



HAL
open science

Machine Learning to Nowcast and Forecast Low-Energy Electron Fluxes in Auroral Regions

Simon Bouriat

► **To cite this version:**

Simon Bouriat. Machine Learning to Nowcast and Forecast Low-Energy Electron Fluxes in Auroral Regions. Other. Université Grenoble Alpes [2020-..], 2023. English. NNT : 2023GRALY061 . tel-04595098

HAL Id: tel-04595098

<https://theses.hal.science/tel-04595098>

Submitted on 30 May 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : PHYS - Physique

Spécialité : Astrophysique et Milieux Dilués

Unité de recherche : Institut de Planetologie et d'Astrophysique de Grenoble

Apprentissage machine pour la modélisation et la prévision des flux d'électrons auroraux de basse énergie

Machine Learning to Nowcast and Forecast Low-Energy Electron Fluxes in Auroral Regions

Présentée par :

Simon BOURIAT

Direction de thèse :

Mathieu BARTHELEMY

PROFESSEUR DES UNIVERSITES, UNIVERSITE GRENOBLE ALPES

Directeur de thèse

Jocelyn CHANUSSOT

Directeur de recherche, INRIA

Co-directeur de thèse

Rapporteurs :

ADELINE PAIEMENT

MAITRESSE DE CONFERENCES HDR, UNIVERSITE DE TOULON

THIERRY DUDOK DE WIT

PROFESSEUR DES UNIVERSITES, UNIVERSITE D'ORLEANS

Thèse soutenue publiquement le **1 décembre 2023**, devant le jury composé de :

ERIK KERSTEL,

PROFESSEUR DES UNIVERSITES, UNIVERSITE GRENOBLE ALPES

Président

ADELINE PAIEMENT,

MAITRESSE DE CONFERENCES HDR, UNIVERSITE DE TOULON

Rapporteuse

THIERRY DUDOK DE WIT,

PROFESSEUR DES UNIVERSITES, UNIVERSITE D'ORLEANS

Rapporteur

SIMON WING,

PROFESSEUR, JOHNS HOPKINS UNIVERSITY

Examineur

FREDERIC PITOUT,

ASTRONOME ADJOINT, UNIVERSITE TOULOUSE III - PAUL SABATIER

Examineur

MINNA PALMROTH,

PROFESSEURE, HELSINGIN YLIOPISTO

Examinatrice

Invités :

PAUL VANDAME

Université Grenoble Alpes



A Gérard, Betty, Jacques & Marie

L'homme et la mer

Homme libre, toujours tu chériras la mer !
La mer est ton miroir ; tu contemples ton âme
Dans le déroulement infini de sa lame,
Et ton esprit n'est pas un gouffre moins amer.

Tu te plais à plonger au sein de ton image ;
Tu l'embrasses des yeux et des bras, et ton cœur
Se distrait quelquefois de sa propre rumeur
Au bruit de cette plainte indomptable et sauvage.

Vous êtes tous les deux ténébreux et discrets :
Homme, nul n'a sondé le fond de tes abîmes ;
Ô mer, nul ne connaît tes richesses intimes,
Tant vous êtes jaloux de garder vos secrets !

Et cependant voilà des siècles innombrables
Que vous vous combattez sans pitié ni remord,
Tellement vous aimez le carnage et la mort,
Ô lutteurs éternels, ô frères implacables !

Charles Baudelaire

Remerciements

La thèse est une activité étrange, et tous ceux qui s'en sont sortis ont beaucoup de choses à en dire (et pas que positives d'ailleurs). Trois ans, c'est à la fois long et court. Long parce que beaucoup de choses changent en trois ans, en particulier si, comme la majorité des thésards, ces années se situent entre vos 20 et vos 30 ans. Et court parce que quand on arrive au bout, tout reste à faire : les sujets ne sont pas assez creusés, les résultats pas assez bons, et seule une petite surface du problème a été effleurée. La thèse c'est du bon, du moins bon, du difficile, de l'impossible (parfois), de la flemme (souvent), du stress (beaucoup) et de l'imprévu (tout le temps). Alors loin de moi l'idée de faire une généralité de mon cas, mais au final, la seule chose que ma thèse n'a pas été c'est solitaire. Ma thèse, c'est les autres. Michel Beaud disait très justement dans *L'art de la thèse. Comment préparer et rédiger un mémoire de master, une thèse de doctorat ou tout autre travail à l'ère du Net* : "Sachez trouver le ton juste : ni emphase ni circonvolutions ; soyez sobre et concis". Je comptais faire l'inverse de ce qu'il dit, parce qu'il dit aussi "l'ère du Net".

Le risque c'est toujours de savoir comment on doit remercier. Au début j'ai lu tout un tas de remerciements de tout un tas de thèses pour m'inspirer. C'était pas inspirant. Puis j'ai tapé "*comment bien dire merci dans une thèse ?*" dans la barre de recherche du Net. La réponse c'est "merci bien" ou encore que ça se prononce "mér-si". Pas inspirant non plus. Donc j'ai plutôt essayé de définir mon remerciement : soit A l'événement "réussir sa thèse" dans un espace probabilisé (Ω, \mathcal{F}, P) , et G l'ensemble des événements "intervention de gens". Alors, si vous êtes mentionnés dans cette thèse, vous appartenez à l'ensemble des événements x tels que : $\{x \subseteq G \mid P(A \mid \bigcup_{E \in x} \bar{E}) = 0\}$. Ou dit sans emphase ni circonvolutions, sobre et concis: sans vous, j'aurais raté ma thèse. Alors merci.

Merci à mes directeurs de thèse, Mathieu et Jocelyn. Merci d'avoir cru en ce projet dès le début et ce malgré les obstacles nombreux comme mon colossal retard en machine learning. Merci Mathieu pour tout ce que tu m'as apporté et pour cet incroyable voyage en Norvège qui m'a ouvert les yeux sur ma vocation. Je suis ravi de t'avoir rencontré et j'espère qu'on se verra encore à l'avenir. Merci aussi à Paul, mon presque directeur de thèse, qui m'a permis de combler en un temps record mes lacunes et qui m'a aidé à chacune des décisions du processus. Merci à Elisa, mon acolyte nouméro ouno, de ta rigueur et d'avoir été ma grande soeur de thèse pendant trois ans. Et merci à Pierre le sang avec qui nous avons défriché nos premières problématiques IA.

Bien sûr, un immense merci à tous les membres du jury. Tout d'abord Adeline PAIEMENT et Thierry DUDOK DE WIT pour avoir accepté d'être rapporteurs et pour m'avoir fourni de si bon retours. Merci d'avoir pris le temps de lire cette (très longue) thèse en détail. Merci également à tous les examinateurs. Merci Dr. Simon WING pour nos innombrables discussions depuis notre rencontre à la conférence *Machine Learning in Heliophysics: ML-Helio* et jusqu'à la publication d'un travail commun. Merci Frédéric PITOUT, depuis les cours du double diplôme SUPAERI/ASEP jusqu'à cette thèse, que de chemin parcouru. Merci Minna PALMROTH. C'est un honneur de vous avoir eu dans mon jury, et d'avoir eu la chance de découvrir Vlasiator. J'espère un jour pouvoir rejoindre une telle équipe et travailler sur un sujet aussi passionnant. J'en profite aussi pour remercier Maxime GRANDIN. Merci pour ton temps, tes impressionnantes connaissances sur l'ionosphère et la magnétosphère et les idées que tu m'as donné lors de la rédaction du manuscrit. Enfin, un immense merci à Erik KERSTEL pour les remarques très intéressantes sur mes travaux, pour nos échanges qui ont suivi, et surtout d'avoir accepté de présider ce jury de thèse.

Merci à SpaceAble et à Julien CANTEGREIL sans qui cette thèse n'aurait pas eu lieu. Merci à tous mes collègues et ex-collègues de SpaceAble. Vous avez tous, parfois sans le savoir, contribué

Remerciements

à la réussite de ce projet. Que de personnes inspirantes tout au long de cette aventure. Merci Angélique, Lionel, Yannis, Céline, Delphine, Isao, Antoine, Arnaud, Louis, Dimitri, Olivier, Benjamin, Fatoumata, Robin, et tous les autres, les "anciens" comme les nouveaux. Difficile de vous mentionner tous sans ajouter trop de lignes à ces remerciements déjà trop longs mais j'espère garder contact avec chacun d'entre vous et avoir la chance de continuer à échanger à l'avenir. C'est grâce à vous si aujourd'hui les mots "vie" et "active" mis bout à bout ne me font plus peur.

Merci à tous les potos, les coropotes, les lads, les supaeriens, les chimistes, les washingtoniens, les potes d'enfance, le XV, science & socio, les forceux. Je sais que vous savez dans quel bloc vous êtes. De Paris à Bruxelles, Bordeaux, Grenoble, Biarritz, le Bassin et tous les endroits où je vous ai croisé. Merci d'avoir supporté cette thèse dont vous avez plus qu'entendu parler (même pour ceux que j'ai rencontré il y a peu). Et double merci à tous ceux (oroken, les Gut on est là) à qui j'ai promis de faire des centaines de choses dès que la thèse s'arrêterait. Ça y est, nous y est.

Merci à mon incroyable famille, sans laquelle rien n'aurait trop d'intérêt de toute façon. Merci à mes parents. Je sais que je vous en ai demandé beaucoup sur ces trois années. Merci Mam d'être tant à l'écoute et de décrocher à chacun de mes appels. Merci Pap pour nos discussions, pour cette aventure au Népal, parenthèse forte qui a précédé et sûrement permis l'écriture du manuscrit. Merci Lul pour la bonne humeur, les memes, ou même d'avoir sacrifié tes révisions pour venir à ma soutenance. Merci ma Chlo pour tes nombreux passages à Paris, épisodes courts et agréables qui me permettaient de souffler (et merci pour la Guadeloupe hehe). Merci à mes cousins et cousines, mes oncles et tantes. Et merci à mes quatre grands-parents à qui je dédis cette thèse. Merci de m'avoir offert quatre visions distinctes du monde, chacune aussi riche que les autres. Merci pour ma liberté de penser et merci pour tout cet amour.

Merci à celle qui m'a accompagné presque chaque jour de cette aventure compliquée. Merci d'avoir enduré mes humeurs et mes anxiétés, d'avoir donné ton avis sur tout, d'avoir su me soutenir et m'aiguiller dans les (nombreux) moments difficiles. Je ne serai pas allé bien loin sans toi et j'espère te le rendre au centuple.

Merci à toutes les personnes inspirantes que j'ai croisé et que je croise encore et merci à celles qui ont gardé le contact avec moi. Merci Stéphanie, merci Giuseppe, thanks Angelique. Thanks to Dr. Camporeale and Dr. McGranaghan for your inspiring work, essential to do this PhD.

Merci aux créateurs de jeux de société (laissez-m'en un peu j'arrive), aux Pioneer XDJ-RR, à Google Scholar, à OpenAi, au Shift Project, et j'en passe. Enfin, mention spéciale à Koji Kondo sans qui le manuscrit serait encore une page blanche.

J'ai l'impression que nous arrivons tous ensemble au bout d'une course que nous avons tous couru et que mon rôle était seulement de tenir le chrono. On a mis 28 464 heures et 31 minutes pour un total de 317 pages. Pas mal. C'est 17 078 fois le temps que va réaliser Colme au marathon la semaine prochaine pour faire seulement 1.32 fois plus de pages que le livre de Paco. Quand même pas mal. Je crois.

Merci infiniment.

Abstract

Space weather is a broad scientific field focused on understanding the causes and consequences of the interactions between Earth and its surrounding space environment. It encompasses studies on solar activity, geomagnetic storms, or cosmic rays, and aims to improve our knowledge of the sun, interplanetary, and planetary environments, as well as to refine predictions and models of the disturbances that affect them. Indeed, phenomena in space weather can have direct implications on space systems, terrestrial structures, and even astronauts. In this context, methods and tools from artificial intelligence could play a pivotal role in modeling, forecasting, and understanding various phenomena.

This PhD thesis is part of a CIFRE program, signifying a collaboration between industry, specifically the company SpaceAble, and academia. The goal is to develop operational products to better apprise satellite operators and insurers of potential risks. The thesis is structured into a theoretical section, serving as a repository of information, and a practical section, where codes and models are delineated. The former is intended to act as a comprehensive reference for the partner company, covering a wide range of topics in space weather and artificial intelligence, and suggesting future avenues of research. The latter section is dedicated to fulfilling our objectives. Here, we model and seek to forecast auroral electron total energy fluxes (in the low-energy range, ≤ 30 keV) as measured by the SSJ/4 and SSJ/5 instruments aboard the DMSP satellites. Two types of machine learning models are employed: Fully Connected Neural Networks (FCNN) and Temporal Convolutional Networks (TCN). Both are trained using measurements of the solar wind, interplanetary magnetic field, and specific near-Earth indices, collected by the ACE satellite and on NASA's OMNIWeb database. Practicality and industrial applicability have guided our choice of the PyTorch and PyTorch-Lightning libraries. Each model and database has been meticulously analyzed and evaluated, with the results compared to two established models: OVATION, widely recognized in the community, and PrecipNet, an FCNN, and the currently most proficient model.

The FCNN and TCN architectures show higher performances than OVATION across all basic metrics (MSE, MAE, and RMSE) employed but are marginally superior than PrecipNet's. However, adapting PrecipNet to new problems is challenging due to its code structure and data pre-processing, making our model a preferred solution in terms of industrialization and transparency. Importantly, while FCNNs have set robust foundations, TCNs have emerged as more promising, managing extensive temporal data ranges, albeit with longer training times and a requirement for continuous data. Several solutions, including specific interpolations, new datasets, and architecture combinations, are considered to address these limitations and will be explored in future work.

A notable advantage of this study is the establishment of a framework of methods and ideas that can be amalgamated or adapted with new datasets to tackle other challenges. A significant insight is the promising integration of TCN with the "integrated gradients" method, enhancing model interpretability by attributing "importance" to each input parameter. This will also enable tracing information such as input-output delay and determining which solar wind parameter exerts the most influence near Earth.

In conclusion, this thesis provides a comprehensive overview of space weather and methodologies used in analyses, with a focus on the applicability of machine learning techniques. It serves both as a reflection of our current understanding of several phenomena and as a stepping stone towards deeper collaborations with industries and private companies, which hold a significant role in protecting and preserving the space environment.

Résumé

La météorologie de l'espace (SWE) explore les causes et conséquences des interactions Terre-environnement spatial, de l'activité solaire aux tempêtes géomagnétiques, en passant par la dynamique ionosphérique et les rayons cosmiques. Elle tente de comprendre l'état du soleil, du milieu interplanétaire et planétaire, et de prévoir et modéliser les perturbations qui les affectent. En effet, les phénomènes en jeu impactent directement les systèmes spatiaux, les structures sol et même les astronautes. Dans ce contexte, les méthodes et outils d'intelligence artificielle (IA) ont un rôle à jouer et peuvent nous aider à modéliser, prévoir et comprendre certains phénomènes.

Cette thèse est une collaboration entre l'industrie et la recherche. Son objectif est de fournir des produits opérationnels utilisés pour mieux informer les opérateurs ou assureurs de satellite des dangers qu'ils encourent. Elle se découpe en une partie théorique, fonctionnant comme un recueil d'informations, et une partie plus pratique, où les codes et les modèles sont présentés. La première sert d'encyclopédie pour l'entreprise partenaire et balaye plus largement les sujets de SWE et d'IA, englobant par la même occasion de futures pistes de recherche pour l'entreprise. La deuxième partie se concentre sur la modélisation et la prévision des flux d'électrons auroraux de basse énergie (≤ 30 keV) tels que mesurés par les instruments SSJ/4 et SSJ/5 du programme DMSP. Pour cela, deux familles de modèles d'IA sont utilisées: les Réseaux Neuronaux Entièrement Connectés (FCNN) et les Réseaux Convolutionnels Temporels (TCN). Pour les entraîner, nous utilisons les mesures de vent solaire, les composantes du champ magnétique interplanétaire et certains indices proche-Terre, mesurés par le satellite ACE et par la base de données OMNI-Web de la NASA. Pour des raisons pratiques et d'applications industrielles, nous avons utilisé les bibliothèques PyTorch et PyTorch-Lightning. Données, modèles et résultats ont été méticuleusement évalués et comparés à OVATION, modèle le plus largement utilisé par la communauté, et PrecipNet, un FCNN, et le plus performant aujourd'hui.

Nos FCNN et TCN obtiennent des performances bien supérieures à OVATION sur les métriques utilisées mais que très légèrement meilleures que PrecipNet. En revanche, PrecipNet est, par construction et prétraitement des données, difficile à adapter à de nouvelles problématiques, ce qui fait de notre produit une meilleure réponse aux besoins d'industrialisation. Enfin, si le FCNN a posé des bases et performances solides, c'est le TCN qui s'est révélé le plus prometteur, par sa capacité à gérer d'importantes plages de données temporelles et ce malgré son temps d'entraînement plus long et son besoin en données continues. Des solutions à ces problèmes telles que des interpolations spécifiques, de nouveaux jeux de données, ou des combinaisons d'architectures ont été envisagées et feront l'objet de travaux futurs.

Un des avantages de cette étude est qu'elle pose une base réadaptable et combinable avec de nouvelles données pour répondre à d'autres problématiques. En particulier, une idée prometteuse est la combinaison du TCN avec la méthode des "gradients intégrés" qui permet de renforcer l'interprétabilité du modèle. Il sera ainsi possible de remonter à des informations comme le délai entre l'entrée et la sortie ou savoir quel paramètre de vent solaire a le plus d'influence sur ce qu'il se passe près de la Terre.

En conclusion, la thèse offre une vue d'ensemble des méthodes et approches utilisées pour l'analyse du SWE, avec une attention particulière portée aux techniques d'apprentissage automatique et à leur applicabilité dans ce domaine. En essence, elle sert à la fois de témoignage de notre compréhension actuelle de certains phénomènes, et marque un pas vers des collaborations plus profondes avec l'industrie et les entreprises privées, qui ont également un rôle important à jouer dans la protection et la préservation de l'environnement spatial.

Acronyms

AACGM	Altitude Adjusted Corrected Geomagnetic
ACR	Anomalous Cosmic Ray
ACE	Advanced Composition Explorer
AE	Auroral Electrojet / Autoencoders
AGU	American Geophysical Union
AI	Artificial Intelligence
AL	Auroral Electrojet Lower Envelope
ANN	Artificial Neural Network
AU	Auroral Electrojet Upper Envelope / Astronomical Unit
BSN	Bow Shock Nose
CHAMP	CHALLENGING Mini-satellite Payload
CIFRE	Conventions Industrielles de Formation par la REcherche (Cifre)
CIR	Corotating Interaction Region
CME	Coronal Mass Ejection
CREAM	Collision Risk Estimation and Automated Mitigation
CRAND	Cosmic Ray Albedo Neutron Decay
CRRES	Combined Release and Radiation Effects Satellite
DCA	Dynamic Calibration Atmosphere
DDD	Displacement Damage Dose
DMSP	Defense Meteorological Satellite Program
DOY	Day of Year
DSCVR	Deep Space Climate Observatory
DST	Disturbance Storm-Time (Index)
DTM	Drag Temperature Models
ECI	Earth Centered Inertial
ELF	Extremely Low Frequency
EMFISIS	Electric and Magnetic Field Instrument Suite and Integrated Science
EMIC	Electromagnetic Ion Cyclotron
ESA	European Space Agency
ESD	Electrostatic Discharge
FCN	Fully Connected Network
FCNN	Fully-Connected Neural Network
GCR	Galactic Cosmic Ray
GEO	Geostationary Orbit
GITM	Global Ionosphere Thermosphere Model
GOCE	Gravity Field and Steady-State Ocean Circulation Explorer

GRACE	Gravity Recovery And Climate Experiment
GRACE-FO	Gravity Recovery And Climate Experiment Follow-On
GUVI	Global Ultraviolet Imager
HASDM	High Accuracy Satellite Drag Model
HF	High Frequency
HEO	Highly Elliptical Orbit
ICCTCT	International Conference on Current Trends towards Converging Technologies
ICET	International Conference on Engineering and Technology
ICME	Interplanetary Coronal Mass Ejection
ICMLA	International Conference on Machine Learning and Applications
IDM	Internal Discharge Monitor
IG	Integrated Gradient
IMF	Interplanetary Magnetic Field
ISGI	International Service for Geomagnetic Indices
ISSN	International Solar Sunspots Number
JAXA	Japan Aerospace Exploration Agency
JB	Jacchia-Bowman
LANL	Los Alamos National Laboratory
LASCO	Large Angle and Spectrometric Coronagraph
LEO	Low Earth Orbit
LPP	Laboratoire de Physique des Plasmas
LRO	Low Resolution OMNI
LSTM	Long Short-Term Memory
LTI	Lower-Thermosphere Ionosphere
MEO	Medium Earth Orbit
MDI	Michelson Doppler Imager
MHD	Magnetohydrodynamics
MLAT	Magnetic Latitude
MLT	Magnetic Local Time
MSE	Mean-Square Error
MSIS	Mass Spectrometer Incoherent Scatter
MUSCAT	Multi-Utility Spacecraft Charging Analysis Tool
MVTS	Multivariate Time Series
NASA	National Aeronautics and Space Administration
NASCAP	NASA Charging Analyzer Program
NENL	Near-Earth Neutral Line
NEO	Near-Earth Object
NMSE	Normalized Mean-Square Error
NOAA	National Oceanic and Atmospheric Administration

NLP	Natural Language Processing
NSF	National Science Foundation
OECD	Organization for Economic Co-operation and Development
OVATION	Oval Variation, Assessment, Tracking, Intensity, and Online Nowcasting
PC	Polar Cap
PCN	Polar Cap North
PCS	Polar Cap South
PFSS	Potential Field Source Surface
POES	Polar Operational Environmental Satellites
POLAR	Potentials Of Large Objects in the Auroral Region
PRESTO	Polar Explorer for Science and Technology Observations
QSPS	Quasi Static Potential Structure
RCN	Research Coordination Network
SCATHA	Spacecraft Charging At High Altitude
SCOSTEP	Scientific Committee on Solar-Terrestrial Physics
SEM	Space Environment Monitor
SEP	Solar Energetic Particles
SET	Single Event Transient
SHIELDS	Space Hazards Induced Near Earth by Large Dynamic Storms
SIR	Stream Interaction Region
SLR	Satellite Laser Ranging
SOHO	Solar and Heliospheric Observatory
SPARCS	Spacecraft Charging Software
SPENVIS	Space Environment Information System
SPIS	Space Plasma Interaction System
SSA	Space Situational Awareness
SSC	Sudden Storm Commencement
SSJ	Special Sensor J
STEREO	Solar Terrestrial Relations Observatory
SWPC	Space Weather Prediction Center
TCN	Temporal Convolutional Network
TEC	Total Electron Content
TID	Total Ionizing Dose
TIMED	Thermosphere Ionosphere Mesosphere Energetics and Dynamics
TLE	Two Line Elements
TOD	True of Date
TP	Trapped Energetic Proton
ULF	Ultra Low Frequency
UV	Ultraviolet
VLF	Very Low Frequency

Contents

Remerciements	i
Abstract	iv
Résumé	v
Acronyms	vii
Contents	xi
Introduction	1
1 Space Weather and its Measure	8
1.1 Introduction to Plasma Physics	9
1.1.1 Plasmas	9
1.1.2 Charged Particle Motions	10
1.1.2.1 Maxwell's Equations	10
1.1.2.2 The Cyclotron & Guiding Center Approximation	11
1.1.2.3 Drift Motions in Electromagnetic Fields	11
1.1.2.4 Adiabatic Invariants	13
1.1.2.5 Summary of Motions in the Radiation Belts	16
1.1.3 Kinetic theory: Vlasov equation	17
1.1.4 Fluid theory: Magnetohydrodynamics	18
1.1.4.1 MHD's equations	19
1.1.4.2 Magnetic field lines "frozen" in the plasma	19
1.1.4.3 Magnetic pressure and magnetic tension	20
1.2 The Sun-Earth Chain	21
1.2.1 Starting point: the Sun	21
1.2.1.1 Structure and characteristics	21
1.2.1.2 A Dynamic Activity	21
1.2.2 Solar Wind in the Interplanetary Medium	27
1.2.2.1 Properties of the Solar Wind	28
1.2.2.2 Idealized Model for Solar Wind	28
1.2.2.3 Influence of the Solar Magnetic Field	31
1.2.2.4 Associated Phenomena	33
1.2.3 Solar Wind's interaction with Earth's Magnetosphere	36
1.2.3.1 Introduction to Earth's Magnetosphere	36
1.2.3.2 Magnetopause & Magnetopause Current	37
1.2.3.3 Magnetic Reconnections & Dungey Cycle	39
1.2.3.4 Polar Cusps	41
1.2.3.5 Ring Current	43
1.2.3.6 Plasma Sheet & Cross-Tail Current	44
1.2.3.7 Plasmasphere	44
1.2.3.8 Radiation Belts	44
1.2.4 Geomagnetic Storms and Substorms	47

1.2.5	Plasma Waves in the Magnetosphere	48
1.2.6	Ionosphere	52
1.2.6.1	Introduction to the Earth's Atmosphere	52
1.2.6.2	Ionosphere	52
1.2.6.3	Field Aligned Currents	54
1.2.7	Auroral Physics	56
1.2.7.1	Related Particle Acceleration Processes	56
1.2.7.2	Auroral Types	57
1.2.7.3	Auroral Morphology & Models	58
1.3	Space Weather Measures and Forecasts	62
1.3.1	Measuring Space Weather through Indices	62
1.3.2	The Role of Modeling and Forecasting	64
1.3.3	Impacts on space and ground systems	67
1.3.3.1	Spacecraft Charging	67
1.3.3.2	Drag	71
1.3.3.3	Geomagnetically Induced Currents	74
1.3.3.4	Radiation Effects: Single Event Effects	74
1.3.3.5	Radiation Effects: Cumulative Effects	76
1.3.4	A Danger for Humanity	77
1.3.4.1	Events From the Past	78
1.3.4.2	Dangers	78
1.3.4.3	Worst Case Scenario	79
1.3.4.4	International Programs	80
1.3.5	The Rise of the New Space	81
2	Machine Learning, Deep Learning and Their Application in Space Weather Research	84
2.1	Introduction to Artificial Intelligence	85
2.2	Machine Learning	87
2.2.1	Machine Learning & Space Weather	88
2.2.2	Supervised Learning	89
2.2.2.1	Regression	91
2.2.2.2	Classification	91
2.2.2.3	Supervised Learning and Space Weather	91
2.2.3	Unsupervised Learning	92
2.2.3.1	Clustering	92
2.2.3.2	Dimensionality Reduction	93
2.2.3.3	Unsupervised Learning and Space Weather	94
2.2.4	Reinforcement Learning	94
2.2.5	Summary	94
2.3	From Supervised Machine Learning to Deep Learning	95
2.3.1	Mathematical Introduction of Supervised Learning Models	96
2.3.2	A First Neural Network	101
2.3.3	Training Neural Networks	102
2.3.3.1	Loss Function	103
2.3.3.2	Backpropagation	104
2.3.3.3	Optimizers	108
2.3.3.4	Schedulers	110
2.3.3.5	Activation Functions	112
2.3.4	Evaluation & Diagnostic of Neural Networks	113
2.3.4.1	Metrics	114
2.3.4.2	Training, Validation and Test Sets	116

2.3.4.3	Overfitting & Loss Curves	116
2.3.5	Fine-tuning Neural Networks	119
2.3.5.1	Regularization techniques	119
2.3.5.2	Tunable Parameters	121
2.3.6	Deep Learning System Design	121
2.3.6.1	Data Analysis	121
2.3.6.2	Data Preprocessing	124
2.3.6.3	Architecture	126
2.4	Libraries & Needed Tools	137
2.4.1	Hardware	137
2.4.2	Libraries	138
2.4.2.1	Getting Started with PyTorch	138
2.4.2.2	PyTorch-Lightning	140
3	Problem Statement, Data Analysis & Preprocessing	143
3.1	Clarifying the Research Problem	144
3.2	Data Description	146
3.2.1	Defense Meteorological Satellite Program (DMSP)	146
3.2.1.1	SSJ/4	147
3.2.1.2	SSJ/5	147
3.2.1.3	Precipitating Electron Data	150
3.2.2	Advanced Composition Explorer (ACE)	151
3.2.3	High-Resolution OMNIWeb	154
3.3	Data Analysis & Preprocessing	155
3.3.1	ACE	155
3.3.2	DMSP	174
3.3.2.1	New DMSP database from Redmon et al. (2017)	174
3.3.2.2	Our database	176
3.3.2.3	Histograms and Characteristics	176
3.3.2.4	Outliers	180
3.3.2.5	Uncertainties	181
3.3.3	High-Resolution OMNIWeb	182
3.3.4	Input-Output Relationship: Justification for AI Implementation	184
3.4	Summary & Final Preprocessing	208
3.4.1	Pre-process Summary	209
3.4.2	Combining inputs and outputs: final datasets	210
3.4.3	Summary of our study	214
4	Algorithmic Implementation, Iterative Procedures, and Results	217
4.1	Organization of our code	218
4.2	Exploring PrecipNet	219
4.2.1	PrecipNet	219
4.2.2	First Remarks	219
4.2.3	Attempt to Simplify PrecipNet	221
4.2.3.1	Reproduction of PrecipNet	221
4.2.3.2	The role of hidden layers and overfitting	222
4.2.3.3	The role of position & time as inputs	222
4.2.4	Conclusion	226
4.3	Introducing Our FCNNs	227
4.3.1	Training Process	227
4.3.2	Visualizing Results	228

4.3.3	Performance Metrics	228
4.3.4	Comparative Analysis Insights	234
4.4	Forecast	238
4.5	Addressing Limitations with the TCN	240
4.5.1	Training Process	240
4.5.2	Visualizing Results	242
4.5.3	Performance Metrics	242
4.5.4	Insights on the Final Product & Comparative Analysis	243
4.5.5	A Step Further: Integrated Gradients	245
4.6	Final Remarks	248
5	Future Directions & Applications	252
5.1	Summary of Key Observations & Results	253
5.2	Perspectives for Future Work	254
5.2.1	Improvements over our Algorithms	254
5.2.2	Combination of Models	255
5.2.3	Integrated Gradients	256
5.3	Implications and Applications for the Research Community	257
5.4	Implications and Applications for the Industry	257
5.5	Conclusion	258
	Conclusion	261
	Résumé en français	i
	Bibliographie	I
	List of figures	XXV
	List of Tables	XXXIV

Introduction (français)

Cette thèse est le fruit de trois ans de collaboration entre l'entreprise SpaceAble et les laboratoires IPAG (Institut de Planétologie et d'Astrophysique de Grenoble) et Gipsa-Lab (Grenoble Images Parole Signal Automatique) dans le cadre d'un dispositif CIFRE (Conventions Industrielles de Formation par la REcherche). Ce dispositif, mis en place par l'ANRT (Association Nationale Recherche Technologie), a pour objectif de "favoriser le développement de la recherche partenariale publique-privée et de placer les doctorants dans des conditions d'emploi"¹. Dans ce cadre, le salarié-doctorant travaille avec les laboratoires et l'entreprise afin d'oeuvrer pour un but commun. Cette thèse a été cosupervisée par les Dr. Mathieu BARTHELEMY et Jocelyn CHANUSSOT, une cosupervision importante où chacun d'eux a su transmettre ses connaissances afin que les deux domaines se mélangent efficacement et, on l'espère, harmonieusement.

SpaceAble est une start up française créée en 2018 par M. Julien CANTEGREIL. Elle est spécialisée dans le domaine du *Space Situational Awareness* et offre donc dans ce cadre un service de gestion du trafic spatial (ou Space Traffic Management - STM), de compréhension et prévision de la météorologie de l'espace (ou Space Weather - SWE) ainsi qu'un service "en-orbite" (In Orbit Services - IOS) pour les orbites basses. L'objectif de ses solutions est d'assurer la sécurité des opérations et de contribuer à la durabilité des satellites en orbite basse. Elle a également pour vocation de contribuer à l'élaboration des normes européennes pour réguler les activités en orbite basse. Au sein de cette entreprise, j'occupe le poste d'ingénieur et scientifique en météorologie de l'espace. J'oeuvre ainsi à la fois à porter une expertise en SWE pour l'entreprise, mais également à développer des solutions opérationnelles utilisables par des clients. De cela découlent diverses compétences : gestion de projet, mise en place de partenariats, analyse de données, état de l'art, développement de modèles de risques, et bien d'autres, ce qui me permet de bénéficier d'une double formation académique et professionnelle.

L'objectif de cette thèse est la modélisation et la prévision des flux d'électrons auroraux de basse énergie (≤ 30 keV) tels que mesurés par les instruments SSJ/4 et SSJ/5 du Defense Meteorological Satellite Program des Etats-Unis pour lequel plus de 20 ans de données viables sont disponibles. Une fois la modélisation fonctionnelle, les résultats sont à la fois comparés à des modèles existants (OVATION Prime, le plus utilisé et PrecipNet, le plus récent et performant) et étendus spatialement afin de produire des cartes polaires qui représentent ces flux d'électrons dits *précipités*. Les paramètres utilisés pour cette modélisation sont les paramètres du vent solaire (vitesse, densité, pression), les composantes X, Y et Z du champ magnétique interplanétaire (en coordonnées GSE), et les indices AL, AU et SYM-H. Les architectures utilisées sont des réseaux de neurones entièrement connectés (FCNN - Fully Connected Neural Network) et des réseaux de neurones convolutionnels temporels (TCN - Temporal Convolutional Network). Nos travaux sont ici présentés en cinq chapitres distincts. Du chapitre 1 au chapitre 3, cette thèse est aussi structurée comme une compilation de connaissances du domaine et sert donc de référence pour les recherches

1. <https://www.anrt.asso.fr/fr/le-dispositif-cifre-7844>

qui sont menées au sein de l'entreprise SpaceAble. Par conséquent, certaines des informations qui s'y trouvent sortent du cadre de notre étude et allonge quelque peu ces trois premiers chapitres.

Au travers du chapitre 1 nous espérons présenter un panorama exhaustif des acteurs de la météorologie de l'espace, révélant leurs interactions captivantes, les phénomènes et théories qui les composent. Depuis notre Soleil et jusqu'au champ magnétique de notre Terre, ce chapitre tente de saisir comment les éruptions solaires, les éjections de masse coronale ou encore les tempêtes géomagnétiques sont liées. Bien entendu, il ne s'agit pas d'un aperçu théorique détaillé et plusieurs sujets sont volontairement omis. Ce chapitre invite plutôt le lecteur à explorer la physique spatiale, à acquérir une compréhension plus profonde de l'influence significative de la météorologie spatiale sur notre vie quotidienne. Comme annoncé, c'est donc un chapitre qui pourra servir de connaissance de base pour l'entreprise SpaceAble et qui permettra à des non-initiés de saisir l'essentiel.

De façon analogue, le chapitre 2 expose le domaine de l'intelligence artificielle (IA), en mettant l'accent sur les branches de l'apprentissage machine (Machine Learning) et profond (Deep Learning). Ce chapitre vise à fournir une compréhension approfondie des mécanismes, des algorithmes et des principes sous-jacents à ces technologies émergentes, établissant ainsi une fondation solide pour les discussions et analyses ultérieures dans le reste de l'ouvrage. Aussi, nous y expliquons les concepts d'apprentissage supervisé, non-supervisé ou de renforcement, que nous ponctuons d'exemple inhérent à la météorologie de l'espace. Nous présentons également les mathématiques qui se cachent derrière des réseaux de neurones simples, les notions d'hyperparamètres et les axes principaux sur lesquels agit le "data scientist". Enfin nous le concluons avec une très brève présentation de la place qu'occupe le choix du matériel (hardware, GPU et CPU), des librairies et des outils d'IA.

Le chapitre 3 marque le début de l'étude spécifique et la fin de la partie théorique. Après avoir posé le contexte et clarifié la problématique, il se consacre entièrement à la présentation et à l'analyse des données à notre disposition. Effectuer une analyse de données avant de se lancer dans des projets d'Intelligence Artificielle (IA) est essentiel car cela permet de comprendre la structure, la qualité et les tendances inhérentes aux données. Cette démarche assure que nos modèles d'IA seront bien adaptés, augmente la précision des prédictions en réduisant le risque d'erreurs liées à des données mal interprétées ou de mauvaise qualité. Cette étape est, selon nous, le coeur de la thèse, l'étape la plus importante et la plus chronophage. Dans un premier temps, nous y présentons les données : les mesures des satellites de DMSP, les données extraites de OMNIWeb, et les données du satellites Advanced Composition Explorer (ACE). Puis nous les analysons *sous le prisme de l'IA* ce qui a, dans le cas des données de ACE, donné lieu à un papier de recherche (Bouriat et al., 2022) dont l'objectif était également de présenter une méthode pour ce genre d'analyses. Enfin, nous tentons de démontrer, hors IA, si un lien existe bien entre les données d'entrée et les données de sortie, ce qui a également donné lieu à une publication (Bouriat et al., 2023). Une fois toutes ces analyses effectuées, nous pouvons conclure en exposant le pré-traitement le plus adéquat pour notre problématique.

Ce qui nous amène au chapitre 4 : l'implémentation des algorithmes et la présentation des résultats. Pour ce faire, initialement, nous nous sommes concentrés sur PrecipNet, compte tenu de ses résultats pertinents. Notre objectif principal y est de reproduire ses résultats en interne, afin d'avoir une meilleure compréhension de son fonctionnement et de ses limitations. Ensuite, à partir de ce que nous avons appris, nous introduisons notre FCNN. Une comparaison approfondie de nos résultats avec ceux de PrecipNet (McGranaghan et al., 2021) et d'OVATION Prime suivra. Par la suite, nous présenterons des prévisions court-terme (10 min) des flux d'électrons précipités afin de valider la faisabilité de futurs travaux de recherche. Enfin nous présentons notre TCN. Il a l'avantage d'être une méthode beaucoup plus récente qui a fait ses preuves avec les séries tem-

porelles (Bai et al., 2018), et il vient corriger l'un des principaux problèmes de PrecipNet et de nos FCNN : le choix subjectif des données historiques à utiliser. Il utilise en effet une plage temporelle entière du passé, dont la taille ne dépend que de nous. Dans cette thèse, nous nous restreignons à 30 minutes pour des raisons de puissance de calcul.

Enfin, nous concluons cette thèse avec le chapitre 5 qui est l'occasion de rassembler nos résultats et de conclure. Ce chapitre présente les améliorations concrètes des algorithmes sur lesquelles nous travaillons déjà, mais il sert également d'ouverture vers les nombreuses recherches qu'il restera à faire, tant dans le domaine académique que dans le domaine industriel. Ce domaine étant étroitement lié avec celui des dangers concrets pour les acteurs du spatial, nous sommes confiants quant à l'avenir qu'il aura auprès des entreprises privées.

Avant d'entamer le manuscrit et pour terminer cette introduction, nous tenions à remercier OMNIWeb² et le Coordinated Data Analysis Web³ de la NASA, ainsi que le ACE Science Center⁴ de l'Université de Caltech pour avoir rendu disponible et publique l'ensemble des données utilisées ici. Je tiens également personnellement à remercier toutes les personnes impliquées dans le projet, l'IPAG, le Gipsa-Lab et, bien sûr, SpaceAble.

2. <https://omniweb.gsfc.nasa.gov/>

3. <https://cdaweb.gsfc.nasa.gov/>

4. <https://izw1.caltech.edu/ACE/ASC/>

Introduction

This dissertation marks the completion of a three-year collaborative project between SpaceAble, the IPAG (Institute of Planetology and Astrophysics of Grenoble) and the Gipsa-Lab (Grenoble Images Speech Signal Automatic) laboratories, carried out within the framework of a CIFRE arrangement (Conventions Industrielles de Formation par la REcherche). This program, established by the ANRT (Association Nationale Recherche Technologie), aims to "promote collaborative research between public and private sectors and to create conditions conducive to employment"⁵. In this setting, the doctoral candidate, while employed, works with both the laboratories and the company towards a common goal. The dissertation was co-supervised by Dr. Mathieu BARTHELEMY and Jocelyn CHANUSSOT, a valuable relationship where each could share his expertise, aiding in an effective, and hopefully, seamless integration of the two fields.

SpaceAble is a French start-up, founded in 2018 by Mr. Julien CANTEGREIL, that specializes in Space Situational Awareness. As such, it offers a wide range of services including Space Traffic Management (STM), Space Weather (SWE) understanding and forecasting, as well as in-orbit services (IOS) for low-Earth orbits. The main goal of these solutions is to ensure operational safety and contribute to the sustainability of satellites in low-Earth orbit. Additionally, the company is committed to contributing to the development of European standards to regulate activities in low-Earth orbit.

In this innovative company, I serve as an Engineer and Scientist in Space Weather. My role is varied, involving both providing SWE expertise to the company and developing operational solutions that are suitable for our clients. This role requires the development of a diverse skill set, including project management, forming partnerships, data analysis, cutting-edge research, development of risk models, and more. This wide range of responsibilities has allowed me to gain a well-rounded mix of academic and professional experience.

The goal of this thesis is to model and forecast the electron total energy fluxes for low-energy auroral electrons (≤ 30 keV) as measured by the SSJ/4 and SSJ/5 instruments of the United States Defense Meteorological Satellite Program, for which over two decades of viable data are available. Once the model is operational, the outcomes are compared to existing models — specifically OVATION Prime, the most commonly used, and PrecipNet, the newest and most efficient. Additionally, the results are spatially expanded to produce polar maps depicting these so-called precipitating electron fluxes. The parameters used for this modeling include solar wind parameters (speed, density, pressure), the X, Y, and Z components of the interplanetary magnetic field (in GSE coordinates), and the AL, AU, and SYM-H indices. The architectural frameworks utilized in this study are Fully Connected Neural Networks (FCNN) and Temporal Convolutional Networks (TCN). Our work is presented in five distinct chapters in this document. Chapters 1 to 3 of this thesis serve as a comprehensive compilation of domain knowledge, acting as a reference for research within SpaceAble. Consequently, some included information extends beyond our study's

5. <https://www.anrt.asso.fr/en/cifre-scheme-7844>

scope, slightly lengthening these initial chapters.

In the scope of Chapter 1, we aim to provide a broad overview of the main players in space meteorology, explaining their interesting interactions and the various phenomena and theories that make up this field. From our powerful Sun to Earth’s complex magnetic field, this chapter seeks to understand the intricate relationships between solar flares, coronal mass ejections, and geomagnetic storms. It is important to mention that this is not intended to be an exhaustive theoretical overview, as several topics are intentionally left out for the sake of brevity and clarity. Instead, this chapter invites the reader to explore the complexities of space physics and to gain a deeper understanding of the widespread impact of space weather on our everyday lives. In this way, this chapter is designed to act as a foundational source of knowledge for SpaceAble in the field, offering an essential understanding of the subject for those previously unfamiliar with it.

Similarly, Chapter 2 sheds light on the field of Artificial Intelligence (AI), highlighting the branches of Machine Learning and Deep Learning. The goal of this chapter is to provide a detailed understanding of the mechanisms, algorithms, and foundational principles driving these growing technologies, thus laying a strong foundation for further discussions and analyses throughout the rest of this manuscript. Additionally, we explain the concepts of supervised, unsupervised, and reinforcement learning, illustrating each with examples directly related to space meteorology. We also uncover the basic mathematics behind elementary neural networks, introduce the idea of hyperparameters, and examine the main areas where a data scientist works. Finally, the chapter concludes with a brief but informative overview of the crucial role of hardware selection (including GPUs and CPUs), libraries, and Artificial Intelligence tools in this field.

Chapter 3 marks the initiation of the specific study and concludes the theoretical portion. With the context established and the problem statement clarified, this chapter is entirely dedicated to the presentation and analysis of the available data. Undertaking a data analysis before diving into Artificial Intelligence (AI) projects is crucial, as it fosters an understanding of the data’s structure, quality, and inherent trends. This strategy ensures that our AI models are aptly tailored, heightens the accuracy of predictions, and reduces the risk of errors stemming from misinterpreted or inferior data. In our viewpoint, this stage is the linchpin of the thesis and represents the most critical and time-intensive phase. Initially, we introduce the data: measurements from the DMSP satellites, data harvested from OMNIWeb, and measures from the Advanced Composition Explorer (ACE) satellite. Following this, we scrutinize them through the AI prism, which, in the case of ACE data, culminated in a research paper (Bouriat et al., 2022). The paper’s goal was also to propose a methodology for such analyses. Ultimately, we seek to determine, beyond the confines of AI, whether a palpable link truly exists between the input and output data, a pursuit that also yielded a publication (Bouriat et al., 2023). Once all these analyses have been conducted, we are positioned to conclude by outlining the most apt preprocessing for our problem statement.

This brings us to Chapter 4, which is dedicated to the implementation of algorithms and the presentation of the resultant findings. Our initial focus is directed towards PrecipNet, attributable to its salient outcomes. Our first objective within this chapter is to internally replicate its results, thereby having a better understanding of its operational mechanisms and inherent limitations. Following this, informed by the insights gained, we introduce our Fully Connected Neural Network (FCNN). A meticulous comparison of our results with those procured by PrecipNet (McGranaghan et al., 2021) and OVATION Prime will ensue. Next, we will present short-term (10-minute) forecasts of precipitating electron fluxes to validate the possibility of future research work. Concluding this chapter, we present our Temporal Convolutional Network (TCN). This approach offers the benefit of being a comparatively recent method that has exhibited proficiency with time series (Bai et al., 2018). Furthermore, it remedies a fundamental concern associated with PrecipNet and our

FCNN – the subjective determination of historical data to incorporate. It makes use of an entire temporal range from the past, the duration of which is determined solely by us. In this thesis, we restrict this to 30 minutes due to limitations in computational capacity.

This thesis ends with Chapter 5, acting as a key moment to combine our findings and form definitive conclusions. This chapter both outlines the advancements in the algorithms we are still refining and opens the door to a multitude of future research opportunities, covering both the academic and industrial fields. Given that this area is closely connected with real risks facing space stakeholders, we maintain a positive view on its likely importance within private enterprises.

Before diving into the manuscript and to conclude this introduction, we would like to express our sincere thanks to NASA's OMNIWeb⁶, Coordinated Data Analysis Web⁷, and to the ACE Science Center⁸ of the Caltech University, for generously making all the used data available and public. On a personal note, I would like to extend heartfelt thanks to everyone involved in the project, including the teams at IPAG, Gipsa-Lab, and, of course, SpaceAble.

6. <https://omniweb.gsfc.nasa.gov/>

7. <https://cdaweb.gsfc.nasa.gov/>

8. <https://izw1.caltech.edu/ACE/ASC/>



Space Weather and its Measure

"On s'apercevra vite que la nuit à la belle étoile est néfaste. La voûte céleste rend insomniaque : trop de beauté, trop de grandeur pour songer à dormir."

Sylvain Tesson - Petit traité sur l'immensité du monde

Contents

1.1	Introduction to Plasma Physics	9
1.1.1	Plasmas	9
1.1.2	Charged Particle Motions	10
1.1.3	Kinetic theory: Vlasov equation	17
1.1.4	Fluid theory: Magnetohydrodynamics	18
1.2	The Sun-Earth Chain	21
1.2.1	Starting point: the Sun	21
1.2.2	Solar Wind in the Interplanetary Medium	27
1.2.3	Solar Wind's interaction with Earth's Magnetosphere	36
1.2.4	Geomagnetic Storms and Substorms	47
1.2.5	Plasma Waves in the Magnetosphere	48
1.2.6	Ionosphere	52
1.2.7	Auroral Physics	56
1.3	Space Weather Measures and Forecasts	62
1.3.1	Measuring Space Weather through Indices	62
1.3.2	The Role of Modeling and Forecasting	64
1.3.3	Impacts on space and ground systems	67
1.3.4	A Danger for Humanity	77
1.3.5	The Rise of the New Space	81

1.1 Introduction to Plasma Physics

The objective of this thesis is to leverage artificial intelligence tools for modeling and predicting particle flows in low Earth orbit. To accomplish this objective, it is crucial to have a thorough understanding of the physical reality of these phenomena and of the equations that governs them. The particle flows are an important aspect and one of the final components in the Sun-Earth chain. Grasping their origin entails comprehending all aspects of this chain. Additionally, comprehending the equations that underlie these phenomena is vital for selecting suitable AI algorithms and developing precise prediction models.

In this section, we will introduce and explain the physical and mathematical concepts underlying plasmas. The aim here is to introduce those necessary for understanding this thesis and the following chapters, and not to provide a comprehensive presentation. If the reader wishes to obtain more information, there are excellent references in French such as [Delcroix and Bers \(1994\)](#) or in English like [Bellan \(2006\)](#), [Cravens \(1997\)](#), or [Kivelson and Russell \(1995\)](#). With this introduction, we aim to provide sufficient background, eliminating the need for equations in parts 2 and 3.

1.1.1 Plasmas

Plasma, often referred to as the fourth state of matter, is an ionized gas that contains a significant fraction of free electrons and protons. Plasma physics is a relatively recent science, which emerged shortly after World War II and was intimately linked to the beginnings of space exploration. Plasma constitutes approximately 99% of the visible matter in the Universe. It exists under very high temperatures, which can arise in various conditions:

- through heating, for temperatures above 10^4 Kelvin, conditions found in stellar atmospheres.
- through radiation, for wavelengths below 100 nanometers, conditions reached by UV radiation, which impact the ionosphere of certain planets.
- through electron or proton bombardment, conditions that are notably found in polar auroras.
- through electrical discharge, as is the case in neon lamps.
- under extreme pressures, such as inside stars.

In neutral gases, interactions between particles, called Van der Waals interactions, occur over very short distances (the influence decreases as $1/r^7$ with r being the distance from the center of the relevant particle). In contrast, in plasmas, a particle is influenced by all other particles even at great distances (the influence decreases as $1/r^2$). It is then said that *collective effects* are dominant over binary collisions, which will only be considered here as minor perturbations. Thus, hot and dilute plasmas can be considered as non-collisional because the *mean free path* (average distance traveled by a particle between successive impacts) of charged particles is much greater than the time and distance scales considered. This is the case for the plasmas involved in the Sun-Earth interactions. It will suffice to consider the movements of the particles and their responses to the forces in which they evolve.

Without going too much into details, it is important to note a fundamental property of plasmas: they are considered electrically neutral on macroscopic scales (from a distance called the Debye length - see below). In other words, at such scales, the density of positive charge particles is equal to the density of negative charge particles. In the next three sections, we will present three different approaches to plasma physics and their limitations:

- The motion of a charged particle in an electromagnetic field (E, B).
- Kinetic theory
- Fluid theory or magnetohydrodynamics (MHD for short)

The choice of one approach over the other is fundamental and depends on the objectives and issues at hand. The fluid approach is more relevant for modeling global events, whereas the kinetic theory is relevant for describing radiative processes, for example.

Finally, let us recall some parameters of a plasma:

- Kinetic energy density: $E_c \approx nk_bT$
- Potential energy density: $E_p \approx ne^2/4\pi\epsilon_0d$, where d is the average distance between each particle
- Plasma parameter: $\Lambda = 4\pi n\lambda_D^3$, also equal to the ratio of potential energy density to kinetic energy density. Space plasmas all have a parameter $\Lambda \ll 1$, meaning their kinetic energy dominates over potential energy. Only metals have $\Lambda > 1$ due to strong electrostatic coupling between particles.
- Debye length: $\lambda_D = \sqrt{\frac{kT}{4\pi ne^2}}$, which can be interpreted as the radius of the sphere of influence of a charge on the quasi-neutrality of a plasma. To simplify, when a positive charge is placed in a sea of electrons, the electrons will clump around it and compensate for the positive charge. The Debye length is the distance from the positive charge at which the overall system appears neutral, i.e., where the positive charge is compensated.

1.1.2 Charged Particle Motions

1.1.2.1 Maxwell's Equations

A stationary charged particle creates an electric field \mathbf{E} around it. When it moves, it carries this electric field with it. However, a variation in the electric field in turn creates a magnetic field \mathbf{H} often approximated (rightly so) as \mathbf{B} . This idea is encapsulated in four fundamental equations of electromagnetism discovered by James Clerk Maxwell (1831 – 1879), aptly named the Maxwell equations. Without these equations, we cannot describe the motion of a charged particle in an electromagnetic field because they allow for the description of the electromagnetic field and its variations in a vacuum. They are as follows:

$$\text{Gauss's law for electric fields} \quad \nabla \cdot \mathbf{E} = \frac{\rho}{\epsilon_0} \quad (1.1)$$

$$\text{Magnetic monopoles} \quad \nabla \cdot \mathbf{B} = 0 \quad (1.2)$$

$$\text{Faraday's law of induction} \quad \nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \quad (1.3)$$

$$\text{Ampère's law with Maxwell's correction} \quad \nabla \times \mathbf{B} = \mu_0 \left(\mathbf{J} + \epsilon_0 \frac{\partial \mathbf{E}}{\partial t} \right) \quad (1.4)$$

The **Gauss' law** tells us that the divergence of the electric field directly comes from the volumetric electric charge density ρ in the considered medium. Simply put, the electric field comes from the electric charges and is divergent.

The **magnetic monopole law** (or local magnetic flux equation) tells us that magnetic field lines do not diverge: for a given surface immersed in a magnetic field, there are as many incoming field lines as outgoing ones. In fact, simply put, this equation tells us that there are no magnetic monopoles, that is, particles carrying a magnetic charge.

In the **Ampere-Maxwell law**, the right-hand side term is separated into two pieces. The first indicates that a charged particle in motion creates a magnetic field, as we explained at the beginning of this section. The second piece, on the other hand, tells us that even in the absence of a charged particle in motion, a variation in the electric field alone creates a magnetic field.

The **Faraday's law** can be seen from the opposite perspective. It tells us that a variation in a magnetic field generates an electric field.

1.1.2.2 The Cyclotron & Guiding Center Approximation

Let's consider a particle with mass m , charge q , and velocity \mathbf{v} , immersed in an electromagnetic field (\mathbf{E}, \mathbf{B}) . The particle then undergoes two forces: the Lorentz force (1.2.1) due to the electromagnetic field, and gravitational forces denoted F_g that we neglect here. We can now apply the fundamental principle of dynamics (FPD, or second law of Newton) stating that mass times acceleration is equal to the sum of forces (1.2.2). This equation is useful for describing the first interesting motion for us: the cyclotron.

$$\text{Lorentz force:} \quad \mathbf{F} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B}) \quad (1.5)$$

$$\text{FPD:} \quad m\mathbf{a} = q\mathbf{E} + q\mathbf{v} \times \mathbf{B} + \mathbf{F}_g \quad (1.6)$$

Let's take a reference frame $((\mathbf{e}_x), (\mathbf{e}_y), (\mathbf{e}_z))$ in which the electric field is zero and the magnetic field is only along (\mathbf{e}_z) , so $\mathbf{B} = (0, 0, B)$. We then project the FPD onto the three axes and obtain two coupled equations that need to be differentiated to solve.

$$\begin{cases} m\ddot{x} = qv_y B \\ m\ddot{y} = -qv_x B \end{cases} \quad \text{then} \quad \begin{cases} \ddot{x} = \omega_c^2 x \\ \ddot{y} = \omega_c^2 y \end{cases} \quad (1.7)$$

where

$$\omega_c = \frac{qB}{m} \quad (1.8)$$

is the *cyclotron frequency*, which depends on the charge. The radius of the circular motion or *Larmor radius* is

$$r_L = \frac{mv_{\perp}}{|q|B} \quad (1.9)$$

Thus, a positively charged particle will rotate clockwise, while a negatively charged particle will rotate counterclockwise. The motion can be more easily visualized directly with the Lorentz force. A positively charged particle launched on a horizontal plane with velocity \mathbf{v} and immersed in an upward-oriented magnetic field will tend to be pushed to the right (right-hand rule for the cross product or *right-handedness*). The particle will then have a constant speed v_{\parallel} along the \mathbf{B} axis and a circular velocity v_{\perp} perpendicular to the field. The resulting motion is shown in Figure 1.1.

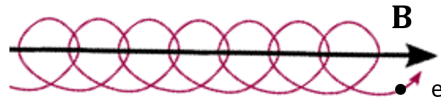


Figure 1.1 – Gyration motion of a particle along a magnetic field \mathbf{B} .

Let's keep in mind a very important approximation pointed out by Alfvén: the *guiding center approximation*, which states that the Larmor radius remains much smaller than the characteristic dimensions of the particle motion and that the field does not change much with the particle motion along the magnetic field during one gyro period (Koskinen and Kilpua, 2022).

1.1.2.3 Drift Motions in Electromagnetic Fields

We have seen the gyrotory motion of a particle in a magnetic field, which is at the basis of the functioning principle of cyclotrons, and also allows for the separation of isotopic charges (a phenomenon present in certain space instruments). However, when we add another uniform force field \mathbf{F} , or when the \mathbf{E} and \mathbf{B} fields vary (as shown by the Maxwell equations), the particle undergoes other drifts. Here are some examples:

- If the \mathbf{F} field is parallel to the \mathbf{B} field, the perpendicular motion is preserved and the force field modifies the parallel velocity, so the effect will be to vary the pitch of the helix. The particle completes turns following \mathbf{B} and covers for each turn a distance that increases along the axis.
- If the \mathbf{F} field is perpendicular to the \mathbf{B} field, the drift occurs in a direction perpendicular to \mathbf{F} and \mathbf{B} . The general motion is therefore a gyrotory motion around the guide center, plus a drift at constant velocity, given by:

$$\mathbf{v}_D = \frac{\mathbf{F} \times \mathbf{B}}{qB^2} \quad (1.10)$$

From this formula, we see that if the force \mathbf{F} does not depend on the charge, a current is generated. An important particular case for us is the presence of an electric field $\mathbf{F} = q\mathbf{E}$. As we have just seen, such a field creates a drift in a direction perpendicular to \mathbf{E} and \mathbf{B} . The drift velocity is the same as above, but the charges cancel out, and we obtain $\mathbf{v}_E = (\mathbf{E} \wedge \mathbf{B})/B^2$. This motion is at the origin of a global convection of the plasma. Moreover, since the kinetic energy is conserved over one gyrotory motion, a parallel electric field is necessary for acceleration.

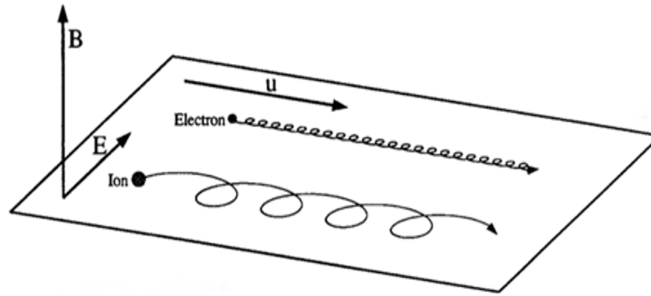


Figure 1.2 – Motion of a positive ion and an electron in a field (\mathbf{E}, \mathbf{B}) from [Kivelson and Russell \(1995\)](#). Accelerated by the \mathbf{E} field, the gyroradius is larger in the direction of \mathbf{E} , causing a drift \mathbf{u} perpendicular to both \mathbf{B} and \mathbf{E} .

- Finally, particles also encounter spatial variations in the magnetic field \mathbf{B} . These also give rise to drifts:
 - A first drift that deserves to be mentioned is the centrifugal force resulting from a particle following a curved field line. The resulting velocity is as follows:

$$\mathbf{v}_c = \frac{mv_{\parallel}^2}{2qB} \frac{\mathbf{n} \times \nabla \mathbf{B}}{R_c}. \quad (1.11)$$

- If \mathbf{B} varies along an axis perpendicular to the field lines, a drift velocity appears according to the formula:

$$\mathbf{v}_g = \frac{1}{2}mv_{\perp}^2 \frac{\mathbf{B} \times \nabla \mathbf{B}}{qB^3}, \quad (1.12)$$

which is actually the same formula seen earlier. Therefore, a drift occurs perpendicular to both \mathbf{B} and $\nabla \mathbf{B}$. This formula is only valid under certain conditions of the ratio between the value of the magnetic field B and its variation, which we will detail a little more in the next section. It will be noted that the drift velocity depends on the charge of the particle.

- If \mathbf{B} varies along an axis parallel to the field lines, the particle then experiences a force opposite to the magnetic gradient, $\mathbf{F} = -\mu \nabla_{\parallel} \mathbf{B}$. One of the consequences of

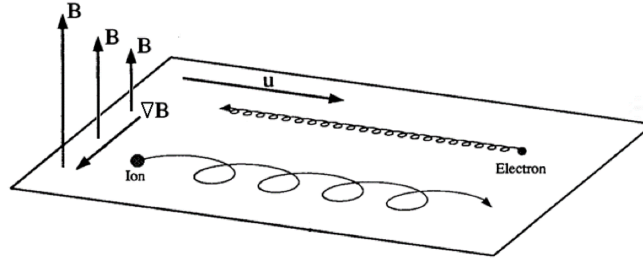


Figure 1.3 – Movement of a positive ion and an electron in a non-uniform magnetic field \mathbf{B} in the plane, with a variation of the gyration radius resulting in a drift \mathbf{u} for positively charged particles. Inspired by [Kivelson and Russell \(1995\)](#).

this physics on particles in the radiation belts is called the *mirror effect*. This effect is simply explained by the fact that a strong enough magnetic field slows down a particle to the point where the parallel velocity module v_{\parallel} becomes zero at a point called the *mirror point*, from which the particle would start moving in the opposite direction. Not all particles can be trapped and the next section (1.1.2.4) explains this effect as well as the first adiabatic invariant μ .

In summary, once the gyration motion is applied, three drifts are generally sufficient to describe the orbital motion of particles alone in electromagnetic fields:

- The $\mathbf{E} \times \mathbf{B}$ drift, also called \mathbf{v}_D above.
- The *gradient drift*: drift due to magnetic field gradients, called \mathbf{v}_g above. This effect is particularly observable when ions (resp. electrons) from the magnetotail heading towards the Earth drift westward (resp. eastward) - see Figure 1.3.
- The *curvature drift*: drift due to the centrifugal force experienced by the particle in its gyration, called \mathbf{v}_c above.

$$\mathbf{v}_D = \frac{\mathbf{F} \wedge \mathbf{B}}{qB^2} \quad (1.13)$$

$$\mathbf{v}_g = \frac{1}{2}mv_{\perp}^2 \frac{\mathbf{B} \wedge \nabla \mathbf{B}}{qB^3} \quad (1.14)$$

$$\mathbf{v}_c = \frac{mv_{\parallel}^2}{2qB} \frac{(\hat{n} \wedge \nabla \mathbf{B})}{R_c} \quad (1.15)$$

1.1.2.4 Adiabatic Invariants

The radiation belts (see Section 1.2.3.8) are made of charged particles that are *trapped* in the inner magnetosphere, following quasi-periodic motions. We usually only describe three of these motions through what we call *adiabatic invariants*. But before diving into the subject and the resulting motions that are of interest to us, we need to quickly jump by the Hamiltonian mechanics. To better understand the concepts and mathematics behind, I highly recommend the reader to read chapters 3.2 to 3.4 of [Bellan \(2006\)](#) where a proper proof is given to the following information.

If a dynamical system has an equation of motion of the form of

$$\frac{d^2x}{dt^2} + \omega^2(t)x = 0 \quad (1.16)$$

then S , the *action integral* over one period, is a constant of the motion (Bellan, 2006).

$$S = \int_{t_0}^{t_0+\tau} L dt \quad \text{with } L \text{ being the Lagrangian} \quad (1.17)$$

Even with a time-dependent ω , one element is conserved if the movement is slow enough: the action integral. This can be extended to general Hamiltonian systems with a general form of the action integral (Eq.1.18) where P, Q are the relevant canonical momentum-coordinate conjugate pair. The action integral is an adiabatic invariant, i.e. a conserved quantity.

$$S = \oint P dQ \quad (1.18)$$

Approximations that should be made here are called *adiabatic approximations* and then depends on very slow variations of the system parameters (which explains the term "adiabatic"). In our context:

- The ratio of the temporal variation of the magnetic field to the magnetic field itself must be much smaller than the gyration period. Mathematically: $\frac{\partial \mathbf{B}}{\mathbf{B} \partial t} \ll \omega_c$.
- The ratio of the spatial variation of the magnetic field to the magnetic field itself must be much smaller than the Larmor radius. Mathematically: $\frac{\partial \mathbf{B}}{\mathbf{B} \partial r} \ll R_L$.

First Adiabatic Invariant & Mirror Effect

In the frame of reference where $v_{\parallel} = 0$, the charged particle motion creates a current I along its circular path, with an associated magnetic moment $\mu = I\pi r_L^2$.

$$\mu = I\pi r_L^2 = \frac{q\omega_c}{2\pi} \pi \frac{v_{\perp}^2}{\omega_c^2} = \frac{mv_{\perp}^2}{2B} \quad (1.19)$$

The magnetic moment tends to create a magnetic field opposed to the background magnetic field, hence weakening it (Koskinen and Kilpua, 2022) and it is our first adiabatic invariant.

If, and only if, the adiabatic approximations are satisfied (in other words, if \mathbf{B} varies slowly enough), then the ratio $\mu = mv_{\perp}^2/2B$ is constant. To show its invariance, we can start from the action integral or from the projected equation of motion along the axis parallel to \mathbf{B} in the case of a varying magnetic field. The only force in the parallel axis is thus due to the variation of \mathbf{B} .

$$m\mathbf{a}_{\parallel} = -\mu \nabla_{\parallel} \mathbf{B} \quad (1.20)$$

By multiplying both sides by the parallel velocity, we obtain on one side the derivative of $mv_{\parallel}^2/2$ and on the other side a gradient not in space but in time.

$$m \left(v_{\parallel} \cdot \frac{dv_{\parallel}}{dt} \right) = -\mu \frac{dz}{dt} \frac{\partial B}{\partial z} \iff \frac{d}{dt} \left(\frac{1}{2} mv_{\parallel}^2 \right) + \mu \frac{dB}{dt} = 0 \iff \frac{d}{dt} \left(\frac{1}{2} mv_{\parallel}^2 + \mu B \right) = B \frac{d\mu}{dt}$$

By conservation of the total kinetic energy, $\frac{d}{dt} \left(\frac{1}{2} mv_{\parallel}^2 + \frac{1}{2} mv_{\perp}^2 \right) = 0$, i.e., $\frac{d}{dt} \left(\frac{1}{2} mv_{\parallel}^2 + \mu B \right) = 0$. Thus, we have $B d\mu/dt = 0$, which implies that μ is constant.

Two consequences directly follow the time invariance of μ .

- First, if μ remains constant, then the magnetic flux through the surface of a gyration must remain the same, which means that if B increases, then the radius of the orbit decreases.

- Second, we can deduce a criterion for a *mirror point*. With α being the angle of attack of the particle, $v_{\perp} = v \sin \alpha$ and then we have:

$$\mu = \frac{mv^2 \sin^2 \alpha}{2B} \quad (1.21)$$

If μ and the kinetic energy are conserved, then the only two parameters that can vary are α and B . A particle following an increasing magnetic field will then see its angle of attack increase. Hence, it might exist a value of the magnetic field B_m for which $\sin \alpha = 1$. As a consequence, the particle will stop moving along B and will turn back. This point is called the *mirror point*.

In line with what we just said, we can explain how a particle can be trapped in a *magnetic bottle*. Let's take the following equality for two different locations of the particle in a varying magnetic field by taking $\mu_1 = \mu_2$:

$$\frac{\sin^2 \alpha_1}{\sin^2 \alpha_2} = \frac{B_1}{B_2} \quad (1.22)$$

The direct consequence of this equality is that, for our magnetic field maximum value B_{max} and for a B_0 being the minimum of the magnetic field (at the center of the "bottle"), it exists a limit angle α_0 at point 0 such that the mirror point will be located where $B = B_{max}$ (at both ends of the bottle). If the angle of attack at the center of the bottle is smaller than the mirror point will happen for a $B_m > B_{max}$ and the particle will not be trapped. Our trapping criterion then corresponds to the particle being outside a cone called the *loss cone* that can be imagined at the injection level. A particle injected into this cone would not experience a mirror point and then will be lost. In our magnetosphere, the magnetic bottle could be thought of as being a magnetic field line connecting North and South poles. A particle would then be lost if it reaches an altitude where it collides with atmospheric particles.

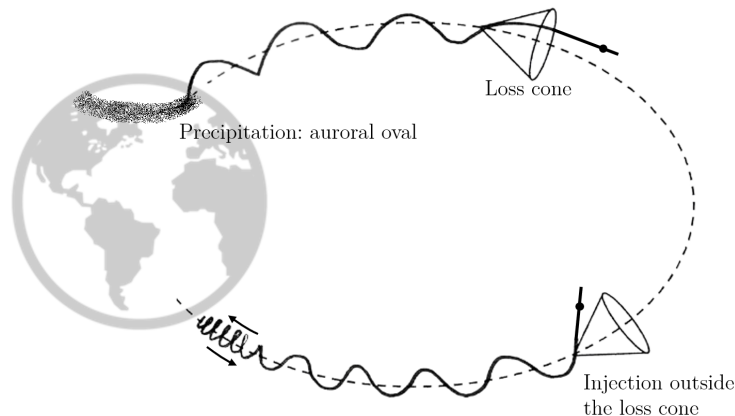


Figure 1.4 – Mirror Effect and Loss Cone Near Earth.

Second Adiabatic Invariant

A trapped particle moving along a magnetic field line experiences the deceleration described earlier and eventually reaches its mirror point, after which it bounces back towards the opposite pole. Assuming conservation of kinetic energy (and hence speed), we observe a nearly periodic motion between two mirror points with a bounce period denoted as τ_b . By employing canonical coordinates, considering momentum, and under the assumption of quasi-periodic motion in the adiabatic approximation, we can establish an adiabatic invariant (initially proven by [Northrop and](#)

Teller (1960)). For the canonical momentum reduced to $p_{\parallel} = mv_{\parallel}$, and the canonical position represented by the projected position of the particle on the magnetic field line (with s denoting the arc length), the second adiabatic invariant can be expressed as follows:

$$J = m \oint v_{\parallel} ds \quad (1.23)$$

where the integral is performed along a field line. It is called the bounce or longitudinal-invariant.

This second adiabatic invariant is associated with the periodic *bouncing motion* of a particle trapped between two mirror points on a magnetic field line. This invariant holds significant importance in understanding particle dynamics. Due to the impact of the solar wind on the magnetosphere, the latter undergoes distortions, acquiring a comet-like shape and losing its axisymmetry. Consequently, it is reasonable to expect that a particle, while orbiting Earth, will transition from one magnetic field line to another. However, all magnetic field line at a specific azimuthal angle correspond to a distinct value of J for a particle with a given energy and momentum. Consequently, the conservation of J prevents the radial movement of particles in or out of the radiation belts as they drift around Earth (Fitzpatrick, 2011).

Third Adiabatic Invariant

We have presented the adiabatic invariants associated with gyro-motion around the field line and the bouncing motion along field lines. The final adiabatic invariant is related to the quasi-periodic precession of particles around Earth and can be expressed as follows:

$$\Phi = \oint \mathbf{A} d\mathbf{l} \quad (1.24)$$

where the integration path represents the trajectory of the middle of the flux tube (at the equator) around Earth, known as the drift path. Here, \mathbf{A} represents the vector potential of the field. It is called the flux invariant or the L-shell. As explained by Koskinen and Kilpua (2022), this invariant is weaker than the first two invariants, as smaller changes can disrupt its invariance. This is primarily because the drift period τ_d should be much larger than both τ_b and τ_L . Considering that the drift period is approximately one hour (for MeV energy protons and electrons), this condition holds true only when the magnetosphere is relatively inactive. A notable consequence of this is the radially inward (outward) motion of the radiation belts in the case of increasing (decreasing) solar wind intensity, assuming that the variation in solar wind intensity occurs on timescales greater than τ_d .

1.1.2.5 Summary of Motions in the Radiation Belts

At this point, we have already mentioned several times the existence of radiation belts, but these are structures that we will revisit a little later in this chapter, during the journey from Earth to the Sun. Here, let's summarize the three different motions of particles in the belts that results from everything said above:

- The guiding center approximation can be applied if r_L is very small in front of the characteristic dimensions. For us here, it means that $r_L \ll R_C$ with R_C the curvature of the magnetic field lines 1.1.2.2. A charged particle has a gyro motion of period τ_L around its center, the axis of the field line. The magnetic moment μ is invariant.
- A particle has an equatorial pitch angle α_{eq} corresponding to a mirror point on the field line. Depending on this angle, the particle is or is not in a loss cone. If it is not, it experiments a bounce motion between the two mirror points (see Section 1.1.2.4), of period τ_b . There is a longitudinal / bounce invariant J if characteristic times are very large compared to τ_b , and $\tau_b \gg \tau_L$.

- As it moves from the equator towards a pole, the particle experiences an increasing $\|B\|$ and a curvature of the magnetic field lines, leading to both a gradient and a curvature drift around the Earth, with electrons drifting to the east and protons drifting to the west, with a period τ_d . This motion is associated with a L-shell invariant stating that the magnetic flux enclosed by the drifting particle is invariant. It is invariant provided that the characteristic times are very large compared to τ_d , and $\tau_d \gg \tau_b \gg \tau_L$.
- It is interesting to notice that low-energy particles motions are dominated by the $\mathbf{E} \times \mathbf{B}$ drift (Russell et al., 2016).

This set of phenomena is what gives rise to the Van Allen belts, the plasmasphere, and the ring current. These structures will be more detailed in Section 1.2.3.8. A summary of these movements can be seen in Figure 1.5.

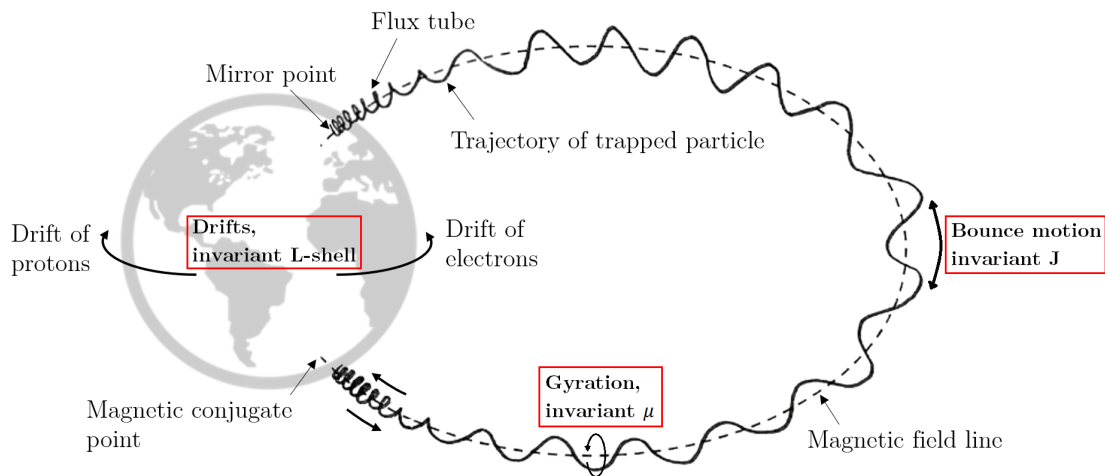


Figure 1.5 – Summary of basic motions of trapped particles in the Earth's magnetic field.

Finally, two accelerations are worth mentioning:

- The betatron accelerations can be decomposed into two phenomenon: *gyro betatron acceleration* and *drift betatron acceleration*. The gyro betatron acceleration occurs when the magnetic field strength gradually increases over time, relative to the gyroperiod, resulting in an increase in the perpendicular energy of the particle. This is due to the fact that the kinetic energy of the particle is not constant due to the presence of electric fields associated with the time-varying magnetic field but μ remains constant due to slow gradual increase.
- Fermi acceleration is a special case of betatron acceleration where the drift of particles bring them where the mirror field increases. This leads to the mirror point getting closer to each other, which is compensated by an increase in parallel energy of the particle.

1.1.3 Kinetic theory: Vlasov equation

The movements of charged particles allow us to understand many phenomena, but when we look at a plasma at larger scales, it is not possible to solve propagation problems by focusing on each individual particle. This is where the kinetic description of the plasma comes into play.

The idea is simple: instead of looking at each individual particle independently, we say that the number of particles located in a 6-dimensional space (3 dimensions for position and 3 dimensions for velocity) follows a distribution law $f_\alpha(\mathbf{r}, \mathbf{v}, \mathbf{t})d^3\mathbf{r}d^3\mathbf{v}$. Thus, the quantity $f_\alpha(\mathbf{r}, \mathbf{v}, \mathbf{t})$ is the particle distribution function. Depending on the problem and constraints, this distribution function can take on different forms, one of the best known and most used being the Maxwell distribution in the case of independent particles at thermal equilibrium. Furthermore, by integrating over velocity space, it is possible to identify moments:

- The moment of order 0 which is the particle density: $n_\alpha(\mathbf{r}, t) = \int f_\alpha(\mathbf{r}, \mathbf{v}, t) d^3\mathbf{v}$
- The moment of order 1 which is the particle flux: $n_\alpha(\mathbf{r}, t)(\mathbf{u}_\alpha(\mathbf{r}, t)) = \int \mathbf{v} f_\alpha(\mathbf{r}, \mathbf{v}, t) d^3\mathbf{v}$
- And so on to obtain the moment of order 2 (the stress tensor) or the moment of order 3 (the heat flux tensor).

To describe the time evolution of the distribution function, for simplicity, we can consider a two-dimensional space with $f(t)$ in the form of $f(x, v, t) dv dx$. We will not go into the details of the calculation here, but by taking the difference between the number of incoming and outgoing particles from a cell of our 2D space at time t and the number of incoming and outgoing particles at time $t + dt$, we obtain the following equation:

$$\frac{\partial f}{\partial t} = -v \frac{\partial f}{\partial x} - \frac{\partial (af)}{\partial v} \quad (1.25)$$

This can be generalized to 3D space:

$$\frac{\partial f}{\partial t} + \mathbf{v} \cdot \frac{\partial f}{\partial \mathbf{r}} + \frac{\partial (\mathbf{f} \cdot \mathbf{a})}{\partial \mathbf{v}} = 0 \quad (1.26)$$

The acceleration \mathbf{a} is determined by external forces via the PFD. If we only consider an electromagnetic field (\mathbf{E}, \mathbf{B}), we only consider the Lorentz force, $\mathbf{a} = \frac{q}{m}(\mathbf{E} + \mathbf{v} \times \mathbf{B})$. Thus, we see that each component of the acceleration (a_x, a_y, a_z) will depend only on v_y and v_z , v_z and v_x , v_x and v_y respectively by the principle of the cross product. We can therefore remove the acceleration from the derivative. This gives us the Vlasov equation:

$$\frac{\partial f}{\partial t} + \mathbf{v} \cdot \frac{\partial f}{\partial \mathbf{r}} + \frac{q}{m}(\mathbf{E} + \mathbf{v} \times \mathbf{B}) \cdot \frac{\partial f}{\partial \mathbf{v}} = 0 \quad (1.27)$$

Although it seems quite simple, it is actually one of the most difficult equations to solve. The Vlasov equation corresponds to the Boltzmann equation without the collision terms. In conclusion, the Vlasov equation here describes the time evolution of the distribution function of particles of mass m and charge q in a plasma, neglecting collisions between particles.

1.1.4 Fluid theory: Magnetohydrodynamics

Magnetohydrodynamics is a branch that aims to combine vacuum electromagnetism and the fluid dynamics equations known as the Navier-Stokes equations. It consists of assimilating a current-carrying conductor medium (such as a plasma) to a fluid. For the study of plasma, we therefore consider larger scales and treat plasma as a continuous medium. It should be recalled that the objective of this chapter is to understand the physical foundations that will allow us to understand the Sun-Earth chain without equations. Several concepts are presented, such as plasma frozen in magnetic fields, convection, magnetic pressure, and tension.

To describe the plasma in this theory, it must always remain close to electroneutrality, a condition that was already mentioned in part 1.2. It is also necessary to define macroscopic quantities that describe the plasma: a velocity V , a temperature T , a pressure P , and a density ρ . Finally, we will use Ohm's law, which gives us a relationship between the current density \mathbf{j} and the electric field \mathbf{E} .

$$\mathbf{j} = \frac{1}{\eta}(\mathbf{E} + \mathbf{V} \times \mathbf{B}) \quad (1.28)$$

Where η is the electrical resistivity.

1.1.4.1 MHD's equations

Reminder of the conditions: we assume that the plasma consists of a single fluid. The assumed variables are ρ for density, \mathbf{V} for the fluid's overall velocity, \mathbf{B} for its magnetic field, P for pressure, its current density \mathbf{j} , and its electric field \mathbf{E} . We will construct the induction equation from the two Maxwell's curl equations and Ohm's law. Macroscopic movements are much slower than microscopic movements.

Equations:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{V}) = 0 \quad \text{Continuity equation} \quad (1.29)$$

$$\rho \frac{\partial \mathbf{V}}{\partial t} + \rho (\mathbf{V} \cdot \nabla) \mathbf{V} = -\nabla P + \mathbf{j} \times \mathbf{B} + \rho \mathbf{F}_g \quad \text{Momentum conservation in a fluid} \quad (1.30)$$

$$\frac{\partial \mathbf{B}}{\partial t} = \nabla \times (\mathbf{V} \times \mathbf{B}) + \frac{\eta}{\mu_0} \nabla^2 \mathbf{B} \quad \text{Induction equation} \quad (1.31)$$

$$\frac{P}{\rho^\gamma} = \text{constant} \quad \text{Thermodynamic equation of state} \quad (1.32)$$

The momentum conservation equation in magnetohydrodynamics corresponds to the equation of motion, which is the direct implication of the Navier-Stokes equation. \mathbf{F}_g represents different gravitational forces.

The induction equation informs us that if the fluid is at rest, then

$$\frac{\partial \mathbf{B}}{\partial t} = \frac{\eta}{\mu_0} \Delta \mathbf{B} \quad (1.33)$$

This is a diffusion equation, so the magnetic field decreases in a uniform sphere. The first term, on the other hand, is a convection term. Thus we see two opposing terms (diffusion and convection). In fact, this equation allows us to highlight what is called the magnetic Reynolds number \mathcal{R}_m , which is the ratio of the magnitudes of these two terms. If $\mathcal{R}_m \ll 1$, the diffusion term dominates, and collisions responsible for the plasma's resistivity dissipate magnetic energy. If $\mathcal{R}_m \gg 1$, the dominant term is convection, and we are in the ideal MHD case.

1.1.4.2 Magnetic field lines "frozen" in the plasma

The easiest approach is the so-called ideal approach, in which $\mathcal{R}_m \gg 1$. It assumes a collisionless moving plasma (which is often a good first approximation for space plasmas (Kivelson and Russell, 1995)), which gives then a conductivity σ so large that we can consider it infinite (or a nul resistivity $\eta = 0$). The induction equation becomes:

$$\frac{\partial \mathbf{B}}{\partial t} = \nabla \times (\mathbf{V} \times \mathbf{B}) \quad (1.34)$$

This also means that Ohm's law becomes:

$$\mathbf{E} + \mathbf{V} \times \mathbf{B} = 0 \quad (1.35)$$

This last Equation 1.35 shows that the electric field is exclusively written as a convective term. We can deduce a property from this: the conservation of magnetic flux, which also implies the notion of plasma and field being "frozen" together.

The conservation of magnetic flux can be easily shown by calculation. Consider a volume of plasma with velocity \mathbf{v} moving in a magnetic field \mathbf{B} . We can then write the temporal variations of magnetic flux as:

$$\frac{\partial \phi}{\partial t} = \iint \frac{\partial \mathbf{B}}{\partial t} \cdot d\mathbf{S} + \oint \mathbf{v} \times \mathbf{B} \cdot d\mathbf{l} \quad (1.36)$$

According to Stokes' theorem, we can also write this equation as:

$$\frac{\partial \phi}{\partial t} = \iint \frac{\partial \mathbf{B}}{\partial t} \cdot d\mathbf{S} - \iint \nabla \times (\mathbf{v} \times \mathbf{B}) \cdot d\mathbf{S} \quad (1.37)$$

And according to the equations resulting from ideal MHD seen above, we arrive at:

$$\frac{\partial \phi}{\partial t} = 0 \quad (1.38)$$

The temporal variation of the flux is zero, therefore the flux is constant. This means that a flux (\mathbf{B}_1) through a surface S_1 will become a higher flux (\mathbf{B}_2) if the surface narrows into a smaller surface S_2 , a phenomenon observable in the solar wind.

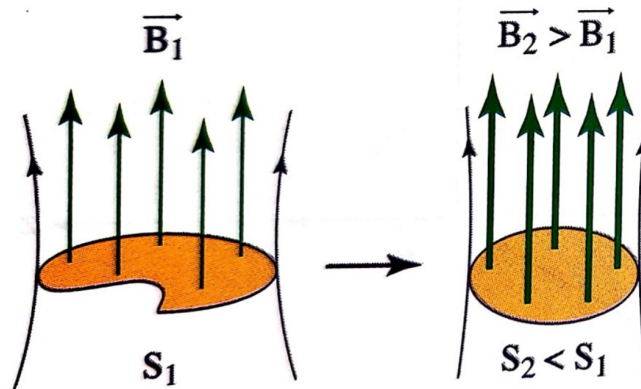


Figure 1.6 – Magnetic flux conservation in a field line from [Lilensten and Bornarel \(2001\)](#).

Thus, the plasma is "frozen" in the field lines, but the field is also "frozen" in the plasma. The particles have their trajectories modified by the fields, and in turn create fields themselves, it is the *frozen-in field condition*. Therefore, all the particles on a given field line will remain on that field line. So, if we have a magnetized plasma that moves, it carries the magnetic field with it. In this case, the line can deform depending on the different movements of the plasma. But if a plasma is not magnetized and encounters a magnetic field of different origin, there will be no mixing and the particles will simply push the encountered magnetic field (Figure 1.7). This is the phenomenon we encounter when particles magnetized by the interplanetary magnetic field encounter the Earth's magnetic field.

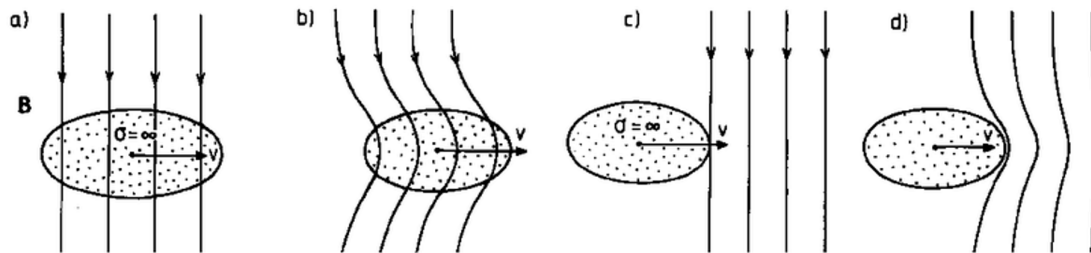


Figure 1.7 – Plasma and frozen-in condition. In (a) and (b), we can see the deformation of field lines caused by the plasma motion. In (c) and (d), we can see the plasma pushing against a flux tube that it cannot cross ([Brekke, 1997](#)).

1.1.4.3 Magnetic pressure and magnetic tension

Without going into detail on the calculations, the magnetic term of the equation of motion can be expressed as:

$$\mathbf{j} \times \mathbf{B} = -\nabla_{\perp} \left(\frac{B^2}{2\mu_0} \right) + \frac{B^2}{\mu_0 r_c} \mathbf{n}_B \quad (1.39)$$

In this new form, we can identify a first term which resembles a pressure gradient term where the magnetic pressure P_m is expressed as: $P_m = \frac{B^2}{2\mu_0}$. The second term on the right-hand side contains r_c , which is the radius of curvature of the magnetic field line, and resembles a tension term like that of a stretched elastic. This notion has interesting consequences for the dynamics of the magnetospheric tail, which we will discuss later. We define the plasma beta parameter as the ratio of thermal and magnetic pressures. Thus, plasmas considered "cold" (low β) are dominated by magnetic forces, while "hot" plasmas (high β) are controlled by thermal effects, and magnetic fields are induced by plasma motions.

1.2 The Sun-Earth Chain

1.2.1 Starting point: the Sun

1.2.1.1 Structure and characteristics

In this section, we will review the characteristics of the Sun. For the sake of clarity, we will try to break down the different zones of the Sun and its surroundings into discernible pieces as it has been done in the book [Lilensten and Blelly \(2008\)](#). However, it should be kept in mind that it is practically impossible to fully describe solar physics this way. The purpose of this thesis is not to delve into the technical details of this subject, but only to provide the reader with the keys to understanding how our star works.

Some key features of our star include:

- Equatorial diameter: 1,392,000 km, approximately 109 times that of Earth.
- Mass: 1.99×10^{30} kg, which is 99.97% of the total mass of the solar system, with 50 to 70% located in the first quarter of the solar radius known as the nuclear core.
- Average density: 1.4×10^3 kg.m⁻³, which is a quarter of that of Earth.
- Total radiated power: 4×10^{26} W, of which Earth receives about 1.743×10^{17} W.
- Components and their percentages ([Lilensten and Blelly, 2008](#)):
Hydrogen (93.96%), Helium (5.919%), Oxygen (0.0648%), Carbon (0.0395%), Nitrogen (0.0082%), Silicon (0.0042%), Magnesium (0.0037%), Neon (0.0035%), Iron (0.0030%), Sulfur (0.0015%), Aluminum (0.0003%), Calcium (0.0002%), Sodium (0.0002%), Nickel (0.0002%), Argon (0.0001%).

1.2.1.2 A Dynamic Activity

The Sun is a hot ball of plasma and a magnetic star with highly dynamic activity. It is tilted at 7.25° to the ecliptic plane and exhibits a differential rotation, which means that the angular velocity decreases with increasing latitude. In fact, the Sun's equatorial regions rotate faster, taking only about 24 days, compared to the polar regions, which rotate once in more than 30 days.

Turbulent motions in the rotating, electrically conducting convective zone generate a chaotic interior dynamo and produce the solar magnetic field. Solar magnetic activity essentially controls the entire outer solar atmosphere, heating the coronal gas to millions of degrees, producing flares

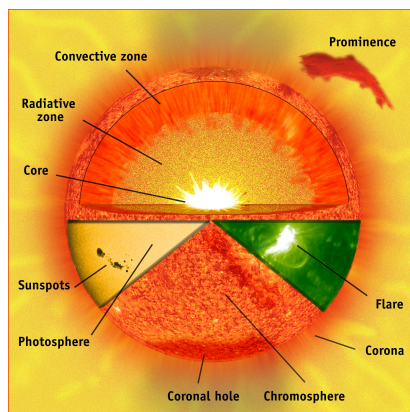


Figure 1.8 – Sun parts from Steele Hill / NASA. Courtesy of SOHO consortium. SOHO is a project of international cooperation between ESA and NASA.

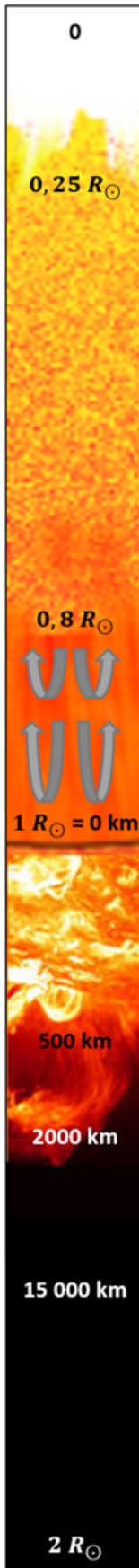
that interact with planetary atmospheres, guiding the solar wind, and protecting our solar system from cosmic rays (Güdel, 2007). This tumultuous magnetic activity follows an 11-year cycle, as evidenced by the number of dark spots on the photosphere (see Figure 1.8), which have been measured for centuries and are known as sunspots.

Solar Magnetic Activity

Solar activity is historically modeled by the presence of these solar spots. These dark regions on the photosphere, with diameters reaching hundreds of thousands of kilometers, are mainly localized between 40° latitude north and 40° latitude south. They contain a magnetic field that is 100 to 5000 times more intense than elsewhere (Lilensten and Bornarel, 2001). Additionally, their lower temperature (1000– 1900 K cooler than the quiet Sun - Solanki (2003)) explains their dark appearance. The formation of these spots falls slightly outside the scope of this thesis, but we can provide a brief explanation inspired by Lilensten and Bornarel (2001).

As mentioned, the Sun's magnetic field originates from the convective zone, which is not at its center. The Sun's differential rotation distorts the magnetic field lines and complexifies the configuration of the solar magnetic field, causing locally intense magnetic field tubes that are perpendicular to the surface. In this zone, matter carried by the field lines hinders heat transfer from neighboring regions, resulting in gas cooling and the appearance of the spot. Subsequently, filaments may form above the spot: dark, long, and narrow meandering features that rise towards the chromosphere.

Solar spots have been observed since the 11th century, but it was not until the 17th century that their numbers were observed with the help of the first telescopes. It was then noticed that between two periods without solar spots, approximately 10 to 13 years would pass, and the Sun would return to its initial state. This is known as the "Schwabe cycle," which lasts about 11 years. Around the 1850s, Rudolf Wolf had the idea of computing the amount of sunspots that is still used today. The formula is the following $R = k_{corr}(10G + T)$ with G the number of groups, T the number of spots and k_{corr} an observatory factor that depends on the instrument used to account for different methods and instruments observing. Following, the very first butterfly diagram was plotted in 1886 (Russell et al., 2016). A butterfly diagram represents the spots or groups of spots as a function of solar latitude. Figure 1.9(b) shows that, over the course of the solar cycle, these spots tend to migrate towards the equator. Figure 1.9(a) also illustrates the solar cycle, depicted by the number of visible solar spots over time. Although the measurement of solar activity through sunspots is debatable due to the significant improvement in observation instruments, it remains one of the longest available data series. As shown in Figure 1.9(a), over the long term, the cy-



Nuclear core: As previously mentioned, the nuclear core contains 50 to 70% of the total mass of the sun, extending from the center to $0.25 R_{\odot}$ with R_{\odot} the radius of the Sun (approx. 696,340 km). Hydrogen is transformed into helium through nuclear fusion, following the proton-proton cycle. The temperature at its core is around 15.6 million Kelvin, and the pressure is 2.2×10^{11} times the average pressure at the surface of the Earth (Lilensten and Blelly, 2008)

Radiative zone: The zone extending from $0.3 R_{\odot}$ to $0.8 R_{\odot}$ is called the radiative zone. From its base to its summit, the pressure and temperature decrease. The pressure drops from a few tens of billions of Earth's atmosphere to about 6 million, while the temperature decreases from a few tens to slightly over one million Kelvin. It takes several million years for the photons emerging from the nuclear reactions in the core to traverse this zone. As they collide with solar matter, they experience a spectral spread up to the domain of X-rays, which explains the strong UV radiation and white light (Lilensten and Blelly, 2008). Note that, unlike the convective zone, the radiative zone rotates uniformly.

Convective zone: The convective zone is a turbulent zone, with differential rotation, composed of ionized matter in convective motion: hot protons and electrons rise to the surface where they cool before descending. This creates convection cells visible from the surface as granulations. They come in varying sizes, but separated into two distinct categories: granules, with a diameter of thousands of kilometers and a lifetime of approx. 5 to 10 minutes, and supergranules, with a diameter around tens of thousands of kilometers and a lifetime of nearly 20 hours (Lilensten and Bornarel, 2001; Russell et al., 2016).

Photosphere: Approx. 500 km thick starting at R_{\odot} , the temperature in the photosphere drops from 6000 to around 4000 K. Various structures can be found here, such as spicules and macropicules, which emerge respectively between granules and supergranules. They expel solar matter and are responsible for what is called slow solar wind.

Chromosphere: Between approximately 500 and 2000 km high lies the chromosphere. Here, the temperature increases from 4200 to nearly 10,000 Kelvin.

Transition region: Initially, the temperature remains relatively low. Then, around 3000 km, it increases suddenly to several million degrees K. It then gradually continues to increase up to 15,000 km altitude, at the level of the corona.

Corona: At this distance, we reach the corona, where we can observe less dense and therefore colder areas called coronal holes. Plumes emerge from these areas, which are responsible for the emission of what is called fast solar wind. From this zone, we are in the high solar atmosphere and are then touching what is called the interplanetary medium, which the sun has filled with solar wind (fast, slow, and even explosive), radiation (EUV, gamma, and others), and energetic particles (protons, electrons, alpha particles). All of these are a result of the dynamics of our star, which will be the subject of next sections.

cles have varying durations, maximum spot counts, periods of activity increase and decrease, and different lengths. Furthermore, during each solar minimum, the polarities between the Northern and Southern hemispheres reverse, which actually results in a 22-year solar cycle (the Hale cycle), contrary to the commonly mentioned 11 years (Hathaway, 2015).

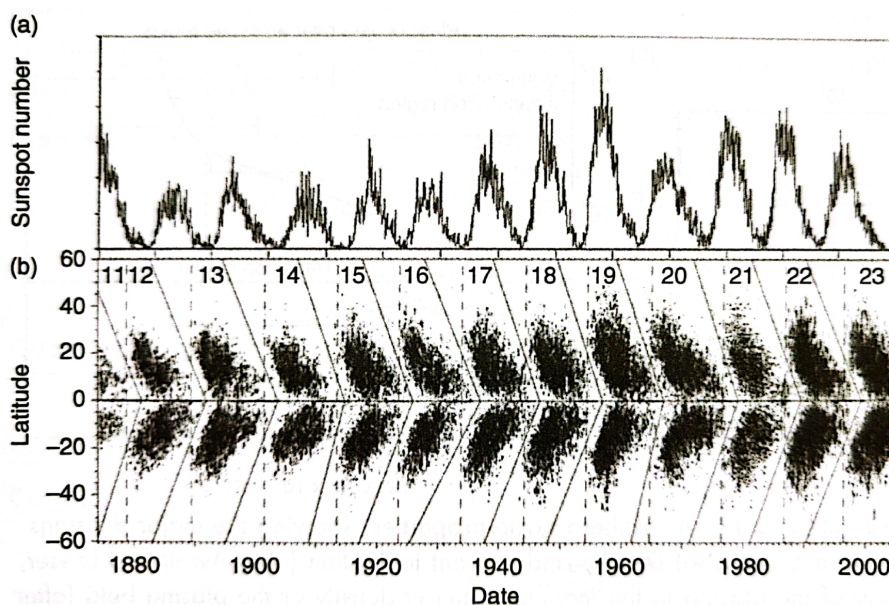


Figure 1.9 – (a) International Sunspot Number (ISSN); (b) butterfly diagram : latitudes of sunspots as a function of time; (Russell et al., 2016).

From Coronal Holes to CMEs

The solar corona is the outermost part of the Sun and is directly linked to the solar wind and the interplanetary magnetic field. It should be noted that to reach this region, the temperature abruptly increases from 10,000 to 100,000 degrees and then suddenly rises to several million degrees. In the solar corona, the solar atmosphere becomes rarified, and the movements of fully ionized constituents are constrained by the magnetic field lines (see Section 1.1.4.2). Thus, the corona is a relatively collisionless medium, and the intense heating in this region appears to result from the generation and absorption of various types of waves, such as mechanical oscillations at its base (Russell et al., 2016).

Observations of the Sun in X-rays or extreme ultraviolet (EUV) reveal dark and cool regions of varying sizes called coronal holes, which can cover up to one-third of the solar surface. They are considered regions of low density where the magnetic field lines are open to space. Because of this density, they are considered to behave as a collisionless plasma. Ionized protons and electrons escape along the open magnetic fields and form the starting point of fast solar winds, which will be discussed in Section 1.2.2. According to Cranmer (2009), this gives us three definitions (the darkest patches seen in UV and X-ray; the lowest-intensity regions observed with a coronagraph; or open-field footpoints of time-steady solar wind flows) which do not completely overlap. For more information on coronal holes, we encourage the reader to look at the vast number of existing reviews on the topic such as Cranmer (2002, 2009); Harvey and Sheeley (1979); Hudson (2002); Jones (2005); Kohl and Cranmer (1999); Ofman (2005); Parker (1991); Suess (1979); Toma and Arge (2005); Wang (2009); Zirker (1977).

During periods of low solar activity, large coronal holes cover the polar caps, while in more active periods, they can exist at all latitudes but evolve over time. Coronal holes are of interest for

theoretical modeling, as they represent a time-steady state and provide insights into collisionless kinetic processes and their dissipation mechanisms. As they are also associated with high-speed solar wind streams, they can contribute to major geomagnetic storms and corotating interaction regions.

The various methods of observing the solar corona, from EUV to X-rays, as well as coronagraph images, have highlighted two other important transient phenomena: solar flares and coronal mass ejections (CMEs). Initially, these two phenomena were not distinct, and it is important to keep in mind that the wide range of eruptive events sometimes makes it difficult to distinguish between flares and CMEs. During both phenomena, which we will describe here, large quantities of particles are emitted, including protons, electrons, and heavy ions.

- Solar flares (Figure 1.11(b)) are phenomena involving significant energies and temperatures (up to 10^{32} ergs (Harra et al., 2023) and theoretically between 6 MK and 100 MK (Shibata and Yokoyama, 2002)). They can be interpreted as the result of the restructuring of the photospheric magnetic field, which becomes twisted during differential rotations. Most flares are associated with active regions that possess intense magnetic fields and complex magnetic polarities. Russell, Luhmann, and Strangeway describe them as follows (Russell et al., 2016):

"A flare is an impulsive brightening in the low corona associated with active regions. Flares produce enhanced (by orders of magnitude) emissions in energetic photon fluxes from their locale, including EUV and X-rays. They are thought to be a signature of the localized release of magnetic stresses resulting in low coronal heating. They are sometimes, especially in large flare cases, accompanied by energetic particle emissions and coronal mass ejections."

The simplest or "standard" model for flares (Figure 1.10) considers that magnetic reconnection occurs at the top of a closed loop of magnetic field or between two close loops, at the chromosphere level. This results in the acceleration of electrons towards both the surface and space. Magnetic reconnections will be discussed in Section 1.2.3.3. Although widely debated, this model provides a good foundation for understanding flares.

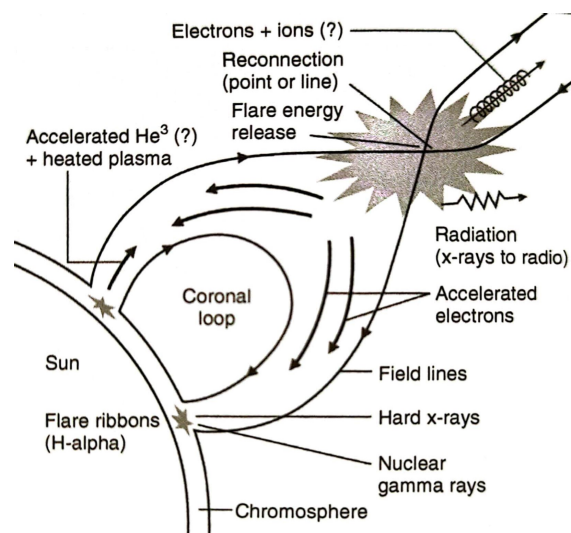


Figure 1.10 – Standard flare model from Russell et al. (2016).

- Coronal Mass Ejections (CMEs) are the most significant solar phenomena in space weather. They are eruptions of matter originally in the corona, expelled into space at an average speed of 300 km/s. However, depending on the CME, velocities can range from 100 to 2000 km/s,

with higher velocities associated with more violent CMEs originating from more active regions and sometimes associated with major flares, such as the example in October 2003 (Figure 1.11). Their size can reach several tens of solar radii (Gruet, 2018). CMEs appear as large loops, magnetic bubbles, or sometimes fine jets (Figure 1.11(c)). They are responsible for most large-scale disturbances in the plasma and magnetic field of the Sun, as well as most geomagnetic storms. It takes several hours to three days for a CME to travel from its origin to Earth. They can evolve slowly, taking several hours to leave the corona, or escape abruptly within minutes (Russell et al., 2016). The duration of the effects of a CME on Earth also depends on its intensity but typically ranges from 24 to 72 hours.

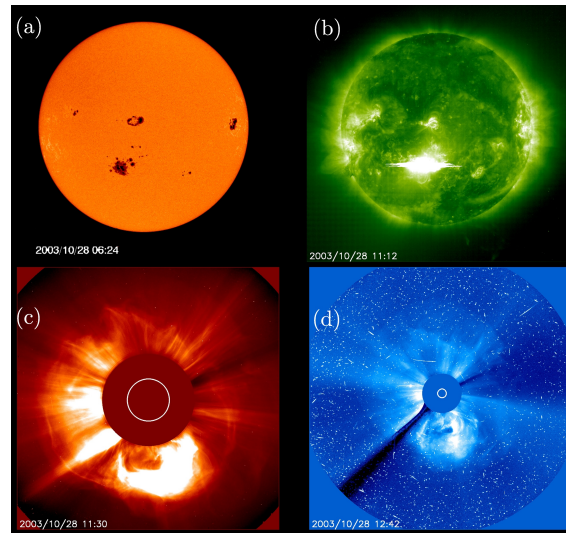


Figure 1.11 – Solar flare and CME from October 2003, during what is known as Halloween solar storms. (a) Michelson Doppler Imager (MDI) image of the Sun’s disk, showing the large group of sunspots of active region 10486; (b) EUV Imaging Telescope (EIT) image of solar flare; (c) C2 camera of the Large Angle and Spectrometric Coronagraph (LASCO) showing the CME cloud; (d) LASCO C3 camera showing the CME cloud. Energetic particles show up in this image as bright points and streaks, when hitting the instrument’s detectors. (Copyright: SOHO/MDI, SOHO/EIT, SOHO/LASCO (ESA & NASA)).

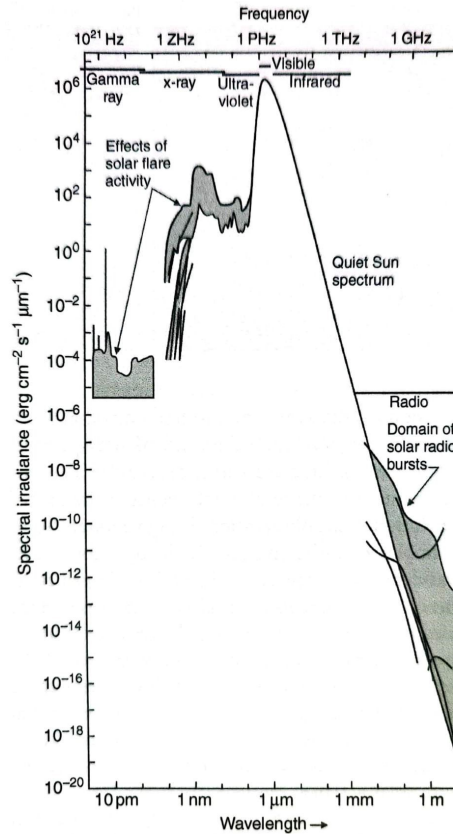


Figure 1.12 – Overall solar emission spectrum from [Russell et al. \(2016\)](#) adapted from [Golub and Pasachoff \(1997\)](#).

Radiations

As explained in the introduction, this thesis focuses on the Sun-Earth chain and aims to understand how solar activity measured in the interplanetary medium provides us with information about precipitating electrons, which are responsible for polar auroras. Since solar radiation is outside the scope of this study, we will provide only a brief introduction here and will not mention certain phenomena such as solar radio bursts.

So far, we have described the structure of the Sun and some of the main phenomena and energy sources originating from the Sun. The energy we receive on Earth can be divided into two categories: particles and electromagnetic radiation. Flares, coronal mass ejections, and the solar wind (the subject of the next section) are accompanied by particle emissions and often a significant increase in electromagnetic radiation. The solar emission spectrum ranges from the gamma domain to the radio domain, including X-rays, ultraviolet, infrared, and visible light. Solar flares are classified in several categories (A, B, C, M, X) based on the intensity of their emission in the 0.1 - 0.8 nm range (X-rays). Figure 1.12 provides an overview of the solar emission spectrum and highlights the importance of X-rays and EUV radiation. The solar radiative output can reach up to 100 billion watts per square meter.

1.2.2 Solar Wind in the Interplanetary Medium

The solar wind is a flow of ionized solar plasma and the remnants of the solar magnetic field that permeate interplanetary space. It is the result of a large difference in gas pressure between the solar corona and the interplanetary medium. Indeed, despite the Sun's gravitational force, the

thermal pressure and ionizing photon flux produce an highly ionized plasma in the corona with a speed above the Sun's escape velocity (Russell et al., 2016). The solar wind is heavily influenced by variations in the coronal magnetic field that cannot contain it globally. Once escaped in the interplanetary medium, the solar wind contains a weak magnetic field mainly oriented parallel to the ecliptic plane at an angle of approximately 45° called the interplanetary magnetic field (IMF). The solar wind is one of the most important elements when it comes to space weather, if not the most important (along with CMEs that can strongly enhance solar wind parameters), producing, among other things, auroras and geomagnetic storms.

In this section, we begin by introducing the general properties of the Solar Wind. In the following section we present the idealized fluid model for the Solar Wind, which serves as a valuable framework for understanding the theoretical basis and spatial configuration of this phenomenon. We then explore the influence of the solar magnetic field, leading us to consider a more comprehensive magnetohydrodynamic (MHD) system of equations. While we do not delve into the mathematical details of these equations here, we discuss the implications and recent perspectives regarding the existence of the Solar Wind. Finally, we provide a comprehensive list of intriguing phenomena associated with the Solar Wind.

1.2.2.1 Properties of the Solar Wind

The Solar Wind, a hot, tenuous, and fast stream of charged particles emitted by the Sun, has been the subject of observations and measurements since the 1970s (Parker, 1959). These observations have been conducted from Earth's orbit and from the Lagrange point 1, located approximately 200 times the radius of the Earth in the direction of the Sun. This Lagrange point, being a stable point in the gravitational potentials of the Earth-Sun system, offers an advantageous vantage point for measurements as it lies upstream of Earth, beyond the influence of the magnetosphere.

The characteristics of the Solar Wind exhibit variations across a range of timescales, spanning from minutes to the duration of a solar cycle. These characteristics primarily include plasma density, bulk speed, ion temperature, and the interplanetary magnetic field and its components. Recent advancements in observations have enabled the identification of ion compositions and the distribution functions for ions and electrons. It is primarily composed of ionized hydrogen with a small percentage ($\sim 5\%$) of ionized helium and even fewer heavier ions (Carbon, Nitrogen, Oxygen, Iron and Silicon). A detailed statistical analysis of solar wind data, which serves as the foundation for this research, will be presented in Chapter 3.

The Solar Wind is commonly described in terms of two distinct classes, each stemming from different physical processes. Observations of bulk speed and density reveal intervals where high bulk speed is associated with low density, and low bulk speed is associated with high density. Although the statistical distribution of solar wind speed highlights a continuity between the peak corresponding to the most common case of slow solar wind (< 350 km/s) and the tail for high speeds (> 600 km/s), it is generally recognized that there are two varieties of solar wind on either side of the threshold of 450 km/s. Table 1.1 provides an overview of the characteristics of these two varieties. Observations suggest that slow solar wind originates from the boundaries of coronal holes, while fast solar wind emanates from the central regions of coronal holes (see Section 1.2.2.3).

1.2.2.2 Idealized Model for Solar Wind

Explaining the standard vision for the existence of solar wind does not take into account the magnetic effect of the corona. It requires us to mainly focus on:

Property	Fast solar wind	Slow solar wind
Time Variation	quasi-steady	typically variable
Speed Range	$\sim 600-800 \text{ km.s}^{-1}$	$\sim 300-500 \text{ km.s}^{-1}$
Density at 1AU	$\sim 1-7 \text{ cm}^{-3}$	$\sim 7-15 \text{ cm}^{-3}$
Proton temperatures	$\sim 4 \times 10^4 \text{ K}$	$\sim 2 \times 10^5$
Electron temperatures	$\sim 1 \times 10^5 \text{ K}$	$\sim 1 \times 10^5 \text{ K}$
Composition	higher He^{++} ($\sim 4\%$)	higher $\text{O}^{+7}/\text{O}^{+6}, \text{Fe}/\text{O}$
Field structure	Alfvén waves	Current sheet(s), rotational discontinuity
Sources	Coronal hole centers	Coronal hole, streamers, boundaries

Table 1.1 – Fast and slow solar wind properties according to [Russell et al. \(2016\)](#)

- A fluid model, explaining the equilibrium state of the solar corona and its relate supersonic continuous flow of plasma from the corona into the interplanetary medium.
- Then a description of the spatial configuration of magnetic field lines that are frozen in the expanding plasma and are therefore carried into the interplanetary medium by the solar wind ([Kivelson and Russell, 1995](#); [Russell et al., 2016](#)).
- And finally the boundary condition: the formation of a shock region in distant regions of the solar system.

Before we delve into this, let's have a quick reminder of some physical concepts. The continuity equation is the equation that accounts for the conservation of mass in a flow. To derive it, we consider an infinitesimal volume element and sum up the inflow and outflow of matter. The resulting equation is as follows:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{u}) = \sigma \quad (1.40)$$

with \mathbf{u} being the flow velocity speed (or vector field describing the movement of the quantity we are looking at, q) and σ gives the net rate of our quantity q (amount per unit volume per unit time). In plasma for instance, numbers of particles can be added by ionization of neutrals or removed by recombination of ions and electrons.

- For a steady flow (no time variation), the left-hand term vanishes.
- For a conservative flow (no sources or sinks), the right-hand term vanishes.
- For an incompressible flow (constant density), the left-hand term vanishes.
- For a conservative and incompressible flow, only $\nabla \cdot (\rho \mathbf{u}) = 0$ remains.

Fluid Model

In fluid mechanics, the principle of mass conservation is described by the continuity Equation 1.41 in several different forms: local conservative (normal time derivative), local non-conservative (time derivative following the particle in its motion), or integral form. The continuity equation represents the conservation of mass. The conservation of momentum in a fluid is reflected in Equation 1.42.

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{u}) = 0 \quad (1.41)$$

$$\rho \frac{\partial \mathbf{u}}{\partial t} + \rho \mathbf{u} \cdot \nabla \mathbf{u} = -\nabla p + \mathbf{j} \times \mathbf{B} + \rho \mathbf{F}_g \quad (1.42)$$

These equations are solved using several approximations:

- The flow is assumed to be at equilibrium, hence independent of time, resulting in zero time derivatives.
- The system is spherically symmetric, and all physical properties are functions of the distance from the center of the Sun, denoted as r . The equations are expressed in spherical coordinates.
- The flow velocity is assumed to be strictly radial.
- Magnetic effects $\mathbf{j} \times \mathbf{B}$ are neglected for now (see Section 1.2.2.3).

This leads to the following equations:

$$\frac{1}{r^2} \frac{d}{dr} (\rho u r^2) = 0 \quad (1.43)$$

$$\rho u \frac{du}{dr} = -\frac{dp}{dr} - \rho \frac{GM_{\odot}}{r^2} \quad (1.44)$$

From here, two main approaches have emerged for solving these equations.

First attempt: Hypothesis of hydrostatic equilibrium. The first and simplest approach is to consider that $u(r) = 0$ everywhere. The first equation is directly satisfied, and the second equation balances the pressure gradient of a static atmosphere with the gravitational forces. By solving for the pressure in a static, isothermal atmosphere, we obtain:

$$p(r) = p_0 \exp\left(\frac{GM_{\odot}m}{2kT} \left(\frac{1}{r} - \frac{1}{R}\right)\right) \quad (1.45)$$

This equation is a generalization of the familiar formula for decreasing pressure with increasing altitude in a static, isothermal atmosphere. However, the problem with this formulation is that as $r \rightarrow \infty$, $p = p_0 \exp\left(-\frac{GM_{\odot}m}{2kTR}\right)$. This pressure value at infinity is several orders of magnitude higher than the pressure of the ambient interplanetary medium, and thus cannot represent an equilibrium between the two.

Second attempt: Towards the solar wind. The second resolution is based on the work of E.N. Parker in the 1950s. It starts with the relation $\rho u r^2 = \text{const.}$ from Equation 1.43. Parker derived the following differential equation for $u(r)$ and du/dr in an isothermal, expanding atmosphere:

$$\left(u^2 - \frac{2kT}{m}\right) \frac{1}{u} \frac{du}{dr} = \frac{4kT}{mr} - \frac{GM_{\odot}}{r^2} \quad (\text{Parker}) \quad (1.46)$$

This is a differential equation for $u(r)$ and du/dr in an isothermal, expanding atmosphere. This equation led to the idea of the existence of a solar wind. For any temperature T of the solar corona, the second term on the right-hand side (GM_{\odot}/r^2) is larger than the first term at the base of the corona, $4kT/mr$. This means that regardless of its temperature, the solar corona is trapped by gravity. At the critical radius $r_c = \frac{GM_{\odot}m}{4kT}$, the right-hand side of the equation becomes positive again. At this critical radius, the left-hand side of the equation must be zero, which can be solved in two ways. The first is to consider that $\frac{du}{dr}|_{r_c} = 0$, which would mean that the solar wind solutions have their minimum or maximum velocity at r_c . As a slight Doppler shift has been detected at the level of the corona, we know that u is smaller in this region. By observing the signs, we would have u increasing until r_c and then decreasing again. All solutions based on this approach share the issue of the first formulation (the value of pressure at infinity). The second way is to consider that at $r = r_c$, we have $u^2 - \frac{2kT}{m} = 0$ meaning that u has the sound speed. It is by exploring this solution that we obtain the correct version, i.e. a wind that smoothly passes from sub to supersonic consistent with both the coronal and outer boundary requirements: the solar wind.

Spatial Configuration of the Interplanetary Magnetic Field in Fluid Model

The plasma outflow from the corona carries the open coronal field embedded within it: it is the frozen-in condition (see Section 1.1.4.2). Hence, the plasma flows is dragging the magnetic structure with it. The intensity of such purely radial magnetic field decreases at a $1/r^2$ rate. However, as mentioned before, the Sun has a differential rotation which varies with latitude. Looking at the equator, the result would be a archimedean spiral pattern of outflow as shown Figure 1.13, often presented by Parker as analogous to a "garden sprinkler". Figure 1.13(b) is showing this result in the case of an equatorial, constant speed flow. The angle between a magnetic field line and the orbit of the Earth located at 1 astronomical unit (AU)¹ is approximately 45° .

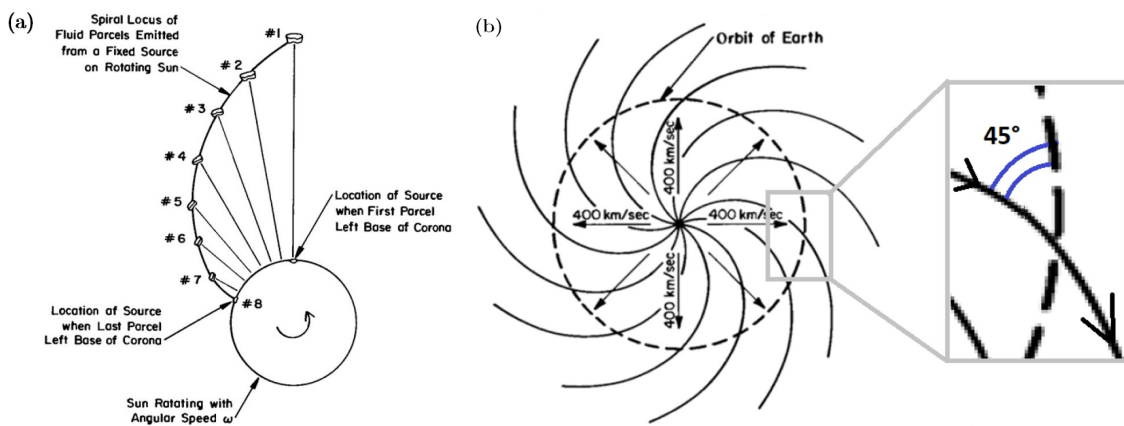


Figure 1.13 – Parker's model of Interplanetary Magnetic Field from Russell et al. (2016). (a) Motion of a single parcel of solar-wind fluid carrying the Sun's open field line (Russell et al., 2016). (b) Parker spiral field in the equatorial plane in case of a constant solar-wind speed of $400\text{km}\cdot\text{s}^{-1}$ from Kivelson and Russell (1995).

Boundary Conditions and Shock

We find ourselves in the opposite problem compared to earlier. At infinity, everything tends toward zero, making it difficult to connect the solar wind with the conditions of the interplanetary medium. In reality, the momentum flux density ρu^2 is a property of the fluid that approximates pressure. Since ρ follows a $1/r^2$ dependence according to the continuity equation, the momentum flux density falls below the pressure value of the interplanetary medium. To address this problem, one would expect the solar wind to slow down, but it is not easy to decelerate it. In fact, the only way to resolve this dilemma is to assume the formation of a shock wave that abruptly decelerates the solar wind. After this shock, the solar wind can gradually adjust to the equilibrium of pressures. Based on calculations, the distance between the Sun and this shock is estimated to be 160 AU.

1.2.2.3 Influence of the Solar Magnetic Field

In order to obtain a more realistic representation of the solar wind, it is necessary to consider the interaction between the expansion of the solar corona and the solar magnetic field, specifically the term $\mathbf{j} \times \mathbf{B}$ in Equation 1.42. However, solving this system of equations becomes considerably more challenging compared to the ideal model. The assumption of a spherically symmetric MHD flow describing the corona is no longer valid, and the assumption of purely radial velocity u needs

1. 1 AU = 149,597,870,700 m

to be abandoned (Kivelson and Russell, 1995).

To simplify the problem, one approach is to impose the magnetic dipole as a boundary condition at the base of the corona, which allows for a more tractable solution within the MHD framework. Pneuman and Kopp (1971) adopted this approach and obtained a configuration consisting of a mixture of closed field lines below a certain latitude and open field lines above it. The open-field lines again provide us a model for solar wind formation. Figure 1.14 illustrates this configuration. We can notice two field lines, very close to each other along the equatorial plane, exhibiting different polarities. This results in a thin layer of high current density normal to the equatorial plane, known as the "interplanetary" or "heliospheric" current sheet (highlighted in yellow in Figure 1.14). Taking into account the dipole's tilt relative to the Sun's rotational axis, we obtain a configuration often referred to as the "ballerina's skirt"-shaped heliospheric current sheet, as depicted in Figure 1.15. It is important to note that modern 3D rendering from MHD models, incorporating realistic source distributions, present a more intricate picture. However, the tilted dipole representation serves as a useful baseline for describing the heliospheric current sheet.

