameter is an adiabatic invariant for a particle in a static dipole field, assignable to any particle on the drift shell and given by:

$$L = \frac{2\pi a^2 B_0}{|\Phi|} \tag{1.49}$$

In the geomagnetic field, the McIlwain L parameter in Equation 1.49 is replaced by Roederer's L^* parameter. The L^* parameter is an adiabatic invariant in the geomagnetic field, with a value equal to L of an adiabatically equivalent particle in a centred dipole field. In this way, it provides a more intuitive version of the third invariant by relating to a drift shell around a centred dipole, but in the geomagnetic field it is a property of a particle, not a point in space. Computation of L^* requires a method such as that suggested by Roederer and Lejosne (2018): i) solving a

complete drift orbit numerically; ii) identifying field lines coinciding with the drift shell and following them onto a reference surface; and iii) calculating Φ through the surface.

A formula for the drift frequency is given by Equation 1.35 of Schulz and Lanzerotti (1974):

$$\frac{\Omega_3}{2\pi} = -\frac{3L}{2\pi\gamma}(\gamma^2 - 1)\left(\frac{c}{a}\right)^2\left(\frac{m_0 c}{qB_0}\right)\left[\frac{6T(y) - Y(y)}{12T(y)}\right] \tag{1.50}$$

The quantities $T(y)$ and $Y(y)$ are given by the simple approximations of Equations 1.41 and 1.45 respectively (see Section 1.1.2.2).

1.2 The Magnetospheric Environment

1.2.1 Charged Particle Populations

The region around Earth in which magnetic topology is controlled by the geomagnetic field is called the magnetosphere. In general, charged particles within the magnetosphere are classified into different populations based on the processes that govern their motion, since they are generally the focus of different fields of study. For example, radiation belt particles have energies high enough to exhibit individualistic motion rather than the bulk convection of a plasma, but energy low enough so as to obey the laws of adiabatic charged particle motion. Figure 1.5 from Schulz and Lanzerotti (1974) shows how charged particles in the magnetosphere can be roughly classified according to energy and L shell, since the type of physical process which is locally dominant depends strongly on these factors. This chapter is about magnetospheric processes that influence the radiation belts, and it will involve brief discussions on some of these neighbouring populations.

1.2.2 Interaction with the Solar Wind

A plasma in which collisions are negligible and conductivity is high behaves according to the laws of ideal magnetohydrodynamics. Ohm's law associates plasma flow

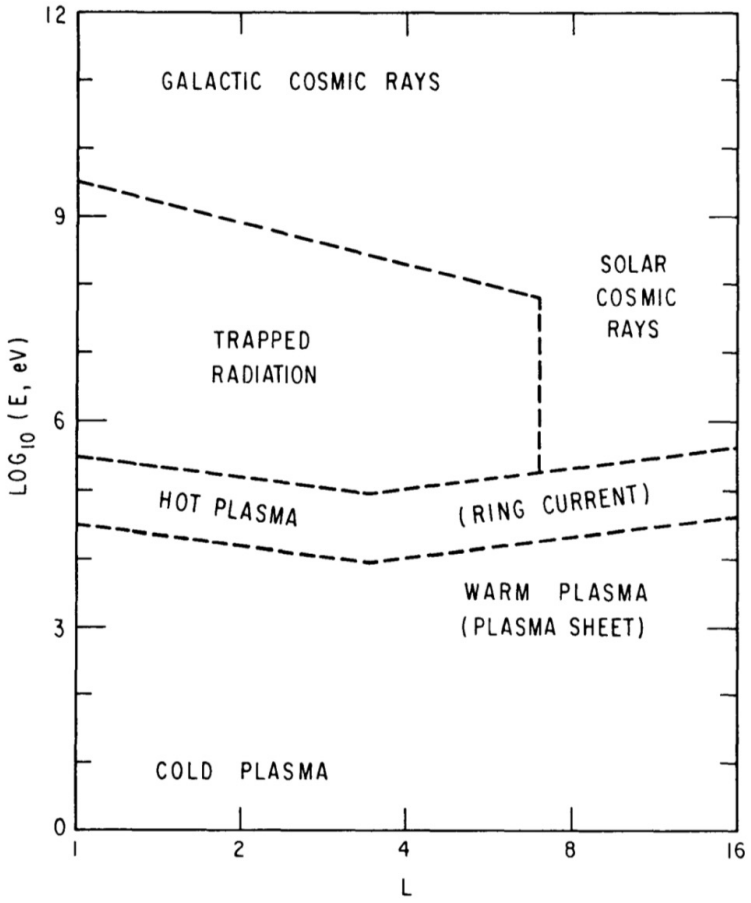


Figure 1.5: Spatial and spectral classifications of charged particles, from Figure 13 of Schulz and Lanzerotti (1974)

$\mathbf{u}$ with an electric field $\mathbf{E}$ according to:

$$\mathbf{E} + \mathbf{u} \times \mathbf{B} = 0 \tag{1.51}$$

As a result, movement of a plasma results in an electric current that in turn produces a net magnetic field such that magnetic field lines appear to move with the plasma. This gives rise to the "frozen in field" condition, whereby motion of a magnetic field line drags along the threaded plasma particles and vice-versa.

The frozen in field condition applies as heated plasma is ejected from the solar corona due to the pressure difference in interstellar space, thus it carries with it solar magnetic field lines. The tubes of magnetic flux frozen to the plasma spread radially outwards at $\sim$ 400km/s whilst the Sun rotates, forming a Parker spiral. The plasma constitutes the solar wind, and the field lines form the interplanetary magnetic field (IMF). By the time the solar wind reaches Earth it has a density of $\sim 10\text{cm}^{-3}$ and field strength of the IMF is $\sim$ 10nT (Kivelson and Russell, 1995).

The solar wind is deflected by the geomagnetic field at the boundary of the magnetosphere. The geomagnetic field is compressed inside this boundary, leading to a higher field strength of $\sim$ 75nT (Lopez and Gonzalez, 2017), and its orientation also differs generally to the solar wind, facing northward at the geomagnetic equator. The two colliding and differently oriented magnetic fields give rise to a current layer as a result of Ampere's law from the non-zero curl $\nabla \times \mathbf{B}$. This current layer is called the magnetopause and separates solar wind from the magnetosphere, illustrated in Figure 1.6, left panel. The magnetopause location is controlled by a balance between ram pressure of the solar wind and internal magnetic pressure from the geomagnetic field. The pressure balance is dynamic, but places the magnetopause surface $\sim 10R_E$ sunward of Earth.

In the direction downstream to the solar wind, the geomagnetic field becomes highly distorted and forms the magnetotail. Field lines in the northern lobe point generally Earthward, whilst field lines in the Southern lobe point tailward. Again, the difference in orientation gives rise to a current layer, illustrated in Figure 1.6, right panel. This current layer is called the neutral sheet, and forms another boundary surface of the magnetosphere.

Plasma also occupies geomagnetic field lines inside the magnetosphere, as shown

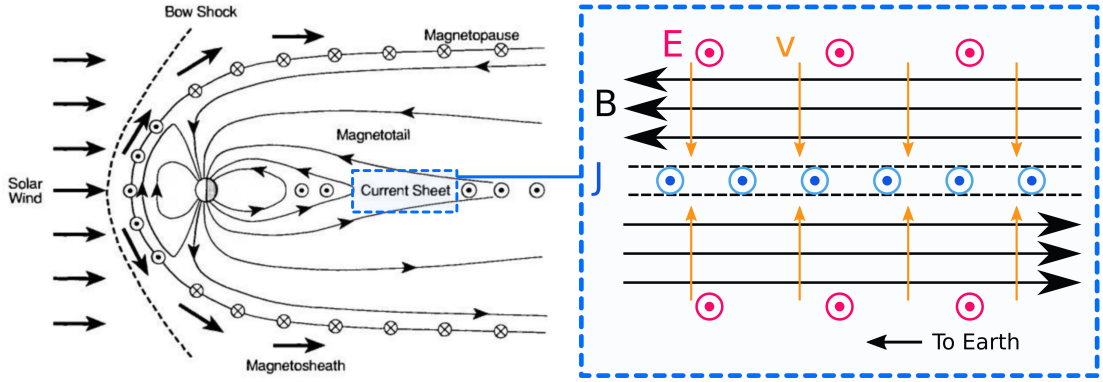


Figure 1.6: Overview of the magnetosphere boundary and the orientation of magnetic fields and currents near the neutral sheet, adapted and modified from Figure 9.1 of Kivelson and Russell (1995)

in Figure 1.5, but in comparison to the solar wind it is more rarefied. The difference in magnetic field orientation across the magnetopause means that incoming solar wind plasma and internal magnetospheric plasma generally do not mix due to the frozen in field condition. The magnetopause boundary therefore traces a demarcation between the two plasma populations. However, in regions where two sets of field lines oppose one another directly, magnetic reconnection can occur, resulting in the two sets of field lines becoming connected and the mixing of adjacent plasma populations. This process can connect field lines at various points across the boundary of the magnetosphere, including: Sunward of Earth (the dayside) through the nose of the magnetopause, between the IMF and geomagnetic field; and tailward of Earth, between oppositely oriented field lines of the magnetotail. Subsequently, magnetic field lines that connect the geomagnetic field to the IMF are said to be "open", whilst geomagnetic field lines that close in the opposite hemisphere are "closed".

Following dayside reconnection, open field lines tend to be swept tailward over Earth's poles as they are dragged by solar wind passing over the flanks of the magnetosphere. This is demonstrated by the progression in panels A, B and C of Figure 1.7 from Eastwood et al. (2015), where blue indicates closed IMF field lines, red indicates closed geomagnetic field lines, and purple indicates open field lines following reconnection. The dragging of field lines this way tends to drive convection of plasma throughout the magnetosphere in a process known as the

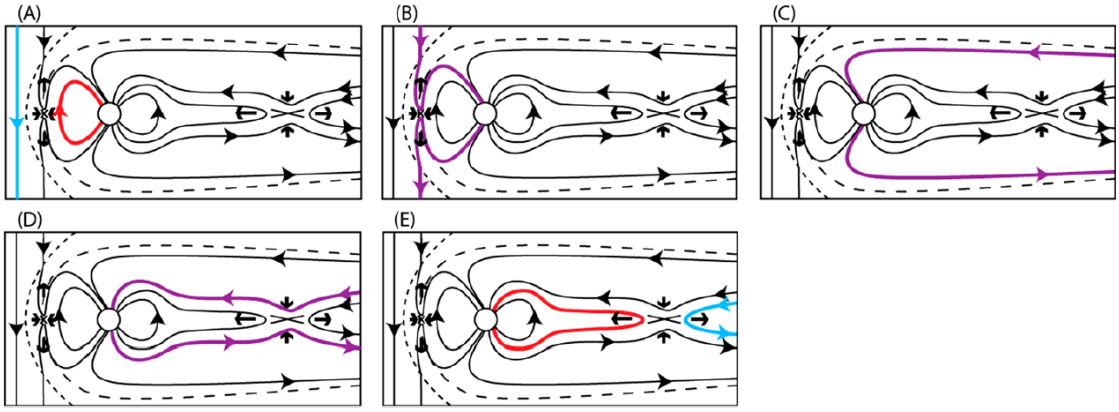


Figure 1.7: Time progression from panels A to E illustrating the convection of magnetic field lines as part of the Dungey cycle, taken from Figure 8 of Eastwood et al. (2015)

Dungey cycle (Dungey, 1961). Panels D and E of Figure 1.7 demonstrate the next part of the Dungey cycle, whereby open field lines convect into the magnetotail and move towards the neutral sheet, also indicated by the orange arrows in Figure 1.6 showing the motion of field lines. A buildup of field lines and plasma either side of the neutral sheet results in further reconnection, forming closed field lines on the nightside of Earth. Due to the frozen in field condition, convection of plasma over the course of the Dungey cycle obeys Equation 1.51 and is therefore accompanied by an electric field, known as the convection electric field.

1.2.3 The Arrival of Geomagnetic Storms

The rate of dayside reconnection is controlled by the orientation of the IMF. Periods of southward IMF result in the strongest coupling between the IMF and geomagnetic field, whereby large amounts of energy are transferred via Dungey cycle convection. This can lead to an imbalance when reconnection in the tail occurs at a slower rate, resulting in the buildup of open field lines and magnetic flux on the nightside of Earth. Over a prolonged period (hours), this constitutes the growth phase of a geomagnetic substorm, whereby the neutral sheet current becomes stronger and the magnetotail becomes stretched due to increased plasma pressure. Following the growth phase, substorm onset occurs; a burst of reconnection between field lines in the northern and southern lobes results in plasma being expelled towards

Earth. This strengthens the ring current as approaching plasma begins to undergo gradient and curvature drift due to the stronger magnetic field. Substorms may last several hours, and end with a recovery phase in which the ring current returns to pre-substorm levels.

During large solar flares, high density magnetic structures can be launched towards Earth, constituting a coronal mass ejections (CMEs). Such events are associated with interplanetary shocks travelling ahead of the ejecta due to the difference in density compared with the solar wind. Corotating interaction regions (CIRs) are another source of shock waves, formed when faster moving regions of the solar wind catch up to slower moving regions. When magnetic disturbances such as these arrive at Earth, a geomagnetic storm can occur.

The initial phase of a geomagnetic storm is marked by compression of the magnetosphere during the arrival of the disturbance. The main phase of a storm is then triggered by enhanced magnetospheric convection that gives rise to frequent substorms, leading to a strengthening of the ring current typically lasting for a few days. The stronger ring current leads to a depression in field strength at Earth's surface on the order of tens to hundreds of nanotesla. Finally, the recovery phase begins as the period of enhanced convection comes to an end, usually because of a change in orientation of the IMF to face Northward which is less conducive to dayside reconnection.

Interplanetary shocks travelling away from the Sun, such as those associated with CMEs and CIRs, also provide an acceleration mechanism for some particles to reach very high energies (keV to GeV), resulting in travel at much higher speeds than the surrounding solar wind (Reames, 2013). This is due to processes such as shock drift and Fermi acceleration that accelerate particles as they reflect or gyrate back and forth across the shock, exiting upstream or downstream. Figure 1.9 shows observations of energetic protons of solar origin by the IMP-8 satellite, which was situated on IMF field lines at $> 20R_E$ distance from Earth (McGuire and von Rosenvinge, 1984). These particles are classified as solar energetic particles (SEPs), and can travel quickly towards Earth along IMF field lines as a focused beam due to the weakening solar magnetic field reducing pitch angle to $\sim$0 (Ryan et al., 2000). The arrival of SEPs at Earth (an "SEP event") may precede the arrival of a shock, but also last several days. To give an idea of timescales, Figure 1.8 shows

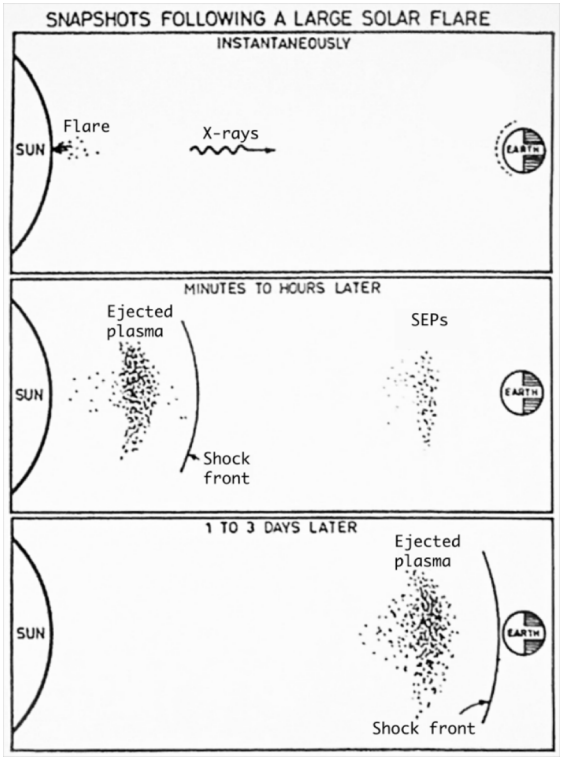


Figure 1.8: Snapshots following a large solar flare, demonstrating the arrival of high energy solar energetic particles before and after the shock front associated with a CME, adapted and modified from MacNamara (1994)

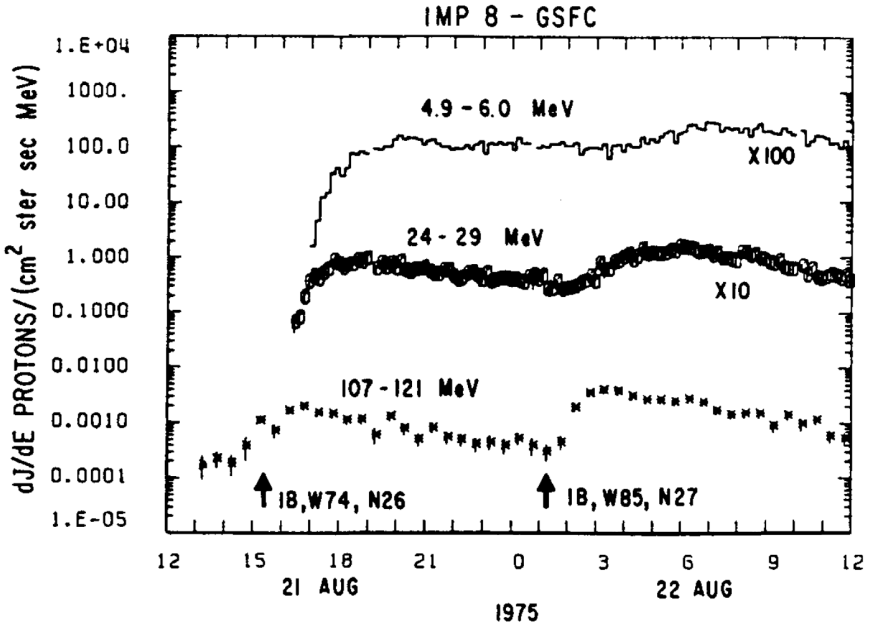


Figure 1.9: Time history of proton flux in three energy bands during two solar energetic particle events on 21-22 August, 1975, taken from Figure 1 of McGuire and von Rosenvinge (1984)

an overview of SEPs arriving at Earth followed by a shock front associated with a dense region of ejected plasma.

1.2.4 Time Variability of the Geomagnetic Field

Intermittent arrivals of geomagnetic storms compress the magnetosphere, inducing time variability in the geomagnetic field associated with stronger driving of magnetospheric current systems. The extent of the proton radiation belt in terms of energy and L depends on constraints imposed by the geomagnetic field; i.e., where it supports adiabatic motion. Time variability can therefore lead to nonadiabatic changes, and dynamic processes driven by magnetic activity can add to, or erode, the belt over short timescales (days or less) at the outer boundary.

The internally generated component of the geomagnetic field is approximately static over timescales of a few years. Therefore, the proton belt is magnetically shielded close to Earth by the internal field, and on this basis the proton belt can

be split into an outer zone ($L \gtrsim 1.7$), and stable inner zone ($L \lesssim 1.7$) wherein variability occurs over decades (Selesnick et al., 2016). Over long timescales (decades), secular variation of the geomagnetic field occurs due to changes in Earth's dynamo, but the effect on the radiation belts is purely adiabatic.

The internal field is modelled as the gradient of a scalar potential V, given by an expansion in spherical harmonics. First order expansion describes a dipole, and higher order terms account for magnetic anomalies in Earth's crust and complexities of Earth's dynamo. When the coordinate system in which V is expressed has its origin at the centre of Earth and is aligned with the polar axis, the first term of the expansion gives the centred dipole model. However, the coordinate system can be transformed such that second order (and higher) terms of V are minimised. First order expansion in this case produces the eccentric dipole model, with a centre not necessarily coincident with that of Earth's, and an axis tilted with respect to the polar axis that is prone to slowly change orientation over the course of secular variations.

Due to secular variation, coefficients used to expand V are updated based on satellite and ground measurements. The International Geomagnetic Reference Field (IGRF) is a time series of coefficients g_n^m and h_n^m that can be used to model the geomagnetic field for a given epoch. Field strength according to a dipole model of the geomagnetic field at radial distance $1R_E$ on the magnetic equator, B_0, is related to the coefficients involved in a first order expansion like so:

$$B_0^2 = (g_1^0)^2 + (g_1^1)^2 + (h_1^1)^2 \tag{1.52}$$

This value relates to the magnetic moment M of the dipole according to

$$M = \frac{4\pi}{\mu_0} B_0 a^3 \tag{1.53}$$

where μ_0 is vacuum permeability, with a value in SI units of 1.25663706e−6H/m (where 1H=1kg m^2 s^{-2} A^{-2}). Figure 1.10 plots the evolution in M and B_0 according to the IGRF coefficients, demonstrating secular variation of the centred dipole model over the last 100 years. The extent of variation in Figure 1.10 means that values of B_0 quoted in previous literature pertain to specific epochs.

Even when the internal field is modelled accurately using a higher order expan-

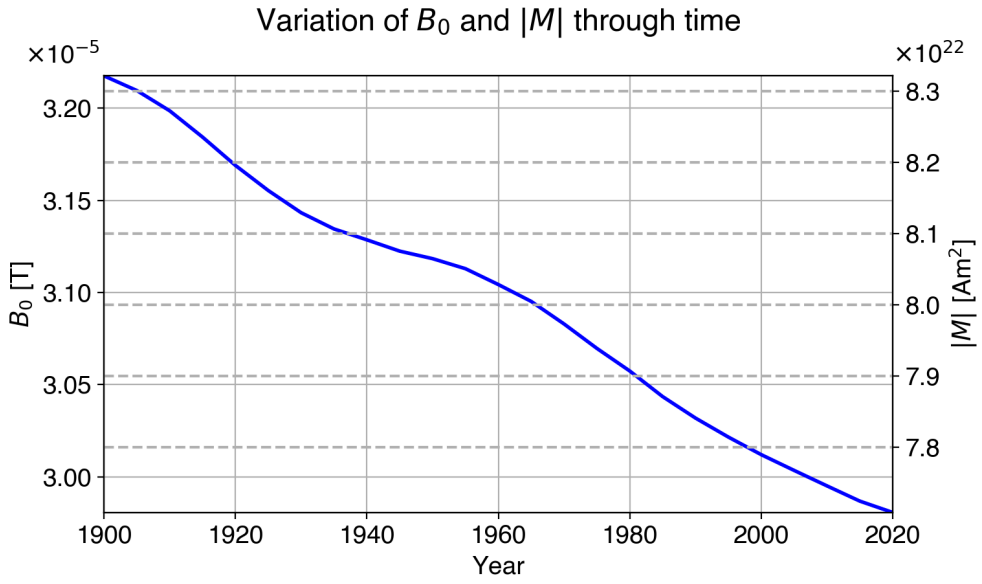


Figure 1.10: The left ordinate shows variation over 120 years in B_0, the field strength according to a dipole model of the geomagnetic field at radial distance $1R_E$ on the magnetic equator. The right ordinate shows the corresponding variation in the magnetic moment M.

sion of V, calculating the magnetic field in the outer zone proton belt is subject to uncertainty because there is an additional "external" component induced by magnetospheric current systems. Time variability necessitates the parameterisation of external magnetic field models by a geomagnetic activity index. In general however, the effect of the external field can be divided around the ring current region; the ring current causes a diamagnetic effect that suppresses the field within, whilst outside the ring current, the external field tends to increase magnetic field strength as the magnetosphere is in a compressed state.

1.3 Variability of Trapped Proton Flux

1.3.1 Flux and Phase Space Density

Spacecraft sensors typically measure particle flux in order to quantify the intensity of Earth's radiation belts at a particular location. Differential, directional flux j, for a given location $\mathbf{r}$, direction $\hat{\boldsymbol{\theta}}$ and energy E, measures the number of particles passing through a unit area perpendicular to $\hat{\boldsymbol{\theta}}$ at $\mathbf{r}$, per unit time, per unit solid angle $\mathrm{d}\Omega$ in the direction $\hat{\boldsymbol{\theta}}$, within the energy range E to $E + \mathrm{d}E$. The units are normally $\mathrm{cm}^{-2}\ \mathrm{s}^{-1}\ \mathrm{sr}^{-1}\ \mathrm{MeV}^{-1}$, and the number of particles is therefore given by

$$\mathrm{d}N(\mathbf{r}, E, \hat{\boldsymbol{\theta}}) = j(E, \hat{\boldsymbol{\theta}})\ \mathrm{d}A\ \mathrm{d}E\ \mathrm{d}\Omega \tag{1.54}$$

Direction is usually relative to the local magnetic field and specified in terms of pitch angle α. Therefore by measuring $j(E, \alpha)$ for $0 \leq \alpha \leq 180°$, the full angular distribution of flux can be summarised as a "pitch angle distribution". The pitch angle distribution of protons measured by the Relativistic Electron Proton Telescope on the Van Allen Probes satellites is shown in Figure 1.11, taken from Selesnick et al. (2014).

Pitch angle distributions summarise the intensity of particles along a portion of a drift shell, and can be used to investigate dynamical processes which may imprint certain signatures upon the distribution. In particular, measurements of pitch angle distributions from the magnetic equator are essential for showing the full population of trapped radiation at a given L. This is because, in contrast to

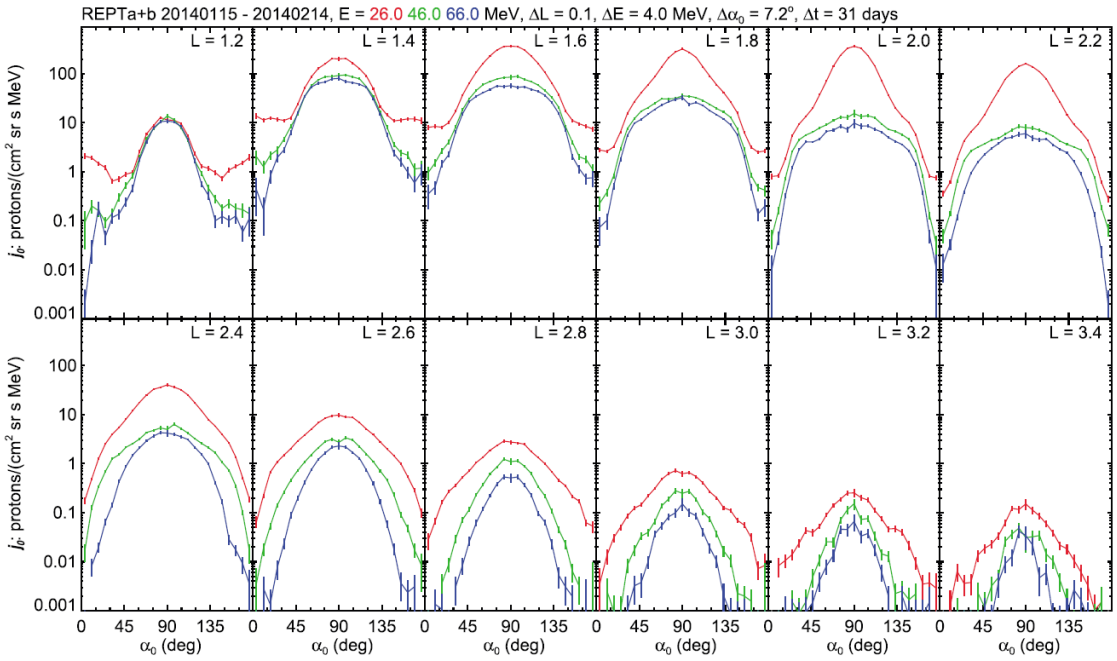


Figure 1.11: Proton equatorial pitch angle distributions measured by the Relativistic Electron Proton Telescope on the Van Allen Probes satellites at various L, showing $j(E, \alpha_{eq})$ at E =26, 46 and 66MeV (red, green and blue) over the period 15 January to 14 February 2014. Taken from Figure 4 of Selesnick et al. (2014)

equatorial measurements, measurements taken at latitudes away from the magnetic equator will not include the portion of particles on the drift shell that have mirrored between the point of observation and magnetic equator.

In order to interpret spacecraft measurements of flux, one must relate them to the coordinates of the population being measured. A coordinate system in terms of three adiabatic invariants provides a natural way to organise the radiation belts, because a set of three fixed invariants describes an individual drift path. However, there is spatial variation in flux over a drift path even in steady state that must be taken into account.

Liouville's theorem states that for a system in steady state, the phase space density of particles following a particular dynamical path is constant along the path. Phase space density is defined in relation to flux by

$$F(\mathbf{x}, \mathbf{p}) = \frac{j(E, \alpha_{eq})}{\rho^2} \tag{1.55}$$

where $\mathbf{x}$, $\mathbf{p}$ describe the spatial coordinates and their conjugate momenta. As a result of Liouville's theorem, time variations in phase space density at a fixed set of coordinates are not due to adiabatic motion, but rather highlight nonadiabatic changes in the population. A useful property of phase space density is

$$F(\mathbf{x}, \mathbf{p}) \propto F(\mu, J, \Phi) \tag{1.56}$$

where $F(\mu, J, \Phi)$ is a distribution function in adiabatic invariant space defined with respect to the canonical action integrals as coordinates. A distribution function f defined in another adiabatic invariant space, for example $f(\mu, J, L)$, can be related to $F(\mu, J, \Phi)$ since each coordinate can be related to the canonical action integrals. Therefore, a relation between f and flux exists via this property and Equation 1.55.

1.3.2 Transport in L

Time variation in the external component of the geomagnetic field gives rise to electromagnetic field perturbations within the radiation belts. As a result, trapped particles may undergo changes to the third adiabatic invariant and scatter across the field to neighbouring drift shells. This requires the perturbation to occur

over a small fraction of the drift period, and be asymmetric along the drift orbit (dependent on local time, Parker, 1960).

This mechanism has a key influence on the spatial distribution of particles. Simultaneous conservation of the first invariant also leads to changes in energy and pitch angle. It is therefore important to understand these processes in order to model the radiation belts, as well as to interpret spacecraft observations which may show signatures of this mechanism caused by past variability.

1.3.2.1 Changes in Energy and Pitch Angle

It can be shown using Faraday's law that changes in the local magnetic field strength induce a change in the momentum of radiation belt particles perpendicular to the magnetic field. When $\partial B/\partial t$ is approximately uniform over a gyration (thus also conserving μ):

$$\begin{aligned} \oint \frac{\partial \mathbf{B}}{\partial t} \cdot \mathrm{d}\mathbf{S} &= -\oint \nabla \times \mathbf{E} \cdot \mathrm{d}\mathbf{S} \\ \frac{\partial B}{\partial t} \pi {r_g}^2 &= -\oint \mathbf{E} \cdot \mathrm{d}\mathbf{l} \end{aligned} \tag{1.57}$$

As the magnetic field is approximately perpendicular to d$\mathbf{S}$, Equation 1.57 shows that $\partial B/\partial t$ results in work done by an induced electromotive force on the particle. The work done in one gyration is given by

$$W_g = -q \oint \mathbf{E} \cdot \mathrm{d}\mathbf{l} = q \frac{\partial B}{\partial t} \pi {r_g}^2 \tag{1.58}$$

For the case of magnetic mirroring, it was shown in Section 1.1.1.2 that a locally converging magnetic field geometry imparts a Lorentz force that deflects the particle away from the convergence. Equations 1.57 and 1.58 imply that the increase in perpendicular velocity of the particle (until the mirror point) is the transfer of work done by the particle against the Lorentz force in the direction parallel to the magnetic field, as it approaches a region of higher magnetic field strength until its parallel momentum is depleted. For the case of radiation belt particles near the magnetic equator, the field strength varies roughly as $B \propto 1/r^3$. Therefore, an inward/outward motion, or increase/decrease in geomagnetic field strength, leads to perpendicular energisation/de-energisation of the particle due to

$\partial B/\partial t$.

When a particle is inwardly transported due to violation of the third invariant, the increase in perpendicular momentum is maintained. A signature of this process may therefore be an energised population of particles with increased pitch angles, leading to a strong peak in the pitch angle distribution near 90°.

1.3.2.2 Radial Diffusion

Repeated, small perturbations cause the smoothing of gradients in proton phase space density as a function of L by scattering particles back and forth. Calculating the exact motion of particles requires that perturbations in the field are known, but this is usually not practical for simulations relying on spacecraft data. Therefore, the time evolution of phase space density is modelled as subject to radial diffusion, with a diffusion coefficient D_{LL}.

Accurately quantifying the effect of radial diffusion is a key challenge for proton belt modelling that will be explored later in this thesis: Chapter 3 begins with a review of the Fokker Planck equation used to describe changes in phase space density; Chapter 4 presents an attempt to constrain the value of D_{LL} for trapped protons by matching modelling results to observations during an era of high solar activity and proton belt enhancements; uncertainty in proton D_{LL} is further explored in Chapter 5, by experimenting with several different values to model the distribution of ∼MeV energy protons.

The magnetically shielded inner zone proton belt $L \lesssim 1.7$ generally exhibits long timescales for radial diffusion (∼years to decades), with shorter timescales in the outer zone. Figure 1.12 shows Van Allen probes measurements of the differential flux of equatorially mirroring protons versus L at two epochs separated by 20 months, from 24 to 76MeV. The peak near $L \sim 2$ appears to have diffused radially inward by $\sim 0.25 R_E$ over this period. At $L > 2$, enhancements in proton belt flux have been observed to form over ∼day timescales, with a suggested mechanism being radial diffusion at an increased rate, caused by a drift-resonant interaction with magnetosonic ULF waves (Lorentzen et al., 2002; Selesnick et al., 2010). Boscher et al. (1998) also suggest the idea of enhanced radial diffusion, occurring during active periods, in order explain how newly injected, low energy protons of

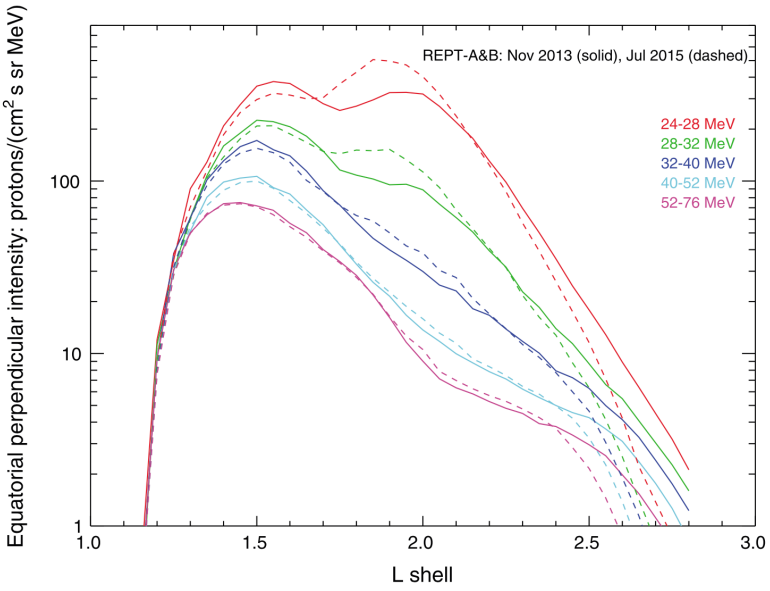


Figure 1.12: Inward radial diffusion of a flux peak of equatorially mirroring protons over a time period of ∼20 months, plotted for various energies. This Figure has been taken from Figure 1 of Selesnick et al. (2016)

solar origin near geostationary orbit are transported inward to the proton belt. Therefore, it is implied that the rate of proton radial diffusion may fluctuate and undergo temporary increases at $L > 2$, but this has yet to be fully explored.

1.3.2.3 Rapid Transport via Electric Impulses

Rapid, interplanetary shock-driven onset of a geomagnetic storm can include sudden compression and relaxation of the dayside magnetosphere, called a storm sudden commencement (SSC). An SSC was observed by the Combined Release and Radiation Effects Satellite (CRRES) on 24th March 1991, temporarily compressing the magnetosphere to within geostationary orbit distance, and resulting in a proton belt enhancement that has been widely studied as an example of an extreme space weather event. The SSC was coincident with the arrival of SEPs, and led to their rapid inward transport to form an enhancement in proton belt flux at $L \sim 2.5$ lasting for hundreds of days (Hudson et al., 1995).

The March 1991 event provides an example of the coherent acceleration of particles due to a drift-resonant interaction with a large electromagnetic field

perturbation. Rapid compression and relaxation of the magnetosphere during the event was measured by CRRES as a bipolar pulse in the electric field with a peak to peak magnitude of 80mV/m and a period of ∼150s, and in the magnetic field as a monopolar pulse with a magnitude of 140nT (Li et al., 1993; Wygant et al., 1994). The electric field pulse was mostly in the azimuthal direction, therefore particles in the longitudinal vicinity of the pulse, and with a drift velocity similar to the azimuthal velocity of the pulse, experienced a steady electric field acceleration over a portion of their drift orbit. Equation 1.18 shows that a force due to an electric field, given by $\mathbf{F} = \mathbf{E}q$, results in a guiding centre drift in the $\mathbf{E} \times \mathbf{B}$ direction. Therefore, particles subject to the electric field underwent radially inward or outward acceleration depending on which phase of the bipolar pulse they experienced. Simulations have shown that this led to inward transport of protons by ∼1-2R_E (Hudson et al., 1997).

1.3.3 Sources

1.3.3.1 Trapping of Solar Energetic Particles

The entry and trapping of SEPs provides an external source of $\gtrsim$MeV trapped protons at $L > 2$, with the solar origin confirmed by measurements of heavy ions (Mazur et al., 2006). Observations from the CRRES satellite shed light on this mechanism, and showed that enhancements in proton belt flux can form on rapid (∼minute) timescales coinciding with the injection of SEPs after a storm sudden commencement (SSC). For example, the 24th March 1991 event resulted in a second >20MeV proton belt forming at $L \sim 2.5$, lasting for hundreds of days (Mullen et al., 1991; Blake et al., 1992). Figure 1.13 shows two snapshots in time of the proton population observed during this event: after the arrival and injection of SEP particles (left panel); and just after the arrival of the shock (right panel). Figure 1.13 shows that inward penetration of SEPs was initially limited to $L \sim 4$ (left panel), but following the SSC, inward transport and energisation led to a dramatic enhancement down to $L \sim 2.5$ (right panel).

Attenuation of incoming SEP particle trajectories due to the geomagnetic field prevents their access to certain altitude-latitude combinations as a function of particle rigidity (momentum divided by charge), an effect known as geomagnetic

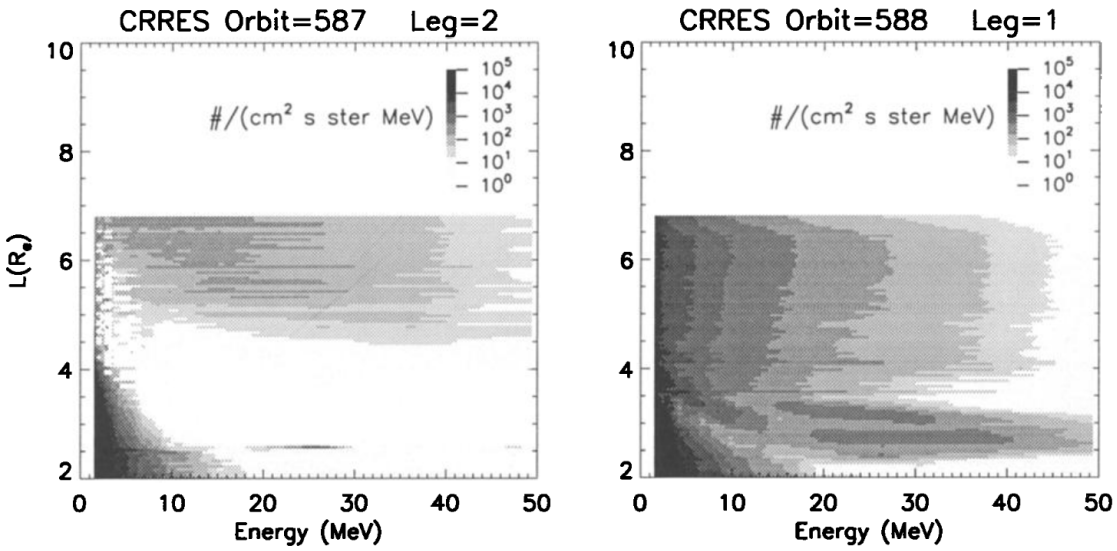
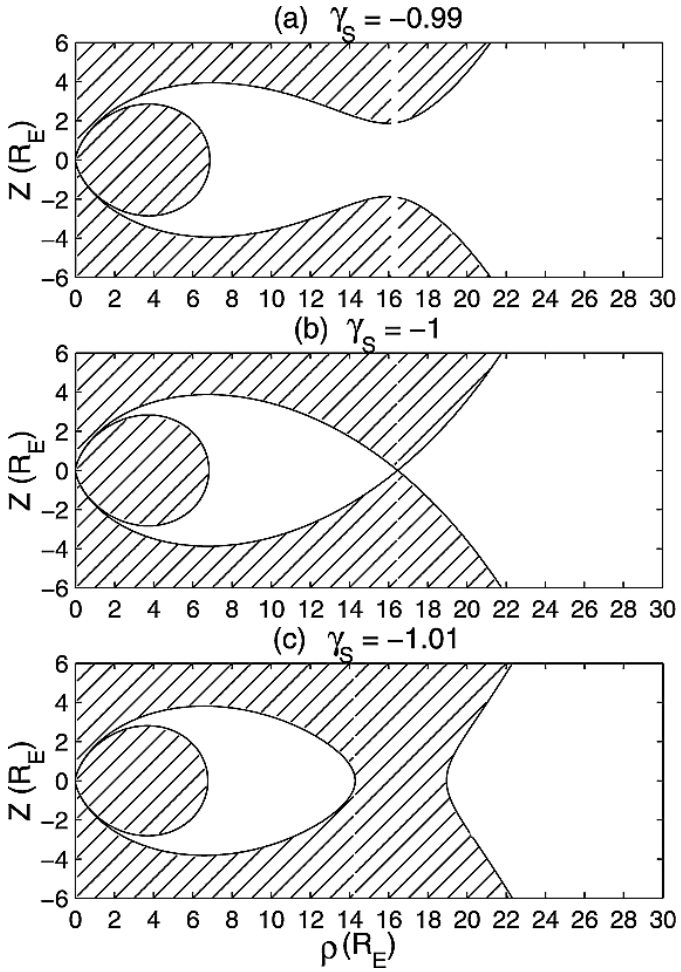


Figure 1.13: Differential flux of equatorially mirroring protons observed by CRRES at $L > 2$ just before (left) and just after (right) the arrival of an SSC in the March 1991 storm. Taken from Figure 1 of Hudson et al. (1996)

cutoff. Störmer theory describes geomagnetic cutoff surfaces as contours for a particular value of magnetic rigidity to represent the access limits for particles of that rigidity. The allowed regions for particle orbits, confined by these surfaces, is illustrated in Figure 1.14 from Kress et al. (2005) for different values of dimensionless parameter γ_S (see Equation 2 of Kress et al., 2005), a function of particle rigidity and dipole moment. In Figure 1.14, an innermost cutoff surface controls the deepest extent of particle access, but an outer cutoff surface may block particles approaching from outside the region, or allow them depending on the value of γ_S.

When Störmer theory is applied to the geomagnetic field, entry to the inner magnetosphere depends on an outer cutoff surface as illustrated in Figure 1.14. By default, entry of incoming SEPs may be blocked, but the dependence of γ implies that when the geomagnetic field is compressed, such as during the arrival of an SSC, the outer cutoff surface can reconfigure to allow entry to the inner region. This effect is known as geomagnetic cutoff suppression, and can allow entry of SEPs through the front-side magnetopause to the radiation belt region.

Kress et al. (2004) and Kress et al. (2005) investigated the trajectories of SEPs at tens of MeV by solving the equation of motion for many individual particles,



. The shaded areas indicate forbidden regions.

Figure 1.14: Three different configurations of outer cutoff surface demonstrated for a dipole model of the geomagnetic field as calculated using Störmer theory, adapted and simplified from Figure 2 of Kress et al. (2005)

and showed that trapping of incoming SEPs is moderated by reconfiguration of the geomagnetic field during the arrival of a geomagnetic storm. Simulation results presented in Figure 1.15 show entry and trapping of 25MeV protons (SEPs) through the front side magnetopause. The arrival of a shock (the region of increased density) leads to geomagnetic cutoff suppression, allowing the particles to penetrate to $L \sim 4$ via low-latitude entry (third panel). As the field decompresses (fourth panel), the subsequent restoration of geomagnetic cutoff (within timescales comparable to one drift orbit) then prevents particles from leaving by trapping the particles between an inner and outer cutoff surface. After gaining access to the inner magnetosphere this way, coherent acceleration of SEPs by an electric field pulse associated with the SSC, via the mechanism described in Section 1.3.2.3, can then lead to further inward transport and the formation of trapped enhancements within the proton belt's outer region ($L > 2$).

The direct injection of SEPs through the front-side magnetopause applied to $\gtrsim$10MeV particles only. However, $\lesssim$1MeV SEPs provide a low energy source of protons, penetrating the magnetosphere via open field lines in the magnetotail (Blake et al., 2019). Injection of such low energy protons into the outermost trapped population near geostationary orbit has been shown to occur during impulsive reconfiguration of the geomagnetic field (i.e. Baker and Belian, 1985; Wang et al., 2008). Some mechanism of enhanced radial diffusion is then inferred to allow inward transport to the proton belt region (Boscher et al., 1998).

1.3.3.2 Cosmic Ray Albedo Neutron Decay

In addition to protons gaining access through the frontside magnetopause and geomagnetic tail, the radiation belts have an internal source of protons produced by the beta decay of neutrons escaping from the atmosphere. This is called the cosmic ray albedo neutron decay source (CRAND), and it is responsible for the distribution of trapped protons at $L \lesssim 1.25$ and $E \gtrsim 50$MeV (Jentsch, 1981).

This process begins with the arrival of galactic cosmic rays at Earth. Figure 1.16 shows modelled energy spectra of arriving H and He cosmic ray particles, and the dependence on solar cycle. The inverse relationship between flux and solar activity is due to the increased attenuation of galactic cosmic rays during solar

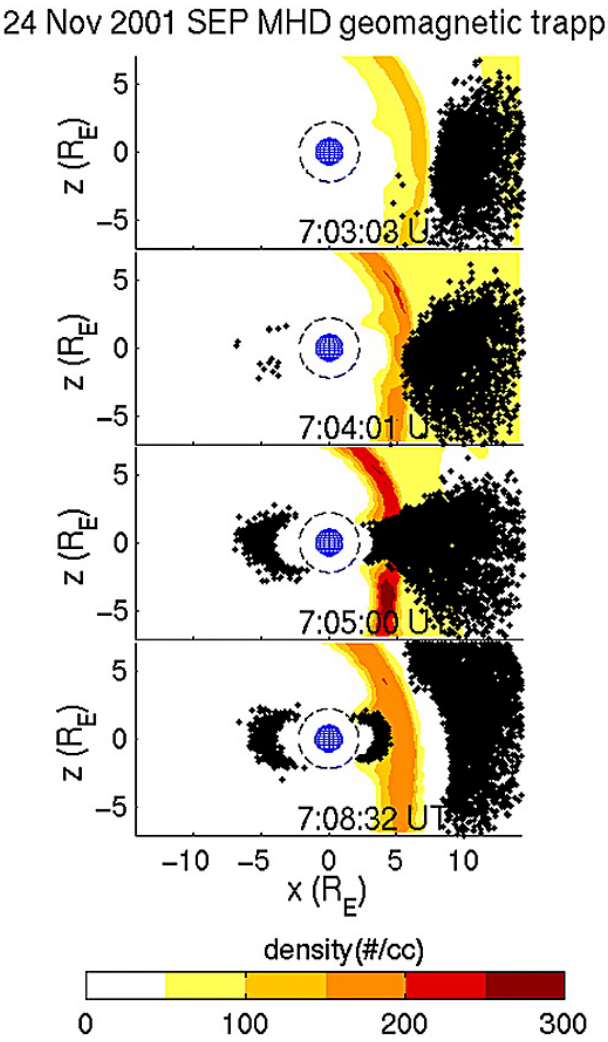


Figure 1.15: The time evolution of a 25MeV proton population, calculated by following particle trajectories in a field calculated using a time dependent MHD code during the simulated arrival of a geomagnetic storm. Taken from Figure 3 of Kress et al. (2005)

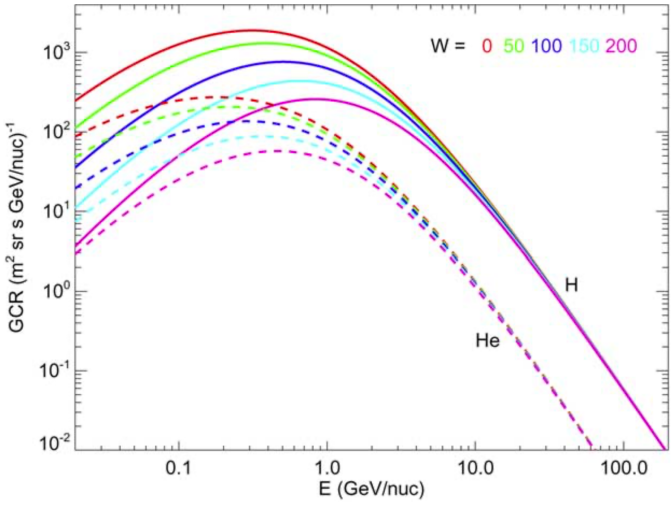


Figure 1.16: Modelled energy spectra of galactic cosmic ray particles arriving at Earth (H, solid; He, dashed) as a function of sunspot number W as a proxy for solar activity, taken from Figure 7 of Selesnick et al. (2007)

maximum, when the heliospheric magnetic field is stronger. Collisions of incoming cosmic rays with atmospheric constituents produce neutrons in all directions leading to an upward flux, seen in high altitude balloon measurements at the top of the atmosphere (Singer, 1958; Vernov et al., 1959).

Neutrons have no charge so their trajectory is not altered by the geomagnetic field. However, after an average time of 887s after production, albedo neutrons (newly produced neutrons with an upward trajectory) undergo beta decay, producing a proton, electron and an antineutrino (Singer, 1958). This average lifetime for albedo neutrons is easily long enough to escape the radiation belt region. However, a small fraction of albedo neutrons decay much sooner than this. Therefore, amongst the fraction of neutrons undergoing early decay, there is a chance that new protons may be produced in the radiation belt region.

Protons produced as a result of albedo neutron beta decay move in approximately the same direction and with the same kinetic energy as the neutron. If production happens to coincide with a region, energy range and pitch angle conducive to adiabatic trapping, the proton becomes part of the proton radiation belt. This is demonstrated in Figure 1.17, which shows the proton produced by beta decay at two different locations, one of which becomes trapped, whilst the other is lost to

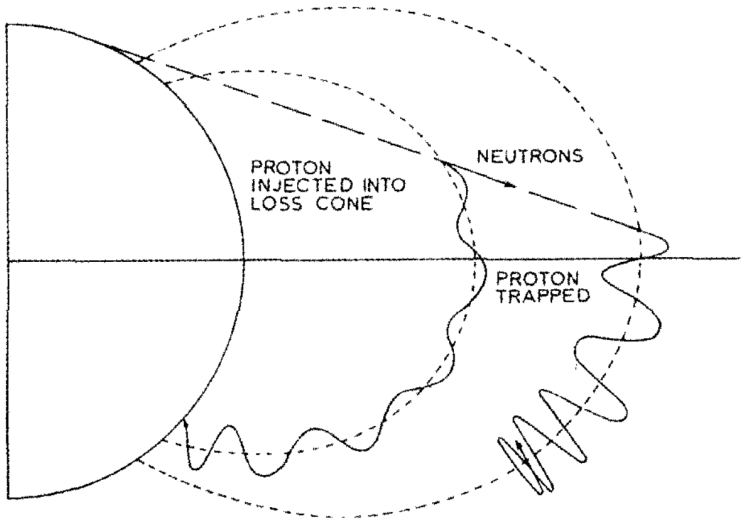


Figure 1.17: Schematic showing two possible outcomes following the beta decay of a neutron within the radiation belt region. Taken from Figure 1 of Singer and Lemaire (2009)

the atmosphere.

1.3.4 Losses

1.3.4.1 Coulomb Collisions

Radiation belt protons undergo coulomb collisions with free and bound electrons present in the atmosphere, ionosphere and plasmasphere. Due to the much higher energy of the proton, this type of collision does not significantly alter pitch angle. However, protons are decelerated and eventually lost from the radiation belts. Stopping power, or energy loss per unit distance, due to coulomb collisions increases as a particle slows down, resulting in a Bragg curve. As a result, the rate of change in the first invariant μ is higher for lower energy protons.

An increase in the rate of coulomb collisional loss occurs during solar maximum. This is due to thermal expansion of the atmosphere, caused by heating from increased extreme ultraviolet radiation (Fuller-Rowell et al., 2004). At fixed altitude in the radiation belts, this leads to higher density and therefore more coulomb collisions occur. The opposite effect applies during solar minimum, leading

to a solar cycle driving with a timescale much shorter than changes due to radial diffusion. Variations in proton intensity at low altitude ($L \lesssim 1.3$) are mostly driven by variations in the rate of coulomb collisional loss due to this effect (Li et al., 2020), and the solar cycle driving is stronger for lower energy protons.

1.3.4.2 Processes Controlling the Outer Boundary

During a geomagnetic storm, build-up of the ring current can cause outward motion of drift orbits associated with conservation of the third invariant, in addition to field line curvature scattering due to breakdown of adiabatic motion along stretched field lines (Anderson et al., 1997). Engel et al. (2015) and Engel et al. (2016) showed that modelling these two effects can account for losses reaching L$\sim$2.5. The outer boundary of trapped flux, controlled by these losses, can therefore move to lower L shell over several hours corresponding to build-up of the ring current. Observations show the time taken for the outer boundary to recover is hundreds of days due to outward radial diffusion (Selesnick et al., 2010). Some trapped enhancements, particularly at high L, may therefore be short-lived because of subsequent variability.

1.4 The Exposure of Satellites

1.4.1 Overview

Irradiation of spacecraft by trapped protons causes damage to solar cells. More specifically, incoming protons penetrate the upward mounted surface of the cell and displace atoms in the crystal lattice structure. This causes defects throughout the device that decrease the lifetime of minority charge carriers. This type of damage is referred to as non-ionising dose, and is caused in particular by protons around $\sim$10MeV in energy, depending on the coverglass shielding thickness, which attenuates the proton before it impacts the solar cell (Messenger et al., 1997).

The end result is a reduction in maximum power availability from the cell (as well as other photovoltaic output parameters), referred to as degradation, which can adversely affect operations because solar cells are usually required to output

close to maximum power. For spacecraft frequently traversing the proton belt, non-ionising dose from trapped protons is a primary cause of solar cell degradation and therefore a key factor influencing mission lifetime. The degradation effect on solar cells of the Akebono, Tacsat-4, Arase and Van Allen Probes satellites are documented in the following example literature: Miyake et al. (2014); Jenkins et al. (2014); Toda et al. (2018); and Maurer et al. (2018).

The following sections present background concepts to understand how the calculation of non-ionising dose and solar cell degradation is performed. Later in this thesis, the NRL method explained in Section 1.4.2.3 will be used.

1.4.2 The Calculation of Non-ionising Dose

1.4.2.1 Overview of the JPL and NRL Methods

The calculation of solar cell degradation due to non-ionising events is set out by Messenger et al. (2001) in terms of two methods. The first has been developed by the US Jet Propulsion Laboratory (JPL method) and second, newer method, by the US Naval Research Laboratory (NRL method). Practically there are two important differences to consider when choosing a method. These are: a.) data availability; and b.) the way resulting damage is quantified. In more detail:

(a) The JPL method relies on data collected by bombarding a solar cell with mono-energetic beams, using eight proton energies and three electron energies (at a minimum) and measuring the degradation under each separately. The JPL method therefore depends on extensive testing. On the other hand, the NRL method requires degradation data for just one proton energy and one electron energy (see Section 1.4.2.3 for why only one electron energy is needed) (Baur et al., 2017).

(b) The JPL method calculates the 1MeV electron equivalent fluence to quantify damage, whereas the NRL method is in terms of displacement damage dose, with units MeV/g. These are explained further in Sections 1.4.2.2 and 1.4.2.3.

After using either method to calculate damage to a solar cell, this can be converted to degradation (the equivalent drop in output power or voltage, etc.). An

input required by both methods is the total fluence (time integrated flux) spectrum of particles incident on the solar cell. In the context of mission planning this is the output given by a radiation belt model.

1.4.2.2 JPL method

Calculating Damage via 1MeV Electron Equivalent Fluence

An explanation of the JPL method invokes the concept of relative damage. This is damage caused by a particle at one particular energy relative to that caused by a particle of the same species at a reference energy. Using this concept, fluence at any energy can be converted to an equivalent fluence at the reference energy that would result in the same amount of degradation.

Relative damage as a function of energy is quantified by a "relative damage coefficient" (RDC). Deriving a RDC at energy E for particle species x begins with measuring "critical fluence" at this energy $\Phi_{x,C}(E)$. This is the fluence of perpendicularly incident monoenergetic particles that, incident on a solar cell, causes a particular photovoltaic parameter (output power, voltage or current) to degrade to 75% of its original value. This critical fluence can be measured for any energy of proton or electron, represented by $\Phi_{p,C}(E)$ and $\Phi_{e,C}(E)$ respectively. For protons, the RDC $D_p(E)$, is then given by the ratio of the critical fluence at a 10MeV reference energy to this critical fluence:

$$D_p\left(E\right) = \frac{\Phi_{p,C}\left(\text{10MeV protons}\right)}{\Phi_{p,C}\left(E\right)} \tag{1.59}$$

The electron RDC, $D_e(E)$, is given by the equivalent ratio, but using a critical fluence at a reference energy of 1MeV instead:

$$D_e\left(E\right) = \frac{\Phi_{e,C}\left(\text{1MeV electrons}\right)}{\Phi_{e,C}\left(E\right)} \tag{1.60}$$

Degradation curves, showing the gradual loss of maximum output power under bombardment at different energies, are shown in Figure 1.18. As an example, it can be seen in Figure 1.18 that the 2.4MeV electron degradation curve falls off more rapidly than the 1MeV electron (reference energy) curve, as the rate of degradation

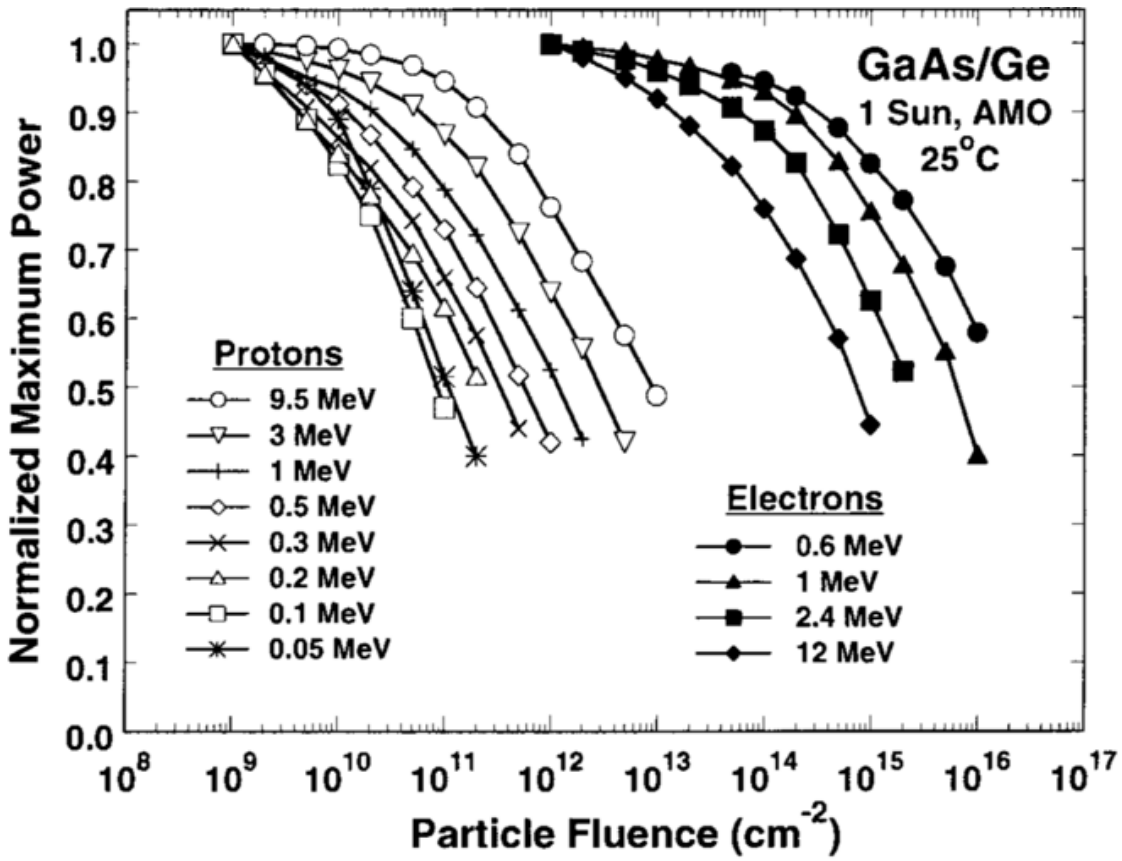


Figure 1.18: The remaining normalised maximum power of an example GaAs solar cell under bombardment at different energies, with measurements taken at various fluence levels for each energy of proton and electron, from Messenger et al. (2001)

is higher. Therefore critical fluence is lower, and the electron RDC at 2.4MeV is greater than unity.

A complete set of RDC data must be determined for this method. This process can be summarised like so:

(a) The RDC is calculated for each test energy, with an example shown in Figure 1.18. The RDC at other energies can then be interpolated/extrapolated to determine RDCs across the whole spectrum.

(b) In a realistic setting, fluence is omnidirectional. The RDCs are therefore converted for omnidirectional fluence. At the same energy, the fluence of omnidirectional particles required to cause the same degradation as perpendicular

particles is higher. The perpendicular critical fluence at the reference energy is still the numerator in Equations 1.59 and 1.60, therefore the omnidirectional RDCs are smaller. The critical fluence for omnidirectional particles can be derived from the critical fluence for perpendicular particles by considering the geometry and does not require re-testing (see Messenger et al., 2001).

(c) In a realistic setting, incident particle energy is also reduced by solar cell coverglass shielding. Further calculations account for the effect of shielding at several thicknesses on the energy of incident particles, to produce a set of omnidirectional RDCs for each shielding thickness.

A set of RDC curves describe the relative damage as a function of energy for each photovoltaic parameter (output power, voltage, etc.), for the solar cell material and structure used in testing paired with the particular coverglass material used to calculate the effect of shielding. As the effect of shielding is calculated for various thicknesses, one can interpolate to find the RDCs for any thickness. Such a dataset is represented in Figure 1.19 for protons.

Using the appropriate set of RDC data, an incident fluence spectrum can be converted to a single value, called the 1MeV electron equivalent fluence. This is the fluence of 1MeV electrons that would cause the same degradation as the original fluence spectrum (thus "equivalent"), given by:

$$\Phi_{\text{1MeV e. equivalent}} = \int \frac{\mathrm{d}\Phi_e\left(E_e\right)}{\mathrm{d}E_e} D_e\left(E_e\right) \mathrm{d}E_e + D_{pe} \int \frac{\mathrm{d}\Phi_p\left(E_p\right)}{\mathrm{d}E_p} D_p\left(E_p\right) \mathrm{d}E_p \quad (1.61)$$

where D_p, D_e represent the proton, electron RDC as a function of energy E, $\mathrm{d}\Phi(E_x)/\mathrm{d}E_x$ is differential fluence at energy E, and D_{pe} represents the proton to electron damage ratio. D_{pe} is the ratio of the critical fluence of perpendicular 1MeV electrons to critical fluence of perpendicular 10MeV protons, given by

$$D_{pe} = \frac{\Phi_{e,C}\left(\text{1MeV electrons}\right)}{\Phi_{p,C}\left(\text{10MeV protons}\right)} \quad (1.62)$$

Proton RDCs relate to 10MeV protons as the reference particle, therefore the integral on the right in Equation 1.61 represents the 10MeV proton equivalent fluence. D_{pe} is therefore used to convert 10MeV proton equivalent fluence to 1MeV

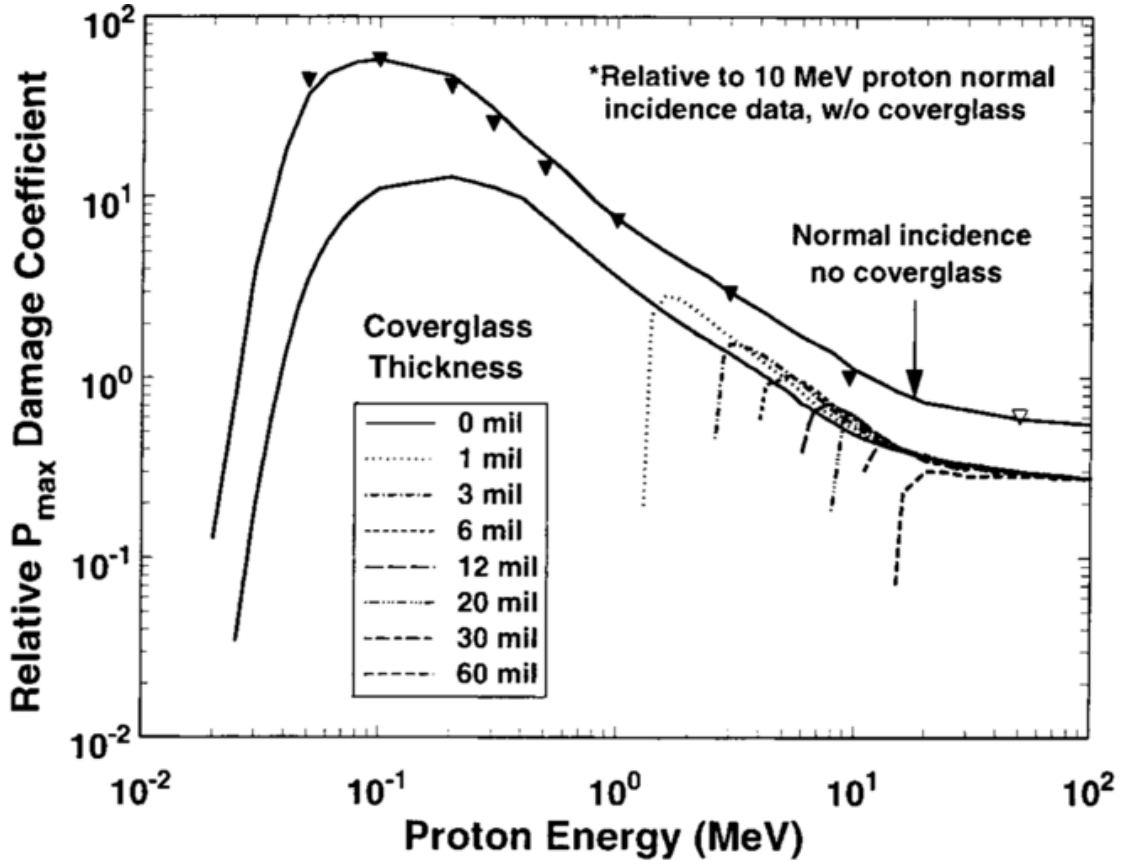


Figure 1.19: Example RDC curves for a GaAs-based solar cell showing the calculation of omnidirectional RDCs and the effect of various shielding thicknesses, compared to the perpendicular unshielded RDC curve (top), from Messenger et al. (2001)

Radiation Performance at 1 MeV Electron Irradiation, EOL/BOL Ratios

Fluence (e/cm²)	Voc	Isc	Vmp	Imp	Pmp(1)
3.00 E+ 13	0.96	0.99	0.98	0.99	0.99
1.00 E+ 14	0.95	0.98	0.97	0.99	0.96
5.00 E+ 14	0.91	0.97	0.93	0.96	0.90
1.00 E+ 15	0.89	0.94	0.91	0.94	0.85
3.00 E+ 15	0.86	0.89	0.87	0.86	0.75
1.00 E + 16	0.82	0.82	0.83	0.74	0.62

(1) Per AIAA-S-111 Standards

Figure 1.20: Response of a SolAero ZTJ solar cell to an equivalent fluence of 1MeV electrons, in terms of degradation of various photovoltaic parameters (SolAero Technologies Corp, 2018)

electron equivalent fluence.

Calculating Degradation

Finally, the 1MeV electron equivalent fluence can be converted to remaining maximum output of any photovoltaic parameter by using the corresponding characteristic degradation curve (degradation versus 1MeV electron fluence). For power output, this is the 1MeV electron curve in Figure 1.18. The characteristic curve can be fit using an equation of the form

$$\frac{Z}{Z_0} = 1 - C\log\left(1 + \frac{F}{F_X}\right) \tag{1.63}$$

where Z/Z_0 is the remaining output ratio of any photovoltaic parameter (such as the normalised maximum power P/P_0) and F is the 1MeV electron fluence. C and F_X are fitting parameters that parameterise the curve, and therefore describe the degradation properties of the solar cell. Alternatively, the remaining output at various 1MeV electron fluence levels can be given to recreate the curve. This is the norm in solar cell spec sheets, an example of which is shown for SolAero ZTJ cells in Figure 1.20 (SolAero Technologies Corp, 2018).