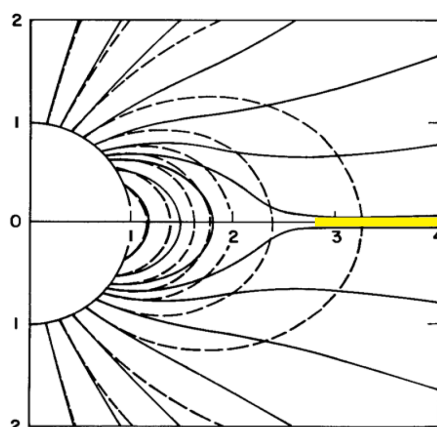


Figure 1.14 – Isothermal coronal-expansion model from Pneuman and Kopp (1971) for a dipole magnetic field considered at the base of the corona. Dashed lines are field lines for classical dipole field. Yellow zone corresponds to a thin layer of high current density. Adapted from Kivelson and Russell (1995).

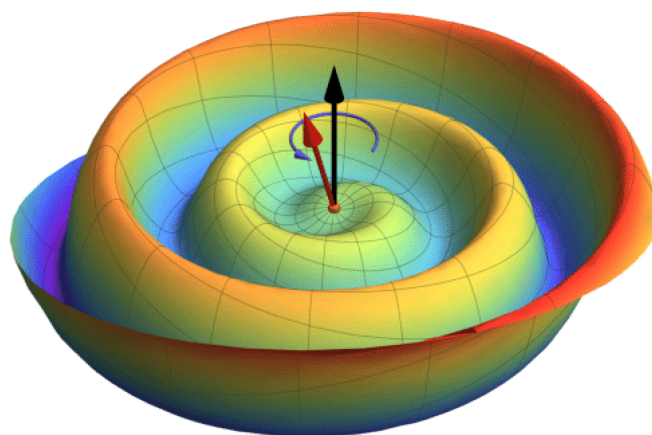


Figure 1.15 – 3D rendering of the interplanetary current sheet in a Parker solar wind model for a tilted dipole, from (Orcinha et al., 2019).

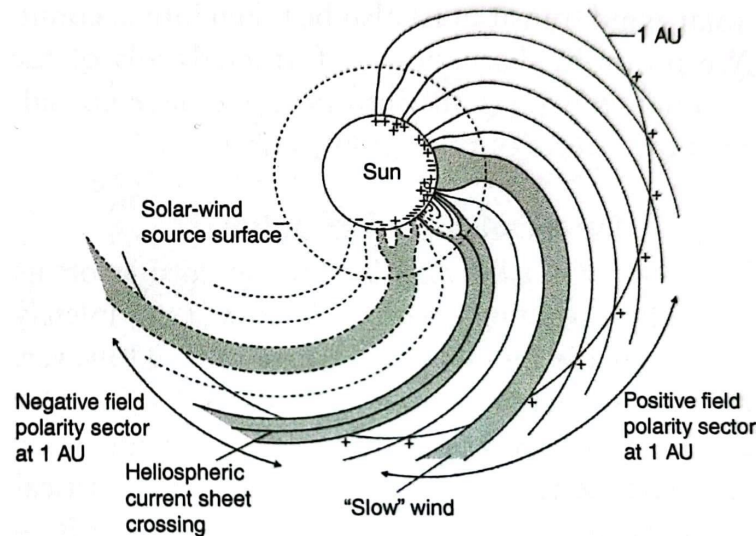


Figure 1.16 – Complete solar-wind extension into interplanetary medium. The + (respect. -) symbol indicates positive outward (respect. negative inward) magnetic field. Region of slow solar wind are shaded (Russell et al., 2016).

Through various models, such as the *potential field source surface* (PFSS) model (Altschuler and Newkirk, 1969; Hoeksema, 1984; Schatten et al., 1969; Wang and Sheeley, 1992), we can determine a critical radius where the combined effects of solar gravity and the corotations of the solar magnetic field diminish significantly, allowing the solar wind to flow outward unimpeded. This radius lies well beyond the region where the last closed coronal loops exist and is theoretically situated around $\sim 10\text{-}20$ solar radii (Russell et al., 2016). One other important finding from these models is the understanding that most of the slower solar wind originates from the boundaries of coronal holes, which serve as sources of higher-speed solar wind streams. This suggests a relationship where the lower-speed solar wind sets a boundary for the higher-speed streams, which is seen in measurements at 1AU.

Figure 1.16 from Russell et al. (2016) (adapted from Schatten et al. (1969)) encapsulates most of the ideas presented here.

- The Solar-wind source surface is the starting point of our roughly spiral-shaped flow in the interplanetary medium.
- The heliospheric current sheet is the boundary between positive field and negative field sectors (see also Figures 1.14 and 1.15). From Earth (1AU) we recurrently see magnetic field change when going above and below the interplanetary current sheet (we can imagine Earth’s orbit in image 1.15). Last Sun’s magnetic field flip happened around December 2013².
- Lower-speed solar wind bounds the open field regions thus separating higher-speed streams from different sources.

1.2.2.4 Associated Phenomena

In this final section about the solar wind, we will present several effects and consequences associated with it. Firstly, we will discuss the corotating interaction regions (CIRs) and the shocks

2. <https://www.nasa.gov/content/goddard/the-suns-magnetic-field-is-about-to-flip/>

they generate. Next, we will examine the propagation of coronal mass ejections (CMEs) in the solar wind, specifically focusing on the resulting interplanetary coronal mass ejections (ICMEs) and their associated shocks. Furthermore, we will explore the particle acceleration mechanisms within these shocks, leading to the formation of solar energetic particle (SEP) events. Finally, we will briefly touch upon the interaction between the solar wind and non-solar particles.

When looking at Figure 1.16, we notice the propagation of the fast and slow solar winds. It is straightforward to understand that a fast solar wind tends to overtake a slow solar wind. However, due to the frozen-in nature of the magnetic field within the plasma, these regions of interaction between the flows correspond to areas where both plasma density and magnetic fields experience compression ahead of the fast solar wind. A clear understanding of this phenomenon is provided by Figure 1.17. These regions, known as corotating interaction regions (CIRs), reappear periodically every 27 days as they corotate with the Sun. Notably, these regions persist in density measurements taken near Earth, even as one moves farther away from the Sun. Moreover, it is important to note that within these regions, the slower solar winds are subject to acceleration while the faster solar winds experience deceleration, primarily due to momentum redistribution. Additionally, the sudden changes in velocity profiles can trigger shock waves, resulting in abrupt alterations in solar wind properties, such as elevated pressure, density, and temperature. The orientation of CIRs also plays a role in the frequency and regularity of shocks. Forward shocks (in front of the high speed-stream) tend to occur more frequently than reverse shocks (behind the high speed stream) (Riley et al., 2012).

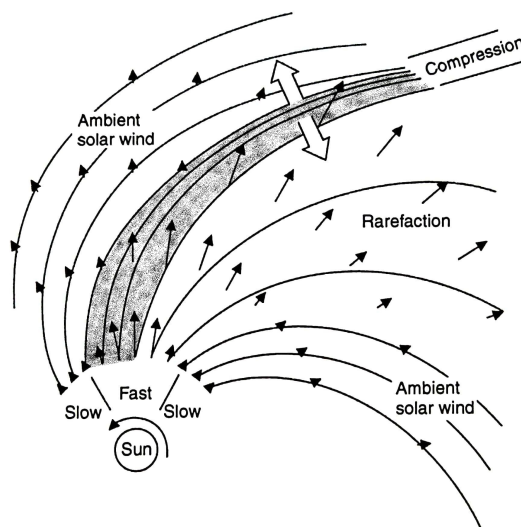


Figure 1.17 – Visual representation of a corotating interaction region (CIR) forming compression and rarefaction zones. From Russell et al. (2016).

The steady-state of the solar wind, as previously described, consistently fills the interplanetary medium. However, it is subject to temporal evolution, influenced by coronal conditions and occasionally exhibiting violent transient behaviors, particularly during interactions with coronal mass ejections (CMEs). As previously mentioned, certain explosive CMEs can attain velocities of several thousand kilometers per second within a few tens of solar radii. These events leave evident signatures in the interplanetary medium when measured at a distance from the Sun. Referred to as interplanetary coronal mass ejections (ICMEs), the term "interplanetary" emphasizes the *in situ* measurements of the CME. As outlined by Russell et al. (2016), ICMEs exhibit the following characteristics:

- The compression and deflection of the solar wind by the ejecta generate a leading shock that imparts additional heat to the solar wind.
- ICMEs typically display normal density, lower ion temperature, and enhanced magnetic field compared to the prevailing solar wind conditions.
- Unusual ion composition signatures may be observed within ICMEs.
- In certain instances, particularly at high heliolatitudes, ICMEs may also induce a reverse shock.
- The occurrence of ICMEs follows the solar cycle, as expected based on the occurrence rates of CMEs.
- Statistics on the speed of CMEs have revealed that CMEs initially moving slower than the solar wind tend to accelerate as they reach 1 AU, whereas CMEs moving faster tend to decelerate. The modeling efforts aimed at understanding the underlying mechanisms behind these effects are still ongoing.

Both shocks generated by ICMEs and CIRs are categorized as interplanetary shocks. While shocks originating from CIRs intensify with radial distance, they remain relatively weak compared to those caused by ICMEs. But overall, both CIR shocks and ICME shocks are considerably weaker compared to the bow shock (see Section 1.2.3.1).

Solar energetic particles (SEPs) are ions, particularly protons, that originate from particle acceleration mechanisms occurring in interplanetary shocks or solar flares and exhibit energy levels ranging from keV to GeV. Hence, they usually are classified as coming from two type of event: "impulsive" for the flares or "gradual" for the ICME shocks. They are distinguished through their duration (hours vs days), their composition (higher proportion of heavy ions such as Fe vs solar-wind ion composition) and the radio burst type that they are associated with (type III vs type II)³.

There are a variety of particles and processes that do not come from the Sun: interstellar neutrals, dust, or galactic and anomalous cosmic rays (GCRs and ACRs). Here, we will not present the complex heliospheric pickup ions (and the ion pickup process) or interactions with dust but we will stick to a brief introduction on GCRs. GCRs come from interplanetary and interstellar space and are present throughout the Solar System. Their energy range span from sub-GeV to beyond TeV. They have astrophysical origins, such as supernovae explosions and active galactic nuclei, which could explain their incredible energy. GCRs consist mainly of protons (~70-90%), helium nuclei (~7-10%), heavy elements (~1%), and electrons (~1%). During periods of heightened solar activity, the interplanetary magnetic field (IMF) is intensified, creating a stronger shield that hampers the penetration of GCRs into the inner heliosphere. Conversely, during solar minimum phases, the weaker IMF allows for greater GCR influx towards Earth's atmosphere (Bothmer and Daglis, 2007).

Let us recall that what we have presented in Sections 1.2.1 and 1.2.2 is a condensed and simplified overview of the Sun's activity and its associated phenomena. We have intentionally excluded comprehensive topics such as *coronal heating*. For a more comprehensive understanding of Sun-related topics, readers are advised to use the many resources available, such as Schrijver

3. A radio burst is a brief period during which the Sun's radio emission is above background level. A type II radio burst is a relatively slow drift from high to low frequencies of around $1 \text{ MHz}\cdot\text{s}^{-1}$ while a type III is a rapid frequency drift from high to low frequencies of around $100 \text{ MHz}\cdot\text{s}^{-1}$

and Siscoe (2010) or Priest (2014).

In summary, the Sun exhibits a very dynamic activity starting at its nuclear core. The convective zone, in particular, plays a crucial role in generating the solar magnetic field through a dynamo-like mechanism. This magnetic field controls the structure of the coronal field, which can be envisioned as a network of interconnected loops. In the corona, the Sun's outermost ionized atmosphere, plasma outflows caused by high pressure gradients are observed. These outflows lead to the formation of coronal holes, members of the family of active regions. The location, distribution, and emergence of active regions change throughout the solar cycle, influencing the coronal field and, consequently, impacting Earth. Solar cycle-related phenomena such as flares and coronal mass ejections (CMEs) generate energetic particles and radiation. Moving to higher altitudes, the solar atmosphere transitions into the solar wind, which can be described using magnetohydrodynamics (MHD). Parker's MHD fluid description, developed in the 1960s, provides an initial framework for understanding the solar wind's continuity from subsonic to supersonic regimes and exhibits the existence of the heliospheric current sheet and its distinctive skirt-like geometry. Other models (e.g., PFSS) have aided in comprehending the sources of slow and fast solar wind outflows. Farther in the interplanetary medium appear interaction regions forming compression and shocks that depend on the ~ 27 day solar rotation (CIRs). We have also examined the propagation of coronal mass ejections (CMEs) within the solar wind, which subsequently lead to the formation of ICMEs. These ICMEs often trigger significant solar energetic particle (SEP) events. Finally, we briefly mentioned galactic cosmic rays (GCRs), an extraordinarily energetic population of particle. All these elements will be actors around the Earth's Magnetosphere and main protagonists of Space Weather.

1.2.3 Solar Wind's interaction with Earth's Magnetosphere

In this section, we shift our focus to the interaction between the solar wind and the Earth's magnetic field, building upon our previous explanation of the solar wind's nature and its propagation in the interplanetary medium. Our objective is to examine how the solar wind influences the Earth's magnetosphere. To lay the groundwork for subsequent discussions, we will first introduce the key features of the magnetosphere, without delving into technical intricacies, thereby establishing a common vocabulary for the remainder of this section. Subsequently, we will delve into a detailed analysis of the interactions between the solar wind and the outer magnetosphere. Lastly, we will explore the processes occurring within the inner magnetosphere, which also arise as a result of its coupling with the solar wind. Our references include Koskinen and Kilpua (2022) and Russell et al. (2016), among others, which guided the structuring of this section. Nonetheless, we encourage readers to explore the extensive literature available on the subject of the Earth's magnetosphere, including notable works like Bothmer and Daglis (2007).

1.2.3.1 Introduction to Earth's Magnetosphere

The Earth's magnetic field, like that of the Sun, originates from a hidden dynamo effect within the planet. This phenomenon is driven by the convective motion of liquid iron in the Earth's liquid core, located more than 1200 km from the planet's center. Initially, we adopt a simplified approach by considering the Earth's magnetic field as a dipole inclined at an 11° angle relative to the Earth's rotational axis. However, this dipole is not perfectly centered but instead offset by approximately 450 km from the rotational axis. As a result, the magnetic field intensity is higher at the magnetic North Pole than at the magnetic South Pole, with a minimum value occurring within a region known as the South Atlantic Anomaly. We refer to the region influenced by this magnetic field as the *magnetosphere*, where particle motion is governed not by the solar wind but by the Earth's magnetic field. The term "magnetosphere" is used when a planetary magnetic field is confined by the solar wind. In the simplified dipolar field model, the equation describing a magnetic field

line is given by $r = r_0 \sin^2 \theta$, where θ represents the magnetic colatitude and r_0 is the distance between the dipole center and the point where the field line crosses the equator. This equation can be transformed into $r = L \cos^2 \lambda$, where λ denotes the magnetic latitude. Here, L is a distance measured in planetary radii and is commonly used to identify field lines and associated field intensities, providing an explanation for the earlier reference to the "L-shell" term.

In the ideal scenario where the Earth's magnetic field and the interplanetary magnetic field (IMF) coexist without collisions or dissipations, they would act as impenetrable barriers to each other. However, reality differs as the solar wind encounters difficulties in penetrating the Earth's magnetosphere, resulting in the formation of a distinct boundary. This boundary, referred to as the *magnetopause*, represents a discontinuity in the flow and serves as the interface between the influence regions of the solar wind and the Earth's magnetic field. The position of the magnetopause directly depends on the equilibrium between the pressure exerted by the solar wind and the magnetic pressure within the magnetosphere. In the Earth's reference frame, the solar wind attains velocities significantly higher than the maximum speed at which disturbances propagate through the fluid. This speed, the magnetosonic velocity ($v_{ms} = \sqrt{v_s^2 + v_A^2}$, where v_A represents the Alfvén velocity, to be discussed later), leads to the formation of a collisionless shock ahead of the magnetopause. This shock converts the substantial kinetic energy of the solar wind into electromagnetic energy and heat. Consequently, a region of high disturbance arises between this shock and the magnetopause, known as the *magnetosheath*. On the day side, the front point of this shock is the *bow shock nose* (BSN), with its distance from Earth varying between 13 and 6 Earth radii during calm and active periods, respectively (with an average distance of approximately 10 Earth radii under typical conditions). On the night side, the interactions cause the initial dipole shape to stretch, resulting in the formation of a long tail known as the *magnetotail*. Observations have revealed that this magnetotail extends far beyond the Moon (Kallio and Facskó, 2015), supporting the notion that a significant amount of energy is transferred from the solar wind to the magnetosphere. Additionally, magnetospheric models emphasize the existence of neutral points known as *polar cusps*, located at the north and south poles. These points persist regardless of the IMF configuration in which the magnetosphere is immersed and provide direct access for magnetosheath plasma to the ionized region of the upper atmosphere, known as the *ionosphere* (Russell, 2000).

A substantial portion of the magnetosphere comprises two regions referred to as the *tail lobes*, which are magnetically connected to the polar caps, themselves bounded by the *auroras*. Furthermore, there is a *plasma sheet* containing an electric current known as the *cross-tail current*, which forms a closed loop around the lobes where it forms the nightside of the *magnetopause current*. Later in this section, we will delve into greater detail regarding additional currents such as the *ring current*. Figures 1.18 and 1.19 provide visual representations of the various elements discussed here.

In the following sections, we will explore in detail various phenomena occurring within the magnetosphere. This will enable us to elucidate the origin of the main magnetospheric currents.

1.2.3.2 Magnetopause & Magnetopause Current

The magnetopause and its associated current serve as the initial indicators of the solar wind's arrival in the vicinity of our planet. As previously discussed, in an ideal scenario of collisionless plasma without any reconnection processes, magnetic plasmas would remain distinct and unable to mix. If there is no activity in the magnetosheath, characterized by a vacuum magnetic field on the magnetosphere side, the boundary between the magnetosphere and the solar wind represents a balance between plasma pressure and magnetic pressure. Within this context, the magnetopause hosts a specific current known as the magnetopause current or Chapman-Ferraro current.

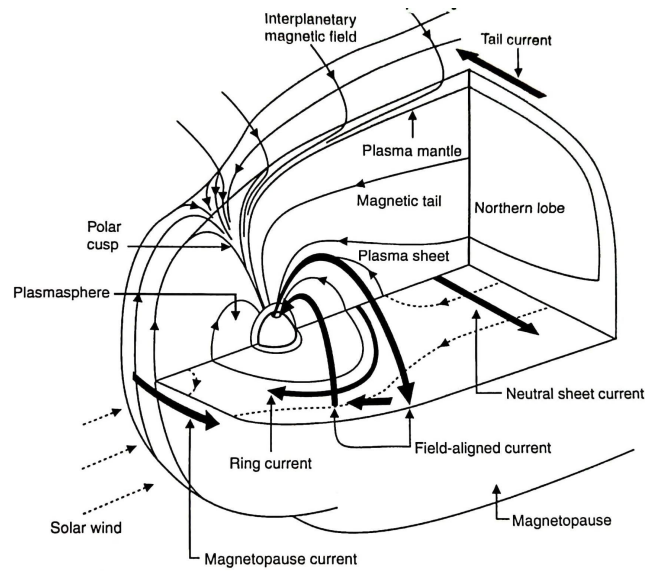


Figure 1.18 – 3D representation of the magnetosphere showing the major regions and currents, from [Russell et al. \(2016\)](#)

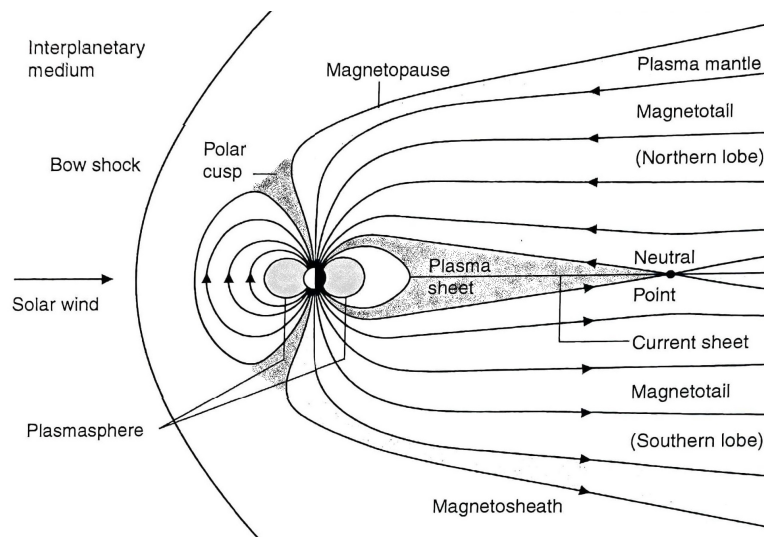


Figure 1.19 – Cross section of the magnetosphere showing the major regions, including the outer magnetosphere and the magnetotail. From [Russell et al. \(2016\)](#).

Conceptualizing this current is relatively straightforward. The magnetopause, situated at the interface between the vacuum magnetic field and a plasma of near-infinite conductivity, signifies the last region where particles undergo gyration. Consequently, certain particles, specifically electrons moving westward and protons moving eastward, become detached from the magnetopause. This collective motion across the entire magnetopause-solar wind interface generates a current flowing in an eastward direction (denoted as j_{mp} in Figure 1.20).

On the dayside, this current manifests approximately 10 Earth radii from the planet and completes its circuit by encircling a neutral point on the magnetopause. The magnetopause current induces noteworthy effects at the Earth's surface, including magnetic disturbances parallel to the terrestrial magnetic field and directed toward the north. Additionally, it amplifies the magnetic field intensity within the magnetosphere and along the dayside surface of the Earth.

While the actual behavior of the magnetopause is more intricate, it is important to acknowledge the existence of this current, and the explanation here is a valuable approximation. In the absence of other influencing processes, the entirety of the Earth's magnetic field would be confined within the boundaries of the magnetopause.

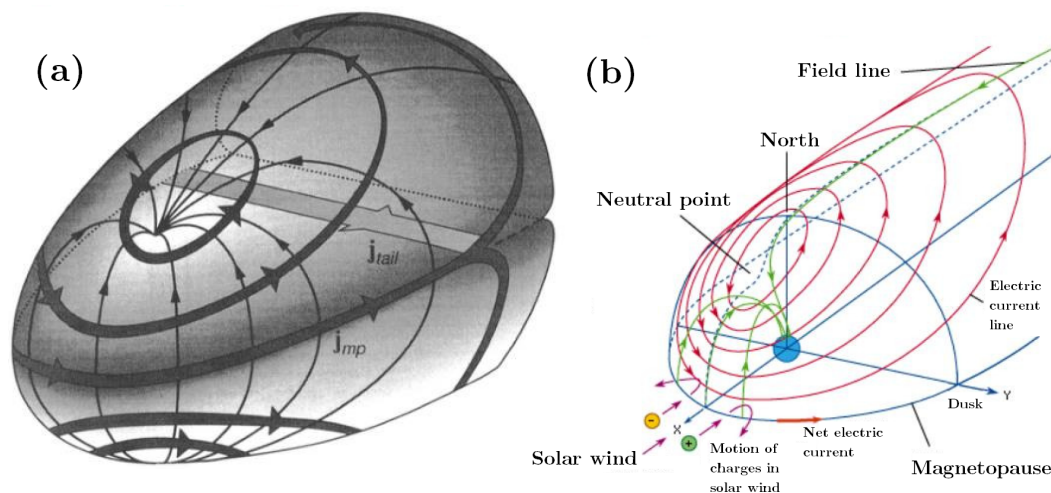


Figure 1.20 – (a) 3-D view from above of the magnetopause and the currents j_{tail} and j_{mp} from Dr. A. Marchaudon's class, found online <https://www.slideserve.com/ion/le-champ-magn-tique-d-origine-externe-aur-lie-marchaudon-lp2e> and (b) identical perspective view of the northern portion of the magnetopause current, as seen from above the ecliptic plane. Charged particles in the solar wind are deflected in opposite directions by Earth's main field, creating a boundary current (from Encyclopædia Britannica, <https://www.britannica.com/science/geomagnetism/field/The-ionospheric-dynamo>, last accessed September 2023).

1.2.3.3 Magnetic Reconnections & Dungey Cycle

The interaction between the solar wind and the Earth's magnetosphere locally involves magnetic reconnection processes. We have previously mentioned the phenomenon of magnetic reconnection (see Section 1.2.1.2), but now we will delve into its mechanism and the pivotal role it plays in the Sun-Earth system. Our discussion will try to provide a comprehensive overview, although no consensus about some fundamental questions "is presently available, thus indicating that the subject of magnetic reconnection remains open for further theoretical, computational, and

observational studies" (Gonzalez and Parker, 2016).

Without delving into the mathematical intricacies of the process, magnetic reconnection is defined as a localized rearrangement of the magnetic field topology, wherein antiparallel fields annihilate each other, converting a fraction of magnetic energy into kinetic, thermal, and particle acceleration energy (see Figure 1.21). This phenomenon primarily occurs in plasma environments where collisions are negligible. Initially distinct antiparallel magnetic fields connect as a result of the instability within the magnetic configuration. The principal reconnection processes in the magnetosphere manifest at the magnetopause, in the magnetotail, and within the solar atmosphere. The first occurs at the interface between the interplanetary magnetic field (IMF) and the magnetospheric field, as elaborated below. The second arises from the stretching of magnetic fields into a thin sheet geometry. The third is driven by diverse dynamic motions of the solar magnetic field, giving rise to phenomena like coronal mass ejections (CMEs). Various models and simulations have observed the existence of this process in magnetohydrodynamic (MHD), hybrid, and fully kinetic scenarios, facilitating an understanding of its underlying mechanisms. For instance, P. A. Sweet's model (Sweet, 1958) restricted diffusion to a localized region, while H. E. Petschek's model (Petschek, 1964) incorporated the influence of MHD waves. RG. Giovanelli and J. W. Dungey, on the other hand, highlighted the three-dimensional nature of reconnection.

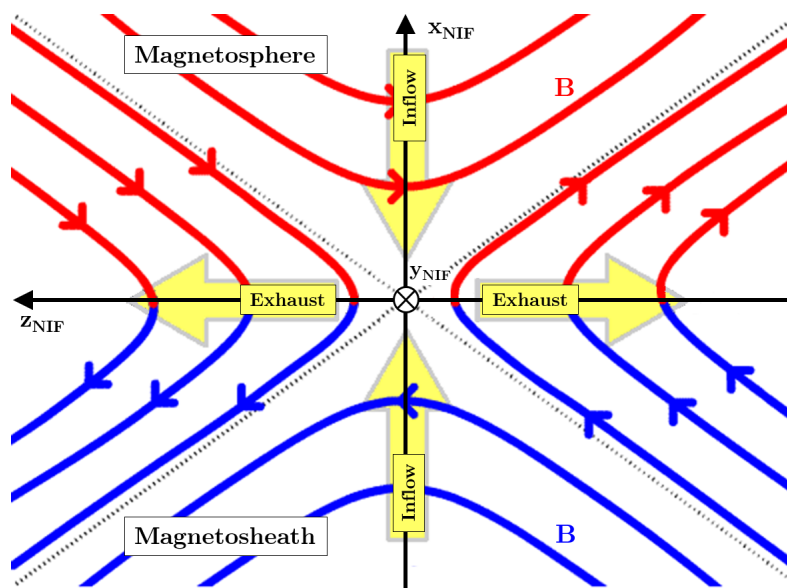


Figure 1.21 – Schematic view of the reconnection geometry at the magnetopause. NIF = normal incidence frame. Adapted from Russell et al. (2016).

The reconnection process is the dominant mechanism responsible for momentum transfer from the solar wind. J.W. Dungey was the first to establish the connection between magnetic reconnection at the magnetosphere's leading edge and this energy transfer, known as the Dungey cycle (Figure 9.19 from Dungey (1961)). He was the first to propose the idea of an "open magnetosphere", where coupling between the geomagnetic field and the interplanetary magnetic field (IMF) leads to a large-scale, global, circulatory flow of magnetic field lines and plasma within the magnetosphere (Sandhu et al., 2019).

Initially, an interplanetary magnetic field (IMF) line (as seen Figure 1.22) connects with a terrestrial magnetic field line, leading to plasma ejection and acceleration on both sides of the reconnection point. The magnetic field lines relax and retreat behind the Earth, resulting in deceleration of the plasma, while the incoming magnetic flux adds energy to the magnetotail. Subsequently,

a secondary reconnection event can occur at the center of the tail, between lines 6 and 6' (Figure 1.22), enabling the plasma to continue its journey. This plasma flow bifurcates, with one portion heading towards the Earth and the other moving in the opposite direction. Upon rapid return, field line 7 can also undergo reconnection within the ionosphere, redirecting plasma back to the dayside. Conversely, field line 7' becomes entirely detached and transforms into an interplanetary field line. The plasma pressure likely propels the high-altitude flux towards the dayside, gradually transforming field line 7 into field line 9. During this process, the footpoint of the field lines, located in the ionosphere, rotates counter to the Earth's rotation, thereby counteracting the ionosphere's drag force. This equilibrium between the ionosphere and the magnetosphere necessitates the entrainment of magnetospheric field lines at high altitudes, leading to shear in the magnetic field between the magnetospheric and ionospheric field lines. A shear in magnetic field results in the bending of magnetic field lines. Consequently, charged particles, primarily electrons and protons, moving along these magnetic field lines experience disparate velocities and relative motions. This relative motion generates electric currents parallel to the magnetic field lines that can accelerate electrons towards the Earth, close within the ionosphere, and form dawn to dusk currents along the polar cap. These currents themselves generate lateral forces through Lorentz force $\mathbf{j} \times \mathbf{B}$, pushing the footpoints of the magnetic field lines.

As represented on the spheres of Figures 1.22 and 1.29, the plasma flow produces two foci in the circulating pattern. The electric potential difference between the two serves as a valuable indicator of magnetospheric activity. The voltage drop measured here is directly proportional to the rate of magnetic reconnection (Russell et al., 2016). This reconnection process enables the transfer of kinetic energy flux from the solar wind (e.g., for a proton density of 10 cm^{-3} , the solar wind typically exhibits a kinetic energy flux of $1 \text{ mJ m}^{-2} \text{ s}^{-1}$) into the magnetosphere, specifically the magnetotail region. Consequently, the kinetic energy of the solar wind undergoes conversion into magnetic energy, which becomes stored in the tail and is available for subsequent release. The thickness of the reconnection layer and the magnetic connectivity play critical roles in determining the magnitude of energy transfer.

In summary:

- The magnetosphere, in conjunction with solar wind interactions, exerts a pulling effect on magnetic field lines, as illustrated by the sequence from 1 to 9 in Figure 1.22.
- Shearing between the magnetosphere at high altitudes and the ionosphere results in relative motions among particles, leading to the formation of currents that accelerate charged particles along magnetic field lines.
- These currents close when the magnetic field lines meet at the ionosphere level along the polar cap (see 1.2.6.3 to better understand all the field-aligned currents in the ionosphere-magnetosphere coupling).
- The consequent $\mathbf{j} \times \mathbf{B}$ force can overcome the ionosphere drag, causing the footpoint of magnetic field lines to shift, creating a circulating pattern for plasma flow.
- Remark: The reconnection also comes with an energy flow into the magnetosphere. The kinetic energy carried by the solar wind is converted into magnetic energy, effectively stored within the magnetotail, and available for subsequent release in various phenomena. We will explore some of them in the following sections.

1.2.3.4 Polar Cusps

We have previously discussed polar cusps, which are regions where magnetosheath plasma has "direct access" to the ionosphere and experiences strong interaction with the solar wind. In a 2D cross-section of the magnetosphere (Figure 1.19), the polar cusps are located at the north and

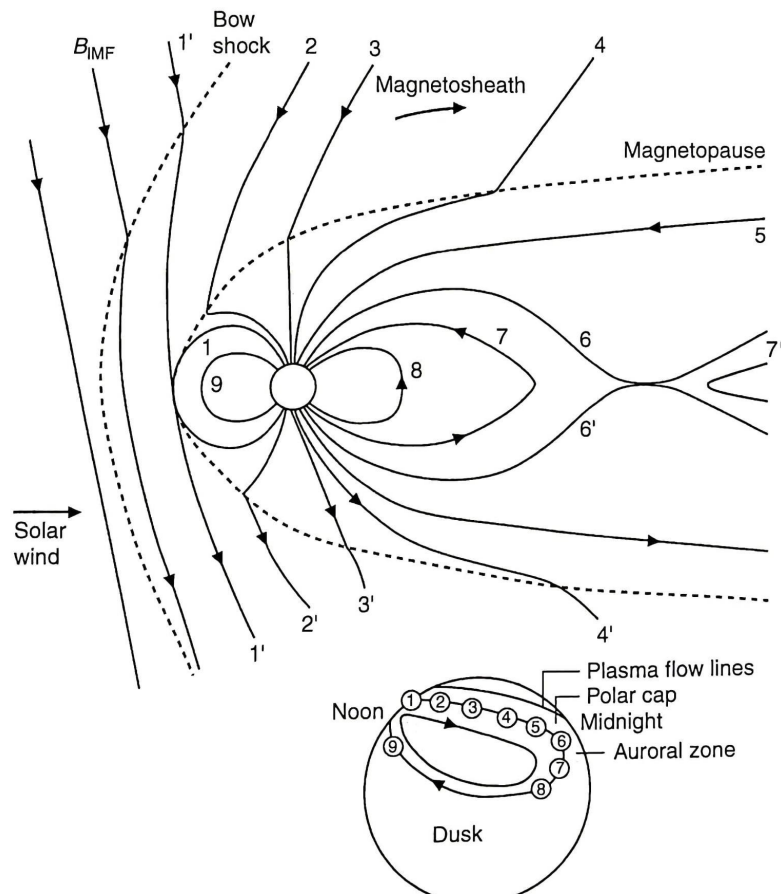


Figure 1.22 – Dungey cycle from *Kivelson and Russell (1995)*. Magnetospheric convection due to the magnetic reconnections. A field line labeled 1 connects with a field line labeled 1' causing a change of topology of the field line. Numbers show the successive configurations of the field line, in the antisunward direction, in the magnetosphere, polar cap and auroral ionosphere.

south, separating magnetic loops directed towards the Sun from those going in the opposite direction. These regions represent "weak points" where the intensity of the magnetic field approaches zero.

The polar cusps are particularly significant in terms of the coupling between the solar wind and the magnetosphere, as they are in direct contact with the magnetosheath plasma. During transient events in the solar wind, the density in the polar cusps can reach similar levels to that of the magnetosheath.

While the existence of polar cusps does not depend on reconnection phenomena or the presence of the solar wind, their properties are directly influenced by these factors. In a non-reconnecting magnetosphere, the location of the cusp is determined by the shape of the magnetopause. However, when the magnetosphere undergoes reconnection with either a southward or northward interplanetary magnetic field, the position of the cusp is altered (Russell, 2000). During significant reconnection events, the polar cusps move towards lower latitudes, closer to the Sun. Some observations from Russell (2000) include the following: the cusp moves equatorward when the IMF turns southward and also shows a tendency to move equatorward for increasingly northward IMF. As the IMF By component becomes more negative, the cusp shifts to earlier local times in the northern hemisphere, indicating a displacement of the reconnection site away from the noon meridian when the IMF is not predominantly southward. Conversely, when the IMF By component is positive, the northern cusp moves to later local times. Additionally, an increase in solar wind dynamic pressure causes the polar cusp to widen both in terms of local time and latitude.

1.2.3.5 Ring Current

When discussing the inner magnetosphere, we generally refer to the region where the Earth's magnetic field exhibits quasi-dipolar characteristics. Within this region, three specific regions of interest can be distinguished: the plasmasphere, the ring current, and the radiation belts. The ring current and radiation belts are directly formed through the trapping of particles within the magnetic field (see Section 1.1.2). In this context, the ring current arises as a direct consequence of particle drift, with protons drifting westward and electrons drifting eastward. It primarily consists of ions drifting in the energy range of 10-200 keV, with electrons playing a lesser role. It encircles the Earth in the equatorial region at a distance of approximately 3 to 4 Earth radii (R_E). The ring current is then one of the main cause of perturbations of the north component of the magnetic field measured on ground at the equator. As we will see in Section 1.3, on-ground magnetic indices such as the *Disturbance Storm-Time Index (Dst)* are then good witness of the ring-current activity, itself a good witness of geomagnetic storms (Lin, 2021). We can define *geomagnetic storms* simply as the energization of the plasma on closed field lines in the magnetosphere. It is mainly the response of the Earth's magnetosphere to a violent phenomenon such as ICMEs, SIRs, and fast solar wind. A storm is characterized through a specific pattern in the *Dst* and hence will be described in more details Section 1.3.

Energetic protons and O⁺ ions serve as the main carriers of the ring current. It is primarily sourced from the ionosphere and the solar wind. While oxygen originates from the ionosphere, protons usually come from both sources. Based on AMPTE/CCE and CRRES observations found in Daglis et al. (1999) solar wind (respect. ionosphere) contributes to approximately 65% (respect. 35%) of species in quiet times and 30% (respect. 70%) during intense storms. Ion outflow from the ionosphere occurs mainly in auroral and polar cap latitudes. Ions are transported to the magnetospheric tail and gradually accelerated before reaching the inner magnetosphere. The heating and acceleration processes occur through multiple steps, including the influence of fluctuating electric fields and magnetic field-aligned electric potential structures. The enhancement of the ring current is a fast process as seen in the *Dst* signatures of geomagnetic storms. However, to recover from this change, the ring current experiences a loss of energetic ions, mainly through collisions

between charged particles and neutral atoms from the *geocorona*. The geocorona, a cloud of hydrogen atoms, extends beyond the Moon's orbit, reaching up to 630,000 km above Earth's surface.

1.2.3.6 Plasma Sheet & Cross-Tail Current

The plasma sheet is a thin, elongated layer within the magnetosphere of Earth, where a dense and hot plasma exists (see Figure 1.19). It is located at the equatorial plane of the magnetosphere, approximately 6 Earth radii away from the planet's surface. The plasma sheet extends outward, reaching distances of 100 to 200 Earth radii. It contains a high concentration of charged particles, including ions and electrons. The energy levels of these particles within the plasma sheet typically range from 2 to 5 keV for ions and 0.5 to 1 keV for electrons. These energetic particles contribute to the dynamic behavior and various phenomena observed in the magnetosphere.

One significant process that occurs is magnetic reconnection as seen Figure 1.22 with field lines quoted ϕ and ϕ' . This reconnection process leads to the accumulation of plasma and energetic particles in the central region of the magnetotail, forming the plasma sheet. Moreover, as particles move within the plasma sheet and travel towards the polar regions, they can undergo further acceleration. Electrons, in particular, can be accelerated to higher energies ranging from 10 to 100 keV. These accelerated particles play a crucial role in generating polar auroras and contribute to the formation of the outer radiation belt, which consists of particles with MeV (mega-electron volt) energies.

The cross-tail current is a circulating current at the interface between the two magnetospheric lobes. It corresponds to j_{tail} in Figure 1.20 and traverses the plasma sheet we just described. It forms a narrow sheet of current flowing westward and diverging at the magnetopause towards the north and south. Its effects include a decrease in the intensity of the magnetic field within the magnetosphere and at the Earth's surface, with a significantly stronger decrease on the night side. In ideal magnetohydrodynamics (MHD), the electric field and plasma velocity are related by $\mathbf{E} = -\mathbf{v} \times \mathbf{B}$. Consequently, in the tail plasma, where the electric field points from dawn to dusk and the magnetic field points northward, particles are directed toward Earth. A portion of these particles then contribute to the formation of the ring current and radiation belts.

1.2.3.7 Plasmasphere

The *plasmasphere* (see Figure 1.19) extends from the ionosphere to the innermost part of the magnetosphere and is primarily composed of low-energy (~ 1 eV) and dense ($\gtrsim 10^3$ cm $^{-3}$) plasma originating from the ionosphere. It was already known that the plasmasphere exists due to ionospheric processes. The plasmasphere is then regarded as a cold plasma with particle motion dominated entirely by the geomagnetic field, hence (mostly) co-rotating (Helmboldt, 2020). The outer boundary of the plasmasphere, known as the *plasmopause*, is characterized by a significant drop in proton density, varying by several orders of magnitude depending on solar activity.

The location of the plasmopause varies widely based on geomagnetic activity. During periods of high activity ($Kp > 4$, see Section 1.3), the plasmopause is more distinct and closer to Earth (refer to Figure 1.24).

1.2.3.8 Radiation Belts

In the various structures of the magnetosphere that we have discussed (polar cusps, ring current, plasma sheet, and plasmasphere), low-energy particles are greatly influenced by variations in the electric and magnetic fields resulting from reconnections at the magnetopause and in the

tail. However, there are high-energy particles that remain unaffected by the energy exchanges and become trapped in specific regions of the magnetosphere known as the *radiation belts*.

The discovery of the radiation belts dates back to February 1958 with the Explorer I mission and its Geiger-Müller instrument, followed by subsequent missions such as Explorer III and Pioneer III. At that time, it was understood that there were two radiation belts as in Figure 1.23, namely the inner belt and outer belt. Observations and measurements quickly provided insights into their main characteristics, which are summarized in Table 1.2. Subsequent missions, such as the *Van Allen Probes*, have revealed an extremely complex and variable structure of the radiation belts.

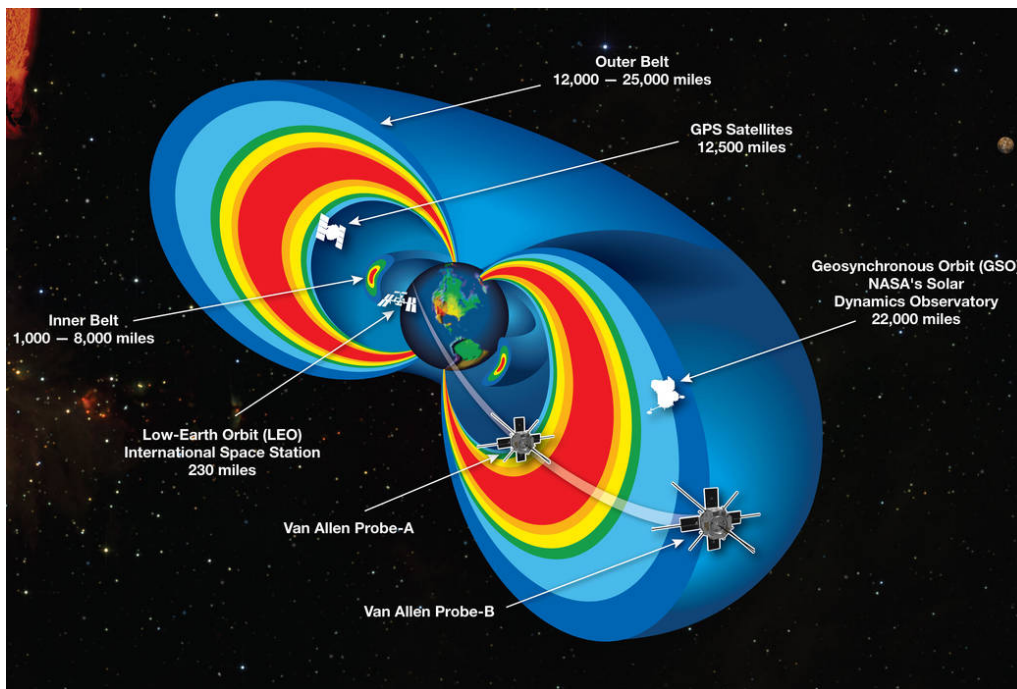


Figure 1.23 – Cutaway model of the radiation belts with the 2 Van Allen Probes satellites flying through them. Credit: NASA (NASA, 2013).

In terms of particle population, the inner belt consists of protons and electrons (see Table 1.2), while the outer belt is predominantly composed of electrons. The particles are trapped and experience the three main motions described in Section 1.1.2 and Figure 1.5: the cyclotron motion, the bounce motion and the drift motion. The inner belt maintains a relatively stable population, while the outer belt can undergo significant changes within minutes, varying by several orders of magnitude. The *Van Allen Probes* have observed a few events where the *slot region* (the region between the two belts) was filled with ultra-relativistic electrons, which remained trapped there for several months. Additionally, the plasmasphere, which is influenced by geomagnetic activity (see Figure 1.24), has a significant impact on the dynamics of the outer radiation belt.

The sources of particles for the radiation belts are also complex. Waves in the magnetosphere, such as chorus waves or ULF waves, are known to scatter and energize electrons depending on factors such as particle energy, wave amplitude, and direction of wave propagation (Koskinen and Kilpua, 2022). Section 1.2.5 provides a detailed description of these wave phenomena.

In the inner belt, the primary source of particles is the CRAND mechanism, which stands for *Cosmic Ray Albedo Neutron Decay*. The CRAND effect occurs when cosmic rays (including

GCRs and SEPs) bombard the upper atmosphere. Highly energetic cosmic rays collide with oxygen and nitrogen atoms, producing multi-MeV neutrons. Most of these neutrons will either escape the magnetosphere or interact with the Earth's atmosphere or surface, but a small fraction will decay into protons, electrons, and neutrinos while still within the inner magnetosphere. Neutrinos can escape, but electrons and protons may become trapped by the magnetic field in the inner Van Allen belt through the mirror effect.

In terms of the electron population in the outer radiation belt, we can categorize them into three groups (Koskinen and Kilpua, 2022): seed electrons, relativistic electrons, and ultra-relativistic electrons. *Seed electrons* represent electrons in the medium energy range that mainly originate from substorm injections. The highest energy populations are solely derived from the acceleration of these electrons through chorus and ULF waves. Therefore, seed electrons serve as the exclusive source, or "seed", for the highest energy electrons. The *core population* encompasses electrons with energies ranging from 0.5 to 2 MeV, while electrons with kinetic energies above 2 MeV are referred to as *ultra-relativistic electrons*.

	Proton Inner Belt	Electron Inner Belt	Outer Electron Belt
Population	Protons	Electrons	(1) Seed electrons (2) Relativistic, core electrons (3) Ultra-relativistic electrons
Location	1.1 to 3 RE	1.1 to 2 RE	3RE to 7-10 R
Energy Range	10 to 100 MeV, and up to 1-2 GeV	30 keV - 200 keV	(1) 200 keV - 500 keV (2) 500 keV - 2 MeV (3) >2MeV
Source	CRAND effect	Substorms, global convection	(1) Substorms, global convection (2) Acceleration by chorus waves, inward transport by ULF waves (3) Acceleration by chorus waves, inward transport by ULF waves

Table 1.2 – Inner and outer belts characteristics

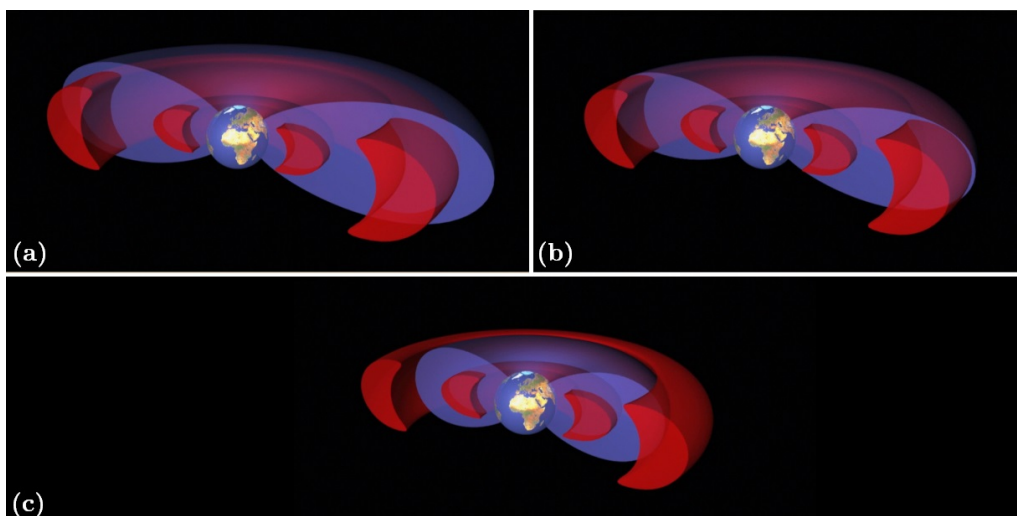


Figure 1.24 – Radiation belts (red) and plasmasphere (blue) under low (a) moderate (b) and high (c) geomagnetic activity. From ESA website - (Carreau, 2013).

1.2.4 Geomagnetic Storms and Substorms

Geomagnetic storms and substorms serve as important indicators of geomagnetic activity but exhibit distinct causes and characteristics. Geomagnetic storms are triggered by phenomena such as coronal mass ejections (CMEs), solar flares, or high-speed solar wind streams. Their effects are observed through measurements of disturbances in the horizontal component of the Earth's magnetic field at the equatorial region, quantified by the Disturbance Storm-Time (Dst) index. In fact Geomagnetic storms are characterized by perturbations at low latitudes. In contrast, substorms primarily arise from the broader influence of the solar wind and manifest as disturbances in the horizontal magnetic field component caused by electric currents in the ionosphere, represented by the auroral electrojet (AE), measured at high latitudes. Both storms and substorms predominantly result from magnetic reconnection processes and involve similar energy transfer mechanisms, albeit occurring at different spatial and temporal scales (Lakhina et al., 2006).

Geomagnetic storms occur during intense solar events (e.g., CMEs, high-speed streams) mostly together with a southward-oriented interplanetary magnetic field (Kamide, 1992; Lakhina et al., 2006). The primary mechanism driving these storms is the reconnection process at the magnetopause, whereby energy transfer is significantly amplified by the southward IMF component, as elucidated in Section 1.2.3.3. Consequently, a substantial amount of plasma is injected into the inner magnetosphere from the magnetotail (Dungey, 1961), leading to intense auroral activity at high latitudes on the nightside (refer to Section 1.2.7). As previously discussed in Section 1.2.3.5, protons move westward while electrons move eastward upon arrival from the magnetotail, resulting in the formation of a ring current. This current system becomes highly charged with particles, exerting a considerable influence on the near-equatorial ground-level magnetic field. Geomagnetic storms are characterized by a distinct pattern observed in the measurement of the horizontal component of the magnetic field (see Figure 1.25): a rapid increase known as the *sudden storm commencement* (SSC), followed by a significant drop referred to as the *main phase*. The subsequent *recovery phase*, spanning several days, signifies the gradual dissipation of excess energy until the system returns to its pre-storm state (Figure 1.25). It is worth noting the SYM-H index also exists, and bears similarity to Dst but provides a temporal resolution of one minute, as opposed to one hour. According to Wanliss and Showalter (2006), it should "be used as a de facto high-resolution Dst index".

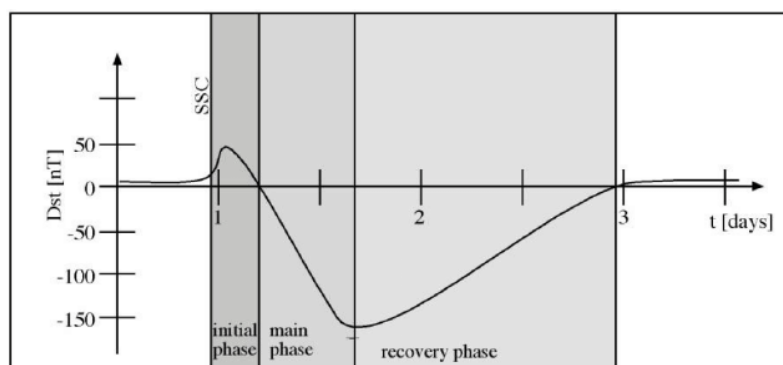


Figure 1.25 – Characteristic signature of a geomagnetic storm in the Dst index [nT] (Amory-Mazaudier et al., 2017).

On the other hand, substorms typically persist for durations ranging from one to several hours (Lakhina et al., 2006), are initiated in the nightside magnetosphere on closed magnetic field lines (Liou et al., 2018). They usually arise from the release of stored energy in the form of particles in the magnetotail, with energies spanning from 5 to 50 keV. This release is accompanied by a fast

plasma flow ranging from 100 to 1000 km/s. The dissipation of this energy manifests as the formation of auroras on the nightside. Substorms are intimately linked to auroral electrojets, which will be further explored in the context of the coupling between the ionosphere and magnetosphere. Two primary theories regarding the triggering of substorms have been proposed (Lui, 1991): the Near-Earth neutral line (NENL) model and the cross-tail current disruption (CD) model. The NENL model suggests that substorms are triggered by magnetic field reconnection in the midtail region (Hones, 1973; Liou et al., 2018; Shiokawa et al., 1998). The CD model suggests that they are triggered by plasma instabilities, such as the cross-tail current instability, in the near-Earth tail resulting in cross-tail current disruption (Liou et al., 2018; Lui, 1991). For a more comprehensive understanding, readers are encouraged to refer to the review by Sandhu et al. (2019).

It is important to emphasize that both phenomena involve multiple aspects of the coupling between the solar wind, magnetosphere, ionosphere, and thermosphere. Currently, no single dataset or model can comprehensively describe these phenomena (Sandhu et al., 2019). Both storms and substorms leave identifiable signatures in ground-based magnetic observations. Furthermore, they have adverse effects on ground power networks, can induce communication blackouts, and pose risks to satellite systems (see Section 1.3.3).

1.2.5 Plasma Waves in the Magnetosphere

We have presented the particle motions within our magnetosphere, specifically the adiabatic invariants, under certain conditions. However, what happens when these conditions are not met? Perturbations and non-adiabatic behaviors in the radiation belts primarily arise from interactions between particles and plasma waves present within the magnetosphere, which facilitate the transfer of energy and momentum. Magnetic storms, as discussed previously, serve as an excellent example of disturbances in the outer belt. They are the main source of a global decrease in the Earth's magnetic field through injections into the ring current. The resulting reduction in magnetic field intensity impacts the adiabatic drift of particles, leading to belt motion, energy loss, particle diffusion into the loss cone, and subsequent precipitation into the atmosphere, among other effects. The following section will present plasma waves, their origin and impacts on particles mainly based on the work by Kletzing et al. (2013); Koskinen and Kilpua (2022).

Plasma waves in our magnetosphere serve multiple purposes, including achieving a balance between the injection and loss of high-energy particles. These waves originate from various mechanisms: external sources, such as VLF transmitters, lightning strokes, or interplanetary shocks impacting the magnetopause, or internal sources, such as plasma instabilities arising from unstable particle distributions or magnetic field configurations. For example, the THEMIS mission has demonstrated that under repeated solar assaults, our magnetopause vibrates like a drum (Johnson-Groh, 2019). Waves can act as both sources and losses of particles, where accelerating electrons can be seen as a loss of low-energy electrons and a source of high-energy electrons. The main wave-particle interaction mechanism is resonant scattering, where resonances provide energy and can increase or decrease the parallel or perpendicular energy of particles. We will now distinguish four main resonances:

- **Gyro resonance:** Resonance of the wave with the particle's gyro-frequency, which alters the particle's momentum.
- **Landau resonance:** Particles traveling at speeds close to the phase speed of the wave can interact with it and exchange energy, akin to a surfer riding a wave. Particles with speeds above the phase speed slow down and transfer energy to the wave, while particles with speeds below accelerate and gain energy from the wave. This resonance is influenced by an electric field parallel to the magnetic field, which either accelerates or decelerates the particle's parallel velocity.

- **Bounce resonance:** Resonance of the wave at the particle's bounce frequency between two mirror points.
- **Drift resonance:** Resonance with the particle's drift frequency around Earth.

Waves in the inner magnetosphere can be regarded as simple oscillations of the particle-field mixture and are observed from both ground-based and space-based platforms. They span a range of frequencies, from ultra-low-frequency (ULF) oscillations in the millihertz (mHz) range to very-low-frequency whistler-mode emissions in the kilohertz range. ULF waves, with periods in the tens of minutes, violate the third adiabatic invariant, leading to radial diffusion and the potential for energetic electron acceleration or loss during storm conditions. Higher-frequency extremely low frequency (ELF) and very low frequency (VLF) waves violate the first two invariants, resulting in pitch angle scattering loss to the atmosphere or local stochastic energy diffusion. During storm conditions, all three adiabatic invariants can be simultaneously violated, necessitating multidimensional diffusion models to differentiate between source and loss processes. The basic characteristics of these waves can be described using plasma theories, including cold plasma theory, magnetohydrodynamics (MHD), or Vlasov theory. We will not go into the details of the mathematical equations here but will present the dominant wave modes found in the radiation belts, as seen Figure 1.26:

- **Whistler-mode chorus waves:** these waves are short, right-hand polarized emissions in the kHz range. We can find them around the equator and outside the plasmasphere up to the magnetopause. They are driven by anisotropic electron populations in the energy range 1-100 keV that have been injected from the magnetotail. Hence, we find them in the midnight to dawn sector as electrons drift eastwards. They propagate away from the equator but are attenuated before reaching the ionosphere, mainly through wave-particle interactions (probably Landau damping by suprathermal electrons around 1 keV) (Koskinen and Kilpua, 2022). These waves can interact with particles through gyro and Landau resonances. Gyro resonance will scatter $\lesssim 100$ keV electrons around the equator, and MeV electrons at higher latitudes ($\lambda \lesssim 15^\circ$), leading to pitch-angle diffusion toward the atmospheric loss cone. This resonance will also break the first adiabatic invariant, accelerating electrons from a few hundred keV to MeV energies. Landau resonance will interact with 30 keV to MeVs electrons (Koskinen and Kilpua, 2022).
- **EMIC waves:** *Electromagnetic ion cyclotron waves* or *Alfvén ion cyclotron waves* also play an important role in the loss of ultra-relativistic radiation belt electrons, and in heating ions. They are observed in the afternoon sector close the plasmopause and beyond. They are left-hand polarized and their frequency range is 0.2 to 5 Hz, close to the ions' gyro-motion frequency, and they are driven by anisotropic proton populations in the energy range 1-100 keV that have been injected from the magnetotail. They interact with particles through gyro, Landau and bounce resonances. During storm time, EMIC waves lead to electron pitch-angle diffusion but do not accelerate them. The gyro resonance affects $\gtrsim 1$ -2 MeV electrons with pitch angles between 30 and 70° . Landau resonance affects 30 keV to MeVs electrons with $\gtrsim 85^\circ$ pitch angles and bounce resonance affects 50-100keV electrons with $\gtrsim 85^\circ$ pitch angles.
- **Hiss waves:** *Plasmasphere hiss waves* are one of the main wave modes inside the plasmasphere and can be found mostly on the dayside (dawn to post-noon, see Figure 1.26). They have a large range of frequencies but mostly interact with radiation belt electrons at frequency below 100 Hz. Hiss waves play a crucial role in electron scattering to the loss cone and the formation of the slot region. They are thought to originate from interaction between particles from the radiation belt and the plasmasphere, or from terrestrial lightning strikes, but their origin remains unclear. They interact with electrons through gyro, Landau and bounce resonances. The gyro resonance affects ~ 100 keV to meV electrons with pitch angles between 30 and 70° . Landau resonance affects 100 keV electrons with pitch angles

between 30 and 70° and bounce resonance affects $\gtrsim 1$ MeV electrons with pitch angles between 30 and 70°.

- **Magnetosonic waves:** *Equatorial magnetosonic noise* are the second main wave modes inside the plasmasphere but can occur inside and outside of it. These waves can resonate with energetic electrons through gyro, Landau and bounce resonances and transfer energy from ring current protons to radiation belt electrons.
- **ULF waves:** The *Ultra-Low Frequency waves* discussed here have frequencies ranging from 2 to around 20 mHz, which are below the local ion gyro frequency. These waves are generated at the magnetopause boundary due to velocity shear, solar wind pressure fluctuations, or inherent plasma instability. ULF waves induce radial diffusion transport, causing changes in the energy distribution of trapped particles. The rate of radial transport depends on the power spectral density of the waves, with faster transport typically observed in the outer magnetosphere (Kletzing et al., 2013).

In table 1.3, we summarized all the info gathered on wave modes and their effect on the particles in the inner magnetosphere. This section was a very broad view to help the reader understand the vocabulary associated. But we encourage to take a look at references concerning this topic, including the two used to build this section (Kletzing et al., 2013; Koskinen and Kilpua, 2022).

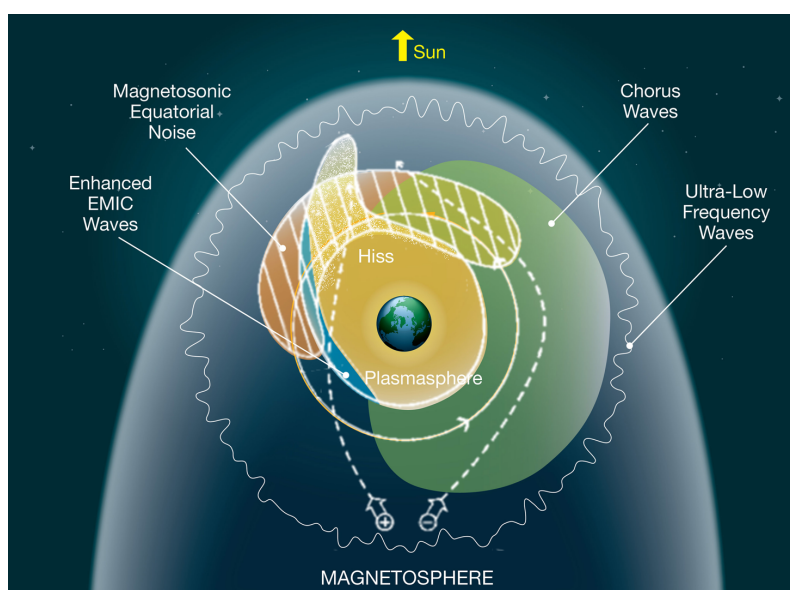


Figure 1.26 – Plasma waves and their region in the inner magnetosphere. Adapted from both NASA’s Goddard Space Flight Center/Mary Pat Hrybyk-Keith image and Kletzing et al. (2013).

Finally, we will briefly introduce the broader category of Alfvén waves, which play a crucial role in the electron acceleration process. Alfvén waves typically have frequencies ranging from a few millihertz to a few hertz and propagate purely as shear waves along the magnetic field lines. When we refer to Alfvén waves, we are primarily describing waves with a dominant shear (transverse) component, although they do not necessarily require a purely transverse component. One way to envision these waves is to imagine vibrations on a guitar string connected to the Earth on both ends. Alfvén waves can originate from various phenomena, including solar activity, reconnection processes, plasma instabilities, particle-wave interactions, and even seismic activity or volcanic eruptions and they propagate perturbations along the oscillating field lines. They globally affect particles of the inner magnetosphere through heating, acceleration, scattering, trapping or loss.

Wave mode	Frequency	Region	Direction of propagation	Origin	Resonances	Main effects
Whistler-mode chorus	0.5-10 kHz (ULF & VLF)	Midnight to dawn (even early dusk) Outside the plasmasphere Up to the magnetopause	Along magnetic field lines near equator and obliquely at higher latitudes	Convective injection of plasma sheet electrons leading to anisotropic distributions of 1-100 keV electrons	Gyro & Landau	Dominant scattering process leading to diffuse auroral precipitation Loss of radiation belt electrons Local acceleration of radiation belt electrons
EMIC	0.1-5 Hz (ULF)	Noon to dusk Dayside outside the plasmasphere up to magnetopause	Along magnetic field lines near equator and obliquely at higher latitudes	Ring current ions injection (during magnetic storm) leading to anisotropic distributions of 1-100 keV protons	Gyro, Landau & Bounce	Rapid scattering and loss of ring current ions Rapid scattering and loss for electrons >0.5 MeV
Plasmaspheric hiss	\lesssim 100 Hz (ULF) & from 100 Hz to several kHz	Dawn to post-noon Inside the plasmasphere only	Along magnetic field lines near equator and obliquely at higher latitudes	Local generation Subset of chorus waves avoiding Landau damping	Gyro, Landau & Bounce	Formation of the quiet time electron slot between inner and outer radiation belts
Equatorial magnetosonic noise	few Hz to few 100h Hz (ULF)	Noon to dusk Inside and outside the plasmasphere	Perpendicular to magnetic field lines	Changes in ion ring distributions	Landau & Bounce	Scattering of electrons ~30 keV - 1 MeV through Landau resonance
ULF Pc5 & Pc4	2-20 mHz	Global, most frequent in the dawn and dusk sector	Can exhibit all directions, from field-aligned to perpendicular	Excited at the magnetopause in response to velocity shear and or solar wind pressure fluctuations	Drift	Radial diffusion transport and change of energy in the population of trapped particles

Table 1.3 – Summary of frequencies, regions, sources and main impacts of different wave modes in the inner magnetosphere. Table built using *Koskinen and Kilpua (2022)* and *Kletzing et al. (2013)*

1.2.6 Ionosphere

This section is dedicated to the ionosphere, which is crucial for understanding the concepts related to polar auroras. We will begin by introducing the Earth's atmosphere to provide context for the ionosphere. Then, we will describe the ionosphere itself. Finally, we will explore the various currents present in the ionosphere, which play a vital role in the ionosphere-magnetosphere coupling.

1.2.6.1 Introduction to the Earth's Atmosphere

In order to discuss the ionosphere, it is important to familiarize ourselves with the different layers of the Earth's atmosphere:

- **Troposphere:** This is the first layer starting from the surface and it extends up to approximately 12 km in altitude. The troposphere is the densest layer of the atmosphere, and the temperature decreases by about 70 degrees until it reaches its upper boundary, known as the tropopause.
- **Stratosphere:** The stratosphere is the second layer, extending from the troposphere up to around 50 km in altitude. This is where the ozone layer is found. In the stratosphere, the temperature increases, and it reaches near-zero degrees at the upper limit called the stratopause.
- **Mesosphere:** The mesosphere is the layer situated between approximately 50 and 80 km in altitude. The temperature gradually decreases in this region and reaches about -80°C (190 K), making it the coldest part of the Earth's system.
- **Thermosphere:** This layer is located just before the exosphere and can extend up to 700 km, depending on the exobase boundary. The temperature in the thermosphere increases rapidly between 80 and 200 km and then stabilizes at what is known as the "thermospheric temperature." This temperature varies between 750 K during periods of low solar activity and 1500 K during periods of high activity (Gruet, 2018). The thermosphere is predominantly composed of atomic oxygen, along with molecular oxygen and nitrogen. It is within this layer that polar auroras can be observed.
- **Exosphere⁴:** The exosphere is the outermost layer of the atmosphere, spanning from approximately 500 (the exobase) to 10,000 km. The atoms in this region are sparsely distributed, and particles tend to escape into space. At the lower part of the exosphere, it is also possible to observe polar auroras.

Above a boundary known as the turbopause, located at around a hundred kilometers in altitude and extending into space, the components of the atmosphere stratify under the influence of gravity, forming what is referred to as the *heterosphere*. The lowest layer of the heterosphere consists of nitrogen, followed by oxygen, helium, and finally hydrogen. Below the turbopause, down to the surface, it is known as the *homosphere*.

1.2.6.2 Ionosphere

In certain regions, the components of the heterosphere can become ionized, leading to the formation of the ionosphere. There are two main processes of ionization:

- **Photoionization**, which occurs when solar emissions in the UV and EUV range ionize the particles.

4. Note that, in general, we categorize atmospheric zones based on their thermal properties, such as Troposphere, Stratosphere, Mesosphere and Thermosphere, and their fluid properties, such as Homosphere, Heterosphere, and Exosphere

- Particle precipitation in polar regions, where particles are injected along the magnetic field lines, often due to storms or geomagnetic substorms. As explained earlier in this chapter (see Section 1.1.2.4), some particles can enter the loss cone through various impacts and "precipitate" into the upper atmosphere by colliding with the molecules present there.

These two processes create distinct regions in the ionosphere, which exhibit differences between day and night. Figure 1.27 and Table 1.4 provide an overview of the presence and characteristics of these layers. During the day, photoionization prevails, exciting particles and atoms in the thermosphere, resulting in the formation of the D, E, F1, and F2 layers, also known as the "Chapman layers." On the other hand, during the night, particle precipitation from the magnetotail leads to ionization, excitation, and heating through the Joule effect. This gives rise to the E and F layers.

Region	Altitude range [km]	Peak altitude [km]	Electron density [m^{-3}]	Recombination coefficient m^3s^{-1}	Major components	Ionization source
D	50-90	75	10^2 (night) to 10^9 (day)	10^{-14}	NO^+ , O_2^+	Solar Lyman alpha (121.5 nm) and hard solar x-rays (1nm)
E	90-150	120	2×10^9 to 10^{11}	5×10^{-14}	NO^+ , O_2^+	Solar x-rays (1-10 nm) and solar UV (80-102.7 nm)
Es	95-105	100	$1-2 \times 10^{11}$	5×10^{-14}	NO^+ , O_2^+	Precipitation electrons and meteorites
F1	120-200	180	— to $2-5 \times 10^{11}$	5×10^{-15}	NO^+ , O_2^+ , O^+	EUV (10-100 nm)
F2	200	300-350	$2-5 \times 10^{11}$ to $1-2 \times 10^{12}$	10^{-16}	O^+ , N^+ , H^+	EUV (10-100 nm)

Table 1.4 – Characteristics of the various ionospheric regions. From *Pisacane (2008)*

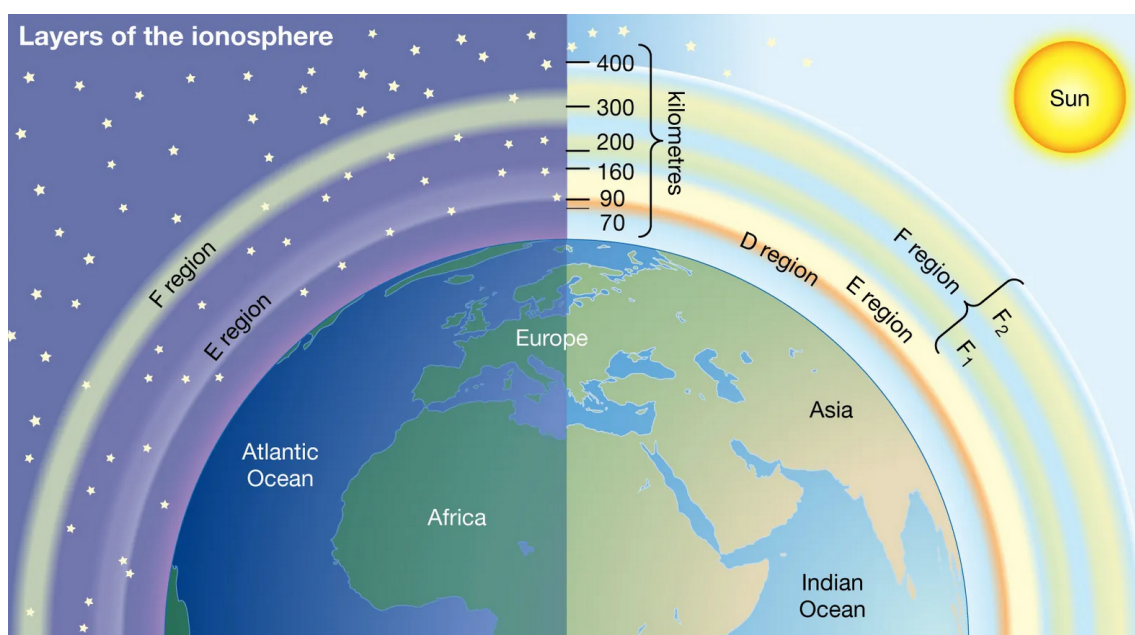


Figure 1.27 – Layers of the ionosphere in our atmosphere by the *Encyclopaedia Britannica, Inc (Britannica, 2012)*.

1.2.6.3 Field Aligned Currents

The field-aligned currents (FACs) were first suggested to connect the ionosphere to space (Birkeland, 1908) and were later observed by satellites (Cummings and Dessler, 1967; Zmuda et al., 1966) through observations of magnetic field variations. A very complete review has recently been made by Palmroth et al. (2021). Particles are accelerated from the magnetotail along magnetic field lines, scattered by plasma waves, subject to complex trajectories and can end up in the loss cone, precipitating in the ionosphere. FACs connect together magnetospheric regions with different controlling parameters, like the plasma sheet or the low-latitude boundary layer, to the auroral zone in the ionosphere (see Figure 1.28a). FAC structures, consisting of upward and downward currents with planar or filamentary geometry (Boström, 1964). The Figure 1.29 shows the full three-dimensional structure of the FACs. The black lines represent the convection from the Dungey cycle, seen when the IMF is oriented southward.

There are essentially two main regions of upward and downward currents called region 1 (R1) and region 2 (R2), as seen in Figure 1.29, based on their location relative to the polar cap (Iijima and Potemra, 1976). R1 is located at higher latitudes (poleward) and R2 at lower latitudes (equatorward). These two regions form thick concentric rings around each magnetic pole (Figure 1.28b). The currents exhibit opposite signs between the two regions and on both sides of the midnight-midday axis. R1 downward in the dawn sector and upward in the dusk sector and R2 flows in the opposite direction. Although they originate from different sources, these regions converge within the ionosphere, creating a circuit that serves as the core of the magnetosphere-ionosphere coupling. R1 is considered flowing along magnetic field lines closing in the flank (low-latitude) magnetopause while R2 is often considered to flow along magnetic field lines closing in the ring current (Figure 1.28a). However, the location of generation regions for these currents is still investigated today (Ebihara and Tanaka, 2022).

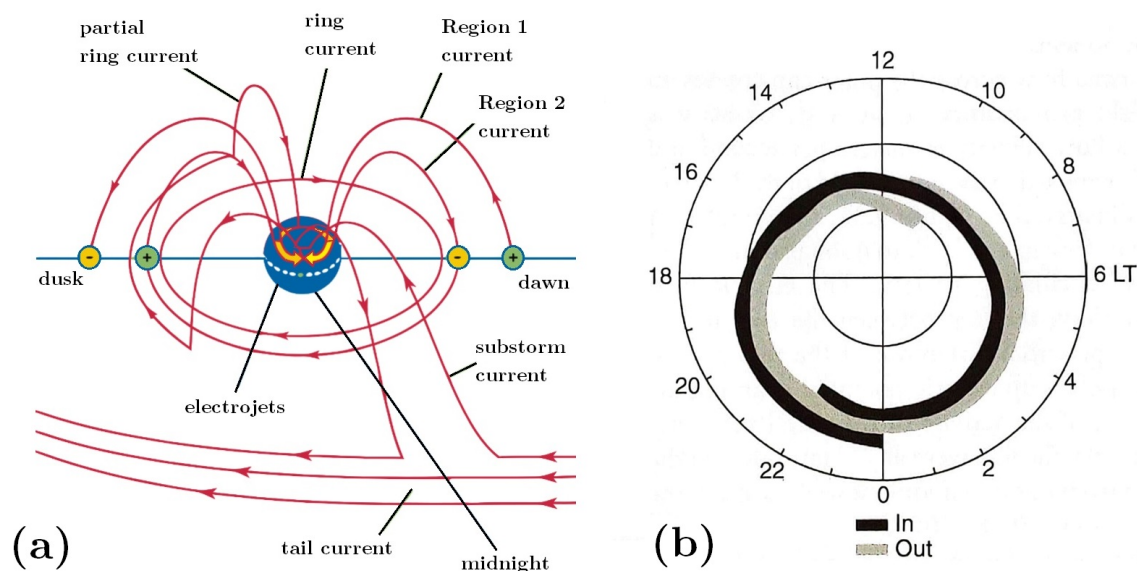


Figure 1.28 – (a) Field-aligned currents system from *Britannica* (1994), (b) Currents towards and away from the ionosphere, from *Russell et al.* (2016).

FACs close in the high latitude ionosphere forming the Pedersen and Hall currents. Pedersen currents close the FACs both from R1 currents to R2 currents in the pole-to-equator direction, and in the polar cap from downward R1 to upward R1. Anisotropic magnetized medium generates another ionospheric current known as Hall currents or electrojets. In the presence of orthogonal

electric field (\mathbf{E}) and magnetic field (\mathbf{B}) (as observed between the field lines and the Pedersen current), electrons and ions drift in the same direction, perpendicular to \mathbf{E} and \mathbf{B} (known as the $\mathbf{E} \times \mathbf{B}$ drift). However, within the ionosphere, ions experience more collisions with neutral atoms, causing them to drift slower compared to electrons, which have fewer collisions. As a result, a current is generated in the opposite direction to the particle drift, known as the Hall current. Thus, two counter-rotating Hall ionospheric current cells are formed (Figure 1.29). These cells flow from midnight to noon in the polar cap, westward on the morning side, and eastward on the evening side, and are referred to as the auroral electrojets.

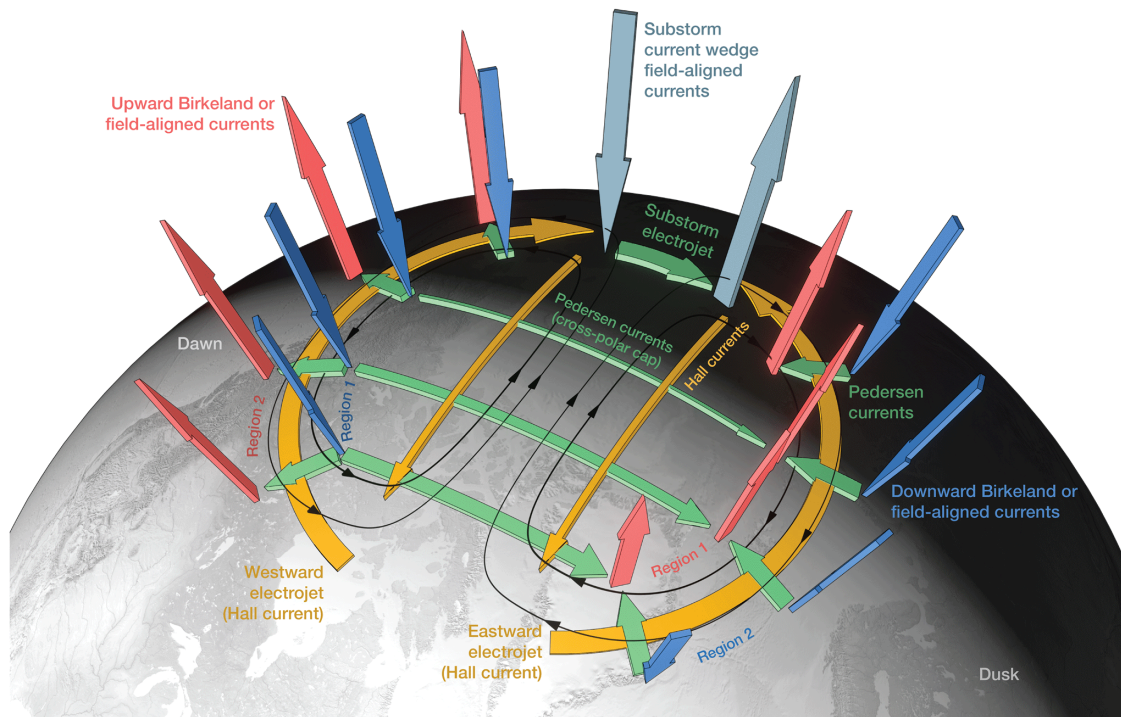


Figure 1.29 – Overview of the field-aligned current systems, showing Hall and Pedersen currents. Figure from Palmroth et al. (2021).

These currents are measured using 12 high-latitude ground-based stations (around 60° latitude). These stations employ magnetometers with a resolution of 1 minute to capture the "auroral electrojet index" (AE) and its components (AL, AU) in nanotesla. This index represents the magnetic signature of eastward and westward auroral electrojets in the northern hemisphere. For more detailed information, readers are advised to refer to the International Service of Geomagnetic Indices⁵. These currents play a crucial role in quantifying geomagnetic activity specifically in the auroral region, particularly during geomagnetic substorms.

As presented by Palmroth et al. (2021) smaller-scale FAC current systems can also be observed, especially within auroral arcs (Partamies et al., 2010). Within the context of auroral arcs, an upward FAC, primarily composed of precipitating electrons, is accompanied by a downward FAC sheet (Elphic et al., 1998), 1998). This mechanism can accelerate electrons downward precipitating them in the ionosphere and creating auroras (see 1.2.7.1). Multiple arcs can comprise numerous pairs of upward and downward FACs, while also exhibiting a distinct shared current system, with all the arcs aligned along the upward FAC leg (Palmroth et al., 2021; Wu et al., 2017).

5. <https://isgi.unistra.fr>

We have provided a brief overview of the fundamental mechanisms involved in the coupling between the magnetosphere and the ionosphere. Moving forward, we will delve into the finer details of polar auroras and conclude by discussing their variations during geomagnetic substorms, along with the current modeling approaches available today.

1.2.7 Auroral Physics

Auroras represent the final stage and visible outcome of a complex chain that connects solar wind, the magnetosphere, and the atmosphere. In this section, we will explore the range of phenomena that give rise to these mesmerizing light displays, as well as their distinctive characteristics. To begin with, we will delve into the primary particle acceleration effects and the resulting classification of auroras. It's worth noting that these particle accelerations are intricately linked to ionospheric currents, which we have discussed in Section 1.2.6.3. We will then provide a broader overview of the different types of auroras, including discrete and diffuse auroras, and their associated emissions. Finally, we will examine the morphology of auroras and their dynamic evolution, particularly during geomagnetic substorms.

1.2.7.1 Related Particle Acceleration Processes

As we saw when we discussed the existence, origins and impacts of plasma waves in the inner magnetosphere, particles can be accelerated through several processes and end up in the loss cone, hence precipitating in the Earth's ionosphere, producing auroras. Usually, we distinguish several type of auroras based on the related precipitation mechanisms in Earth's auroral zone. We usually consider that three precipitation mechanisms predominate: *quasi-static potential structure* (QSPS, also called *inverted-V* or *monoenergetic*) acceleration, *Alfvénic* (or *broadband*) acceleration, and *wave scattering* of plasma sheet electrons into the loss cone (also called *diffuse* precipitation) (Dombeck et al., 2018). Here are short introductions to these mechanisms:

- The first mechanism considered is the *potential drop* or *quasi-static potential structure* (QSPS). The resulting auroras are called *inverted-V auroras* (1.2.7.2). They correspond to regions of upward field-aligned region carried by electrons that have been accelerated through an electrostatic potential (Russell et al., 2016). The parallel acceleration due to parallel electric field then enhance the energy flux of precipitated electrons. A QSPS with a large potential drop usually results in a narrow intense electron energy spectrum that appears almost monoenergetic (Dombeck et al., 2018).
- The second mechanism considered is the *Alfvénic* or *broadband* acceleration. Alfvénic acceleration occurs when the electric field of the wave has a component parallel to the magnetic field, hence accelerating electrons with parallel velocities similar to the local Alfvén speed (Dombeck et al., 2018). Acceleration by Alfvén waves is dependent on the relative velocities of the waves and particles. Since electrons are not uniformly accelerated, it leads to a broad and often intense energy spectrum. The resulting auroras are called *alfvénic auroras* (1.2.7.2).
- The third mechanism is the wave scattering of plasma sheet particles into the loss cone, generally by chorus whistler waves as discussed in 1.2.5. Diffusion through whistler mode is associated with an energy spectrum that closely resembles a Maxwellian distribution and is comparatively less intense. The resulting auroras are the *diffuse auroras* 1.2.7.

Acceleration of particles in the magnetosphere occurs through various processes, leading to precipitation in the Earth's ionosphere and the production of auroras. Here we have only discussed three predominant precipitation mechanisms: quasi-static potential structure (QSPS) acceleration, Alfvénic acceleration, and wave scattering of plasma sheet electrons. These mechanisms result

in inverted-V auroras, Alfvénic auroras, and diffuse auroras, respectively. In the following sections, we will shortly present the different types of auroras and their emissions and end up with a summary of these two Sections [1.2.7.1](#) and [1.2.7.2](#).

1.2.7.2 Auroral Types

Even if we will not delve into detailed descriptions of auroras (as our study focuses solely on the measurement of precipitated particles without considering the specific types of auroras they produce) we would like to highlight a few key points that might be of interest to the reader:

- The most significant consequence of auroras is the enhancement of ionospheric electric currents. The associated magnetic fields can be readily measured using magnetometers. The strength of these electric currents, in turn, can be utilized to quantify the magnitude of geomagnetic activity during a geomagnetic storm.
- The emissions are predominantly concentrated around an altitude of 100 km. Generally, the arcs exhibit an east-to-west orientation, but they can exhibit complex twisting patterns.
- *Discrete* auroras are characterized by highly intense arcs that can be narrower than 100 m ([Hui and Seyler, 1992](#); [Knudsen et al., 2021](#)). They originate directly from ionization and excitation caused by accelerated electrons through two mechanisms:
 - Alfvén waves, wherein electrons with energies between 1 and 5 keV generate dynamic auroras and arcs spanning approximately 1 to 10 km. These phenomena are commonly referred to as *Alfvénic auroras*. The acceleration of electrons is believed to occur due to the parallel electric field resulting from the interaction of electrons with Alfvén waves ([Hui and Seyler, 1992](#)).
 - Quasi-Static Potential Structure (QSPS), which involves lower-energy electrons (5 to 10 keV) producing more stable waves (than alfvénic ones - [Karlsson et al. \(2020\)](#)) with arcs on the order of 100 km. The characteristic pyramid-shaped signature in energy fluxes leads to their designation as *inverted-V auroras*.
- The *diffuse* auroras ([Nishimura et al., 2020](#)) can extend over widths exceeding 100 km in the north-south direction and stretch up to 1000 km in the east-west direction along magnetic field lines. These auroras lack well-defined structures and are associated with the *whistler mode* discussed in Section [1.2.5](#). The electrons responsible for the diffuse auroras primarily originate from the plasma sheet with energies between 100 eV and 10 keV ([Ni et al., 2016](#)).
 - A subset of diffuse auroras is the *pulsating aurora*, characterized by visible flickering patches lasting approximately 2 to 20 seconds, with series lasting several tens of minutes. The electrons responsible for these auroras have energies ranging from 1 to 100 keV. Pulsating auroras thickness varies from 10 to 200 km, and emissions are predominantly green. They are typically located in the equatorward part of the auroral oval.
- The colors of the emissions directly correlate with the types of ionized molecules:
 - Oxygen atoms emit mainly yellow-green light (557.7 nm) and, at higher altitudes, red light (630.0 nm and 636.4 nm).
 - Nitrogen molecules N_2 emit mainly in the dark red light (first positive band: 650-680 nm).
 - N_2^+ ions emit dark blue (391.4 nm), violet (427 nm) and intense red-near infrared (Meinel Band ([Piper et al., 1986](#))) light.

As a conclusion, presented below is table [1.5](#) summarizing the aurora types discussed.

Auroras	Acceleration process	Electron energy range	Size	Energetic distribution
Discrete	QSPS	5-10 keV	≥ 100 km	Monoenergetic
Discrete	Alfvén waves	1-5 keV	$\sim 1 - 10$ km	Broad and intense spectrum
Diffuse	Whistler mode	~ 100 eV - 10 keV	can reach ≥ 100 km wide can reach ≥ 1000 km in the east-west direction	Maxwellian
Pulsating	Whistler mode	few keV - 100 keV	10 - 200 km	Maxwellian

Table 1.5 – Summary of aurora types and their characteristics

1.2.7.3 Auroral Morphology & Models

What about the shape and morphology of polar auroras? We have already mentioned the oval shape of auroras in Section 1.2.6.3, but as one would expect, it evolves with solar activity. Auroras form on both sides of the Earth in ovals centered around the geomagnetic poles (not the geographic poles). On each side, the oval doesn't extend all the way to the pole but stops short. The region without auroras, centered on the pole, is called the *polar cap*. Occasionally, auroras can be observed within this region during *theta auroras* (named after their shape, where a north-south auroral band connects two edges of the oval, passing through the pole). This pattern is typically observed from space, but space-based instruments have the disadvantage of not capturing the fine structures of the auroras. Ground-based instruments compensate for this but have the drawback of intermittent coverage (due to clouds, moonlight, etc.). The shape of the auroral oval varies with geomagnetic activity (assessed through the *Kp* index). For low activity, the oval is generally situated at latitudes close to 75° on the day side and 70° on the night side, but during active periods, it descends to around 70° on the day side and drops below 60° on the night side (Feldstein and Starkov, 1967).

When the morphology of auroras abruptly changes, primarily due to activity in the magnetosphere, we refer to it as *auroral substorms*. They follow the cycle described below (Figure 1.30), first observed by Akasofu in the 1960s (Akasofu, 1968):

- Initially, there is a calm interval where the arcs are not very intense.
- Suddenly, an arc near the equator lights up. This is known as the *substorm onset*.
- It then rapidly moves poleward while expanding westward, referred to as the *westward traveling surge*.
- After several tens of minutes, the auroral activity diminishes, and the substorm enters its *recovery phase*.

To wrap up our exploration of the Sun-Earth chain, let's briefly mention common models for polar auroras or, more generally, for precipitated particles. Full reviews of these models can be found in the literature such as Machol et al. (2012); McGranaghan (2016); McGranaghan et al. (2021); Newell et al. (2015).

- **Feldstein-Starkov:** In 1994, Starkov (1994) started working on an auroral oval model based on the *Kp* index (1.3.1). It is a mathematical framework that relates the *Kp* index to the location of the auroral oval boundaries, especially their latitudes. The model starts by deriving the AL index from the *Kp* index, then uses polynomial equations to obtain coefficients, used

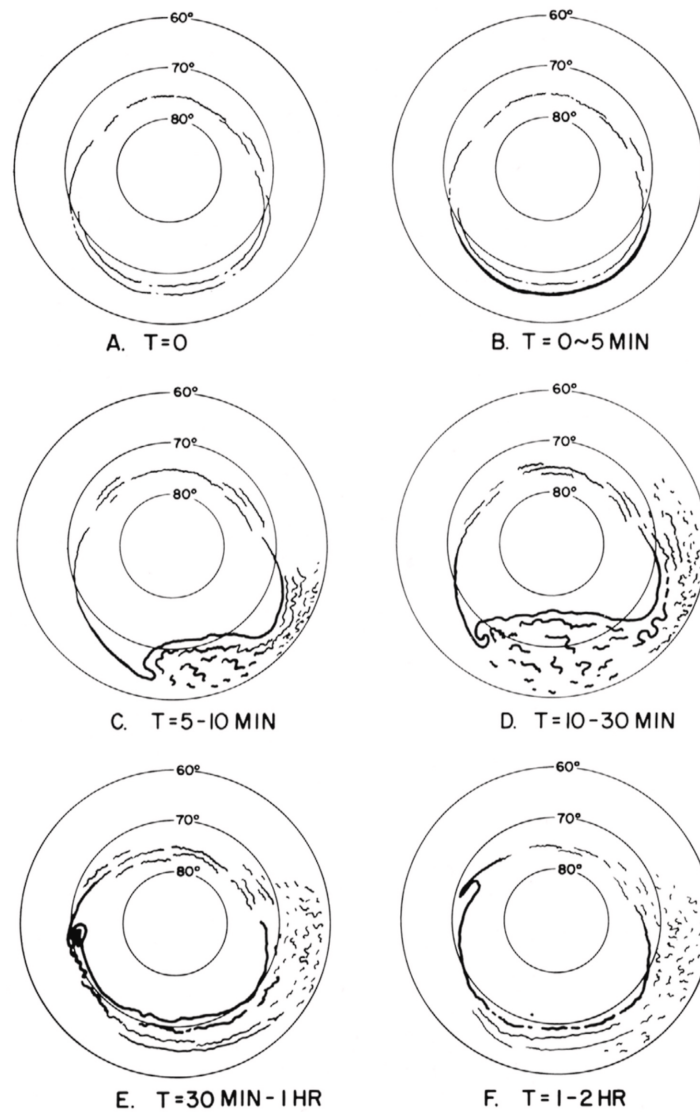


Figure 1.30 – First observation of the cycle of auroral substorm in the pole from *Akasofu (1968)*.

to determine the shape and position of the auroral oval. As the K_p index increases, the model predicts that the auroral oval expands smoothly, indicating higher geomagnetic activity.

- **Zhang-Paxton:** In 2008, [Zhang and Paxton \(2008\)](#) developed an auroral oval model based on measures by the Ultraviolet Imager (GUVI) on board the TIMED (Thermosphere Ionosphere Mesosphere Energetics and Dynamics) satellite. This model uses Epstein functions to calculate either the electron flux or the mean energy flux for precipitating electrons based on data from the GUVI on board the TIMED satellite. The Figure 1.31 from [Sigernes et al. \(2011\)](#) compares simulations from the two models at different K_p levels.

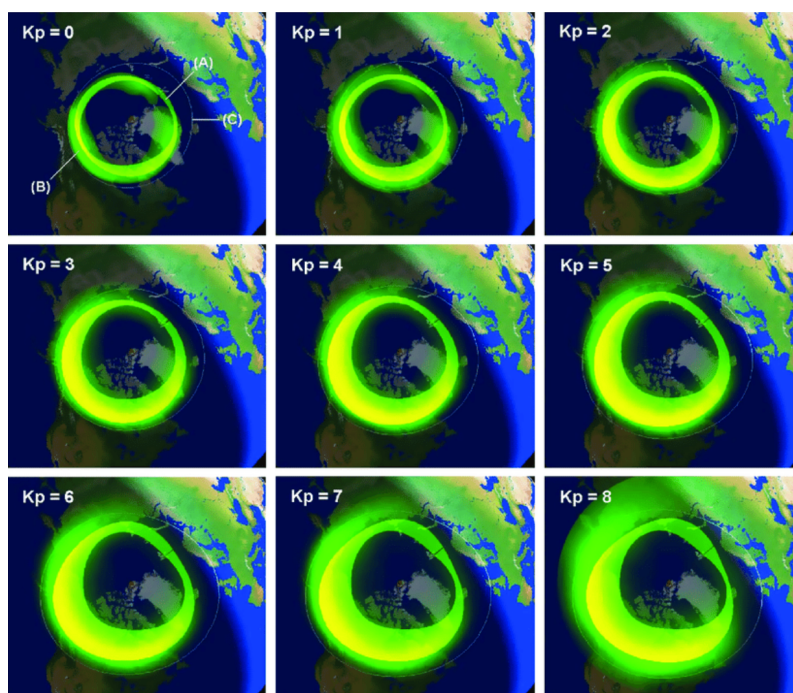


Figure 1.31 – "Animated model aurora ovals as a function of K_p index for 24th December 2009 at 08:50 UT. The transparent polygons represent Feldstein-Starkov ovals (A). The faint white outer ring is the equatorward boundary of the diffuse aurora (C). The Zhang-Paxton ovals are displayed on top with green intensity values scaled according to the electron energy flux (B). The yellow scaled intensity areas are the intersection $(A \cap B)$ between the two models." from [Sigernes et al. \(2011\)](#).

- [Hardy et al. \(1985, 1987\)](#) presented a statistical model of auroral electron precipitation based on the Defense Meteorological Satellite Program F2 and F4 and the Satellite Test Program P78-1 satellites. They determined the statistical characteristics of auroral electron precipitation as a function of magnetic local time, magnetic latitude, and geomagnetic activity as measured by K_p . They did so using a 2° latitude 0.5 h local time grid.
- [Fuller-Rowell and Evans \(1987\)](#): The Polar-orbiting NOAA spacecrafts monitor particle influx into the atmosphere since 1978. They used these data to create statistical patterns of height-integrated conductivities and ionization rates. The observations are organized based on an auroral activity index, providing valuable information for ionospheric and thermospheric research.
- **OVATION Prime:** OVATION (Oval variation, assessment, tracking, intensity, and online nowcasting) Prime is an auroral precipitation model parameterized by solar wind driving, the most used and widely available aurora model today. First introduced in 2002 ([Newell et al., 2002](#)), and improved in 2010 ([Newell et al., 2010](#)), the recent upgraded version comes from [Newell et al. \(2014\)](#). According to it, "distinguishing features of the model include

an optimized solar wind-magnetosphere coupling function which predicts auroral power significantly better than Kp or other traditional parameters, the separation of aurora into categories (diffuse aurora, monoenergetic, broadband, and ion), the inclusion of seasonal variations, and separate parameter fits for each magnetic latitude (MLAT) \times magnetic local time (MLT) bin, thus permitting each type of aurora and each location to have differing responses to season and solar wind input—as indeed they do." We will come back on this model during our study as it will serve as a comparison to the results we will produce when modeling precipitations. Below, Figure 1.32 is an example of OVATION's result as presented on the NOAA website⁶. Machol et al. (2012) evaluates OVATION prime by comparing them with Polar UVI (Horwitz et al., 1998) images and concludes that it provides accurate forecasts 77% of the time.

- PrecipNet (McGranaghan et al., 2021): Researchers compiled 51 years of Defense Meteorological Satellite Program (DMSP) observations to create an improved particle precipitation database. They developed PrecipNet, a neural network that effectively integrated diverse information from solar wind and geomagnetic activity, including their time histories. PrecipNet achieved a significant reduction of over 50% in errors compared to the current state-of-the-art model (considered to be OVATION Prime). It also better captured dynamic changes in the auroral flux and demonstrated effective reconstruction of mesoscale phenomena. Our work in this thesis builds upon the advancements made by the PrecipNet team, using the same comprehensive particle precipitation database.

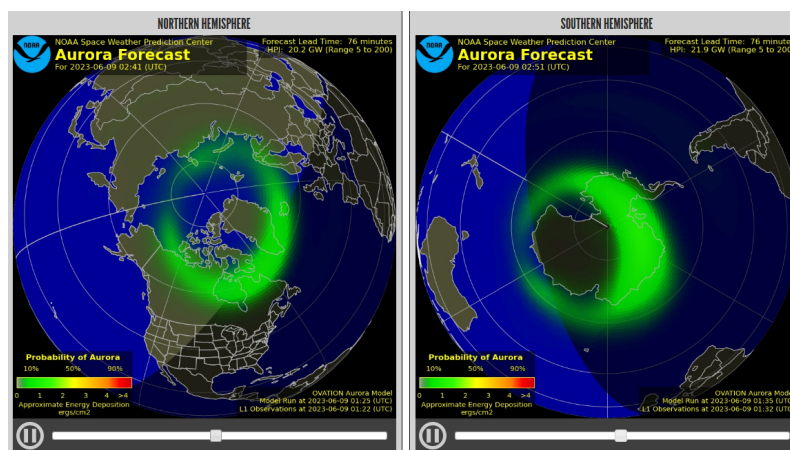


Figure 1.32 – Short-term 30 to 90 minute forecast of the location and intensity of the aurora based on the OVATION model. Credits: NOAA Space Weather Prediction Center; <https://www.swpc.noaa.gov/products/aurora-30-minute-forecast>.

In Part 1.2, we explored the Sun-Earth chain, which encompasses the interconnected phenomena among the Sun, the solar wind, Earth's magnetosphere, ionosphere, and auroras. We began by examining the Sun as the starting point and learned about the solar wind and its interaction with the interplanetary medium. We then delved into the solar wind's interaction with Earth's magnetosphere, discussing geomagnetic storms, substorms, and the role of plasma waves in the magnetosphere. Additionally, we explored the ionosphere and its coupling with the magnetosphere, highlighting its significance in the formation of auroras. Finally, we showed how auroras are the visible manifestations of these complex interactions. These topics provide a foundation for understanding the intricate dynamics of space weather, but they alone do not suffice, and readers are encouraged to delve deeper into the subjects. To conclude our discussion on Space Weather,

6. <https://www.swpc.noaa.gov/products/aurora-30-minute-forecast>

we must now explore the methods used for measuring it over the years and examine its impacts on our systems and everyday life. We will emphasize the role of modeling and forecasting in mitigating these effects, particularly in the context of New Space.

1.3 Space Weather Measures and Forecasts

All the main actors of Space Weather have been introduced, from the Sun to the IMF, the magnetosphere and the ionosphere. A definition of what Space Weather is has been properly given by [Lilensten and Blelly \(2008\)](#) as follows:

“Space weather is the physical and phenomenological state of natural space environments. The associated discipline aims, through observation, monitoring, analysis and modelling, at understanding and predicting the state of the sun, the interplanetary and planetary environments, and the solar and non-solar driven perturbations that affect them; and also, at forecasting and nowcasting the possible impacts on biological and technological systems.”

The Sun’s dynamic and complex activity conditions the properties of electromagnetic fields and space particles and has direct and indirect influences on space and ground-based infrastructures. Geomagnetically induced currents, spacecraft charging, single-event effects, erosion, or even atmospheric drag are examples of the dangerous consequences coming from the Sun. Its activity coupled with the increasing dependence of our society on critical space technologies emphasizes the need to better understand and prevent these risks.

Space weather as a whole poses a danger to us, but how can we quantify it? Most satellites positioned throughout the Sun-Earth chain measure phenomena that are considered part of "space weather". Examples include satellites like Parker Solar Probe, Solar Orbiter, STEREO, and Ulysses near the Sun, satellites like WIND, ACE, DSCVR, and SOHO at Lagrange point 1 in the interplanetary medium, and satellites like NOAA POES, DMSP, and MetOp in low Earth orbit near our planet. However, as we have seen, the complexity of these phenomena and their interconnected relationships make it difficult to establish a direct link between measurements and the associated risks. While each of these measurements plays a role in risk prediction and infrastructure protection, the development of more comprehensive and representative indices became necessary. These indices serve as indicators of solar and geomagnetic activities. We have already mentioned some of them, such as the Dst, and the Auroral Electrojet, but let’s make a brief overview and present them once again.

1.3.1 Measuring Space Weather through Indices

There are several measurements that provide scientists with information about the state of space weather, but indices offer a quick and effective way to get a comprehensive understanding. Of course, indices do not replace direct data but rather serve as a summary. In the following sections, we provide a brief description and then summarize all the indices in a table. To accomplish this, we primarily relied on the platform provided by the International Service of Geomagnetic Indices⁷ (ISGI). It’s important to note that we are presenting only the main geomagnetic and solar indices in this discussion.

7. <https://isgi.unistra.fr/>

- ***AE, AU, AL***: The *AE* (for Auroral Electrojet) index monitors the magnetic signature of the eastward and westward auroral electrojets in the Northern hemisphere by measuring the horizontal component disturbances in the geomagnetic field in nanotesla, through 12 observatories in the Northern auroral zone (namely, BRW, CMO, YKC, FCC, SNK, NAQ, LRV, ABK, DIK, CCS, TIK, PBK, Figure 1.33). According to ISGI, the magnetograms of the horizontal components from the AE stations are superimposed: the upper envelope defines the *AU* (for "upper") index, and the lower envelope defines the *AL* (for "lower") index and we have $AE = (AU - AL)$.
- ***PCN, PCS***: The Polar Cap North (*PCN*) and South (*PCS*) indices monitor the geomagnetic activity over the polar caps caused by changes in the interplanetary magnetic field (IMF) and solar wind, driven by the geoeffective interplanetary electric field irrespective of time, season and solar cycle. They do so by measuring the horizontal component disturbances in the geomagnetic field with two different stations: one in the north (THL, see Figure 1.33) and one in the south (VOS).
- ***Kp***: We already mentioned *Kp* several times but did not explain what it is. The purpose of *Kp* is to characterize the intensity of geomagnetic activity on a planetary (hence, the "p") scale by computing the arithmetic mean of the 3-hour standardized K-indices for 13 *Kp*-observatories (11 northern and 2 southern stations between 44° and 60° northern or southern geomagnetic latitude). According to ISGI, a K-index is an integer between 0 and 9 corresponding to a class that contains the largest range of geomagnetic disturbances (a_x and a_y) in the two horizontal components (X and Y) during a 3-hour UT interval. Because of the historical context at the time of its creation, the *Kp* network is heavily weighted towards Europe and Northern America. The observatories as located on the Figure 1.33 are SIT, MEA, OTT, FRD, LER, ESK, HAD, BFE, UPS, NGK, WNG. *Kp* is usually used to characterize the severity of geomagnetic storms. The NOAA Space Weather Scales describe them as follows: minor storm (*Kp* = 5), moderate (*Kp* = 6), strong (*Kp* = 7), severe (*Kp* = 8 including 9-) and extreme (*Kp* = 9). We can also quickly mention the half-hourly *Hp30* and hourly *Hp60* (or the *Hpo* indices) developed at GFZ during the Horizon 2020-project SWAMI (Yamazaki et al., 2022). *Hpo* can be used as a higher time-resolution *Kp*, and is not limited to 9 (i.e., *open-ended*) hence better characterizing the severe geomagnetic storms.
- ***ap***: The *ap* index is obtained from *Kp* through a conversion table (that can be found on ISGI's website) and with a linear scale in unit 2nT.
- ***Km, am***: The *am* index is supposed characterize the global geomagnetic activity using a large set of 24 stations representing all longitudes and possible hemispheric discrepancies (more than *Kp* and *ap*). The stations as seen in Figure 1.33 are VIC, NEW, TUC, OTT, FRD, HAD, CLF, NGK, ARS, NVS, IRT, MMB, PET, MGD around 50°N and PST, AIA, KEP, HER, CZT, PAF, AMS, GNG, CNB, EYR around 50°S. An average of *an* for the north and *as* for the south is made ($am = (an+as)/2$). As for the link between *ap* and *Kp*, *Kpm* is a quasi-logarithmic scale as a third of *am* units (28 values) through a conversion table that can be found on ISGI's website.
- ***Dst, ASY/SYM***: *Dst* stands for Disturbance Storm-Time Index. It consists in 1-hour values of the horizontal component disturbances in the geomagnetic field through 4 low latitudes stations (HON, SJG, HER, KAK in Figure 1.33). The World Data Center for Geomagnetism in Kyoto (WDC Kyoto) is responsible for monitoring and reporting the *Dst* index in near real-time. The *Dst* is also largely used to evaluate the presence and severity of geomagnetic storms. Usually, a geomagnetic storm is defined with a *Dst* index of less than -50 nT. According to Park et al. (2021), an event start time is defined as the maximum *Dst* time at the main phase and the recovery phase (hence the storm) ends when *Dst* exceeds -30 nT (1.2.4). The severity is linked to the minimum reached by the *Dst* during the storm. A weak storm ranges from -30nT to -50nT, moderate from -50nT to -100nT, strong from -100nT to

-200nT, severe from -200nT to -350nT, and a minimum Dst index below -350nT is classified as a great storm (Zhao et al., 2022). Alternatively, some researchers consider a minimum *Dst* index below -200nT as a great storm. Finally, we can quickly mention the *ASY/SYM* indices (such as *SYM-H* already mentioned 1.2.4). They measure 1-minute geomagnetic disturbances at mid-latitudes in terms of longitudinally asymmetric (*ASY*) and symmetric (*SYM*) disturbances for components parallel (H) and perpendicular (D) to the dipole axis, hence creating four indices: *ASY-H*, *ASY-D*, *SYM-H*, *SYM-D*. According to ISGI, the *SYM-H* index and the *DST* index are more or less the same and differ only in terms of their time resolution.

- $F_{10.7}$: As defined by NOAA, $F_{10.7}$ is the solar radio flux at 10.7 cm (2800 MHz) and is an excellent indicator of solar activity. It has been measured continuously for over seven decades and has proven very valuable in specifying and forecasting space weather (Space Weather Prediction Center, 2023). The measurements for $F_{10.7}$ are obtained from ground-based observatories and satellite data that monitor the solar radio emissions and they are not specifically associated with individual stations or sources. However, the National Research Council of Canada is measuring it since 1947 and has become the international responsible for it. It goes from approximately 70 when the Sun's activity is low to around 300 for a high activity and the overall variations are very close to the solar spot number mentioned Section 1.2.1.2
- F_{30} : F_{30} , the solar radio flux at 30 cm, has been recorded since the 1950s (more details can be found in Dudok de Wit et al. (2014)). It has proven to be a valuable alternative to $F_{10.7}$ and has shown superior performance in density modeling, as observed in Dudok de Wit and Bruinsma (2017), particularly when integrated into the DTM model (which we describe in Section 1.3.3.2) Bruinsma and Boniface (2021).

Extensive research is being conducted in the field of space weather indices, aiming to enhance their accuracy, refine their parameters, develop new indices, and establish their correlations with specific events and their impacts. The ultimate goal is to improve our monitoring of the space environment and deepen our understanding of it. One significant advantage of such indices lies in their retrospective applicability. Some of them have been measured for a considerable period (e.g., ISSN, $F_{10.7}$), offering extensive time series data encompassing multiple solar cycles. Others help determine if we are experiencing a geomagnetic storm or not. When combined with actual measurements, all these indices serve as ideal candidates for modeling and predicting the dynamics of our space environment.

1.3.2 The Role of Modeling and Forecasting

The space weather community is dedicated to understanding and quantifying the threats associated with space weather, mitigating them, and ideally preventing them altogether. A recent scientific program called PRESTO, led by the Scientific Committee on Solar-Terrestrial Physics (SCOSTPE) and detailed by Daglis et al. (2021), aims to predict the variability of the Solar-Terrestrial coupling. This program has shed light on remaining questions regarding the understanding of the connection between the Sun and Earth.

One of the key inquiries focuses on how different solar wind conditions, such as IMF components, speed, density, and turbulence levels, along with various large-scale drivers, control the efficiency of coupling and the transfer of energy and mass from the solar wind to the magnetosphere. Additionally, researchers seek to understand how solar wind conditions influence the occurrence frequency and location of different magnetospheric plasma waves. These questions highlight the significance of the solar wind in predicting the space environment around Earth. To gain a better understanding, it is crucial to conduct studies that integrate space- and ground-based data analysis with models, particularly in relation to the solar wind and interplanetary magnetic

Index	Description	Time resolution	Unit	Source / Stations
AE / AU / AL	Auroral index; magnetic horizontal component disturbances	1-minute	nT	12 observatories in the Northern auroral zone (above 60°N)
PCN / PCS	Polar Cap index; magnetic horizontal component disturbances, PCN in the North and PCS in the South	1-minute	mV/m	2 polar cap stations (~80°N and ~80°S)
am	Global geomagnetic activity north and south combined (average between north and south)	3-hour	nT	14 north stations and 10 south stations
K _{pm}	Same as am but different scale obtained from am through a conversion table ⁸	3-hour	quasi-logarithmic scale as a third of K units (28 values): 0o, 0+, 1-, 1o, 1+, 2-, ... , 8+, 9-, 9o	
K _p	K _p is the arithmetic mean of the 3-hour standardized K-indices for the 13 K _p -observatories.	3-hour	quasi-logarithmic scale as a third of K units (28 values): 0o, 0+, 1-, 1o, 1+, 2-, ... , 8+, 9-, 9o	11 northern and 2 southern stations between 44° and 60° northern or southern geomagnetic latitude
ap	ap is obtained from K _p through a conversion table ⁹	3-hour	linear scale in unit ~2nT	
Dst	Equatorial index; magnetic horizontal component disturbances	1-hour	nT	4 low latitude, near-equator stations
ASY/SYM	Geomagnetic longitudinally asymmetric and symmetric horizontal component disturbances	1-hour	nT	6 stations evenly distributed in longitude (11 observatories whose data are interchangeable depending on their availability)
F _{10.7}	Called solar radio flux at a wavelength of 10.7 centimeters, even if it is a flux density	1-day	Solar flux units (sfu) = 10 ²² W.m ⁻² .Hz ⁻¹	
F ₃₀	Solar radio flux at a wavelength of 30 centimeters	1-day	Solar flux units (sfu) = 10 ²² W.m ⁻² .Hz ⁻¹	

Table 1.6 – Summary of main geomagnetic and solar indices

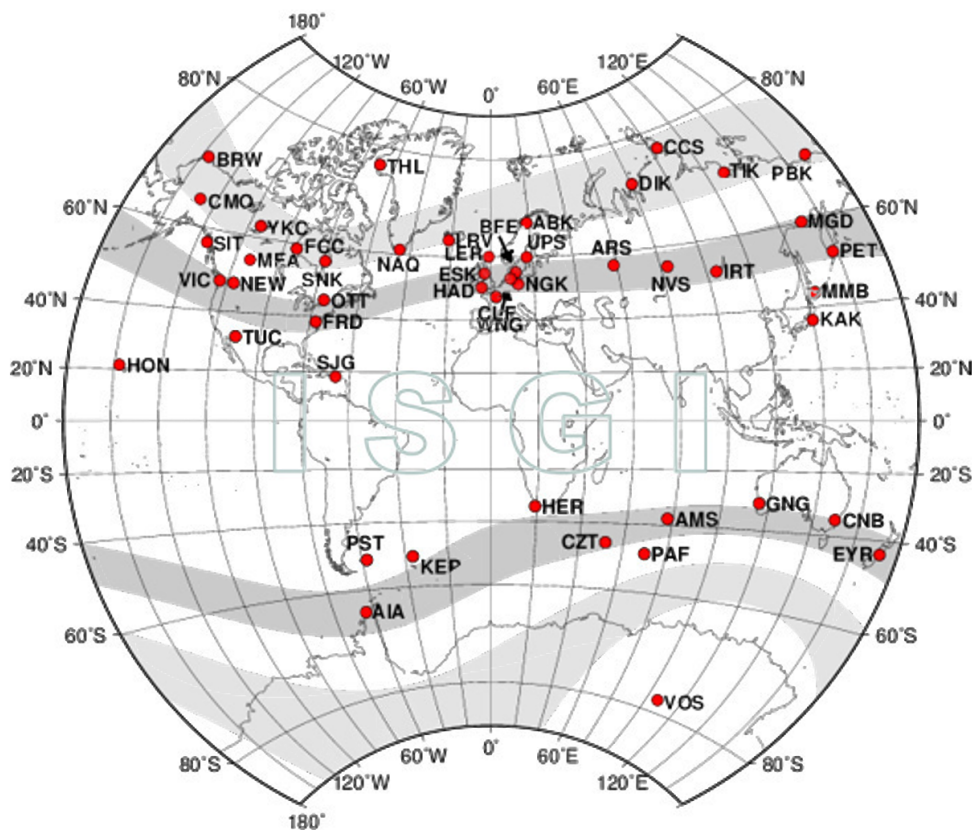


Figure 1.33 – Most stations and observatories used for geomagnetic indices. ABK = Abisko, Sweden; AIA = Argentine Islands, Antarctica (Ukraine); AMS = Martin De Vivies-Amsterdam Island, French Southern and Antarctic Lands (France); ARS = Arti, Russia; BFE = Brorfelde, Denmark; BRW = Barrow, United States of America; CCS = Cape Chelyuskin, Russia; CLF = Chambon La Foret, France; CNB = Canberra, Australia; CMO = College, United States of America; CZT = Port Alfred, French Southern and Antarctic Lands (France); DIK = Dixon, Russia; ESK = Eskdalemuir, United Kingdom; EYR = Eyrewell, New Zealand; FCC = Fort Churchill, Canada; FRD = Fredericksburg, United States of America; GNG = Gingin, Australia; HAD = Hartland, United Kingdom; HER = Hermanus, South Africa; HON = Honolulu, United States of America; IRT = Irkutsk, Russia; KAK = Kakioka, Japan; KEP = King Edward Point, United Kingdom; LER = Lerwick, United Kingdom; LRV = Leirvogur, Iceland; MEA = Meanook, Canada; MGD = Magadan, Russia; MMB = Memambetsu, Japan; NAQ = Narsarsuaq, Greenland (Denmark); NEW = Newport, United States of America; NGK = Niemegk, Germany; NVS = Novosibirsk, Russia; OTT = Ottawa, Canada; PAF = Port-Aux-Francais, French Southern and Antarctic Lands (France); PBK = Pebek, Russia; PET = Paratunka/Petropavlovsk, Russia; PST = Port Stanley, Falkland Islands (United Kingdom); SIT = Sitka, United States of America; SJG = San Juan, United States of America; SNK = Sanikiluaq, Canada; THL = Qaanaaq (Thule), Greenland (Denmark); TIK = Tixie Bay, Russia; TUC = Tucson, United States of America; UPS = Uppsala, Sweden; VIC = Victoria, Canada; VOS = Vostok, Antarctica (Russia); WNG = Wingst, Germany; YKC = Yellowknife, Canada. From ISGI [isgi.unistra.fr/](http://isgi.igsi.unistra.fr/).

field.

Machine learning algorithms have emerged as a promising solution in the field of space weather, allowing for the nowcasting and forecasting of space weather phenomena. Recent studies have demonstrated the potential of machine learning, including deep learning, in this domain. Notably, papers such as those by [Reiss et al. \(2021\)](#), [Zewdie et al. \(2021\)](#), [Stumpo et al. \(2021\)](#), and [Reep and Barnes \(2021\)](#) have showcased the utilization of machine learning in space weather research. These developments have provided valuable insights into the application of machine learning algorithms for improved space weather predictions (see Chapter 2).

In 2018, the National Science Foundation (NSF) initiated a Research Coordination Network (RCN) known as "Towards Integration of Heliophysics Data, Modelling, and Analysis Tools" (@HDMIEC). This network aimed to advance our understanding of physical mechanisms occurring on the Sun, as well as the modeling, accessibility, and analysis of heliophysics data. Within this framework, workshops and discussions were conducted to explore the intersection of machine learning and space weather. Opinions were gathered from the community, and several noteworthy outcomes from the Q&A sessions, as highlighted by [Nita et al. \(2020\)](#), emerged. Attendees at these sessions expressed concerns regarding the heliophysics community's limited understanding of machine learning capabilities and limitations. It was generally agreed upon that there is currently no substantial collaboration between machine learning and heliophysics. However, it was recognized that machine learning methods have shown greater success in handling the extensive data environment of heliophysics compared to traditional physics-based methods. Despite this, there was no consensus on specific areas where machine learning outperforms physics-based approaches. A significant majority of attendees strongly advocated for the combination of physics-based and machine learning models to enhance space weather predictions. Moreover, attendees generally did not perceive machine learning as a "bubble" that is at risk of bursting, indicating ongoing confidence in its potential for advancements in space weather research.

1.3.3 Impacts on space and ground systems

Since the launch of Sputnik in 1957, scientists have observed that the natural environment can pose risks to satellites. This realization gave birth to the idea of analyzing, quantifying, and mitigating these risks. It is important to note that this thesis was developed in partnership with the company SpaceAble, which aligns perfectly with this objective. Consequently, extensive research has been conducted to gain a better understanding of the various risks associated with space weather. In the following sections, we will provide an overview of the main risks identified by the scientific community and their consequences. We will give more details for spacecraft charging and drag as they can be directly related to precipitated particles. Most of what is explained here comes from researches made with SpaceAble's scientists, especially Dr. Elisa Robert, and these results can also be found in her PhD thesis ([Robert, 2023](#)).

1.3.3.1 Spacecraft Charging

Spacecraft charging refers to the accumulation of electric charges on the surface (surface charging) or inside (internal charging) of a spacecraft due to the spacecraft's direct contact with a charged medium like the ionosphere and poses significant risks for satellites. In Low Earth Orbit (LEO), the risk resides mainly in surface charging, which is influenced by particle precipitation. Overall, spacecraft charging can impact scientific measurements, electronic instruments, telemetry, navigation systems, and even lead to spacecraft termination. Monitoring it and its intensity is crucial for measurement accuracy and safety. For more detailed information, we recommend

referring to [Mikaelian \(2009\)](#) and [Lai \(2011\)](#).

Surface & Internal Charging

How exactly do we dissociate internal and surface charging? *Surface charging* results in an accumulation of charges (keV electrons and ions) around and on the surface of the spacecraft which travels through plasma. *Internal charging* occurs when high-energy electrons and ions (MeV and higher) penetrate satellite shielding materials and deposit charge on internal spacecraft components. In the literature, we often see “deep dielectric” or “bulk” charging. This refers to charge densities that accumulate within insulating (or dielectric) materials when exposed to penetration radiation. The more generic term “internal charging” includes charge densities that accumulate on the surfaces of conducting materials within the shielding afforded by the outer structure of a spacecraft.

Both processes result in electric fields within spacecraft structures and materials and represent a threat to arcing which can damage spacecraft components. The natural fluxes of the ambient electrons and ions responsible for surface charging are orders of magnitude higher than those for deep dielectric charging. Indeed, surface charging responds almost instantaneously to the ambient flux temperature and can be easily measured.

What causes Surface Charging?

Spacecraft charging is a complex phenomenon that can occur in the orbits of most satellites (GEO, LEO, polar, interplanetary and so on). It is influenced by various environmental factors, including day-night variations, semi-annual variations, eclipses, as well as solar and geomagnetic activity. Additionally, it is directly impacted by the characteristics and orbital parameters of the satellite, such as its orbit, attitude, design, geometry, as well as the properties of the materials used in its construction. Consequently, modeling and predicting the interaction between the spacecraft and its environment is a highly intricate task.

In LEO orbit, spacecraft charging can be attributed to two primary sources: the cold plasma environment, suprathermal electrons and free protons, and solar radiation.

- The cold plasma in LEO has a density of 10^2 to 10^6 cm^{-3} and a temperature of approximately 0.1 eV, contributing to spacecraft charging. In polar orbits, the ionosphere is additionally influenced by charged particles from the solar wind, including electrons. The precipitation of higher-energy electrons (1-100 keV) in this region intensifies spacecraft charging effects.
- According to [Mikaelian \(2009\)](#), the sources beside the cold plasma environment are suprathermal electrons and free protons of both low energy (eV-10 keV) and high energy (> 10 keV). These particles originate from precipitating electrons in polar regions but are also protons and electrons generated by the CRAND effect (discussed in Section 1.2.3.8). A given energy will mean a different particle penetration depth (see Figure 1.34) hence we often differentiate surface and internal charging regarding the related energy of particles.
- The flux of UV and EUV solar radiation, specifically in the range of 100nm to 400nm, exhibits seasonal variations and is influenced by the 11-year solar cycle. During periods of high solar activity, the radiation flux is more intense compared to periods of low activity. When these EUV and UV photons interact with a spacecraft, they induce the photoelectric effect, leading to the ejection of electrons from the surface of metallized materials, referred to as photoelectrons. The presence of photoelectrons creates a current that flows out of the spacecraft's surface, potentially mitigating the negative effects of surface charging.

eV			keV			MeV			GeV			...
1	10	100	1	10	100	1	10	100	1	10	100	
Surface charging						Internal charging			Radiation effects			

Figure 1.34 – Table from the presentation of Dr. Linda Neergaard Parker (Parker, 2017) representing surface and internal charging based on particle energy.

Figure 1.35 gives a good overview on the different phenomena at stake (upcoming photon, low and high energies protons, low and high energies electrons). We can note that most currents on the satellite usually come from the deposition of negative charges. The positive currents generally result from emission of low energy secondary or backscattered electrons and photoelectrons (ejected by the photoionisation) but the positive potentials that can be attained are relatively modest.

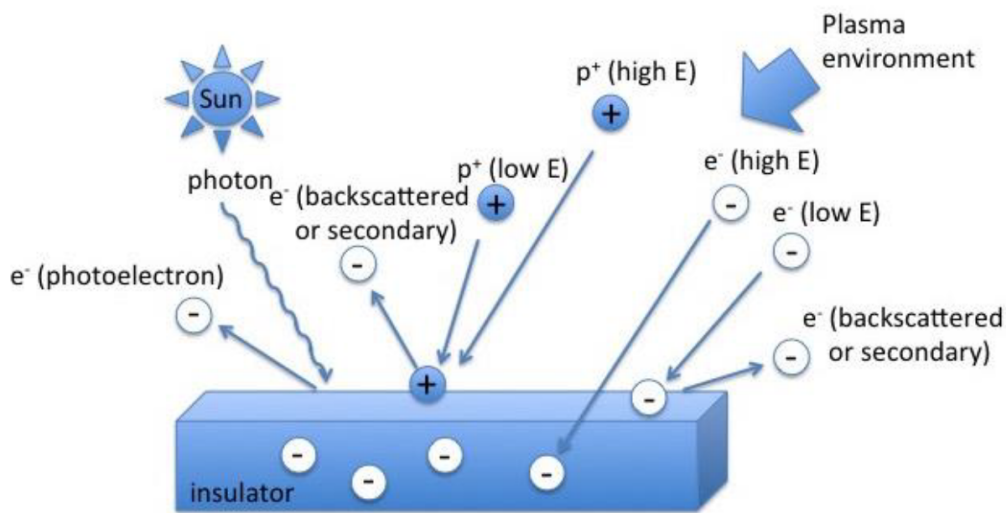


Figure 1.35 – Overview of the different processes during spacecraft charging. Figure from the presentation of Dr. Linda Neergaard Parker (Parker, 2017).

Consequences of Surface Charging

One of the consequences of spacecraft charging is *Electrostatic Discharge* (ESD), one of the prominent sources of anomalies on spacecraft. It's the sudden flow of electricity between two electrically charged objects caused by contact, an electrical short, or dielectric breakdown. Its effects are a combination of the electron environment and its interaction with specific spacecraft surfaces and components, resulting in current flow through wires to sensitive instruments and electromagnetic wave interference to telemetry. This discharge is usually presented into three categories:

- *Flashover*: Discharge from one surface to another.
- *Discharge to space*: Discharge from spacecraft to the surrounding plasma.
- *Punch-through* : Discharge from the interior structure of a spacecraft through its surface.

Main impacts of the electrostatic discharges are compromised function or destruction of sensitive electronics. This includes damage to solar arrays, power loss and failures. System states may experience un-commanded changes. Timing circuits can lose synchronization, leading to disruptions in operations. Additionally, there can be spurious mode switching, power-on resets, or erroneous sensor signals that can further result in telemetry noise or loss of data.

Beside electrostatic discharges, spacecraft charging can lead to other to more simple electromagnetic interferences or power drains from parasite currents on or in the satellites.

Studying and Modeling Spacecraft charging

In order to study spacecraft charging, numerous missions have been launched by the scientific community. Among them is SCATHA (Spacecraft Charging At High Altitude), a mission conducted by NASA and the US Air Force, which confirmed the presence of spacecraft charging in geostationary orbit (an orbit approximately 36,000 km above the Earth's surface where a satellite remains stationary). Another mission, CRRES (Combined Release and Radiation Effects on Satellite), also carried out by NASA and the US Air Force, successfully detected over 4,000 electrostatic discharges using its internal discharge monitor (IDM). Furthermore, the Defense Meteorological Satellite Program (DMSP) and POES missions, both situated in low Earth orbit, played crucial roles. DMSP, equipped with instruments such as SSJ/4 and SSJ/5, and POES, with the Space Environment Monitor (SEM), were able to establish discharge initiation thresholds, which will be discussed in detail later on in this thesis.

In addition to in-situ measurements that aid in understanding and localizing the phenomenon, scientists also endeavor to model spacecraft charging. Below is a non-exhaustive list of existing models that simulate spacecraft charging:

- NASCAP-2K - Maxwell Technologies: NASCAP, developed by Maxwell Technologies, is an American model with various variants designed for specific orbits: NASCAP-LEO (Mandell and Davis, 1990) for LEO, NASCAP-GEO (Katz et al., 1977) for GEO, and POLAR (Potentials Of Large objects in the Auroral Region) (Lilley Jr et al., 1989) for the auroral region. NASCAP-2K, the most recent model, combines NASCAP-GEO, NASCAP-LEO, and POLAR (Mandell et al., 2004).
- Space Plasma Interaction System (SPIS) - ESA: SPIS is a model developed by the European community SPINE (Roussel et al., 2008).
- SPacecraft Charging Software (SPARCS) - Thales Alenia Space (Clerc et al., 2003).
- Space Environment Information System (SPENVIS) - ESA: SPENVIS, the most widely used model in the European community, extends beyond spacecraft charging modeling (Heynderickx et al., 2000).
- Space Hazards Induced Near Earth by Large Dynamic Storms (SHIELDS) - LANL: SHIELDS, a recent model introduced by Los Alamos National Lab (LANL) (Jordanova, 2017), focuses on modeling and predicting space hazards, particularly surface charging, and facilitates forensic analyses of space-system failures.
- Multi-Use Spacecraft Charging Analysis Tool (MUSCAT) - JAXA: Developed in collaboration between JAXA and the Kyushu Institute of Technology (Muranaka et al., 2008).

Mitigating Spacecraft Charging Effects

Surface charging can be mitigated using various techniques. Electron emission is one such method where electron guns, sharp spikes, or hot filaments are utilized to reduce the negative voltage level. Plasma emission is considered one of the most effective mitigation technique as it allows electrons to escape while positive ions return to the highly charged areas. Another approach is through the use of plasma contactors, which create a large, localized plasma cloud to establish good electrical contact with the surrounding plasma. This effectively grounds the satellite structure to the ambient plasma. Polar molecule emission and surface materials with high secondary emission coefficients can also help mitigate surface charging. Autonomous devices such as sharp cones can facilitate field emission, removing excess electrons from spacecraft surfaces connected to the devices.

Internal charging, on the other hand, can be mitigated using different strategies. Materials of finite conductivity are required to prevent the buildup of internal charge while still serving insulation purposes. The choice of conductivity depends on the specific system requirements. Partially conductive paint, like indium oxide, can also be used. However, covering the entire spacecraft with a uniform paint to prevent differential charging may obstruct instruments and solar arrays. Thin materials are also beneficial in mitigating internal charging as high-energy electrons or ions can pass through them without leaving significant deposits. Finally, shielding is another approach, where different energies of ions and electrons penetrate materials to varying depths. By employing shielding, electrons or ions below a certain desired energy level can be prevented from reaching sensitive instruments. However, there is a tradeoff as excessive shielding can impede instrument functionality and add weight at launch.

Space weather forecasting plays a crucial role in mitigating both surface charging and internal charging. Major magnetic storms often lead to internal charging in dielectrics one to several days later. As we now know, the occurrence of a magnetic storm serves as a warning of the potential presence of high-energy (MeV) electrons in the radiation belts. Forecasting methods have been developed using linear prediction filters and neural networks to anticipate environmental conditions that may lead to electrostatic discharge (ESD). However, these prediction methods are limited because they often assume stationary time series. Coronal mass ejections (CMEs) can also cause a rapid increase in energetic (MeV) electrons upon arrival at the Earth's magnetosphere, resulting in internal charging. Finally, monitoring solar activity also helps us understand substorms and injection of particles in polar areas, as well as precipitated particles that are often responsible for surface charging.

1.3.3.2 Drag

The phenomenon of atmospheric drag or friction occurs primarily in the high-altitude atmosphere. Despite the decrease in density with altitude, there is still sufficient matter to induce aerodynamic friction on the satellite, resulting in trajectory alteration. The layer responsible for this braking effect is the thermosphere, composed of neutral components consisting of non-ionized atoms and molecules. With an altitude range of 95 km to 500-1000 km, depending on the thermopause limit, the thermosphere is predominantly composed of O, O², N², and He, with oxygen and helium being the most responsible for the drag effect (Thayer et al., 2012).

At this altitude, the thermosphere is intertwined with the ionosphere, the ionized portion of the thermosphere referred to as plasma. While the ionosphere also contributes to drag through ionospheric aerodynamic effects (Capon et al., 2019), its detailed discussion is beyond the scope of this section.

The calculation of drag can be easily performed using Equation 1.47 (Zheng et al., 2019), which incorporates the thermospheric density (ρ), satellite velocity (u), the area subjected to friction (A), and the ballistic coefficient (C_D). Most of these parameters are contingent on the satellite's design or orbit, with atmospheric density being the sole factor dependent on the medium traversed. As such, the density plays a central role in the drag calculation and is the primary focus of investigations when we try to model the drag forces.

$$F_D = \frac{1}{2} \rho u^2 C_D A \quad (1.47)$$

Thermosphere Density & Solar Activity

There are two primary sources responsible for the deceleration of satellites in orbit and their impact on the thermospheric density:

1. Solar activity, specifically the EUV solar flux at 170 nm.
2. Geomagnetic activity, involving particle precipitation at the poles and the Joule effect caused by ionospheric electric currents.

These two processes exert varying influences on the thermosphere depending on solar activity levels. During periods of low activity, the energy deposition from precipitating particles amounts to one-fourth of the energy deposition from the EUV solar flux. Conversely, during periods of high activity, the contribution from precipitating particles can reach up to twice that of the EUV solar flux. In such scenarios, the thermospheric density becomes approximately ten times greater compared to periods of low activity (Doornbos, 2012).

The EUV solar flux and precipitating particles have a specific impact on the thermosphere, causing heating, excitation, and dissociation of the atoms and molecules present. Dissociation leads to an increase in the atmospheric atom density, resulting in higher thermospheric density. Hence solar activity significantly influences the concentration of species in the thermosphere, leading to changes in thermospheric density.

The temperature of the thermosphere is primarily driven by solar radiation. However, heating occurs indirectly, as it is not directly caused by the interaction between the atmosphere and solar radiation. Instead, it arises from the frictional forces between the excited particles and their non-excited counterparts. Additionally, chemical reactions contribute to atmospheric heating. In the thermosphere, the absence of convective and conductive heat transfer mechanisms maintains a constant temperature known as the thermospheric temperature. This temperature ranges from 750 K during periods of low solar activity to as high as 1,500 K during periods of high activity. The term "thermosphere" derives from the pronounced temperature gradient observed in this region (Vallado et al., 2006).

Studying & Modeling Drag

The calculation of atmospheric drag necessitates an accurate density estimation. First, a model should acquire satellite position measurements to observe orbital perturbations, either through in-situ or remote sensing techniques. Then it estimates or, in some cases, predicts the thermospheric density based on these measurements, employing various methodologies. Subsequently, the drag calculation uses this density measurements, along with satellite-specific parameters such as the friction coefficient and reference area. In practice, operators and space agencies often employ modeling tools or simulation software, such as the STELA software developed by CNES.

During the 1990s, the increasing interest in thermospheric density measurements led to the development of diverse methods: ground-based satellite tracking techniques, including radar observations, GPS measurements, and direct utilization of Two-Line Elements (TLEs), or onboard in-situ instruments, such as accelerometers.

To get the positions of satellites and recover the thermospheric density, there are four precise tracking systems: GPS, Satellite Laser Ranging (SLR), one-way Doppler radio tracking (DORIS), and ground-based radars used for space surveillance. However, accessing accurate raw data, such as radar measurements, is difficult because most available, public data have already been processed and averaged. This poses challenges for precise density measurements, and there are political obstacles to sharing information. Resolving these issues would require partnerships and agreements.

Another way of getting the thermosphere density is through in-situ measurements. However, to date, no instrument for direct, in-situ measurement of thermospheric density has been designed.

Instead, accelerometers are used to indirectly retrieve the density. Satellites such as CHAMP (Reigber et al., 1999), GRACE (Tapley et al., 2004), GRACE-FO (Kornfeld et al., 2019), GOCE (Drinkwater et al., 2003), and SWARM (Friis-Christensen et al., 2008), can indirectly provide us with access to the frictional force. However, the use of an accelerometer as an instrument poses limitations due to its size and high sensitivity. Consequently, alternative approaches have been explored in recent years. A promising, but not yet achieved, candidate is the mass spectrometer, which enables direct measurement of thermospheric density. This possibility is being considered with the development of the Cosmorbitrap instrument as part of the ICARUS project (Selliez-Vandernotte, 2018).

Finally, some models exist to approximate or predict the thermosphere density (and hence, the atmospheric drag) and they are the main source of information for operators. We can mention the following models: the Drag Temperature Models (DTM), the Jacchia-Bowman (JB) model, the Mass Spectrometer Incoherent Scatter (MSIS) model and the US Space Force High Accuracy Satellite Drag Model (HASDM).

The HASDM (Storz et al., 2005) developed by the US Air Force is currently the only model capable of real-time estimation and prediction of thermospheric density up to 72 hours in advance. Implemented within the Space Battlelab project, HASDM assimilates trajectories of over 75 inactive satellites and space debris in LEO, providing atmospheric corrections every 3 hours using the Dynamic Calibration Atmosphere (DCA) algorithm. By extrapolating density correction coefficients from the past 27 days and estimating ballistic coefficients through the Segmented Solution for Ballistic coefficient (SSB) technique, HASDM achieves precise density predictions. However, it should be noted that HASDM is not available in open-source.

The JB2008 (Bowman et al., 2008), DTM-2020 (Bruinsma and Boniface, 2021), and NRLMSISE-00 (Picone et al., 2002) models are the latest versions within each family and differ in terms of historical databases, solar and geomagnetic indices used as inputs, and parametric equations for deriving thermospheric density from exospheric temperature. These models primarily utilize accelerometer-based drag coefficient and density data, with JB2008 and DTM2020 assimilating density values from satellites for higher precision. All models employ the $F_{10.7}$ solar index, and JB2008 stands out by using additional solar indices to capture a broader range of thermospheric heating. Geomagnetic indices (K_p or A_p - daily average of eight ap values, see 1.3.1) are used by all models, with DTM-2020 introducing the A_m index to cover disturbances in the Northern and Southern hemispheres. Finally, the parametric equations for density calculation vary among the models, with JB2008 incorporating latitude-dependent parameters for semi-annual and seasonal variations.

Consequences of Drag

The main consequence associated with drag is the alteration of satellite trajectories. An unforeseen deviation in trajectory can lead to satellite loss for operators, thereby substantially impeding preemptive avoidance maneuvers and amplifying the collision risk. The collision risk is also highly increased because of unpredictable movements of already existing debris in LEO.

Moreover, the increase in thermospheric density can impose a greater fuel consumption for satellites trying to reach their designated orbits, consequently diminishing the mission's lifespan. In February 2022, the Starlink satellites operated by SpaceX encountered this issue related to atmospheric drag, resulting in the disintegration of several satellites before they could reach their intended final orbits. These satellites were launched right after a geomagnetic storm which led to an elevated density in the thermosphere and consequently an increase in atmospheric drag. As

a consequence, out of the 49 satellites launched, 40 were unable to escape their transfer orbits, situated at an altitude of approximately 210 km, and ultimately disintegrated within the Earth's atmosphere at lower altitudes (Fang et al., 2022).

1.3.3.3 Geomagnetically Induced Currents

Geomagnetically Induced Currents (GICs) are electric currents generated on the ground through the process of electromagnetic induction caused by rapid changes in the geomagnetic field, particularly during Coronal Mass Ejections (CMEs). As discussed in Section 1.2.6.3, the auroral electrojets in the ionosphere follow circular paths around the geomagnetic poles. During periods of low magnetic activity, the electrojet remains within the auroral oval. However, during disturbed periods, the electrojet intensifies, expanding to both higher and lower latitudes, which can cause fluctuations in the geomagnetic field. Therefore, to assess the generation of GICs, it is crucial to measure the time derivative of the geomagnetic field, rather than relying solely on indices like Kp or Dst , as they are not directly correlated with GICs production.

Severe geomagnetic storms can result in Earth-surface potential values ranging from 1 to 7 V/km, particularly in high-latitude regions with low earth conductivity, such as areas with igneous rock formations or coastal regions. Regions in North America, in particular, are at higher risk due to the positioning of the north magnetic pole in relation to the north geographic pole, which increases their vulnerability to elevated Earth-surface potential values. Consequently, electric power systems in these areas are more prone to disturbances in the geomagnetic field. The Earth-surface potential acts as an ideal voltage source between grounded neutrals of specific transformers within a power system, causing geomagnetically induced currents to flow between the neutrals (see Figure 1.36). Power networks with long transmission lines spanning hundreds of kilometers are more susceptible to magnetic field disturbances, while smaller grids are less exposed to such phenomena. Table 1.7 summarizes the different risks from GICs.

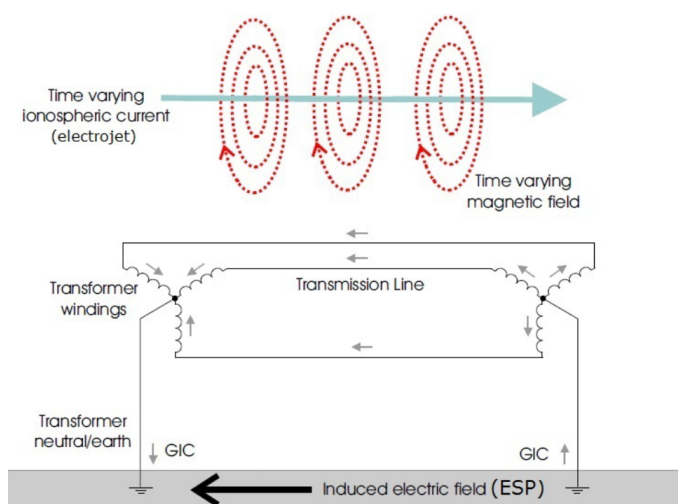


Figure 1.36 – Formation of the induced electric field and the GICs as a consequence. Figure from McKay (2004).

1.3.3.4 Radiation Effects: Single Event Effects

A *single event effect* (SEE) results from, as the term suggests, a single, highly energetic particle. It is an electrical disturbance that disrupts the normal operation of a circuit and can be either

Transformer saturation	GIC acts as a DC current source which determines a bias in transformer operation, shifting it to the saturated region of the magnetization curve.
Reactive power losses	Saturated operation yields augmented current demand which generates linearly increasing (i.e. with the current) re- active power consumption.
Harmonics	Operation in saturated region is highly non-linear and dis- torted, harmonics generated in current/voltage waveforms.
Transformer overheating	The excess flux induced by saturation flows externally to the core into the transformer tank and can generate localized heating spots of over 170°C.
Generator overheating	Generators are subject to harmonics and voltage unbalance caused by transformer saturation. The ensuing harmonic currents can potentially generate excessive heating and mechanical vibrations, although no serious generator damage due to GICs has been documented.
Protective relaying issues	Relays reacting to the peak value of the current are sensitive to the harmonics injected by saturated transformers and can erroneously trip.
Telecom- munication systems	Geomagnetic disturbances affect phone lines and Internet cables, however optical fibres (increasingly used nowadays in the high- bandwidth lines of SCADA networks) are immune to electro-magnetic interference. Satellite systems are directly prone to interference from geomagnetic disturbances.

Table 1.7 – Summary of effects of GICs and overall disturbances in the geomagnetic field on power systems, from [Beccutti \(2013\)](#)

“destructive” (cause permanent damage) or “non-destructive” ([Gomez Toro et al., 2014](#)). The possibility of single-event upsets was first postulated by [Wallmark and Marcus \(1962\)](#). There are three main sources of SEEs ([Poivey, 2019](#)):

1. Solar Energetic Particles (SEP, see Section 1.2.2.4): mostly protons ($\sim 96.4\%$), alpha particles ($\sim 3.5\%$), and heavy ions ($\sim 0.1\%$), in an energy range between 10 and 100 MeV ([Bothmer and Daglis, 2007](#))
2. Galactic Cosmic Rays (GCR): mostly protons ($\sim 90\text{-}95\%$), helium ($\sim 7\text{-}10\%$), heavy ions ($\sim 1\%$) and electrons ($\sim 1\%$) from interplanetary and interstellar space, all in the GeV to TeV energy range ([Bothmer and Daglis, 2007](#)). GCRs are the most energetic particles (energies up to 10^{21} eV) found in our Solar System and are fully ionized: that is, they consist of nuclei only.
3. Trapped energetic proton (TP) from the inner Van Allen belt produced by the CRAND effect, itself caused by the GCRs ([Bothmer and Daglis, 2007](#)). When GCRs population increase, so are the trapped protons populations. Usually, the inner radiation zone is avoided by operators because the total radiation dose is very large there. However, some low-altitude missions still experience effects from the energetic protons in the *South Atlantic Anomaly*. The asymmetries in the geomagnetic field cause the radiation belts to “dip” closer to the Earth in the south Atlantic regions, and satellites that pass through this region experience more single-event effects ([Koons and Fennell, 2006](#)).

The fundamental process is the following: first, an incident ion loses energy, ionizes and interacts with material along its track in the device, producing free charge carriers (electrons and holes). Then, electrons and holes move by diffusion and drift through the material (oxides and semiconductors) to sensitive node while they also recombine. Finally, the additional charge on the node alters the voltage that ultimately leads to SEEs. Voltage glitches may propagate through a circuit. An ion can also undergo a nuclear interaction with the atoms in the device. This generates a shower of energetic nuclei that then suffer ionization losses.

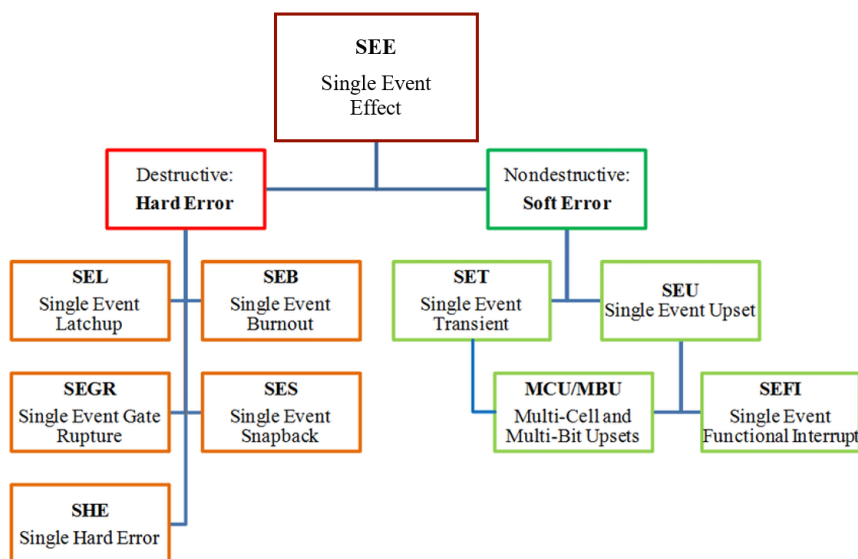


Figure 1.37 – Classifications of SEEs, from Gomez Toro et al. (2014).

The term SEEs is used to encompass the various effects resulting from the interaction of an energetic ion with a device. Historically, these different types of SEEs have been categorized based on the response of the microelectronic device, as shown in Figure 1.37. While we won't discuss all of them in detail, let's touch upon a few. Among the non-destructive phenomena, we have the Single Event Upset (SEU), which arises from transient radiation effects in electronics and leads to changes in the memory or bit states. According to Koons and Fennell (2006), SEUs account for the second most common cause of anomalies in satellites, representing 28.4% of cases. The Single Event Transient (SET) introduces voltage issues in circuits and can induce changes in logical states in both combinational and sequential logic circuits. On the other hand, among the destructive phenomena, the Single Event Latchup (SEL) causes an undesired short-circuit in an integrated circuit, typically triggered by heavy ions or protons. Lastly, the Single Event Burnout (SEB) poses a risk of device destruction due to high current in a power transistor.

1.3.3.5 Radiation Effects: Cumulative Effects

The second category in which radiation can affect electronics is called *cumulative effects* and encompasses two subcategories: *Total Ionizing Dose* (TID) and *Displacement (or Atomic) Damage Dose* (DDD).

Total Ionizing Dose

The total ionizing dose arises from the satellite's continuous exposure to and absorption of space radiation, especially electrons and protons. This dose involves the ionization of the materials that constitute the satellite. It is measured in Gray (Gy), where 1 Gy corresponds to the absorbed energy in exposed material of 1J/kg. The old unit, rad, is still widely used in the community (100 rad = 1 Gy). When the satellite is exposed to these particles, it generates a number of electron-hole pairs proportionally to the energy transfer. The term "electron-hole pair" refers to the phenomenon where each moving electron leaves behind a vacant position or hole.

The primary sources of particles responsible for TID are high-energy protons (MeV) from Solar Energetic Particle Events, typically observed during solar flares, and the South Atlantic Anomaly.

According to [Poivey \(2019\)](#), the Total Ionizing Dose primarily affects semiconductor oxides. Within the semiconductor oxide, electron-hole pairs are capable of mobility and can recombine, which does not result in any damage. The rate of recombination is influenced by the electric field applied to the oxide, as well as the type and energy of the incident particle. When a device is biased, electrons are expelled from the oxide, while holes remain, leading to the accumulation of trapped charges within the oxide or interface traps at the oxide-silicon interface. The degradation of components is highly dependent on the device technology, the manufacturing process, and the bias conditions.

Total Ionizing Dose leads to substantial degradation of components, especially semiconductors and insulators, and in certain instances, even their destruction. The flow of current, stemming from electron-hole pairs, can result in heightened power consumption within the device, decreased component gain, or changes in the threshold voltage of a metal-oxide-semiconductor gate ([Pisacane, 2008](#)).

Displacement Damage Dose

Displacement Damage Dose, also known as Total Non-Ionizing Dose (TNID), is a form of damage caused by the energy accumulated inside the satellite due to high-energy particles. These particles displace atoms within the crystalline structure of components, resulting in weakened structures with imperfections, reduced lifespans, changes in the electrical properties of the affected region (see [Figure 1.38](#)).

Displacement damage (DDD) can have various effects on electronic devices. In gate-oxide breakdown, accumulated defects can lead to a complete short circuit, melting the insulating layer and causing structural destruction. Solar panels are also particularly vulnerable to DDD that reduces their overall performances. In lasers, irradiation-induced defects act as recombination centers, increasing the threshold current and broadening the lasing wavelength. In image sensors, DDD can cause bright spots (clusters of defects) and signal streaks (defects acting as charge traps), reducing pointing accuracy and resolution in star trackers and Earth-observation detectors, respectively.

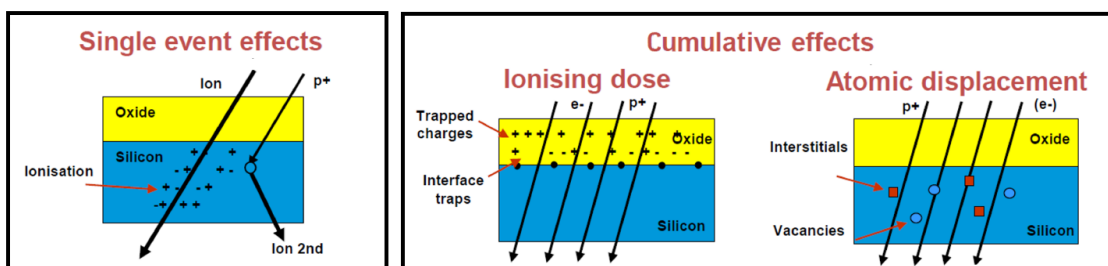


Figure 1.38 – All three categories of radiation effects, from [Poivey \(2019\)](#).

1.3.4 A Danger for Humanity

In order to conclude our discussion on the impacts of Space Weather, we aimed to provide a comprehensive overview of past events and emphasize the different sectors affected by space meteorology. To compile this section, we drew upon the insights and research presented in [Lilensten et al. \(2021\)](#), which not only explores historical events associated with Space Weather but also

serves as an excellent, accessible, resource for understanding the subject. We highly recommend this book to all readers.

1.3.4.1 Events From the Past

Before delving into the numerous human domains that can be impacted by Space Weather, we wanted to first present few impressive events that already happened in our close past.

- To this date, the "Carrington Event" that happened in September 1859, during solar cycle 10, is the most intense geomagnetic storm ever recorded, and resulted from a CME impact on Earth. At first, Richard Carrington observed very intense flashes emitted on the surface of the Sun ([Carrington, 1859](#)). These flashes were actually several solar eruptions that ultimately caused disruptions in telegraph communications ([Bothmer and Daglis, 2007](#)). Auroras could be seen in regions close to the equator.
- In March 1989, during solar cycle 22, a series of severe solar storms hit Earth. Among them, the worst one struck Earth on March 13 and caused, in less than 90 seconds, a nine hour outage of the Hydro-Québec's electricity transmission system, plunging people into darkness.
- In July 2000, during solar cycle 23, Earth experienced the impact of a severe solar flare, accompanied by a solar particle event and a coronal mass ejection (CME). This event, known as the "Bastille Day Event," marked the first major solar storm following the launch of the Solar and Heliospheric Observatory (SOHO). The groundbreaking data collected by the pioneering SOHO satellite provided researchers with rapid insights into the physics of extreme flares. The effects of this event were felt not only by satellites like the Astro-D (ASCA), which experienced drag and spacecraft charging issues ([Cannon et al., 2013](#)), but also by individuals aboard commercial flights at high latitudes.
- In October 2003, Earth experienced a series of solar flares and coronal mass ejections (CMEs) that are commonly referred to as the "Halloween Storms" ([Gopalswamy et al., 2005](#)). These events had a significant impact on our planet. One consequence was the disruption of civil aviation radiocommunications above 57°N due to variations in ionospheric density ([Bothmer and Daglis, 2007](#)). Additionally, there was a one-hour shutdown of a section of the high-voltage electricity transmission system in Malmö, southern Sweden ([Pulkkinen et al., 2017](#)).
- In 2012, during solar cycle 24, a large CME missed passed close to Earth but missed it. Several studies tried to evaluate the risk that this CME represented for Earth and concluded that it was an equivalent to the event from 1989 and 2003 ([Ngwira et al., 2013](#)).
- As we previously mentioned in Section 1.3.3.2, there was a notable incident in 2022 where SpaceX experienced the loss of 40 satellites during a geomagnetic storm. This unfortunate event occurred as a result of an upsurge in thermospheric density, causing a significant increase in drag force. Consequently, the satellites used all their fuel in their attempt to reach their designated orbit, ultimately leading to their incineration upon reentry into the atmosphere ([Fang et al., 2022](#)).

1.3.4.2 Dangers

Space Weather represents a significant hazard, as evidenced by the various events that have occurred in the past. Numerous human domains are susceptible to solar activity. Below, we have compiled a non-exhaustive list of domains that can be impacted, along with a few examples, thanks to the work by [Lilensten, Jean et al. \(2021\)](#).

- Human health: The most severe solar flares are lethal to humans. In 1972, the estimated dose received on the Moon was 7 sieverts per hour (averaging 0.0024 per year, with a lethal

dose at 2). At conventional aircraft altitudes, the dose is approximately 20 microsieverts per hour, but a solar event could increase these values by a factor of 1000.

- **Constellations:** A "Carrington-type" event could potentially disrupt satellite constellations, resulting in a catastrophe with costs 10 to 100 times higher than the most powerful terrestrial cyclones. Clients, operators, and especially insurers would be severely impacted, potentially leading to insurer bankruptcies and destabilizing the banking system.
- **Debris:** A solar flare can increase ultraviolet radiation, which excites the upper atmosphere, causing it to heat up and expand (section 1.3.3.2). This leads to increased friction on Low Earth Orbit (LEO) satellites and carries away debris. Consequently, the two global tracking centers (NASA Langley & ESA Darmstadt) lose track of the debris.
- **Corrosion:** Satellites below 1000 km are exposed to atomic oxygen, which ionizes during solar activity and corrodes spacecraft surfaces more effectively. This could potentially lead to a runaway Kessler effect.
- **GPS:** Solar activity affects the electronic content of the ionosphere, leading to signal deviations and inaccuracies in GPS navigation. This can result in misguidance for missiles, triggering dam sirens to evacuate cities, and causing disruptions to autonomous vehicles.
- **Embedded Electronics:** During periods of low solar activity, cosmic radiation, including gamma rays, muons, positrons, and mesons, can penetrate the atmosphere and damage embedded electronics in various systems, such as trains and aircraft.
- **Telecommunications:** The military uses radars that rely on ionospheric bounce to detect missiles.
- **Power Plants:** Geomagnetically induced currents (GICs) generate low-intensity continuous currents in conductive ground, which can flow into power plants and melt transformers.
- **Oil Exploration:** Underground drilling locations rely on Earth's magnetic field for positioning. A magnetic storm can cause pointing inaccuracies.
- **Airspaces:** Airports use primary and secondary radars for aircraft detection and communication within airspace. On November 4, 2015, increased solar activity resulted in a series of eruptions accompanied by radio emissions within the relevant frequency band, impacting Western Europe. The effects included flight delays, passenger diversions, traffic congestion, and other disruptions. These radio bursts and their frequency range remain unpredictable to this day.

1.3.4.3 Worst Case Scenario

To persuade operators, states or the private sector of the importance of allocating resources to protect against an imminent catastrophe, it is crucial to emphasize the potential risks and financially devastating consequences associated with geomagnetic storms, particularly Carrington-type events.

In 2013, Lloyd's published a report emphasizing the inevitability of a new Carrington-type storm (Maynard et al., 2013). According to this report, Quebec-like storms occur every 50 years, while Carrington-like event could occur every 150 years. The consequences could be catastrophic, with an estimated 20 to 40 million Americans enduring prolonged power outages lasting from 16 days to 2 years, resulting in costs estimated between \$600 billion and \$2.6 trillion. Furthermore, the Swiss Academy of Sciences suggests that a complete recovery after such a storm would take between 4 and 10 years. This evaluation illustrates the magnitude of potential financial losses, highlighting the importance of investing in adequate protective measures. IN 2011, the Organization for Economic Co-operation and Development (OECD) classified geomagnetic storms as one of the five major global risks, alongside financial risks, cyber-risks, social unrest, and pandemics.

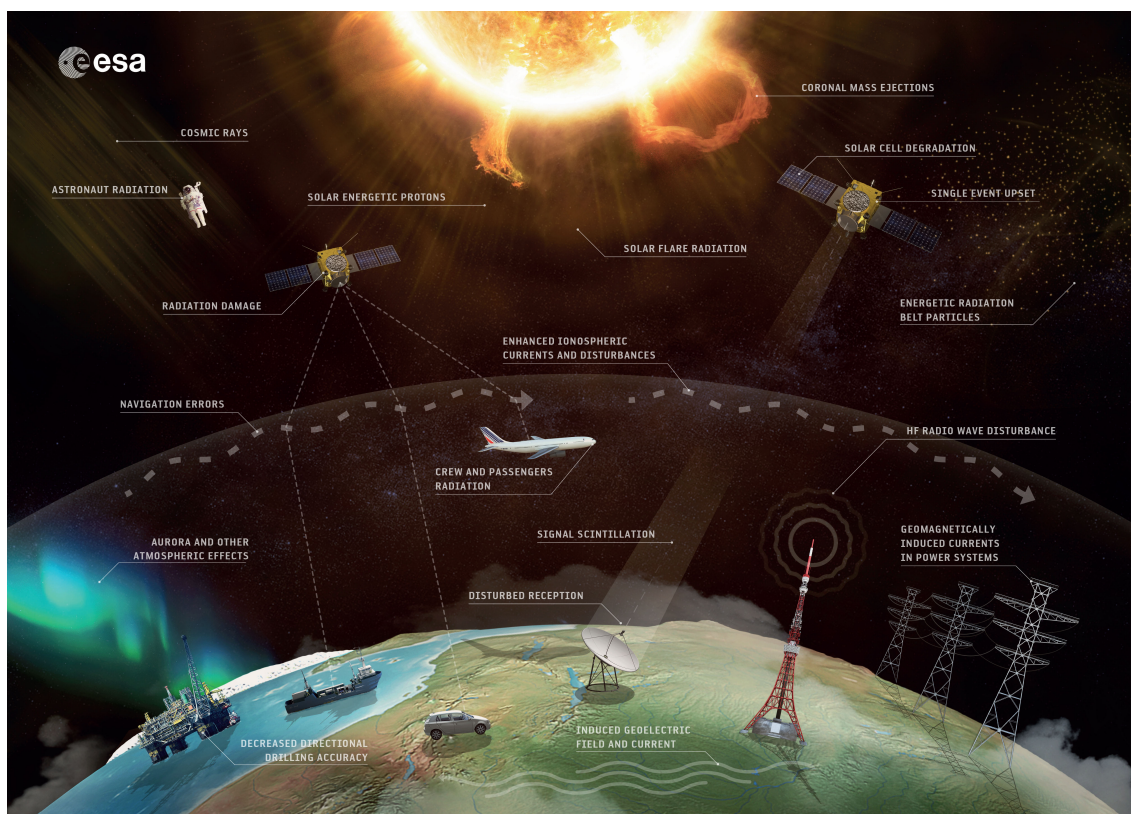


Figure 1.39 – Space Weather Effects Overview. Credits: ESA.

An additional study conducted in England concluded that the probability of a Carrington-like event occurring within the next decade is approximately 12% (Riley, 2012). Moreover, researchers, through the observation of solar-type stars, have estimated that a storm 1000 times more powerful than Carrington could arise approximately once every 5000 years, while a storm 100 times more energetic may occur about once every 800 years (Maehara et al., 2012). Additional evidence from tree-ring observations by Dr. Fusa Miyake and colleagues also suggests that such events have happened in our past (Miyake et al., 2012). Although these findings remain subjects of ongoing discussion and controversy, they highlight the gravity of addressing the risks associated with geomagnetic storms earnestly. The repercussions of such an event would be catastrophic, including potential fatalities among astronauts, core meltdowns in power plants, and fatal radiation exposure for airline passengers.

1.3.4.4 International Programs

Several international programs appeared, such as the recent E-SWAN (*European Space Weather and Space Climate Association*, initially Quo Vadis) community¹⁰, an international non-profit association established in 2022 which mission is to "unite, sustain, and develop Space Weather and Space Climate activities in Europe" (Lilensten, Jean et al., 2021).

ESA's *Space Safety Program* (S2P), a two-aspect programme, successor of the ESA's *Space Situation Awareness* (SSA) program. On one hand, research and development activities and pre-operational services aimed at warning users and protecting them against various threats from the space environment. Three threats are considered: space weather, near-earth object (NEO) including threats of asteroid collisions with our planet or our infrastructure and space surveillance

10. <https://eswan.eu/>

and tracking (SST) including all risks of collisions between satellites or with debris. On the other hand, the S2P Program focuses on the development of observational satellites or ground stations to study the three aforementioned threats. We can mention several missions:

- In space weather: The Vigil mission¹¹ mission is being developed to monitor the trajectory of solar ejecta from the Sun towards Earth. Within the S2P Program, instrumentation is being developed for observing space weather phenomena, which will be integrated into planned satellite launches. The Program also funds the operations of the PROBA-2 satellite, made in Belgium, which has been observing the solar atmosphere since 2009.
- In NEO: The HERA mission will observe and characterize the asteroid Dimorphos. This mission is supposed to help us determine how much we can alter orbits of asteroids that could potentially collide with Earth.
- In SST: the development of radar systems to monitor orbits of satellites and debris. Missions like ClearSpace-1 for instance are part of the larger *Active Debris Removal/In-Orbit Servicing* missions and will remove large piece of debris from space. An automatic system called CREAM (*Collision Risk Estimation and Automated Mitigation*) is also under development to prevent collisions.

1.3.5 The Rise of the New Space

Since the dawn of the space age, the space industry has predominantly been driven by major space agencies such as NASA, ESA, and Russian Roscosmos, along with prominent aerospace players like Airbus, Thales, Boeing, and Dassault. This situation was primarily shaped by the exorbitant costs associated with space missions, requiring extensive expertise and advanced technologies. For instance, the Apollo mission alone employed over 400,000 individuals, with a total cost equivalent to over 100 billion dollars in today's terms (Vernile, 2018). However, in recent decades, a wave of young actors and entrepreneurs has entered the arena, seeking their share of the space market. Technological advancements, increased private sector investments, and the soaring demand for space-related data have provided opportunities for young private companies to carve out their niche (with some even becoming major players) and allowed non-space companies, including tech giants like Google and Facebook, to venture into this domain and explore synergies between ICT and space applications (Vernile, 2018). Notable figures in this emerging landscape include Elon Musk with SpaceX, Jeff Bezos with Blue Origin, and Richard Branson with Virgin Galactic. In 2021, a report by SpaceTech Analytics¹² (Analytics, 2021) identified over 10,000 companies (52% of which were American) operating in the space sector, a number that continues to grow. Consequently, space, once solely governed by government institutions, has gradually come under the influence of American giants (Pasco, 2017).

The development of reusable launch vehicles, SmallSats, and CubeSats has significantly reduced the cost of space system development and payload deployment into space. Private sector investments in the space industry have soared (with an average of \$1.5 billion invested annually in space start-ups in the US during the period 2010-2015), resulting in heightened competition, innovation, and the emergence of novel business models like mega-constellations. Noteworthy projects such as Starlink, OneWeb, Kuiper, Lynk, SatRevolution, Sfera, and Guowan plan to launch between 600 to 12,000 satellites per constellation, primarily focusing on satellite communications (Kodheli et al., 2021). Furthermore, space data is now considered a valuable resource¹³, and the integration of advanced technologies such as AI, big data, and blockchains with new satellite capabilities has given rise to innovative applications. As a result, there is a significant surge in activities

11. https://www.esa.int/Space_Safety/Vigil

12. <https://analytics.dkv.global/spacetech/SpaceTech-Industry-2021-Report.pdf>

13. <https://www.mews-partners.com/space-data-the-golden-age/>

related to space data, as well as the development of ancillary products and services such as on-orbit servicing. Specialized companies now provide high-quality data directly to customers, making it possible for non-space companies to send their own satellites into orbit by leveraging comprehensive service providers throughout the entire value chain, from design and manufacturing to launch.

The imperative to develop a more extensive space ecosystem has never been greater, necessitating collaboration among various stakeholders, including space companies, government agencies, non-space enterprises, and academia. As an example, in order to nurture the most innovative initiatives on the continent and support startups at every stage of their development, European institutions have launched the Cassini Fund¹⁴. This initiative was unveiled during the fourteenth European Space Conference held in Brussels on January 25, 2022. With a budget of one billion euros, the Cassini Fund will be backed by the European Investment Fund and the European Investment Bank.

Nevertheless, the space sector faces numerous challenges, as highlighted in a recent report by Deloitte¹⁵: supply chain disruptions, developing cost-competitive space-grade products, regulatory requirements, a shortage of qualified talent, reduced capital investment, mass production, miniaturization of electronic components, government acquisition timelines, funding, shifting defense priorities, security concerns, greenhouse gas emissions, environmental impact at launch sites, radiofrequency congestion, and perhaps the most emblematic challenge of all, space debris. Numerous companies, including SpaceAble, LeoLabs, and ShareMySpace, have emerged to tackle these challenges and contribute to the growth and sustainability of the space sector.

14. https://defence-industry-space.ec.europa.eu/eu-space-policy/space-entrepreneurship-initiative-cassini_en

15. <https://www2.deloitte.com/us/en/insights/industry/aerospace-defense/future-of-space-economy.html>

2

Machine Learning, Deep Learning and Their Application in Space Weather Research

"I am putting myself to the fullest possible use, which is all I think that any conscious entity can ever hope to do."

HAL 9000 - 2001: A Space Odyssey - Stanley Kubrick

Contents

2.1	Introduction to Artificial Intelligence	85
2.2	Machine Learning	87
2.2.1	Machine Learning & Space Weather	88
2.2.2	Supervised Learning	89
2.2.3	Unsupervised Learning	92
2.2.4	Reinforcement Learning	94
2.2.5	Summary	94
2.3	From Supervised Machine Learning to Deep Learning	95
2.3.1	Mathematical Introduction of Supervised Learning Models	96
2.3.2	A First Neural Network	101
2.3.3	Training Neural Networks	102
2.3.4	Evaluation & Diagnostic of Neural Networks	113
2.3.5	Fine-tuning Neural Networks	119
2.3.6	Deep Learning System Design	121
2.4	Libraries & Needed Tools	137
2.4.1	Hardware	137
2.4.2	Libraries	138

2.1 Introduction to Artificial Intelligence

Everyone has heard the term "Artificial Intelligence" at least once. Although these terms and the associated research topics have been around for several decades, AI is now a topic of discussion in various contexts. It represents either a promising future or invokes fear, being seen as a threat to employment (Huang and Rust, 2018) and even speculated to bring about the end of the human species (Cellan-Jones, 2014) by some, while others view it as an extension of human intelligence (Quiñonero Candela and LeCun, 2016) and an opportunity to improve our everyday lives.

In recent years, there has been a great increase in the use of AI across diverse domains. Today, thousands of applications and software incorporate AI methods, ranging from social networks, online shopping and security systems to arts, politics, science, or literature. Industry giants like Google, Meta, and OpenAI have established themselves as leaders in the field. Well-known algorithms include Google's FaceNet and DeepDream, respectively capable of identifying faces and generating psychedelic-like images from real-world images. We can also mention OpenAI's DALL-E algorithm which generates images based on textual descriptions; OpenAI's ChatGPT, "a state-of-the-art conversational AI model"¹; DeepMind's AlphaGo that garnered attention for defeating the world's top Go player in 2016 (see Figure 2.1). Based on McKinsey's projections, AI is anticipated to make a contribution of \$13 trillion to the global economy by 2030.



Figure 2.1 – (a) AlphaGo's victory over Go champion Lee Sedol in the Google DeepMind Challenge Match, featured in Nature (January 28th, 2016). (b) DALL-E's impressive image extension capability demonstrated on Johannes Vermeer's "Meisje met de parel." (c) DeepDream output example.

To delve into the understanding of Machine Learning, we first need to take a step back and talk about the broader topic of Artificial Intelligence (AI). AI can be defined as "a system's ability to correctly interpret external data, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation." (Kaplan and Haenlein, 2019). As explained by Quiñonero Candela and LeCun (2016), "AI is a rigorous science focused on designing intelligent systems and machines, using algorithmic techniques somewhat inspired by what we know about the brain". The word "intelligence" comes more from what we would like AI to do instead

1. Description generated by ChatGpt on 20 June 2023

of what it is really doing, which is closer to "learning".

Artificial intelligence is a field of computer science that seeks to simulate the abilities of the brain. To put it simply, it functions like a large funnel where we input information and from which new information, desired by the user, emerges. Although this description closely aligns with the idea of a model, AI encompasses a wide range of techniques and methods that go beyond simple modeling. Overall, it is focused on developing systems that can perform tasks requiring human-like intelligence. Figure 2.2 provides an overview of the various research branches within AI. We will now briefly describe these branches, but it's important to note that this categorization does not represent a consensus, and there are many different approaches:

- *Machine Learning*: This is the branch we will focus on throughout this thesis. Its goal is to analyze data, learn from it, and discover relationships, patterns, or models to make intelligent decisions or perform tasks.
- *Natural Language Processing (NLP)*: NLP involves understanding and generating human language, whether in textual or audio form. It includes tasks such as language translation, sentiment analysis, or creating chatbots capable of conversation (e.g., ChatGPT²).
- *Computer Vision*: This branch focuses on enabling machines to understand, analyze, and interpret images or videos (e.g., FaceNet, Schroff et al. (2015)). It encompasses facial recognition, object detection, and has numerous applications in areas like autonomous vehicles and medical imaging.
- *Robotics*: This field involves designing, developing, and programming physical robots capable of sensing real-world data (such as temperature, movement, sound), interacting with their environment, and autonomously performing tasks.
- *Expert Systems*: In AI, an expert system is a program that aims to replicate the behavior and decision-making of a human expert (or organization) in a specific domain.
- *Planning and Decision Making*: This branch focuses on developing strategies or sequences of actions that are typically executed by robots or autonomous vehicles. It often works in conjunction with decision theory.
- *Knowledge Representation*: This field aims to represent information and data in a form that a computer system can understand and utilize to solve complex tasks.
- *AI Ethics*: With the advent of AI, a new branch that is not strictly part of AI has emerged. Its objective is to question and ensure ethical considerations in the use of AI.

When considering an AI solution to address a problem, its implementation typically follows a similar procedure: data preparation, model creation, design of the system on which the model will run, and deployment on hardware or enterprise systems. Throughout this process, the data preparation stage undoubtedly demands the most time and resources. It requires domain expertise in dealing with often massive amounts of data, and it is generally this stage that determines the feasibility of the mission. We will further discuss the importance of data preparation within the context of machine learning, which is the focus of our interest.

2. <https://openai.com/chatgpt>

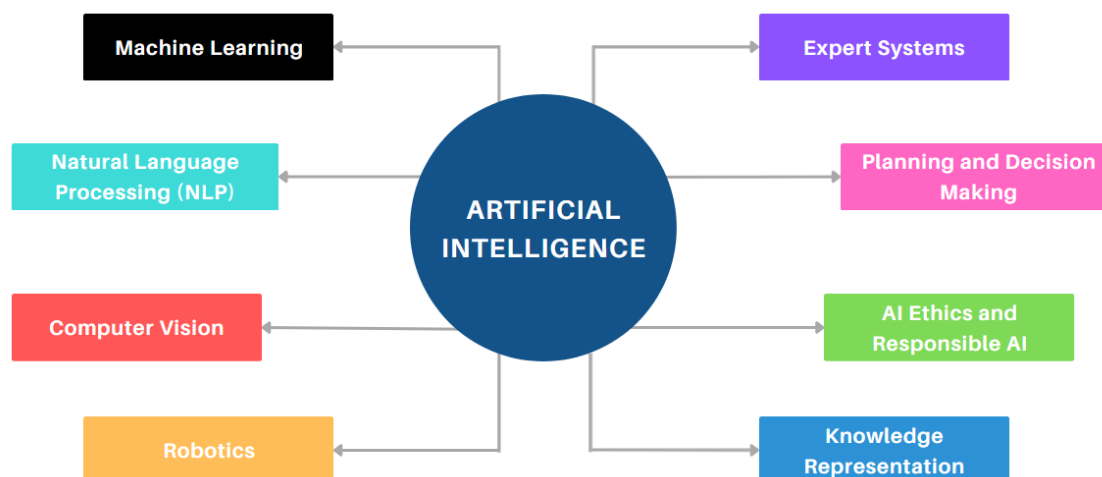


Figure 2.2 – Different branches of AI. Note that "AI Ethics and Responsible AI" could be represented differently from other branches, as it is not exactly a branch in itself but rather a cross-cutting consideration across all branches.

2.2 Machine Learning

Machine Learning techniques comprise a collection of algorithms employing statistics and mathematics to extract knowledge from data. According to the Oxford Dictionary, machine learning is defined as "the use and development of computer systems capable of learning and adapting without explicit instructions, employing algorithms and statistical models to analyze and draw inferences from patterns in data." The outcomes of machine learning encompass models, forecasts, as well as the identification of patterns, anomalies, relationships, and even causalities within the data. Generally, machine learning involves supervised and unsupervised learning, although modern approaches like reinforcement learning and semi-supervised learning are also worth mentioning. Figure 2.3 illustrates the different categories of machine learning algorithms.

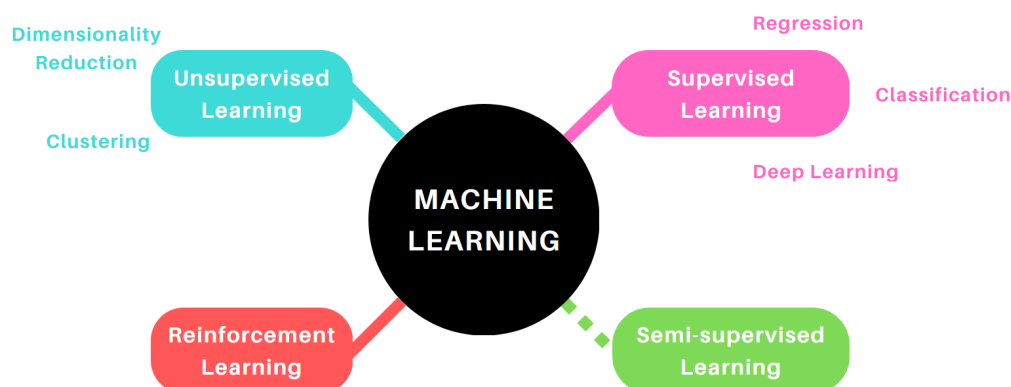


Figure 2.3 – Categories of Machine Learning algorithms.

In the subsequent sections, we will provide an introductory overview of the interplay between machine learning and space weather, elucidating the emergence of novel opportunities stemming from the availability of large datasets in the field of Space Weather and recent advancements in machine learning. Subsequently, we will delve into comprehensive descriptions of supervised, unsupervised, and reinforcement learning techniques, accompanied by selected examples showcasing the application of these algorithms within the domain of space weather. For most of the

descriptions, we rely on the very comprehensive resources from IBM Jones (2017)³.

2.2.1 Machine Learning & Space Weather

In the field of space weather, there have been numerous endeavors to utilize machine learning since the 1990s, particularly the application of neural networks for predicting geomagnetic indices. Dr. Enrico Camporeale conducted a comprehensive review of machine learning works in space weather (Camporeale, 2019), which yielded two primary reasons to continue exploring machine learning techniques. Firstly, not all possibilities have been exhausted. For instance, convolutional neural networks (to be discussed in Section 2.3.6.3), one of the most successful applications in machine learning (LeCun et al., 2015), have received limited attention within our community. Secondly, the recent success of ML can be attributed to three significant factors: the increased availability of large datasets, the advancements in software with open-source libraries, and the improved hardware capabilities, particularly powerful GPUs. Consequently, it is an opportune time to reevaluate ideas proposed 10 or 20 years ago, as approaches that previously seemed ineffective may now yield remarkable results. As such, the amount of papers in the field of machine learning applied to space weather has drastically increased. Figure 2.4 reflects this trend in AGU Space Weather Journal.

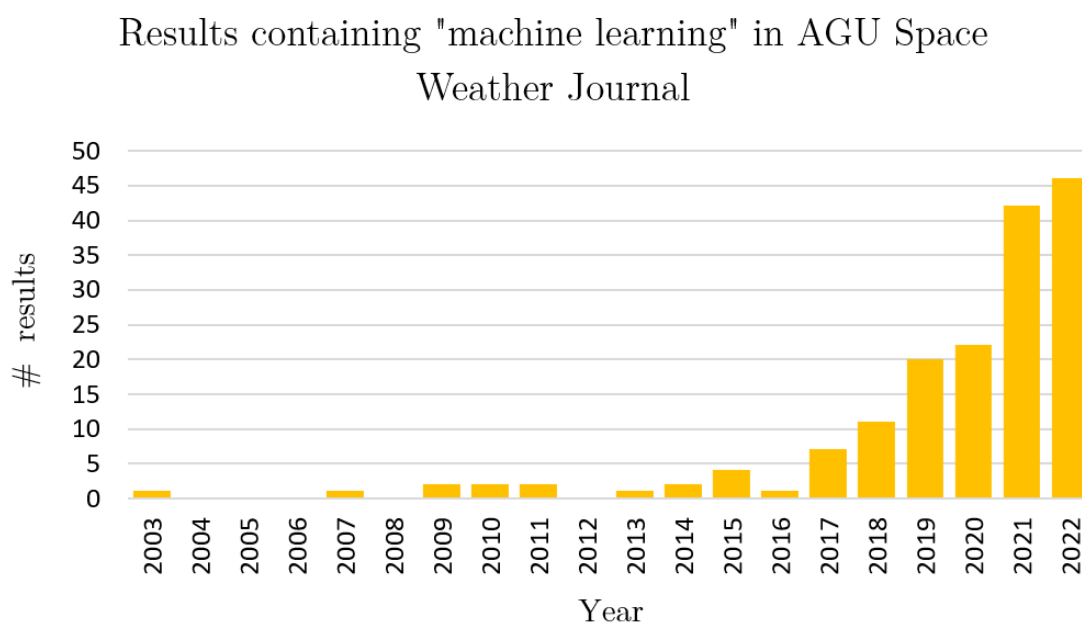


Figure 2.4 – Search results highlighting the occurrence of the words 'machine learning' in the AGU Space Weather Journal website's search bar <https://agupubs.onlinelibrary.wiley.com/journal/15427390>. 183 results that can be research articles (158), technical articles (7), editorial (5), issue information (4), commentary (3), commissioned manuscript (1), feature (2), meeting report (1), news article (1). Words can appear in either the title, keywords, abstract, author affiliation or funding agency of the results.

Regarding datasets, as indicated in Camporeale (2019), the Space Weather community offers an extensive and openly accessible dataset comprising decades of in situ and remote observations. The utilization of large datasets holds significant importance in the development of successful

3. <https://developer.ibm.com/articles/cc-models-machine-learning/>, last accessed June 22 2023

neural network models. Depending on one's objectives and given the 11-year duration of the solar cycle, it can be imperative to possess sizable datasets that encompass several solar cycles. In the following tables (2.1 & 2.2) multiple examples of significantly large datasets suitable for AI applications in Space Weather are presented, as well as libraries, software or already processed datasets for machine learning. This list is non-exhaustive but gives a great glimpse of the vast work done on the topic. However, as we elaborate in Chapter 3, it's crucial to recognize that while these data may be available, their suitability for direct application in machine learning tasks can be limited or non-trivial, requiring preprocessing or transformation steps to enhance usability.

Mission	Website
ACE	http://www.srl.caltech.edu/ACE/
Wind	https://wind.nasa.gov/
DSCOVR	https://www.nesdis.noaa.gov/content/dscovr-deep-space-climate-observatory
SOHO	https://sohowww.nascom.nasa.gov/
STEREO	https://stereo.gsfc.nasa.gov/
SDO	https://sdo.gsfc.nasa.gov/
OMNI	https://omniweb.gsfc.nasa.gov/index.html
VAP	http://vanallenprobes.jhuapl.edu/
GOES	https://www.goes.noaa.gov
POES	https://www.ospo.noaa.gov/Operations/POES/index.html
GPS	https://www.ngdc.noaa.gov/stp/space-weather/satellite-data/satellite-systems/gps/
DMSP	https://www.ngdc.noaa.gov
On-ground magnetometers	http://www.intermagnet.org
GONG	https://gong.nso.edu/

Table 2.1 – Examples of large datasets used in Space Weather, from *Camporeale (2019)*. Note: ACE = Advanced Composition Explorer; DSCOVR = Deep Space Climate Observatory; SOHO = Solar and Heliospheric Observatory; STEREO = Solar Terrestrial Relations Observatory; SDO = Solar Dynamics Observatory; VAP = Van Allen Probes; GOES = Geostationary Operational Environmental Satellite system; POES = Polar Operational Environmental Satellites; DMSP = Defense Meteorological Satellite Program; GONG = Global Oscillation Network Group.

Finally, it is worth mentioning that while not the primary approach utilized here, Space Weather is ideally suited for what is known as grey-box models (*Bohlin, 2006; Camporeale, 2019*). A grey-box model is a combination of a white-box model and a black-box model. In our field, a white-box model refers to a physics-based model that operates on established formulas and physics approximations, where the relationship between input and output is known. On the other hand, a black-box model is characterized by its unknown internal functioning, typically limited to specific datasets, as seen in most neural networks. In our context, we define a grey-box model as one that leverages physics-based insights to transform inputs and/or outputs of black-box models (*Kroll, 2000*). Space Weather is well-suited for these models due to the inherent challenges posed by vast spatial and temporal scales, the limited time lag between causes and effects, and the computational power needs. The Space Weather community recognizes the limitations of a first-principle approach for forecasting and, instead, explores the potential of combining data-driven methods (black-box) with identified gaps in physics-based models (white-box) to enhance prediction capabilities.

2.2.2 Supervised Learning

Supervised learning relies on annotated training data, and is referred to as "supervised" because there is a 'supervisor' that guides the learning system by providing labels for training examples (class labels in classification tasks for instance) (*Cunningham et al., 2008*). By using this

Name	Description	Url
pyHC	A community knowledge base for performing heliophysics research in Python	https://heliopython.org/
sunpy	The community-developed, free and open-source solar data analysis environment for Python	https://sunpy.org/
SolarNet	Deep learning research toolbox focusing on self-supervised learning on solar data.	https://jdonzallaz.gitlab.io/solarnet/
SpaceML	A machine learning toolbox and developer community building open science AI applications for space science and exploration.	https://spaceml.org/
pysat	The Python Satellite Data Analysis Toolkit (pysat) is a flexible package for handling and analyzing various scientific measurements. It supports diverse types of ground- and space-based data, providing a user-friendly interface.	https://pypi.org/project/pysat/
HelioML	This book features interactive Jupyter notebooks in Python that showcase the use of machine learning, statistics, and data mining techniques on heliophysics datasets to reproduce published results.	https://helioml.org/Introduction/title.html
AIDapy	AIDapy is a high level Python package for the analysis of spacecraft data from heliospheric missions using modern techniques	https://gitlab.com/aidaspace/aidapy
SDOML	A curated data set from the NASA Solar Dynamics Observatory (SDO) mission in a format suitable for machine-learning research.	https://sdoml.github.io/
DMSP AI-ready	A DMSP ready-to-use dataset for AI, linked to the work by	https://zenodo.org/record/4281122
SDO AI-ready	A Machine learning-ready dataset prepared from the NASA's SDO mission	https://doi.org/10.3847/1538-4365/aba82f
MVTS	Multivariate time series (MVTS) data extracted for space weather data analytics	https://doi.org/10.1038/s41597-020-0548-x

Table 2.2 – Examples of libraries (top) and AI-ready datasets (bottom) for applying machine learning in space weather.

labeled data, supervised learning algorithms generate models that can classify unlabeled data. Supervised learning mainly consists of *regression algorithms* and *classification algorithms*, each serving different purposes.

Deep learning, as we will see Section 2.3 can be considered as a subfield or an extension of supervised learning. While supervised learning encompasses various algorithms and techniques, deep learning is a specific approach that utilizes neural networks with multiple layers (hence the term "deep") to learn and extract representations from data.

2.2.2.1 Regression

Regression tasks focus on identifying relationships between input and output data, usually predicting results within a continuous output. They encompass a variety of techniques and methods such as linear and polynomial regressions, decision trees, neural networks, or ensemble methods. Regressions are commonly used to approximate functions or predict future values of continuous functions. For instance, they can be employed to forecast weather, estimate housing prices, or model temperature trends, and hence are largely used within the space weather community (Bouriat et al., 2022).

2.2.2.2 Classification

Classifications are designed to map input variable to specific classes or categories. Algorithms to perform this task assign labels or class memberships to input instances based on their features. Examples of algorithms used to perform classifications include support vector machines (Vapnik, 1999), discriminant analysis, naive Bayes, and k-nearest neighbors. Classifications are frequently used to solve problems such as spam detection, sentiment analysis, or medical diagnosis, where the goal is to assign a predefined class to each input sample such as a True/False problem.

Among the algorithms that exist, we find the *logistic regression*, detailed Section 2.3.1, that simply models the relationship between the input variables and the probability of belonging to a particular class. This type of classification may fail when the relationship between input and output is nonlinear. In such cases, algorithms like decision trees can be used. Tree-based models make binary decisions at each node, leading to the "splitting" of data, classifying instances by following a sequence of rules (Figure 2.5). During the training process, the decision tree algorithm searches for the features and cutoff values that provide the best division of the data based on the objective.

2.2.2.3 Supervised Learning and Space Weather

Supervised learning is the most commonly used method in space weather meteorology. As discussed in Section 2.2.1, machine learning has made significant advancements in the field of space weather meteorology in recent years. However, as early as the 1990s, applications of these methods were already found in nowcasting and forecasting certain variables, such as geomagnetic indices using data located at the Lagrange point L1 (Gleisner et al., 1996; Lundstedt and Wintoft, 1994; Macpherson et al., 1995), or the solar cycle (Ashmall and Moore, 1997; Calvo et al., 1995; Fessant et al., 1996). The review by Camporeale (2019) provides numerous citations and examples of supervised algorithms applied to space weather meteorology, which will not be reiterated here. To complement this, we can present some recent research that has been published since that review. Similar to previous years, advancements have been made in predicting geomagnetic indices such as SYMH (Bhaskar and Vichare, 2019; Siciliano et al., 2021), Kp (Sexton et al., 2019), or Dst (Hu et al., 2023; Park et al., 2021). Furthermore, studies have been conducted on modeling and predicting solar wind properties, such as speed (Brown et al., 2022) and magnetic field (Reiss et al., 2021), as well as geomagnetically induced currents (GICs) (Bailey et al., 2022; Smith et al.,

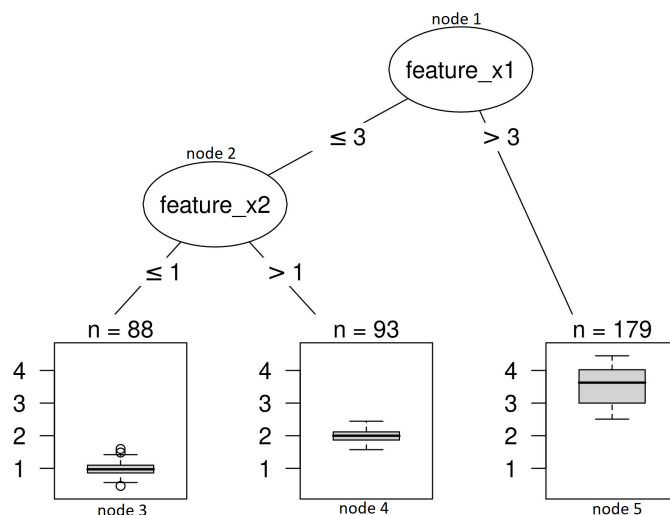


Figure 2.5 – Decision tree with artificial data. Instances with a value greater than 3 for feature $x1$ end up in node 5. All other instances are assigned to node 3 or node 4, depending on whether values of feature $x2$ exceed 1. Credits: Christophe Molnar, *Interpretable Machine Learning*, <https://christophm.github.io/interpretable-ml-book/>.

2021), solar proton events (Stumpo et al., 2021), solar energetic particles (SEPs) (Kasapis et al., 2022), electron fluxes in radiation belts (Tang et al., 2022), and Total Electron Content (TEC) (Lee et al., 2021; Liu et al., 2020; Zewdie et al., 2021).

2.2.3 Unsupervised Learning

In unsupervised learning, we have limited or no prior knowledge of the desired outcome (Barlow, 1989). Unlike supervised learning, where we provide labeled data and specify expected results to the algorithm, unsupervised learning involves analyzing and clustering unlabeled datasets. Through these algorithms, hidden patterns or data groupings can be discovered without requiring human intervention. The capacity to identify similarities and differences in information makes unsupervised learning an ideal solution for exploratory data analysis, cross-selling strategies, customer segmentation, and image recognition (Ghahramani, 2004). Common examples of unsupervised learning include *clustering* and *dimensionality reduction* algorithms.

2.2.3.1 Clustering

Clustering algorithms are used to group input data into specific categories based on identifiable structures or patterns within the data, such as segmenting customer profiles. They fall under the category of unsupervised learning because the desired output, such as the number of groups, is unknown (IBM, 2021).

One prominent example of a clustering algorithm is the widely known and extensively used *k-means clustering*. The *k-means* algorithm aims to group similar data points together by iteratively adjusting cluster centers until the points are effectively organized into meaningful clusters. Although *k-means* is an unsupervised algorithm, it requires prior knowledge of the desired number of clusters. To address this limitation, a solution called *U-k-means* has been proposed in (Sinaga and Yang, 2020). Figure 2.6 illustrates the results of the *U-k-means* algorithm.

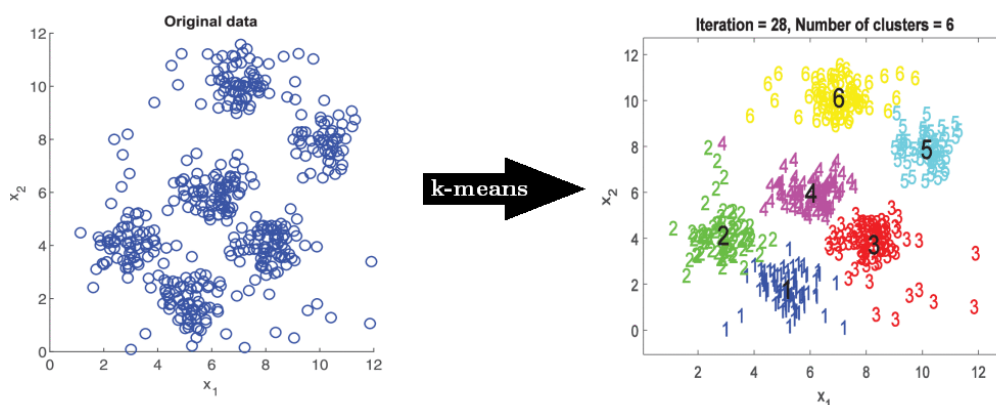


Figure 2.6 – Example of the U-k-means clustering ending in a 6-cluster dataset, from [Sinaga and Yang \(2020\)](#). Unlike the classical k-means algorithm, the amount of clusters is here an output.

2.2.3.2 Dimensionality Reduction

As we can see in Figure 2.6, the number of inputs or *features* also represents the dimensionality of our data "space." Each feature denotes a distinct aspect or property of the data, and the dimensionality is determined by the total number of these features. In Figure 2.6, we only have two features (x_1 and x_2), representing a 2D space in which we seek clusters. Intuitively, one might think that having more features would result in more accurate outcomes. However, an excessive number of features (or dimensions) can lead to a decline in model performance, even if all features are relevant to the task at hand. This issue is particularly pronounced in empirical modeling within machine learning, impacting the performance of machine learning algorithms (e.g., leading to overfitting - see Section 2.3.4.3) and making datasets challenging to visualize. This challenge underscores the importance of dimensionality reduction.

Dimensionality reduction aims to mitigate this issue by reducing the number of data inputs to a manageable size while preserving the dataset's integrity as much as possible. It is a common preprocessing technique. Notable methods include *principal component analysis* (PCA), and *Autoencoders*. PCA is a linear method, while Autoencoders, being neural networks, can capture non-linear relationships as well.

Additionally, beyond these methods, classical (non-deep learning) non-linear techniques play a significant role in dimensionality reduction. Some examples include:

- Diffusion Map ([Coifman et al., 2005](#)): This method is particularly useful for capturing the underlying geometric structure of high-dimensional data by modeling the diffusion process on the data manifold.
- Locally Linear Embedding (LLE) ([Roweis and Saul, 2000](#)): LLE aims to preserve local relationships between data points, capturing the intrinsic geometry of the data manifold. It works by reconstructing each data point as a linear combination of its neighbors.
- Isometric Mapping or ISOMAP ([Tenenbaum et al., 2000](#)): ISOMAP is a technique for dimensionality reduction that seeks to preserve the geodesic distances between all pairs of data points. It does so by approximating the intrinsic geometry of the data manifold.
- t-Distributed Stochastic Neighbor Embedding (t-SNE) ([Hinton and Roweis, 2002](#)): t-SNE is particularly effective for visualizing high-dimensional data by mapping them to a lower-dimensional space while preserving the local structure of the data. It is often used for exploratory data analysis and visualization.

Integrating these classical non-linear methods into dimensionality reduction discussions provides a more comprehensive understanding of the techniques available and their respective strengths and weaknesses. They offer valuable alternatives to PCA and Autoencoders, especially when dealing with complex data distributions and non-linear relationships.

2.2.3.3 Unsupervised Learning and Space Weather

A significant majority of machine learning projects in space weather applications belong to the category of supervised learning. However, there are a few noteworthy projects in the unsupervised domain as well. These include the automatic classification of plasma regions or distributions (Bakrania et al., 2020; Olshevsky et al., 2021) and magnetospheric particle distributions (Souza et al., 2018). More recently, unsupervised learning has also been employed to identify coronal holes in solar images (Inceoglu et al., 2022), classify solar wind (Amaya, 2019; Amaya et al., 2020; Heidrich-Meisner and Wimmer-Schweingruber, 2018; Teichmann et al., 2023), and recognize specific space weather events (Bals et al., 2022; Marlowe, 2022; Yeakel et al., 2022).

2.2.4 Reinforcement Learning

Reinforcement learning is a powerful learning model that goes beyond mapping individual inputs to outputs. It learns to map sequences of inputs to outputs, considering dependencies as seen in Markov decision processes. In reinforcement learning, the focus is on states within an environment and the available actions at each state. The learning process involves exploring state-action pairs randomly to build a table of these pairs. Subsequently, the algorithm utilizes the acquired knowledge to exploit state-action pair rewards and select the optimal action for a given state, ultimately aiming to reach a desired goal state. Unlike supervised learning, where a user grades each output from the algorithm, in reinforcement learning, the user may only provide a grade when a goal state is achieved. Figure 2.7 provides a visual representation of how it works.

A notable example is showcased in IBM's resources (referenced in the caption of Figure 2.7) using a blackjack player scenario. In blackjack, an algorithm or agent learns to play by considering the sum of its cards as the state and deciding whether to hit or stand. The agent undergoes training through numerous hands of blackjack, receiving rewards for winning or losing. Each hand is assigned an arbitrary state number. Here are some examples:

- For a state of "10," hitting is the optimal choice with a reward of 1.0, while standing has a reward of 0.0.
- In the state of "20," standing is the best option with a reward of 1.0, while hitting has a reward of 0.0.
- For a more complex situation, a state of 17 may have action values of 0.95 for standing and 0.05 for hitting. Consequently, the agent would predominantly stand (95% of the time) and occasionally hit (5% of the time) based on probabilities.

The rewards are acquired over multiple poker hands, aiding the agent in making optimal choices for different states or hands, with the goal of reaching a total card sum of 21. In this context, similar to other reinforcement learning methods, the agent must determine the actions that led to receiving rewards or punishments.

2.2.5 Summary

With these three types of machine learning models, we usually cover most of what exists in the field. Despite recent advancements in techniques and capacities, newly developed algorithms typically fall into one of these three categories. The structures for the three types are summarized in Figure 2.8. We can synthesize their functioning as follows:

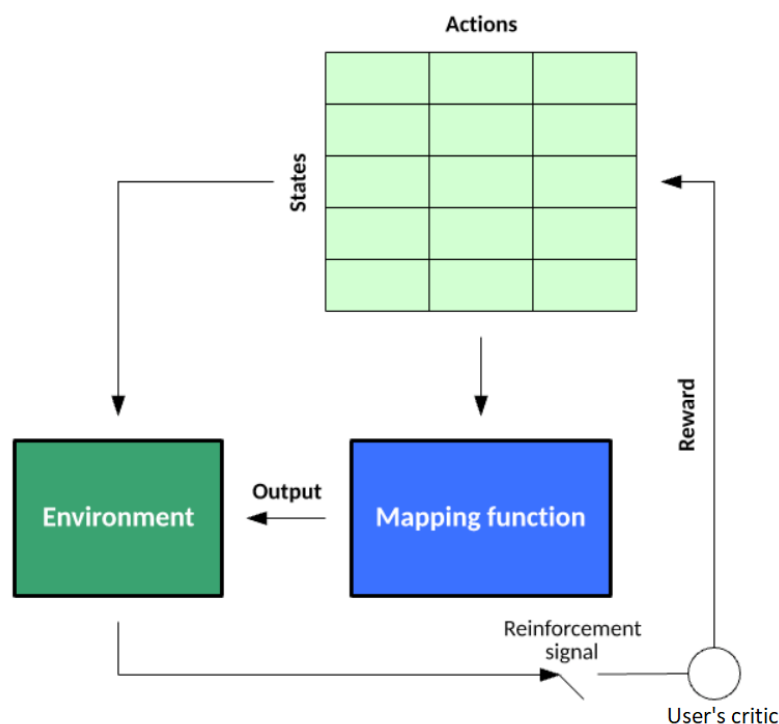


Figure 2.7 – Simplified structure of reinforcement learning algorithms, adapted from IBM's resources, <https://developer.ibm.com/articles/c-models-machine-learning/> (Jones, 2017).

- Supervised learning: A dataset includes desired outputs called labels. By comparing the predicted output to the desired output, the algorithm learns from the resulting error and adjusts its mapping.
- Unsupervised learning: No labels, so the algorithm aims to segment the dataset into classes, grouping similar data based on common features.
- Reinforcement learning: Here, the algorithm learns actions for states that lead to a goal state. Errors are not provided after each example but are received through reinforcement signals, similar to how humans learn with feedback based on rewards.

Finally, it is worth mentioning that other types exist such as *semi-supervised learning*. Semi-supervised learning uses both supervised and unsupervised learning. During training, it employs a smaller labeled dataset to guide classification or feature extraction from a larger unlabeled dataset. This approach can address issues like limited amount of labeled data for supervised learning algorithms.

2.3 From Supervised Machine Learning to Deep Learning

Deep learning is a field of artificial intelligence that empowers computers to learn and make intelligent decisions by mimicking the workings of the human brain through the utilization of deep neural networks. It "allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction" (LeCun et al., 2015). We will delve into the mathematical aspects of deep learning, exploring its fundamental foundations and principles. Starting with a simple supervised learning technique called linear regression, we will

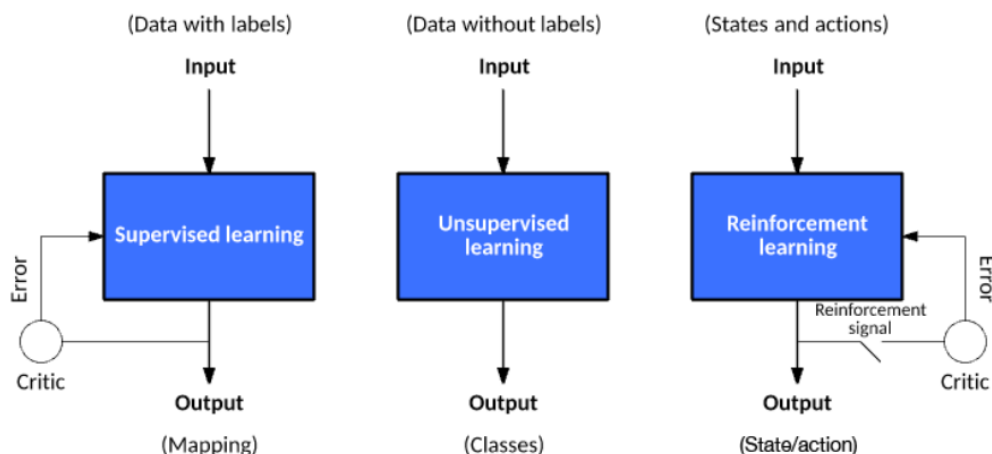


Figure 2.8 – Simplified and summarized structures for all main three types of machine learning algorithms: supervised, unsupervised and reinforcement learning, adapted from IBM’s resources, <https://developer.ibm.com/articles/cc-models-machine-learning/>, last accessed June 22, 2023 (Jones, 2017).

progressively expand our knowledge, eventually culminating in a comprehensive understanding of neural networks.

2.3.1 Mathematical Introduction of Supervised Learning Models

Let’s introduce the fundamental mathematical concepts underlying some basic supervised learning models. We specifically focus on supervised learning, as it is the predominant approach in machine learning for Space Weather applications. Furthermore, the mathematical principles presented in this section will serve as a foundation for understanding the mathematics behind neural networks (NNs) and temporal convolutional networks (TCNs), which will be elaborated upon in Section 2.3. These two algorithms constitute the primary focus of this thesis.

We will start with linear regression and subsequently extend our exploration to encompass linear regression with multiple variables and then logistic regression. This will give us the keys to understanding the concept of neural networks.

Linear Regression

In a supervised learning algorithm we have a set of the feature x containing i values (our inputs) and a corresponding set y containing i values (our outputs) to train our algorithm (see Section 2.2.2). Hence, one training example is denoted (x^i, y^i) and our training set is (x, y) containing, let’s say, m training examples. Our goal from there is to approximate a function $h : X \rightarrow Y$ such that $h(x)$ is a good approximation of y . The letter h stands for *hypothesis*.

Let’s imagine a linear approximation for h :

$$h_{\theta}(x^i) = \theta_0 + \theta_1 x^i \quad (2.1)$$

The goal there is to find the values of the set θ_0, θ_1 such that for all i , $h_{\theta}(x^i)$ (also written \hat{y}^i) is as close as possible to y^i . This is actually an optimization problem. We want to minimize a *cost function* J :

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^i) - y^i)^2 \quad (2.2)$$

Please disregard the specific parameters in the function notation. We only wrote θ_0, θ_1 as parameters, but the cost function obviously depends on the expected and predicted outputs (and so on inputs), as well as on regularization parameters (explained Section 2.3.4). One possible way to minimize such function is to use the *gradient descent* method. The method is quite straightforward. Imagine being on a surface in a 3-D space where the x and y directions represent the values of θ and z represents the altitude corresponding to the value of $J(\theta_0, \theta_1)$. We want to find the minimum on this surface. First, we choose a starting point for θ_0, θ_1 . Then by differentiating the cost function at this specific location, we obtain the slope direction. We then modify θ_0, θ_1 so that we go in this direction to reduce the cost function. We can see this as taking a step going down in the surface. The "length" of our step is called the *learning rate* α . We then obtain a new location, and do the all process again.

Mathematically we repeat until convergence (i.e. until we are close enough to a local or global minima) for $j = 0$ and $j = 1$ simultaneously:

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1) \quad (2.3)$$

The first θ_j on the left side is the new value, while the θ_j on the right side was the old value. As we approach a local minimum, the gradient descent must slow down i.e. the learning rate must decrease. We can easily imagine that with a large learning rate (big "steps") we can overshoot the minimum in our surface. On the other hand, with small learning rate, the gradient descent can be too slow and stay stuck in a local minimum.

In our case of linear regression with one variable, we obtain:

$$\begin{cases} \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=1}^m (\theta_0 + \theta_1 x^i - y^i) \\ \frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=1}^m (\theta_0 + \theta_1 x^i - y^i) x^i \end{cases}$$

It is usually more efficient to use linear algebra to write down these equations. Our equations are the following:

$$\begin{bmatrix} h_\theta(x^1) \\ h_\theta(x^2) \\ \dots \end{bmatrix} = \begin{bmatrix} 1 & x^1 \\ 1 & x^2 \\ \dots & \dots \end{bmatrix} \times \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} \quad (2.4)$$

Now, let's do everything again in the case of multiple variables. Let's first imagine one simple problem that we are trying to resolve: find the age of a tree based on several characteristics (trunk diameter, height, etc.) that are called *features*. Let's say that we have n features and m training examples. They are denoted x_j . As we previously wrote, one training example is then denoted x_j^i which is the i^{th} example of the j^{th} feature (let's say for instance the trunk diameter is the feature and the example is 25 cm). Each tree is represented by a set of features and represent for us one training example. As we have n features we can now write the hypothesis function for a given example as follow:

$$h_\theta(x^i) = \theta_0 + \theta_1 x_1^i + \dots + \theta_n x_n^i \quad (2.5)$$

If we write, $x_0^i = 1$ for all i , then we can write:

$$x^i = \begin{bmatrix} x_0^i & x_1^i & \dots & x_n^i \end{bmatrix} \in \mathbb{R}^{n+1} \text{ and } \Theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \dots \\ \theta_n \end{bmatrix} \in \mathbb{R}^{n+1} \quad (2.6)$$

And if we write

$$X = \begin{bmatrix} x^1 \\ x^2 \\ \dots \\ x^m \end{bmatrix} = \begin{bmatrix} x_0^1 & x_1^1 & x_2^1 & \dots & x_n^1 \\ x_0^2 & x_1^2 & x_2^2 & \dots & x_n^2 \\ \dots & \dots & \dots & \dots & \dots \\ x_0^m & x_1^m & x_2^m & \dots & x_n^m \end{bmatrix} \text{ and } Y = \begin{bmatrix} y^1 \\ y^2 \\ \dots \\ y^m \end{bmatrix}$$

Then we have

$$h_\theta(X) = X\Theta \quad (2.7)$$

All values in Θ have to be found such that we minimize the difference between $X\Theta$ and Y . From there, we have two possibilities to solve the problem: using the normal equation, or using the gradient descent as we previously explained.

Several gradient descent methods are detailed in Section 2.3.3.2 (and in Ruder (2016)) but the idea remains exactly the same as for one variable: we update weights by iteratively adjusting them in the direction opposite to the gradient of the cost function, aiming to minimize the loss.

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} \frac{1}{m} \sum_{i=1}^m (h_\theta(x^i) - y^i) \quad (2.8)$$

with x^i now being an ensemble of n values, for all corresponding n features.

Concerning the normal equation, our cost function is the squared difference between our output $X\Theta$ that we will note \hat{Y} and what we want Y . If we take ε as the *error* such that $\varepsilon = Y - \hat{Y}$ (a vector), we can write the sum of the squared $E = \varepsilon^T \varepsilon$. By replacing with $X\Theta$, and by linear algebra, we have:

$$\begin{aligned} E &= (Y^T - (X\Theta)^T) (Y - X\Theta) \\ \iff E &= Y^T Y - 2(X\Theta)^T Y - (X\Theta)^T (X\Theta) \\ \iff E &= Y^T Y - 2\Theta^T X^T Y - (X\Theta)^T (X\Theta) \end{aligned}$$

Note that both $(X\Theta)$ and Y are vectors with same dimension so when we multiply one by another, it doesn't matter what the order is (i.e., $(X\Theta)^T Y = Y^T (X\Theta)$). Then, we derive by each component of the vector, and combine the derivatives into a new vector again, which corresponds to:

$$\begin{aligned} \frac{\partial E}{\partial \Theta} &= 0 - 2X^T Y + 2X^T X\Theta = 0 \\ \iff X^T X\Theta &= X^T Y \end{aligned}$$

And hence, we obtain the normal equation:

$$\Theta = (X^T X)^{-1} X^T Y \quad (2.9)$$

The Normal Equation is an analytical solution to the linear regression problem with a least-squares cost function and is, in some cases, more effective than applying gradient descent. We can use it to directly compute the parameters of a model that minimizes the sum of the squared difference between the actual and predicted terms. This method is quite useful when the dataset is small but may not be able to give us the best parameter of the model for large dataset. The inverse operation involved in solving for the parameters has a runtime complexity of $O(n^3)$, making it slow for large values of n . Using the normal equation can be advantageous for datasets with 10,000 features or less, as it eliminates the need to select a learning rate, reducing the number of hyperparameters to tune. Table 2.3 wraps up all these information.

Gradient Descent	vs	Normal Equation
Need to chose a learning rate α		No need to chose a learning rate
Needs many iterations		Don't need to iterate
Works well even when n is large		Difficult if n is very large (ok for $n=100$ or 1000 but difficult starting $10\ 000$)
Complexity $O(kn^2)$		Complexity $O(n^3)$
Converges iteratively, potentially slower (but suitable for large datasets)	Converges immediately (but computationally expensive for large datasets due to matrix inversion)	

Table 2.3 – Brief comparison between the gradient descent and normal equation methods

Logistic Regression

With what we just presented, we have the foundations to understand the *logistic regression*. As a reminder, while the name contains the word "regression", the logistic regression is a classification algorithm which means that the output will belong to a class. The simplest problem is the binary classification, with $y \in \{0, 1\}$. The notion that we miss is the *activation function*.

The role of the activation function is described in section 2.3.3.5 and its main role is to introduce non-linearity in a network. Here, it is also used to change the output of our continuous linear regression in a binary output (either category A or category B - say 0 and 1 respectively). We want $0 \leq h_{\theta}(x) \leq 1$ and we want to fix a threshold classifier output at 0.5 and say: if $h_{\theta}(x) \geq 0.5$ then $y = 1$ and if $h_{\theta}(x) < 0.5$ then $y = 0$. The "classical" function used in this case is the sigmoid or *logistic* function (see Figure 2.9). It is defined as follows:

$$a(z) = \frac{1}{1 + \exp(-z)} \quad (2.10)$$

From this, we then change the hypothesis function into:

$$h_{\theta}(X) = g(X\Theta) = \frac{1}{1 + \exp(-X\Theta)} \quad (2.11)$$

Now, our prediction is not simply a binary value, but rather a probability indicating the likelihood of belonging to either category A or category B. Taking our example with trees, let's consider

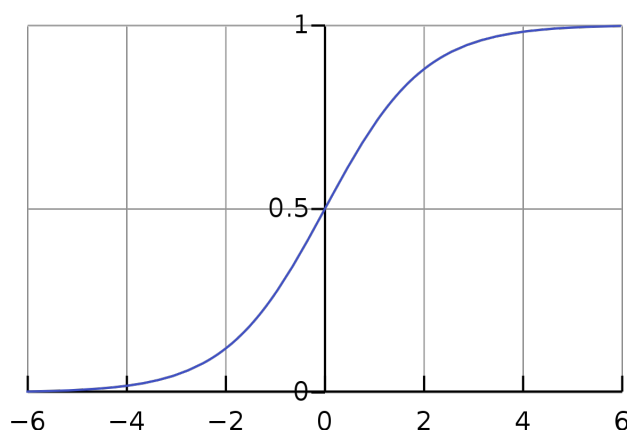


Figure 2.9 – Sigmoid function.

that our input consists of certain characteristics, and our output is either being a deciduous tree (0) or a coniferous tree (1). In this case, our output represents the probability of an observation (a set of features) belonging to one of the two categories.

To convert these probabilities into actual class predictions, we can introduce a threshold function. This function helps determine whether the output should be classified as one of the two classes, rather than a probability value. While we won't delve further into this concept here, it's worth noting that besides the logistic function used in the logistic regression model, various activation functions can be employed, as we will explore in Section 2.3.3.5.

To provide an overview of logistic regression, refer to Figure 2.10 for the schematic diagram illustrating the overall process.

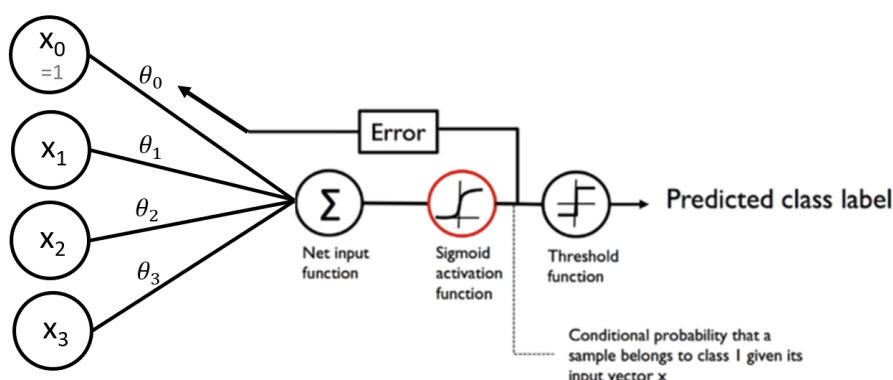


Figure 2.10 – Schematic diagram of the logistic regression classification, adapted from Raschka and Mirjalili (2017).

With the appropriate parameters Θ , we can determine the corresponding category for a given set of features. How do we train our model to learn these parameters? There are multiple methods to train a Logistic Regression model. Essentially, we aim to fit the sigmoid function to our data, although sometimes another activation function may be more suitable. To learn the parameters, we can employ iterative optimization algorithms like gradient descent to minimize a cost function, as we have previously explained. For logistic regression, our goal is to minimize the *cross-entropy loss* (see Section 2.3.3.1). Alternatively, probabilistic methods such as Maximum Likelihood can

be utilized⁴. The training process is the central focus of Section 2.3.3, where we will delve into all the relevant aspects.

However, the information presented thus far may raise a question. When we stack several linear regressions one after the other, taking the output of one linear regression as the input for another, nothing remarkable occurs as it simply results in a linear combination of linear regressions, ultimately yielding one large linear regression. However, with the introduction of activation functions, such as in logistic regression, non-linear combinations arise. So, what happens when we combine multiple logistic regressions together? In such cases, we create a network where each logistic regression can be seen as a neuron — a neural network.

2.3.2 A First Neural Network

As soon as we talk about neural networks, we also talk about *Deep Learning*. Let's delve into what a neural network is mathematically, starting with a simple *two layer neural-network*.

First, let's go back to the example represented in Figure 2.10 and let's introduce some new notations. From the figure we see we have the following multiplication at the first node:

$$z(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3$$

After the sigmoid now , we obtain:

$$g(z) = \frac{1}{1 + e^{-z(x)}} = \frac{1}{1 + e^{-\theta_0 - \theta_1 x_1 - \theta_2 x_2 - \theta_3 x_3}}$$

Now, let's add what we call a *second layer* that is *fully connected* to the first one, meaning that all starting node (our inputs) will be connected to the nodes of the second layer called the *hidden layer*. The diagram will then evolve in Figure 2.11.

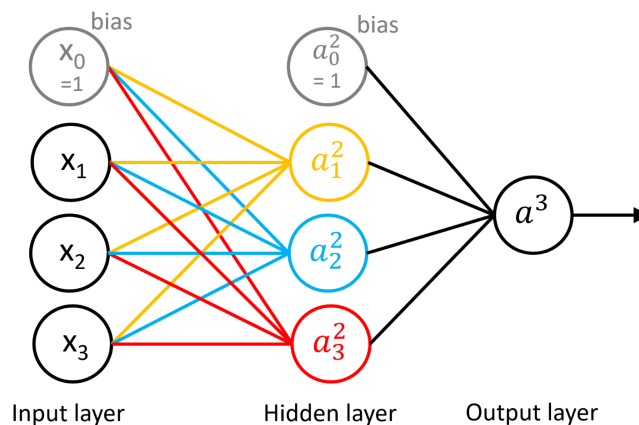


Figure 2.11 – Schematic diagram of a 2-layer neural network.

Here are some information needed to fully understand the diagram Figure 2.11:

4. For those familiar with the field of Machine Learning, it is worth noting that we intentionally omitted discussing certain probabilistic methods, as they are more complex for beginners and were not employed in this thesis

- Our input X can be written as:

$$X = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

- We now have 12 links between the first four inputs and the 3 nodes of the hidden layer. All this links represent one parameter. We previously had 4 parameters (θ_0 to θ_3) and we now have 4 parameters times 3 new nodes. The new matrix Θ is now $\Theta^{(1)}$ to express the idea that they are the parameters to go to from layer 1 to layer 2 (more generally $\Theta^{(j)}$ from layer j to layer $j + 1$). Using the color-code found on Figure 2.11, we can write the matrix⁵:

$$\Theta^{(1)} = \begin{bmatrix} \theta_{\text{orange } 0} & \theta_{\text{orange } 1} & \theta_{\text{orange } 2} & \theta_{\text{orange } 3} \\ \theta_{\text{blue } 0} & \theta_{\text{blue } 1} & \theta_{\text{blue } 2} & \theta_{\text{blue } 3} \\ \theta_{\text{red } 0} & \theta_{\text{red } 1} & \theta_{\text{red } 2} & \theta_{\text{red } 3} \end{bmatrix}$$

- We note $z_i^{(j+1)}$ the different results of the linear combinations and $z^{(j+1)}$ the corresponding matrix to account for the fact that we are now in the next layer, hence:

$$z^{(2)} = \Theta^{(1)}X \quad (2.12)$$

- We note $a_i^{(j)}$ the result of $g(z_i^{(j)})$ and $a^{(j)}$ the corresponding vector such that:

$$a^{(2)} = g\left(z^{(2)}\right) \quad (2.13)$$

Now with a given input X what is the whole process? From X we get $z^{(2)} = \Theta^{(1)}X$, then we obtain $a^{(2)} = g(z^{(2)})$ (our non-linear transformation). Then from $a^{(2)}$, we obtain $z^{(3)} = \Theta^{(2)}a^{(2)}$, and finally $a^{(3)} = g(z^{(3)})$ our final result. This process is called *forward propagation*, and corresponds to how the neural network computes an output from an input.

We just created a simple fully connected neural network. Our next step is to *train* it to identify the best parameters, i.e., the optimal values in $\Theta^{(l)}$ with $l = 1, 2, \dots, L \in \mathbb{N}$ (L the number of hidden layers) in order to approximate the known outputs.

2.3.3 Training Neural Networks

Let's take a moment to revisit a few concepts we have already discussed. In machine learning algorithms, learning occurs through the utilization of a cost or loss function. The primary objective of these algorithms is to identify the optimal parameters that minimize this function. Typically, a cost function is constructed by measuring the difference between the expected output y (remember, we are in a supervised learning scenario) and the predicted output (often denoted as \hat{y}), which represents the error. Beginning from an initial random point in the parameter space, the algorithm navigates its way towards finding a local minimum.

5. Note that these new Θ and X matrices are the transposed version of the previous Θ and X such that the linear combination is now ΘX

As we have seen in the case of neural networks, a given input can yield one or multiple outputs. The loss function is responsible for quantifying the disparity between these outputs and the expected values (i.e., the ground truth). Through the utilization of optimization methods, such as gradient descent, and by performing a process known as *backpropagation*, the network's parameters are adjusted to minimize this error. Through iterative steps, our aim is to observe a reduction in the computed loss function.

2.3.3.1 Loss Function

Determining the most suitable loss function for machine learning algorithms is a task that depends on various factors, such as the specific machine learning algorithm employed, the ease of computing derivatives, and the potential presence of outliers in the dataset. Loss functions can be broadly categorized into two main types, namely Regression losses and Classification losses, which are based on the type of learning task at hand. The loss function represents the gap between what you desire and what you actually achieve, acting as a measure of distance. Ideally, it should satisfy the triangular inequality, and at the minimum, it should always be positive.

Let's begin by introducing a widely used classification loss known as the cross-entropy loss. This loss function is particularly common in classification problems, such as logistic regression. The cross-entropy function, denoted as C , can be expressed as follows:

$$C(y, \hat{y}) = -\frac{1}{m} \sum_{i=1}^m (y^i - 1) \log(1 - \hat{y}^i) - y^i \log(\hat{y}^i) \quad (2.14)$$

with m the amount of training example depending on the parameters hidden in \hat{y}^i . As evident from the formula, the loss function diverges to infinity when the error is 1 and approaches zero when the error tends to 0. A noteworthy characteristic of the cross entropy loss is its significant penalty on predictions that are both confident and incorrect. Similarly, another noteworthy classification loss is the *hinge loss* or *multi-class SVM loss*. While we won't delve into its details here, we encourage the reader to explore this loss function independently. It provides an alternative approach to classification and can offer valuable insights into the realm of machine learning.

In this thesis, our focus is primarily on regression problems. Therefore, the losses that we will present in the following table 2.4 hold significant importance for our analysis and study.

Regression Losses (deterministic)	Formula
Mean-Square Error (MSE)	$\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$
Root Mean-Square Error (RMSE)	\sqrt{MSE}
Normalized Mean-Square Error (NMSE)	$\frac{1}{n} \sum_{i=1}^n \left(\frac{y_i - \hat{y}_i}{\max(y) - \min(y)} \right)^2$
Mean Absolute Error (MAE)	$\frac{1}{n} \sum_{i=1}^n y_i - \hat{y}_i $

Table 2.4 – Example of loss functions for regression in the case of deterministic forecasts

MSE, or Mean Square Error, measures the average squared difference between predicted and actual values. MSE emphasizes larger errors due to its squaring operation, and hence "punishes" more large errors during the training. It is computationally efficient and disregards error direction.

RMSE, the square root of MSE, provides an interpretable measure of the average magnitude of the errors. It is commonly used when the scale of the error needs to be presented in the original

units of the target variable.

NMSE normalizes the MSE by dividing it by the squared range of the target variable. This allows for a standardized comparison of the error across different datasets with varying scales. *NMSE* is particularly useful when comparing performance between different regression models or datasets with different ranges of the target variable.

MAE, or Mean Absolute Error, calculates the average absolute difference between predictions and actual values, making it more robust to outliers. It does not involve squaring and requires more complex gradient computations. The choice between MSE and MAE depends on the specific problem and desired model characteristics.

Now that we have a grasp of how to measure the error between the ground truth and the prediction, our goal is to enable our algorithm to utilize this loss function to update its parameters. This is accomplished through the application of optimization methods, such as gradient descent, along with the implementation of backpropagation. To gain insight into their workings, let's revisit the neural network example discussed in Section 2.3.2 and incorporate the cross-entropy loss that we have just introduced.

2.3.3.2 Backpropagation

From what we have seen, what does it mean for an algorithm to learn? It means updating the parameters or *weights* (denoted θ but often denoted w in the literature) to minimize the loss function, computing the error. The weights are optimized through a chosen method called the *optimizer* such as *batch* or *stochastic gradient descent* as presented in the Section 2.3.4 (Ruder, 2016).

Before delving any further in the backpropagation explanation, let's put some notations we are about to use in table 2.5.

We already presented the gradient descent when presenting the linear regression. It is a well-known method used to modify the weight in order to minimize the loss function by finding the slope of the cost function with respect to each parameters. From there, we modify the weight by going down each slope by a certain amount called the learning rate. The weights are updated as such:

$$\theta := \theta - \alpha \frac{\partial J}{\partial \theta}, \forall \theta \text{ (see linear regression)}$$

Recall that during *forward propagation*, data travels through the network from the input to the output layer. Nodes in each layer receive input from the preceding layer, consisting of a weighted sum of connections multiplied by the previous layer's output. This input is then passed through an activation function to generate the output for that node. This output serves as the input for the nodes in the subsequent layer. This iterative process continues until the data reaches the output layer.

To update the weight, the optimizer computes the derivative (or the gradient) of the loss function with respect to the weights in the model. Intuitively, the *optimizer* understands if values in the output nodes should each increase or decrease compared to what is expected. For instance for a classification algorithm, we will want one output node value to increase and all the other to decrease.

Symbol	Definition
L	Number of layers in the network
m	Number of training samples
n	Number of features
y_j	The value of node j in the output layer L for a single training sample
$x^{(i)}$	Vector of n values, the i^{th} training input
C	Individual cost function for a single training sample
$J(\Theta)$	Cost / Error / Loss function. J also depends on inputs, outputs and hyperparameters.
$\theta_j^{(l)}$	The vector of weights connecting all nodes in layer $l - 1$ to node j in layer l
$\theta_{jk}^{(l)}$	The weight that connects node k in layer $l - 1$ to node j in layer l
$z_j^{(l)}$	The input for node j in layer l
$g^{(l)}$	The activation function used for layer l
$a_j^{(l)}$	The activation output of node j in layer l

Table 2.5 – All notation used to understand the training process of an algorithm. Some notations, such as the

The values of the output nodes are determined by the weighted sum of connections in the output layer, multiplied by the output of the previous layer and passed through the activation function. To update the output node values as discussed earlier, we can modify the weights connected to the output layer or alter the activation output of the previous layer. Although we cannot directly change the activation output since it depends on weights and the previous layer’s output, we can indirectly influence it by updating the weights in a similar manner as we did for the output layer.

This process continues backwards through the network until reaching the input layer. It’s crucial to keep the values of the nodes in the input layer unchanged because they represent the actual input data. As we move backward, we update the weights from right to left to nudge the values of the output nodes in the direction they should be heading to minimize the loss.

Let’s consider the simpler case of a squared difference for the loss function. We apply this difference to the outputs $a_j^{(L)}$ for all j . To avoid excessive indices, we do not specify the number of elements in the output layer, but it corresponds to the number of expected outputs (e.g., 2 for a classification problem of deciduous vs coniferous trees). We sum over all these elements, which is:

$$C = \sum_j \left(a_j^{(L)} - y_j \right)^2 \tag{2.15}$$

This can also be expressed as:

$$C = \sum_j \left(g^{(L)}(z_j^{(L)}) - y_j \right)^2 \tag{2.16}$$

or even:

$$C = \sum_j \left(g^{(L)} \left(\sum_k \theta_{jk}^{(L)} a_k^{(L-1)} \right) - y_j \right)^2 \tag{2.17}$$

We can easily understand the transition from Equation 2.15 to 2.17 by referring to what we discussed in Section 2.3.2 and examining table 2.5.

Now, based on Equation 2.15, we can infer that C depends on $a_j^{(L)}$. Additionally, we know that $a_j^{(L)}$ is the result of applying the activation function to the output of the weighted sum $z_j^{(L)}$, which, in turn, depends on it. Lastly, $z_j^{(L)}$ depends on $\theta_{jk}^{(L)}$. To differentiate the loss function C with respect to *one* specific weight $\theta_{jk}^{(L)}$, which connects node k in layer $L - 1$ to node j in the output layer, we can apply the *chain rule*:

$$\frac{\partial C}{\partial \theta_{jk}^{(L)}} = \left(\frac{\partial C}{\partial a_j^{(L)}} \right) \left(\frac{\partial a_j^{(L)}}{\partial z_j^{(L)}} \right) \left(\frac{\partial z_j^{(L)}}{\partial \theta_{jk}^{(L)}} \right) \quad (2.18)$$

Let's break down each of the three terms in this equation:

- For the first term, we know that the loss function has the form seen in Equation 2.15 and we want to derive it with respect to only one element in $a^{(L)}$, let's say $a_1^{(L)}$ for clarity.

$$\frac{\partial C}{\partial a_1^{(L)}} = \frac{\partial}{\partial a_1^{(L)}} \left(\sum_j \left(a_j^{(L)} - y_j \right)^2 \right)$$

If we expand the sum, we will have terms that do not depend on $a_1^{(L)}$ except for $(a_1^{(L)} - y_1)^2$. Thus, the result is:

$$\frac{\partial C}{\partial a_1^{(L)}} = 2 \left(a_1^{(L)} - y_1 \right) \quad (2.19)$$

- For the second term, since $a_j^{(L)} = g^{(L)}(z_j^{(L)})$ for each node j in the output layer L , we have a straightforward expression:

$$\frac{\partial a_j^{(L)}}{\partial z_j^{(L)}} = g'^{(L)} \left(z_j^{(L)} \right) \quad (2.20)$$

- As we have previously seen, $z_j^{(L)} = \sum_k \theta_{jk}^{(L)} a_k^{(L-1)}$. To compute the derivative of $z_j^{(L)}$ with respect to a specific index $\theta_{jk}^{(L)}$, let's consider the example of $\theta_{12}^{(L)}$ (the weight connecting node 2 in layer $L - 1$ to node 1 in the output layer L). In the derivative calculation:

$$\frac{\partial z_1^{(L)}}{\partial \theta_{12}^{(L)}} = \frac{\partial}{\partial \theta_{12}^{(L)}} \left(\sum_k \theta_{1k}^{(L)} a_k^{(L-1)} \right)$$

we observe that the only term in this sum that depends on $\theta_{12}^{(L)}$ is when $k = 2$. The derivative for the other terms will be zero. Therefore, we obtain:

$$\frac{\partial z_1^{(L)}}{\partial \theta_{12}^{(L)}} = \frac{\partial}{\partial \theta_{12}^{(L)}} \left(\theta_{12}^{(L)} a_2^{(L-1)} \right) = a_2^{(L-1)} \quad (2.21)$$

Taking altogether Equations 2.19, 2.20, 2.21, and for a given weight $\theta_{jk}^{(L)}$ we obtain the following:

$$\frac{\partial C}{\partial \theta_{jk}^{(L)}} = 2 \left(a_j^{(L)} - y_j \right) g'^{(L)} \left(z_j^{(L)} \right) a_k^{(L-1)} \quad (2.22)$$

Please note that, up until now, we have only examined the derivative of the loss function for the final parameters just before the output. How can we generalize this process, and where does the concept of "back" propagation originate? Let's start by considering the derivative of the loss function with respect to a single weight that connects layer $L - 2$ and $L - 1$, for instance $\theta_{42}^{(L-1)}$. If we visualize it, this parameter corresponds to the connection between node 2 in layer $L - 1$ and node 4 in layer $L - 1$. The calculation is as follows:

$$\frac{\partial C}{\partial \theta_{42}^{(L-1)}} = \left(\frac{\partial C}{\partial a_4^{(L-1)}} \right) \left(\frac{\partial a_4^{(L-1)}}{\partial z_4^{(L-1)}} \right) \left(\frac{\partial z_4^{(L-1)}}{\partial \theta_{42}^{(L-1)}} \right) \quad (2.23)$$

- Here, the second and third terms on the right-hand side will be computed using the same approach as before and we will obtain:

$$\frac{\partial C}{\partial \theta_{42}^{(L-1)}} = \left(\frac{\partial C}{\partial a_4^{(L-1)}} \right) g'^{(L-1)}(z_4^{(L-1)}) a_2^{(L-2)} \quad (2.24)$$

- The first term requires our attention. The loss function C is what we saw in the three Equations 2.15, 2.16 and 2.17. Equation 2.17 clearly shows how the loss function is a composition of functions. When we looked at the derivative with respect to a parameter linking to the output layer we just had to look at the corresponding output node. Now, it is different. When we look at one parameter linking layers $L - 2$ and $L - 1$, we look at the output of a node in the layer $L - 1$ (namely here $a_4^{(L-1)}$) and this output will then go (by multiplication with all parameters linking $L - 1$ and L) in all the final outputs of layer L . Hence, for all j , all $a_j^{(L)}$, and hence all $z_j^{(L)}$ depends on our $a_4^{(L-1)}$. This will add a sum over all j 's in our computation. So, we use again the chain rule and the derivative of the loss with respect to the activation output for node 4 in layer $L - 1$ is:

$$\frac{\partial C}{\partial a_4^{(L-1)}} = \sum_j \left[\left(\frac{\partial C}{\partial a_j^{(L)}} \right) \left(\frac{\partial a_j^{(L)}}{\partial z_j^{(L)}} \right) \left(\frac{\partial z_j^{(L)}}{\partial a_4^{(L-1)}} \right) \right] \quad (2.25)$$

So finally we have:

$$\frac{\partial C}{\partial \theta_{42}^{(L-1)}} = \sum_j \left[\left(\frac{\partial C}{\partial a_j^{(L)}} \right) \left(\frac{\partial a_j^{(L)}}{\partial z_j^{(L)}} \right) \left(\frac{\partial z_j^{(L)}}{\partial a_4^{(L-1)}} \right) \right] g'^{(L-1)}(z_4^{(L-1)}) a_2^{(L-2)} \quad (2.26)$$

What does the very first term in the sum on the right side of the Equation 2.26 mean? To compute the gradient of the cost function with respect to any weight, such as $\theta_{42}^{(L-1)}$ in this case, we then need the gradient of the cost function with respect to all of the activation outputs between the cost computation *at the end* and the weight of interest. By starting at the end and going back layer by layer, we recursively have all these gradients. This process is known as *backpropagation*. We start by computing the initial gradients of the cost function for all the weights in the last layer. Then, we move one layer back and utilize these gradients to compute the new gradients, and so on. This backpropagation journey allows us to determine the gradient of the cost function with respect to all the weights in the network.

Before summarizing the information, it is important to address a well-known issue that can arise during algorithm training: the *vanishing gradient*. As demonstrated earlier, when computing gradients during the backpropagation process, the derivatives of the cost function with respect to the parameters in each layer are recursively obtained using the chain rule. In layers preceding the current one, the gradient is multiplied by the derivative of the activation function (represented by the term g' in Equation 2.26). Activation functions like the sigmoid or hyperbolic tangent

function have derivatives that are less than 1 in magnitude (e.g., the sigmoid derivative is given by $g'(x) = g(x)(1 - g(x))$ and has a maximum around 0.25). This means that small derivatives are repeatedly multiplied together during backpropagation, resulting in exponentially diminishing gradient values as they propagate backward. Consequently, the vanishing gradient problem occurs when the gradient becomes extremely small, impeding the weight values from changing effectively. In severe cases, this can halt the training of the neural network altogether (Basodi et al., 2020). A major factor contributing to the vanishing gradient problem is then an excessively large number of layers in a network. To address this issue, various techniques, including careful weight initialization, activation function choices, and architectural modifications, have been developed.

Now, let's summarize the information and outline the process of algorithm training.

1. Randomly initialize weights $\theta_{jk}^{(l)}$ for all j, k and l .
2. Implement the forward propagation to get $h_{\Theta}(x^i)$ for all i from 0 to m , where $x^{(i)}$ is here a vector of features corresponding to one training sample over m available. So we compute all outputs for all inputs. $h_{\Theta}(x) \in \mathbb{R}^K$ if we have K outputs, or classes for instance.
3. Compute the cost function $J(\Theta)$. In the example of logistic regression we would get:

$$J(\Theta) = -\frac{1}{m} \sum_{i=1}^m \sum_{k=1}^K y_k^i \log((h_{\Theta}(x^i))_k) + (1 - y_k^i) \log((h_{\Theta}(x^i))_k)$$

where the k indices here refers to output number k among K outputs.

4. Implement the backpropagation to compute all partial derivatives:
 - Perform the forward propagation for one sample, hence obtaining $a^{(l)}$ for all l .
 - Perform the backpropagation to obtain the gradient of the cost function with respect to any weight.
 - Iterate these two steps for all training samples and obtain not only the gradient of the cost function but the average gradient of the cost function with respect to any weight over all samples
5. Update all the weight using an optimizer (several optimizers will be presented in the Section 2.3.4) such as the gradient descent.
6. Iterate steps 2 to 5 and observe how the overall cost function (the average of all cost functions) behaves

Now that we have a general understanding of how an algorithm can train a model, there are two important aspects that we need to delve into to gain a complete understanding of the training process: *optimizers* and *schedulers*. These two methods play crucial roles in enhancing the training process and optimizing model performance. Finally, we will end by giving a brief overview of activation functions and some examples.

2.3.3.3 Optimizers

The "optimizer" is a central aspect of deep learning algorithms. It constitutes the "search" technique to find the optimization values of parameters to minimize the loss function. These values "cannot be reasonably obtained by using a deterministic optimization technique" (Okewu et al., 2019). This highlights the need for an iterative approach that randomly selects data segments and assigns initial values to optimization parameters. Through this process, error functions are continuously calculated until an acceptable level of error is achieved, as deterministic optimization techniques alone cannot feasibly obtain these values. Previously, we discussed the gradient descent algorithm as the typical stochastic optimization approach for training deep neural networks. It treats the training process as a non-convex optimization problem. Now, we will delve into

several (but not all) extensions of gradient descent, including Stochastic Gradient Descent, Momentum, Adagrad, RMSProp, or ADAM. These variations have been developed to further improve accuracy, convergence rate, and training time (Okewu et al., 2019). To date, there is no theory that adequately explains how to choose among all the existing optimizers (Choi et al., 2019). Instead, the community relies on empirical studies (Wilson et al., 2017) and benchmarking (Schneider et al., 2019).

- The traditional gradient descent method we have previously discussed is not ideal when dealing with large datasets, as it requires computing the gradient for every individual sample. SGD addresses this issue by computing and updating the gradient based on a single randomly selected training example in each iteration. It randomly selects a training example, calculates the gradients of the loss function with respect to the weights, and updates the weights using the gradient descent rule. As only a single data point is used instead of the entire dataset, the loss curve will appear more erratic, and SGD may require more iterations to converge to a minimum. However, it still offers lower overall computational cost.
- Momentum: The momentum optimization algorithm builds upon SGD by introducing a momentum term. It accumulates a weighted average of the past gradients and uses it to update the parameters. This helps to accelerate convergence, especially in scenarios with high curvature or noisy gradients.
- AdaGrad (Adaptive Gradient Descent, Duchi et al. (2011)): Adagrad is a specific gradient descent method as it adapts the learning rate for each parameter individually. It maintains a separate learning rate for each parameter based on the historical gradient information. The formula is as follows:

$$\theta_{i,t} = \theta_{i,t-1} - \alpha_{i,t} \frac{\partial J}{\partial \theta_{i,t-1}}$$
$$\alpha_{i,t} = \frac{\alpha}{\sqrt{G_{i,t} + \epsilon}}$$

With

- The learning rate for the parameter θ_i at iteration t is $\alpha_{i,t}$
 - ϵ a small value added for numerical stability to prevent division by zero
 - ϵ a small value added for numerical stability to prevent division by zero
 - $G_{i,t}$ is the accumulated sum of squared gradients for parameter θ_i up to iteration t .
 - α is the initial learning rate (a hyperparameter set by the user).
- RMSProp (Tieleman and Hinton, 2012): RMSProp modifies Adagrad by introducing an additional parameter to control the accumulation of historical gradients. By using an exponentially decaying average of squared gradients, RMSProp mitigates the diminishing learning rate issue and improves convergence, it reduces the monotonically decreasing learning rate in Adagrad. RMSprop penalizes the parameter causing excessive oscillation in the cost function. For example, if a fish classification model heavily relies on "color" and makes many errors, RMSprop discourages over-reliance on "color" and encourages consideration of other features. It converges faster and requires less tuning than traditional gradient descent algorithms (see Gupta (2023) article).
 - ADAM (Adaptive Moment estimation, Kingma and Ba (2014)): It combines the benefits of both momentum-based methods and adaptive learning rate methods. ADAM maintains an exponentially decaying average of past gradients and their squared values, which are used to adaptively update the learning rates for each parameter. In other words, it uses first

moment (mean, m) and second moment (v) of the gradients to update the parameters:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) \frac{\partial J}{\partial \theta_{i,t}}$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) \left(\frac{\partial J}{\partial \theta_{i,t}} \right)^2$$

with then a bias correction that we don't present here that gives \hat{m}_t, \hat{v}_t . β_1 is a hyperparameter that controls the decay rate of the first moment estimate (typically 0.9) and β_2 a hyperparameter that controls the decay rate of the second moment estimate (typically 0.999). Then the update is as follows:

$$\theta_{i,t} = \theta_{i,t-1} - \alpha \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t + \epsilon}}$$

Where: θ represents the parameters to be optimized, α is the learning rate, \hat{m}_t is the bias-corrected first moment estimate, \hat{v}_t is the bias-corrected second moment estimate, and ϵ is a small constant for numerical stability.

Although the Adam optimizer combines the strengths of various algorithms and is highly regarded, it may not always be the optimal choice for every application. Algorithms like stochastic gradient descent prioritize individual data points and can provide better generalization, albeit with slower computation speed. The selection of an optimization algorithm should consider specific requirements and the characteristics of the data being analyzed. While optimizers adapt the learning rate to each parameter, the learning rate still remains an input to the optimizer. This is where *schedulers* come into play, dynamically updating the learning rate during training. Optimizers handle the actual parameter updates using algorithms like Adam, while schedulers focus on optimizing the learning process by adapting the learning rate. Together, these components work to improve training and model performance.

2.3.3.4 Schedulers

Schedulers are an ensemble of functions that dynamically update the learning rate during the training (as a function of the epochs or steps). There are several commonly used learning rate schedulers in deep learning. Some of the main ones include StepLR, ReduceLROnPlateau, CosineAnnealingLR, OneCycleLR, ExponentialLR. In the following description, the variable *epoch* corresponds to the current epoch at which we update the learning rate. All schedulers curve (except for ReduceLROnPlateau) can be seen in Figure 2.12.

- StepLR: The learning rate is multiplied by a factor $\gamma < 1$ every K epochs.

$$\alpha = \alpha_0 \times \gamma^{\lfloor \frac{\text{epoch}}{K} \rfloor}$$

- ReduceLROnPlateau: The learning rate is reduced when a monitored metric has stopped improving for a certain number of epochs that we call the *patience*. For instance we can take the validation loss as this metric, and the patience to be 3, this would mean that if for three consecutive epochs the validation loss does not improve, the learning rate is then multiplied by a factor (< 1).

if metric < threshold: $\alpha := \alpha \times \text{factor}$

- CosineAnnealingLR: The learning rate follows a cosine annealing schedule that gradually decreases from the maximum value to the minimum value over T_{max} epochs (see Figure 2.12)

$$\alpha = \alpha_{min} + (\alpha_{max} - \alpha_{min}) \times \frac{1}{2} \left(1 + \cos \left(\frac{\text{epoch}}{T_{max}} \times \pi \right) \right)$$

- OneCycleLR (Smith and Topin, 2019): The learning rate follows a cyclical schedule where the user has to specify several parameters such as `max_lr`, `div_factor` and `final_div_factor`, among others (see PyTorch documentation on this function⁶). The learning rate that starts at $\frac{\text{max_lr}}{\text{div_factor}}$ and enters a warm-up phase where it increases to reach a maximum value (`max_lr`), and then decays for the remaining epochs to finally reach $\frac{\text{max_lr}}{\text{final_div_factor}}$.
- ExponentialLR: The learning rate decreases exponentially by a factor of γ every K epochs (see Figure 2.12).

$$\alpha := \alpha \times \gamma^{\left(\frac{\text{epoch}}{K}\right)}$$

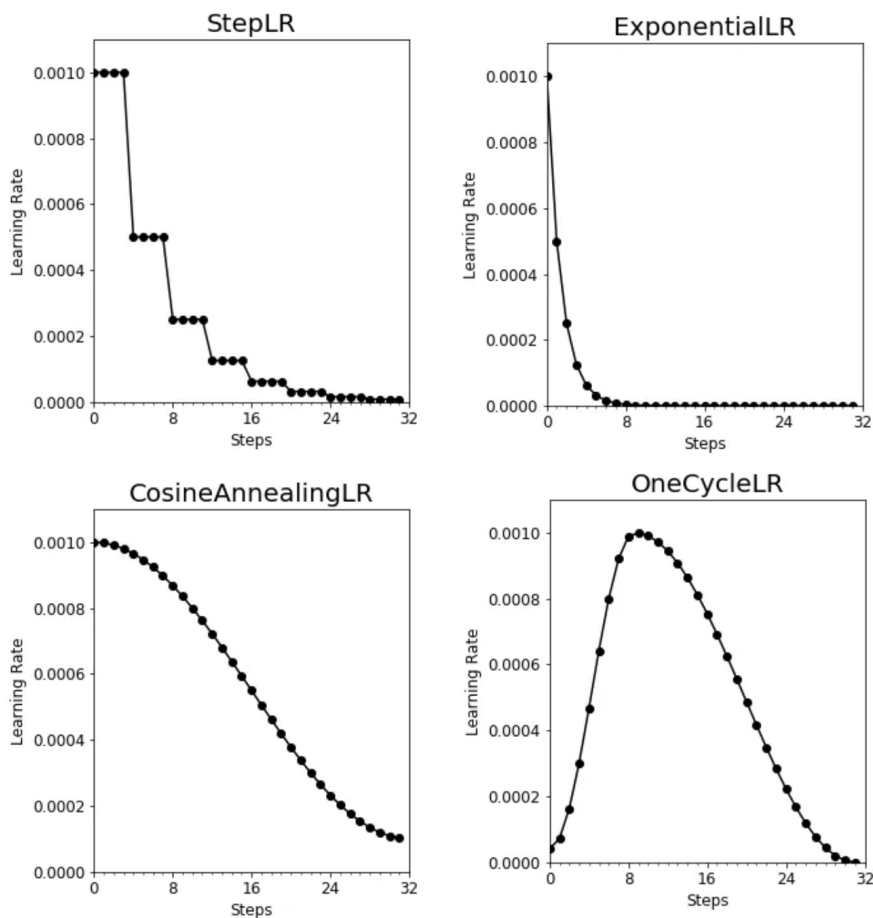


Figure 2.12 – Lineplots of learning rate evolution as a function of the epoch, from Monigatti (2022). Of course, the shape of the curve depends on the parameters specified by the user.

Several reasons justify the use of schedulers. The main reason usually invoked is the speed of convergence. A learning rate scheduler can help accelerate the training process by converging faster to a local minima. A high learning rate means quick movement in the loss space. It is classic to start with high learning rates and then reduce their values to slow down before reaching the minima. Indeed, one of the risk is to "miss" the minima and start increasing again if the steps we are taken in our space are too big (when we are close to a minimum, which can be visualized as a hole, taking a too large step can cause us to end up on the other side, potentially higher than our initial position).