1.4.2.3 NRL Method

Calculating Damage via Displacement Damage Dose

In the NRL method, a key concept is non-ionising energy loss (NIEL). In non-ionising events NIEL is the rate at which energy is transferred from incident particle to target atoms due to various types of interactions (Rutherford scattering as well as nuclear elastic and inelastic), with units MeV cm^2/g. NIEL is a function of energy, and can be thought of as the (non-ionising) damage-causing energy applied per mass of target material in a collision.

A convenient feature of NIEL is that it is proportional to the RDC for a given species (see Section 1.4.2.2 to understand the concept of relative damage). Therefore, by multiplying a monoenergetic fluence at a given energy by the corresponding NIEL value for that energy, the product is a normalised measure of damage (with respect to the species of particle) with units MeV/g. This product is called the displacement damage dose, given by

$$D_d = \Phi_x(E)\, S_x(E) \tag{1.64}$$

for a particle of species x, where $S_x(E)$ is NIEL at energy E. The reason for the linear dependence of NIEL on RDCs is that damage occurs via a similar mechanism over the whole energy range of interest. This is indicated in Figure 1.18 because the degradation curves have approximately the same shape. In other words, for any chosen energy, the effect of many non-ionising collisions in the target lattice can be reproduced using any other energy, the only difference being the amount of fluence required which is accounted for by the NIEL rate for that species. However, perhaps the most important difference between NIEL and a RDC is that NIEL can be calculated theoretically, as opposed to being empirically deduced from test data.

Displacement damage dose is somewhat analogous to the 1MeV electron equivalent fluence used in the JPL method, as it can be calculated for fluence at each energy across an incident spectrum and summed, taking into account the relative damage of each energy. Equation 1.65 gives total D_d imparted by a fluence of protons and electrons by integrating across all energies:

$$D_{d,\text{total}} = \int \frac{\mathrm{d}\Phi_p\left(E_p\right)}{\mathrm{d}E_p} S_p\left(E_p\right) \mathrm{d}E_p + \frac{1}{R_{ep}} \int \frac{\mathrm{d}\Phi_e\left(E_e\right)}{\mathrm{d}E_e} S_e\left(E_e\right) \mathrm{d}E_e \tag{1.65}$$

Quantities $\mathrm{d}\Phi_p(E_p)/\mathrm{d}E_p$ and $\mathrm{d}\Phi_e(E_e)/\mathrm{d}E_e$ are the omnidirectional proton and electron fluences after the effects of coverglass shielding, $S_p(E)$ and $S_e(E)$ are proton and electron NIEL at energy E, and R_{ep} is the electron to proton damage equivalency factor. Displacement damage dose is normalised with respect to a particular species, but by default it is implied with respect to protons. Therefore, the left and right integrals in Equation 1.65 (without the factor R_{ep}) actually give proton displacement damage dose and electron displacement damage dose separately. This subtlety leads to the proton damage equivalency factor R_{ep}, which converts electron displacement damage dose to equivalent proton displacement damage dose, analogous to the proton to electron damage ratio in Equation 1.61. A practical way to calculate R_{ep} is given in Section 1.4.2.3 in terms of solar cell degradation parameters.

Calculating Degradation

This occurrence of similar damage type at every energy is a key conclusion because it allows the solar cell degradation as a function of displacement dose to be calculated using just one test energy using the theoretically derived NIEL rate. For example, by multiplying any of the proton fluence curves in Figure 1.18 by the NIEL value corresponding to their energy, one obtains the same characteristic degradation curve in terms of maximum output versus displacement damage dose. This occurs separately for protons and electrons resulting in two separate characteristic degradation curves for proton and electron-induced dose.

As in Section 1.4.2.2, characteristic curves for displacement damage dose can be fitted in terms of two parameters, C and D_X using

$$\frac{Z}{Z_0} = 1 - C\log\left(1 + \frac{D_d}{D_X}\right) \tag{1.66}$$

where D_X is equivalent to F_X in Equation 1.62, but with the new units of MeV g^{-1}.

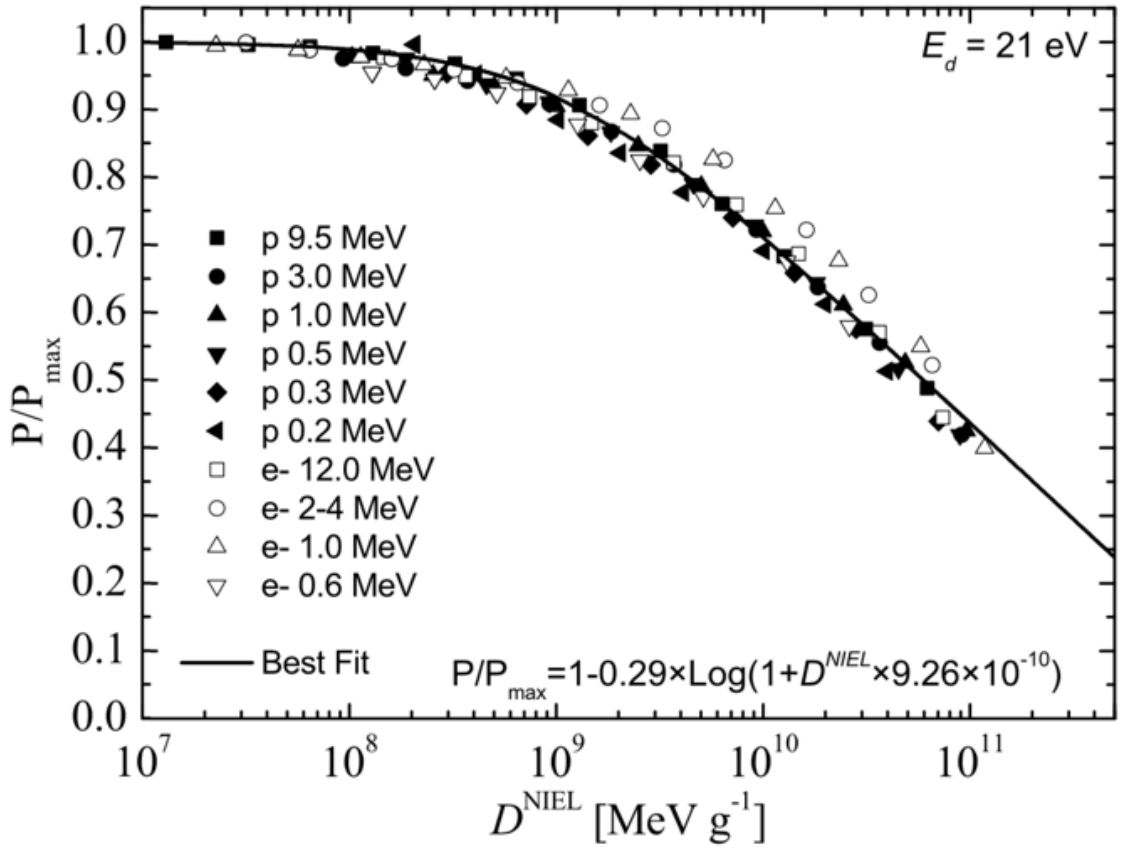


Figure 1.21: Characteristic degradation curve in terms of non-ionising displacement damage dose for a GaAs single junction solar, obtained using NIEL values with a 21eV displacement energy threshold, from Baur et al. (2014)

By deriving one proton degradation curve (as described above) and one electron degradation curve, then fitting Equation 1.66 to each, one can derive C and D_X in both cases. The value of C is identical in either the proton or electron case, but D_X is not. This leads to the quantities D_{eX} and D_{pX}, which are fitting parameters for the solar cell characteristic degradation curves for electrons and protons respectively for use in Equation 1.66.

The electron to proton damage equivalency factor, used to convert displacement damage caused by electrons to an equivalent displacement damage in terms of protons as in Equation 1.65, can be derived using these two fitting parameters:

$$R_{ep} = D_{eX}/D_{pX} \tag{1.67}$$

When the electron fluence curves in Figure 1.18 are multiplied by the NIEL value corresponding to their energy, then by R_{ep}, they align with the proton curves. This process is shown by the collapse of proton and electron fluence curves in Figure 1.21 to the same characteristic curve from Baur et al. (2014).

It is also useful to know that, for a solar cell of interest, 1MeV electron equivalent fluence data such as that shown in Figure 1.20 can be converted to electron displacement damage dose by multiplying with the 1MeV electron NIEL value, then the fitting parameters C and D_{eX} can be found by fitting Equation 1.66.

Classic NIEL and Changes to the Method

In Messenger et al. (2001), Equation 1.65 is instead written:

$$D_{d,\text{total}} = + \int \frac{\mathrm{d}\Phi_p\left(E_p\right)}{\mathrm{d}E_p} S_p\left(E_p\right) \mathrm{d}E_p + \frac{1}{R_{ep}} \int \frac{\mathrm{d}\Phi_e\left(E_e\right)}{\mathrm{d}E_e} S_e\left(E_e\right) \left[\frac{S_e\left(E_e\right)}{S_e\left(1\text{MeV electron}\right)}\right]^{n-1} \mathrm{d}E_e \tag{1.68}$$

Compared to this, Equation 1.65 above is simplified because it does not contain any non-linear dependence on electron NIEL values. This is because $S_p(E)$ and $S_e(E)$ represent NIEL calculated under the assumption of a minimum displacement energy $E_d = 21\text{eV}$ in the target material, versus the NIEL for $E_d = 10\text{eV}$ shown in Messenger et al. Minimum displacement energy is the minimum incident particle energy required to cause a permanent displacement of an atom in a particular

material. The increase in E_d for the calculation of NIEL accounts for annealing effects, and has been shown for GaAs-based cells to result in NIEL coefficients that produce the same degradation curve across all electron energies, hence the linear dependence (Baur et al., 2014; Pellegrino et al., 2020). Therefore, for the solar cell technologies under investigation in this work, the NIEL curves for $E_d = 21$eV allow a value of $n = 1$ to be used, and Equation 1.65 is the correct calculation of displacement damage dose.

Shielding calculation in the NRL method

Although RDCs take longer to derive experimentally, they take into account the effect of shielding through the extra calculation explained in Section 1.4.2.2. To calculate the effect of shielding in the NRL method, a transport code must be used to acquire the slowed down spectrum of fluence after attenuation by shielding. It is this slowed down spectrum which is used as input to calculate D_d.

The transport code MULASSIS (Lei et al., 2002) is an example code that was used to simulate the effect of coverglass shielding. Although this greatly enhances execution time it is much more flexible because any thickness of shielding can be considered and one is not constrained by the availability of experimental data, unlike the JPL method.

1.4.3 The TacSat-4 Solar Cell Experiment

The Tacsat-4 Solar Cell Experiment was launched on 27th September 2011 into highly elliptical orbit (700 × 12050km at 63.4° inclination). It provided measurements of solar cell degradation which could be compared with the predictions of various models.

Jenkins et al. (2014) compared normalised remaining output power P/P_0 after two years in orbit to model predictions, and results are shown in Figure 1.22. Figure 1.22 shows the actual degradation data from Tacsat-4's BTJM solar cells (black), alongside the degradation calculated using the measured fluence spectrum (purple). The agreement of these two values implies that the calculation of remaining power is subject only to a small error when the spectrum of incident flux is correctly predicted. Calculations of P/P_0 based on the predictions of various radiation belt

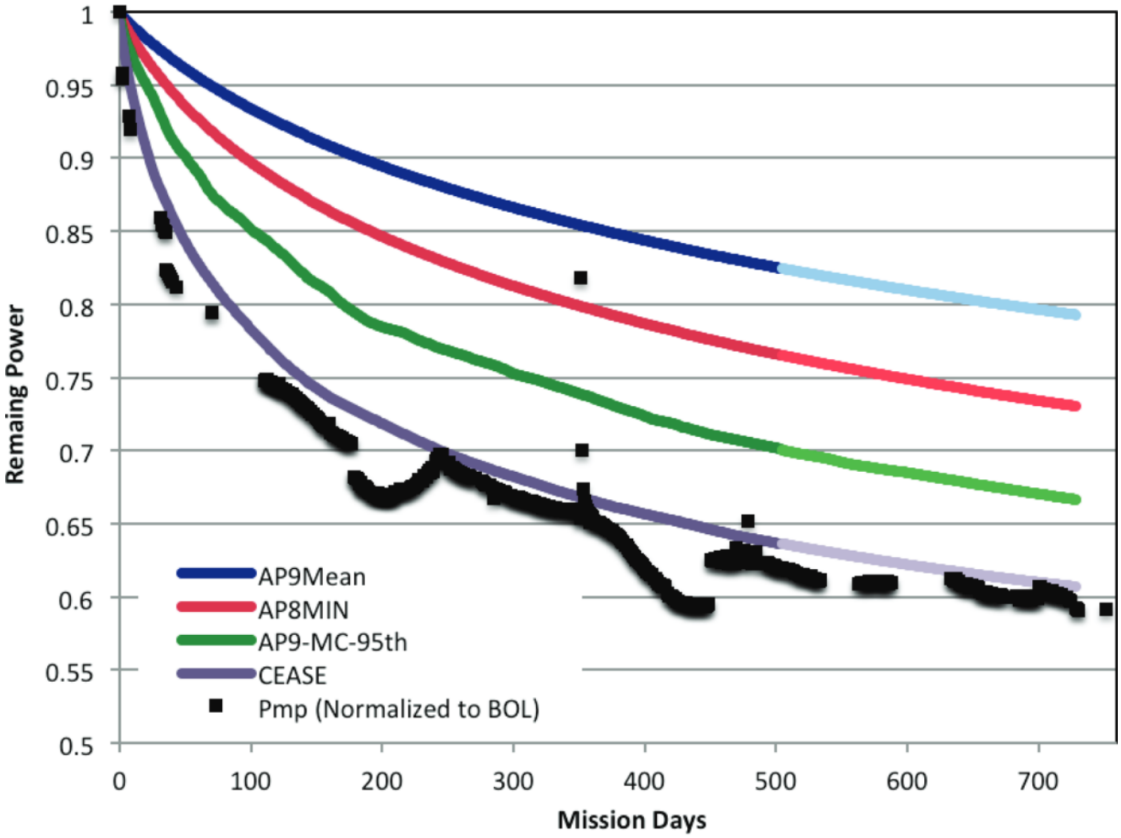


Figure 1.22: Remaining power fraction of the Tacsat-4 BTJM solar cells (black) compared with the values calculated using the observed fluence spectrum of the onboard CEASE spectrometer (purple), and the predictions of various radiation belt models before launch, from Jenkins et al. (2014)

models are shown alongside (green, red and blue). Measured P/P_0 is >10% lower than predicted with AP-8, and ∼20% lower than predicted with AP-9 Mean. These results imply that each model under-predicted proton belt flux significantly, and demonstrate in general that statistical radiation belt models have the potential to under-predict degradation when protons are the dominant contributor. Tacsat-4 data has since been used to update AP-9 in version 1.20 (Johnston et al., 2015).

Chapter 2

Solar Cell Degradation during Electric Orbit Raising to GEO

This chapter is based on a research article:

Solar Cell Degradation due to Proton Belt Enhancements During Electric Orbit Raising to GEO

Space Weather, July 2019, Volume 17, Issue 7

https://doi.org/10.1029/2019SW002213

Alexander R. Lozinski[ab], Richard B. Horne[a], Sarah A. Glauert[a], Giulio Del Zanna[b], Daniel Heynderickx[c], Hugh D. R. Evans[d]

[a]British Antarctic Survey, Cambridge, UK
[b]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK
[c]DH Consultancy BVBA, Leuven, Belgium
[d]European Space Agency/European Space Research and Technology Centre, Noordwijk, Netherlands

Satellites intended for geostationary earth orbit (GEO) can now be fitted with all-electric propulsion, which enables lower-cost access to space by replacing chemical propellant and reducing wet mass. This type of mission involves "electric orbit raising" (EOR), whereby the satellite's electric thrusters are used to raise the satellite from its initial geostationary transfer orbit (GTO) to GEO. However, the electric thrusters exert a smaller impulse per manoeuvre and the raising process

therefore takes ∼200 days, in comparison to just a few days for chemical propulsion (Horne and Pitchford, 2015). During this time, an EOR orbit involves multiple passes through the radiation belts, and has been shown to significantly increase non-ionising radiation dose from trapped protons (Messenger et al., 2014). Electric orbit raising to GEO therefore provides a useful case study to understand the radiation risks that proton belt enhancements pose to orbiting spacecraft. Furthermore, EOR was first performed using fully electric commercial GEO satellites in 2015 (see review by Lev et al., 2019), and it is important to understand the potential penalty of this fairly new technique in terms of radiation exposure. In this chapter, the risk to EOR missions posed by dynamic enhancements in proton flux is investigated by calculating solar cell degradation over the course of an EOR mission during both active and quiet times, using an environment model based on observations by the CRRES satellite. The contribution of electron flux to solar cell degradation is also taken into account in order to compare.

2.1 Modelling an Enhanced Environment

As shown in Figure 1.13, measurements from the Combined Release and Radiation Effects Satellite (CRRES) captured a large enhancement coinciding with the arrival of a SSC on 24th March 1991. The 24th March 1991 storm is one of the largest SEP trapping events for which equatorial observations of protons are available, and the enhancement in proton belt flux following the storm was sustained for at least six months. CRRES's onboard Proton Telescope (PROTEL) and High-Energy Electron Fluxmeter (HEEF) measured 1 - 100MeV protons and 1 - 10MeV electrons throughout its elliptical orbit (350 × 33000km) at 18° inclination. PROTEL data have previously been used to construct the time-averaged CRRESPRO proton belt model (Gussenhoven et al., 1993; Meffert and Gussenhoven, 1994), which includes Quiet and Active versions corresponding to conditions averaged over ∼ 200 days before and after the March 1991 storm. The difference between Quiet and Active versions therefore gives an example of the variation in trapped proton flux associated with a large enhancement.

In order to assess the impacts of an enhanced proton belt on solar array degradation, these two models have been used to compare degradation in a quiet

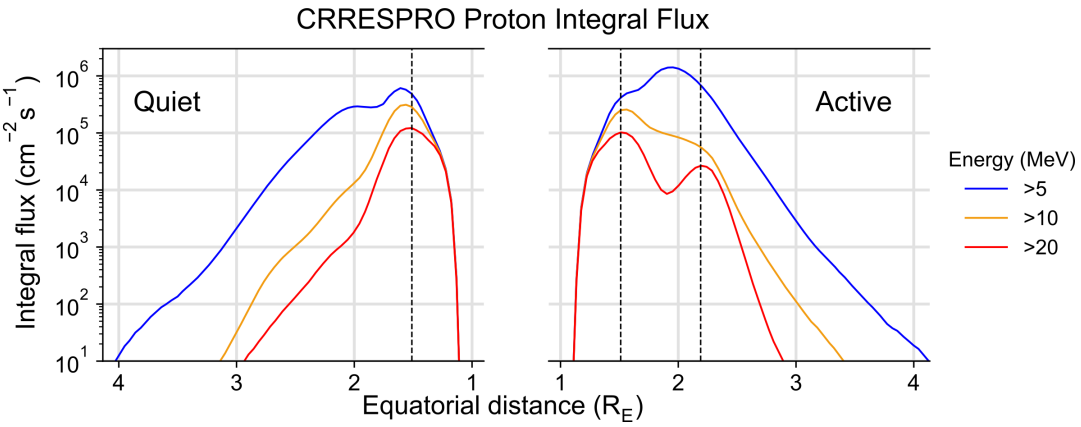


Figure 2.1: Proton integral flux on the geographic equator according to CRRESPRO Quiet (left panel) and Active (right panel). Integral flux is shown for: >5MeV (blue), >10MeV (yellow) and >20MeV (red). Vertical dashed lines represent the peaks of >20MeV flux, showing two peaks in active conditions but only one peak in quiet conditions.

environment (specified by CRRESPRO Quiet) with degradation in an active environment (specified by CRRESPRO Active). Both environments are shown in Figure 2.1 in terms of integral flux. Above 20MeV (red curve in Figure 2.1), the active state exhibits a second peak at 2.2 R_e due to newly trapped SEPs. At lower energies above 5MeV (blue curve), high integral flux ($\gtrsim 10^5 cm^{-2} s^{-1}$) persists until $L \sim 2.5$ in both quiet and active states.

Trapped electrons also contribute towards solar cell damage (Hands et al., 2018). However, the electron model (CRRESELE; see Brautigam and Bell, 1995) corresponding to CRRESPRO does not include electrons at $L < 2.5$ due to potential contamination of the data. This excludes a potentially important region for EOR, therefore CRRESELE was not used to measure the electron contribution. In the interest of understanding the final electron contribution to non-ionising dose, proton and electron damage are later compared when proton and electron environments are specified at an equivalent level of activity, using the AP-8/AE-8 MAX and AP-9/AE-9 statistical model pairs. This gives a sense of the extra contribution to non-ionising dose missing from the calculations that include only CRRESPRO.

Satellite	Launch date (D/M/Y)	Launch site	Designation	EOR duration (days)
SES-15	18/05/2017	Guiana Space Centre (5.2°N)	EOR-1	185
ABS-2A	14/06/2016	Kennedy Space Centre (28.6°N)	EOR-2	200
SES-14	25/01/2018	Guiana Space Centre (5.2°N)	EOR-3	188

Table 2.1: Summary of Electric Orbit Raising Trajectories used

2.2 Satellite Trajectories

The calculation of optimal trajectories for EOR/low-thrust transfers has been a topic of active research since the mid-1970s (Messenger et al., 2014). A particularly significant challenge for the optimisation process is taking into account diminishing thrust due to power loss (addressed recently by Kluever and Messenger, 2019). To illustrate the range of approaches used so far, three trajectories are considered based on previous EOR missions to geostationary orbit. The missions represented are: SES-15, launched on 18th May 2017 from Guiana Space Centre (5.2°N); ABS-2A, launched on 14th June 2016 from Kennedy Space Centre (28.6°N); and SES-14, launched on 25th January 2018 from Guiana Space Centre. These scenarios are hereby referred to as EOR-1, EOR-2, and EOR-3 respectively and summarised in Table 2.1. A measure of the EOR duration (in terms of trajectory) has been specified for each mission in Table 2.1 (column 5), corresponding to the day on which each satellite's longitude stopped increasing, and the satellite became parked at GEO altitude.

Each trajectory has been extrapolated from two line element (TLE) orbit data. The PyEphem library (Rhodes, 2011) was used to convert TLE data to position at a given time, using the most recently available TLE. The spacecraft position was extracted at regular intervals for 200 days, starting from the time of the first TLE available after launch. This encompassed the EOR period for all three trajectories. The time step used to sample position throughout the 200 days was chosen based on the satellite's instantaneous speed at each previous position, such that the spacecraft traversed $\sim$ 600km or less between samples. As the spacecraft's velocity

was higher at perigee than at apogee, this method was used to ensure a consistent spatial resolution.

The trajectories produced as a result of the above process are shown in Figure 2.2 on the X-Y geographic equatorial plane. The locations of the two peaks in >20MeV proton integral flux from the active environment model (shown in Figure 2.1) are indicated by two solid black concentric circles to show the location of the proton belt. A key characteristic of each orbit shown by Figure 2.2 is the initial apogee of the satellite as it enters a geostationary transfer orbit (GTO) after launch. The EOR-2 scenario (second panel) has an apogee that extends well past geostationary orbit altitude, whereas the apogee of EOR-1 (first panel) remains well within. From higher apogee, it is theoretically possible to raise perigee by a higher amount for a manoeuvre exerting the same impulse, potentially enabling a faster rate of raising.

2.3 Calculating Non-ionising Dose and Degradation

To calculate proton non-ionising dose, the total fluence was first calculated along each trajectory using the CRRESPRO Quiet or Active model via the European Space Agency's Spenvis interface (Heynderickx et al., 2005). Non-ionising dose was then calculated in terms of the displacement damage dose, D_d, using the MC-SCREAM tool (Messenger et al., 2010). This parameter is derived according to the method developed by the US Naval Research Laboratory (Messenger et al., 2001). The effect of shielding was taken into account by the integrated MULASSIS transport code (Lei et al., 2002).

To calculate total dose as a function of time in Spenvis, the total fluence spectrum was re-calculated for every day of EOR by repeatedly uploading larger and larger fractions of the 200 day trajectory. The calculation of D_d was repeated for each daily total fluence spectrum using MC-SCREAM. This method resulted in lower error compared to summing D_d values obtained for incremental fluence spectrums.

For each trajectory and both environments, the calculation of dose through

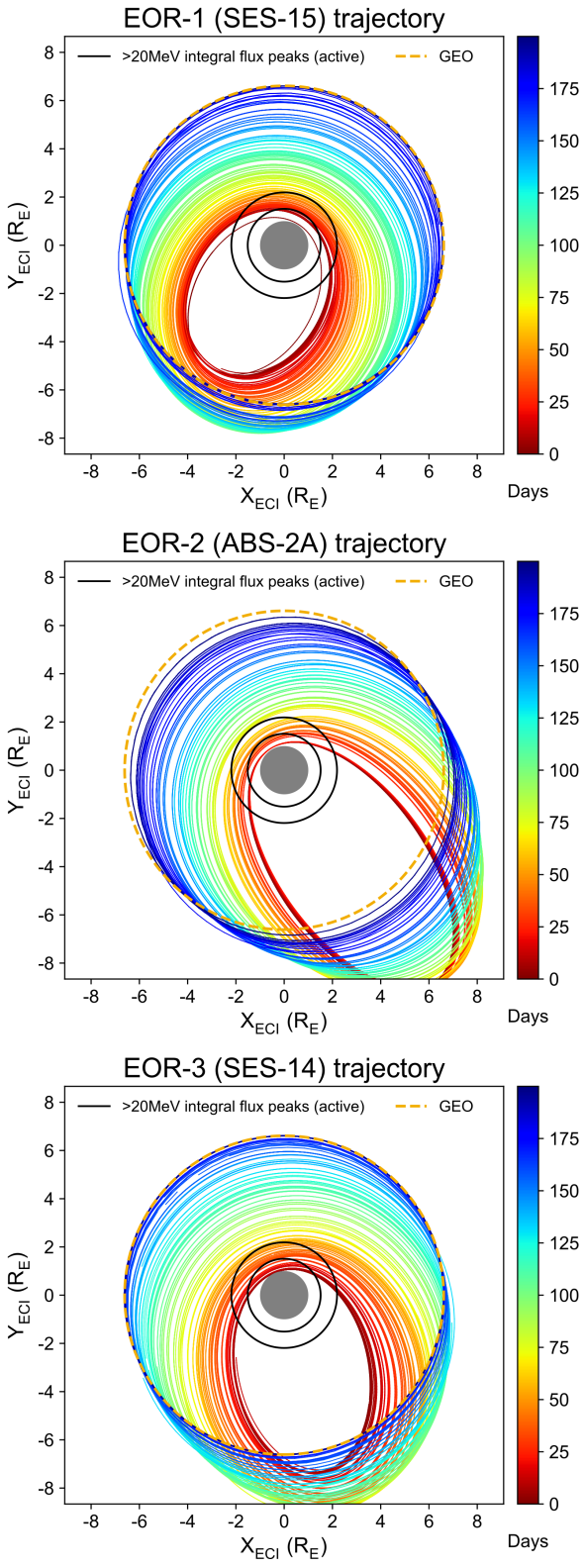


Figure 2.2: Trajectories of EOR-1 (SES-15), EOR-2 (ABS-2A), and EOR-3 (SES-14) over the first 200 mission days. The colour coding indicates the number of days after launch. Peaks in >20MeV integral flux during the active environment, shown in Figure 2.1, are indicated by black concentric rings for reference. GEO orbit is indicated by a yellow dashed line.

time was repeated for three thicknesses of solar cell coverglass: 100μm, 150μm, and 200μm. This range includes levels of shielding previously used on satellites in highly exposed orbits such as Tacsat-4 (150μm), as well as for solar panels qualified for GEO (typically 150μm or less: Messenger et al., 2006; Spectrolab, 2010). Material density was kept constant with a typical value of $2.6 gcm^{-3}$ (Qioptiq Space Technology, 2015). Calculations assumed that the back side of the solar cell was perfectly shielded. This may imply a small underestimate in dose depending on panel structure, estimated to be less than 5% at 200 days for any EOR scenario.

Each value of D_d was converted to the ratio of remaining solar cell output power relative to beginning of life: P/P0. This photovoltaic parameter indicates the level of corresponding solar cell degradation and is a key indicator of remaining lifetime. P/P0 is given by the characteristic equation:

$$\frac{P}{P_0} = 1 - C\log\left(1 + \frac{D_d}{D_{pX}}\right) \tag{2.1}$$

where C and D_{pX} are experimentally-determined fitting parameters for power degradation. This equation is a specific case of Equation 1.66 applied to remaining power. Conversion from D_d to P/P0, for all calculations herein, was based on the characteristics of an Azur Space 3G30 triple junction solar cell, representative of current-generation technology. The C and D_{pX} parameters for this solar cell are built into Spenvis (0.306 and 3.63×10^9 respectively for proton-equivalent D_d).

To investigate the balance between solar cell damage caused by protons and electrons at the end of EOR, the displacement damage dose from each species was calculated at day 200 for EOR-1, 2 and 3. A similar process was followed for each trajectory and coverglass thickness, but instead using the AP-8/AE-8 MAX and AP-9/AE-9 models to specify the same level of activity in proton and electron environments. AP-9/AE-9 models were run in percentile mode using the 95% setting. The effects of local time variation have not been included, having been found to cause a negligible difference for the orbits used.

2.4 Results

Figure 2.3 shows the displacement damage dose (D_d, top panels) and degradation in remaining power output (P/P0, bottom panels) as a function of time for all three EOR trajectories, for both quiet (left panels) and active (right panels) environments using $150\mu m$ coverglass thickness. Only the first 100 days are shown, after which there was no significant increase in degradation until the end of EOR. This is primarily because the perigee of all three satellites had increased to a region of lower proton flux beyond $3R_e$. Figure 2.1 shows that an altitude of $3R_e$ corresponds to a drop by over 2 orders of magnitude in >5MeV integral flux compared to any value between 1.5 - $2R_e$, for both quiet and active environments.

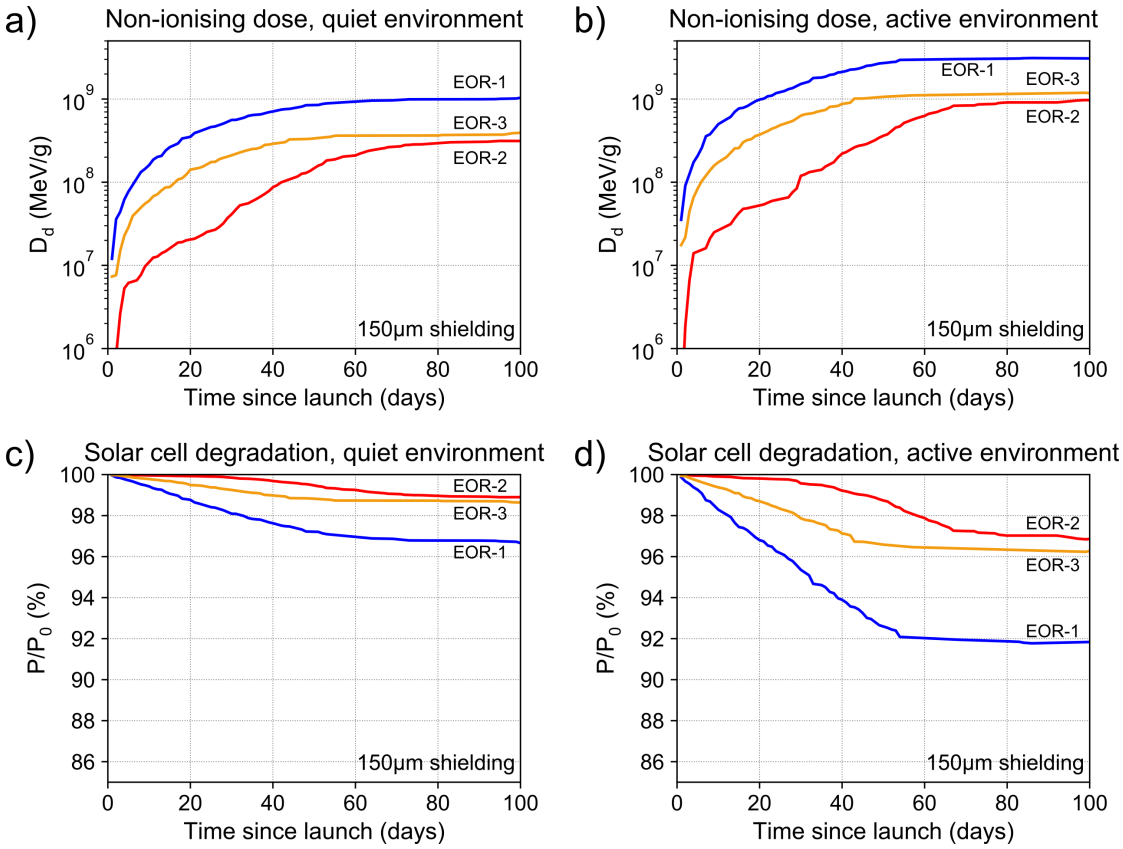


Figure 2.3: Displacement damage dose, D_d, (top panels) and remaining power, P/P0, (bottom panels) for three EOR trajectories for quiet (left panels) and active (right panels) conditions. Coverglass thickness is kept constant at 150μm.

For the quiet environment, Figure 2.3a shows that for EOR-2 (red curve) only minimal dose was accrued before day 30, after which dose rose steadily until around day 80. Figure 2.3c shows this resulted in a $\sim 1\%$ drop in P/P0 after 100 days. In contrast, for EOR-1 (blue curve), the dose accumulated for the most part between days 1-60, leading to a final power loss of $\sim 3\%$.

For the active environment, Figure 2.3b shows the dose-time curves have a very similar shape to those for the quiet environment (panel a); dose increased primarily between days 30-80 for EOR-2 and days 1-60 for EOR-1/EOR-3. However, Figure 2.3d shows that the largest power drop in the active environment, occurring for EOR-1 (blue curve), was about 8%. This is somewhat larger than the 3% drop in power for quiet conditions (Figure 2.3c, blue curve). Both EOR-2 (red curve) and EOR-3 (amber curve) also show a larger drop in power during active conditions (3% and 4% respectively, up from $\sim 1\%$).

To demonstrate the importance of solar cell coverglass shielding, the analysis of solar cell degradation in the active environment was repeated for three different thicknesses. Figure 2.4 shows P/P0 for all three EOR trajectories in an active environment, for coverglass thicknesses of 100, 150, and $200\mu m$. The results show that for EOR-1, as the coverglass thickness is increased from $100\mu m$ to $150\mu m$ to $200\mu m$, the amount of degradation is reduced, and the remaining power increases from 85% to 92% to 95%. A similar trend is shown for EOR-2 (94% to 97% to 98%) and EOR-3 (92% to 96% to 98%), although the effect is more pronounced for the EOR-1 trajectory with a higher exposure to non-ionising dose.

Figure 2.4 shows that when only $100\mu m$ thick coverglass is used, up to 15% degradation in power output can occur due to non-ionising dose from energetic protons during the first 100 days of electric orbit raising. However, the above results do not include the added contribution towards non-ionising dose from energetic electrons, or other effects such as coverglass darkening, and therefore are a lower limit on the power reduction.

To understand the extra impact from trapped electrons, the non-ionising dose caused by the presence of both species was calculated for the EOR-1, EOR-2, and EOR-3 scenarios using environments specified by the AE-8/AP-8 MAX models, and again using the AE-9/AP-9 models at 95% percentile setting. Table 2.2 shows the non-ionising dose caused by protons and electrons separately after 200 days

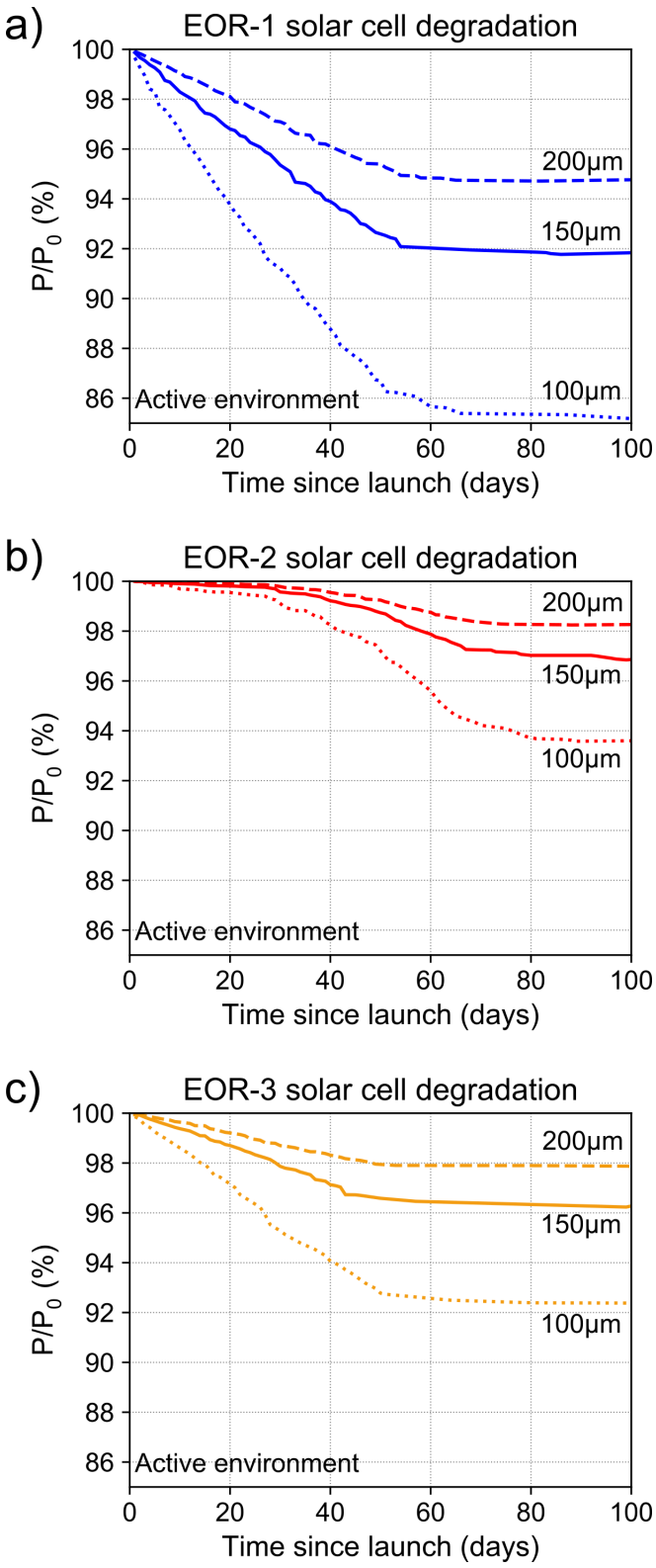


Figure 2.4: Solar cell degradation for all three EOR trajectories, calculated for different coverglass thicknesses in the active, post-storm environment.

Orbit	Shielding (µm)	Proton D_d (MeV/g)	Electron D_d (MeV/g)	Electron D_d frac.	P/P_0 (excl.→ inc. electrons)
	100	$1.05 \pm 0.04 \times 10^{10}$	$1.63 \pm 0.08 \times 10^{8}$	1.5%	81.9 → 81.8%
EOR-1	150	$4.59 \pm 0.21 \times 10^{9}$	$1.36 \pm 0.04 \times 10^{8}$	2.9%	89.1 → 88.9%
	200	$2.79 \pm 0.18 \times 10^{9}$	$1.28 \pm 0.04 \times 10^{8}$	4.4%	92.4 → 92.2%
	100	$4.55 \pm 0.34 \times 10^{9}$	$7.93 \pm 0.18 \times 10^{7}$	1.7%	89.2 → 89.1%
EOR-2	150	$1.67 \pm 0.12 \times 10^{9}$	$7.24 \pm 0.35 \times 10^{7}$	4.2%	95.0 → 94.8%
	200	$8.32 \pm 0.49 \times 10^{8}$	$6.05 \pm 0.15 \times 10^{7}$	6.8%	97.3 → 97.1%
	100	$4.98 \pm 0.26 \times 10^{9}$	$8.66 \pm 0.26 \times 10^{7}$	1.7%	88.5 → 88.4%
EOR-3	150	$1.91 \pm 0.19 \times 10^{9}$	$7.91 \pm 0.25 \times 10^{7}$	4.0%	94.4 → 94.2%
	200	$9.68 \pm 0.47 \times 10^{8}$	$7.67 \pm 0.36 \times 10^{7}$	7.3%	96.9 → 96.6%

Errors shown relate to the calculations performed using the MC-SCREAM tool

Table 2.2: EOR Non-ionising Dose After 200 Days with AE-8/AP-8 MAX

according to the AE-8/AP-8 MAX model. Table 2.3 shows the AE-9/AP-9 95% equivalent. The D_d caused by electrons is shown in terms of proton-equivalent dose so that it can be directly compared. In the fifth column of Table 2.2 and 2.3, the proportion of non-ionising dose attributed to electrons is shown as a percentage. The sixth column of Table 2.2 and 2.3 shows P/P0 before and after taking the electron contribution into account.

Table 2.2 shows that according to AE-8/AP-8 MAX, electrons make up 1.5% to 7.3% of the total non-ionising dose depending on coverglass and trajectory. Table 2.3 shows that according to AE-9/AP-9 95%, the contribution is slightly higher, from 2.3% to 13.1%. Although electrons may therefore cause up to ∼10% of the total displacement damage dose during EOR, the sixth column in Tables 2.2 and 2.3 show that the additional contribution causes a very small decrease in remaining power (less than 1%). This is because P/P0 falls off logarithmically with dose according to Equation 1. Therefore, when the dose from protons is already high, extra exposure to electrons over the EOR duration causes a minor change.

Table 2.2 and Table 2.3 also show that the relative contributions from protons and electrons to total dose during EOR depends on coverglass thickness, with the electron contribution (fifth column) becoming more significant as coverglass thickness increases. This is because non-ionising dose from protons decreases by

Orbit	Shielding (µm)	Proton D_d (MeV/g)	Electron D_d (MeV/g)	Electron D_d frac.	P/P_0 (excl.→ inc. electrons)
EOR-1	100	$1.32 \pm 0.04 \times 10^{10}$	$4.30 \pm 0.34 \times 10^{8}$	3.2%	79.6 → 79.3%
	150	$5.38 \pm 0.30 \times 10^{9}$	$3.32 \pm 0.14 \times 10^{8}$	5.8%	87.9 → 87.4%
	200	$2.54 \pm 0.12 \times 10^{9}$	$3.02 \pm 0.12 \times 10^{8}$	10.6%	93.0 → 92.3%
EOR-2	100	$5.65 \pm 0.27 \times 10^{9}$	$2.08 \pm 0.10 \times 10^{8}$	3.6%	87.5 → 87.2%
	150	$1.69 \pm 0.08 \times 10^{9}$	$2.06 \pm 0.14 \times 10^{8}$	10.9%	94.9 → 94.4%
	200	$1.10 \pm 0.11 \times 10^{9}$	$1.66 \pm 0.09 \times 10^{8}$	13.1%	96.5 → 96.0%
EOR-3	100	$1.05 \pm 0.07 \times 10^{10}$	$2.42 \pm 0.12 \times 10^{8}$	2.3%	81.9 → 81.7%
	150	$3.07 \pm 0.14 \times 10^{9}$	$2.31 \pm 0.19 \times 10^{8}$	7.0%	91.9 → 91.4%
	200	$1.26 \pm 0.10 \times 10^{9}$	$1.81 \pm 0.06 \times 10^{8}$	12.6%	96.0 → 95.6%

Errors shown relate to the calculations performed using the MC-SCREAM tool

Table 2.3: EOR Non-ionising Dose After 200 Days with AE-9/AP-9 95%

75% or more in each scenario when coverglass is increased from 100μm to 200μm (third column of Table 2.2 and Table 2.3). In contrast, the decrease in non-ionising dose from electrons is about 30% at most (fourth column of Table 2.2 and Table 2.3). A reason for this result is that increasing coverglass, in general, absorbs more of the low energy portion of the incident spectrum. For electrons, unlike protons, it is the high energy particles that do more damage per collision, and these fluxes are less affected (Messenger et al., 2001).

2.5 Discussion

2.5.1 The Influence of Orbit

The dose-time curves in Figure 2.3a and b show that after just 20 days, the solar arrays on EOR-1 had accrued a similar level of non-ionising dose to those on EOR-2 and EOR-3 after 100 days. It is important to understand why this occurred in order to be able to avoid such damage when possible. To investigate, the total time spent in each 130km wide bin from R = 1 to 3R_e has been plotted, where R is the distance from the centre of Earth to satellite. This bin sizing was found to be the most effective at highlighting certain features of each orbit discussed later. The

time spent inside a bin after one pass is given by

$$\Delta t_b = \int_{R_{enter}}^{R_{leave}} \frac{\mathrm{d}t}{\mathrm{d}R}\,\mathrm{d}R \tag{2.2}$$

where R_{enter}, R_{leave} are the distances at which the trajectory enters, leaves bin b respectively. Each contribution given by Equation 2.2 has been summed over every pass through the bin to find the total time spent by the satellite within 65km of the bin centre. Figure 2.5 shows the total time spent in each bin after the first 10, 20, 40, and 60 days (panels a to d) of electric orbit raising for each scenario (left ordinate). The >5MeV integral flux given by the CRRESPRO quiet model is also shown for the same range of R (right ordinate).

Several peaks are apparent for each trajectory in Figure 2.5. These peaks occur due to the shape of the orbit: at perigee, dR/dt is small despite the satellite having higher overall speed, and fewer radial bins are crossed. Therefore, more time is accumulated in R bins near to orbit perigee. Conversely, away from perigee, the satellite passes quickly across different radii whilst changing altitude, spending less time in each bin. The peaks in Figure 2.5 therefore indicate the location of orbit perigee.

As orbit raising progresses for each satellite, the altitude of perigee is slowly increased by manoeuvres. This is shown by the continuous addition of peaks in Figure 2.5 as the mission progresses from 10 to 60 days (panels a through d). The perigee at later times in each panel, indicated by the rightmost peak, is indicated by an arrow for each satellite.

Panel a shows that after only 10 days, the perigee of EOR-1 (blue curve) is raised up to R $\sim 1.3R_e$, marked by the blue arrow. As this is a region of high >5MeV flux (dotted black line), EOR-1 begins accruing a large dose almost immediately. In contrast, the perigee of both EOR-2 (red curve) and EOR-3 (amber curve) remained low within this time period, indicated by the red and amber arrows.

Figure 2.5, panel b shows that even after 20 days in orbit, the perigee of EOR-2 (red arrow) did not increase. This post-launch period of a few weeks was instead used to reduce inclination, bringing the spacecraft into the equatorial plane. Therefore, EOR-2 stayed in its initial high-apogee GTO orbit, accruing very little dose because its perigee was beneath the region of high flux. Following this, EOR-2

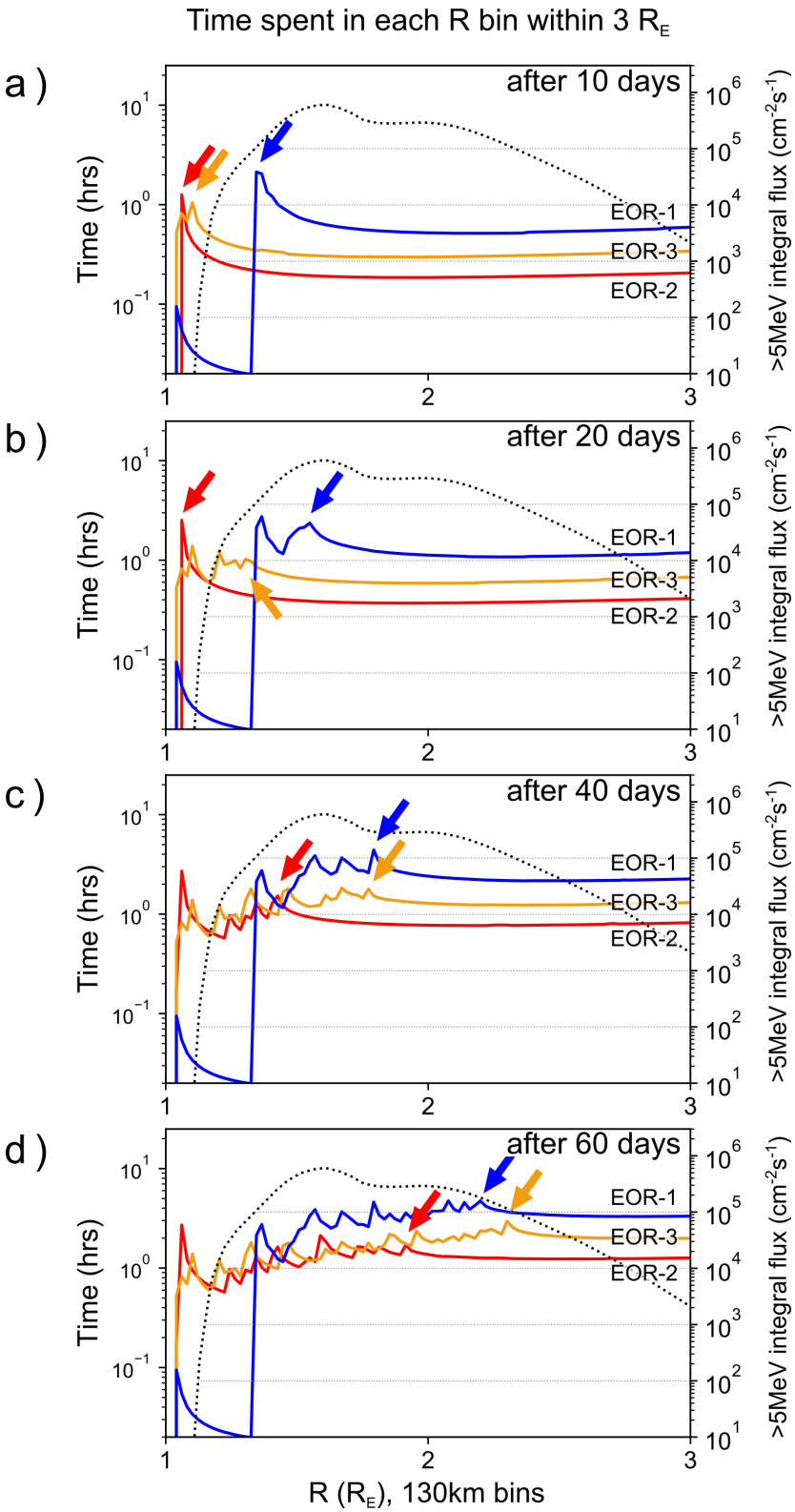


Figure 2.5: Total time spent in each 130km-wide radius bin between RE = 1 to 3 for each orbit, after: (a) 10 days; (b) 20 days; (c) 40 days; and (d) 60 days. Coloured arrows highlight the location of orbit perigee at each of these times for each satellite. The >5MeV integral flux according to the quiet time CRRESPRO model is also included (right hand scale).

began raising through the proton belt at a similar rate to EOR-3, indicated by the progression of the red arrow from day 20 to day 60 (panel b to d). This is faster than the rate at which EOR-1 perigee was raised, with EOR-3 perigee overtaking EOR-1 perigee by day 60 (panel d). Figure 2.5 thus shows that after 60 days, EOR-1 (blue curve) has spent more time within the region of high ($\gtrsim 10^4 cm^{-2}s^{-1}$) >5MeV flux, leading to a higher fluence and associated drop in P/P0. This is because it was placed into the high flux region early, then raised slowly. In contrast, EOR-2 (red curve) and EOR-3 (amber curve) were able to traverse the region of high >5MeV trapped flux quickly. This demonstrates the utility of using an initial GTO with a high apogee that, in general, would allow faster raising of perigee. The small amount of dose EOR-2 accumulates within the first 20 days also shows the advantage of having a perigee beneath the proton belt when in GTO. This highlights the importance of modelling the location of innermost trapped flux accurately, shown by the steep gradient in the dotted line on the left of Figure 2.5, in order to understand exposure at low perigee.

2.5.2 Dependence of Dose on Shielding and Energy

For protons impacting solar cells without shielding, the highest damage per collision is caused by sub-MeV particles. This energy dependence is described both by experimentally-determined relative damage coefficients and by calculated non-ionising energy loss coefficients (Messenger et al., 2001). Figure 2.4 shows that when coverglass thickness is increased, power loss is reduced. This demonstrates a change to the spectrum of particles after they have traversed the coverglass, caused by their initial energy being reduced. The three EOR scenarios together with the proton environment observed by CRRES provide an opportunity to test the dependence of non-ionising dose on the energy of incident flux before it impacts shielding.

To investigate, calculated the average differential flux spectrum for EOR-2 has been calculated after the first 100 days of EOR in the active environment. The spectrum was then modified by setting the flux to zero above a 'cut-off energy', and using MC-SCREAM to compute total dosage. Figure 2.6 shows that when the spectrum was set to zero above 3MeV for $100\mu m$ coverglass, the dose was zero (red

curve). However, when the cutoff was increased to 7MeV, the dose was the same as for the unmodified spectrum. Thus for $100\mu m$, the important energy range is 3 to 7MeV. As the coverglass thickness was increased, the important energy range was shifted up to between 3 to 10MeV for $150\mu m$, and 5 to 10MeV for $200\mu m$. Therefore, for the range of shielding considered, only protons between 3 and 10MeV make a significant contribution towards non-ionising dose. The dose imparted to shielded devices by different energies in an incident proton spectrum is addressed by Messenger et al. (1997), using a simulated solar proton event, and by Summers et al. (1997), for simulated inclined circular orbits. Results from these analyses show similar features, whereby a large fraction of total non-ionising dose can be attributed to a small energy range around 1 to 10MeV.

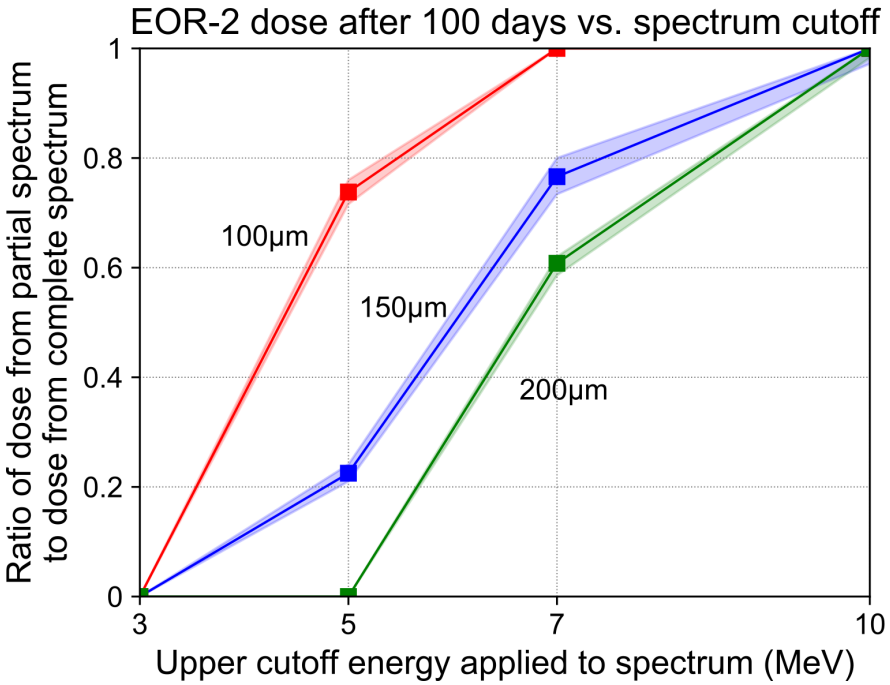


Figure 2.6: Ratio of total dosage calculated using a 100 day average differential flux spectrum for EOR-2, cut off after a certain energy, to the same calculation using the complete spectrum. The coloured borders indicate error in the MULASSIS transport code calculation but do not take into account other limitations such as radiation belt model error or interpolation in energy.

A comparison between dose-time curves in the quiet and active environments

(Figure 2.3a and b) shows a similar time period over which dose increases. Given that only 3 - 10MeV flux contributes significantly, the reason for this similarity is explained by comparing quiet and active conditions at this energy. Figure 2.1 shows that high values of >5MeV integral flux ($\gtrsim 10^5 cm^{-2}s^{-1}$) persist until $R \sim 2.5R_e$ in both quiet and active times. Therefore, fluxes near 3 - 10MeV do not show a significant change in radial extent between active and quiet conditions, and degradation continues for roughly the same amount of time until the satellite reaches higher altitude. The addition of a >20MeV flux peak within the slot region does not cause prolonged degradation in active times because the contribution from high energy towards non-ionising dose is low.

Figure 2.4 shows how power loss through time can be reduced with the application of thicker coverglass. An interesting result of increasing coverglass thickness is the shortening of the time window over which power loss occurs. For example, EOR-2 (middle panel) stops accumulating dose after $\sim$ 70 days for $200\mu m$ compared with $\sim$ 90 days for $100\mu m$, with a subtle but similar effect seen for EOR-1 and EOR-3. As the coverglass thickness is increased, the damaging part of the spectrum is shifted to higher energies within the 3 to 10MeV range (Figure 2.6), but as Figure 2.1 shows, higher energy fluxes tend to fall off more rapidly with radial distance. This reduces the time window during which the solar panel is subject to significant non-ionising dose.

The significance of 3 to 10MeV environmental fluxes also highlights the importance of understanding the physical processes behind enhancements in this energy range. The radial transport of trapped particles to low L (in both a diffusive and shock-induced manner) is associated with significant energisation. Therefore, SEPs contributing to the formation of $\lesssim$10MeV enhancements may enter the magnetosphere at considerably lower energy. Particle tracing simulations by Richard et al. (2002) show that entry of SEPs to the magnetosphere near 1MeV can be via the cusp or flank regions and is strongly influenced by IMF orientation, whereas direct entry and trapping through the front-side magnetopause generally applies to SEPs at $\gtrsim$10MeV. The various types of entry may therefore complicate modelling of source populations.

Although thicker coverglass is useful to reduce non-ionising dose, this is one area where various engineering and cost requirements may take priority. For example,

there may also be a knock-on effect due to the increase in mass of thicker coverglass, such as the need to upgrade solar array drive mechanisms and structure. In this vein, it is useful to consider that increasing coverglass thickness tends to follow a law of diminishing returns, in terms of reducing total solar panel degradation during EOR, as demonstrated in Figure 2.4 for all three missions.

2.6 Conclusions

In this chapter, an analysis of non-ionising dose from trapped protons accrued over 200 days during the course of electric orbit raising has been presented. Results show the variability caused by realistic changes to environment, trajectory and coverglass thickness. Several key conclusions can be drawn from this work, numbered below.
1) For a typical coverglass thickness of $150\mu m$, launching in an active environment can increase solar cell degradation due to trapped protons by 2 to 5% before start of service compared to a quiet environment depending on trajectory. These values are in terms of remaining power, normalised to beginning of life.
2) The crucial energy range for enhancements in proton flux is 3 - 10MeV for solar cells with a level of shielding between 100-200μm.
3.) For a coverglass thickness of $150\mu m$, solar cell degradation in an active environment can vary by $\sim$5% for different EOR scenarios. In the EOR-1 scenario, solar cell degradation in an active environment can vary by $\sim$10% based on the choice of coverglass thickness between 100-200μm. For EOR-2, this variation is around 5%.
4.) In the worst case tested (active environment, $100\mu m$ coverglass, EOR-1), degradation of up to 15% is possible within the EOR period, before taking into account other effects such as electron dose.
5.) In addition to the degradation caused by trapped proton flux, evaluations of non-ionising dose at the end of the EOR period indicate an extra contribution from trapped electrons. This contribution is on the order of $\sim$ 10% or less in terms of the displacement damage dose. However, due to the dose already being high and because power decreases logarithmically with dose, adding this contribution has only a minor effect (less than 1%) on the remaining power predicted at the end of the raising period. This also casts estimates of further solar cell degradation due to trapped electrons at GEO in a new light, since operators will need to recalibrate

P/P0 estimates based on dose accrued during EOR.
6.) A higher initial orbit apogee generally implies that perigee can be raised faster during EOR, allowing the satellite to skip through the belt in fewer passes, whilst also implying a higher velocity at perigee. These two factors mean less time is spent within the proton belt during raising from a high apogee GTO. It is therefore recommended that EOR missions begin this way, such as in the case of EOR-2 (and EOR-3 to a lesser extent).

Several effects not taken into account in this analysis may reduce the transmission efficiency of solar cell coverglass, causing a further drop in performance. This can occur due to coverglass darkening from radiation damage, as well as the deposition of ions ejected from electric thrusters (Horne and Pitchford, 2015). Arc-induced contamination may also contribute towards solar cell power degradation throughout LEO, MEO and GEO environments depending on the grounding of conducting surfaces (Ferguson et al., 2016).

The importance of considering dynamic enhancements in trapped proton flux suggests a role for physics-based modelling to help assess radiation damage, and address the increasing utilisation of low and medium earth orbits. However, more real-time information is required on the transient nature of the proton belt's outer region to understand these processes. In particular, the demonstrated importance of enhancements near 3 - 10MeV near the equator at low L, which may not show signatures at high latitude due to transport energisation, highlights the need for improved observational capability.

Chapter 3

Constructing a Physics-based Numerical Model

3.1 Describing the Time Evolution of Proton Distributions

To build on the work from Chapter 2, a physics-based model of the proton belt was constructed in order to make theoretical calculations of solar cell degradation and to investigate dynamic variability. For the practical purpose of modelling Earth's proton belt, the phase-averaged distribution of radiation belt protons is considered. The stochastic nature of forces acting on radiation belt particles leads to a time evolution of the distribution that can be described as if subject to:

- diffusion, arising from acceleration by many minute electromagnetic fluctuations;
- friction, arising from systematic deceleration over many small deflections via coulomb collisions; and
- sources or sinks, arising from changes of identity to a particle such as ion pickup in the case of CRAND, or sudden loss of energy in the case of charge exchange.

Early work on the proton belt evaluated the stationary distribution arising from a

balance between the CRAND source and coulomb collisions (for example, Lenchek and Singer, 1962; Dragt et al., 1966), until it was realised that transport by radial diffusion and a source of solar protons were also essential considerations in order to reproduce observations. Eventually, Fokker Planck formulations were invoked to account for each of the three types of processes listed above. This chapter includes a brief review of how the Fokker Planck equation was developed and used to describe nonadiabatic transport of radiation belt protons.

In the 1943 work "Stochastic Problems in Physics and Astronomy", Chandrasekhar considers the Brownian motion of a free particle, with acceleration in the absence of an external field described by the Langevin equation:

$$\Delta \mathbf{u} = -\beta \mathbf{u} \Delta t + \mathbf{B}(\Delta t) \tag{3.1}$$

where $\mathbf{u}$ represents velocity, $-\beta\mathbf{u}$ represents a dynamical friction and $\mathbf{B}(\Delta t)$ represents a net acceleration arising from fast acting and small amplitude fluctuations during an interval Δt. Considering the distribution of particles in velocity space $W(\mathbf{u}, t)$, the Fokker Planck equation is derived in its most general form (see Equation 224, p.33, Chandrasekhar, 1943). As the stochastic differential equation for a radiation belt particle is analogous to Equation 3.1, Davis and Chang (1962) used a one dimensional expansion of the Fokker Planck equation to study the radial transport of particles along the geomagnetic equator due to repeated geomagnetic disturbances in a dipole field. Their equation is

$$\frac{\partial \varphi^*}{\partial n} = -\frac{\partial}{\partial r}\left[D_1\ \varphi^*\right] + \frac{1}{2}\frac{\partial^2}{\partial r^2}\left[D_2\ \varphi^*\right] \tag{3.2}$$

where $\varphi^*(r; n)\mathrm{d}r$ is the number of particles on the geomagnetic equator within $\mathrm{d}r$ of radial distance r after n geomagnetic disturbances, and the two coefficients $D_1 = \langle \Delta r \rangle$ and $D_2 = \langle (\Delta r^2) \rangle$ represent the mean radial displacement and mean square radial displacement caused by the average effect of a storm. This transport equation does not include the effect of collisions, and as such there is no frictional term.

Fälthammar (1965) derived a general formula for the mean square radial displacement term D_2 for an electric potential field disturbance. An equivalent

formula for evaluating the mean radial displacement term D_1 directly was not available, but Davis and Chang (1962) and Nakada and Mead (1965) did derive relations between D_1 and D_2 for electromagnetic perturbations in special cases. However, Dungey (1965) separately derived a transport equation with only one coefficient to encompass the statistical properties of a geomagnetic disturbance. The transport equation of Dungey (1965) was not derived from a Fokker-Planck equation, and suggested D_1 and D_2 must be related more generally in order for the Fokker Planck approach to agree. The explicit relation between D_1 and D_2 was then derived by Fälthammar (1966) from the one dimensional Fokker-Planck equation, reproducing the transport equation derived by Dungey (1965), as well as the results of Davis and Chang (1962) and Nakada and Mead (1965). The relation, from Equation 5 of Fälthammar (1966), is given by

$$D_1 = \frac{r^2}{2}\frac{\partial}{\partial r}\left(\frac{D_2}{r^2}\right) \tag{3.3}$$

and can be substituted into Equation 3.2 to write

$$\frac{\partial \varphi^*}{\partial n} = \frac{1}{2}\frac{\partial}{\partial r}\left[\frac{D_2}{r^2}\,\frac{\partial}{\partial r}\left(r^2\varphi^*\right)\right] \tag{3.4}$$

Farley and Walt (1971) developed the transport equation further by adding additional source and loss terms, and writing it in terms of the distribution function $\bar{f}(\mu, J, L; t)$, where $\bar{f}\mathrm{d}\mu\mathrm{d}J\mathrm{d}L$ gives the number of particles in the interval $\mathrm{d}\mu\mathrm{d}J\mathrm{d}L$. The transport equation thus becomes (Equation 1 of Farley and Walt, 1971):

$$\frac{\partial \bar{f}}{\partial t} = \frac{\partial}{\partial L}\left[\frac{D_{LL}}{L^2}\frac{\partial}{\partial L}\left(L^2\bar{f}\right)\right] + \text{sources} - \text{losses} \tag{3.5}$$

By switching variable from r to L, Farley and Walt's equation describes the transport of particles across drift shells, applying also to particles mirroring away from the equator with non-zero J. The distribution function $\bar{f}$ is valid to describe particles at relativistic energies. Equation 3.5 follows the convention of rewriting in terms of a diffusion coefficient $D_{LL} = D_2/2$. The D_2 coefficient is the same as in Equation 3.4, but because the independent variable has changed from n to t, it is thought of as representing the effect of disturbances over a time interval Δt, which is the time for n to increment.

In more recent works, the transport equation also appears with the following form:

$$\frac{\partial f}{\partial t} = L^2 \frac{\partial}{\partial L}\left[\frac{D_{LL}}{L^2}\frac{\partial}{\partial L}\left(f\right)\right] + \text{sources} - \text{losses} \tag{3.6}$$

The difference between Equation 3.5 and Equation 3.6 (taken from Equation 1 of Claflin and White, 1974) arises because $f \propto L^2\bar{f}$, and one can change the form of Equation 3.5 to that of Equation 3.6 simply by multiplying both sides by L^2 then substituting for f. Both distribution functions are proportional to the number of particles per unit volume of a space defined in adiabatic invariant coordinates. However, the advantage of using f with Equation 3.6 is that the unit volume is effectively transformed to be in terms of the canonical action variables μ, J and Φ. This is because $\mathrm{d}L/\mathrm{d}\Phi \propto L^2$ via Equation 1.49, and so one can consider the number of particles N in unit volumes described by $\bar{f}$ and f as:

$$\begin{aligned} N &\propto \bar{f}\ \mathrm{d}\mu\mathrm{d}J\mathrm{d}L \\ &= \mathrm{d}L/\mathrm{d}\Phi\bar{f}\ \mathrm{d}\mu\mathrm{d}J\mathrm{d}\Phi \\ &\propto L^2\bar{f}\ \mathrm{d}\mu\mathrm{d}J\mathrm{d}\Phi \\ &\propto f\ \mathrm{d}\mu\mathrm{d}J\mathrm{d}\Phi \end{aligned} \tag{3.7}$$

Therefore, as discussed in Section 1.3.1, f is proportional to phase space density $F(\mathbf{x}, \mathbf{p})$. The direction of currents due to radial diffusion are directed away from peaks in phase space density and towards troughs, and this means that a plot of f versus L is simpler to interpret in terms of the expected time evolution.

A general diffusion equation was formulated by Haerendel (1970) which allows for diffusion in more than one invariant. For a distribution function expressed in terms of an arbitrary space $F(X_1, X_2, X_3)$, the diffusive transport equation is:

$$\frac{\partial F}{\partial t} = \frac{1}{\mathcal{J}}\frac{\partial}{\partial X_i}\left(\mathcal{J}D^{ij}\frac{\partial F}{\partial X_j}\right) \tag{3.8}$$

where the Jacobian matrix $\mathcal{J}$ relates each coordinate to the three canonical action variables, given by

$$\mathcal{J} = \frac{\partial\left(\mu, J, \Phi\right)}{\partial\left(X_1, X_2, X_3\right)} \tag{3.9}$$

This equation is useful context to understand the transport term in Equation 3.6, which is an expansion of Equation 3.8 for diffusion in the L coordinate, with $\mathcal{J} \propto L^{-2}$. For modelling the proton belt (within the energy range of interest) only pure third invariant diffusion is required, and the diffusion tensor D^{ij} can therefore be transformed such that it has only one non-zero element, which is on the diagonal. This is not necessarily the case for the electron belt, where other diffusion mechanisms operate.

Farley and Walt (1971) further developed Equation 3.5 to account for atmospheric collisional losses. Coulomb collisions occur over timescales short enough to violate the first invariant, and the resultant energy degradation slightly decreases a particle's μ and J value. Changes to the distribution function at a fixed coordinate manifests as convection, and this effect is included via the loss term appearing in Equations 3.5 and 3.6 above by

$$-\text{losses} = -\frac{\partial}{\partial\mu}\left[\frac{\mathrm{d}\mu}{\mathrm{d}t}_{\text{fric}} f\right] - \frac{\partial}{\partial J}\left[\frac{\mathrm{d}J}{\mathrm{d}t}_{\text{fric}} f\right] \tag{3.10}$$

The quantities $\mathrm{d}\mu/\mathrm{d}t_{\text{fric}}$ and $\mathrm{d}J/\mathrm{d}t_{\text{fric}}$ represent the changes in μ and J at a particular set of coordinates due to the cumulative effect of many small deflections by free and bound electrons in the atmosphere, ionosphere and plasmasphere. However, it can be shown that

$$\frac{\mathrm{d}J}{\mathrm{d}t}_{\text{fric}} = \frac{J}{2\mu}\frac{\mathrm{d}\mu}{\mathrm{d}t}_{\text{fric}} \tag{3.11}$$

and this is demonstrated in Appendix A. The full definition of $\mathrm{d}\mu/\mathrm{d}t_{\text{fric}}$ is given at the end of this section.

Using the form of Fokker Planck equation given by Equation 3.6 with the addition of frictional loss provides the starting point for a comprehensive description. Diffusional and frictional changes to the distribution are accounted for, and the extra sources and sinks (as mentioned in the third bullet point near the beginning of this section) can be included via additional terms to represent changes due to

the CRAND process as well as nuclear scattering. Putting this together, the full equation including these additional terms is:

$$\frac{\partial f}{\partial t} + \frac{\partial}{\partial \mu}\left[\frac{\mathrm{d}\mu}{\mathrm{d}t}_{\text{fric}} f\right] + \frac{\partial}{\partial J}\left[\frac{\mathrm{d}J}{\mathrm{d}t}_{\text{fric}} f\right] = L^2 \frac{\partial}{\partial L}\left[\frac{D_{LL}}{L^2}\frac{\partial f}{\partial L}\right] + S_n - \Lambda f \tag{3.12}$$

where $f\ \mathrm{d}\mu \mathrm{d}J \mathrm{d}L$ gives a quantity proportional to the number of particles in a unit volume of phase space, S_n is the rate of change in f due to CRAND (discussed in Section 3.3), and Λ is a loss term for nuclear scattering with dimensions of inverse time. The full definition of Λ is given at the end of this section. Making use of Equation 3.11, this can be simplified to eliminate $\mathrm{d}J/\mathrm{d}t_{\text{fric}}$, giving the 3D "master equation":

$$\frac{\partial f}{\partial t} + \frac{\partial}{\partial \mu}\left[\frac{\mathrm{d}\mu}{\mathrm{d}t}_{\text{fric}} f\right] + \frac{\partial}{\partial J}\left[\frac{J}{2\mu}\frac{\mathrm{d}\mu}{\mathrm{d}t}_{\text{fric}} f\right] = L^2 \frac{\partial}{\partial L}\left[\frac{D_{LL}}{L^2}\frac{\partial f}{\partial L}\right] + S_n - \Lambda f \tag{3.13}$$

The equation for $\mathrm{d}\mu/\mathrm{d}t_{\text{fric}}$ is

$$\frac{\mathrm{d}\mu}{\mathrm{d}t}_{\text{fric}} = \left\langle \frac{\mathrm{d}T}{\mathrm{d}x} \right\rangle \frac{\left(T^2 + 2E_0 T\right)^{\frac{1}{2}}}{B_e m_0 c} \sin^2\left(\alpha_{eq}\right) \tag{3.14}$$

where $\left\langle \frac{\mathrm{d}T}{\mathrm{d}x} \right\rangle$ is the drift averaged stopping power. The concept of a drift averaged quantity will be explained in Section 3.2. This quantity has contributions due to ions and electrons which are added together. The equation, in SI units, is given by:

$$\begin{aligned}\left\langle \frac{\mathrm{d}T}{\mathrm{d}x} \right\rangle = \frac{4\pi}{m_e v^2}\left(\frac{e^2}{4\pi\epsilon_0}\right)^2 \Big(\langle n_e \rangle \left[\beta^2 - \ln\left(\lambda_D m_e v/\hbar\right)\right] + \\ \sum_{n=1}^{N} \langle n_i \rangle Z_i \left[\beta^2 - \ln\left(\left(\gamma^2 - 1\right) 2m_e c^2 / I_i\right)\right] \Big)\end{aligned} \tag{3.15}$$

where $\langle n_e \rangle$ and $\langle n_i \rangle$ refer to the drift averaged densities of electrons and other constituents i of the atmosphere, ionosphere and plasmasphere that a radiation belt protons interacts with along an orbit. In Equations 3.14 and 3.15, symbol T

has been used for kinetic energy, but the use of E is now resumed.
The equation for Λ is

$$\Lambda = v \sum_i \sigma_i \langle n_i \rangle \tag{3.16}$$

where $\langle n_i \rangle$ is the drift averaged density of constituent i, and $\sigma_i(v)$ is the scattering cross section for a collision between a proton at velocity v and a nuclei of constituent i. For modelling purposes, Equation 3.16 is evaluated over i = H, He, N, O and Ar.

As a final aside, one can see from Equation 3.13 that the rate of loss at any coordinate depends on number density via $\mathrm{d}\mu/\mathrm{d}t_{\mathrm{fric}}$. However by expanding the loss terms of Equation 3.13, one can see that numerous other factors also affect loss rates, such as the local spectrum $\partial f/\partial \mu$:

$$\begin{aligned}
-\text{losses} &= -\frac{\partial}{\partial \mu}\left[\frac{\mathrm{d}\mu}{\mathrm{d}t}_{\text{fric}} f\right] - \frac{\partial}{\partial J}\left[\frac{\mathrm{d}J}{\mathrm{d}t}_{\text{fric}} f\right] \\
&= -\frac{\partial}{\partial \mu}\left[\frac{\mathrm{d}\mu}{\mathrm{d}t}_{\text{fric}} f\right] - \frac{\partial}{\partial J}\left[\frac{J}{2\mu}\frac{\mathrm{d}\mu}{\mathrm{d}t}_{\text{fric}} f\right] \\
&= -\left[\frac{\partial}{\partial \mu}\left[\frac{\mathrm{d}\mu}{\mathrm{d}t}_{\text{fric}}\right] f + \frac{\mathrm{d}\mu}{\mathrm{d}t}_{\text{fric}}\frac{\partial f}{\partial \mu}\right] - \left[\frac{1}{2\mu}\frac{\mathrm{d}\mu}{\mathrm{d}t}_{\text{fric}} f + \frac{J}{2\mu}\frac{\partial}{\partial J}\left[\frac{\mathrm{d}\mu}{\mathrm{d}t}_{\text{fric}} f\right]\right] \\
&= -\left[\frac{\partial}{\partial \mu}\left[\frac{\mathrm{d}\mu}{\mathrm{d}t}_{\text{fric}}\right] f + \frac{\mathrm{d}\mu}{\mathrm{d}t}_{\text{fric}}\frac{\partial f}{\partial \mu}\right] - \left[\frac{1}{2\mu}\frac{\mathrm{d}\mu}{\mathrm{d}t}_{\text{fric}} f + \frac{J}{2\mu}\frac{\mathrm{d}\mu}{\mathrm{d}t}_{\text{fric}}\frac{\partial f}{\partial J} + \frac{J}{2\mu}\frac{\partial}{\partial J}\left[\frac{\mathrm{d}\mu}{\mathrm{d}t}_{\text{fric}}\right] f\right]
\end{aligned} \tag{3.17}$$

Even for the 2D case (with $J = 0$), Equation 3.17 shows that the loss rate has a complicated dependence on number density and f.

Equations 3.13 to 3.15 are fully relativistic and form the basis of the physics-based model developed by the end of this section. Throughout this work, the effort to advance proton belt modelling focused generally on: firstly improving the empirical evaluation of S_n and $\mathrm{d}\mu/\mathrm{d}t_{\mathrm{fric}}$ relative to previous work, and secondly; driving the model with new observational data to investigate uncertainty in the diffusion coefficients.

3.2 How to Compute Drift Averages for a 3D Numerical Model

3.2.1 Definition of a Drift Average

Fundamentally, the need for drift averaging arises because a distribution function describes the phase-averaged intensity of particles, but may depend on a physical quantity A that varies strongly with phase. To allow the distribution function to be calculated, dependence on A is replaced by $\langle A \rangle$, a phase-averaged quantity that approximately accounts for the influence of A on the distribution. $\langle A \rangle$ is given by

$$\langle A \rangle = \frac{\int_{s_1}^{s_2} A \, \mathrm{d}s}{\int_{s_1}^{s_2} \mathrm{d}s} \tag{3.18}$$

where $\mathrm{d}s$ is the path element along a particular trapped proton trajectory S bounded by points s_1 and s_2, which must be solved numerically by integrating the equation of motion of a proton in space. To capture the average variation in A, the integral in Equation 3.18 should be over an integer number of complete drift orbits, and hence $\langle A \rangle$ is known as a drift average.

Quantities such as $\langle n_i \rangle$ in Equation 3.15 are "drift averaged", and so are $\mathrm{d}\mu/\mathrm{d}t_{\mathrm{fric}}$ and Λ in a sense because they depend on $\langle n_i \rangle$. The evaluation of terms in the master equation using drift averages is an essential capability in order to build a competitive numerical model.

3.2.2 High-Level Process Design

The Problem

The master equation of a numerical radiation belt model must be solved on a grid that represents a discretised coordinate space defined with respect to adiabatic invariants ("the solution grid"). Solving a master equation involves evaluating coefficients of its finite difference approximation on the solution grid. As can be seen from the example master equation of Equation 3.13, drift averaged quantities such as $\langle n_i \rangle$ will appear within terms of the finite difference approximation, and must

therefore be known at coordinates lying on the solution grid. An association between drift averaged quantities like $\langle n_i \rangle$ and a set of adiabatic invariant coordinates can easily be made, since a drift averaged quantity is a function of the integration path S, which corresponds to the path of a radiation belt particle, which can be parameterised by a set of adiabatic invariants. The problem is *how* to obtain drift averaged quantities during a simulation, since Equation 3.18 is tricky to evaluate.

The Solution

One solution is to solve the proton equation of motion for S, then solve Equation 3.18 for $\langle n_i \rangle$, etc., at each point on the solution grid during the simulation. However, this would take a lot of time and memory, and involve a totally separate set of equations. A more efficient solution is to pre-determine each drift averaged quantity over a range of coordinates beforehand, then load in and use this calculation during every simulation.

To implement this solution, each drift averaged quantity $\langle A \rangle$ must first be calculated as a function of the adiabatic invariants. The results at specific coordinates can be arranged on a grid with axes defined with respect to the adiabatic invariants ("the drift average grid"). This grid is stored on disk, then loaded into memory before a numerical simulation. At each point on the solution grid, $\langle A \rangle$ can be interpolated from the loaded drift average grid.

A schematic illustration of a drift average grid is shown in Figure 3.1, shaded in yellow, for the 2D case in terms of μ and L. An example numerical model solution grid is also shown, shaded in green. A conceptual difference between the two grids is that the drift average grid is of fixed size, since it is a pre-computation stored on disk, whilst the solution grid can change size and shape since it is configured by the user at the start of a simulation. A drift average $\langle A \rangle$ must always be interpolable on the solution grid (green), and so it must be pre-determined at surrounding grid points. Therefore, the drift average grid must be large enough to encompass any potential region of interest for numerical modelling.

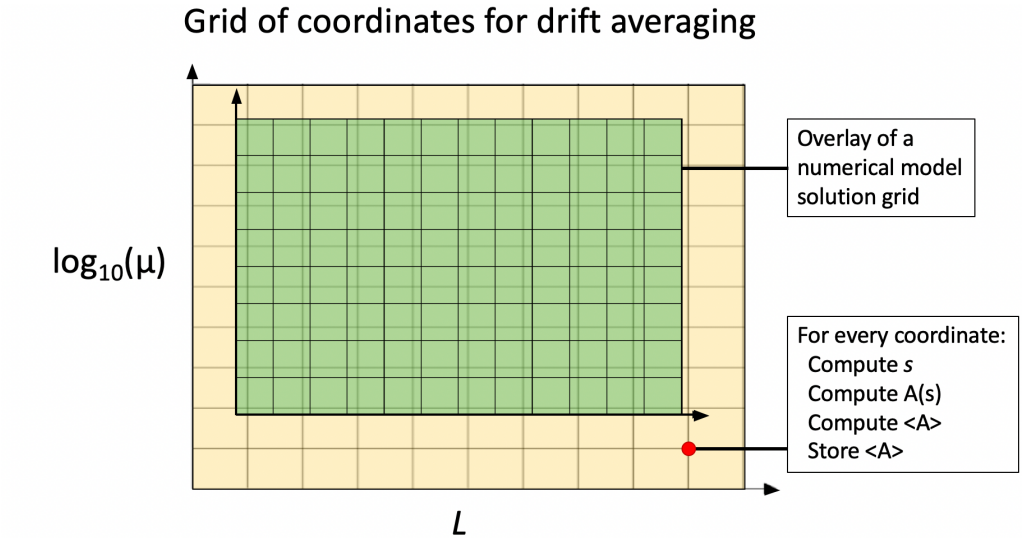


Figure 3.1: Overlap of the region of interest for numerical modelling (green) and the region where drift averaged quantities must be calculated (yellow). This way, a drift averaged quantity can be interpolated from previous calculations anywhere in the green model region.

Implementation

An extensive set of code was developed from scratch to pre-determine drift paths and drift averaged quantities. The final version of this code implements the above solution by producing a 3D grid of drift averaged quantities in terms of the coordinates μ, α_{eq} and L with the following dimensions:

- Dimension #1 (45 intervals):
  $0.1 \leq x \leq 4.0$, $\Delta x = 13/150$, where $x = \log_{10}(\mu \ / \ 1\text{MeV G}^{-1})$
- Dimension #2 (79 intervals):
  $11 \leq \alpha_{eq} \leq 90°$, $\Delta\alpha_{eq} = 1°$
- Dimension #3 (45 intervals):
  $1.1 \leq L \leq 2.0$, $\Delta L = 0.02$

For 2D modelling of equatorially mirroring protons, only quantities at $\alpha_{eq} = 90°$ are required, but since the drift average grid is constant between simulations it must encompass all regions of interest. The choice of α_{eq} as a coordinate instead of J is

explained in Section 3.2.3. This grid can store 169280 quantities, but many of the coordinates correspond to the loss cone so some calculations can be skipped (less than half, though). Outside the grid coordinate range, $\langle A \rangle$ can be extrapolated if necessary (for example, down to the low energy boundary), at the potential cost of accuracy.

The first overall step towards producing this solution was to calculate proton trajectories (abbreviated as PTs in Figures 3.2 to 3.4) at each set of invariant coordinates outside the loss cone on the drift average grid. A high level overview of the process used to compute each trajectory is shown in Figure 3.2, involving two scripts. The process is represented by a flowchart, with a "caller script" handling processes highlighted in blue, and an "equation of motion (EOM) solver" handling processes highlighted in red. A configuration file (top centre of Figure 3.2) was used to specify the grid dimensions shown in the list above, along with other information. The solution output (near centre of Figure 3.2) is the file containing the collection of solved trajectories, written using the HDF5 (Hierarchical Data Format) binary file format.

The internal structure of a .HDF5 file is similar to the hierarchical tree structure of a computer file system, consisting of: groups (analogous to directories); datasets (analogous to files); and descriptive metadata attached to both groups and datasets. Each trajectory is stored as two datasets within a group: the datasets contain position and time along a trajectory, and the group name is a unique ID for the trajectory. The .HDF5 file format has two key features that made it a good choice for storing the large number of proton trajectories required: firstly, the data slicing feature allows a single dataset, or even subsets of that dataset, to be extracted without loading the entire file, reducing loading times; and secondly, the Python interface compresses/decompresses datasets upon closing/opening the file object, saving disk space. An example of the internal structure of the .HDF5 solution file is illustrated in Figure 3.3, for a small collection of 5 proton trajectories.

The process shown in Figure 3.2 can take a week or longer to complete, so an essential feature of the caller script was being able to resume calculations after an interruption, by loading the partially complete solution file and continuing from the ID of the most recently solved trajectory. This minimised the impact of potential crashes due to network outages, etc.

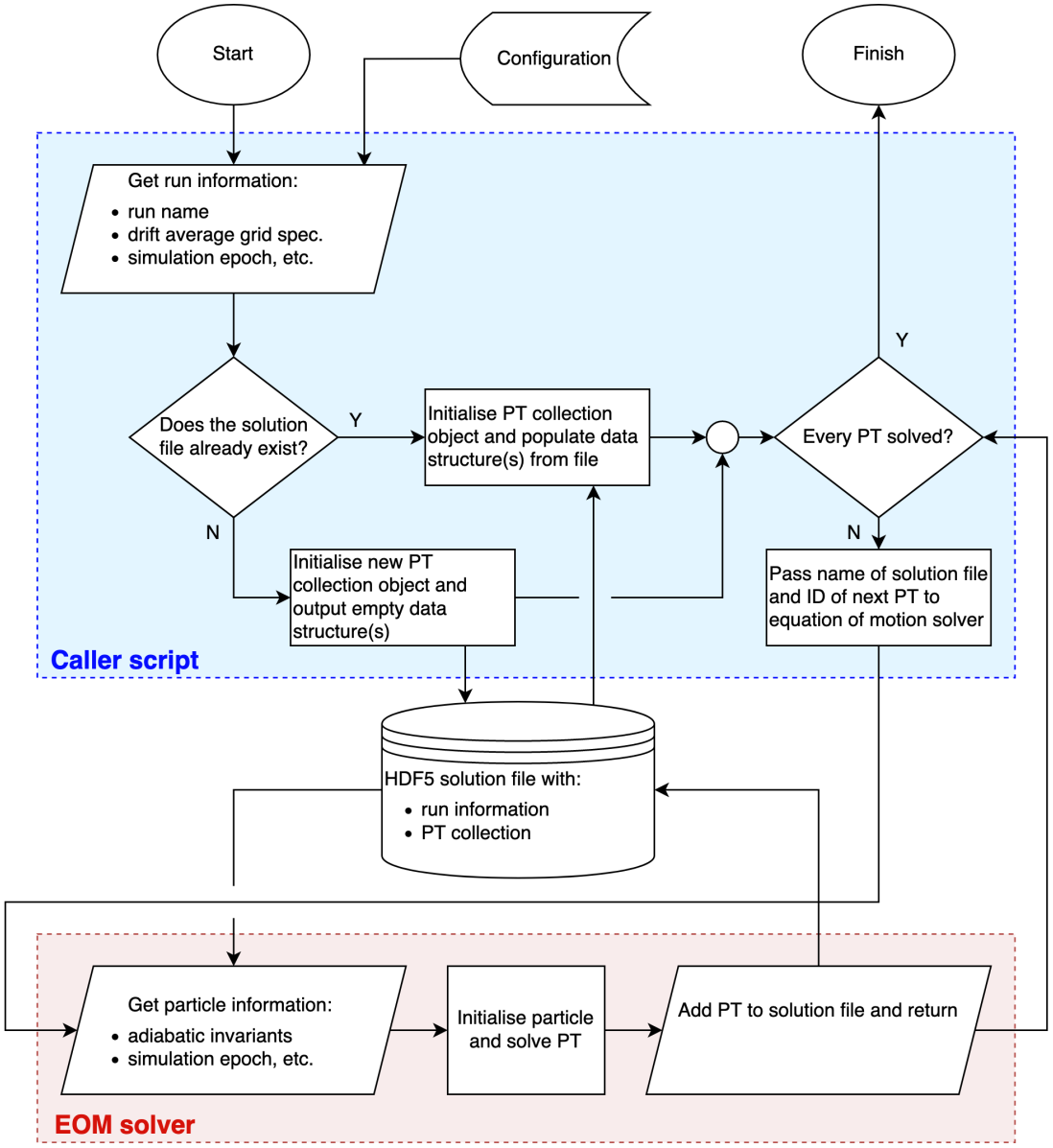


Figure 3.2: Overview of the process for calculating a grid of drift averaged quantities (part 1 of 2):

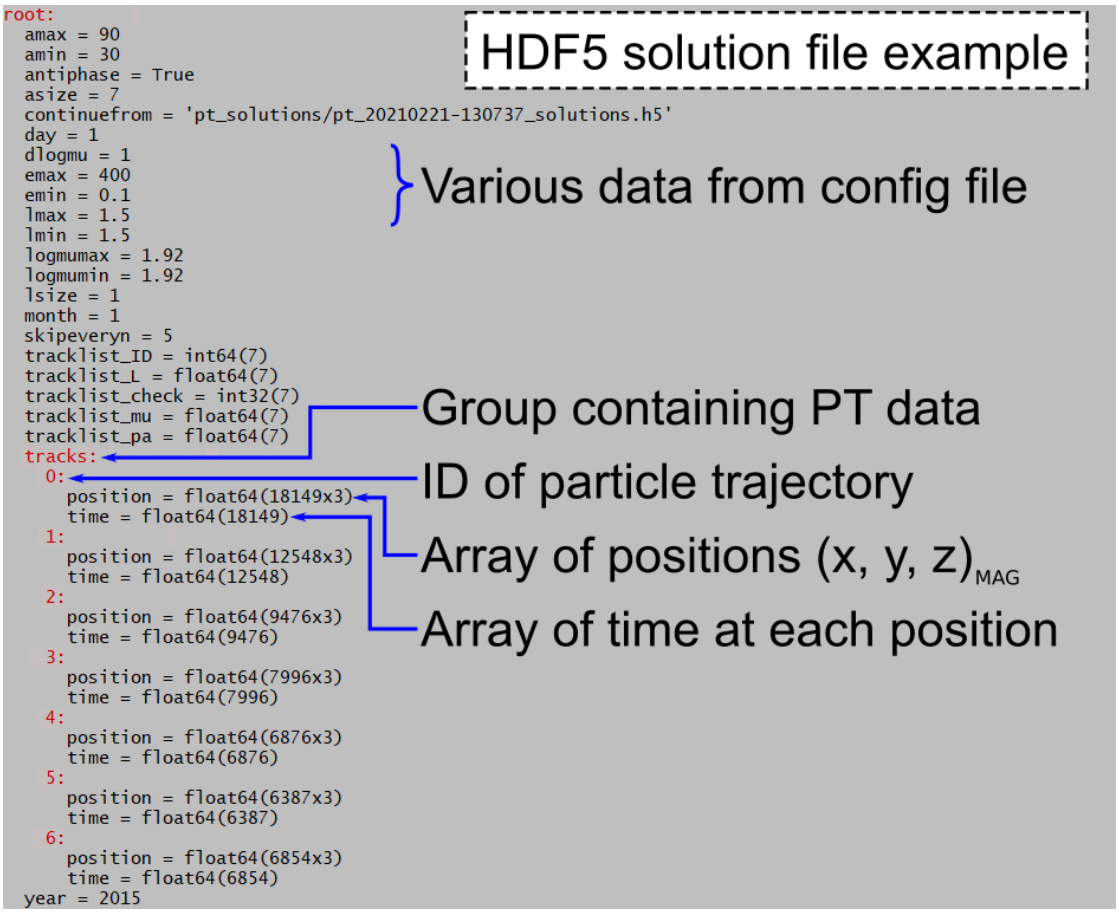

root:
amax = 90
amin = 30
antiphase = True
asize = 7
continuefrom = 'pt_solutions/pt_20210221-130737_solutions.h5'
day = 1
dlogmu = 1
emax = 400
emin = 0.1
lmax = 1.5
lmin = 1.5
logmumax = 1.92
logmumin = 1.92
lsize = 1
month = 1
skipeveryn = 5
tracklist_ID = int64(7)
tracklist_L = float64(7)
tracklist_check = int32(7)
tracklist_mu = float64(7)
tracklist_pa = float64(7)
tracks:
0:
position = float64(18149x3)
time = float64(18149)
1:
position = float64(12548x3)
time = float64(12548)
2:
position = float64(9476x3)
time = float64(9476)
3:
position = float64(7996x3)
time = float64(7996)
4:
position = float64(6876x3)
time = float64(6876)
5:
position = float64(6387x3)
time = float64(6387)
6:
position = float64(6854x3)
time = float64(6854)
year = 2015
HDF5 solution file example
Various data from config file
Group containing PT data
ID of particle trajectory
Array of positions (x, y, z)MAG
Array of time at each position

Figure 3.3:

The finished .HDF5 solution file contains solved proton trajectories across the portion of 169280 drift average grid coordinates outside the loss cone. It is required as input for the second overall step, which is to evaluate Equation 3.18 for each drift average quantity. A high level overview of this process is shown in Figure 3.4, for the pretend physical quantity A, also involving two scripts. The script to evaluate $\langle A \rangle$, handling processes highlighted in teal, is different for every physical quantity A. For example, if evaluating $\langle n_i \rangle$, the script needs to find ion density n_i along each drift path, and depends on input from atmospheric, ionospheric and plasmaspheric density models. Therefore, it works very differently to the equivalent script required to evaluate drift averaged CRAND, for example. On the other hand, the second script to format results in a grid, highlighted in yellow, is the same for any drift averaged quantity, since the table of values written by the previous script is in a standardised format. The script to evaluate $\langle A \rangle$ shown in Figure 3.4 also includes the ability to resume from a previously interrupted calculation, but the flow chart has been simplified since this methodology is shown already in Figure 3.2. The drift average grid produced can be loaded directly into the proton belt numerical model, and uses the proprietary ".mrda" file extension, an abbreviation of "model-ready drift average".

There are two steps shown in Figures 3.2 and 3.4 that merit elaboration:

- "Initialise particle and solve particle trajectory" as part of the EOM solver script (Figure 3.2); and
- "Evaluate [Equation 3.18]" as part of the script to evaluate $\langle A \rangle$ (Figure 3.4).

The former of these is explained next, in Section 3.2.3. The latter depends on modelling the physics of quantity A, and is explained in Section 3.3 for the case of CRAND, and Section 3.4 for the case of atmospheric, ionospheric and plasmaspheric densities.

3.2.3 Modelling a Relativistic Radiation Belt Proton

A core function of the drift average code is to solve proton trajectories for the path S in Equation 3.18. This task is handled by the EOM solver script shown in Figure 3.2, which models the motion of one proton situated in the vicinity of the

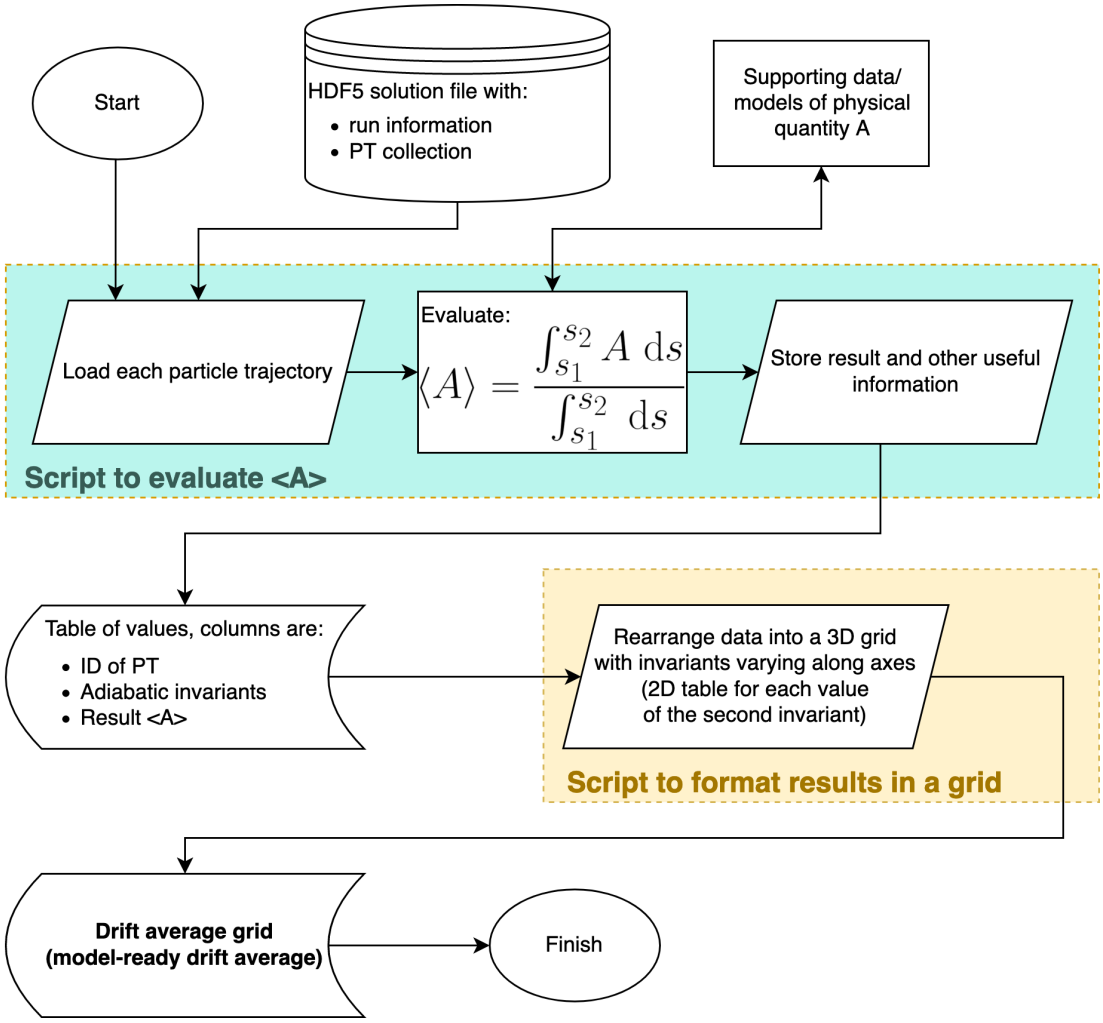


Figure 3.4: Overview of the process for calculating a grid of drift averaged quantities (part 2 of 2):

geomagnetic field via the method described in this section. The model can also be repurposed to simulate electrons or other ions simply by re-specifying the mass and charge of an instantiated particle object.

Coordinate System

Three-dimensional space is described in this model with Cartesian geometry, and coordinates are specified with respect to the Geomagnetic Coordinate (MAG) frame, which is defined with a z axis parallel to the geomagnetic dipole axis, and a centre co-located with that of the GEO frame.

Modelling the Geomagnetic Field

The geomagnetic field is represented by a dipole. The dipole field is given in spherical coordinates by Equation 1.35. The value of B_0 is derived for a given epoch using Equation 1.52, making use of first order coefficients from the IGRF-13 magnetic field model, which are accessed using the pyIGRF Python package (available at `https://pypi.org/project/pyIGRF/`). Epoch is specified by the configuration file shown in Figure 3.2. Figure 1.10 shows the importance of epoch-dependence, since proton drift trajectories change due to secular variation of the geomagnetic field.

Solving The Proton Equation of Motion

To calculate the motion of a particle is to calculate position through time. The change in position $\Delta\mathbf{x}$ over a time interval Δt can be guessed if one knows the momentum of the particle at the start of the time interval. However, the Lorentz force is always applying a change in momentum. Therefore, to calculate the motion of a particle, the time evolution of both position and momentum must be solved from an initial condition.

Together, position and momentum form a state vector $\mathbf{Y} = (\mathbf{x}, \boldsymbol{\rho}) = (x, y, z, \rho_x, \rho_y, \rho_z)$. The evolution of the state vector $\mathbf{Y}$ is given by an ordinary differential equation

(ODE) of the form

$$\frac{\mathrm{d}\mathbf{Y}}{\mathrm{d}t} = (v_x, v_y, v_z, \frac{\mathrm{d}\rho_x}{\mathrm{d}t}, \frac{\mathrm{d}\rho_y}{\mathrm{d}t}, \frac{\mathrm{d}\rho_z}{\mathrm{d}t}) = f(\mathbf{Y}) \tag{3.19}$$

The first step was to derive $f(\mathbf{Y})$ for a radiation belt proton. To begin with, the Lorentz force on a charged particle is given by Equation 1.1 as a vector. By expanding the cross product, the following set of three scalar equations is produced:

$$\begin{aligned} \frac{\mathrm{d}\rho_x}{\mathrm{d}t} &= q(E_x + v_y B_z - v_z B_y) \\ \frac{\mathrm{d}\rho_y}{\mathrm{d}t} &= q(E_y + v_z B_x - v_x B_z) \\ \frac{\mathrm{d}\rho_z}{\mathrm{d}t} &= q(E_z + v_x B_y - v_y B_x) \end{aligned} \tag{3.20}$$

An expression for $\mathbf{v}$ is required to evaluate this set of equations. One can be derived considering the total energy of a particle, given by

$$E_{tot} = E_0 + T = \sqrt{m_0^2 c^4 + \rho^2 c^2} = \gamma m_0 c^2 \tag{3.21}$$

Rearranging for γ gives the expression

$$\gamma = \sqrt{1 + \left(\frac{\rho}{m_0 c}\right)^2} \tag{3.22}$$

Since $\boldsymbol{\rho} = \gamma m_0 \mathbf{v}$, then $\mathbf{v}$ can be expressed by substituting in Equation 3.22 to give

$$\mathbf{v} = \frac{\boldsymbol{\rho}}{m_0}\left(1 + \left(\frac{\rho}{m_0 c}\right)^2\right)^{-1/2} \tag{3.23}$$

Just as the Lorentz force is split up in Equation 3.20, Equation 3.23 is separated

into three scalar equations for velocity:

$$
\begin{aligned}
v_x &= \frac{\rho_x}{m_0}\left(1+\left(\frac{\rho}{m_0 c}\right)^2\right)^{-1/2} \\
v_y &= \frac{\rho_y}{m_0}\left(1+\left(\frac{\rho}{m_0 c}\right)^2\right)^{-1/2} \\
v_z &= \frac{\rho_z}{m_0}\left(1+\left(\frac{\rho}{m_0 c}\right)^2\right)^{-1/2}
\end{aligned}
\tag{3.24}
$$

$f(\mathbf{Y})$ can now be written out in full by substituting in the scalars in Equations 3.20 and 3.24 into Equation 3.19. A numerical ODE solver from the SciPy Python package was used to integrate $f(\mathbf{Y})$ at each timestep over a the drift trajectory (Virtanen et al., 2020). The solver chosen was "DOP853", which is an explicit Runge-Kutta method of order 8.

Initialising a Trapped Proton

The initial state vector $\mathbf{Y_0}(\mathbf{x_0}, \boldsymbol{\rho_0})$ at time t_0 (the epoch) is determined from the set of coordinates (μ, α_{eq}, L) provided by the caller script. Since a set of invariant coordinates do not specify phase, the initial state vector $\mathbf{Y_0}$ does not have a unique solution; a proton can be initialised at any phase of gyration, and at a non-specific magnetic latitude, longitude combination that places it anywhere on the drift shell. The following algorithm was implemented to initialise protons on the magnetic equator. The steps to determine $\mathbf{Y_0}$ from coordinates (μ, α_{eq}, L) are:

(a) calculate equatorial magnetic field strength B_e at radius La

(b) calculate $\rho_{0\perp} = \sqrt{2\mu B_e m_0}$

(c) calculate $\rho_{0\parallel} = \rho_{0\perp} / \tan(\alpha_{eq})$

(d) choose the starting momentum arbitrarily to be either
$\boldsymbol{\rho_0} = (0, -\rho_{0\perp}, \rho_{0\parallel})$, or
$\boldsymbol{\rho_0} = (0, \rho_{0\perp}, -\rho_{0\parallel})$

(e) calculate γ and v from $\boldsymbol{\rho_0}$ using Equations 3.22 and 3.24

(f) approximate the gyroradius via $r_g = m_0\gamma v^2/F$, where F is the magnitude of the Lorentz force for a guiding centre $\mathbf{x}_{\mathrm{GC}} = (La, 0, 0)$, given by Equation 3.20

(g) choose the starting position to be either
$\mathbf{x_0} = (La + r_g, 0, 0)$ if the first starting momentum above was used, or
$\mathbf{x_0} = (La - r_g, 0, 0)$ if the second starting momentum was used

Figure 3.5 shows a proton trajectory initialised at $\mathbf{x_0}$ with μ =316MeV/G, α_{eq}=40° and $L = 1.4$, solved using the particle tracing code and then truncated to one bounce period. Specifying the second invariant of a particle in terms of α_{eq} allows the initial state vector to be determined via the above method. This was the main motivation for choosing invariant coordinates (μ, α_{eq}, L) to define the drift average grid dimensions. In contrast, for a given value of J, a proton's initial state vector can be approximated (using Equation 1.45, etc.) but not exactly determined and vice versa. However, an alternative method to initialise proton drift orbits is performed by Selesnick et al. (2007), whereby particles are initialised at their approximated mirror points for a given value of J.

One minor insufficiency of parameterising a drift averaged quantity by three adiabatic invariant coordinates is a small dependence on the initial phase along a gyration or bounce that traces out S, or in other words a dependence on $\mathbf{Y_0}$. This dependence can become important when S corresponds to a particle that is not well described by adiabatic motion or "quasi-trapped"; for example, a particle with high energy close to the trapping limit that may escape from the radiation belts after a few drift orbits. It can also be important when the drift average is taken over a short path S. A technique used by Selesnick et al. (2007) to check that a particle is properly trapped is to calculate S several times by initialising a particle with the same properties but at different phases, and allowing S to extend over multiple drift orbits to check the invariants are conserved. This technique could also be applied to investigate any dependence of $\langle A \rangle$ on initial phase, and eliminate it by re-evaluating $\langle A \rangle$ over multiple paths corresponding to the same set of invariants then averaging the result.

The above algorithm implements a basic version of this capability by offering the two different combinations of starting position and momentum. There is a 180°

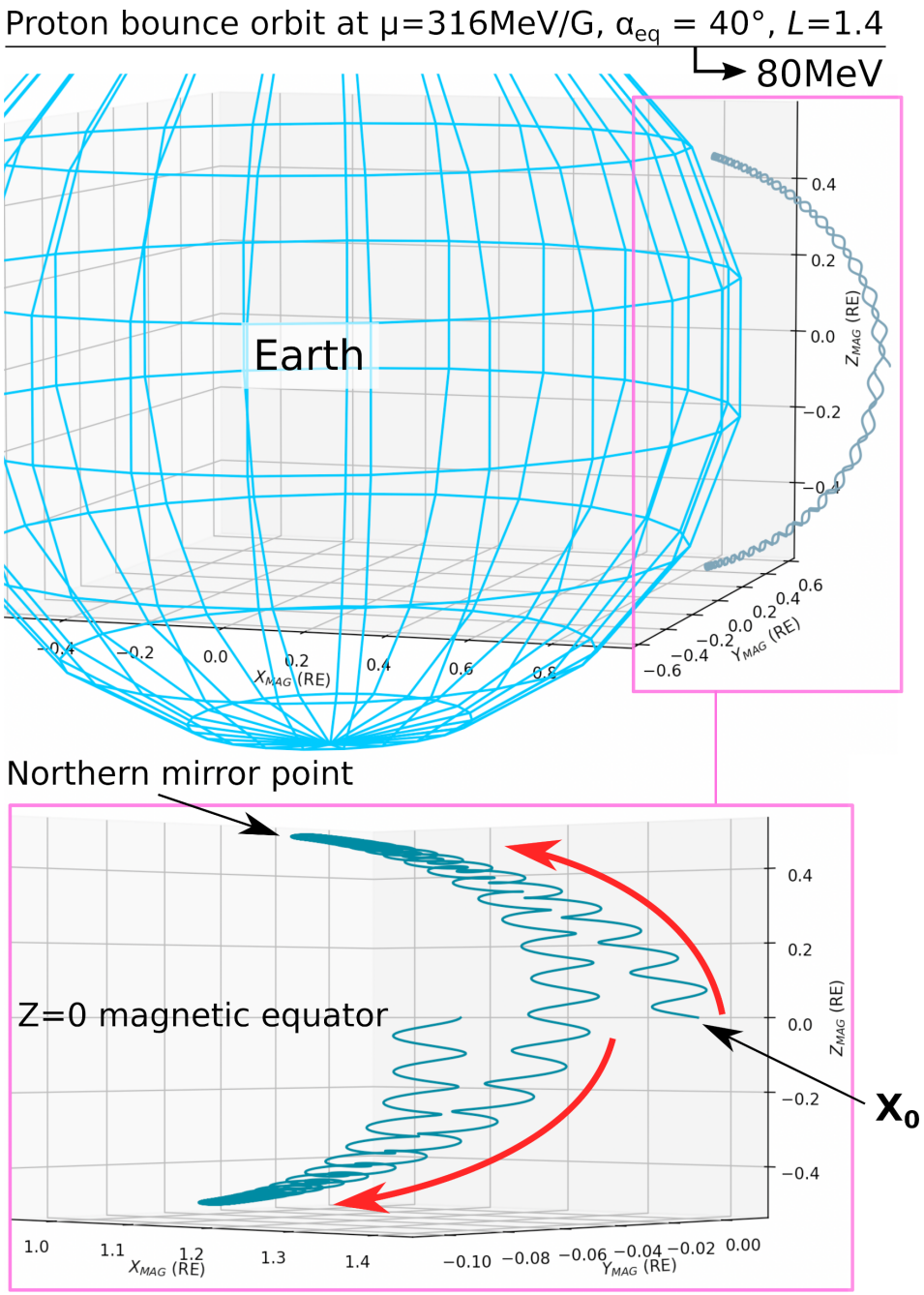


Figure 3.5: Example bounce trajectory produced using the drift averaging code. The particle trajectory shown is at 80MeV, 40° equatorial pitch angle and $L = 1.4$. The trajectory is truncated to exactly one bounce, from $\mathbf{X_0}$ as shown until the magnetic equator is reached again.

difference in gyrophase between the two options, and the initial vertical component of velocity is oppositely directed. Each drift average can thus be calculated twice using both options for $\mathbf{Y_0}$ to investigate the dependence on phase, and average if deemed necessary.

Controlling the Timestep

The timestep Δt was chosen to be a constant fraction of the local gyroperiod T_g, related by

$$\Delta t = \frac{T_g}{N_{\Delta t}} \propto \frac{1}{B} \tag{3.25}$$

A value of $N_{\Delta t} = 125$ timesteps per gyroperiod was found to provide a good balance between accuracy and execution time. Since T_g depends on local magnetic field strength, the timestep is dynamically adjusted by the solver over the course of a particle trajectory, becoming smaller where the magnetic field becomes stronger.

Certain quantities such as γ are physically conserved along S and therefore should also be numerically conserved. Determining these quantities periodically along S from the simulation allowed the value of $N_{\Delta t}$ to be tested. Since many timesteps are calculated, it is not necessary to store all of them on disk in the .HDF5 output file, and the fraction saved is configurable.

3.2.4 A Trick to Reduce Computation Time in a Dipole Field

The drift average grid, as described in Section 3.2.2, holds 169280 coordinates (μ, α_{eq}, L). Many of these coordinates correspond to trajectories inside the loss cone where particle tracing is skipped, but the majority are coordinates that lead to trapping. The aim was to compute a proton trajectory S over one or more drift periods at each coordinate, so that each drift averaged quantity required could be evaluated across the entire grid.

Computing the drift orbits of this many particles is computationally expensive, and indeed an issue encountered during the project was that computations were predicted to require an unreasonably long time. Therefore, a shortcut was devised to drastically cut the required computation time. The principal behind the shortcut

is that particle trajectories are solved in a dipole field, and therefore a given trajectory should be longitudinally symmetric around the centre of the MAG frame, except for variations in gyrophase. Therefore, the orbit over a single bounce period can be computed, then "copied and pasted" around Earth, i.e. rotated to different magnetic longitudes, in order to achieve a similar-looking trajectory.

This shortcut was implemented, and the aim was therefore to compute proton trajectories S over a single bounce period at each of the 169280 coordinates. When it came to drift averaging a quantity A as per Equation 3.18, the longitudinal dependence of A was taken into account by evaluating $\langle A \rangle$ along the single bounce trajectory S, then rotating S in longitude via a simple matrix rotation of the stored particle trajectory, and repeating the calculation of $\langle A \rangle$. Each value of $\langle A \rangle$ was then averaged.

To implement this, an algorithm was written to solve the proton equation of motion for a length of time equal to $1.1\times$ the bounce period given by Equations 1.39 and 1.40. Since these equations are approximations, the factor of 1.1 ensured that each proton completed at least one bounce without needing to evaluate conditional statements during timestepping. The trajectory was then followed backwards from its finish point and truncated just before its last crossing of the magnetic equator (located at $z = 0$). The trajectory was then solved forward in time for a fraction of the usual timestep depending on v_z, such that the final position was located almost exactly at $z = 0$. The result of this algorithm is demonstrated in Figure 3.5, where the proton trajectory is truncated at the magnetic equator.

One caveat of this method was that it perhaps exacerbated any dependence $\langle A \rangle$ had on the initial phase of the "drift" orbit. As discussed in Section 3.2.3 however, the solver allowed each particle trajectory to be initialised using an alternative state vector $\mathbf{Y_0}$ at a different gyrophase. This allowed drift averaged quantities susceptible to phase dependence to be re-calculated along each trajectory and averaged with the result derived from the alternative $\mathbf{Y_0}$. This was found to make a difference for drift averaged CRAND at high energy, and CRAND drift averages were therefore calculated twice then averaged using this method.

3.3 Modelling CRAND

3.3.1 Main Equation

Cosmic ray albedo neutron decay is the primary source of protons at $\gtrsim$50MeV at low L (Selesnick et al., 2007), occurring when interactions between the atmosphere and incoming galactic cosmic rays produce upward-moving neutrons, which undergo beta decay to produce protons that become trapped. Following the definition of phase space density $f = m_0^3 j/p^2$, the rate of change in model phase space density due to CRAND is given by

$$S_n = \frac{\partial}{\partial t}\left(m_0^3 \frac{j_p}{p^2}\right) = \frac{m_0^3}{p^2}\left(\frac{\partial j_p}{\partial t}\right) \tag{3.26}$$

where $\partial j_p/\partial t$ is the rate of change in proton flux at a set of adiabatic coordinates caused by the pickup of newly produced protons. The analytical relation between $\partial j_p/\partial t$ and neutron albedo flux $\boldsymbol{j_n}$ is given by

$$\frac{\partial j_p}{\partial t}_{n} = \frac{\oint_S \boldsymbol{j_n} \cdot \mathrm{d}\mathbf{s}}{\gamma\tau_n \oint_S \mathrm{d}s} \tag{3.27}$$

where the integral takes place along a proton drift orbit S, $\mathrm{d}\mathbf{s}$ is an element of length along the trajectory, $\boldsymbol{j_n} \cdot \hat{\mathrm{d}\mathbf{s}}$ is the flux of albedo neutrons leaving the atmosphere in a direction coinciding with the section of drift orbit, and $\gamma\tau_n$ is the relativistic neutron lifetime (Dragt et al., 1966). Selesnick et al. (2007) gives a value $\tau_n = 887$s. Equation 3.27 assumes that the proton produced moves in the same direction of the decaying neutron and absorbs all its kinetic energy (Selesnick et al., 2007).

To evaluate the integral in Equation 3.27, albedo neutron flux j_n must be determined at each position along S in the instantaneous direction of the particle. However, this flux is produced by interactions within the atmosphere, so j_n along S is the result of neutrons travelling into space then happening to align directly with the particle's instantaneous velocity.

3.3.2 Injection Coefficient Method

To deal with the difficult computation of Equation 3.27, Dragt et al. (1966) introduced an injection coefficient given by

$$\chi = \frac{\partial j_p / \partial t_n}{J_{2\pi} / \gamma \tau_n} \tag{3.28}$$

where $J_{2\pi}$ is globally averaged neutron escape flux. This quantity is defined by

$$J_{2\pi} = (2\pi A)^{-1} \int j_n \cos z \; \mathrm{d}A \; \mathrm{d}\Omega \tag{3.29}$$

where the integral is taken over the surface area of Earth A, and over the upper half of the unit solid sphere, with zenith angle z from the local vertical. To understand Equations 3.28 and 3.29, it may help to relate the terms to physical data. Figure 3.6 shows measurements of albedo neutron flux $j_n(E, z)$ from a balloon experiment at $\sim$ 36.5km altitude and 40° latitude, as presented by White et al. (1972). The x-axis of Figure 3.6 is zenith angle z, with measurements plotted at $z < 90°$ showing upwards flux (moving past the balloon towards space), and measurements at $z > 90°$ showing downwards flux (towards Earth). Therefore, the integrand $j_n \cos z$ of Equation 3.29 is simply the vertical component of each flux measurement. The integral of Equation 3.29 is performed over half a solid sphere, meaning $\mathrm{d}\Omega = 2\pi \sin(z) \; \mathrm{d}z$, with corresponding limits of integration $0 \leq z \leq 90°$. Therefore, the integration is just over the upward flux measurements shown in the left half of Figure 3.6. After averaging across the surface area of Earth, $J_{2\pi}$ is then the total vertically upward flux of neutrons that would be measured at any point on the surface of an isotropic source equivalent to Earth's atmosphere. The quantity χ is then the ratio of the rate of change in proton flux from j_n, to the rate of change in proton flux from $J_{2\pi}$.

The quantity χ allows the approximation of $\partial j_p / \partial t$ based on some simplifying assumptions. This "injection coefficient method" was used by Claflin and White (1974) to evaluate CRAND for equatorially mirroring protons, and their source function was then re-used by Albert et al. (1998) and Albert and Ginet (1998). For the special case of equatorially mirroring protons, the injection coefficient was

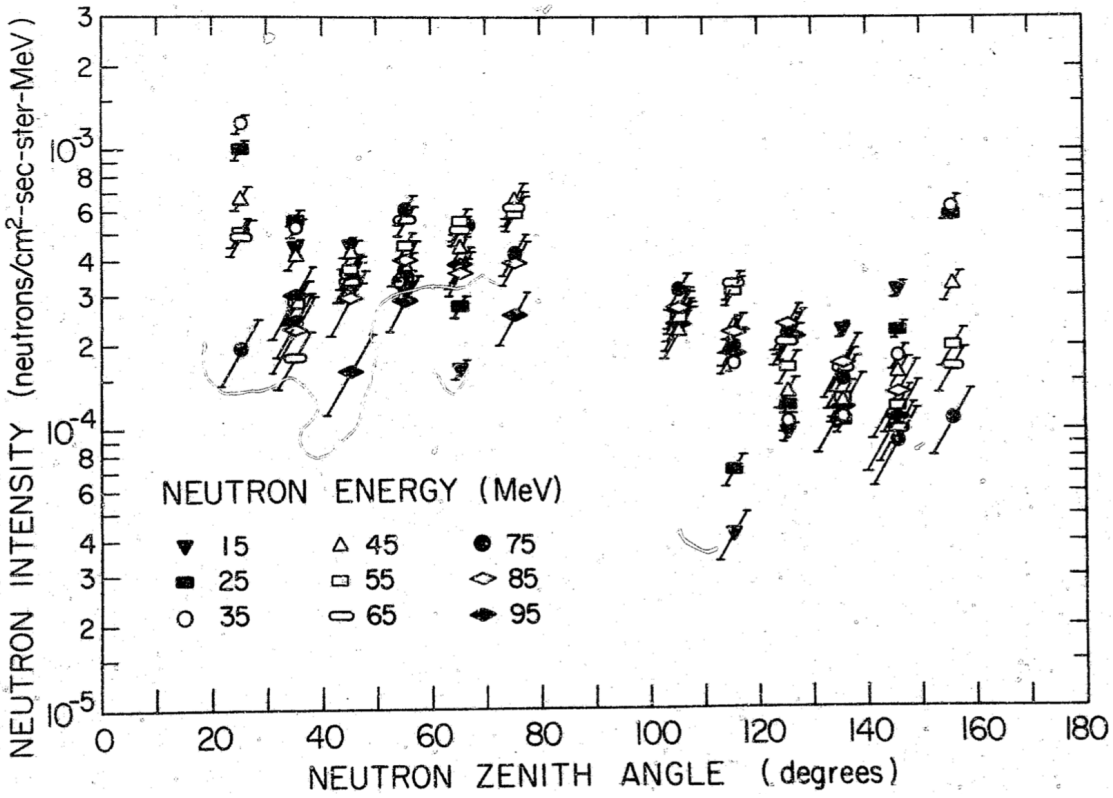


Figure 3.6: Balloon measurement of neutron flux made at 120,000ft and 40° latitude as a function of zenith angle (zero zenith is vertically upwards), from Figure 2 of White et al. (1972)

written as

$$\chi(L,E) = \frac{2}{LJ_{2\pi}(E)} \int_0^{\pi/2} \frac{j_n(\phi, E, \lambda = 0) \cos\phi \, \mathrm{d}\phi}{(1 - L^{-2}\sin^2\phi)^{1/2}} \tag{3.30}$$

from Equation 8 of Claflin and White (1974, with an extra factor of 2π in the authors' source term S versus Equation 3.26 above).

During initial development of the proton belt numerical model, it was decided to validate against the 2D numerical model of Albert et al. (1998). The same master equation of Albert et al. (1998) was therefore implemented independently, described in Seciton 3.5.1. The model aimed to solve for non-relativistic proton phase space density as function of μ and L. Since Albert et al. (1998) use the same CRAND source as (Claflin and White, 1974), Equation 3.30 was therefore also implemented and evaluated by digitising the neutron flux measurements presented in Figure 3.6 from White et al. (1972). Once the calculation of the injection coefficient had been validated (against Figure 2 of Claflin and White (1974)), the neutron data used to evaluate it was updated to that shown in Figures 9 and 12 of Preszler et al. (1976), which includes more corrections. As a result of this process, it was discovered that the evaluation of CRAND in the numerical model of Albert et al. (1998) contained an error causing S_n to be overestimated by a factor of π. Consultation with the original author of the paper confirmed this.

As development of the numerical model continued, it was decided to use a more accurate approach to evaluate CRAND so as not to rely on balloon datasets or rely on the approximate injection coefficient calculation. Selesnick et al. (2007) demonstrate a more thorough approach, but it was only until after the drift average code was developed and tested that the new approach, explain below, could be attempted.

3.3.3 Drift Averaging Method

Equation 3.27 can be computed more rigorously using a drift average method. It has the form as Equation 3.18, and the drift averaged quantity $\langle j_n \rangle = \oint_S \boldsymbol{j_n} \cdot \mathrm{d}\mathbf{s} / \oint_S \mathrm{d}s$ represents average atmospheric neutron albedo flux in the direction of a proton along its drift. The integral can be evaluated along the path S of a proton trajectory to find $\langle j_n \rangle$ for a set of adiabatic invariant coordinates corresponding to S.

An application of this method was first published for the Selesnick et al. (2007)

numerical proton belt model. For this previous work, the interactions of incoming galactic cosmic rays throughout the atmosphere were modelled to find directional neutron albedo flux at the top of the atmosphere $j_n = j_n(E_n,\mathrm{z},\mathcal{R}_{cv},\mathrm{F10.7})$, where E_n is kinetic energy, z is zenith angle, and $\mathcal{R}_{cv}$ is vertical geomagnetic cutoff rigidity. This dataset can be used to evaluate the neutron flux along a proton trajectory at position $\boldsymbol{\sigma}$, by following the negative tangent vector to S at $\boldsymbol{\sigma}$ in the $-\hat{\mathrm{d}}\mathbf{s}$ direction, back to a point on top of Earth's atmosphere that it intersects. This geometry is shown in Figure 3.7, with the point $\boldsymbol{\sigma}$ in blue. For a known atmospheric intersection point (shown as a yellow dot in Figure 3.7), the independent variables z and $\mathcal{R}_{cv}$ can be evaluated, giving j_n via the dataset. For positions along S where there is no such intersection, $j_n = 0$.

The dataset produced by Selesnick et al. (2007) giving $j_n = j_n(E_n,\mathrm{z},\mathcal{R}_{cv},\mathrm{F10.7})$ was kindly provided at request by Richard Selesnick to use in this work, and allowed the computation of Equation 3.27. Aside from this neutron albedo dataset, the drift averaging method to evaluate CRAND shown here was developed independently. It involved numerically evaluating $\langle j_n \rangle$ across the set of proton drift orbit trajectories calculated in Section 3.2, resulting in $\langle j_n \rangle$ at each set of adiabatic invariant coordinates on the drift average grid. For each drift orbit trajectory, $\langle j_n \rangle$ was re-computed for five values of F10.7 (60, 100, 140, 180 and 220 sfu), allowing the solar cycle dependence of S_n to be taken into account. The method is broken down into its requisite steps below, which describe each stage of implementation.

Transformations Between GEO and MAG Frames

Since particle trajectories solved by the drift averaging process (Section 3.2) are written with respect to the MAG frame, but Earth's atmosphere is most easily modelled with respect to the GEO frame, it is necessary to convert vectors between MAG and GEO frames when calculating interactions between the two based on geometry. A 3-vector in the GEO frame $\mathbf{V}_{\mathrm{GEO}}$ can be converted to the MAG frame

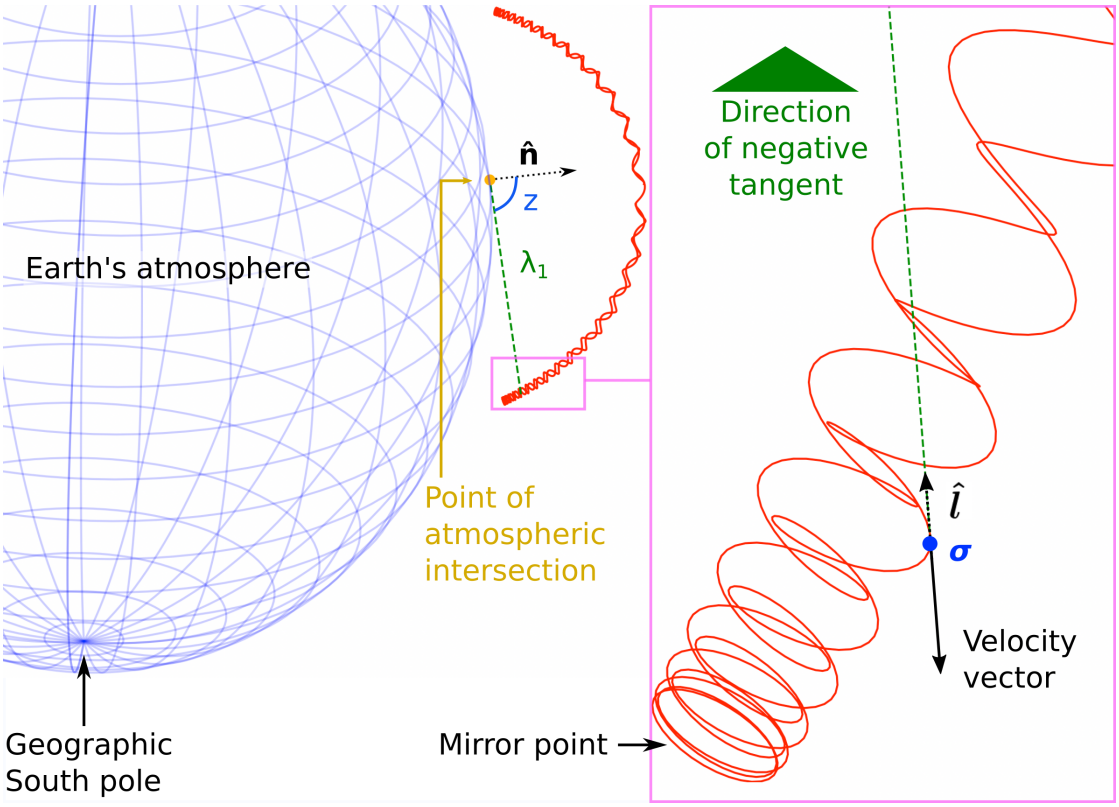


Figure 3.7: Illustration of an intersection between the negative tangent to the particle trajectory (green dashed line) and the top of Earth's atmosphere (blue surface). The particle trajectory, shown in red, is the same as shown in Figure 3.5. The distance from the particle, along the negative tangent, until the atmospheric intersection is denoted λ_1. The yellow dot represents the point of intersection.

via the transformation matrix R, which was defined like so:

$$R\ \mathbf{V}_{\text{GEO}} = \begin{bmatrix} 0.29448006 & 0.95434236 & 0.05012142 \\ -0.9408124 & 0.29871502 & -0.1601292 \\ -0.1677901 & 0. & 0.98582274 \end{bmatrix} \mathbf{V}_{\text{GEO}} = \mathbf{V}_{\text{MAG}} \tag{3.31}$$

Similarly, a 3-vector in the MAG frame $\mathbf{V}_{\text{MAG}}$ can be converted to the GEO frame via the transformation matrix R', which was defined according to:

$$R'\ \mathbf{V}_{\text{MAG}} = \begin{bmatrix} 0.29448006 & -0.9408124 & -0.1677901 \\ 0.95434236 & 0.29871502 & 0. \\ 0.05012142 & -0.1601292 & 0.98582274 \end{bmatrix} \mathbf{V}_{\text{MAG}} = \mathbf{V}_{\text{GEO}} \tag{3.32}$$

The matrices R and R' shown above correspond to the epoch of new year 2015; the MAG and GEO frames are both Earth-fixed, so the general transformations R and R' are epoch-dependent insofar as to be influenced by secular changes in Earth's field. Both were calculated using tools available via the Python SpacePy package (Morley et al., 2011). R was initially calculated for the year 1965 to validate against calculations given by Russell (1971), but R and R' are re-calculated by the CRAND evaluation code for whichever epoch the particle trajectory corresponds to (specified within the .HDF5 solution file). Consistency of epoch is important, since both particle trajectories as well as transformation matrices used to evaluate CRAND are functions of secular variation.

Modelling Earth's Atmosphere

The top of the atmosphere was approximated by modifying the WGS84 ellipsoid (Decker, 1986) model of Earth's surface, with an added 100km of radius. The WGS84 model is defined with respect to the GEO frame. In this frame the Earth is described as an oblate spheroid, since it is symmetrical in the x and y directions, with a semimajor axis length of $a = 6378137$m and semiminor axis length of $b = 6356752.3142$m. This geometry is shown in Figure 3.8.

With the extra 100km to account for height of the atmosphere, the top of the atmosphere was thus modelled using an oblate spheroid with semimajor axis

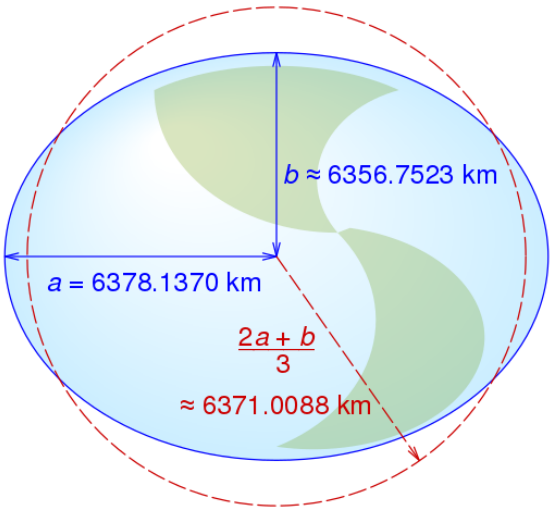


Figure 3.8: The WGS84 model of Earth with the semimajor axis a and semiminor axis b labelled (not to scale), taken directly from Wikipedia (2021).

$a' = 6478137$m and semiminor axis $b' = 6456752.3142$m, with a centre at the origin of the GEO frame. The equation for a point on this surface, derived from the standard equation for an ellipsoid, is

$$\frac{x^2}{a'^2} + \frac{y^2}{a'^2} + \frac{z^2}{b'^2} = 1 \tag{3.33}$$

where x, y and z are components of a coordinate $\mathbf{x}_{\text{GEO}}$ in the GEO frame.

This model of Earth's atmosphere can be written as a vector equation by defining a matrix $\mathcal{E}$ given by:

$$\mathcal{E} = \begin{bmatrix} \frac{1}{a'^2} & 0 & 0 \\ 0 & \frac{1}{a'^2} & 0 \\ 0 & 0 & \frac{1}{b'^2} \end{bmatrix} \tag{3.34}$$

so that Equation 3.33 is given by ${\mathbf{x}_{\text{GEO}}}^T \, \mathcal{E} \, \mathbf{x}_{\text{GEO}} = 1$. The transformation matrix R in Equation 3.31 can then be used to transform the model to be used with coordinates $\mathbf{x}$ defined in the MAG frame:

$$(\mathbf{x} - \mathbf{c})^T \; R^T \mathcal{E} R \; (\mathbf{x} - \mathbf{c}) = 1 \tag{3.35}$$

where $\mathbf{c}$ denotes the centre of the ellipsoid in the MAG frame. Since it is centred on the origin in the GEO frame, and the origins of the MAG and GEO frames coincide, then $\mathbf{c} = (0, 0, 0)$, but it is included in Equation 3.35 to indicate that an additional translation step may be necessary for other frame transformations.

Alignment of $\mathbf{j_n}$ with the Proton Trajectory

Using the ellipsoidal model of Earth's atmosphere in the MAG frame from Equation 3.35, one can find whether or not the negative tangent to the proton trajectory intersects with the top of the atmosphere by solving the problem of a line intersecting with an ellipsoid. In this problem, the point of intersection is denoted by the same symbol $\mathbf{x}$ as in Equation 3.35 since it exists on the ellipsoid surface. In the case that an intersection does exist, it is also given by the following equation for a line:

$$\mathbf{x} = \boldsymbol{\sigma} + \lambda \hat{\boldsymbol{l}} \tag{3.36}$$

where $\boldsymbol{\sigma}$ is the coordinate along the proton trajectory under consideration, and $\hat{\boldsymbol{l}}$ is a unit vector pointing in the direction of the negative tangent to the proton trajectory at $\boldsymbol{\sigma}$. The tangent to the proton trajectory is illustrated by the black arrow for an example point in Figure 3.7, and $\hat{\boldsymbol{l}}$ would therefore point in the opposite direction along the green dashed line as indicated.

The unit vector $\hat{\boldsymbol{l}}$ can be found numerically at the ith position along a solved proton trajectory by

$$\hat{\boldsymbol{l}} = -\|\boldsymbol{\sigma}_{i+1} - \boldsymbol{\sigma}_i\| \tag{3.37}$$

The intersection point $\mathbf{x}$ thus depends on the remaining unknown λ. It can be solved for by substituting Equation 3.36 into Equation 3.35, yielding:

$$(\boldsymbol{\sigma} + \lambda \hat{\boldsymbol{l}} - \mathbf{c})^T \; R^T \mathcal{E} R \; (\boldsymbol{\sigma} + \lambda \hat{\boldsymbol{l}} - \mathbf{c}) = 1 \tag{3.38}$$

Setting $\mathbf{c} = 0$ and $\mathcal{F} = R^T \mathcal{E} R$, Equation 3.38 can be simplified to

$$(\boldsymbol{\sigma} + \lambda \hat{\boldsymbol{l}})^T \ \mathcal{F} \ (\boldsymbol{\sigma} + \lambda \hat{\boldsymbol{l}}) - 1 = 0 \tag{3.39}$$

Expanding this leads to the follow quadratic equation:

$$\lambda^2 \hat{\boldsymbol{l}}^T \mathcal{F} \ \hat{\boldsymbol{l}} + \lambda(\hat{\boldsymbol{l}}^T \mathcal{F} \ \boldsymbol{\sigma} + \boldsymbol{\sigma}^T \mathcal{F} \ \hat{\boldsymbol{l}}) + (\boldsymbol{\sigma}^T \mathcal{F} \ \boldsymbol{\sigma} - 1) = 0 \tag{3.40}$$

Equation 3.40 may have two real and unique solutions λ_1 and λ_2, where $\lambda_1 < \lambda_2$. In this case:

- λ_1 is equal to the distance from $\boldsymbol{\sigma}$ along $\hat{\boldsymbol{l}}$ to a point on the top surface of the atmosphere, as labelled in Figure 3.7;
- λ_2 is equal to the distance starting from $\boldsymbol{\sigma}$, through the point $\boldsymbol{\sigma} + \lambda_1 \hat{\boldsymbol{l}}$, and continuing through the ellipsoid volume to a point on the opposite side of the atmosphere facing away.