6. https://pytorch.org/docs/stable/generated/torch.optim.lr_scheduler.OneCycleLR.html

In complex optimization landscapes, a fixed learning rate can lead to oscillations or getting stuck in suboptimal solutions and it can then be essential to use a scheduler to help the model navigate and avoid local minima to look for a global minima.

Overall, learning rate schedulers can lead to faster convergence, improve generalization, and increase robustness. They contribute to the optimization process by providing better control over learning dynamics in a loss landscape that can be very complex. We have explored some popular schedulers, such as StepLR, ReduceLRonPlateau, CosineAnnealingLR, OneCycleLR, and ExponentialLR, each offering unique strategies for adapting the learning rate. Now, let's shift our focus to another fundamental component of deep learning models: activation functions.

2.3.3.5 Activation Functions

To conclude our exploration of various aspects of training neural networks, let's delve into the topic of activation functions. We have previously mentioned the sigmoid function and briefly discussed its role in introducing non-linearity to models, enabling them to capture complex representations.

By employing a diverse range of activation functions, we can unleash the full capacity of the network to model non-linear and intricate phenomena. Each activation function possesses unique properties and characteristics that profoundly influence the flow of information within the network. Some functions, such as the sigmoid or hyperbolic tangent, compress the input into a specific range, while others, like the rectified linear unit (ReLU), allow positive values to pass through unchanged.

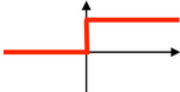
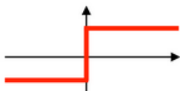
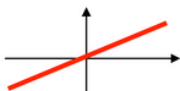
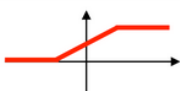
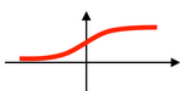
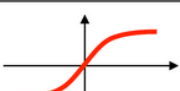

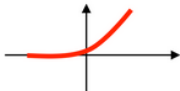
Furthermore, different activation functions exhibit distinct computational properties that can impact training dynamics, convergence speed, and the network's ability to handle gradients that vanish or explode. Therefore, the selection of an appropriate activation function relies on the specific task, dataset, and network architecture.

Here is a non-exhaustive list of several activation functions and their properties. For further information, the reader can find several surveys on the topic such as [Dubey et al. \(2022\)](#).

- The **sigmoid function** is often preferred for binary classification problems as it can simulate Heaviside functions. However, it may struggle to effectively separate closely spaced data points or handle overlapping classes in a narrow range. Normalizing the input is crucial for optimal performance.
- **Linear activation functions** maintain the linearity of the input and are generally not recommended to be stacked consecutively. Multiple linear layers can be mathematically factored into a single layer, resulting in redundant computations.
- The **hyperbolic tangent function**, similar to the sigmoid function, provides a symmetric range of output values from -1 to 1. However, it is less commonly used compared to other activation functions.
- **ReLU (Rectified Linear Unit)** is widely adopted due to its simplicity and efficiency (see Figure 2.13). It divides the parameter space into segments, activating when inputs are positive and putting to zero when inputs are negative. Dead neurons can be a concern during training if they consistently produce zero output, rendering a portion of the network unused.
- **LeakyReLU** addresses the issue of dead neurons in ReLU by introducing a small slope for negative input values. This non-zero slope ensures that neurons with negative inputs contribute some information to the overall computation.

- **SiLU (Sigmoid Linear Unit)**, also known as Swish, combines properties of sigmoid and linear activation functions. It provides a slightly smoothed slope, preventing gradient instability at zero and offering improved training dynamics compared to the traditional sigmoid function.
- **Softmax activation function** is commonly employed in multiclass classification tasks. It transforms the outputs into probabilities, amplifying larger values and suppressing smaller ones, ensuring that the output values sum up to 1 and represent class probabilities.

Each activation function has its own advantages and disadvantages, and the choice should align with the specific problem that the user is facing. They can effectively contribute to the performance of the model.

Activation function	Equation	Example	1D Graph
Unit step (Heaviside)	$\phi(z) = \begin{cases} 0, & z < 0, \\ 0.5, & z = 0, \\ 1, & z > 0, \end{cases}$	Perceptron variant	
Sign (Signum)	$\phi(z) = \begin{cases} -1, & z < 0, \\ 0, & z = 0, \\ 1, & z > 0, \end{cases}$	Perceptron variant	
Linear	$\phi(z) = z$	Adaline, linear regression	
Piece-wise linear	$\phi(z) = \begin{cases} 1, & z \geq \frac{1}{2}, \\ z + \frac{1}{2}, & -\frac{1}{2} < z < \frac{1}{2}, \\ 0, & z \leq -\frac{1}{2}, \end{cases}$	Support vector machine	
Logistic (sigmoid)	$\phi(z) = \frac{1}{1 + e^{-z}}$	Logistic regression, Multi-layer NN	
Hyperbolic tangent	$\phi(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$	Multi-layer Neural Networks	
Rectifier, ReLU (Rectified Linear Unit)	$\phi(z) = \max(0, z)$	Multi-layer Neural Networks	
Rectifier, softplus	$\phi(z) = \ln(1 + e^z)$	Multi-layer Neural Networks	

Copyright © Sebastian Raschka 2016
(<http://sebastianraschka.com>)

Figure 2.13 – Non-exhaustive list of activation functions for artificial neural networks. Credits: Sebastian Raschka, 2016 <http://sebastianraschka.com>.

2.3.4 Evaluation & Diagnostic of Neural Networks

We have discussed the training process of an algorithm, from the use of loss functions to the useful optimizers, schedulers or activation functions, but how can we assess the quality of a given training? How can we determine if it is performing well? As we have observed the existence of

potential issues such as the vanishing gradient, it becomes crucial to monitor and diagnose a neural network in order to enhance its training. Therefore, the question arises: how can we effectively evaluate the performance of a neural network and identify areas for improvement?

2.3.4.1 Metrics

To determine the success of a training process, it is crucial to have a measurable outcome to evaluate. This outcome, referred to as a *metric*, serves as an objective criterion that enables users or data scientists to assess the performance and effectiveness of their algorithm. While certain loss functions can also be used as metrics since they quantify errors, it's important to note that loss functions and metrics serve distinct purposes in the field of machine learning. The loss function is employed by the algorithm during training to optimize its performance, whereas the metric is utilized solely for evaluating and grading the algorithm's performance. Metrics facilitate model comparisons, progress tracking, identification of areas for improvement, and informed decision-making regarding model selection or optimization strategies. The table 2.6 showcases various metrics, including accuracy, confusion matrix, F1 score, and the already-familiar MAE, MSE, and RMSE. Precision and Recall are also listed and presented below.

Metric	Formula	Description
Classification Accuracy	$\frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$	Measures the proportion of correctly classified instances in a classification task.
Precision	$\frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$	Precision measures the accuracy of positive predictions. It is the ratio of correctly predicted positive observations to the total predicted positives.
Recall	$\frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$	Recall, also known as sensitivity, measures the ability of a classifier to find all positive instances. It is the ratio of correctly predicted positive observations to all actual positives.
Confusion Matrix	N/A	A table that summarizes the performance of a binary classification model by displaying the counts of true positive, true negative, false positive, and false negative predictions.
F1 Score	$\frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$	Combines precision and recall into a single metric, providing a balanced measure of a classifier's performance.
Mean Absolute Error (MAE)	$\frac{1}{N} \sum_{i=1}^N y_i - \hat{y}_i $	Measures the average absolute difference between predicted and true values, commonly used in regression tasks.
Mean Squared Error (MSE)	$\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$	Calculates the average squared difference between predicted and true values, commonly used in regression tasks.
Root Mean Squared Error (RMSE)	$\sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2}$	Provides the square root of the mean squared difference between predicted and true values, offering a more interpretable measure than MSE.

Table 2.6 – Example of metrics

We can provide more detailed explanations for the confusion matrix and the F1 score. The confusion matrix is a table that summarizes the performance of a binary classification problem, showing the counts of true positives, true negatives, false positives, and false negatives. An example of

a confusion matrix is shown in Figure 2.14. From the values in the confusion matrix, we can calculate the following metrics:

- The sensitivity or true positive rate

$$\text{TruePositiveRate} = \frac{\text{TruePositive}}{\text{FalseNegative} + \text{TruePositive}}$$

- The specificity or true negative rate

$$\text{TruePositiveRate} = \frac{\text{TrueNegative}}{\text{TrueNegative} + \text{FalsePositive}}$$

- The precision

$$\text{Precision} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalsePositive}}$$

- The recall

$$\text{Recall} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}}$$

When a classifier has high precision but lower recall, it means that it is very accurate in classifying instances, but it may miss a large number of instances that are difficult to classify. The F1 score addresses this by computing the harmonic mean between precision and recall, resulting in a value between 0 and 1. A higher F1 score indicates better performance of the model. It provides an evaluation of both the precision (how many instances are classified correctly) and the robustness (how many instances are not missed). For more information about metric selection, the reader can refer to [Molnar \(2023\)](#).

$$F1 = 2 \frac{1}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}}$$

N = 1000	Predicted: <i>True</i>	Predicted: <i>False</i>
Expected: <i>True</i>	350 (TP)	180 (FN)
Expected: <i>False</i>	58 (FP)	412 (TN)

Figure 2.14 – Example of confusion matrix for 1000 training samples in a binary classification. Note: TP = True Positives; TN = True Negatives; FN = False Negatives; FP = False Positives.

Now, we have a way to assess the performance of a model, but an important question arises. After training our model extensively on a large dataset and achieving improved metrics, we may encounter a situation where the model fails to perform well on a new dataset with similar information. What could be the reason behind this discrepancy? The issue arises when the model, due to intensive training, is biased towards our dataset and gives a false impression of understanding the underlying patterns and relationships between inputs and outputs. This phenomenon is known as *overfitting*. To avoid this problem, we need to implement techniques that promote generalization and prevent the model from over-relying on the training data. The first essential technique to address this issue is to split our dataset into three distinct subsets: the training set, the validation set, and the test set. Up until now, we have been referring to the entire dataset as the training set. However, in order to properly evaluate and optimize our model, we need to allocate a portion of the data specifically for validation and testing purposes.

2.3.4.2 Training, Validation and Test Sets

Splitting our data into training, validation, and test sets is actually considered part (and actually one of the last steps) of the data preprocessing, developed in Section 2.3.6. It allows the user to observe phenomena such as the overfitting (section 2.3.4.3) and improve the training. So what are the purposes behind each of these sets:

- The *training set* comprises the data that will be used to train the algorithm. It forms the basis for the model to learn and adjust its parameters and undergoes the whole process that we have seen already.
- The *validation set* is used to evaluate the performances and make the necessary adjustments. It consists of the same loss computation, but the weights are not updated according to it and it never goes through the backpropagation process. It allows the user to assess the accuracy, speed, and effectiveness of the model on a dataset never seen by the algorithm. This enables the user to take informed decisions and to fine-tune the algorithm (architecture, hyperparameters, optimizers, etc.). The performance of the algorithm on the validation set, which is not used in the cost function minimization, serves as a gauge of quality, indicating whether the model has learned generic features that are not specific to the training set.
- The *test set* is a dataset that remains untouched until the final stages (Bouriat et al., 2022). It serves as a benchmark to assess the model's accuracy on new, unseen data and then gives an unbiased final model performance metric in terms of accuracy, precision, etc. The result of applying our model to the test set is the answer to the question "how well does the algorithm model/predict?"

There is no universally optimal split size for training, validation, and test sets, but a common practice is to allocate 60% for training, 20% for validation, and 20% for testing. The chosen allocation aims to strike a balance by utilizing a significant portion of data for training the model while still ensuring sufficient data in the validation and test sets for effective model evaluation. This brings the idea of *dataset equilibrium*. During dataset analysis and preprocessing (Section 2.3.6), it is crucial to maintain well-balanced datasets, preserving all relevant information such as outliers, distributions, means, term frequencies, classes, or backgrounds after the split (Bouriat et al., 2022).

A technique that can be introduced here is the cross-validation technique. Cross-validation partitions the available dataset into multiple subsets, typically referred to as "folds." The model is then trained on a subset of the data and evaluated on the remaining data. This process is repeated multiple times, with each fold serving as the validation set exactly once. The results from each iteration are then averaged to provide a more robust estimate of the model's performance, helping to mitigate the potential bias introduced by a single train-test split. Cross-validation can assist in hyperparameter tuning and model selection.

2.3.4.3 Overfitting & Loss Curves

As we said, overfitting appears when the algorithm is not learning generic features and not understanding how inputs and outputs are linked. Instead, it starts memorizing the training dataset. In Figure 2.15, we can see two examples of overfitting, the first one in a support vector machine classification and the second one in a regression model. Figure 2.15 also showcases underfitting although it is rare to end up in this case. Overfitting is one of the main issues that we are trying to avoid when training machine learning algorithms and is often due to either a too complex model that memorizes subtle patterns only happening in the training set, or when the training set is too small or contains too many irrelevant data points.

As we said, splitting the data allows the user to actually see the overfitting. How so? By plotting the curves of the changing loss functions during the training. If you are a bit familiar with

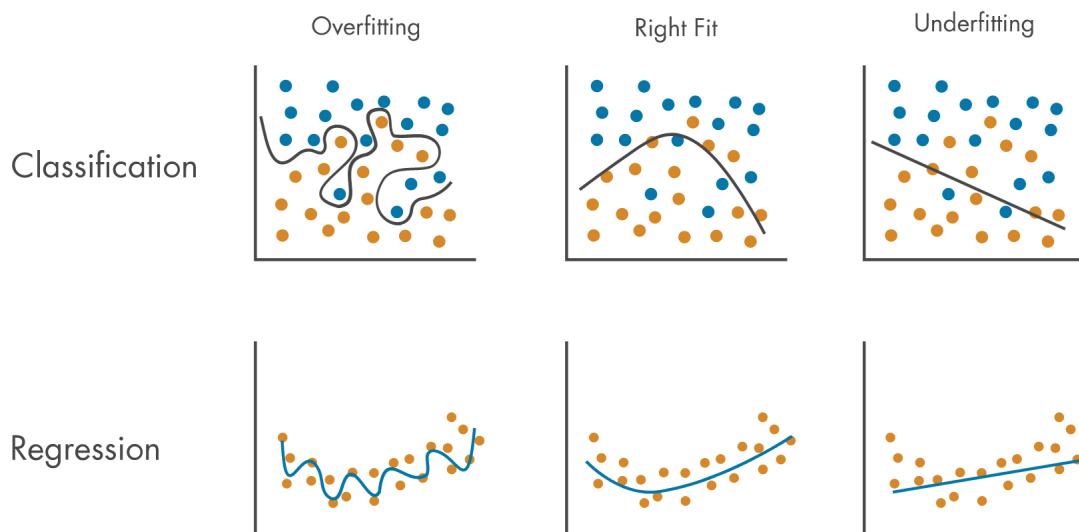


Figure 2.15 – Example of overfitting and underfitting fits on simple cases. Credits: MathWorks courses <https://www.mathworks.com/discovery/overfitting.html>, last accessed in June 2023.

machine learning you might have seen these lineplots, essential to diagnose one’s algorithm performances. Usually, we plot both the *training loss* (i.e., the loss function computed on the training set) and the *validation loss* (i.e., the loss function computed on the validation set) on the y-axis and the *epochs* or *steps* on the x-axis. To understand the difference between steps and epochs, we need to introduce the notion of batch.

A *batch* is a subset of the training set that is processed together during each iteration of the training algorithm. This means that instead of updating the model’s weights after each training sample, we update the weights once b training samples went through the forward propagation. b is called the *batch size*. In this context, one *step* corresponds to one update of the weights after b training samples passed in the network. After 100 steps, we updated the weights 100 times. One *epoch* represents a complete iteration over the entire training dataset. hence for a training set of size N , we have N/b steps to obtain one epoch. The total amount of epochs, as well as the batch size, are examples of hyperparameters that the user has to choose before training. The batch size allows for a trade-off between computational efficiency and convergence speed. A larger batch size (e.g., using the entire training set as a single batch) provides a more accurate estimate of the gradients but requires more memory and computational resources. On the other hand, a smaller batch size (e.g., a mini-batch) introduces more noise in the gradient estimation but can converge faster and allows for parallel processing. As an example, in the loop presented at the end of Section 2.3.3.2 the batch size would be m (but as we did not consider batch yet, one epoch corresponds to one step and the batch size is the size of the dataset).

Now equipped with the knowledge we have gained, we are prepared to comprehend and, more importantly, interpret the following plots Figure 2.16 illustrating training and validation losses. From there, we will introduce several methods and tools such as the early stopping criterion, the schedulers, dropout, or the L1 and L2 regularizations.

Let’s briefly interpret the curves shown in Figure 2.16:

- Panel (a): The curves in panel (a) can be considered as underfitting. The curves are still descending, indicating that the training was stopped prematurely. We are facing underfit-

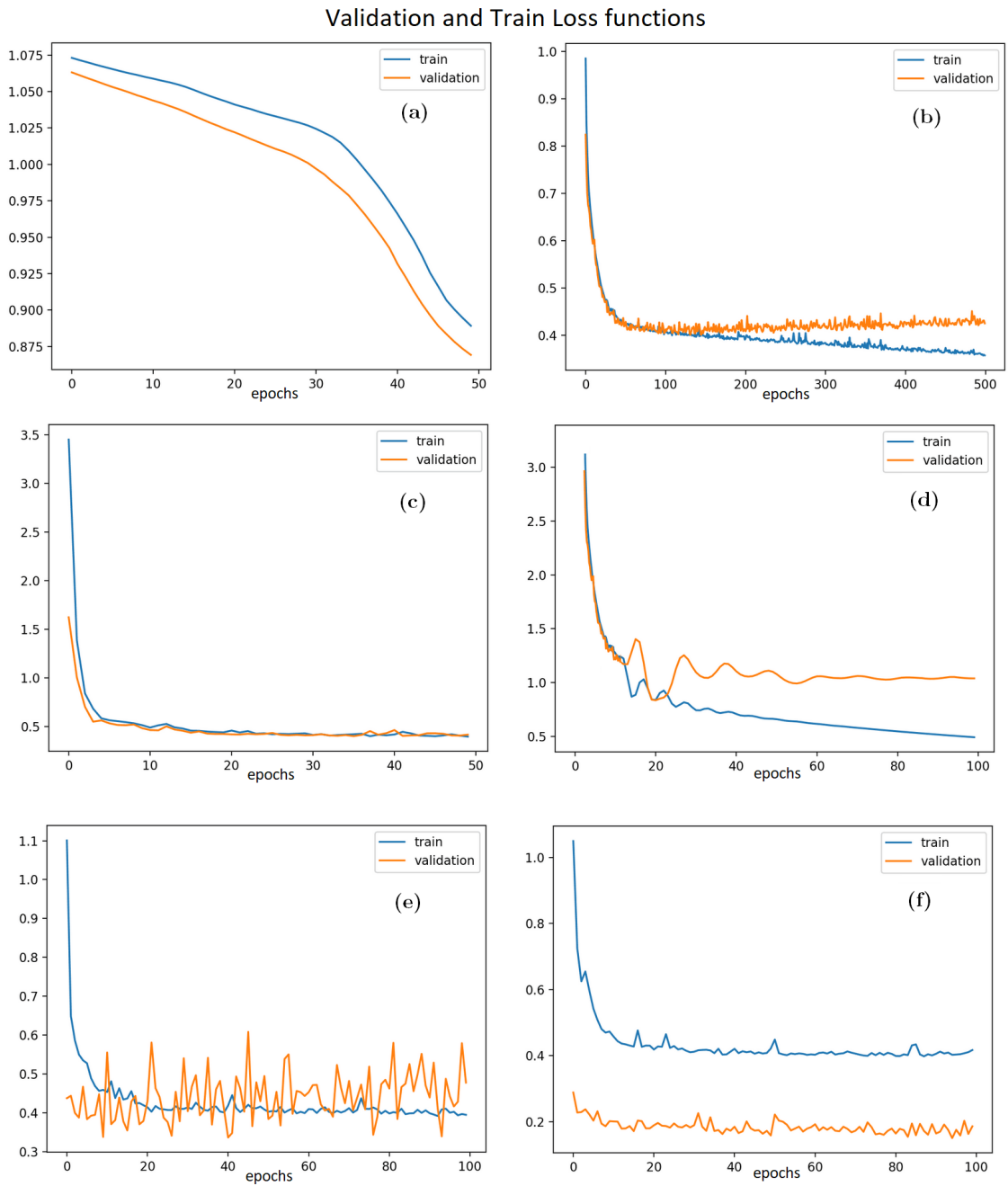


Figure 2.16 – Example of validation and training loss curves from *Brownlee (2019)*.

ting, not due to model limitations, but rather because the model requires further training. Typically, underfitting is identified solely by observing the training loss, which may exhibit a noisy or flat line with relatively high values. This suggests that the model failed to learn from the training dataset.

- Panel (b): Panel (b) illustrates a classic example of overfitting. Initially, both the training and validation losses decrease, indicating that the model is successfully learning meaningful patterns. However, at some point, the validation loss starts to rise while the training loss continues to decrease. This implies that the model performs well on the training dataset but struggles to generalize on the validation set. In other words, it is memorizing the training set which is an example of overfitting. One possible solution is to stop the training at the inflection point, using a technique called the *early stopping criterion*.
- Panel (c): This example in panel (c) represents a good learning curve where both the training loss and validation loss decrease initially and then reach a plateau. This indicates that the algorithm is no longer learning new information with each epoch.
- Panel (d): In panel (d), both curves are initially decreasing (indicating successful training) but then level off with a gap between them. This suggests that our training set is not representative of the validation set. In other words, there are too few examples in the training set that reflect what occurs in the validation set. This is an issue of data balance.
- Panels (e) and (f): Both panels highlight the same issue, which is a validation set that is not representative of the information we are seeking. In panel (e), the algorithm is learning from the training set, but no meaningful conclusions can be drawn from the validation set. This could be due to the amount of data used in the validation set or another data balance issue. In panel (f), the validation loss is lower than the training loss, suggesting that perhaps some challenging patterns still exist in the training set but not in the validation set. This makes the validation set easier to predict. Again, this reflects the issue of the validation set's lack of representativeness.

This provides the reader with an understanding of how to interpret and diagnose the loss curves. However, once we identify an issue like overfitting, what tools and techniques can be used to address it?

We can now shift our focus to the crucial step of fine-tuning a deep learning model. This entails wrapping up the various methods and hyperparameters already presented that require attention in order to optimize the model's performance. This discussion will provide an opportunity to explore the concept of regularization and its role in enhancing the model's generalization capabilities.

2.3.5 Fine-tuning Neural Networks

In this section, we will explore the topics that have been covered thus far and provide a comprehensive summary of how users can modify the loss curves to achieve desirable outcomes. Prior to this summary, we will introduce regularization techniques, a topic that we intentionally omitted in Section 2.3.3, but is crucial for enhancing a model's performance and effectively combating overfitting. Once we have presented regularization, we will then proceed to discuss what we refer to as "tunable parameters." This part will provide an insightful list of the various factors that have already been presented, which users can leverage to fine-tune their models. The summary will encompass essential hyperparameters and useful functions, ranging from the learning rate to the activation functions.

2.3.5.1 Regularization techniques

The objective behind regularization is to reduce the complexity of the model by adding some constraints. Regularization techniques are widely used in order to prevent overfitting (by prevent-

ing the model from becoming overly complex and memorizing training data), improve generalization by promoting simpler weights configuration (avoiding noise and focusing on trends), select features (pushing some weights to zero to ignore redundant features) and, as we said, control the trade-off between model complexity and data fitting.

Here, we present three regularizations techniques and how they are used:

- **Dropout:** In dropout regularization, each neuron outputs only a single value. During the training loop, dropout randomly deactivates some neuron outputs by setting them to zero. This technique is employed to mitigate overfitting by introducing noise and promoting redundancy in the network. By randomly disabling neurons, dropout encourages other neurons to develop multiple patterns to compensate, thus reducing the reliance on any individual neuron. This helps prevent the network from relying too heavily on specific features and improves its generalization ability. Dropout is particularly effective for larger networks with a higher risk of overfitting, but it can also be beneficial for smaller networks. It is important to note that dropout is applied during training and is typically turned off during inference or testing.
- **L1 Regularization:** L1 regularization, also known as Lasso regularization, adds a penalty term to the loss function proportional to the L1 norm of the weight vector. This regularization technique promotes sparsity in the model by driving some weights to exactly zero. It encourages feature selection and can be beneficial when dealing with high-dimensional datasets, as it helps identify the most relevant features for the task at hand. In order to perform it, we add a term to the loss function:

$$J(\Theta) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^i) - y^i)^2 + \lambda \sum_{j=1}^m |\theta_j|$$

During the gradient descent, a term will appear:

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^i) - y^i) - \alpha \lambda \text{sign}(\theta_j)$$

The term added penalizes the loss function and we call λ the *regularization parameter*. It is another hyperparameter that the user will have to choose and it controls the trade-off between minimizing the chosen cost function and reducing the magnitude of the weights.

- **L2 Regularization:** L2 regularization, also known as Ridge regularization, adds a penalty term to the loss function proportional to the L2 norm of the weight vector. The L2 regularization works the same way as the L1 does mathematically but encourages smaller weights without driving them to exactly zero. It helps prevent overfitting and reduces the impact of individual weights on the overall model. L2 regularization is commonly used and can improve the generalization performance of the model.
- **Combining L1 and L2** is known as Elastic Net regularization and leverages the benefit of both regularizations (Jin et al., 2009). The L1 component promotes sparsity and feature selection, while the L2 component encourages small but non-zero coefficients. This can be particularly useful when dealing with high-dimensional datasets where there may be many irrelevant or redundant features.

By incorporating dropout and regularization techniques such as L1 and L2 regularization, machine learning models can achieve better performance, handle overfitting, and enhance their ability to generalize well to unseen data. Even if they're not considered to be stricto sensu regularization techniques, we can mention two other techniques usually used to solve issues of overfitting or bad training in general: the early stopping criterion and the data augmentation. We already explained

the early stopping criterion, it then determine when to stop the training process based on a pre-defined condition. Usually, it monitors a specific metric, such as validation loss, and stops the training if the metric does not improve or starts to deteriorate.

Data augmentation is an ensemble of techniques to artificially increase the size and diversity of the training dataset by applying various transformations. The transformations can be operations such as rotation, flipping, or cropping for images, as well as scaling or adding noise. the data then created are considered to be other data and the size of the whole dataset is increased. Data augmentation helps improve the model's ability to generalize by exposing it to different variations of the training examples, thereby reducing overfitting and improving its performance on unseen data. Data augmentation techniques is a wide field in the artificial intelligence domain still improved everyday (Maharana et al., 2022; Shorten and Khoshgoftaar, 2019; van Dyk and Meng, 2001).

2.3.5.2 Tunable Parameters

Henceforth, the reader possesses all the foundational knowledge necessary to refine a model to address the majority of challenges they may encounter. In the subsequent table, we encapsulate all of this information, which represents the entirety of parameters upon which the user can exert influence. All values that require predefined settings, such as the learning rate or the number of layers, are referred to as *hyperparameters*. Indeed, the term "parameter" is generally reserved for the weights of the model. Consequently, all elements in table 2.7 must be selected by the user and can subsequently be adjusted to overcome specific learning challenges.

One might observe that certain elements, such as the proportions of the dataset designated for the training, validation, and test subsets, have not been expounded upon in this section. Indeed, these constitute decisions that fall within the user's purview. We have, in actuality, reserved the discussion related to data preparation and all its associated facets for the subsequent section. Efforts will be made to outline the principles that constitute sound analysis and proficient preprocessing of the data.

2.3.6 Deep Learning System Design

2.3.6.1 Data Analysis

The primary objective of a learning algorithm is to achieve high performance, and sometimes, it is easier to achieve high performance by picking up on peculiarities in the data rather than learning what the user intends (Ribeiro et al., 2016). A well-known example in the AI community involves a neural network trained to differentiate between wolf and husky pictures. Despite achieving high performance during training, the algorithm had a tendency to misclassify clear and obvious images. Further investigation revealed that the model had developed a bias and learned to classify based on the presence of snow in the background. This bias emerged due to all the training images of wolves featuring snow, while the images of huskies did not. Analyzing, understanding, and preparing a dataset is the cornerstone of building a successful machine learning model.

There are several risks associated with a dataset that need to be carefully checked for errors:

- Firstly, insufficient data for analysis can pose a significant challenge. In the case of an Artificial Neural Network (ANN), a rule of thumb is to have a number of samples (N_{sample}) that exceeds twice the number of parameters ($2N_{parameters}$).
- Secondly, it is crucial for the samples to exhibit sufficient correlation with the underlying physical problem being investigated. Training on noisy data can lead to inaccurate results.
- Furthermore, excessive correlation among samples should be avoided, as it can introduce bias and hinder the generalizability of the model.

Notation	Hyperparameter	Description
α	Learning rate	Controls the step size during gradient descent and affects the speed of convergence.
N_{epochs}	Number of epochs	Determines the number of times the entire dataset is passed through the network during training.
b or N_{batch}	Batch size	Specifies the number of training examples processed in each iteration, or each <i>step</i> .
g	Activation function	Determines the non-linearity applied to the output of each neuron.
N_{layers}	Number of layers	Defines the depth of the neural network by specifying the number of hidden layers, i.e. the number of "columns" of neurons.
N_{neurons}	Number of neurons per layer	Specifies the number of neurons or units in each hidden layer.
J	Loss function	Measures the discrepancy between the predicted and actual values during training.
λ_1 and/or λ_2	For regularization techniques	Parameters for the L1 and L2 regularizations, used to prevent overfitting.
D and the chosen layer	Dropout	In the dropout strategy, the dropout rate D (between 0 and 1) is the proportion of neurons randomly "dropped out" during training to prevent overfitting, and the user has to choose on which layer to put the dropout
	Weight initialization strategy	Determines how the initial weights of the network are set. We usually initialize them randomly.
	Optimizer	Selects the optimization algorithm used to update the network weights, such as Adam, RMSProp, or SGD.
	Learning rate scheduler	Controls the dynamic adjustment of the learning rate during training.
	Early stopping criterion	Stops training if the model's performance on a validation set fails to improve after a certain number of epochs.

Table 2.7 – Non-exhaustive list of the main hyperparameters and factors which can be modified by the user.

- Data corruption is yet another potential pitfall, encompassing issues like incorrect formats or conversion errors, which need to be addressed.
- Errors in labeling the data can also have a substantial impact on the performance of machine learning models.

Overall, a high-quality dataset should be flexible, allowing for easy manipulation and modification. When working on a problem, the dataset should establish the boundaries within which your algorithm operates, while also ensuring that the data accurately represents the problem domain. This highlights the importance of data preprocessing, which involves addressing measurement uncertainties, mitigating noise originating from the measurement chain, ensuring appropriate data formatting, and verifying the quality of the collected information.

A well-curated dataset provides the necessary context for your problem, facilitating seamless analysis within a specific environment, such as Python. To assess the robustness of a dataset, several techniques can be employed:

- Direct visualization such as histograms or lineplots.
- Construction of a correlation matrix to evaluate relationships between different samples.
- Examination of the correlation matrix between different classes.
- Utilization of Principal Component Analysis (PCA) on the dataset to discern the relevance of specific variables (Bro and Smilde, 2014). PCA will be introduced in Bouriat et al. (2022), presented in Chapter 3.
- Conducting statistical analyses to explore inter-sample correlations, such as spatial correlation for images, temporal correlation for time series data, or radiometric correlation for multi/hyperspectral training.
- Study of the noise present in the dataset.
- Study of the presence and frequency of missing values.

By following this non-exhaustive list, valuable insights can be gained into the dataset's reliability, the relationships between variables, and the impact of noise on the data. Typically, a *pipeline* is established—a methodical and automated procedure for managing and refining data from its raw form to its ultimate state of readiness for machine learning tasks. It is crucial to bear in mind the importance of not constructing a machine learning algorithm based on an unverified and unstructured dataset. In this context, it is crucial to engage in discussions with domain experts who are familiar with the field from which the data originates. This collaboration allows for valuable insights, as they may already possess knowledge about potential issues, ideas, or processing techniques relevant to the analysis. As a data scientist, it is essential to grasp and comprehend any findings that arise during the analysis, benefiting from the expertise and understanding shared within the community. Two fundamental principles to uphold are prioritizing result visualization and consistently scrutinizing the data prior to the models to address any potential malfunctions. We will not develop any further the notion of data analysis because a full analysis of the ACE dataset, which is specific to our problem, is given in Chapter 3, by Bouriat et al. (2022).

By discussing data analysis, we gained valuable insights into our dataset. The point was to explore its characteristics through visualization, identify biases or correlations, gain a deeper understanding of the relationships between variables or observe the presence of missing values. This knowledge serves as the foundation to perform effective data preprocessing. By understanding the data through this kind of analysis, we can make informed decisions about how to clean, transform, and prepare the dataset. The idea is to enhance its quality and relevance for a given machine learning tasks (we will not prepare time series as we prepare images). Data analysis provides the necessary context and understanding that helps us to perform effective preprocessing and build accurate and robust models.

2.3.6.2 Data Preprocessing

In the context of our discussion on data analysis, data preprocessing is directly influenced by the user's desired objective. The effectiveness of our model in addressing a specific problem greatly hinges on the quality of data preprocessing, which must be tailored to suit the particular problem at hand. Data preprocessing encompasses various techniques that depend on the dataset itself, and although there are numerous approaches, several key domains frequently emerge:

- **Data Cleaning:** This category encompasses methods dedicated to cleaning the data, such as handling outliers or resolving inconsistencies that may arise within the dataset. It is not uncommon for data to undergo different processes prior to analysis, leading to variations in units within a single feature (e.g., velocity data expressed in both meters per second and kilometers per hour). Therefore, data cleaning becomes a crucial initial step following analysis.
- **Feature Selection and Extraction:** Not all features in a dataset necessarily contribute meaningfully to the problem at hand. As mentioned before, an excess of features can potentially result in overfitting or hinder the model's convergence during training. Thus, identifying and selecting the most relevant features, as well as reducing dimensionality when appropriate (e.g., extracting features exhibiting high correlations), are essential steps. This process often follows techniques like PCA or correlation analysis.
- **Missing Values:** Dealing with missing data is a challenging task that requires careful consideration. While it falls under the umbrella of data cleaning, it warrants specific attention. Thoroughly analyzing the quantity, location, and significance of missing data can be particularly challenging, especially in fields like space meteorology and time-series in general. In some cases, missing data indicates instrument measurement saturation, and for extreme value identification, missing data itself becomes valuable information. After analysis, selecting the appropriate method for handling missing values, whether through simple imputation or more advanced techniques like spline interpolation, becomes essential. For further insights, the paper by [Bouriat et al. \(2022\)](#) in Chapter 3 can be referenced.
- **Scaling and Normalization:** Normalization is a specific type of scaling technique. It aims to bring each feature to a comparable scale, ensuring that the model assigns equal importance to each of them during training. Maintaining proportional contributions from each feature is crucial. Consequently, the responsibility of performing normalization or scaling lies with the user, except in specific scenarios. Scaling, in a broader sense, refers to transforming the data while preserving its shape and distribution. Among scaling and normalization methods, we can mention:
 - **Min-Max Scaling:** Rescales the dataset X to a specific range, typically between 0 and 1. The minimum value of the feature is mapped to 0, and the maximum value is mapped to 1, with other values scaled proportionally in between. Values $x \in X$ are changed to x_{scaled} following:

$$x_{scaled} = \frac{x - \min(X)}{\max(X) - \min(X)}$$

- **Z-Score Normalization:** It transforms the data to have zero mean and unit standard deviation. Each data point x is subtracted by the mean of the feature X and divided by its standard deviation as follows:

$$x_{normalized} = \frac{x - \text{mean}(X)}{\text{std}(X)}$$

- **Standardization:** It transforms the data to have zero mean and unit variance, similar to z-score normalization.

$$x_{\text{standardized}} = \frac{x - \text{mean}(x)}{\text{std}(x)}$$

However, standardization does not restrict the data to a specific range (scaling method, not normalization).

- **Robust Scaling:** It scales the data using statistical measures that are robust to outliers. Robust scaling utilizes measures like median and interquartile range to reduce the influence of extreme values on the scaling process.

$$x_{\text{scaled}} = \frac{x - \text{median}(X)}{\text{Q3}(X) - \text{Q1}(X)}$$

- **Time-Series Data:** Time-series data requires a specific set of treatments based on the desired objective, such as temporal resolution changes or temporal correlations.
- **Imbalanced Data:** As previously mentioned, a dataset should be representative of the real-world scenario. In classification problems, it often happens that one class is overrepresented compared to others, which can negatively impact the algorithm's performance in modeling that class. Various techniques exist to address this issue, such as undersampling (deliberately removing instances from the majority class), oversampling (repeating instances from the minority class while potentially applying transformations), and, more generally, data augmentation methods. Readers can refer to existing reviews in the field, such as [Kotsiantis et al. \(2005\)](#); [Spelman and Porkodi \(2018\)](#), or explore literature on data augmentation more broadly ([Maharana et al., 2022](#); [Shorten and Khoshgoftaar, 2019](#); [Shorten et al., 2021](#)).
- **Splitting into Training, Validation, and Test Sets:** It is crucial to split the dataset into separate subsets for training, validation, and testing, and we consider this step to be part of the data preprocessing. As we already explained, this ensures that the model's performance is evaluated on unseen data. Common splitting techniques include random sampling, stratified sampling, or time-based splitting for temporal data ([James et al., 2021](#)). This might also be a good moment to mention cross-validation, a method that doesn't simply split the dataset into three subsets. Firstly, a portion of the dataset is kept aside as the test set. Instead of just dividing the data into a single training-validation pair, cross-validation repeats this process multiple times, each time with different data partitions. One of the most commonly used techniques is k-fold cross-validation. It involves dividing the dataset into k equally sized parts, or "folds". The model is then trained k times, using k-1 folds as the training set and the remaining fold as the validation set. The model's performance is evaluated by averaging the results obtained from the k iterations. This method provides a more reliable estimate of the model's performance and its ability to generalize to new data.
- **Transformation of Categorical Data:** Sometimes, categories need to be explicitly modified to match the model's expected outputs. For example, in the case of MNIST digit recognition (Modified National Institute of Standards and Technology - a collection of handwritten digits represented as grayscale images), the model's output should not be a single digit from 0 to 9, but rather a vector of size 10. This vector consists of nine zeros and a single 1, where the position of the 1 corresponds to the digit. This transformation is commonly referred to as one-hot encoding (for instance, transforming the digit 3 into the vector $[0, 0, 0, 1, 0, 0, 0, 0, 0, 0]$).
- **Distributions:** Asymmetry in the distribution of certain data can present challenges during training. Some statistical models assume that the target values follow a normal distribution, as is the case with linear regression. Certain loss functions used in neural networks, such as

mean squared error (MSE) or mean absolute error (MAE), implicitly assume symmetric errors and may perform better with normal distributions. However, it's important to note that deep learning models are flexible and can handle a wide range of data distributions without explicit assumptions (Goodfellow et al., 2016). This highlights the importance of understanding the distribution we are working with and the occasional need to adapt them. Note that, currently we are considering the possibility of conducting a research paper, exploring the quantification of the impact of various distributions in input features on model efficiency and training time.

In general, it is also possible to explore the generation of new data by combining existing datasets, aggregating them, adjusting their granularity, and more. The possibilities are limitless. Each user should leverage their data analysis skills and consult domain experts to evaluate the optimal preprocessing techniques. For example, let's consider a scenario where the goal is to predict extreme earthquakes. The dataset would likely consist of time series data, and extreme earthquake events would be relatively rare. In this context, a data scientist could focus their preprocessing efforts on determining the appropriate temporal resolution, addressing the imbalanced nature of the dataset, handling missing data, and seeking guidance from seismic experts on intelligent data fusion techniques.

After completing the data analysis and preprocessing stages, the next crucial step is selecting an appropriate architecture for the model. It is important to consider that certain aspects of data preprocessing are intertwined with the choice of architecture. The preprocessing techniques applied should align with and complement the selected architecture to ensure optimal performance and meaningful results.

2.3.6.3 Architecture

How to choose an architecture?

The choice of architecture is a critical aspect of problem-solving, but there is no universal method, and finding the right architecture can be quite challenging. A significant area of current research focuses on automating architecture selection. This area is known as Neural Architecture Search (NAS) and encompasses techniques such as reinforcement learning-based methods, evolutionary algorithms, and parameter sharing techniques (Elsken et al., 2019; Liu et al., 2017; Pham et al., 2018; Zoph and Le, 2016).

Currently, the search for the optimal architecture is largely empirical. When starting out in deep learning, it is easier to focus on well-established architectures that have proven to be effective in their respective domains. In the following sections, we will introduce some of the most well-known architecture families, including Convolutional Neural Networks (CNNs), Autoencoders, Gated Recurrent Units (GRUs), and Long-Short Term Memory (LSTMs). We will also delve into the Temporal Convolutional Network (TCNs), as it is the main architecture used in this thesis. It is important to note that the fully-connected neural network (FCNN), which we have already discussed and used as a starting point for explanations, is also an architecture employed in this thesis.

The architecture depends on the type of data we're using. CNNs are specifically designed to analyze and process image data and are great at identifying patterns in images. Recurrent Neural Networks (RNNs) such as GRUs and LSTMs are designed to process sequential data by maintaining an internal state that allows them to remember previous inputs. This makes them well-suited for tasks such as speech recognition, language translation, and natural language processing. For classification tasks, FCNNs can be used. In case we have missing data, Generative Adversarial

Networks (GANs), an architecture that pits a generator network against a discriminator network, working in tandem to generate realistic synthetic data, are commonly used for generating new data similar to the ones we already have.

The architecture also depends on the amount of available data. As we have seen, applying a complex model to a small dataset can result in overfitting. It is crucial to provide sufficient data to the algorithm so that it can comprehend the intricate relationships between inputs and outputs. Moreover, the number of parameters in the model needs to be chosen carefully. In the case of linear regression, if the number of parameters exceeds or equals the number of training samples, overfitting is guaranteed. However, when it comes to neural networks, [Zhang et al. \(2016\)](#) have shown that a simple two-layer neural network with $2n + d$ parameters can perfectly fit any dataset consisting of n samples with a dimension of d (indicating overfitting). Nevertheless, deep neural networks often perform well despite the potential overfitting issues, thanks to regularization effects inherent in the optimization algorithm and the network architecture. Additionally, explicit regularization methods such as dropout or data augmentation are commonly employed. Neural networks model highly complex relationships, and using a small network (i.e., making the data appear larger by employing a smaller model) can lead to the problem of the network being too simplistic and incapable of representing the desired mapping. These considerations emphasize the significance of selecting the appropriate architecture (alongside a well-preprocessed dataset). Lastly, when choosing a neural network architecture, it is important to examine existing models and benchmarks for the targeted task. This can provide valuable insights into commonly used neural network architectures for similar tasks and their performance. In the subsequent subsections, we will introduce several renowned architectures; however, it is important to note the existence of numerous additional architectures.

Convolutional Neural Networks

As we mentioned, Convolutional Neural Networks widely used for image and video recognition and are designed to used hierarchical pattern recognition to automatically learn and extract features from input data. A lot of literature exist on the topic ([Krizhevsky et al., 2012](#); [LeCun et al., 2015](#)), and we'll try to quickly explain their functioning here.

The basic building blocks of a CNN are convolutional layers, non-linearity layers, pooling layers and fully-connected layers ([Albawi et al., 2017](#)). We already known what fully connected layers are, but let's explain convolutional and pooling layers. For this, let's imagine that our purpose is to recognize dogs in pictures. Our dataset is then a set of images and yes/no labels associated.

- Convolutional layers: Convolutional layers play a crucial role in detecting local patterns or features in the input data by applying learnable filters, also known as kernels. These filters are matrices of weights that the algorithm learns. The objective is to identify patterns in images. The basic principle of a convolutional layer is illustrated in [Figure 2.17](#).

In this process, the filter slides across the image and performs element-wise multiplications and summations, including the bias term, to generate a feature map. Initially, the filter is positioned at the top-left corner of the input image, where it undergoes element-wise multiplication with the corresponding values of the image patch it covers. The results are then summed to obtain a value that becomes the top-left entry of a new matrix. Subsequently, the filter is shifted one position to the right, and this procedure continues. [Figure 2.18](#) illustrates the first two steps of this computation. Note that the RGB channels of the image are separated, and each filter comprises three matrices.

The amount by which the filter moves horizontally and vertically is determined by the *stride*, which can be adjusted. Additionally, *padding* can be introduced by adding a row and column

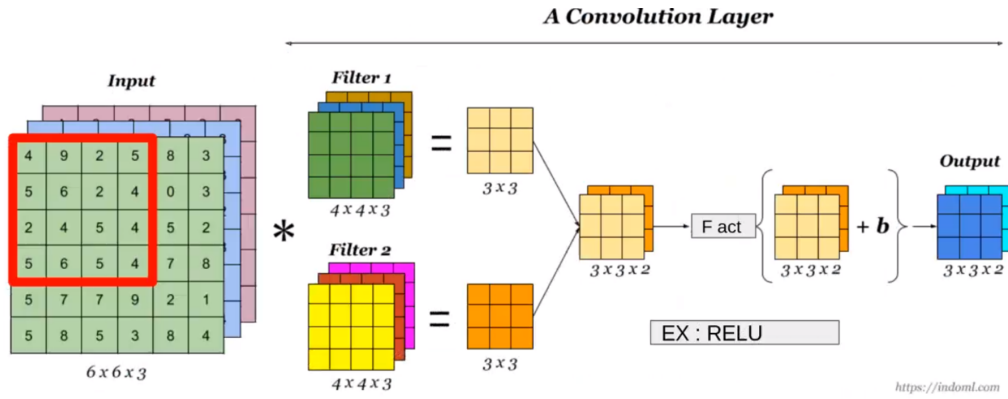


Figure 2.17 – Example of a convolutional layer applied to a 6x6 input with 3 channels, using 4x4 filters (3 channels). No padding and a stride of one are applied. Adapted from <https://indoml.com>.

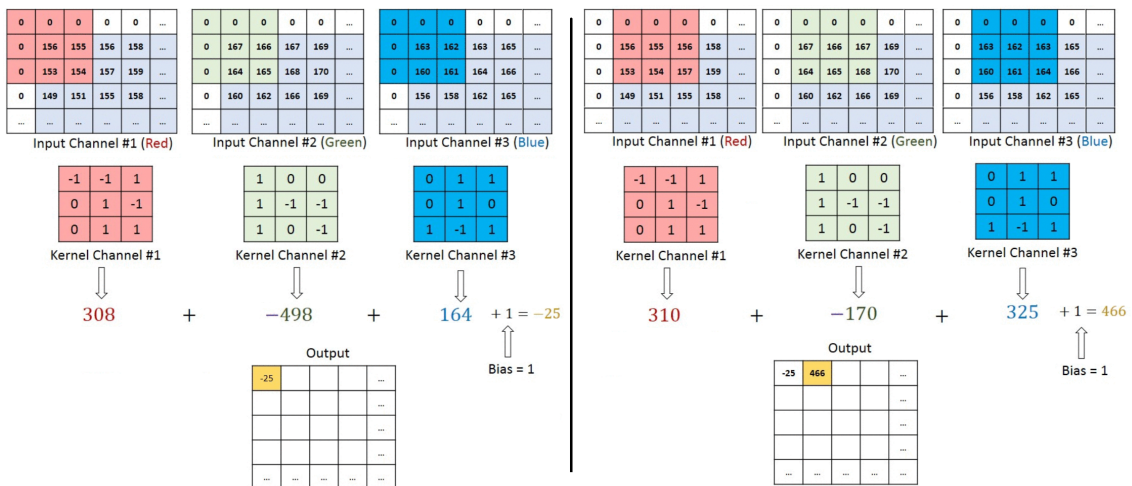


Figure 2.18 – Example of the filter sliding over the three channels. Padding has been applied, and the stride is set to one.

of zeros to the image, as seen in Figure 2.18. Padding helps maintain the spatial dimensions of the input data, ensuring that the output has the same size as the input. Finally, a dilation factor can be defined that dilates each pixel when the filters is patched over the image (see Figure 2.19).

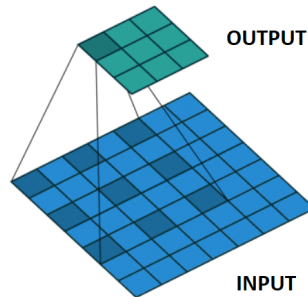


Figure 2.19 – Example of the a dilation factor of 2 where the filter skips one pixel in between each value, creating a sparse sampling pattern. Here, the filter is patched on the input (in blue) to create the first value of the output matrix (in cyan).

When an image with height H_{in} and width W_{in} passes through a convolutional layer, the output dimensions are calculated using the following formulas:

$$H_{out} = \left\lfloor \frac{H_{in} + 2P - \text{dilation} \times (K_H - 1) - 1}{S} + 1 \right\rfloor$$

$$W_{out} = \left\lfloor \frac{W_{in} + 2P - \text{dilation} \times (K_W - 1) - 1}{S} + 1 \right\rfloor$$

Here, P represents the padding, K_H and K_W denote the height and width of the kernel, S represents the stride, and dilation indicates the dilation factor. The stride and padding values can differ for the height and width dimensions.

- Pooling layer: Pooling layers are commonly utilized to reduce the spatial dimensions of the input feature maps. Here are three pooling techniques:
 - Max pooling: Max pooling scans the values within a pooling window, similar to a kernel, and selects the highest value. It replaces the entire window in the output feature map with this maximum value. Figure 2.20 illustrates an example of max pooling with a window size of 2x2 and a stride of 2, resulting in non-overlapping regions.

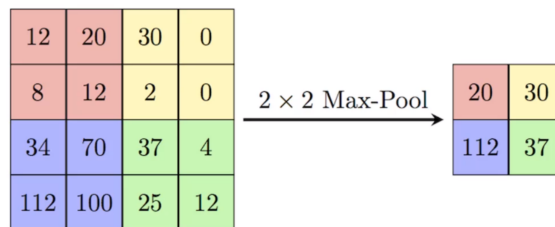


Figure 2.20 – Example of a pooling layer using the max pooling method. Source: Computer Science Wiki <https://computer-science-wiki.org/index.php/File:MaxpoolSimplified.png>, last accessed in July 2023.

- Average Pooling: Similar to max pooling, average pooling operates within a pooling window but computes the average of all the values instead of selecting the maximum value.

- Global Pooling: Global pooling performs average or max pooling over the entire matrix for each channel. Consequently, the output is a single value per channel. In the example shown in Figure 2.18, there would be three values corresponding to the Red, Green, and Blue channels.

Now, we can understand the family of convolutional neural networks, as with the example depicted in Figure 2.21. In this illustration, the convolutional and pooling layers are utilized to extract and identify features within the image. The output from these layers consists of condensed vectors containing the relevant information. The subsequent fully-connected layers then function as a classifier to determine whether the image is that of a dog or not.

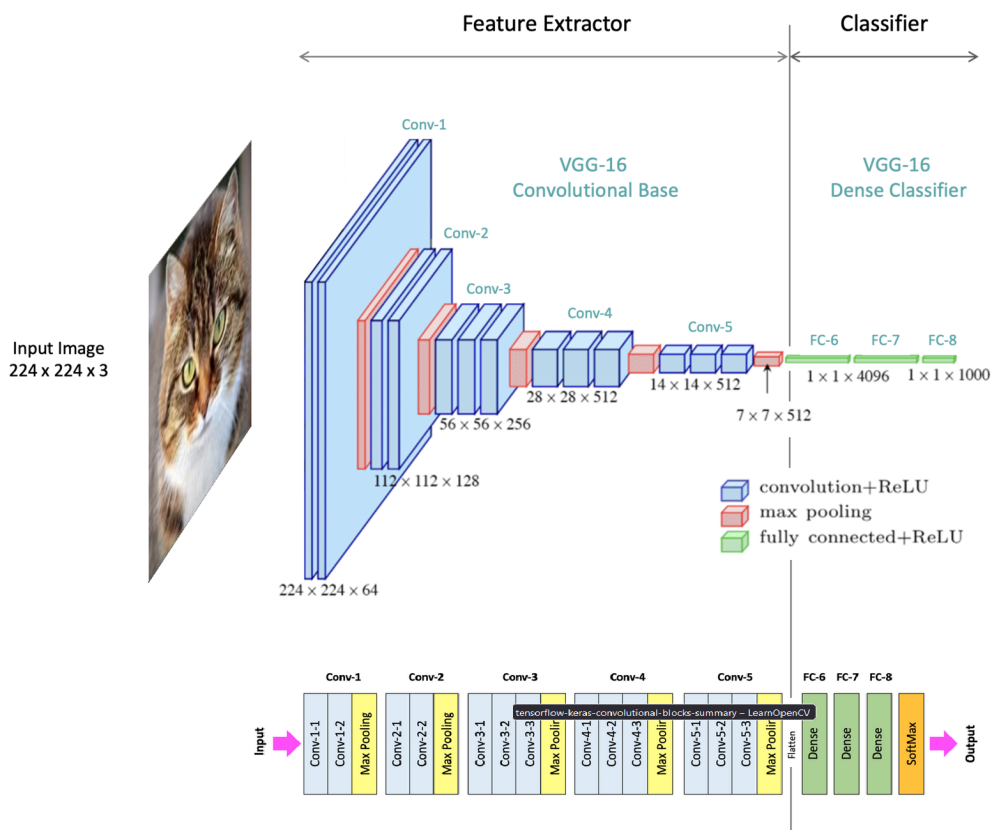


Figure 2.21 – Example of a convolutional neural network with 5 convolutional layers and 5 pooling layers, culminating in three fully-connected layers. These final layers can be employed because the pixel values have been transformed into vectors through a "flatten" operation, allowing us to return to the familiar concept of classic neural networks. Image from <https://learnopencv.com/understanding-convolutional-neural-networks-cnn/#Convolutional-Blocks-and-Pooling-Layers>, last accessed in July 2023.

Convolution is a mathematical operation, and since convolutions are linear, we can represent them here as a matrix product. CNNs are particularly effective and efficient in locating specific features in images. Furthermore, CNNs exhibit translation invariance, enabling them to recognize patterns within images regardless of their position. They can even recognize high-dimensional patterns, such as handwriting, with minimal image preprocessing (Lecun et al., 1998). In CNNs, the initial layers seem to capture simple features like edges or textures, while deeper layers capture object parts or semantic information (Zeiler and Fergus, 2013). This hierarchical structure allows CNNs to gain a better understanding of the relationships among different pieces of data.

On the other hand, the large number of parameters involved makes these architectures com-

putationally demanding. Moreover, CNN architectures require an extensive amount of data for training, meaning that a large number of inputs have to be labeled. Additionally, CNNs are prone to overfitting, meaning that they can memorize the noise and details of the training data, which may hinder their ability to generalize to new and diverse data. Lastly, these models offer limited interpretability and are often considered black boxes, which is an important limitation when we want to look towards grey-box models.

Autoencoders

Autoencoders (AE) are unsupervised algorithms that compress data into a lower-dimensional representation (the *encoder* part) and reconstruct it back to the initial dimension (the *decoder* part) (Hinton and Salakhutdinov, 2006; Vincent et al., 2010), as illustrated in Figure 2.22.

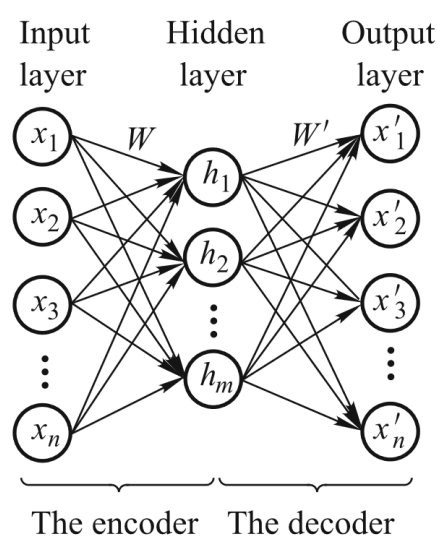


Figure 2.22 – Architecture of a basic autoencoder with W representing the weights of the encoder part and W' representing the weights of the decoder part. Figure adapted from Zhang et al. (2020).

Over the past decades, AEs have proven to be highly efficient in data dimensionality reduction, feature extraction, and data reconstruction (Zhang et al., 2020). They can be integrated into architectures such as CNNs or FCNNs. By reducing the dimension before reconstructing the data, AEs might lose some information from the original data but often also reduce noise and keep the "high-level" information. As well, they can be trained independently to mitigate noise, following the scheme presented in Figure 2.23, and subsequently utilized to clean a dataset.

In summary, AEs serve multiple purposes, such as dimensionality reduction, denoising, unsupervised learning, and even anomaly detection. In fact, AEs can identify anomalies or outliers in the data by computing the error between the input and the reconstructed output. Anomalous data points often yield higher reconstruction errors.

LSTM & GRU

There are plenty of accessible descriptions and articles out there that delve into the workings of Gated Recurrent Units (GRUs) and Long Short-Term Memory (LSTM) (Cho et al., 2014a,b; Hochreiter and Schmidhuber, 1997). Both of these belong to what we call recurrent neural networks (RNNs) (Tsoi, 1997), which are neural network setups used for handling sequential or time-based data.

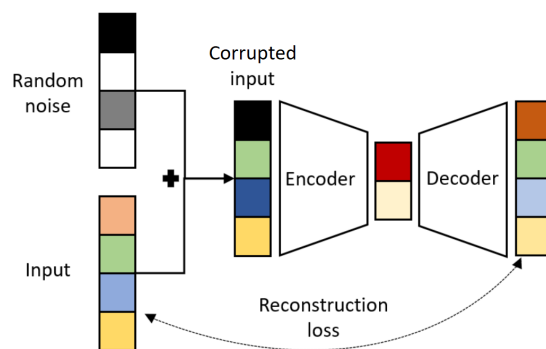


Figure 2.23 – Example of a denoising autoencoder where the corrupted input image is encoded to a representation and then decoded. Figure adapted from [Bank et al. \(2020\)](#).

Unlike the usual feedforward networks, RNNs come with recurrent connections. This lets them hold onto an internal memory and make use of past contextual information when dealing with new input, as you can see in Figure 2.24. This knack for retaining information across various time points makes RNNs pretty potent models for tasks like predicting sequences, generating text, auto-translation, and natural language processing. But, they can run into snags, like vanishing gradients over lengthy sequences. This hitch led to the creation of variations like LSTMs and GRUs, which aim to tackle these limitations head-on.

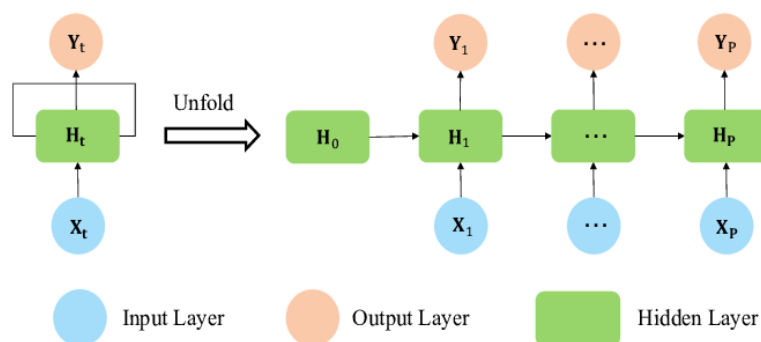


Figure 2.24 – Schematic view of the functioning of a Recurrent Neural Network. From [Ye et al. \(2022\)](#).

LSTM networks belong to the subset of RNNs and are designed to capture long-term dependencies. First introduced by [Hochreiter and Schmidhuber \(1997\)](#), one of their standout features is the gate mechanism that effectively controls the information flow within the network. This mechanism comprises three primary gates: the forget gate, the input gate, and the output gate (refer to Figure 2.25). These gates empower the LSTM to handle both historical and current information, as well as the generated output. This unique capability enables the LSTM to selectively retain or discard information over extended sequences, thereby proving to be exceptionally efficient in tackling long-term dependencies.

GRU stands for Gated Recurrent Unit and is another type of RNN designed to tackle the issue of vanishing or exploding gradients. Like LSTM, GRU aims to solve this challenge. However, GRU takes a different approach by simplifying the architecture. It combines the input and forget gates into a single "update gate," and it merges the output gate and cell memory into a unified

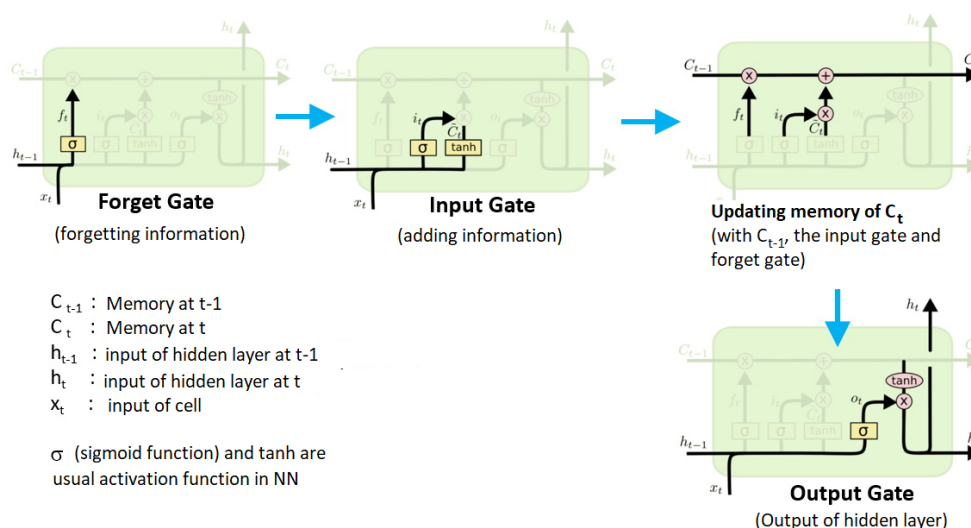


Figure 2.25 – Schematic view of a LSTM cell. Several cells are then put together to form the LSTM architecture. Figure adapted from Hairy (2021).

component known as the "hidden state."

GRU introduces an adaptive update mechanism, which determines how much of the past information should be forgotten or utilized based on new input data. This streamlined design of GRU makes it easier to train and involves fewer parameters compared to LSTM. Remarkably, GRU maintains competitive performance in various natural language processing tasks, highlighting its effectiveness despite its architectural simplicity.

Temporal-Convolutional Network

In the realm of deep learning, recurrent neural networks (LSTM and GRU) have conventionally been the go-to for sequence modeling. However, a recent development by Bai et al. (2018) suggests that convolutional networks should also be explored for handling sequential data. In their study, they demonstrated that convolutional networks can actually outperform RNNs in various tasks, while sidestepping the common issues often associated with recurrent models, such as the vanishing or exploding gradient problems, as well as challenges with memory retention. What's more, using convolutional networks enables parallel computation of outputs, which could potentially lead to performance boosts. This led them to propose an architecture later coined the Temporal Convolutional Network (TCN) (Bai et al., 2018; van den Oord et al., 2016). In what follows, much of the explanation draws heavily from a highly elucidating article penned by Francesco Lässig available on the <https://unit8.com>⁷ website, from which we gathered figures 26 to 30.

Starting with the basics, let's dive into the concept of 1D convolution, as depicted in Figure 2.26. This mirrors the workings of 2D convolutional layers. Imagine a kernel or filter in play, which slides along the input vector(s) (one filter assigned to each input variable). The dot product between this filter and the observed window is computed, and then the window shifts by a set stride. To ensure the output sequence maintains the same length as the input sequence, zero-

7. Last accessed on July 6, 2023: <https://unit8.com/resources/temporal-convolutional-networks-and-forecasting/>

padding comes into play, illustrated in Figure 2.27. Crucially, the convolution must retain a causal nature. In simpler terms, the i -th element of the output sequence can solely hinge on elements ranging from 0 to i in the input sequence. For time series data, the algorithm should steer clear of training on future values. Therefore, zero-padding is exclusively applied to the left side of the input sequence. Without delving into dilation, it's worth noting that the number of added zeros always equals the filter length minus one.

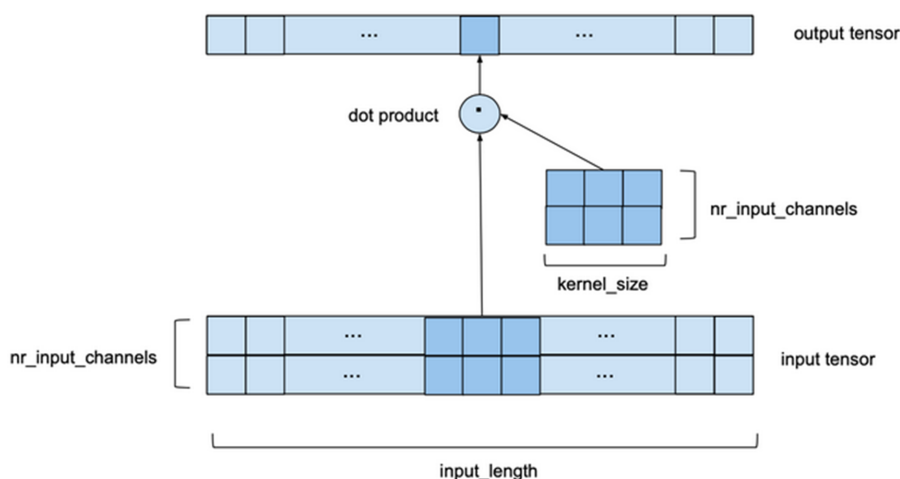


Figure 2.26 – This figure illustrates 1D convolution for $nr_input_channels$, the number of input features, on a sequence of length $input_length$ (for example, 30 minutes of data for two temporal variables).

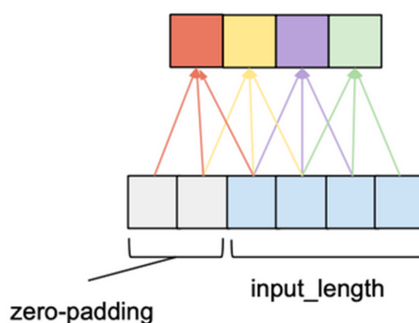


Figure 2.27 – Explanation of zero padding in the case of no dilation.

As we reach the end of our TCN, we'll end up with just a single value for a given input sequence. So, it's crucial that the entire input sequence is effectively utilized to produce this output – a concept referred to as "full history coverage". Building on the previous point, we can infer that a 1D convolutional network with n layers (the number of "stages"), employing a filter of length k yields a receptive field size of $r = 1 + n(k - 1)$ as showcased in Figure 2.28. The key notion is to set $r = input_length$ to ensure a comprehensive full history coverage, which necessitates a minimum of n layers. This leads to the following expression:

$$n = \frac{input_length - 1}{k - 1}$$

Now, introducing dilation into the picture – following a pattern resembling what we see in Figure 2.19, albeit in 1D – enables us to decrease the number of layers needed for full history

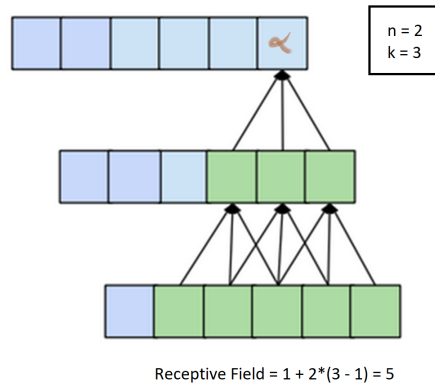


Figure 2.28 – Schematic view of the receptive field. A full history coverage would mean no blue squared in the bottom row.

coverage. Yet, a linear input tensor length remains a prerequisite for complete coverage. To tackle this, a gradual adjustment of dilation as layers are added can be employed, illustrated in Figure 2.29. In the example illustrated in this figure, the process begins with a dilation factor of 2 and increments it using $d = b^i$, where i denotes the particular layer and b represents the dilation base (the initial value). The width of the receptive field is then:

$$n = 1 + \sum_{i=0}^{n-1} (k - 1) \cdot b^i = 1 + (k - 1) \cdot \frac{b^n - 1}{b - 1}$$

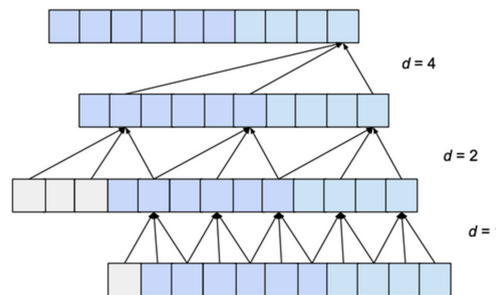


Figure 2.29 – Example of dilation 2^i over several layers.

Important note: In a general context, in order to prevent any "gaps" in a receptive field (missing values), the kernel size k should be at least as large as the dilation base b . Ultimately, to ensure complete coverage, our TCN must adhere to the following rule:

$$1 + (k - 1) \cdot \frac{b^n - 1}{b - 1} \geq \text{input_length} \tag{2.27}$$

Lastly, the necessary amount of zero-padding becomes:

$$P = b^i(k - 1)$$

Now, with all the necessary components at hand, we can delve into understanding the TCN, illustrated in Figure 2.30. It consists of a sequence of what we term as "residual blocks," each composed of convolutional layers. Instead of individual layers, these residual blocks are connected to one another, distinguished by their filter size and dilation (which increases as b^i). These blocks replace the standalone "layers" we've encountered thus far. Moreover, there are a few noteworthy changes:

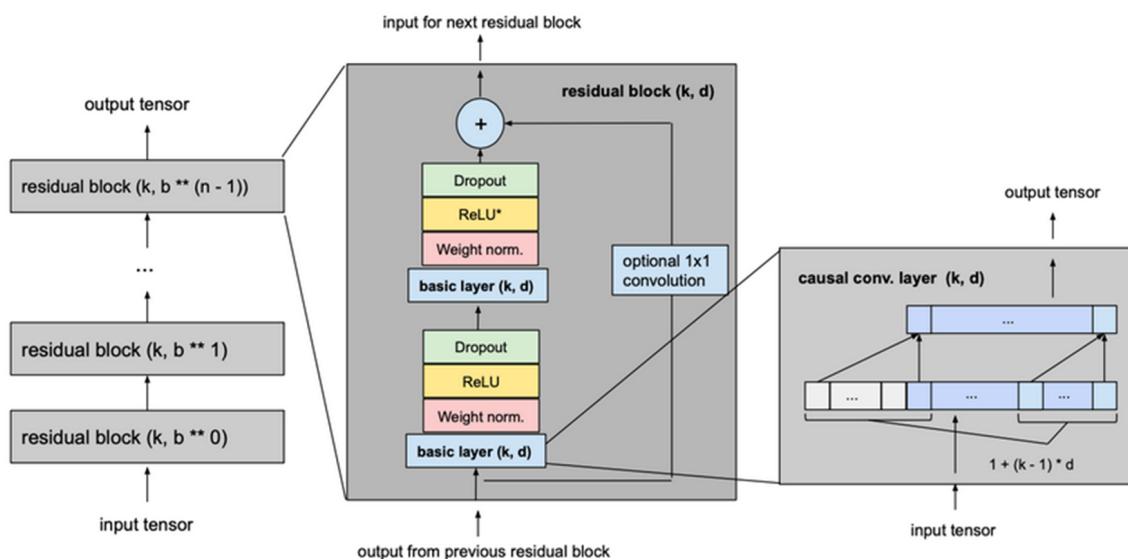


Figure 2.30 – Architecture of a TCN.

- Convolution is performed twice, as opposed to a single time. The first convolution acts as a local receptive field, capturing proximate dependencies within a specific context window. The subsequent convolution layer amalgamates these local features further to capture higher-level dependencies and patterns across a wider context window.
- To introduce non-linearity and sidestep a purely linear regression model, activation functions like ReLU are applied to the convolutional layers within the TCN.
- Weight normalization is utilized to standardize the input of hidden layers, addressing issues like the exploding gradient problem.
- Dropout regularization is incorporated after each convolutional layer in the residual blocks to prevent overfitting.
- A 1x1 convolution is included in both the first and last residual blocks (optional for the rest) to maintain consistent sizes.

In conclusion, TCNs present a robust alternative approach for sequence modeling. They outshine traditional recurrent networks by circumventing problems associated with gradient explosions and memory retention. By utilizing 1D convolutional layers, they capture temporal dependencies and intricate patterns within sequential data. Thanks to their capacity for parallel computations, they enhance computational efficiency. Through non-linear activations, weight normalization, and dropout regularization, they can be fine-tuned for various tasks. All these factors make TCNs a prime candidate for our study, displaying significant potential as an architecture for handling sequential data. Their recent emergence means they've never been applied to space weather problems before. We'll return to this point. It's also worth highlighting that the choice of employing 1D temporal convolution enables the input not only of a chosen set of past data but also of a vector of past data, facilitating a more comprehensive exploration of connections between phenomena that may have unfolded over time, leading to delayed effects.

2.4 Libraries & Needed Tools

To successfully implement all of the above, a combination of computational power and high-level libraries is essential. Calculations need to be swift and efficient. The choice of library or hardware has, over the past few years, become a crucial part of problem-solving. A few misguided decisions in this realm can render certain problems insurmountable. Therefore, let's briefly discuss the importance of hardware. We'll wrap up by introducing the libraries that have been employed throughout this study: PyTorch⁸ and PyTorch-Lightning⁹.

2.4.1 Hardware

There are two types of computing units: central processing units (CPUs) and graphics processing units (GPUs), each serving distinct roles.

What is a CPU?

The CPU serves as the central processor of a computer, whether it's referred to as Core 2 Duo or Athlon 64. Its acronym stands for Central Processing Unit. The CPU can feature multiple processing cores and is often referred to as the "brain" of the computer. It is essential for all modern computer systems, as it executes commands and processes necessary for the computer and operating system to function. Additionally, the CPU plays a crucial role in determining program execution speed, whether it involves web browsing or spreadsheet creation.

What is a GPU?

A GPU, on the other hand, is the processor responsible for powering the graphics card. GPU stands for Graphics Processing Unit. The primary distinction lies in the tasks they handle. The GPU consists of numerous smaller and specialized cores. Working in unison, these cores deliver exceptional performance when processing tasks that can be divided and executed across multiple cores. Originally, the GPU's main responsibilities included rendering pixels, textures, shapes on the screen, and video processing.

What sets a CPU apart from a GPU?

Although CPUs and GPUs share many similarities as essential processing engines, they possess different architectures and serve distinct purposes.

CPUs are well-suited for a broad range of workloads, particularly those that demand low latency and strong per-core performance. Operating as a powerful execution engine, the CPU allocates its fewer cores to individual tasks, emphasizing speedy execution. This makes it highly effective for tasks spanning from sequential processing to database operations.

Initially developed as specialized Application-Specific Integrated Circuits (ASICs) to accelerate specific 3D rendering tasks, GPUs have evolved over time to become more programmable and flexible. While their primary function remains handling graphics and increasingly realistic visuals in modern games, GPUs have also transformed into more versatile parallel processors, capable of handling an expanding array of applications.

In a broader context, the central distinction lies in the operational characteristics of CPUs and GPUs. CPUs are designed to execute a single operation per thread, implying limited parallelization capabilities. In contrast, GPUs excel at concurrently handling multiple, simpler operations, often

8. <https://pytorch.org/>

9. <https://lightning.ai/>

numbering in the hundreds or even thousands. By harnessing the computational power of GPUs, one can achieve a speedup factor defined by Amdahl's Law:

$$Speedup = \frac{1}{(1 - P) + \frac{P}{N}}$$

Here, "Speedup" represents the total acceleration anticipated through task parallelization, P denotes the proportion of the task amenable to parallelization (ranging from 0 to 1), and N signifies the quantity of available processors.

As we said, deep learning models require intensive computations, often involving large datasets and complex operations. Libraries like TensorFlow and PyTorch are optimized to leverage the full power of hardware resources, such as GPUs, in order to accelerate model training and inference. A GPU can handle several thousands of operations per seconde but managing the memory should be handled by a good library.

2.4.2 Libraries

There are many libraries, and we won't be able to cover them all. Among them, TensorFlow and PyTorch are two extremely popular libraries in the field of deep learning. Both of them serve as matrix and tensor managers, providing advanced features to simplify differential calculations on both CPUs and GPUs. Learning these libraries can be challenging, but it is absolutely essential for efficient work in these domains.

Two primary languages are commonly used to utilize these libraries: C++ and Python. C++ is a lower-level, more technical language that is often employed for model production. In this realm, a manager called Eigen, developed by INRIA, is widely used. On the other hand, Python is primarily utilized for artificial intelligence development and rapid prototyping of models.

PyTorch is an exceedingly flexible Python library that enables easy tensor management. With PyTorch, you can directly manipulate your variables and construct models in a more localized manner. In contrast, TensorFlow is more focused on classes and provides a more industrial and standardized approach. TensorFlow employs a graph system to define and execute computations, making it efficient for large-scale deployments.

In terms of philosophy, PyTorch is often favored in the research field, as it offers greater freedom and a more à la carte approach to building custom models, which we will briefly explain in the next section. On the other hand, TensorFlow is frequently preferred in industrial applications where standardization and scalability are crucial.

In summary, TensorFlow and PyTorch are powerful libraries for deep learning, offering advanced features for tensor manipulation and model construction. The choice between the two often depends on the specific use case, with PyTorch being favored in research and TensorFlow in industrial applications.

2.4.2.1 Getting Started with PyTorch

In this section, we'll provide an overview of how PyTorch functions are organized and how the library operates in general. This serves as a brief introduction that is necessary to understand PyTorch-Lightning. Subsequently, we'll delve into the different aspects of PyTorch-Lightning,

which form the fundamental building blocks of our code. The code is freely available on GitHub¹⁰ for anyone to access.

According to the PyTorch website, organizing code for data sample processing can become convoluted and difficult to maintain. It is preferable to have separate and modular code for dataset handling, independent from the model training code, to enhance readability and modularity. PyTorch provides two useful data primitives: `torch.utils.data.DataLoader` (shown in Figure 2.31) and `torch.utils.data.Dataset`. These tools allow you to work with pre-loaded datasets or your own custom data.

The `Dataset` object stores the samples along with their corresponding labels, and a `Dataset` object in PyTorch represents the data to be loaded. On the other hand, the `DataLoader` object creates an iterable interface to the `Dataset`, making it convenient to access the samples easily. To define a dataset class, we need to define three functions: `__init__`, `__len__`, and `__getitem__`. With these three functions, the `Dataset` class allows us to retrieve the features and labels of our dataset one sample at a time.

When coding in PyTorch (and most other libraries), it is recommended to follow these seven steps:

1. Import the necessary modules.
2. Define the model class by inheriting from `torch.nn.Module` (or `LightningModule` from PyTorch-Lightning, as we will see). In the example shown in Figure 2.31, the model has one input layer, a hidden layer with a ReLU activation function, and an output layer. Layers can be called through the "Layers" module of PyTorch. We can observe in the `forward` method that our input x passes through all the layers.
3. Instantiate the model by specifying the size of the layers (variables when initializing the `SimpleNet` class need to be defined). `input_size` and `output_size` represent the size of our input sample and the size of our prediction vector (e.g., 10 for MNIST).
4. Choose a loss function and an optimizer.
5. Prepare the training data by using the `Dataset` and `DataLoader` modules from PyTorch. Here, we did not create a custom `Dataset` or `DataLoader` and did not redefine `__init__`, `__len__`, and `__getitem__`.
6. Write the training loop, which involves forward propagation, loss computation, backpropagation to obtain gradients, and weight updates. You can refer to the PyTorch documentation for detailed explanations of `loss.backward()` and `optimizer.step()`. In this loop, PyTorch iterates through the `train_loader` object and knows what to do.
7. Finally, apply the model to an input to obtain a prediction.

Then, during model training, it is more efficient to pass samples in batches, shuffle the data at each epoch to prevent overfitting, and utilize multiprocessing in Python to speed up the data retrieval process. To simplify this process, PyTorch provides the `DataLoader` class. It acts as an iterable and abstracts away the complexities mentioned above, offering a user-friendly API for handling batching, shuffling, and multiprocessing.

To visualize and track the training progress of our model, we utilize Tensorboard¹¹, a visualization tool provided by TensorFlow. It offers graphical functionalities for visualizing learning curves, histograms, model graphs, images, embeddings, as well as performance profiling. Tensorboard greatly simplifies the exploration and understanding of our model's results.

10. <https://github.com/simonbouriat/DMSP-Auroral-Forecast>

11. <https://www.tensorflow.org/tensorboard>

```

import torch
import torch.nn as nn
import torch.optim as optim

class SimpleNet(nn.Module):
    def __init__(self, input_size, hidden_size, output_size):
        super(SimpleNet, self).__init__()
        self.fc1 = nn.Linear(input_size, hidden_size)
        self.relu = nn.ReLU()
        self.fc2 = nn.Linear(hidden_size, output_size)

    def forward(self, x):
        x = self.fc1(x)
        x = self.relu(x)
        x = self.fc2(x)
        return x

input_size = 10
hidden_size = 20
output_size = 2

model = SimpleNet(input_size, hidden_size, output_size)

criterion = nn.CrossEntropyLoss()
optimizer = optim.SGD(model.parameters(), lr=0.01)

# Assume you have input data x_train and corresponding labels y_train
# Convert your data to PyTorch tensors
x_train = torch.Tensor(x_train)
y_train = torch.LongTensor(y_train)

# Create a DataLoader to handle data loading
train_dataset = torch.utils.data.TensorDataset(x_train, y_train)
train_loader = torch.utils.data.DataLoader(train_dataset,
                                           batch_size=32, shuffle=True)

num_epochs = 10

for epoch in range(num_epochs):
    for inputs, labels in train_loader:
        # Reset gradients to zero
        optimizer.zero_grad()

        # Perform forward pass and predictions
        outputs = model(inputs)

        # Calculate the loss
        loss = criterion(outputs, labels)

        # Backpropagation and weight updates
        loss.backward()
        optimizer.step()

    # Print the average loss for each epoch
    print(f"Epoch {epoch+1}/{num_epochs}, Loss: {loss.item():.4f}")

# Assume you have new data x_test
# Convert the data to PyTorch tensors
x_test = torch.Tensor(x_test)

# Make predictions
predictions = model(x_test)

# Observe the results
print(predictions)

```

Figure 2.31 – Example of a very simple network with PyTorch, demonstrating the 7 essential steps to create a functioning machine learning algorithm. All 7 steps are explained in Section 2.4.2.1.

2.4.2.2 PyTorch-Lightning

PyTorch Lightning is a library built on top of PyTorch (and is also open-source) that aims to streamline the model development and training process by providing a structured framework and additional functionalities. PyTorch Lightning notably abstracts the training code, allowing for a clear separation between model-specific components and the training code itself. This enhances code readability, comprehension, and maintainability. Additionally, it offers a predefined training loop that facilitates the implementation of the standard training process, including training, validation, and testing steps. An example illustrating this can be seen in Figure 2.32. This reduces the amount of repetitive code and enables developers to focus more on the specific aspects of the model. Furthermore, PyTorch Lightning provides built-in features for parallelism and distribution, simplifying accelerated training on large datasets using multiple GPUs or nodes.

Thus, we conclude this comprehensive introduction to Machine Learning and Deep Learning. We have covered various topics, from the presence of AI in the space weather community to the mathematical reasoning behind neural networks, while also introducing the roles of CPUs and GPUs in the process. We concluded this section by presenting the general functioning of the PyTorch library and PyTorch Lightning, which allowed us to introduce the organization of our code, publicly accessible. In this way, we hope that the vocabulary, methods, and results will make more sense to the reader. We also hope that through these explanations, the reader will be able to reproduce the results presented throughout this thesis without much difficulty. Finally, the broader objective of this section, as well as the previous one, is to create a repository of information from which the reader can draw what interests them. Some topics may have slightly deviated from the scope of our study, but were necessary to open a (small) window into the world of artificial intelligence to make it accessible, while demonstrating its complexity.

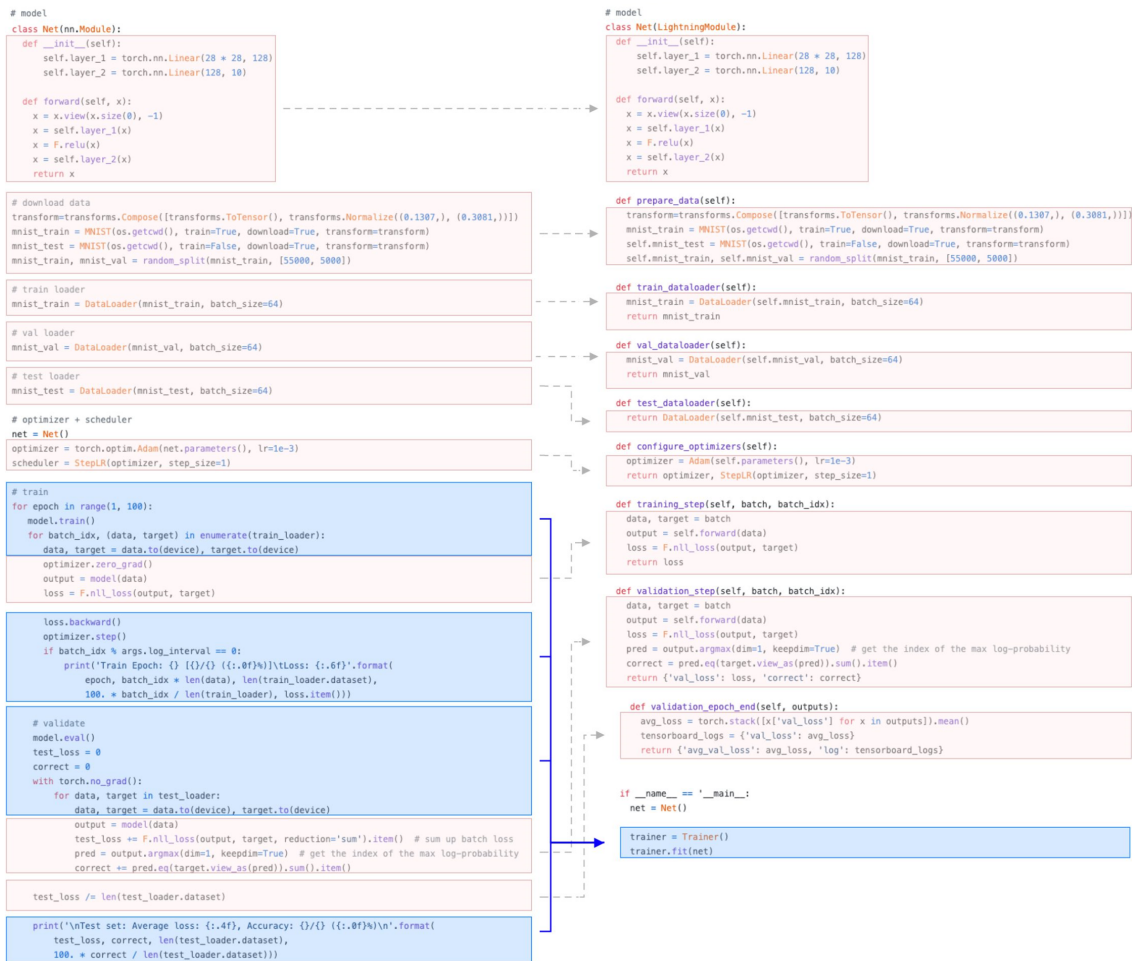


Figure 2.32 – Comparison between PyTorch and PyTorch Lightning, highlighting key differences. Image sourced from the PyTorch-Lightning documentation https://pytorch-lightning.readthedocs.io/en/0.7.1/introduction_guide.html, last accessed in July 2023.

3

Problem Statement, Data Analysis & Preprocessing

"Like I always say, can't find a door? Make your own."

Edward Elric - Full Metal Alchemist - Hiromu Arakawa

Contents

3.1	Clarifying the Research Problem	144
3.2	Data Description	146
3.2.1	Defense Meteorological Satellite Program (DMSP)	146
3.2.2	Advanced Composition Explorer (ACE)	151
3.2.3	High-Resolution OMNIWeb	154
3.3	Data Analysis & Preprocessing	155
3.3.1	ACE	155
3.3.2	DMSP	174
3.3.3	High-Resolution OMNIWeb	182
3.3.4	Input-Output Relationship: Justification for AI Implementation	184
3.4	Summary & Final Preprocessing	208
3.4.1	Pre-process Summary	209
3.4.2	Combining inputs and outputs: final datasets	210
3.4.3	Summary of our study	214

3.1 Clarifying the Research Problem

Now that we have set the framework and provided a detailed introduction to the two domains at the core of this thesis, we can present the problem we have chosen to address.

As we discussed in Chapter 1, electron precipitation into the ionosphere is driven by various mechanisms. Electrons precipitate as they enter the loss cone, traveling along magnetic field lines and often undergoing acceleration through mechanisms such as Quasi Static Potential Structure (QSPS) and Alfvénic. Some of the precipitating particles have been trapped in various magnetospheric regions (e.g., radiation belts, magnetospheric tail, etc.) and others can originate directly from the solar wind (e.g., via polar cusps). In this intricate series of events, particles interact with various factors like plasma waves, magnetic reconnection, and magnetospheric currents. These interactions lead the particles to carry current, transfer energy, and ultimately precipitate into the ionosphere.

Particle precipitation is a crucial link between the ionosphere and the magnetosphere, (McGranaghan et al., 2021), widely used in most Global Circulation Models (GCMs), such as the Global Ionosphere Thermosphere Model (GITM) (Ridley et al., 2006), the Thermosphere Ionosphere Electrodynamics General Circulation Model (TIE-GCM) (Roble et al., 1988), and the Whole Atmosphere Model-Ionosphere Plasmasphere Electrodynamics (WAM-IPE) model (Fuller-Rowell et al., 2008; Millward et al., 1996). This suggests that electron precipitation is a direct indicator of abrupt and intense electrical events in the ionosphere, which, in turn, leads to various impacts as we discussed in Section 1.3.3 of Chapter 1. Among these impacts, surface charging (Section 1.3.3.1) represents the greatest risk for LEO satellites in polar orbits (i.e., passing over the poles, where particles precipitate).

Several models of particle precipitation at the poles exist, such as Hardy et al. (1985), Fuller-Rowell and Evans (1987), and OVATION (Newell et al., 2014) (see Section 1.2.7.3). The first two models produce two-dimensional spatial distributions (2D) of energy and particle flux based on certain assumed parameters. They are generally limited to the northern hemisphere and are constrained by their input parameters. For example, the Hardy model relies on Kp and thus has a temporal resolution limited to 3 hours, preventing it from predicting changes on shorter time scales. The arrival of the third model, OVATION, brought a fresh perspective to precipitation modeling by using solar wind parameters as the sole driver of precipitation, a piece of information derived from outside the magnetosphere. OVATION became the first tool "to locate the auroral oval or to quantify its intensity" (Newell et al., 2002). Its success was evident, and OVATION became one of the most downloaded products from the National Oceanic and Atmospheric Administration (NOAA) Space Weather Prediction Center (SWPC)¹ (McGranaghan et al., 2021).

Subsequently, Newell et al. (2015) conducted a detailed analysis, and evaluated these models, leading to proposed improvements. They emphasized that the efficiency and usefulness of auroral particle precipitation models are heavily reliant on their input parameters and the choices made for their representation and organization. These limited choices often fail to capture the complexity of the system, including non-linearities. Consequently, as (McGranaghan et al., 2021) points out, existing models lack the ability to provide crucial information to users and fully understand the underlying physics. Additionally, merely aggregating different data is insufficient to produce reliable results, and it is sometimes essential to use direct in-situ observations of precipitation for model development (such as DMSP, instead of indices like Kp). McGranaghan et al. (2021) addressed these issues with their PrecipNet model, considering that Machine Learning meets the

1. <https://www.swpc.noaa.gov/products/aurora-30-minute-forecast>

expectations regarding the choice and representation of parameters.

Developing models with greater expressive capacity is imperative to effectively capture and represent the intricacies of precipitated electrons. PrecipNet has been a remarkable success in this regard, but it does exhibit some limitations concerning data representation. Given our approach closely aligns with the fundamental objectives of PrecipNet (i.e., contributing to a framework for employing machine learning models along the Sun-Earth chain), this thesis dedicates an entire section, in Chapter 4, to replicating, analyzing, and enhancing PrecipNet’s outcomes. There, we will delve into the limitations and challenges inherent in utilizing the PrecipNet network.

Drawing from the various limitations identified by [Newell et al. \(2015\)](#) and [McGranaghan et al. \(2021\)](#), our approach aims to address needs, both using data and algorithms. Indeed, machine learning allows us to:

- Integrate a wide range of features and analyze them to identify the most relevant ones for better precipitation modeling.
- Integrate data with different temporal resolutions and measured from distinct locations (e.g., near-Earth geomagnetic indices and solar wind characteristics in the interplanetary medium).
- Utilize a variety of algorithms, including recurrent or convolutional neural networks, to tackle diverse challenges and even combine them synergistically.
- Fine-tune the model through carefully chosen hyperparameters tailored to our specific problem.

As we said, our work builds upon and runs parallel to the research done by [McGranaghan et al. \(2021\)](#), but covers two main objectives that are the core of this CIFRE (Convention industrielle de formation par la recherche) project. A CIFRE is a tight research collaboration between a laboratory and a French company, accomplished through the thesis of a doctoral candidate working within the company. In this case, the project serves two main purposes: first, it involves purely fundamental research objectives to tackle scientific inquiries, and second, it aims at product development (and subsequent commercialization) for SpaceAble.

- As part of our research project, our approach involves modeling the electron energy flux using a wide range of information, including solar wind parameters and geomagnetic indices. The research project’s importance lies in advancing the understanding of electron precipitation in LEO and of the magnetosphere-ionosphere system by developing a more robust and accurate model for electron energy flux. It also has practical applications in various fields, including mitigating the potential impacts on modern technological systems.
- As part of SpaceAble’s² objectives, our approach focuses on creating a highly adaptable and robust AI model with flexible inputs, outputs, and methods. Beyond modeling the electron total energy flux, we prioritize practicality, security, tangible value delivery, adaptability to new data over time, interpretability, explainability, optimization, and low latency. Cost-effectiveness is also paramount, ensuring a reasonable and justifiable investment for the company. This collaborative process necessitates a perfect understanding of every line of code, python file, and library used, aligning the AI model precisely with the company’s vision.

Hence, this thesis strives to be descriptive, as a thorough understanding is essential to distinguish the algorithmic aspects from problem-specific considerations. Yet, the algorithmic choices

2. <https://spaceable.org>

and modern tools raise broader questions about the application of machine learning in modeling and predicting phenomena along the Sun-Earth chain. How crucial is data quality for predictive capabilities? How should we tackle spatio-temporal challenges in satellite data? Can we effectively model highly dynamic phenomena intertwined with cyclical events? Is it even feasible to model such complex phenomena? This PhD also serves as a starting point for addressing some of these open questions.

In this chapter, you'll find a detailed analysis of the different data we want to compare. On one side, we have information from outside the magnetosphere (solar wind characteristics), and on the other side, we have measurements of particle precipitation spanning over 50 years (DMSP). Following that, we'll describe the final choices we made regarding the data and the preprocessing applied.

3.2 Data Description

3.2.1 Defense Meteorological Satellite Program (DMSP)

To begin with, we will introduce the satellite family of the Defense Meteorological Satellite Program (DMSP) and the measurements that our algorithm will utilize, specifically the data from the SSJ4 and SSJ5 instruments within the DMSP family. These data will serve as the "ground truth" or reference for our analysis. It's crucial to comprehend, analyze, and ascertain their applicability. A concise description of the DMSP satellites and the data used can also be found in Section 3.3.4 of the paper [Bouriat et al. \(2023\)](#). For the description of DMSP and its instruments, we drew upon information from the dedicated website of Boston College³ as well as the site of the NCEI (US National Centers for Environmental Information)⁴.

The DMSP (Defense Meteorological Satellite Program, [Dickinson \(1974\)](#); [Nichols \(1975\)](#)) spacecrafts are a series of LEO, polar-orbiting spacecraft whose primary mission is to observe the tropospheric weather (through the Operational Linescan System⁵). Its secondary mission is to monitor the space environment and mainly the in-situ plasma environment. DMSP satellites belong to the US DoD (Department of Defense), are managed by USAF (United States Air Force) and operated by the 6th Satellite Operations Group at Offutt AFB (Air Force Base), Nebraska. They are polar orbiting, Sun synchronous (i.e., with a fixed local time), axis stabilized (since DMSP 5A/F1) and have an orbital period of 101 min, an inclination of 98.9°, and an altitude of 840 km ([Redmon et al., 2017](#)).

Lots of studies use DMSP data to observe the space environment and its effects such as spacecraft charging. This is mainly done through the Space Environment Monitoring (SEM) sensors. Among them we can find:

- SSJ/4 & SSJ/5 (Special Sensor Precipitating Electron and Ion Spectrometer - 4 & 5)
- SSM (Special Sensor Microwave)
- SSIE (Special Sensor Imager Experiment) on the first two DMSP satellites and then SSIES (Special Sensor Imager for Environmental Sensing)

3. <https://dmsp.bc.edu/>

4. <https://www.ncei.noaa.gov/products/dmsp-j4-precipitating-electron-ion-spectrometer>

5. https://dmsp.bc.edu/html2/dmsp_ols.html

- SSULI (Special Sensor Ultraviolet Limb Imager)
- SSUSI (Special Sensor Ultraviolet Spectrographic Imager)

In our study, we are interested in SSJ/4 and 5 precipitating particle sensors. DMSP SSJ/4 and 5 data provide a complete energy spectrum of the low energy particles that cause the aurora and other high latitude phenomena. The data set consists of electron and ion particle fluxes between 30 eV and 30 KeV recorded every second, satellite ephemeris and magnetic coordinates where the particles are likely to be absorbed by the atmosphere.

3.2.1.1 SSJ/4

The Precipitation Electron/Proton Spectrometer, developed by Ampek Inc. based in Bedford, MA, is an advanced instrument designed to measure the transfer energy, mass, and momentum of charged particles as they pass through the magnetosphere-ionosphere within Earth's magnetic field. The SSJ/4 is a significant upgrade from its predecessor, the SSJ/3, and provides valuable data for missions that require a thorough understanding of the polar and high-latitude ionosphere. Applications of this data, as we have seen in Chapter 1, include enhancing communication systems, surveillance operations, and detection systems that rely on energy propagation either off or through the ionosphere.

Mounted on the Flight 6 (F6) to F15 satellites, the SSJ/4 utilizes 20 energy channels logarithmically distributed from 30 eV to 30 keV (see Figure 3.2), employing four cylindrical curved plate electrostatic analyzers arranged in two pairs. Each pair consists of an analyzer with a radius of curvature of 127° and another with a radius of curvature of 60° . These analyzers operate at varying voltage levels, with the 127° (60°) analyzers capturing particles within energy channels ranging from 1 keV to 30 keV (30 keV to 1 keV).

To ensure stability, the detector remains in each energy channel for 98 milliseconds, with a 2-millisecond interval between steps to stabilize the voltage. The two analyzers are synchronized, enabling the retrieval of a comprehensive 20-point ion and electron spectrum every second. Detailed illustrations of the SSJ/4 analyzer's aperture and curved plate configuration are provided in Figure 3.1.

While the analyzers for ions and electrons are identical in design, the voltage polarity on the plates and the size of the low-energy apertures differ. The low-energy ion apertures are larger compared to the low-energy electron apertures. Compact and efficient, the SSJ/4 sensor package weighs only 5 pounds and consumes a mere 0.25 watts of power.

In order to improve measurement consistency, the CEMs employed in the refurbished F8 sensors differ from those utilized in the DMSP series F6 – F10 sensors described in [Hardy et al. \(1985\)](#). The new design replaces the previous two overlapping CEM cones with a single CEM featuring a larger collecting cone. This modification ensures a more consistent coverage of the effective particle collecting area behind the detector's exit aperture.

3.2.1.2 SSJ/5

SSJ/5 is an electrostatic analyzer detector developed and constructed by Amptek Inc. Its initial deployment occurred on the DMSP F16 satellite, launched on October 18, 2003. The SSJ/5 represents an enhanced version of the SSJ/4 instrument. Just like the SSJ/4, the collected data aids in missions that necessitate an understanding of the polar and high-latitude ionosphere and supports

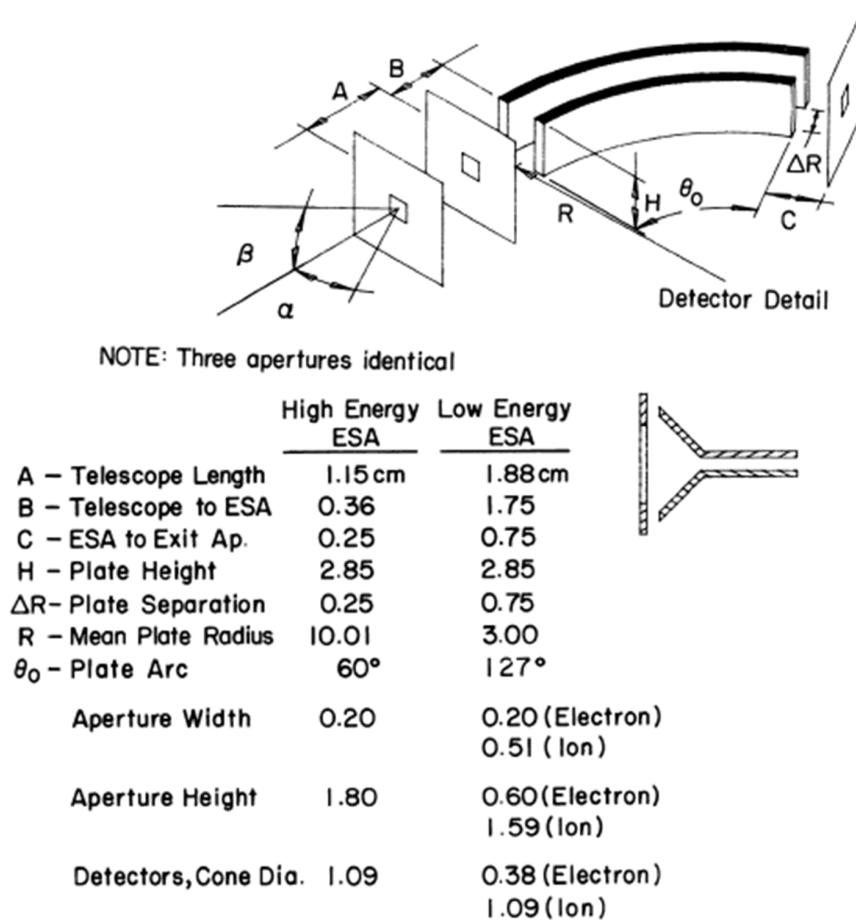


Figure 3.1 – Details of the configuration of the aperture and curved plates of an SSJ/4 Electrostatic Analyzer (ESA), from Schumaker (1988).

Channel	E_{peak} (eV)	ΔE (eV)	$G(E)_{\text{peak}}$ ($\text{cm}^2 \text{ster}$)	G ($\text{cm}^2 \text{ster eV}$)
1	31,300	3050	1.70×10^{-4}	5.35×10^{-1}
2	21,100	2000	2.30×10^{-4}	4.60×10^{-1}
3 [*]	14,300	1330	2.65×10^{-4}	3.60×10^{-1}
4	9720	860	3.10×10^{-4}	2.65×10^{-1}
5 [*]	6610	615	3.80×10^{-4}	2.30×10^{-1}
6	4500	430	4.50×10^{-4}	1.85×10^{-1}
7 [*]	3050	284	5.35×10^{-4}	1.50×10^{-1}
8	2070	184	6.70×10^{-4}	1.25×10^{-1}
9	1400	125	7.25×10^{-4}	9.30×10^{-2}
10	950	88	9.05×10^{-4}	7.90×10^{-2}
11	950	85	7.10×10^{-4}	5.60×10^{-2}
12	640	63	7.10×10^{-4}	4.15×10^{-2}
13	440	42	7.10×10^{-4}	2.60×10^{-2}
14 [*]	310	29	7.10×10^{-4}	1.85×10^{-2}
15	210	20	7.10×10^{-4}	1.35×10^{-2}
16 [*]	144	13	7.10×10^{-4}	8.30×10^{-3}
17 [*]	98	9.1	7.10×10^{-4}	5.80×10^{-3}
18 [*]	68	6.3	7.10×10^{-4}	4.00×10^{-3}
19 [*]	45	4.2	7.10×10^{-4}	2.70×10^{-3}
20 [*]	31	2.9	7.10×10^{-4}	1.85×10^{-3}

* Channel response curve not determined by electron beam

Figure 3.2 – Adjusted channel response characteristics of the SSJ/4 F8 electron detector, taken as an example to understand energy channels in the SSJ/4 instrument, from Schumaker (1988).

various applications, that rely on energy propagation off or through the ionosphere.

The primary objective of the instrument is the same: detect and analyze electrons and ions that precipitate in the ionosphere, ultimately leading to the generation of an aurora display. It also operates in the low-energy range of 0.3 to 30 keV.

To achieve its goals, the instrument utilizes a nested spherical deflection plate system, enabling simultaneous analysis of electrons and ions across a 90° field of view (see Figure 3.3). Furthermore, it incorporates a space-qualified microprocessor, facilitating customization of data rates, measurement ranges, on-board storage, and specific analysis algorithms. These algorithms can include tasks such as auroral boundary detection or real-time charging measurements.

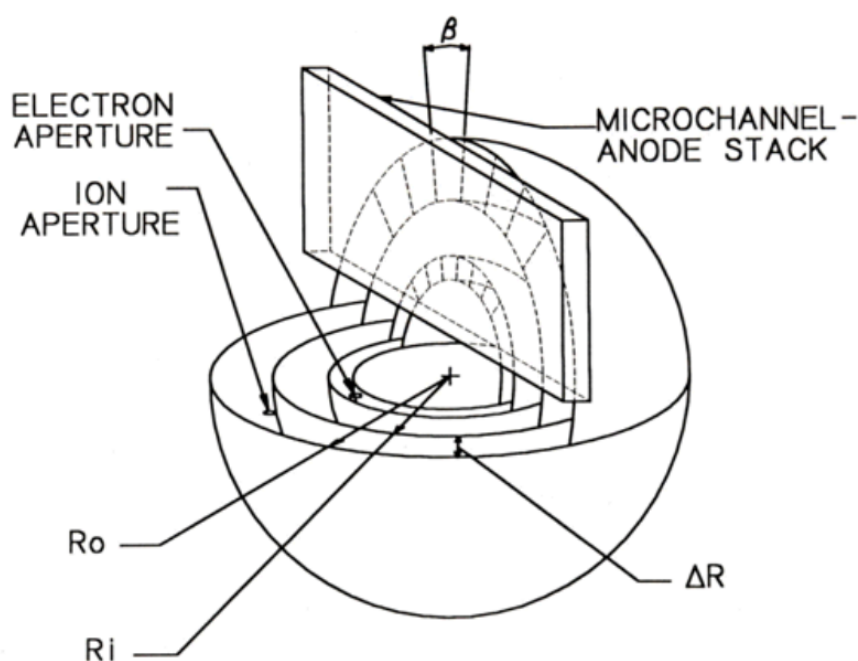


Figure 3.3 – Details of the configuration of the SSJ/5 Electrostatic Analyzer (ESA) field of view. Taken from Boston College DMSP website, https://dmsp.b.c.edu/html2/ssj5_inst.html, last accessed September 2023.

3.2.1.3 Precipitating Electron Data

Typically, the data obtained from precipitating electrons (and ions) can be effectively visualized using a color spectrogram, as depicted in the example shown in Figure 3.4. This spectrogram provides a graphical representation of the differential number flux, which represents the number of particles falling onto a square centimeter per second within each steradian of solid angle per eV of energy range. The spectrogram displays this flux in relation to both the energy of the particles and the corresponding time of particle observation. By presenting the data in this format, patterns and trends in the particle behavior can be easily identified and analyzed.

In order to obtain a comprehensive understanding of the data collected by the instrument, it is often beneficial to aggregate or integrate the data across the entire energy range covered by the instrument and this is what we are doing in this study. This integration allows for the calculation

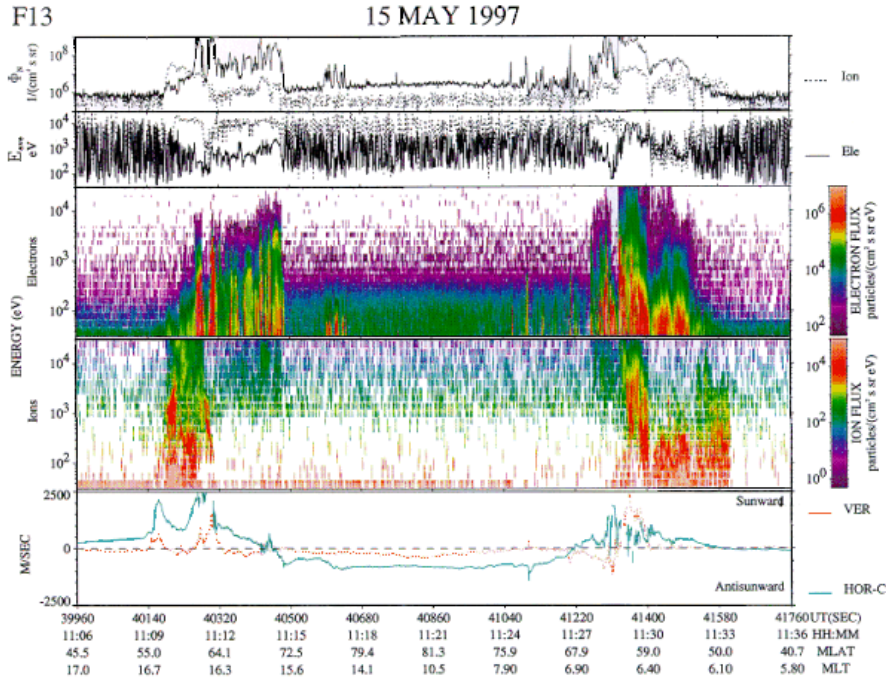


Figure 3.4 – Typical color spectrogram of the SSJ4 data obtained in one polar pass of a DMSP F13 spacecraft during the crossing of the northern polar region on 97 May 15 during a geomagnetic storm. Credits: Boston College website, http://dm.sp.b.c.edu/html2/dmspssj4_data.html, last accessed in July 2023.

of three key parameters: average energy (E_{AVE}), total energy flux (J_{ETOT}), and total number flux (J_{TOT}). The formulas used to calculate J_{TOT} and J_{ETOT} are as follows:

$$J_{TOT} = \sum_{i=1}^{19} J(E_i) \frac{E_{i-1} - E_{i+1}}{2}$$

$$J_{ETOT} = \sum_{i=1}^{19} E_i J(E_i) \frac{E_{i-1} - E_{i+1}}{2}$$

$$E_{AVE} = \frac{J_{ETOT}}{J_{TOT}}$$

In these equations, $J(E_i)$ represents the differential number flux for the channel with a central energy of E_i . J_{TOT} is measured in units of particles per cm^2 per steradian per second, while J_{ETOT} is measured in units of eV per cm^2 per steradian per second. Note that the integration is performed across only 19 channels (although 20 channels can be seen on Figure 3.2). This is because channels 10 and 11 of the instrument are set at the same energy peak (950 eV), and the data at this energy should be used only once to avoid duplication. The average energy, E_{AVE} , is measured in units of eV.

As we explained in Section 3.1, we will only study here the **electron total energy flux** J_{ETOT} , but our code is made such that it is also possible to change the input to the **electron average energy**.

3.2.2 Advanced Composition Explorer (ACE)

The Advanced Composition Explorer (Stone et al., 1998) satellite observes and measures magnetic fields and particles in space from the L1 (Lagrange 1) libration point (Farquhar, 1969) be-

tween the Earth and the Sun, about 1,500,000 km forward of Earth. Launched in 1997, ACE carries nine instruments seen Figure 3.5: six high-resolution spectrometers measuring the elemental, isotopic, and ionic charge-state composition of nuclei (within the solar wind and the galactic cosmic rays) and three instruments that provide the heliospheric context. We are introducing all the instruments below. ACE records different types of radiation, including bursts of particles from the Sun that can potentially impact near-Earth space and cause disruptions to satellites, radio communication, and navigation systems. ACE detects these bursts 20 to 60 minutes before they reach Earth, allowing us to assess their strength and prepare for potential impacts. The spacecraft's observations are still utilized by the US NOAA's SWPC to give advance warning of geomagnetic storms⁶.

ACE is a compact, eight-faceted cylindrical satellite, measuring 1.6 meters in diameter and 1 meter in height. It weighs 785 kg, with 156 kg dedicated to scientific instruments and 185 kg of hydrazine fuel for orbital insertion and maintenance. The satellite boasts four solar panels attached to its upper deck, while six scientific instruments are strategically positioned around it to ensure an unobstructed field of view. Additionally, two more instruments are fixed to its sides, and the magnetometer masts extend from two of the solar panels. ACE maintains a constant alignment with the line connecting the Sun and Earth, keeping its upper deck pointed towards the Sun while gently rotating around its vertical axis to maintain orientation. Communication with Earth is facilitated through the S-band, allowing for data transmission rates of 7 kilobits in real-time and 78 kilobits in deferred mode. Data can be conveniently stored in a 2-gigabit mass memory (Margolies and von Rosenvinge, 1998).

The instruments on ACE are the following:

- Six spectrometers all optimized for a given energy range:
 - CRIS (Cosmic Ray Isotope Spectrometer) measures the isotopic composition of cosmic rays, ranging from helium to nickel with energies of 100 to 600 MeV per nucleon.
 - SIS (Solar Isotope Spectrometer) identifies and measures the isotopic composition of the same atomic nuclei with energies of 10 to 100 MeV. It detects particles emitted by the Sun during solar storms, capturing their trajectory and energy.
 - ULEIS (Ultra Low Energy Isotope Spectrometer) idem with energies ranging from 45 keV to a few MeV per nucleon. It can also detect heavier ions with energies around 0.5 MeV per nucleon.
 - SEPICA (Solar Energetic Particle Ionic Charge Analyzer) determines the electric charge, element type, and energy of ions emitted by the Sun with energies from 0.5 to 10 MeV per nucleon.
 - SWIMS (Solar Wind Ion Mass Spectrometer) measures the composition and speed of the solar wind.
 - SWICS (Solar Wind Ion Composition Spectrometer) measures the charge, temperature, and speed of ions in the solar wind, ranging from 145 km/s (protons) to 1,532 km/s (iron).
- Three standard instrument for the context of the interplanetary medium. The description here come from the OSCAR⁷ (Observing Systems Capability Analysis and Review Tool) database:
 - SWEPAM (Solar Wind Electron, Proton, and Alpha Monitor): SWEPAM serves the purpose of measuring electron and ion fluxes in the low energy solar wind range.

6. <https://www.swpc.noaa.gov/products/ace-real-time-solar-wind>

7. <https://space.oscar.wmo.int/>

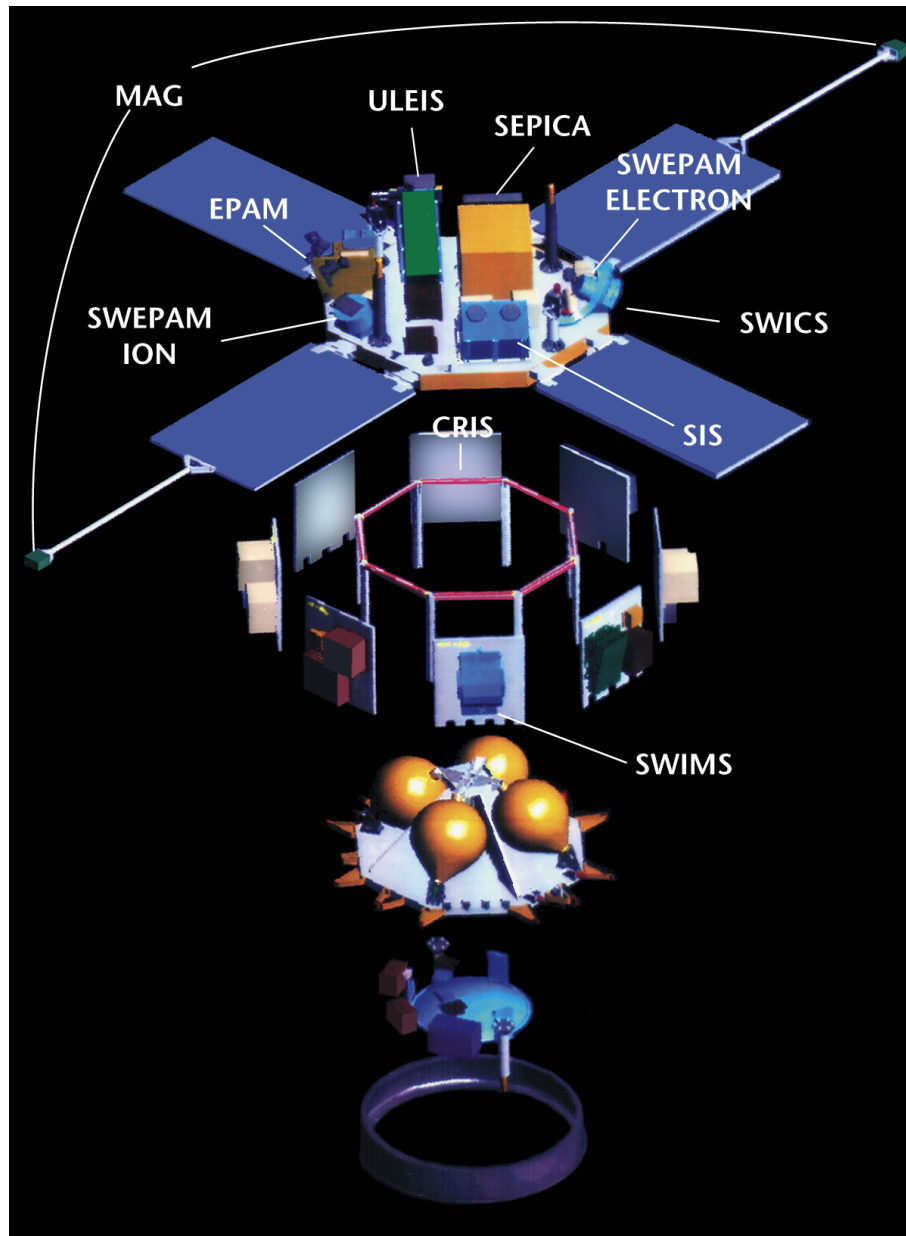


Figure 3.5 – Schematic details of the ACE satellite and its instruments (image credit: NASA, accessed through the ESA's eoportal, <https://www.eoportal.org/satellite-missions/ace#rf-communications>, last accessed September 2023).

With two separate sensors, it conducts simultaneous and independent electron and ion spectroscopy. The instrument covers a wide energy range, detecting electrons with energies from 0.0016 to 1.35 keV, protons with energies from 0.26 to 36 keV, and alpha-particles with energies from 0.52 to 72 keV. Positioned to view the Sun from the L1 Lagrange libration point, SWEPAM continuously observes the full solar disk, providing valuable data for solar observation.

- EPAM (Electron, Proton, and Alpha-particle Monitor): EPAM is designed to measure the flux and energy of protons, alpha particles, and electrons emitted during solar storms. It can process particles with energies ranging from 30 keV to 4 MeV per nucleon for elements from helium to iron.
- MAG (Magnetic Field Monitor) in ACE: The MAG's primary function is to measure the three components of the solar magnetic field. This instrument includes twin, triaxial fluxgate magnetometers mounted on a boom. With eight dynamic ranges ranging from ± 4 nT to $\pm 65,536$ nT, MAG can precisely measure the solar magnetic field's strength. Situated at the L1 Lagrange libration point and pointed towards the Sun, MAG continuously scans the full solar disk at intervals of 0.09 seconds, providing essential data for solar observation and magnetic field analysis.

As part of this study, we specifically narrowed our focus to the solar wind data (i.e., its speed, temperature, and density) and the components of the interplanetary magnetic field (i.e., the components along X, Y, and Z, as well as the overall magnitude of the magnetic field). Consequently, the only instruments we considered for this analysis are SWEPAM for the solar wind parameters and MAG for the IMF components.

3.2.3 High-Resolution OMNIWeb

OMNIWeb⁸ is a WWW-based system that enables users to generate plots and access data seamlessly, facilitating researchers in identifying trends and obtaining data efficiently (Mathews and Towheed, 1995). It is updated regularly with new data and is widely used in the heliospheric community (Papitashvili et al., 2014). Today, it contains two categories of spacecraft-interspersed, near-Earth solar wind data according to the OMNI Website descriptions:

- Low resolution OMNI (LRO) is Hourly "Near-Earth" solar wind magnetic field and plasma data, energetic proton fluxes (>1 to >60 MeV), and geomagnetic and solar activity indices. There are Daily, 27-day, and Yearly resolution derived from hourly data also. All details about that data descriptions and about the data access can be found at <https://omniweb.gsfc.nasa.gov/ow.html>.
- High 1-min and 5-min resolution OMNI (HRO) : Solar wind magnetic field and plasma data at Earth's Bow Shock Nose, and geomagnetic activity indices. 5-min resolution derived from 1-min data with added 5-min energetic proton fluxes. All details about high resolution OMNI data descriptions and about data access can be found at https://omniweb.gsfc.nasa.gov/ow_min.html

OMNI provides solar wind parameters at the Earth's Bow Shock Nose (BSN) using measurements from L1 King and Papitashvili (2006). To do that, it combines and propagates data from different spacecraft such as ACE, Wind, IMP 8, and Geotail, to the Earth's bow shock nose. To shift the data, it considers the magnetic field "frozen" in the plasma. The assumption made is that solar wind magnetic field values observed by a spacecraft at a particular time and location form a planar surface (phase front) that moves along straight wave fronts at the plasma velocity between L1 and the BSN. The Earth's orbital motion around the Sun is also taken into account. This is used

8. <https://omniweb.gsfc.nasa.gov/>

to calculate the time difference between when an parameter is observed at its original location (L1) and when it arrives at the target location (BSN). This propagation time is determined through the formula given in Equation 3.1 (King and Papitashvili, 2006), where \vec{n} represents the wave front normal, and \vec{R}_d, \vec{R}_O indicate the positions of the upstream measuring satellite and the bow shock nose, while \vec{V} denotes the plasma velocity.

$$\Delta t = \frac{\vec{n} \cdot (\vec{R}_d - \vec{R}_O)}{\vec{n} \cdot \vec{V}} \quad (3.1)$$

In this context, we are only interested in high-resolution data. As we'll look into the data analysis, some difficulties arise with ACE satellite data, which led us to consider the data from OMNIWeb only. However, OMNI high resolution data also face some difficulties and limitations that will be explained.