The value λ_1 is taken as the solution, since in most cases a neutron cannot make its way from the point $\boldsymbol{\sigma} + \lambda_2 \hat{\boldsymbol{l}}$ to $\boldsymbol{\sigma}$ without colliding with Earth (although it may be possible if this only included traversal through a thin section of the atmosphere above Earth). In the case that only one unique solution exists, then λ_1 is the distance to a point on the top of the atmosphere that is barely skimmed by a vector following the negative tangent to the particle trajectory. In the case that no real solution exists, the vector following the negative tangent does not intersect with the atmosphere.

Knowing the point $\mathbf{x} = \boldsymbol{\sigma} + \lambda_1 \hat{\boldsymbol{l}}$, the next step is to find the albedo flux of neutrons $\boldsymbol{j_n}$ in the integrand of Equation 3.27, which emanates from $\mathbf{x}$. It can be determined using the dataset describing directional neutron albedo flux at the top of the atmosphere $j_n = j_n(E_n, \mathrm{z}, \mathcal{R}_{cv}, \mathrm{F10.7})$. To do this requires 4D interpolation from the dataset in terms of the independent variables. Energy E_n was approximated as equal to the kinetic energy of the proton, so this quantity is known, and F10.7 is also known since it was prescribed for the calculation in order to calculate CRAND for a specific level of solar activity. Therefore, the zenith angle of intersection z and vertical geomagnetic cutoff rigidity $\mathcal{R}_{cv}$ must next be found at $\mathbf{x}$.

Vertical Geomagnetic Cutoff Rigidity

Equation 4 of Smart and Shea (2005) gives geomagnetic cutoff rigidity $\mathcal{R}_c$ as a function of zenith angle and latitude using mixed CGS units. This equation was adapted to find vertical geomagnetic cutoff rigidity $\mathcal{R}_{cv}$ (zenith equal to zero) in SI units, giving

$$\mathcal{R}_{cv} = \frac{\mu_E \cos^4(\lambda)}{4d^2 (100a)^2} \frac{3 \times 10^5}{1 \times 10^9} \tag{3.41}$$

where $\mathcal{R}_{cv}$ has units GV, μ_E is Earth's dipole moment (Am^2), d is distance from the dipole centre in Earth radii, λ is magnetic latitude, a is Earth's radius (m), and the various other factors arise from the non-SI to SI conversion (see Smart and Shea, 2005).

Zenith Angle of Intersection

The normal to the surface defining the top of the atmosphere is a vector pointing up into space aligned with local zenith $z = 0°$. To find this vector, one can consider the equation

$$\epsilon = \frac{x^2}{a'^2} + \frac{y^2}{a'^2} + \frac{z^2}{b'^2} - 1 = 0 \tag{3.42}$$

which is true at any point on the surface, where, as for Equation 3.33, x, y and z are components of a coordinate $\mathbf{x}_{\mathrm{GEO}}$ in the GEO frame. Use of the gradient operator ∇ on ϵ then gives the normal vector to the surface, which can then be transformed to the MAG frame like so:

$$\begin{aligned} \hat{\mathbf{n}}_{\mathrm{MAG}} &= R\ \hat{\mathbf{n}}_{\mathrm{GEO}} \\ &= R(\nabla\epsilon) \\ &= R\left(\frac{2x}{a'^2}, \frac{2y}{a'^2}, \frac{2z}{b'^2}\right)^T \end{aligned} \tag{3.43}$$

The local zenith angle of the vector intersecting the atmosphere in the direction of the negative tangent from S is thus given by

$$z = \arccos(\hat{\boldsymbol{l}} \cdot \hat{\mathbf{n}}_{\mathrm{MAG}}) \tag{3.44}$$

3.3.4 Demonstrating the Effect of CRAND

Calculations of S_n via the above method extended from 2.31MeV (the minimum energy for which neutron albedo data was available from the $j_n(E_n,\mathrm{z},\mathcal{R}_{cv},\mathrm{F10.7})$ dataset) up to 500MeV. In the numerical model, the CRAND source was extrapolated below 1MeV by linearly extrapolating the gradients in S_n with respect to energy at each model L and K, then smoothing the extrapolated data to produce approximate values covering the range of low energy coordinates.

It is generally regarded that the CRAND source is not important at such low energy (i.e., Jentsch, 1981). To demonstrate its effect, Figure 3.9 shows a steady state solution of proton phase space density for equatorially mirroring protons, calculated using the diffusion coefficients of Jentsch (1981) for magnetic fluctuations, and boundary data from the MagEIS instrument aboard Van Allen Probes (discussed in Section 5). The solution has been computed with (red) and without (blue) the CRAND source for four different values of the first invariant μ as indicated. The corresponding energy of an equatorially mirroring proton is labelled in grey along the top of each plot as a function of L.

Figure 3.9 shows that CRAND is important for forming the distribution even at a few MeV at $L < 1.3$. However this is mainly because phase space density is flowed to these coordinates from higher energy where S_n provides a source comparable to radial diffusion. The difference made by CRAND is particularly pronounced at higher energy or μ, and CRAND is especially important at 500MeV/G for the region $L \lesssim 1.6$ (bottom right). More tests showed that extrapolating below 2.31MeV had a negligible effect on the ∼1MeV distribution since S_n is very small at this energy, but this was done anyway to avoid discontinuities in the model.

3.4 Drift Averaging Density and Temperature

3.4.1 Model Dependence on Density and Temperature

By considering the loss terms in the 3D model master equation (Equation 3.13), one can see that the loss rate depends on the number density of various constituents via the $\mathrm{d}\mu/\mathrm{d}t_{\mathrm{fric}}$ and Λ terms. Equation 3.15 also shows that stopping power depends

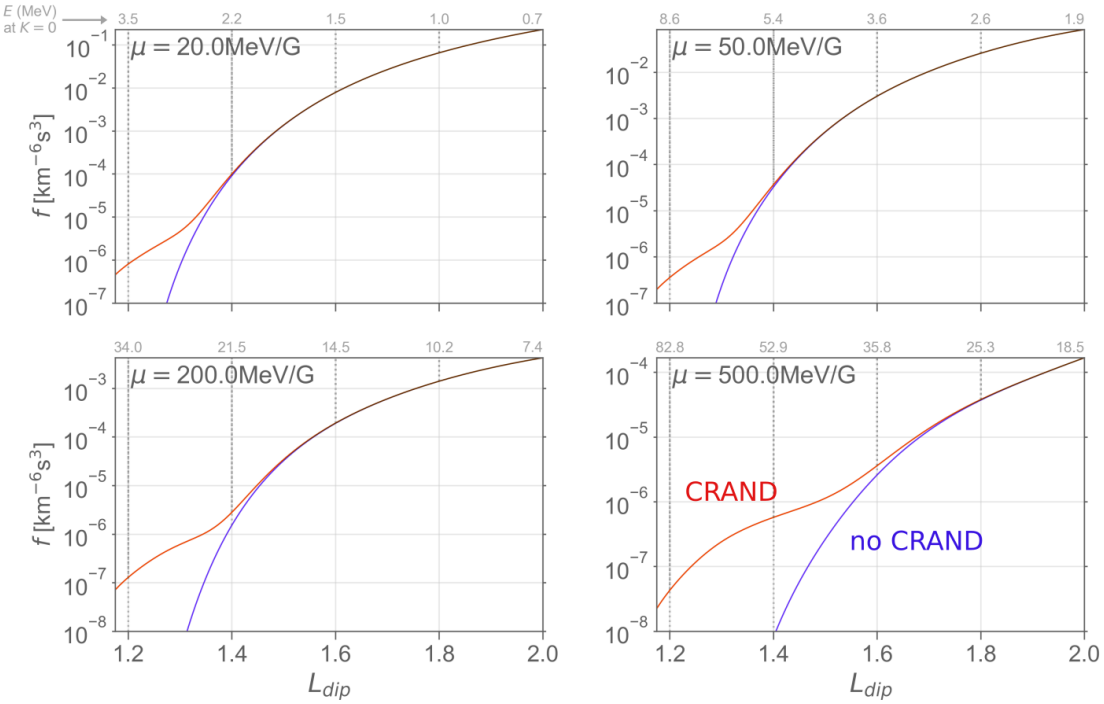


Figure 3.9: Steady state phase space density of equatorially mirroring protons as a function of L, shown in red, for four values of μ (one per panel), calculated using the magnetic diffusion coefficients of Jentsch (1981). The solution has been re-computed after disabling the CRAND source, shown in blue, to indicate where the CRAND source is responsible for the distribution of particles.

on Debye length λ_D, which is a function of ion and electron temperature given by

$$\lambda_D = \sqrt{\frac{\epsilon_0 k/e^2}{n_e/T_e + \sum_{ij} j^2 n_{ij}/T_i}} \tag{3.45}$$

where ϵ_0 is the permittivity of free space, k is Boltzmann's constant, T_e and T_i are electron and ion temperature, and n_{ij} is the density of constituent n_i with ionic charge number j. Therefore, a numerical proton belt model should incorporate calculations of density and temperature within the atmosphere, ionosphere and plasmasphere. Since Equation 3.15 only depends on $\ln(\lambda_D)$, an approximate calculation of temperature was considered acceptable and greater importance was placed on accurately determining density.

Terms depending on density and temperature must be evaluated by: a.) using previously published density and temperature models to produce datasets; then b.) reading the datasets into the model during execution and interpolating at particle coordinates. However, both density and temperature are subject to significant time variability, which should be included to some extent in dynamic simulations to produce realistic results. Furthermore it is difficult to test accuracy considering that the output of density and temperature models can include significant uncertainty, and there are usually no spacecraft measurements of density available in the inner proton belt to validate them for the era one wishes to model. A key challenge was therefore finding a way to accurately prescribe time varying density and temperature within the model.

3.4.2 Variability of Density

To understand variations in density, it is useful to consider the number density n_i of a constituent i at a test point $\mathbf{x}$ located somewhere in the atmosphere. At a given time, there are spatial variations in n_i over latitude, longitude and altitude if $\mathbf{x}$ moves over the surface of Earth. To exclude this component of variation, $\mathbf{x}$ must be fixed to the GEO or MAG frame (rotating with Earth). The time evolution of density n_i at the fixed test point includes independent variations over different timescales. These include:

- diurnal variation, as the point rotates through different MLT;

- seasonal variation, as Earth's rotational axis tilts back and forth between summer and winter solstices;
- solar cycle variation; and
- occasional dynamic changes driven by magnetic activity (see for example, Oyama et al., 2005).

Two examples have been selected from literature to illustrate observational evidence of some of these variations. As the first example, Figure 3.10 from Kakinami et al. (2011) shows ionospheric electron density at 600km altitude calculated across longitude-latitude bins from observations of the Hinotori satellite. The top panel of Figure 3.10 shows data collected between 0900-1100 hours local time, whilst the bottom panel shows data from 1300-1500. Both panels show electron density decreasing towards higher latitudes, but the density is $\sim 2\times$ higher at the equator at 1300-1500 hours compared to 0900-1100 hours, demonstrating large diurnal variability. The authors were careful to exclude other components of variability from the data in Figure 3.10, for example, by using data only from magnetically quiet conditions (Kp<4).

As a second example, Figure 3.11 from Clilverd et al. (2007) shows observed seasonal variations in plasmaspheric equatorial electron density (left ordinate) at $L = 2.5$ between June (dashed line) and December (solid line). Figure 3.11 shows that this seasonal variation depends strongly on (geographic) longitude. As in the case of Figure 3.10, the authors were careful to exclude other components of variability besides the longitudinal and seasonal variations they were trying to highlight. In this case, for each data point shown, diurnal changes in density were averaged over by incorporating observations uniformly distributed in magnetic local time, so that the longitudinal variation shown is distinct from this diurnal effect. The density of He+ is also plotted (right ordinate) in each season (December: crosses and squares; June: diamonds and triangles). It shows an anti-correlation with electron density, and the basic reason is that charge exchange processes moderate the production and loss of ions and electrons in the ionosphere and plasmasphere, tending to conserve overall charge.

A detailed discussion of the physics driving each type of variation listed above is outside the scope of this work. The key takeaway message is that isolating specific

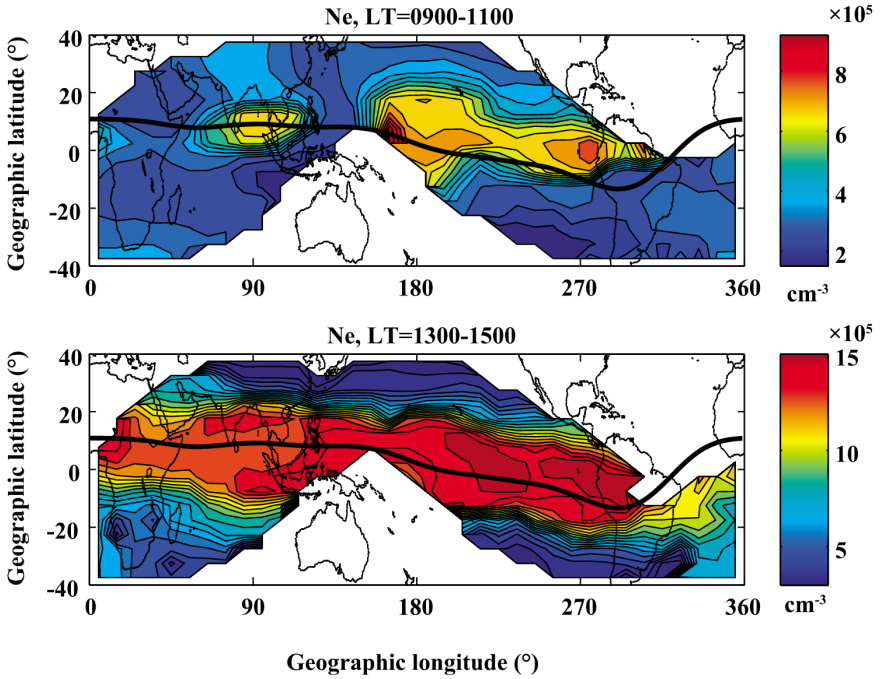


Figure 3.10: Ionospheric electron density at 600km based on observations from the Hinotori satellite. The top panel shows data from observations collected at 0900-1100 hours local time, and the bottom panel shows 1300-1500 hours. The top and bottom panels have been adapted from Figure 1 and 2 respectively of Kakinami et al. (2011)

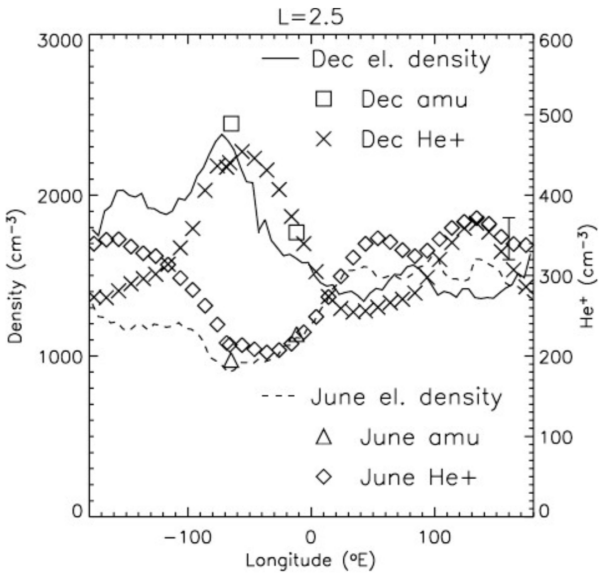


Figure 3.11: Plotted on the left ordinate: equatorial electron density variation with longitude at $L = 2.5$, for December (solid) and June (dashed). Plotted on the right ordinate: He+ density variation with longitude at the same L, also in December (crosses and squares) and June (diamonds and triangles) but in a different year, taken directly from Clilverd et al. (2007)

variations in atmospheric, ionospheric and plasmaspheric density from a model or data requires careful consideration; one must sometimes "zoom in" or out to study different time or spatial scales to provide context for an observed variation. The same applies to temperature, which can show the same types of independent variabilities.

3.4.3 Initial Attempts to Model Electron Density using the GCPM

Modelling results are sensitive to changes in density because loss timescales increase (decrease) as density decreases (increases). In early versions of the numerical model (until ∼summer 2020), the Global Core Plasma Model of (GCPM Gallagher et al., 2000, updated to version 2.4 on June 2009) was used to determine ionospheric and plasmaspheric density. The GCPM provides electron density and temperature from 90km altitude up to the plasmaspheric trough region by integrating several region-specific models: it includes a modified version of the plasmaspheric density model by Carpenter and Anderson (1992), and bridges it to the International Reference Ionosphere model (Bilitza and Reinisch, 2008) using an exponential function. The GCPM provides density as a function of time, MLT, and Kp-index. The time input is a proxy for including seasonal variation and solar cycle variation, though the GCPM does not model other time-dependent changes.

The numerical model was still 2D at the time of using the GCPM, giving phase space density of equatorially mirroring particles as a function of (μ, L). Therefore, density from the GCPM could be read into the model very simply without using drift averages. Electron density given by the GCPM for a specific set of inputs was averaged in longitude and determined at regular intervals along the magnetic equator to produce a profile of density versus L. The profile of electron density versus L was read directly into the numerical model and used to calculate loss. However, the contribution to loss from all other constituents of the atmosphere, ionosphere and plasmasphere was ignored.

Figure 3.12 shows three different profiles of electron density versus L (red, yellow and blue lines) derived using the GCPM in the left hand column. In the right hand column, calculations of steady state phase space density at $\mu = 100$MeV/G

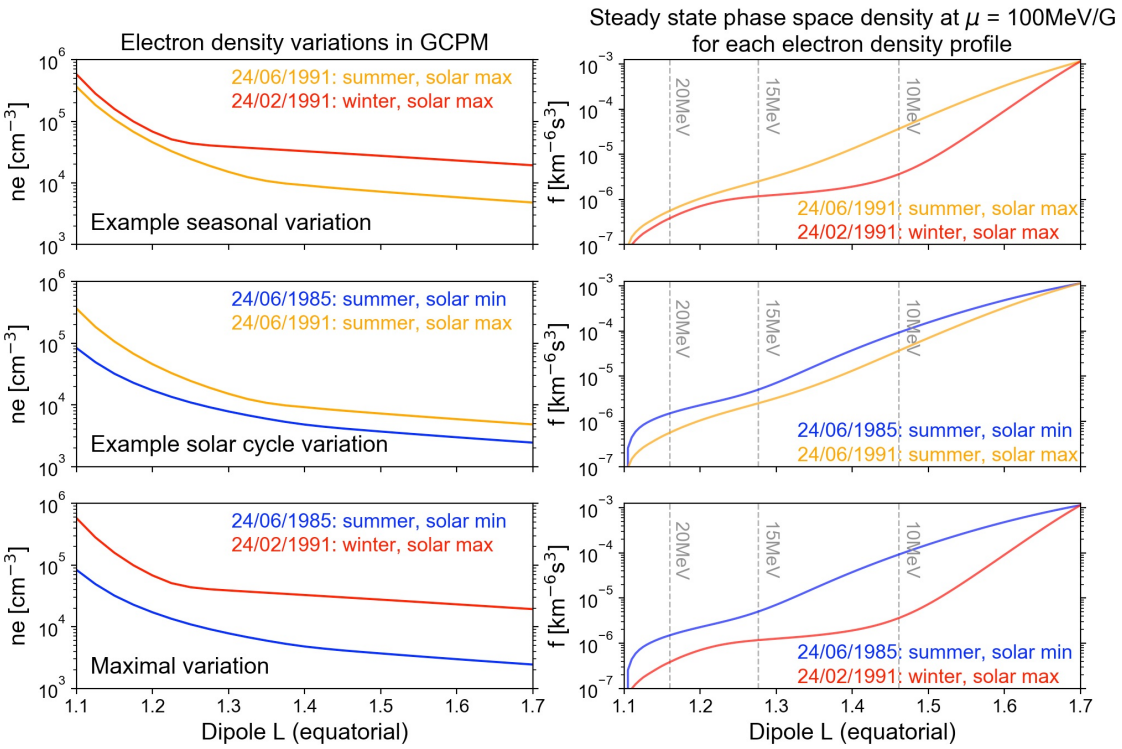


Figure 3.12: Plasmaspheric electron density according to the GCPM model of Gallagher et al. (2000) (left panels) demonstrating seasonal (top) and solar cycle (middle) components of variation along with both components together (bottom), plotted next to corresponding steady state solutions in phase space density (right panels) incorporating those values of density but leaving all other model inputs constant. Phase space density shown is for equatorially mirroring ($\alpha_{eq} = 90°$) particles. Grey vertical lines indicate selected energies corresponding to different L.

using the early 2D numerical model are shown. The colour of each model solution corresponds to the electron density profile used to perform the calculation, with all other model parameters unchanged between solutions. The model outer boundary was set at $L = 1.7$ using proton flux measurements from the PROTEL instrument on the CRRES satellite. Each panel on the left compares two electron density profiles to show seasonal, solar cycle, and maximum possible variation according to the GCPM results.

Figure 3.12 demonstrates that, according to the GCPM, significant changes in density can occur due to seasonal and solar cycle variation. This, in turn, causes

large changes in steady state phase space density. It was found that the region where a steady state solution exhibits the most variation due to a change in density depends upon the diffusion coefficients used (with higher diffusion rates pushing the region of maximum variability inwards). However, at this stage there were significant issues with model accuracy: even by varying the diffusion coefficients freely using an optimisation algorithm, no model solution could be generated that agreed well with PROTEL data describing flux in the region $\sim 1.2 \leq L \leq 1.7$, when using the "period correct" density profile according to the GCPM.

It was impossible to know for sure why the model was unable to match the data. One reason could be that the PROTEL data was not in steady state, and another reason could be because ionospheric/plasmaspheric constituents aside from electrons were not taken into account by the model. However, with hindsight a likely contributor was significant overestimation of electron density given by the GCPM. Figure 3.13 compares equatorial electron density predicted by the GCPM with a least squares fit to measurements taken by the Radio Plasma Imager instrument aboard the IMAGE spacecraft, presented in Ozhogin et al. (2012). The GCPM profile shown in Figure 3.13 is the time average over the same period as the measurements, with the red shaded area indicating standard deviation. The modelled profile shows an overestimation of 2-3× compared with data at $L \sim 1.5$. Ozhogin et al. (2012) suggest that at high L, the GCPM may be based on measurements of low electron density from outside the plasmapause that were included erroneously, leading to an overly steep gradient as density increases towards low L. Attempts were made to make empirical corrections to the density profiles derived using the GCPM, and whilst this somewhat improved agreement between model and data, it was decided that such modifications could not be justified on a scientific basis.

3.4.4 Composing a Global Model of Drift Averages

Following the difficulties caused by uncertainty in ionospheric/plasmaspheric electron density, the numerical model was improved to evaluate loss terms $\mathrm{d}\mu/\mathrm{d}t_{\mathrm{fric}}$ and Λ as accurately as possible. For 3D modelling, this requires determining the drift averaged number density of electrons $\langle n_e \rangle$ and other neutrals/ions $\langle n_i \rangle$, as well

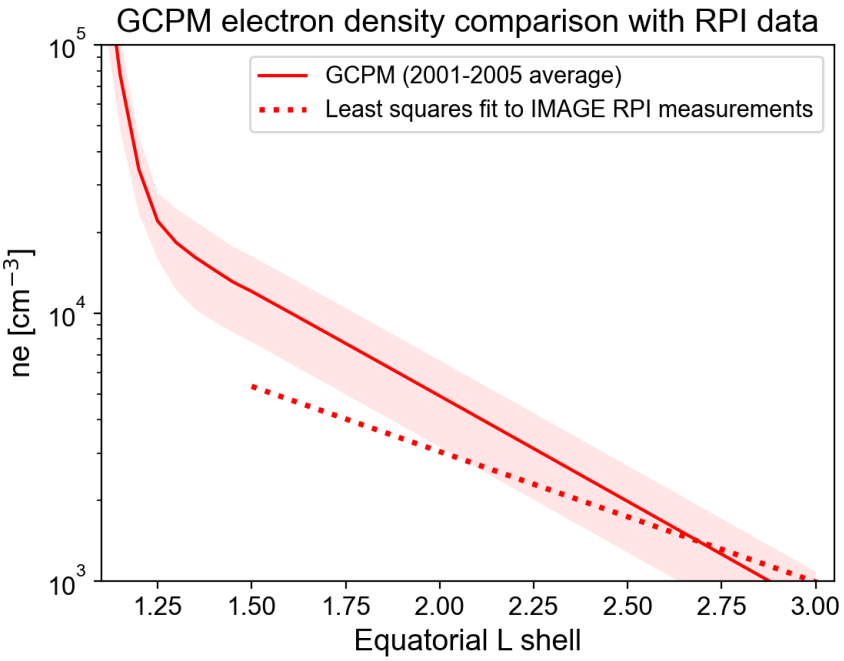


Figure 3.13: Comparison between equatorial electron density as modelled by the GCPM versus a fit to measurements from the RPI instrument on IMAGE. Shaded red indicates the standard deviation of GCPM results around the time averaged profile.

as drift averaged electron and ion temperatures $\langle T_e \rangle$ and $\langle T_i \rangle$ at adiabatic invariant coordinates. Furthermore, to produce realistic modelling results, it was decided that the solar cycle and seasonal dependence of these terms should be included in the proton belt model. Since models of the proton belt from previous literature have parameterised density only by solar cycle (i.e., Selesnick and Albert, 2019), the inclusion of seasonal variations would be somewhat novel.

However, to begin calculating $\langle n_e \rangle$, $\langle n_i \rangle$, $\langle T_e \rangle$ and $\langle T_i \rangle$ for a given solar cycle and seasonal phase, one must first have the capability to find density and temperature along a drift path in the region of interest (i.e. as a function of position in the MAG frame). No existing model was able to provide density and temperature over the entire proton belt region in order to permit this computation. Therefore, it was decided to compose a new "global" model that could compute average density and temperature along a drift orbit. This would be achieved by integrating together previously published models describing different parts of the atmosphere, ionosphere and plasmasphere.

The newly composed global model was integrated with drift averaging code, so as to output $\langle n_e \rangle$, $\langle n_i \rangle$, $\langle T_e \rangle$ and $\langle T_i \rangle$ when a proton trajectory is supplied as input. The procedure to calculate a 3D grid of values for some drift averaged quantity $\langle A \rangle$ covering adiabatic invariant coordinate space is explained conceptually in Section 3.2.2. Within this process, the global drift averaged density and temperature model constitutes the "script to evaluate $\langle A \rangle$" shown in Figure 3.4.

3.4.4.1 Using Data from Existing Models

To construct a global map of density and temperature for drift averaging, model data was utilised to obtain the following 16 quantities:

- neutral density of He, O, N2, O2, Ar, H, N and anomalous O, given by the Mass Spectrometer and Incoherent Scatter Radar model (NRLMSISE-00) (Picone et al., 2002);
- ion density of O+, H+, He+, O2+ and NO+, given by the International Reference Ionosphere model (IRI) up to 2000km (Bilitza et al., 2017);
- electron density, given by IRI up to 2000km, and above 2000km by the

Ozhogin et al. (2012) plasmasphere model;

- electron and ion temperatures, given by IRI up to 2000km.

The NRLMSISE-00 and IRI models are written in Fortran, but the Python module PyGlow wraps them (and several other climatological models) so that they can be called from Python. The PyGlow method for calling these models requires universal time as an input. Time is a proxy for including seasonal, solar cycle and magnetic variability, and is converted internally to an index corresponding with each type of variation modelled, such as F10.7 for short term solar cycle variation. This method for calling the models was convenient because changes in density and temperature could easily be plotted through time, the NRLMSISE-00 and IRI models could be called using the same object input, and because the drift averaging code and CRAND evaluation code was written in Python, so Python code could be more easily integrated into the workflow. Therefore, the PyGlow library was used to access the NRLMSISE-00 and IRI models. The inputs controlling each of the three models were as follows.

- NRLMSISE-00 density depends on universal time, geographic latitude, longitude and altitude.
- IRI density and temperature depends on universal time, geographic latitude, longitude and altitude.
- The Ozhogin et al. (2012) plasmasphere model depends on magnetic latitude λ and L.

The latter model is based on the fit to Radio Plasma Imager data shown in Figure 3.13, with an added field line dependence. The model gives electron density as a function of magnetic latitude and L, according to

$$\begin{aligned} n(L,\lambda) =& n_{eq}(L)\ \cos^{-0.75}\left(\frac{\pi}{2}\frac{1.01\lambda}{\lambda_{\text{INV}}}\right) \\ n_{eq}(L) =& 10^{4.4693-0.4903L} \end{aligned} \tag{3.46}$$

where λ_{INV} is invariant latitude. Equation 3.46 is as presented in Ozhogin et al. (2012, Equation 2). It is simple to implement, and valid at >2000km up to the

plasmapause. In regions where $1.01\lambda > \lambda_{\mathrm{INV}}$ (high λ and low altitude), the model returns unphysical density values as the cosine dependence breaks down.

It was decided to use a day of the year variable (DOY) to parameterise seasonal variation, and 81-day averaged solar radio flux (F10.7a) to parameterise long term solar cycle variation. Since the new density-temperature model would ultimately be needed for drift averaging, the aim was therefore to unify the models such that density and temperature could be found as a function of constituent, spatial position in the MAG frame, DOY and F10.7a anywhere in the proton belt, i.e. $n_i = N(\mathbf{x}, i, \mathrm{F10.7a}, \mathrm{DOY})$ and $T_i = T(\mathbf{x}, i, \mathrm{F10.7a}, \mathrm{DOY})$. However, there were challenges to address in order to unify the models this way. These included:

- The density and temperature data given by IRI was only available up to 2000km. An altitude profile has been plotted in Figure 3.14, showing the density of a few selected constituents returned by IRI, with the 2000km limit visible.

- As per Equation 3.46, the plasmaspheric electron density model does not contain seasonal or solar cycle dependence (since the spacecraft dataset on which it is based only includes five years of data). Figure 3.15 shows electron density plotted 20 times by IRI along the same magnetic equatorial altitude profile but using different solar cycle/ seasonal conditions. Also plotted on Figure 3.15 is the electron density given by the Ozhogin et al. (2012) model. These models were to somehow be combined in such a way as to preserve the solar cycle and seasonal dependence of IRI, but use the high altitude density from the Ozhogin model.

- When drift averaging to find density/temperature over a proton trajectory, values cannot be returned directly from IRI because the computation takes too long given the many points per trajectory. Therefore, for certain constituents where modelling involved IRI data, pre-calculated altitude profiles of density and temperature were to be derived at different latitudes across the modelling region and interpolated from, reducing execution time.

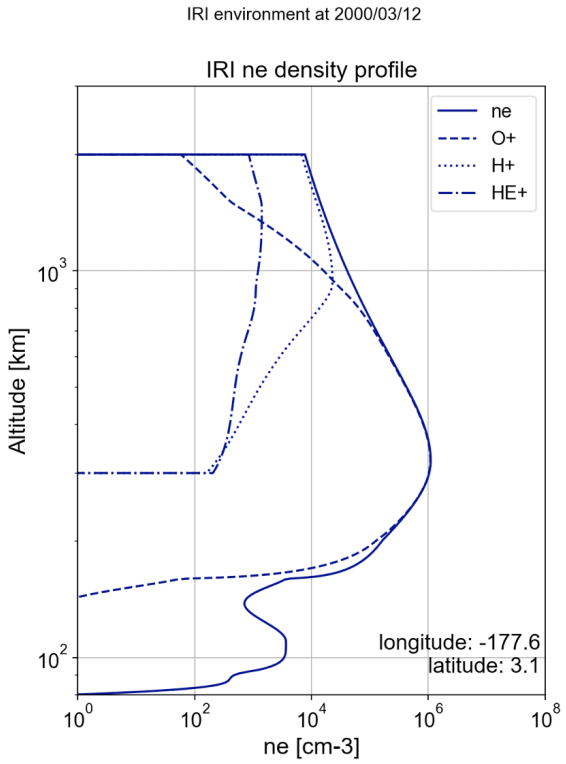


Figure 3.14: An example, near equatorial, altitude profile of electron, O+, H+ and He+ density produced using the IRI model for a given epoch (on 12/03/2000). The abrupt zeroing of densities at 2000km is due to the altitude range of the model being exceeded.

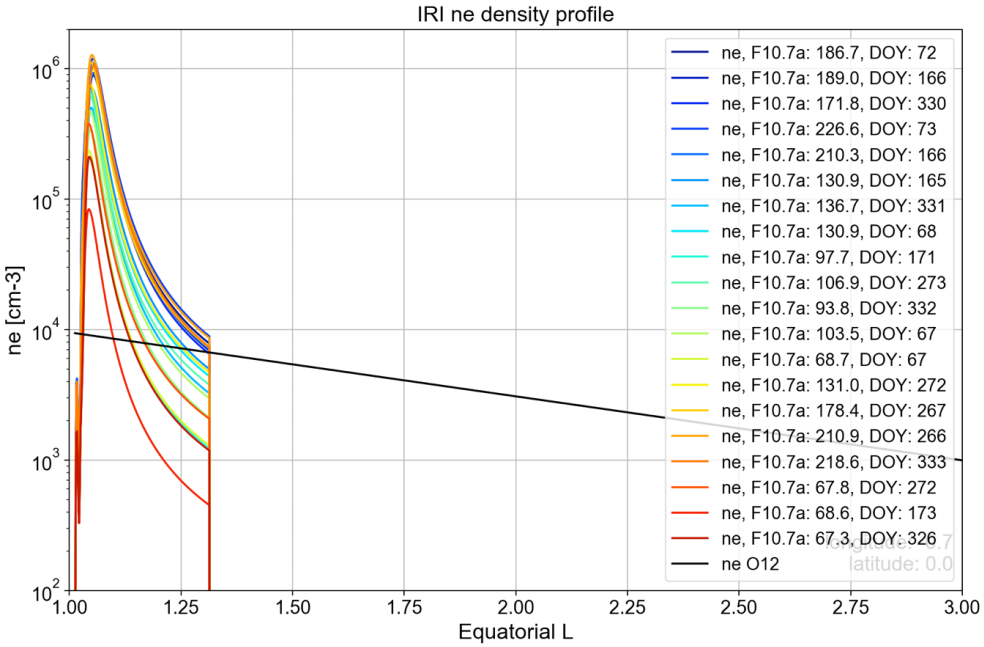


Figure 3.15: 20 plots of IRI electron density along the same altitude profile located at the magnetic equator using different combinations of seasonal/ solar cycle inputs as shown, along with the single plot of electron density given by the Ozhogin et al. (2012) model.

3.4.4.2 Isolating Solar Cycle and Seasonal Dependence

It is possible to make drift averaged density and temperature a function of solar cycle and season by re-calculating it several times using specific values of F10.7a and DOY as inputs to the NRLMSISE-00 and IRI models. One can then interpolate between the results for any combination of F10.7a, DOY. A minor complication is that, since the input (via PyGlow) to the NRLMSISE-00 and IRI models is universal time, the F10.7a and DOY variables cannot be directly controlled, but instead correspond to the time entered via a built-in time history of indices. To overcome this complication and call the models for specific combinations of F10.7a and DOY, one can go backwards in time from a starting epoch in small steps, check the F10.7a and DOY corresponding to each new epoch using the time history of indices, and repeat until the desired combination of F10.7a, DOY is found at some actual point in history (allowing for some error since a very specific F10.7a may never have been recorded exactly). However, as discussed in Section 3.4.2, density and temperature can be affected by numerous other types of independent variability at any one time. Many of these are modelled and dependent on the epoch, so one must be careful to avoid capturing an unintended effect from the model this way.

To parameterise solar cycle and seasonal dependence in the drift average density model, five different values of F10.7a were considered (60, 100, 140, 180, 220 solar flux units, sfu) along with four days of the year (DOY; day 70, 170, 270, 330). These sets of values constitute a parametric matrix with 20 elements: one for each possible combination of F10.7a and DOY. Using the trick above, a corresponding universal time for each element was found by searching through a time history of indices (backwards from the present) until a date was found where F10.7a matched within a margin of ± 10 sfu, and the DOY matched within a margin of ± 4 days. Using these 20 universal times, the NRLMSISE-00 and IRI models could be called for a specific combination of F10.7a and DOY from the parametric matrix.

Whilst investigating the time evolution of densities and temperatures in each model, it was found that density returned by the NRLMSISE-00 and IRI models at an atmospheric test point occasionally showed large time-dependent changes. Figure 3.16 shows numerous examples of this occurring over a year, for a test

point with a fixed latitude, longitude and altitude of 30°, -80° and 1000km. In the top panel, various magnetic (left ordinate) and solar indices (right ordinate) are plotted. In the middle panel, the total mass density of every atmospheric constituent according to NRLMSISE-00 is shown. In the bottom panel, the electron mass density is shown according to IRI. The variation in Figure 3.16, appearing as ∼days long blips in atmospheric mass density and electron density, appears to correlate with temporary jumps in daily AP index (top panel, grey), indicating magnetic activity. Note that the variation is only identifiable because a 24 hour time step was used to derive the time series in each panel, and this had the effect of excluding day-night variation in density, which is of a comparable magnitude. If the time of each data point were shifted by a few hours, the blips would still be visible since a range of local times are affected, but each plot of density would have a different trend/ shape.

It was important to ensure these rapid changes in density did not coincide with any of the universal times used as proxies for each set of F10.7a and DOY coordinates in the parametric matrix. This is because, to effectively parameterise each drift average by solar cycle and season, changes caused by magnetic activity must be excluded to isolate solar cycle and seasonal variability. Figure 3.16 shows enhancements in total mass density (middle panel) but decreases in electron density (bottom panel) during the disturbance. To investigate further, Figure 3.17 shows the IRI electron density profile over six days during one of the magnetic disturbances. The decrease in electron density appears to begin from ∼150km, near the base of the ionosphere F region, and could be a response to a composition change in the ionosphere driven by magnetic activity.

Effects on density due to this type of magnetic variability were excluded from calculations of drift averages at each F10.7a, DOY combination by manually checking that each of the 20 universal times corresponded to magnetically quiet conditions and by checking for any blips in density similar to those shown in Figure 3.16. Finally, to test whether this method of capturing solar cycle and seasonal variability was effective, density at a fixed test point was then probed at each of the 20 times for different solar cycle and seasonal conditions, for two MLTs on opposite sides of Earth. The results of this test are shown in Figure 3.18.

Figure 3.18 shows that there is a clear solar cycle and seasonal dependence in

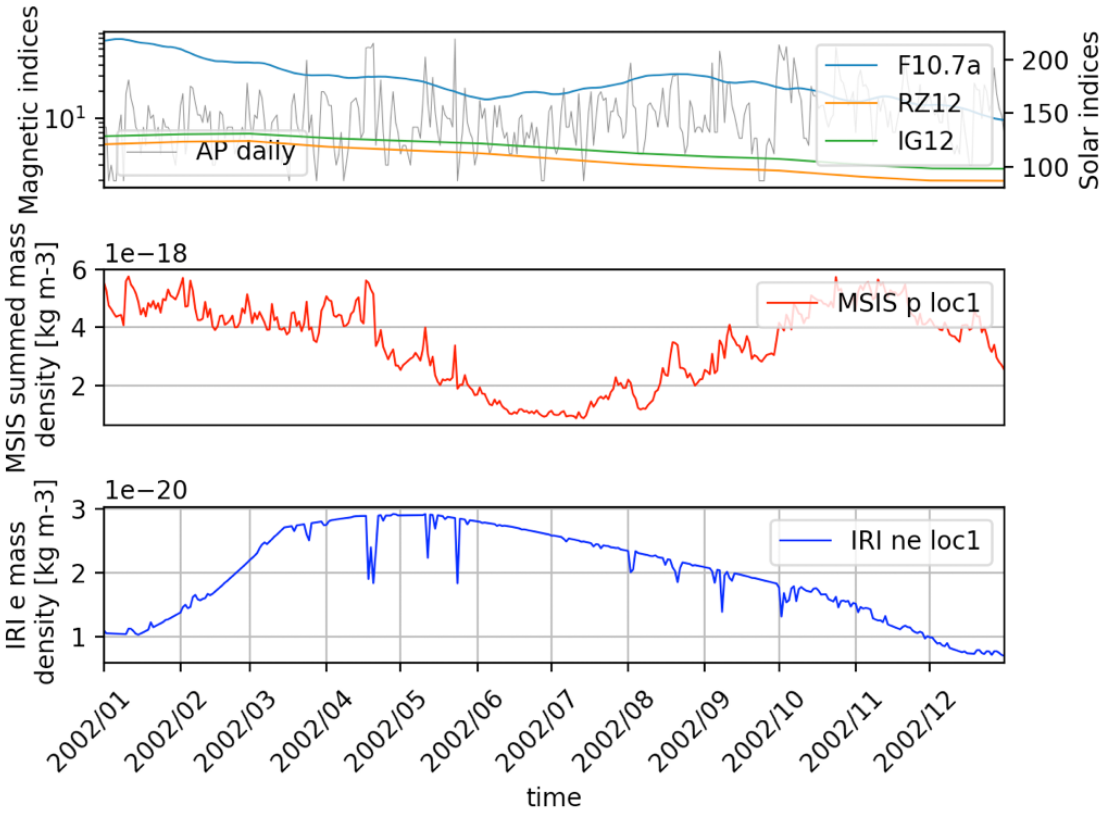


Figure 3.16: Overview of time variations in density over a one year period, as modelled at a test point with a fixed longitude, latitude and altitude in Earth's atmosphere. The top panel plots magnetic (left ordinate) and solar indices (right ordinate) against time. The middle panel shows total mass density of every atmospheric constituent according to NRLMSISE-00. The bottom panel shows electron density according to IRI. The fast ∼day long variations in both density profiles are caused by magnetic activity, correlating with magnetic AP index.

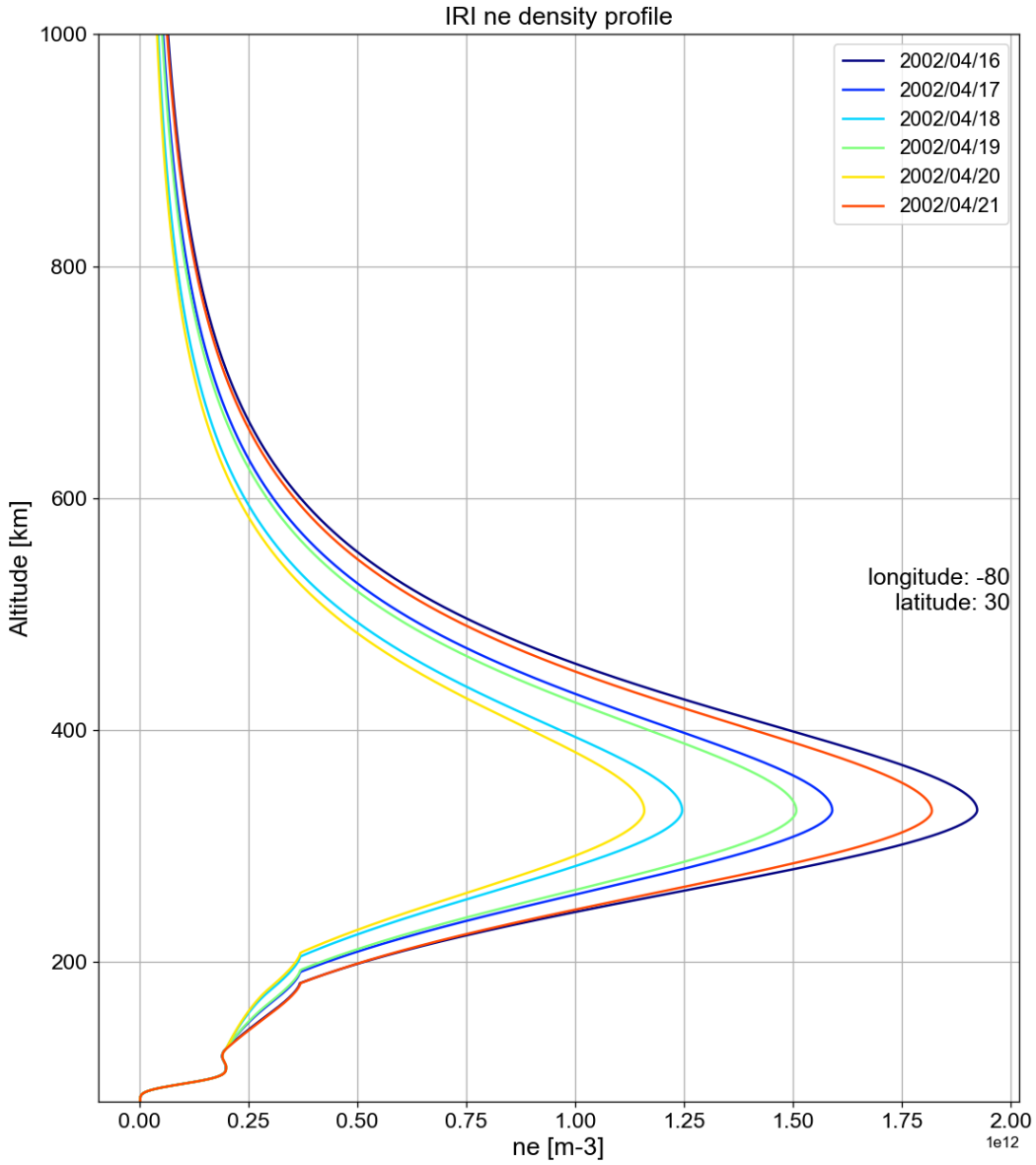


Figure 3.17: Electron density modelled by IRI along an altitude profile at a fixed latitude and longitude, over the duration of a dynamic change correlating with a jump in magnetic AP index (shown in Figure 3.16).

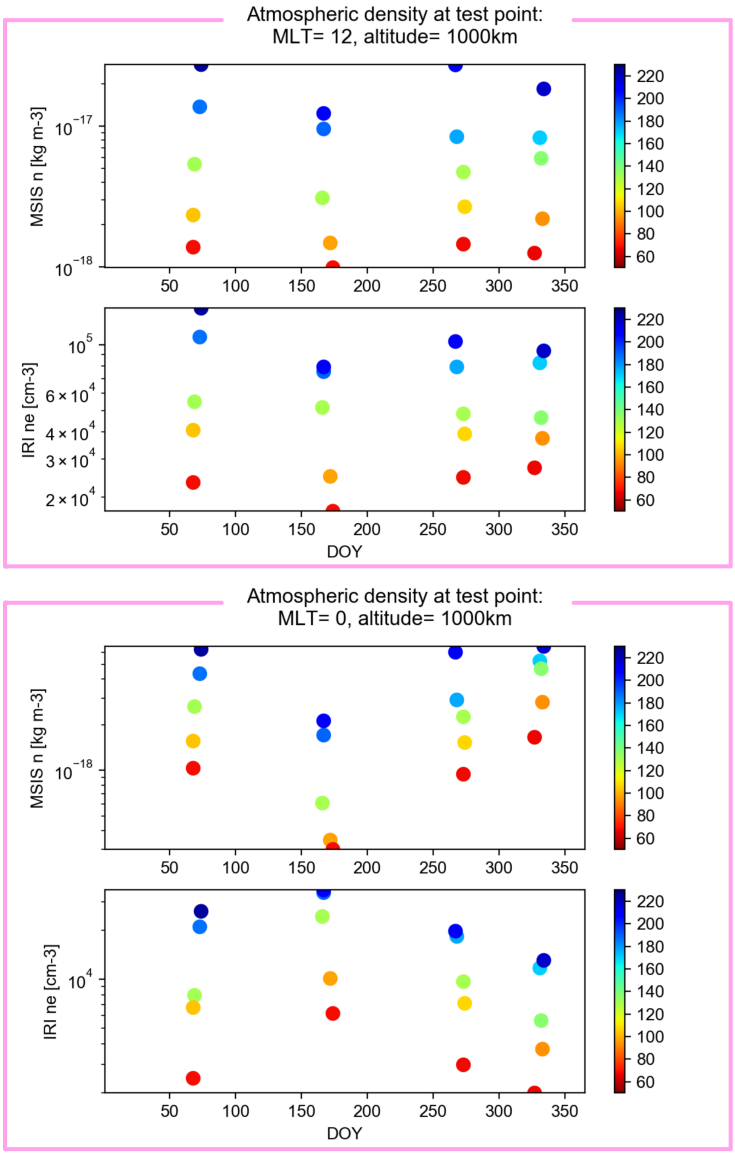


Figure 3.18: Density at an atmospheric test point fixed in the MAG frame with 1000km altitude and 0° magnetic latitude is considered. In the top two panels, the test point is at MLT= 12. The top panel shows total mass density of neutral atmospheric constituents at the test point, calculated using the NRLMSISE-00 model. The panel beneath shows electron density at the test point, calculated using IRI. There are 20 values of density given in each plot, each corresponding to one of the 20 elements of the parametric matrix specifying values F10.7a (colour coded) and DOY (along x axis). In the bottom two panels, 12 hours have passed and the test point is at MLT= 0. The third and fourth panels down show the same information as top two panels but for MLT= 0.

density at the test point, regardless of MLT. For example, for each fixed F10.7a (colour), density has the same up-down variation throughout the year at the test point. Likewise, for a fixed DOY, density always increases as F10.7a increases.

3.4.4.3 Interpolating and Extrapolating Density and Temperature

After establishing a method for isolating solar cycle and seasonal dependence from the NRLMSISE-00 and IRI models, the global drift average model could be developed. The method for calculating drift averaged density and temperature, described below, was repeated for each of the 20 environments defined using the parametric matrix, with each repetition giving a result for a particular phase over solar cycle and season.

The top left panel of Figure 3.19 illustrates the first step in the method: altitude profiles were constructed at regular intervals in magnetic latitude, each ranging from 90km up to the $L = 3.25$ field line. The aim was to pre-calculate density and temperature along these altitude profiles, so that density and temperature could be interpolated anywhere within the $L = 3.25$ field line. The higher limit of $L = 3.25$ encompasses most of the proton belt, and was dictated by reliability of electron density measurements from the Ozhogin et al. (2012) model, discussed later in this section. The altitude profiles shown in Figure 3.19 were fixed at the following magnetic latitudes: 0, ±10.0, ±20.0, ±30.0, ±35.0, ±40.0, ±42.5, ±45.0, ±47.5, ±50.0, and ± ∼50.5184°. The higher and lower magnetic latitude limits are where the $L = 3.25$ field line intersects 2000km altitude in a dipole field.

Along each altitude profile, neutral densities of He, O, N2, O2, Ar, H, N and anomalous O were determined by NRLMSISE-00 below 2000km. Above this height, they were extrapolated as straight lines of $\log10\,(n_i)$ with altitude, up to the $L = 3.25$ field line. Ion densities of O+, H+, He+, O2+ and NO+, as well as electron and ion temperature, are given by IRI below 2000km and were extrapolated to higher altitudes along each profile via the same method. Ion and electron temperatures exhibit a similarly complex dependence on local time, etc., as density, but generally do not exceed 10000K in the region of interest (see for example the plots in Kutiev et al., 2002), and therefore a hard limit was set so that T_e and T_i did not exceed 10000K.

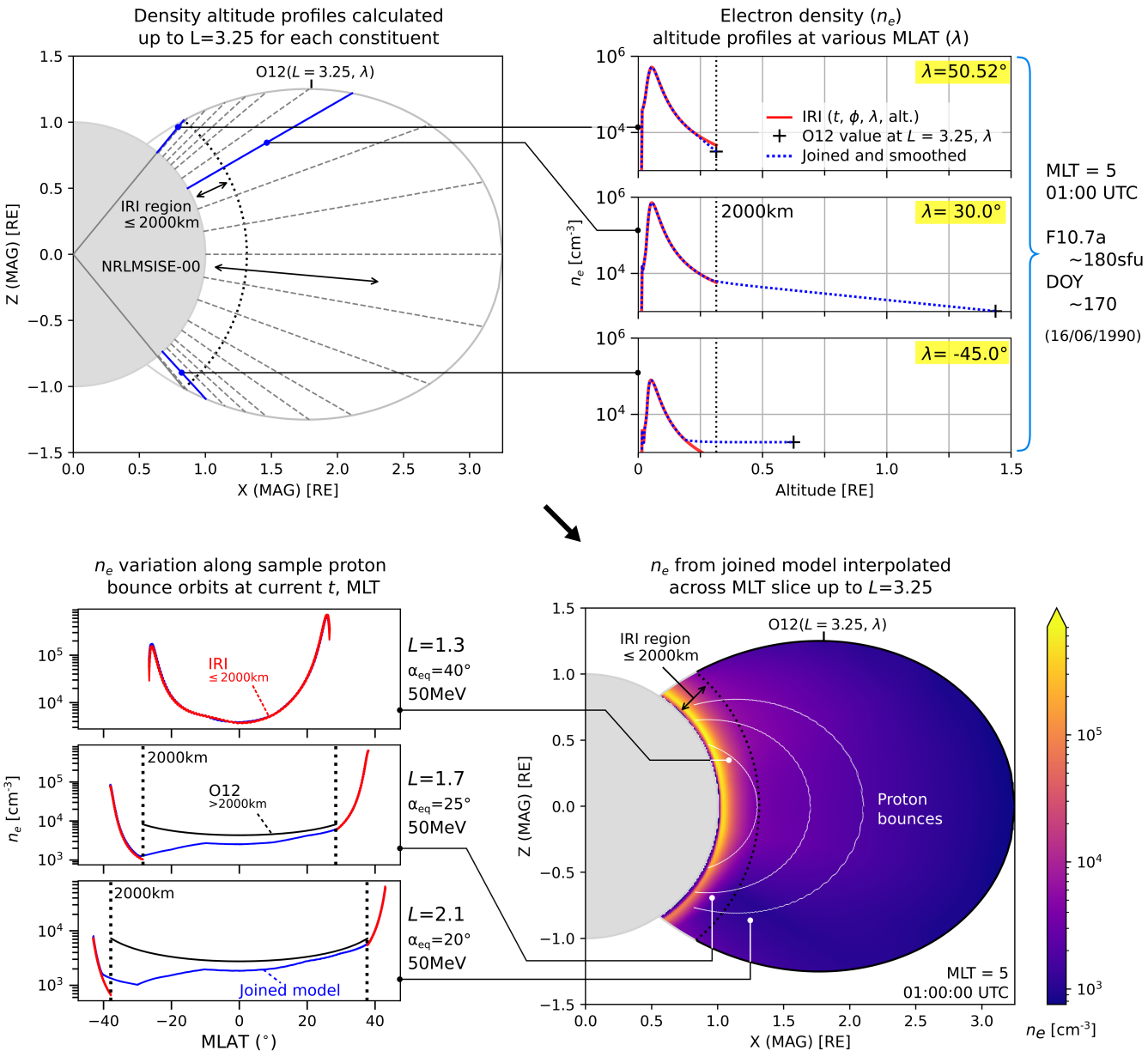


Figure 3.19: An overview of the construction of the drift averaged density and temperature model (for the case of electron density). The top left panel shows density profiles calculated at regular intervals in magnetic latitude, derived by interpolating and smoothing between the values given by IRI at $\leq$ 2000km and the Ozhogin et al. (2012) model at $L = 3.25$, as shown in the top right panel for three selected altitude profiles. The bottom right panel shows the result of interpolating between these altitude profiles of electron density across an MLT slice, and the bottom left panel shows the variation in electron density along three example proton bounce orbits at this MLT as a function of magnetic latitude (each bounce orbit is also drawn on the bottom right panel).

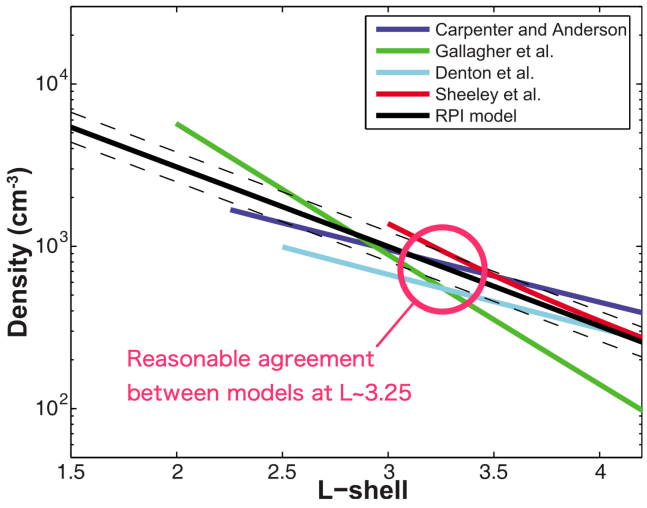


Figure 3.20: Comparison of equatorial density models, adapted from Figure 8 of Ozhogin et al. (2012) to highlight the relatively good agreement at $L \sim 3.25$. Dashed lines represent statistical uncertainties of the Ozhogin et al. model (shown in black).

In the case of electron density, extrapolation was avoided to minimise uncertainty, since uncertainty in electron density had been the cause of problems in the earlier model version as discussed in Section 3.4.3. Historical measurements indicate relatively consistent electron density at $L = 3.25$ below the plasmapause. This is demonstrated by good agreement between empirical models, shown in Figure 3.20, adapted from Ozhogin et al. (2012). This is also demonstrated by Park et al. (1978), who plot profiles of n_e derived using whistler wave observational data for the month of June 1959, 1965 and 1973, shown in Figure 3.21. The authors interpret this data to assert that density of the plasmasphere is not sensitive to solar cycle variations beyond $L \sim 3$. The suggestion that plasmaspheric electron density is relatively stable at $L = 3.25$ formed an important assumption upon which to construct the density model: electron density was given by IRI below 2000km, but above 2000km was interpolated between IRI and the O12 model value at $L = 3.25$.

However, interpolating electron density between the IRI and Ozhogin et al. (2012) models is not necessarily simple. The disagreement between the models is shown in Figure 3.15, and a straight line interpolation between IRI at 2000km and the Ozhogin et al. (2012) model at $L = 3.25$ may result in a positive gradient with

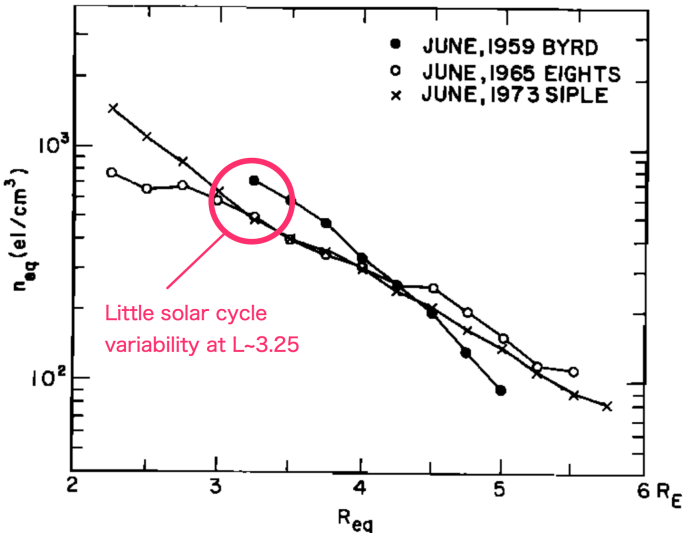


Figure 3.21: Equatorial electron density derived from whistler wave observations for June 1959, 1965 and 1973, adapted from Figure 10 of Park et al. (1978)

L (unphysical). The second step in the modelling process was dealing with this by smoothing each electron density profile across the transition between models, whilst ensuring a physical profile. This was achieved by defining a transition region near 2000km and using Bézier curves to connect the logarithm of electron density profile from IRI to the static value of Ozhogin et al. (2012). The top right panel in Figure 3.19 demonstrates this, showing electron density along three altitude profiles at different magnetic latitudes. The dashed blue line in each of these profiles represents the final electron density profile as a result of smoothing the values between IRI (shown in red) and the static Ozhogin et al. (2012) model at L=3.25 (black + symbol). The field line dependence of the Ozhogin et al. (2012) model (Equation 3.46) was made use of to derive the electron density along $L = 3.25$. All values of electron density shown in Figure 3.19 correspond to the environmental indices displayed to the right of the top right panel in Figure 3.19 (via IRI dependence).

Pre-calculating altitude profiles of density in this way for each constituent allowed faster execution time when drift averaging, and overcame the challenge of IRI being slow to call at every point over a bounce path. To demonstrate the effect of smoothing electron density as described, and to demonstrate how density can be interpolated between the altitude profiles, electron density has been interpolated

across magnetic latitude to construct a density slice in the meridian plane for a specific magnetic local time (MLT), shown in the bottom right panel of Figure 3.19. It is important to highlight that electron density shown in the bottom right panel is calculated for a specific time of day, MLT, F10.7a and DOY.

The next stage was to drift average density and temperature along a provided particle trajectory. The bottom left panel of Figure 3.19 gives an insight into this process, showing the variation of electron density versus magnetic latitude along three proton bounce orbits, derived using density across the MLT slice shown in the bottom right panel. The energy, equatorial pitch angle and L of the proton bounce is labelled next to each profile, and each orbit is also drawn onto the MLT slice in the bottom right panel. Density returned directly from the IRI model along each bounce path is shown in red up to 2000km, the maximum height of IRI. Likewise, density returned directly from the Ozhogin et al. (2012) model is shown in black above 2000km, which is the minimum height of this model. The second and third bounce orbits shown in the panel cross 2000km, and for these two orbits Figure 3.19 shows the mismatch in electron density at 2000km (red versus black) according to IRI and the Ozhogin et al. (2012) model. The discrepancy depends on MLT, time, etc., and would introduce steep gradients into the drift averaging results if not smoothed. The blue line in each profile represents the smoothed electron density derived as described above by combining IRI values up to 2000km and Ozhogin et al. (2012) model values at $L = 3.25$. Each bounce orbit mirrors at low altitude close to the loss cone, and therefore density is high at the mirror points as each proton penetrates the atmosphere.

Since density and temperature depend on MLT, and the proton trajectories supplied to the global drift average model are bounce trajectories due to the shortcut explained in Section 3.2.4, the densities and temperatures stored in each altitude profile via the above method were re-calculated and averaged across the following eight MLTs: 0, 3, 6, 9, 12, 15, 18 and 21 hours. This is equivalent to rotating a bounce orbit around the centre of the MAG frame in 45° intervals and averaging drift averaged density and temperature across it until the initial phase is reached to recreate the effect of a whole drift orbit. To eliminate day-night variation in density (Figure 3.10), this entire process was then repeated six times throughout the day at the following selected times: 00:00, 04:00, 08:00, 12:00, 16:00 and 20:00.

In summary, altitude profiles were:

- determined at 21 magnetic latitudes,
- for eight different MLT
- for six times throughout the day.

This produced altitude profiles that were MLT and diurnally-averaged as a result of 48 re-calculations.

Drift averages were then performed for a given particle trajectory for 20 different universal times corresponding to specific F10.7a, DOY combinations. This process required recalculating each of the 21 MLT and diurnally-averaged altitude profiles 20 times. Furthermore, there were 14 constituent densities and two temperatures modelled, which each required separate storage during this process, but drift averages for each of these 16 quantities were calculated and output simultaneously for a given F10.7a, DOY.

The final result of this method was a set of $\langle n_e \rangle$, $\langle n_i \rangle$, $\langle T_e \rangle$ and $\langle T_i \rangle$ for each environment specified in the parametric matrix. This produced 320 .mrda drift average files in total (16 constituents $\times$ 20 combinations of F10.7a, DOY), each containing the result of drift averages along the same 169280 adiabatic invariant trapped particle coordinates. As payoff for this long-winded method, drift averaged density and temperature was thus made available to the proton belt numerical model as a function of particle coordinate μ, α_{eq} and L, as well as F10.7a and DOY, for each of the 14 constituents and two temperatures, based on data from up to date models and some simple assumptions.

One useful feature of the drift averaged density and temperature model is that it can easily be expanded to cover higher L in the future, since the electron density at the outer limit $L = 3.25$ is given by the simple function defining the Ozhogin et al. (2012) model (Equation 3.46). This expansion may be necessary in the future to study magnetic time variability in density outside the plasmasphere, and a simple approach would be applying an empirical correction to density extrapolated past $L = 3.25$ from within the numerical model to recreate the effect of time variability at higher L.

3.5 Numerical Scheme

The proton belt numerical model was written in modern Fortran. In this section, a fully implicit numerical scheme is presented to solve the 3D master equation given by Equation 3.13. However, early development of the model focused on solving a more basic 2D master equation, and the 2D case will therefore be the starting point for discussion.

3.5.1 First Attempts

The Fokker Planck equation appears in Albert et al. (1998) with the following form based on the work of Cornwall (1972):

$$\frac{\partial f}{\partial t} = L^2 \frac{\partial}{\partial L}\left[\frac{D_{LL}}{L^2}\frac{\partial f}{\partial L}\right] + \frac{G(L)}{\mu^{1/2}}\frac{\partial f}{\partial \mu} - \Lambda f + S_n \tag{3.47}$$

where phase space density $f(\mu, L)$ describes the distribution of equatorially mirroring particles, given by $f = m_0^3 j/p^2$ with units $\mathrm{km}^{-6}\mathrm{s}^3$, where j is unidirectional differential proton flux, p is non-relativistic momentum and m_0 is proton rest mass. $G(L)$ in Equation 3.47 is the coulomb energy degradation factor given by Cornwall (1972) as

$$G(L) = 50\sqrt{2}\pi \frac{e^4 m_p^{1/2}}{m_e B^{3/2}} n_e \tag{3.48}$$

where (for the required units of $\mathrm{MeV}^{3/2}\mathrm{G}^{-3/2}\mathrm{s}^{-1}$) e is the electron charge (in CGS units), n_e is electron number density (cm^{-3}), m_e and m_p are the electron and proton mass (g), and B is the magnetic field strength (G). Comparing with Equation 3.13, the $G(L)/\mu^{1/2}$ term is equivalent to $\mathrm{d}\mu/\mathrm{d}t_{\mathrm{fric}}$.

The proton belt numerical model was initially based on this master equation, in order to use the work of Albert et al. (1998) as a starting point and for validation. Unfortunately, the equation is non-relativistic. It is also missing a term where $\partial/\partial\mu$ operates on $\mu^{-1/2}$, and therefore coulomb collisional loss is not fully modelled. This is explicitly justified by Cornwall (1972) on the grounds that "charge-exchange losses dominate", because that work focused on modelling energy ranges of "$\lesssim 1$ MeV/nucleon at $L = 3$". The authors also claim that this extra term cancels

non-relativistically but do not show this. Coulomb collisional loss is also not taken into account fully for the separate reason that the loss term G only includes a contribution from plasmaspheric electron density. Furthermore, Equation 3.47 is missing a term proportional to $\mathrm{d}\mu/\mathrm{d}t_{\mathrm{fric}} f$, shown by Equation 3.49 in the next section.

With hindsight, it can be suggested that Equation 3.47 is inappropriate for building a numerical model to be used at relativistic energies, due to non-relativistic assumptions and incomplete modelling of coulomb collisional loss. The work presented in Chapter 4 was initially carried out using this version of the model, until comments from first round reviewers prompted an overhaul.

3.5.2 2D Case

The 2D version of Equation 3.13 can be derived by setting $J = 0$ to consider equatorial particles only, leading to the following simplification, where $f(\mu, L)$ describes 2D relativistic proton phase space density:

$$\frac{\partial f}{\partial t} + \frac{\partial}{\partial \mu}\left[\frac{\mathrm{d}\mu}{\mathrm{d}t}_{\mathrm{fric}} f\right] + \frac{1}{2\mu}\frac{\mathrm{d}\mu}{\mathrm{d}t}_{\mathrm{fric}} f = L^2 \frac{\partial}{\partial L}\left[\frac{D_{LL}}{L^2}\frac{\partial f}{\partial L}\right] + S_n - \Lambda f \tag{3.49}$$

Equation 3.49 was solved on a 2D grid with linearly spaced increments in $\log_{10}(\mu)$ and L. Each grid point is represented here by an i and k index increasing in each dimension respectively. The first step towards solving Equation 3.49 was deriving its finite difference approximation (FDA) so that it could be expressed in terms of discrete grid space.

Equation 3.49 contains a convection term (second from left). The quantity $\mathrm{d}\mu/\mathrm{d}t_{\mathrm{fric}}$ is always negative, so this term represents a flow of f from top to bottom across the grid in the i dimension. Therefore, new values for each point must come from above (upwind) rather than below (downwind), otherwise information will be moving in the wrong direction and the numerical method will be unstable. This upwind scheme requires for the finite difference approximation that $\frac{\partial f}{\partial \mu} \rightarrow \frac{f_{i+1} - f_i}{\Delta \mu}$, where $\Delta\mu = \mu_{i+1} - \mu_i$.

The finite difference approximation for Equation 3.49 is therefore

$$af_{k-1}^{n+1} + bf^{n+1} + ef_{k+1}^{n+1} + gf_{i+1}^{n+1} = f^n + \Delta t S_n \tag{3.50}$$

where the upper index n represents the current timestep, and $n+1$ is the next timestep. The coefficients a, b, e and g are given by:

$$\begin{aligned} a &= -\frac{L^2\ \Delta t}{{L_{k-\frac{1}{2}}}^2\ \Delta L^2} D_{LL_{k-\frac{1}{2}}} \\ e &= -\frac{L^2\ \Delta t}{{L_{k+\frac{1}{2}}}^2\ \Delta L^2} D_{LL_{k+\frac{1}{2}}} \\ g &= \frac{\Delta t}{\Delta \mu} \frac{\mathrm{d}\mu}{\mathrm{d}t}_{i+1} \\ b &= 1 - a - e - \frac{\Delta t}{\Delta \mu} \frac{\mathrm{d}\mu}{\mathrm{d}t} + \frac{\Delta t}{2\mu} \frac{\mathrm{d}\mu}{\mathrm{d}t} + \Delta t \Lambda \end{aligned} \tag{3.51}$$

Equation 3.50 can be re-written as the system of linear equations $M_{2m}\mathbf{f} = \mathbf{r}$, where M_{2m} is a square matrix containing coefficients a, b, e and g, $\mathbf{f}$ is a vector containing all f_{k-1}^{n+1}, f^{n+1}, f_{k+1}^{n+1} and f_{i+1}^{n+1}, and $\mathbf{r}$ is a vector containing the known quantities on the right hand side. This system is arranged as shown in the Figure 3.22 schematic, where non-zero diagonals of M_{2m} have been indicated. The position of elements a is sometimes referred to as the "lower diagonal", and the position of e as the "upper diagonal", with the lower, centre and upper diagonals forming the tridiagonal group a, b and e.

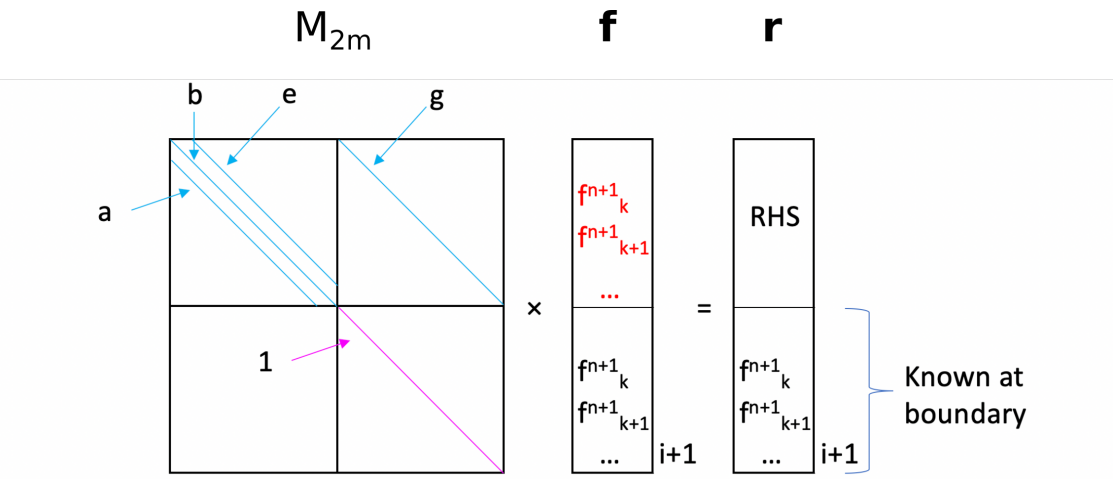


Figure 3.22: A system of linear equations expressing the 2D model equation.

To expand on the schematic in Figure 3.22, Equation 3.52 gives M_{2m} in terms of the coefficients a, b, e and g. It has been written for a grid of general size, with m grid points in the L direction so that the index variable k goes from 1 to m:

$$M_{2m} = \left[\begin{array}{ccc|ccc} \multicolumn{3}{c|}{M_A} & \multicolumn{3}{c}{M_B} \\ \hline \multicolumn{3}{c|}{\begin{array}{ccc} b_1 & e_1 & 0 \\ a_2 & b_2 & e_2 \\ 0 & a_3 & b_3 \\ & \vdots & \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{array} \quad \cdots \quad \begin{array}{ccc} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ & \vdots & \\ b_{m-2} & e_{m-2} & 0 \\ a_{m-1} & b_{m-1} & e_{m-1} \\ 0 & a_m & b_m \end{array}} & \multicolumn{3}{c}{\begin{array}{ccc} g_1 & 0 & 0 \\ 0 & g_2 & 0 \\ 0 & 0 & g_3 \\ & \vdots & \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{array} \quad \cdots \quad \begin{array}{ccc} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ & \vdots & \\ g_{m-2} & 0 & 0 \\ 0 & g_{m-1} & 0 \\ 0 & 0 & g_m \end{array}} \\ \hline \multicolumn{3}{c|}{\begin{array}{ccc} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ & \vdots & \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{array} \quad \cdots \quad \begin{array}{ccc} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ & \vdots & \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{array}} & \multicolumn{3}{c}{\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ & \vdots & \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{array} \quad \cdots \quad \begin{array}{ccc} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ & \vdots & \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{array}} \\ \multicolumn{3}{c}{M_D} & \multicolumn{3}{c}{M_E} \end{array}\right] \tag{3.52}$$

The next value at each grid point depends on the current value, the two current values at surrounding k indices, and the current value at $i+1$. Therefore, for a given i on the model grid, it is possible to solve $M_{2m}\mathbf{f} = \mathbf{r}$ for f^{n+1} at all k simultaneously if:

- f^{n+1} is already known at the smallest and largest k, and
- each ${f^{n+1}}_{i+1}$ is already known at every k.

The first condition is fulfilled at any i by the following boundary conditions:

$$f\left(\mu, L_{\min}\right) = 0 \tag{3.53}$$

$$f\left(\mu, L_{\max}\right) = f_{\text{boundary data}}\left(\mu, K\right) \tag{3.54}$$

Physically, these equations dictate that flux is zero at $L_{\min}$, and flux is given by an outer boundary function at $L_{\max}$. These are both true when $L_{\min}$ is close enough to Earth and when spacecraft data is available to drive the outer boundary at $L_{\max}$. The second condition above is fulfilled at the second to last row in i by the following boundary condition:

$$f\left(\mu_{\max}, L\right) = 0 \tag{3.55}$$

Physically, this equation represents the high energy trapping limit; since particles with first invariant $\mu_{\max}$, for any L, are very high energy, they do not undergo adiabatic motion and cannot be trapped. By beginning on the row $i = n_\mu - 1$, where n_μ is the size of the grid in the i dimension, then moving down in the i index from top to bottom across the solution grid, the grid can be solved for f^{n+1}.

To solve the matrix equation $M_{2m}\mathbf{f} = \mathbf{r}$ at a given row i, the technique of LU decomposition was used. To demonstrate this method with a simple example, a matrix A can be permuted by left-multiplying with a permutation matrix P so that

$$PA = LU \tag{3.56}$$

where L is a lower triangular matrix and U is an upper triangular matrix. A 3×3 matrix A may be broken down like so in the case that no reordering is necessary ($P = I$):

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix}$$

When P reorders the rows of A, this technique is said to include "partial pivoting". When P reorders columns too, it is referred to more generally as "pivoting". Application of this technique is limited here to cases where pivoting is not required, as in the example above.

Using this technique, a system of linear equations given by $A\mathbf{f} = \mathbf{r}$ can be solved for $\mathbf{f}$. The method of solution is illustrated by writing out the system of equations in terms of L and U like so:

$$\begin{aligned} A\mathbf{f} &= \mathbf{r} \\ A &= LU \\ \therefore LU\mathbf{f} &= \mathbf{r} \\ \therefore L\mathbf{y} &= \mathbf{r} \quad \text{where } U\mathbf{f} = \mathbf{y} \end{aligned} \tag{3.57}$$

The solution can then be found in two steps: solve $L\mathbf{y} = \mathbf{r}$ for $\mathbf{y}$ using forward substitution; then solve $U\mathbf{f} = \mathbf{y}$ for $\mathbf{f}$ using back substitution.

LU decomposition of M_{2m} can therefore be used to solve the system $M_{2m}\mathbf{f} = \mathbf{r}$ by considering the decomposition $L_{2m}U_{2m} = M_{2m}$. However, the shape of M_{2m} allows a further simplification to be made: non-zero elements of L_{2m} are limited to the diagonal and lower diagonal. The general forms of L_{2m} and U_{2m} are thus given by Equations 3.58 and 3.59 respectively.

$$L_{2m} = \left[\begin{array}{ccccccc|ccccccc}
1 & 0 & 0 & & 0 & 0 & 0 & 0 & 0 & 0 & & 0 & 0 & 0 \\
l_2^1 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\
0 & l_3^2 & 1 & & 0 & 0 & 0 & 0 & 0 & 0 & & 0 & 0 & 0 \\
 & \vdots & & \ddots & & \vdots & & & \vdots & & \ddots & & \vdots & \\
0 & 0 & 0 & & 1 & 0 & 0 & 0 & 0 & 0 & & 0 & 0 & 0 \\
0 & 0 & 0 & \cdots & l_{m-1}^{m-2} & 1 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\
0 & 0 & 0 & & 0 & l_m^{m-1} & 1 & 0 & 0 & 0 & & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & & 0 & 0 & l_{m+1}^m & 1 & 0 & 0 & & 0 & 0 & 0 \\
0 & 0 & 0 & \cdots & 0 & 0 & 0 & l_{m+2}^{m+1} & 1 & 0 & \cdots & 0 & 0 & 0 \\
0 & 0 & 0 & & 0 & 0 & 0 & 0 & l_{m+3}^{m+2} & 1 & & 0 & 0 & 0 \\
 & \vdots & & \ddots & & \vdots & & & \vdots & & \ddots & & \vdots & \\
0 & 0 & 0 & & 0 & 0 & 0 & 0 & 0 & 0 & & 1 & 0 & 0 \\
0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & l_{2m-1}^{2m-2} & 1 & 0 \\
0 & 0 & 0 & & 0 & 0 & 0 & 0 & 0 & 0 & & 0 & l_{2m}^{2m-1} & 1
\end{array}\right] \tag{3.58}$$

$$U_{2m} = \left[\begin{array}{ccccccc|ccccccc}
u_1^1 & u_1^2 & u_1^3 & & u_1^{m-2} & u_1^{m-1} & u_1^m & u_1^{m+1} & u_1^{m+2} & u_1^{m+3} & & u_1^{2m-2} & u_1^{2m-1} & u_1^{2m} \\
0 & u_2^2 & u_2^3 & \cdots & u_2^{m-2} & u_2^{m-1} & u_2^m & u_2^{m+1} & u_2^{m+2} & u_2^{m+3} & \cdots & u_2^{2m-2} & u_2^{2m-1} & u_2^{2m} \\
0 & 0 & u_3^3 & & u_3^{m-2} & u_3^{m-1} & u_3^m & u_3^{m+1} & u_3^{m+2} & u_3^{m+3} & & u_3^{2m-2} & u_3^{2m-1} & u_3^{2m} \\
 & \vdots & & \ddots & & \vdots & & & \vdots & & \ddots & & \vdots & \\
0 & 0 & 0 & & u_{m-2}^{m-2} & u_{m-2}^{m-1} & u_{m-2}^m & u_{m-2}^{m+1} & u_{m-2}^{m+2} & u_{m-2}^{m+3} & & u_{m-2}^{2m-2} & u_{m-2}^{2m-1} & u_{m-2}^{2m} \\
0 & 0 & 0 & \cdots & 0 & u_{m-1}^{m-1} & u_{m-1}^m & u_{m-1}^{m+1} & u_{m-1}^{m+2} & u_{m-1}^{m+3} & \cdots & u_{m-1}^{2m-2} & u_{m-1}^{2m-1} & u_{m-1}^{2m} \\
0 & 0 & 0 & & 0 & 0 & u_m^m & u_m^{m+1} & u_m^{m+2} & u_m^{m+3} & & u_m^{2m-2} & u_m^{2m-1} & u_m^{2m} \\
\hline
0 & 0 & 0 & & 0 & 0 & 0 & u_{m+1}^{m+1} & u_{m+1}^{m+2} & u_{m+1}^{m+3} & & u_{m+1}^{2m-2} & u_{m+1}^{2m-1} & u_{m+1}^{2m} \\
0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & u_{m+2}^{m+2} & u_{m+2}^{m+3} & \cdots & u_{m+2}^{2m-2} & u_{m+2}^{2m-1} & u_{m+2}^{2m} \\
0 & 0 & 0 & & 0 & 0 & 0 & 0 & 0 & u_{m+3}^{m+3} & & u_{m+3}^{2m-2} & u_{m+3}^{2m-1} & u_{m+3}^{2m} \\
 & \vdots & & \ddots & & \vdots & & & \vdots & & \ddots & & \vdots & \\
0 & 0 & 0 & & 0 & 0 & 0 & 0 & 0 & 0 & & u_{2m-2}^{2m-2} & u_{2m-2}^{2m-1} & u_{2m-2}^{2m} \\
0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & u_{2m-1}^{2m-1} & u_{2m-1}^{2m} \\
0 & 0 & 0 & & 0 & 0 & 0 & 0 & 0 & 0 & & 0 & 0 & u_{2m}^{2m}
\end{array}\right] \tag{3.59}$$

Multiplying L_{2m} and U_{2m} leads to the following equation for M_{2m}:

$$M_{2m} = L_{2m}U_{2m} = \begin{bmatrix} M_A & M_B \\ M_D & M_E \end{bmatrix} \tag{3.60}$$

By equating components of M_{2m} as defined in Equations 3.52 and 3.60, beginning with the final row and working upwards, it is possible to solve for one unknown at a time and determine every coefficient l and u in terms of a, b, e and g, and thereby fully determine L and U. With a small extension to this method, it is possible to go even further and solve the three dimensional problem.

3.5.3 3D Case

For the 3D case, the master equation of Equation 3.13 was used, repeated below:

$$\frac{\partial f}{\partial t} + \frac{\partial}{\partial \mu}\left[\frac{\mathrm{d}\mu}{\mathrm{d}t}_{\text{fric}} f\right] + \frac{\partial}{\partial J}\left[\frac{J}{2\mu}\frac{\mathrm{d}\mu}{\mathrm{d}t}_{\text{fric}} f\right] = L^2 \frac{\partial}{\partial L}\left[\frac{D_{LL}}{L^2}\frac{\partial f}{\partial L}\right] + S_n - \Lambda f \tag{3.61}$$

Equation 3.61 was solved on a 3D grid with linearly spaced increments in $\log_{10}(\mu)$, K and L. Each grid point is represented here by an i, j and k index increasing in each dimension respectively. Instead of transforming the master equation to be in terms of K, it is more convenient to keep the canonical action angle variable J for the second invariant and simply convert K to J internally whenever needed. The conversion between K and J is like so:

$$\begin{aligned} K &= \sqrt{B_m} I = LaY(y)\sqrt{B_m} \\ J &= 2\rho LaY(y) \\ \therefore J &= \frac{2\rho}{\sqrt{B_m}} K \end{aligned} \tag{3.62}$$

The conversion between K and J for any point on the model grid was made quickly by using a pre-calculated mapping between α_{eq}, K and B_m at each model L. Kinetic energy at every model coordinate is also stored, which is used to calculate ρ quickly.

There are two convectional terms in 3.61, and an upwind scheme requires for the finite difference approximation that $\frac{\partial f}{\partial \mu} \to \frac{f_{i+1} - f_i}{\Delta \mu}$, where $\Delta\mu = \mu_{i+1} - \mu_i$, and that $\frac{\partial f}{\partial J} \to \frac{f_{j+1} - f_j}{\Delta J}$, where $\Delta J = J_{j+1} - J_j$. The finite difference approximation for Equation 3.61 is therefore

$$a f_{k-1}^{n+1} + b f^{n+1} + e f_{k+1}^{n+1} + g f_{i+1}^{n+1} + h f_{j+1}^{n+1} = f^n + \Delta t S_n \tag{3.63}$$

with coefficients given by:

$$
\begin{aligned}
a = & -\frac{L^2\ \Delta t}{{L_{k-\frac{1}{2}}}^2\ \Delta L^2} D_{LL_{k-\frac{1}{2}}} \\
e = & -\frac{L^2\ \Delta t}{{L_{k+\frac{1}{2}}}^2\ \Delta L^2} D_{LL_{k+\frac{1}{2}}} \\
g = & \frac{\Delta t}{\Delta \mu} \frac{\mathrm{d}\mu}{\mathrm{d}t}_{i+1} \\
h = & \frac{\Delta t}{\Delta J} \frac{J_{j+1}}{2\mu} \frac{\mathrm{d}\mu}{\mathrm{d}t}_{j+1} \\
b = & 1 - a - e - \frac{\Delta t}{\Delta \mu} \frac{\mathrm{d}\mu}{\mathrm{d}t} - \frac{\Delta t}{\Delta J} \frac{J}{2\mu} \frac{\mathrm{d}\mu}{\mathrm{d}t} + \Delta t \Lambda
\end{aligned}
\tag{3.64}
$$

As in the 2D case, Equation 3.63 can be re-written as a system of linear equations $M_{3m}\mathbf{f} = \mathbf{r}$. This system is arranged as shown in the Figure 3.23 schematic with non-zero diagonals of M_{3m} indicated.

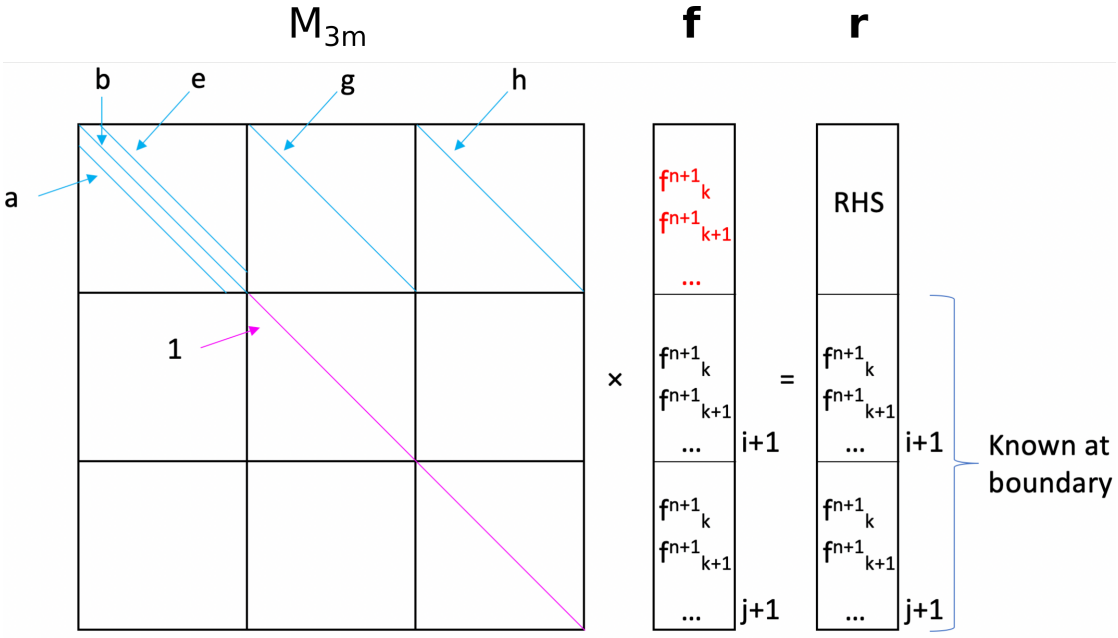


Figure 3.23: A system of linear equations expressing the 3D model equation

To expand on the schematic in Figure 3.23, Equation 3.65 gives M_{3m} in terms of the coefficients a, b, e, g and h. It has been written for a grid of general size, with m elements in the L direction:

$$
M_{3m} = \left[\begin{array}{ccc}
M_A & M_B & \begin{matrix}
h_1 & 0 & 0 & \cdots & 0 & 0 & 0 \\
0 & h_2 & 0 & \cdots & 0 & 0 & 0 \\
0 & 0 & h_3 & \ddots & 0 & 0 & 0 \\
\vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \\
0 & 0 & 0 & & h_{m-2} & 0 & 0 \\
0 & 0 & 0 & \cdots & 0 & h_{m-1} & 0 \\
0 & 0 & 0 & & 0 & 0 & h_m
\end{matrix} \quad M_C \\
M_D & M_E & \begin{matrix}
1 & 0 & 0 & \cdots & 0 & 0 & 0 \\
0 & 1 & 0 & \cdots & 0 & 0 & 0 \\
0 & 0 & 1 & \ddots & 0 & 0 & 0 \\
\vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \\
0 & 0 & 0 & & 1 & 0 & 0 \\
0 & 0 & 0 & \cdots & 0 & 1 & 0 \\
0 & 0 & 0 & & 0 & 0 & 1
\end{matrix} \quad M_F \\
\begin{matrix}
0 & \cdots & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & 0
\end{matrix} & \begin{matrix}
0 & \cdots & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & 0
\end{matrix} & \begin{matrix}
1 & 0 & 0 & \cdots & 0 & 0 & 0 \\
0 & 1 & 0 & \cdots & 0 & 0 & 0 \\
0 & 0 & 1 & \ddots & 0 & 0 & 0 \\
\vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \\
0 & 0 & 0 & & 1 & 0 & 0 \\
0 & 0 & 0 & \cdots & 0 & 1 & 0 \\
0 & 0 & 0 & & 0 & 0 & 1
\end{matrix} \quad M_I \\
M_G & M_H &
\end{array}\right]
\tag{3.65}
$$

In the 3D case, the next value at each grid point depends on the current value, the two current values at surrounding k indices, the current value at $i+1$ and the current value at $j+1$. Therefore, for a given i and j, it is possible to solve $M_{3m}\mathbf{f} = \mathbf{r}$ for f^{n+1} at all k simultaneously if

- f^{n+1} is already known at the smallest and largest k,
- ${f^{n+1}}_{i+1}$ is already known at every k, and
- ${f^{n+1}}_{j+1}$ is already known at every k.

The first two conditions require boundary conditions similar to the 2D case:

$$f\left(\mu, K, L_{\min}\right) = 0 \tag{3.66}$$

$$f\left(\mu, K, L_{\max}\right) = f_{\text{boundary data}}\left(\mu, K\right) \tag{3.67}$$

$$f\left(\mu_{\max}, K, L\right) = 0 \tag{3.68}$$

In addition to these, the following boundary condition is needed:

$$f(\mu, K > K_{\max}(L), L) = 0 \tag{3.69}$$

where $K_{\max}(L)$ is the largest K outside the loss cone at L. Physically, this equation describes phase space density inside the loss cone as zero.

The LU decomposition technique from Section 3.5.2 for the 2D case is extended to solve the system $M_{3m}\mathbf{f} = \mathbf{r}$, by considering the decomposition $L_{3m}U_{3m} = M_{3m}$.

Multiplying L_{3m} and U_{3m} leads to the general form of M_{3m}:

$$M_{3m} = L_{3m}U_{3m} = \begin{bmatrix} M_A & M_B & M_C \\ M_F & M_E & M_F \\ M_G & M_H & M_I \end{bmatrix} \tag{3.70}$$

$$M_G = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & & 0 & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix} \qquad M_H = \begin{bmatrix} 0 & 0 & \cdots & 0 & l_{2m+1}^{2m}u_{2m}^{2m} \\ 0 & 0 & & 0 & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix}$$

$$M_A = \begin{bmatrix} u_1^{2m+1} & u_1^{2m+2} & u_1^{2m+3} & \cdots & u_1^{3m-2} & u_1^{3m-1} & u_1^{3m} \\ l_2^1u_1^{2m+1} + u_2^{2m+1} & l_2^1u_1^{2m+2} + u_2^{2m+2} & l_2^1u_1^{2m+3} + u_2^{2m+3} & \cdots & l_2^1u_1^{3m-2} + u_2^{3m-2} & l_2^1u_1^{3m-1} + u_2^{3m-1} & l_2^1u_1^{3m} + u_2^{3m} \\ l_3^2u_2^{2m+1} + u_3^{2m+1} & l_3^2u_2^{2m+2} + u_3^{2m+2} & l_3^2u_2^{2m+3} + u_3^{2m+3} & & l_3^2u_2^{3m-2} + u_3^{3m-2} & l_3^2u_2^{3m-1} + u_3^{3m-1} & l_3^2u_2^{3m} + u_3^{3m} \\ \vdots & \vdots & & & & \vdots & \\ l_{m-2}^{m-3}u_{m-3}^{2m+1} + u_{m-2}^{2m+1} & l_{m-2}^{m-3}u_{m-3}^{2m+2} + u_{m-2}^{2m+2} & l_{m-2}^{m-3}u_{m-3}^{2m+3} + u_{m-2}^{2m+3} & & l_{m-2}^{m-3}u_{m-3}^{3m-2} + u_{m-2}^{3m-2} & l_{m-2}^{m-3}u_{m-3}^{3m-1} + u_{m-2}^{3m-1} & l_{m-2}^{m-3}u_{m-3}^{3m} + u_{m-2}^{3m} \\ l_{m-1}^{m-2}u_{m-2}^{2m+1} + u_{m-1}^{2m+1} & l_{m-1}^{m-2}u_{m-2}^{2m+2} + u_{m-1}^{2m+2} & l_{m-1}^{m-2}u_{m-2}^{2m+3} + u_{m-1}^{2m+3} & & l_{m-1}^{m-2}u_{m-2}^{3m-2} + u_{m-1}^{3m-2} & l_{m-1}^{m-2}u_{m-2}^{3m-1} + u_{m-1}^{3m-1} & l_{m-1}^{m-2}u_{m-2}^{3m} + u_{m-1}^{3m} \\ l_m^{m-1}u_{m-1}^{2m+1} + u_m^{2m+1} & l_m^{m-1}u_{m-1}^{2m+2} + u_m^{2m+2} & l_m^{m-1}u_{m-1}^{2m+3} + u_m^{2m+3} & & l_m^{m-1}u_{m-1}^{3m-2} + u_m^{3m-2} & l_m^{m-1}u_{m-1}^{3m-1} + u_m^{3m-1} & l_m^{m-1}u_{m-1}^{3m} + u_m^{3m} \end{bmatrix}$$

$$M_B = \begin{bmatrix} l_{m+1}^m u_m^{2m+1} + u_{m+1}^{2m+1} & l_{m+1}^m u_m^{2m+2} + u_{m+1}^{2m+2} & l_{m+1}^m u_m^{2m+3} + u_{m+1}^{2m+3} & \cdots & l_{m+1}^m u_m^{3m-2} + u_{m+1}^{3m-2} & l_{m+1}^m u_m^{3m-1} + u_{m+1}^{3m-1} & l_{m+1}^m u_m^{3m} + u_{m+1}^{3m} \\ l_{m+2}^{m+1}u_{m+1}^{2m+1} + u_{m+2}^{2m+1} & l_{m+2}^{m+1}u_{m+1}^{2m+2} + u_{m+2}^{2m+2} & l_{m+2}^{m+1}u_{m+1}^{2m+3} + u_{m+2}^{2m+3} & & l_{m+2}^{m+1}u_{m+1}^{3m-2} + u_{m+2}^{3m-2} & l_{m+2}^{m+1}u_{m+1}^{3m-1} + u_{m+2}^{3m-1} & l_{m+2}^{m+1}u_{m+1}^{3m} + u_{m+2}^{3m} \\ l_3^{m+2}u_{m+2}^{2m+1} + u_{m+3}^{2m+1} & l_3^{m+2}u_{m+2}^{2m+2} + u_{m+3}^{2m+2} & l_3^{m+2}u_{m+2}^{2m+3} + u_{m+3}^{2m+3} & & l_3^{m+2}u_{m+2}^{3m-2} + u_{m+3}^{3m-2} & l_3^{m+2}u_{m+2}^{3m-1} + u_{m+3}^{3m-1} & l_3^{m+2}u_{m+2}^{3m} + u_{m+3}^{3m} \\ \vdots & \vdots & & & & \vdots & \\ l_{2m-2}^{2m-3}u_{2m-3}^{2m+1} + u_{2m-2}^{2m+1} & l_{2m-2}^{2m-3}u_{2m-3}^{2m+2} + u_{2m-2}^{2m+2} & l_{2m-2}^{2m-3}u_{2m-3}^{2m+3} + u_{2m-2}^{2m+3} & & l_{2m-2}^{2m-3}u_{2m-3}^{3m-2} + u_{2m-2}^{3m-2} & l_{2m-2}^{2m-3}u_{2m-3}^{3m-1} + u_{2m-2}^{3m-1} & l_{2m-2}^{2m-3}u_{2m-3}^{3m} + u_{2m-2}^{3m} \\ l_{2m-1}^{2m-2}u_{2m-2}^{2m+1} + u_{2m-1}^{2m+1} & l_{2m-1}^{2m-2}u_{2m-2}^{2m+2} + u_{2m-1}^{2m+2} & l_{2m-1}^{2m-2}u_{2m-2}^{2m+3} + u_{2m-1}^{2m+3} & & l_{2m-1}^{2m-2}u_{2m-2}^{3m-2} + u_{2m-1}^{3m-2} & l_{2m-1}^{2m-2}u_{2m-2}^{3m-1} + u_{2m-1}^{3m-1} & l_{2m-1}^{2m-2}u_{2m-2}^{3m} + u_{2m-1}^{3m} \\ l_{2m}^{2m-1}u_{2m-1}^{2m+1} + u_{2m}^{2m+1} & l_{2m}^{2m-1}u_{2m-1}^{2m+2} + u_{2m}^{2m+2} & l_{2m}^{2m-1}u_{2m-1}^{2m+3} + u_{2m}^{2m+3} & & l_{2m}^{2m-1}u_{2m-1}^{3m-2} + u_{2m}^{3m-2} & l_{2m}^{2m-1}u_{2m-1}^{3m-1} + u_{2m}^{3m-1} & l_{2m}^{2m-1}u_{2m-1}^{3m} + u_{2m}^{3m} \end{bmatrix}$$

$$M_C = \begin{bmatrix} l_{2m+1}^{2m}u_{2m}^{2m+1} + u_{2m+1}^{2m+1} & l_{2m+1}^{2m}u_{2m}^{2m+2} + u_{2m+1}^{2m+2} & l_{2m+1}^{2m}u_{2m}^{2m+3} + u_{2m+1}^{2m+3} & \cdots & l_{2m+1}^{2m}u_{2m}^{3m-2} + u_{2m+1}^{3m-2} & l_{2m+1}^{2m}u_{2m}^{3m-1} + u_{2m+1}^{3m-1} & l_{2m+1}^{2m}u_{2m}^{3m} + u_{2m+1}^{3m} \\ l_{2m+2}^{2m+1}u_{2m+1}^{2m+1} & l_{2m+2}^{2m+1}u_{2m+1}^{2m+2} + u_{2m+2}^{2m+2} & l_{2m+2}^{2m+1}u_{2m+1}^{2m+3} + u_{2m+2}^{2m+3} & & l_{2m+2}^{2m+1}u_{2m+1}^{3m-2} + u_{2m+2}^{3m-2} & l_{2m+2}^{2m+1}u_{2m+1}^{3m-1} + u_{2m+2}^{3m-1} & l_{2m+2}^{2m+1}u_{2m+1}^{3m} + u_{2m+2}^{3m} \\ 0 & l_{2m+3}^{2m+2}u_{2m+2}^{2m+2} & l_{2m+3}^{2m+2}u_{2m+2}^{2m+3} + u_{2m+3}^{2m+3} & & l_{2m+3}^{2m+2}u_{2m+2}^{3m-2} + u_{2m+3}^{3m-2} & l_{2m+3}^{2m+2}u_{2m+2}^{3m-1} + u_{2m+3}^{3m-1} & l_{2m+3}^{2m+2}u_{2m+2}^{3m} + u_{2m+3}^{3m} \\ \vdots & \vdots & & & & \vdots & \\ 0 & 0 & 0 & & l_{3m-2}^{3m-3}u_{3m-3}^{3m-2} + u_{3m-2}^{3m-2} & l_{3m-2}^{3m-3}u_{3m-3}^{3m-1} + u_{3m-2}^{3m-1} & l_{3m-2}^{3m-3}u_{3m-3}^{3m} + u_{3m-2}^{3m} \\ 0 & 0 & 0 & & l_{3m-1}^{3m-2}u_{3m-2}^{3m-2} & l_{3m-1}^{3m-2}u_{3m-2}^{3m-1} + u_{3m-1}^{3m-1} & l_{3m-1}^{3m-2}u_{3m-2}^{3m} + u_{3m-1}^{3m} \\ 0 & 0 & 0 & & 0 & l_{3m}^{3m-1}u_{3m-1}^{3m-1} & l_{3m}^{3m-1}u_{3m-1}^{3m} + u_{3m}^{3m} \end{bmatrix}$$

$$M_E = \begin{bmatrix} l_{2m+1}^{2m}u_{2m}^{2m+1} + u_{2m+1}^{2m+1} & l_{2m+1}^{2m}u_{2m}^{2m+2} + u_{2m+1}^{2m+2} & l_{2m+1}^{2m}u_{2m}^{2m+3} + u_{2m+1}^{2m+3} & \cdots & l_{2m+1}^{2m}u_{2m}^{3m-2} + u_{2m+1}^{3m-2} & l_{2m+1}^{2m}u_{2m}^{3m-1} + u_{2m+1}^{3m-1} & l_{2m+1}^{2m}u_{2m}^{3m} + u_{2m+1}^{3m} \\ l_{2m+2}^{2m+1}u_{2m+1}^{2m+1} & l_{2m+2}^{2m+1}u_{2m+1}^{2m+2} + u_{2m+2}^{2m+2} & l_{2m+2}^{2m+1}u_{2m+1}^{2m+3} + u_{2m+2}^{2m+3} & & l_{2m+2}^{2m+1}u_{2m+1}^{3m-2} + u_{2m+2}^{3m-2} & l_{2m+2}^{2m+1}u_{2m+1}^{3m-1} + u_{2m+2}^{3m-1} & l_{2m+2}^{2m+1}u_{2m+1}^{3m} + u_{2m+2}^{3m} \\ 0 & l_{2m+3}^{2m+2}u_{2m+2}^{2m+2} & l_{2m+3}^{2m+2}u_{2m+2}^{2m+3} + u_{2m+3}^{2m+3} & & l_{2m+3}^{2m+2}u_{2m+2}^{3m-2} + u_{2m+3}^{3m-2} & l_{2m+3}^{2m+2}u_{2m+2}^{3m-1} + u_{2m+3}^{3m-1} & l_{2m+3}^{2m+2}u_{2m+2}^{3m} + u_{2m+3}^{3m} \\ \vdots & \vdots & & & & \vdots & \\ 0 & 0 & 0 & & l_{3m-2}^{3m-3}u_{3m-3}^{3m-2} + u_{3m-2}^{3m-2} & l_{3m-2}^{3m-3}u_{3m-3}^{3m-1} + u_{3m-2}^{3m-1} & l_{3m-2}^{3m-3}u_{3m-3}^{3m} + u_{3m-2}^{3m} \\ 0 & 0 & 0 & & l_{3m-1}^{3m-2}u_{3m-2}^{3m-2} & l_{3m-1}^{3m-2}u_{3m-2}^{3m-1} + u_{3m-1}^{3m-1} & l_{3m-1}^{3m-2}u_{3m-2}^{3m} + u_{3m-1}^{3m} \\ 0 & 0 & 0 & & 0 & l_{3m}^{3m-1}u_{3m-1}^{3m-1} & l_{3m}^{3m-1}u_{3m-1}^{3m} + u_{3m}^{3m} \end{bmatrix}$$

$$M_F = \begin{bmatrix} l_{2m+1}^{2m}u_{2m}^{2m+1} + u_{2m+1}^{2m+1} & l_{2m+1}^{2m}u_{2m}^{2m+2} + u_{2m+1}^{2m+2} & l_{2m+1}^{2m}u_{2m}^{2m+3} + u_{2m+1}^{2m+3} & \cdots & l_{2m+1}^{2m}u_{2m}^{3m-2} + u_{2m+1}^{3m-2} & l_{2m+1}^{2m}u_{2m}^{3m-1} + u_{2m+1}^{3m-1} & l_{2m+1}^{2m}u_{2m}^{3m} + u_{2m+1}^{3m} \\ l_{2m+2}^{2m+1}u_{2m+1}^{2m+1} & l_{2m+2}^{2m+1}u_{2m+1}^{2m+2} + u_{2m+2}^{2m+2} & l_{2m+2}^{2m+1}u_{2m+1}^{2m+3} + u_{2m+2}^{2m+3} & & l_{2m+2}^{2m+1}u_{2m+1}^{3m-2} + u_{2m+2}^{3m-2} & l_{2m+2}^{2m+1}u_{2m+1}^{3m-1} + u_{2m+2}^{3m-1} & l_{2m+2}^{2m+1}u_{2m+1}^{3m} + u_{2m+2}^{3m} \\ 0 & l_{2m+3}^{2m+2}u_{2m+2}^{2m+2} & l_{2m+3}^{2m+2}u_{2m+2}^{2m+3} + u_{2m+3}^{2m+3} & & l_{2m+3}^{2m+2}u_{2m+2}^{3m-2} + u_{2m+3}^{3m-2} & l_{2m+3}^{2m+2}u_{2m+2}^{3m-1} + u_{2m+3}^{3m-1} & l_{2m+3}^{2m+2}u_{2m+2}^{3m} + u_{2m+3}^{3m} \\ \vdots & \vdots & & & & \vdots & \\ 0 & 0 & 0 & & l_{3m-2}^{3m-3}u_{3m-3}^{3m-2} + u_{3m-2}^{3m-2} & l_{3m-2}^{3m-3}u_{3m-3}^{3m-1} + u_{3m-2}^{3m-1} & l_{3m-2}^{3m-3}u_{3m-3}^{3m} + u_{3m-2}^{3m} \\ 0 & 0 & 0 & & l_{3m-1}^{3m-2}u_{3m-2}^{3m-2} & l_{3m-1}^{3m-2}u_{3m-2}^{3m-1} + u_{3m-1}^{3m-1} & l_{3m-1}^{3m-2}u_{3m-2}^{3m} + u_{3m-1}^{3m} \\ 0 & 0 & 0 & & 0 & l_{3m}^{3m-1}u_{3m-1}^{3m-1} & l_{3m}^{3m-1}u_{3m-1}^{3m} + u_{3m}^{3m} \end{bmatrix}$$

$$M_I = \begin{bmatrix} u_1^{2m+1} & u_1^{2m+2} & u_1^{2m+3} & \cdots & u_1^{3m-2} & u_1^{3m-1} & u_1^{3m} \\ l_2^1u_1^{2m+1} + u_2^{2m+1} & l_2^1u_1^{2m+2} + u_2^{2m+2} & l_2^1u_1^{2m+3} + u_2^{2m+3} & & l_2^1u_1^{3m-2} + u_2^{3m-2} & l_2^1u_1^{3m-1} + u_2^{3m-1} & l_2^1u_1^{3m} + u_2^{3m} \\ l_3^2u_2^{2m+1} + u_3^{2m+1} & l_3^2u_2^{2m+2} + u_3^{2m+2} & l_3^2u_2^{2m+3} + u_3^{2m+3} & & l_3^2u_2^{3m-2} + u_3^{3m-2} & l_3^2u_2^{3m-1} + u_3^{3m-1} & l_3^2u_2^{3m} + u_3^{3m} \\ \vdots & \vdots & & & & \vdots & \\ l_{m-2}^{m-3}u_{m-3}^{2m+1} + u_{m-2}^{2m+1} & l_{m-2}^{m-3}u_{m-3}^{2m+2} + u_{m-2}^{2m+2} & l_{m-2}^{m-3}u_{m-3}^{2m+3} + u_{m-2}^{2m+3} & & l_{m-2}^{m-3}u_{m-3}^{3m-2} + u_{m-2}^{3m-2} & l_{m-2}^{m-3}u_{m-3}^{3m-1} + u_{m-2}^{3m-1} & l_{m-2}^{m-3}u_{m-3}^{3m} + u_{m-2}^{3m} \\ l_{m-1}^{m-2}u_{m-2}^{2m+1} + u_{m-1}^{2m+1} & l_{m-1}^{m-2}u_{m-2}^{2m+2} + u_{m-1}^{2m+2} & l_{m-1}^{m-2}u_{m-2}^{2m+3} + u_{m-1}^{2m+3} & & l_{m-1}^{m-2}u_{m-2}^{3m-2} + u_{m-1}^{3m-2} & l_{m-1}^{m-2}u_{m-2}^{3m-1} + u_{m-1}^{3m-1} & l_{m-1}^{m-2}u_{m-2}^{3m} + u_{m-1}^{3m} \\ l_m^{m-1}u_{m-1}^{2m+1} + u_m^{2m+1} & l_m^{m-1}u_{m-1}^{2m+2} + u_m^{2m+2} & l_m^{m-1}u_{m-1}^{2m+3} + u_m^{2m+3} & & l_m^{m-1}u_{m-1}^{3m-2} + u_m^{3m-2} & l_m^{m-1}u_{m-1}^{3m-1} + u_m^{3m-1} & l_m^{m-1}u_{m-1}^{3m} + u_m^{3m} \end{bmatrix}$$

Analogous to the 2D case, components of M_{3m} can be equated between Equations 3.65 and 3.70, beginning with the final row and working upwards, to solve every coefficient l and u in terms of a, b, e, g and h, and thereby fully determine L and U.

3.5.4 Solving Algorithm

The code snippet in Appendix B was written to perform LU decomposition of a matrix M_{nm} for the n dimensional problem. This applies for systems like those dealt with in this chapter that have:

- a tridiagonal inside the first square $m \times m$ portion of M_{nm};
- and a single diagonal in the remaining $n - 1$ dimensions, arranged in groups of m rows and columns after the tridiagonal as shown in Figures 3.22 and 3.23.

A Fortran version of this code was used to solve Equations 3.50 and 3.63. It is shown in Appendix B written in Python for readability, and since Python is easier to validate by using matrix operations from the NumPy package. The `m` and `dimensions` variables set in the code (lines 21 and 22) can be changed to solve on grids of different sizes or dimensions, and the code includes validation near the end.

3.5.5 Diagonal Dominance

An extra complication arises in the 3D case: although the numerical scheme is fully implicit, the solution is no longer unconditionally stable. For stability, the diagonal elements of M_{3m} must be greater than or equal to the sum of the off-diagonal elements in the same row, which is known as diagonal dominance. The 3D case therefore has the following stability criterion:

$$|a| + |e| + |g| + |h| > |b| \tag{3.71}$$

This condition can met by choosing a timestep small enough, but when the condition in Equation 3.71 is not met this is known as a "diagonal dominance violation". An equivalent condition applies to the 2D case too, but it can easily be shown that

$|a| + |e| + |g| > |b|$ is always true (for any Δt) when a, b, e and g are components of M_{2m} given by the set of equations in Equation 3.51.

In the 3D case, diagonal dominance violations can be triggered by steep gradients in $\mathrm{d}\mu/\mathrm{d}t_{\mathrm{fric}}$ at neighbouring grid points, combined with high timesteps. Re-arranging for the maximum timestep that will not cause a diagonal dominance violation yields:

$$\Delta t_{\max} = \left(\frac{|a| + |e| + |g| + |h|}{\Delta t} - \frac{(|b| - 1)}{\Delta t} \right)^{-1} \tag{3.72}$$

where the Δt term on the right hand side cancels out from the definitions of a, b, e, g and h in Equation 3.64 to make the right hand side independent of Δt.

To prevent model crashes due to instabilities, a system was built into the model to detect diagonal dominance violations, calculate $\Delta t_{\max}$, then automatically lower the timestep to $0.95\Delta t_{\max}$ and continue with the solution. For dynamic simulations not in steady state, the solar cycle and seasonal parameterisation of drift averaged quantities such as $\mathrm{d}\mu/\mathrm{d}t_{\mathrm{fric}}$ leads to a dependence on time-varying indices such as F10.7a, which can vary up or down in subsequent timesteps. This may happen to change the outcome of the diagonal dominance condition in Equation 3.71 and increase the minimum timestep required. The factor of 0.95 was found to help prevent diagonal dominance violations from recurring on following timesteps due to this type of variation, optimising execution time. Furthermore, the model system performs checks by periodically re-evaluating $\Delta t_{\max}$ even when there is no violation. If the timestep is below $\Delta t_{\max}$, it will be increased up to $0.95\Delta t_{\max}$ or the limit specified in the configuration file (whichever is lower).

3.6 Mapping Between K and Equatorial Pitch Angle

Figure 3.24 illustrates the mapping between K and α_{eq} for several fixed values of L. A dashed horizontal line is used to indicate $K_{\max}$ for each L, which represents the boundary K outside the loss cone. Figure 3.24 shows that with regular spacing in K, the solution becomes more detailed near the loss cone, because there are more grid points per small change in α_{eq} due to the shape of the curve. Having a high

grid resolution in this region is important due to sharp gradients in loss timescales, and in order to track the outer boundary of the loss cone closely. Therefore, this represents a convenient feature of the choice of adiabatic invariant coordinates (μ, K, L).

However, for model grids that include a wide range in L, there is a large difference between $K_{\max}$ at opposite sides of the grid in L. For example, $K_{\max}$ at L =1.7 is $\sim$1.15G$^{1/2}$ R_E, whereas $K_{\max}$ at L =1.1 is just less than 0.1G$^{1/2}$ R_E. This leads to a rather inconvenient feature of choosing adiabatic invariant coordinates (μ, K, L), because if the model grid has regular intervals in K, shown by the grey lines in Figure 3.24, then relatively few grid lines will be inside the loss cone at low L. One can increase the number of intervals in K to deal with this problem, but inefficient utilisation of the model grid in this case becomes a memory issue.

To solve this problem, a non-regular grid spacing in the K direction was used, with extra grid lines inserted at low K (near $\alpha_{eq} = 90°$). This measure results in smoother pitch angle distributions produced by the model near 90°, as well as a higher resolution solution at low L when the outer boundary is set far away in L. This solution was implemented such that extra grid lines can be inserted based on configuration file options.

3.7 Overcoming Model Instabilities when Computing 3D Steady State

Sometimes it is useful to calculate a steady state proton belt distribution. To perform this simulation, the solution grid can be initialised as zero, except at the outer boundary which must be prescribed to allow protons to diffuse inward. However, the simulation time required to form the steady state proton belt from nothing (not the real time required for the simulation to finish) may be over one hundred years, depending on the rate of radial diffusion.

On a 3D model grid, adiabatic coordinates show positive gradients in $\mathrm{d}\mu/\mathrm{d}t_{\mathrm{fric}}$ versus K which get rapidly steeper approaching $K_{\max}$ at the loss cone boundary. This is driven by sharply increasing drift averaged densities as the associated mirror point gets closer to the atmosphere. Although the numerical scheme is fully

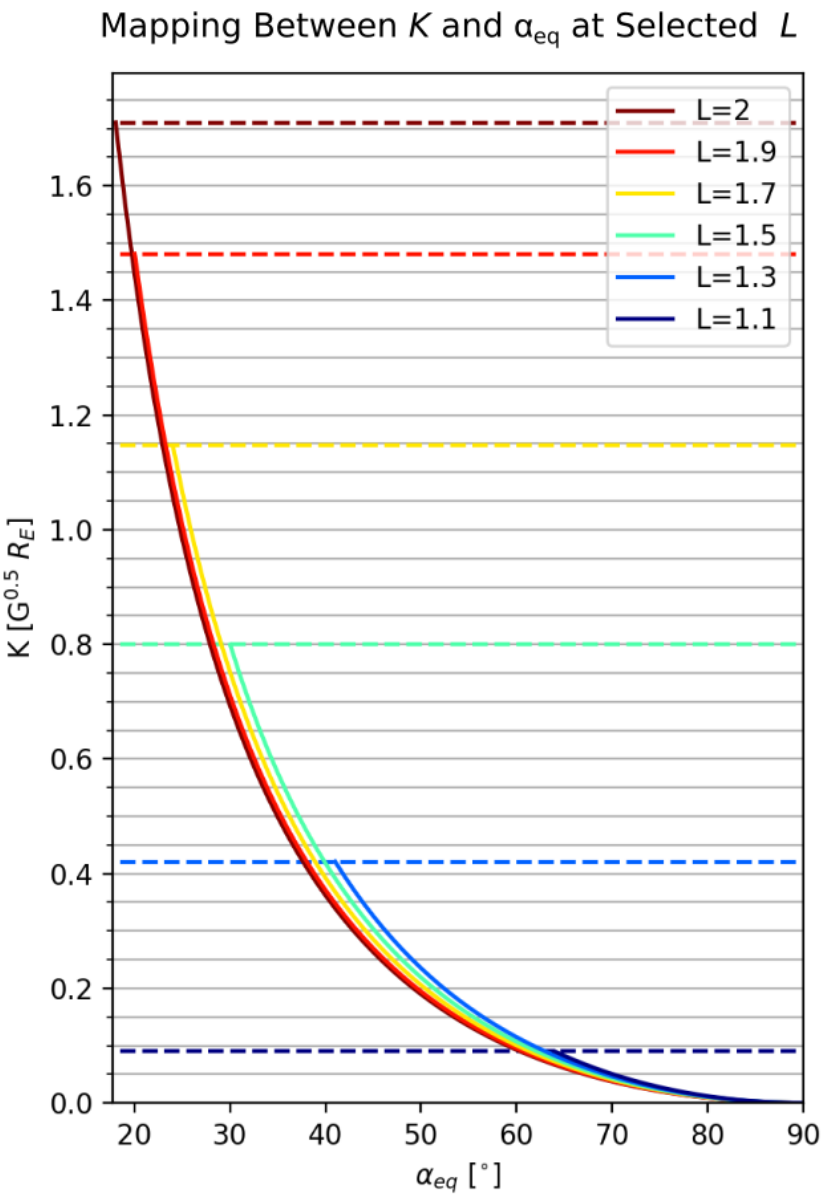


Figure 3.24: Plot of the mapping between α_{eq} and K at various L, where α_{eq} has been extended from 90° down to the dipole loss cone for a 300km atmospheric scale height. The maximum K at each L, corresponding to the loss cone, has been indicated with a dashed horizontal line for each L.

implicit, these gradients can cause model instability at high timesteps, in the form of diagonal dominance violations. This problem is exacerbated by using higher resolution grids, but can be mitigated by reducing the timestep. Therefore, when 3D simulations require lots of detail, a small timestep is required and the time required to calculate steady state can easily exceed one week.

However, one other strategy to decrease the runtime is to compute a solution using a high "loss cone altitude" - the minimum altitude that a particle may mirror at, below which it is considered lost to the atmosphere. This is a critical altitude controlling the boundary condition $f(\mu, K > K_{\max}(L), L) = 0$; when a higher loss cone altitude is set in the configuration file, $K_{\max}$ is reduced at all L, restricting the physical domain of the model to particles mirroring at higher altitude (lower K). As discussed, regions of the model grid with the highest gradients in $\mathrm{d}\mu/\mathrm{d}t_{\mathrm{fric}}$ versus K tend to be near the loss cone. When $K_{\max}$ is made smaller, these regions fall beyond the $f = 0$ boundary condition, outside the range of coordinates needing to be solved. The model is therefore more stable since the gradients in loss timescales across neighbouring grid points are smaller. As a result, the model timestep can be increased - sometimes more than doubled for a $\sim$ 50km change in loss cone altitude.

When the loss cone altitude is increased as described above the solution at low pitch angles, close to the $f = 0$ boundary condition, may be underestimated. However, once steady state is approached using the high loss cone altitude, the solution can be used to initialise a new simulation with a lower altitude loss cone. A smaller timestep will be required of course, but the solution takes less time to reach steady state since radial diffusion will have finished supplying regions of the proton belt during the first simulation. This process can be repeated for ever decreasing loss cone altitudes to approach a detailed 3D steady state much more quickly.

3.8 Reading Drift Averaged Quantities into the Model

Each drift averaged quantity is stored on disk in a 3D grid in terms of the coordinates (μ, α_{eq}, L) as a result of the drift averaging process. This grid is contained within a "model-ready drift average" file, as discussed in Section 3.2.2, with multiple .mrda files for each drift average quantity describing it over different coordinates of a parameter space. For example, since CRAND was calculated for five different F10.7 values, there are five CRAND .mrda files. For each of 14 densities and two temperatures, there are 20 files, which leads to 325 files in total.

The choice of coordinates (μ, α_{eq}, L) for the drift average grid was convenient because an initial state vector $\mathbf{Y_0}$ could be calculated exactly for a trapped proton, which allowed the trajectory S to be solved and the drift average to be evaluated. However, when drift averages must be read into the model, a conversion from α_{eq} to K must be made. Secondly, each quantity must be readily available at each coordinate on the model grid to avoid interpolation whilst timestepping. However, the drift average grid does not necessarily align with the model grid in the μ and L dimensions, so the process involves interpolation in all three dimensions.

For large 3D simulations, the process of reading in every drift averaged quantity across parameter space, then interpolating each quantity onto a grid in memory that aligns with the numerical solution grid, can require several hours and significant amounts of memory. To overcome this problem, the methodology shown in Figure 3.25 was developed, involving the creation of "solution-ready environment averages".

To summarise Figure 3.25, after reading reading in each of the drift averaged quantities once, they are used to calculate $\mathrm{d}\mu/\mathrm{d}t_{\mathrm{fric}}$, Λ and S_n from Equation 3.13 everywhere across the model grid for each environmental parameterisation. In doing so, the 17 drift averaged quantities (CRAND, 14 densities and two temperatures) are combined to just three quantities ($\mathrm{d}\mu/\mathrm{d}t_{\mathrm{fric}}$, Λ and S_n). This requires less memory to store. Each quantity is then output to a separate ".srea" file, incorporating all environmental parameterisations. When a new simulation is started, the same pre-calculated $\mathrm{d}\mu/\mathrm{d}t_{\mathrm{fric}}$, Λ and S_n quantities can be loaded directly from the .srea files into the model from 3D grids aligning with the model grid, providing the new

simulation has the same dimensions as the previous solution used to create each .srea file. Therefore, no interpolation is required, and drift averaged quantities from .mrea files do not have to be loaded. Using this technique, the load time is reduced from several hours to several minutes.

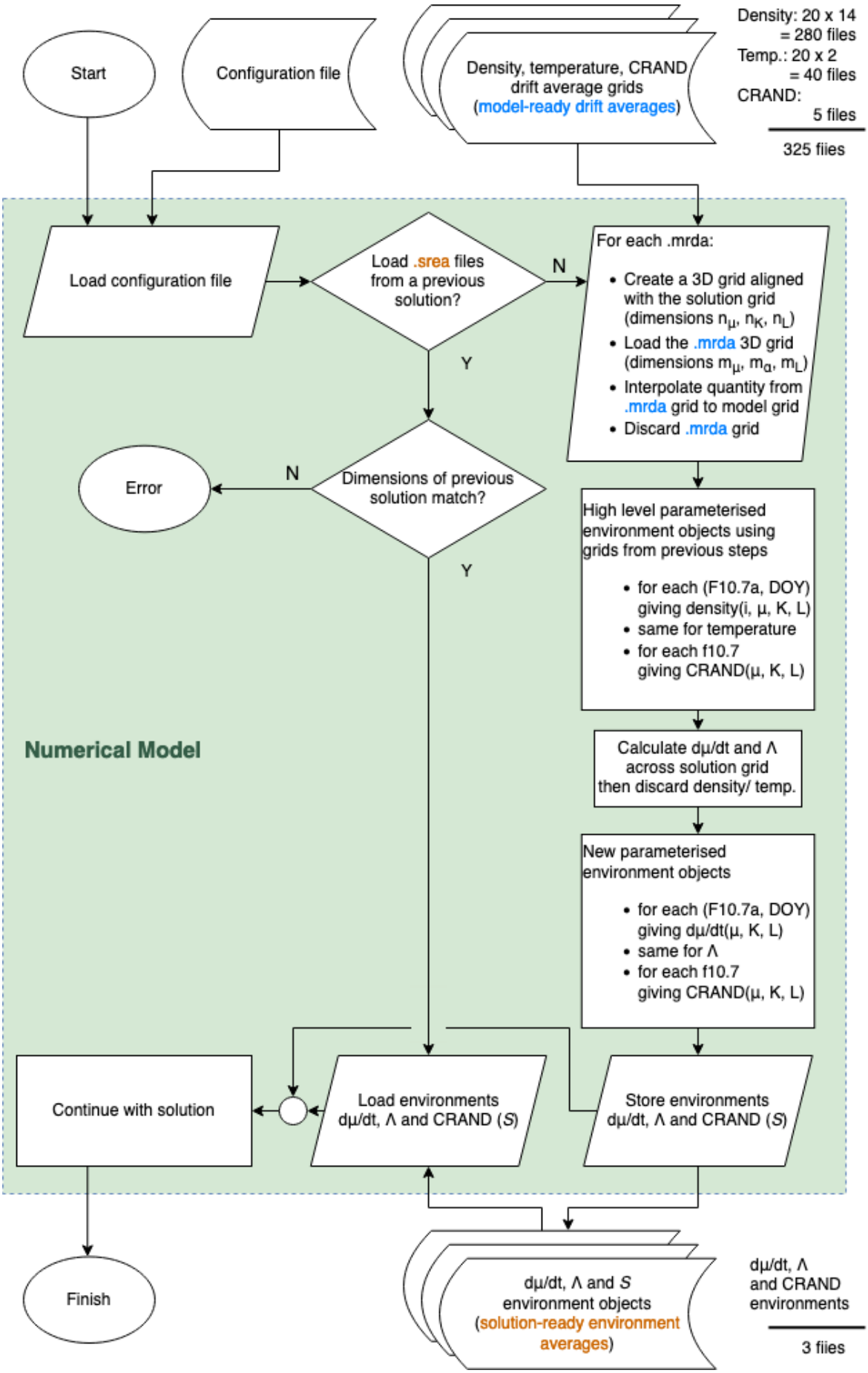


Figure 3.25: Conversion of pre-calculated drift averages to solution-ready environment average files on model startup. This process allows subsequent model runs using the same size solution grid to quickly load $\mathrm{d}\mu/\mathrm{d}t_{\mathrm{fric}}$, Λ and S_n across the model grid from the .srea files produced.

Chapter 4

2D Model Application: Optimisation of Proton Radial Diffusion Coefficients

This chapter is based on a research article:

Optimisation of Radial Diffusion Coefficients for the Proton Radiation Belt During the CRRES Era

JGR Space Physics, March 2021, Volume 126, Issue 3

https://doi.org/10.1029/2020JA028486

Alexander R. Lozinski[ab], Richard B. Horne[a], Sarah A. Glauert[a], Giulio Del Zanna[b], Jay M. Albert[c]

[a]British Antarctic Survey, Cambridge, UK

[b]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK

[c]Space Vehicles Directorate, Air Force Research Laboratory, Kirtland AFB, Albuquerque, NM, USA

4.1 Introduction

Theoretical work has resulted in formal expressions for proton radial diffusion coefficients incorporating parameters that depend on the power spectrum of elec-

tromagnetic fluctuations (Fälthammar, 1965; Fälthammar, 1968). Steady state models have varied these parameters to fit spacecraft observations, allowing determination of diffusive timescales with μ and L dependence (Farley and Walt, 1971; Fischer et al., 1977; Claflin and White, 1974; Jentsch, 1981). However, Albert et al. (1998) demonstrated that such optimisations may give different results based on the empirical approximations used to model source and loss. To show this, the authors computed theoretical equatorial phase space density using a radial diffusion model and compared the results to measurements of 1-100MeV protons throughout the inner zone provided by the Combined Release and Radiation Effects Satellite (CRRES). When the authors switched between different CRAND and density models, the radial diffusion coefficients found to provide the best fit with data were prone to significant variation, underscoring the dependence of this technique on modelling other source and loss processes accurately. Since the work of Albert et al., a more thorough approach to evaluating CRAND (Selesnick et al., 2007), as well as the development of new plasmaspheric density models (i.e. Gallagher et al., 2000; Denton et al., 2006; Ozhogin et al., 2012; Chu et al., 2017), allows better approximations. In addition, there is a strong need for physics-based modelling of the inner proton belt given its increasing utilisation by commercial spacecraft (Lozinski et al., 2019; Horne and Pitchford, 2015).

The work in this chapter follows a similar method to Albert et al. (1998) and presents modelling of the flux of equatorially mirroring protons within the CRRES era, revisiting measurements taken by CRRES's proton telescope (PROTEL). A 2D version of the radial diffusion model was used. Compared with Albert et al. (1998), it includes more modern evaluations of key source/loss processes, as well as improved theoretical modelling of coulomb collisional loss. A steady state optimisation method is used to derive new estimates of proton radial diffusion coefficients for the inner zone. Modelling results are presented and discussed, along with the optimised diffusion coefficients which are compared to other works. This leads to a number of findings and recommendations for future attempts at steady state modelling.

4.2 PROTEL Data

4.2.1 Data Overview

The Proton Telescope (PROTEL) instrument on-board the CRRES satellite made measurements of differential flux on 24 energy channels in the 1 to 100 MeV range with full pitch angle resolution (Violet et al., 1993). Measurements were made from elliptical orbit (350km perigee, 36000km apogee) at 18° inclination. Data for the present study is extracted from the '.pad' files made available by the NASA Space Physics Data Facility (`https://spdf.gsfc.nasa.gov/pub/data/crres/`), which contain flux averaged over one minute intervals as a function of equatorial pitch angle, from 15th August 1990 until 11th October 1991. These measurements were mapped from local to equatorial pitch angle using the IGRF85 internal and Olson and Pfitzer (1974) external magnetic field models. After excluding orbits which exhibit bad data (Brautigam, 2001), this dataset spans 979 orbits.

CRRES observed dynamic changes in trapped flux due to numerous magnetic disturbances, the most significant being the 24th March 1991 storm occurring roughly half way through the mission. In this event, a large interplanetary shock compressed the magnetosphere causing a storm sudden commencement (SSC). This was accompanied by the arrival of an SEP event, in which solar protons were able to penetrate the magnetosphere and become trapped in the outer zone down to $L \sim 2$. This trapping was enabled by the combination of two main factors: firstly, the fast suppression, and subsequent restoration, of geomagnetic cutoff limits over the course of the SSC. These limits define access regions for incoming particles due to attenuation by the geomagnetic field, and their temporary suppression allowed particles of the same magnetic rigidity closer towards Earth. Secondly, in tandem, the shock's compression of the magnetosphere induced an azimuthal electric field pulse which led to inward acceleration and transport of trapped particles drifting in time with the pulse (Li et al., 1993; Hudson et al., 1997). This led to a large enhancement in trapped flux over the timescale of a drift orbit which persisted until at least the end of the CRRES mission. Both rapid and diffusive inward transport (to lower L) lead to betatron acceleration of trapped particles due to violation of the third adiabatic invariant and simultaneous conservations of the first and

second, as discussed in Section 1.3.2. One signature of the CRRES enhancement is therefore highly anisotropic pitch angle distributions over an affected L range. The event followed a period of magnetically quiet conditions, and therefore separates the CRRES data into a quiet and active era. Other magnetic disturbances during the CRRES mission are summarised in Table 1 of Hudson et al. (1997).

PROTEL flux measurements averaged over two $\sim$200 day periods encompassing the quiet and active era have been used to build the CRRESPRO Quiet and Active static radiation belt models (Meffert and Gussenhoven, 1994). The method used to prepare flux maps for the CRRESPRO models is described by Gussenhoven et al. (1993), and data for the current study has been processed using a similar method but for different time average periods. These steps are described in Section 4.2.2.

An overview of PROTEL data is presented in Figure 4.1 as a time series of weekly averaged differential flux at 90° equatorial pitch angle (left panels), along with a measure of anisotropy (n) of each pitch angle distribution (right panels), at L bins from 1.3 to 2.3 ($\Delta L = 0.05$). Both quantities have been calculated by fitting each weekly averaged pitch angle distribution using a $\sin^n$ fitting function, described in more detail in Section 4.2.2. The narrowness of each pitch angle distribution centre peak is therefore controlled by n, with a higher n indicating a more anisotropic (narrower) distribution. The time axis spans the whole period for which data is available. The quiet and active periods used for the CRRESPRO quiet and active models are indicated in Figure 4.1 by two vertical dashed white lines, which show the end of the quiet period average and beginning of the active period average. Magnetic disturbances are also indicated in Figure 4.1 by dotted black vertical lines, marking the arrival of storm sudden commencements as listed in Table 1 of Hudson et al. (1997).

Figure 4.1 demonstrates the large increase in intensity throughout the outer zone (left panels) following the March 1991 storm (and subsequent enhancements), and the effect on particle pitch angle distributions (right), which become more anisotropic during the active period. Some uncertainty, particularly in the n fitting parameter at $L \lesssim 1.45$, arises because one week is not a sufficiently long average period, leading to some fits based on only a few measurements. These poor fits have been excluded from Figure 4.1, and longer flux average periods are used in later analysis. From the 9.7MeV channel and below in Figure 4.1, increases in

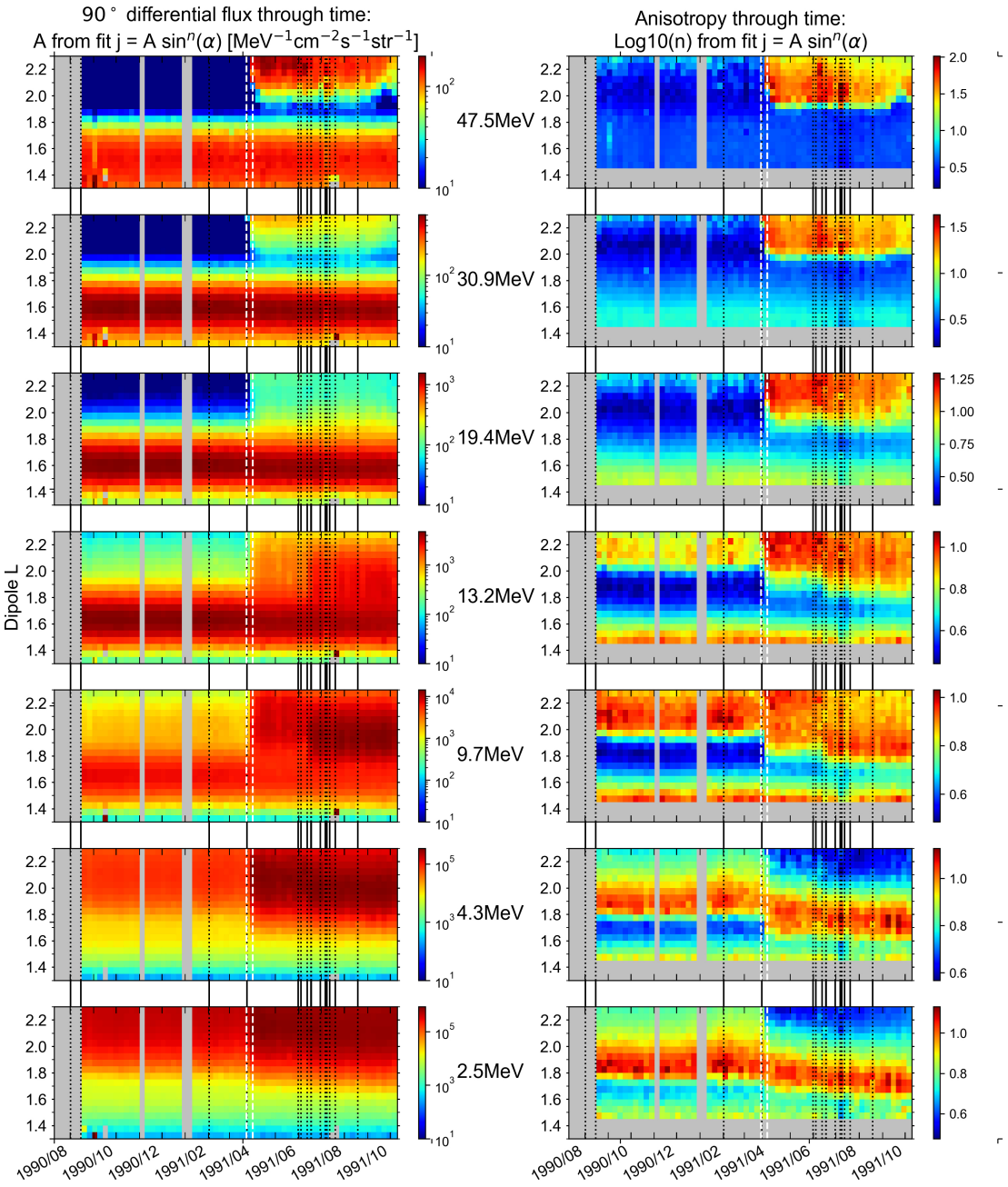


Figure 4.1: Fitted PROTEL data at selected energy channels, showing 90° flux (left) and anisotropy of the pitch angle distribution (right) through time. To produce these values, data was averaged over one week intervals then fit using Equation 4.1

anisotropy at $L \sim 1.7$ are shown during the active period. New particles were injected into the region $2 < \mathrm{L} < 3$ (Blake et al., 1992), but this change at lower L could be the result of inward transport and betatron acceleration leading to a higher proportion of 90° particles, illustrating the extent of the region affected by enhancements.