3.3 Data Analysis & Preprocessing

3.3.1 ACE

To begin with, we'll delve into the data analysis of the ACE satellite. This is because the in-depth analysis carried out in the article Bouriati et al. (2022), which follows, sets the groundwork for sound data analysis practices, and its conclusions will be further utilized in the subsequent sections.

In the paper, we present an analysis of the Level-2 ACE SWEPAM and MAG measurements from 1998 to 2021 by the ACE Science Center. The study focuses on the challenges and potential issues encountered in this dataset, widely used within the Space Weather community. The paper addresses the data's quality and the impact it has on artificial intelligence models, exploring issues like non-uniform distributions in histograms, data reproducibility, rounding errors, missing values, and the presence of non-linear relationships. The findings provide valuable insights and offer suggestions to overcome these challenges, enabling the dataset's effective usage despite its complexities. The objective behind such paper is also to give an overview on how a data analysis should be conducted.



OPEN ACCESS

EDITED BY

Fadil Inceoglu,
National Centers for Environmental
Information (NCEI) at National Atmospheric
and Oceanographic Administration
(NOAA), United States

REVIEWED BY

Reinaldo Roberto Rosa,
National Institute of Space Research (INPE),
Brazil
Amy Keesee,
University of New Hampshire, United States

*CORRESPONDENCE

S. Bouriat,
simon.bouriat@spaceable.org

SPECIALTY SECTION

This article was submitted to Astrostatistics,
a section of the journal Frontiers in
Astronomy and Space Sciences

RECEIVED 28 June 2022

ACCEPTED 04 October 2022

PUBLISHED 23 November 2022

CITATION

Bouriat S, Vandame P, Barthélémy M and
Chanussot J (2022), Towards an AI-based
understanding of the solar wind: A critical
data analysis of ACE data.
Front. Astron. Space Sci. 9:980759.
doi: 10.3389/fspas.2022.980759

COPYRIGHT

© 2022 Bouriat, Vandame, Barthélémy and
Chanussot. This is an open-access article
distributed under the terms of the [Creative
Commons Attribution License \(CC BY\)](#). The
use, distribution or reproduction in other
forums is permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original publication in
this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

Towards an AI-based understanding of the solar wind: A critical data analysis of ACE data

S. Bouriat^{1,2,3,4*}, P. Vandame³, M. Barthélémy^{1,2} and J. Chanussot³

¹CNRS, IPAG, University of Grenoble Alpes, Grenoble, France, ²CSUG, University of Grenoble Alpes, Grenoble, France, ³GIPSA-Lab, Grenoble INP, CNRS, University of Grenoble Alpes, Grenoble, France, ⁴SpaceAble, Paris, France

All artificial intelligence models today require preprocessed and cleaned data to work properly. This crucial step depends on the quality of the data analysis being done. The Space Weather community increased its use of AI in the past few years, but a thorough data analysis addressing all the potential issues is not always performed beforehand. Here is an analysis of a largely used dataset: Level-2 Advanced Composition Explorer's SWEPAM and MAG measurements from 1998 to 2021 by the ACE Science Center. This work contains guidelines and highlights issues in the ACE data that are likely to be found in other space weather datasets: missing values, inconsistency in distributions, hidden information in statistics, etc. Amongst all specificities of this data, the following can seriously impact the use of algorithms:

Histograms are not uniform distributions at all, but sometime Gaussian or Laplacian. Algorithms will be inconsistent in the learning samples as some rare cases will be underrepresented. Gaussian distributions could be overly brought by Gaussian noise from measurements and the signal-to-noise ratio is difficult to estimate.

Models will not be reproducible from year to year due to high changes in histograms over time. This high dependence on the solar cycle suggests that one should have at least 11 consecutive years of data to train the algorithm.

Rounding of ion temperatures values to different orders of magnitude throughout the data, (probably due to a fixed number of bits on which measurements are coded) will bias the model by wrongly over-representing or under-representing some values.

There is an extensive number of missing values (e.g., 41.59% for ion density) that cannot be implemented without pre-processing. Each possible pre-processing is different and subjective depending on one's underlying objectives

A linear model will not be able to accurately model the data. Our linear analysis (e.g., PCA), struggles to explain the data and their relationships. However, non-linear relationships between data seem to exist.

Data seem cyclic: we witness the apparition of the solar cycle and the synodic rotation period of the Sun when looking at autocorrelations.

Some suggestions are given to address the issues described to enable usage of the dataset despite these challenges.

KEYWORDS

data analysis, solar wind, MAG, SWEPM, machine learning, ACE

1 Introduction

The space weather community aims to understand and quantify the associated threats, mitigate them, and in the best cases, prevent them altogether. Recently, [Daglis et al. \(2020\)](#) detailed a new scientific program of the Scientific Committee on Solar-Terrestrial Physics (COSTEP) called PRESTO, for Predictability of the variable Solar-Terrestrial coupling. Such a study highlighted the remaining questions surrounding the understanding of the Sun-Earth coupling. Among these open questions, we can find:

- How do various solar wind conditions (e.g., IMF components, speed, density, level of turbulence) and different large-scale drivers control the coupling efficiency and the energy/mass transfer from the solar wind to the magnetosphere?
- How do solar wind conditions control the occurrence frequency and location of different magnetospheric plasma waves?

These questions emphasize the role of the solar wind as it is indeed one of the key issues in the predictability of the Earth space environment. Studies to better understand both solar wind and the interplanetary magnetic field using coordinated space- and ground-based data along with models are of essential importance. Recently, the emergence of machine learning algorithms in space weather [Camporeale et al. \(2018\)](#), [Camporeale \(2019\)](#), [Camporeale \(2020\)](#) appeared as one of the most promising solutions to nowcast and forecast phenomena in space weather. More and more papers using machine learning, and especially deep learning, are published in the field of space weather [Reiss et al. \(2021\)](#), [Zewdie et al. \(2021\)](#), [Stumpo et al. \(2021\)](#), [Reep and Barnes \(2021\)](#).

Initiated in 2018, a Research Coordination Network (RCN) supported by the National Science Foundation (NSF) named “Towards Integration of Heliophysics Data, Modelling, and Analysis Tools” (@HDMIEC) planned to make progress in the understanding of physical mechanisms in the Sun and on modelling and the data accessibility and analysis. In this regard, workshops, and discussions around the topic

of Machine Learning in Space Weather were held and the opinion of the community was gathered. Several outcomes from the Q&A sessions are worth to be noticed from [Nita et al. \(2020\)](#):

- Half of the attendees (46.7%) agreed that the heliophysics community does not even have a fair understanding of machine learning capabilities and limitations.
- There was a consensus that cooperation between ML and heliophysics does not exist.
- ML methods are more successful regarding the Big Data environment behind heliophysics than physics-based methods. But there is no consensus around which areas could ML methods outperform physics-based ones.
- The overwhelming majority of attendees strongly agreed (73.3%) that there is a need to combine physics-based and ML models.
- Most of the attendees did not feel that the ML was a “bubble” ready to burst.

In this paper, we decided to discuss the use of solar wind data in the context of artificial intelligence. Firstly, because the solar wind is a central data as seen through PRESTO. Secondly, because most of the space weather community is not so familiar with AI and its good practices but seems ready to use it more in the future [Nita et al. \(2020\)](#). Hence, we present here a complete data analysis of the ACE solar wind and IMF measurements, an essential and largely used data when forecasting on-Earth events, even today ([Myagkova et al. \(2020\)](#), [Wintoft et al. \(2015\)](#), *etc.*). While we will not expand on this in this paper, it is interesting to notice that a lot of studies use the NASA’s OMNIWeb dataset (see https://omniweb.gsfc.nasa.gov/html/ow_data.html) such as [Wihayati et al. \(2021\)](#) or [Gombosi et al. \(2018\)](#) for instance. High-Resolution OMNIWeb data are made of ACE, IMP 8, Wind and Geotail satellites data gathered and time-shifted to the Bow Shock Nose. Although they are really interesting data, we did not want to add here any complexity through the fact that this time-shifting was based on several assumptions and needed an intercalibration between satellites. This data preparation is largely documented on OMNIWeb (<https://omniweb.gsfc.nasa.gov/html/HROdocum.html>).

These kinds of analyses are “required to correct for scattering, baselines changes, peak shifts, noises, missing values and several other artefacts so that the “true” relevant underlying structure can be highlighted and/or, if required, the property of interest can be predicted correctly” [Mishra et al. \(2020\)](#). The chosen data go from 1998 to 2021, including a large part of the 23rd and the full 24th solar cycles (for a schematic view, see the [Supplementary Figure S1](#), showing the Solar radio flux index at 10.7 cm, a good representation of the solar activity). The objective of this paper is to extract all possible useful information that can be found in solar wind data and highlight the issues that could arise when applying machine learning algorithms and techniques.

Before diving into the subject, it is worth noticing that impressive work has been done by [Smith et al. \(2022\)](#) on a similar topic. Their paper consists of an analysis of the quality and continuity of the data that are available in Near-Real-Time from the Advanced Composition Explorer and Deep Space Climate Observatory (DSCOVR) spacecraft. Part five (Discussion and Conclusion) of our work details how our two studies differ.

2 A quick introduction to machine learning concepts

In order to better understand the data analysis presented here, we first need to quickly introduce some concepts in Machine Learning. According to Oxford Dictionary, Machine Learning is “the use and development of computer systems that are able to learn and adapt without following explicit instructions, by using algorithms and statistical models to analyse and draw inferences from patterns in data”. The products include models, forecasts, identification of patterns, anomalies, or even relationships among data. Machine learning is usually described through two categories of algorithms: *supervised learning* and *unsupervised learning* (although it exists a weakly-supervised algorithm that embraces both ideas).

- Supervised learning [LeCun et al. \(2015\)](#) includes *regression algorithms* and *classification algorithms*. The first aims at discovering connections between input and output data and is often employed to approximate functions or predict future values of continuous functions. It comprises linear regressions, decision trees, most of neural networks, and ensemble methods, among others. The second aims at mapping input data to classes and is therefore usually employed to classify data (e.g., True/False problems). It includes support vector machines (SVM), discriminant analysis, naïve Bayes classifiers, and K-nearest neighbour, among others. Such algorithms are called *supervised* because

they require to be fed examples of output data to train.

- Unsupervised Learning [Storrs et al. \(2021\)](#) includes *clustering algorithms*, which group data together. Clustering algorithms map input data to a set of categories initially identified by the system (e.g., Gaussian Mixture Model, K-mean). They are referred to as *unsupervised* because one does not know their output. A simple example is the grouping of customer profiles, where the final quantity of groups is unknown at first.

The procedure is often the same: access, scrap, analyse and pre-process the data (format, missing values, *etc.*), choose and compute the features that will be used for the model and train the model thanks to a loss function taking or not the label into account (subjective choice from the user). We then iterate the last three steps to find the best model. As introduced in this paper, the key and most time-consuming parts are the data analysis and the pre-processing, as we need to build a scalable and efficient ready-to-use dataset to answer a given problem. The pre-process ends with a split of our data into three groups:

- The train set: a dataset that will be used by our algorithm to train itself.
- The validation set: a dataset used by the algorithm to test itself. The accuracy of the model on this dataset allows the user to see how good it is at predicting, how fast and how well it is training and allows him to make some changes accordingly.
- The test set: a dataset that will never be used by the user and the model until the very last moment. The user applies his trained and ready-to-use model to this dataset to ultimately know the final accuracy of his model and to avoid a human/user bias from hyperparameter tuning.

Finally, it is worth noticing that the train and validation sets have to be “well-balanced”. This means that all possible cases should appear in both datasets and, ideally, in the same quantities. A model will easily get used to recurrent cases. If we only have one or two samples of fast solar wind in our train set, and thousands of slow solar wind samples, the model will not be able to accurately predict fast-solar wind cases. One possibility to address this issue is to perform what we call *data augmentation*, but we will not expand it here [Shorten and Khoshgoftaar \(2019\)](#), [Chen et al. \(2020\)](#).

3 Data description

The Advanced Composition Explorer (ACE) satellite is located in the Lagrange point 1 (L1), a stable point in space, between the Earth and the Sun, where the gravitational attraction

from both bodies and the centrifugal force all balance each other. Satellites at this location are at the front line to see phenomena coming from the Sun.

ACE Solar Wind Data are Level-2 Real-Time Solar Wind (RTSW) data. “Level 2” means that raw data from the instruments have been processed by the instrument teams. According to the Ace Science Center Level 2, it includes such operations as calibration, organization into energy and time bins, or application of ancillary data. The frequency of measurements for instruments MAG (Magnetometer) and SWEAPAM (Solar Wind Electron Proton Alpha Monitor) are respectively 16-s and 64-s, from 1998 to 2020. The data have been gathered from the following link: srl.caltech.edu/ACE/ASC/level2/where they are considered to be official and verified¹. A lot of research needing solar wind data also uses OMNIWeb 1-min and 5-min solar wind datasets mathematically time-shifted from the Lagrange one point to the Earth’s bow shock nose [King and Papitashvili \(2006\)](#). Choosing these manually propagated data as input to nowcast or forecast near-Earth data [Shprits et al. \(2019\)](#), [McGranaghan et al. \(2021\)](#), [Bentley et al. \(2018\)](#) is a good idea to prevent a machine-learning-made propagation which can be subject to unidentified errors. However, as the point of this paper is to highlight the dangers of using *in situ* data, it was more relevant to take *in situ* solar wind values.

We focus on the data from two main instruments of the ACE satellite:

- SWEAPAM [McComas et al. \(1998\)](#), for Solar Wind Electron, Proton and Alpha Monitor measures rates of electron and ion flows with two distinct electrostatic analyzers with fan-shaped fields of view that use the spacecraft’s rotation to observe in all directions. The first one observes electrons in the 1 eV–1.35 keV energy range and the second one ions in the 0.26–36 keV energy range. For this instrument, we only focus on ion data, spanning 23 years from 1998 to 2020 with a 64-s resolution. This corresponds to 11, 299, 710 measurements.
- MAG, for Magnetic Field Monitor, consists of a set of twin sensors (triaxial fluxgate magnetometers, [Stone et al. \(1998\)](#)) measuring the three components of the interplanetary magnetic field at L1. For this instrument, we have 25 years of 16-s data from 1997 to 2021. We removed the years 1997 and 2021 to have the same time range as the SWEAPAM instrument. In the end, we have 45, 365, 393 data points for this instrument. For the first part of our analysis, we decided to subsample the dataset every 64 s. With both years removed, we obtain 11, 341, 349 measurements. However, the corresponding times of each

sample do not correspond to SWEAPAM’s ones, and another post-process (presented further) had to be done to compare data between the two instruments.

Here are the analyzed *in situ* measurements and their unit:

- IMF X, Y and Z-component, GSE coordinates [nT]
- Solar wind proton density [cm^{-3}]
- Solar wind proton speed [km.s^{-1}]
- Solar wind ion temperature [K]

For the interplanetary magnetic field, X, Y and Z-component are in the GSE (Geocentric Solar Ecliptic) coordinates instead of the GSM. By definition [Russell \(1971\)](#) the X-axis points from the Earth towards the Sun, the Y-axis is chosen to be in the ecliptic plane opposing the planetary motion, and the Z-axis is parallel to the ecliptic pole. This system has been chosen instead of GSM because the aberration of the solar wind due to the orbital motion of Earth around the Sun representing a 30 km/s vector oriented in the minus Y direction axis is easier to remove [Russell \(1971\)](#). According to [Russell \(1971\)](#), GSE coordinates have been widely used to display satellite trajectories, interplanetary magnetic field observations, and solar wind velocity data.

4 Data analysis

In this part, we present the full analysis along with related conclusions.

- In the first part, statistical distributions of the data are plotted and explained, and every variable will be looked at independently of others. In all the datasets, there are some missing values that perturb the statistics computations. We removed all of them in this first part.
- The second part is an example of how to handle the aforementioned missing and extreme values.
- The third part will study the classical linear relationships between the different variables. Aside from being important to better understand our data, it is worth reminding that too many intercorrelated input features may give redundant information to an AI algorithm and then lower its performance. The topic of interdependencies in solar wind data has already been looked at in the literature (e.g., in [Bentley et al. \(2018\)](#)) but will be done here in the light of neural networks and deep learning.

4.1 Linear analysis of the IMF, and Plasma’s parameters

Before studying neural networks, it is important to begin with a simple linear analysis. These analysis allow to reveal

¹ A special thanks to Andrew Davis from the ACE Science Center for his answers and advice on the use of data.

TABLE 1 Mean, median, 0.005th and 99.995th percentile from ACE MAG and SWEPAM data.

Variables	Mean	Median	0.005th percentile	99.995th percentile	% Of missing data
Bx (GSE) [nT]	6.93×10^{-2}	8.4×10^{-2}	-36.6	25.5	0.128
By (GSE) [nT]	2.98×10^{-2}	-9.00×10^{-3}	-30.7	38.7	0.128
Bz (GSE) [nT]	9.34×10^{-3}	2.20×10^{-2}	-43.5	32.3	0.128
Bt (GSE) [nT]	5.76	5.04	0.32	54.5	0.128
Proton density [p/cc]	5.88	4.54	0.1	80.0	41.59
Proton speed [km/s]	4.30×10^2	4.08×10^2	2.38×10^2	1.03×10^3	6.80
Ion temperature [K]	9.20×10^4	7.05×10^4	2.84×10^3	1.00×10^6	20.10

some important information and features about data with simple computations which will help you save a lot of time during the deep learning study.

Observing various parameters of the Solar Wind and the interplanetary field gives us a good insight into their nature. The first step is to look at their histogram and statistical parameters such as mean, median, maximum or the standard deviation, globally, yearly and potentially in a shorter time period. Both solar wind and the IMF are influenced by the solar activity which evolves on 11-year cycles. Recall that all the statistics in this part are computed on **non-missing values** only.

It is essential to understand how values can fluctuate, evolve, or change in time when we are dealing with time series. The following **Table 1** highlights two interesting things: the great number of missing values in SWEPAM data, and the large distance between the 99.995th percentiles and mean values for Bt and the Ion Temperature. Such spread values seem dangerous to implement in a deep-learning algorithm without a pre-process.

Figure 1A shows the yearly standard deviation of the three components of the IMF. It is a direct witness to the obvious dependencies of some of our parameters over the solar cycle because it follows the global trend of the solar activity index F10.7 throughout the year. All possible figures to detect dependence of distribution parameters over time have been plotted. Only some of them are shown in this paper.

Figure 1B is another example of how values can change over time and shows that the evolution of the yearly average temperature and speed of the solar wind already suggests a dependence between the two. In other words, different periods in our dataset imply different distributions depending on the solar activity. Although it may seem obvious for a space weather expert, it is information of prime importance for the data scientist dealing with these datasets. Such observations suggest that the solar activity in the name of F10.7 has to be part of the inputs as we will have to know where we are in the solar cycle. Moreover, this highlights the need to have at least one full solar cycle in our training set to span all possible cases.

4.1.1 Histograms

Distributions are essential for the data scientist to assess the information contained in a dataset. For instance,

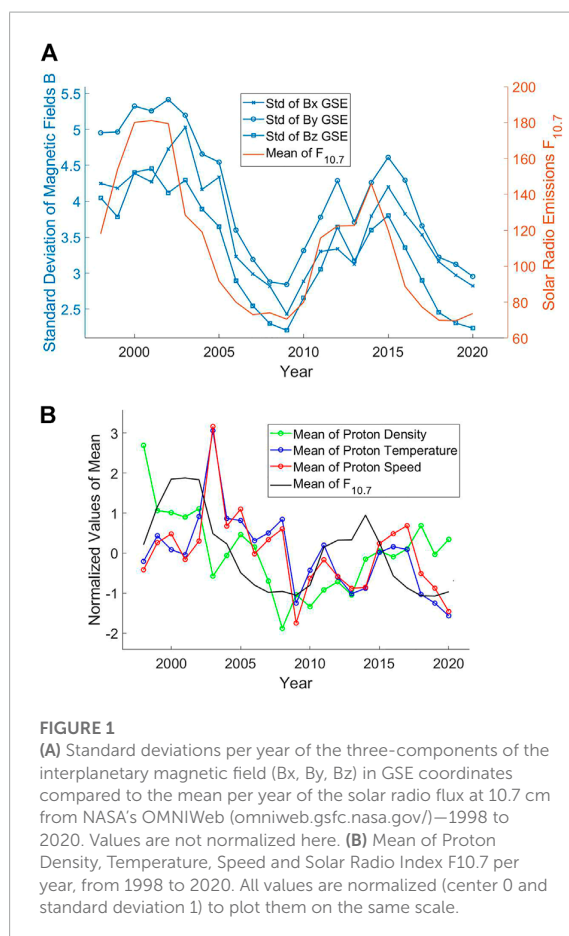
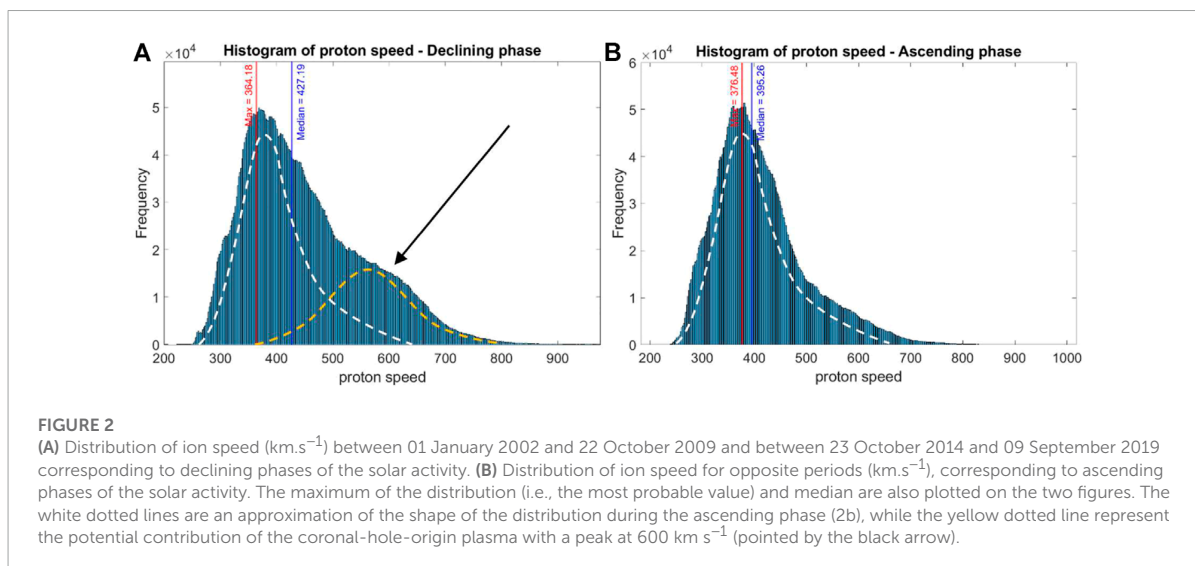


FIGURE 1 (A) Standard deviations per year of the three-components of the interplanetary magnetic field (Bx, By, Bz) in GSE coordinates compared to the mean per year of the solar radio flux at 10.7 cm from NASA's OMNIWeb (omniweb.gsfc.nasa.gov/)—1998 to 2020. Values are not normalized here. (B) Mean of Proton Density, Temperature, Speed and Solar Radio Index F10.7 per year, from 1998 to 2020. All values are normalized (center 0 and standard deviation 1) to plot them on the same scale.

under-represented values will have a more important high error on average than most-represented values. Although a limited number of plots are shown here, all histograms have been plotted and analyzed.

4.1.1.1 SWEPAM

Most of SWEPAM variables distributions (i.e., plasma parameters) were close to lognormal distributions [Burlaga and Lazarus \(2000\)](#). Hence, for clarity purposes and to enhance our



understanding, we also plot the distribution of the logarithm applied to these variables.

Ion velocity is the only plasma parameter that differs from a lognormal distribution. The most probable value is 364 km s^{-1} , lower than its median value 408 km s^{-1} . Most conclusions from Veselovsky et al. (2010) still hold when adding all the data until 2021. Although solar wind speed can reach values such as $1,000 \text{ km s}^{-1}$, 94.2% of all values are contained in the 300 km s^{-1} to 700 km s^{-1} window. Moreover, 500 km s^{-1} seems to be a breaking point, suggesting that two different distributions could overlap with a local maximum of around 600 km s^{-1} . According to Burlaga and Lazarus (2000), this could be due to corotating interactions regions where fast solar wind catches up slow solar wind, when corotating streams from coronal holes are numerous. As these phenomena appear more during declining solar activity, we plotted distribution for 2002–2009, 2015–2020 (two cumulated declining phases of solar activity) and distribution on the remaining dates (ascending phases). If needed as a comparison, the full distribution can be observed in the Supplementary Figure S2.

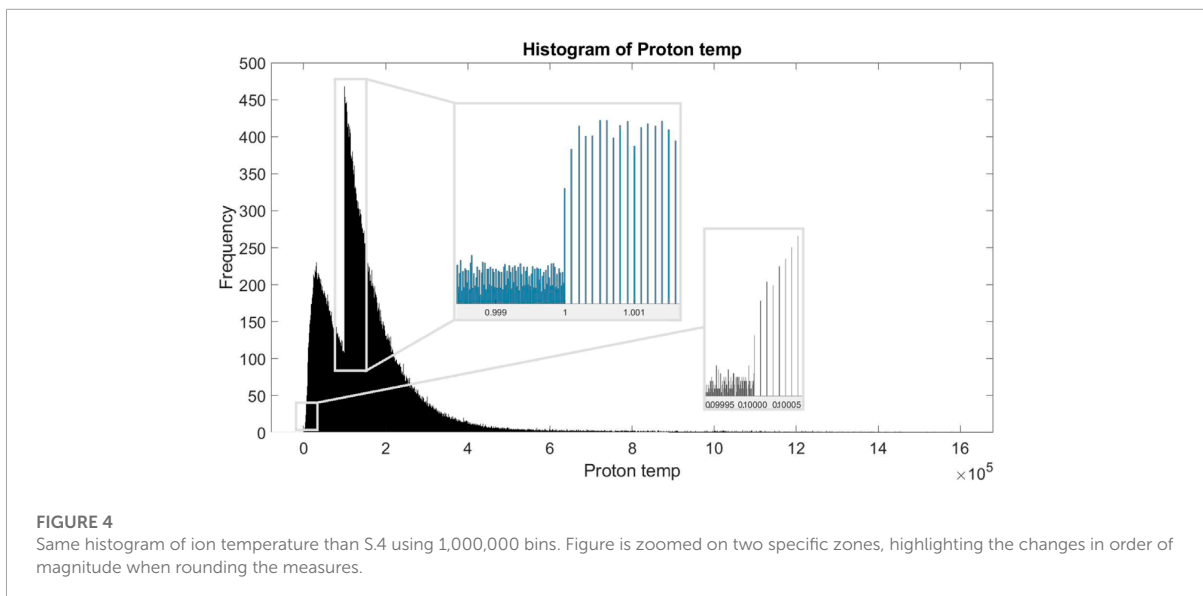
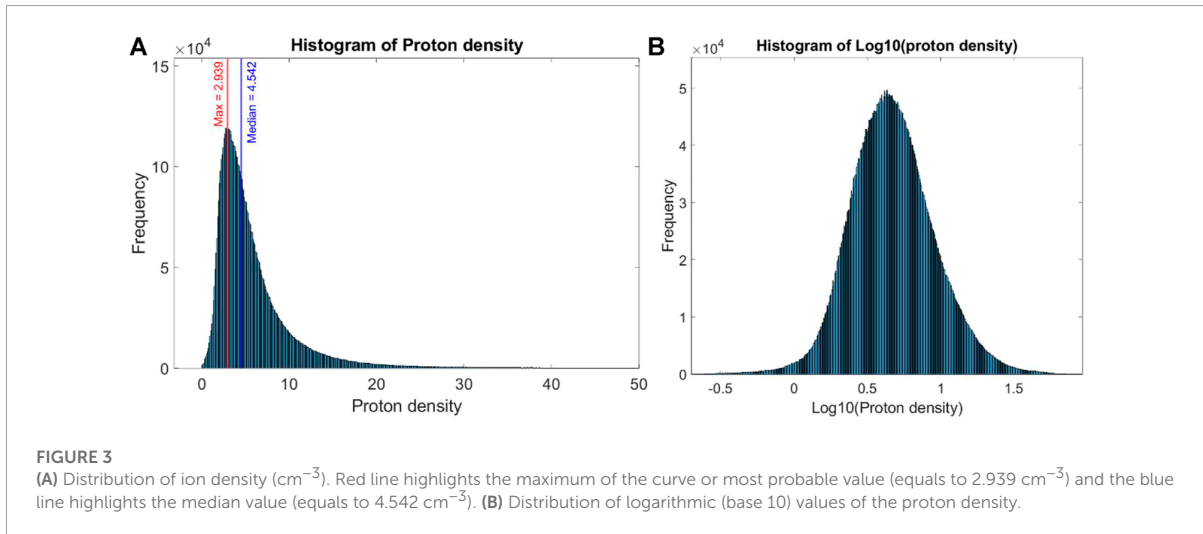
Figures 2A, B confirm a 600 km s^{-1} peak during the declining phases of the solar activity, as suggested by Burlaga and Lazarus (2000) with data from 1995 to 1998 (here confirmed with new data from 1998 to 2020). If we approximate the distribution of speed as lognormal, it then appears that the two declining phases of the solar activity are adding a Gaussian distribution of speed centered around $580\text{--}600 \text{ km s}^{-1}$. This is confirmed when looking at Figure 8 in Xu and Borovsky (2015). They classified solar wind into four plasmas: coronal-hole-origin plasma, streamer-belt-origin plasma, sector-reversal-region plasma, and ejecta. We see in Figure 2A the coronal-hole-origin plasma (black arrow). Let's keep in mind that a lot of work is being done

to have the solar wind classified (e.g., Camporeale et al. (2017)). This information is of prime importance if we want to use the solar wind to forecast other parameters.

Proton density follows a lognormal distribution with the most probable value being 2.94 cm^{-3} (Figures 3A, B). As seen Figure 1B, this value is highly influenced by the solar cycle. The peak of the density distribution is moving through our 23-years period, roughly following the solar cycle (represented here by F10.7 averaged each year).

Unlike Burlaga and Lazarus (2000), we do not observe any double peak in our ion density distributions, and our most probable value is far from their 8.0 cm^{-3} value (our is approximately 2.94 cm^{-3}). One can imagine this to be related to the global decline in solar activity since solar cycle 22 (solar cycles can be seen in the Supplementary Figure S1) but we do not have a proper explanation. This suggests, again, that we should not use less than 11 years of data to train our algorithm (ideally at least three cycles).

Temperature is well approximated by a lognormal distribution, and the most probable value is around $30,000 \text{ K}$. The two distributions have a similar shape to the ones in Figure 3 (if needed, this can be seen in the Supplementary Figure S4 for normal and logarithmic histograms). However, our computation of the most probable value shows an offset to the right instead of representing the peak of the distribution. This is one of the major issues that we will have with SWEPAM data in general, mentioned in Veselovsky et al. (2010): rounding of numbers is performed to different orders, with different significant digits. To highlight this issue, we plotted another histogram of the ion temperature with a higher number of bins (1,000,000) and zoomed on corresponding zones (Figure 4). As an example, when the temperature reaches $10,000 \text{ K}$, the measures start to



be rounded every 1 K (instead of every 0.1 K for values below 10,000 K). The same goes when reaching 100,000K, the measures start to be rounded every 10 K. The distribution when taking the logarithm values of the ion temperature is an even better view of the “jumps” in scale. Although these changes are anecdotic in most astrophysical applications, they are far from negligible in the AI context. Such changes in scale multiply the amount of data having identical values (as seen in [Figure 4](#), where the maximum is shifted to the right). Yet, a deep-learning algorithm will wrongly interpret these values as being more probable and will give them more importance during the training although they are not supposed to be so (maximum probability of temperature should stay around 30,000 K). As a consequence, the algorithm

will only focus on the most-probable value and the others will not be able to lead to coherent and correct results.

4.1.1.2 MAG

Histograms of the X, Y, and Z-components of the IMF seem close to Gaussian distributions. The norm of the IMF magnetic field vector seems close to a lognormal distribution. All plots can be found in the [Supplementary Figures S5–S8](#)). Some observed characteristics:

- X and Y-components could be interpreted as two superposed Gaussian, with two different most probable results each. X and Y components seem to have opposite

values and are linked by the orientation of the IMF when coming from the Sun (i.e., magnetic field lines are either oriented towards or away from the Sun). In addition, plotting the median of all values each year for these components suggest also suggests a strong relationship between the two, that should be considered before implementing them as input. The yearly median of both distributions seems to evolve in opposite directions over time and this is in line with the investigation shown in part 4.3. (report to part 4.3. for a better understanding but this can still be checked [Supplementary Figure S3](#)).

- The Z-component of the IMF is strangely following a perfect Gaussian curve with a center close to 0. Without any additional information from the space weather scientific community, one might assimilate the Z-component to a white-noise signal i.e., consider Bz as random. However, it is known (see for example [Kivelson et al. \(1995\)](#)) that the Bz-component orientation is responsible for magnetic reconnection at the front of the magnetosphere. When pointing southward, the IMF can connect to the Earth's northward magnetic field, allowing plasma to enter the dayside magnetosphere. When using ACE data to nowcast or forecast possible impacts of solar phenomena on in-space and on-ground systems, it is not possible to exclude the Bz-component. In general, analyzing data to answer a specific need using Machine Learning cannot be properly done without including the physical systems and phenomena responsible for the observations. The physics lying behind the data has to be addressed and understood to avoid absurd solutions and errors.
- Finally, the total IMF—B— distribution seems very close to a Laplacian distribution.

As a conclusion on histograms:

- Data shown here cannot be put in the algorithm as such. Distributions are everything but uniform and will lead to unequal training over samples. A possible consequence is having an algorithm incapable of dealing with rare cases (tails of the Gaussian and Laplacian curves).
- Gaussian noise is inherent to instruments. It might be very difficult (but useful) to evaluate the signal-to-noise ratio. A possible consequence on the training loss curve is to observe a steep drop followed by a flat trend, meaning that the algorithm quickly trained on the information it has and then started training on noise.
- Relation (linear or not) between data cannot be overlooked (we investigate them in part 4.3.).
- Particular attention is required on data values, as shown by the changes in the order of magnitude in the ion temperature (which, furthermore, could not be seen without manually increasing the number of bins).

4.1.2 Autocorrelations

The autocorrelation function (ACF) gives the data analyst indications on how future values are influenced by past values in time series. It helps identify randomness or periodic patterns, seasonality, and trends. When plotting ACF on the different features here, no autocorrelation is noticed, except for two trends.

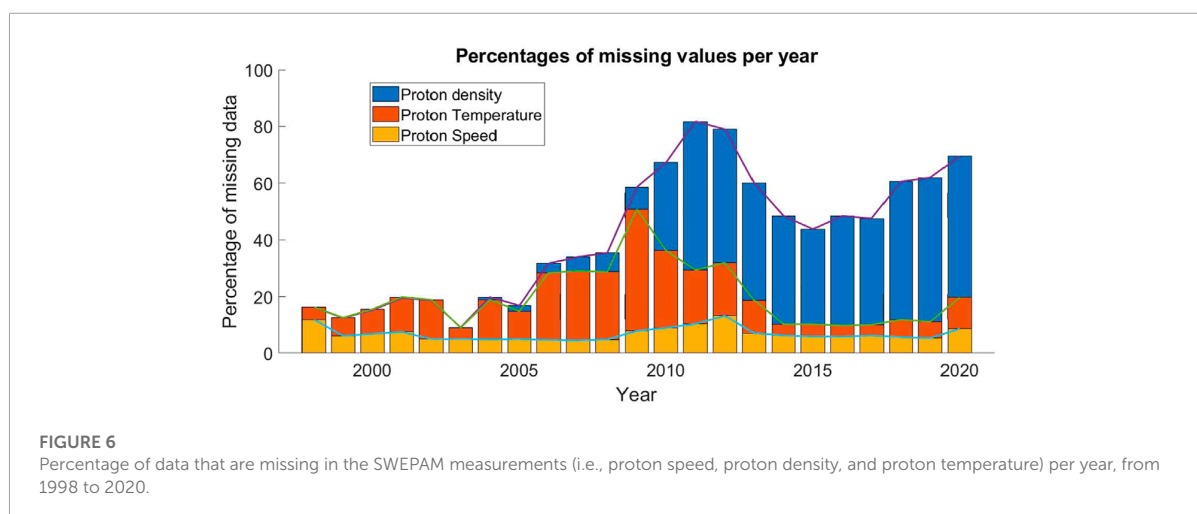
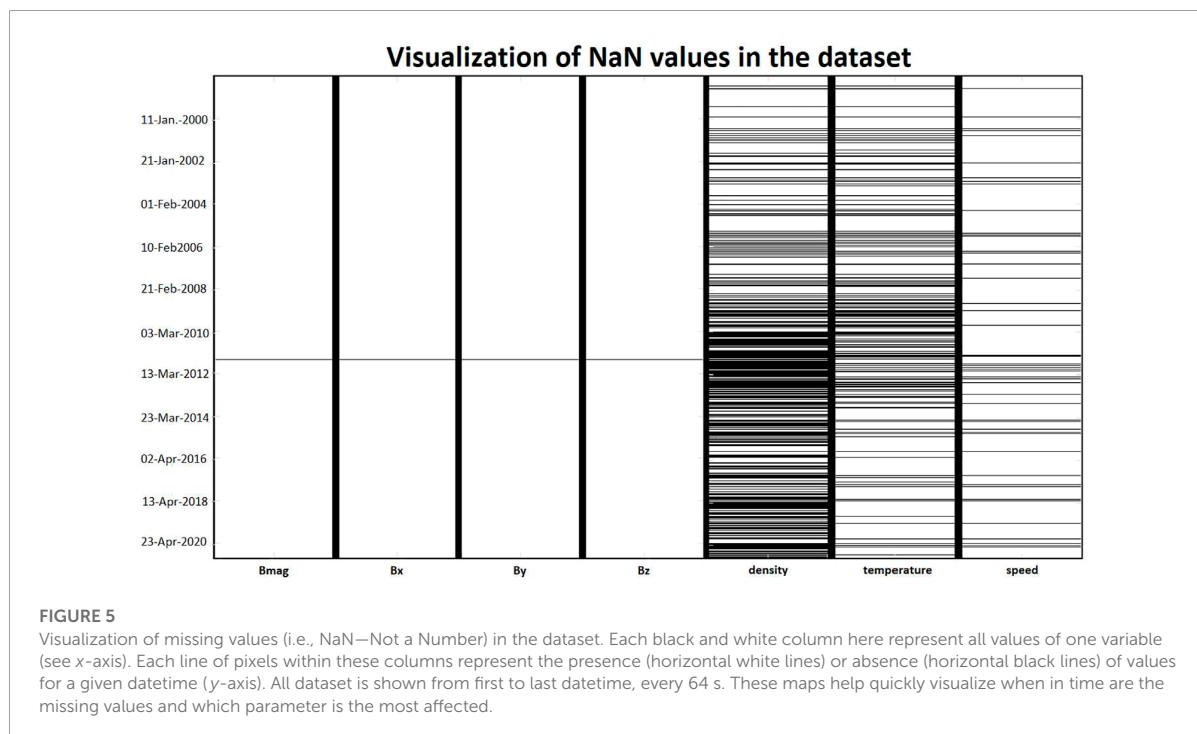
- IMF norm, X and Y-components reveal a 27-day periodicity, corresponding either to the Carrington synodic rotation period of the Sun or the Bartels Rotation Number. Solar rotation varies with latitude, with a maximum of 38 days at the poles and less than 25 days at the equator. In this context, the synodic period is 27.2753 days [Wilcox \(1972\)](#), and the Bartels Rotation Number is chosen to be exactly 27 days [Bartels \(1934\)](#) (the number of apparent rotations of the Sun as viewed from Earth and from L1 in our case).
- IMF norm also reveals the 11-year solar activity cycle.
- Density, temperature, and speed reveal the same two periodicities.

Graphs do not bring enough information and will just appear in the [Supplementary Figures S9–S12](#) for the reader's curiosity. It is important to notice that the lack of a clear autocorrelation is good news to apply machine learning technics. Time series data tend to be autocorrelated by having consecutive data with quite similar values. The risk when trying to forecast the next value is to end up using a persistence model, where the algorithm just picks the last value as being the best approximation for the next one. This is avoided when we have almost no autocorrelation, which is our case here.

4.2 Missing and extreme values

Missing and extreme values is a real struggle when working on AI algorithms but are more than usual in the Space Weather field. [Table 1](#) also presents the percentages of missing data in our dataset and [Figure 5](#) represents through a black and white image the missing values in our dataset. On the left side of [Figure 5](#) are the MAG measurements and on the right side are the SWEPAM measurements. At a glance, we can see that the time tags for missing values in the interplanetary magnetic field are the same, meaning that the instrument measures the components of the IMF altogether and that one missing value on a component means missing values on all components.

What we noticed from MAG does not hold for SWEPAM. Although most of the missing data happen at the same moment, they are still not distributed in the exact same way. However, it seems like when speed is missing, all of them are. This type of visualization is easy to create, and very useful to make a first opinion on how missing values are organized (in which variables, around which year, *etc.*).



Finally, there are much more missing data in the SWEAP dataset than in the MAG one, and almost half of the proton density data is missing: a very high amount that cannot be ignored when dealing with AI. Such high percentages require that we take a closer look as done in [Figure 6](#) (and, later, in [Table 2](#)).

Data here are Level-2 data, meaning that a group of experts analyzed them and kept reliable measures. Starting from 2009/2010, the amount of missing data is greatly increasing.

This information seen in [Figure 6](#) is confirmed when looking at the data status update of the ACE Science Center on 23 October 2012:

“The SWEAP observations, in particular the proton density and to a lesser extent the temperature, became increasing sparse starting in 2010 as the primary channel electron multiplier (CEM) detectors have aged. [...] In response, the ACE science team has developed and implemented, starting 23 Oct 2012, an innovative mission operations concept that more frequently

TABLE 2. Size of the biggest gap (number of consecutive missing values in the data) and number of gaps having a certain size (e.g., size 1 = one missing data surrounded by non-missing values, size 3 = 3 consecutive missing data) in the X, Y and Z-components of the IMF, and in the solar wind density, speed and temperature.

Variable	Size of biggest gap	# Gaps of size 1	# Gaps of size 2	# Gaps of size 3	# Gaps of size > 10	# Gaps of size > 100	# Gaps of size > 1350
Bx GSE	2255	1075	39	43	72	17	4
By GSE	2255	1075	39	43	72	17	4
Bz GSE	2255	1075	39	43	72	17	4
Bt GSE	2255	1075	39	43	72	17	4
Density	75182	299046	19794	6378	7413	2700	752
Speed	7007	446233	23214	6157	2418	286	41
Temperature	20037	413177	28534	8217	5223	1,616	416

repoints the ACE spacecraft’s spin axis further away from the Sun.” (Skoug et al. (2012)).

This information is of high value for Machine Learning scientists. As we saw, when working with an AI algorithm, we split the data into a train, a validation, and a test dataset. What is usually done in AI applied to Space Weather (and even broader when dealing with time series) is to pick a whole period (e.g., an entire year) as the validation or test set (McGranaghan et al. (2021)). A random choice would be dangerous as we might end up with a year with 81% of missing values (e.g., 2010).

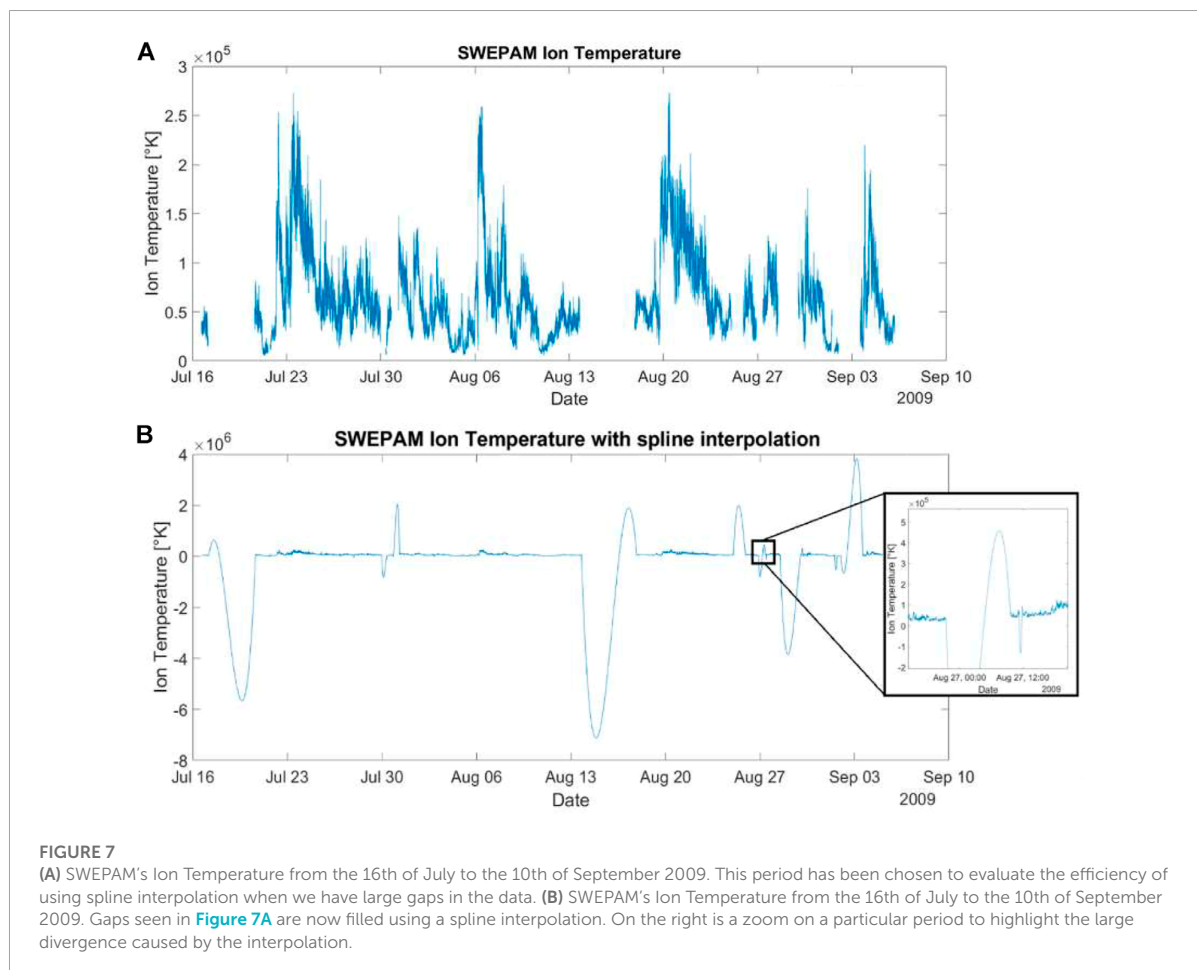
The next question to answer is how these missing values will be processed. First, let’s check the gaps of consecutive missing data and their size.

Table 2 shows the variety of gaps in our data. The biggest gap in the SWEFAM instrument is made of 75,182 consecutive missing data, approximately 55.6 consecutive days in the proton density. The density also has 752 gaps longer than a day. The longest gap in the speed is 5.2-day long, and the longest gap in the temperature is almost 15-day long. They have respectively 41 and 416 gaps more than 24 h long. Once again, we confirm the issue already seen for the density and temperature measurements: more than a lot of missing data, there are a lot of consecutive missing data.

Concerning the interplanetary magnetic field, this table also goes in the same direction as Figure 5. It confirms that the missing data are at the same time for all components of the magnetic field: only four gaps longer than a day and the longest gap is approximately 40-h long. Several processes exist to deal with missing data. Here are some examples:

- Removing all the rows containing missing data. The main advantage of this method is the robustness of the resulting model. However, using this method usually also removes some non-missing data. Here, the total loss of rows will be based on the ion density’s data, as it has 41.51% of missing data. It will result in a loss of almost four million proton speed data and 2.4 million proton temperature data points.
- Imputing missing values (especially for time series) with mean, median, last seen value, or through linear, spline or other interpolations. Such methods are quite easy to implement but might result in unpalatable results. In the following (Figures 10, 11), we applied the spline interpolation (as seen in some literature concerning AI in Space Weather - e.g., Gruet (2018)) on a few hours’ gaps in our SWEFAM’s ion temperature data around November 2020.

Figure 7B represents Figure 7A with gaps filled with spline interpolation. As expected, a spline interpolation cannot be used when a gap is too large, it fills the dataset with values at different



orders of magnitudes. The risk lies in the divergences such as the one between 13 August and 20 August 2009, giving very high values compared to the initial curve that now seems flat. Such extreme values will highly disturb algorithms, especially neural networks and can restrain them from learning. Even more, neural networks will tend to give high importance to these values, that were not even in our dataset at first.

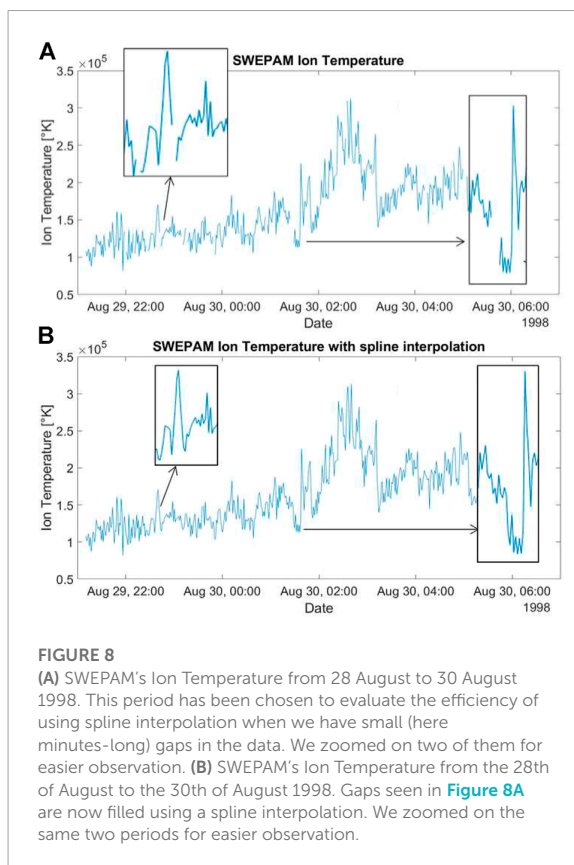
However, it is efficient with smaller gaps in (around 10 to 15 missing values according to [Andriambahoaka \(2008\)](#)). The gaps seen in [Figure 8A](#) between 28 August and 30 August 1998, are good examples on how efficient the spline interpolation can be for small gaps. [Figure 8B](#) shows the result when interpolating with spline.

It is essential to keep in mind this dependence on gaps' sizes when trying to impute values to missing data. The best way to deal with it is to have a detailed analysis of missing data (as we saw in [Table 2](#), or in [Figures 5, 6](#)), and use the best available method by first isolating characteristic gaps and testing methods on them independently.

- Finally, it is worth noticing that in astrophysics, gaps in the data could be filled by using other instruments and satellites that are measuring the same variables. In our case, satellites such as DSCOVR, also located in L1, represent viable solutions. However, inter-calibration between instruments will then have to be double-checked and can become critical if not considered.

As a conclusion on missing values:

- Missing data cannot be left aside and have to be looked at and processed, especially when dealing with time series.
- An analysis of the missing data should at least include percentages per variable, amount of missing data in time, size and number of gaps, few plots along with the data. It is advised to also consult the data suppliers and experts to better understand the analysis.
- While a large number of processes exist (e.g., removing rows or interpolating), they are not equivalent, and their use should depend on the aforementioned dataset analysis.



4.3 Interdependencies between variables

After analyzing every data independently, we now focus on comparing them together through three assessments:

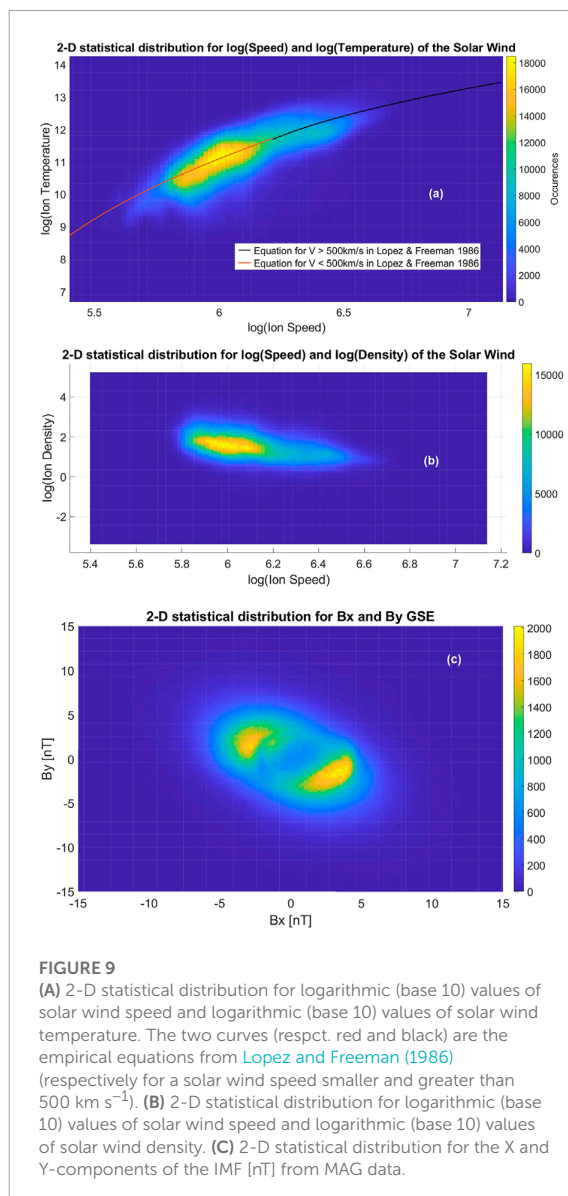
- Two-dimensional statistical distributions
- Correlation matrices
- Principal Component Analysis

4.3.1 Two-dimensional statistical distributions

We analyzed the two-dimensional statistical distributions of values for the logarithm of the solar wind's speed, temperature, and velocity. We are using the logarithm as an answer to the lognormal distributions observed in part 4.1.1.1. Here are the figures for speed and temperature (**Figure 9A**) and for speed and density (**Figure 9B**). The distribution for temperature and density did not highlight anything interesting.

These 2D statistical distributions highlighted well-known results:

- Proton temperature increases with solar wind speed and a linear correlation appears between the two (**Figure 9A**). In 1986, this linear correlation has been approximated by

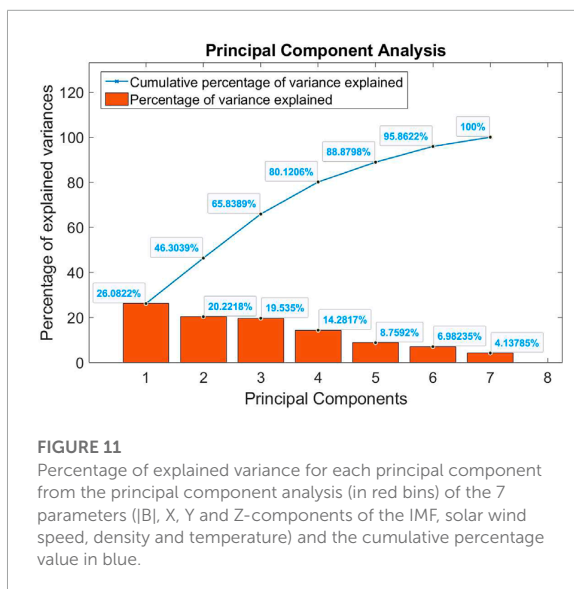
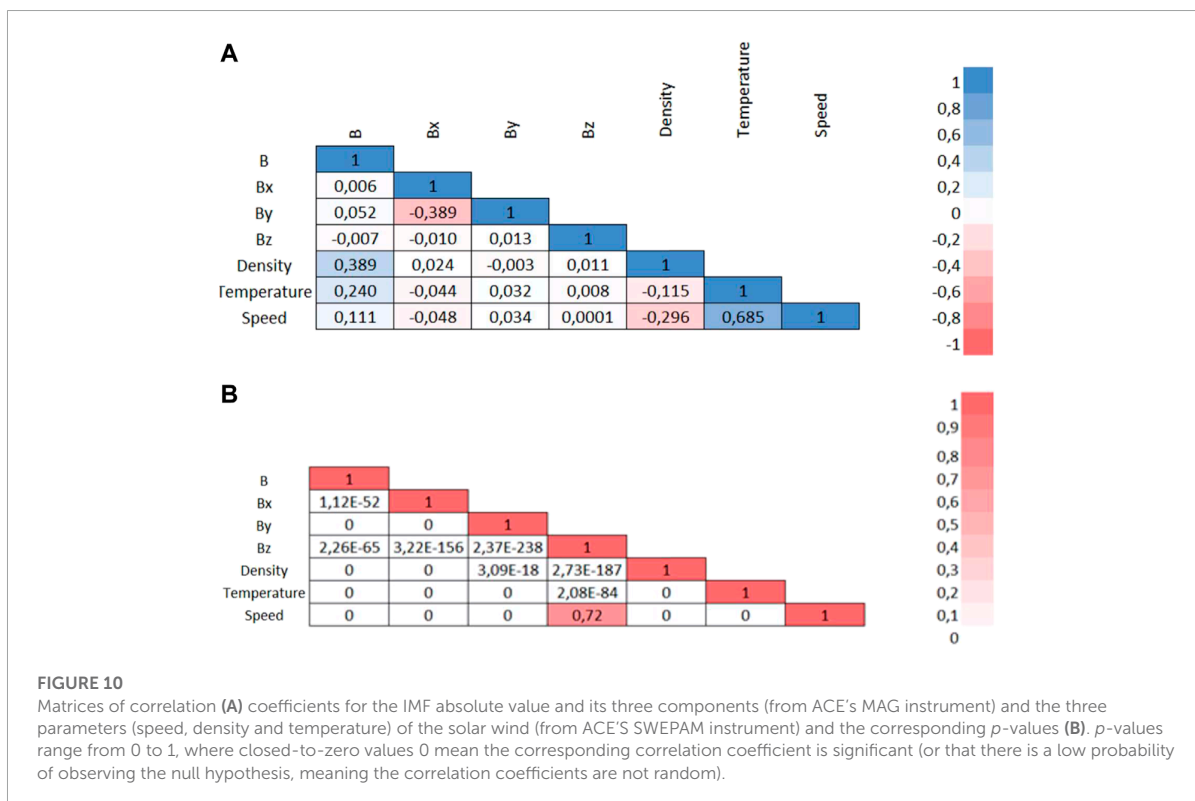


[Lopez and Freeman \(1986\)](#) with a difference for speeds above and below 500 km s^{-1} . We verified the accuracy of their following equations in the graph:

$$T = (0.77 \pm 0.021) V - (265 \pm 12.5) \text{ for } V > 500 \text{ km.s}^{-1} \quad (1)$$

$$T^{0.5} = (0.031 \pm 0.002) V - (4.39 \pm 0.08) \text{ for } V < 500 \text{ km.s}^{-1} \quad (2)$$

The first one appears in black in **Figure 9A** while the second appears in red. It is interesting to notice that a model built in 1986 seems quite valid on these data from 1998 to December



2020. During CMEs however, temperature is usually lower Richardson and Cane (1995).

- Density and speed are well correlated. Fast solar wind is usually less dense, and slow solar wind varies a lot

Geiss et al. (1995). Recall that fast solar wind can catch up slow solar wind and compress it creating what is called corotating interaction regions Jian et al. (2006). This is a known result but the corresponding figure is shown here (Figure 9B)

Concerning MAG, the two-dimensional statistical distribution for the X and Y-components of the IMF has two maxima and shows the approximate 45° angle between the IMF vector and the radial Sun-Earth direction. This angle is the direct consequence of the characteristic Parker's spiral (first theoretically predicted by Chapman) flow of the solar wind Parker, 1963; Kivelson et al. (1995). These two maxima can be found in Figure 9C. Two-dimensional plots including the Bz were not adding new information and are not shown here.

4.3.2 Linear correlation matrices

Until now, we subsampled MAG data to obtain one data point every 64 s. Now, to compare MAG and SWEPAM together, it is important to have the same timestamp for every data. The choice made was to keep SWEPAM data and its corresponding timestamps and, for every data point of SWEPAM, take the closest MAG 16sec-data point in time and change its timestamp

to the SWEPAM's one². The result is a dataset of 11, 282, 160 data points, from the 04th of February 1998 to the 22nd of December 2020. After removing all rows where data were missing, we end up with 6,559,840 samples from which we can compute the correlations and the corresponding p -values matrices (Figure 10).

As expected, the correlations between the proton's speed and temperature, and between density and temperature (although smaller) appear. Two small negative correlations (between proton's speed and density and between X and Y-components of the IMF) also appear. Oddly enough, there is a high p -value for the correlation coefficient between speed and the Z-component of the IMF, but the correlation is inexistent (Figure 10). Finally, let's recall that the correlation coefficient is none other than the cosine of the angle between the two centered vectors and that the cosine function is not linear. Hence, a 0.685 (our higher correlation coefficient here) corresponds to a 46.76° angle between the two vectors. In other words, no significant enough correlation has been obtained here. From an AI point of view, and without nonlinear pre-processing, this means that we want to keep all the parameters as they might contain different relevant information. But would it be possible to combine parameters together to reduce the total amount of parameters needed? The principal component analysis will answer this question.

4.3.3 Principal component analysis (PCA)

What is the idea behind the PCA? As an example, let's assume that we have a dataset made of p different variables and let's suppose that each observation is close to a specific n -dimension hyperplane in $\mathbb{R}_p (n \leq p)$. The idea of the PCA is to find this possible "best plane" (the plane such that the sum of the distances of the points to that plane is the smallest). PCA then gives us this new coordinate system (or affine space of dimension n) and data are projected in it. Note that the distances between observations in this new system best reflect the distances between observations in the starting space \mathbb{R}_p . PCA answers the problem of finding the n -dimensional linear space which best represents the observations in the sense that the orthogonal projection on this space moves them as little as possible. In AI, it is widely used when preprocessing the data either to reduce the number of features needed or to target the most relevant features in a given dataset.

Following in Figure 11 is the PCA applied to our dataset (IMF $|B|$, X, Y and Z-components, proton density, temperature, and speed). In the output of the PCA are the principal components, which are the vectors of the new coordinate system. The first component is such that it contains the greatest

variance of some scalar projection of the data points on it. For further understanding, a handmade example (for which data have nothing to do with ours) can be seen in the Supplementary Figure S13.

Hence, from Figure 11, it appears that the PCA does not find any good coordinate system in which to project the data points, justifying the use of more complex models for data analysis and data processing (e.g., non-linear models, Hinton's t-distributed stochastic neighbour embedding—Van der Maaten and Hinton (2008)). This can be seen in the quasi-linear augmentation of the cumulative percentage of variance explained. An ideal case would have been to have a major (90%) percentage of the variance explained by the first three principal components but almost all components here explain the same amount of variance.

As a conclusion on dependencies between variables:

- There is a dependency between speed and temperature that will need further observations.
- There is a non-linear relationship between the X and the Y-components of the IMF. Hence, they cannot be considered independently.
- Most of the graphs did not show any linear dependency (this will be checked further with correlation matrices) and hence might imply the use of non-linear models. This has been confirmed by the correlation matrix and the PCA.

5 How to use the data for AI

Once we observed and analyzed the data, we need to preprocess it, which usually means:

- Choosing the final set of input features and labels. The selection of variables to pass as input into the model is essential. The model must be informed of the possible relationships between inputs and outputs. Some information might not be sufficient for the model to understand these relationships and it is then highly recommended to discuss the underlying objectives with experts from the field. In astrophysics problematics, the physical relationships between variables have to be used to construct the set of features McGranaghan et al. (2021). However, too many features carrying the same information might also impact the performance of the model. It is better to avoid redundant information Khalid et al. (2014). Intercorrelations and PCA are quite useful to remove some unwanted features. Moreover, the final samples will be built as a vector containing all the input features and the label (labelled data appear in supervised learning algorithms only) and, in the case of time series, one has to choose the temporal resolution for it (usually the resolution of the labels).

² Special thanks to Pierre Porchet and his generous help in preparing this massive dataset (processing 45 million data points for MAG and 11 million for SWEPAM - respectively 6.3 and 2.6 Gigabytes of data).

Features with a lower resolution will have missing values and features with a higher resolution will be transformed (e.g., mean, max, standard deviation, *etc.*). Indeed, the set of features can include transformed variables (e.g., the square of the density). Of course, it might also contain passed values of variables (e.g., the magnetic field B, and the same magnetic field B 1 hour ago, or 1 day ago). In this case, autocorrelations are useful to identify redundant information in time. Overall, choosing the input features will depend on our objectives (whether it is forecasting or classifying for example) and our knowledge of the underlying physical phenomena (depending on our aims, the algorithm might find better solutions with the density squared or with the past 3 h of magnetic field).

- Handling the missing values. Null values are quite a challenge as they are abundant in the space weather field. Removing entire rows of data will result in significant information loss, and we just saw that interpolation depends on the sizes of gaps in the data. In our context, a good response might be to find another satellite or data source when talking with experts (e.g., DSCOVR) and fill the gaps using interpolation with these new data points. If we do not have other data sources, a compromise should be found between removing and interpolating.
- Standardizing or normalizing the data. We will not detail here the differences between these two, but rescaling the data is required for the model to compare inputs together. It avoids placing too much emphasis on variables with large values (e.g., speed would be considered more important than density). The field of astrophysics also faces observations with high variance and a large number of outliers (defined as extreme values far from the initial distribution, often thought to be generated by a different mechanism—Hawkins (1980)). Outliers are particularly problematic when located in the labels of the dataset. The extent to which a label-outlier disrupts the training will be discussed in a subsequent paper. One way to remove these outliers is to remove entire samples where labels are behind certain quantiles in their probability distribution. For instance, McGranaghan et al. (2021) removed all samples where the labels' values were out of the 99.995th percentiles. However, in space weather we are often concerned about the extreme values since they pose the most risk. A user must be able to differentiate between real anomalies and extreme values that are accounting for extreme phenomenon and treat them differently. The algorithm cannot distinguish them by itself. Another possibility would be to do anomaly detection (another field of Machine Learning) but, again, it is impossible to assess the efficiency of the algorithm without an expert able to differentiate anomalies and relevant extreme values. Finally, a user could adapt the loss function to account for physical phenomenon. Loss

functions are functions allowing the algorithm to learn, they are cost function Wang et al. (2022) such as the Mean Squared Error function. The difficulty in adapting a loss function is that one must very well understand the physics behind the phenomena, but trying to understand these phenomena is often the very point of using AI in the first place.

6 Discussion and Conclusion

In the field of Space Weather, the use of AI is progressively gaining importance. First mentions of machine learning techniques or neural networks at the European Space Weather Week appeared around 2011 and dedicated “Machine learning and statistical inference techniques” sessions only appeared in 2016. In this context, proper understanding and pre-processing of the data is central. Here, we decided to focus on ACE satellite data as it has been widely used by the community and considered a good indicator to forecast the near-Earth phenomena. Its location (L1) and measurements (IMF, solar wind parameters and particle fluxes) made it the perfect candidate for our study. Obviously, the methods presented here have to be adjusted depending on the dataset and one's objectives. Concerning our dataset, the conclusions are the following:

- Some parameter distributions are well approximated by Gaussian distributions while others are closer to lognormal laws. As said in Veselovsky et al. (2010) lognormal laws can testify of “multiple multiplicative transformations of local characteristics at intermitting random intensifications and attenuations of waves, compression and rarefaction of irregularities in turbulent processes of transporting mass, energy, and momentum on the Sun and in the heliosphere”. Overall, histograms are not uniform distributions at all. If we use data as such, algorithms will perform well on more frequent samples and poorly on rare cases. For example, if the purpose is to forecast events related to very fast and dangerous solar winds, our algorithm will struggle to obtain anything interesting. Moreover, it is important to keep in mind the possible noise in our measurements. The signal-to-noise ratio seems difficult to estimate here and the interesting information may be hidden in noisy data.
- Histograms are not steady and change from year to year, maybe due to dependence to the solar cycle. This means that a model built on a single (or limited number of) year(s) might not be reproducible and usable in the future. At best, one should know the origin of such changes. In any case, the training set has to be well-balanced and has to include several different years of data (e.g., both ascending and descending phases of the solar cycle).

- We must pay special attention to rounded measurements when there are changes in the order of magnitude within the data, as seen with the ion temperature data. The consequence could be an over-attention of the algorithm on higher values as they would appear more frequent. Two possible solutions here: either round all the data to the highest order of magnitude, or artificially re-distribute values following the closest Gaussian distribution (when looking at the logarithm of proton temperature).
- The number of missing values in our dataset is significant and has to be addressed (e.g., 41.59% of proton density data missing). For the analysis, we removed the corresponding samples, but it is not a solution for the training when the number of missing values is very high. The best solution here would be to use DSCOV data. Either way, when filling missing data, sizes of gaps have to be looked at to choose a corresponding interpolation method for instance.
- Even if we noticed the well-known linear relationship between speed and temperature of the solar wind, a linear model might not be enough to accurately model the data. It seems that non-linear relationships between data exist (e.g., X-component and Y-component of the interplanetary magnetic field). PCA, correlation matrix and 2-D statistical distributions suggest that all parameters should be kept and that non-linear models should be preferred.
- Overall, some cycles appear in the dataset. Proton speed seems highly dependent on the solar cycle and the synodic rotation period of the Sun appears in most of the autocorrelations. We advise having several solar cycles included in the training set to avoid biases. Solar cycle could also be part of our input features through the solar radio flux at 10.7 cm or the sunspots number.

As mentioned in the introduction, [Smith et al. \(2022\)](#) study is very complementary to ours. The differences lie in the methods and data chosen.

- First, [Smith et al. \(2022\)](#) take into account both ACE and DSCOV data while we only focused on level-2 ACE data.
- They compare together Near-Real-Time (NRT) raw data to the same data post-processed by the scientific community. On our side, we do not assess the quality and relevance of raw data as we considered level-2 data as the entry point of any AI study in this field. [Smith et al. \(2022\)](#) indeed show that the NRT values are subject to short-term variability and anomalous values, confirming our choice.
- Concerning missing values, [Smith et al. \(2022\)](#) again compare NRT and scientific data. They draw some conclusions about the amount data gaps, but we go slightly beyond in [Table 2](#) and through the testing of filling methods. However, their analysis on windowed data validity (part

3.2.2.) is very interesting. Indeed, some AI algorithms need windows of consecutive data (e.g., Temporal Convolutional Network) to learn properly. Here, as shown in their study: “if 2 hours (120 min) of continuous input are required then [...] approximately 1% of plasma data are available.” Missing values is then an even bigger problem and it is required to choose a method to deal with missing values.

- Finally, concerning autocorrelations, the difference lies in the use of data. [Smith et al. \(2022\)](#) do autocorrelations on NRT data and only on 1 h-long windows of consecutive data without missing values. On our side, we do autocorrelations on level-2 ACE data and we take all the data as input and omit the computation for missing values.

Data analysis goes hand in hand with the field's expertise. Some of the solutions suggested here will not be ideal depending on one's objectives and the conclusions one might have when looking only at the statistics could also be wrong. As an example, even if Gaussian distributions are often associated with random processes, we know that the mechanisms lying behind the values of the IMF and solar wind are everything but random. We also know that very fast and powerful CMEs can saturate instruments and create missing values, hence changing how we would consider replacing them. Knowing how AI algorithms work can give us clues on what to focus on when analyzing a dataset and where a problem might arise. However, it is the understanding of these data and a space weather expertise together that will allow us to favor one solution over another.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: <https://izw1.caltech.edu/ACE/ASC/level2/>.

Author contributions

SB: Conceived and designed the analysis; Collected and organized the data; Contributed data or analysis tools; Contributed astrophysics analysis; Performed the analysis and created graphs; Wrote the paper PV: Contributed to the data science part of the analysis; Conceived and designed the analysis; Contributed data or analysis tools; Helped writing paper; Contributed to the redaction with major corrections and changes MB: Contributed to the astrophysics part of the analysis; Corrected several parts of the manuscript JC: Correction and guidance All authors contributed to manuscript revision, read, and approved the submitted version.

Acknowledgments

The authors would like to acknowledge the ACE Science Center for providing the data and especially Andrew Davis for answering our concerns. In addition, the authors would like to thank Data Science Expert and the PNST (Programme National Soleil-Terre). A special thanks to Pierre Porchet for his precious help in understanding, processing, and drawing conclusions from the data. Special thanks to Elisa Robert and Angélique Woellflé for their support and expertise. The authors would also like to thank SpaceAble for their support and expertise. The company was not involved in the study design, collection, analysis or interpretation of data.

Conflict of interest

Author SB was employed by SpaceAble.

The remaining authors declare that the research was conducted in the absence of any commercial or financial

relationships that could be construed as a potential conflict of interest.

Publisher's note

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

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fspas.2022.980759/full#supplementary-material>

References

- Andriambahoaka, Z. (2008). "Modélisation régionale du champ magnétique terrestre et établissement de cartes magnétiques détaillées appliqués à Madagascar." (Strasbourg, France: Université Louis Pasteur). Ph.D. thesis.
- Bartels, J. (1934). Twenty-seven day recurrences in terrestrial-magnetic and solar activity, 1923–1933. *J. Geophys. Res.* 39 (3), 201–202a. doi:10.1029/TE039i003p00201
- Bentley, S., Watt, C., Owens, M., and Rae, I. (2018). ULF wave activity in the magnetosphere: Resolving solar wind interdependencies to identify driving mechanisms. *J. Geophys. Res. Space Phys.* 123 (4), 2745–2771. doi:10.1002/2017JA024740
- Burlaga, L., and Lazarus, A. (2000). Lognormal distributions and spectra of solar wind plasma fluctuations: Wind 1995–1998. *J. Geophys. Res.* 105 (A2), 2357–2364. doi:10.1029/1999JA900442
- Camporeale, E., Carè, A., and Borovsky, J. E. (2017). Classification of solar wind with machine learning. *J. Geophys. Res. Space Phys.* 122 (11), 10–910. doi:10.1002/2017JA024383
- Camporeale, E., and S. O. C. of ML-Helio (2020). ML-Helio: An emerging community at the intersection between heliophysics and machine learning. *JGR. Space Phys.* 125 (2), e2019JA027502. doi:10.1029/2019JA027502
- Camporeale, E. (2019). The challenge of machine learning in space weather: Nowcasting and forecasting. *Space weather.* 17 (8), 1166–1207. doi:10.1029/2018SW002061
- Camporeale, E., Wing, S., and Johnson, J. (2018). *Machine learning techniques for space weather.* Amsterdam, Netherlands: Elsevier.
- Chen, S., Dobriban, E., and Lee, J. H. (2020). A group-theoretic framework for data augmentation. *J. Mach. Learn. Res.* 21 (1), 9885–9955. doi:10.48550/arXiv.1907.10905
- Daglis, I., Chang, L., Dasso, S., Gopalswamy, N., and Khabarova, O. (2020). Predictability of the variable solar-terrestrial coupling. *Annales Geophysicae.* doi:10.5194/angeo-39-1013-2021
- Geiss, J., Gloeckler, G., and Von Steiger, R. (1995). Origin of the solar wind from composition data. *Space Sci. Rev.* 72 (1), 49–60. doi:10.1007/BF00768753
- Gombosi, T. I., Chen, Y., Manchester, W., Zou, S., Hero, A. O., Landi, E., et al. (2018). *Machine learning and the "holy grail" of space weather forecasting.* SM54A–02.
- Gruet, M. (2018). "Intelligence artificielle et prévision de l'impact de l'activité solaire sur l'environnement magnétique terrestre." (Toulouse, ISAE). Ph.D. thesis.
- Hawkins, D. M. (1980). *Identification of outliers.* Berlin, Germany: Springer.
- Jian, L., Russell, C., Luhmann, J., and Skoug, R. (2006). Properties of stream interactions at one AU during 1995–2004. *Sol. Phys.* 239 (1), 337–392. doi:10.1007/s11207-006-0132-3
- Khalid, S., Khalil, T., and Nasreen, S. (2014). "A survey of feature selection and feature extraction techniques in machine learning," in Proceedings of the 2014 Science and Information Conference, London, UK, 27–29, Aug. 2014, 372–378. doi:10.1109/SAI.2014.6918213
- King, J., and Papitashvili, N. (2006). *One min and 5-min solar wind data sets at the Earth's bow shock nose.* Greenbelt, Md: NASA Goddard Space Flight Cent.
- Kivelson, M. G., Kivelson, M. G., and Russell, C. T. (1995). *Introduction to space physics.* Cambridge, United Kingdom: Cambridge University Press.
- LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature* 521 (7553), 436–444. doi:10.1038/nature14539
- Lopez, R. E., and Freeman, J. W. (1986). Solar wind proton temperature-velocity relationship. *J. Geophys. Res.* 91 (A2), 1701–1705. doi:10.1029/JA091iA02p01701
- McComas, D., Bame, S., Barker, P., Feldman, W., Phillips, J., Riley, P., et al. (1998). "Solar wind electron proton alpha monitor (SWEPAM) for the Advanced Composition Explorer," in *The advanced composition explorer mission* (Berlin, Germany: Springer), 563–612. doi:10.1007/978-94-011-4762-020
- McGranaghan, R. M., Ziegler, J., Bloch, T., Hatch, S., Camporeale, E., Lynch, K., et al. (2021). Toward a next generation particle precipitation model: Mesoscale prediction through machine learning (a case study and framework for progress). *Space weather.* 19 (6), e2020SW002684. doi:10.1029/2020SW002684
- Mishra, P., Biancolillo, A., Roger, J. M., Marini, F., and Rutledge, D. N. (2020). New data preprocessing trends based on ensemble of multiple preprocessing techniques. *TrAC Trends Anal. Chem.* 132, 116045. doi:10.1016/j.trac.2020.116045
- Myagkova, I., Shirokii, V., Vladimirov, R., Barinov, O., and Dolenko, S. (2020). "Comparative efficiency of prediction of relativistic electron flux in the near-earth space using various machine learning methods," in *International conference on neuroinformatics* (Berlin, Germany: Springer), 222–227. doi:10.1007/978-3-030-60577-325

- Nita, G., Georgoulis, M., Kitiashvili, I., Sadykov, V., and Camporeale, E., 2020. Machine learning in heliophysics and space weather forecasting: A white paper of findings and recommendations. *arXiv preprint arXiv:2006.12224*. doi:10.48550/arXiv.2006.12224.
- Parker, E. N. (1963). The Solar-Flare Phenomenon and the Theory of Reconnection and Annihilation of Magnetic Fields. *The Astrophysical Journal Supplement Series* 8, 177
- Reep, J. W., and Barnes, W. T. (2021). Forecasting the remaining duration of an ongoing solar flare. *Space weather*. 19 (10), e2021SW002754. doi:10.1029/2021SW002754
- Reiss, M. A., Möstl, C., Bailey, R. L., Rüdiger, H. T., Amerstorfer, U. V., Amerstorfer, T., et al. (2021). Machine learning for predicting the Bz magnetic field component from upstream *in situ* observations of solar coronal mass ejections. *Space weather*. 19 (12), e2021SW002859. doi:10.1029/2021SW002859
- Richardson, I., and Cane, H. (1995). Regions of abnormally low proton temperature in the solar wind (1965–1991) and their association with ejecta. *J. Geophys. Res.* 100 (A12), 23397–23412. doi:10.1029/95JA02684
- Russell, C. T. (1971). Geophysical coordinate transformations. *Cosm. Electrodyn.* 2 (2), 184–196.
- Shorten, C., and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *J. Big Data* 6 (1), 60–48. doi:10.1186/s40537-019-0197-0
- Shprits, Y. Y., Vasile, R., and Zhelavskaya, I. S. (2019). Nowcasting and predicting the Kp index using historical values and real-time observations. *Space weather*. 17 (8), 1219–1229. doi:10.1029/2018SW002141
- Skoug, R., McComas, D., and Elliott, H. (2012). *Effect of ACE spacecraft repointing on SWEPAM calculated moments*.
- Smith, A., Forsyth, C., Rae, I., Garton, T., Jackman, C., Bakrania, M., et al. (2022). On the considerations of using near real time data for space weather hazard forecasting. *Space weather*. 20 (7), e2022SW003098. doi:10.1029/2022sw003098
- Stone, E. C., Frandsen, A., Mewaldt, R., Christian, E., Margolies, D., Ormes, J., et al. (1998). The advanced composition explorer. *Space Sci. Rev.* 86 (1), 1–22. doi:10.1023/A:1005082526237
- Storrs, K. R., Anderson, B. L., and Fleming, R. W. (2021). Unsupervised learning predicts human perception and misperception of gloss. *Nat. Hum. Behav.* 5 (10), 1402–1417. doi:10.1038/s41562-021-1097-6
- Stumpo, M., Benella, S., Laurenza, M., Alberti, T., Consolini, G., and Marcucci, M. F. (2021). Open issues in statistical forecasting of solar proton events: A machine learning perspective. *Space weather*. 19 (10), e2021SW002794. doi:10.1029/2021SW002794
- Van der Maaten, L., and Hinton, G. (2008). Visualizing non-metric similarities in multiple maps. *Mach. Learn.* 9 (11), 33–55. doi:10.1007/s10994-011-5273-4
- Veselovsky, I., Dmitriev, A., and Suvorova, A. (2010). Algebra and statistics of the solar wind. *Cosm. Res.* 48 (2), 113–128. doi:10.1134/S0010952510020012
- Wang, Q., Ma, Y., Zhao, K., and Tian, Y. (2022). A comprehensive survey of loss functions in machine learning. *Ann. Data Sci.* 9 (2), 187–212. doi:10.1007/s40745-020-00253-5
- Wihayati, Purnomo, H. D., and Trihandaru, S. (2021). “Disturbance storm time index prediction using long short-term memory machine learning” in 2021 4th International Conference of Computer and Informatics Engineering (IC2IE), Depok, Indonesia, 14–15 Sep. 2021 (IEEE), 311–316. doi:10.1109/IC2IE53219.2021.9649119
- Wilcox, J. M. (1972). “Divers solar rotations,” in *Cosmic plasma physics* (Berlin, Germany: Springer), 157–164. doi:10.1007/978-1-4615-6758-520
- Wintoft, P., Wik, M., and Viljanen, A. (2015). Solar wind driven empirical forecast models of the time derivative of the ground magnetic field. *J. Space Weather Space Clim.* 5, A7. doi:10.1051/swsc/2015008
- Xu, F., and Borovsky, J. E. (2015). A new four-plasma categorization scheme for the solar wind. *J. Geophys. Res. Space Phys.* 120 (1), 70–100. doi:10.1002/2014ja020412, Available at: <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2014ja020412>.
- Zewdie, G. K., Valladares, C., Cohen, M. B., Lary, D. J., Ramani, D., and Tsidu, G. M. (2021). Data-Driven forecasting of low-latitude ionospheric total electron content using the random forest and LSTM machine learning methods. *Space weather*. 19 (6), e2020SW002639. doi:10.1029/2020SW002639, Available at: <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2020SW002639>.

To summarize the conclusion of our paper, proper data understanding and preprocessing are crucial, and we focused here on ACE satellite data, known for forecasting near-Earth phenomena. The conclusions about these ACE data are numerous: parameter distributions are not uniform; histograms change over time, likely due to the solar cycle; attention must be paid to rounded measurements to avoid bias; significant missing values require addressing; linear models may not be sufficient, and non-linear relationships exist; cycles appear in the dataset, with proton speed highly dependent on the solar cycle. Including multiple solar cycles in the training set is recommended to avoid biases.

As a consequence of this analysis (and after several trials of applying AI) to the ACE satellite data, it became evident that the challenges posed by the dataset necessitated a change in approach. Due especially to the extensive number of missing values in regard of our deep learning approach, we made the decision to focus solely on the OMNIWeb dataset for its better suitability and potential for addressing these issues. Utilizing the OMNI data allowed us to proceed with greater confidence in achieving meaningful results. However, as we will see in the following section, OMNI web dataset also face some difficulties.

3.3.2 DMSP

Regarding our target data, the primary analyses and preprocessing that need to be performed actually concern the raw physical data itself: Are there too many measurement errors and noise in the data? Can it be quantified? Has the instrument remained reliable throughout all these years? Do we have a solid understanding of its position? Are the instruments comparable and properly intercalibrated? As it turns out, most of these analyses, along with appropriate data treatments to make it more usable and reliable, were conducted by [Redmon et al. \(2017\)](#) when they created a new database for precipitated auroral electrons and ions from DMSP. This database is considered robust, highly reliable, and will be the one we use. Hence, in this section, we will start by giving a brief overview of the work done by [Redmon et al. \(2017\)](#). Afterward, we will proceed to a concise analysis of the data distributions and check for the presence of any outliers.

3.3.2.1 New DMSP database from [Redmon et al. \(2017\)](#)

[Redmon et al. \(2017\)](#) established a comprehensive public database featuring SSJ instrument data onboard DMSP spacecraft. Spanning over 30 years, it includes data from DMSP missions launched between 1982 and 2009 (F06 to F18). The database encompasses crucial information on precipitating electrons and ions, encompassing original counts, calibrated differential fluxes accounting for penetrating radiation, total kinetic energy flux, characteristic energy, and uncertainty estimates. Additionally, precise ephemerides (spacecraft positions and orientations) are provided after improved estimation. Accessible through the National Centers for Environmental Information and the Coordinated Data Analysis Web (NCEI, CDAWeb), this database is poised to facilitate diverse space science research, spanning from individual observatory investigations to comprehensive system science studies. In their article, [Redmon et al. \(2017\)](#) provide their ephemeris calculations, the calculations used to arrive at particle fluxes, and the known issues, that we summarized here:

- Ephemeris: The ephemeris refers to the spacecraft's position and orientation in three different coordinate frames: Earth Centered Inertial (ECI) True of Date (TOD) Epoch, geographic (GEO), and Altitude Adjusted Corrected Geomagnetic (AACGM). The authors employed two methods to calculate the spacecraft's ephemeris, which refers to its position and orientation in different coordinate frames. The first method involved using the Simplified General

Perturbations (SGP, [Vallado et al. \(2006\)](#)) theory to propagate two line elements (TLE). The second method interpolated ECITOD estimates from the NASA Space Physics Data Facility (SPDF). The latter approach was their standard processing method and utilized an eight-order interpolation to align with the timestamps of environmental measurements, resulting in an accurate RMSE of less than 6 meters per axis. The authors then rotated the ECITOD locations to the geographic (GEO) frame using the IDL Astronomy User's Library (IDLAstro) "eci2geo" routine, followed by transforming GEO to AACGM latitude, longitude, and magnetic local time (MLT) using the Super Dual Auroral Radar Network IDL AACGM library. A minor adjustment was made for the geocentric radius discrepancy between the two calculators (6378.137 km in IDLAstro and 6371.2 km in AACGM). To account for the lack of time-varying magnetic field coefficients in the version of AACGM used, the authors linearly interpolated the AACGM estimates at the two nearest 5-year epochs onto the instrument timestamps. The ephemeris parameters provided in their data repository include ECITOD in Cartesian coordinates (in kilometers), geographic latitude, longitude, geocentric radius (in kilometers), and AACGM latitude, longitude, and local time. Limitations include discrepancies between estimated and retrospectively computed ephemerides and the lack of time-varying magnetic field coefficients in the AACGM version used. However, the AACGM magnetic local time (MLT) and magnetic latitude (MLAT) are the coordinates used in our study.

- **Particle Fluxes:** The authors used two methods to calculate particle fluxes from instrument counts, also seen in [Hardy et al. \(2008\)](#). They adjusted the original observed (telemetered) counts (O) by estimating and subtracting background counts (B) to account for contamination by penetrating protons and electrons, obtaining corrected counts (C). O and B (and hence, C) are considered Poisson distributed by the authors and hence compute the associated 1 sigma uncertainty. The uncertainties in the computed fluxes arise mainly from Poisson counting statistics and telemetry compression. The relative uncertainty in the measurement of corrected counts is primarily due to Poisson uncertainty, with telemetry compression playing a minor role. The uncertainties in the differential energy and number fluxes are considered identical. Additionally, calibration uncertainty dominates the effective uncertainty under significant particle flux, estimated to be approximately 20% for electrons and 50% for ions. The total number and energy fluxes are obtained by integrating the differential fluxes over energy. As we already said Section 3.2.1, the characteristic energy is computed as the ratio of the total energy flux to the total number flux. The uncertainties are smallest under significant auroral signal and increase dramatically outside the auroral zone due to low count Poisson uncertainty.
- **Known issues:**
 - On-orbit degradation factors for F6 and F7 cannot be estimated before 1987.
 - The accuracy of the lowest energy channels is affected by ground calibration challenges and on-orbit spacecraft charging, which were not accounted for in their error analysis.
 - The low-energy ion detectors on F13 and F15 are suspected to be less sensitive than originally planned, resulting in slightly higher uncertainties, which are considered in the error analysis.
 - In January 2000, F15's low-energy ion detector became insensitive, further affecting the data quality.

3.3.2.2 Our database

The DMSP data we needed was directly obtained from NASA CDAWeb, accessible at: <https://cdaweb.gsfc.nasa.gov/pub/data/dmsp/>⁹. These data consisted of measurements from the SSJ/4 and SSJ/5 instruments, taken every second, for each of the satellites F06 to F09 and F12 to F18, in eV/cm²/sec/ster (using AACGM coordinates - Baker and Wing (1989); Shepherd (2014)). In Section 3.3.4 of the paper Bouriat et al. (2023), Table 1 provides an overview of the data availability based on years, instruments, and satellites. From the CDAWeb, we selected data points with magnetic latitudes greater than 45° or less than -45°, focusing specifically on the polar regions. The complete set of measurements amounted to 54.5 Gigabytes, comprising over 555 million lines of values¹⁰.

To ensure better compatibility with ACE data, sampled every 64 seconds, or OMNI data, sampled every minute, we needed to reduce the temporal resolution to one minute. We considered two options: either subsampling the data to retain only one measurement per minute or averaging the data at the minute level (median could also be an option). We opted for subsampling for two primary reasons: first, it preserved the original data integrity without introducing further alterations, and second, it allowed us to compare our results with existing studies, particularly the work by McGranaghan et al. (2021), who also followed this approach.

After eliminating missing data and subsampling the DMSP data, we obtained 7,121,301 measurements of total electron energy flux for analysis. It is crucial to note that this dataset serves as our target data. To create a supervised learning dataset for training our algorithm, we need to associate these "labels" with corresponding input data (in our case, solar wind data or indices). As a result, the final size of our dataset may be further reduced if there are missing input data during the sample creation process.

3.3.2.3 Histograms and Characteristics

Understanding the distribution of our training data and making adjustments accordingly are key steps in creating a quality model as we will see when analyzing the ACE dataset. Imbalanced datasets are especially likely to occur when trying to predict something infrequent, which might be the case for extreme events in our case (e.g., high-speed streams, CMEs). Rare or unusual events in our datasets (e.g., eruptive events) will be harder to model than frequent, background events (e.g., slow solar wind). However, regardless our domain, we always need to assess the distribution of our target, here, the DMSP SSJ precipitating electron observations, upgraded by Redmon et al. (2017). We will not spend too much time on this analysis as it has also been done by McGranaghan et al. (2021).

During our analysis, we plotted the overall distribution as seen in Figure 3.6, along with the distributions for each individual satellite. We also explored the distributions for each year, although we don't display them here to avoid clutter. Additionally, we examined the data distribution in the magnetic latitude and magnetic local time space. Here are several observations and decisions made based on this analysis:

- As we can observe from the distribution, the data spreads over such a large range (in powers of 10) that it is preferable to take the logarithm (base 10) of the flux data. The initial distribution exhibits heavy-tailed characteristics with a significant right-skew. Algorithms can

9. Last accessed: July 24, 2023

10. The number of measurement points (accounting for missing data in the CDAWeb dataset) were: 4,480,416 for F06; 8,320,803 for F07; 3,983,147 for F08; 878,177 for F09; 5,891,846 for F12; 135,339,042 for F13; 46,907,479 for F14; 109,423,643 for F15; 65,638,107 for F16; 91,348,090 for F17; 83,088,504 for F18

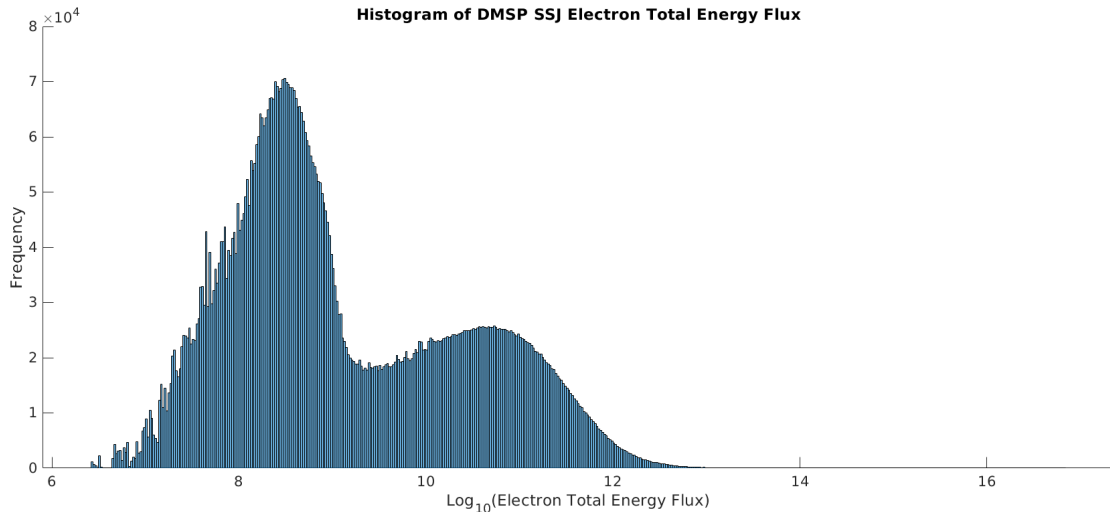


Figure 3.6 – Histogram of the DMSP SSJ Electron total energy flux, from F06 to F18, after applying a base-10 logarithm.

be sensitive to such value distributions and may underperform if not properly normalized. Therefore, we applied a base 10 logarithm to the data, resulting in the distribution shown in Figure 3.6. This logarithmic transformation stabilizes the data variance and can improve the model’s performance (Lütkepohl and Xu, 2012).

- The distribution is neither uniform nor Gaussian. As a result, it may pose a challenge for the algorithm, which might initially try to approximate it with a normal distribution. To handle this, we may need to adjust the loss function to account for extremely high values, and we may also need to experiment with the number of epochs and network complexity to capture the non-linear dependencies that seem to exist.
- By plotting the year-wise distributions of observations, we noticed a correlation between the years and the position of the first peak in the double Gaussian distribution. As expected, the first peak shifts to the left gradually between 2000 and 2010, and then gradually returns to the right after 2010. This trend corresponds to the decline and growth phases of the solar cycle (see Figure 3.7). This strong dependence between the distribution and the observed period reinforces the idea of a relationship between our inputs and outputs (which we will confirm in Section 3.3.4). To maintain dataset balance, we decided to remove data from the green region in Figure 3.7. These data points do not belong to the main red period, are quite sparse (with only a few months having scattered data), and may disproportionately represent a specific solar cycle phase (specifically, the growth phase). Moreover, they account for only 2.8% of the entire dataset, and after some experimentation, we found that their absence did not significantly impact the training quality. The dependence to the location in the solar cycle implies that we should avoid isolating a single year of data in any of the datasets (train, validation, or test).
- Since the data distribution in space (Figure 3.8) is not evenly distributed, we decided to combine the polar regions. To achieve this, we simply took the absolute value of the magnetic latitude for each measurement. It’s important to note that while this choice helps maintain a sufficiently large dataset, it may not be ideal. Combining the polar regions could result in certain effects, especially those related to the X and Y components of the IMF, canceling out and becoming unobservable. For instance, negative B_x tends to cause stronger polar rain in the northern hemisphere, while positive B_x does the same in the southern hemisphere (Fairfield and Scudder, 1985; Newell et al., 2009; Yeager and Frank, 1976). Similar opposite effects have been observed in the northern and southern lobes for B_y positive or negative

(Baker et al., 1986; Gosling et al., 1985, 1986; Newell et al., 2009).

- As for the magnetic local time histogram (Figure 3.9), it illustrates the orbit followed by each satellite, with the two main peaks consistently separated by 12 hours, corresponding to the same two hours traversed in both poles. The graph shown here represents the contribution of each satellite, with the final peak positions in the combined histogram being dominated by F13, F17, and F18. Lastly, the magnetic latitude histogram of this Figure 3.10 displays two peaks, representing the two latitudes most frequently covered by the satellite - one in the northern hemisphere (around 75° MLAT) and the other in the southern hemisphere (around -68° MLAT). The presented histogram combines both mentioned hemispheres. Figures 3.9 and 3.10 highlight the spatial bias in the DMSP measurements, which varies across different satellites. This bias needs to be considered while modeling sparsely represented areas and implies that we should avoid isolating a single satellite in any of the datasets (train, validation, or test).

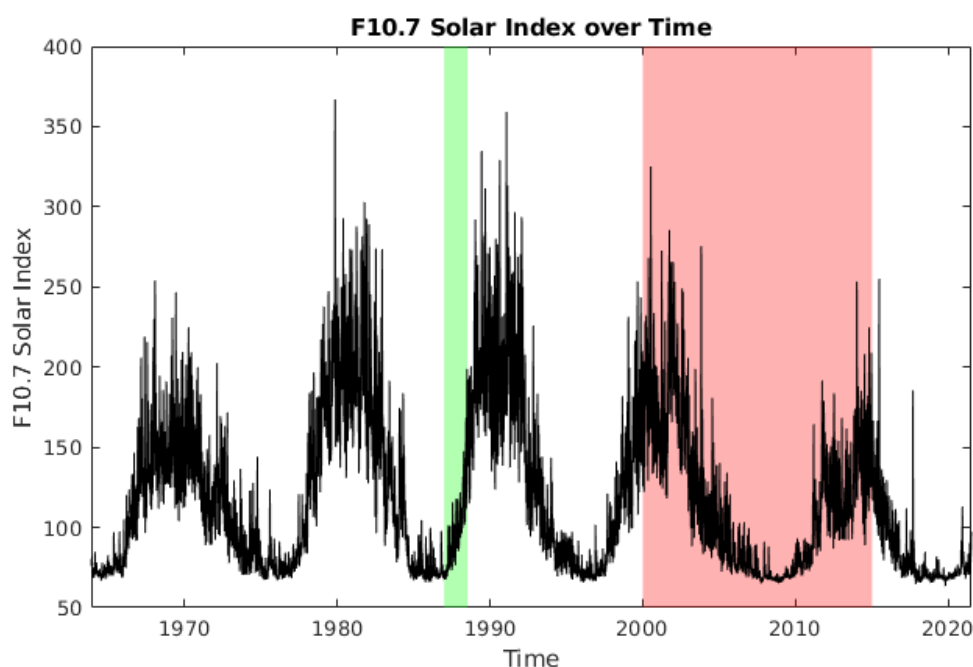


Figure 3.7 – Solar cycle represented by $F_{10.7}$ measures in solar flux units. The green region represents the measures by F06 to F09 in 1987 and 1988. The red region represents the measures by F12 to F18.

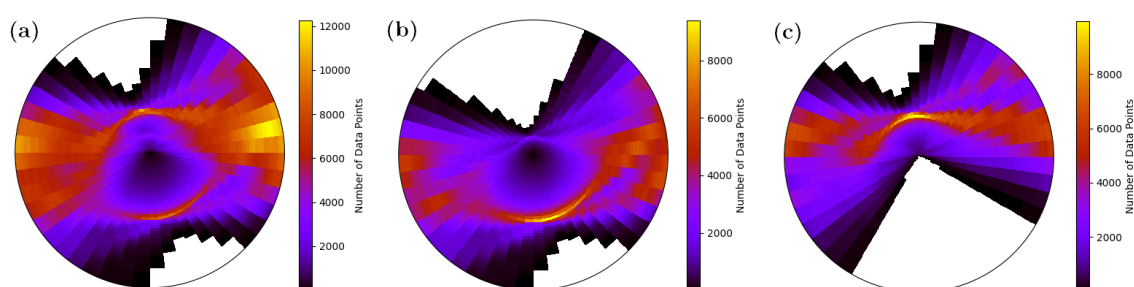


Figure 3.8 – 2D 24x24 polar grid plotting DMSP observational density with (a) the combined north and south pole, (b) the south pole data only and (c) the north pole data.

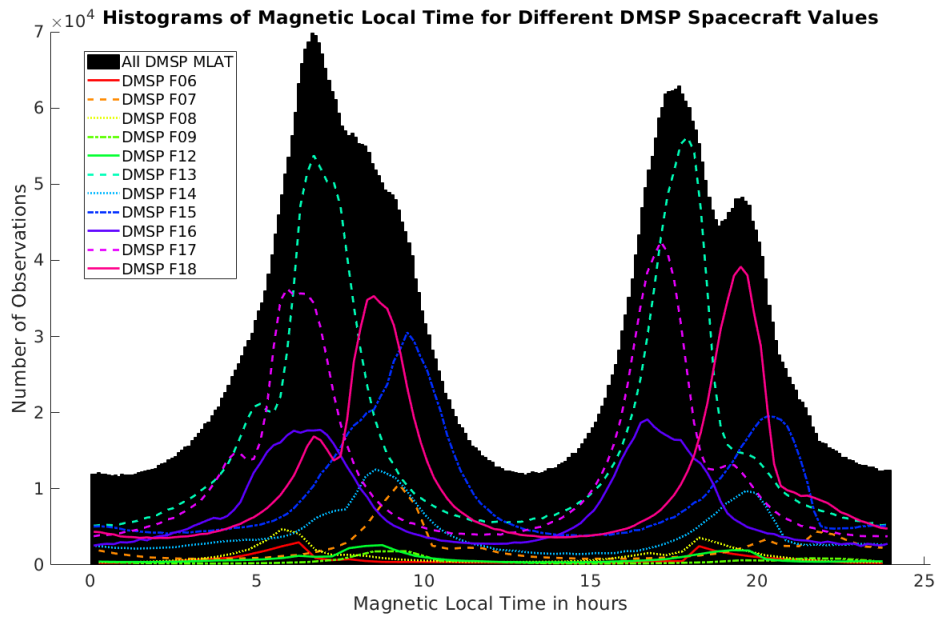


Figure 3.9 – Histogram of the Magnetic Local Time observations (in hours) for all satellites combined (black) and separately (lineplots). Differences arise from the slightly different orbits of the satellites, as well as their progressive shifts over time.

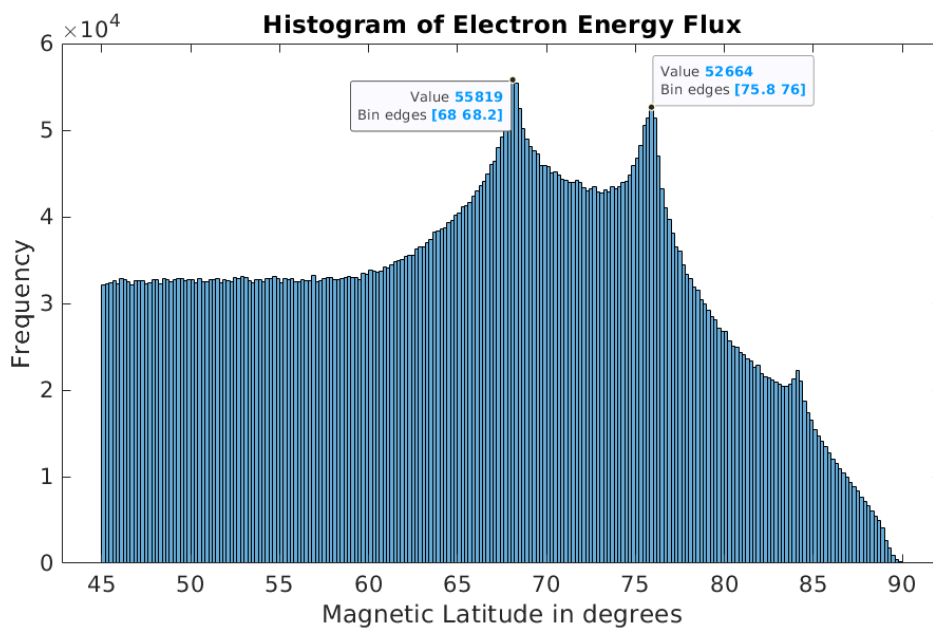


Figure 3.10 – Histogram of the Magnetic Latitude observations (in degrees).

3.3.2.4 Outliers

Outliers are data points that significantly deviate from the majority of the data points and can skew statistical measures and affect the accuracy of predictive models. A first quick observation of outliers involves plotting the data (in this case, its logarithm in base 10) over time (see Figure 3.11 where we plot the Electron Energy Flux for each satellite in time, separately). We can notice the presence of horizontal patterns for the smaller flux values. These patterns may arise due to the instrument's resolution limitations, where it cannot detect very fine variations, resulting in value rounding for the lower ranges. Notably, the positions of these patterns vary across satellites, indicating differences between sensors. Additionally, even for the same satellite (e.g., F15), the lower limit changes regularly. This might be due to regular recalibrations performed by operators, as the measurements are plotted for each satellite over time.

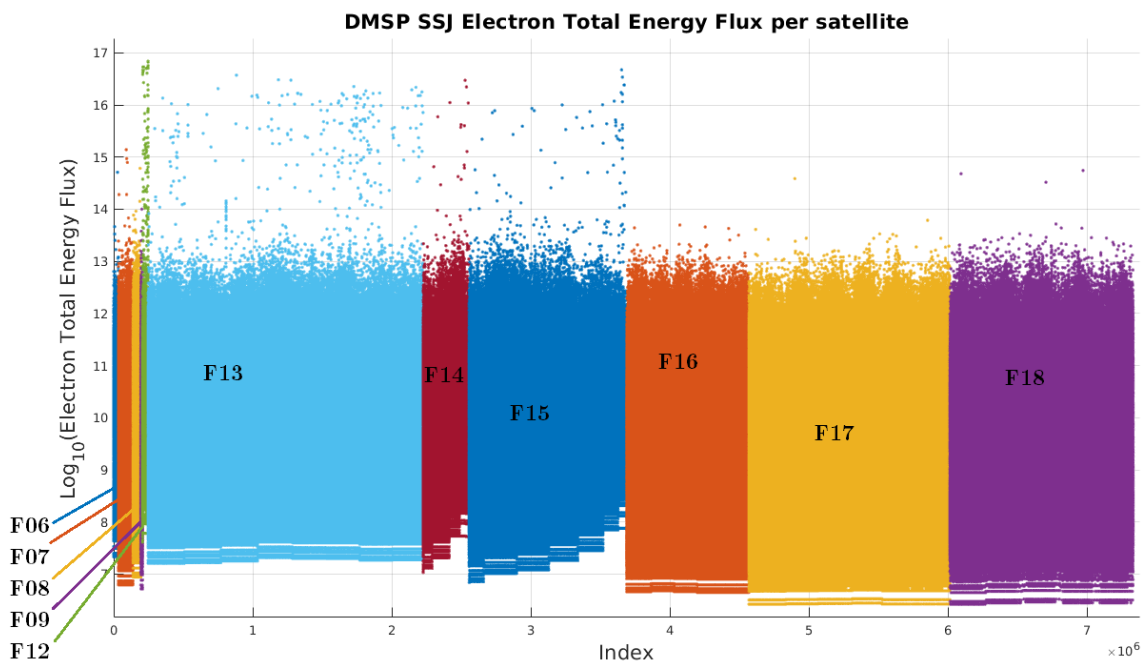


Figure 3.11 – DMSP SSJ data points ordered per satellite. Data values in $\log_{10}(eV/cm^2/sec/ster)$.

As we can see in Figure 3.11, many data points appear to be significantly higher than the "normal" range ("normal" based on visually inspecting the main area where the data is concentrated). There don't seem to be any outliers in the lower values, as expected based on the histogram. However, some very high and rare values appear to be unrealistic (McGranaghan et al., 2021), and they could potentially cause issues for the algorithm. Several methods are available to handle these outliers. Here are two examples:

- Using the *z-score*, a measure that indicates how many standard deviations a data point deviates from the mean of our dataset. This method, which also helps standardize datasets, allows us to compare data points with each other.

$$z = \frac{x - \mu}{\sigma}$$

where μ is the mean of our dataset, σ is the standard deviation of the dataset, and x is the observed data point. The z-score shows the "distance from the mean" in terms of the number of standard deviations. By setting a threshold on the absolute value of the z-score, we can identify data points that significantly deviate from the mean and consider them as outliers.

- Using the IQR (InterQuartile Range), a statistical measure of dispersion that provides information about the spread of the middle half of the data. It is calculated as the difference

between the 75th percentile ($Q3$) and the 25th percentile ($Q1$) of the data. To identify outliers using the IQR method, a common approach is to define lower and upper thresholds, generally considering data points below $Q1 - 1.5 \times IQR$ or above $Q3 + 1.5 \times IQR$ as outliers. The values falling above and below this range represent approximately 13% each.

Both of these methods are commonly used and it appears empirically that the problematic data points are those exceeding $14 \log_{10}(\text{eV}/\text{cm}^2/\text{sec})$, which represent less than 1% of the data. Therefore, we chose to use the z-score approach, which simply involves removing data points beyond a certain quantile threshold. After analysis, this threshold corresponds approximately to a z-score of 4.7. This choice is also based on empirical evidence from [McGranaghan et al. \(2021\)](#), as it corresponds to the 99.995th quantile, which, for us, is 1.4435×10^{14} eV/cm²/sec. Figure 3.12 shows the data points that were removed. This approach significantly reduces the number of lost data points (only 356 out of 7,121,301).

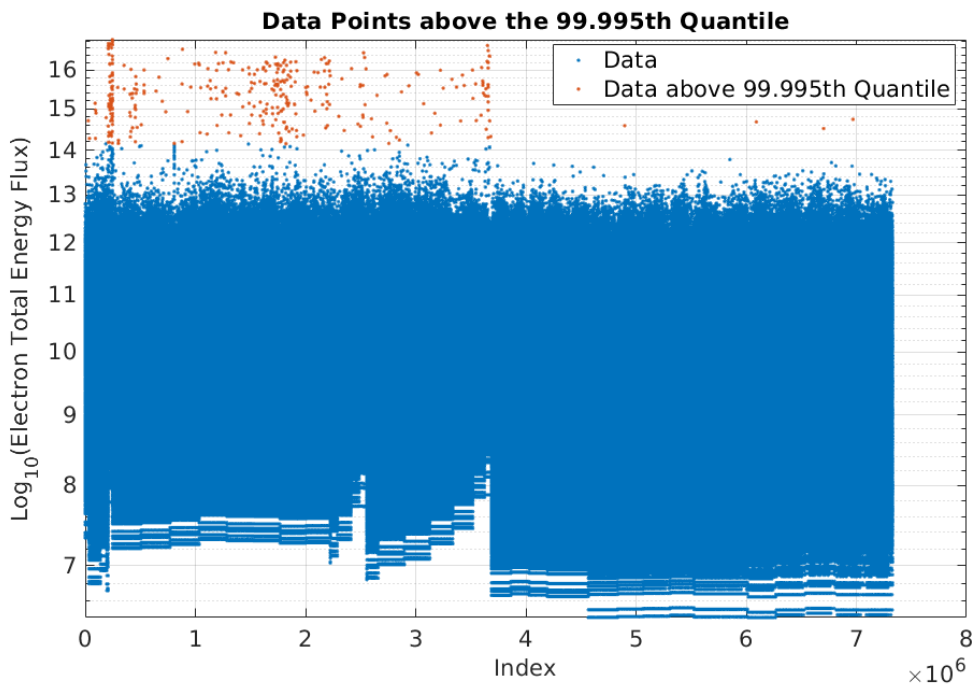


Figure 3.12 – DMSP SSJ data points in $\log_{10}(\text{eV}/\text{cm}^2/\text{sec}/\text{ster})$ ordered per satellite, and in time as in Figure 3.11, with data points above the 99.995th quantile highlighted in red.

3.3.2.5 Uncertainties

Regarding the uncertainties in the DMSP data, the best that can be done today has been addressed in [Redmon et al. \(2017\)](#) and was mentioned in the section above. Essentially, at low fluxes (counts), the estimated Signal-to-Noise Ratio (SNR) is dominated by Poisson statistics, and at higher fluxes, it is influenced by overall instrument calibration estimates. However, this does not fully account for other sources of uncertainties, such as telescope pointing, potential increased noise due to on-orbit degradation, and geophysical variations.

An interesting study by [Kilcommons et al. \(2017\)](#) focused on magnetic field measurements from DMSP. According to their findings, there are two main types of undesired fields in the data: on-off jumps and a long-period baseline. On-off jumps occur due to intermittently operating high-current equipment on the spacecraft and can be corrected using an automatic algorithm they developed. The long-period baseline is a slowly evolving, sinusoid-like variation caused by

a combination of geophysical and instrumental effects. They address it using a process called "MFIT," which fits polynomials to the baseline and subtracts them from the perturbations.

The text also mentions the challenge of quantifying the reliability of space-based observations and presents calibration efforts to reduce uncertainties. These efforts involve correcting known sources of error, including core misalignment/non-orthogonality, scalar offset, and time slips. Calibration against high-order models of the Earth's main field significantly reduces average residuals between observed and model fields, thereby improving the accuracy and reliability of the magnetometer data. These meticulous calibration efforts contribute to the overall reduction of uncertainties in the measurements, ensuring more robust and reliable space-based observations. All these considerations highlight the complexity of the datasets and the difficulty of dealing with uncertainties. Note that each instrument has its own issues and requires tailored tools. We encourage the reader to

3.3.3 High-Resolution OMNIWeb

The primary source of errors in OMNI arises from the interaction of a slow solar wind stream followed by a fast solar wind stream. OMNI tends to predict that the faster wave front will reach the bow shock nose first, but do not take into account the compression phenomena that arise. [Vokhmyanin et al. \(2019\)](#) conducted a detailed study to assess OMNI data quality (through the Pearson correlation coefficient and the precision efficiency), specifically comparing it to solar wind measurements from satellites (Geotail) passing the bow shock.

The article's findings reveal that 42% of the data exhibit excellent agreement, 33% show relatively good consistency, 10% exhibit the right trend but inaccurate absolute values, and 15% are considered very poor data. Additionally, the study establishes a correlation between the satellite's distance from the Earth-Sun axis orbiting at $L1$ and the quality of the data.

Overall, authors identified several physical factors that can introduce errors in the database, such as the size of the solar wind flow tube interacting with the magnetosphere and the evolution of solar wind structures. Additionally, they find that data measured far from the Sun-Earth line are often wrong, indicating that the database's quality improves when spacecraft are closer to this line.

Here are all the variables from the HR-OMNIWeb dataset that we kept for our study. The period :

- Year: Year of the data record.
- Day: Day of the year (1-365 or 366).
- Hour: Hour of the day (0-23).
- Minute: Minute of the hour (0-59 at the start of average).
- B_x , nT (GSE, GSM): Magnetic field component (X) in GSE and GSM coordinates.
- B_y , nT (GSE): Magnetic field component (Y) in GSE coordinates.
- B_z , nT (GSE): Magnetic field component (Z) in GSE coordinates.
- B_y , nT (GSM): Magnetic field component (Y) in GSM coordinates.
- B_z , nT (GSM): Magnetic field component (Z) in GSM coordinates.
- Flow speed, km/s: Solar wind flow speed in kilometers per second.
- Proton Density, n/cc: Proton density in particles per cubic centimeter.
- Flow pressure, nPa: Flow pressure in nanoPascal.
- AL-index, nT: AL-index in nanoTesla.

- AU-index, nT: AU-index in nanoTesla.
- SYM/H index, nT: SYM/H index in nanoTesla.

The main ideas behind these choices are: minimizing the amount of inputs, to have the main inputs used by [McGranaghan et al. \(2021\)](#) and to use variables that can be found when using the ACE (or other L1-located) satellites. Beside the IMF, speed, density and pressure, we decided to also take AL, AU (and by linearity, AE and AO) and the SYM/H index to account for geomagnetic storms and substorms [Vorobjev et al. \(2013\)](#). However, as we will see, a lot of tests have been realised without them. The idea behind is that if we obtain good results with only BSN's data, this will give us a small forecasting capability, equivalent to the delay between what is happening at the BSN and the near-Earth environment.

As said before, we favored this data over the ACE dataset for several reasons: ACE dataset had too many missing values, the OMNI dataset combines data from several other satellites, the "solar wind observations at ACE are not a good indication of the solar wind reaching the Earth" ([Ashour-Abdalla et al., 2008](#)). However, ultimately, our code should be able to use ACE (or other) datasets allowing to avoid the debatable OMNIweb processing ([King and Papitashvili, 2006](#); [Vokhmyanin et al., 2019](#)) and to have an even larger delay between inputs and outputs, increasing the forecasting performance. For now, this choice remains subjective: the ACE dataset is free of heavy process like OMNIweb's one (and several satellites combined makes data inhomogeneous, which makes the analysis much trickier), but taking ACE dataset implies that the model performs the propagation.

However, it's important to mention that for the IMF components, and the speed and density of the solar wind, our conclusions remain consistent with those presented in Section 3.3.1. The distributions exhibit similar shapes. As for the behavior of the solar wind pressure, no detailed analysis uncovers any attributes that might have been overlooked after reviewing the speed and density of the solar wind. Notably, the "ram" pressure is proportional to the product of density and the square of speed. The observed distribution doesn't reveal anomalies or additional insights that challenge its application. It follows a log-normal pattern with a pronounced peak and an extensive tail to the right. After applying a base 10 logarithm, this distribution becomes a clean Gaussian, a transformation we employ for these three data sets in our study. The primary issue we addressed with OMNI was the abundance of missing data, which has been significantly reduced compared to ACE data.

Analyzing the OMNI data over the same timeframe as the DMSP data (from late 2000 to the end of 2014), we found:

- For the X, Y, and Z components of the IMF in both GSM and GSE coordinates, 6.73% of the data is missing, which translates to 495,557 missing data points out of 7,368,089.
- For the solar wind flow speed, 18.87% of the data is missing, or 1,390,458 missing data points out of 7,368,089.
- The solar wind proton density is missing 18.87% of its data, again 1,390,458 data points out of 7,368,089.
- The solar wind pressure has 18.41% of its data missing, or 1,356,399 missing data points out of 7,368,089.
- AL, AU, and SYM/H datasets have no missing entries.

Given this analysis, we opted to implement an interpolation method to account for certain missing data. This approach is inherent to the TCN's operation and isn't employed in a fully-connected neural network. For the TCN, input data should be sequential time measurements (in our case, 30-minute intervals, thus 30 data points, per parameter). A single missing data

point in this series renders the entire sample unusable. Hence, one absent piece of data can effectively diminish the sample size by 30. During hyperparameter selection, one can opt for an OMNI dataset where gaps (i.e., sequences of missing data) have been interpolated using a selected method. It's ideal to fill smaller gaps (1 to 4 data points) to maximize sample availability. However, there's an increasing risk of inaccurate interpolation as the gap size expands (Bouriat et al., 2022). While our code is adaptable and allows different interpolation methods, we consistently utilized the *Piecewise Cubic Hermite Interpolating Polynomial* ("pchip")¹¹. This method guarantees shape-preserving piecewise cubic interpolation (Yang and Huiyan, 1996). A cubic polynomial is a polynomial of degree 3, hence its terms reach up to x^3 . Hermite polynomials are specialized types that let us define both function values and derivatives at segment endpoints. This ensures smooth, continuous curves, offering superior results, especially when interpolating data characteristics of the solar wind. As a preparatory step for our research, we curated two datasets using this method, filling gaps of size 1 (first set) and those up to size 4 (second set). The subsequent missing data metrics are:

- For the X, Y, and Z components of the IMF in both GSM and GSE coordinates: from an initial 6.73% missing data, it's reduced to 5.30% with gaps of size 1 interpolated, and further to 4.34% when gaps up to size 4 are interpolated.
- For the solar wind flow speed: from an initial 18.87% missing data, it's reduced to 15.25% with gaps of size 1 interpolated, and down to 9.08% when gaps up to size 4 are interpolated.
- For the solar wind proton density: similarly, from an initial 18.87% missing data, it drops to 15.25% with gaps of size 1 interpolated, and to 9.08% with gaps up to size 4 interpolated.
- For the solar wind pressure: from an initial 18.41% missing data, it decreases to 14.89% with gaps of size 1 interpolated, and to 8.79% when interpolating gaps up to size 4.

The literature presents numerous intriguing interpolation methods for addressing gaps in solar wind data (e.g., singular spectrum analysis from Kondrashov et al. (2010) or the use of the Lomb-Scargle periodogram from Hocke and Kämpfer (2009)). While we opted for the simpler pchip interpolation due to time constraints, a comprehensive exploration of these methods is warranted for future stages of this project. Further investigation is essential to determine the most suitable approach.

Extensive statistical analyses already exist on AL, AU and SYM/H (Amariutei and Ganushkina, 2012; Bergin et al., 2023; Makarov, 2022; Nakamura et al., 2015; Pulkkinen et al., 2011; Wanliss and Showalter, 2006) so we will not spend time on it. The only remark we can make on the corresponding distributions (seen Figure 3.13) is that they should not represent difficulties like the bimodal B_X or B_Y or the complex DMSP SSJ distributions. AL and AU distributions both look like log-normal distributions (with a right and left skewed distribution) and the SYM-H index has an approximately normal distribution.

We will now perform an analysis of the relationship between both inputs and outputs, namely OMNIWeb HRO and DMSP measurements, focusing only on the data that are in the interplanetary medium (IMF, speed, density and pressure). It is very important for data scientists to be sure that there is a causal effect between the two. Proving the existence of this causal effect is the purpose of the next section.