4.2.2 Data Processing

The one minute averaged pitch angle data files introduced in Section 4.2.1 make available average differential flux measurements binned by equatorial pitch angle and L. Each average flux measurement in each pitch angle/L bin corresponds to a number of observations made within a one minute interval, and this number is included in the data along with average ephemeris for each interval. The following processing steps are already implemented on the available .pad files: noise spike removal (see Brautigam, 2001); mapping of the data into equatorial pitch angle bins (5° bins from 0 to 90° for 18 bins in total); binning of the data in L ($\Delta L = 0.05$); and a loss cone correction to remove background flux. There is some uncertainty about the effectiveness of the loss cone correction in particular, with implications for data accuracy that are discussed in the next section.

Follow a convention of previous works dealing with this dataset, the 8.4MeV channel is ignored due to its overlap, and the 15.2MeV channel is ignored due to sensitivity issues. The first two channels at 1.5 and 2.1MeV were also ignored because these were below the energy range of the model version used for this work; the model energy range began at 2.31MeV since the method to extrapolate the CRAND source below this energy described in Section 3.3 had not yet been implemented.

Loss Cone Correction

During computer simulations it was found that >60MeV protons could pass through PROTEL shielding and be erroneously counted, leading to a background contamination (Hein, 1993). To correct for this, Gussenhoven et al. (1993) describe a loss cone correction whereby the flux at the pitch angle bin just inside the loss cone is treated as the background noise level. Flux at the pitch angle bin of the loss

cone and below is set to zero, whilst the background is subtracted from all higher pitch angle bins. Gussenhoven et al. validated the results of this correction using a Monte Carlo ray tracing code to derive the PROTEL response function. However, the correction has been shown to be inadequate for the 6.8 and 8.5MeV channels (too large) and for energies >40MeV (too small). Furthermore, the correction is similar in magnitude to the measurements themselves at $L < 1.4$.

To deal with these caveats, in addition to the measures listed in Section 4.2.2, the 6.8MeV channel was ignored, since it appears to cause a particularly strong deviation in the spectrum at low L. Data driving the model at >40MeV may represent a slight ($\sim$20%) overestimate owing to limitations of the loss cone correction, but as the majority of energy channels influencing the optimisation method are below 40MeV, this limitation was not expected to have a significant effect on results. All data below $L = 1.35$ was ignored, and uncertainty was quantified by calculating the standard deviations of PROTEL measurements (in terms of phase space density) and including them in plots of results. The potential uncertainties in the data at $L = 1.35$ were also expected to have a minimal influence on results, as the comparison is mostly with measurements at higher L.

Filtering and Fitting

The next step required to process PROTEL .pad files is the filtering out of anomalous records of flux caused by SEP events, to ensure measurements are indicative of trapped flux only. To address this, the method of Gussenhoven et al. (1993) was followed: for any time average period, one minute average flux values are firstly averaged, weighted by the number of observations; the mean flux and standard deviation in each pitch angle and L bin is then calculated; and finally, for each bin, flux measurements outside two standard deviations are excluded, and the mean/standard deviation is then recomputed.

In addition to occasional periods of data unavailability, there is a shortage of measurements at 85 and 90° equatorial pitch angle during two $\sim$month long phases of the mission, beginning around February 1991 and September 1991, and in general there are fewer measurements of equatorial flux at $L \lesssim 1.5$ compared to higher L. The process of averaging equatorial flux over time periods, or ranges in L,

coinciding with this shortage of measurements, gives averages associated with few observations. The final processing step is taken to mitigate this issue, and involves fitting all time averaged equatorial pitch angle distributions using the function

$$j = A \sin^n(\alpha) \tag{4.1}$$

where j is differential, unidirectional flux, α is equatorial pitch angle, and A, n are fitting parameters. Combined with the loss cone correction which has already been applied, this fit has been shown to work well over this range in L (Valot and Engelmann, 1973).

Figure 4.2 shows the application of this fit to equatorial pitch angle distributions at $L = 1.7$. Time averaged flux distributions are shown at 2.9, 5.7, 10.7 and 30.9MeV (rows 1 to 4), for three average periods during the CRRES mission (columns 1 to 3), taken before the March 1991 storm (Quiet), and at two intervals after (Active 1 and Active 2). The average periods are used in later analysis and explained in Section 4.4. Average flux values in each pitch angle bin (black crosses) are plotted along with the corresponding standard deviation (vertical blue lines). The best fit (red curve) and fitting parameters are also shown for each distribution. Figure 4.2 shows the advantage of using this fit to get a stable measurement of 90° flux, especially for the Active 2 period where standard deviation is higher. The fit is weighted by flux at lower pitch angle, for which many more observations exist given that dwell time off-equator is comparatively high. One disadvantage of this fit is that for highly anisotropic distributions ($n > 10$), the fit can sometimes seem to over-predict 90° flux. Figures 4.1 and 4.2 show that during the geomagnetically active period following the March 1991 storm, time averaged pitch angle distributions generally become more anisotropic, with the n fitting parameter increasing through time. Peak flux indicated by the A fitting parameter also undergoes changes at a given L depending on energy channel, with some channels undergoing a flux increase at L=1.7 (i.e. 2.9 and 5.7MeV in rows 1 and 2 of Figure 4.2), and some showing a decrease in flux (i.e. 30.9MeV in row 4 of Figure 4.2).

4.3 Numerical Modelling

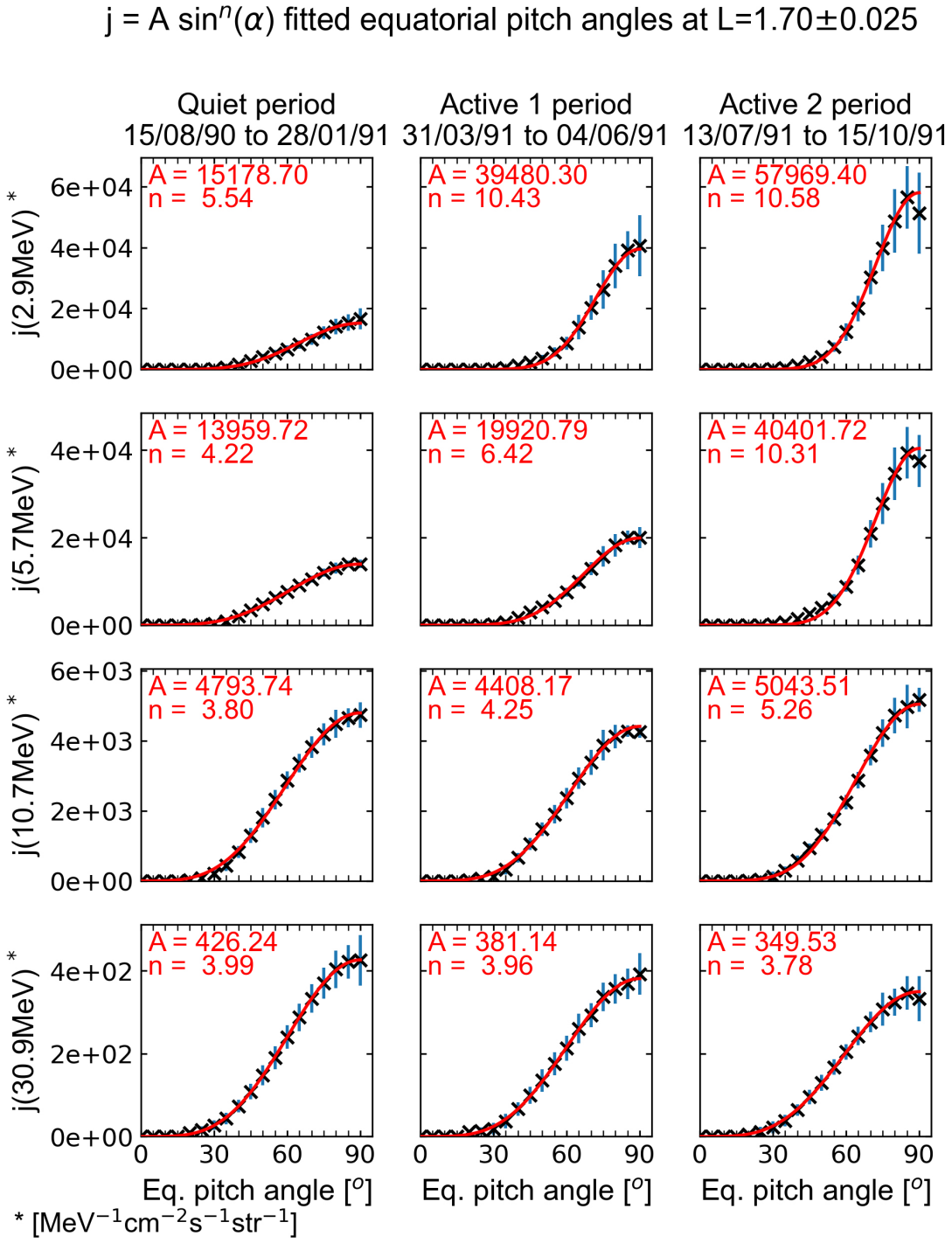


Figure 4.2: sin^n fitting of pitch angle distributions at four energy channels (rows) over three time averages (columns). Data is shown as black crosses, standard deviation of each flux measurement is shown in blue, and the fit using Equation 4.1 is shown in red.

4.3.1 Model Overview

Results presented in this chapter are in terms of relativistic 2D phase space density of equatorially mirroring protons $f(\mu, L)$. This quantity is given by $f = m_0^3 j/p^2$, which allows direct comparison with the results of Albert et al. (1998) later on in this study, who work in terms of a non-relativistic phase space density with the same dimensions and units of $\mathrm{km}^{-6}\mathrm{s}^3$. Calculation of f is performed using the 2D version of the proton belt model, described in Section 3.5.2. The source and loss mechanisms considered were the CRAND source (S_n), and loss from coulomb collisions as well as inelastic nuclear collisions (Λf), and therefore the full model equation is given by Equation 3.49.

Drift averaged calculations performed to calculate S_n and the densities of various constituents are as described in Sections 3.3 and 3.4 respectively. However, since only equatorial particles were considered, it was only necessary to load a small section of each 3D grid holding a pre-calculated drift average. The obvious mapping from $K = 0$ to $\alpha_{eq} = 90°$ also made the method for loading drift averages, described in Section 3.8, much simpler.

Three boundary conditions were applied to the model: $f(L_{min}) = 0$, $f(\mu_{max}) = 0$, and $f(L_{max}, \mu) = f_b(\mu)$, where $L_{min} = 1.1$, $L_{max} = 1.65$, and f_b is the outer boundary spectrum derived from PROTEL data at L_{max}. Inside the energy range of the included PROTEL channels, f_b was linearly interpolated. Outside this range, f_b was extrapolated by fitting a 2nd order polynomial function P to the outer boundary spectrum such that $\log j = P(E)$, where E is energy. The gradient $\mathrm{d}P/\mathrm{d}E$ at the lowest (highest) energy channel was then used to linearly extrapolate $\log j$ to lower (higher) energies, giving a continuous spectrum. The magnetic field was given by a magnetic dipole model with dipole moment $7.83\mathrm{Am}^2$, representative of the era. Whereas Albert et al. (1998) set their model's data-driven outer boundary at $L = 1.7$, the choice of $L_{max} = 1.65$ was made based on an analysis of the variation in fluxes (see Section 4.4.2), which motivated this work to be more selective about the region steady state modelling can be applied to.

4.3.2 Diffusion Coefficients

In this study, radial diffusion coefficients for equatorial protons were parameterised according to the following equation from Claflin and White (1974):

$$D_{LL} = A_m L^{6+2\alpha} \mu^{2-\alpha} + A_e L^{6+2\beta} \mu^{-\beta} \tag{4.2}$$

where A_m and A_e are constants proportional to the power spectrum of certain features in the magnetic and electric field perturbations. This formulation allows for the L and μ dependence giving the best fit to data to be determined via optimisation of the free parameters α, β, A_m and A_e.

4.3.3 The Influence of Plasmaspheric Density

The drift averaged density computation, described in Section 3.4, incorporates solar cycle and seasonal variations in plasmaspheric electron (and other) densities. Figure 4.3 demonstrates the impact this variation has on steady state phase space density, calculated using the PROTEL data. In the top left panel, the density profiles on four days of the year are shown for a fixed F10.7a value over the model L range. In the bottom left panel, density at five values of F10.7a are shown for a fixed day of the year. These plots show the seasonal and solar cycle dependence of electron density respectively. Right-hand panels show the corresponding steady state solutions, calculated with the model using a fixed CRAND source and using the total D_{LL} values for protons of Selesnick and Albert (2019) as an example.

Each profile of electron density in Figure 4.3 (left panels) is fixed at $L = 3.25$ to the Ozhogin et al. (2012) model value, but variation across the L range plotted arises from the ionospheric density given by IRI. The dominant component of variation is from solar cycle effects and reflects the IRI dependence on F10.7a as a model input. To produce the steady state phase space density profiles in Figure 4.3 (right panels), the outer boundary is held constant using PROTEL data averaged over the pre-storm quiet era. In reality the outer boundary flux may vary, but the solutions show some differences and highlight the importance of accurately modelling density both for static and dynamic modelling, given the rather short timescale for seasonal changes. The figure suggests the difference in phase space

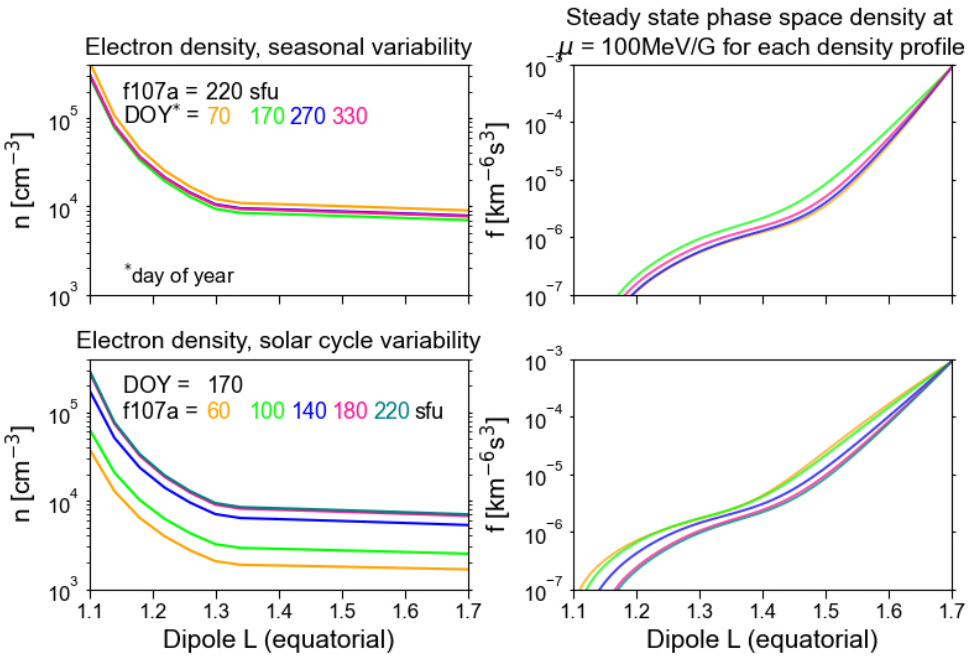


Figure 4.3: Plasmaspheric drift averaged electron density (left panels) demonstrating seasonal (top) and solar cycle (bottom) components of variation, plotted next to corresponding steady state solutions in phase space density (right panels) incorporating those values of density but leaving all other model inputs constant.

density as a result of density variation is relatively uniform in L but becomes particularly pronounced near L_{min}. In fact the corresponding change in phase space density at any L also depends on the diffusion coefficients used, as higher D_{LL} tends to limit the region at which coulomb collisions dominate to lower L.

4.4 Method

4.4.1 Selecting Average Periods

Using the model described in Section 4.3.1, Equation 3.49 was solved for steady state f, assuming $\partial/\partial t = 0$. The aim for this work was to find diffusion coefficients that minimised the difference between f and some time-averaged data via optimisation of the four parameters α, β, A_m and A_e in Equation 4.2. Performing this optimisation involved minimising the mean square deviation in f between the model and PROTEL data over all utilised energy channels from L_{min} to L_{max}, given by

$$\Upsilon(A_m, \alpha, A_e, \beta) = \frac{1}{N} \sum_{L,E} (\log f - \log f_{\text{data}})^2 \tag{4.3}$$

There are several assumptions associated with deriving period-correct diffusion coefficients this way, including that i) all sources and losses are accounted for via accurate modelling; ii) the time averaged data to which the model is compared represents true steady state; and iii) the constraints on D_{LL} imposed via its defining parameters in Equation 4.2 allow for the correct solution.

One must pay particular attention to the first assumption when calculating steady state for a finite average period. A steady state calculation must approximate source and loss processes controlling flux with static inputs over that period, when in fact these processes may be time-varying. For example, seasonal changes in density occur, which means that averaging density over a longer time period may be averaging over variation, and therefore a fixed average is not necessarily representative of the period. For best practise, one should ensure that any time averaged inputs used to calculate empirical source or loss terms are representative across the whole average period. The ∼200 day long CRRESPRO quiet period may seem like an appropriate time average period over which to optimise model

fit, based on the minimal intensity variation shown in Figure 4.1 before the 24th March storm suggesting steady state. However, it may be that the proton belt is in a dynamic equilibrium, whereby changes in source and loss rates do occur, but rebalance to maintain the same level of flux. A dynamic equilibrium could also involve time-dependent radial diffusion, brought about by frequent changes in geomagnetic activity, but the steady state optimisation can only derive time-independent period-average diffusion coefficients.

The selection of an average period in which data represents steady state was therefore influenced by wanting to eliminate variability in this manner. The main concern was plasmaspheric density, given that seasonal changes can affect the steady state solution as shown in Figure 4.3. The variability in the CRAND source is driven by solar cycle and CRAND is therefore relatively constant over the CRRES mission, and inelastic nuclear scattering (driven by neutral density) is known to have only a minor influence on the distribution at the energy range of interest (Albert et al., 1998). To gain insight into proton belt variability over the entire CRRES mission, Figure 4.4 plots electron density through time (first and second panel), along with the change in weekly average flux (plotted as logarithm of the ratio j to j_{t0}) throughout the mission according to several PROTEL energy channels at selected L (third and fourth panels). Figure 4.4 also marks the epoch of all flux measurements near the model outer boundary L over the course of the CRRES mission (bottom panel) to give an idea of data availability. As a first step, these plots were used to select suitable steady state optimisation periods. The time variation in electron density shown in Figure 4.4 can be seen for other constituents, but electron density has been highlighted as it is the main driver of coulomb collisional loss near the model outer boundary.

Figure 4.4 (top panel) shows a gradual increase in electron density through time until around March 1991 where it peaks before a gradual decrease. The main changes in flux (third and fourth panels) coincide with geomagnetic disturbances during the active period. The third panel, showing flux at $L = 1.7$, shows two main enhancements: once following the 24th March storm, and once again coinciding with various SEP/SSC events that occurred near the beginning of June 1991. These enhancements are interesting because they occur below the region of newly injected SEPs and are limited to <10MeV particles. In fact, the flux in the 13.2MeV channel

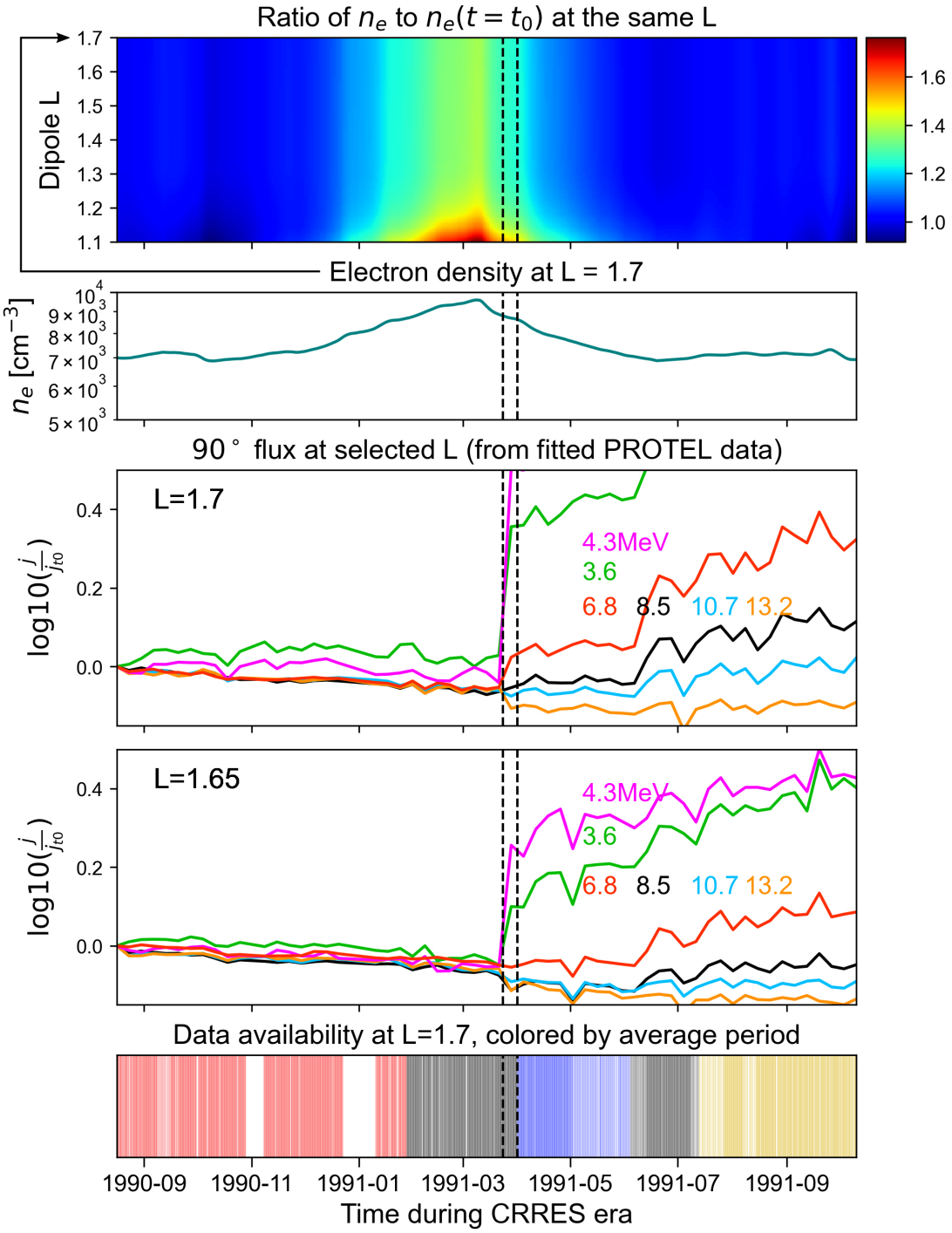


Figure 4.4: Plasmaspheric electron density through time as a ratio to initial value at start of mission (top panel), electron density at $L = 1.7$ (second panel), logarithm of normalised flux through time at selected energy channels (third and fourth panels for $L = 1.7$ and $L = 1.65$ respectively); and data availability at $L = 1.7$ (bottom panel). Changes in density and intensity through time are avoided as much as possible to select suitable average periods, which are indicated in the bottom panel, named Quiet (red), Active 1 (blue), and Active 2 (amber).

and higher appears to decrease following these events.

The ∼five month quiet period following the start of mission was chosen as one average period, limiting it up to 28 January 1991 to keep electron density variation within 25%. During the active period following the March 1991 storm, it is likely that the proton belt is not in steady state, making data here less suitable to drive the model or optimise against. However, if the deviation from steady state throughout the inner zone during the active period is limited enough (in terms of the range of affected L and energy channels), the minimisation of Υ may still be dominated by comparison with steady state f_{data}, and similar D_{LL} may be derived. By filtering out specific energy channels affected by active period enhancements, it is therefore possible to select additional data average periods following the March 1991 storm and repeat the steady state optimisation method to get similar results as for the quiet time average.

Three average periods were selected in total, indicated by the red, blue and amber colours in the bottom panel of Figure 4.4. The first of these average periods, expected to produce the best fit using a steady state model, is the Quiet period (red) from 15/08/1990 to 28/01/1991. The second average period, Active 1 (blue), spans from 31/03/1991 to 04/06/1991, and the third period, Active 2 (amber), spans from 13/07/1991 to 15/10/1991. Time averaged density over each of these periods was used to compute the corresponding loss timescale for a given μ and L, and time averaged 90° flux was used to fix the model outer boundary. To filter out data that is likely to be out of steady state, the Active 1 and Active 2 optimisation of D_{LL} was limited to only include data at $\mu \geq$105MeV/G, because this corresponds to ∼6.8MeV at L_{max}=1.65, which is the highest energy channel indicating a major disturbance during the active period (red line in Figure 4.4, fourth panel). At the next energy channel and above, flux is relatively unaffected at L_{max} and below.

Υ was minimised for each period to derive three sets of diffusion coefficient parameters α, β, A_m and A_e. The minimisation was performed using the Nelder–Mead method implemented in the SciPy library (Virtanen et al., 2020), with the same initial guess parameters as used by Albert et al. (1998), taken from Schulz (1991): $A_{m0} = 7\text{e}{-9}$, $A_{e0} = 1\text{e}{-4}$, $\beta_0 = 2$, $\alpha_0 = 2$. Convergence of the method on consistent results was verified by modifying the initial guess and repeating a set of runs for the Quiet period.

4.4.2 Selection of Outer Boundary L

Another interesting feature of Figure 4.4 is the observed decrease in flux across all energy channels at $L = 1.65$ during the quiet period (fourth panel). Given the region and lack of enhancements during this time, these fluxes are expected to remain near steady state. This change coincides with the gradual increase in electron density given by the drift average density model (second panel), which is the primary driver of loss at this L. Taken together, this appears to show the type of response demonstrated in Figure 4.3, whereby steady state flux can change due to seasonal changes in density. A decrease in flux would be expected to accompany an increase in density (increasing loss), and the observed fluxes therefore appear to corroborate the results of the F10.7a and day of year-dependent drift averaged density calculation.

The gradual decrease in flux is also visible at $L = 1.7$ (panel three), except in the lowest two energy channels shown (3.6 and 4.3MeV), where no systematic decrease is observed. The flux of these two energy channels is much less steady and seems to suggest that proton belt flux is subject to dynamic changes near $L = 1.7$ even during a quiet period. This highlights the approximation of considering an "inner" and "outer" zone, given the region of flux affected by dynamic changes is energy dependent. The systematic decrease in flux observed at $L = 1.65$ motivated the choice of $L_{max} = 1.65$ as the model outer boundary. Although the changes in 3.6 and 4.3MeV flux at $L = 1.7$ appear to be small, changes in outer boundary flux can have a large impact on steady state calculations considering this flux is radially diffused inward to form the profile of the inner belt. Therefore, choosing $L_{max} = 1.65$ strengthened the steady state assumption compared to using $L_{max} = 1.7$ or higher.

4.5 Results and Discussion

4.5.1 Optimisation Results

The final converged values of α, β, A_m and A_e for the Quiet, Active 1 and Active 2 periods are shown in Table 1, along with the corresponding value of Υ indicating how closely the steady state solution was able to fit the data. Υ is higher for

the Quiet optimisation, but this is partly due to the exclusion of $\mu \leq$105MeV/G data during the Active 1 and Active 2 optimisations, meaning the summation in Equation 4.3 was performed over $N = 105$ values of f_{data} during the Quiet period, compared with only $N = 60$ during the Active 1 and Active 2 periods. The decrease in Υ from Quiet to Active optimisations is despite a similar fit being achieved, and can be explained by the grouping of excluded data: in the Active 1 and 2 case, energy channels across the low energy range of the model are ignored, and the optimisation is therefore able to better fit channels at higher energy, lowering the average mean squared difference. In previous investigations before the exclusion of $\mu \leq$105MeV/G data, the Active 1 and Active 2 optimisations were found to produce higher Υ for the same $N = 105$, indicating a closer fit for the Quiet period. This variation in Υ implies potential caveats when comparing the performance of runs based on the value of a minimised parameter calculated with a different set of f_{data}.

The optimisation path from initial guess to converged values is shown for the Quiet period in Figure 4.5. The process was found to complete fastest when optimising in terms of log10 (A_m) and log10 (A_e), and Figure 4.5 shows a logarithmic axis for A_m and A_e to reflect this. During convergence, Υ reduces quickly at first, and after $\sim$80 iterations the minimisation continues but makes negligible changes to the resultant D_{LL}. A similar phenomenon was observed for both the Active 1 and Active 2 optimisation runs.

Using the optimised parameters, Equation 4.2 gives D_{LL} as a combination of an electromagnetic and electrostatic component. According to the optimised values the electrostatic component does not contribute significantly in any case, being a few order of magnitudes lower than the electromagnetic component, and a similar result is shown by the optimised parameters of Albert et al. (1998). However, the optimisation process relies on Equation 4.2 only as a means to derive the μ and L dependence of total D_{LL} to best fit the data. As such there may be a limit on the physical interpretability of this result, because errors in the data, or evaluation of source/loss terms, may slightly alter the D_{LL} that gives best fit, and may therefore also alter the balance between electromagnetic and electrostatic diffusion indicated by the optimised parameters which appears to be quite sensitive.

Figure 4.6 shows the solution in steady state phase space density (blue) for each

Period	A_m	A_e	β	α	Υ
Quiet	1.536e−6	3.554e−4	2.657	2.740	0.00553
Active 1	1.381e−6	1.472e−4	2.332	2.641	0.00374
Active 2	8.775e−7	4.582e−5	2.147	2.546	0.00353

Table 4.1: Optimised diffusion coefficient parameters for the Quiet, Active 1 and Active 2 time periods along with the corresponding minimised Υ

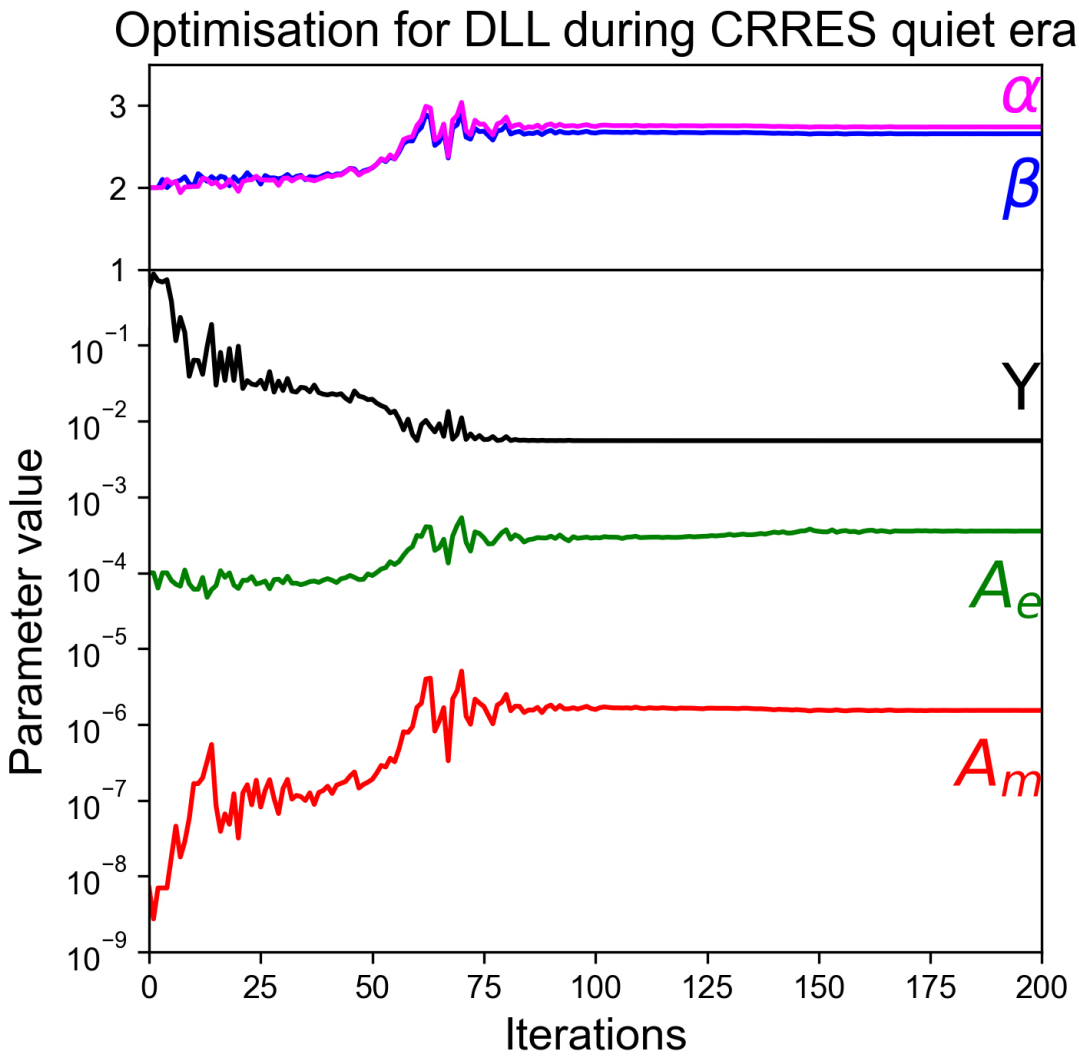


Figure 4.5: Optimisation path of A_m, A_e, β, α for steady state optimisation over the Quiet period, showing convergence from the initial guess values to final values as the model vs. data fit improved, minimising Υ (black line)

period as radial profiles, plotted alongside PROTEL data (crosses). Data included in (excluded from) each optimisation is marked in blue (red). Excluded data fulfils one of the following criteria: i) located at $L < 1.35$; ii) has $\mu \leq$105MeV/G during the Active 1 or 2 period; or iii) has E>45MeV (this criterion is explained below). The Quiet, Active 1 and Active 2 results occupy columns 1, 2 and 3 respectively, with different values of μ shown in each row. Figure 4.6 generally shows a good match between the model result and included PROTEL data for each average period, meaning that a set of diffusion coefficients could be found in each case to explain the data assuming steady state. The Quiet optimisation (left column) appears to produce the best fit in general, though the difference between average periods is small. In order to give an idea of data stability, the standard deviation is plotted (black bars) for each data point. This represents the standard deviation in 90° flux from each equatorially mapped pitch angle distribution at fixed energy channels, converted to phase space density, and then interpolated at fixed μ. It is therefore an approximate measure, and in fact the data is somewhat more stable than it indicates because the $\sin^n$ fitting used by the model to read pitch angle distributions is not heavily influenced by a high level of variation in the 90° values alone. Nevertheless, the standard deviations shown indicate that data from the Active periods contain higher levels of variation, particularly at low μ. This is to be expected because the data corresponds to an average following an enhancement, and data availability is also more scattered (bottom panel of Figure 4.4). In general, the highest standard deviations for data in each period corresponds to $L < 1.4$, where the data is known to be less reliable due to the loss cone correction (see Section 4.2.2). Comparison between the Quiet and Active 2 period data also shows the enhanced boundary f caused by the series of SEP/SSC events in 1991 at μ=50MeV/G. The Active 2 profile at μ =50MeV/G appears out of steady state because of the enhanced boundary, and excluding $\mu \leq$105MeV/G data has stopped the optimisation process from trying to fit this feature.

Data at μ=500MeV/G (row 6 of Figure 4.6) forms a plateau in the region $1.3 < L < 1.5$, and all three solutions show a deviation from data at μ=500MeV/G near this feature. Further investigation indicates this deviation is better described in terms of energy, happening at ∼45MeV and above, and can thus also be seen at μ=400MeV/G at low L (row 5 of Figure 4.6). The same observation was noted

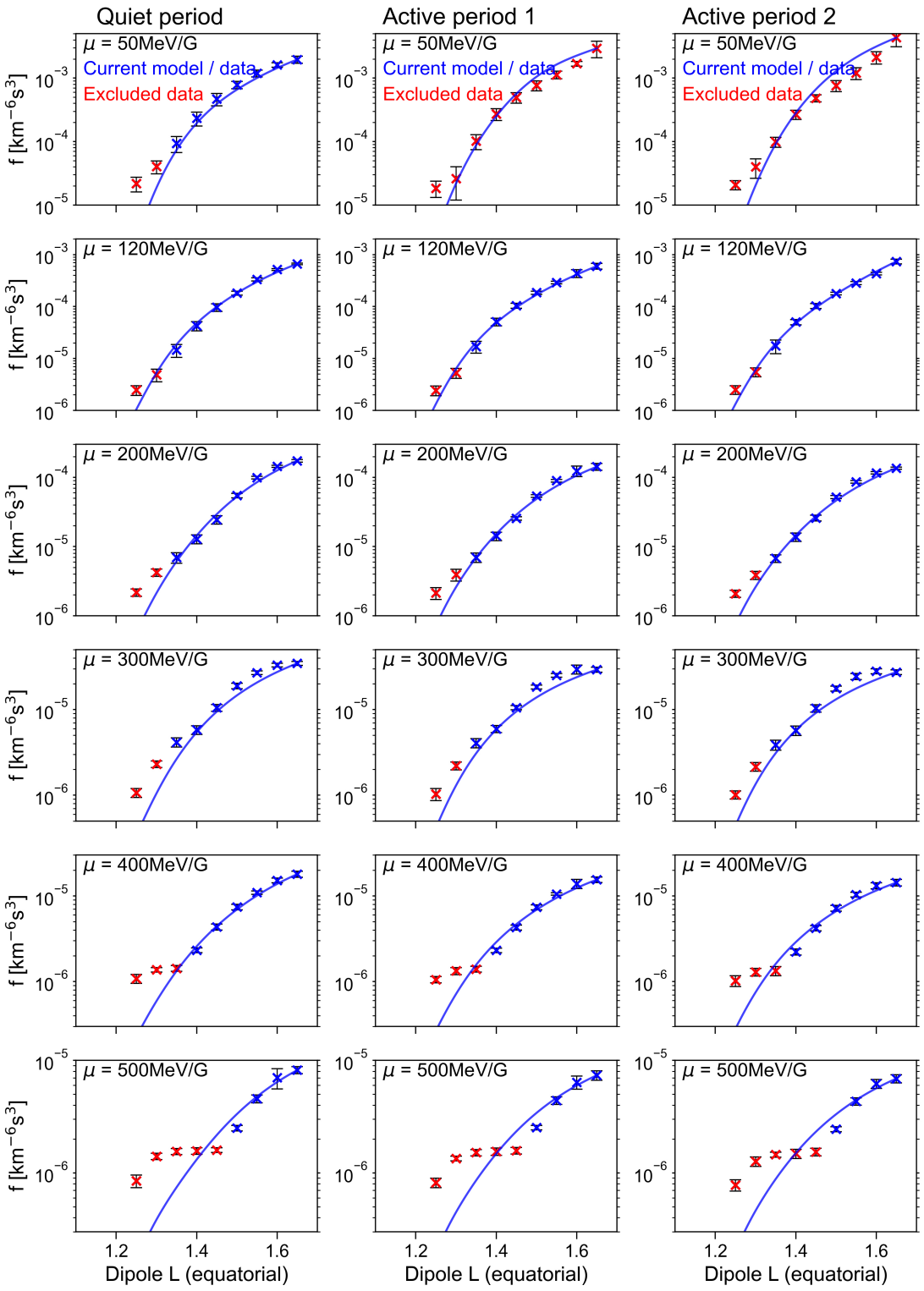


Figure 4.6: Results of the three optimisations over the Quiet, Active 1 and Active 2 periods (columns) showing model steady state phase space density (blue) plotted against PROTEL data (red crosses) for different values of μ(rows). Standard deviation in phase space density is approximated based on the available observations of flux close to 90°, and plotted as black bars surrounding each data point.

by Albert et al. (1998), and the difficulty that both steady state models have in reproducing this feature could be due to flux at E$\gtrsim$45MeV being out of steady state. Therefore, the four energy channels in this range have been excluded from the optimisation process to weight the optimisation using steady state data as much as possible. The feature could be the result of an injection prior to the modelled period which has diffused inwards, or alternatively could be the result of incorrect data/processing. As Albert et al. (1998) note, the profile at this μ appears to be stable and quite unaffected by the storm on 24th March, 1991.

4.5.2 Comparison to Previous Work

Figure 4.7 shows a direct comparison between the new steady state profiles derived for the Quiet period versus the solution calculated by Albert et al. (1998). As Albert et al. calculated six solutions using different combinations of CRAND and density models, the solution appearing to give the best fit of these six has been chosen for comparison in Figure 4.7. This solution relies on the CRAND source of Claflin and White (1974) and the Parameterised Ionospheric Model for electron density (Daniell et al., 1995). The data processing steps taken and average window used were not the same as used by Albert et al. to prepare the data, and therefore some disagreement is expected between the data itself. Figure 4.7 therefore shows both the data and modelling results from both works. There is a particularly noticeable discrepancy between the data at high μ, which may arise from the different definitions of phase space density. Figure 4.7 indicates that a closer fit to the data is achieved by the current model (blue) compared to the Albert et al. model (red), especially at $\mu \leq$200MeV/G, although region to which the current model was applied is slightly more limited, having an outer boundary at $L = 1.65$ compared to $L = 1.7$. At μ=500MeV/G, a slightly closer fit is achieved in general using the current model, but neither model is able to recreate the plateau noted previously.

The diffusion coefficients derived during the steady state optimisations carried out over all three time average periods can be compared to diffusion coefficients derived by other works. Figure 4.8 shows the diffusion coefficients given by the optimised values in Table 1 at μ=100MeV/G and μ=500MeV/G (red, blue and

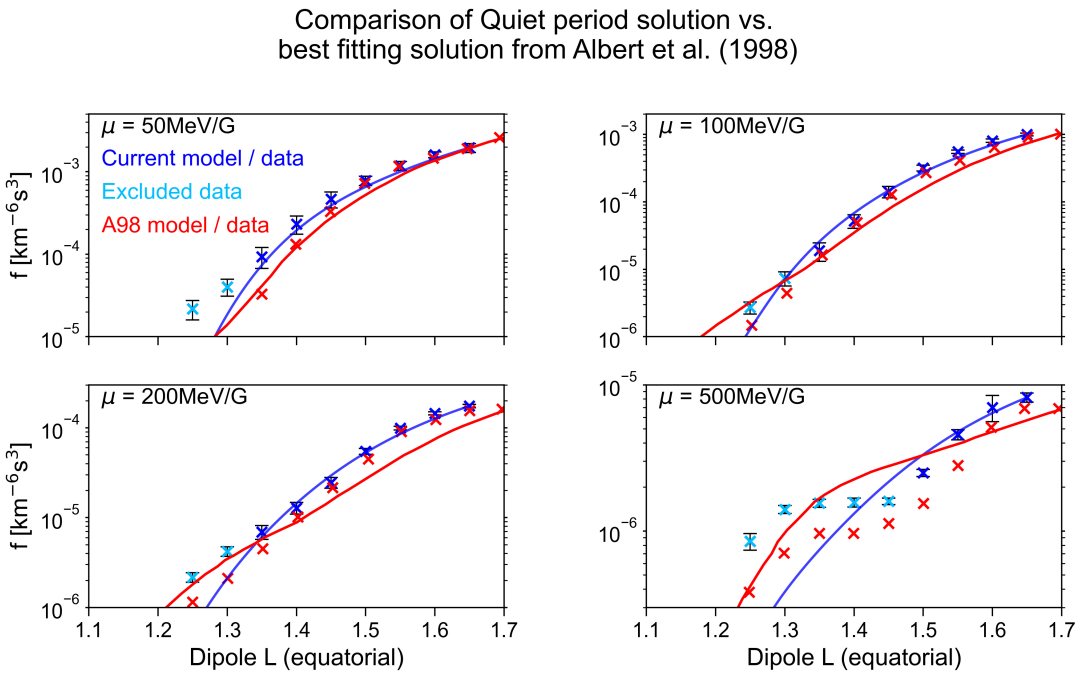


Figure 4.7: Steady state profiles with optimised diffusion coefficients derived using the current model (blue lines), and the model run by Albert et al. (1998, , red lines) which uses the CRAND source of Claflin and White (1974) and electron density from the Parameterised Ionospheric Model of Daniell et al. (1995). Data is plotted as blue and red crosses, corresponding to each model. Data used by the current model represents a time average over the Quiet period of this work (see Figure 4.4), whereas data used by Albert et al. (1998) represents the time average from the CRRESPRO Quiet model.

amber lines). These D_{LL} are compared to the overall D_{LL} derived by Ozeke et al. (2014) for electrons at $L > 2.5$ using measurements of ULF wave power, which depend on activity level parameterised by Kp index. As Lejosne et al. (2013a) show that the energy dependence of radial diffusion is weak, there should be a close correspondence between proton and electron diffusion coefficients. Another difference is that the coefficients given by Ozeke et al. (2014) are derived using the method described by Fei et al. (2006), whereby diffusion rates are split into a magnetic and overall electric component, and this is shown to result in an underestimation by a factor of $\sim$2 (Lejosne, 2019a). In Figure 4.8, these coefficients are labelled as O14 and are shown for three different values of Kp index (light blue, plotted only at $L \geq 2.5$). In addition, the result of Albert et al. (1998) from Figure 4.7 is plotted, labelled as A98 (dashed red line).

Figure 4.8 shows that the newly derived D_{LL} is higher than results by Albert et al. (1998) by a factor of $\sim$2 to $\sim$5 depending on μ. Figure 4.8 plots results up to $L = 3$, despite the fact coefficients were only optimised over the inner zone. Analytical expressions for D_{LL} given by Ozeke et al. (2014) were derived using data at $L > 2.5$. Therefore, it is interesting to note the region of overlap at $1.65 < L < 2.5$, into which both sets of diffusion coefficients can be extrapolated. Two key features of the diffusion coefficients given by Ozeke et al. (2014) are energy independence, compared with the fairly weak energy dependence derived in this work, in addition to the dependence on Kp. For the innermost proton belt region, this dependence on Kp may be different because the inner zone is for the most part shielded from the geomagnetic activity at larger L that can lead to faster diffusion rates. However, the optimisation method is only able to quantify the effects of radial diffusion over a long time scale, as required to form the steady state distribution of protons over the data average period. If the timescale of radial diffusion varies significantly with activity level, as implied by the results of Ozeke et al. (2014) in Figure 4.8, the newly derived radial diffusion coefficients should be considered averages near solar maximum of a time-varying process. This could be another reason that a perfect fit was not achieved against the data, as parts of the distribution may have deviated from steady state during each average period due to enhanced radial diffusion over short timescales, given the higher levels of geomagnetic activity around solar maximum. Therefore, an interesting question is

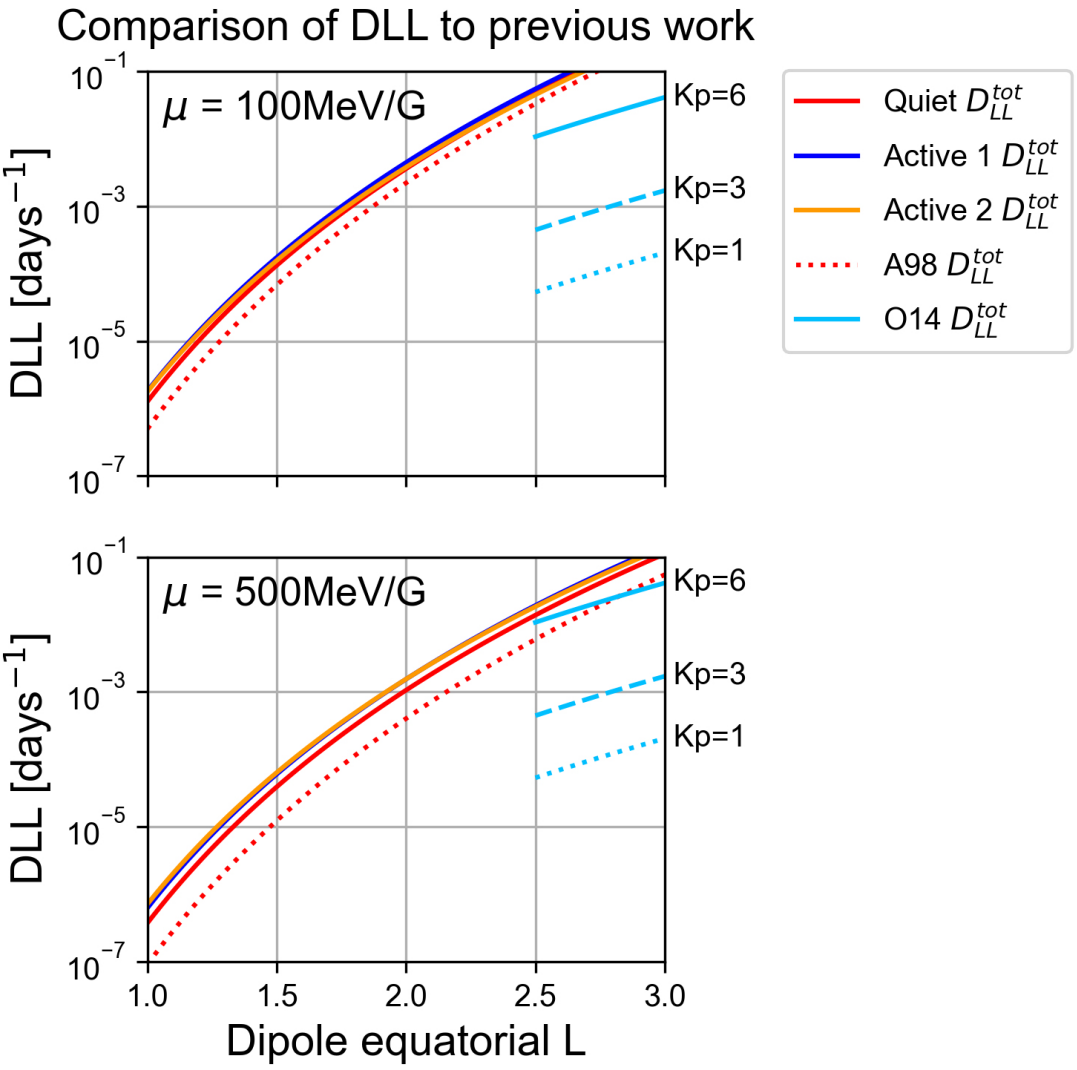


Figure 4.8: Diffusion coefficients derived from each optimisation over the three average periods, plotted next to those derived in Ozeke et al. (2014) and the model run from Albert et al. (1998) shown in Figure 4.7 (labelled O14 and A98 respectively). Top and bottom panels correspond to $\mu = 100$ and 500MeV/G respectively.

how activity dependence should be taken into consideration when extrapolating diffusion coefficients to higher or lower L than the region they were derived in, and how they can be applied to examine shorter timescales. A comparison of results to diffusion coefficients derived at low L during solar minimum may be a useful starting point to better understand how long term averages can vary.

4.6 Conclusions

Steady state phase space density was solved for using a model that includes a physics-based evaluation of CRAND, and a drift averaged density model to drive coulomb collisions capturing solar cycle and seasonal variation, with the outer boundary set by PROTEL data. The fit between the model and PROTEL data in the region $1.1 \leq L \leq 1.65$ was optimised following a similar method to Albert et al. (1998) to recalculate time averaged radial diffusion coefficients that govern the transport of relativistic protons in the proton belt during solar maximum.

Measures were taken to improve the modelling process conceptually, by selecting time averages that exhibited minimal variability, and by carefully excluding data to strengthen the assumption of steady state. It was found that a suitable time average for steady state optimisation should be less than six months to avoid potential seasonal variations in plasmaspheric density. Diffusion rates required to explain steady state were sensitive to the density, but density is not well constrained in general by measurements between the topside ionosphere out to $L \sim 1.7$ which leads to some uncertainty.

The similarity between diffusion coefficients derived in this study for three time periods indicates that although the 24th March storm caused a large enhancement, steady state optimisation could still be performed during the Active 1 and 2 periods with reasonable results. As only a short interval of a few months or less is required to average data, steady state optimisation could be performed at regular intervals throughout a long dataset to derive updated diffusion coefficients. In general, radial diffusion coefficients derived here are comparable to but higher than the previous work of Albert et al. (1998) by a factor of 2 to 3 at μ =100MeV/G and a factor of 3 to 5 at 500MeV/G, and provide a better fit to PROTEL data.

Due to the long time averages, the optimisation method for deriving diffusion

coefficients is not able to uncover activity-dependent variability as is suggested by diffusion rates derived from in situ and ground based measurements for electrons at $L > 2.5$, which vary by an order of magnitude or more for a small change of $\sim$2 in geomagnetic activity index Kp. Further investigation is required to fully understand the activity dependence of proton radial diffusion coefficients in this region of interest.

Chapter 5

3D Model Application: Modelling Inner Proton Belt Variability at Energies 1 to 10MeV

This chapter is based on a research article:

Modeling Inner Proton Belt Variability at Energies 1 to 10 MeV Using BAS-PRO

JGR Space Physics, December 2021, Volume 126, Issue 12

https://doi.org/10.1029/2021JA029777

Alexander R. Lozinski[ab], Richard B. Horne[a], Sarah A. Glauert[a], Giulio Del Zanna[b], Seth G. Claudepierre[cd]

[a]British Antarctic Survey, Cambridge, UK

[b]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK

[c]Space Sciences Department, The Aerospace Corporation, El Segundo, CA, USAfn:note3

[d]Department of Atmospheric and Oceanic Sciences, UCLA, Los Angeles, CA, USAfn:note4

5.1 Introduction

Irradiation by trapped proton flux in the 1-10MeV range is a primary cause of solar cell degradation for spacecraft traversing the proton belt, and a key factor influencing mission lifetime (Miyake et al., 2014; Jenkins et al., 2014). Predicting variability at this energy is therefore of practical importance for evaluating the operational risk to satellites. One key challenge is modelling the effect of radial diffusion at this energy. Chapter 4 made use of theoretical work by Fälthammar (1965) to express the radial diffusion coefficient D_{LL} as the contribution of magnetic and electrostatic terms. The magnetic term is often assumed to have a L^{10} dependence and decreases by up to $\sim$90% with decreasing equatorial pitch angle (Walt, 1971), whilst the electrostatic term varies with L^6 and is nearly independent of pitch angle (Lejosne and Kollmann, 2020). Analytical expressions for D_{LL} by Lejosne (2019b) include an inverse energy dependence, which has been previously demonstrated for electrons by modelling asymmetric field perturbations to derive D_{LL} analytically (Lejosne et al., 2013b). For protons, an energy-dependence was also inferred by the optimisation process in Chapter 4 by matching the results against spacecraft measurements of flux. In general however, the dependencies of proton D_{LL} are not well constrained, especially at $\lesssim$10MeV where data is mostly unavailable.

The strong L dependence of D_{LL} means variations in MeV proton phase space density at low altitude ($L \lesssim 1.3$) are mostly driven by coulomb collisional loss, whereby free and bound electrons in the atmosphere, ionosphere and plasmasphere decelerate protons and reduce the first invariant μ (see Section II.2, Schulz and Lanzerotti, 1974). During solar maximum the atmosphere undergoes thermal expansion from increased extreme ultraviolet radiation (Fuller-Rowell et al., 2004). At fixed altitude in the radiation belts, this leads to higher density and therefore higher loss. Li et al. (2020) show this effect, plotting cyclical variations in proton flux near $\sim$40MeV lagging behind changes in sunspot number, with a delay of hundreds of days just below $L = 1.2$ for equatorial particles. The relatively short timescales for variation indicate how trapped flux levels quickly rebalance changes in loss rather than conserve a previous state, in contrast to higher altitudes ($L \sim 1.6$) where radial diffusion controls variability over decades. However, at $L \gtrsim 2$, months

long enhancements in proton belt flux have been recorded forming over ∼day timescales or less (Hudson et al., 1995; Lorentzen et al., 2002) due to trapping of incoming solar energetic particles (SEPs, see Kress et al., 2004, 2005). The radiation environment of satellites orbiting at $L \lesssim 2$ may therefore be subject to long term increases, driven by solar cycle or magnetic activity, where little data is available for monitoring.

In this Chapter, the 3D model developed in Chapter 3 is applied to investigate variability in ∼MeV proton phase space density at $1.15 \leq L \leq 2$ as a function of the three adiabatic invariants μ, K and L. To drive the model, a dynamic outer boundary at $L = 2$ is constructed using proton flux data down to 0.7MeV from the RBSPICE and MagEIS instruments on the Van Allen Probes mission (Mitchell et al., 2013; Blake et al., 2013). This data, described in Section 5.2, allows modelling variability over the period from January 2014 to March 2018. Section 5.4 presents results of dynamic models run initialised from steady state for three sets of D_{LL} values taken from literature which exhibit various energy dependence, in order to highlight the sensitivity of results. Several features of the results are then discussed in Section 5.5. In particular, by showing the sensitivity of flux levels to D_{LL} at low energies relevant for satellite solar cell degradation, a key practical impact of uncertainty in D_{LL} is demonstrated.

5.2 Proton Data

The Van Allen Probes pair of satellites (formerly known as the Radiation Belt Storm Probes, RBSP) were launched into elliptical orbit (∼600km pergiee to $\sim 5.8 R_E$ apogee) at 10° inclination on 30 August 2012 (Kessel et al., 2013). This work makes use of proton flux measurements collected by three instruments on board: the Radiation Belt Storm Probes Ion Composition Experiment (RBSPICE, Mitchell et al., 2013); the Magnetic Electron Ion Spectrometer (MagEIS, Blake et al., 2013); and the Relativistic Electron-Proton Telescope (REPT, Baker et al., 2012). At $L < 2$, certain measurements were contaminated by the unintended counting of electrons and higher energy protons. Therefore, processed data from all three instruments were combined to derive a spectrum at $L = 2$, the innermost region where this interference could be avoided. This section describes processing of

RBSPICE and MagEIS measurements from Van Allen Probe B (RBSP-B) to derive ∼0.7 to 10MeV proton flux from January 2014 to March 2018. Proton flux data at >19MeV shown by Selesnick and Albert (2019), based on REPT measurements and covering the same period, was then used to extend the spectrum.

5.2.1 RBSPICE Measurements up to 1MeV

5.2.1.1 Processing

The RBSPICE instrument measures ions from ∼20keV to several MeV. The type of data collected by the instrument sensor, and subsequently the data product generated, depends on the selected "hardware mode" at any given time. The availability of RBSPICE data products is therefore determined in part by which hardware modes were enabled at the time of data collection. The data availability from several products was investigated, and it was found that extracting proton flux measurements at $L = 2$ over the model period was only feasible via the Ion Species High Energy Resolution Low Time Resolution (ISRHELT) product measuring ion spectra (Manweiler and Mull, 2017). However, despite its better availability, ISRHELT measurements are susceptible to electron contamination in certain regions due in part to the reduced accuracy of the hardware mode and, in addition, there is no discernment between ion species. This section describes the processing steps performed on ISRHELT data, and these two potential caveats are addressed in more detail throughout Section 5.2.1.2.

ISRHELT data is contained within the level 3 Common Data Format (CDF) files obtainable online at `http://rbspiceb.ftecs.com/Level_3/ISRHELT/`. This data (from Van Allen Probe B) was used to derive a time series of equatorial pitch angle distributions for each energy channel over the modelling period, as described below.

The CDF files provide proton differential unidirectional flux as a 3D array of values, with dimensions epoch, energy channel and telescope. There were six telescopes recording simultaneously, and each measurement of flux was taken in the instantaneous look direction of the corresponding telescope, rotating with the spacecraft. To allow for angular resolution of measurements, the CDF files also provide a 2D array of the telescope look directions in terms of local pitch angle,

with dimensions epoch and telescope number. In order to capture variability at sufficient time resolution, the modelling period was first split into intervals six days long. Within each interval, data was then preprocessed according to the following steps: i) the Python interface to IRBEM provided by the spacepy package (Morley et al., 2011) was used to calculate B/B_e (the ratio of local magnetic field strength to magnetic field strength at the equator along the local field line) at each measurement epoch, using the IGRF internal and Olson-Pfitzer quiet external magnetic field (Alken et al., 2021; Olson and Pfitzer, 1982); ii) the 2D array of telescope look directions was converted from local pitch angle α to equatorial pitch angle α_{eq} using the well-known relation

$$\frac{\sin^2(\alpha)}{B} = \frac{\sin^2(\alpha_{eq})}{B_e} \tag{5.1}$$

derived from conservation of the first invariant; iii) values of equatorial pitch angle were placed in one of 15 equally spaced bins spanning 0° to 180° and of width $\Delta\alpha_{eq} = 12°$, with the first bin centre at $\alpha_{eq} = 6°$; iv) each flux measurement was associated with an equatorial pitch angle bin via the recording telescope's look direction, and a cadence was applied to average the flux measurements in each bin across one minute intervals, resulting in a 4D array of flux with dimensions of time (at the centres of each one minute interval), energy channel, telescope and equatorial pitch angle bin; v) for a given L, data outside $L \pm 0.02$ were filtered out using the spacecraft L location at each epoch provided within the CDF files.

This method was used to examine the equatorial pitch angle distribution of a given energy, formed by one minute-averaged flux measurements collected across the six day period for which data was extracted. Measurements from the first telescope were ignored because they were found to cover only a narrow range in equatorial pitch angle. This was due to the telescope being centred close to the spacecraft spin axis and showing little spin modulation. Measurements from the five remaining telescopes were combined, and the equatorial pitch angle distributions at each energy were fitted using the function

$$j = A \sin^n(\alpha_{eq}) + c \tag{5.2}$$

where j is unidirectional flux at equatorial pitch angle α_{eq}, and A, n and c are the fitting parameters. By repeating the above process at each six day interval over the modelling period, the time series of fitted data for each instrument channel was derived.

Figure 5.1, left side of panel a, shows an example fitted pitch angle distribution at $L = 2$ from early June 2014 in the 0.69MeV channel of ISRHELT, with the different colours corresponding to one minute-averaged flux measurements taken by different telescopes. The standard deviations of the one minute-averaged fluxes in each pitch angle bin are shown by the black bars in Figure 5.1 and indicate data variability as well as reliability. Non-zero flux at loss cone pitch angles was assumed to be a consequence of insufficient angular resolution of the measurements rather than penetrating background, because the telescope look directions given in the CDF files represent the centre of a few degrees range in local pitch angle, and were then binned after being converted to equatorial pitch angle as part of preprocessing, leading to some loss of precision. Some channels exhibit high standard deviations depending on the date and region. To give an overall sense of data availability and quality, Figure 5.1, right side of panel a, shows the ratio of the standard deviation to the absolute flux value in the 90° bin, taken from fits to the data at $L = 2$ over each time interval. Periods where data is unavailable are left unshaded (white). This plot corresponds to the 0.69MeV channel but shows results representative for all channels of ISRHELT used in the study. Data is generally useable at $L = 2$ where the standard deviation of flux is $\leq$50% of the bin average.

5.2.1.2 Data Issues and Validation

The ISRHELT data product was collected using the "energy" mode of the RBSPICE instrument. In this mode, incident particle energy is measured by ion solid state detectors. However, there is no magnet in the RBSPICE detector to sweep out electrons, and so ISRHELT measurements may record the arrival of both species leading to contamination of ion flux. Another useful product from RBSPICE is the Time of Flight by Energy Ion Species Rates (TOFxEIon), which also provides flux of ions. However, in the hardware mode used for TOFxEIon, "time of flight" data is collected whereby a microchannel plate detects secondary electrons produced

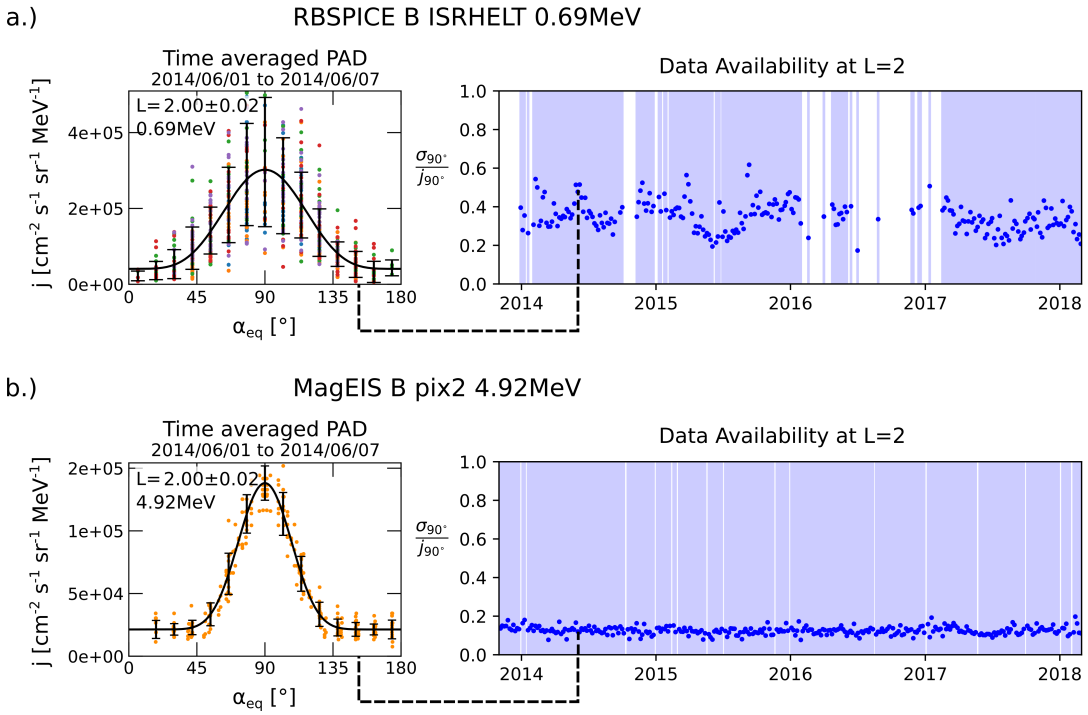


Figure 5.1: Summary of the data availability and quality of fit for preprocessed pitch angle distributions given by the RBSPICE ISRHELT data (panel a) and MagEIS pix2 data (panel b). Equatorial pitch angle distributions fitted using Equation 5.2 are shown on the left, along with the standard deviation of flux in each pitch angle bin. On the right, the ratio between the standard deviation in 90° flux measurements versus the mean flux value is shown for fitted pitch angle distributions at six day intervals at $L = 2$. White unshaded regions indicate a lack of data.

by the passage of an ion through two thin foils prior to the solid state detector, allowing detector counts to be validated as ions (Mitchell et al., 2013).

There are only a few short periods of time over which the TOFxEIon product is available down to $L = 2$, nevertheless they allowed a comparison to be made against ISRHELT. The susceptibility of ISRHELT data to electron contamination was therefore tested by preprocessing the TOFxEIon measurements using a similar method, and comparing the datasets at $L \geq 2$ over several months from April to mid-August 2017. When comparing 90° differential flux at $L = 2$ over this period, a very good agreement was found between the fitted flux distributions in the 0.84 and 0.93MeV channels between instruments, and a reasonable (within factor of 2) agreement for the 0.63 and 0.76MeV channels. Electron contamination issues in TOFxEIon data were not expected due to the time of flight data providing better accuracy, suggesting that ISRHELT was also not subject to major electron contamination in the region of interest. It was assumed that protons dominate the ion population in the region of interest, and therefore that ISRHELT and TOFxEIon ion measurements represent protons.

To continue checking for electron contamination, the fitted pitch angle distributions derived for ISRHELT (Section 5.2.1.1) were compared at different altitudes. When examining the radial profile of 90° flux versus L in energy channels up to 1MeV, two peaks were observed in some low energy channels. A peak at high altitude ($L > 3$) was expected, given the nominal distribution of flux according to previous measurements (i.e., Figure 5 of Stassinopoulos and Raymond, 1988). However, channels at 0.52MeV and below exhibited a secondary flux peak at $L < 2$ which interfered with measurements at $L = 2$. These features were attributed to interference at $\lesssim$0.5MeV, possibly from electrons. Finally, the energy spectrum at $L = 2$ derived from the fitted pitch angle distributions was found to show increasing flux with increasing energy, starting from the 1.85 and 2.03MeV channels. This leads to a factor of $\sim$2 disagreement compared with the 2.05MeV channel on MagEIS pix2 (see Section 5.2.2.1), and suggests that ISRHELT data at $\gtrsim$2MeV may also be unreliable. Therefore, use of RBSPICE ISRHELT data was limited to the six energy channels covering 0.69 - 1.13MeV at $L = 2$, where no signs of contamination were found.