3.3.4 Input-Output Relationship: Justification for AI Implementation

The proper identification and understanding of input-output relationships play an essential role in ensuring accurate and meaningful analysis. Before delving into any AI-driven investigation, it is essential to observe and comprehend the relationship between the input parameters and the

11. <https://www.mathworks.com/help/matlab/ref/pchip.html>

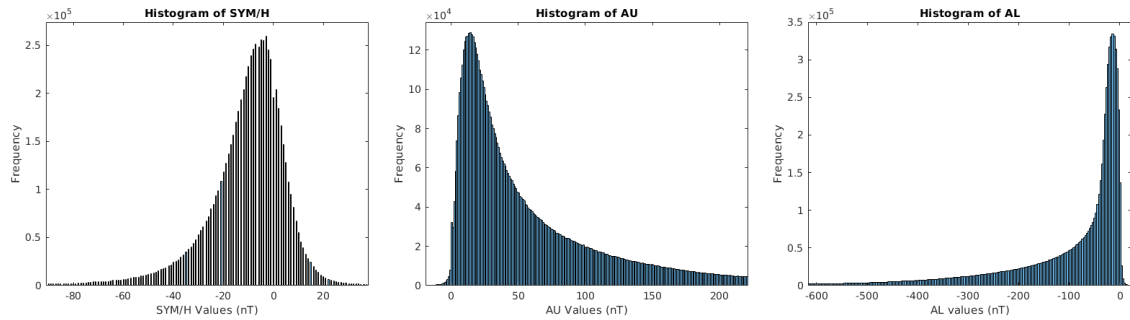


Figure 3.13 – Histograms of SYM-H (nT), AL(nT) and AU (nT).

corresponding output, be it apparent to the human eye, suggested by domain experts, or inferred through rigorous dataset testing. This subsection introduces our paper that addresses precisely this fundamental aspect, focusing on the analysis of the relationship between Lagrange 1 solar wind data and DMSP SSJ electron total energy flux.

In Chapter 1 of this PhD, evidence has been provided that there exists a correlation between solar wind parameters and electron low-energy fluxes measured by DMSP in LEO on the magnetic poles. However, acknowledging the significance of thoroughly understanding this relationship, the present study endeavors to further investigate and expand on the identified link between solar wind conditions and precipitated electron.

The paper shown below, entitled "Electron Aurora and Polar Rain dependencies on Solar Wind Parameters" presents the findings of an extensive data analysis conducted on DMSP SSJ/4/5 data with Dr. Simon Wing from John Hopkins University and Dr. Mathieu Barthélémy. The primary objective of this investigation is to characterize the relationship between solar wind parameters, including velocity, density, dynamic pressure, and B_z of the interplanetary magnetic field, and the corresponding median of electron energy flux for each magnetic latitude-magnetic local time (MLAT-MLT) pair. By examining this relationship, the study aims to shed light on the implications of high solar wind velocity, density, and pressure on electron energy flux variations, polar rain energy fluxes, and oval shape.

JGR Space Physics



RESEARCH ARTICLE

10.1029/2023JA031598

Electron Aurora and Polar Rain Dependencies on Solar Wind Parameters

S. Bouriat^{1,2,3} , S. Wing⁴ , and M. Barthélémy^{2,3}¹SpaceAble, Paris, France, ²CNRS, IPAG, University of Grenoble Alpes, Grenoble, France, ³CSUG, University of Grenoble Alpes, Grenoble, France, ⁴Applied Physics Lab, Johns Hopkins University, Baltimore, MD, USA

Key Points:

- We produced 2D maps of the electron precipitation in the auroral oval and polar cap
- Electron flux peaks at midnight to the prenoon sector; consistent with curvature and gradient drifts and very low frequency wave induced pitch angle scattering
- Polar rain increases with IMF B_z and the precipitating electron flux in the auroral oval increases as the solar wind velocity enhances

Correspondence to:

S. Bouriat,
simon.bouriat@spaceable.eu

Citation:

Bouriat, S., Wing, S., & Barthélémy, M. (2023). Electron aurora and polar rain dependencies on solar wind parameters. *Journal of Geophysical Research: Space Physics*, 128, e2023JA031598. <https://doi.org/10.1029/2023JA031598>Received 22 APR 2023
Accepted 17 AUG 2023

Abstract Data analysis was performed using 17 years of Defense Meteorological Satellite Program SSJ/4/5 data to characterize the relations between the solar wind parameters and the electron low-energy fluxes measured on both magnetic poles (magnetic latitude above 55°). Inputs are solar wind velocity, density, dynamic pressure, and B_z of the interplanetary magnetic field. The median of electron energy flux for each MLAT-MLT pair has been computed for given values of solar wind condition parameters. Results highlight that high velocity, density or pressure implies higher energy flux overall, higher polar rain energy fluxes, and wider nightside oval. There seems to be a positive correlation between polar rain and solar wind density, contrary to a previous study. As a function of B_z , the oval width has a “U” shape and the polar cap activity a “V” shape, with their minimum at B_z around zero.

Plain Language Summary Auroral precipitations are indicators of the magnetosphere-ionosphere coupling. Here, data analysis was performed using 17 years of Defense Meteorological Satellite Program SSJ/4/5 data to characterize the relations between the solar wind condition parameters and the electron low-energy fluxes measured on both magnetic poles (magnetic latitude above 55°). Inputs are solar wind velocity, density, dynamic pressure, and B_z of the interplanetary magnetic field, from NASA's OMNIWeb database. Median of electron energy flux for each MLAT-MLT pair has been computed for given values of solar wind parameters. Results highlight that high velocity, density or pressure implies higher energy flux overall, higher polar rain energy fluxes, and wider nightside oval. There seems to be a positive correlation between polar rain and solar wind density, contrary to a previous study. As a function of B_z , the oval width has a “U” shape and the polar cap activity a “V” shape, with their minimum at B_z around zero. This work is a unique contribution to the field as it put together a global picture of the electron precipitation that scientific community can use as a reference for how the oval and polar rain vary at different magnetic local time (MLTs) values as a function of solar wind parameters.

1. Introduction

The aurora, also known as the northern lights, has fascinated people since ancient times. The area above the Earth's poles where auroras are visible is commonly referred to as the auroral zone. However, it was not until the early 1960s that scientists realized that this region has a distinct shape resembling a ring or an oval. As a result, it was given the name auroral oval (Feldstein, 2016, see review).

The aurora light emission is primarily caused by ions and electrons. These particles originate from the solar wind and the magnetosphere and interact with the neutral components of the upper atmosphere, causing them to ionize and excite. The majority of the precipitating particles on the dayside oval, including the cusp, mantle, low-latitude boundary layer (LLBL), open-field line LLBL (open-LLBL), and high-latitude boundary layer regions, come from the solar wind (Fujimoto et al., 1998; Lockwood et al., 1993; Lyons et al., 1994; Newell & Meng, 1992; Shi et al., 2013, 2009; Wing et al., 1996, 2001), while those at the equatorward portion of the oval such as boundary plasma sheet (BPS) and central plasma sheet (CPS) come from the magnetosphere (Newell & Meng, 1992; Newell et al., 2004). On the nightside oval, the majority of precipitating particles are magnetospheric in origin (Newell et al., 2004, 1991). The dayside and nightside particle precipitation regions have different characteristics, but there are similarities and connections between the two regions. For example, the particles in the BPS and CPS are the nightside plasma sheet particles that have curvature and gradient-drifted to the dayside and then precipitate (Newell & Meng, 1992; Wing et al., 2023). However, it should be noted that these magnetospheric particles whether on the dayside or nightside, originate from solar

© 2023. The Authors.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial License](https://creativecommons.org/licenses/by-nc/4.0/), which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

wind plasma that has entered the magnetosphere several hours to over 15 hr earlier, depending on the location within the magnetosphere and the prevailing solar wind conditions (Berchem et al., 2014, 2016; Borovsky et al., 1998; Sorathia et al., 2019; Wing et al., 2014, 2006, 2005; Wing & Newell, 1998). Hence, the particle precipitation in the polar region, whether on open or closed field lines, should have dependencies on the solar wind properties.

The position, structure, intensity, and latitudinal width of the auroral oval can vary significantly due to fluctuations in the solar wind. Previous studies investigating the drivers of auroral oval variability have primarily focused on the micro or mesoscale structures within the auroral oval, typically limited to a few magnetic local time (MLT) (Gabrielse et al., 2021; Johnson et al., 2021; Sergeev et al., 2004; Zhu et al., 2018). Fewer studies examined the global scale of electron precipitation within the entire auroral oval or polar region across all MLTs. For example, Wing et al. (2013) studied the evolution of the electron aurora oval as a function of the substorm phase at the global scale. Likewise, Newell, Sotirelis, and Wing (2009) studied the dependencies of the electron aurora on the strength of solar wind driving, as determined by solar wind-magnetosphere coupling functions, and seasonal variations at a global scale. Seasonal dependence of global-scale auroral particle precipitation has been studied both with models (Wiltberger et al., 2009) and with observations from the Defense Meteorological Satellite Program (DMSP) satellites (Newell et al., 2010). Finally, Liou et al. (2007) presented a case study of the auroral oval response to a long duration of high solar wind dynamic pressure.

The precipitating electrons are the field-aligned plasma sheet electrons that are in the loss cone and precipitate in the ionosphere. Most of these electrons would not mirror back to the magnetosphere. However, in the magnetosphere, the electrons in the loss cone can be replenished by pitch angle scattering of the non-field aligned electrons through wave-electron interactions. Energy exchange between electrons and waves can occur when the wave frequency and the frequency of the electron periodic motion match, resulting in violation of adiabatic invariant and diffusion in phase space, which can effectively alter the electrons' pitch angles. Very low frequency (VLF) whistler-mode chorus waves have been proposed as a leading mechanism for pitch angle scattering of the plasma sheet electrons (Ni et al., 2016; Summers et al., 1998). Thorne (2010) showed that the VLF whistler-mode chorus waves are particularly active from midnight to noon (see their Figure 1). The local time distribution of these waves and their intensities should have an impact on the local time distribution and intensity of the precipitating electrons, particularly diffuse electrons.

The present study investigates statistically the global scale of the position, structure, intensity, and latitudinal width of the precipitating electrons within the entire auroral oval and polar cap due to solar wind velocity, density, dynamic pressure, and the B_z component of the interplanetary magnetic field (IMF).

2. Data and Methodology

2.1. DMSP Satellites and Data

The DMSP satellites are in Sun-synchronous nearly circular polar orbits at about 845 km altitude, with orbital inclinations of 98.7°. The areas with the least amount of coverage occur around post-noon and post-midnight local time, with the exception of regions at high magnetic latitudes where the coverage is more uniform. A significant number of measurements obtained from DMSP are concentrated within the intervals of 5–10 MLT and 16–21 MLT (as seen in Figures A5–A8).

The SSJ/4 and SSJ/5 instruments (Special Sensor Precipitating Electron and Ion Spectrometer [SESS]) are respectively part of the Space Environment Monitor and SESS packages. They measure the flux of precipitating electrons and ions through a curved plate electrostatic analyzer for electrons protons and alpha-particles in the energy range 0.03–30 keV, with one complete spectrum each obtained per second (Hardy et al., 1984). The SSJ/4 instrument was deployed on spacecraft belonging to the DMSP series, specifically from F6 to F15, while the SSJ/5 instrument was deployed on satellites F16–F19. The satellites are three-axis stabilized, and the detector apertures always point toward local zenith. At the latitudes of interest in this paper, this means that only highly field-aligned particles well within the atmospheric loss cone are observed.

Data from the SSJ/4 and SSJ/5 were used to highlight changes in polar auroras. The satellites (and corresponding years and instrument) were available on the Coordinated Data Analysis Web (CDAWeb) interface of the Goddard Space Flight Center, spanning 17 different years as shown in Table 1. The data collected include the

Table 1

This Table Shows the Available Data Years for Each Defense Meteorological Satellite Program Satellite

	1987	1988	...	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
SSJ4																		
F06																		
F07																		
F08																		
F09																		
F12																		
F13																		
F14																		
F15																		
SSJ5																		
F16																		
F17																		
F18																		

Note. Green cells indicate data availability.

1 s resolution total electron energy flux (in $\text{eV}/\text{cm}^2/\text{ster}/\text{s}$), which is obtained by integrating the differential energy fluxes measured by DMSP across the energy range. Data are displayed in AACGM coordinates (Baker & Wing, 1989).

As pointed out by Newell, Sotirelis, and Wing (2009), the operational lifespan of several detectors has led to degradation of sensitivity and consequently, decreased reliability in boundary identification in recent years. Furthermore, the low-energy ion head of certain SSJ/4 detectors has been launched in a sub-optimal condition, thereby diminishing the quality of ion precipitation data.

2.2. Solar Wind Data

The solar wind data gathered on NASA's OMNIWeb database are not in situ measurements. They consist of high-resolution (1 min) solar wind and IMF data at the Earth's magnetopause: data from ACE, WIND, IMP 8, and Geotail spacecraft that have been processed and time-shifted to the Bow Shock Nose (BSN) (King & Papitashvili, 2006). The solar wind and IMF data used in this paper are only the IMF B_z in GSM coordinates (in nT), the solar wind proton density (in cm^{-3}), velocity, (in $\text{km}\cdot\text{s}^{-1}$), and ram pressure (in nPa and derived from particles' densities, speeds, and masses $\propto n v^2$).

In this study, we combined the data from both the north and south hemispheres, as we considered it more appropriate for our research objectives. Therefore, we removed the X - and Y -components of the magnetic field, assuming that their effects are symmetric in both hemispheres and will cancel each other out.

2.3. Methodology

As previously mentioned, we combined data from the North and South poles (above and below 45° and -45° MLAT, respectively) and averaged it instead of subsampling to maintain a 1 min resolution. We obtained 1 min resolution solar wind, magnetic field, and plasma data from the OMNI interface at Earth's BSN and applied a 30 min wide moving average. Specifically, we replaced each data point at time T_0 with the mean value of all data points in the time window $[T_0 - 29 : T_0]$. We required a minimum of 10 points in the window for us to accept the average and discard other data points.

We binned the four solar parameters (solar wind flow speed, density, temperature, and dynamic pressure) into eight bins each, with each bin having specified limits. Here are the limits for each parameter:

- IMF B_z in GSM coordinates: $[-\infty, -9, -6, -3, 0, 3, 6, 9, +\infty]$ nT.
- Solar wind pressure: $[0, 1, 2, 3, 4, 5, 8, 15, +\infty]$ nPa.
- Solar wind proton density: $[0, 2, 4, 6, 8, 10, 15, 20, +\infty]$ cm^{-3} .
- Solar wind flow speed: $[0, 300, 400, 500, 600, 700, 800, 900, +\infty]$ $\text{km}\cdot\text{s}^{-1}$.

In these limits, the right endpoint is not included. For example, the first bin for solar wind flow speed is from 0 to 300 km·s⁻¹, including all values between 0 and 300 km·s⁻¹, but not including 300 km·s⁻¹.

Finally, we grouped DMSP data falling into a given bin (e.g., all DMSP measurements for B_z between -9 and -6 nT). Then, we computed the median of these DMSP data for all given MLAT-MLT pairs. We plotted the resulting medians in a polar plot, where each compartment represents a $1^\circ(\text{MLAT}) \times 1 \text{ hr}(\text{MLT})$ area and shows the corresponding median value of DMSP measurements using a base-10 logarithm for clarity. Essentially, we generated a polar plot of DMSP measurements for each instance where the average of a solar wind driver in the past 30 min fell within a particular range (e.g., when the average of B_z in the past 30 min was between -9 and -6 nT—see Figure 4b).

The outcome was one graph per solar wind parameter per bin, with each compartment on the polar graphs measuring $1^\circ(\text{MLAT}) \times 1 \text{ hr}(\text{MLT})$ and displaying the corresponding median value of DMSP measurements. To improve clarity, the graphs shown in Figures 1–4 display only MLAT values above 55° .

Throughout the following descriptions, when we refer to the electron energy flux, we are referring to the base-10 logarithm of the total electron energy flux, which represents the amount of energy carried by the electrons and is measured in eV/cm²/ster/s.

3. Results

3.1. Dependence of Electron Energy Flux on Solar Wind Speed

Figure 1 shows the base-10 logarithm of the total electron energy flux within the entire auroral oval as a function of solar wind speed. Several things are worth noting. First, the electron energy flux is higher around midnight to noon than from noon to midnight. This can be explained by the electron gradient and curvature drifts and the VLF whistler-mode chorus waves. Electrons coming earthwards following reconnections in the magnetotail would also curvature and gradient drift eastward toward dawn. Electrons that are field-aligned (pitch angle 0°) are quickly lost through precipitation, but the field-aligned electrons are replenished by pitch angle scattering. The leading mechanism for pitch angle scattering is the electron interactions with the VLF whistler-mode chorus waves, which have been shown to be active at midnight-noon local time (Ni et al., 2016; Reeves et al., 2009; Summers et al., 1998; Thorne, 2010). Once we enter the post-noon region, the whistler-mode chorus wave's activity is reduced and we see less pitch angle scattering, and hence a reduction in electron energy flux. The MLT profile and the dawn-dusk asymmetry seen in Figure 1 are similar to those of the diffuse electron precipitation in Wing et al. (2013), which is not surprising because most of the electrons are diffuse electrons.

As the solar wind speed increases, the electron energy flux also increases, from the smallest velocities to the largest velocities. This relationship can be attributed to the increased occurrence of substorms and subsequent wave activities, which would increase with higher solar wind speeds (Newell et al., 2016). Substorm injections would energize and transport particles from the plasma sheet inward, resulting in ion temperature anisotropy and the growth of VLF whistler-mode chorus waves. These waves, in turn, can enhance electron pitch angle scattering. Additionally, Figure 1 shows that the auroral oval extends equatorward to smaller latitudes as the solar wind velocity increases.

Understanding the increase in the width of the auroral oval is more challenging since it is not consistent across different MLT regions. By setting the boundary of the oval to 10 eV/cm²/ster/s, we can see that the approximate width of the auroral oval generally increases with as the solar wind velocity enhances, except for the 11–16 MLT region, where it appears to remain roughly the same and increase only in the last two panels. Note that whenever we mention MLT regions in the format $X-X'$, we include the X' bin (meaning including the zone between line X' and $X' + 1$). Figure A1c in appendix confirms several trends: the oval width in 20–7 MLT region mostly displays a linear increase as the solar wind velocity enhances, the 17–19 MLT region somehow shows an exponential increase, and the 8–10 MLT region exhibits a rapid increase in the oval width from panel A to panel C, followed by a slower increase from panel C to panel F. Finally, computing the oval width over all MLT regions except for region 1 MLT (due to missing values), we observe an increase from an average width of 6.4 MLAT for a median speed of 288 km/s (panel A) to 11.7 MLAT for a median speed of 733 km/s (panel F), corresponding to an 83% increase.

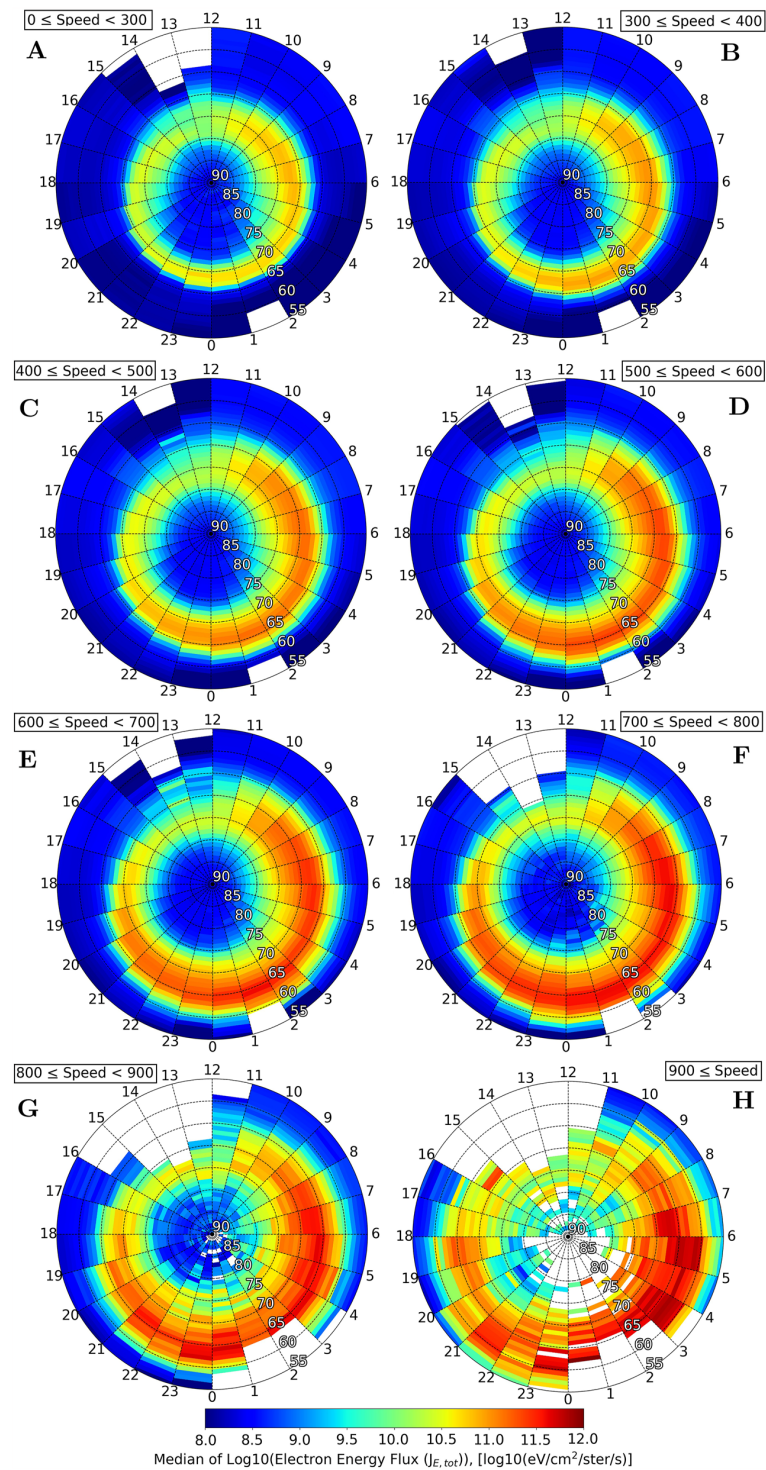


Figure 1. Median of $\text{Log}_{10}(J_{E,\text{tot}})$ [eV/cm²/ster/s] for solar wind speed respectively between 0 and 300 km·s⁻¹ (a), 300 and 400 km·s⁻¹ (b), 400 and 500 km·s⁻¹ (c), 500 and 600 km·s⁻¹ (d), 600 and 700 km·s⁻¹ (e), 700 and 800 km·s⁻¹ (f), 800 and 900 km·s⁻¹ (g), 900 km·s⁻¹ and infinity (h).

The configuration of the magnetic field lines within the polar cap (poleward of the auroral oval) is believed to be open, that is, connected to the solar wind. Also we set the boundary of the oval to 10 eV/cm²/ster/s, the polar cap is the region of MLAT above the poleward boundary of the oval, extending from this limit to 90 MLAT. The electron precipitation in the polar cap, which is known as polar rain, originates from the suprathermal (strahl) component of the solar wind electrons (Fairfield & Scudder, 1985). Figure 1 and Figure A1b show that the polar rain electron flux appears to increase with the solar wind velocity, but this effect is not strong. However, we have taken great care to ensure the reliability of the data by retaining only the bins with a minimum of 10 data points. Additionally, we have computed the median of these data points to mitigate any influence from individual anomalous events.

In the polar cap, few solar wind ions can enter the magnetosphere and hence a parallel electric field or potential arises to prevent most solar wind core electrons from entering in order to maintain charge quasi-neutrality (Fairfield et al., 2008; Wing et al., 2015, 1996). However, the solar wind suprathermal electrons (strahl), having higher temperature/energy, can overcome the parallel electric potential and enter the magnetosphere. Borovsky (2021) showed that solar wind velocity is correlated with solar wind strahl. This correlation is consistent with the increase of polar cap electron precipitation (polar rain) energy flux with solar wind velocity seen in Figure 1, although the trend is not strong.

3.2. Dependence of Electron Energy Flux on Solar Wind Density

Figure 2 depicts the relationship between the base-10 logarithm of the electron energy fluxes and the solar wind density. Several observations are evident from Figure 2. First, the same asymmetry observed in Figure 1 is also evident in Figure 2. Specifically, the energy flux is higher at midnight-noon compared to noon-midnight. This can be attributed to the electron curvature and gradient drifts, as well as VLF whistler mode chorus waves, as discussed in Section 3.1.

Second, it appears that an increase in the solar wind density would increase the energy flux in the polar cap, but this effect is weak. When plotting this increase, it seems to behave as a logarithmic function. Figure 2 and Figure A2b show that the mean value of energy flux in the polar cap for panel A is lower than for panel H. When computing it, we see an increase from 8.82 to 9.09 log₁₀(eV/cm²/ster/s). Borovsky (2021) showed that solar wind density is correlated with solar wind strahl. The increase of polar cap electron precipitation (polar rain) energy flux with solar wind density seen in Figure 2 is consistent with Borovsky (2021) result. Riehl and Hardy (1986) analyzed 262 DMSP passes and found no correlation between polar rain flux and solar wind density. It is not clear why they found no correlation, but their study used a much smaller data set than the present study. One of the explanations given for the Riehl and Hardy (1986) result was that the origin of the polar rain is the solar wind suprathermal electrons, rather than the solar wind core electrons and hence little or no correlation can be expected (Newell, Liou, & Wilson, 2009). However, the positive correlation between solar wind density and strahl intensity found in the more recent study by Borovsky (2021) can help explain why the positive correlation between polar rain electron fluxes and the solar wind density can be expected albeit this effect is rather weak.

However, there is an anomaly. It can be seen that going from panels A (solar wind density 0–2 cm⁻³) to B (solar wind density 2–4 cm⁻³), the polar cap energy flux actually decreases slightly from 8.82 to 8.76 eV/cm²/ster/s on average rather than increases. This seems to be due to the depression in energy flux in the polar cap observed in panel B between 20:00 and 02:00 MLT. It is not clear what causes this anomaly.

Third, as the solar wind density increases, the energy flux in the auroral oval also increases, but the change is not as significant as it is with the velocity. In the case of low density, as in panel A, there is already a high energy flux in the auroral oval. This can be explained by the fact that solar wind density is generally negatively correlated with solar wind velocity, so low solar wind density would correspond to high solar wind velocity (Borovsky, 2020; Maggiolo et al., 2017; Wing et al., 2016, 2022). However, high solar wind density can also result in high solar wind dynamic pressure, which can lead to storms and substorms, thus creating a competing effect that can be observed in this study.

Fourth, as the solar wind density increases, the behavior of the auroral oval width becomes more complex. While Figure 2 provides a visual representation, the full range of variations can be better understood by referring to Figure A2c. Observations show that the 20–7 MLT region experiences a slight decrease in width, followed by an increase. The 8–13 MLT region also shows a small decrease in width. On the other hand, the width in the

14–19 MLT region appears to increase as the solar wind density enhances. Overall, the thickness of the auroral oval seems to transfer from the morning-to-noon region to the evening-to-midnight region, as observed in Figure 2.

3.3. Dependence of Electron Energy Flux on Solar Wind Dynamic Pressure

Figure 3 demonstrates that the electron energy flux increases almost monotonically as the solar wind dynamic pressure enhances. This result is expected as dynamic pressure (or so-called ram pressure) is proportional to nv^2 , where n is the solar wind density and v is the solar wind velocity. This pattern for dynamic pressure is consistent with what we have observed for both density and velocity, as shown in Figures A1a and A2a in appendix. In particular, we can see in Figure 3 that the electron energy flux starts at higher values in panel A than in panel A of Figure 1, and does not reach the high values observed in panel H of Figure 1 for high solar wind velocity. Density and velocity anti-correlation can explain this trend.

The effect of dynamic pressure on the width of the auroral oval is also apparent in Figure 3, as we can observe an increase in width and a significant extension of the oval equatorward from panels A to H.

Finally, Figure 3 shows that the polar rain increases as the dynamic pressure enhances. This trend is even clearer than that for solar wind velocity (Figure 1) and solar wind density (Figure 2). This is perhaps unsurprising since the energy fluxes of polar rain increase with n and v , as discussed in Sections 3.1 and 3.2. This increase is further amplified by the dynamic pressure of the solar wind.

3.4. Dependence of Electron Energy Flux on B_z GSM

Figure 4 displays the electron energy flux as a function of the southward component of the IMF B_z . Studies have shown that as the southward B_z increases, the magnetosphere can become more active due to substorms or storms, resulting in particle injections and energization of the particle population in the magnetotail (Kamide et al., 1977; Wing & Johnson, 2009). In Figure 4, a clear dependence can be observed between the north-south component of the IMF and the shapes of both the oval and the polar cap. The polar cap area is delimited by the open-closed field line boundary (OCB), and the B_z component of the IMF is often responsible for magnetic reconnections that impact the OCB's shape (Tulegenov et al., 2023). As demonstrated here, a large southward component of the IMF generally causes the boundary to move equatorward, while a northward component moves the boundary poleward (Tulegenov et al., 2023). Moreover, for positive B_z , a boundary layer can form poleward of the cusp (Shi et al., 2013, 2009), which can shift the poleward edge of the oval to higher latitudes. On the other hand, a large negative B_z corresponds to a lower-latitude average position for the oval (Burch, 1979) and a higher activity in the polar cap. The observations suggest that substorms occurring in isolation or during storms can increase the width and intensity of the oval, as shown in Figure 4, panel A. When the width of the oval is plotted from panels A to H (see Appendix A, Figure A4c), a “U” shape is suggested, with the exception of the 8–16 MLT region, which shows an increasing trend. On average, the width of the oval appears to decrease from panels A to D, with a reduction of the poleward boundary, and then increase again from panels E to H, with an extension of the poleward boundary. From panel A to panel H, the oval evolves from a very asymmetric shape with a thin oval on the dayside and a wide oval on the night side to an approximately symmetric and wide oval centered on the magnetic pole.

The peak value of electron energy flux, which is located between 22:00 and 06:00 MLT in panel A, gradually shifts to the interval between 05:00 and 10:00 MLT as the southward component of the IMF B_z increases from negative to positive values. This phenomenon may be analogous to the effect of solar wind velocity on substorm probability, where high solar wind velocity tends to increase the likelihood of substorms and shifts the peak substorm occurrence toward the nightside. One possible explanation for this behavior is that a strong southward IMF can enhance the generation of whistler mode waves in the magnetosphere, which tend to peak in intensity between 22:00 and 06:00 MLT during substorms. During quieter times, the peak of whistler mode waves may shift to the morning sector, although this cannot be confirmed without direct wave measurements which are not available in our study. The total electron energy flux in the oval exhibits a decreasing trend from panels A to D, corresponding to the shift of the poleward boundary toward lower latitudes. However, it appears that all MLT regions have relatively stable activity levels from panels E to H, on average. These observations are supported

by Figure A4a in appendix, which depicts the median electron energy flux as a function of MLT for various B_z intervals.

The polar cap activity exhibits a distinct V-shape pattern as a function of B_z , as shown in Figure A4b. According to Gussenhoven et al. (1984), the polar rain number and energy fluxes increase as the geomagnetic activity enhances and as IMF B_z reduces when IMF $BB_z < 0$. However, their investigation of IMF B_z only used two bins: IMF $B_z < -2.5$ and $0 \text{ nT} < \text{IMF } B_z < -2.5 \text{ nT}$. By using a smaller bin size and more data points, Figure 4 confirms that the polar rain energy fluxes indeed increase when IMF $B_z < 0$ and becomes more negative.

It appears that the polar rain energy flux increases with the $|\text{IMF } B_z|$. For the case of IMF $B_z < 0$, as IMF B_z becomes more negative, the reconnection strength and rate would increase, which would increase the polar rain electron flux (Newell, Liou, & Wilson, 2009), as shown in Figure 4. For the case of IMF $B_z > 0$, it appears that an electron flux also increases with increasing B_z perhaps for the same reason, but there could be other reasons as well. An increase in B_z can reduce the polar cap size (open-closed boundary moves to higher latitude) (Milan et al., 2004; Newell et al., 1997; Tulegenov et al., 2023). Moreover, an increase in IMF B_z can also increase the occurrence of the polar cap arcs, which could be considered an extension of the auroral oval and which have higher fluxes than polar rain (Newell et al., 1997; Troshichev et al., 1988). The effect of the polar cap arcs, whose locations can vary depending on solar wind conditions, would be smeared out in the statistical map shown in Figure 4. All these effects can complicate the determination of the electron flux in the polar cap in Figure 4.

4. Summary and Discussion

In this study, we investigated the global-scale position, structure, intensity, and latitudinal width of the precipitating electrons above 55° MLAT (within both the auroral oval and polar cap) due to solar wind velocity, density, dynamic pressure, and the Z-component of the IMF. Here is a summary of the observations made for each solar wind driver considered.

For solar wind velocity, density, and dynamic pressure, the electron energy flux is always observed to be higher from midnight to noon than from noon to midnight. This phenomenon can be attributed to the electron curvature and gradient drifts, as well as VLF whistler mode chorus waves. This is also true for positive B_z values. However, for negative B_z , this asymmetry seems to appear on either side of the 19:00–07:00 MLTs line.

As a consequence of increasing the solar wind velocity at the BSN:

- The electron energy flux within the auroral oval increases by 83% on average, with the highest flux from midnight to noon. The increase is due to electron gradient and curvature drifts, as well as VLF whistler-mode chorus waves.
- The auroral oval extends equatorward to smaller latitudes as the solar wind velocity increases.
- The polar cap energy flux increases as the solar wind velocity increases, due to an increase in the strahl component of solar wind electrons entering the magnetosphere.
- The approximate width of the auroral oval generally increases as the solar wind velocity enhances, except for the 11–16 MLT region, where it remains roughly the same and increases only in the last two panels.

As a consequence of increasing the solar wind density at the BSN:

- The energy flux in the auroral oval also increases, but the change is not as significant as it is with the velocity. It is higher at midnight-noon compared to noon-midnight, attributed to electron curvature and gradient drifts, as well as VLF whistler mode chorus waves.
- The energy flux in the polar cap increases as the solar wind density enhances, behaving as a logarithmic function. There seems to be a correlation between polar rain and solar wind density, as opposed to what Riehl and Hardy (1986) found, explained by the positive correlation between density and strahl intensity (Borovsky, 2021). However, there is a small decrease when moving from panels A to B, mainly located between 20:00 and 02:00 MLT.

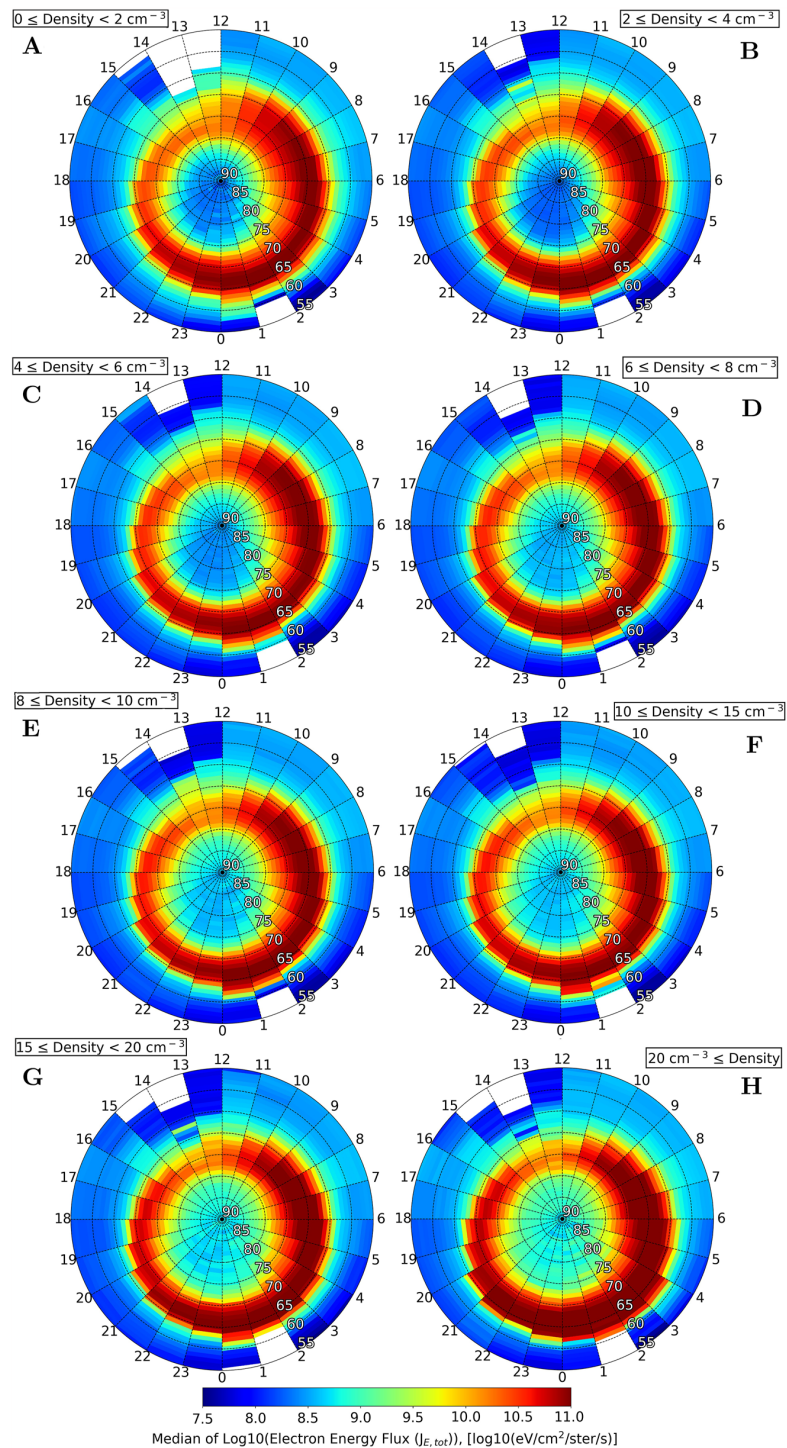


Figure 2. Median of $\text{Log}_{10}(J_{E,\text{tot}})$ [eV/cm²/ster/s] for proton density respectively between 0 and 2 cm⁻³ (a), 2 and 4 cm⁻³ (b), 4 and 6 cm⁻³ (c), 6 and 8 cm⁻³ (d), 8 and 10 cm⁻³ (e), 10 and 15 cm⁻³ (f), 15 and 20 cm⁻³ (g), 20 cm⁻³ and infinity (h).

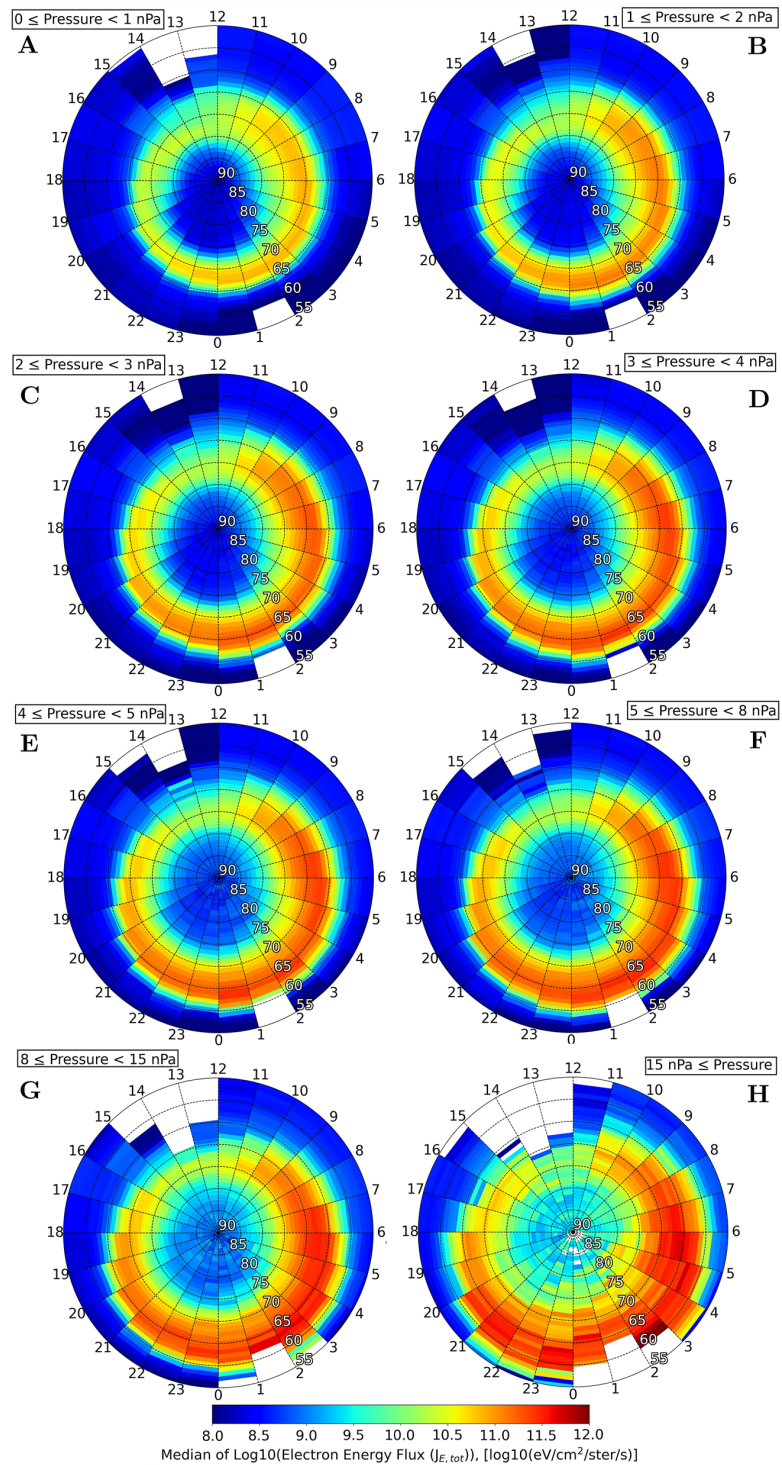


Figure 3. Median of $\text{Log}_{10}(J_{E,\text{tot}})$ [eV/cm²/ster/s] for solar wind pressure respectively between 0 and 1 nPa (a), 1 and 2 nPa (b), 2 and 3 nPa (c), 3 and 4 nPa (d), 4 and 5 nPa (e), 5 and 8 nPa (f), 8 and 15 nPa (g), 15 nPa and infinity (h).

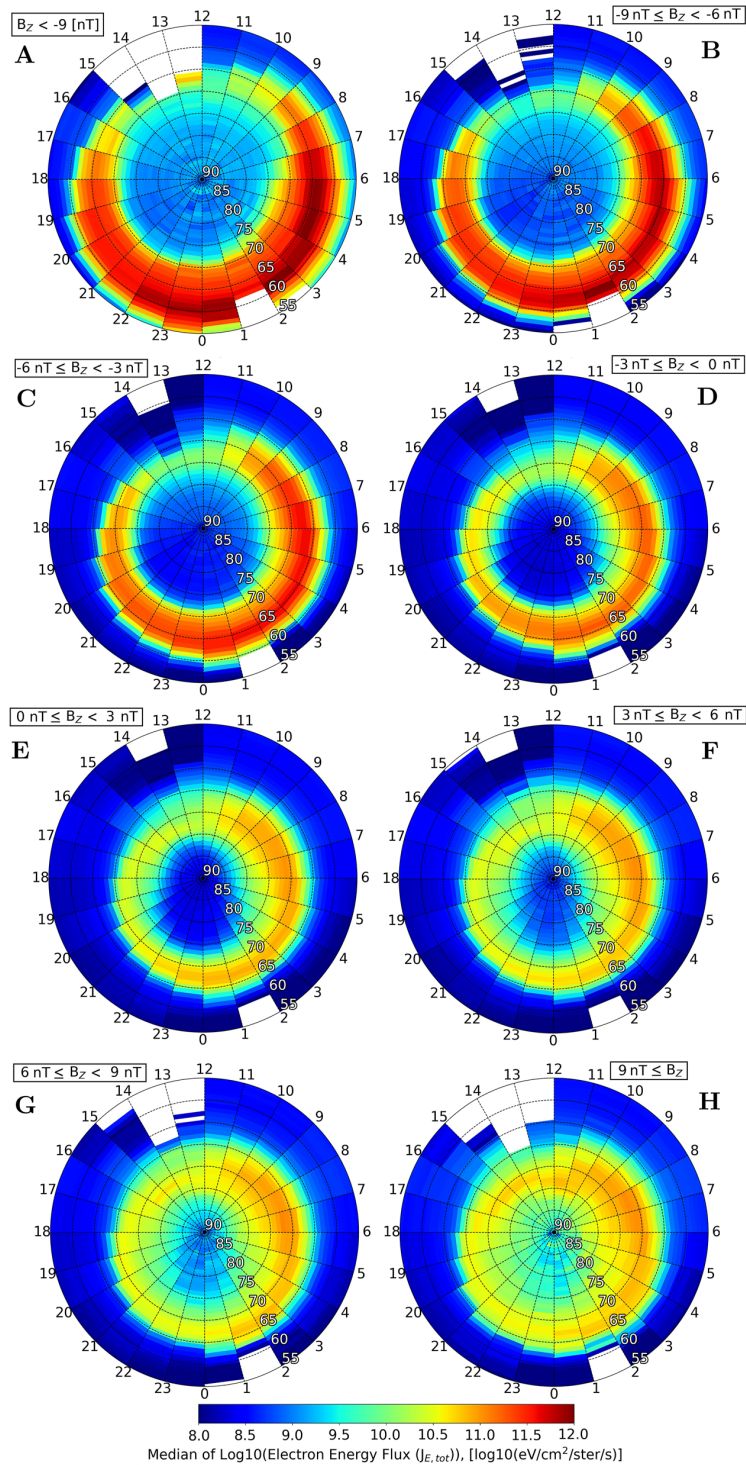


Figure 4. Median of $\text{Log}_{10}(J_{E,\text{tot}})$ [$\text{eV}/\text{cm}^2/\text{ster}/\text{s}$] for B_z GSM respectively below -9 nT (a), between -9 and -6 nT (b), -6 and -3 nT (c), -3 and 0 nT (d), 0 and 3 nT (e), 3 and 6 nT (f), 6 and 9 nT (g), and above 9 nT (h).

- As the solar wind density increases, the behavior of the auroral oval width becomes more complex, with variations observed in different regions at different times.

As a consequence of increasing the solar wind dynamic pressure at the BSN:

- Electron energy flux increases almost monotonically, as expected by the correlation between velocity and pressure.
- Auroral oval width increases and extends equatorward.
- Polar rain increases as the dynamic pressure enhances. The increase is even clearer than that for solar wind velocity and density.

Concerning the IMF Z-component:

- As the southward component of the IMF B_z increases, the magnetosphere can become more active due to substorms or storms, resulting in particle injections and energization of the particle population in the magnetotail. The peak value of electron energy flux gradually shifts to the interval between 05:00 and 10:00 MLT as the southward component of the IMF B_z increases from negative to positive values. Overall energy flux increases with $|B_z|$ but large negative B_z means more intense oval than for large positive values.
- A clear dependence can be observed with the shapes of both the oval and the polar cap. A large southward component of the IMF generally causes the boundary to move equatorward, while a northward component moves the boundary poleward. The oval width as a function B_z seems to have a “U” shape.
- The polar cap activity as a function B_z has a “V” shape.
- The polar rain energy fluxes increase when $|B_z|$ increases.

These results can be useful for comparisons to electron precipitation models (Newell et al., 2014, 2002; Wiltberger et al., 2009; Zhu et al., 2021) and to electron precipitation reconstructions based on ionospheric simulations (Simon Wedlund et al., 2013). As a follow-up study, we will examine the effects of IMF B_y and B_x on the auroral oval and polar cap (polar rain) electron precipitation, and we will investigate the dependence of solar wind clock angle ($\arctan(\text{IMF-}B_y/\text{IMF-}B_z)$), cone angle ($\arctan(\text{IMF-}B_z/\text{IMF-}B_x)$), and azimuthal angle ($\arctan(\text{IMF-}B_y/\text{IMF-}B_x)$) on the polar cap electron flux enhancements.

Appendix A: Additional Figures

Figures A1–A4 presented in this appendix show line plots of three subfigures for each solar wind parameter considered. These subfigures depict the median of the total electron flux inside the auroral oval, the median of the total electron flux inside the polar cap, and the approximate width of the auroral oval. It is important to note that we defined the boundary of the auroral oval arbitrarily as $10 \log_{10} (\text{eV/cm}^2/\text{s/ster})$. In each of the line plots shown in the four figures presented in this appendix, the x -axis represents the median value of either the total electron flux inside the auroral oval or the approximate width of the auroral oval, depending on the subfigure. The y -axis represents the median value of the solar wind parameter considered in the bin considered, as described in the main text of the paper. Thus, each point on the line plots represents a specific combination of the solar wind parameter and the electron flux. By including these additional figures in the appendix, we aim to provide a more comprehensive understanding of the complex interactions between the solar wind parameters and the Earth's

magnetosphere. We hope these figures provide a clear and more detailed visualization of the relationship between the polar zone activity, the DMSP electron flux in LEO, and the solar wind parameters.

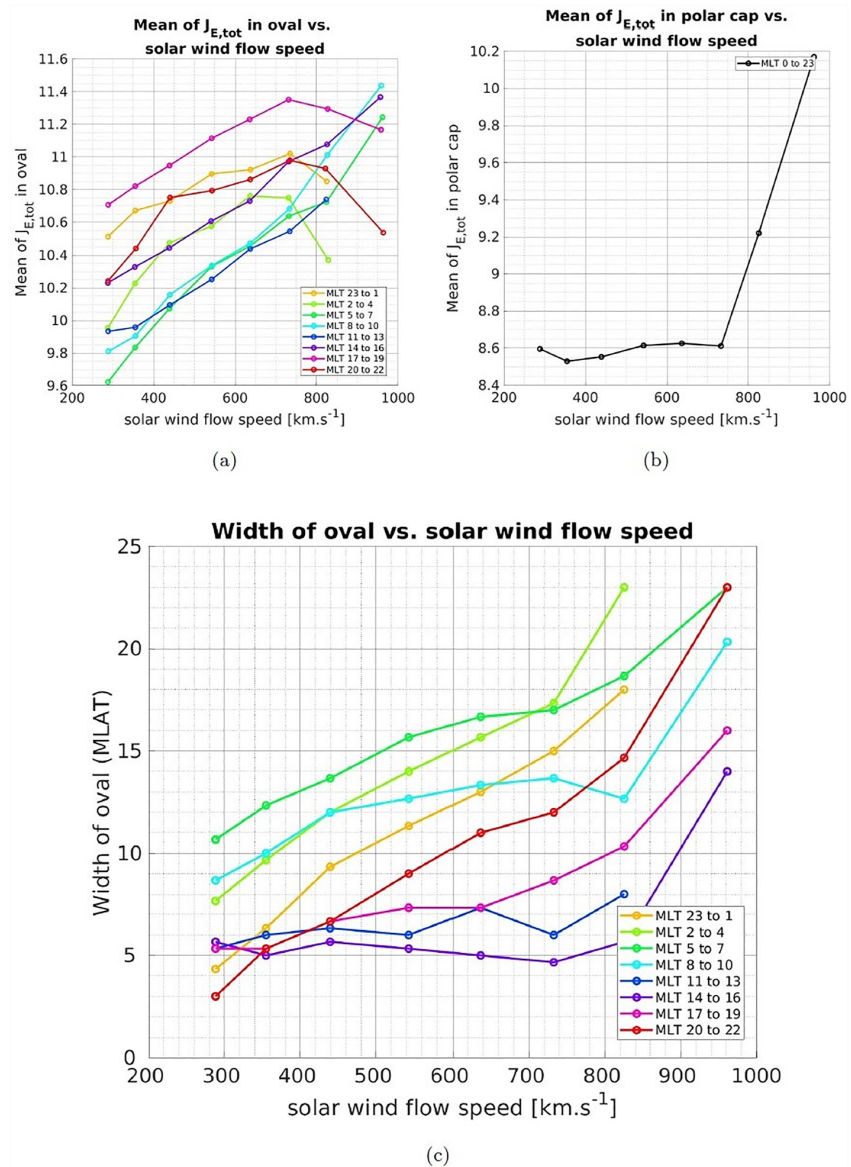


Figure A1. These line plots depict three measures as a function of solar wind flow speed. Panel (a) shows the mean of all values observed in the polar plots, where MLTs were grouped in sets of three. Each point in the polar plot corresponds to the median total electron energy flux for a specific MLAT-MLT pair. Panel (b) displays the same average, but only for values located within the auroral oval. The auroral oval is defined as the region where values exceed $10 \log_{10}$ ($\text{eV}/\text{cm}^2/\text{s}/\text{ster}$). Finally, panel (c) shows the approximate width of the auroral oval in terms of the number of MLAT.

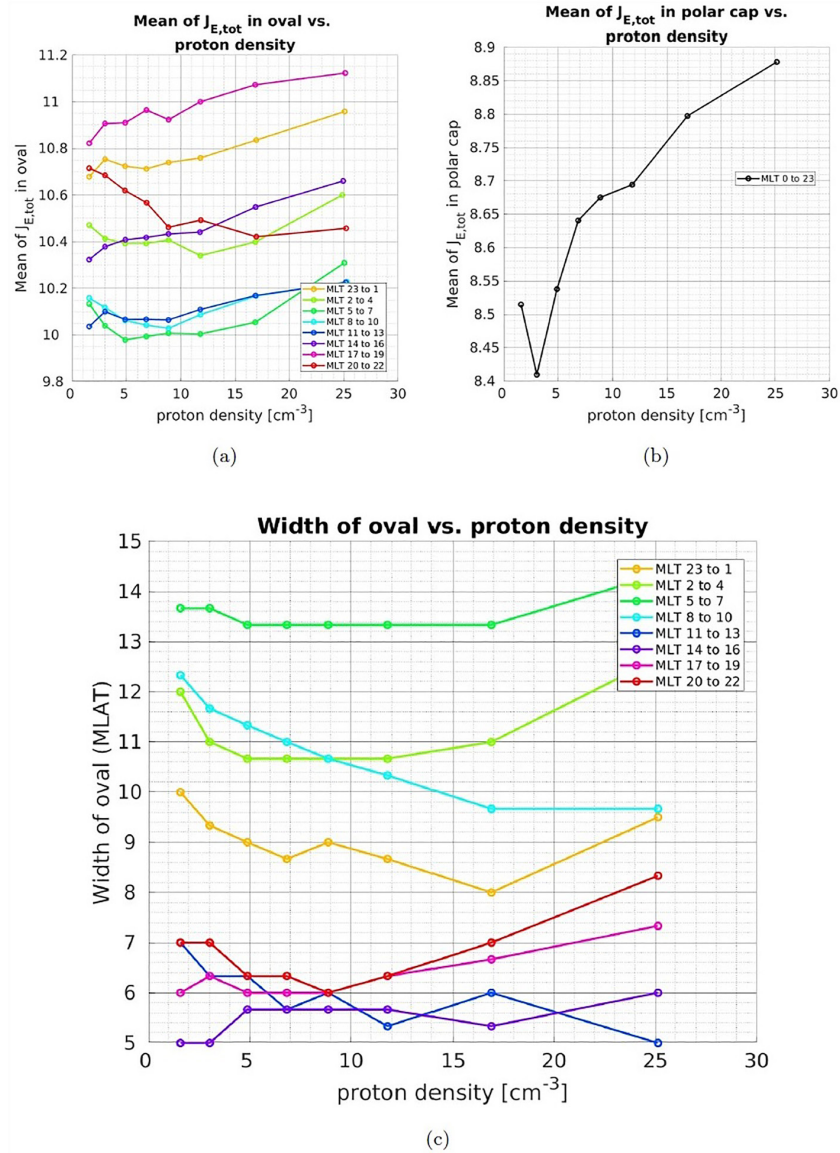


Figure A2. These line plots depict three measures as a function of solar wind proton density. Panel (a) shows the mean of all values observed in the polar plots, where MLTs were grouped in sets of three. Each point in the polar plot corresponds to the median total electron energy flux for a specific MLAT-MLT pair. Panel (b) displays the same average, but only for values located within the auroral oval. The auroral oval is defined as the region where values exceed $10 \log_{10} (eV/cm^2/s/ster)$. Finally, panel (c) shows the approximate width of the auroral oval in terms of the number of MLAT.

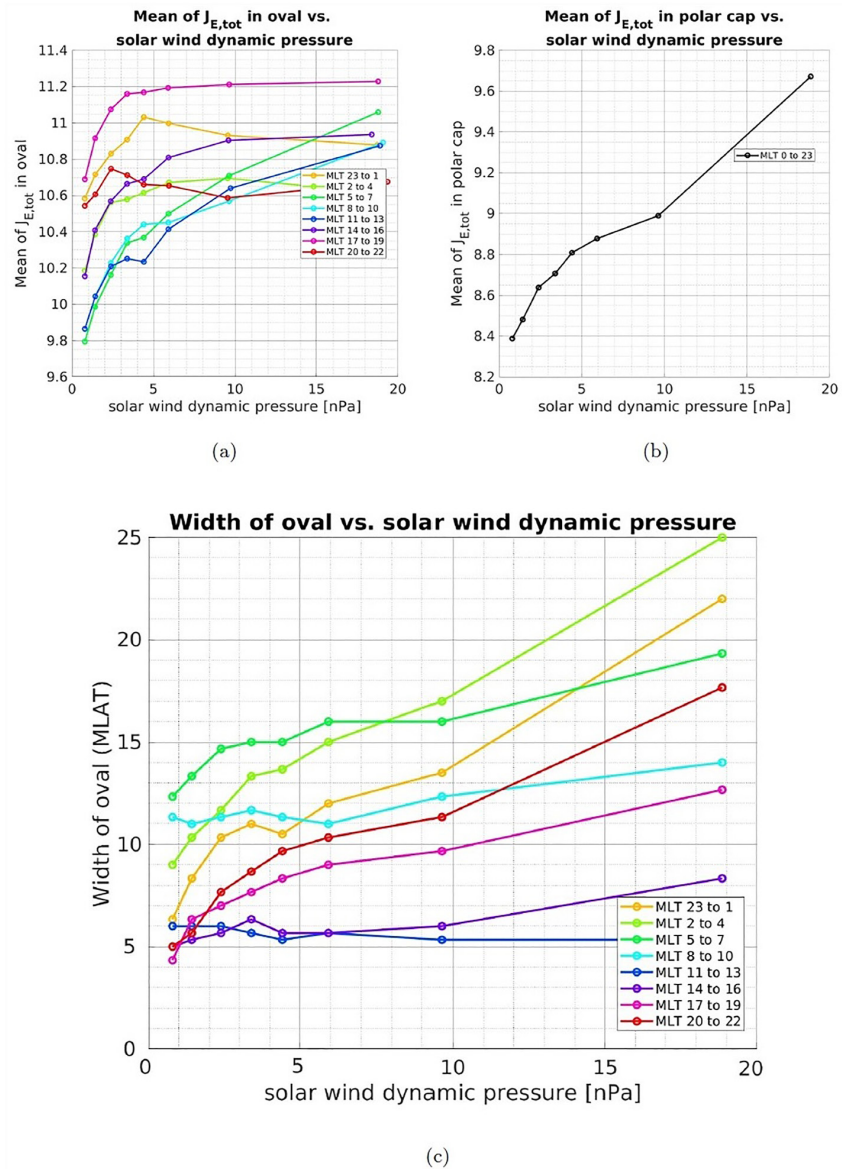


Figure A3. These line plots depict three measures as a function of solar wind dynamic pressure. Panel (a) shows the mean of all values observed in the polar plots, where MLTs were grouped in sets of three. Each point in the polar plot corresponds to the median total electron energy flux for a specific MLAT-MLT pair. Panel (b) displays the same average, but only for values located within the auroral oval. The auroral oval is defined as the region where values exceed $10 \log_{10} (\text{eV}/\text{cm}^2/\text{s}/\text{ster})$. Finally, panel (c) shows the approximate width of the auroral oval in terms of the number of MLAT.

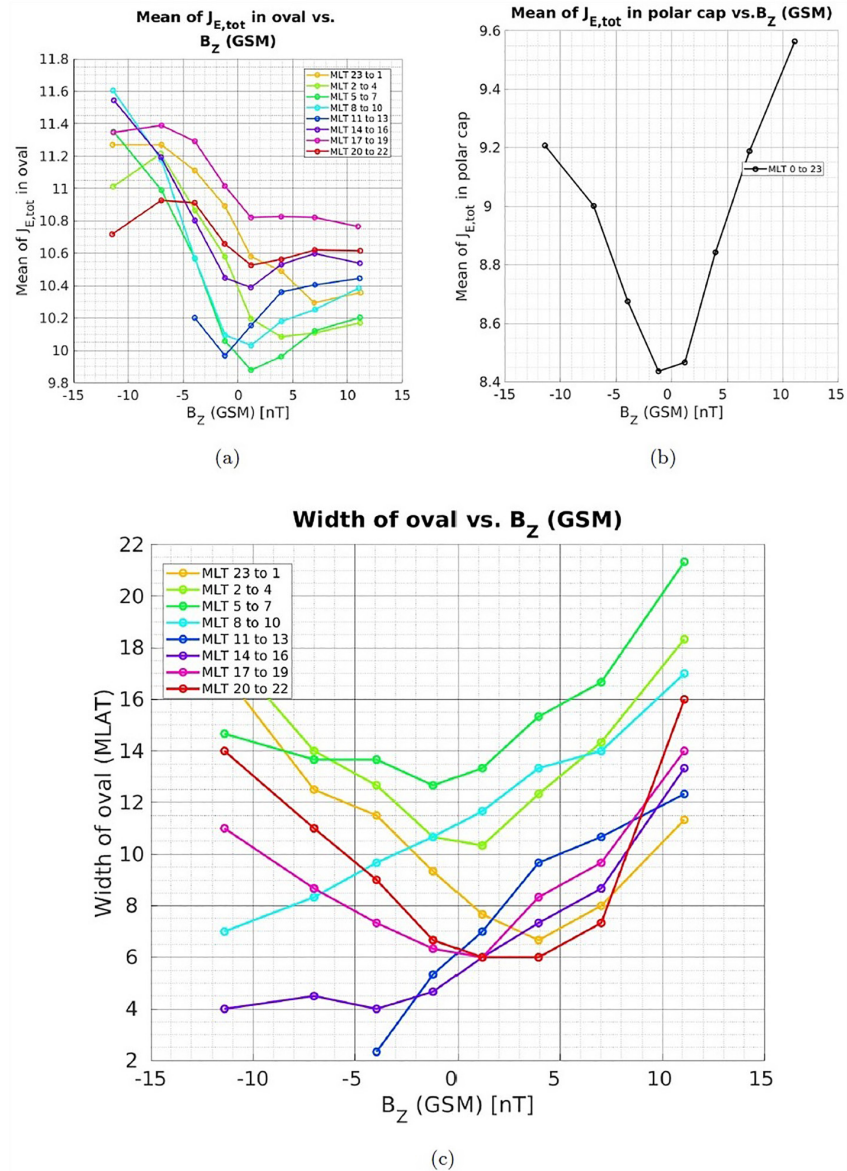


Figure A4. These line plots depict three measures as a function of IMF B_z (in GSM coordinates). Panel (a) shows the mean of all values observed in the polar plots, where MLTs were grouped in sets of three. Each point in the polar plot corresponds to the median total electron energy flux for a specific MLAT-MLT pair. Panel (b) displays the same average, but only for values located within the auroral oval. The auroral oval is defined as the region where values exceed $10 \log_{10} (\text{eV}/\text{cm}^2/\text{s}/\text{ster})$. Finally, panel (c) shows the approximate width of the auroral oval in terms of the number of MLAT.

The last four figures are representations of the magnetic latitude/magnetic longitude grid for each of the four solar wind parameters studied in the article. However, unlike the figures in the main body of the article, these figures show the density of data available for each point on the grid. Specifically, each point on the grid is color-coded according to the number of data points available for that particular combination of magnetic latitude, magnetic longitude, and solar wind parameter. These figures can help identify any gaps or biases in the data set, which can inform future research and data collection efforts. Overall, these figures are an important complement to the main

body of the article, providing additional information about the data used in the study and helping to ensure the validity and accuracy of the research findings (Figures A5–A8).

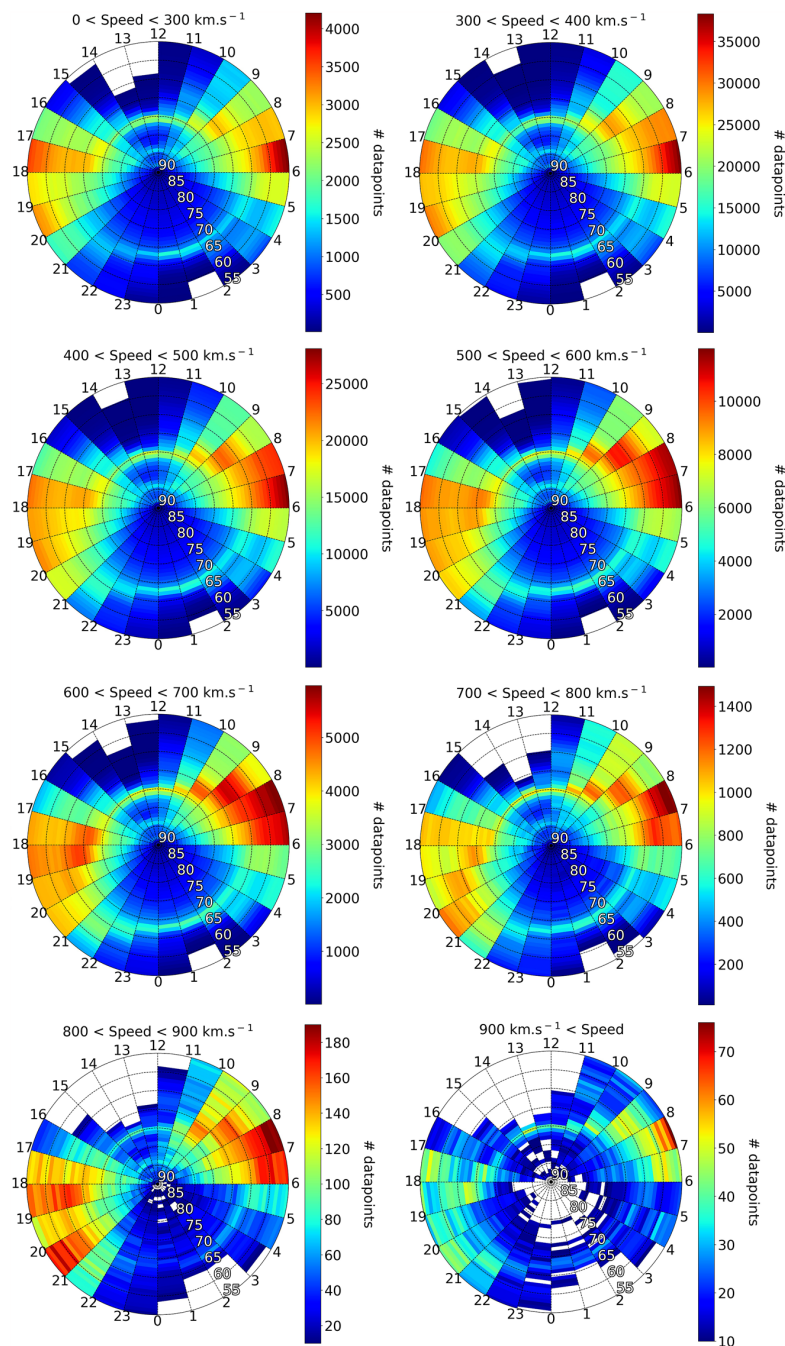


Figure A5. Number of data points available for solar wind speed ($\text{km}\cdot\text{s}^{-1}$) in the same ranges than for the polar plots.

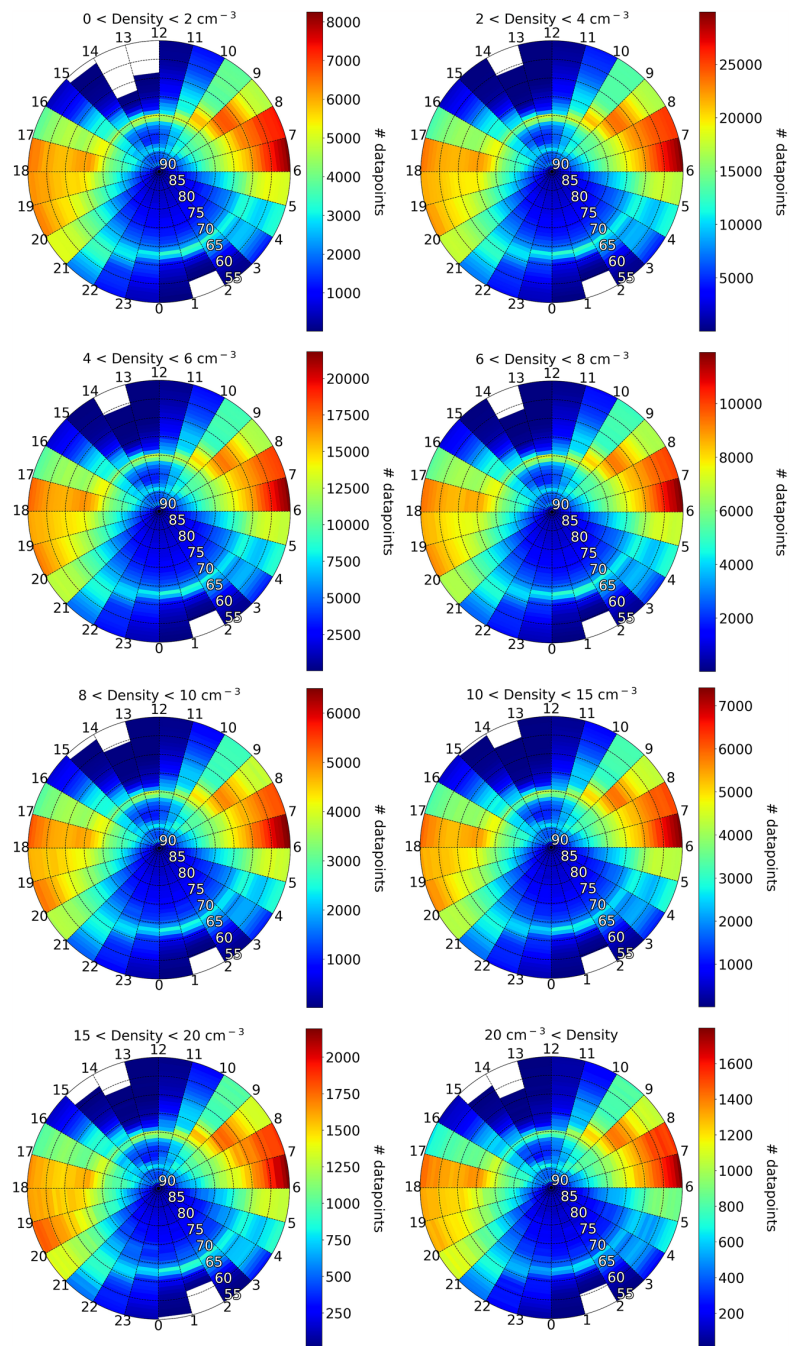


Figure A6. Number of data points available for proton density (cm^{-3}) in the same ranges than for the polar plots.

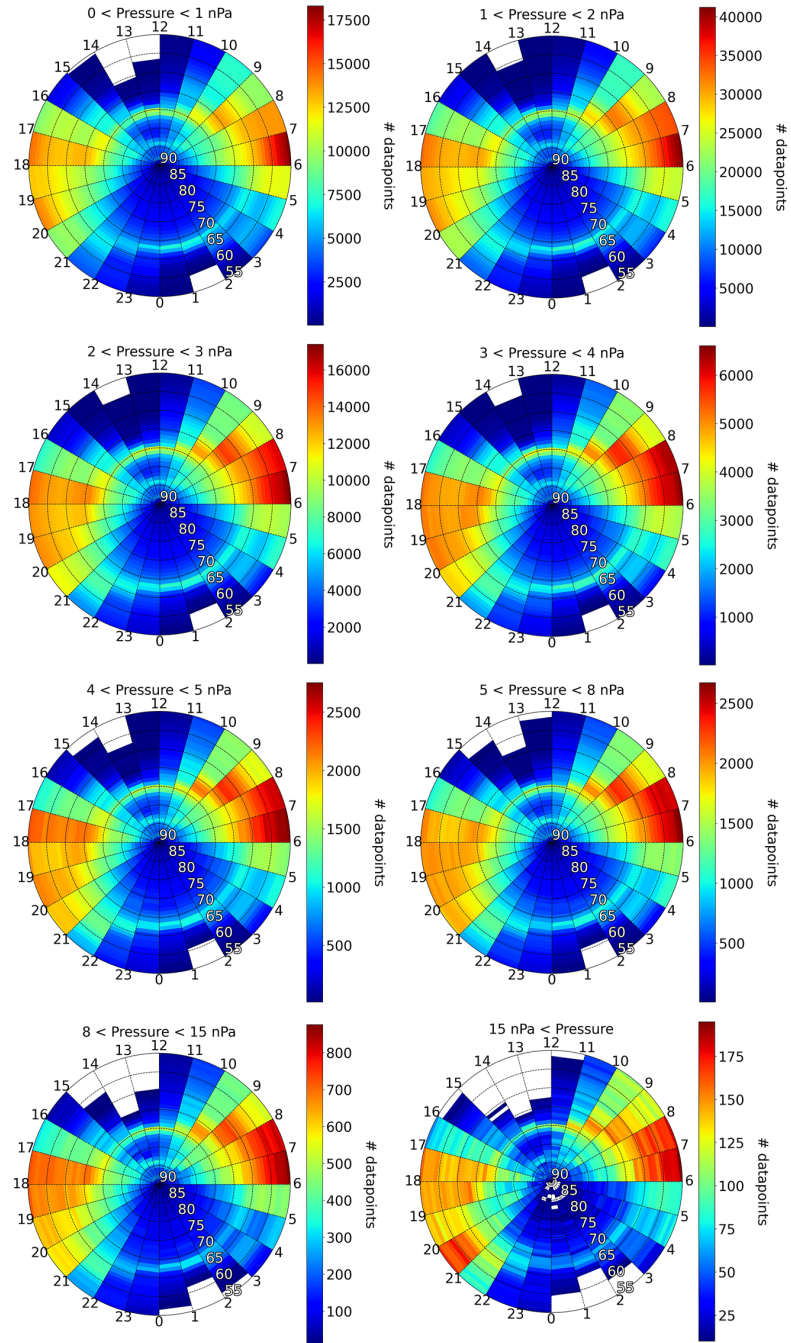


Figure A7. Number of data points available for dynamic pressure (nPa) in the same ranges than for the polar plots.

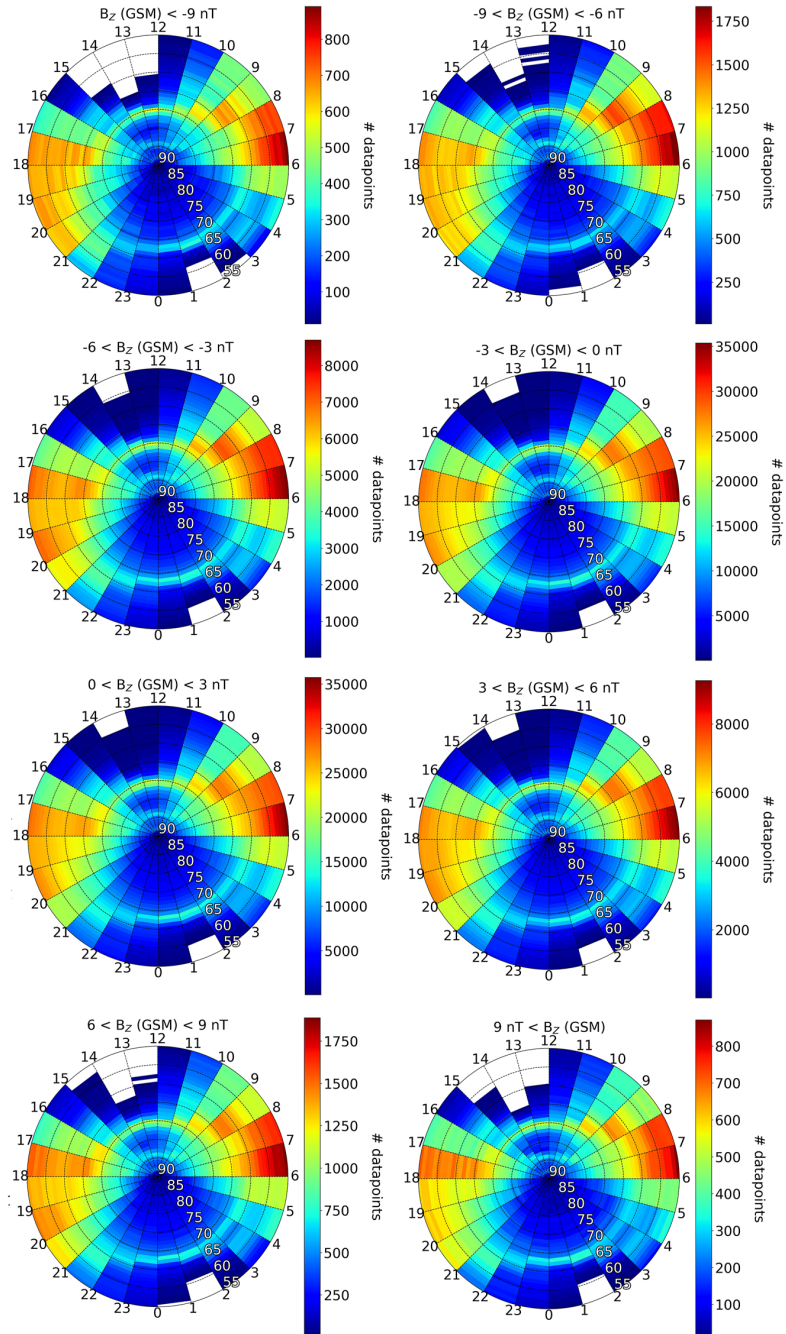


Figure A8. Number of data points available for B_z -component IMF (nT) in the same ranges than for the polar plots.

Data Availability Statement

The 1 min resolution solar wind data are available at the OMNIWeb <https://omniweb.gsfc.nasa.gov/>. The DMSP data are available at CDAWeb <https://cdaweb.gsfc.nasa.gov/>.

Acknowledgments

The authors would like to acknowledge SpaceAble and the experts who helped understand, prepare, and manage the data. The authors would like to acknowledge the NASA's CDWeb interface and OMNIWeb database experts for providing quality and ready-to-use data. Simon Wing acknowledges the support of NASA Grants 80NSSC20K0704, 80NSSC21K1678, 80NSSC22K0515, 80NSSC21K1321, 80NSSC22K0304, 80NSSC19K0843, and 80NSSC20K0188.

References

- Baker, K. B., & Wing, S. (1989). A new magnetic coordinate system for conjugate studies at high latitudes. *Journal of Geophysical Research: Space Physics*, *94*(A7), 9139–9143. <https://doi.org/10.1029/JA094iA07p09139>
- Berchem, J., Richard, R., Escoubet, P., Wing, S., & Pitout, F. (2014). Dawn-dusk asymmetry in solar wind ion entry and dayside precipitation: Results from large-scale simulations. *Journal of Geophysical Research: Space Physics*, *119*(3), 1549–1562. <https://doi.org/10.1002/2013JA019427>
- Berchem, J., Richard, R. L., Escoubet, C. P., Wing, S., & Pitout, F. (2016). Asymmetrical response of dayside ion precipitation to a large rotation of the IMF. *Journal of Geophysical Research: Space Physics*, *121*(1), 263–273. <https://doi.org/10.1002/2015JA021969>
- Borovsky, J. E. (2020). What magnetospheric and ionospheric researchers should know about the solar wind. *Journal of Atmospheric and Solar-Terrestrial Physics*, *204*, 105271. <https://doi.org/10.1016/j.jastp.2020.105271>
- Borovsky, J. E. (2021). Exploring the properties of the electron strahl at 1 AU as an indicator of the quality of the magnetic connection between the Earth and the Sun. *Frontiers in Astronomy and Space Sciences*, *8*. <https://doi.org/10.3389/fspas.2021.646443>
- Borovsky, J. E., Thomsen, M. F., & Elphic, R. C. (1998). The driving of the plasma sheet by the solar wind. *Journal of Geophysical Research: Space Physics*, *103*(A8), 17617–17639. <https://doi.org/10.1029/97JA02986>
- Burch, J. (1979). Effects of the interplanetary magnetic field on the auroral oval and plasmapause. *Space Science Reviews*, *23*(3), 449–464. <https://doi.org/10.1007/BF00172250>
- Fairfield, D. H., & Scudder, J. D. (1985). Polar rain: Solar coronal electrons in the earth's magnetosphere. *Journal of Geophysical Research: Space Physics*, *90*(A5), 4055–4068. <https://doi.org/10.1029/JA090iA05p04055>
- Fairfield, D. H., Wing, S., Newell, P. T., Ruohoniemi, J. M., Gosling, J. T., & Skoug, R. M. (2008). Polar rain gradients and field-aligned polar cap potentials. *Journal of Geophysical Research: Space Physics*, *113*(A10). <https://doi.org/10.1029/2008JA013437>
- Feldstein, Y. I. (2016). The discovery and the first studies of the auroral oval: A review. *Geomagnetism and Aeronomy*, *56*(2), 129–142. <https://doi.org/10.1134/S0016793216020043>
- Fujimoto, M., Terasawa, T., Mukai, T., Saito, Y., Yamamoto, T., & Kokubun, S. (1998). Plasma entry from the flanks of the near-earth magnetotail: Geotail observations. *Journal of Geophysical Research: Space Physics*, *103*(A3), 4391–4408. <https://doi.org/10.1029/97JA03340>
- Gabrielse, C., Nishimura, T., Chen, M., Hecht, J. H., Kaeppler, S. R., Gillies, D. M., et al. (2021). Estimating precipitating energy flux, average energy, and hall auroral conductance from THEMIS all-sky-imagers with focus on mesoscales. *Frontiers in Physics*, *9*. <https://doi.org/10.3389/fphy.2021.744298>
- Gussenhoven, M. S., Hardy, D. A., Heinemann, N., & Burkhardt, R. K. (1984). Morphology of the polar rain. *Journal of Geophysical Research: Space Physics*, *89*(A11), 9785–9800. <https://doi.org/10.1029/JA089iA11p09785>
- Hardy, D., Schmitt, L., Gussenhoven, M., Marshall, F., & Yeh, H. (1984). Precipitating electron and ion detectors (SSJ/4) for the block 5D/flights 6–10 DMSP (defense meteorological satellite program) satellites: Calibration and data presentation. Rep. AFGL-TR-84-0317.
- Johnson, J. R., Wing, S., Delamere, P., Petrincec, S., & Kavosi, S. (2021). Field-aligned currents in auroral vortices. *Journal of Geophysical Research: Space Physics*, *126*(2), e2020JA028583. <https://doi.org/10.1029/2020JA028583>
- Kamide, Y., Perreault, P. D., Akasofu, S. I., & Winningham, J. D. (1977). Dependence of substorm occurrence probability on the interplanetary magnetic field and on the size of the auroral oval. *Journal of Geophysical Research*, *82*(35), 5521–5528. <https://doi.org/10.1029/JA082i035p05521>
- King, J., & Papitashvili, N. (2006). *One min and 5-min solar wind data sets at the Earth's bow shock nose*. NASA Goddard Space Flight Center.
- Liou, K., Newell, P. T., Shue, J.-H., Meng, C.-I., Miyashita, Y., Kojima, H., & Matsumoto, H. (2007). "Compression aurora": Particle precipitation driven by long-duration high solar wind ram pressure. *Journal of Geophysical Research: Space Physics*, *112*(A11). <https://doi.org/10.1029/2007JA012443>
- Lockwood, M., Denig, W. F., Farmer, A. D., Davda, V. N., Cowley, S. W. H., & Lühr, H. (1993). Ionospheric signatures of pulsed reconnection at the Earth's magnetopause. *Nature*, *361*(6411), 424–428. <https://doi.org/10.1038/361424a0>
- Lyons, L. R., Schulz, M., Pridmore-Brown, D. C., & Roeder, J. L. (1994). Low-latitude boundary layer near noon: An open field line model. *Journal of Geophysical Research: Space Physics*, *99*(A9), 17367–17377. <https://doi.org/10.1029/94JA00867>
- Maggiolo, R., Hamrin, M., De Keyser, J., Pitkanen, T., Cessateur, G., Gunell, H., & Maes, L. (2017). The delayed time response of geomagnetic activity to the solar wind. *Journal of Geophysical Research: Space Physics*, *122*(11), 11109–11127. <https://doi.org/10.1002/2016JA023793>
- Milan, S. E., Cowley, S. W. H., Lester, M., Wright, D. M., Slavin, J. A., Fillingim, M., & Singer, H. J. (2004). Response of the magnetotail to changes in the open flux content of the magnetosphere. *Journal of Geophysical Research: Space Physics*, *109*(A4), A04220. <https://doi.org/10.1029/2003ja010350>
- Newell, P. T., Liou, K., Gjerloev, J., Sotirelis, T., Wing, S., & Mitchell, E. (2016). Substorm probabilities are best predicted from solar wind speed. *Journal of Atmospheric and Solar-Terrestrial Physics*, *146*, 28–37. <https://doi.org/10.1016/j.jastp.2016.04.019>
- Newell, P. T., Liou, K., & Wilson, G. R. (2009). Polar cap particle precipitation and aurora: Review and commentary. *Journal of Atmospheric and Solar-Terrestrial Physics*, *71*(2), 199–215. <https://doi.org/10.1016/j.jastp.2008.11.004>
- Newell, P. T., Liou, K., Zhang, Y., Sotirelis, T., Paxton, L. J., & Mitchell, E. J. (2014). Ovation Prime-2013: Extension of auroral precipitation model to higher disturbance levels. *Space Weather*, *12*(6), 368–379. <https://doi.org/10.1002/2014SW001056>
- Newell, P. T., & Meng, C.-I. (1992). Mapping the dayside ionosphere to the magnetosphere according to particle precipitation characteristics. *Geophysical Research Letters*, *19*(6), 609–612. <https://doi.org/10.1029/92GL00404>
- Newell, P. T., Ruohoniemi, J. M., & Meng, C.-I. (2004). Maps of precipitation by source region, binned by IMF, with inertial convection streamlines. *Journal of Geophysical Research: Space Physics*, *109*(A10), A10206. <https://doi.org/10.1029/2004JA010499>
- Newell, P. T., Sotirelis, T., Ruohoniemi, J. M., Carbary, J. F., Liou, K., Skura, J. P., et al. (2002). Ovation: Oval variation, assessment, tracking, intensity, and online nowcasting. *Annales Geophysicae*, *20*(7), 1039–1047. <https://doi.org/10.5194/angeo-20-1039-2002>
- Newell, P. T., Sotirelis, T., & Wing, S. (2009). Diffuse, monoenergetic, and broadband aurora: The global precipitation budget. *Journal of Geophysical Research: Space Physics*, *114*(A9). <https://doi.org/10.1029/2009JA014326>
- Newell, P. T., Sotirelis, T., & Wing, S. (2010). Seasonal variations in diffuse, monoenergetic, and broadband aurora. *Journal of Geophysical Research: Space Physics*, *115*(A3). <https://doi.org/10.1029/2009JA014805>
- Newell, P. T., Wing, S., Meng, C.-I., & Sigillito, V. (1991). The auroral oval position, structure, and intensity of precipitation from 1984 onward: An automated on-line data base. *Journal of Geophysical Research: Space Physics*, *96*(A4), 5877–5882. <https://doi.org/10.1029/90JA02450>
- Newell, P. T., Xu, D., Meng, C.-I., & Kivelson, M. G. (1997). Dynamical polar cap: A unifying approach. *Journal of Geophysical Research: Space Physics*, *102*(A1), 127–139. <https://doi.org/10.1029/96JA03045>
- Ni, B., Thorne, R. M., Zhang, X., Bortnik, J., Pu, Z., Xie, L., et al. (2016). Origins of the Earth's diffuse auroral precipitation. *Space Science Reviews*, *200*(1), 205–259. <https://doi.org/10.1007/s11214-016-0234-7>
- Reeves, G. D., Chan, A., & Rodger, C. (2009). New directions for radiation belt research. *Space Weather*, *7*(7). <https://doi.org/10.1029/2008SW000436>

- Riehl, K. B., & Hardy, D. A. (1986). Average characteristics of the polar rain and their relationship to the solar wind and the interplanetary magnetic field. *Journal of Geophysical Research: Space Physics*, 91(A2), 1557–1571. <https://doi.org/10.1029/JA091iA02p01557>
- Sergeev, V. A., Liou, K., Newell, P. T., Ohtani, S.-I., Hairston, M. R., & Rich, F. (2004). Auroral streamers: Characteristics of associated precipitation, convection and field-aligned currents. *Annales Geophysicae*, 22(2), 537–548. <https://doi.org/10.5194/angeo-22-537-2004>
- Shi, Q. Q., Zong, Q.-G., Fu, S., Dunlop, M., Pu, Z., Parks, G., et al. (2013). Solar wind entry into the high-latitude terrestrial magnetosphere during geomagnetically quiet times. *Nature Communications*, 4(1), 1466. <https://doi.org/10.1038/ncomms2476>
- Shi, Q. Q., Zong, Q.-G., Zhang, H., Pu, Z. Y., Fu, S. Y., Xie, L., et al. (2009). Cluster observations of the entry layer equatorward of the cusp under northward interplanetary magnetic field. *Journal of Geophysical Research: Space Physics*, 114(A12). <https://doi.org/10.1029/2009JA014475>
- Simon Wedlund, C., Lamy, H., Gustavsson, B., Sergienko, T., & Brändström, U. (2013). Estimating energy spectra of electron precipitation above auroral arcs from ground-based observations with radar and optics. *Journal of Geophysical Research: Space Physics*, 118(6), 3672–3691. <https://doi.org/10.1002/jgra.50347>
- Sorathia, K. A., Merkin, V. G., Ukhorskiy, A. Y., Allen, R. C., Nykyri, K., & Wing, S. (2019). Solar wind ion entry into the magnetosphere during northward IMF. *Journal of Geophysical Research: Space Physics*, 124(7), 5461–5481. <https://doi.org/10.1029/2019JA026728>
- Summers, D., Thorne, R. M., & Xiao, F. (1998). Relativistic theory of wave-particle resonant diffusion with application to electron acceleration in the magnetosphere. *Journal of Geophysical Research: Space Physics*, 103(A9), 20487–20500. <https://doi.org/10.1029/98JA01740>
- Thorne, R. M. (2010). Radiation belt dynamics: The importance of wave-particle interactions. *Geophysical Research Letters*, 37(22). <https://doi.org/10.1029/2010gl044990>
- Troschichev, O., Gusev, M., Nickolashkin, S., & Samsonov, V. (1988). Features of the polar cap aurorae in the southern polar region. *Planetary and Space Science*, 36(5), 429–439. [https://doi.org/10.1016/0032-0633\(88\)90102-X](https://doi.org/10.1016/0032-0633(88)90102-X)
- Tulegenov, B., Raeder, J., Cramer, W. D., Ferdousi, B., Fuller-Rowell, T. J., Maruyama, N., & Strangeway, R. J. (2023). Storm time polar cap expansion: Interplanetary magnetic field clock angle dependence. *Annales Geophysicae*, 41(1), 39–54. <https://doi.org/10.5194/angeo-41-39-2023>
- Wiltberger, M., Weigel, R. S., Lotko, W., & Fedder, J. A. (2009). Modeling seasonal variations of auroral particle precipitation in a global-scale magnetosphere-ionosphere simulation. *Journal of Geophysical Research: Space Physics*, 114(A1). <https://doi.org/10.1029/2008JA013108>
- Wing, S., Berchem, J., Escoubet, C. P., Farrugia, C., & Lugaz, N. (2023). Multispacecraft observations of the simultaneous occurrence of magnetic reconnection at high and low latitudes during the passage of a solar wind rotational discontinuity embedded in the April 9–11, 2015 ICME. *Geophysical Research Letters*, 50(9), e2023GL103194. <https://doi.org/10.1029/2023GL103194>
- Wing, S., Fairfield, D. H., Johnson, J. R., & Ohtani, S.-I. (2015). On the field-aligned electric field in the polar cap. *Geophysical Research Letters*, 42(13), 5090–5099. <https://doi.org/10.1002/2015GL064229>
- Wing, S., Gkioulidou, M., Johnson, J. R., Newell, P. T., & Wang, C.-P. (2013). Auroral particle precipitation characterized by the substorm cycle. *Journal of Geophysical Research: Space Physics*, 118(3), 1022–1039. <https://doi.org/10.1002/jgra.50160>
- Wing, S., & Johnson, J. R. (2009). Substorm entropies. *Journal of Geophysical Research: Space Physics*, 114(A9). <https://doi.org/10.1029/2008JA013989>
- Wing, S., Johnson, J. R., Camporeale, E., & Reeves, G. D. (2016). Information theoretical approach to discovering solar wind drivers of the outer radiation belt. *Journal of Geophysical Research: Space Physics*, 121(10), 9378–9399. <https://doi.org/10.1002/2016JA022711>
- Wing, S., Johnson, J. R., Chaston, C. C., Echim, M., Escoubet, C. P., Lavraud, B., et al. (2014). Review of solar wind entry into and transport within the plasma sheet. *Space Science Reviews*, 184(1), 33–86. <https://doi.org/10.1007/s11214-014-0108-9>
- Wing, S., Johnson, J. R., & Fujimoto, M. (2006). Timescale for the formation of the cold-dense plasma sheet: A case study. *Geophysical Research Letters*, 33(23), L23106. <https://doi.org/10.1029/2006GL027110>
- Wing, S., Johnson, J. R., Newell, P. T., & Meng, C.-I. (2005). Dawn-dusk asymmetries, ion spectra, and sources in the northward interplanetary magnetic field plasma sheet. *Journal of Geophysical Research: Space Physics*, 110(A8). <https://doi.org/10.1029/2005JA011086>
- Wing, S., Johnson, J. R., Turner, D. L., Ukhorskiy, A. Y., & Boyd, A. J. (2022). Untangling the solar wind and magnetospheric drivers of the radiation belt electrons. *Journal of Geophysical Research: Space Physics*, 127(4), e2021JA030246. <https://doi.org/10.1029/2021JA030246>
- Wing, S., & Newell, P. T. (1998). Central plasma sheet ion properties as inferred from ionospheric observations. *Journal of Geophysical Research: Space Physics*, 103(A4), 6785–6800. <https://doi.org/10.1029/97JA02994>
- Wing, S., Newell, P. T., & Onsager, T. G. (1996). Modeling the entry of magnetosheath electrons into the dayside ionosphere. *Journal of Geophysical Research*, 101(A6), 13155–13167. <https://doi.org/10.1029/96JA00395>
- Wing, S., Newell, P. T., & Ruohoniemi, J. M. (2001). Double cusp: Model prediction and observational verification. *Journal of Geophysical Research: Space Physics*, 106(A11), 25571–25593. <https://doi.org/10.1029/2000JA000402>
- Zhu, Q., Deng, Y., Maute, A., Kilcommons, L. M., Knipp, D. J., & Hairston, M. (2021). Ashley: A new empirical model for the high-latitude electron precipitation and electric field. *Space Weather*, 19(5), e2020SW002671. <https://doi.org/10.1029/2020SW002671>
- Zhu, Q., Deng, Y., Richmond, A., & Maute, A. (2018). Small-scale and mesoscale variabilities in the electric field and particle precipitation and their impacts on joule heating. *Journal of Geophysical Research: Space Physics*, 123(11), 9862–9872. <https://doi.org/10.1029/2018JA025771>

To summarize, the paper reveals interesting observations, indicating that high solar wind velocity, density, or pressure correspond to higher energy flux levels overall, as well as broader night-side ovals and increased polar rain energy fluxes. Interestingly, we uncover a positive correlation between polar rain and solar wind density, challenging the findings of a prior study. Additionally, the paper highlights how variations in the Bz component of the interplanetary magnetic field lead to distinctive shapes in the oval width and polar cap activity curves, emphasizing the need for a comprehensive understanding of this relationship.

By rigorously exploring the connection between input solar wind parameters and output electron energy fluxes, this study serves as a step forward in advancing our knowledge of space weather and its implications for Low Earth Orbit environments. Moreover, it underscores the vital role of identifying input-output relationships as a crucial preliminary step in any AI-driven analysis, ensuring the accuracy and reliability of the results obtained. For us, it serves as the ultimate justification on why AI should work on this issue (beside difficulties arising from dataset itself).

3.4 Summary & Final Preprocessing

Regarding the input data, our primary focus was on selecting the parameters that commonly arise during auroral precipitation modeling. This included solar wind parameters, the Interplanetary Magnetic Field (IMF), and a handful of geomagnetic indices (McGranaghan et al., 2021; Vorobjev et al., 2013). A detailed analysis of data from a relevant satellite (ACE) (Bouriat et al., 2022) indicated that processing the data would be quite intensive, making it unsuitable for direct use in our framework. Consequently, like many other researchers in this field, we opted for the high-resolution OMNI data, even though they come with their own limitations and aren't considered perfect (Ashour-Abdalla et al., 2008; Samara et al., 2021; Vokhmyanin et al., 2019). AU, AL, and SYM/H, which we also incorporated, have shown their effectiveness in modeling auroral precipitations (Vorobjev et al., 2013). We plan to conduct tests both with and without them to gauge the feasibility of modeling precipitations while circumventing near-Earth data. Notably, studies leveraging artificial intelligence have demonstrated the potential to reconstruct some of these indices from data external to the magnetosphere (Amariutei and Ganushkina, 2012; Bhaskar and Vichare, 2019; Siciliano et al., 2021).

When it comes to our output data, it's worth noting that electron precipitation data exhibits bimodal or multimodal distribution patterns, contingent on the location and activity level (Hardy et al., 2008). This dataset's inherent imbalance presents a challenge for our network. Several hyperparameters, such as the number of epochs, the loss function, and the network's size, will play pivotal roles in effectively handling this complexity. Obtaining constructive feedback on our training through Tensorboard will also hold significant value. Given the correlation between the distribution and solar activity, our strategy involves shuffling the data comprehensively before splitting it into training, validation, and testing sets. This approach aims to ensure an equitable distribution of cases across each set, thereby avoiding any potential bias towards specific periods of solar activity or particular satellites. Lastly, as is often the case with satellite measurements, uncertainties are expected in DMSP data. These uncertainties encompass aspects such as instrument calibration, statistical methodologies, and noise present in telemetry or measurements (e.g., radiations, photons). A more in-depth discussion on how we handle and account for these uncertainties can be found in the works of Hardy et al. (2008); Redmon et al. (2017).

Lastly, it's worth mentioning the dataset provided by [McGranaghan \(2019\)](#); [McGranaghan et al. \(2021\)](#), which can be directly accessed online via the following link: <https://zenodo.org/record/4281122>. This dataset has undergone similar transformations (explained further below) as our dataset number 2 (refer to Figure 3.15). The key difference is that the creation of histories (along with time-averages) of OMNI data was accomplished using the `timehist2` function¹². This dataset is indispensable for us to effectively compare the outcomes yielded by PrecipNet and our own results.

In the subsequent part of this section, we will begin by providing a concise overview of the preprocessing steps undertaken on our measurements. We'll cover both the processes involving OMNIweb and DMSP data. Following that, we'll introduce the three final datasets resulting from the integration of inputs and outputs (as illustrated from Figures 3.14, 3.15, to Figures 3.16 and 3.18).

3.4.1 Pre-process Summary

Let's now sum up all the processes we've implemented so far for the inputs (*OMNIWeb High-Resolution*) and the outputs (*DMSP SSJ Electron Total Energy Flux*).

For *OMNIweb* (i.e., the IMF components, solar wind velocity, density, pressure, as well as AL, AU, and SYM/H indices):

- We switched out temperature and opted for pressure instead.
- Velocity, density, and pressure data underwent a base-10 logarithmic transformation, replacing the original data.
- We created two additional datasets using *PCHIP* interpolations. The first one filled all gaps of size 1, while the second one filled gaps of size 4 or smaller.

In total, we have three usable OMNIWeb-HR input datasets, as depicted in Figure 3.14.

Turning to *DMSP* (i.e., electron total energy flux):

- We started with the data from [Redmon et al. \(2017\)](#).
- We retained only the data where $|\text{MLAT}| \geq 45$ (focusing on the poles).
- We subsampled at the minute level.
- Data from 1987 and 1988 were excluded, and our focus narrowed down to the period from 2000 to 2014 (encompassing over 7 million measurements).
- We combined data from both the northern and southern hemispheres.
- Outliers, identified as those with a z-score exceeding 4.7, were excluded based on empirical assessment.
- We also preserved the subset of DMSP data used by [McGranaghan et al. \(2021\)](#).

In total, we have two output datasets: ours and the one from [McGranaghan et al. \(2021\)](#) (which is a subset of ours), as presented in Figure 3.14.

Furthermore, an analysis encompassing the causal connection between the inputs and outputs ([Bouriat et al., 2023](#)) has been undertaken, affirming the existing correlation between solar wind

12. https://github.com/rmcgranaghan/ISSI_geospaceParticles/blob/master/time_hist2.py, last accessed on August 2, 2023.

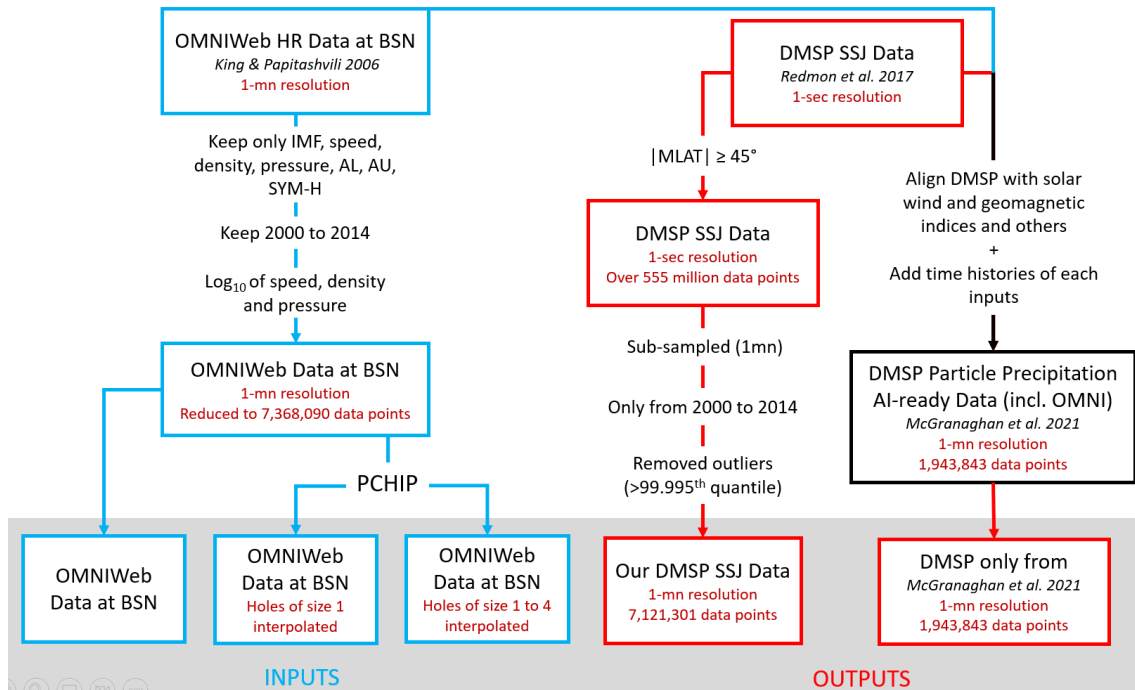


Figure 3.14 – Diagram illustrating the preparation of the available datasets (inputs in blue and outputs in red). The referenced articles are King and Papitashvili (2006), Redmon et al. (2017), and the AI ready dataset (here <https://zenodo.org/record/4281122>) from McGranaghan (2019); McGranaghan et al. (2021).

and the observable trends within electron precipitation. The model under development promises an enhanced level of insight, enabling precise modeling of temporal variations on a diminutive scale.

3.4.2 Combining inputs and outputs: final datasets

The concluding stage of our preprocessing involves the integration of input and output datasets to construct comprehensive samples. This entails aligning the temporal sequences of input and output data based on corresponding dates. Within our study, two distinct methods are employed for combining input and output datasets, contingent on the chosen algorithm: either a TCN or an FCNN. As a result of this amalgamation process, certain data points may be lost, notably in cases where data points are present on a given date in one dataset but not in the other. Nonetheless, we have meticulously ensured that the post-analysis implications we derived for each parameter, encompassing both inputs and outputs, remain coherent after this integration process. A detailed illustration of the formulation of these two final datasets, pivotal to our study, is provided in Figure 3.15.

Therefore, with three distinct input datasets and two distinct output datasets at hand, a total of twelve possible combinations arise. However, it's important to note that we won't be examining all twelve complete datasets in this context. Nevertheless, each of these combinations may offer valuable insights.

- The first dataset, as depicted in Figure 3.15, is designated for use with the TCN. It consistently comprises an index list, a high-resolution OMNIWeb dataset (specifically the downloaded data), and a dataset containing total electron energy flux information.

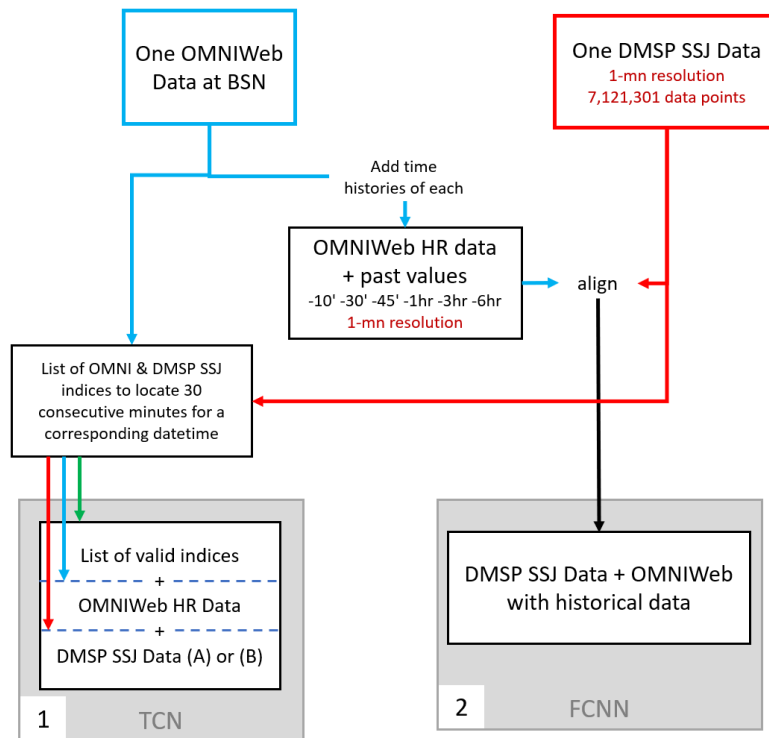


Figure 3.15 – Diagram illustrating the integration of inputs and outputs to produce the ultimate combined datasets employed in our study. Blue arrows delineate the transition of OMNI data, red arrows denote the total electron energy flux data, and black arrows represent the joint movement of both OMNI and DMSP data (often following synchronization of the two). The green arrow corresponds to the index list.

As detailed in the TCN description (Section 2.3.6.3), these networks mandate consecutively arranged temporal data and their corresponding labels. In our scenario, we've opted for a 30-minute historical window. This choice arises from the understanding that any event reaching the Bow Shock Nose (BSN) wouldn't require more than 30 minutes to affect Earth. To illustrate, let's consider a DMSP data point at 8:00 AM on January 12, 2010. For each DMSP data point, we aggregate the preceding 30 minutes of each parameter (e.g., solar wind velocity) leading up to that specific time (from 7:30 AM to 8:00 AM on January 12, 2010). This procedure transforms each training sample into a tensor featuring 30 instances of 9 parameters (at most) for each DMSP value.

However, the size of such a dataset can grow unwieldy and pose computational challenges. To address this, instead of generating a fresh dataset, we adopt a strategy to identify, for each DMSP value, the corresponding data positions within the OMNIWeb dataset (housing environmental condition data). This is realized through the creation of a *list of valid indices*.

For instance, if the DMSP value at 8:00 AM on January 12, 2010, corresponds to row 1,252,145 in the DMSP dataset, we search the OMNIWeb dataset for data from the same date and time. We record the position only if there are no missing data points between 8:00 AM and 7:30 AM on that day. If there are gaps in the data, we omit the addition to the "valid indices" list. However, if all data points are present, we note the position of the 8:00 AM data point on January 12, 2010 (for instance, row 5,526,689).

Consequently, an entry is appended to the list of valid indices, in the format: "01/12/2010 08:00 1,252,145 5,526,689." Employing this list, the algorithm seamlessly retrieves data from both the OMNIWeb and DMSP datasets. This approach circumvents the need to generate or modify a new dataset and mitigates memory-related complexities tied to dataset management. The algorithm efficiently employs the index data to directly access necessary information within both datasets, thus optimizing efficiency. The list is structured with three columns: datetime, corresponding index in the OMNI dataset, and corresponding index in the DMSP dataset.

This approach yields the input for the TCN using three datasets: any OMNI dataset, any DMSP dataset, and a list of valid indices employed by the algorithm. Utilizing this list, the TCN input (comprising 17 vectors of size 30, illustrated in Figure 3.16) can be seamlessly retrieved by the algorithm.

- To create dataset number 2, we followed a similar methodology as detailed in the work by McGranaghan et al. (2021). Initially, we accessed OMNIWeb data using the *nasaomnireader*¹³ package. For these data, we enriched each parameter with an additional 7 values, resulting in a maximum of 63 inputs (for 9 parameters) from the OMNI dataset. Specifically, for each data point at a given time and date T, we gathered values from T-10 minutes, T-30 minutes, and T-45 minutes earlier. Additionally, we computed averages around T-1 hour, T-3 hours, and T-6 hours prior. These averaging intervals spanned 30 minutes around T-10 and T-30, and 1 hour around T-60. The process is conceptually depicted in Figure 3.17 in McGranaghan et al. (2021). Notably, the authors initially intended to incorporate historical data from intervals of 5, 10, 15, 30, and 45 minutes, as well as 1, 3, 5, and 6 hours. However, following analysis, historical data at intervals of 5 minutes, 15 minutes, and 5 hours were

13. <https://github.com/lkilcommons/nasaomnireader> developed by Liam M. Kilcommons at CU Boulder, last accessed on August 2, 2023

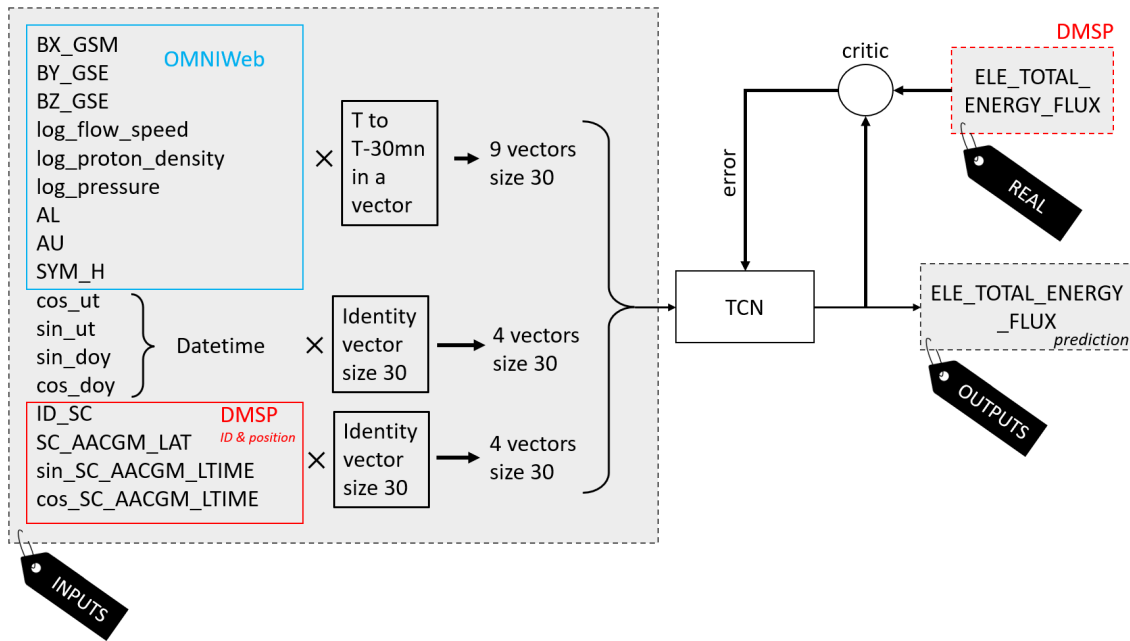


Figure 3.16 – Operational schema of the TCN with presentation of selected inputs and outputs.

excluded.

In essence, the sequential steps involved the following:

- Retrieval of DMSP data from CDWeb¹⁴.
- Data extraction at one-second intervals, encompassing parameters such as magnetic latitude in AACGM coordinates (labeled as SC_AACGM_LAT), magnetic local time in AACGM (noted as SC_AACGM_LTIME), and electron total energy flux (ELE_TOTAL_ENERGY_FLUX).
- Removal of outlier values (above the 99.995th quantile).
- Temporal alignment of data with the previously established OMNIWeb dataset.
- Transformation of cyclical variables through their sine and cosine values. This transformation is applied to Universal Time (UT), Day of Year (DOY), and Magnetic Local Time (SC_AACGM_LTIME).

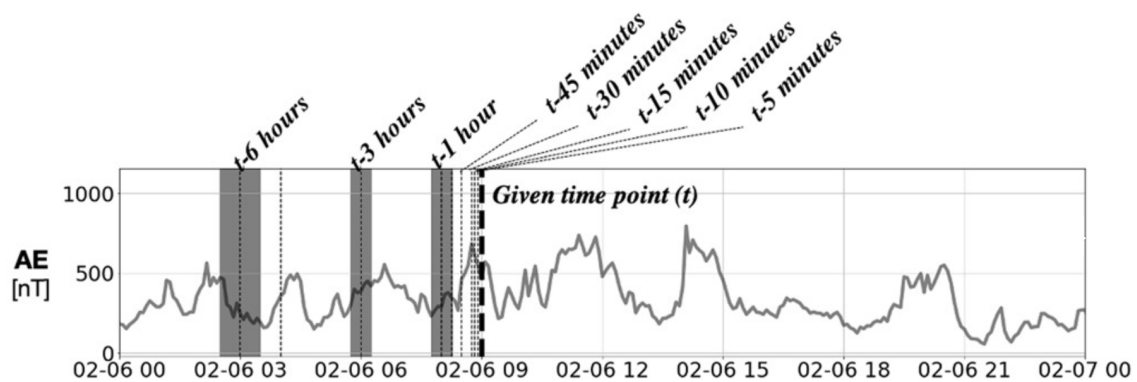


Figure 3.17 – Schematic representation of the historical data retrieval process for each parameter. Adapted from McGranaghan et al. (2021).

14. <https://cdweb.gsfc.nasa.gov/pub/data/dmsp/>, last accessed on August 2, 2023.

This approach results in the construction of our dataset tailored specifically for use with the FCNN. For instance, focusing solely on the OMNIWeb data without interpolation, our DMSP dataset yields over 3 million data points, whereas the dataset from [McGranaghan et al. \(2021\)](#) provides nearly 2 million. Our output exclusively encompasses "ELE_TOTAL_ENERGY_FLUX," while the potential inputs encompass the comprehensive array of data illustrated in Figure 3.18. This figure succinctly captures the utilization of our dataset 2, meticulously designed for seamless integration with the fully connected neural network.

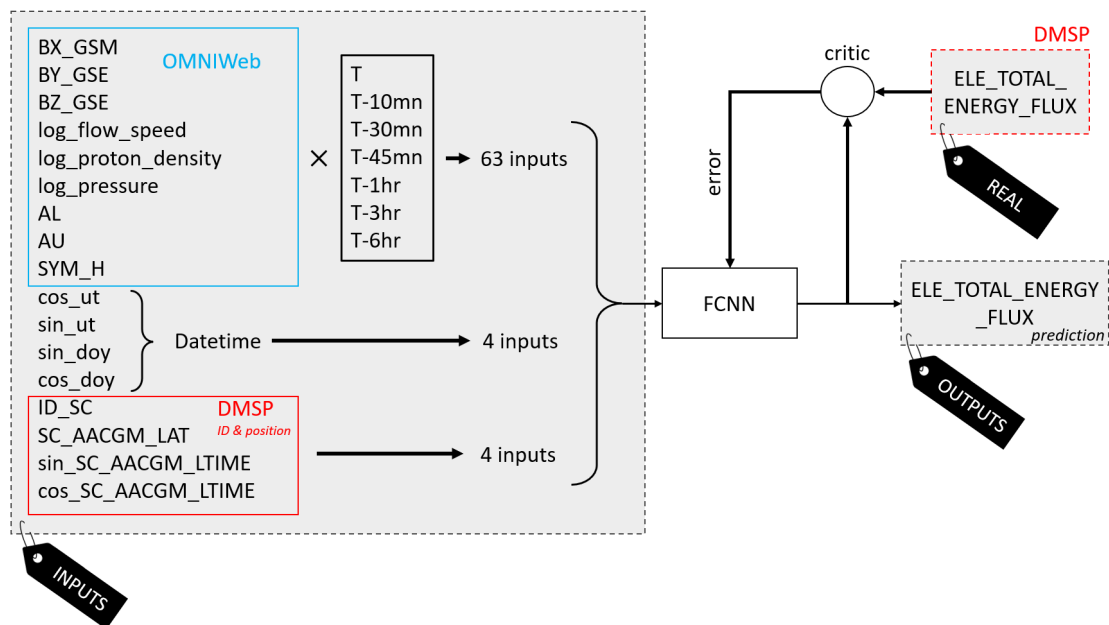


Figure 3.18 – Illustration of inputs and outputs within the context of a fully connected neural network using our dataset number 2 (we may selectively utilize data from the "INPUTS" category).

3.4.3 Summary of our study

To conclude this chapter, we would like to recap the key aspects of our study:

- Exploring PrecipNet:** Initially, our focus will be on PrecipNet, given its relevant findings. To achieve this, we will prepare our dataset specifically for the FCNN (number 2 on Figure 3.15). We will not take an interpolated input, and we will restrict the outputs to the DMSP values used by PrecipNet. Consequently, the inputs will not be strictly identical for our training and that of PrecipNet (since the latter also includes $F_{10.7}$ and the Polar Cap index). Our primary objective is to attempt the replication and comprehensive understanding of the limitations of PrecipNet.
- Introducing our FCNN:** To delve further and rectify certain aspects of PrecipNet, we will proceed to apply our customized FCNN. This time, the model will be trained on the comprehensive dataset comprising over 3 million samples from "our" DMSP SSJ dataset. A comprehensive comparison of our results with those of PrecipNet and OVATION Prime will ensue.
- Improvement and Forecast:** Subsequently, we will investigate the effects of varying input choices. By excluding the data at T_0 and retaining only past data (historical), we will substantiate the relevance and feasibility of future research endeavors concerning short-term forecasting of precipitating electrons.
- Addressing Limitations with the TCN:** However, as we will observe, both PrecipNet and our FCNN exhibit limitations. Particularly, the modeling of precipitating electrons relies

on the use of arbitrarily chosen historical data points. It is reasonable to consider that the dynamic and intricate movements of the solar wind contradict the notion that pertinent information for modeling our phenomenon would be confined to fixed temporal points in the past (e.g., T-10 mins, T-30 mins, etc.). To address this, in the fourth phase, we will introduce a TCN (dataset number 1) to tackle the challenge of subjectively selecting "past" data points. Unlike PrecipNet's arbitrary historical data selection, our TCN approach incorporates all data within a specified past interval. It's important to note that, due to time and computational constraints, we have confined our dataset for this section to the DMSP data within the records of [McGranaghan et al. \(2021\)](#).

4

Algorithmic Implementation, Iterative Procedures, and Results

"Every path is the right path. Everything could've been anything else. And it would have just as much meaning."

Nemo Nobody - Mr. Nobody - Jaco van Dormael

Contents

4.1	Organization of our code	218
4.2	Exploring PrecipNet	219
4.2.1	PrecipNet	219
4.2.2	First Remarks	219
4.2.3	Attempt to Simplify PrecipNet	221
4.2.4	Conclusion	226
4.3	Introducing Our FCNNs	227
4.3.1	Training Process	227
4.3.2	Visualizing Results	228
4.3.3	Performance Metrics	228
4.3.4	Comparative Analysis Insights	234
4.4	Forecast	238
4.5	Addressing Limitations with the TCN	240
4.5.1	Training Process	240
4.5.2	Visualizing Results	242
4.5.3	Performance Metrics	242
4.5.4	Insights on the Final Product & Comparative Analysis	243
4.5.5	A Step Further: Integrated Gradients	245
4.6	Final Remarks	248

4.1 Organization of our code

This section immediately follows the conclusion of Chapter 2. In this part, we will outline the overall structure of our code, a structure that will remain consistent for all the algorithms we will implement.

Figure 4.1 presents an overview of how our code is organized. The majority of classes and functions in this code are to be implemented by the user. Here, we will take the time to describe the different classes that need to be completed and their roles in the code structure.

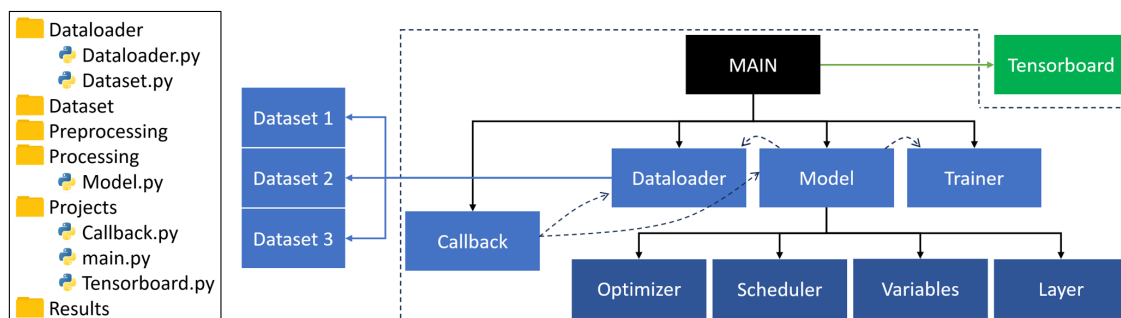


Figure 4.1 – Organization of our code using the PyTorch Lightning framework. A solid arrow pointing from box X to box Y indicates that X instantiates Y . A dashed arrow pointing from box X to box Y indicates that X refers to Y (i.e., in X , we can perform the operation $self.Y$).