5.2.2 MagEIS Measurements up to 10MeV

5.2.2.1 Processing

There are four MagEIS instruments on each Van Allen Probe. The "low" and "medium" energy units are electron spectrometers and do not measure ions. The "high" unit electron spectrometer also houses an ion range telescope with three silicon detectors. On Van Allen Probe B, ∼2 to 20MeV protons are measured by the 2500-micron detector in this arrangement (Blake et al., 2013). This data is accessible via the "FPDU_pix2" variable in the Level 3 RBSP-B CDF files available online at the RBSP-ECT Science & Data Portal (`https://rbsp-ect.newmexicoconsortium.org/data_pub/rbspb/`). The excellent data continuity of "pix2" data over the modelling period allowed RBSPICE data to be supplemented with these higher energy measurements. However, as pix2 data was collected by a separate instrument, different processing steps were required as described below. Different contamination issues also arose, addressed in Section 5.2.2.2, but were comparatively minor.

The CDF files provide a 3D array of differential unidirectional flux with dimensions epoch, local pitch angle and energy channel. Local pitch angle is in terms of 15 equally spaced bins spanning 0° to 180° and of width $\Delta\alpha_{eq} = 12°$, with the first bin centre at $\alpha_{eq} = 6°$. A time series of equatorial pitch angle distributions was derived for each energy channel by splitting the modelling period into six day long intervals (the same as used to process RBSPICE ISRHELT data). Preprocessing the data was somewhat simpler because measurements of the full local pitch angle distribution were available at each epoch. Within each interval, the method used to preprocess data was as follows: i) B/B_e was calculated at each epoch using Equation 5.1, again using the IGRF internal and Olson-Pfitzer quiet external magnetic field via spacepy; ii) local pitch angles at the centre of each of the 15 bins were mapped to equatorial pitch angle at each epoch, and the mapped values stored directly (not re-binned) so that flux values in the 3D array were associated with equatorial pitch angle; iii) for a given L, data outside $L \pm 0.02$ were filtered out using the spacecraft L from the CDF files. This method resulted in every observation of flux within the six day interval being associated with equatorial pitch angle. Equation 5.2 was then used to fit an average distribution over all observations, and the process was

repeated for each six day interval over the modelling period.

Figure 5.1, left side of panel b, shows an example fitted pitch angle distribution at $L = 2$ from early June 2014 in the 4.92MeV channel. The standard deviations of flux are calculated across the distribution by binning observations in equatorial pitch angle, using the same 15 bins used to specify local pitch angle in the CDF file. The standard deviations are shown by the black bars in Figure 5.1. Compared with the standard deviations shown for ISRHELT data in panel a, the standard deviations in pix2 flux is low. This indicates that higher energies exhibit less variability over the same six day window. Figure 5.1, right side of panel b, also shows the ratio of the standard deviation to the absolute flux value in the 90° bin, taken from fits to the data at $L = 2$ over each time interval. The data shown for MagEIS pix2 at 4.92MeV in Figure 5.1 is representative for all channels.

5.2.2.2 Data Issues and Validation

Using the 4.92MeV and 22.5MeV pix2 data to calculate omnidirectional flux results in a strong correlation in values between the two channels at $L < 1.9$, indicating that data is highly contaminated in this region (H. D. R. Evans, personal communication, 2019), presumably from energetic (100s MeV) inner belt proton contamination. In addition, there are periodic spikes in intensity in channels below ∼4 MeV at $L > 3$, indicative of contamination by Bremsstrahlung in the presence of multi-MeV electrons. However, at $L = 2 \pm 0.02$, pitch angle distributions do not show obvious signs of contamination and have low intensity in the loss cone. At $L = 2$ below ∼5MeV, there is a reasonably close agreement (in general by a factor of ∼2 or less) between this data and fluxes modelled using AP9 V1.50 mean (Ginet et al., 2013). At > 10MeV, this data has worse agreement and, on the highest energy channel (22.5MeV), directional flux at 90° pitch angle is much lower than recorded by the low energy channels on the Relativistic Electron-Proton Telescope (REPT) instrument. Use of MagEIS pix2 data was therefore limited to the seven energy channels covering 2.05 - 9.38MeV at $L = 2$.

5.2.3 Energy Spectrum

After restricting the data as described above, flux was available at six energy channels from the RBSPICE ISRHELT product and seven channels from the MagEIS pix2 detector. Fitted equatorial pitch angle distributions were used to derive a time-varying energy spectrum at $L = 2$ across the ∼four year modelling period, which could then be used as the outer model boundary in a numerical simulation. The spectrum is shown in Figure 5.2 for two values of equatorial pitch angle (45° and 90°), with the energy of each data channel indicated by vertical coloured bars for each instrument. In addition to ISRHELT (red lines in Figure 5.2) and MagEIS pix2 (blue lines in Figure 5.2), proton data from the REPT instrument has been included to help constrain the spectrum at higher energies (amber lines in Figure 5.2). The inclusion of this data was approximate; time-varying equatorial pitch angle distributions were derived by digitising the data shown in Figure 7 of Selesnick and Albert (2019). This figure shows unidirectional proton flux at $E \geq$19MeV for five epochs covering the modelling period derived from REPT measurements. Data is shown for equatorial pitch angles of 90° and 60°, and these two data points were used at $L = 2$ to derive the pitch angle distribution by assuming a distribution of the form $j = A \sin^n(\alpha_{eq})$ and solving for the two unknowns A and n. Extending the spectrum to higher energies was important because coulomb collisional loss leads to a convection of phase space density to lower values of μ, meaning that uncertainty at high energies affects lower energies too. However, after comparing different spectrum fits at high energy to understand this sensitivity, it was found that changes in the high energy ($\gtrsim$30MeV) spectrum did not introduce significant changes at the $\lesssim$10MeV energy range that is the focus of this study.

The colour bar in Figure 5.2 shows how flux varies through time at the outer boundary of the modelling region. An interesting feature of the 90° spectrum is that over the intermediate energies (from ∼2 until 20MeV), flux starts high in 2014 (blue) and decreases towards the end of the modelling period in 2018 (red), but outside this energy range there is an increase in flux. In contrast, the data at 45° shows an increase in flux through time at all energies. These two trends indicate that, at intermediate energies where flux decreases at 90°, there is a

steady reduction in anisotropy over the four year period leading to wider equatorial pitch angle distributions. Throughout the four year modelling period, there are times where boundary data is unavailable at certain energy channels, and this is indicated by the white unshaded regions in Figure 5.1. This was dealt with by simply interpolating from pitch angle distributions surrounding the data outage period, allowing a continuous spectrum.

5.3 Numerical Modelling

5.3.1 Model Overview

The 3D numerical model described in Section 3.5.3 solves Equation 3.13 for relativistic phase space density as a function of the three adiabatic invariants μ, K and L. In this work, the geomagnetic field is modelled as a dipole with $B_0 = 2.9867 \times 10^{-5} T$ (calculated for the year 2015 using coefficients of the IGRF magnetic field model, Alken et al., 2021). The quantity $f(\mu, K, L)$ is modelled, given by $f = m_0^3 j / p^2$, where m_0 is the proton rest mass. This quantity is proportional to phase space density by a constant.

Four boundary conditions were applied to the model: i) $f(\mu, K, L_{min}) = 0$ where $L_{min} = 1.15$, ii) $f(\mu_{max}, K, L) = 0$ where $\mu_{max} = 5000$MeV/G ($\sim$170MeV at $L = 2$), iii)$f(\mu, K > K_{\max}, L) = 0$ where $K_{\max}(L)$ is K corresponding to just inside the loss cone at L, and iv) $f(\mu, K, L_{max}, t) = f_b(\mu, K, t)$ where $L_{max} = 2.0$ and $f_b(\mu, K, t)$ is the time-varying outer boundary spectrum specified from the Van Allen Probe measurements and extrapolated across the range in μ.

A close fit to the spectrum data was achieved by making f_b a combination of two polynomial fitting functions P_1 and P_2, which are re-derived for every value of K on the boundary to fit the curve of $\log j(E, \alpha_{eq})$ versus energy. This curve is calculated by taking the logarithm of the data shown in Figure 5.2. P_1 is a first order polynomial (straight line fit) derived to fit data points $\log j(E \leq 7\text{MeV})$, and P_2 is a second order polynomial derived to fit data points $\log j(E \geq 5\text{MeV})$. The gradient $\mathrm{d}P_2/\mathrm{d}E$ at the highest energy channel is used to linearly extrapolate $\log j$ to higher energies outside the data range, thereby transitioning P_2 into a straight

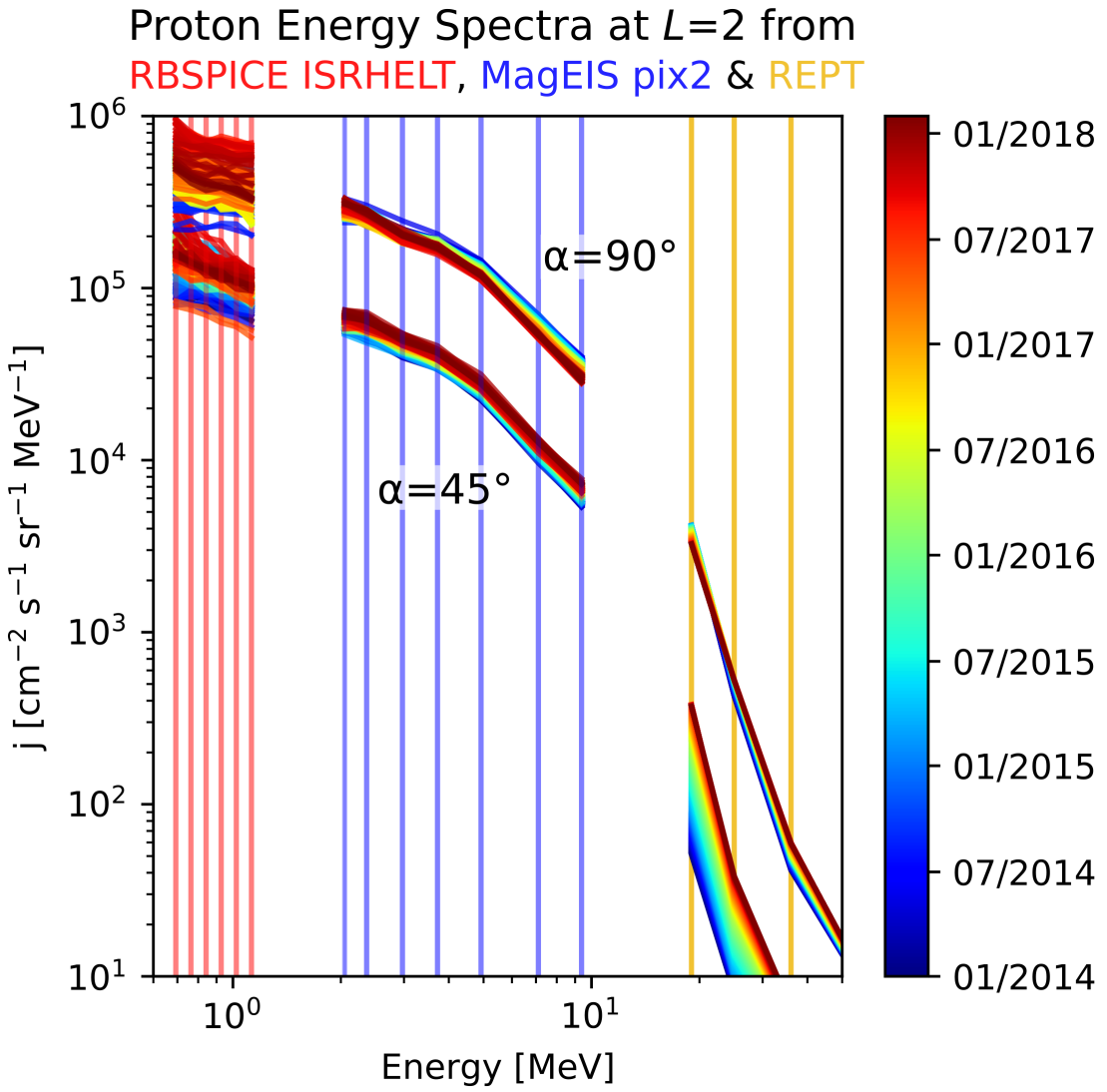


Figure 5.2: The proton energy spectrum at $L = 2$ derived from fitting a time series of pitch angle distributions to each energy channel. Energy channels are shown by vertical lines and correspond to three separate instruments: RBSPICE (0.69 - 1.13MeV, red); MagEIS (2.05 - 9.38MeV, blue) and REPT (19 - 60MeV, amber). The REPT data was derived approximately by fitting pitch angle distributions to the data in Figure 7 of Selesnick and Albert (2019).

line fit. f_b (in terms of flux) is then given according to:

$$f_b = \begin{cases} \mathrm{e}^{P_1}; E \leq 5\mathrm{MeV} \\ A\mathrm{e}^{P_1} + B\mathrm{e}^{P_2}; 5 < E < 7\mathrm{MeV} \\ \mathrm{e}^{P_2}; E \geq 7\mathrm{MeV} \end{cases} \tag{5.3}$$

where A and B vary linearly with energy from $A = 1$, $B = 0$ at 5MeV to $A = 0$, $B = 1$ at 7MeV so that f_b is linearly interpolated in this range from the two fitting functions.

5.3.2 Variation in Loss Rates

To understand loss timescales over the modelling period, the global drift averaged density model described in Section 3.4 was used to calculate the characteristic timescales of coulomb collisional loss and inelastic nuclear scattering at two different epochs over the modelling period. The results are plotted in Figure 5.3 for proton energies 1, 10 and 35MeV (first, second and third rows respectively). The timescale for coulomb collisional loss was approximated as $\tau_{cc} = (\mathrm{d}\mu/\mathrm{d}t_{fric})/\mu$, and the timescale for inelastic nuclear scattering was given by Λ. The left hand column of Figure 5.3 shows loss timescales for particles with $\alpha_{eq} = 90°$ calculated for 18 June 2017. This corresponds to an epoch where F10.7a=75.7sfu and DOY=170, and so each value of $\mathrm{d}\mu/\mathrm{d}t_{fric}$ and Λ has been interpolated at these conditions from pre-calculated values at each of the 20 F10.7a, DOY drift average coordinates. The timescale τ_{cc} has been split into three components, representing the individual contributions from coulomb collisions with atmospheric neutral constituents (blue), ionospheric constituents excluding electrons (amber), and electrons throughout the ionosphere and plasmasphere (red).

For comparison, the central column of Figure 5.3 shows loss timescales for equatorially mirroring particles calculated for 27 November 2014, when F10.7a peaked during the simulation near solar maximum with a value of 162.4sfu. The right hand column of Figure 5.3 also shows loss timescales on 27 November 2014, but for particles with $\alpha_{eq} = 50°$, mirroring at latitudes away from the equator. The grey shaded region in the right hand column indicates coordinates in the loss cone.

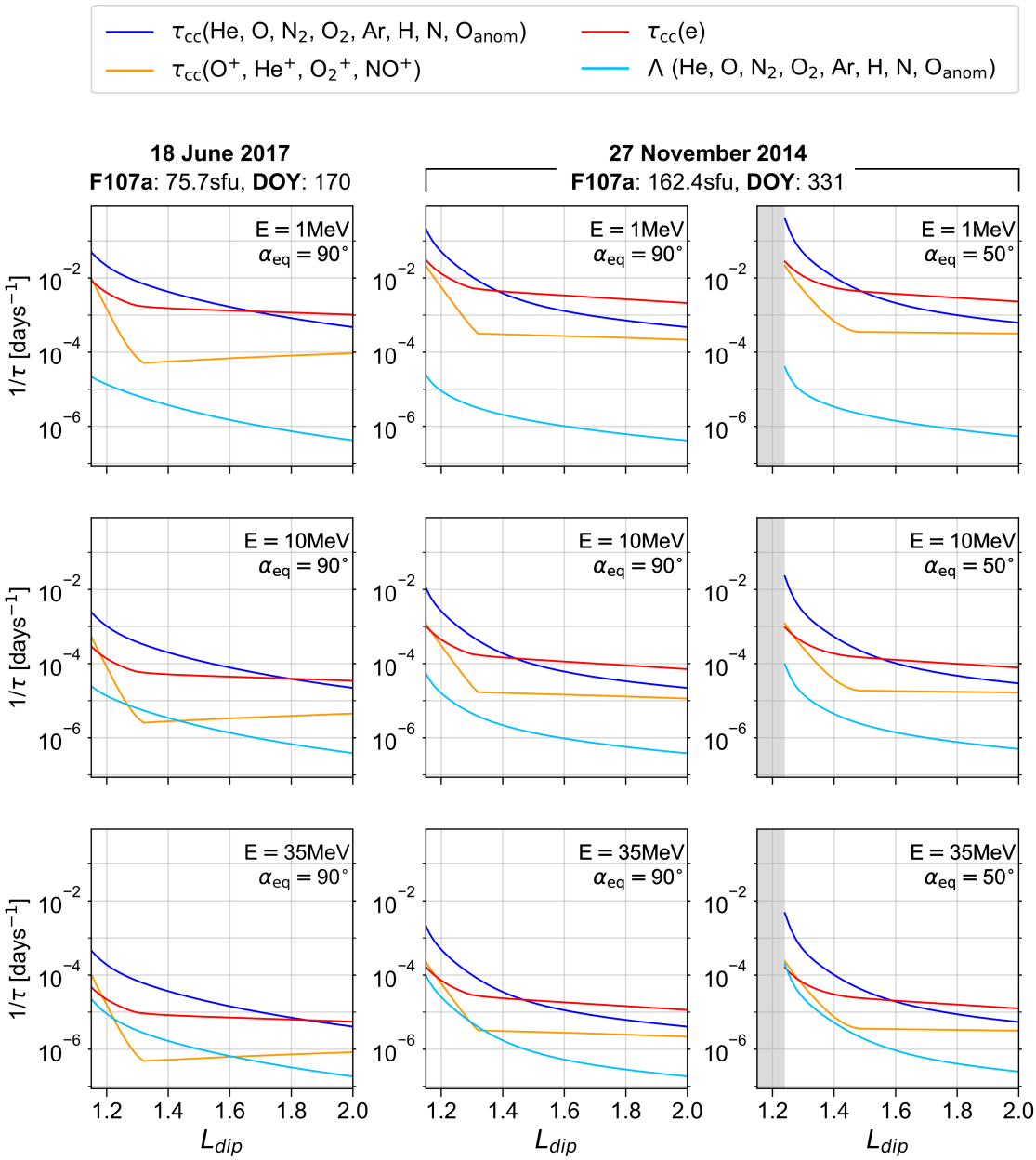


Figure 5.3: Timescales for coulomb collisional loss τ_{cc} and inelastic nuclear scattering Λ are shown as a function of energy (rows) and for two different epochs (left column versus centre and right column). Left and centre columns plot loss timescales for equatorially mirroring protons, whilst the right column shows loss timescales for particles with equatorial pitch angle $\alpha_{eq} = 50°$. The timescale for coulomb collisional loss has been separated into three components (blue, amber and red), corresponding to the contribution from each group of constituents indicated by the legend (top of figure).

Figure 5.3 shows that coulomb collisions are the dominant loss process for all energies under investigation, and that loss due to inelastic nuclear collisions (light blue curve) has a relatively small impact. In particular, coulomb collisions with free electrons in the plasmasphere (red curve) are dominant at $L \sim 2$, whilst coulomb collisions with neutral constituents (blue curve) become dominant at some lower L that depends on solar cycle and season. Comparing the left and central columns for a given energy also shows that during the transition from solar maximum (central column) towards solar minimum (left column), the loss timescale for coulomb collisions with the neutral atmosphere falls to around half its prior value at $L \lesssim 1.2$, but is relatively unaffected at $L = 1.4$. This is due to cooling and shrinking of the atmosphere.

Figure 5.3 illustrates two more important features of coulomb collisional loss timescales. The first is seen by comparing τ_{cc} in the central column and right hand column: protons mirroring at higher latitudes experience a higher rate of loss compared with equatorially mirroring protons of the same energy and L. This is a density-driven effect, because particles mirroring at higher latitude pass through denser regions of the atmosphere at lower altitude. The second feature is seen by comparing τ_{cc} in the top, middle and bottom rows of a given column: lower energy protons are subject to higher loss. This is not a density driven phenomenon, it occurs due to the energy dependence of $\langle \mathrm{d}\mu/\mathrm{d}t \rangle$. Li et al. (2020) show the variation in 36MeV integral flux driven by solar cycle density variation, and note how this effect becomes very weak at $L > 1.2$. However a comparison between the top and bottom rows in Figure 5.3 shows that 1MeV particles are subject to significantly higher loss rates than at 35MeV. This implies that lower energy particles are also more sensitive to solar cycle variations in density, and therefore solar cycle variations in density may drive variability at $L > 1.2$ in flux at some energies below 36MeV.

5.3.3 Diffusion Coefficients

In this study, D_{LL} in Equation 3.13 is given according to empirically-derived expressions used in previous literature. There are considerable differences between both the magnitude of D_{LL} and its dependencies between works, reflecting different

applications. Three sets of D_{LL} were chosen to demonstrate this range. Each is presented in this section, then used to produce simulation results for comparison.

The first equation for D_{LL} is from Equation 5 of Selesnick and Albert (2019). In the original work it is applied to model high energies ($\geq$ 19MeV) after 1 January 2015. It is modified slightly here, with the $y = 1 - K/(0.58\mathrm{G}^{1/2}\mathrm{R}_E)$ term in the original work being replaced by $\sin(\alpha_{eq})$:

$$D_{LL}(\alpha_{eq}, L) = 1.4 \times 10^{-13} L^{10} \sin^{1.6}(\alpha_{eq}) \mathrm{s}^{-1} \tag{5.4}$$

The second D_{LL} represents the magnetic diffusion coefficient required to produce a good comparison with experimental data at E >10MeV by Jentsch (1981). It was also used by Selesnick et al. (2007):

$$D_{LL}(E, \alpha_{eq}, L) = 3.75 \times 10^{-12} L^{9} \left(\frac{1 \text{ MeV}}{E} \right) \sin^{2.7}(\alpha_{eq}) \mathrm{s}^{-1} \tag{5.5}$$

The third D_{LL} is the same as Equation 12 of Selesnick et al. (2016) except for the dipole dependence term, which is set equal to 1 here because Earth's dipole was considered fixed over the modelling period. This D_{LL} was originally used to produce a fit with REPT data at $E \geq$ 24MeV.

$$D_{LL}(E, L) = 6 \times 10^{-11} L^{9} \left(\frac{1 \text{ MeV}}{E} \right)^{3/2} \mathrm{s}^{-1} \tag{5.6}$$

The three sets of D_{LL} are each used to calculate $f(\mu, K, L)$ whilst keeping all other model parameters constant. The three corresponding model runs are hereon referred to as "SA19", "J81" and "S16" respectively, based on abbreviations of the works from which each D_{LL} was taken. In these original works, each D_{LL} value was found to produce agreement between measured and computed values. However, compared to the original works, these values are being applied to model lower energy protons. This has a varying effect on the value of each D_{LL}, as the D_{LL} of Selesnick and Albert (2019) does not include energy dependence, whereas the other D_{LL} do, with the energy dependence of D_{LL} from Selesnick et al. (2016) being strongest.

5.4 Modelling Variability

5.4.1 Method

The 3D numerical model was used to solve for $f(\mu, K, L)$ over the period from 1 January 2014 to 1 March 2018, using the outer boundary spectrum shown in Figure 5.2. Simulations of the model period were performed for the three sets of D_{LL} in Equations 5.4, 5.5 and 5.6, with results denoted "SA19", "J81" and "S16' respectively. Prior to this, the initial condition for each model run was formed by computing the steady state solution at the model start time (1 January 2014). Each steady state solution was also calculated using the D_{LL} values corresponding to the dynamic simulation.

In addition to D_{LL}, the initial state of the proton belt was an uncertain aspect of the simulation. The proton belt is unlikely to be in steady state at any time due to the long timescales required by radial diffusion to rebalance changes in boundary flux that have been observed to occur at $L \geq 2$ on much shorter timescales (see for example, Figure 1, Selesnick et al., 2016). In Chapter 2, solutions of equatorial steady state phase space density during the CRRES satellite era were found to deviation from steady state strongly at $\mu \gtrsim 400$MeV/G. More rigorous methods of initialising the proton belt, such as in Selesnick et al. (2007), require integrating changes in phase space density over a time history of SEP injections, geomagnetic secular variations, and other causes of long term variation. Regardless, such methods still depend on knowledge of the diffusion coefficients, which are poorly constrained at the energies investigated here. Therefore, being somewhat limited by the current capabilities of the numerical model, the method of initialising the proton belt in steady state in a sense goes further to demonstrate the consequences of uncertainty in D_{LL}, which may lead to uncertain initialisations either way.

Although the model is numerically implicit, numerical instabilities can be caused by large gradients in $\mathrm{d}\mu/\mathrm{d}t_{fric}$ across the model grid. As discussed in Section 3.7, one way to avoid these instabilities is to increase the effective loss cone altitude. This works because large gradients tend to occur near the loss cone where gradients in atmospheric density are the highest. Therefore, a steady state solution was initially computed using a high altitude loss cone with a large timestep. A new

simulation was then initialised using the previous solution but decreasing the loss cone altitude, requiring a reduced timestep for stability but less time to reach steady state. This process was repeated, with the final resolution corresponding to a dipole loss cone altitude of 585km, where the boundary condition specifies that $f = 0$ at the equivalent value of K. This is somewhat higher than the ∼300km loss cone altitude predicted in Table 1 of Fischer et al. (1977), but was found to provide a good balance between the required run time and solution detail. Results of the three runs SA19, J81 and S16 are presented in the next section using this loss cone altitude.

5.4.2 Results

Figure 5.4 shows the radial profile of phase space density f over the modelling period at fixed values of $\mu = 20$MeV/G and $\mu = 50$MeV/G (left and right columns respectively), repeated for each of the three model runs (rows one to three show SA19, J81 and S16 respectively). The solution is shown for two values of the second invariant K, with $K = 0$ representing equatorial particles. Energy $E(\mu, K, L)$ corresponding to the values of μ on the horitzontal axis for $K = 0$ is labelled at the top of each plot in grey.

The value of D_{LL} increases in the model runs from SA19 to J81 to S16. Figure 5.4 shows that as a result, the phase space density f increases by up to three orders of magnitude at $L \sim 1.4$ for $\mu = 20$MeV/G. At $L \sim 1.2$ for $\mu = 20$MeV/G, f is very similar between the SA19 and J81 runs, and around one and a half orders of magnitude higher for S16. At $\mu = 50$MeV/G, the increase in f at $L \sim 1.4$ from SA19 to J81 to S16 is still significant (around two orders of magnitude) but not as large as at $\mu = 20$MeV/G. This indicates an increasing divergence in simulation results at lower energy, as each set of D_{LL} is extrapolated further away from the energy range it was originally derived to model. These large variations between model runs are primarily caused by differences in the steady state initial condition of each simulation. However, the differences in D_{LL} between runs also affects the region of time variability, with phase space density at $L \sim 1.5$ staying relatively constant over the four years during the SA19 run, but increasing by a factor of ∼2 at 20MeV/G in the S16 run.

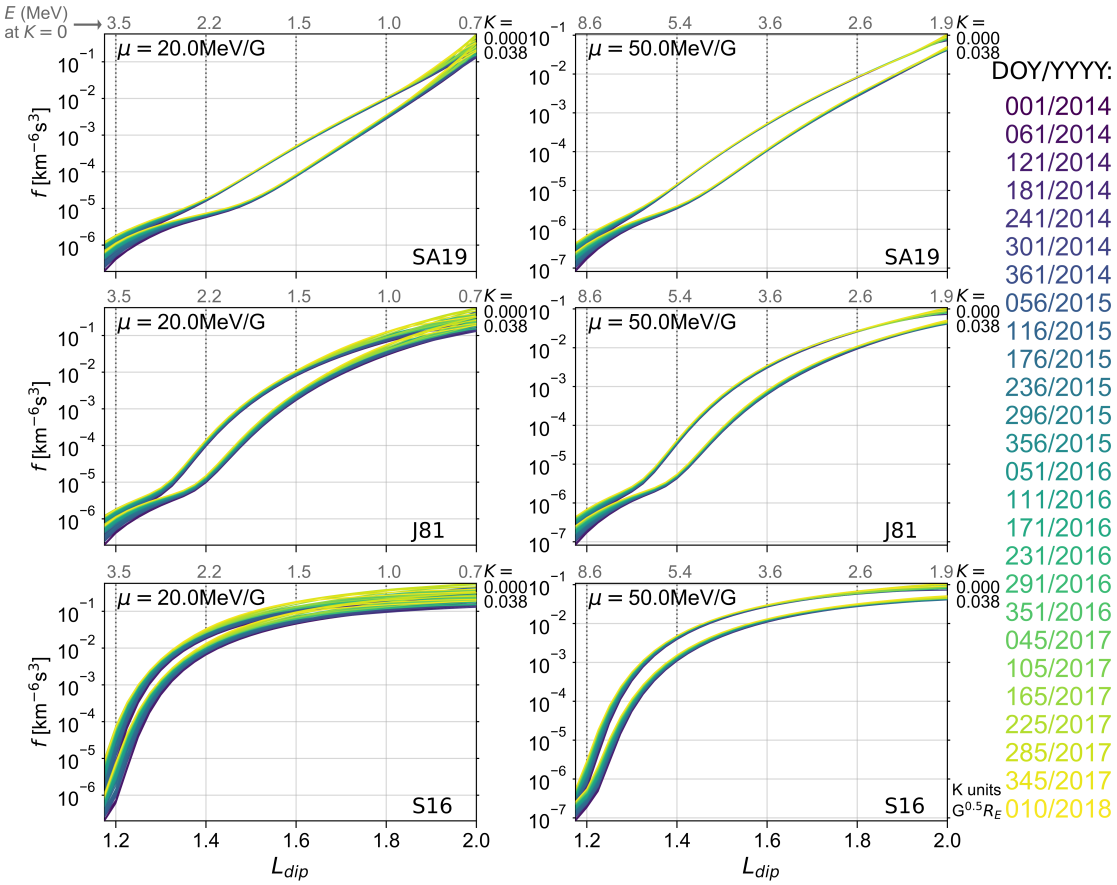


Figure 5.4: Solutions for relativistic phase space density $f = m_0^3 j/p^2$ at 30 day intervals over the modelling period, with colours corresponding to the dates shown on the right hand side. The column on the left shows f at a fixed value of μ = 20MeV/G, with μ = 50MeV/G solutions shown on the right. Each row corresponds to a different model run, characterised by the different sets of D_{LL} used (labelled within each panel).

Figures 5.5 and 5.6 show the evolution of pitch angle distributions during the J81 and S16 model runs respectively at selected L (increasing in columns left to right), and at selected fixed energies (increasing in rows top to bottom). Figures 5.5 and 5.6 are plotted in terms of unidirectional flux at each energy rather than phase space density. The evolution of pitch angle distributions during the SA19 model run is not shown because the results are somewhat similar to the J81 results, with the main difference demonstrated by Figure 5.4: at $L \sim 1.5$, flux is over an order of magnitude lower than the J81 result, and shows less time variation.

Figures 5.5 and 5.6 highlight time variability during both the J81 and S16 model runs as opposed to changes between model runs caused by different initial conditions. One striking feature of both figures is the time evolution of distributions at $L = 1.2$. For example, at 2.5MeV and $L = 1.2$ (top left panels), 90° flux approximately doubles over the modelling period during the J81 model run (Figure 5.5), and increases by nearly tenfold during the S16 model run (Figure 5.6).

Loss cone flux is fixed at zero, so the increases in 90° flux at $L = 1.2$ shown in Figures 5.5 and 5.6 give the impression of sharpening distributions through time. However, flux also increases at lower pitch angles such that the ratio of 90° flux to ∼70° flux does not change much, and so anisotropy (as defined by the n parameter for a fit like $j \propto \sin^n \alpha_{eq}$) is relatively stable at a fixed L and energy during both model runs. To demonstrate this, Figure 5.7 plots the anisotropy n versus L at the start and end time of each model run (solid blue and red curves respectively). The anisotropy is shown for energies 1, 10 and 45MeV (left, centre and right columns respectively).

Figure 5.7 shows that the largest change in anisotropy over time was during the S16 model run at $L = 1.2$ and ∼1MeV (solid curves in the bottom left panel): n decreases from ~ 55 to ~ 40. However, comparison of n between the SA19, J81 and S16 runs suggests that changes in anisotropy with L and energy are sensitive to the choice of diffusion coefficients. A key difference between each model run is the pitch angle and energy-dependence of D_{LL}. For example, in the J81 run $D_{LL}(E, \alpha_{eq}, L)$ decreases toward lower pitch angle, but in the S16 run $D_{LL}(E, L)$ is independent of pitch angle.

In order to better understand the dependence of anisotropy on D_{LL}, four extra model runs were performed. Two of these were variations of the SA19 and J81

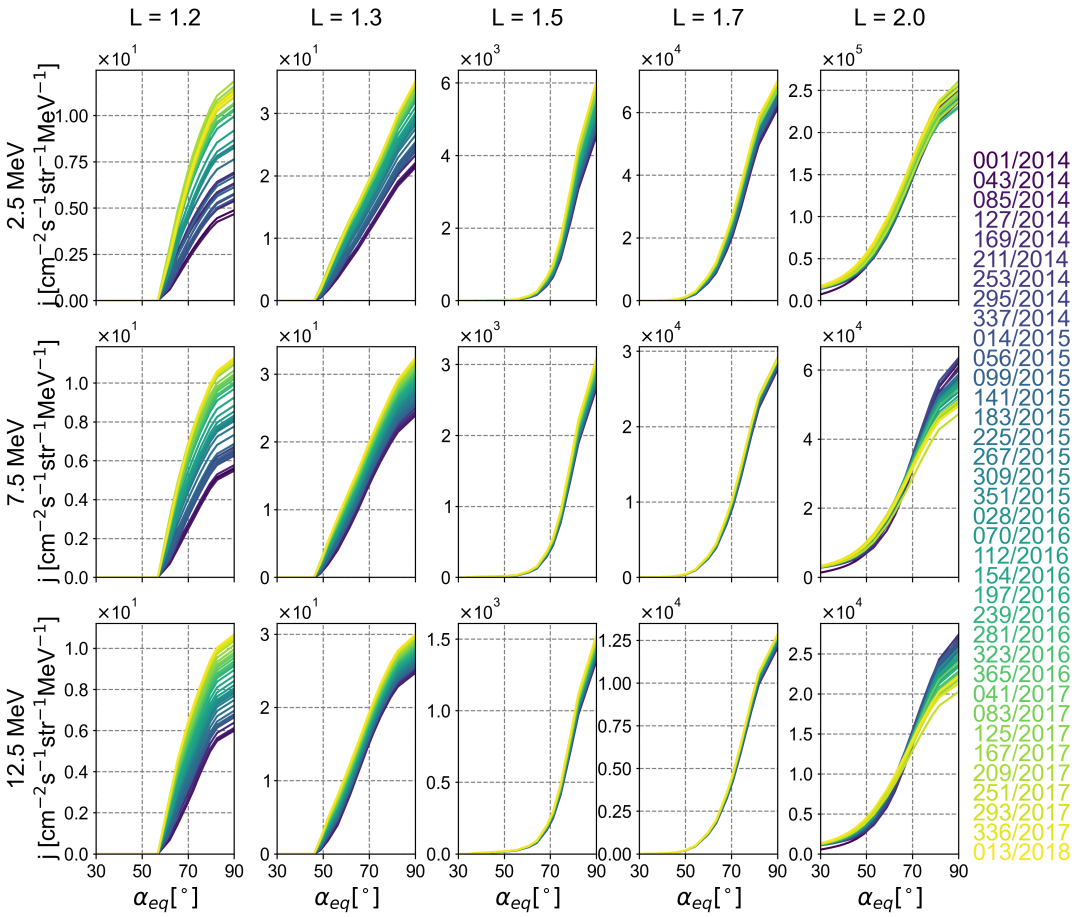


Figure 5.5: Solutions for equatorial pitch angle distributions of flux at fixed energies (rows) and L (columns) at 30 day intervals over the modelling period, with colours corresponding to the dates shown on the right hand side. These solutions were calculated for the J81 set of D_{LL}.

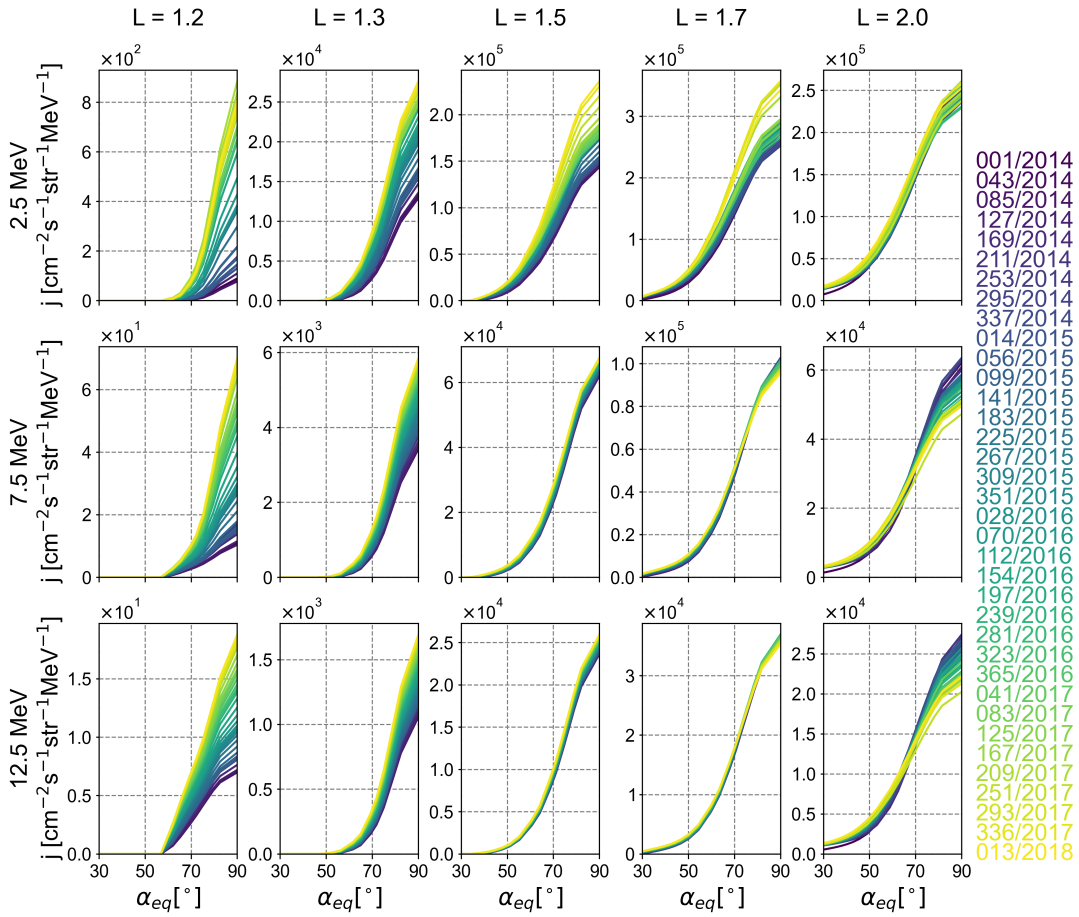


Figure 5.6: Solutions for equatorial pitch angle distributions of flux at fixed energies (rows) and L (columns) at 30 day intervals over the modelling period, with colours corresponding to the dates shown on the right hand side. These solutions were calculated for the S16 set of D_{LL}.

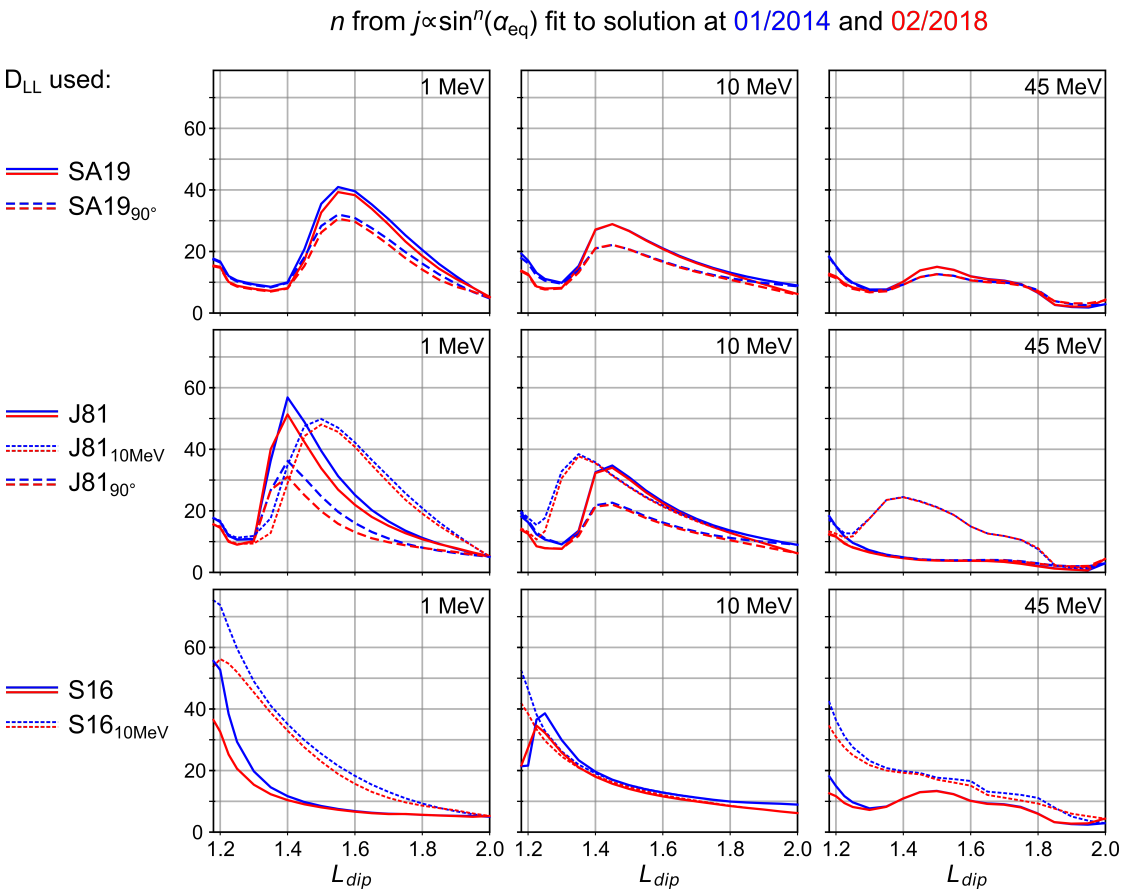


Figure 5.7: Pitch angle distribution anisotropy, quantified by the parameter n from the fit $j \propto \sin^n(\alpha_{eq})$, plotted against L for each of the SA19, J81 and S16 solutions (top, middle and bottom rows respectively, solid lines). Anisotropy n is plotted for 1, 10 and 45MeV (left, middle and right columns respectively). Variations of the SA19 and J81 solutions have also been computed by modifying the corresponding D_{LL} to eliminate dependence on pitch angle (n shown by dashed lines). Additional variations of the J81 and S16 solutions have been computed by modifying the corresponding original D_{LL} to eliminate dependence on energy (n shown by dotted lines).

model runs, in which pitch angle dependence of the original diffusion coefficients was eliminated. This was done by setting $\alpha_{eq} = 90°$ in Equations 5.4 and 5.5 for variations in each D_{LL} respectively. The other two runs were variations of the J81 and S16 model runs, in which energy dependence of the original diffusion coefficients was eliminated. This was done by fixing $E = 10\text{MeV}$ in Equations 5.5 and 5.6, to derive each new D_{LL} respectively. The anisotropy n of each of these four extra model runs is plotted in Figure 5.7 alongside the original results. The extra runs are labelled $\text{SA19}_{90°}$ and $\text{J81}_{90°}$ (using D_{LL} independent of pitch angle as described), and $\text{J81}_{10\text{MeV}}$ and $\text{S16}_{10\text{MeV}}$ (using D_{LL} independent of energy).

5.5 Discussion

Phase space density and flux levels plotted in Figures 5.4 to 5.6 are primarily controlled by a balance between inward transport via radial diffusion, and coulomb collisional losses to the atmosphere/ionosphere/plasmasphere. It was found that variations in the CRAND source exert negligible influence over the distribution at ∼MeV energy. The increase in D_{LL} from model runs SA19 to J81 to S16 hence increases the flux at lower L. Over the course of each model run, time variability arises because the balance between coulomb collisional loss and inward radial diffusion shifts. This is mostly caused by a decrease in atmospheric density, driven by a transition from solar maximum towards solar minimum, leading to increased timescales of coulomb collisional loss shown by Figure 5.3. Radial diffusion therefore increases phase space density by supplying protons from higher L. Changes in outer boundary flux also drive time variability, but this effect is small in the SA19 and J81 model runs except near $L = 2$.

In the S16 run, diffusion exerts more influence over time variability because the low altitude belt can be supplied with protons from higher altitude more quickly, and because changes in outer boundary flux are able to diffuse more quickly inward. This leads to the highest amount of time variability out of all the model runs, shown in Figure 5.6. In contrast, variability is minimised in the SA19 run, where D_{LL} is lowest. The SA19 value of D_{LL} has no energy dependence but was applied at much lower energies than it was derived for, perhaps leading to an underestimation.

Another factor controlling the balance between coulomb collisional loss and

radial diffusion is the energy range. Figure 5.4 shows the extent of variability is less at 50MeV/G (right panels) compared to 20MeV/G (left panels), and Figures 5.5 and 5.6 show the same trend, with higher variability at lower energy. This is because changes in coulomb collisional loss are more effective at lower energy where this process exerts greater influence, shown in Figure 5.3 by the difference in loss timescales between a 1 and 35MeV proton. Because of this effect, solar cycle variability is able to drive significant changes in flux at $L = 1.3$, shown in Figures 5.5 and 5.6 for the J81 and S16 model runs. For example, the increase in 7.5MeV flux at $L = 1.3$ over four years is around 30% and 75% respectively for each model run. This result can be compared with work by Li et al. (2020) that indicates there is no obvious solar cycle variation at $L > 1.2$ in >36MeV integral flux measurements made near the magnetic equator by the POES-18 satellite (Figure 6 of Li et al., 2020). As 7.5MeV is within the key energy range responsible for solar cell degradation, this implies that solar cycle effects can also drive changes in the rate of non-ionising dose over a typical mission lifetime for a range of D_{LL}.

Figures 5.5 and 5.6 show that pitch angle distributions at $L = 1.2$ appear flat near the beginning of the modelling period, and become more peaked through time due to large increases in 90° flux. The J81 and S16 model runs both show a strong sharpening of pitch angle distributions at $L \lesssim 1.3$, but the increase in flux is significantly higher in the S16 case (Figure 5.6) due to the higher D_{LL}. Despite this, Figure 5.7 shows that anisotropy of each distribution is relatively stable throughout the duration of each model run when quantified using the fitting factor n for a fit where $j \propto \sin^n(\alpha_{eq})$. At $L = 2$, Figure 5.7 shows that the outer boundary evolves to become less anisotropic over the modelling period (blue to red) at $\sim$ 10MeV (centre column), but this only seems to drive time variations in n at $L \gtrsim 1.7$.

Figure 5.7 shows significant variations in the anisotropy of pitch angle distributions across L. For example, one feature of the J81 model run is a clear increase in the anisotropy of distributions from $L = 1.3$ to 1.5 at 1-10MeV (centre panel of Figure 5.7, also shown in columns two and three of Figure 5.5). Stable n during the modelling period leads to this feature persisting over four years of time variation. However, this is somewhat at odds with a general trend suggested by previous work. For example, Figure 7a and 7b of Fischer et al. (1977) show only a decrease in n towards higher L at $L \leq 1.35$ using data at tens of MeV from the Dial satellite

(collected March to May 1970, which is similar to the modelling period considered in this work in terms of solar cycle). A decrease in n towards higher L is also shown in Figure 8 of Gussenhoven et al. (1993) at 36.3MeV using data from the CRRES satellite. Figure 5.7 (right column) shows that the numerical model results somewhat agree with the observed trend for n to decrease with L at much higher energy (45MeV).

One reason why these results show a trend that disagrees with previous observations at 1 to 10MeV may be the steady state initialisation of the proton belt. Steady state was calculated near solar maximum with high coulomb collisional loss which led to distributions being flattened at 90°, perhaps reducing anisotropy over many years. In reality, this particular balance between diffusion and loss may be too short lived to cause such changes. However, there is little data to compare with in the energy range of interest, and it can therefore be suggested that trends highlighted in previous literature may not be as general as expected, and may not apply at lower energy.

The extra runs presented in Figure 5.7 somewhat indicate the effect of pitch angle and energy dependence on anisotropy. Removing the pitch angle dependence effectively increased D_{LL} for particles at low equatorial pitch angles, and the effect is shown by comparing n between the SA19 and SA19_{90° results, as well as comparing the J81 and J81_{90° results. Anisotropy decreased at $L \sim 1.5$ when the pitch angle dependence of D_{LL} was removed, especially at lower energies, but did not change at $L \sim 1.2$. In contrast, the chosen method of removing the energy dependence effectively led to a decrease in D_{LL} at E< 10MeV, but an increase in D_{LL} at E> 10MeV, compared with the original values. Comparing n between the J81 and $\text{J81}_{10\text{MeV}}$ results, and S16 and $\text{S16}_{10\text{MeV}}$ results, shows that removing the energy dependence did not have a consistent effect. However, the increase in D_{LL} at $>$ 10MeV generally resulted in unrealistically high values of phase space density at low L. Therefore, diffusion coefficients without energy dependence seem to be less applicable in simulations at 1-10MeV.

5.6 Conclusions

This chapter has presented physics-based calculations of proton belt phase space density and flux at $1.15 \leq L \leq 2.0$ using the 3D numerical model. Results show variability over the ∼4 year period from 2014 to 2018, with an outer boundary driven by data derived from the RBSPICE, MagEIS and REPT instruments. Particular attention was paid to variations in the proton belt at low energies relevant to spacecraft solar cell degradation.

The model was applied over a coordinate range where processes known to cause variability are well constrained compared with the timescales for radial diffusion. Therefore, simulations were run for three different sets of D_{LL} taken from previous literature and exhibiting various dependencies. The initial state of the proton belt was approximated as a steady state solution. An analysis of the simulation results lead to a number of conclusions:

1) The proton belt is formed by inward radial diffusion from a source outside $L = 2$ balanced by coulomb collisional losses to the atmosphere, ionosphere and plasmasphere.

2) The steady state solution of phase space density can vary by three orders of magnitude at $\mu = 20$MeV/G at $L \sim 1.4$. The variation is due to uncertainty in extrapolating the radial diffusion coefficient to energies of 1-10MeV. Since this is a very important energy range for assessing solar array degradation, more work is required to reduce the uncertainty in D_{LL}.

3) Due to the increased importance of collisional loss at low energies, solar cycle variability is able to drive up to a ∼75% increase in 7.5MeV flux at $L = 1.3$ over four years for the D_{LL} tested, a crucial energy for solar cell degradation.

4) At $L < 1.5$, certain model solutions indicate that the anisotropy of pitch angle distributions may increase towards higher L. This is somewhat at odds with previous work showing a tendency for anisotropy to decrease towards higher L. However, results show this trend is sensitive to D_{LL}, and in particular the dependence on pitch angle.

Spacecraft measurements of flux are useful to validate theoretical calculations, but there are very few satellites equipped with detectors to measure the proton radiation belt in the 1-10MeV energy range. However, for many spacecraft traversing

the proton belt, changes in solar cell output power are dominated by the effect of proton-induced non-ionising dose affecting a fairly narrow range of energies around $\sim$ 10MeV. Therefore, the solar cell power fluctuations experienced by low and medium Earth orbit satellites could be compared with theoretical predictions based on 1-10MeV modelling results, in order to validate the results and provide some constraints on uncertain model parameters. For this reason, the sharing of operational data would help improve physics-based proton belt models.

Chapter 6

Solar Cell Degradation at 1200km

6.1 Introduction

Changes to levels of trapped proton belt flux, such as the time variability demonstrated in Figures 5.4 to 5.6, will vary the rate of damage imparted to spacecraft solar cells passing through the region. This chapter, the final piece of work in this thesis, continues from Chapter 5 by using the physics-based model results to calculate proton-induced non-ionising dose and solar cell degradation for an example satellite in ∼1200km circular orbit over the ∼four year modelling period from 01/2014 to 02/2018. The dependence of solar cell degradation on diffusion coefficients used to model the environment is also explored.

The orbital and degradation characteristics of the OneWeb 0063 satellite (NORAD catalogue number 45448) were used to calculate solar cell degradation over the four year modelling period. OneWeb 0063 was launched on 21/03/2020, so the calculation of solar cell degradation from 01/2014 to 02/2018 corresponds to a hypothetical mission, as if the satellite had been launched on 01/01/2014 instead.

6.2 OneWeb Orbit Characteristics

The model results from Chapter 5 make available phase space density $f(t_m, \mu, K, L)$, where t_m is model epoch varying across the range $01/01/2014 \leq t_m \leq 01/02/2018$. This can be converted to directional flux via Equation 1.55, then interpolated in

terms of energy and equatorial pitch angle. Therefore, the model results also make available equatorial pitch angle distributions as the function $j(t_m, E, \alpha_{eq}, L)$.

Calculating non-ionising dose involves calculating flux at locations along a satellite's orbit. Modelled equatorial pitch angle distributions $j(t_m, E, \alpha_{eq}, L)$ can be mapped to local pitch angle distributions at a location specified in terms of magnetic coordinates L and B/B_e using Equation 5.1. Therefore, the first goal was to parameterise the position s of OneWeb 0063 in terms of magnetic coordinates by time t, such that $s(t) = [L(t), B/B_e(t)]$, where t varies from t_0 to $t_0 + T$, and T is some orbit timescale. This would enable the local pitch angle distribution to be derived along an orbit from modelling reults. Note that t is a different coordinate to t_m. The former parameterises position of the satellite, whilst the latter is used to specify the model epoch and therefore parameterises time variability of the environment.

To begin, a two line element (TLE) was obtained using the Satellite Catalogue provided by CelesTrak (`https://celestrak.com/`). Figure 6.1 presents a visualisation of the orbit in 3D space over a 24 hour period relative to the location of Earth, generated by Spenvis from the TLE. The small variation in altitude is indicated by the colour scale.

Two key points summarising the orbit are:

- OneWeb 0063 is in circular orbit at an average altitude of 1234km, at 88° inclination; and
- OneWeb 0063 returns to approximately the same location after a period of 24 hours, during which it completes 13 laps around Earth at varying longitudes.

The second point is important because the magnetic coordinates L, B/B_e differ significantly along different passes around Earth due to precession of the orbital plane in longitude, combined with asymmetries in the geomagnetic field such as the South Atlantic Anomaly. To demonstrate this, Figure 6.2 plots the satellite's magnetic coordinates $L(t)$ (blue) and $B/B_e(t)$ (orange) against magnetic latitude $\lambda(t)$ over the 13 orbital passes shown in Figure 6.1. Figure 6.2 was generated by propagating the TLE forward from an arbitrarily chosen t_0 to $t_0 + 24$ hours, and the Python interface to Irbem provided by the spacepy package (Morley et al., 2011)

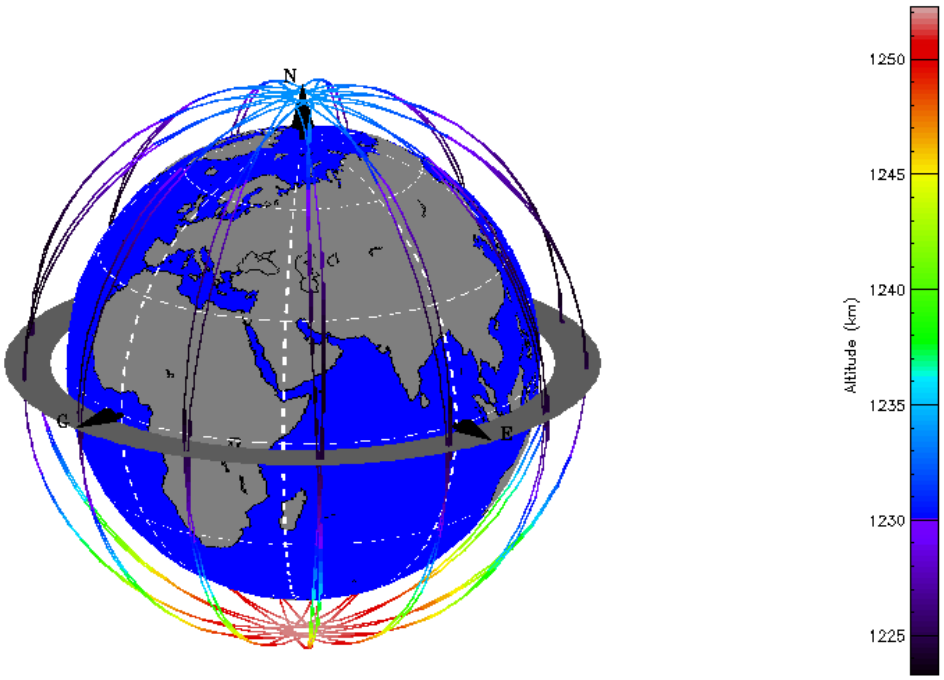


Figure 6.1: Visualisation of the OneWeb 0063 orbit over 24 hours based on TLE data, and generated by the Spenvis online interface (Heynderickx et al., 2005). Colour along the orbit track is used to indicate altitude. The axes of the GEO frame are shown as black arrows, and the geographic equator is indicated by the grey disc.

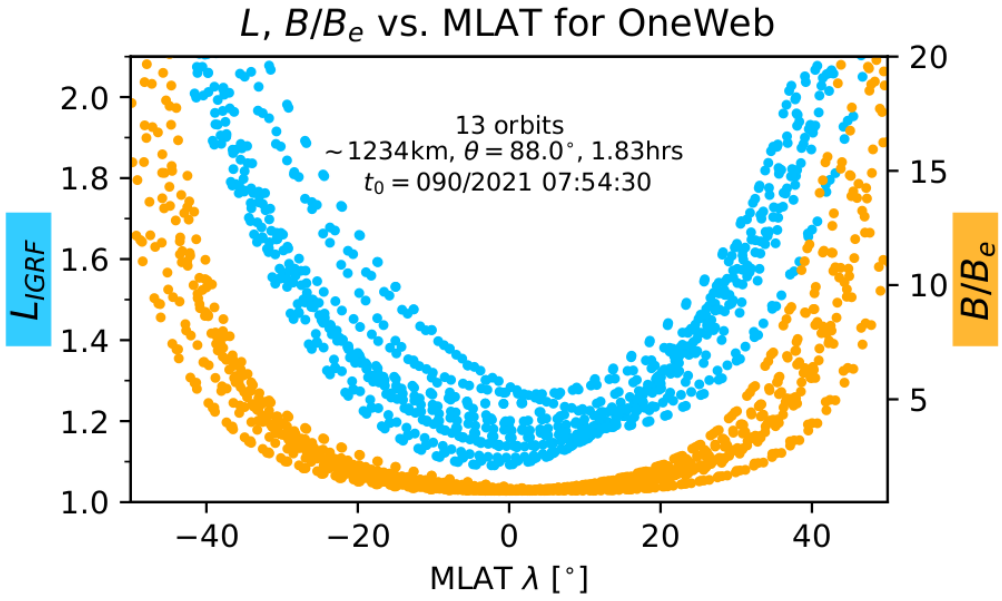


Figure 6.2: Magnetic coordinates L (light blue) and B/B_e (orange) calculated for the OneWeb-0063 satellite during 20 orbits around Earth as a function of magnetic latitude, with the first orbit beginning at the displayed time t_0.

was then used to sample L, B/B_e and λ at regular intervals in time (Δt ~44.3s) according to the IGRF internal magnetic field. At high latitudes where L is not defined a fill value was collected.

Figure 6.2 shows that, although the altitude of the satellite is fixed at $\sim 1.19 R_E$, during 13 passes over Earth the OneWeb 0063 satellite crosses the magnetic equator anywhere from $L \sim 1.1$ to $L \sim 1.3$ (blue curve). The orbital period for OneWeb 0063 is 1.83 hours according to the TLE information, corresponding to one of these passes. However, a choice of $T = 1.83$ hours would not provide a sufficiently long timescale over which to parameterise $s(t) = [L(t), B/B_e(t)]$, because the variation in $L(t), B/B_e(t)$ over a time period t_0 to $t_0 + 1.83$ hours depends on longitude. Since the satellite returns to approximately the same location every 24 hours, a choice of $T = 24$ hours is appropriate, as the variation in coordinates $L, B/B_e$ is cyclical over this timescale. A mapping was therefore constructed between orbit time t and magnetic coordinates L, B/B_e over the range $t_0 \leq t \leq t_0 + 24$ hours using the time series of points calculated with spacepy, giving $s(t) = [L(t), B/B_e(t)]$.

Figure 6.2 also shows that the L coordinate of the satellite goes above the model outer boundary at $L = 2$ when $|\lambda| \gtrsim 40°$. The component of non-ionising dose accrued at $L > 2$ cannot be calculated due to the limited model range, so the current investigation is limited to non-ionising dose accrued during passage at $L \leq 2$. The satellite spends ∼43.1% of time at $L \leq 2$, and ∼57.8% of time at $L \leq 3.25$, meaning only ∼14.7% of the orbit is spent at $2 < L \leq 3.25$. Since the proton belt does not extend much beyond $L = 3.25$, the percentage of time spent at $2 < L \leq 3.25$ gives some indication of the contribution from non-ionising dose that is ignored, but whether or not this contribution is important depends on the levels of flux reaching the satellite along field lines in this L range.

6.3 Mapping Model Flux to an Orbit

The second goal was to use the modelling results which provide equatorial pitch angle distributions of flux $j(t_m, E, \alpha_{eq}, L)$, along with the parameterisation $s(t) = [L(t), B/B_e(t)]$ derived in Section 6.2, to calculate average omnidirectional flux along an orbit $j(t_m, E)$.

The first step towards achieving this was to implement a method to determine the local pitch angle distribution of directional flux at the satellite location $j(t_m, E, s(t), \alpha)$. Since s was now available in terms of L and B/B_e for a given t, local pitch angle α could be converted to an equivalent equatorial pitch angle α_{eq} using Equation 5.1, which just depends on B/B_e. Therefore, the local pitch angle distribution of flux at the satellite for a given epoch t_m, energy E and time t along the orbit is extracted directly from the model results by re-writing

$$j(t_m, E, s(t), \alpha) = j(t_m, E, \alpha_{eq} = \sin^{-1}\sqrt{\frac{\sin^2(\alpha)}{B/B_e(t)}}, L = L(t)) \tag{6.1}$$

Substituting local pitch angle into Equation 6.1 at regular intervals across the range $0 \leq \alpha \leq 90°$ results in a local pitch angle distribution of directional flux which can be integrated to calculate omnidirectional flux. Omnidirectional flux at

the satellite position $j(t_m, E, s(t))$ is thus given by

$$\begin{aligned} j(t_m, E, s(t)) &= \int_{\Omega} j(t_m, E, s(t), \alpha) \mathrm{d}\Omega \\ &= \int_0^{\pi} j(t_m, E, s(t), \alpha) 2\pi \sin(\alpha) \mathrm{d}\alpha \end{aligned} \tag{6.2}$$

with units $\mathrm{cm}^{-2}\mathrm{s}^{-1}\mathrm{MeV}^{-1}$, where Ω is solid angle (Walt, 1994).

Figure 6.3 shows $j(t_m, E, s(t))$ calculated at the position of OneWeb 0063 along 13 passes of the satellite around Earth, corresponding to a time period of 24 hours. Results derived using a model epoch t_m =01/01/2014 are shown in the left hand column, and results derived for t_m =01/01/2018 are shown in the right hand column. Results are also shown for each of the SA19, J81 and S16 model runs performed in Chapter 5 (first, second and third rows), which correspond to the different diffusion coefficients discussed in Section 5.3.3. Since the longitude of the satellite varies from pass to pass, each line has been colour coded to represent geographic longitude of the satellite, indicating the longitudinal dependence of flux arising due to asymmetries in the geomagnetic field. Since model results only cover the region $1.15 \leq L \leq 2$, flux outside this range is not shown.

For any of the model solutions considered (fixed rows), Figure 6.3 shows that as the modelling period advances from 2014 (left) to 2018 (right), omnidirectional flux reaching the satellite increases due to the time variability of flux at low L. Figure 6.3 shows that exposure of the satellite is highest according to the S16 model solution (bottom row), which used diffusion coefficients with a higher value to solve for the environment. Figure 6.3 also shows that flux is highest at $\sim -50°$ longitude (green) and at southern latitudes, which is when the satellite passes over the South Atlantic Anomaly. This corresponds to a region where the satellite is able to reach higher L because the geomagnetic field is weaker.

Whilst performing this calculation, it was found that calculating $j(t_m, E, s(t))$ at higher latitudes near the model outer boundary $L = 2$ is particularly prone to error when the model outer boundary spectrum is derived from equatorial pitch angle distributions with non-zero flux at the loss cone. This is illustrated in Figure 5.1 for both the MagEIS and RBSPICE instruments, which show non-zero trapped flux in the loss cone due to freedom in the fitting parameter c of Equation 5.2.

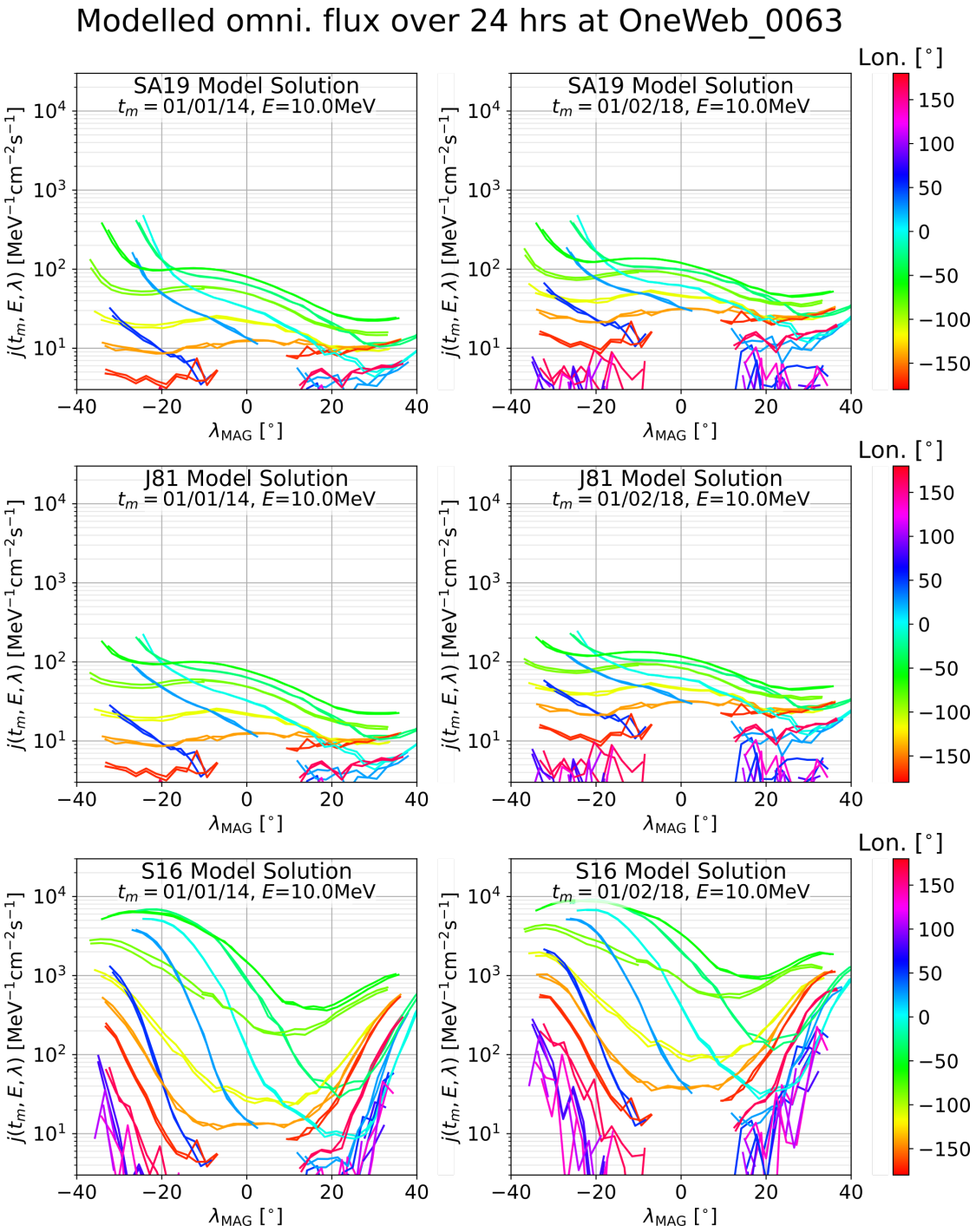


Figure 6.3: Omnidirection flux $j(t_m, E = 10\text{MeV}, s(\lambda))$ as a function of magnetic latitude λ over a time period of 24 hours, calculated for two different model epochs t_m (left and right columns) and three different sets of D_{LL} (rows). Each pass of the satellite is colour coded by geographic longitude, as indicated by the colour scale.