The "Dataset" folder contains all the data, usually divided into two groups: "Raw" for raw data and "Processed" for data that has undergone processing using the functions located in the "Preprocessing" folder. The "Results" folder will contain all the results and functions that aid in result visualization. The "Trainer" block, which is not covered here, actually corresponds to a class in PyTorch Lightning that handles the training loop entirely, as shown in Figure 2.32. Let's describe the remaining files:

- `Dataset.py`: This Python file will contain the "Dataset" class, which inherits from `torch.utils.data.Dataset`. Its role is to call and instantiate the data. It contains three functions: an `__init__` function that runs once when instantiating the Dataset object, loading the data from the appropriate folder; a `__len__` function that returns the number of samples in the dataset, usually by using the `len` function in Python; and a `__getitem__` function that loads and returns a sample located at index `idx`, a parameter passed as input.
- `Dataloader.py`: This Python file will contain the "Dataloader" class, which inherits from `pytorch_lightning.LightningModule`. This class will prepare the training and validation datasets and return PyTorch DataLoader objects.
- `model.py`: This Python file will contain the model class, named after the model itself, which also inherits from `pytorch_lightning.LightningModule`. It contains the following essential functions for training: `forward` to define the forward method, such as the chaining of linear layers and activation functions; `training_step` to compute and return the training loss and/or additional metrics (the same for `validation_step` and `test_step`); `configure_optimizer` to configure the optimizer; and a few optional additional functions, such as `train_epoch_end`, which allows executing functions at the end of an epoch.
- `Callback.py`: Contains all the callback functions. In our case, it mainly involves saving the model's training progress to display it on Tensorboard.
- `Tensorboard.py`: In this final Python file, we implement everything necessary for Tensorboard to function.

- `main.py`: This Python file prepares everything for training and is the one that the end user will utilize. Inside, the user can modify all the hyperparameters.

The "main.py" file contains all the hyperparameters defined by the user. Firstly, it instantiates the PyTorch Lightning DataLoader (corresponding to step 5 in Figure 2.31). To do this, it first creates a Dataset object that loads the data, and then a DataLoader object based on the Dataset object. Next, the model is instantiated with all its associated hyperparameters (similar to step 3 in Figure 2.31). We then instantiate Tensorboard, and finally, the Trainer that will train our model. We conclude by saving the model, allowing it to be applied to new data for making predictions.

Now that we presented the overall organization, we will try to reproduce and understand PrecipNet. Then, we will try to improve its results.

4.2 Exploring PrecipNet

4.2.1 PrecipNet

PrecipNet is a Fully-Connected Neural Network implemented by [McGranaghan et al. \(2021\)](#) with the aim of modeling precipitated electrons as measured by DMSP satellites. Its characteristics are as follows:

- 72 features: X, Y, and Z components of the IMF, AL, AU, SymH, PC, speed & density of the solar wind, each for (T0, T0-6hr, T0-3hr, T0-1hr, T0-45mn, T0-10mn, see Figure 3.17), and F10.7, as well as the spacecraft ID, spacecraft position (AACGM's latitude, and cosine and sine of AACGM's local time), and elements of the datetime (cosine and sine of Universal Time, and of the day of the year).
- Epochs: 1000 (with an early stopping criterion)
- Batch size: 32,768
- Dropout value: 0.1 (the dropout layer is positioned after the first hidden layer of the network)
- Network Architecture: 8 hidden layers [256, 64, 32, 256, 1024, 256, 32, 4]
- Learning rate: 0.001

We have already introduced the datasets in Chapter 3. Here are some reminders and remarks about them:

- Label: $\log_{10}(\text{Electron Total Energy Flux} \times \pi)$ (multiplying by pi compensates for steradians)
- Training set: 1,890,579 data points
- Validation set: All 2010 data from the satellite F16 = 55,210 data points
- Test set: No test set
- Optimizer: Adam

An essential point to mention is the choice of a Robust Scaler for scaling the data. As discussed in Section 2.3.6.2, the data must undergo scaling or normalization. The method we and the authors of PrecipNet have chosen here is Robust Scaling (see 2.3.6.2).

4.2.2 First Remarks

Several observations can be made upon an initial examination of PrecipNet: there is no test set, a specific and potentially unrepresentative satellite was selected for the validation set, and the

architecture seems overly intricate.

Addressing the first point, the authors split the dataset into two subsets: the training and validation sets. As previously mentioned in Chapter 2, we believe that a test set is crucial for evaluating the model's accuracy. In the article by [McGranaghan et al. \(2021\)](#), the results shown in the table (which we've included below, Figure 4.2) represent the performance on the validation set. Yet, a data scientist's choice of hyperparameters to optimize their algorithm is based on performance on the validation set. In other words, while they strive to optimize the algorithm in general, they gradually tune it to perform well on the validation set. A test set is needed to provide an unbiased performance metric. That's the approach we'll adopt.

Metric	PrecipNet (mean \pm 1 σ)
PE	0.751 \pm 0.007
Intercept of linear fit	2.632 \pm 0.248
Slope of linear fit	0.717 \pm 0.027
RMSE	0.764 \pm 0.011
MAE	0.558 \pm 0.019

Figure 4.2 – Model Evaluation Metrics for the machine Learning Model (PrecipNet) from [McGranaghan et al. \(2021\)](#).

Our data analysis revealed that data distribution was heavily skewed across satellites (differences between satellites). For training PrecipNet, the authors chose the entire 2010 data from satellite F16 as the validation set. When segmenting time series data into training, validation, and test sets, it's imperative to maintain their temporal order. This replicates real-world scenarios where the model is trained on past data and tested on future data. For independent and identically distributed data, cross-validation is the typical approach [Geisser \(1975\)](#). However, with time series, it's often advisable to use "time-based splitting" to prevent "data leakage". Data leakage occurs when future data (which the model shouldn't have access to) accidentally seep into the training, validation, or test datasets. This can bias the model's performance and may not reflect its ability to make predictions on new data.

For instance, if an algorithm uses 10 consecutive time points as inputs to predict the 11th, and by random separation, the label at T1 ends up in the training set while T0 is in the test set, then the 10 data points preceding T1 will be in the training set as inputs. This includes the value at T0. Thus, the 9 points preceding T0 as well as T0 are both in the training and test sets: the test set "leaks" into the training set. In certain contexts, using time-based splitting for training and validation datasets can minimize this data leakage, making it more appropriate than a random split ([Lyu et al., 2021](#)). Unfortunately, we noticed quite late that we could experience some data leak with our random split. To ensure the viability of our results, we performed some trainings doing a time-based splitting and observed no difference with our current results. The choice made by [McGranaghan et al. \(2021\)](#) is intriguing since they chose an entire year as the validation set. However, Figure 4.3 indicates that the data used for validation isn't representative of the overall set. Consequently, the algorithm will struggle to evaluate itself on the validation set. As a result, a noticeable and irremediable gap will be evident between the validation loss and training loss curves. Figure 4.4 (left) displays the evolution of the Loss curves provided by [McGranaghan et al. \(2021\)](#) for the training and validation sets. As anticipated, the validation loss remains significantly above the training loss.

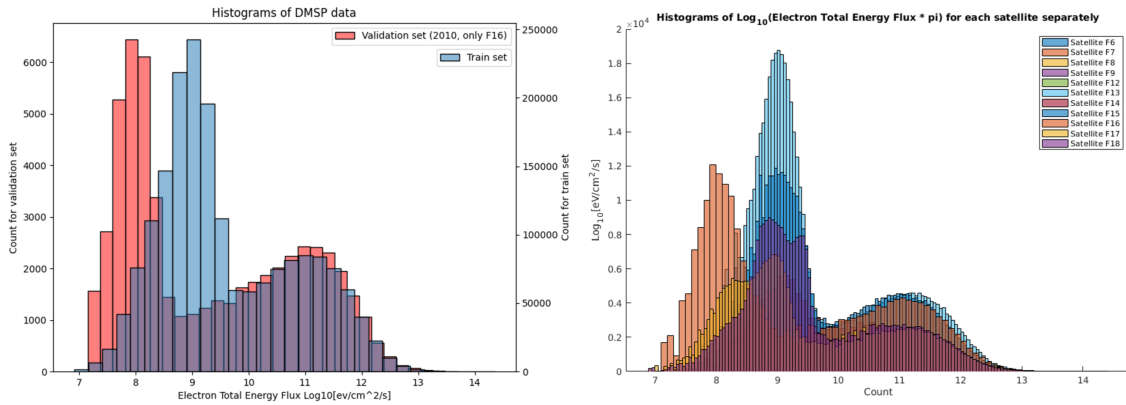


Figure 4.3 – Left: Histograms of precipitated electron flux values on the training set and the validation set where the validation set are data from the satellite F16 in the year 2010. Right: Histograms of data from all satellites separately, showing a clear difference between F16 and the others.

In the following section, we will present our internal reproduction of PrecipNet and its results, which will allow us to address the theme of architectural complexity.

4.2.3 Attempt to Simplify PrecipNet

4.2.3.1 Reproduction of PrecipNet

From this point onward, we internally reproduced PrecipNet. For this, we retained the DMSP data from the AI-Ready dataset (<https://zenodo.org/record/4281122>) and trained the same algorithm with the same inputs as PrecipNet. The distinction lies only in how we sourced the OMNIWeb values. This corresponds to dataset number 2 from Figure 3.15. The decision to exclude the OMNIWeb portion of the "AI-Ready" dataset by McGranaghan et al. (2021) arises from its construction and, in part, from the utilization of the function `time_hist2`¹ to compute the time histories for each parameter. This function employs linear interpolation across the entire dataset, and as discussed in Section 3.3.1, this approach poses challenges. Thus, the tests and conclusions drawn from our version of PrecipNet apply solely to our rendition. However, given that the data is approximately the same, we are confident that these conclusions generalize to PrecipNet itself. Figure 4.4 displays both outcomes side by side (PrecipNet and our reproduction). Our reproduction achieved results that deviated by less than one percent (0.54%) from PrecipNet (refer to Figure 4.4), which we deemed accurate. Henceforth, we will refer to our internal replication of PrecipNet as **PrecipNet-R**.

As mentioned in the previous section, we initially trained PrecipNet-R on different data, selected randomly to avoid using satellite F16's data in the validation set. Figure 4.5 highlights this modification and confirms that the disparity between validation and training in the results from McGranaghan et al. (2021) is indeed due to this specific data split choice. After executing a random split, we obtained a final MSE of 0.389 on a test set set aside beforehand, and we recorded an MSE of 0.506 for the 2010 data from satellite F16. With PrecipNet-R, we had achieved 0.555 (see Figure 4.4), translating to an improvement of 8.7%. It's worth noting that some data from the 2010 year of satellite F16 might have ended up in the training set, which could bias our evaluation towards a better outcome. However, we clearly demonstrate here that the validation set choice by McGranaghan et al. (2021) complicates the convergence of the validation loss towards a minimum.

1. https://github.com/rmcgranaghan/ISSI_geospaceParticles/blob/master/time_hist2.py

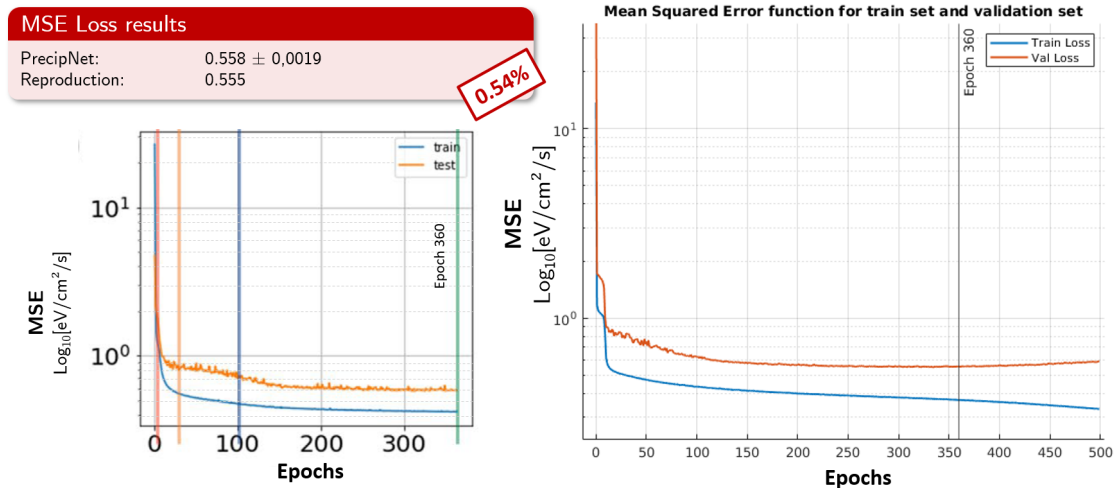


Figure 4.4 – Training and Validation Loss from *McGranaghan et al. (2021)* on the left and from our reproduction on the right. The Loss function is a mean squared error (MSE).

4.2.3.2 The role of hidden layers and overfitting

We trained PrecipNet-R for a higher number of epochs (surpassing the early stopping criterion set by *McGranaghan et al. (2021)*). After waiting for 1000 epochs, we observed that the validation loss began to increase while the training loss continued to decrease (Figure 4.6). This is indicative of overfitting, suggesting that the network is overparameterized. We then questioned if it was possible to reduce the size of the network, and we examined how it performed while keeping the total number of parameters constant.

For a given parameter count (approximately 539k), we increased the number of hidden layers and juxtaposed the loss curves. The results and the architecture are illustrated in Figure 4.7. The conclusion is clear: adding more layers introduces unnecessary complexity, leading to overfitting, even when the parameter count remains constant.

Subsequently, we attempted to simplify PrecipNet-R’s architecture to demonstrate that comparable results could be achieved with reduced complexity. A simplified design results in shorter runtimes and reduced computational resources. Figure 4.8 shows that with roughly ten times fewer parameters (52,105 vs. 579,593), the outcome diverges by only 2.5%: yielding an MSE of 0.569 compared to 0.555 for PrecipNet-R, with a computational time reduced by a factor of 3.5. For industrial applications, sacrificing this small amount of precision seems reasonable if it results in more than a threefold reduction in processing time.

4.2.3.3 The role of position & time as inputs

Lastly, in this section, we evaluated the importance and role of inputs related to position and date. As observed, precipitated electrons form the auroral oval (a combination of the northern and southern hemispheres for our study). The shape of this oval changes based on solar wind parameters (*Bouriat et al., 2023*) and also evolves throughout the 11-year solar cycle (a next study could be to perform a sensitivity study and compare the oval for solar min and solar max, all other parameters remaining identical). However, its overall form remains generally consistent: an oval centered around the poles. Consequently, we hypothesize that the position and time values

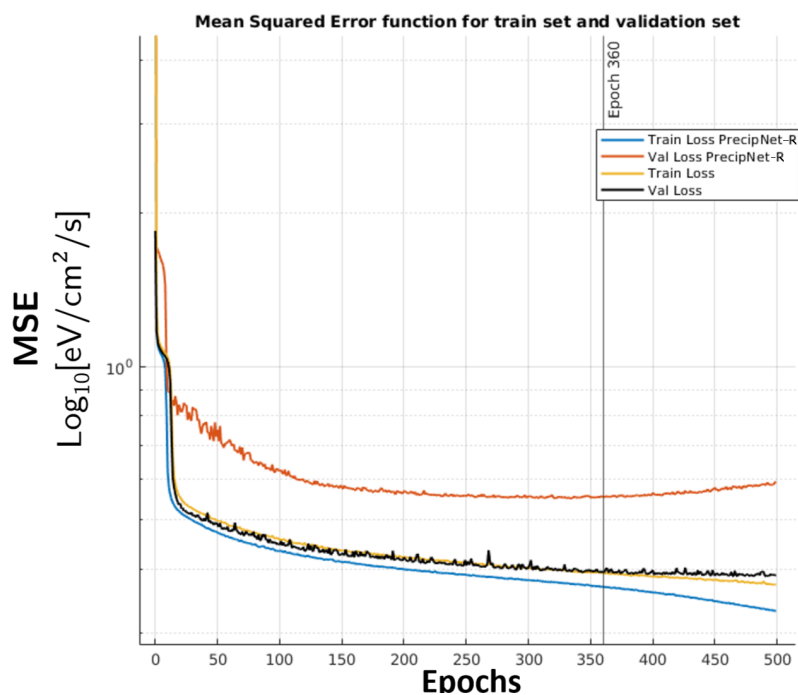


Figure 4.5 – Loss curves for the PrecipNet-R training, both for random sampling to produce the training (yellow curve) and validation sets (black curve), and the choice of the 2010 year of satellite F16 (blue and red curves for training and validation sets, respectively).

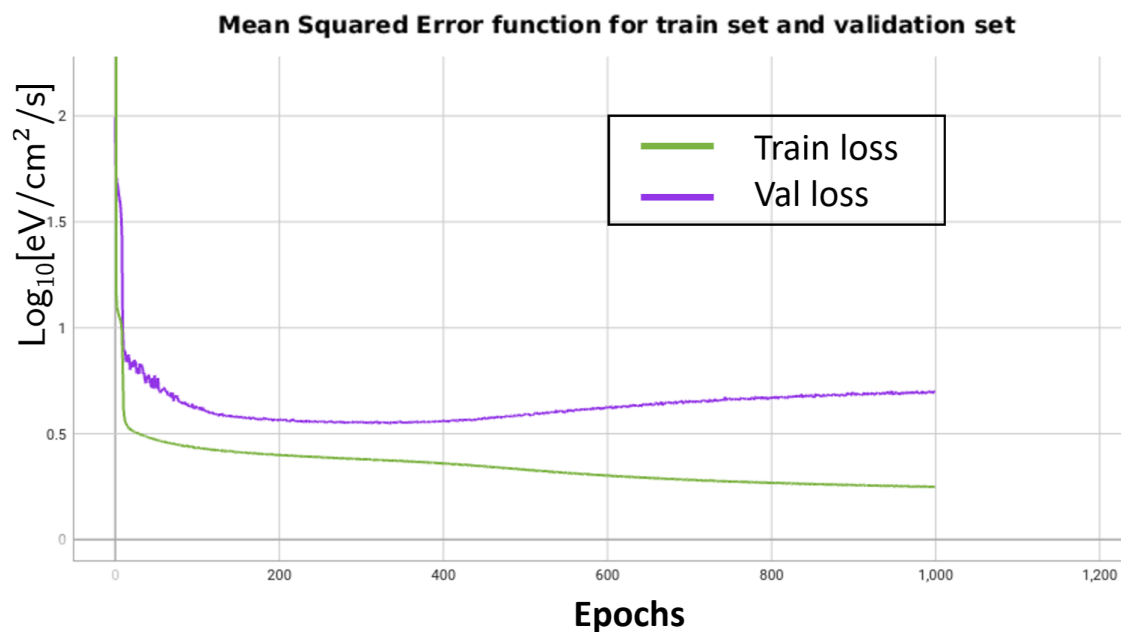


Figure 4.6 – Training and Validation loss functions for PrecipNet-R during training over 1000 epochs without the early-stopping criterion.

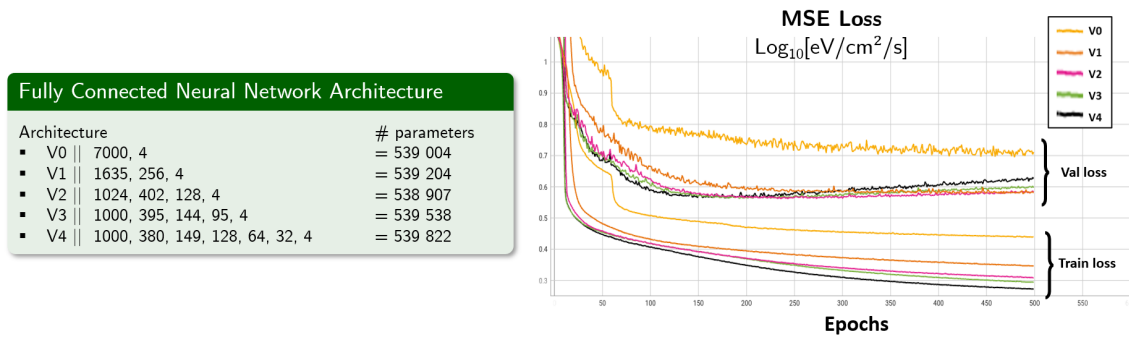


Figure 4.7 – PrecipNet-R training for various architectures without Dropout. Numbers found in the description of the architecture correspond to the amount of neurons in each layer (e.g., V0 is made of to 2 hidden layers containing respectively 7000 and 4 neurons).

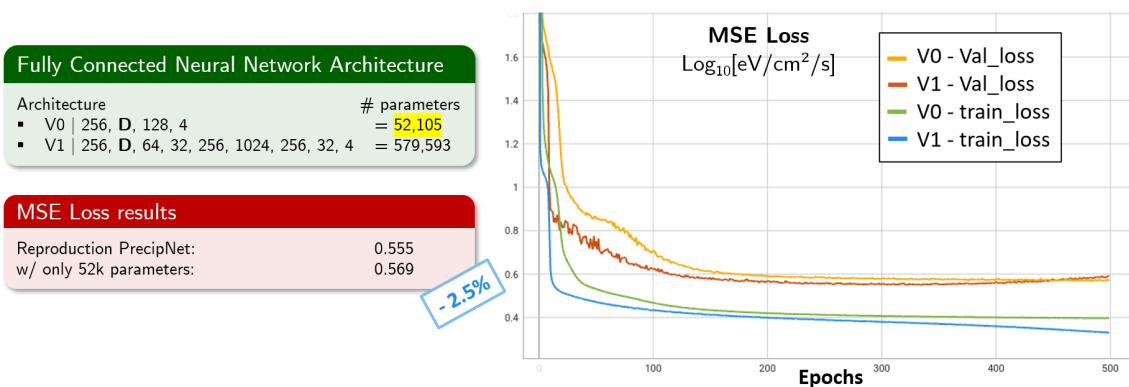


Figure 4.8 – Original PrecipNet-R (V1) vs. a simpler architecture (V0) loss curves.

of DMSP satellites are vital inputs for the algorithm. With this information, it should primarily model the energetic flux values. Knowing the position, it can determine if it is, on average, within the oval. Inside or outside, it would then have a ballpark estimate of the response to produce. This poses an issue as we are most interested in the dynamic variations of the highest values within the oval. To verify this, we trained and made predictions using PrecipNet-R, retaining only position values as inputs (AACGM's latitude, and cosine and sine of the AACGM's local time), and datetime elements (cosine and sine of Universal Time, and day of the year). Figures 4.10 & 4.9 showcase the results of these simulations.

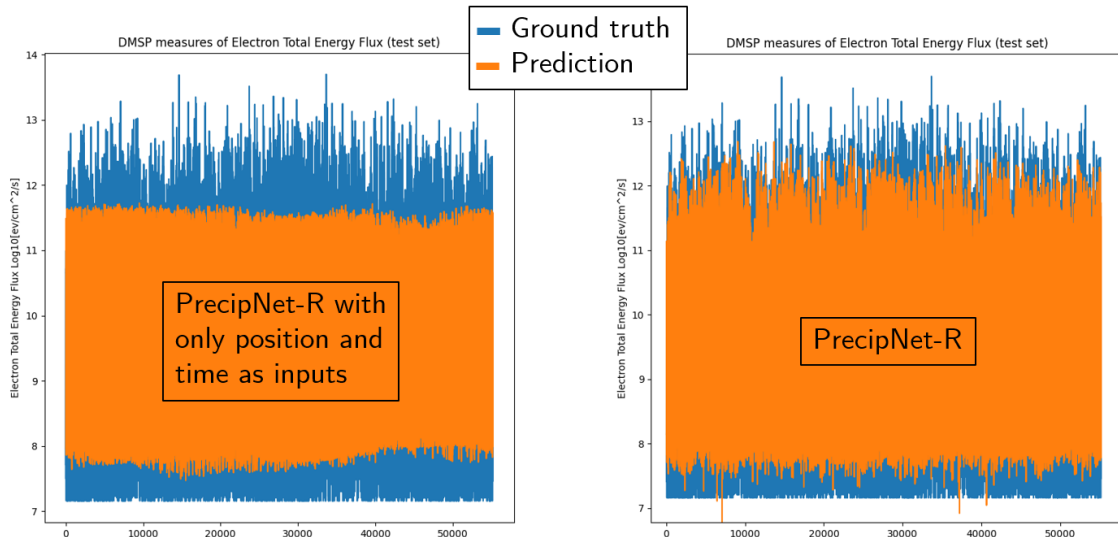


Figure 4.9 – Actual values compared with model predictions for the set-aside test set. The model is PrecipNet-R with position and date inputs only shown on the left and the original PrecipNet-R on the right.

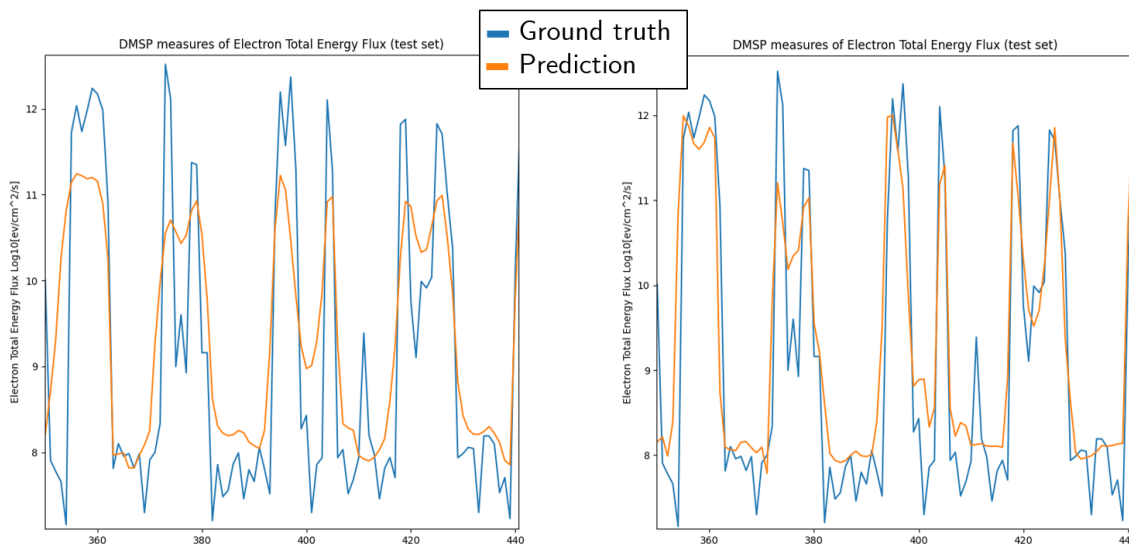


Figure 4.10 – A closer look at the curves observed in Figure 4.9.

As anticipated, nearly 60% of our predictive capabilities solely rely on position and time data. As evident in Figures 4.10 & 4.9, the model that relies only on position and time data does not predict extreme data points (high or low) as accurately. However, the primary trends are well-

captured by the algorithm (especially evident in Figure 4.10). It has grasped the regularity with which a satellite enters and exits the auroral oval.

Ideally, it would be beneficial to eliminate this information. Indeed, the extreme values, which are harder to predict, are the ones of most interest to us. If it were possible to subtract the average flux values from the "average" oval over eleven years, we would compel the algorithm to focus not on these medium-term trends but on local flux variations. Crudely put, a significant portion of PrecipNet-R's (and therefore, of PrecipNet's) computational power is devoted to replicating these periodic trends evident in Figure 4.10. However, these trends only correspond to the average position of the oval and the fact that the satellite periodically orbits within it. They do not reflect a physical reality of a periodicity in the electron energy flux. These are merely consequences of an observer (the satellite) orbiting the phenomenon.

One solution we contemplated was data isolation and the multiplication of the networks to use. In other words, we began to create neural networks for each MLAT-MLT pair. Each network focuses on a unique 1° MLAT-1h MLT cell within a grid. In Figure 4.11, we present results for the cell at 12:00 MLT (from 11h to 12h MLT) and 75° MLAT (magnetic latitude between 74° and 75°).

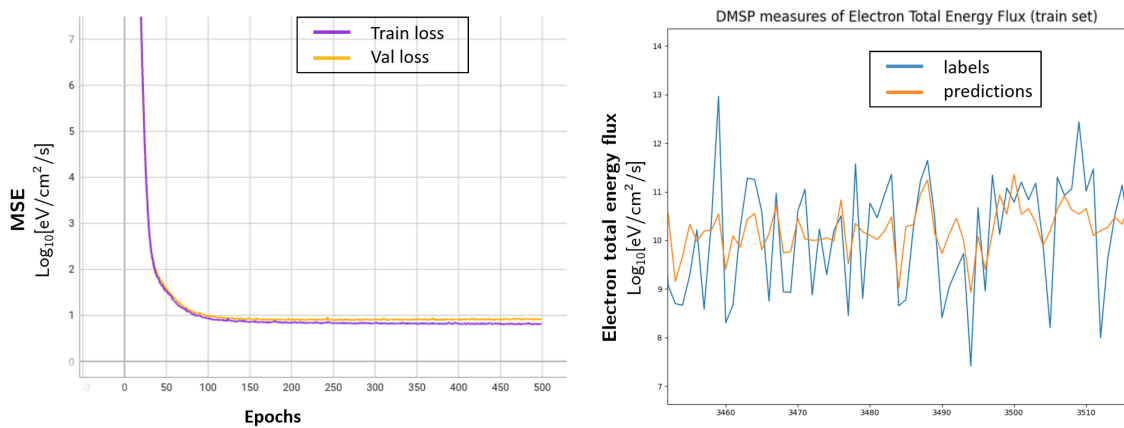


Figure 4.11 – On the left: training and validation loss functions for the 12:00 MLT cell (from 11h to 12h MLT) and 75° MLAT (magnetic latitude between 74° and 75°). On the right: a closer comparison between labels and predictions on the training set.

The results displayed here appear promising (total MSE: 0.895). Our network trains rapidly and employs an architecture with two layers of size 4. Unfortunately, the training set for this grid cell contains only 5084 data points. Consequently, a more complex architecture (with more parameters) quickly results in overfitting. Moreover, this is the cell with the most training points in our entire dataset. We quickly reached the limits of this model. This is why we decided to revert to the idea behind PrecipNet-R and include position and time data inputs.

4.2.4 Conclusion

To summarize this section, let's review the identified challenges and difficulties:

- We reproduced PrecipNet internally with the same architecture and DMSP data (labels). Only the input data differed (OMNI) as we wished to avoid linear interpolation (included in *time_hist2*) to preserve the integrity of the original data. As a result, we have slightly fewer training points, but we effectively approximate PrecipNet's results. We've named this collective PrecipNet-R.

- Using the 2010 data from the F16 satellite for validation seems odd since it's the only satellite whose data is challenging to compare with the rest. Using random sampling rectifies the divergence between validation and training curves.
- A higher number of layers adds complexity to the network, causing overfitting even with a constant parameter count. Therefore, minimizing the number of layers is essential to avoid overfitting and reduce computational time and resource requirements.
- With an architecture containing ten times fewer parameters and computational time almost reduced by the same factor, we achieve results that are only 2.5% worse than before. A simple architecture is preferable, especially for industrial applications.
- By adopting PrecipNet-R's architecture and only considering position and time parameters as inputs, we achieve 60% of the initial predictive capabilities. Therefore, the majority of the information the algorithm focuses on is the position and shape of the auroral oval (and thus, of the satellite). This is where much of its computational capacity is directed.
- Sadly, isolating and modeling through a straightforward neural network (two layers of size 4) for a specific region of the oval quickly reveals its limitations: not enough data to extend the method across the entire oval, rapid overfitting, and challenges in applying deep learning.

Based on these findings and our existing knowledge, we aim to develop our neural network with the following concepts in mind:

- Focus on a simple architecture, keeping in mind the subsequent industrial application, and to avoid overfitting.
- Employ random sampling as the dataset split method.
- Devise strategies to predict extreme values without the algorithm overly concentrating on average values, which are easily predictable using only position and time data.
- Utilize our DMSP dataset, comprising over 3 million samples, hoping this will enhance predictive capacity.
- Implement a more comprehensive display to track the algorithm's modeling quality.

One lingering issue we've yet to address pertains to the dynamic delay between input and output. Indeed, we've retained historically relevant data points as deemed by [McGranaghan et al. \(2021\)](#), but nothing assures us that all pertinent information indeed exists at these timestamps in the past. This problem has only two solutions: to include even more data points from the past (which would exponentially increase feature count and, in turn, model complexity) or to change the algorithm. We'll revisit this when implementing the TCN.

4.3 Introducing Our FCNNs

4.3.1 Training Process

As touched upon at the end of the previous chapter, we will now utilize our dataset. This dataset contains 3,367,245 records, is prepared using Method 2 from Figure 3.15, and contains 71 inputs as depicted in Figure 3.18. For this, we've set aside 25%, or 841,811 records, for testing. The remainder will be used with 25% for validation (631,358 records) and 75% for training (1,894,076 records).

A few preliminary remarks: given the challenges associated with satellite measurements, our dataset is inherently noisy and thus, hard to train without inducing overfitting. Additionally, given

the histograms of flux measurements, we can assume the algorithm might focus on central data more than on extreme, rarer data. Thus, with a larger dataset than PrecipNet-R, we anticipate a slightly lower cost function. We've indeed added more central data points than rare, extreme ones (this could be compensated later on, for example by giving a larger weight in the cost function to extreme values).

The four primary models retained after numerous trials are:

- A basic FCNN model with approximately one-third of PrecipNet-R's parameters.
- An FCNN model with an autoencoder architecture.
- And the same two architectures but with a specialized loss function: the Tail Weighted Loss (Qi and Majda, 2020; Ziegler and Mcgranaghan, 2021).

Recall that our goal is modeling the energetic electron fluxes in low Earth orbit. However, our industrial application at SpaceAble emphasizes the importance of achieving maximum accuracy during extreme events, which we know are the hardest to predict due to their rarity (see Figure 4.9). Traditional mean squared error (MSE) loss might not effectively penalize errors for these extreme values. The Tail Weighted loss function is initially a classic MSE function, but penalizes the neural network when it underpredicts values above a certain threshold. This helps the network prioritize accurate predictions for extreme values, crucial in many real-world scenarios. The formula is as follows (Qi and Majda, 2020):

$$\text{Tail Weighted Loss} : \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 (1 + a) \quad (4.1)$$

In this formula, $a = 0$, except when the true value y exceeds a threshold y_r and the prediction \hat{y} falls below this threshold y_r . In such cases, a increases the total loss, penalizing the network. After several tests, the best results were achieved with pairs (y_r, a) of: $(2.5, 12)$, $(5, 12.5)$, $(10, 13)$, $(12, 13.25)$, $(15, 13.5)$. In this context, y_r represents $\log_{10}(\text{Electron Total Energy Flux} \times \pi)$.

The following table outlines the chosen hyperparameters for these four models, juxtaposed against PrecipNet's architecture for reference. The "architecture" row shows the hidden layers and their neuron counts. The letter "D" indicates where dropout was applied in the network.

4.3.2 Visualizing Results

To provide a comprehensive view of our training and results, we offer several visualization figures below, which we'll discuss in the subsequent section. We'll showcase training and validation curves, a comparison between the test dataset's histograms and the original ones, and certain curves to assess the accuracy of our predictions, such as displaying the density estimates of model predictions against the DMSP test data. Throughout, we'll also display results from OVATION, which serves as our baseline since it is currently the most widely adopted algorithm in the community.

4.3.3 Performance Metrics

The tables below detail our test dataset results using the chosen metrics. We believe the most relevant metrics for our purpose are the Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and training time.

The MAE gauges the average discrepancy between predictions and actual values. It treats all errors uniformly, without emphasizing any specific one. Essentially, it answers the question: "On

Aspect	Model 1 (FC)	Model 2 (AE)	Model 3	Model 4	PrecipNet architecture
Architecture	[512 D 256 20]	[512 D 32 16 32 512]	[512 D 256 20]	[512 D 32 16 32 512]	[256 D 64 32 256 1024 256 32 4]
Number of Parameters	173,353	71,761	173,353	71,761	579,337
Epochs	200	150	200	150	360 (with no early stopping)
Batch Size	10,000	10,000	10,000	10,000	32,768
Learning Rate	0.001	0.001	0.001	0.001	0.001
Loss Function	MSE	MSE	Tail Weighted Loss	Tail Weighted Loss	MSE
Optimizer	Adam	Adam	Adam	Adam	Adam
Lambda 1 & 2	None	None	None	None	None
Scheduler	None	None	None	None	None
Dropout	0.1	0.1	0.1	0.1	0.1
Training Time	45.10 min	37.07 min	43.58 min	37.01 min	1.501 hr

Table 4.1 – Table presenting the hyperparameters of the four final models used. Note: FC = Fully-Connected. AE = Autoencoder. The hardware used was from a single MSI computer (GPU: nvidia Quadro RTX; CPU: Intel i7-10875H @3GHz).

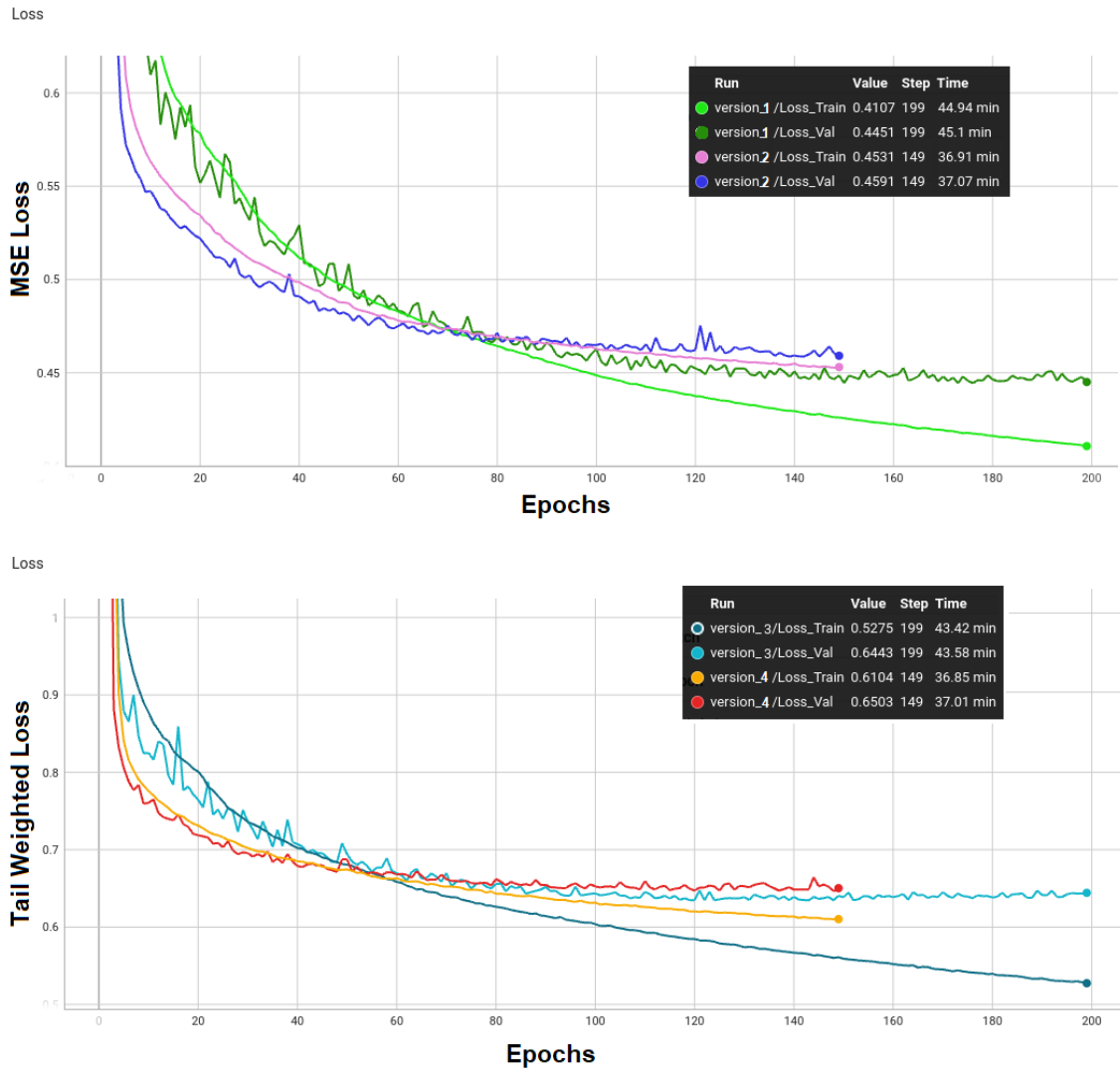


Figure 4.12 – Top: Training and validation curves for models 1 and 2. Bottom: Training and validation curves for models 3 and 4.

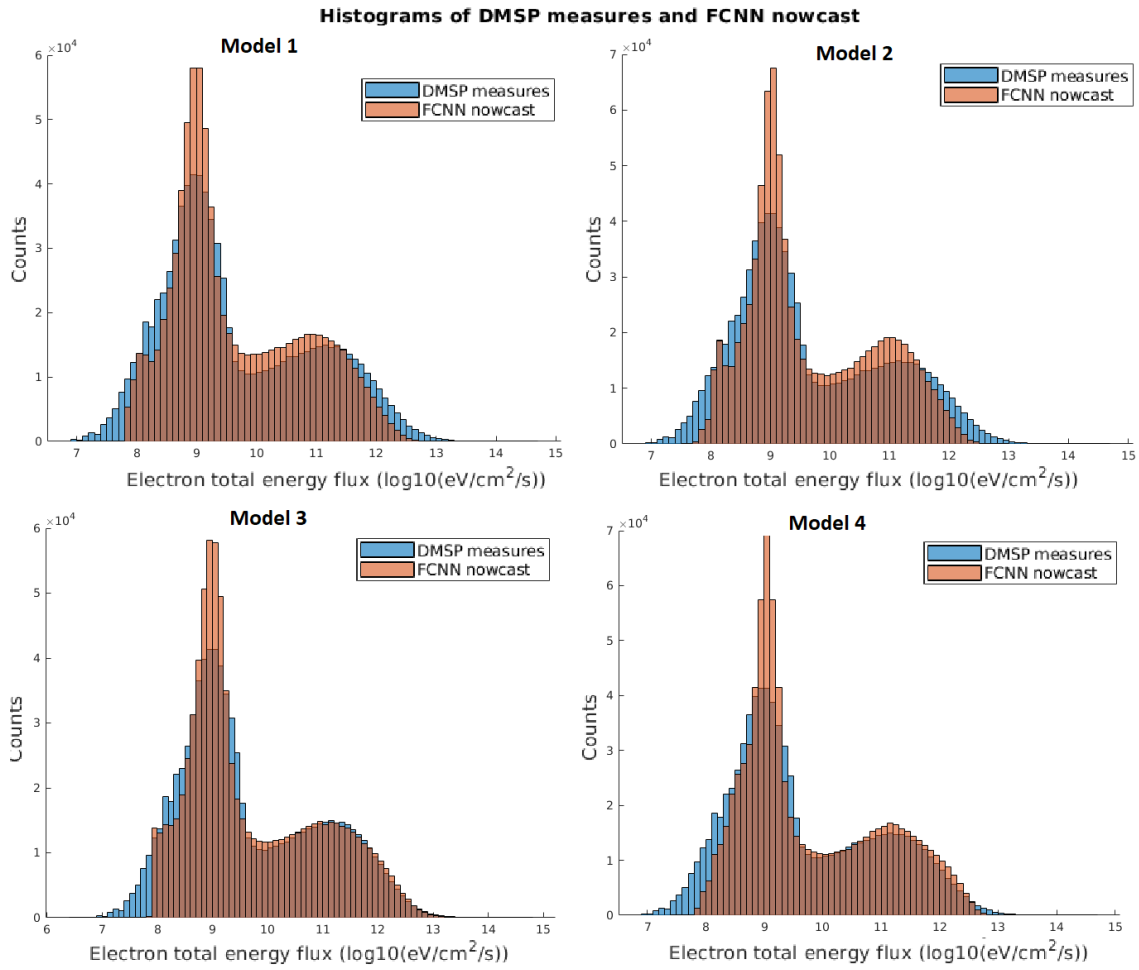


Figure 4.13 – Histogram comparison of the test dataset versus the predictions made by trained models 1 to 4.

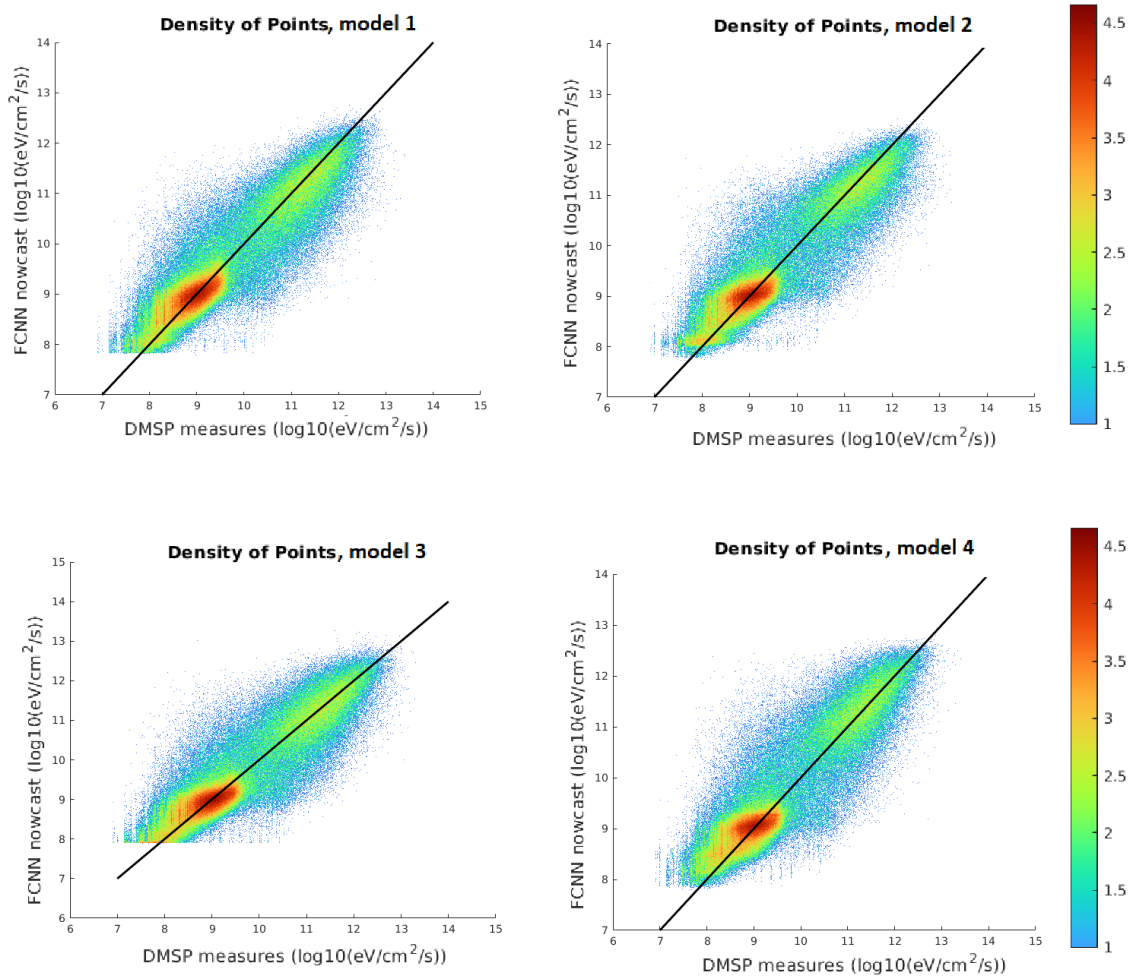


Figure 4.14 – Density displays of the test dataset data points versus predictions made by the three models. Density is shown in logarithmic scale for visualization clarity, with 1 added to the result to represent gaps in white. Thus, no point has a value between 0 and 1. The line $x = y$ is highlighted in black.

average, what's the difference between reality and prediction?"

On the other hand, MSE emphasizes larger errors by squaring them. This means it heavily penalizes predictions that deviate significantly from actual values. Given our interest in accurate predictions for extreme events, this is particularly relevant. We also present the RMSE, which is the square root of MSE. Unlike the MSE, which is presented in squared units, RMSE is in the same units as the output variable, making it more directly interpretable by showing the average standard deviation between the model predictions and actual values.

Additionally, we provide results for the MSE metric applied solely to data points above the 90th, 95th, and 99th percentile of the concerned dataset (the test set here). This is essential since the most intense events are likely to cause satellite damage. Given SpaceAble's intended industrial application, we assess the model's predictive capability on the highest data values. All results and visualizations will be discussed in the next section.

	MAE	RMSE	MSE	MSE 90th	MSE 95th	MSE 99th	Time
Model 1	0.467	0.665	0.442	0.816	1.059	2.117	45.10 min
Model 2	0.480	0.676	0.457	0.815	1.082	2.188	37.07 min
Model 3	0.491	0.702	0.493	0.626	0.731	1.358	43.58 min
Model 4	0.506	0.718	0.515	0.531	0.632	1.225	37.01 min

Table 4.2 – Values of the selected metrics on the test dataset, which the algorithm has not previously encountered. "MSE Xth" refers to the MSE value for data outside the Xth percentile. "Time" indicates the training duration. It is for the hardware from a single MSI computer (GPU: nvidia Quadro RTX; CPU: Intel i7-10875H @3GHz).

Table 4.3 displays results for the validation dataset from [McGranaghan et al. \(2021\)](#), specifically the data from the F16 satellite for the year 2010. However, it's essential to note that 34% of the data in this dataset have not been encountered by our algorithm before, while 66% were previously part of either the training or validation datasets. This skews our results in favor of our algorithm. Importantly, as this is PrecipNet's validation dataset, the results are inherently biased. A separate test dataset should have been set aside during PrecipNet's training to ensure an unbiased basis. We'll further discuss these results in the following section.

	MAE	RMSE	MSE
Model 1	0.533	0.749	0.560
Model 2	0.554	0.772	0.596
Model 3	0.551	0.774	0.597
Model 4	0.585	0.802	0.644
PrecipNet	0.558	0.764	0.584
OVATION	1.574	1.887	3.560

Table 4.3 – MAE, RMSE and MSE for satellite F16 and year 2010 data

4.3.4 Comparative Analysis Insights

Most of the information that guided our architectural choices has been discussed. However, let's review some decisions here and their impact on performance.

- Numerous tests were carried out with L1 and L2 regularizations. No significant findings were observed.
- Similarly, various schedulers were used, but none drastically improved the results.
- Both SGD and Adam optimizers were tested. Although Adam can be more complex, it almost always converged faster, reducing the number of necessary epochs (and therefore the run time).
- Using an autoencoder often improved results. The bottleneck architecture helped reduce noise in the data. However, the best results were achieved with larger architectures and a high number of parameters, which we aimed to avoid due to performance considerations and computational capacity.
- Numerous changes were made, especially concerning the learning rate and batch size. A very small batch size makes each epoch tedious, whereas a too large batch doesn't add much and slows the algorithm's convergence. A slightly higher learning rate than 0.001, combined with an ExponentialLR scheduler, proved effective but sometimes pushed the machine to its limits. The final choices here remain straightforward in both implementation and understanding.

The training and validation curves are typical, indicating stable and progressive learning. The results on the metrics provide an average performance for each model.

- Observing the training and validation curves for model 1 shows a growing gap between the decreasing training curve and the stabilizing validation curve. As long as the latter does not rise, we are not overfitting. This gap might be due to the peculiarities of the validation dataset, or perhaps overfitting would only become evident with prolonged training. Model 1 is also our best-performing model, with an average deviation of 0.467 in $\log_{10}(\text{Electron Total Energy Flux} \times \pi)$, representing an average deviation of approximately $5.09 \times 10^{10} \text{ eV.cm}^{-2}.\text{s}^{-1}.\text{ster}^{-1}$.
- For model 2, training could have been slightly extended. The learning curve is good, but the phenomenon visible in model 1 seems to be emerging gradually. However, initial learning is much faster. The autoencoder architecture, intended to best reduce noise in the data, doesn't seem to have identified trends not already noticed by model 1.
- Models 3 and 4 exhibit similar curves and trends to models 1 and 2, respectively. As expected, they perform much better on higher values (see table 4.2), since the cost function was designed with this purpose. The training curves seen in Figure 4.12 end around 0.5 because the chosen cost function penalizes the result, but the selected metrics accurately reflect performance (table 4.2). Model 3 underperforms model 1 by about 5%, and model 4 underperforms by about 8%. However, model 4 (the best performer on high values) outperforms model 1 by about 53% for data above the 90th percentile and by about 73% for data above the 99th percentile. Model 4 is also the quickest to train.
- The density curve displays show that all four models are effective and align well with the $y=x$ curve. Additionally, models 1 and 3, which have the simplest architecture, struggle to predict values below 8. This trend is also confirmed by the histograms.

Thus, models 1 and 4 appear to be the most effective, each relevant depending on the specific goal. To conclude our analysis, let's compare the results of these two models with the state of the art. We randomly selected 66,882 dates from the test set for which we computed predictions

using OVATION and our models 1 and 4. Below, in Figures 4.15, 4.16, and table 4.4, histograms, densities, and metrics for these predictions are shown. We did not compute OVATION values for the entire test set due to internal computational power constraints.

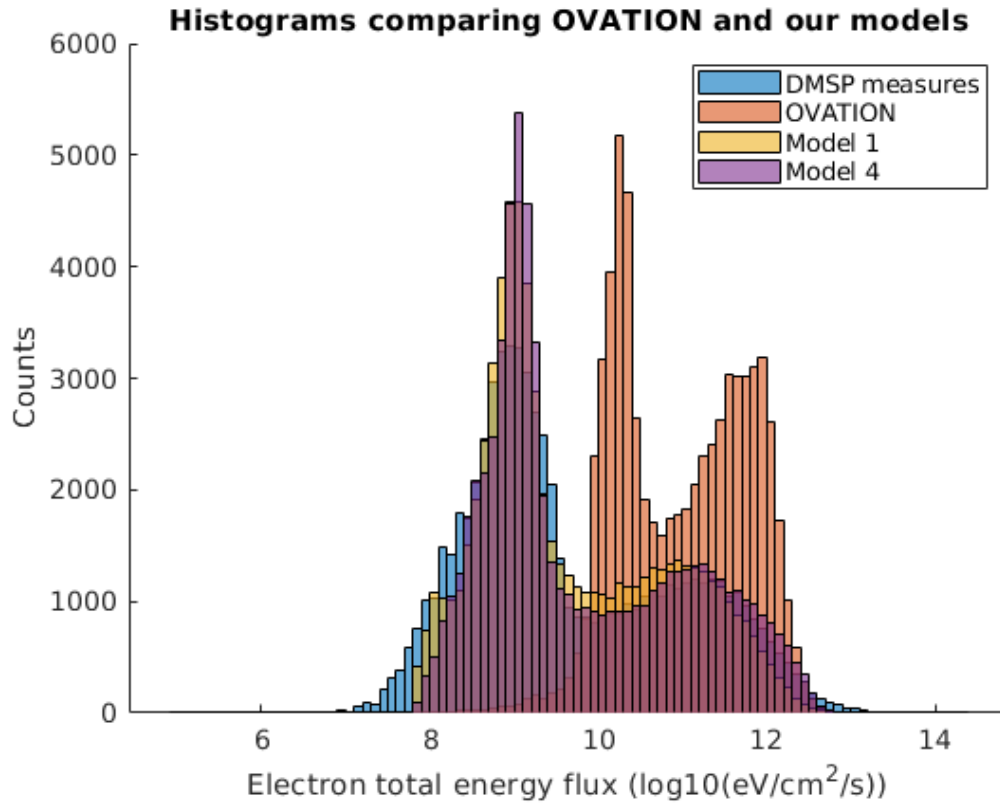


Figure 4.15 – Histograms of OVATION, models 1 and 4 predictions over 66,882 samples taken randomly from the test set.

	MAE	RMSE	MSE	MSE 90th	MSE 95th	MSE 99th
Model 1	0.465	0.662	0.439	0.807	1.045	2.122
Model 4	0.504	0.715	0.511	0.524	0.623	1.256
OVATION	1.399	1.693	2.867	0.327	0.458	1.098

Table 4.4 – Comparison of metrics results between models 1, 4 and OVATION over 66,882 samples taken randomly from the test set.

We also plotted the forecasts for a representative period: that of November 14, 2010, around 19:30. This date allows us to this time compare our results directly to those of PrecipNet presented in the 2021 paper, see Figure 4.17. We limited ourselves to models 1 and 4 for this display.

After comparing with OVATION and PrecipNet in Figures 4.15, 4.16, 4.17 and tables 4.3 and 4.4, we observe that:

- Model 1 slightly outperforms PrecipNet as seen in table 4.3 and the curves in Figure 4.17.
- Model 4 outperforms PrecipNet for high values, thus during violent events (visible in 4.17).
- Our models 1 and 4 outperform OVATION in terms of the MAE, RMSE, and MSE metrics.

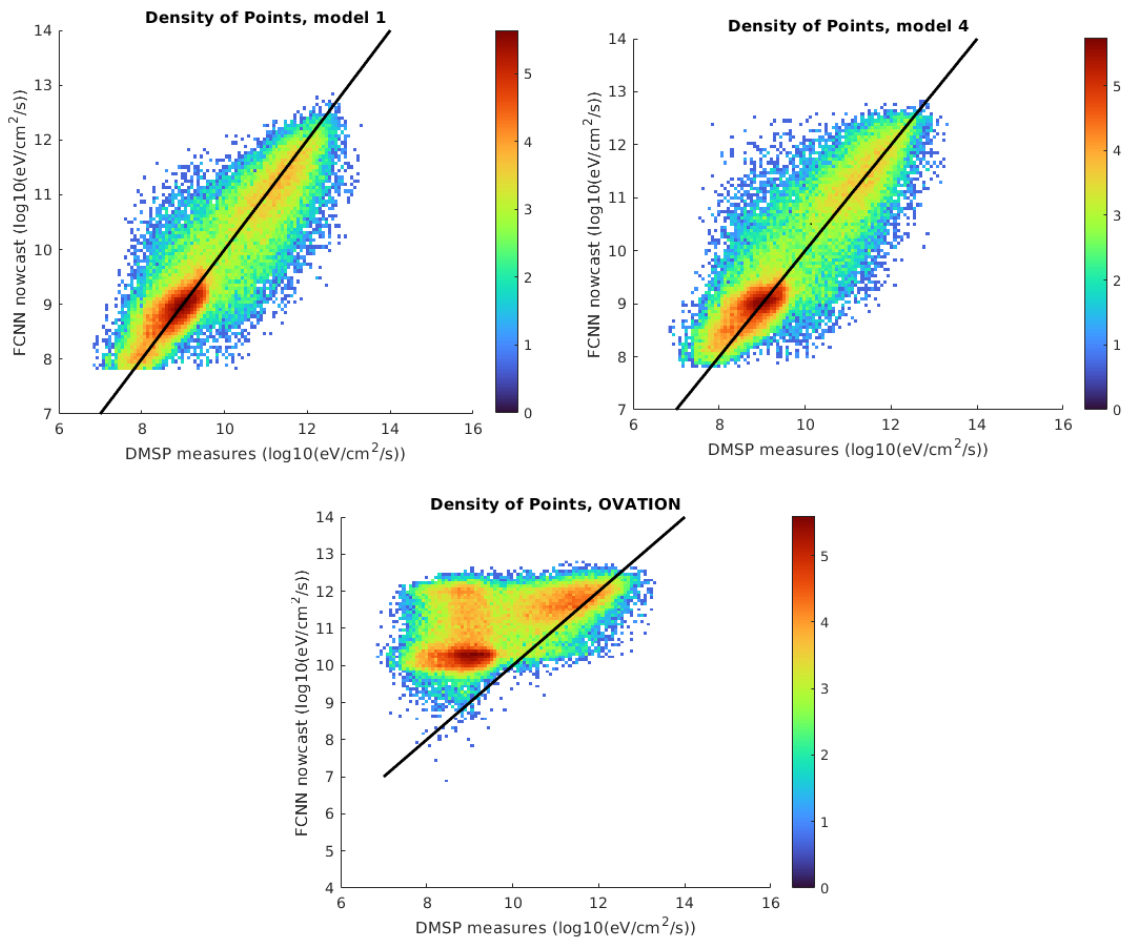


Figure 4.16 – Density estimates between DMSP real measurements and OVATION, models 1 and 4 predictions over 66,882 samples taken randomly from the test set. The density is displayed in logarithmic scale for visualization purposes, and we added 1 to the results, making any gaps appear white. Thus, no point has a value between 0 and 1. The line $x = y$ appears in black.

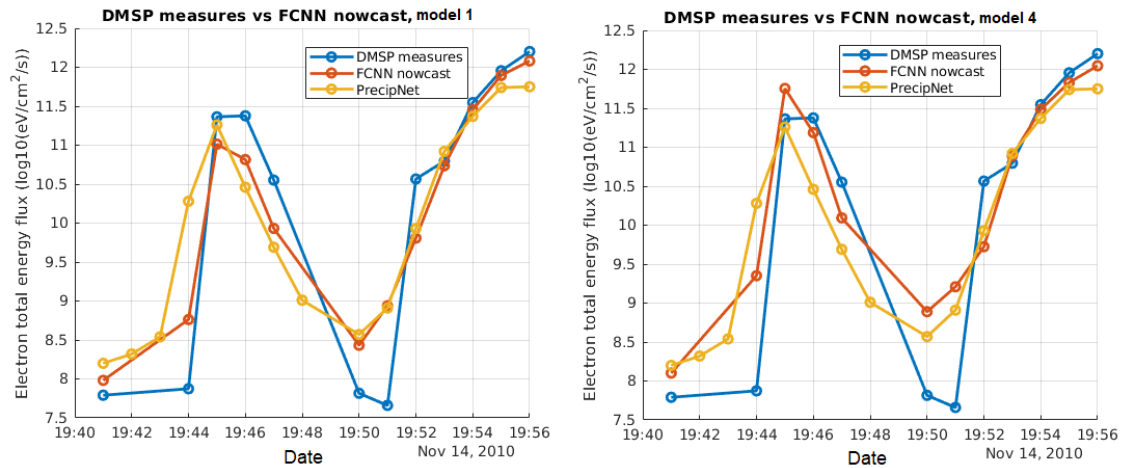


Figure 4.17 – Display of the results obtained by PrecipNet sourced from the article [McGranaghan et al. \(2021\)](#) compared with our two most relevant models: our model 1 for its simplicity and our model 4 for its results on extreme values. Note: some PrecipNet data that we do not have appear here. They may correspond to missing points obtained in PrecipNet by interpolation, which we therefore do not have. Some minor differences are observed between our datasets, the cause of which we have not identified.

Model 1 improves OVATION’s results by 85% for MSE (and respectively 20% for MAE and 61% for RMSE).

- OVATION consistently overestimates the actual values, as can be seen in the histogram in Figure 4.15 and in the density estimates in Figure 4.16. However, this makes it very effective for rare cases and extreme events, as shown in table 4.4. In the most extreme cases (above the 99th percentile), it outperforms our model 4 by about 14%. However, it is still far less effective than models 1, 4, or even PrecipNet.
- If we consider that all metrics are equally important to us, our model 4 is the most balanced, slightly underperforming OVATION on extreme values and slightly underperforming model 1 on standard metrics.
- We do not strictly use the same inputs as PrecipNet. Indeed, we omitted $F_{10.7}$ and the Polar Cap Index. Adding them does not significantly improve our results (difference in results is equivalent to changing the seed value).

The goal of this study is primarily to establish a methodology, code, and reusable, adaptable, and easy-to-interpret product. Although our results far surpass the community’s state-of-the-art (OVATION Prime), they only slightly surpass the very recent results of PrecipNet. However, they differ significantly in their approach, which is more industrial. The code, prepared using PyTorch-Lightning, can be directly implemented in an industrial setting. It is based on an organizational method that the company SpaceAble can quickly integrate. Internally, future work will only aim to improve results through more thorough benchmarking or innovative ideas.

One example to be explored in future studies is the possibility of blending the algorithms used. Indeed, model 1 can focus on average values and model 4 on extreme values. In this way, and with a wisely chosen threshold, it would be possible to use both algorithms and switch between the two depending on the predicted value. For example, if the value predicted by model 1 exceeds 11, model 4 will likely be closer to the actual value, where the latter would be less accurate for values between 8 and 10. To go further, the switch could be replaced by a progressive, continuous mix.

Finally, some issues have been intentionally left out, such as refining forecasts for low values (between 7 and 8). The idea of an adapted cost function to push the algorithm to its limits when forecasting these values is entirely possible. However, since these values represent non-hazardous cases for space and ground systems, it was not a priority to address them.

Two questions remain unresolved, to which we wanted to provide a preliminary answer. With these architectures and tools, would short-term forecasting be possible? We had mentioned that one of the challenges related to FCNN was the arbitrary choice of past data; would it be possible to overcome this difficulty using a TCN? We address these two questions in Sections 4.4 and 4.5 respectively.

4.4 Forecast

Now that a primary study has been conducted on applying neural networks to our problem, the idea was to remove the data at T0 and thus reduce the number of inputs to 62. Therefore, the modeling capacity of our network becomes a predictive capability at 10 minutes. Indeed, the first historical data used is the one at T0-10mn. By setting T0-10mn as T0, we find ourselves in a situation where we predict the energy fluxes at T0+10mn. Position and time data (cos_ut, sin_ut, sin_doy, cos_doy) can be set at T0+10mn (as they are not measurements) just like latitude and longitude data. To provide a first glimpse (a "proof of concept"), we retrained model 1 following this idea. Here are the results:

	MAE	RMSE	MSE	MSE 90th	MSE 95th	MSE 99th	Time
Model 1	0.466	0.667	0.444	0.802	1.045	2.098	33.90 m

Table 4.5 – Results when training the architecture of model 1 with only historical values of the inputs B_x , B_y , B_z in GSE coordinates, flow speed, proton density, and pressure (in logarithm) and AL, AU, and SYM-H. Time is for the hardware from a single MSI computer (GPU: nvidia Quadro RTX; CPU: Intel i7-10875H @3GHz).

The histogram figures, density (figure 4.18), and learning curves remain approximately the same, as can be seen below. As we have already mentioned, a large part of the prediction lies in the date and observed position (latitude and magnetic longitude), but we still note that our 10-minute forecast is relevant and justifies future studies for potential improvement of this predictive capability. Moreover, unlike the results of model 1 visible in Figure 4.13, it seems that the modeling was better for low values, below 8 in $\log_{10}(\text{eV}/\text{cm}^2/\text{s})$.

However, we are still limited to the chosen historical timestamps. Ideally, we know that the progression of solar wind and the delay between what is observed at the bow shock and what arrives on Earth is dynamic. It's too crude to assume that all observations at the poles come from events at the bow shock level that occurred a fixed time ΔT before. To circumvent this problem, several solutions are available, such as taking dynamic inputs based on certain values. One approach would be to prepare a primary model capable of providing the approximate delay between input and output (possibly based on rare and easily identifiable events on both sides) and then build a dataset using this delay. The models thus created would therefore provide a value output as well as a delay ΔT between the time of the input data and that of the output data.

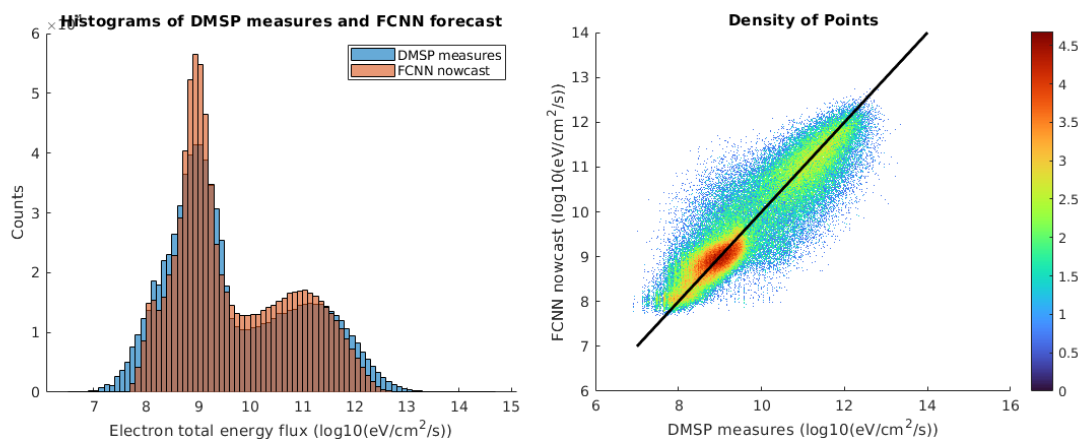


Figure 4.18 – Display of the histogram and density estimates obtained by forecast using model 1’s architecture. The density is displayed in logarithmic form for visualization purposes, and we added 1 to the result, making any gaps appear white. Thus, no point has a value between 0 and 1. The line $x = y$ appears in black.

A second, more interesting solution in our opinion, is to turn to other categories of networks that can take into account all past data. This is the case for some memory networks like LSTMs. But the most relevant seemed to us to be the temporal convolutional network. As a reminder, key features of the TCNs include causal and dilated convolutions. Causal convolutions ensures that the model’s prediction at time t is only dependent on information from time t or earlier, maintaining the temporal order. Dilated Convolutions allows the network to have a larger receptive field, meaning it can take into account information from further back in the past without increasing the number of parameters significantly. Its advantages over RNNs and LSTMs can be summarized through three main known points:

- **Parallelization:** Unlike RNNs, TCNs process all data points simultaneously, allowing for faster training times.
- **Stable Gradients:** TCNs avoid the vanishing and exploding gradient problems often encountered in RNNs.
- **Flexible Receptive Field Size:** The use of dilated convolutions allows for adjustable receptive fields to capture relevant temporal context.

In the project’s early stages, we faced uncertainties about how far back we wanted to retrieve data and were keen to avoid being restricted by our choice of architecture. While LSTMs generally outperform RNNs, their effectiveness diminishes as historical sequences lengthen, leading to increased complexity and reduced performance. Moreover, the computation time is higher for LSTMs. Despite the potential advantages of exploring various architectures with more time, we initially focused on TCNs.

In 2020, we participated in a machine learning challenge online, hosted by NOAA (MagNet Challenge²). We ended 28th over 84 participants. In this competition, one of the competing models utilized an LSTM with 4 million parameters, a substantial contrast to our own model, which had only 50k parameters. This significant difference in parameter count make the LSTM impractical to store the required information in our case. While we considered employing a transformer architecture, previous research in image processing has indicated that transformers tend to perform optimally with a large number of parameters, surpassing what we had available. Additionally, given the novel nature of transformers and the constraints of our thesis project, which had

2. <https://ngdc.noaa.gov/geomag/mag-net-challenge.html>

already incorporated a fully connected (FC) network, integrating a transformer model seemed too ambitious and potentially disruptive to the project's scope and timeline.

4.5 Addressing Limitations with the TCN

Like analyzing all the patterns on an image to identify its contents (the CNNs), a TCN will analyze the entirety of a past time range to identify what occurred to then provide an output value. And thus, just as a conventional convolutional network can identify an object regardless of its location in an image, we hope that the TCN will dynamically look at the past time range to make the connection between observed trends and output values.

4.5.1 Training Process

The dataset used here is the one containing only DMSP values that we find in PrecipNet's training. This choice stems from a desire to more easily compare our results with what was done by PrecipNet since the run time for a TCN is much longer than for training an FCNN (as we will see), but also for timing reasons, the bulk of the study having been already conducted on this subset and the creation of an initial valid index list (reexplained below) being time-consuming.

The dataset we create to run the TCN is named dataset 1, which we detailed in Section 3.4.2 and Figure 3.15. From DMSP data and OMNIWeb data, we retrieve a list of indices considered "valid", i.e., for which the n previous data do not contain any NaNs (Not a Number, the code equivalent of missing data). The value of n corresponds to the size in minutes of the historical range we take as input.

In this study, we chose only the 30 minutes preceding T_0 . Firstly because the longer the time range, the higher the probability of finding a missing data point (making the sample unusable). Next, because the longer the time range, the longer the training duration. Lastly, because we believe that 30 minutes is a sufficiently long delay between the bow shock's nose and Earth to account for most phenomena, particularly the most violent ones.

The index list thus indicates the line of the OMNI dataset corresponding to the date T_0 (and for which the last 30 values do not contain missing data) and the line of the DMSP dataset that matches. In this way, the TCN will pick from both datasets without us having to preconstruct a dataset 30 times larger than planned (30 values per input parameter). Thus, a TCN input sample corresponds to 17 vectors of size 30. The OMNIWeb parameters, i.e., the components of the IMF, the speed, density, and pressure of the solar wind, as well as AL, AU, and SYM-H, are therefore taken from T_0 to T_0-30mn . Date and time parameters, as well as DMSP's position parameters, however, remain the same and do not vary over the last 30 minutes (we are looking at a single place and time). To be taken as such by the TCN, we multiply them by the unit vector of size 30, thus obtaining a repetition of the same value 30 times. This is equivalent to providing a single value to the TCN.

Thus, after preparing the index list, we obtained 478,669 valid samples. To maximize the number of samples sent for training, this time we kept 20% of the data (or 95,734) for the test set and 80% for validation and training. Let's note something important: we can arbitrarily increase our dataset size by using the data whose gaps have been interpolated using the PCHIP algorithm,

as explained in Section 3.4, but we did not have time to explore this avenue, which we will return to in the prospects of the next chapter.

Before starting our benchmark, i.e., the set of standardized tests and measures used to evaluate and compare the performance of algorithms between them, we set the seed value to 3, arbitrarily. The seed value, in computer science, is an initial value used to "control" the generation of random numbers and thus make it reproducible. By always providing the same value to the seed, the random number generation remains the same for all training sessions. As the random distribution of data between training and validation sets is random, we ensure by fixing the seed that the same samples are consistently sent to the same data sets. Runs are therefore comparable between them.

It seems important to us to start our benchmark by varying the seed multiple times. Indeed, if the results remain nearly the same despite varying the seed, it indicates that the model's performance is stable concerning the distribution of data between the training and validation sets. In this case, the choice of seed has little impact on the results, which is desirable to have a reliable performance evaluation of the model. So, we first confirmed that for several similar trainings and only varying the seed, there was a negligible difference in the results (less than 0.1% in our case).

After conducting an extensive benchmark (several hundred trials), we settled on a final model, in two versions, corresponding—as previously—to a change in the loss function (MSE for Model 1 and Tail Weighted Loss for Model 2). The parameters of the model used are as follows:

Aspect	Model 1	Model 2
Channels	[30 30 30 30]	[30 30 30 30]
Number of Parameters	54,991	54,991
Epochs	50	50
Batch Size	500	500
Learning Rate	0.1	0.1
Loss Function	MSE	Tail Weighted Loss
Optimizer	Adam	Adam
Lambda 1 & 2	None	None
Scheduler	Exponential LR 0.9	Exponential LR 0.9
Dropout	None	None
Dilation Factor	2	2
Kernel	8	8
Training Time	12h 6mn 3sec	12h 10mn 41sec

Table 4.6 – Hyperparameters of the two versions of our Temporal Convolutional Network

Some notes regarding these hyperparameters:

- After several tests, both L1 and L2 regularization (as shown in table 4.6) did not make a significant difference, so we left them out.
- The hyperparameters "channels", "dilation factor", and "kernel" are specific to the TCN, which we discussed in Section 2.3.6.3.
- This time, we used a scheduler, specifically the exponential LR set at 0.9. This accounts for the high "learning rate", which represents the starting value.

- The smaller number of epochs is due both to the training curve stabilizing quickly and the lengthy training duration (over 12 hours for 50 epochs).
- Dropout did not improve the results in our tests, so we removed it.

4.5.2 Visualizing Results

In this section, we display a set of curves that represent both the training process and the results. Figure 4.19 shows the training and validation curves for Model 1. Model 2 exhibited a similar trend but with a higher loss value by default. As we can observe, the curve has almost stabilized and shows no signs of overfitting. We could have allowed it to train for a few more epochs, but numerous tests conducted so far have shown us that the marginal gains were equivalent to a change in seed, which remains negligible.

Figure 4.20 illustrates the evolution of our learning rate, which follows an exponential decay and has a value of $0.1 \times (0.9)^n$ at epoch n . It is evident that when the learning rate becomes very low, the training stabilizes, and the noise decreases in the validation curve (gray) as seen in Figure 4.19.

Figures 4.21 and 4.22 showcase results similar to those we previously observed in the FCNN section. These depict histograms of the results compared to the actual values and an estimation of the point density by plotting our results against the actual values.

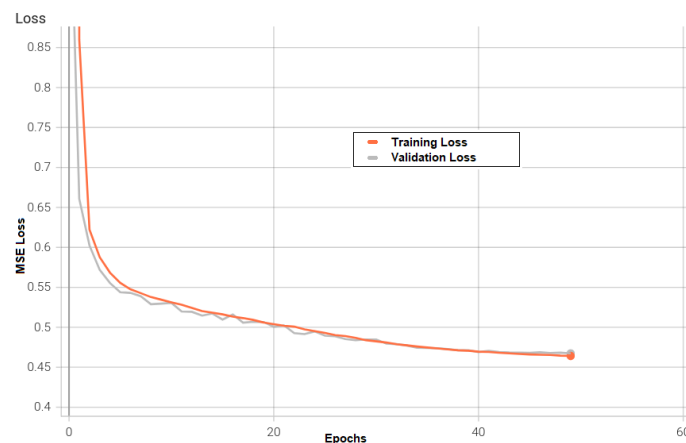


Figure 4.19 – Training and validation loss curves over 50 epochs.

4.5.3 Performance Metrics

The performance of this new model can be seen in Table 4.7.

	MAE	RMSE	MSE	MSE 90th	MSE 95th	MSE 99th	Time
Model 1	0.492	0.684	0.468	1.015	1.305	2.433	12h 6min 3sec
Model 2	0.507	0.712	0.507	0.801	0.966	1.638	12h 10min 41sec

Table 4.7 – Results for our metrics on versions 1 and 2 of the TCN

There isn't much new to comment on regarding these results, except for the training time, which increased more than 19-fold with this new method. It will thus be challenging to regularly update this algorithm with new data.

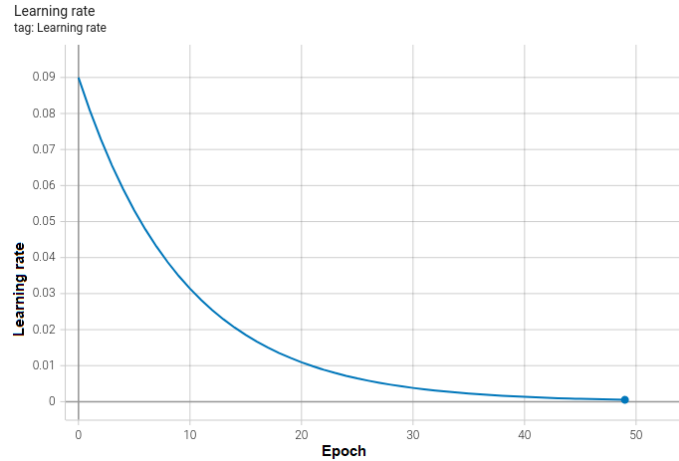


Figure 4.20 – Evolution curves of the learning rate, following the exponential LR set at 0.9 (the learning rate is multiplied by 0.9 at each epoch).

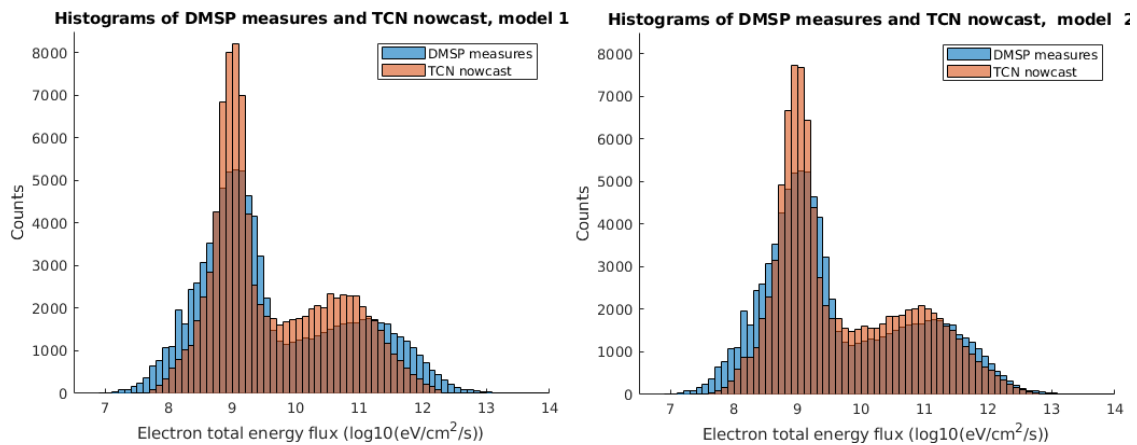


Figure 4.21 – Histograms of Model 1 and 2 results compared to the histogram of actual DMSP values on the test set.

However, the metrics from the other results closely resemble previous results and remain superior to those from OVATION. The approach we’ve taken with this seems more promising than with a traditional FCNN (which we’ll discuss in Section 4.6). Yet, in the case of the MSE, for instance, the algorithm underperforms by 5.8%. We still see a marked improvement for rare and extreme cases when using the Tail Weighted Loss. From this point on, we’ve developed the **final product** of our study based on this TCN.

4.5.4 Insights on the Final Product & Comparative Analysis

Before presenting a comparative analysis between the state of the art and our TCN, we’ll first introduce our product and then use it to compare our forecasts with those of OVATION, which remains the community’s gold standard today. The reasons for selecting the TCN for our product are summarized in Section 4.6.

The final product can be used in two distinct ways and is applied once the model is trained and finalized. The first method to use the code is by inputting a starting date and a number of points to plot (along with a few other inputs, as shown in Figure 4.23). This will plot the DMSP satellites’ paths as well as the values predicted by the code. Let’s delve into the visible inputs in 4.23 and

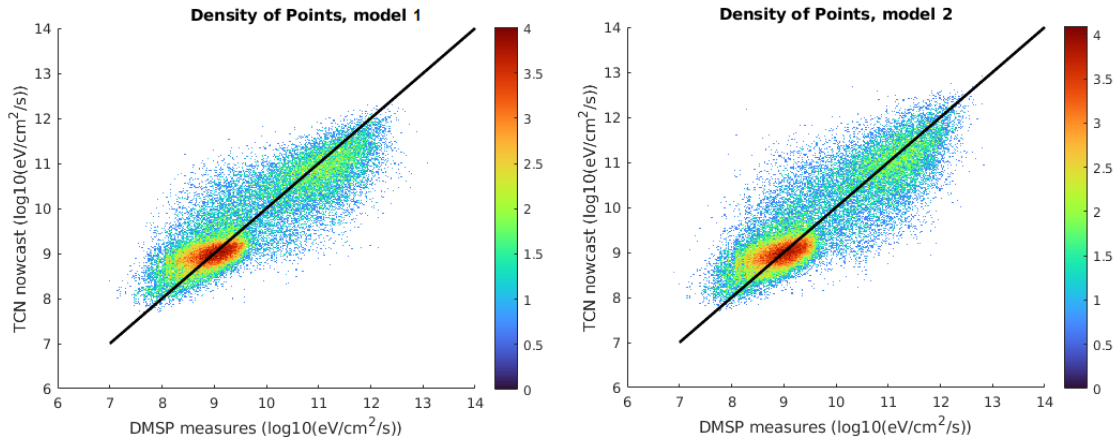


Figure 4.22 – Density of points when plotting predicted values from the TCN models against actual values on the test set. The density is displayed in logarithm for visualization purposes, and we added 1 to the result and made voids appear in white. Thus, no point has a value between 0 and 1. The line $x = y$ is shown in black.

how they work.

- **path_to_project** indicates the path to the project, structured as we've outlined.
- **start_datetime** refers to the date in the format 'yyyy-mm-dd HH:MM:SS' from which to start plotting the energy flux values.
- **version_model** indicates which model to use, in our case either 1 or 2.
- **unit** specifies the unit in which to plot the curves. It can be in $\log_{10}(\text{eV}/\text{cm}^2/\text{s})$, in $\text{erg}/\text{cm}^2/\text{s}$, or in $\text{eV}/\text{cm}^2/\text{ster}/\text{s}$.
- **interpolation_gap_size** determines whether to use or not the interpolated OMNI data for plotting, which allows more data points to be plotted, even if the algorithm hasn't trained with them.
- **amount_data_plot** indicates the number of data points to plot from the given date.
- **plot_OVATION** is a boolean to decide whether or not to plot OVATION results for the same period.

Using these inputs, the algorithm fetches the corresponding OMNI data online for the provided date. If there's no data available for that date, it automatically retrieves the nearest available date. It then plots the "amount_data_plot" data from this date. Data is taken every minute, and if there's missing OMNI data to get a result, the code moves to the next point, ensuring the total number of plotted points is exactly "amount_data_plot". The code also checks which satellites measured data during this period and plots the TCN results for each of these satellites (recall, the spacecraft's identity is an input for our model). For example, considering March 18, 2010, at 02:00 (a period of high activity), as seen in Figure 4.24, we indeed get 15 points corresponding to the F16 satellite's pass, minute by minute. We've chosen interpolation here, where gaps of size 4 or less have been filled using *PCHIP*, providing us with 15 points from 02:00 to 02:14. Without interpolation, we only have a point at 01:59 followed by points from 04:26 to 04:39, primarily for the F18 satellite. The code also displays a top-down view of the pole, showing the satellite's path, with point colors matching their values. For this specific example, we also obtain points for the F17 and F18 satellites, which we aren't displaying here.

This first feature of the product allows us to swiftly plot the paths of DMSAP satellites during crucial moments, while consistently comparing them to OVATION. A notable advantage is that we don't rely on the existence of data points measured by DMSAP. For comparison purposes, we've


```

from ResultsUtils import *
import seaborn as sns
sns.set()

#-----
# Inputs
#-----
path_to_project = '/home/simon/These/Algorithms/tcn-swe-forecast'
start_datetime = '2010-03-18 02:00:00'
version_model = '2'
unit = "log10(eV/cm2/s)" # or "erg/cm2/s" or "eV/cm2/ster/s"
interpolation_gap_size = '4'
amount_data_plot = 15
#-----

result = create_TCN_previsions( path_to_project = path_to_project,
                               start_datetime = start_datetime,
                               version_model = version_model,
                               unit = unit,
                               interpolation_gap_size = interpolation_gap_size,
                               amount_data_plot = amount_data_plot)

plot_TCN_outputs(result)

```

Figure 4.23 – Overview of the code used to display what is visible in Figures 4.24 and 4.25.

displayed in Figure 4.24 the actual DMSP data alongside. Yet, only the OMNI data, coupled with a point in space and time, are necessary for plotting these curves. Hypothetically, we could predict what the F08 satellite would measure in 2012 at a particular latitude and longitude. This brings us to the second and final utility of our product: displaying auroral maps derived from our TCN for a specific date.

Just as we can retrieve OMNI values for a particular date and time, resulting in a value for electron energy flux at a specific magnetic latitude and longitude, it's entirely possible to generate a display of all flux values for every position at a specified date and time. Therefore, Figure 4.26 offers two examples of what our product yields. By providing a date, time, model version to use, and the identity of the DMSP satellite to "emulate", the result is an aurora map. It's also possible to specify the desired precision in magnetic local time and magnetic latitude. Additionally, producing such a map for a particular date takes roughly 2 minutes. In these examples and generally, we observe an irregularity near 15:00 MLT. We believe this stems from the lack of data in this region, making predictions challenging. Indeed, the DMSP data density (found in the appendix of our paper 3.3.4) reveals a gap in magnetic local times between 12:00 and 15:00 for lower latitudes (around 50°).

Before summarizing all these results, we wanted to highlight a more recent aspect of our research (and AI research in general) which will be the focus of a subsequent study: integrated gradients.

4.5.5 A Step Further: Integrated Gradients

Interpretability techniques are vital for a better understanding and interpretation of Machine Learning algorithms. A significant research push is happening in this domain (Carvalho et al., 2019). Some algorithms, like linear or logistic regression, decision trees, k-nearest neighbors, generalized additive models, and Bayesian models, are inherently explainable and transparent. They offer clear insights into their decision-making rationale and are easily interpreted by humans. However, others like random forests, support vector machines (SVMs), and deep neural networks require an explainability method to make their decisions understandable to humans. In applications like medical, political, or legal decision-making, understanding how a decision is made is vital, and a lack of transparency can be a severe setback. Consequently, methods have

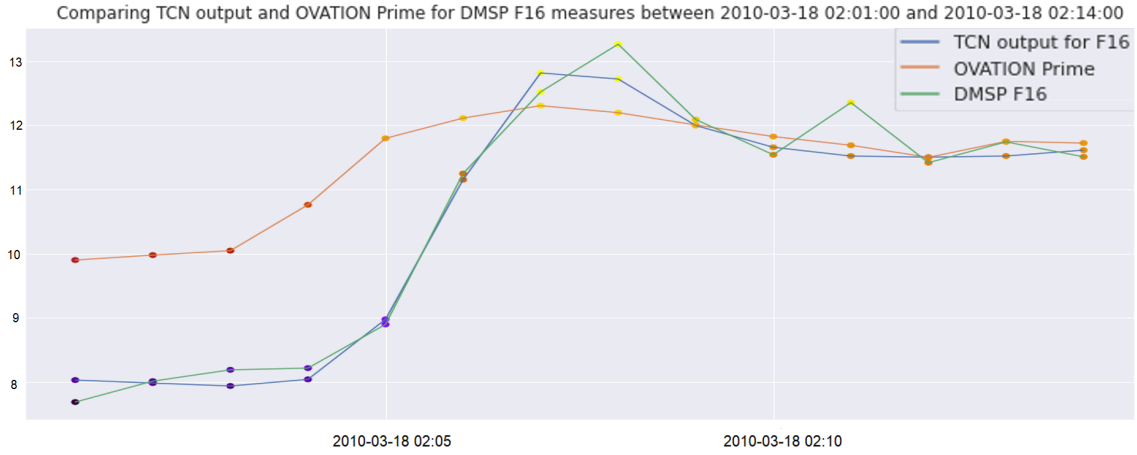


Figure 4.24 – Output of the code plotting a certain number of points "amount_data_plot" (here 14) after a given date "start_datetime" (here March 18, 2010, at 02:01) as output from the trained TCN. Also displayed here are the true DMSP values as well as OVATION's forecasts. The color code of the points comes from Figure 4.25.

been developed to improve transparency, such as generating explanations for decisions, visualizing internal processes, or using simpler models to approximate their behavior.

In this scenario, Integrated Gradient (IG) is primarily a method used with deep learning neural networks, including convolutional and recurrent ones. It pinpoints which input features contribute the most to a prediction. A significant advantage is that it doesn't modify the original network. It was initially introduced in [Sundararajan et al. \(2017\)](#). The fundamental idea behind it is to compute the model's output gradient relative to its input features, integrating the gradient along a path from a baseline (or reference) input to the actual input. This integration method offers a measure of feature importance by looking at how each feature contributes to the model's output change.

$$\text{IntegratedGrads}_i(x) ::= (x_i - x'_i) \times \int_{\alpha=0}^1 \frac{|\partial F(x' + \alpha \cdot (x - x'))|}{\partial x_i} d\alpha$$

Where:

- **IntegratedGrads_i(x)** is the integrated gradient for the feature i of the input x .
- x_i is the value of feature i in the actual input x .
- x'_i is the value of feature i in the baseline input x' .
- F is the model's output (prediction) function.
- $\frac{\partial F(x' + \alpha \cdot (x - x'))}{\partial x_i}$ is the partial derivative of the model's output with respect to feature i at the point $x' + \alpha \cdot (x - x')$ along the path from the baseline x' to the actual input x .
- The integral is taken over a linear path from the baseline x' to the actual input x with α varying from 0 to 1.

The advantage that IG has over other existing methods is that it satisfies both the axioms of sensitivity and implementation invariance (which is not the case for other methods such as *DeConvNets* or *Guided back-propagation*):

- **Sensitivity:** This axiom states that if a feature is not influential in the prediction of a model, then the attribution assigned to that feature should be zero. In other words, if a feature has no impact on the model's output, it should not receive any attribution.

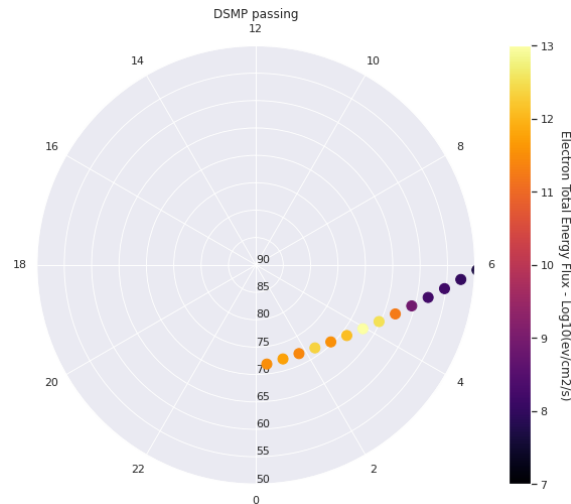


Figure 4.25 – Same data as in Figure 4.24 displayed with a color code on a polar graph representing the poles. The passage points represent 15 values for the F16 satellite between 02:01 and 02:14 on March 18, 2010.

- **Implementation Invariance:** An attribution method such as IGs adheres to the implementation invariance axiom when the attributions remain consistent for two functionally equivalent networks. Two networks are functionally equivalent if their outputs are the same for any input data, even if their architectures are different.

Towards the end of our work, we were able to implement this method and get a glimpse of its results and ensuing possibilities, as shown in Figure 4.27. These results are intriguing because they reflect the "importance" that a given input has on the output *relative to a baseline*. In our case, we arbitrarily chose to send zero vectors as the baseline, which is debatable. The results we obtained were also small and were rescaled for display.

It's crucial to remember that the scale of values alone doesn't necessarily signify if the attributions are meaningful. What's more pivotal is how these attributions contribute to understanding the model's behavior and whether they align with our domain knowledge or the problem's expectations. They're often employed to explain the model's predictions by attributing the importance of each feature to the final prediction. In this context, we decided to scale them between 0 and 1, with 1 being the max feature contribution.

In the attributions (our results after applying integrated gradients), we obtained both positive and negative values. Negative values denote a negative correlation between input and output and are equally significant as positive values. In the display Figure 4.27 we do not show the entries for which the past is not investigated (like the satellite's position) and we normalized by dividing by the absolute maximum value. We then took the absolute value of the attributions. Hence, we get a value between 0 and 1, with a strong correlation near 1. Several tests were conducted to see if displaying on a logarithmic scale would be more insightful, but that wasn't the case. Three results are visible in Figure 4.27 for three randomly selected samples: May 23, 2011, at 05:38; September 20, 2004, at 07:08; and March 18, 2010, at 01:59.

As we can observe, it seems for May 23, 2011, at 05:38 that features like AL and AU had more significance. Moreover, the core information required for modeling seemed to be around 05:10 to 05:14, about 28 to 24 minutes prior. For September 20, 2004, at 07:08, we notice a high correlation with the values at 06:40, precisely 28 minutes earlier. Furthermore, it seems that B_x and B_y played prominent roles in this modeling. Finally, for March 18, 2010, at 01:59, it's challenging to

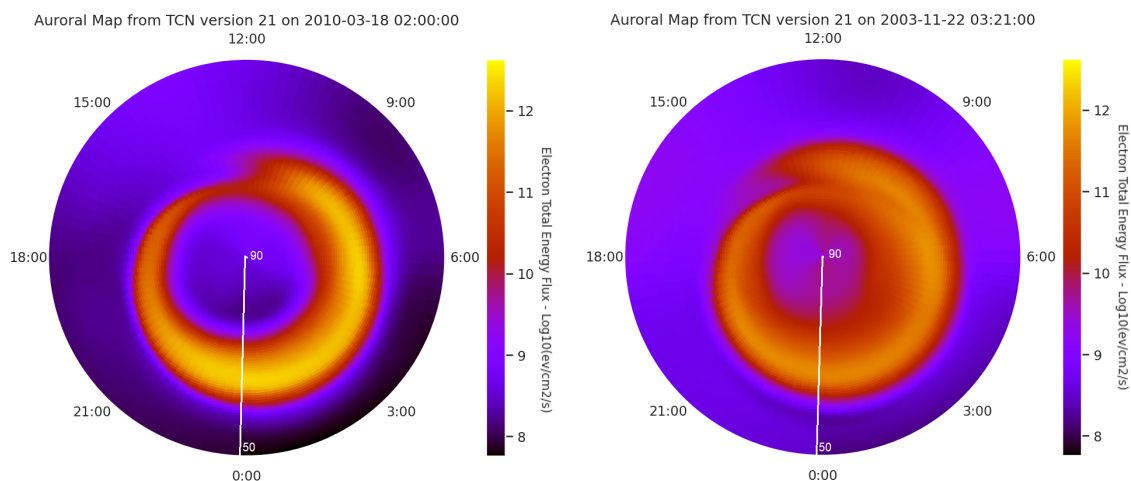


Figure 4.26 – Display of the final product output, based on the TCN model, for two arbitrary dates (March 18, 2010, at 02:00 and November 22, 2003, at 03:21) showing a tumultuous period (left) and a calmer one (right). The precision in magnetic local time is 0.25h, and in magnetic latitude, it's 0.5°.

distinguish a preferred hour, except for 01:59 itself. However, we can see that AL played a more significant role than the others among the OMNI data.

For now, it's hard to draw many more conclusions, as this research is still in its early stages. However, one can already see its potential for all AI applications, especially in similar contexts where modeling (or prediction) is done with data measured at various locations (and thus different times).

4.6 Final Remarks

The goal behind this study wasn't merely to achieve the best possible result on our metrics. It was primarily about laying the groundwork for AI research conducted within SpaceAble in this field. Remember, SpaceAble's objective is to provide its clients with expertise on the risks their assets face, be they on the ground or in space, in both the short and long term. Hence, SpaceAble is dedicated to developing its AI branch tailored to these challenges. Thus, this thesis was mostly crafted to be self-contained, yielding adaptable algorithms. This thesis presents itself as an encyclopedia, and from the coding perspective, as a flexible product from which it's easy to move forward.

In Chapter 4, we detailed our code's organization, presented the used architectures, the achieved results, and compared them with the most advanced code on the topic (PrecipNet) and the most used by the community (OVATION).

Regarding PrecipNet, our research has revealed many insights. First, repurposing PrecipNet to adapt it to a new challenge is quite challenging. A significant number of subjective and crucial decisions were made by the PrecipNet team, and these can't be changed. This includes the choice of historical data points (we can't take data from T0-6min, for example) and some preprocess-

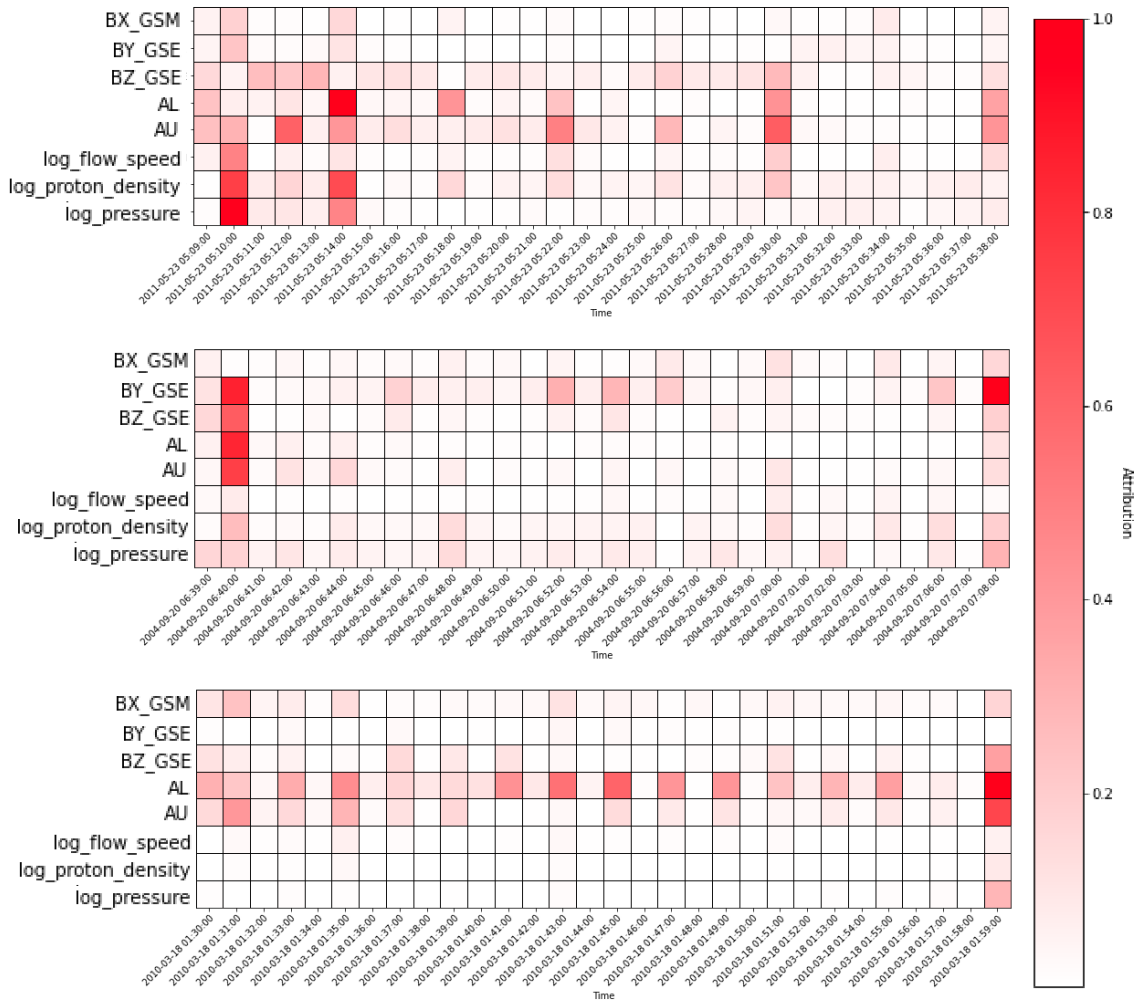


Figure 4.27 – Three displays showing the attribution values from the integrated gradients method for three distinct examples. The first example (top) corresponds to the sample from May 23, 2011, at 05:38. The second (middle) corresponds to the sample from September 20, 2004, at 07:08. The third (bottom) corresponds to March 18, 2010, at 01:59.

ing done on OMNIWeb (using `time_hist2` generates new OMNI data through linear interpolation). While all these decisions are good and relevant, it’s handy to modify them to test other algorithms and methods (like TCNs or a more sophisticated interpolation like PCHIP).

PrecipNet’s lack of a test set makes it difficult to compare with our results. Typically, our results are better on the validation set than the test set, as the user gradually modifies the network’s architecture to enhance the cost function on the validation set. This inadvertently biases the training to be effective on a given validation set. The metrics we deem relevant in our study are those calculated on the test set. Additionally, PrecipNet’s use of the F16 satellite in 2010 as the validation set makes learning challenging, as this set isn’t very representative of the rest of the data. On our side, we opted for a random distribution of data into training, validation, and test sets.

During our final choices, we omitted $F_{10.7}$ and the polar cap index compared to PrecipNet. This was an arbitrary decision aimed at reducing the number of input parameters. The results indicate that their absence didn’t significantly impact the study (difference in results by including it was equivalent to a change in the seed value).

The efforts made to incorporate time and position data as inputs suggest a challenge: the algorithm focuses on modeling the average auroral oval over time. Indeed, we achieve nearly 60% of our predictive capabilities with only these parameters. However, the remaining percentages are most critical, containing dynamic and local variations due to violent eruptions that can damage satellites. One envisioned solution was to focus on a single cell in a predefined grid of latitude and longitude (specifically MLAT and MLT). This solution proved unfruitful due to a lack of data for each grid cell and was abandoned.

Therefore, the chosen solution was to place much more importance on high-value data than others. This is the rationale behind using the Tail Weighted Loss function and specific metrics for high values (like the MSE 90, 95, and 99). The presented results are those obtained for traditional neural networks and for TCNs, with and without this loss function. In the end, we conclude that using this loss function significantly improves our results on extreme values (a reduction of 0.491 MSE on average across all our tests for the MSE 90, 95, and 99 metrics) but slightly degrades them overall (an increase of 0.055 MSE on average across our tests).

All comparisons made between our FCNN and OVATION indicate significant improvements by our algorithms: an enhancement of 20% in MAE, 61% in RMSE, and 85% in MSE. OVATION tends to consistently overestimate actual values, making it better for extreme values. It improves our model containing the tail weighted loss by an average of 33% on "extreme" MSE metrics.

The promising results of our TCN, which also significantly improves upon OVATION's results, prompted us to select it as our primary algorithm for the product. First, it requires far fewer parameters (ten times fewer than PrecipNet with 54,991), thus demanding reduced computational capacity. It's therefore *less expensive* to train. Additionally, the use of convolution allows us to try new methods, such as integrated gradients. We hope to leverage such methods to uncover the time lags between input and output in the future. They will then be usable later in predicting other near-Earth phenomena.

Regarding its drawbacks, we can mention its training time, implying that it won't be updated frequently with new data. Also, it demands data for 30 consecutive minutes, without gaps. But this issue can be circumvented with carefully chosen interpolations, as we've seen.

In conclusion, our FCNN and TCN both yield satisfactory results, and the groundwork has been laid for improving these results within SpaceAble. The data organization, preprocessing done, and the scientific details put into this thesis will serve as an efficient starting point to explore a broader range of methods. For instance, the TCN provides more opportunities in the analysis of longer temporal spans. Its use, paired with other algorithms, might allow analyzing the entire past for a single prediction. The arbitrary 30-minute limit was merely to validate a *proof of concept*, but it can be modified.

Moreover, we'll conclude this thesis by looking at the potential perspectives related to this work, both within the collaboration with SpaceAble and for the broader Space Weather community.

5

Future Directions & Applications

"Поехали!"

Yuri Gagarin - 06:07 UT, 12 April 1961

Contents

5.1	Summary of Key Observations & Results	253
5.2	Perspectives for Future Work	254
5.2.1	Improvements over our Algorithms	254
5.2.2	Combination of Models	255
5.2.3	Integrated Gradients	256
5.3	Implications and Applications for the Research Community	257
5.4	Implications and Applications for the Industry	257
5.5	Conclusion	258

5.1 Summary of Key Observations & Results

As we said at the beginning, from Chapter 1 to Chapter 3, this thesis is structured as a compilation, and will serve as a foundational resource for all research conducted within the SpaceAble company in this domain. It covers a wide array of topics, encompassing various physical phenomena associated with the Sun-Earth interaction, discussions on risks and hazards, an overview of essential artificial intelligence concepts, and several practical algorithm tests relevant to our specific problem. Furthermore, this thesis can be regarded as a stepping stone for exploring novel applications of artificial intelligence in predicting near-Earth phenomena.

Now that we've reached the conclusion of our code and its functionality in the previous chapter, let's pinpoint the most significant observations. Firstly, the code's construction and organization, particularly our choice to implement a library like PyTorch-Lightning, make it an ideal candidate for integration as a product into the SpaceAble platform. In this context, we've already conducted initial trials, including participation in an AI challenge hosted on the Driven Data platform¹, organized by NOAA, titled "MagNet: Model the Geomagnetic Field"². Our research, employing an LSTM with the PyTorch-Lightning library, enabled us to secure the 28th position out of more than 500 participants and laid the groundwork for integration into SpaceAble.

Secondly, we hope our study offers valuable insights into data analysis, utilization, and integration. The datasets used may exhibit significant variations from one test to another, whether due to comparison with existing data or variations in preprocessing methodologies. We, on our part, diversified our data sources (ACE, OMNI), experimented with various interpolation techniques for handling missing data, and conducted extensive research on the reliability of databases. In this context, an article has been published (Bouriat et al., 2022), and we've formed a partnership with the Laboratory of Plasma Physics (LPP) in Paris and Ecole Polytechnique to assess the reliability of OMNI data. We initiated a mentoring project with school students in 2022-2023 aimed at enhancing the quality of the OMNI database through the use of artificial intelligence. Whether it pertains to the selection of historical data for FCNN, the preparation of indices for TCN, data interpolation, outlier detection, merging North and South pole data, or excluding specific years of measurements, the choices are multifaceted, often subjective, and tailored to each specific problem. Our overarching conclusion is that, for effective industrial application, preprocessing should remain accessible and customizable by the user. The product can and, in most cases, should streamline the implementation of these preprocessing steps but should refrain from implementing "built-in" preprocessing that the user might not have considered.

Thirdly, we can confirm that leveraging AI in the field of space weather is a viable avenue. AI has demonstrated its efficacy across numerous space weather applications and reaffirms its relevance in this research domain. The utilization of hybrid libraries that offer greater flexibility in reconciling theoretical physics with AI models represents a promising direction for short, medium, and long-term predictions. Through our work, we aim to have opened up new possibilities in this application domain. The use of TCN, an architecture that might be considered more unconventional, in space weather had not been explored previously and validates the existence of a multitude of options beyond the conventional recurrent networks for time series forecasting. This marks a promising initial step. Combined with other contemporary techniques such as integrated gradients, we aim to further enhance our results and bridge the gap between white-box and black-box models.

1. <https://www.drivendata.org/>
2. <https://www.drivendata.org/competitions/73/noaa-magnetic-forecasting/page/279/>, last accessed on September 13th

5.2 Perspectives for Future Work

This section covers upcoming improvements and practical algorithm tests for the next year. It's divided into three parts. The first part discusses quick and easily interpretable tests for future use. In the second part, we explore potential by combining different models. We introduce some possible algorithm combinations and outline a stacking method involving multiple AI algorithms. Finally, we conclude by discussing the potential applications of integrated gradients.

5.2.1 Improvements over our Algorithms

To begin, a pivotal way for improvement involves the utilization of our dataset by expanding the size of the training set. We can extend TCN training to the dataset comprising 3 million data points. However, it's worth noting that this could necessitate a significantly extended training period, likely in the order of 36 hours. The expansion of our dataset has the potential to empower the TCN model to enhance its generalization capabilities by capturing more intricate data nuances. An expansion of our dataset also means the use of other other data. We did not used the $F_{10.7}$ index like PrecipNet but it might be an interesting path to explore. It could be even more interesting to use the F_{30} index.

Furthermore, as previously mentioned, there's the possibility of incorporating a substantially larger time window as input for our TCN, perhaps in the realm of 2 to 3 hours. This decision comes with its own set of challenges, such as handling missing data. Nevertheless, we have access to a preprocessed dataset where the data has been meticulously interpolated and remains untapped. As illustrated by the examples related to Integrated Gradients at the conclusion of the preceding chapter, the sources of variations in low Earth orbit electron energy fluxes may lie further upstream than the 30-minute window we have considered.

Another way for exploring pertains to the design of our cost function. Currently, there exists potential for its revision to more effectively constrain the algorithm during training on less frequent data. One approach could involve drawing inspiration from the data distribution, possibly employing a transformation to render the distribution histogram more uniform. Additionally, we could pinpoint the data samples for which the algorithm performs optimally within an already uniform distribution, and work to comprehend the underlying physical factors contributing to this prediction ease. Such insights could be used to craft an even more efficient cost function.

A supplementary perspective involves enhancing our predictions in specific geographical regions. Our investigations have shown that our modeling of auroral ovals exhibits diminished efficacy in areas characterized by limited coverage by DMSP. To address this, we can aim to have data evenly spread out across different locations in terms of latitude and longitude, which is not the case with our current dataset. We can use methods to add more data points where needed, and we can also expand some satellite measurements using interpolation to fill in gaps. Naturally, it's important to carefully examine these approaches and back them up with physical equations.

"Feature engineering" represents another promising avenue for exploration, involving the creation of novel data from existing features. Analogous to the manner in which solar wind pressure is computed from velocity and density, we could mix features to derive fresh insights more pertinent to our algorithm.

Addressing data noise presents various challenges, each requiring tailored solutions. Autoencoders offer one potential approach, particularly effective for certain types of noise. However, it's crucial to explore alternative methods for reducing noise or detecting anomalies to enhance over-

all data quality. Additionally, investigating regularization merits attention. While its effectiveness has been limited in our tests, reevaluating its parameters and conducting comprehensive research efforts may yield valuable insights into mitigating specific types of noise.

Finally, and of course, considering new model designs is something to think about for future testing. We haven't used recurrent models like RNNs or LSTMs yet, for instance.

5.2.2 Combination of Models

In this study, we conducted a thorough comparative analysis between our prediction model and the OVATION Prime model to evaluate their respective performances in predicting precipitated electron energy flux values in astrophysics. Overall, our model exhibited significant improvements over OVATION Prime, delivering more precise predictions for most values. However, a critical observation is that OVATION Prime outperforms our model in predicting exceptionally high values, despite its inaccuracies in other value ranges. These high-value predictions hold particular significance in our research context, prompting us to consider retaining OVATION Prime predictions for this specific range.

To address this issue, one potential approach involves combining both models using a strategic threshold system. For instance, we could establish a threshold, denoted as "X," beyond which our model's predictions would be replaced by those of OVATION Prime. Nevertheless, it is crucial to note that this approach may lead to a slight underestimation for values just below the threshold X, as we would retain our model's predictions. This combination approach necessitates meticulous consideration to ensure optimal accuracy while acknowledging the critical importance of high values in our research domain. To achieve this, we plan to conduct a study to identify the optimal threshold X, where both our model and OVATION would exhibit equivalent performance. Another intriguing avenue, showcasing the versatility of possibilities, would involve concurrently running both algorithms and weighting their results, potentially leveraging other algorithms.

Furthermore, as illustrated in Figure 5.1, we propose a stacking method that integrates various machine learning architectures. Our approach encompasses several key steps: information encoding using a TCN, medium-term prediction via an LSTM, and decoding through a fully connected network or another recurrent network.

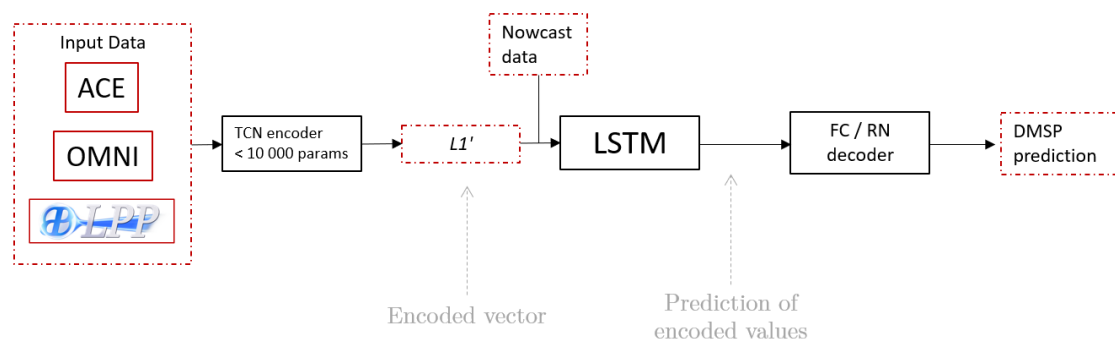


Figure 5.1 – Idea of a stacking method to combine several AI algorithms to forecast DMSP values.

Primarily, we intend to employ the TCN as the encoder. TCNs have proven effective in capturing long-term temporal dependencies while maintaining relative lightweight in terms of parameters (less than 10,000). This balance preserves computational efficiency while upholding robust

modeling capabilities for time sequences. When encoding the data, we generate an encoded representation, often referred to as an "embedding," designed to encapsulate temporal characteristics or significant trends within the data. This encoded representation can be viewed as a compressed and abstract rendition of the input data, encapsulating all the information required for subsequent algorithms to make predictions. However, it may not be readily interpretable in a conventional sense; its optimization is geared towards prediction rather than direct human comprehension.

LSTMs are renowned for their proficiency in handling intricate sequential dependencies. Introducing our encoded representation into an LSTM allows us to harness the TCN's aptitude for capturing long-term dependencies while leveraging the LSTM's capability to model nonlinear sequences. Notably, at this juncture, we have the flexibility to directly incorporate "nowcast" data into the LSTM input, such as K_p values. This inclusion of data, for which a retrospective analysis may not be imperative, nonetheless provides insights into the immediate near-Earth space environment and can be seamlessly integrated into the model without disrupting the input sequence's continuity.

The selection between a Fully Connected Network (FCN) and an alternate recurrent network (e.g., another LSTM) as the decoding layer hinges on the nature of the output data. FCNs are aptly suited for regression tasks where the relationship between input and output data tends to be generally linear. In contrast, recurrent networks excel in managing more intricate relationships and sequential dependencies within output data. The decision should be predicated on empirical assessments to ascertain the optimal architecture.

In summary, our stacking method offers a smart combination of machine learning techniques. It starts with a simple TCN encoder and adds an LSTM to effectively handle intricate sequential patterns. The TCN's encoded representation skillfully captures important timing aspects in the data. Additionally, we can smoothly include "nowcast" data in the LSTM input, providing the model with timely information about the near-Earth space environment.

5.2.3 Integrated Gradients

While TCNs and FCNNs are effective at capturing temporal dependencies in data, like many algorithms, their predictions can often be seen as "black boxes," meaning we don't fully understand how they use input data to generate outputs. By using the Integrated Gradients method, we can actually measure the relative importance of each time point in the input sequence for each prediction. This allows us to highlight the temporal patterns that had the most impact on the model's results for a specific prediction. In essence, it makes the model's predictions more transparent, and that's the whole point.

Moreover, it's also possible to determine the time delay between the most "influential" input and the output. This can be crucial for understanding the underlying physical phenomena and adjusting the model accordingly. For instance, if the delay is identified as significant, we can consider shifting the input data to account for this delay within the model. After training on the entire dataset, we can provide histograms of the identified delays. Further analysis can help us find correlations between these delays and specific inputs. We expect a reasonably strong correlation with solar wind speed, but there might be some interesting surprises.

5.3 Implications and Applications for the Research Community

We hope that this work will make a valuable contribution to the Space Weather community, and we see two primary directions for its potential use. The first involves employing it as an enhanced tool for forecasting, flux calculations, or other relevant parameters. The second entails integrating our models into existing physical codes to improve them. This integration also allows for a comparison of our results with those obtained from these physical codes.

Throughout this study, our fruitful collaboration with Dr. Maxime Grandin from the University of Helsinki has sparked ideas for joint projects. The Vlasiator code, developed by the University of Helsinki, is a six-dimensional Vlasov theory-based simulation. Its aim is to model the entire near-Earth space on a global scale using a hybrid kinetic Vlasov approach, enabling the study of fundamental plasma processes (reconnection, particle acceleration, shocks) and enhancing our understanding of space weather. Initially, we plan to evaluate our models by comparing our results to those of Vlasiator. In the future, we may even consider closely collaborating on topics involving modeling the delay between solar wind and its impact on Earth, speeding up Vlasiator's computational times using AI, or extending medium-term predictions to the entire magnetosphere.

Simultaneously, within the community, there are numerous potential synergies with our work. Although our research primarily focuses on one aspect of the Sun-Earth chain, it can be integrated with other elements, such as solar wind forecasts derived from solar images or ionospheric modeling. Some research conducted by CSUG and IPAG, such as the Transsolo code connecting particle flux to auroral brightness through ground-based measurements, could be meaningfully incorporated. Furthermore, CSUG collaborates with ESA to develop a hyperspectral instrument called WFAI, which includes 9 visible hyperspectral imagers and 3 UV wide-field imagers for studying polar auroras, as part of ESA's Distributed Space Weather Sensor System mission aimed at establishing space weather monitoring (Kraft et al. (2019)). This offers additional opportunities to integrate our results, either to improve measurement quality or by directly incorporating this new data into our code to enhance modeling.

In conclusion, our primary focus will be on combining and comparing our algorithm with existing physical codes, using it to support space missions, harnessing new data to refine its results, and adapting it to new challenges (predicting other phenomena). The key lies in future collaborations and the integration of our algorithm with existing resources. To achieve this, it will be important to leverage the latest AI advancements to advance space weather research even further.

5.4 Implications and Applications for the Industry

This thesis holds a special significance for us, especially in a time when the space weather industry is undergoing notable changes. We're witnessing a growing presence of startups specializing in Space Situational Awareness, marking the early stages of Space Weather integration into the private sector. In this context, our work will be utilized by SpaceAble to enhance space activity management and bolster the resilience of technological systems against space weather events.

As previously discussed, the immediate application of our models lies in space environment prediction. We'll be using data on electron energy fluxes in astrophysics to assess risks, particularly for satellites in low Earth orbit. Integrating our space environment prediction models into

private space operations will assist companies in making informed decisions to safeguard their space assets and mitigate disruptions caused by space weather. Moreover, these prediction models offer the potential to enhance mission planning. Space agencies can utilize our predictions to evaluate expected environmental conditions during launches and operations, thereby ensuring astronaut safety and spacecraft reliability.

To operationalize these models in real-world industrial settings, a multi-tiered approach is essential. During the research and development phase, researchers have full control over every aspect of the model, from its configuration to the underlying code. However, when transitioning these models to a business environment, simplification is required to facilitate flexible adjustments. This entails enabling companies to modify settings seamlessly, retrain the model, and provide user-friendly tools for end-users. We believe that adopting a PyTorch-based approach, particularly leveraging PyTorch Lightning, offers significant advantages. It streamlines hyperparameter management and integrates smoothly into production workflows, enhancing model scalability to meet business requirements.

Companies in the energy sector can also leverage our models to enhance the security of their electrical grids by proactively addressing space weather-induced disruptions. Similarly, telecom companies can incorporate these models into their operations to minimize service interruptions resulting from space radiation. Emerging startups focusing on space communications can ensure the reliability of their networks. Furthermore, insurance companies have a vested interest in improving predictive capabilities across the board.

It's important to note that due to confidentiality considerations, specific details regarding SpaceAble's applications and partnerships in the realm of space weather cannot be fully disclosed in this thesis. The private space industry is continually evolving, with numerous collaborative initiatives and projects in progress. Confidential information related to specific commercial applications and strategic partnerships cannot be publicly divulged at this juncture. Nevertheless, we remain optimistic that these developments will exert a significant impact on the future of the private space industry, fortifying its ability to confront the challenges posed by space weather.

5.5 Conclusion

In wrapping up this final chapter, this thesis signifies a step forward in bridging the gap between space weather and artificial intelligence, especially when it comes to predicting phenomena related to the interaction between the Sun and Earth. The observations and findings shared in this document have laid some groundwork for using artificial intelligence in forecasting events near Earth, which holds promise for both the research community and the space industry.

We've demonstrated that our model, built on architectures like TCNs and FCNNs, not only performs well in predicting energy flux values of precipitated electrons in astrophysics but also offers interpretability advantages through the use of Integrated Gradients. This transparency in predictions can be crucial for understanding the underlying physical processes and making adjustments to the model accordingly.

Our research has also affirmed that artificial intelligence has a legitimate role in the field of space weather, paving the way for the use of hybrid libraries that combine theoretical physics with

AI models. Furthermore, integrating our models into existing physical codes presents an exciting opportunity to enhance the quality of forecasts and advance space research in general.