The model always sets loss cone flux to zero, but the flanks of these distributions just outside the loss cone are somewhat preserved, and flux near the loss cone at $L \sim 2$ is therefore likely to be overestimated. As the spacecraft crosses the $L = 2$ field line at high latitudes, the flux of particles reaching the satellite is mapped from the flanks of these equatorial pitch angle distributions, leading to a potential overestimation of flux reaching the satellite. This overestimation was partly mitigated by fitting the MagEIS and RBSPICE data more carefully, but it is one reason that flux curves upwards towards the lower and upper magnetic latitude ranges in Figure 6.3. Dealing with this issue will be very important for any future operational physics-based predictions of satellite exposure.

As the next step, omnidirectional flux was integrated over orbit time t from $t = t_0$ to $t_0 + T$, giving total omnidirectional fluence per time T. Since the orbit timescale of the satellite, $T = 24$ hours, is small compared to the timescale of variations in $f(\mu, K, L)$ over the ∼four year modelling period, t_m can be considered a constant of integration. Omnidirectional fluence per time period T at a given modelling epoch is therefore given by

$$\Phi_T(t_m, E) = \int_{t_0}^{t_0+T} j(t_m, E, s(t))\mathrm{d}t \tag{6.3}$$

where $j(t_m, E, s(t))$ is provided by Equation 6.2 when the satellite is located inside the model region. At times when the satellite is outside the modelling region, $j(t_m, E, s(t))$ was set to zero to ignore the contribution. Since t_m is constant in Equation 6.3, the integrand varies only due to the motion of the satellite through the environment.

Finally, Equation 6.3 was evaluated for different energies to give $\Phi_T(t_m, E)$ from E =1 to 19MeV at 0.5MeV increments. This set of values constitutes the spectrum of fluence incident on the satellite per time period T, across the range of energies primarily responsible for non-ionising dose (Figure 6, Messenger et al., 1997). The fluence spectrum was calcuated this way at regular intervals in time from t_m =01/01/2014 until 01/02/2018.

6.4 Non-ionising Dose Results

Non-ionising dose D_d can be calculated using the NRL method presented in Section 1.4.2.3. Equation 1.65 gives D_d as a function of the fluence spectrum $\mathrm{d}\Phi_p(E_p)/\mathrm{d}E_p$ incident on the cell, as well as proton NIEL coefficients $S_p(E)$ of the target material. In this case, the electron contribution is ignored, and only frontside exposure to the cell is considered. The latter assumption potentially excludes a small contribution towards dose from particles passing through the array substrate material.

One last step is required before applying Equation 1.65 to calculate D_d: a transport code is required to simulate the effect of coverglass shielding, which attenuates the fluence spectrum incident along the satellite orbit to give the spectrum directly incident on the solar cell. The attenuation of the spectrum depends on the coverglass shielding thickness and the density of coverglass material, both of which must be determined.

SolAero IMM-α solar cells are used on OneWeb satellites such as OneWeb 0063, but specific information on the coverglass thickness in use was not publicly available. However, the density of coverglass can be determined from the IMM-α datasheet (SolAero Technologies Corp, 2021): it specifies a coverglass interconnected cell mass of 83.3 and 96.0mg/cm^2 for versions of the cell with 4 and 6mil thick coverglass respectively, or an increase of 6.35mg/cm^2 per mil, equal to 2.5g/cm^3. The standard availability of 4 and 6mil thick coverglass (1mil = 25.4μm) also suggests that either of these values is a possibility for shielding thickness. 6mil was chosen as an estimate, based on the highly exposed orbit type.

With this information, the MULASSIS transport code (Lei et al., 2002) was used to convert each 24-hour fluence spectrum $\Phi_T(t_m, E)$ derived in Section 6.3 to the slowed-down spectrum represented by $\mathrm{d}\Phi_p(E_p)/\mathrm{d}E_p$ in Equation 1.65. Figure 1 of Haas et al. (2018) shows that SolAero IMM-α cells are comprised of junction materials $In_{0.5}Ga_{0.5}P$, GaAs, $In_{0.3}Ga_{0.7}$ and $In_{0.65}Ga_{0.35}$ from top to bottom respectively. Therefore, the NIEL coefficients for GaAs (with E_d = 21eV) were used to give appropriate values for $S_p(E)$ in Equation 1.65. Equation 1.65 was then evaluated using the slowed down fluence spectrum to give ΔD_d, the non-ionising dose accrued per time interval $\Delta t = T$. ΔD_d was calculated at ∼14 day intervals throughout the modelling period, and the result of this calculation is shown in

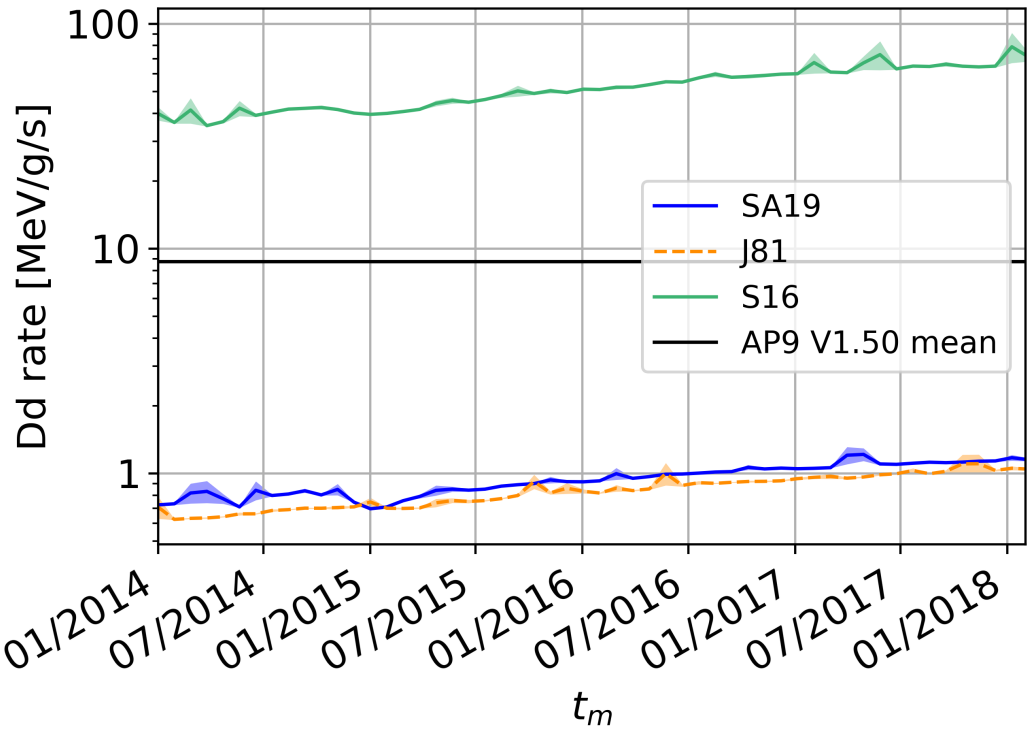


Figure 6.4: The rate at which non-ionising dose D_d is accrued by the spacecraft versus time throughout the ∼4 year model period, according to each model solution corresponding to different values of D_{LL}. Results are compared with the constant rate of dose calculated using the AP9 V1.50 mean model.

Figure 6.4 which plots the rate of non-ionising dose $\Delta D_d/T$ against model epoch t_m.

The black line in Figure 6.4 shows the constant rate of dose accrued by the satellite according to the AP9 V1.50 mean statistical proton belt model, which does not take into account time variation. By contrast, the rate of non-ionising dose calculated for each of the SA19, J81 and S16 model runs as described above increases throughout the modelling period due to the modelled time variability. Figure 6.4 shows that the J81 and SA19 model results gave very similar dose curves (orange and blue respectively). The similarity between these curves is because levels of flux at $L \lesssim 1.4$ are similar in both model runs over the range of energies responsible for non-ionising dose, as shown in Figure 5.4 in terms of phase space

density. This region is where the satellite accrues most of its non-ionising dose. On the other hand, the S16 results (green) show a much higher rate of non-ionising dose because phase space density is significantly higher in this region.

The results shown in Figure 6.4 can be used to calculate total non-ionising dose at time $t > t_0$ according to

$$D_d = \frac{1}{T} \int_{t0}^{t} \Delta D_d(t)\, \mathrm{d}t + D_{d0} \tag{6.4}$$

where D_{d0} is the initial dose at t_0.

6.5 Deriving Degradation Characteristics from Available Data

Cumulative non-ionising displacement damage dose D_d can be converted to a fraction of remaining solar cell output power P/P_0, representing degradation. For this calculation, the two unknowns C and D_{pX} in Equation 2.1 must be determined. For the solar cells studied in Chapter 2, these numbers were available via the online Spenvis interface. However, for the SolAero IMM-α cells onboard OneWeb 0063, they are not publicly available. Fortunately, online datasheets and scientific papers contain enough data to derive approximations.

Figure 4 of Haas et al. (2018), copied to the top panel of Figure 6.5 below, shows P/P_0 as a function of 1MeV electron fluence for a SolAero IMM-α cell in testing. This data is shown under two conditions: with and without post-radiation annealing; the former agrees well with the data points in the IMM-α datasheet (SolAero Technologies Corp, 2021). Figure 5 of Haas et al. (2018), copied to the bottom panel of Figure 6.5 below, also shows P/P_0 as a function of 3MeV proton fluence but only without post-radiation annealing. The data shown in these two figures can be traced over, digitised, and then used to approximate D_{pX} and C in Equation 2.1 as follows. Normalised dose from a 1MeV electron fluence can be written as $D_d = \phi_e(1\mathrm{MeV}) S_e(1\mathrm{MeV}) / R_{ep}$. Substituting this into Equation 2.1, one can perform an optimisation fit to the P/P_0 vs. $\phi_e(1\mathrm{MeV})$ post-radiation annealing electron curve in Figure 4 of Haas et al. (2018), solving for C and

$D_{pX}R_{ep}/S_e(1\text{MeV})$. By then comparing the electron and proton curves without post-radiation annealing, one can consider a 1MeV electron fluence $\phi_e(1\text{MeV})_{0.8}$ and 3MeV proton fluence $\phi_p(3\text{MeV})_{0.8}$ that cause the same drop in P/P_0 to 0.8, and therefore deliver the same dose. Equating the dose gives

$$D_d = \phi_e(1\text{MeV})_{0.8}\frac{S_e(1\text{MeV})}{R_{ep}} = \phi_p(3\text{MeV})_{0.8}S_p(3\text{MeV}) \tag{6.5}$$

If one assumes that R_{ep} and NIEL are consistent between solar cells pre- and post-radiation annealing (or specifically, that $\phi_e(1\text{MeV})_{0.8}$ and $\phi_p(3\text{MeV})_{0.8}$ still deliver equal dose even if $\Delta\text{P}/\text{P}_0$ is different), then rearranging Equation 6.5 gives $R_{ep}/S_e(1\text{MeV})$ in terms of $S_p(3\text{MeV})$. Proton NIEL at $\gtrsim$0.1MeV is consistent across the range of materials in each junction of the solar cell (including InGaP, GaAs, InGaAs; see for example Figure 3, Messenger et al., 2006), and therefore, one can use $S_p(3\text{MeV}) = 1.8421\text{e}{-2}\text{MeVcm}^2/\text{g}$, calculated for GaAs using the online calculator at `www.sr-niel.org`. Solving for $S_e(1\text{MeV})/R_{ep}$ and in turn D_{pX}, the estimates $D_{pX} \sim 4.99\text{e}9$ and $C \sim 0.303$ were obtained, allowing P/P_0 to be calculated for a given non-ionising dose using Equation 2.1.

6.6 Degradation Results

Finally, the time evolution of P/P_0 was calculated over the modelling period using the non-ionising dose results in Figure 6.4 and the degradation parameters derived in Section 6.5. Results are shown in Figure 6.6 for each of the SA19, J81 and S16 environments, assuming an initial P/P_0 of one at t_m =01/01/2014 in each case.

Figure 6.6 shows that P/P_0 falls by only ∼5% during the SA19 and J81 model runs (blue and orange), but falls by ∼80% for the S16 model run. These drops represent the solar cell degradation that the OneWeb 0063 satellite would have undergone if it were in orbit during the modelled period, depending on D_{LL}. For comparison, the black line in Figure 6.6 gives the degradation predicted by the AP9 V1.50 mean proton belt model, a drop of ∼25% after four years.

The results in Figure 6.6 represent an end-to-end physics-based calculation of solar cell degradation, incorporating the effects of proton belt time variability. The large variation in P/P_0 predicted by these modelling results is partly a result of

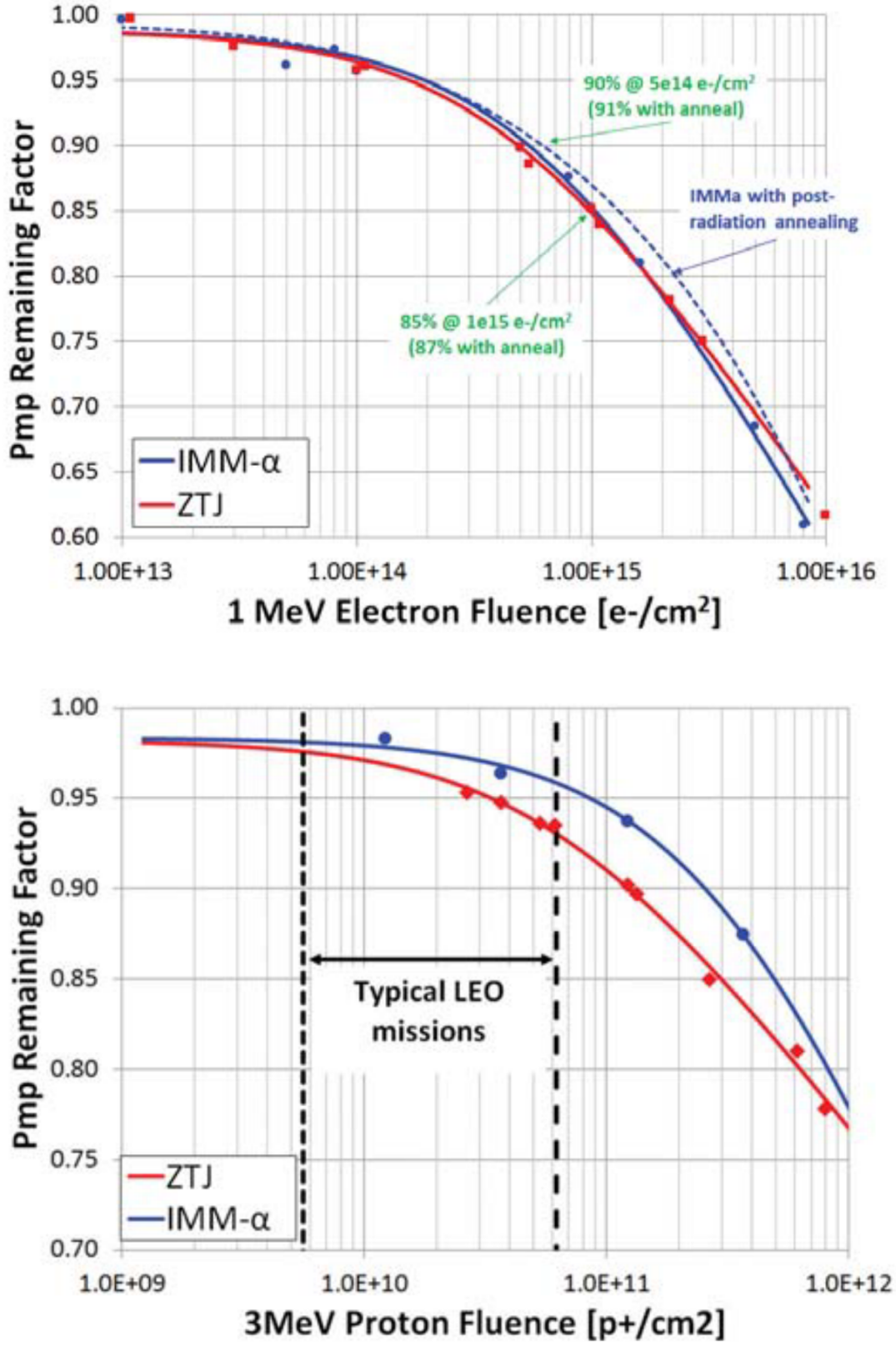


Figure 6.5: Figure 4 (top panel) and 5 (bottom panel) of Haas et al. (2018), comparing the remaining power P/P_0 between a SolAero IMM-α cell versus an older SolAero ZTJ cell both subject to the same monoenergetic fluence of 1MeV electrons (top) and 3MeV protons (bottom).

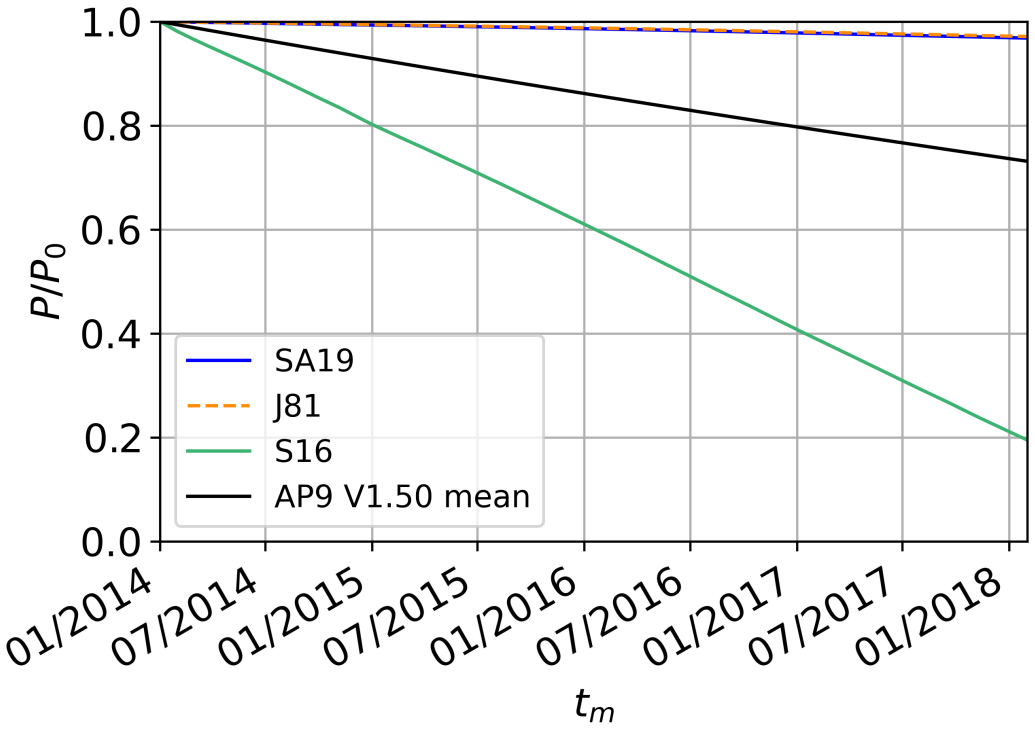


Figure 6.6: Remaining output power P/P_0 calculated for a SolAero IMM-α cell with 6mil coverglass in orbit on board OneWeb-0063, modelled for three environments corresponding to each choice of D_{LL}. Results are compared with P/P_0 calculated using the AP9 V1.50 mean model.

the uncertainty introduced by extrapolating radial diffusion coefficients to the low energies responsible for non-ionising dose. As noted in Chapter 5, the steady state solutions used to initialise each model run vary in phase space density by three orders of magnitude at $\mu = 20$MeV/G at $L \sim 1.4$. If D_{LL} were better constrained, the physics-based modelling of P/P_0 may have produced results in better agreement with AP9 V1.50, although there is uncertainty in these results too.

A conclusion reached at the end of Chapter 5 was that the sharing of operational data would help improve physics-based proton belt models. Figure 6.6 emphasises this: if boundary data were available to drive the model during the current era, the power curves shown in Figure 6.6 could be recomputed over the actual mission timeline and compared with P/P_0 data recorded by the satellite operator. An optimisation technique such as that demonstrated in Chapter 4 could then be used to infer characteristics about the energy dependence of D_{LL}, by optimising D_{LL} to produce power curves that agree with the recorded P/P_0. This technique would enable D_{LL} to be constrained at 1-10MeV without the need to validate against data from a proton telescope instrument.

Chapter 7

Summary and Conclusion

This thesis began with a review of Earth's proton radiation belt and the methods for calculating solar cell degradation. Previous observations of dynamic enhancements in trapped proton flux following solar energetic particle events indicated a key risk to orbiting satellites, suggesting a need for physics-based modelling to predict the impact of such changes. Section 1.4.3 reviewed the unexpectedly high degradation of the Tacsat-4 satellite, with a remaining P/P_0 which was >10% lower than predictions made using the AP-8 statistical model, and ∼20% lower than predicted with AP-9 Mean after two years in orbit. These results demonstrated in general that statistical radiation belt models have the potential to under-predict degradation.

In light of this, the main objective of this project was established: to address the need for physics-based modelling of the proton belt, and to quantify the impact of proton belt variability on solar cell degradation. Research began with the work in Chapter 2, which quantified the impact of proton belt variability on satellites undergoing electric orbit raising (EOR) to geostationary orbit. This chapter presented an analysis of non-ionising dose from trapped protons accrued over 200 days during the course of EOR, using three example trajectories. Conclusions from this work included that:

- for a typical coverglass thickness of $150\mu m$, launching into an enhanced proton environment can increase solar cell degradation due to trapped protons by 2 to 5% before the start of service compared to a quiet environment;
- for the same typical coverglass thickness of $150\mu m$, solar cell degradation in

an active environment can vary by ∼5% between different EOR orbits; and

- in the worst case tested, degradation of up to 15% is possible within the EOR period, before taking into account other effects such as electron dose.

These conclusions showed that dynamic enhancements have the potential to shorten the lifespan of a mission, but also that EOR orbits can be optimised to reduce the level of non-ionising dose accrued. The importance of considering enhancements in trapped proton flux ("active environments") supported the suggestion that physics-based modelling should play a role in helping to assess radiation damage, and address the increasing utilisation of low and medium Earth orbits.

Following this, work began on constructing a numerical model to address the need for physics-based modelling. Through the methods explained in Chapter 3, this involved designing and implementing a process to calculate drift averaged quantities. Using this process, up to date measurements and empirical models were used to derive a CRAND source with solar cycle dependence, and evaluations of coulomb collisional loss and nuclear inelastic scattering that depend on solar cycle and seasonal phase. A fully implicit numerical scheme was presented, along with methods to optimise execution time.

In Chapter 4, a 2D version of the model was applied to derive proton radial diffusion coefficients for a period of solar maximum. This was achieved by varying parameters controlling the rate of radial diffusion in order to optimise the fit between model and data from the PROTEL instrument aboard the CRRES satellite at $1.1 \leq L \leq 1.65$, under the assumption of steady state. Results were compared with diffusion coefficients derived in other literature, and the validity of the steady state assumption underlying this technique was discussed. Some results from this work included:

- a set of new radial diffusion coefficients derived for a period of solar maximum, which are higher than the results of previous work by a factor of 2 to 3 at μ =100MeV/G and a factor of 3 to 5 at 500MeV/G, but provide a better fit to CRRES PROTEL data;

- finding that a suitable time averaging period for steady state optimisation should be less than six months to avoid potential seasonal variations in

plasmaspheric density; and

- finding that steady state optimisation can be performed even when the proton belt is not in steady state at certain energy ranges, by making careful observations and filtering out measurements not in steady state from the data used for optimising.

However, there were some limitations of this work. Firstly, the method was not able to investigate the potential dependence of radial diffusion coefficients on magnetic or solar activity. Secondly, the optimised fit between modelled steady state phase space density and the PROTEL data was not as good as expected, despite taking measures to improve the assumption of steady state by considering only a limited selection of PROTEL data. This suggested that steady state is not a good assumption for the proton belt even during prolonged quiet periods.

Improvements were made to the model following this work, and in Chapter 5 the full 3D model was applied to simulate time variability in the inner zone at $1.15 \leq L \leq 2$. The time evolution of phase space density was modelled at energies of 1-10MeV over the four year period 2014-2018, which as demonstrated by the work of Chapter 2 covers the crucial energy range for solar cell degradation. Results were repeated for three sets of diffusion coefficients from previous literature, and the sensitivity of modelling results to the choice of diffusion coefficients was explored. Some conclusions from this work were:

- the steady state solution of proton phase space density can vary by three orders of magnitude at $\mu = 20$MeV/G at $L \sim 1.4$ due to uncertainty in radial diffusion coefficients at this μ;

- solar cycle variability was able to drive up to a $\sim$75% increase in 7.5MeV flux at $L = 1.3$ over the four year model period, associated with increasing timescales for collisional loss; and

- the anisotropy of 1-10MeV pitch angle distributions may increase towards higher L, at odds with previous work showing a tendency for anisotropy to decrease towards higher L, and this trend is particularly sensitive to the dependence of D_{LL} on equatorial pitch angle.

Accurately simulating the effect of radial diffusion remained a key challenge because the timescales of proton radial diffusion are not well constrained. All three sets of radial diffusion coefficients applied in Chapter 5 were derived to produce good fits between models and data, but only at $E > 10$MeV. Chapter 5 once again highlights the need for improved observational capability that would allow tighter constraints to be placed on diffusive timescales at ∼MeV energies.

Finally, these modelling results were applied in Chapter 6 to calculate solar cell degradation over the same four year period (2014-2018) for an example satellite in 1200km inclined circular orbit. Approximate shielding and degradation characteristics of the solar cells aboard OneWeb 0063 were derived from previous literature by applying principles of the NRL method reviewed in Chapter 1. The predicted rate of non-ionising dose differed significantly between model solutions, as expected due to the large changes in phase space density noted previously (up to three orders of magnitude). However, time variability in the rate of non-ionising dose was successfully modelled, and the methodology used to calculate dose and solar cell degradation for a given orbit and solar cell technology can be applied in the future.

The objective of this project has then been partially met, but work is yet required to model outer zone variability at $L > 2$, where timescales for variation decrease until the outer edge of the proton belt at $L \sim 3.5$. In its current state, the 3D model does not take into account several outer zone processes. Two examples are: the effect on loss timescales of magnetically-driven changes in plasmaspheric density and movement of the plasmapause; and losses due to field line curvature scattering combined with adiabatic expansion of drift orbits near the proton belt outer edge (Section 1.3.4.2). However, there is an opportunity now to continue this work: the model, now known as the British Antarctic Survey Proton Belt Model BAS-PRO, is being further developed under a contract between an academic and government organisation. This is with the aim to develop an operational space weather forecasting system.

References

Albert, J. M. and Ginet, G. P. (1998). Crres observations of radiation belt protons: 2. time-dependent radial diffusion. *Journal of Geophysical Research: Space Physics*, 103(A7):14865–14877.

Albert, J. M., Ginet, G. P., and Gussenhoven, M. S. (1998). Crres observations of radiation belt protons: 1. data overview and steady state radial diffusion. *Journal of Geophysical Research: Space Physics*, 103(A5):9261–9273.

Alken, P., Thébault, E., Beggan, C. D., Amit, H., Aubert, J., Baerenzung, J., Bondar, T., Brown, W., Califf, S., Chambodut, A., et al. (2021). International geomagnetic reference field: the thirteenth generation. *Earth, Planets and Space*, 73(1):1–25.

Anderson, B. J., Decker, R. B., Paschalidis, N. P., and Sarris, T. (1997). Onset of nonadiabatic particle motion in the near-earth magnetotail. *Journal of Geophysical Research: Space Physics*, 102(A8):17553–17569.

Baker, D. and Belian, R. (1985). Impulsive ion acceleration in earth's outer magnetosphere. Technical report.

Baker, D. N., Kanekal, S. G., Hoxie, V. C., Batiste, S., Bolton, M., Li, X., Elkington, S. R., Monk, S., Reukauf, R., Steg, S., Westfall, J., Belting, C., Bolton, B., Braun, D., Cervelli, B., Hubbell, K., Kien, M., Knappmiller, S., Wade, S., Lamprecht, B., Stevens, K., Wallace, J., Yehle, A., Spence, H. E., and Friedel, R. (2012). The relativistic electron-proton telescope (REPT) instrument on board the radiation belt storm probes (RBSP) spacecraft: Characterization of earth's radiation belt high-energy particle populations. *Space Science Reviews*, 179(1-4):337–381.

Baur, C., Campesato, R., Casale, M., Gervasi, M., Gombia, E., Greco, E., Kingma, A., Rancoita, P. G., Rozza, D., and Tacconi, M. (2017). Displacement damage dose and dlts analyses on triple and single junction solar cells irradiated with electrons and protons.

Baur, C., Gervasi, M., Nieminen, P., Pensotti, S., Rancoita, P., and Tacconi, M. (2014). Niel dose dependence for solar cells irradiated with electrons and protons. *Astroparticle, Particle, Space Physics and Detectors for Physics Applications*.

Bilitza, D., Altadill, D., Truhlik, V., Shubin, V., Galkin, I., Reinisch, B., and Huang, X. (2017). International reference ionosphere 2016: From ionospheric climate to real-time weather predictions. *Space Weather*, 15(2):418–429.

Bilitza, D. and Reinisch, B. (2008). International reference ionosphere 2007: Improvements and new parameters. *Advances in Space Research*, 42(4):599 – 609.

Blake, J. B., Carranza, P. A., Claudepierre, S. G., Clemmons, J. H., Crain, W. R., Dotan, Y., Fennell, J. F., Fuentes, F. H., Galvan, R. M., George, J. S., Henderson, M. G., Lalic, M., Lin, A. Y., Looper, M. D., Mabry, D. J., Mazur, J. E., McCarthy, B., Nguyen, C. Q., O'Brien, T. P., Perez, M. A., Redding, M. T., Roeder, J. L., Salvaggio, D. J., Sorensen, G. A., Spence, H. E., Yi, S., and Zakrzewski, M. P. (2013). The magnetic electron ion spectrometer (MagEIS) instruments aboard the radiation belt storm probes (RBSP) spacecraft. In *The Van Allen Probes Mission*, pages 383–421. Springer US.

Blake, J. B., Fennell, J. F., Turner, D. L., Cohen, I. J., and Mauk, B. H. (2019). Delayed arrival of energetic solar particles at mms on 16 july 2017. *Journal of Geophysical Research: Space Physics*, 124(4):2711–2719.

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.

Boscher, D., Bourdarie, S., Friedel, R., and Korth, A. (1998). Long term dynamic radiation belt model for low energy protons. *Geophysical research letters*, 25(22):4129–4132.

Brautigam, D. (2001). *Combined Release and Radiation Effects Satellite (CRRES) High Energy Electron Fluxmeter (HEEF) and Proton Telescope (PROTEL).* AFRL/VSBXR, 29 Randolph Road, Hanscom AFB, MA 01731.

Brautigam, D. H. and Bell, J. T. (1995). CRRESELE documentation. Technical report.

Carpenter, D. and Anderson, R. (1992). An isee/whistler model of equatorial electron density in the magnetosphere. *Journal of Geophysical Research: Space Physics*, 97(A2):1097–1108.

Chandrasekhar, S. (1943). Stochastic problems in physics and astronomy. *Rev. Mod. Phys.*, 15:1–89.

Chen, F. F. (1984). *Introduction to plasma physics and controlled fusion*, volume 1. Springer.

Chu, X., Bortnik, J., Li, W., Ma, Q., Denton, R., Yue, C., Angelopoulos, V., Thorne, R., Darrouzet, F., Ozhogin, P., et al. (2017). A neural network model of three-dimensional dynamic electron density in the inner magnetosphere. *Journal of Geophysical Research: Space Physics*, 122(9):9183–9197.

Claflin, E. S. and White, R. S. (1974). A study of equatorial inner belt protons from 2 to 200 mev. *Journal of Geophysical Research*, 79(7):959–965.

Clilverd, M. A., Meredith, N. P., Horne, R. B., Glauert, S. A., Anderson, R. R., Thomson, N. R., Menk, F. W., and Sandel, B. R. (2007). Longitudinal and seasonal variations in plasmaspheric electron density: Implications for electron precipitation. *Journal of Geophysical Research: Space Physics*, 112(A11).

Cornwall, J. M. (1972). Radial diffusion of ionized helium and protons: A probe for magnetospheric dynamics. *Journal of Geophysical Research*, 77(10):1756–1770.

Cravens, T. E. (1997). *Single particle motion and geomagnetically trapped particles*, page 43–89. Cambridge Atmospheric and Space Science Series. Cambridge University Press.

Daniell, R. E., Brown, L., Anderson, D., Fox, M., Doherty, P. H., Decker, D., Sojka, J. J., and Schunk, R. W. (1995). Parameterized ionospheric model: A global ionospheric parameterization based on first principles models. *Radio Science*, 30(5):1499–1510.

Davidson, G. T. (1976). An improved empirical description of the bounce motion of trapped particles. *Journal of Geophysical Research (1896-1977)*, 81(22):4029–4030.

Davis, L. and Chang, D. B. (1962). On the effect of geomagnetic fluctuations on trapped particles. *Journal of Geophysical Research (1896-1977)*, 67(6):2169–2179.

Decker, B. L. (1986). World geodetic system 1984. volume AD-A167 570, page 22. Fourth International Geodetic Symposium on Satellite Positioning, University of Texas, Austin.

Denton, R., Takahashi, K., Galkin, I., Nsumei, P., Huang, X., Reinisch, B., Anderson, R., Sleeper, M., and Hughes, W. (2006). Distribution of density along magnetospheric field lines. *Journal of Geophysical Research: Space Physics*, 111(A4).

Dragt, A. J., Austin, M. M., and White, R. S. (1966). Cosmic ray and solar proton albedo neutron decay injection. *Journal of Geophysical Research*, 71(5):1293–1304.

Dungey, J. (1965). Effects of electromagnetic perturbations on particles trapped in the radiation belts. *Space Science Reviews*, 4(2):199–222.

Dungey, J. W. (1961). Interplanetary magnetic field and the auroral zones. *Phys. Rev. Lett.*, 6:47–48.

Eastwood, J., Hietala, H., Toth, G., Phan, T., and Fujimoto, M. (2015). What controls the structure and dynamics of earth's magnetosphere? *Space Science Reviews*, 188(1):251–286.

Engel, M. A., Kress, B. T., Hudson, M. K., and Selesnick, R. S. (2015). Simulations of inner radiation belt proton loss during geomagnetic storms. *Journal of Geophysical Research: Space Physics*, 120(11):9323–9333.

Engel, M. A., Kress, B. T., Hudson, M. K., and Selesnick, R. S. (2016). Comparison of van allen probes radiation belt proton data with test particle simulation for the 17 march 2015 storm. *Journal of Geophysical Research: Space Physics*, 121(11):11,035–11,041.

Fälthammar, C.-G. (1965). Effects of time-dependent electric fields on geomagnetically trapped radiation. *Journal of Geophysical Research (1896-1977)*, 70(11):2503–2516.

Farley, T. A. and Walt, M. (1971). Source and loss processes of protons of the inner radiation belt. *Journal of Geophysical Research (1896-1977)*, 76(34):8223–8240.

Fei, Y., Chan, A. A., Elkington, S. R., and Wiltberger, M. J. (2006). Radial diffusion and mhd particle simulations of relativistic electron transport by ulf waves in the september 1998 storm. *Journal of Geophysical Research: Space Physics*, 111(A12).

Ferguson, D., Crabtree, P., White, S., and Vayner, B. (2016). Anomalous global positioning system power degradation from arc-induced contamination. *Journal of Spacecraft and Rockets*, 53(3):464–470.

Fischer, H. M., Auschrat, V. W., and Wibberenz, G. (1977). Angular distribution and energy spectra of protons of energy $5 \leq e \leq 50$ mev at the lower edge of the radiation belt in equatorial latitudes. *Journal of Geophysical Research (1896-1977)*, 82(4):537–547.

Fuller-Rowell, T., Solomon, S., Roble, R., and Viereck, R. (2004). Impact of solar euv, xuv, and x-ray variations on earths's atmosphere. *Geophysical Monograph Series*, 141.

Fälthammar, C.-G. (1966). On the transport of trapped particles in the outer magnetosphere. *Journal of Geophysical Research (1896-1977)*, 71(5):1487–1491.

Fälthammar, C.-G. (1968). Radial diffusion by violation of the third adiabatic invariant. *Earth's particles and fields*, page 157.

Gallagher, D. L., Craven, P. D., and Comfort, R. H. (2000). Global core plasma model. *Journal of Geophysical Research: Space Physics*, 105(A8):18819–18833.

Ginet, G. P., O'Brien, T. P., Huston, S. L., Johnston, W. R., Guild, T. B., Friedel, R., Lindstrom, C. D., Roth, C. J., Whelan, P., Quinn, R. A., Madden, D., Morley, S., and Su, Y.-J. (2013). AE9, AP9 and SPM: New models for specifying the trapped energetic particle and space plasma environment. In *The Van Allen Probes Mission*, pages 579–615. Springer US.

Ginet, G. P., O'Brien, T. P., Huston, S. L., Johnston, W. R., Guild, T. B., Friedel, R., Lindstrom, C. D., Roth, C. J., Whelan, P., Quinn, R. A., Madden, D., Morley, S., and Su, Y.-J. (2014). *AE9, AP9 and SPM: New Models for Specifying the Trapped Energetic Particle and Space Plasma Environment*, pages 579–615. Springer US, Boston, MA.

Gussenhoven, M., Mullen, E., Violet, M., Hein, C., Bass, J., and Madden, D. (1993). Crres high energy proton flux maps. *IEEE transactions on Nuclear Science*, 40(6):1450–1457.

Haas, A., McPheeters, C., Bittner, Z., Cho, B., Cruz, S., Derkacs, D., Hart, J., Kerestes, C., Miller, N., Patel, P., Riley, M., Sharps, P., Stavrides, A., Steinfeldt, J., Struempel, C., and Whipple, S. (2018). Progress in the development, qualification, and productization of imm-α. In *2018 IEEE 7th World Conference on Photovoltaic Energy Conversion (WCPEC) (A Joint Conference of 45th IEEE PVSC, 28th PVSEC 34th EU PVSEC)*, pages 3347–3351.

Haerendel, G. (1970). On the balance between radial and pitch angle diffusion. In McCormac, B. M., editor, *Particles and Fields in the Magnetosphere*, pages 416–428, Dordrecht. Springer Netherlands.

Hands, A. D. P., Ryden, K. A., Meredith, N. P., Glauert, S. A., and Horne, R. B. (2018). Radiation effects on satellites during extreme space weather events. *Space Weather*, 16(9):1216–1226.

Hein, C. (1993). The protel contamination code. Technical report.

Heynderickx, D., Quaghebeur, B., Wera, J., Daly, E., and Evans, H. (2005). Esa's space environment information system (spenvis): A web-based tool for assessing radiation doses and effects in spacecraft systems.

Horne, R. B. and Pitchford, D. (2015). Space weather concerns for all-electric propulsion satellites. *Space Weather*, 13(8):430–433.

Hudson, M., Elkington, S., Lyon, J., Marchenko, V., Roth, I., Temerin, M., Blake, J., Gussenhoven, M., and Wygant, J. (1997). Simulations of radiation belt formation during storm sudden commencements. *Journal of Geophysical Research: Space Physics*, 102(A7):14087–14102.

Hudson, M. K., Elkington, S. R., Lyon, J. G., Marchenko, V. A., Roth, I., Temerin, M., and Gussenhoven, M. S. (1996). *MHD/Particle Simulations of Radiation Belt Formation During a Storm Sudden Commencement*, pages 57–62. American Geophysical Union (AGU).

Hudson, M. K., Kotelnikov, A. D., Li, X., Roth, I., Temerin, M., Wygant, J., Blake, J. B., and Gussenhoven, M. S. (1995). Simulation of proton radiation belt formation during the march 24, 1991 ssc. *Geophysical Research Letters*, 22(3):291–294.

Jenkins, P. P., Bentz, D. C., Barnds, J., Binz, C. R., Messenger, S. R., Warner, J. H., Krasowski, M. J., Prokop, N. F., Spina, D. C., O'Neill, M., et al. (2014). Tacsat-4 solar cell experiment: Two years in orbit. In *10th European Space Power Conference, Noordwijkerhout, Netherlands*, volume 14.

Jentsch, V. (1981). On the role of external and internal source in generating energy and pitch angle distributions of inner-zone protons. *Journal of Geophysical Research: Space Physics*, 86(A2):701–710.

Johnston, W. R., O'Brien, T. P., and Ginet, G. P. (2015). Release of ae9/ap9/spm radiation belt and space plasma model version 1.20.002. *Space Weather*, 13(6):368–368.

Kakinami, Y., Lin, C. H., Liu, J. Y., Kamogawa, M., Watanabe, S., and Parrot, M. (2011). Daytime longitudinal structures of electron density and temperature in the topside ionosphere observed by the hinotori and demeter satellites. *Journal of Geophysical Research: Space Physics*, 116(A5).

Kessel, R., Fox, N., and Weiss, M. (2013). The radiation belt storm probes (rbsp) and space weather. *Space Science Reviews*, 179(1-4):531–543.

Kivelson, M. G. and Russell, C. T. (1995). *Introduction to Space Physics*. Cambridge University Press.

Kluever, C. A. and Messenger, S. R. (2019). Solar-cell degradation model for trajectory optimization methods. *Journal of Spacecraft and Rockets*, 56(3):844–853.

Kovtyukh, A. S. (2020). Earth's radiation belts' ions: patterns of the spatial-energy structure and its solar-cyclic variations. In *Annales Geophysicae*, volume 38, pages 137–147. Copernicus GmbH.

Kress, B. T., Hudson, M. K., Perry, K. L., and Slocum, P. L. (2004). Dynamic modeling of geomagnetic cutoff for the 23–24 november 2001 solar energetic particle event. *Geophysical Research Letters*, 31(4).

Kress, B. T., Hudson, M. K., and Slocum, P. L. (2005). Impulsive solar energetic ion trapping in the magnetosphere during geomagnetic storms. *Geophysical Research Letters*, 32(6).

Kutiev, I., Oyama, K., and Abe, T. (2002). Analytical representation of the plasmasphere electron temperature distribution based on akebono data. *Journal of Geophysical Research: Space Physics*, 107(A12):SMP 24–1–SMP 24–11.

Lei, F., Truscott, R., Dyer, C., Quaghebeur, B., Heynderickx, D., Nieminen, R., Evans, H., and Daly, E. (2002). Mulassis: a geant4-based multilayered shielding simulation tool. *IEEE Transactions on Nuclear Science*, 49(6):2788–2793.

Lejosne, S. (2019a). Analytic expressions for radial diffusion. *Journal of Geophysical Research: Space Physics*, 124(6):4278–4294.

Lejosne, S. (2019b). Analytic expressions for radial diffusion. *Journal of Geophysical Research: Space Physics*, 124(6):4278–4294.

Lejosne, S., Boscher, D., Maget, V., and Rolland, G. (2013a). Deriving electromagnetic radial diffusion coefficients of radiation belt equatorial particles for different

levels of magnetic activity based on magnetic field measurements at geostationary orbit. *Journal of Geophysical Research: Space Physics*, 118(6):3147–3156.

Lejosne, S., Boscher, D., Maget, V., and Rolland, G. (2013b). Deriving electromagnetic radial diffusion coefficients of radiation belt equatorial particles for different levels of magnetic activity based on magnetic field measurements at geostationary orbit. *Journal of Geophysical Research: Space Physics*, 118(6):3147–3156.

Lejosne, S. and Kollmann, P. (2020). Radiation belt radial diffusion at earth and beyond. *Space Science Reviews*, 216(1).

Lenchek, A. M. and Singer, S. F. (1962). Geomagnetically trapped protons from cosmic-ray albedo neutrons. *Journal of Geophysical Research*, 67(4):1263–1287.

Lev, D., Myers, R. M., Lemmer, K. M., Kolbeck, J., Koizumi, H., and Polzin, K. (2019). The technological and commercial expansion of electric propulsion. *Acta Astronautica*, 159:213–227.

Li, X., Roth, I., Temerin, M., Wygant, J. R., Hudson, M. K., and Blake, J. B. (1993). Simulation of the prompt energization and transport of radiation belt particles during the march 24, 1991 ssc. *Geophysical Research Letters*, 20(22):2423–2426.

Li, X., Xiang, Z., Zhang, K., Khoo, L., Zhao, H., Baker, D. N., and Temerin, M. A. (2020). New insights from long-term measurements of inner belt protons (10s of mev) by sampex, poes, van allen probes, and simulation results. *Journal of Geophysical Research: Space Physics*, 125(8):e2020JA028198.

Lopez, R. E. and Gonzalez, W. D. (2017). Magnetospheric balance of solar wind dynamic pressure. *Geophysical Research Letters*, 44(7):2991–2999.

Lorentzen, K. R., Mazur, J. E., Looper, M. D., Fennell, J. F., and Blake, J. B. (2002). Multisatellite observations of mev ion injections during storms. *Journal of Geophysical Research: Space Physics*, 107(A9):SMP 7–1–SMP 7–11.

Lozinski, A. R., Horne, R. B., Glauert, S. A., Del Zanna, G., Heynderickx, D., and Evans, H. D. (2019). Solar cell degradation due to proton belt enhancements during electric orbit raising to geo. *Space Weather*, 17(7):1059–1072.

MacNamara, L. F. (1994). *Radio amateurs guide to the Ionosphere*. Krieger.

Manweiler, J. W. and Mull, H. (2017). *RBSPICE Science Data Handbook (Revision: d)*.

Maurer, R. H., Goldsten, J. O., Butler, M. H., and Fretz, K. (2018). Five-year results from the engineering radiation monitor and solar cell monitor on the vanallen probes mission. *Space Weather*, 16(10):1561–1569.

Mazur, J., Blake, J., Slocum, P., Hudson, M., and Mason, G. (2006). The creation of new ion radiation belts associated with solar energetic particle events and interplanetary shocks.

McGuire, R. and von Rosenvinge, T. (1984). The energy spectra of solar energetic particles. *Advances in Space Research*, 4(2):117–125.

McIlwain, C. E. (1961). Coordinates for mapping the distribution of magnetically trapped particles. *Journal of Geophysical Research (1896-1977)*, 66(11):3681–3691.

Meffert, J. D. and Gussenhoven, M. (1994). Crrespro documentation. Technical report.

Messenger, S. R., Burke, E. A., Walters, R. J., Warner, J. H., Summers, G. P., and Morton, T. L. (2006). Effect of omnidirectional proton irradiation on shielded solar cells. *IEEE Transactions on Nuclear Science*, 53(6):3771–3778.

Messenger, S. R., Jackson, E. M., Warner, J. H., and Walters, R. J. (2010). Scream: A new code for solar cell degradation prediction using the displacement damage dose approach. In *2010 35th IEEE Photovoltaic Specialists Conference*, pages 001106–001111.

Messenger, S. R., Summers, G. P., Burke, E. A., Walters, R. J., and Xapsos, M. A. (2001). Modeling solar cell degradation in space: A comparison of the nrl displacement damage dose and the jpl equivalent fluence approaches†. *Progress in Photovoltaics: Research and Applications*, 9(2):103–121.

Messenger, S. R., Wong, F., Hoang, B., Cress, C. D., Walters, R. J., Kluever, C. A., and Jones, G. (2014). Low-thrust geostationary transfer orbit (lt2geo) radiation environment and associated solar array degradation modeling and ground testing. *IEEE Transactions on Nuclear Science*, 61(6):3348–3355.

Messenger, S. R., Xapsos, M. A., Burke, E. A., Walters, R. J., and Summers, G. P. (1997). Proton displacement damage and ionizing dose for shielded devices in space. *IEEE Transactions on Nuclear Science*, 44(6):2169–2173.

Mitchell, D. G., Lanzerotti, L. J., Kim, C. K., Stokes, M., Ho, G., Cooper, S., Ukhorskiy, A., Manweiler, J. W., Jaskulek, S., Haggerty, D. K., Brandt, P., Sitnov, M., Keika, K., Hayes, J. R., Brown, L. E., Gurnee, R. S., Hutcheson, J. C., Nelson, K. S., Paschalidis, N., Rossano, E., and Kerem, S. (2013). Radiation belt storm probes ion composition experiment (RBSPICE). In *The Van Allen Probes Mission*, pages 263–308. Springer US.

Miyake, W., Miyoshi, Y., and Matsuoka, A. (2014). On the spatial extent of the proton radiation belt from solar cell output variation of the akebono satellite. *Advances in Space Research*, 53(11):1603–1609.

Morley, S. K., Koller, J., Welling, D. T., Larsen, B. A., Henderson, M. G., and Niehof, J. T. (2011). Spacepy - a python-based library of tools for the space sciences. In *Proceedings of the 9th Python in science conference (SciPy 2010)*, Austin, TX.

Mullen, E., Gussenhoven, M., Ray, K., and Violet, M. (1991). A double-peaked inner radiation belt: cause and effect as seen on crres. *IEEE Transactions on Nuclear Science*, 38(6):1713–1718.

Nakada, M. P. and Mead, G. D. (1965). Diffusion of protons in the outer radiation belt. *Journal of Geophysical Research (1896-1977)*, 70(19):4777–4791.

Olson, W. P. and Pfitzer, K. A. (1974). A quantitative model of the magnetospheric magnetic field. *Journal of Geophysical Research*, 79(25):3739–3748.

Olson, W. P. and Pfitzer, K. A. (1982). A dynamic model of the magnetospheric magnetic and electric fields for july 29, 1977. *Journal of Geophysical Research: Space Physics*, 87(A8):5943–5948.

Oyama, K., Lakshmi, D., Kutiev, I., and Abdu, M. (2005). Low latitude n e and t e variations at 600 km during 1 march 1982 storm from hinotori satellite. *Earth, planets and space*, 57(9):871–878.

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.

Ozhogin, P., Tu, J., Song, P., and Reinisch, B. W. (2012). Field-aligned distribution of the plasmaspheric electron density: An empirical model derived from the image rpi measurements. *Journal of Geophysical Research: Space Physics*, 117(A6).

Park, C. G., Carpenter, D. L., and Wiggin, D. B. (1978). Electron density in the plasmasphere: Whistler data on solar cycle, annual, and diurnal variations. *Journal of Geophysical Research: Space Physics*, 83(A7):3137–3144.

Parker, E. N. (1960). Geomagnetic fluctuations and the form of the outer zone of the van allen radiation belt. *Journal of Geophysical Research (1896-1977)*, 65(10):3117–3130.

Pellegrino, C., Gagliardi, A., and Zimmermann, C. G. (2020). Defect spectroscopy and non-ionizing energy loss analysis of proton and electron irradiated p-type gaas solar cells. *Journal of Applied Physics*, 128(19):195701.

Picone, J. M., Hedin, A. E., Drob, D. P., and Aikin, A. C. (2002). Nrlmsise-00 empirical model of the atmosphere: Statistical comparisons and scientific issues. *Journal of Geophysical Research: Space Physics*, 107(A12):SIA 15–1–SIA 15–16.

Preszler, A. M., Moon, S., and White, R. S. (1976). Atmospheric neutrons. *Journal of Geophysical Research (1896-1977)*, 81(25):4715–4722.

Qioptiq Space Technology (2015). Solar cell coverglasses. `http://www.qioptiq.com/download/QST_2015_03_Datasheet_CoverGlass%20v3[1].pdf`. Accessed: 2019-01-25.

Reames, D. V. (2013). The two sources of solar energetic particles. *Space Science Reviews*, 175(1-4):53–92.

Richard, R. L., El-Alaoui, M., Ashour-Abdalla, M., and Walker, R. J. (2002). Interplanetary magnetic field control of the entry of solar energetic particles into the magnetosphere. *Journal of Geophysical Research: Space Physics*, 107(A8):SSH 7–1–SSH 7–20.

Roederer, J. G. and Lejosne, S. (2018). Coordinates for representing radiation belt particle flux. *Journal of Geophysical Research: Space Physics*, 123(2):1381–1387.

Russell, C. T. (1971). Geophysical coordinate transformations. *Cosmic Electrodynamics*, 2(2):184–196.

Ryan, J. M., Lockwood, J. A., and Debrunner, H. (2000). Solar energetic particles. *Space Science Reviews*, 93(1):35–53.

Sawyer, D. M. and Vette, J. I. (1976). AP-8 trapped proton environment for solar maximum and solar minimum. NASA STI/Recon Technical Report N.

Schulz, M. (1991). The magnetosphere. In *Geomagnetism*, pages 87–293. Elsevier.

Schulz, M. and Lanzerotti, L. (1974). Physics and chemistry in space. *Particle Diffusion in the Radiation Belts*, 7.

Selesnick, R. S. and Albert, J. M. (2019). Variability of the proton radiation belt. *Journal of Geophysical Research: Space Physics*, 124(7):5516–5527.

Selesnick, R. S., Baker, D. N., Jaynes, A. N., Li, X., Kanekal, S. G., Hudson, M. K., and Kress, B. T. (2014). Observations of the inner radiation belt: Crand and trapped solar protons. *Journal of Geophysical Research: Space Physics*, 119(8):6541–6552.

Selesnick, R. S., Baker, D. N., Jaynes, A. N., Li, X., Kanekal, S. G., Hudson, M. K., and Kress, B. T. (2016). Inward diffusion and loss of radiation belt protons. *Journal of Geophysical Research: Space Physics*, 121(3):1969–1978.

Selesnick, R. S., Hudson, M. K., and Kress, B. T. (2010). Injection and loss of inner radiation belt protons during solar proton events and magnetic storms. *Journal of Geophysical Research: Space Physics*, 115(A8).

Selesnick, R. S., Looper, M. D., and Mewaldt, R. A. (2007). A theoretical model of the inner proton radiation belt. *Space Weather*, 5(4).

Singer, S. and Lemaire, J. (2009). Geomagnetically trapped radiation: Half a century of research.

Singer, S. F. (1958). Trapped albedo theory of the radiation belt. *Physical Review Letters*, 1(5):181.

Smart, D. and Shea, M. (2005). A review of geomagnetic cutoff rigidities for earth-orbiting spacecraft. *Advances in Space Research*, 36(10):2012–2020. Solar Wind-Magnetosphere-Ionosphere Dynamics and Radiation Models.

SolAero Technologies Corp (2018). Ztj space solar cell: 3rd generation triple-junction solar cell for space applications. Accessed on 10 October 2018.

SolAero Technologies Corp (2021). Imm-α space solar cell - solaero technologies. Accessed on 1 August 2021.

Spectrolab (2010). Space solar panels. `https://www.spectrolab.com/DataSheets/Panel/panels.pdf`. Accessed: 2019-01-25.

Stassinopoulos, E. and Raymond, J. (1988). The space radiation environment for electronics. *Proceedings of the IEEE*, 76(11):1423–1442.

Summers, G. P., Messenger, S. R., Burke, E. A., Xapsos, M. A., and Walters, R. J. (1997). Low energy proton-induced displacement damage in shielded gaas solar cells in space. *Applied Physics Letters*, 71(6):832–834.

Toda, H., Miyake, W., Miyoshi, Y., Toyota, H., Miyazawa, Y., Shinohara, I., and Matsuoka, A. (2018). Spatial distribution of radiation belt protons deduced from solar cell degradation of the arase satellite. *International Journal of Astronomy and Astrophysics*, 08(04):306–322.

Valot, P. and Engelmann, J. (1973). Pitch angle distribution of geomagnetically trapped protons for 1.2< l< 2.1. *spre*, 2:675–681.

Van Allen, J. A. (1959). The geomagnetically trapped corpuscular radiation. *Journal of Geophysical Research (1896-1977)*, 64(11):1683–1689.

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.

Vernov, S., Grigorov, N., Ivanenko, I., Lebedinskii, A., Murzin, V., and Chudakov, A. (1959). Possible mechanism of production of" terrestrial corpuscular radiation" under the action of cosmics rays. *SPhD*, 4:154.

Vette, J. I. (1991). *The NASA/National Space Science Data Center: Trapped Radiation Environment Model Program (1964-1991)*, volume 91. National Space Science Data Center (NSSDC), World Data Center A for Rockets ....

Violet, M., Lynch, K., Redus, R., Riehl, K., Boughan, E., and Hein, C. (1993). Proton telescope (protel) on the crres spacecraft. *IEEE Transactions on nuclear science*, 40(2):242–245.

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

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris,

C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Walt, M. (1971). The radial diffusion of trapped particles induced by fluctuating magnetospheric fields. *Space Science Reviews*, 12(4):446–485.

Walt, M. (1994). *Introduction to Geomagnetically Trapped Radiation*. Cambridge University Press.

Wang, C.-P., Lyons, L. R., Angelopoulos, V., Larson, D., McFadden, J., Frey, S., Auster, H.-U., and Magnes, W. (2008). Themis observations of penetration of the plasma sheet into the ring current region during a magnetic storm. *Geophysical research letters*, 35(17).

White, R., Moon, S., Preszler, A., and Simnett, G. (1972). Earth albedo, and solar neutrons. Technical report, University of California, Riverside Institute of Geophysics and Planetary Physics.

Wikipedia (2021). World geodetic system.

Wygant, J., Mozer, F., Temerin, M., Blake, J., Maynard, N., Singer, H., and Smiddy, M. (1994). Large amplitude electric and magnetic field signatures in the inner magnetosphere during injection of 15 mev electron drift echoes. *Geophysical Research Letters*, 21(16):1739–1742.

Appendix A

Relating Changes in the First and Second Invariant due to Coulomb Collisions

The first invariant can be written in terms of the kinetic energy T and proton rest energy E_0 like so:

$$\begin{aligned}\mu =& \frac{y^2 L^3 \rho^2}{2 m_0 B_e} = \frac{y^2 L^3 \rho^2 c^2}{2 E_0 B_e} \\ =& \frac{y^2 L^3 T (T + 2E_0)}{2 E_0 B_e}\end{aligned} \tag{A.1}$$

Differentiating Equation A.1 with respect to time leads to:

$$\begin{aligned}\frac{\mathrm{d}\mu}{\mathrm{d}t} =& \frac{y^2 L^3}{2 E_0 B_e} \frac{\mathrm{d}}{\mathrm{d}t} \left[T (T + 2E_0) \right] \\ =& \frac{y^2 L^3}{2 E_0 B_e} \left[\frac{\mathrm{d}T}{\mathrm{d}t} (T + 2E_0) + T \frac{\mathrm{d}T}{\mathrm{d}t} \right] \\ =& \frac{y^2 L^3}{2 E_0 B_e} \frac{\mathrm{d}T}{\mathrm{d}t} (2T + 2E_0) \\ =& \frac{y^2 L^3}{E_0 B_e} \frac{\mathrm{d}T}{\mathrm{d}t} (T + E_0) \\ =& \frac{2\mu}{T (T + 2E_0)} \frac{\mathrm{d}T}{\mathrm{d}t} (T + E_0)\end{aligned} \tag{A.2}$$

The second invariant can be written in terms of the kinetic energy T and proton rest energy E_0 like so:

$$\begin{aligned} J =& 2La\rho Y(y) \\ =& 2LaY(y)\frac{\sqrt{T(T+2E_0)}}{c} \end{aligned} \tag{A.3}$$

Differentiating Equation A.3 with respect to time leads to:

$$\begin{aligned} \therefore \frac{\mathrm{d}J}{\mathrm{d}t} =& \frac{2LaY(y)}{c}\frac{\mathrm{d}}{\mathrm{d}t}\left[(T+2E_0)T\right]^{1/2} \\ =& \frac{2LaY(y)}{c}\left[T^{1/2}\frac{\mathrm{d}}{\mathrm{d}t}(T+2E_0)^{1/2} + (T+2E_0)^{1/2}\frac{\mathrm{d}}{\mathrm{d}t}T^{1/2}\right] \end{aligned} \tag{A.4}$$

Expanding the two differential terms inside the square brackets of Equation A.4 leads to:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}(T+2E_0)^{1/2} =& \frac{1}{2}(T+2E_0)^{-1/2}\frac{\mathrm{d}}{\mathrm{d}t}(T+2E_0) \\ =& \frac{1}{2(T+2E_0)^{1/2}}\frac{\mathrm{d}T}{\mathrm{d}t} \end{aligned} \tag{A.5}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}T^{1/2} = \frac{1}{2}T^{-1/2}\frac{\mathrm{d}T}{\mathrm{d}t} \tag{A.6}$$

Substituting Equations A.5 and A.6 into Equation A.4, then simplifying, leads to:

$$
\begin{aligned}
\frac{\mathrm{d}J}{\mathrm{d}t} =& \frac{2LaY(y)}{c}\left[\frac{T^{1/2}}{2(T+2E_0)^{1/2}}\frac{\mathrm{d}T}{\mathrm{d}t} + \frac{1}{2T^{1/2}}\frac{\mathrm{d}T}{\mathrm{d}t}(T+2E_0)^{1/2}\right] \\
=& \frac{2LaY(y)}{c}\frac{\mathrm{d}T}{\mathrm{d}t}\left[\frac{T^{1/2}}{2(T+2E_0)^{1/2}} + \frac{(T+2E_0)^{1/2}}{2T^{1/2}}\right] \\
=& \frac{2LaY(y)}{c}\frac{\mathrm{d}T}{\mathrm{d}t}\left[\frac{T+E_0}{[T(T+2E_0)]^{1/2}}\right] \\
=& \frac{J}{[T(T+2E_0)]^{1/2}}\frac{\mathrm{d}T}{\mathrm{d}t}\left[\frac{T+E_0}{[T(T+2E_0)]^{1/2}}\right] \\
=& \frac{J}{T(T+2E_0)}\frac{\mathrm{d}T}{\mathrm{d}t}\left[T+E_0\right]
\end{aligned}
\tag{A.7}
$$

From Equations A.7 and A.2, one can write:

$$
\frac{\mathrm{d}J}{\mathrm{d}t} = \frac{J}{2\mu}\frac{\mathrm{d}\mu}{\mathrm{d}t} \tag{A.8}
$$

The quantities $\mathrm{d}\mu/\mathrm{d}t$ and $\mathrm{d}J/\mathrm{d}t$ are theoretical. In reality collisional loss is a statistical phenomenon, and particles starting at the same energy may have different ranges even in a (macroscopically) homogeneous medium. This variation is known as range straggling. To describe real world behaviour for radiation belt particles, Equation A.8 can be phase averaged over a drift orbit like so:

$$
\left\langle\frac{\mathrm{d}J}{\mathrm{d}t}\right\rangle = \frac{J}{2\mu}\left\langle\frac{\mathrm{d}\mu}{\mathrm{d}t}\right\rangle \tag{A.9}
$$

Equation A.9 assumes that the length scale of a drift orbit is large enough to average over the straggling effect, so that $\langle\mathrm{d}\mu/\mathrm{d}t\rangle$ and $\langle\mathrm{d}J/\mathrm{d}t\rangle$ can be determined for a set of adiabatic coordinates. However, treating this quantity as constant over a drift orbit is an additional assumption, and requires that friction occurs slowly compared to the drift time. In Section 5.3.2, this is shown to be true since the timescales for collisional loss are much higher than one drift orbit for trapped radiation belt particles.

Appendix B

Numerical Solver Code

```
import numpy as np
from numpy import random

random.seed(1)

# Solution method:
#  determine beforehand f on relevant neighbouring grid points/ boundaries
#  determine beforehand M, R
#  Mf = R
#  LUf = R, y = Uf
#  Ly = R
#  Uf = y
#  then find f via back substitution

#-------------------------------------------------------------------------------+
# ::: setup :::
#-------------------------------------------------------------------------------+

m = 10 #size of the model grid in direction of k indicies
dimensions = 3
n = dimensions*m

w, h = n, n; #width and height of M

# several variables are pre-allocated below, just to make code more portable

#pre-allocate matrices L and U:
mat_L = [[0 for x in range(w)] for y in range(h)]
mat_U = [[0 for x in range(w)] for y in range(h)]

#set diagonal elements of L = 1:
for d in range(0,n):
    mat_L[d][d] = 1


#pre-allocate M:
# store A, B, E, and upper diagonals as random numbers:
dia_A = np.random.rand(m) #value between 0 and 1, first element unused
dia_B = (2+(dimensions))*np.ones(m) + np.random.rand(m)
# value between 2+(dimensions-1) and 2+(dimensions)
# larger numbers are to ensure diagonal dominance of M
dia_E = np.random.rand(m) #value between 0 and 1, last element unused

```

```
# store each upper diagonal in a list
udia_list = []
for dim in range(dimensions-1): #s, t, etc.
    udia_list.append(np.random.rand(m)) #between 0 and 1


#pre-allocate y1, f and R:
y1 = [0]*m #solve for this first via. forward-substituion

# form the solution vector f
f1 = [0]*m #solve for this second via. back-substituion
f_known = np.random.rand((dimensions-1)*m) #already solved for previously

# form the product vector R
RHS = np.random.rand(m) #between 0 and 1
R = np.array(list(RHS) + list(f_known))


#-------------------------------------------------------------------------------+
# ::: solver algorithm :::
#-------------------------------------------------------------------------------+

#first row:
mat_U[0][0] = dia_B[0]
mat_U[0][1] = dia_E[0]
for dim in range(dimensions-1):
    mat_U[0][(1+dim)*m] = udia_list[dim][0]

y1[0] = RHS[0]

#rows until j = m-1:
for j in range(1,m):
    mat_L[j][j-1] = dia_A[j]/mat_U[j-1][j-1]

    mat_U[j][j] = dia_B[j] - mat_L[j][j-1] * mat_U[j-1][j]

    mat_U[j][j+1] = dia_E[j]

    #elements between the upper diagonals and columns at multiplies of m in U:
    for dim in range(dimensions-1):
        for k in range(1,j+1):
            mat_U[j][(1+dim)*m+j-k] = - mat_L[j][j-1] * mat_U[j-1][(1+dim)*m+j-k]
        #upper diagonals in U:
        mat_U[j][(1+dim)*m+j] = udia_list[dim][j]
```

```
88
89
90      #forward substitute using L so far derived:
91      y1[j] = RHS[j] - mat_L[j][j-1] * y1[j-1]
92      #multiply element wise the first j-1 columns of the current row
93
94
95  #subsequent rows which can be pre-empted:
96  for j in range(m,dimensions*m):
97      mat_U[j][j] = 1
98
99
100 #back substitute to solve Uf = y:
101 for i in range(m-1, -1, -1):
102     bb = 0
103
104     #part stored in f_known
105     for j in range (m, n):
106         bb += mat_U[i][j]*f_known[j-m]
107
108     #part stored in f1
109     for j in range (i+1, m):
110         bb += mat_U[i][j]*f1[j]
111
112     f1[i] = (y1[i] - bb)/mat_U[i][i]
113
114
115 #-------------------------------------------------------------------------------+
116 # ::: validation :::
117 #-------------------------------------------------------------------------------+
118
119 #re-calculate M from the L, U solution:
120 mat_L = np.array(mat_L)
121 mat_U = np.array(mat_U)
122 mat_M = np.matmul(mat_L, mat_U)
123
124 #compare the A, B and E diagonals with M re-calculated from the L, U solution:
125 print("Validating tridiagonals via M = LU:")
126 print("")
127 dia_A_chk = np.array([mat_M[i+1][i] for i in range(m-1)])
128 dia_B_chk = np.array([mat_M[i][i] for i in range(m)])
129 dia_E_chk = np.array([mat_M[i][i+1] for i in range(m-1)])
130 print("dia_A original",dia_A)
131 print("dia_A check",dia_A_chk)
```

```
print(" Error (%):", max(100*(dia_A[1:] - dia_A_chk)/dia_A[1:]))
print("")
print("dia_B original",dia_B)
print("dia_B check",dia_B_chk)
print(" Error (%):", max(100*(dia_B - dia_B_chk)/dia_B))
print("")
print("dia_E original",dia_E)
print("dia_E check",dia_E_chk)
print(" Error (%):", max(100*(dia_E[:-1] - dia_E_chk)/dia_E[:-1]))
print("")
print("")
#compare the upper diagonals with M re-calculated from the L, U solution:
print("Validating upper diagonal(s) via M = LU:")
print("")
for dim in range(dimensions-1):
    S = udia_list[dim]
    S_chk = np.array([mat_M[i][i+(dim+1)*m] for i in range(m)])
    print("upper diagonal #",dim,"original",S)
    print("upper diagonal #",dim,"check",S_chk)
    print(" Error (%):", max(100*(S - S_chk)/S))
    print("")
print("")
print("Validating calculation of y via Ly = R:")
print("")
y = np.array(y1 + list(f_known))
R_chk = np.matmul(mat_L, y.T)
print("R original",R)
print("R check",R_chk)
print(" Error (%):", max(100*(R - R_chk)/R))
print("")
print("")
print("Validating calculation of f via Mf = R")
print("")
f = np.array(f1 + list(f_known))
R_chk = np.matmul(mat_M, f.T)
print("R original",R)
print("R check",R_chk)
print(" Error (%):", max(100*(R - R_chk)/R))
print("")
```