Looking ahead, we've identified several promising areas for further study. These include expanding our training datasets, refining our cost functions, exploring new methods for reducing noise and detecting anomalies, as well as delving into previously unexplored model architectures like RNNs and LSTMs.

In essence, this thesis doesn't mark the end but rather is part of this new era in space weather research, where artificial intelligence and physical modeling can work in tandem to provide more accurate forecasts and deeper insights into space-related phenomena. We hope that this work will serve as a catalyst for future fruitful collaborations, applications in industry, and significant progress in our understanding of our space environment. As space weather continues to evolve, we eagerly anticipate how this research will contribute to its ongoing development.

Conclusion

This marks the culmination of three years of CIFRE collaboration involving SpaceAble, the University of Grenoble-Alpes, CSUG, and IPAG. CIFRE theses serve as a bridge between practical applications and research domains, and in this particular case, they have played a pivotal role in consolidating crucial knowledge within the domains of Space Weather, space operation risk, and artificial intelligence. This thesis serves as a comprehensive reference to underpin forthcoming multidisciplinary endeavors at SpaceAble. Simultaneously, it has yielded operational codes and algorithms meticulously crafted for seamless integration into the company's operational activities. These tools will be employed to deliver services to a diverse clientele, encompassing space operators and insurance entities alike. This thesis resides within the broader context of Space Situational Awareness (SSA), with its primary objective being the assurance of operational security within the realm of space. It mirrors the escalating integration of artificial intelligence into the transition process from research to practical operations (R2O) within the domain of space meteorology.

Throughout this study, we have cultivated a suite of immediate prediction models employing fully connected convolutional neural networks (FCNNs) and temporal convolutional networks (TCNs). These models have been nourished with data stemming from the bow shock nose (OMNI), encompassing IMF components, solar wind attributes (velocity, pressure, density), AL, AU, SYM-H, along with temporal and spatial attributes. Notably, these models have succeeded in effectively capturing and modeling the total electron energy flux within the auroral region—a pivotal metric furnished by the Defense Meteorological Satellite Program. The outcomes attained have markedly surpassed the capabilities of the incumbent model, OVATION Prime, which is still utilized in operational contexts. Furthermore, our investigation into contemporary techniques like TCNs (2018) and interpretive methodologies such as "integrated gradients" (2017) has ushered in new horizons, with profound implications for industrial applications and ongoing research endeavors.

Furthermore, we've explored the idea of blending physics with machine learning, paving the way for potential future applications in heliophysics. Drawing inspiration from the recent accomplishments of Vlasiator, a software dedicated to simulating particle behavior in the Earth's magnetosphere, we anticipate that our discoveries could serve as a valuable framework for studying auroral phenomena. This framework might also contribute to the development of electrodynamic models for polar auroras at high latitudes, seamlessly integrating them into global circulation models. Our upcoming research endeavors will focus on investigating synergies between various AI approaches, aiming to deepen our understanding of space processes.

Our journey has led to some progress in Space Situational Awareness (SSA) and space meteorology but has mainly also highlighted the transformative impact of artificial intelligence in tackling complex challenges within space science. The outcomes of this thesis usher in a new era of potential applications, spanning from auroral phenomena prediction to effective risk management in space operations. These applications can harmoniously align with existing AI initiatives. Additionally, we hope that our efforts underscore the importance of translating scientific discoveries into practical solutions. As we move forward, it remains crucial to continue seeking novel

Conclusion

connections between physics and machine learning, constructing a sturdy bridge between a fundamental comprehension of space processes and their real-world applications. This bridge not only reshapes how we utilize space but also enhances our understanding of the universe.

Conclusion (français)

C'est ainsi que nous concluons le fruit de trois ans de collaboration CIFRE entre l'entreprise SpaceAble, l'Université Grenoble-Alpes, le CSUG et l'IPAG. Les thèses CIFRE ont le rôle de faire le lien entre la sphère opérationnelle et le domaine de la recherche, et dans ce cas, elle a servi à consolider des connaissances essentielles dans les domaines de météorologie de l'espace, des risques pour les opérations spatiales et de l'intelligence artificielle. Cette thèse agit comme un recueil théorique pour soutenir les futurs travaux multidisciplinaires de SpaceAble. De l'autre côté, elle a généré des codes et des algorithmes opérationnels conçus pour s'intégrer directement dans les activités de l'entreprise. Ces outils seront utilisés pour fournir des services à des clients, qu'ils soient des opérateurs spatiaux ou des assureurs. Cette thèse s'inscrit dans le contexte plus large du Space Situational Awareness (SSA), dont l'objectif est de garantir la sécurité des opérations spatiales, et elle reflète l'essor de l'utilisation de l'intelligence artificielle dans le processus de transition de la recherche vers les opérations (R2O³) en météorologie de l'espace.

Dans cette étude, nous avons développé plusieurs modèles de prévision immédiate, en utilisant des réseaux de neurones convolutionnels entièrement connectés (FCNN) et des réseaux convolutionnels temporels (TCN). Ces modèles ont été alimentés en données provenant du nez du choc d'étrave (OMNI), notamment les composantes du champ magnétique interplanétaire, les caractéristiques du vent solaire (vitesse, pression et densité), les indices géomagnétiques AL, AU, SYM-H, ainsi que des informations temporelles et spatiales. Les sorties de ces modèles ont permis de modéliser avec succès le flux énergétique total des électrons dans la région aurorale, une mesure essentielle fournie par le Defense Meteorological Satellite Program. Les résultats obtenus ont montré une nette amélioration par rapport au modèle existant, OVATION Prime, encore utilisé dans les opérations. De plus, l'exploration de techniques récentes telles que les TCN (2018) et les méthodes d'explication telles que les "integrated gradients" (2017) a ouvert de nouvelles perspectives pour les applications industrielles et la recherche en cours.

Nous avons également mis en lumière les opportunités d'intégrer la physique dans la conception des modèles d'apprentissage automatique, ouvrant ainsi de nouvelles voies pour l'utilisation future de l'apprentissage automatique en héliophysique. Tout comme les récents résultats de Vlasiator⁴, un code physique qui propage les particules dans la magnétosphère, nous espérons que nos résultats offriront un modèle précieux pour l'étude des phénomènes auroraux, l'alimentation des modèles électrodynamiques des aurores polaires à haute latitude, et l'intégration dans les modèles de circulation globale. Nos futures recherches se concentreront sur l'exploration de synergies

3. "R2O" est un acronyme utilisé pour décrire le processus de transition de la recherche (Research) vers les opérations (Operations) dans divers domaines, notamment les sciences, la technologie et les services gouvernementaux. Il s'agit d'un concept qui implique de prendre des résultats de recherche ou des innovations technologiques et de les mettre en oeuvre dans des applications pratiques ou opérationnelles.

4. L'Université d'Helsinki développe Vlasiator, une simulation basée sur la théorie Vlasov en six dimensions visant à modéliser l'ensemble de l'espace proche de la Terre à l'échelle mondiale en utilisant une approche cinétique hybride-Vlasov pour étudier les processus fondamentaux des plasmas et mieux comprendre les phénomènes météorologiques spatiaux.

entre différentes méthodes d'IA pour une compréhension plus approfondie des processus spatiaux.

Nous avons non seulement abouti à des avancées intéressantes dans le domaine du Space Situational Awareness (SSA) et de la météorologie de l'espace, mais nous avons également souligné la puissance de l'intelligence artificielle dans la résolution de problèmes complexes en sciences spatiales. Les résultats obtenus dans cette thèse ouvrent la voie à de nouvelles applications potentielles, de la prévision des phénomènes auroraux à la gestion des risques liés aux opérations spatiales, qui pourront se combiner avec de nombreux travaux d'IA existants. De plus, nous espérons avoir également souligné l'importance de mettre en oeuvre des solutions pratiques basées sur des découvertes scientifiques. À mesure que nous avançons, il est essentiel de continuer à explorer de nouvelles synergies entre la physique et l'apprentissage automatique, créant ainsi un pont entre la compréhension fondamentale des processus spatiaux et les applications pratiques qui façonnent notre utilisation de l'espace et notre compréhension de l'univers qui nous entoure. Cette thèse représente une pierre à cet édifice et nous sommes impatients de voir l'évolution future de ces recherches.

Résumé en français

Ce résumé substantiel en français récapitule les concepts qui sont abordés au fil de chacun des chapitres, ainsi que les modèles développés et les résultats obtenus. Ce résumé ne rentre cependant pas dans les détails et nous conseillons fortement au lecteur de se reporter au manuscrit pour plus d'informations et une meilleure compréhension.

Les chapitres 1 et 2 servant d'encyclopédie dans ce manuscrit, la plupart des concepts présentés sont rappelés succinctement ici, rendant ainsi les résumés de ces deux chapitres plus longs. Les chapitres 3 à 5 englobent notre problématique, de sa pose à sa résolution. Enfin, il n'y a pas d'introduction et de conclusion dans ce résumé, ces dernières pouvant être trouvées, en français, respectivement aux pages 1 et 263.

Chapitre 1: La météorologie spatiale et sa mesure

Ce premier chapitre déroule une présentation générale des concepts qui forment la météorologie de l'espace. La première partie de ce chapitre se concentre sur la physique des plasmas dans laquelle nous détaillons le mouvement des particules dans les champs électromagnétiques ainsi que les écoulements de plasma. La deuxième partie se concentre, elle, sur la chaîne soleil-terre, que nous décrivons de son point de départ, le soleil, jusqu'à la magnétosphère et l'ionosphère terrestre, en passant par une description du milieu interplanétaire et des phénomènes qui s'y trouvent. Enfin, nous terminons par une dernière partie sur la mesure de la météorologie de l'espace. Dans cette dernière partie, nous détaillons les impacts négatifs de la météorologie de l'espace sur les systèmes spatiaux, sur les systèmes sols et sur notre société plus généralement. Nous y soulignons également l'importance et les techniques de mesures et de prévision existantes.

1.1. Introduction à la physique des plasmas

Le plasma, quatrième état de la matière, constitue 99% de la matière visible de l'Univers. Il est étudié via trois approches principales détaillées dans le manuscrit: le *mouvement des particules* chargées dans des champs électromagnétiques, basé sur les équations de Maxwell et la force de Lorentz, essentiel pour comprendre la magnétosphère terrestre et les mouvements cyclotron; la *théorie cinétique*, représentée par l'équation de Vlasov, qui analyse l'évolution de la distribution des particules; et la *magnétohydrodynamique* (MHD), une approche macroscopique qui combine l'électromagnétisme et la dynamique des fluides pour étudier le plasma comme un milieu continu.

La mouvement des particules chargées piégées dans les ceintures de radiations est en général décrit grâce à ce qu'on appelle les invariants adiabatiques, qui sont des quantités conservées sous des changements lents. Les particules subissent des mouvements gyroscopiques, rebondissent entre les pôles terrestres sous certaines conditions et dérivent autour de la Terre en raison de la variation des champs magnétiques rencontrés et de la courbure des lignes. Certaines accélérations comme les accélérations bêatron et Fermi peuvent aussi modifier l'énergie de ces particules "piégées".

L'objectif de la théorie cinétique, de son côté, est d'étudier l'évolution temporelle de la fonction de distribution des particules dans un espace à 6 dimensions, en négligeant les collisions. L'équation de Vlasov, représentant l'évolution temporelle de la fonction de distribution dans les champs électromagnétiques, est au cœur de cette approche, bien qu'elle soit complexe et difficile

à résoudre.

La MHD, elle, étudie les plasmas à grande échelle, les traitant comme un fluide unique. Elle établit des grandeurs macroscopiques et utilise un ensemble d'équations pour décrire la dynamique des plasmas. Des concepts clés en MHD sont introduits dans cette thèse :

- Le nombre de Reynolds magnétique : Il détermine le rapport entre les termes de convection et de diffusion, indiquant si le plasma est dans le cas idéal de la MHD ou dominé par la diffusion.
- La condition de Champ "Gelé" : Cette condition signifie que le plasma et le champ magnétique sont interconnectés, avec le mouvement du plasma entraînant les lignes de champ magnétique.
- Le paramètre bêta : Il permet la distinction entre plasma "froid" (dominé par les forces magnétiques) et "chaud" (contrôlé par les effets thermiques).

Suivant l'objectif, chacune de ces approches présente des avantages et des inconvénients. Dans cette première section de la thèse, elle sont présentées en détail, ainsi que les équations qui correspondent.

1.2. La chaîne Soleil-Terre

Le point de départ : notre Soleil

Le Soleil, pivot de la chaîne Soleil-Terre, possède un diamètre équatorial de 1 392 000 km et une masse de $1,99 \times 10^{30}$ kg, rayonnant 4×10^{26} W, dont la Terre capte environ $1,743 \times 10^{17}$ W. Il est principalement composé d'hydrogène et d'hélium.

Le Soleil se structure en plusieurs zones de son centre à sa surface. Le cœur nucléaire, où la fusion nucléaire transforme l'hydrogène en hélium, représente 50 à 70 % de sa masse. La zone suivante est la zone dite "radiative" qui est traversée lentement par les photons. La zone convective qui suit est, elle, caractérisée par des mouvements turbulents et la formation de granules et de supergranules. Ensuite vient la photosphère où la température décroît et qui abrite diverses structures. Puis, la chromosphère, et enfin la couronne, composent les dernières régions solaires.

L'activité solaire, dictée par la rotation du Soleil, son champ magnétique et sa dynamo, suit un cycle de 11 ans. Le champ magnétique solaire influence l'atmosphère externe, génère des éruptions, dirige le vent solaire, et fait barrière aux rayons cosmiques. Cette activité est marquée par des phénomènes comme les taches solaires (dont l'observation a mis en avant la présence de ces cycles), les éruptions solaires, ou encore les éjections de masse coronale (ou CME, qui sont des éjections de grande quantité de matière depuis la couronne), qui influent sur notre Terre. Ces phénomènes sont des exemples porteurs de l'énergie solaire qui nous parvient. L'analyse du spectre d'émission du Soleil, qui va des rayons gamma au rayonnements radios et qui atteint jusqu'à 100 milliards de watts par mètre carré, révèle notamment l'importance des rayons X et du rayonnement EUV.

Le vent solaire dans le milieu interplanétaire

Le vent solaire est un flux de particules chargées émanant du Soleil et découvert en 1957 par E.N Parker. Parker a résolu les équations de continuité et de conservation du moment dans un fluide, établissant ainsi que la couronne solaire s'expandait de manière continue. Cette expansion forme le vent solaire, qui transporte avec lui le champ magnétique du Soleil dans l'espace interplanétaire, connu sous le nom de champ magnétique interplanétaire (IMF).

L'IMF, dont l'orientation est influencée par la spirale de Parker, joue un rôle crucial dans l'interaction entre le vent solaire et la magnétosphère terrestre. Les vents solaires sont classés en trois catégories en fonction de leur vitesse et de leur densité :

- Le vent solaire lent, provenant des spicules solaires, se propage à une vitesse moyenne d'environ 250-400 km/s avec une densité de flux d'énergie totale de 1,55 erg/cm²/s.
- Le vent solaire rapide est émis par les trous coronaires aux pôles nord et sud du Soleil, atteignant une vitesse d'environ 400-800 km/s, avec une densité de flux d'énergie totale de 1,43 erg/cm²/s.
- Le vent solaire issu d'événements éruptifs tels que les éjections de masse coronale (CMEs) peut atteindre des vitesses allant jusqu'à 3000 km/s.

Ces vents solaires, lents et rapides, sont principalement composés d'électrons et de protons, avec une petite fraction de noyaux d'hélium. Les particules solaires énergétiques (SEP) accompagnent souvent les vents solaires sporadiques et sont constituées principalement de protons, d'alpha particules et d'ions lourds, avec des énergies allant de 10 MeV à 100 MeV, et pouvant atteindre jusqu'à 20 GeV. L'interaction complexe entre le vent solaire et le champ magnétique solaire nécessite l'usage des équations magnétohydrodynamiques (MHD) et aboutit à divers phénomènes, comme la feuille de courant héliosphérique. De manière générale, une grande partie des phénomènes de météorologie spatiale est directement liée au vent solaire, que ce soit les régions d'interaction co-rotative (CIR), les éjections de masse coronale interplanétaires (ICMEs) ou l'intensité des rayons cosmiques galactiques (GCR) qui affecte la Terre.

Le vent solaire en interaction avec notre magnétosphère

Dans cette section, nous visons à explorer l'influence du vent solaire sur la magnétosphère terrestre. Pour faciliter les discussions, nous introduisons les caractéristiques de la magnétosphère, tout en évitant les détails techniques, afin d'établir un vocabulaire commun. Nous explorons ensuite les processus résultant du couplage vent solaire - magnétosphère. Nos références incluent les oeuvres de Koskinen & Kilpua ou Russell & Luhmann. Les lecteurs sont encouragés à consulter les nombreuses littératures sur le sujet.

Le champ magnétique terrestre est le produit du mouvement de convection du fer liquide dans le noyau externe de la Terre. Ce mouvement génère un phénomène appelé "effet Dynamo", qui produit un champ magnétique global. Ce champ, en forme de dipôle nord/sud, est légèrement incliné par rapport à l'axe de rotation de la Terre, avec son centre décalé par rapport au centre géométrique de la planète. Cette configuration crée une région de protection autour de la Terre appelée magnétosphère.

La magnétosphère est la région de l'espace autour de la Terre où le champ magnétique terrestre est dominante sur le champ magnétique interstellaire. Elle est principalement formée par l'interaction entre le vent solaire, un flux continu de particules éjectées par le Soleil, et le champ magnétique terrestre. Lorsque le vent solaire rencontre la magnétosphère, il exerce une pression sur le côté ensoleillé (ou "jour") de la magnétosphère, déformant le champ magnétique en une sorte de "queue" du côté opposé à la lumière solaire. Sur le côté jour, le nez du choc d'étrave (BSN) marque le point avant de ce choc.

La magnétopause est la frontière externe de la magnétosphère, où le vent solaire rencontre directement le champ magnétique terrestre. Sa taille et sa forme varient en fonction de l'intensité du vent solaire. En période calme du Soleil, le BSN peut être situé à environ 13 rayons terrestres du côté ensoleillé de la Terre, tandis qu'en période d'activité solaire, elle peut se rapprocher à environ 6 rayons terrestres.

Les différentes régions de la magnétosphère décrites dans cette thèse, comme la magnétogaine, les cornets polaires, la magnétoqueue, les lobes magnétosphériques, etc., ont des caractéristiques distinctes en termes de composition plasma et de comportement magnétique. Ces régions ne sont pas statiques et peuvent se déformer ou se déplacer en réponse aux variations du vent solaire.

Les particules du vent solaire peuvent pénétrer la magnétosphère de différentes manières. La reconnexion magnétique, où les lignes de champ magnétique du Soleil et de la Terre s'alignent, est l'un des mécanismes clés. Lorsque cette reconnexion se produit du côté ensoleillé de la magnétosphère, elle permet aux particules chargées du vent solaire de pénétrer plus profondément dans la magnétosphère.

En plus de la reconnexion magnétique, les particules du vent solaire peuvent également pénétrer la magnétosphère par friction le long des flancs de la magnétopause ou à travers des régions de faiblesse dans la magnétosphère, comme le feuillet de plasma. Ces particules transportées le long des lignes de champ magnétique interagissent avec l'atmosphère terrestre, notamment dans les régions polaires, où elles causent des phénomènes tels que les aurores polaires.

Les aurores polaires constituent ce qui nous intéresse dans cette étude et nous détaillons un peu plus loin la physique associée. Cependant, nous pouvons revenir sur les trois sources principales de particules précipitées, à l'origine des aurores :

- Les ceintures de radiations de Van Allen, découvertes par James Van Allen en 1959, représentent une source majeure d'aurores polaires. Ces ceintures, composées de particules piégées par le champ magnétique terrestre, sont situées dans la magnétosphère interne de la Terre. Elles se présentent sous forme de zones toroïdales où les particules, principalement des électrons dans la ceinture externe et des protons dans la ceinture interne, sont maintenues en orbite entre les hémisphères grâce à un effet de miroir magnétique. Les ceintures de Van Allen interne et externe ont des altitudes variables et sont alimentées différemment : la ceinture interne est principalement alimentée par l'effet CRAND, tandis que la ceinture externe est alimentée par le feuillet de plasma. Une particularité à noter est l'anomalie de l'Atlantique Sud, où le champ magnétique terrestre est affaibli, permettant aux particules piégées d'atteindre des altitudes très basses, ce qui peut poser des risques pour les satellites en orbite basse. Les particules des ceintures de radiations peuvent parfois s'échapper de leur confinement et précipiter vers les pôles, contribuant ainsi à la formation des aurores polaires.
- Le feuillet de plasma, situé côté nuit autour du plan équatorial de la magnétosphère à environ 6 rayons terrestres, est la principale source des aurores polaires. Il contient un plasma dense et chaud, avec des ions énergétiques (2 à 5 keV) et des électrons énergétiques (0,5 à 1 keV). Les particules dans le feuillet de plasma subissent une reconnexion magnétique, similaire à celle observée à la magnétopause, qui les réoriente vers la Terre. Ces particules, principalement les électrons, sont accélérées jusqu'à des énergies de 10 à 100 keV lors de leur trajet vers les pôles terrestres. À leur arrivée, elles précipitent dans les régions polaires, alimentant les aurores et la ceinture externe de radiation. Une ceinture de radiation supplémentaire, appelée anneau de courant, est également alimentée au niveau de l'équateur, contenant des ions et des électrons de basses énergies d'environ 50 keV.
- Les cornets polaires, situés aux pôles terrestres, agissent comme une source alternative d'alimentation des aurores polaires. Ces zones, ouvertes sur l'espace interplanétaire et éloignées des points de reconnexion magnétique, permettent aux particules, appelées d'accéder directement à la haute atmosphère terrestre. Contrairement aux particules accélérées par la reconnexion magnétique, celles provenant des cornets polaires ne subissent généralement pas d'accélération. Elles génèrent des aurores diffuses à haute latitude, avec des énergies de

l'ordre de 100 eV et un flux faible. Bien que leur flux soit souvent insuffisant pour exciter les molécules responsables des aurores, des particules à plus haute énergie, de l'ordre du keV, peuvent créer des aurores en dehors de l'ovale auroral. Le processus précis à l'origine de ces particules de haute énergie appelées "polar rain" reste encore mal compris.

Tempêtes géomagnétiques et sous-orages

Les tempêtes géomagnétiques sont des perturbations majeures du champ magnétique terrestre, déclenchées par des événements solaires intenses tels que les éjections de masse coronale et les éruptions solaires. Elles se caractérisent par une augmentation rapide suivie d'une chute significative du champ magnétique, avec une phase de récupération progressive. Ces tempêtes résultent de processus de reconnexion magnétique, entraînant une injection substantielle de plasma dans la magnétosphère interne et la formation d'un courant annulaire.

Les sous-orages géomagnétiques sont plutôt initiés dans la magnétosphère nocturne et résultent de la libération d'énergie stockée dans la queue magnétique. Ils conduisent à la formation d'aurores et peuvent durer de une à plusieurs heures. Deux théories principales expliquent leur déclenchement : le modèle de la ligne neutre proche de la Terre et le modèle de la rupture du courant transversal. Ces phénomènes sont essentiels à étudier en raison de leur impact sur les réseaux électriques, les communications et les systèmes satellitaires, et impliquent une interaction complexe entre le vent solaire, la magnétosphère, l'ionosphère et la thermosphère.

Les ondes de plasma dans la magnétosphère

Cette section que nous ne détaillerons pas dans ce résumé, permet de comprendre le vocabulaire associé aux ondes de plasma dans la magnétosphère. Plusieurs types d'ondes y sont décrites : les ondes de chorus en mode siffleur, les ondes électromagnétiques à cyclotron ionique, les ondes d'Alfvén ou encore les ondes ultra-basses fréquences. Toutes celles-ci sont cruciales car elles interagissent avec les particules, facilitant le transfert d'énergie et de quantité de mouvement, conduisant à divers phénomènes tels que les tempêtes magnétiques. Ces ondes peuvent provenir de sources externes telles que les éclairs et les émetteurs VLF, ainsi que de sources internes telles que les instabilités du plasma. Elles peuvent agir à la fois comme sources et comme pertes de particules. Diverses résonances telles que la résonance Gyro, Landau, Bounce et Drift jouent un rôle crucial dans la modification de l'énergie et de la quantité de mouvement des particules.

L'ionosphère

Dans la thèse, cette section approfondit l'étude de l'ionosphère et son importance pour comprendre les aurores polaires. Nous détaillons rapidement d'abord l'interaction entre magnétosphère et ionosphère à travers les courants alignés et horizontaux. Puis nous détaillons les différentes régions de l'ionosphère.

Les particules magnétosphériques suivent les lignes de champ magnétique vers les régions polaires, situées autour de 65 degrés de latitude magnétique. Ces particules, appelées précipitées, créent des courants ionosphériques. Ces courants se divisent en deux types principaux :

- Les courants alignés dans les zones polaires se répartissent en deux régions distinctes : la région 1 et la région 2. La région 1 provient des mouvements des ions le matin et des électrons le soir dans la magnétosphère interne, engendrant des courants montants et descendants. La région 2 est formée par la modulation du champ électrique de la magnétosphère interne par celui de la magnétosphère externe, produisant des courants montants et descendants opposés.

- Les courants horizontaux, tels que les courants de Perdersen et de Hall, naissent de la circulation de ces courants alignés, créant un nouveau champ électrique dans la région E de l'ionosphère. Les courants de Perdersen, parallèles au champ électrique, et les courants de Hall, perpendiculaires au champ électrique et magnétique, alimentent deux cellules de convection opposées, réchauffant le plasma par effet Joule.

Les électrojets auroraux, comme l'électrojet auroral dans la calotte polaire et l'ovale auroral, sont des exemples de ces courants de Hall, modifiés lors de sous-orages géomagnétiques et orages géomagnétiques. L'ovale auroral, où se trouve l'électrojet auroral, est la région privilégiée d'interaction entre les particules précipitées et les composants de la haute atmosphère, offrant ainsi des observations privilégiées des aurores polaires.

La thermosphère est principalement composée d'oxygène atomique, de dioxygène et de diazote. Elle présente une forte augmentation de température de 80 à 200 km environ, s'élevant jusqu'à 750 K en période calme et 1500 K en période d'activité solaire intense.

L'ionosphère, la partie ionisée de l'hétérosphère, est créée par l'ionisation des molécules et atomes de la thermosphère, principalement par le flux EUV & UV du Soleil et les précipitations de particules lors d'orages géomagnétiques. C'est dans cette couche que se forment les aurores polaires.

La région D, entre 50 et 90 km, est influencée par le rayonnement solaire et cosmique, principalement composée d'ions NO^+ et O_2^+ . La région E, entre 90 et 150 km, est caractérisée par une ionisation due au rayonnement solaire X et UV, avec une densité d'électrons variant entre le jour et la nuit. La région F, de 150 à 1000 km, est la plus dense de l'ionosphère, principalement composée d'oxygène atomique et de diazote, avec une ionisation principalement induite par le rayonnement solaire EUV.

Les orages géomagnétiques intenses maintiennent les différentes couches ionosphériques pendant la nuit, contribuant à la formation des aurores polaires. Ces précipitations de particules, également appelées "suprathermiques", ont des énergies supérieures à celles thermalisées dans l'ionosphère, influençant ainsi le processus d'ionisation et d'excitation.

En conclusion, dans la thèse, cette section offre une compréhension fondamentale de l'ionosphère et de son interaction avec la magnétosphère, préparant le terrain pour une exploration plus détaillée des aurores polaires et de leurs variations lors des sous-orages géomagnétiques, ainsi que des approches actuelles de modélisation.

La physique des aurores

Les aurores polaires sont le résultat visible final d'un processus complexe qui implique le vent solaire, la magnétosphère et la haute atmosphère. Pour débiter, nous abordons les effets de l'accélération des particules et les différentes catégories d'aurores qui en découlent. Il est important de noter que ces accélérations de particules sont étroitement liées aux courants ionosphériques que nous avons discutés précédemment. Ensuite, nous fournissons un aperçu des différents types d'aurores, incluant les aurores discrètes et diffuses, ainsi que les émissions qui leur sont associées. Enfin, nous examinerons la morphologie des aurores et leur évolution dynamique, en particulier pendant les sous-orages géomagnétiques.

Débutons en décrivant les mécanismes d'accélération des particules, dont découlent les aurores. Le transport des particules sur les lignes de champ magnétique implique plusieurs processus d'accélération magnétohydrodynamiques internes, comprenant :

- L'accélération par potentiel électrique quasi-statique (QSPS), qui crée un champ électrique ascendant et accélère les électrons et ions magnétosphériques.
- L'accélération par onde d'Alfvén dans la magnétosphère, où les électrons sont accélérés le long des lignes de champ magnétique.
- La précipitation des particules via les ondes sifflantes chorales dans le cône de perte, bien que le mécanisme d'accélération associé soit encore débattu.

Ces processus conditionnent l'énergie des particules précipitées, influençant les types d'aurores observées. Comprendre ces phénomènes est essentiel pour étudier l'interaction magnétosphère-ionosphère et les processus magnétosphériques complexes.

Les aurores discrètes se présentent sous forme d'arcs ou de bandes lumineuses intenses dans le ciel nocturne, résultant de l'interaction entre les constituants de la haute atmosphère et les électrons accélérés par les ondes d'Alfvén ou le QSPS. Elles peuvent persister de quelques secondes à plusieurs heures et surviennent lors des sous-orages géomagnétiques. Ces arcs se divisent en deux types : ceux formés par des électrons monoénergétiques accélérés par le QSPS, et ceux générés par des électrons de moindre énergie provenant de l'interaction de la magnétosphère avec le vent solaire à faible latitude, typique des ondes d'Alfvén.

D'autre part, les aurores diffuses se caractérisent par des formes non structurées et étendues dans le ciel, résultant du mécanisme de précipitation "Whistler mode". Contrairement aux aurores discrètes, elles ne sont pas associées à un processus d'accélération spécifique mais plutôt à un mécanisme de diffusion. Ces aurores, moins lumineuses, ne sont pas visibles à l'œil nu. Une sous-catégorie des aurores diffuses, les "pulsating aurora", se distinguent par des taches clignotantes et des séries de pulsations. Elles se caractérisent par une dominante de couleur verte et peuvent être observées pendant quelques secondes à plusieurs dizaines de minutes.

La morphologie des aurores est dynamique et varie avec l'activité solaire. Elles se forment des deux côtés de la Terre dans des ovales centrées autour des pôles géomagnétiques. La forme de l'ovale auroral change avec l'activité géomagnétique, et la région sans aurores, centrée sur le pôle, est appelée calotte polaire. Les aurores peuvent subir des sous-orages, pendant lesquels leur morphologie change brusquement en raison de l'activité magnétosphérique. Ces sous-orages suivent un cycle spécifique, commençant par un intervalle calme, suivi du début du sous-orage, de la progression vers l'ouest et de la phase de récupération.

Plusieurs modèles ont été développés pour étudier et prédire le comportement auroral. Le modèle Feldstein-Starkov relie l'indice Kp à l'emplacement de l'ovale auroral, tandis que le modèle Zhang-Paxton calcule le flux d'électrons basé sur les données satellitaires. D'autres modèles notables incluent ceux de Hardy, Fuller-Rowell, OVATION Prime, et PrecipNet, chacun offrant des perspectives uniques sur le comportement auroral et contribuant à notre compréhension de ces phénomènes. Ces modèles sont essentiels pour les chercheurs car ils fournissent des informations précieuses pour la recherche ionosphérique et thermosphérique et peuvent prédire l'activité aurorale.

1.3. La mesure de la météorologie de l'espace et sa prévision

La météorologie spatiale est définie comme l'état physique et phénoménologique des environnements spatiaux naturels, et elle englobe l'étude et la prédiction de l'activité solaire, des environnements interplanétaires et planétaires, des diverses perturbations les affectant ainsi que des conséquences sur les systèmes spatiaux et biologiques. De nombreux mécanismes peuvent affecter les systèmes, des courants induits géomagnétiquement aux phénomènes de charges sur les

engins spatiaux, en passant par l'érosion, le freinage atmosphérique ou les effets des événements singuliers. De plus, la dépendance croissante de la société vis-à-vis des technologies spatiales souligne la nécessité de comprendre et de mitiger ces risques.

Pour quantifier les risques posés par la météorologie spatiale, une multitude de satellites sont positionnés tout au long de la chaîne Soleil-Terre, tels que Parker Solar Probe, Solar Orbiter, STEREO et Ulysses près du Soleil, ou encore Wind, ACE, DSCVR, SOHO, NOAA POES, et même DMSP et MetOp plus près de la Terre. De plus, des indicateurs ont été développés au fur et à mesure des années afin de renseigner sur les activités solaires et géomagnétiques. On les appelle les indices.

Les indices solaires sont des outils incontournables pour évaluer l'activité du Soleil à tout moment, crucial pour comprendre et modéliser ses divers processus. Par exemple, l'indice R , basé sur le nombre de taches solaires visibles dans la photosphère, offre une perspective historique sur l'activité solaire remontant au XVII^e siècle. En parallèle, plusieurs autres indices solaires sont utilisés pour étudier le spectre du flux solaire émis par différentes régions du Soleil, comme $F_{10.7}$ pour le rayonnement radio et $MgII$ pour le flux solaire extrême ultraviolet (EUV). Ces indices sont intégrés dans divers modèles empiriques pour estimer la densité thermosphérique, ce qui est crucial pour la prévision météorologique spatiale et la compréhension des interactions Soleil-Terre. De manière générale, les indices solaires fournissent une vue complète de l'activité solaire, allant de la radio au rayonnement X, permettant une évaluation exhaustive de son impact sur notre environnement spatial. Par exemple, l'indice $Xb10$ capture les émissions de rayons X des éruptions solaires, tandis que l'indice $H Ly\alpha$ reflète l'émission Lyman- α de l'hydrogène dans la haute chromosphère. Ces indices fournissent une vision précieuse de l'activité solaire à différentes longueurs d'onde, jouant un rôle central dans la modélisation des conditions spatiales et la protection des infrastructures terrestres et spatiales.

Les indices géomagnétiques sont des outils cruciaux pour évaluer l'activité du champ magnétique terrestre au fil du temps, avec les indices Kp et Dst étant les plus couramment utilisés dans les modèles de simulation. L'indice Kp , calculé à partir des données de magnétomètres répartis dans le monde entier, offre une évaluation des perturbations géomagnétiques à l'échelle planétaire, tandis que l'indice Dst quantifie le champ magnétique induit par l'anneau de courant terrestre. De nouveaux indices dérivés de Kp , tels que $Hp90$, $Hp60$ et $Hp30$, offrent des résolutions temporelles plus courtes pour une meilleure modélisation des interactions ionosphère-thermosphère. De plus, des indices comme aa , am , et $ASY/SYM-H$ and D fournissent des informations sur l'amplitude du champ magnétique et des perturbations géomagnétiques à différentes latitudes et longitudes.

Pour quantifier l'activité géomagnétique au niveau de l'anneau de courant, l'indice Dst est calculé à partir de magnétomètres répartis sur le plan équatorial, tandis que les indices $ASY/SYM-H$ and D fournissent des données sur les perturbations géomagnétiques à moyennes latitudes. Les indices PCN et PCS surveillent l'entrée énergétique du vent solaire dans les cornets polaires grâce à des magnétomètres spécifiquement situés dans ces régions. Enfin, l'indice AE est essentiel pour quantifier l'activité aurorale dans l'hémisphère nord, offrant une mesure de l'électrojet auroral et des sous-orages géomagnétiques. Ces indices géomagnétiques permettent une surveillance de l'activité magnétique terrestre à différentes échelles spatiales et temporelles.

1.4. Risques et impacts

Depuis le début de l'ère spatiale avec le lancement de Spoutnik en 1957, les satellites ont été confrontés à des anomalies, révélant ainsi les risques associés à leur déploiement dans l'espace.

Les opérateurs de satellites ont tenté de répertorier et de comprendre ces incidents pour renforcer la fiabilité de leurs engins, mais les bases de données existantes sont souvent incomplètes en raison du manque de données précises et de leur documentation lacunaire. Actuellement, cinq principales bases d'anomalies sont identifiées, avec la base de la NOAA qui reste la plus détaillée malgré son absence de mises à jour depuis 1993. D'autres sources gratuites comme Astranix et Satellite News Digest existent mais souffrent également de lacunes en termes de précision et de validation des données. Deux autres bases, Atrium Space Insurance Corporation et Seradata, offrent une couverture plus complète mais sont privées et nécessitent un accès payant. Les experts recommandent la création d'une base de données unifiée et normalisée pour les anomalies de satellites, mais sa mise en place est freinée par des défis liés à la confidentialité des données et aux ressources limitées.

Freinage atmosphérique

Le freinage atmosphérique, surtout présent en orbite basse, est principalement causé par le frottement dans la haute atmosphère, notamment la thermosphère où l'oxygène et l'hélium jouent un rôle. Ce phénomène peut être calculé en prenant en compte la densité atmosphérique, la vitesse du satellite, la surface de frottement et le coefficient balistique. Les outils de modélisation comme les Drag Temperature Models (DTM), le US Space Force High Accuracy Satellite Drag Model (HASDM), et le Semi-analytic Tool for End of Life Analysis (STELA) sont largement utilisés pour estimer ce freinage. Les instruments comme les accéléromètres sont utilisés pour mesurer les variations d'accélération du satellite causées par le freinage, mais leur utilisation est contraignante en raison de leur taille et de leur sensibilité élevée.

Des alternatives sont explorées, comme le spectromètre de masse, pour mesurer directement la densité atmosphérique. L'activité solaire et géomagnétique sont les principales sources de variation de cette densité, impactant le freinage des satellites. Par exemple, lors d'événements géomagnétiques exceptionnels, l'énergie déposée par les particules précipitées peut être jusqu'à deux fois supérieure à celle du flux solaire EUV. Ces variations de densité sont dues à des processus comme la dissociation de l'oxygène moléculaire et le chauffage par effet Joule.

La principale conséquence de la traînée est la modification des trajectoires, qui peut entraîner la perte de satellites et augmenter le risque de collision. Un freinage accru peut également entraîner une consommation de carburant plus importante, réduisant la durée de vie de la mission. Bien que ce phénomène soit souvent utilisé pour désorbiter les satellites en fin de vie, des désintégrations non contrôlées peuvent survenir. En février 2022, plusieurs satellites Starlink se sont désintégrés avant d'atteindre leurs orbites prévues en raison d'une augmentation de la traînée atmosphérique après une tempête géomagnétique.

Charges interne et de surface

L'histoire de l'étude du spacecraft charging remonte au lancement de Spoutnik en 1957, suivi de missions et de modèles pour comprendre ce phénomène. Des données in situ des satellites comme SCATHA, CRRES, DMSP, et Freja ont considérablement enrichi notre compréhension du spacecraft charging.

Les sources du spacecraft charging incluent le plasma ionosphérique, les électrons suprathermiques, les protons, et le flux solaire UV et EUV. La principale conséquence de la charge des engins spatiaux est la décharge électrostatique (ESD), qui peut compromettre ou détruire des composants électroniques sensibles, provoquer des pertes de puissance et des pannes, et perturber les

opérations. En plus de l'ESD, la charge des engins spatiaux peut également entraîner des interférences électromagnétiques ou des drains de puissance dus aux courants parasites.

Des seuils de dangerosité en fonction de divers facteurs environnementaux et orbitaux ont été établis au sein de SpaceAble suite à de nombreux travaux. Par exemple, le spacecraft charging est plus fréquent lors d'orages géomagnétiques, pendant les périodes de minimum solaire, et en période d'éclipse. Des modèles ont été développés pour détecter automatiquement les conditions de chargement en fonction du type de satellite et de l'orbite.

Les modèles visent à calculer le potentiel d'un satellite pour prédire les décharges électrostatiques potentielles, notamment NASCAP-2K, SPIS, SPARCS, SPENVIS, SHIELDS, et MUSCAT. Ils intègrent la géométrie du satellite, les interactions avec le plasma environnant, et les émissions de particules chargées depuis la surface. Ces modèles sont utilisés pour diverses applications, telles que l'étude du chargement différentiel, la dérivation de conditions expérimentales en laboratoire, l'analyse des anomalies en orbite, et l'interprétation des données scientifiques. Les stratégies d'atténuation pour la charge de surface comprennent l'émission d'électrons, l'émission de plasma, l'utilisation de contacteurs plasma, l'émission de molécules polaires et des dispositifs autonomes. Pour la charge interne, les stratégies comprennent l'utilisation de matériaux à conductivité finie, de peinture partiellement conductrice, de matériaux fins et de blindage. Les prévisions de la météorologie spatiale jouent également un rôle crucial dans l'atténuation des charges de surface et internes

Courants induits géomagnétiquement

Les courants induits géomagnétiquement (GIC) sont un autre risque, générés au sol pendant les changements rapides dans le champ magnétique terrestre. Ils peuvent provoquer des perturbations dans les systèmes électriques, surtout dans les régions à faible conductivité terrestre. Les GIC peuvent entraîner la saturation des transformateurs, des pertes de puissance réactive, des harmoniques, la surchauffe des transformateurs, la surchauffe des générateurs, des problèmes de relais de protection, et des perturbations dans les systèmes de télécommunication.

Evenements singuliers

Les événements singuliers, issus de l'impact d'une particule énergétique unique sur un système, peuvent induire des perturbations électriques. Trois sources principales d'événements singuliers sont répertoriées : les particules solaires énergétiques (SEP), les rayons cosmiques galactiques (GCR), et les protons énergétiques piégés (TP).

Ces événements peuvent avoir des conséquences destructives ou non destructives sur les satellites. Parmi les conséquences non destructives, les Single Event Upsets (SEU) et les Single Event Transients (SET) sont les plus courantes. Les SEU se produisent lorsque des particules énergétiques induisent des changements d'état dans la mémoire ou les bits électroniques, tandis que les SET provoquent des variations de tension dans les circuits électroniques.

En revanche, les conséquences destructives comprennent les Single Event Latchups (SEL) et les Single Event Burnouts (SEB). Les SEL se produisent lorsqu'un SEU provoque un court-circuit dans un circuit intégré, tandis que les SEB peuvent entraîner la destruction de composants électroniques.

Ces événements singuliers peuvent perturber le fonctionnement normal des satellites et sont responsables d'un pourcentage significatif d'anomalies observées. Les SEU, en particulier, représen-

tent la deuxième cause la plus fréquente d'anomalies, selon certaines études.

Effets cumulatifs

Les effets cumulatifs des radiations spatiales, comprenant la dose ionisante totale (TID) et le dommage de déplacement (DDD ou TNID), résultent d'une exposition prolongée. La TID, générée par l'interaction des particules de haute énergie avec les matériaux du satellite, induit la formation de paires électron-trou, entraînant des dommages en profondeur dans les composants. Les sources de cette dose comprennent les protons de haute énergie de la ceinture interne de Van Allen, les particules solaires énergétiques (SEP), les rayons cosmiques galactiques (GCR), ainsi que les électrons de haute énergie des ceintures internes et externes de Van Allen.

Mesurée en Gray (Gy), cette dose cause une détérioration avancée des composants, notamment des semi-conducteurs et des isolants, et peut même les détruire. Les pertes de courant résultant des interactions électron-trou peuvent augmenter la consommation d'énergie, modifier les constantes de temps des dispositifs, réduire le gain des composants ou altérer les propriétés électriques des matériaux.

Quant au dommage de déplacement, également appelé TNID, il est provoqué par les mêmes sources que la TID. L'énergie accumulée à l'intérieur du satellite par les particules de haute énergie déplace les atomes dans les réseaux cristallins des composants, fragilisant ainsi leur structure. Cela se traduit principalement par une réduction significative de la durée de vie du composant.

Conclusion

La météorologie spatiale constitue une menace significative pour l'humanité, comme l'illustrent divers événements passés ayant eu des impacts potentiels variés. Des incidents notables, tels que l'Événement Carrington de 1859, la tempête géomagnétique de 1989 affectant Hydro-Québec, et l'incident de perte de satellites SpaceX en 2022, mettent en évidence des impacts concrets sur les systèmes de communication et la technologie. De nombreux secteurs, tels que la santé humaine, les opérations satellitaires, la navigation GPS, les télécommunications et les centrales électriques, sont vulnérables aux effets de la météorologie spatiale, présentant le risque de perturbations généralisées et de pertes financières. La possibilité d'événements similaires à Carrington dans le futur, avec des conséquences dévastatrices et des temps de récupération s'étendant sur des années, souligne l'importance des mesures de protection. Des programmes internationaux tels que E-SWAN et le Programme de Sécurité Spatiale de l'ESA s'efforcent de traiter et d'atténuer ces risques par le biais de la recherche, du développement et d'initiatives de surveillance.

Parallèlement, l'industrie spatiale a subi des changements significatifs avec l'entrée de nouveaux acteurs tels que SpaceX, Blue Origin et de nombreuses startups. C'est l'avènement du New Space. Les progrès technologiques, la réduction des coûts et l'augmentation des investissements privés ont stimulé l'innovation et le développement de projets tels que Starlink et OneWeb. L'intégration de technologies avancées a entraîné une augmentation des activités spatiales et la création d'un nouvel écosystème spatial. Des entreprises émergentes, telles que SpaceAble et LeoLabs, sont apparues pour aider cet écosystème à relever les défis et garantir la durabilité du secteur spatial.

Chapitre 2: Apprentissage machine, apprentissage profond, et leurs applications

2.1. Introduction à l'intelligence artificielle

La section "Introduction à l'Intelligence Artificielle" offre un aperçu de la prééminence de l'IA et de son impact dans la société. Elle met en avant les vastes applications de l'IA dans divers domaines et son potentiel pour façonner positivement l'avenir (ou parfois susciter des inquiétudes). Dans ce contexte, des entreprises telles que Google, Meta ou OpenAI ont été pionnières, développant des algorithmes innovants tels que FaceNet, DeepDream, DALL-E et AlphaGo. Il est aujourd'hui anticipé que l'IA contribuera à hauteur d'environ 13 milliards de dollars à l'économie mondiale d'ici 2030.

L'expression "intelligence artificielle" englobe un large éventail de techniques et méthodes. Son objectif principal est de simuler les capacités de notre cerveau. Cela passe parfois par l'analyse de données, l'identification de modèles et la prise de décisions "intelligentes" ou l'accomplissement de tâches spécifiques. Dans ce résumé, nous présentons à nouveau, très brièvement diverses branches de l'IA.

L'implémentation de solutions "IA" nécessite une méthodologie structurée comprenant par exemple la préparation des données, la création de modèles, la conception de systèmes et leur déploiement. La phase de préparation des données est souvent la plus exigeante en ressources et la plus critique pour évaluer la faisabilité du projet. Dans ce chapitre l'apprentissage machine (*Machine Learning*) est détaillé en présentant les différences entre les apprentissages supervisé, non-supervisé et par renforcement. Puis, les notions de neurone et d'apprentissage profond (*Deep Learning*) sont présentés en partie en présentant les concepts mathématiques, que nous ne représenterons pas ici. Il est important de souligner que les méthodes d'apprentissage présentées ici sont pour la plupart assez datées afin de rester dans une démarche d'introduction au domaine. Cependant, le domaine de l'IA n'est pas constitué d'une suite d'outils figée et est en constante évolution avec une recherche très active et nous invitons le lecteur à se renseigner sur les dernières avancées.

2.2. Apprentissage machine

Depuis les années 1990, l'apprentissage automatique (ML) joue un rôle significatif dans la météorologie spatiale. Ce changement a été notamment mis en lumière par la revue complète du Dr Enrico Camporeale ([Camporeale, 2019](#)). Les avancées technologiques, la disponibilité accrue des données et l'amélioration du matériel ont relancé l'intérêt pour ce domaine et ont permis le développement d'approches qui étaient précédemment jugées inefficaces. La disponibilité de jeux de données étendus, librement accessibles, couvrant parfois plusieurs cycles solaires, est cruciale pour le développement de modèles réussis. Et aujourd'hui, de nombreuses nouvelles ressources et outils deviennent disponibles pour les applications ML (des bibliothèques, des jeux de données pré-traités, etc. - dont une liste, non-exhaustive, se trouve dans ce manuscrit).

Les modèles entièrement basés sur la physique théorique (ou la résolution d'équation connues) sont qualifiés de modèles *white-box*. Ils sont "transparents" et permettent à la communauté de maîtriser et comprendre ce qui est fait. D'autres modèles dits *black-box* correspondent, eux, à des modèles exclusivement basés sur la donnée et sur les relations entre jeux de données. Ils sont considérés comme des boîtes noires car ils ne donnent pas ou peu d'informations sur le fonctionnement interne. C'est le cas par exemple des réseaux de neurones pour lesquels une branche de la recherche s'est développée afin de pouvoir extraire l'information et le fonctionnement (XAI ou

explainable artificial intelligence). La météorologie spatiale est particulièrement propice à la mise en œuvre de modèles dits *grey-box*, qui mélangent des méthodes dites *white box* et des méthodes dites *black-box*). C'est le cas des *physics-informed neural networks* qui intègrent certains comportements physiques connus afin d'orienter les réseaux de neurones dans la bonne direction.

Les méthodes d'apprentissage machine se divisent en général en trois catégories: l'apprentissage supervisé, non supervisé, et par renforcement.

- L'apprentissage supervisé, guidé par un 'superviseur' fournissant des étiquettes, utilise des données annotées pour générer des modèles capables de classer des données non étiquetées. Il englobe la régression, pour prédire des résultats continus, et la classification, pour mapper des entrées à des classes spécifiques. L'apprentissage profond, se concentrant sur les réseaux de neurones multicouches, est un sous-domaine spécialisé. La régression est utilisée pour des tâches comme la prévision météorologique, tandis que la classification gère des problèmes comme la détection de spam. L'apprentissage supervisé a trouvé de nombreuses applications en météorologie spatiale, avec des avancées dans la prédiction des indices géomagnétiques et la modélisation des propriétés du vent solaire.
- L'apprentissage non supervisé analyse des jeux de données non étiquetés pour découvrir des motifs cachés, le rendant approprié pour l'analyse exploratoire de données et la segmentation des clients. Il inclut des algorithmes de clustering, comme k-means, qui regroupent des points de données similaires, et des techniques de réduction de dimensionnalité, comme l'analyse en composantes principales, qui réduisent le nombre de caractéristiques tout en maintenant l'intégrité des données. Bien que l'apprentissage non supervisé soit moins courant dans les applications de météorologie spatiale, il est utilisé pour classifier des régions plasmatiques et identifier des événements météorologiques spatiaux spécifiques.
- L'apprentissage par renforcement mappe des séquences d'entrées à des sorties, se concentrant sur des états au sein d'un environnement et des actions disponibles, avec pour objectif d'atteindre un état désiré. L'agent d'apprentissage explore des paires état-action et reçoit des récompenses, ajustant ses actions en conséquence, comme illustré dans un scénario de blackjack où l'agent apprend des stratégies de jeu optimales. L'apprentissage par renforcement est distinct en fournissant des retours basés sur des signaux de renforcement, similaires à l'apprentissage humain.

Ces trois modèles d'apprentissage automatique englobent la plupart des algorithmes existants. L'apprentissage supervisé nécessite des données étiquetées, l'apprentissage non supervisé travaille avec des données non étiquetées pour trouver des regroupements de données, et l'apprentissage par renforcement apprend à travers des récompenses et des punitions pour atteindre des objectifs. On peut mentionner de plus l'apprentissage semi-supervisé qui combine des éléments des apprentissages supervisé et non supervisé, abordant des défis comme la disponibilité limitée de données étiquetées.

2.3. Apprentissage profond

Les réseaux de neurones artificiels sont la pierre angulaire de l'apprentissage profond. Ils sont constitués de plusieurs couches de neurones interconnectés. Chaque neurone est un nœud mathématique qui prend des entrées, effectue une transformation linéaire (multiplication des entrées par des poids et addition d'un biais) et applique une fonction d'activation non linéaire à la sortie. Les couches peuvent être de différents types : une couche d'entrée pour les données initiales, des couches cachées pour l'extraction des caractéristiques et une couche de sortie pour la prédiction.

Les fonctions d'activation mentionnées sont utilisées pour introduire de la non-linéarité dans le réseau neuronal. Elles permettent aux réseaux de neurones d'apprendre des relations complexes

dans les données, entre entrées et sorties. On compte parmi elles la fonction sigmoïde, adaptée à la classification binaire mais limitée par des points de données proches ; les fonctions linéaires ; la tangente hyperbolique ; ou encore les fonctions ReLU, LeakyReLU, SiLU et Softmax que nous avons décrit en détail dans la thèse. Le choix d'une fonction d'activation dépend de la tâche, du jeu de données et de l'architecture, chacune ayant des propriétés distinctes influençant la performance du réseau.

Le passage des données de l'entrée vers la sortie, en passant par tous les neurones et donc toutes les transformations, est appelée la *propagation avant*. Chaque couche contenant des neurones effectue des calculs sur les données d'entrée à l'aide de ses poids et de sa fonction d'activation respective pour générer des sorties qui serviront d'entrées à la couche suivante. Lorsqu'on parle de *réseaux de neurones entièrement connectés* (FCNN), la sortie de chaque neurone d'une couche est envoyée dans tous les neurones de la couche suivante. Ainsi pour une couche de 10 neurones, nous obtenons 11 sorties (les 10 des 10 neurones et la donnée "1" pour qui sert de biais). Celles-ci seront envoyées dans la couche suivante. Si cette dernière est faite de 5 neurones, il y aura 5 combinaisons linéaires des 11 données soit 55 poids. L'objectif du réseau est de trouver l'ensemble des poids de l'ensemble du réseau tel qu'il puisse obtenir une certaine sortie.

Lors du processus d'apprentissage, le réseau de neurones est exposé à des données d'entraînement qui contiennent des exemples avec des entrées et des sorties attendues (dans le cadre de l'apprentissage supervisé). Il apprend à partir de ces données en ajustant ses poids et ses biais. L'objectif est pour lui de minimiser une fonction de perte, qui mesure la différence entre les sorties réelles et les sorties attendues.

La rétropropagation est l'algorithme clé utilisé pour entraîner le réseau. Il fonctionne en calculant les gradients de la fonction de perte par rapport aux poids du réseau, puis en ajustant ces poids dans la direction qui minimise la fonction de perte. Ce processus est répété sur de nombreuses itérations avec différents exemples d'entraînement jusqu'à ce que le réseau atteigne des performances acceptables. Une fois entraîné, il peut être utilisé pour faire des prédictions sur de nouvelles données en appliquant simplement les opérations apprises pendant l'entraînement. Ces prédictions peuvent être utilisées dans une variété d'applications, telles que la classification d'images, la traduction automatique, la reconnaissance vocale, etc.

En résumé, l'apprentissage profond fonctionne en utilisant des réseaux de neurones artificiels pour apprendre à partir de données en ajustant leurs poids internes via la rétropropagation, afin de réaliser des tâches complexes de prédiction et de classification.

Avec ces informations nous avons la base pour comprendre comment fonctionne un réseau de neurone. Par la suite, nous présentons ici certains des concepts qui gravitent autour du fonctionnement d'un réseau de neurone

- **Optimiseurs:** L'optimiseur fonctionne en utilisant les gradients de la fonction de perte par rapport aux paramètres du modèle (c'est-à-dire les dérivées partielles de la fonction de perte par rapport à chaque paramètre) pour mettre à jour ces paramètres dans une direction qui minimise la fonction de perte. Différents optimiseurs utilisent des stratégies différentes pour mettre à jour les paramètres, ce qui peut influencer la vitesse de convergence et la stabilité de l'apprentissage. Le choix d'un optimiseur se fait souvent par essai et erreur, s'appuyant sur des études empiriques. Parmi les optimiseurs clés, on retrouve par exemple SGD, qui actualise les gradients de façon aléatoire, réduisant la courbe de perte mais de manière erratique; Momentum, qui améliore SGD par l'ajout d'un terme de moment, accélérant la convergence; AdaGrad, qui adapte les taux d'apprentissage mais peut faire face à une diminution de ceux-ci; RMSProp, qui ajuste AdaGrad et pénalise les paramètres os-

cillants; et ADAM, qui combine les avantages de Momentum et AdaGrad pour des mises à jour adaptatives.

- **Schedulers:** Les schedulers planifient dynamiquement le taux d'apprentissage (par exemple, le pas qui est fait dans la direction de plus grande pente) pendant l'entraînement, optimisant la vitesse de convergence du modèle et assurant robustesse et meilleure généralisation. Parmi eux, on trouve StepLR, qui ajuste le taux d'apprentissage à intervalles réguliers ; ReduceLRonPlateau, qui réagit à un plateau dans l'amélioration d'une métrique surveillée ; CosineAnnealingLR, utilisant un calendrier de recuit cosinus ; OneCycleLR, avec un calendrier cyclique et une phase d'échauffement ; et ExponentialLR, réduisant le taux d'apprentissage de manière exponentielle. Chacun propose des stratégies uniques pour adapter le taux d'apprentissage et naviguer dans des paysages d'optimisation complexes, aidant finalement à éviter les minima locaux et à obtenir une meilleure convergence.
- **Les métriques:** Une métrique est une mesure quantitative utilisée pour évaluer les performances d'un modèle par rapport à ses prédictions sur un ensemble de données. Elle fournit une indication de la précision, de la robustesse ou d'autres aspects de la performance du modèle dans la résolution d'un problème spécifique. Par exemple, la *précision* mesure la proportion de prédictions correctes parmi l'ensemble des prédictions du modèle, tandis que le *rappel* (ou *recall*) quantifie la capacité du modèle à identifier correctement tous les cas positifs dans un ensemble de données. Dans le cas de la régression, des métriques telles que l'erreur quadratique moyenne (MSE) et sa racine (RMSE) ou l'erreur absolue moyenne (MAE) fournissent une mesure de la précision des prédictions numériques du modèle
- **La séparation en entraînement, validation et test:** Diviser les données en trois ensembles: d'entraînement, de validation et de test est une étape cruciale de la préparation des données, aidant à observer des phénomènes tels que le surapprentissage et à améliorer l'entraînement du modèle. L'ensemble d'entraînement sert à l'apprentissage et à l'ajustement des paramètres du modèle, l'ensemble de validation évalue les performances et permet un réglage fin de l'algorithme sans influencer les poids du modèle, et l'ensemble de test offre une évaluation finale impartiale des performances sur des données inédites. Bien qu'il n'y ait pas de taille de division universelle, une pratique courante est 60% pour l'entraînement, 20% pour la validation, et 20% pour les tests, visant un équilibre dans l'allocation des données. Il est essentiel de maintenir l'équilibre des jeux de données, en préservant les informations pertinentes telles que les valeurs aberrantes et les distributions dans tous les sous-ensembles pour une évaluation et un entraînement efficaces du modèle.
- **Le surapprentissage:** Celui-ci se produit lorsque l'algorithme a tendance à mémoriser l'ensemble d'entraînement au lieu d'apprendre les motifs sous-jacents. Cela peut être dû à un modèle trop complexe, à un petit ensemble d'entraînement ou à des données non pertinentes. Pour détecter le surapprentissage, nous traçons souvent les courbes de perte pendant l'entraînement. Ces courbes montrent l'évolution de la fonction de coût pendant l'entraînement et la validation au fil du temps. La fonction de coût d'entraînement mesure la qualité de l'ajustement du modèle aux données d'entraînement, tandis que celle de validation évalue sa performance sur des données inconnues. Comprendre ces courbes aide à diagnostiquer les problèmes et guide l'optimisation du modèle. Par exemple, si la courbe d'entraînement continue de descendre et que la courbe de validation remonte, nous sommes en général face à un surapprentissage (bon sur le jeu d'entraînement et mauvais sur les données nouvelles). Pour le résoudre, des techniques telles que l'arrêt précoce, la régularisation et le dropout peuvent être utilisées.
- **Les techniques de régularisation,** telles que le dropout, la régularisation L1 (Lasso) et la régularisation L2 (Ridge), sont essentielles pour prévenir le surajustement et améliorer la généralisation. Le dropout introduit du bruit en désactivant certaines sorties de neurones pendant l'entraînement, la régularisation L1 favorise la parcimonie en conduisant certains

poids vers zéro, et la régularisation L2 encourage des poids plus petits sans les forcer à devenir nuls. De plus, l'arrêt précoce et les techniques d'augmentation de données sont des outils essentiels. L'arrêt précoce interrompt l'entraînement lorsque les performances se détériorent, et la data augmentation accroît la diversité de l'ensemble de données par des transformations telles que la rotation ou le changement d'échelle dans le cas d'images. Ces techniques améliorent collectivement les performances du modèle, réduisent le surapprentissage et renforcent la généralisation du modèle à des nouvelles données.

Enfin, nous pouvons conclure cette section en présentant les étapes clés lors du design d'un réseau de neurones : l'analyse de la donnée, le prétraitement de cette même donnée, et le choix de l'architecture. Suivant l'architecture l'ensemble des points ci-dessus seront ensuite choisis. C'est pourquoi nous terminerons cette section en présentant les deux architectures utilisées dans cette thèse: le réseau de neurones entièrement connecté (FCNN) et le réseau de neurones convolutionnel temporel (TCN).

L'analyse des données en apprentissage automatique garantit que l'ensemble des données est adapté à l'entraînement. Elle consiste à vérifier les erreurs et les problèmes tels qu'un manque de données, du bruit, une corrélation excessive, une corruption des données et des erreurs d'étiquetage. Un ensemble de données de haute qualité doit être flexible, précis et bien structuré. Sa robustesse peut être évaluée grâce à des techniques telles que la visualisation, l'analyse de corrélation, l'analyse en composantes principales (ACP), les analyses statistiques ou l'évaluation du bruit. Une chaîne de traitement des données est souvent établie pour raffiner systématiquement les données brutes pour les tâches d'apprentissage automatique. Enfin, la collaboration avec des experts du domaine est essentielle pour obtenir des informations difficiles à décoder.

Le prétraitement des données, par la suite, est important pour le succès de l'apprentissage automatique. Il peut impliquer diverses tâches qui doivent être en adéquation avec l'architecture du modèle choisi pour garantir des performances optimales, comme par exemple :

- Le nettoyage des données : Supprimer les valeurs aberrantes, corriger les incohérences et garantir l'uniformité.
- La sélection et extraction des caractéristiques : Choisir les caractéristiques pertinentes et réduire la dimensionnalité lorsque cela est nécessaire.
- La gestion des valeurs manquantes : Imputer les données manquantes avec soin.
- La mise à l'échelle et normalisation : Rendre les données comparables en échelle.
- La prise en compte des aspects temporels pour les séries temporelles.
- La gestion des déséquilibres de classe dans les données (une classe sureprésentée par rapport à une autre).
- La division des données pour l'entraînement et la validation comme nous l'avons vu.
- La transformation des données catégorielles : Convertir les données catégorielles selon les besoins.
- La compréhension des distributions : Adapter les données aux hypothèses du modèle.

Et enfin, il faut avoir en tête une architecture qui sera adaptée à notre problématique. Le choix de la bonne architecture d'apprentissage automatique peut être un défi, et il n'y a pas de solution universelle. Par exemple, les réseaux de neurones convolutionnels (Convolutional Neural Networks - CNNs) fonctionnent bien pour les images. Les réseaux de neurones récurrents (Recurrent Neural Networks - RNNs), eux, tels que les unités récurrentes à portes (Gated Recurrent Units - GRUs) et les réseaux de mémoire à court et long terme (Long Short-Term Memory - LSTMs) sont adaptés aux séquences. Les FCNN sont adaptés à la classification et les réseaux antagonistes

génératifs (Generative Adversarial Networks - GANs) aident à générer des données synthétiques, ce qui est utile pour traiter les données manquantes. Il s'agit là d'exemples et non de conseils d'applications, chaque problématique étant unique.

Ci-dessous les deux architectures qui ont été utilisées pour obtenir des résultats dans ce manuscrit:

- **FCNN** : Dans un *fully-connected neural network*, chaque neurone d'une couche est connecté à chaque neurone de la couche suivante. Il est couramment utilisé pour des tâches de classification et de régression, traitant les données d'entrée à travers une série de couches cachées pour effectuer des prédictions.
- **TCN** : Un TCN est un type d'architecture de réseau neuronal spécialement conçu pour les données séquentielles, telles que les séries temporelles. Contrairement aux RNN traditionnels, les TCN utilisent des couches de convolution pour capturer les motifs et les dépendances dans les données sur plusieurs pas de temps simultanément, en 1-D. Cette approche de traitement parallèle permet aux TCN de modéliser efficacement les dépendances à long terme et s'est avérée plus performante que certains réseaux récurrents comme les LSTM dans la prévision de séries temporelles.

2.4. Librairies et matériel

Pour mettre en œuvre avec succès l'apprentissage automatique et l'apprentissage profond, nous avons besoin d'une combinaison de puissance de calcul et de bibliothèques de haut niveau. Le choix du matériel approprié, tel que les CPU et les GPU, est crucial. Les CPU servent de processeurs centraux des ordinateurs et conviennent bien aux tâches nécessitant une latence faible et des performances par cœur élevées. Les GPU, ou unités de traitement graphique, sont spécialisés dans le traitement parallèle et excellent dans la gestion de plusieurs opérations plus simples simultanément.

Les modèles d'apprentissage profond font souvent intervenir des GPU pour accélérer les tâches. Les bibliothèques telles que TensorFlow et PyTorch sont optimisées pour exploiter les capacités des GPU. PyTorch est préféré pour la recherche en raison de sa flexibilité et de sa facilité d'expérimentation, grâce à son architecture dynamique de graphes computationnels. TensorFlow est mieux adapté à l'industrie en raison de sa robustesse et de sa scalabilité, avec une architecture de graphe statique idéale pour le déploiement efficace des modèles en production. Bien sûr, ce choix dépend également de la problématique et ces deux librairies sont interchangeables.

PyTorch Lightning est une extension de PyTorch qui simplifie le processus de développement et de déploiement des modèles. Il offre une structure de code modulaire et standardisée, ce qui permet aux chercheurs de se concentrer davantage sur la conception de modèles plutôt que sur la gestion des détails d'implémentation. Cela en fait un choix attrayant pour la recherche, car il combine la flexibilité de PyTorch avec une meilleure organisation du code et des outils intégrés pour la répétabilité des expériences.

Dans cette étude, nous avons utilisé Pytorch-Lightning. Il offre une abstraction efficace pour le développement de modèles, tout en capitalisant sur les avantages de PyTorch en termes de flexibilité et d'exploration rapide des idées.

Chapitre 3: Problématique, analyse et préparation de la donnée

La précipitation des électrons dans l'ionosphère est entraînée par divers mécanismes. Les électrons précipitent lorsqu'ils entrent dans le cône de perte, se déplaçant le long des lignes de champ magnétique et subissant souvent des accélérations complexes et interagissant fortement avec divers facteurs tels que les ondes de plasma, la reconnexion magnétique et les courants magnétosphériques. Certaines des particules précipitantes ont été piégées dans diverses régions magnétosphériques là où d'autres proviennent directement du soleil. Dans ces enchaînements d'évènements, les particules transportent du courant, transfèrent de l'énergie et, finalement, précipitent dans l'ionosphère.

La précipitation de particules est essentielle pour les modèles de circulation globale tels que GITM, TIE-GCM, et WAM-IPE, et indique souvent des événements électriques intenses dans l'ionosphère, ayant divers impacts comme la charge de surface que nous avons présenté. Ainsi, elles représentent un risque majeur pour les satellites LEO en orbite polaire. C'est pourquoi plusieurs modèles de précipitation de particules aux pôles existent, notamment Hardy (1985), Fuller-Rowell (1987), et OVATION (Newell et al., 2014). Les deux premiers modèles produisent des distributions spatiales bidimensionnelles (2D) d'énergie et de flux de particules en fonction de paramètres préalablement définis. Ils se limitent généralement à l'hémisphère nord et sont contraints par les paramètres d'entrée. OVATION, quant à lui, a révolutionné la modélisation de la précipitation en utilisant exclusivement les paramètres du vent solaire comme entrée. Il est devenu l'outil de référence pour localiser l'ovale auroral et quantifier son intensité (Newell, 2002) et est devenu l'un des produits les plus téléchargés du Centre de Prévision Météorologique Spatiale (SWPC) de la National Oceanic and Atmospheric Administration (NOAA) source (McGranaghan, 2021).

Par la suite, il a été observé que l'efficacité de ces modèles dépendait largement des paramètres d'entrée et des choix de représentation, mais que ces choix étaient souvent incapables de capturer la complexité du système. Autrement dit, les modèles actuels ne fournissent pas suffisamment d'informations cruciales en entrées et ne n'intègrent pas pleinement la physique sous-jacente. McGranaghan et ses collègues ont abordé ces problèmes en 2021 avec le modèle PrecipNet, utilisant l'apprentissage automatique pour améliorer la sélection et la représentation des paramètres d'entrée.

Notre travail réalisé dans le cadre d'un projet CIFRE avec SpaceAblen, s'est appuyé sur les recherches de [McGranaghan et al. \(2021\)](#), et a deux objectifs majeurs. Le premier est d'approfondir la compréhension de la précipitation d'électrons en LEO et du système magnétosphère-ionosphère via une modélisation robuste du flux d'énergie d'électrons. Le second objectif, en harmonie avec les ambitions de SpaceAble, est le développement d'un modèle qui soit adaptable et économique, privilégiant la praticité, l'adaptabilité aux nouvelles données, l'interprétabilité, et s'alignant avec la vision stratégique de l'entreprise.

Cette thèse explore donc les enjeux de l'apprentissage automatique dans la modélisation Soleil-Terre, s'interrogeant sur la qualité des données, les défis spatio-temporels et la possibilité de modéliser des phénomènes complexes. Le chapitre que nous résumons ici présente une analyse des données, incluant des informations hors magnétosphère et des mesures de précipitation de particules de plus de 50 ans (DMSP), et détaille nos choix de prétraitement.

Description des données

A ce stade de l'étude, les données d'entrées envisagées sont les données des satellites ACE, ainsi que les données de la base de données OMNIWeb de la NASA. Les données de sortie, qui serviront d'étiquettes pour entraîner l'algorithme, sont les données des satellites du programme DMSP.

1. Advanced Composition Explorer (ACE) :

- **Localisation et Mission :** Situé au point de Lagrange L1, il mesure les champs magnétiques et les particules dans l'espace, détectant les éruptions de particules du Soleil qui peuvent impacter l'espace proche de la Terre.
- **Instruments :** Neuf instruments au total. Six spectromètres à haute résolution et trois instruments (SWEPAM, EPAM, MAG) fournissant le contexte du milieu interplanétaire.
- **Utilisation des Données :** De notre côté, nous nous concentrons spécifiquement sur les données du vent solaire (vitesse, densité, pression) et sur les composants du champ magnétique interplanétaire provenant respectivement des instruments SWEPAM et MAG.

2. Defense Meteorological Satellite Program (DMSP) :

- **Satellites :** Les satellites DMSP sont en orbite polaire, synchrones au soleil. Ils ont pour mission principale d'observer la météorologie de la troposphère et pour mission secondaire de surveiller l'environnement spatial.
- **Instruments SSJ/4 et SSJ/5 :** Ces instruments mesurent l'énergie, la masse, et la quantité de mouvement des particules chargées dans le champ magnétique terrestre, fournissant des données sur les particules de basse énergie qui causent des aurores et d'autres phénomènes de haute latitude. Ils enregistrent les flux de particules électroniques et ioniques entre 30 eV et 30 KeV chaque seconde, ainsi que l'éphéméride du satellite et les coordonnées magnétiques.
- **Utilisation des Données :** Les données sont utilisées pour comprendre l'ionosphère polaire et de haute latitude et pour améliorer les systèmes de communication, de surveillance et de détection. Notre étude se concentre principalement sur le flux *total* énergétique des électrons.

3. High-Resolution OMNIWeb :

- **Fonctionnalité :** OMNIWeb est un outil en ligne géré par le Space Physics Data Facility de la NASA qui permet aux utilisateurs d'accéder à un ensemble de données et de les afficher.
- **Données utilisées:** Au sein d'OMNI, les données que nous avons utilisées sont les données à haute résolution (1-min) de champ magnétique et de plasma.

Analyse des données

Notre première analyse s'est portée sur les mesures de niveau 2 des instruments SWEPAM et MAG du satellite ACE de 1998 à 2021, ce qui a fait l'objet d'une publication en 2022 ([Bouriat et al., 2022](#)). L'objectif était à la fois d'analyser le jeu de données sous le prisme de l'utilisation des modèles d'IA et à la fois de présenter les pratiques "saines" d'analyse de données dans ce contexte. Pour chaque jeu de données nous avons réalisé des histogrammes, des analyses des données manquantes et extrêmes, des mesures de corrélations et d'autocorrélations ou encore des analyses en composantes principales.

Nos résultats concernant les données de ACE ont indiqué des distributions de paramètres non uniformes, un impact important du cycle solaire sur les valeurs, un grand nombre de données manquantes et disparates et la présence de relations non linéaires. En raison du grand nombre de défis

posés par ces données (voir les conclusions de l'article de [Bouriat et al. \(2022\)](#)), en particulier le nombre significatif de valeurs manquantes, nous avons décidé de nous orienter essentiellement vers le jeu de données OMNIWeb, qui s'est avéré plus adapté à notre problématique.

Concernant OMNIWeb, le manque de fiabilité des données provient principalement des interactions entre les flux de vent solaire lents et rapides. Une étude de [Vokhmyanin et al. \(2019\)](#) comparant les données OMNI aux mesures par satellite a révélé que 42% montraient un excellent accord, 33% une bonne cohérence, 10% la bonne tendance mais des valeurs inexactes, et 15% très inexactes. L'étude a identifié plusieurs facteurs physiques introduisant des erreurs et a indiqué que les données mesurées loin de la ligne Soleil-Terre étaient souvent inexactes. De plus, les données OMNI ont montré des pourcentages variables de données manquantes pour différents composants, que nous avons abordés en implémentant une méthode d'interpolation, spécifiquement le Polynôme d'Interpolation Hermite Cubique Par Morceaux (pchip). Cette méthode a été appliquée à deux reprises, comblant respectivement des trous de tailles 1 et jusqu'à 4, réduisant significativement les pourcentages de données manquantes. Les deux jeux de données contenant ces trous comblés n'ont pas été utilisés pour notre étude mais sont en cours d'utilisation pour améliorer les résultats obtenus dans cette thèse.

Pour notre étude, nous avons retenu pour l'entrée des variables spécifiques du jeu de données de haute résolution (5 minutes) de OMNIWeb: les composants de l'IMF, la vitesse du vent solaire, la densité, la pression et plusieurs indices (AL, AU, et SYM-H) qui s'étaient montrés pertinents dans de précédentes études. A noter que bien que nous ayons choisi les données OMNIweb plutôt que celles de ACE, notre code est, par construction, adaptable et pourra prendre n'importe quelle nouvelles données en entrée.

Concernant DMSP, les principales préoccupations étaient la qualité, les erreurs de mesure, le bruit et l'intercalibration des instruments entre les différents satellites. Une référence clé dans ce domaine est le travail de [Redmon et al. \(2017\)](#), qui a abordé ces sujets et établi une base de données robuste et fiable couvrant plus de 30 ans à partir des engins spatiaux DMSP. Cette base de données présente des limitations connues (telles que des imprécisions pour les canaux de basse énergie) mais sert d'excellente base pour des analyses plus approfondies.

Une étape critique se situe dans la gestion de la distribution non uniforme des données utilisée pour l'entraînement, qui révèle de nombreuses informations dont des dépendances aux cycles solaires. Des stratégies telles que la transformation logarithmique des caractéristiques du vent solaire ont été employées pour stabiliser la variance et améliorer les performances du modèle à venir. De plus, nous avons combiné les régions polaires afin d'obtenir plus de données et donc d'améliorer le déséquilibre et la variabilité dans la distribution spatiale des données.

Enfin, les observations visuelles des données sur le flux d'énergie des électrons ont révélé des motifs indiquant de nombreuses valeurs aberrantes et des incohérences d'un satellite à un autre. Ces observations suggèrent des limitations potentielles des instruments et des disparités de recalibrage. Pour traiter les valeurs aberrantes, nous avons évalué les méthodes z-score et InterQuartile Range (IQR) et finalement adopté l'approche z-score. Cette méthode s'est avérée efficace pour identifier et éliminer moins de 1% des points de données qui déviaient significativement de la moyenne, améliorant ainsi la fiabilité de l'ensemble de données.

Résumé et pré-traitement final

Une fois notre analyse menée à bien, voici un résumé du déroulé de notre étude et les pré-traitements associés :

1. **Exploration de PrecipNet** : Initialement, notre attention se porte sur PrecipNet, l'algorithme de [McGranaghan et al. \(2021\)](#), en raison de ses résultats pertinents. Pour ce faire, nous préparons notre ensemble de données spécifiquement pour le FCNN. Nous n'utilisons pas les jeux de données interpolées et nous limitons les sorties aux valeurs DMSP utilisées par PrecipNet (un peu plus d'un million d'échantillons). Nos entrées ne sont pas strictement identiques à celles de PrecipNet (puisque ce dernier inclut également $F_{10.7}$ et l'indice Polar Cap) mais notre objectif principal est de tenter d'obtenir des résultats équivalents afin de comprendre les limitations qui y sont associées. Notre reproduction légèrement éloignée de PrecipNet est notée PrecipNet-R pour la suite.
2. **Introduction de notre FCNN** : Pour approfondir et rectifier certains aspects de PrecipNet-R, nous appliquons notre FCNN personnalisé. Cette fois, le modèle est entraîné sur l'ensemble des données de DMSP que nous avons, soit plus de 3 millions d'échantillons. Une comparaison complète de nos résultats avec ceux de PrecipNet et OVATION Prime suit.
3. **Amélioration et Prévision** : Par la suite, en excluant les données à T_0 et en ne retenant que les données passées (historiques), nous établissons la pertinence et la faisabilité de futurs travaux de recherche concernant la prévision à court terme des électrons précipitants.
4. **Traitement des Limitations avec le TCN** : Cependant, PrecipNet et notre FCNN présentent tous deux des limitations. En particulier, la modélisation des électrons précipitants repose sur l'utilisation de points de données historiques choisis de manière arbitraire. Il est raisonnable de penser que les mouvements dynamiques et complexes du vent solaire contredisent l'idée que l'information pertinente serait confinée à des points temporels fixes dans le passé. Pour y remédier, dans la quatrième phase, nous introduisons un TCN. Contrairement à la sélection arbitraire de données historiques par PrecipNet, notre approche TCN intègre toutes les données d'un intervalle passé spécifié. Il est important de noter que, en raison de contraintes de temps et de calcul, nous limitons notre ensemble de données pour cette section aux données DMSP contenues dans les archives de [McGranaghan et al. \(2021\)](#) (à savoir un peu plus d'un million d'échantillons).

Chapitre 4: Implémentation des algorithmes, itérations et résultats

Exploration de PrecipNet

Nous avons d'abord étudié le modèle "PrecipNet", un FCNN conçu pour modéliser les électrons précipités mesurés par les satellites DMSP ([McGranaghan et al., 2021](#)). Ce modèle comporte 72 paramètres d'entrée, 1000 époques, un "batch" de 32 768 et un taux d'apprentissage de 0,001, avec une architecture de réseau comprenant 8 couches cachées. N'y ayant pas accès, nous avons décidé dans un premier temps de le reproduire le plus fidèlement possible, afin d'avoir une base de travail et de comparaison à nos résultats. PrecipNet-R est le nom que nous avons donné à la reproduction, en interne, de PrecipNet. Les résultats de celle-ci sur le jeu de validation (F16, 2010) diffèrent de 0.54% par rapport à ceux de PrecipNet. Nous avons donc considéré que notre reproduction était suffisamment fidèle à l'original pour pouvoir tirer des conclusions sur PrecipNet de nos observations de PrecipNet-R.

Nous avons réexaminé les jeux de données utilisés, et avons remarqué que le modèle a été entraîné sur 1 890 579 points de données et validé sur 55 210 points de données de 2010 du satellite F16. En revanche, aucun ensemble de test n'a été défini. L'absence de celui-ci et le choix d'un satellite spécifique non représentatif (F16) sur une année, elle-même non représentative des tendances (2010) pour la validation sont des points de discussion importants. Nous avons de notre côté tenté d'éviter cet écueil. En effet, nous pensons qu'un ensemble de test pour une évaluation non biaisée d'un modèle est nécessaire. Le risque sous-jacent est que l'ajustement répétitif des hyperparamètres basé essentiellement sur les performances du modèle sur un jeu de validation ne présuppose pas de l'efficacité du modèle sur un jeu de données inconnu. De plus, comme nous le disions un peu plus tôt, notre analyse des données a révélé que l'année 2010 du satellite F16 était peu représentative de l'ensemble des données à disposition. Il s'agit d'une période quelque peu unique pour l'activité solaire. Cela soulève des questions sur les performances exactes de PrecipNet. L'utilisation de données non représentatives pour la validation entraîne un écart significatif entre les courbes de perte de validation et d'entraînement qui se constate dans leurs figures. PrecipNet demeure en revanche largement plus performant que l'état de l'art (OVATION Prime).

Afin de comprendre et préparer l'implémentation de nos propres FCNNs, nous avons effectué quelques tests en faisant varier le nombre et la taille des couches cachées, en triant les entrées utilisées ou encore en changeant le choix des jeux d'entraînement, validation et test.

Pour résumer cette section, voici ce qui a été identifié :

- Pour rappel, nous avons reproduit PrecipNet en interne avec la même architecture et les mêmes données DMSP (étiquettes). Seules les données d'entrée différaient (OMNI) car nous souhaitions éviter une interpolation linéaire faite dans [McGranaghan et al. \(2021\)](#) pour préserver l'intégrité des données originales. En conséquence, nous avons légèrement moins de points d'entraînement, mais nous approchons efficacement des résultats de PrecipNet (moins de 0,6% de différence). Nous avons nommé cet algorithme PrecipNet-R.
- Utiliser les données de 2010 du satellite F16 pour la validation semble étrange car c'est le seul satellite dont les données sont difficiles à comparer avec les autres. L'utilisation d'un échantillonnage aléatoire corrige la divergence entre les courbes de validation et d'entraînement observée dans l'article de McGranaghan.
- Un plus grand nombre de couches ajoute de la complexité au réseau, provoquant un sur-apprentissage même avec un nombre de paramètres constant. Par conséquent, minimiser le nombre de couches est essentiel pour éviter le sur-apprentissage et réduire le temps de calcul et les exigences en ressources.
- Avec une architecture contenant dix fois moins de paramètres et un temps de calcul presque réduit du même facteur, nous obtenons des résultats qui ne sont que 2,5% moins bons qu'auparavant. Une architecture simple est préférable, surtout pour les applications industrielles.
- En adoptant l'architecture de PrecipNet-R et en ne considérant que les paramètres de position et de temps comme entrées, nous atteignons tout de même 60% des capacités prédictives initiales. Par conséquent, la majorité des informations sur lesquelles l'algorithme se concentre sont la position et la forme de l'ovale auroral. C'est là que se dirige une grande partie de sa capacité de calcul.
- Malheureusement, isoler et modéliser à travers un réseau neuronal simple (deux couches de taille 4) une région spécifique de l'ovale (une latitude et longitude magnétique fixes) révèle rapidement ses limites : pas assez de données pour étendre la méthode à l'ensemble de l'ovale, sur-apprentissage rapide et difficultés à appliquer l'apprentissage profond.

Sur la base de ces constatations et de nos connaissances existantes, nous visons à développer un

réseau neuronal en gardant à l'esprit les concepts suivants :

- Se concentrer sur une architecture simple, en gardant à l'esprit l'application industrielle ultérieure, et éviter le sur-apprentissage.
- Utiliser l'échantillonnage aléatoire comme méthode de division de l'ensemble de données, plutôt qu'un choix fixe d'année et de satellite.
- Concevoir des stratégies pour prédire les valeurs extrêmes sans que l'algorithme se concentre excessivement sur les valeurs moyennes, qui sont facilement prédictibles en utilisant uniquement les données de position et de temps.
- Utiliser notre ensemble de données DMSP, comprenant plus de 3 millions d'échantillons, en espérant que cela améliorera la capacité prédictive.
- Mettre en place un affichage plus complet pour suivre la qualité de modélisation de l'algorithme.

Un dernier problème que nous n'avons pas encore abordé concerne le délai dynamique entre l'entrée et la sortie. En effet, nous avons conservé les points de données historiquement pertinents comme le recommande [McGranaghan et al. \(2021\)](#), mais rien ne nous assure que toutes les informations pertinentes existent réellement à ces moments dans le passé. Ce problème n'a que deux solutions : inclure encore plus de points de données du passé (ce qui augmenterait exponentiellement le nombre de caractéristiques et, par conséquent, la complexité du modèle) ou changer l'algorithme. Nous reviendrons sur ce point lors de la mise en œuvre du TCN.

Nos réseaux de neurones entièrement connectés

Nos FCNNs ont été entraînés sur un ensemble de données composé de 3 367 245 mesures, avec 71 entrées. Sur l'ensemble des points, 25%, soit 841 811, sont mis de côté pour les tests, 25%, soit 631 358, pour la validation, et les 50% restants, soit 1 894 076 enregistrements, pour l'entraînement. Une schématisation complète des entrées et sorties se trouve à la figure 3.18 de ce manuscrit, page 214.

Étant donné le bruit inhérent associé aux mesures, notre ensemble de données pose des défis, pouvant potentiellement conduire à du sur-apprentissage. De plus, en raison de la nature des mesures de flux, l'algorithme peut privilégier les points de données centraux par rapport à ceux plus extrêmes et plus rares. Par conséquent, une fonction de coût légèrement inférieure est anticipée par rapport à PrecipNet-R, puisque davantage de points de données centraux ont été inclus.

Quatre modèles principaux ont été considérés après de nombreux essais :

1. Un modèle FCNN de base avec environ un tiers des paramètres de PrecipNet-R.
2. Un modèle FCNN avec une architecture d'autoencodeur.
3. Les deux architectures mentionnées ci-dessus, mais incorporant une fonction de perte spécialisée appelée Tail Weighted Loss, conçue pour pénaliser davantage le réseau de neurones lorsqu'il sous-estime les valeurs au-dessus d'un certain seuil, mettant ainsi l'accent sur la précision des prédictions lors d'événements extrêmes. Cette fonction a montré des résultats optimaux avec des paires de seuils et de pénalités spécifiques.

Nos modèles utilisent tous l'optimiseur Adam, un taux d'apprentissage de 0,001 et une valeur de dropout de 0,1. Les temps d'entraînement pour ces modèles varient, tous étant notablement inférieurs aux 1,501 heures de PrecipNet.

Nous privilégions MSE, RMSE, MAE et le temps d'entraînement comme étant les métriques les plus pertinentes pour évaluer nos modèles. La MSE et la RMSE sont des mesures de la précision des prédictions qui pénalisent les erreurs plus importantes de manière significative. Cela

les rend particulièrement pertinentes pour la prédiction d'événements extrêmes. La RMSE est particulièrement utile car elle est dans la même unité que notre cible, ce qui la rend plus facile à interpréter. Nous nous concentrons également sur les métriques MSE appliquées uniquement aux points de données situés au-delà de percentiles spécifiques (90e, 95e, 99e). Il s'agit des événements les plus intenses pertinents et donc les plus pertinents pour l'utilisation de SpaceAble (prévoir les risques sur satellite).

Les détails des modèles 1 à 4 se trouvent sur le tableau 4.1 qui se trouve à la page 229. Les résultats en entraînant les 4 modèles se trouvent dans les tableaux 4.2 et 4.3 de la page 233. Le tableau 4.2 présente les résultats sur les jeux de test des 4 modèles et le tableau 4.3 compare les performances des différents modèles (dont OVATION et PrecipNet) sur le satellite F16 année 2010 (le jeu de validation de [McGranaghan et al. \(2021\)](#)). Voici les résultats de notre analyse :

- De nombreux tests ont été effectués avec des régularisations L1 et L2. Aucune constatation significative n'a été observée.
- De même, divers schedulers ont été utilisés, mais aucun n'a considérablement amélioré les résultats.
- Les optimiseurs SGD et Adam ont été testés. Bien qu'Adam puisse être plus complexe, il converge presque toujours plus rapidement, réduisant le nombre d'époques nécessaires (et donc le temps d'exécution).
- L'utilisation d'un autoencodeur a souvent amélioré les résultats. L'architecture en bouteille d'étranglement a contribué à réduire le bruit dans les données. Cependant, les meilleurs résultats ont été obtenus avec des architectures plus grandes et un grand nombre de paramètres, que nous avons cherché à éviter en raison de considérations de performance et de capacité de calcul.
- De nombreux changements ont été apportés, notamment concernant le taux d'apprentissage et la taille du lot. Une très petite taille de lot rend chaque époque fastidieuse, tandis qu'un lot trop grand n'ajoute pas beaucoup et ralentit la convergence de l'algorithme. Un taux d'apprentissage légèrement supérieur à 0,001, combiné à un planificateur ExponentialLR, s'est avéré efficace mais a parfois poussé la machine à ses limites. Les choix finaux ici restent simples tant en termes d'implémentation que de compréhension.

Après comparaison avec OVATION et PrecipNet (voir figures 4.15, 4.16, 4.17 et tableaux 4.3 et 4.4), nous observons que :

- Le modèle 1 surpasse légèrement PrecipNet comme le montre le tableau 4.3 et les courbes de la figure 4.17.
- Le modèle 4 surpasse PrecipNet pour les valeurs élevées, donc pendant les événements violents (visibles dans 4.17).
- Nos modèles 1 et 4 surpassent OVATION en termes de MAE, RMSE et MSE. Le modèle 1 améliore les résultats d'OVATION de 85% pour le MSE (et respectivement de 20% pour la MAE et de 61% pour le RMSE).
- OVATION surestime systématiquement les valeurs réelles, comme on peut le voir dans l'histogramme de la figure 4.15 et dans les estimations de densité de la figure 4.16. Cependant, cela le rend très efficace pour les cas rares et les événements extrêmes, comme le montre le tableau 4.4. Dans les cas les plus extrêmes (au-dessus du 99ème percentile), il surpasse notre modèle 4 d'environ 14%. Cependant, il est toujours beaucoup moins efficace que les modèles 1, 4 ou même PrecipNet.
- Si nous considérons que toutes les métriques sont également importantes pour nous, notre modèle 4 est le plus équilibré, sous-performant légèrement OVATION sur les valeurs extrêmes et sous-performant légèrement le modèle 1 sur les métriques standard.

- Nous n'utilisons pas strictement les mêmes entrées que PrecipNet. En effet, nous avons omis $F_{10.7}$ et l'indice de la calotte polaire. Les ajouter n'améliore pas significativement nos résultats (la différence de résultats est équivalente à changer la valeur de la graine).

Ainsi, les Modèles 1 et 4 se sont révélés être les plus efficaces, leur pertinence variant en fonction des objectifs spécifiques. Les aperçus des divers tests et la comparaison avec OVATION ont fourni des informations précieuses sur les points forts du modèle et les domaines à améliorer, jetant les bases pour un développement et un raffinement ultérieurs.

Un exemple à explorer dans des études futures est la possibilité de combiner les algorithmes utilisés. En effet, le modèle 1 peut se concentrer sur les valeurs moyennes et le modèle 4 sur les valeurs extrêmes. De cette manière, et avec un seuil judicieusement choisi, il serait possible d'utiliser les deux algorithmes et de passer de l'un à l'autre en fonction de la valeur prédite. Par exemple, si la valeur prédite par le modèle 1 dépasse 11, le modèle 4 sera probablement plus proche de la valeur réelle, où ce dernier serait moins précis pour les valeurs entre 8 et 10. Pour aller plus loin, le passage de l'un à l'autre pourrait être remplacé par un mélange progressif et continu.

Enfin, certaines questions ont été intentionnellement laissées de côté, telles que l'affinement des prévisions pour les valeurs basses (entre 7 et 8). L'idée d'une fonction de coût adaptée pour pousser l'algorithme à ses limites lors de la prévision de ces valeurs est tout à fait possible. Cependant, comme ces valeurs représentent des cas non dangereux pour les systèmes spatiaux et terrestres, ce n'était pas une priorité de les aborder.

Deux questions restent non résolues, auxquelles nous voulions fournir une réponse préliminaire. Avec ces architectures et outils, serait-il possible de prévoir à court terme ? Un des défis liés aux FCNN est le choix arbitraire des données passées ; serait-il possible de surmonter cette difficulté en utilisant un TCN ?

Prévisions

L'idée pour réaliser une prévision était de supprimer les données à T_0 en réduisant le nombre d'entrées à 62. Par conséquent, la capacité de modélisation de notre réseau devient une capacité prédictive à 10 minutes. En effet, les premières données historiques utilisées sont celles à T_0-10mn . En définissant T_0-10mn comme T_0 , nous nous retrouvons dans une situation où nous prédisons les flux d'énergie à T_0+10mn . Pour donner un premier aperçu (une "preuve de concept"), nous avons re-entraîné le modèle 1 en suivant cette idée. Les résultats sont visibles sur le tableau 4.5 page 238 du manuscrit.

Les figures d'histogramme, de densité (figure 4.18), et les courbes d'apprentissage restent approximativement les mêmes. Comme nous l'avons déjà mentionné, une grande partie de la prédiction repose sur la date et la position observée (latitude et longitude magnétique), mais nous notons tout de même que notre prévision de 10 minutes est pertinente et justifie des études futures pour une amélioration potentielle de cette capacité prédictive. De plus, contrairement aux résultats du modèle 1 visibles dans la Figure 4.13, il semble que la modélisation fut meilleure pour les valeurs basses, en dessous de $8 \text{ en } \log_{10}(\text{eV}/\text{cm}^2/\text{s})$.

Cependant, nous sommes toujours limités aux horodatages historiques choisis. Idéalement, nous savons que la progression du vent solaire et le délai entre ce qui est observé au choc de l'arc et ce qui arrive sur Terre est dynamique. Il est trop grossier de supposer que toutes les observations aux pôles proviennent d'événements au niveau du choc de l'arc qui se sont produits un temps fixe ΔT auparavant. Pour contourner ce problème, plusieurs solutions sont disponibles, telles que

l'utilisation d'entrées dynamiques basées sur certaines valeurs.

Une deuxième solution, plus intéressante à notre avis, consiste à se tourner vers d'autres catégories de réseaux capables de prendre en compte toutes les données passées. C'est le cas de certains réseaux de mémoire comme les LSTMs. Mais le plus pertinent nous a semblé être le réseau de convolution temporelle (TCN). Les caractéristiques clés des TCNs incluent les convolutions causales et dilatées. Les convolutions causales garantissent que la prédiction du modèle au temps t dépend uniquement des informations du temps t ou antérieur, maintenant l'ordre temporel. Les convolutions dilatées permettent au réseau d'avoir un champ récepteur plus large, ce qui signifie qu'il peut prendre en compte des informations provenant de plus loin dans le passé sans augmenter significativement le nombre de paramètres. Ses avantages par rapport aux RNNs et aux LSTMs peuvent être résumés à travers trois points principaux connus :

- **Parallélisation** : Contrairement aux RNNs, les TCNs traitent tous les points de données simultanément, ce qui permet des temps d'entraînement plus rapides.
- **Gradients stables** : Les TCNs évitent les problèmes de gradients disparaissants et explosifs souvent rencontrés dans les RNNs.
- **Taille de champ récepteur flexible** : L'utilisation de convolutions dilatées permet d'avoir des champs récepteurs ajustables pour capturer le contexte temporel pertinent.

Dans les premières étapes du projet, nous avons été confrontés à des incertitudes quant à la profondeur à laquelle nous voulions récupérer les données et nous avons été désireux d'éviter d'être limités par notre choix d'architecture. Bien que les LSTMs surpassent généralement les RNNs, leur efficacité diminue à mesure que les séquences historiques s'allongent, entraînant une complexité accrue et des performances réduites. De plus, le temps de calcul est plus long pour les LSTMs. Malgré les avantages potentiels d'explorer différentes architectures avec plus de temps, nous nous sommes donc concentrés sur les TCNs.

Répondre à certaines contraintes via les TCN

Le TCN, semblable aux CNNs en reconnaissance d'image, évalue les plages de temps passées pour identifier des motifs, visant à relier dynamiquement les tendances observées et les valeurs de sortie. Après avoir réalisé un long benchmark, nous avons créé deux modèles différents pour nos TCN qui se retrouvent dans le tableau 4.6.

L'ensemble de données utilisé est créé en utilisant les données DMSP et OMNIWeb. On génère une liste d'indices valides qui ne contiennent aucune donnée manquante dans les n points de données précédents, n représentant la taille en minutes de la plage historique. Une plage de 30 minutes précédant chaque T_0 a été choisie. Cette plage ne doit contenir aucune donnée manquante et c'est pourquoi nous la prenons si petite. Cet aspect est voué à évoluer dans de prochaines études.

Au final, 478 669 échantillons valides ont été obtenus, avec 20% des données réservées pour l'ensemble de test et 80% pour l'ensemble de validation. Nous notons la possibilité d'augmenter la taille de l'ensemble de données en utilisant les données interpolées que nous avons créées, mais cette voie n'a pas été explorée par manque de temps.

Suite à un benchmark approfondi, deux modèles finaux ont été retenus, différant par leur fonction de perte (MSE pour le Modèle 1 et Tail Weighted Loss pour le Modèle 2). Certains hyperparamètres spécifiques ont été adaptés au TCN, tels que "channels", "dilation factor", et "kernel". Un scheduler LR exponentiel a été utilisé, expliquant le "taux d'apprentissage" initial élevé. Le nombre d'époques a été limité en raison de la stabilisation rapide de la courbe d'apprentissage et

de la longue durée de l'entraînement. Les régularisations et le dropout ont été testés mais n'ont pas eu d'impact significatif sur les résultats, ils ont donc été omis.

Notre TCN est présenté est évalué en fonction de diverses métriques, notamment la MAE, la MSE, la RMSE et la MSE sur des données supérieures à certains percentiles (90e, 95e et 99e). Le temps d'entraînement de ce nouveau modèle a considérablement augmenté par rapport aux FCNN, ce qui pose un défi pour les mises à jour régulières une fois industrialisé. Mais malgré certaines lacunes en termes de MSE, il présente une amélioration dans la gestion des cas rares et extrêmes, en particulier lors de l'utilisation de la Tail Weighted Loss.

Nous avons effectué une analyse comparative des résultats du modèle par rapport à ceux d'OVATION et notre modèle présente des performances supérieures. Les résultats se trouvent dans le tableau 4.7 page 242.

Une fois notre modèle entraîné, nous avons tenté de développer son application "pratique" afin qu'il devienne un produit utilisable par l'entreprise SpaceAble. Le produit ainsi créé permet la visualisation des trajectoires des satellites DMSP et des valeurs prédites (voir les figures 4.24, 4.25 et 4.26). Le modèle utilise plusieurs paramètres d'entrée, tels que la date de début, la version du modèle et l'unité, entre autres, pour récupérer de lui-même les données OMNI correspondantes et générer le nombre souhaité de points de données. Il permet également des comparaisons directe avec OVATION et les données DMSP réelles, facilitant la visualisation des données même en l'absence de points de données mesurées.

La polyvalence de ce "produit" est également démontrée par sa capacité à générer des cartes aurorales pour des dates, des heures et des emplacements spécifiques en émulant les satellites DMSP. Ces cartes peuvent être générées en environ deux minutes et peuvent représenter à la fois des périodes tumultueuses et plus calmes avec une précision réglable en termes d'heure magnétique locale et de latitude. Certaines irrégularités dans les prédictions sont notées, pouvant être attribuées à des lacunes dans les données disponibles.

De plus, nous introduisons le concept de Gradients Intégrés (IG), une technique d'interprétabilité conçue pour comprendre les contributions des caractéristiques d'entrée aux prédictions d'un modèle. IG est mis en avant pour sa compatibilité avec les réseaux neuronaux d'apprentissage profond, sa nature non intrusive et sa capacité à identifier les caractéristiques influentes sans modifier le réseau d'origine. Il respecte les axiomes de sensibilité et d'invariance à la mise en œuvre, en faisant un outil précieux pour assurer la transparence et l'explicabilité des modèles.

Gradients Intégrés

La méthode des "integrated gradients" est une technique d'*interprétabilité* des modèles de machine learning, pour comprendre quelles caractéristiques des données d'entrée ont le plus d'influence sur les prédictions du modèle. Elle repose sur l'idée d'attribuer une importance à chaque caractéristique (appelée *attribut*) des données d'entrée en calculant l'intégrale des gradients le long du chemin de la ligne droite reliant un point de référence (généralement un point de référence neutre comme le point de départ) à l'exemple d'entrée actuel.

Pour un modèle de prédiction donné, les gradients représentent la pente de la fonction de prédiction par rapport à chaque attribut des données d'entrée. L'intégration des gradients le long du chemin entre le point de référence et l'exemple d'entrée actuel permet de mesurer l'impact cumulatif de chaque attribut sur la prédiction finale du modèle.

En calculant ces intégrales, on peut attribuer une importance relative à chaque attribut des données d'entrée. Cette importance relative permet de mieux comprendre quels attributs ont le plus d'influence sur les prédictions du modèle et comment ces attributs contribuent aux résultats obtenus.

Vers la fin de notre étude, nous avons été en mesure de mettre en œuvre cette méthode et d'entrevoir ses résultats et les possibilités qui en découlent. Nos résultats sont intrigants car ils reflètent l'"importance" qu'une entrée donnée a sur la sortie mais cela *par rapport à une "ligne de base"*. Dans notre cas, nous avons choisi arbitrairement d'envoyer des vecteurs nuls comme ligne de base, ce qui est discutable.

Ce qui est important, c'est de comprendre comment ces attributions contribuent à la compréhension du comportement du modèle et si elles sont en accord avec nos connaissances du domaine ou les attentes du problème. Dans les attributions (nos résultats après l'application des gradients intégrés), nous avons obtenu à la fois des valeurs positives et négatives. Les valeurs négatives indiquent une corrélation négative entre l'entrée et la sortie et sont tout aussi significatives que les valeurs positives. Dans l'affichage de la Figure 4.27, nous ne montrons pas les entrées pour lesquelles le passé n'est pas utilisé en entrée (comme la position du satellite qui est une entrée fixe) et nous avons normalisé en divisant par la valeur maximale absolue. Nous avons ensuite pris la valeur absolue des attributions. Ainsi, nous obtenons une valeur entre 0 et 1, avec 1 correspondant à une forte corrélation. Trois résultats sont visibles dans la Figure 4.27 pour trois échantillons choisis au hasard : le 23 mai 2011 à 05h38 ; le 20 septembre 2004 à 07h08 ; et le 18 mars 2010 à 01h59.

Pour l'instant, il est difficile de tirer beaucoup plus de conclusions que la simple observation car nous devons encore approfondir. Cependant, on peut déjà en voir le potentiel, surtout dans des contextes similaires où la modélisation (ou la prédiction) est effectuée avec des données mesurées à divers endroits (et donc à différents moments). En effet, dans le cas où cette méthode fonctionnerait parfaitement, il serait par exemple possible de remonter à la "période d'intérêt" dans le passé. Nous pourrions donc observer dans les données quelles données historiques dans l'espace interplanétaire sont à l'origine des observations près de la Terre. Nous pourrions ainsi remonter à la durée (dynamique) qui sépare nos entrées de nos sorties. De nombreux tests pourraient ainsi nous donner un équivalent de loi empirique sur ce ΔT entre entrée et sortie qui pourrait être réutilisée rétroactivement afin d'améliorer à nouveau nos algorithmes, ainsi de suite.

Remarques finales

Cette étude ne vise pas uniquement à optimiser les métriques, mais cherche à établir une base pour la recherche en IA au sein de SpaceAble dans le domaine de l'évaluation des risques pour les objets spatiaux (et dans une certaine mesure, sols). La thèse se construit donc comme une encyclopédie et un ensemble d'algorithmes flexibles, structurés pour former un produit utilisable dans la prévision des risques.

Au chapitre 4, nous détaillons l'organisation de notre code, les architectures utilisées, les résultats obtenus et les comparaisons effectuées avec PrecipNet et OVATION, les codes de référence dans ce domaine. Notre recherche a mis en lumière des défis dans la réadaptation de PrecipNet en raison de son architecture rigide et de décisions, telles que le choix des points de données historiques et les méthodes de prétraitement, qui n'étaient pas facilement modifiables pour tester des algorithmes et des méthodes alternatifs.

L'absence d'un ensemble de test dans PrecipNet a rendu les comparaisons difficiles. Nous avons adopté une approche différente en répartissant aléatoirement les données en ensembles d'entraînement, de validation et de test. Notre étude a également omis intentionnellement certains paramètres d'entrée, qui n'avaient pas d'impacts significatifs sur les résultats.

L'intégration des données de temps et de position posait un défi, l'algorithme modélisant principalement l'ovale auroral moyen. Bien que nous ayons atteint la majorité de nos capacités prédictives grâce à ces paramètres, se concentrer sur les variations dynamiques et locales était crucial. Nous avons exploré diverses solutions, privilégiant finalement les données à haute valeur en utilisant la fonction Tail Weighted Loss et des métriques spécifiques pour les valeurs élevées, améliorant les résultats sur les valeurs extrêmes mais dégradant légèrement le résultat global.

Les comparaisons entre notre FCNN, OVATION et PrecipNet ont révélé des améliorations significatives. OVATION a tendance à surestimer les valeurs réelles, le rendant plus adapté pour les valeurs extrêmes. La performance prometteuse de notre TCN, nécessitant moins de paramètres et offrant de la flexibilité grâce à la convolution, nous a amenés à le sélectionner comme algorithme principal pour notre produit. Malgré son temps d'entraînement et ses exigences en matière de données, des problèmes tels que les lacunes dans les données pourraient être contournés par interpolation.

Nos FCNN et TCN ont montré des résultats prometteurs, jetant les bases pour un développement futur au sein de SpaceAble. Les données organisées, le prétraitement et l'approche scientifique détaillée servent de point de départ solide pour explorer une gamme plus large de méthodes au sein des équipes de SpaceAble. Le TCN, en particulier, offre des opportunités pour analyser des périodes plus longues et pourrait être utilisé en conjonction avec d'autres algorithmes pour des prédictions complètes.

Chapitre 5: Perspectives

Résumé des observations clés et des résultats

Cette thèse sert de ressource pour la recherche au sein de SpaceAble dans le domaine de la météorologie spatiale. Elle couvre divers sujets allant des interactions Soleil-Terre, aux risques, aux concepts derrière l'apprentissage machine jusqu'aux tests d'algorithmes. Parmi les observations significatives de la thèse, on note la construction et l'organisation réussies du code, en utilisant spécifiquement PyTorch-Lightning, une librairie idéale pour l'intégration dans la plateforme SpaceAble. Notre étude a également utilisé un LSTM avec PyTorch-Lightning dans la compétition "MagNet: Model the Geomagnetic Field", obtenant la 28e position sur plus de 500, ouvrant la voie à son intégration dans SpaceAble.

L'étude fournit également des perspectives sur l'analyse, l'utilisation et l'intégration des données, mettant en évidence l'importance des variations de données et de la diversification des sources de données. Un partenariat avec le Laboratoire de Physique des Plasmas (LPP) et l'École Polytechnique, ainsi qu'un projet de mentorat, ont visé à évaluer et à améliorer la fiabilité des données OMNI. Nous pensons que le prétraitement devrait rester accessible et personnalisable par l'utilisateur pour une application industrielle efficace, tandis que les produits finaux proposés ne

devraient pas imposer de prétraitement intégré.

Enfin, la recherche confirme la viabilité de l'utilisation de l'IA dans la météorologie spatiale, l'IA s'avérant efficace dans de nombreuses applications. La thèse explore l'utilisation de bibliothèques hybrides et d'architectures non conventionnelles comme TCN pour la prévision de séries temporelles, marquant une première étape prometteuse. Combiné à des techniques comme les gradients intégrés, l'objectif est d'améliorer les résultats et de combler l'écart entre les modèles boîte blanche et boîte noire.

Travaux futurs et perspectives

Sous forme de liste ci-dessous, nous avons récapitulé les perspectives envisageables et les travaux à venir, en particulier au sein de SpaceAble.

Piste 1 : Améliorations des Algorithmes

- **Expansion de l'Ensemble de Données** : Amélioration des capacités de généralisation du modèle TCN en le formant sur un ensemble de données élargi comprenant trois millions de points de données, tout en reconnaissant la potentielle prolongation de la période de formation.
- **Ajustement de la Fenêtre Temporelle** : Exploration de l'expansion de la fenêtre temporelle d'entrée du TCN à 2-3 heures, en utilisant des ensembles de données prétraités de manière méticuleuse et en abordant les défis associés.
- **Révision de la Fonction de Coût** : Exploration de révisions de la fonction de coût, inspirées par la distribution des données et la compréhension des facteurs physiques sous-jacents, pour optimiser l'efficacité.
- **Spécificité Géographique** : Amélioration des prédictions dans des régions géographiques spécifiques en abordant la répartition des données et en utilisant l'interpolation pour combler les lacunes.
- **Ingénierie des Caractéristiques et Réduction du Bruit** : Développement de nouvelles caractéristiques à partir de données existantes, réduction du bruit, détection des anomalies et reconsidération de la régularisation.
- **Exploration de Nouveaux Modèles** : Proposition d'explorer des modèles récurrents tels que les RNN ou les LSTM pour des tests futurs.

Piste 2 : Combinaison de modèles

- **Intégration de Modèles** : Suggestion d'un système de seuil stratégique pour combiner les modèles, avec des études prévues pour identifier les seuils optimaux tout en équilibrant la précision et l'importance des valeurs élevées. Cette idée vient notamment des performances très élevées de OVATION Prime sur les valeurs extrêmes.
- **Méthode de Stacking** : Proposition d'une méthode de stacking versatile intégrant TCN, LSTM, et potentiellement FCN, pour capturer les dépendances à long terme et gérer des motifs séquentiels complexes.

Piste 3 : Pousser plus loin la méthode des gradients intégrés

- **Transparence du Modèle** : Utilisation des gradients intégrés pour dévoiler la nature « boîte noire » des modèles, améliorant la transparence des prédictions en mesurant l'importance relative des points temporels dans les séquences d'entrée.
- **Compréhension du Délai Temporel** : Utilisation de la méthode pour identifier et analyser les délais temporels entre les entrées et les sorties influentes, avec des attentes de révélation de corrélations, particulièrement avec la vitesse du vent solaire.

Implications et applications pour la recherche et pour l'industrie

Cette thèse marque une avancée significative dans la fusion de la recherche en météorologie spatiale avec l'intelligence artificielle, se concentrant sur la prédiction des phénomènes d'interaction Soleil-Terre. Le travail est prêt à apporter des contributions notables à la fois à la communauté de recherche et à l'industrie naissante de la météorologie spatiale, en particulier dans le secteur de la Conscience de la Situation Spatiale. Le modèle développé, utilisant des architectures comme TCNs et FCNNs, prédit avec succès les valeurs de flux d'énergie et offre une interprétabilité améliorée grâce aux Gradients Intégrés, ajoutant une transparence cruciale pour la compréhension et l'ajustement du modèle.

Dans la communauté de recherche, le travail suggère deux applications potentielles : comme outil de prévision avancé et comme amélioration des codes physiques existants. Une collaboration avec le Dr Maxime Grandin et l'Université de Helsinki a suscité des idées pour de futurs projets conjoints, y compris la possibilité d'intégrer le code Vlasiator. De plus, il existe de nombreuses synergies au sein de la communauté, permettant l'intégration avec d'autres éléments tels que les prévisions de vent solaire, la modélisation ionosphérique, la prédiction d'indices solaires ou géomagnétiques, etc. La collaboration avec des entités comme le CSUG et l'ESA ouvre également ces opportunités via l'exploitation de nouvelles données jusqu'alors inaccessibles.

Pour l'industrie, la thèse revêt une importance particulière à un moment de croissance et de changement, les modèles étant essentiels pour la prédiction de l'environnement spatial et l'évaluation des risques pour les satellites. L'opérationnalisation de ces modèles nécessite une approche simplifiée mais flexible, ce que nous tentons via l'utilisation de PyTorch Lightning. Au-delà de l'espace, le secteur de l'énergie, les entreprises de télécommunications et les compagnies d'assurance peuvent également tirer parti de ces modèles pour renforcer la sécurité et la fiabilité.

Malgré les limitations concernant la divulgation de détails spécifiques sur les applications et partenariats de SpaceAble, la thèse est optimiste quant à l'impact significatif sur l'avenir de l'industrie spatiale privée. Plusieurs avenues prometteuses pour de futures recherches ont été identifiées, notamment l'expansion des ensembles de données, le choix de nouvelles fonctions de coût, l'exploration de méthodes de réduction du bruit et l'investigation de diverses architectures de modèles.

Conclusion

En conclusion, nous pensons que ces travaux s'inscrivent parfaitement dans cette nouvelle ère de la recherche en météorologie spatiale, qui favorise l'intégration de l'IA et de la modélisation physique pour des prévisions précises et des insights plus profonds sur les phénomènes spatiaux. On espère que le travail sera un catalyseur pour de futures collaborations, applications industrielles et avancées dans la compréhension de l'environnement spatial, contribuant de manière significative au domaine en évolution de la météorologie spatiale.

Bibliography

- Akasofu, S.-I., 1968. Polar and Magnetospheric Substorms. New York: Springer-Verlag.
- Albawi, S., T. A. Mohammed, and S. Al-Zawi, 2017. Understanding of a convolutional neural network. In 2017 International Conference on Engineering and Technology (ICET), 1–6. 10.1109/ICEngTechnol.2017.8308186.
- Altschuler, M. D., and G. Newkirk, 1969. Magnetic fields and the structure of the solar corona. *Solar Physics*, **9**(1), 131–149. 10.1007/BF00145734, URL <https://doi.org/10.1007/BF00145734>.
- Amariutei, O. A., and N. Y. Ganushkina, 2012. On the prediction of the auroral westward electrojet index. *Annales Geophysicae*, **30**(5), 841–847. 10.5194/angeo-30-841-2012, URL <https://angeo.copernicus.org/articles/30/841/2012/>.
- Amaya, J., 2019. Automatic unsupervised classification of the solar wind using Self-Organizing Maps. In Machine Learning in Heliophysics 2019, 6.
- Amaya, J., R. Dupuis, and G. Lapenta, 2020. Unsupervised classification of the solar wind using Self-Organizing Maps. In EGU General Assembly Conference Abstracts, EGU General Assembly Conference Abstracts, 15568. 10.5194/egusphere-egu2020-15568.
- Amory-Mazaudier, C., M. Menvielle, J.-J. Curto, and M. Le Huy, 2017. Recent Advances in Atmospheric, Solar-Terrestrial Physics and Space Weather From a North-South network of scientists [2006-2016] PART A: TUTORIAL. *Sun and Geosphere*, **12**, 1–19.
- Analytics, S., 2021. Space Tech Industry 2021/Q2. *Landscape Overview*, May.
- Ashmall, J., and V. Moore, 1997. Long-term prediction of solar activity using neural networks. *Proceedings of AI Applications in Solar-Terrestrial Physics*, Lund, Sweden, 117–122.
- Ashour-Abdalla, M., R. J. Walker, V. Perroomian, and M. El-Alaoui, 2008. On the importance of accurate solar wind measurements for studying magnetospheric dynamics. *Journal of Geophysical Research: Space Physics*, **113**(A8). <https://doi.org/10.1029/2007JA012785>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2007JA012785>.
- Bai, S., J. Zico Kolter, and V. Koltun, 2018. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *arXiv e-prints*, arXiv:1803.01271. 10.48550/arXiv.1803.01271.
- Bailey, R. L., R. Leonhardt, C. Möstl, C. Beggan, M. A. Reiss, A. Bhaskar, and A. J. Weiss, 2022. Forecasting GICs and Geoelectric Fields From Solar Wind Data Using LSTMs: Application in Austria. *Space Weather*, **20**(3), e2021SW002,907. E2021SW002907 2021SW002907, doi.org/10.1029/2021SW002907, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021SW002907>.
- Baker, D. N., S. J. Bame, W. C. Feldman, J. T. Gosling, R. D. Zwickl, J. A. Slavin, and E. J. Smith, 1986. Strong electron bidirectional anisotropies in the distant tail: ISEE 3 observations of polar rain. *Journal of Geophysical Research: Space Physics*, **91**(A5), 5637–5662.

- <https://doi.org/10.1029/JA091iA05p05637>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/JA091iA05p05637>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/JA091iA05p05637>.
- Baker, K. B., and S. Wing, 1989. A new magnetic coordinate system for conjugate studies at high latitudes. *Journal of Geophysical Research: Space Physics*, **94**(A7), 9139–9143. <https://doi.org/10.1029/JA094iA07p09139>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/JA094iA07p09139>.
- Bakrania, M. R., I. J. Rae, A. P. Walsh, D. Verscharen, and A. W. Smith, 2020. Using Dimensionality Reduction and Clustering Techniques to Classify Space Plasma Regimes. *Frontiers in Astronomy and Space Sciences*, **7**. 10.3389/fspas.2020.593516, URL <https://www.frontiersin.org/articles/10.3389/fspas.2020.593516>.
- Bals, A.-M., C. Thakrar, and K. B. Deshpande, 2022. Creating a Database to Identify High-Latitude Scintillation Signatures With Unsupervised Machine Learning. *IEEE Journal of Radio Frequency Identification*, **6**, 240–249. 10.1109/JRFID.2022.3163913.
- Bank, D., N. Koenigstein, and R. Giryes, 2020. Autoencoders. *arXiv e-prints*, arXiv:2003.05991. 10.48550/arXiv.2003.05991, 2003.05991.
- Barlow, H., 1989. Unsupervised Learning. *Neural Computation*, **1**(3), 295–311. 10.1162/neco.1989.1.3.295, <https://direct.mit.edu/neco/article-pdf/1/3/295/811863/neco.1989.1.3.295.pdf>, URL <https://doi.org/10.1162/neco.1989.1.3.295>.
- Basodi, S., C. Ji, H. Zhang, and Y. Pan, 2020. Gradient amplification: An efficient way to train deep neural networks. *Big Data Mining and Analytics*, **3**(3), 196–207. 10.26599/BDMA.2020.9020004.
- Beccuti, G., 2013. Impact of Solar Storms on the Swiss Transmission Network. [Online; last accessed June 2023], URL https://ethz.ch/content/dam/ethz/special-interest/mavt/energy-science-center-dam/events/frontiers-presentations/131106_FiER_Beccuti.pdf.
- Bellan, P. M., 2006. Fundamentals of Plasma Physics. Cambridge University Press. 10.1017/CBO9780511807183.
- Bergin, A., S. C. Chapman, N. W. Watkins, N. R. Moloney, and J. W. Gjerloev, 2023. Extreme Event Statistics in Dst, SYM-H, and SMR Geomagnetic Indices. *Space Weather*, **21**(3), e2022SW003,304. E2022SW003304 2022SW003304, <https://doi.org/10.1029/2022SW003304>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2022SW003304>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2022SW003304>.
- Bhaskar, A., and G. Vichare, 2019. Forecasting of SYMH and ASYH indices for geomagnetic storms of solar cycle 24 including St. Patrick’s day, 2015 storm using NARX neural network. *Journal of Space Weather and Space Climate*, **9**, A12. 10.1051/swsc/2019007.
- Birkeland, K., 1908. The Norwegian aurora polaris expedition 1902-1903, vol. 1. H. Aschelhoug & Company.
- Bohlin, T. P., 2006. Practical grey-box process identification: theory and applications. Springer Science & Business Media.

-
- Boström, R., 1964. A model of the auroral electrojets. *Journal of Geophysical Research (1896-1977)*, **69**(23), 4983–4999. <https://doi.org/10.1029/JZ069i023p04983>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/JZ069i023p04983>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/JZ069i023p04983>.
- Bothmer, V., and I. A. Daglis, 2007. *Space weather: physics and effects*. Springer Science & Business Media. <https://doi.org/10.1007/978-3-540-34578-7>.
- Bouriat, S., P. Vandame, M. Barthélémy, and J. Chanussot, 2022. Towards an AI-based understanding of the solar wind: A critical data analysis of ACE data. *Frontiers in Astronomy and Space Sciences*, **9**. 10.3389/fspas.2022.980759, URL <https://www.frontiersin.org/articles/10.3389/fspas.2022.980759>.
- Bouriat, S., S. Wing, and M. Barthélémy, 2023. Electron Aurora and polar rain dependencies on Solar Wind Parameters. *submitted to Journal of Geophysical Research: Space Physics*.
- Bowman, B., W. K. Tobiska, F. Marcos, C. Huang, C. Lin, and W. Burke. A New Empirical Thermospheric Density Model JB2008 Using New Solar and Geomagnetic Indices, 2008. 10.2514/6.2008-6438, <https://arc.aiaa.org/doi/pdf/10.2514/6.2008-6438>, URL <https://arc.aiaa.org/doi/abs/10.2514/6.2008-6438>.
- Brekke, A., 1997. *Physics of the polar upper atmosphere*. Wiley-Praxis Series in Atmospheric Physics, John Wiley & Sons Inc.
- Britannica, E., 1994. field-aligned current system. [Online; last accessed June 2023], URL <https://www.britannica.com/science/geomagnetic-field/Field-aligned-currents#/media/1/229754/1168>.
- Britannica, E., 2012. layers of Earth's ionosphere. [Online; last accessed June 2023], URL <https://www.britannica.com/science/ionosphere-and-magnetosphere#/media/1/1369043/167048>.
- Bro, R., and A. K. Smilde, 2014. Principal component analysis. *Analytical methods*, **6**(9), 2812–2831.
- Brown, E. J. E., F. Svoboda, N. P. Meredith, N. Lane, and R. B. Horne, 2022. Attention-Based Machine Vision Models and Techniques for Solar Wind Speed Forecasting Using Solar EUV Images. *Space Weather*, **20**(3), e2021SW002976. E2021SW002976 2021SW002976, <https://doi.org/10.1029/2021SW002976>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2021SW002976>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021SW002976>.
- Brownlee, J., 2019. How to use learning curves to diagnose machine learning model performance. [Online; last accessed July 2023], URL <https://machinelearningmastery.com/learning-curves-for-diagnosing-machine-learning-model-performance/>.
- Bruinsma, S., and C. Boniface, 2021. The operational and research DTM-2020 thermosphere models. *Journal of Space Weather and Space Climate*, **11**, 47. 10.1051/swsc/2021032.
- Calvo, R. A., H. A. Ceccato, and R. D. Piacentini, 1995. Neural Network Prediction of Solar Activity. *Astrophysical Journal*, **444**, 916. 10.1086/175661.
- Camporeale, E., 2019. The Challenge of Machine Learning in Space Weather: Nowcasting and Forecasting. *Space Weather*, **17**(8), 1166–1207. <https://doi.org/10.1029/2018SW002061>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2018SW002061>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018SW002061>.
-

- Cannon, P., M. Angling, L. Barclay, C. Curry, C. Dyer, et al., 2013. Extreme space weather: impacts on engineered systems and infrastructure. Royal Academy of Engineering.
- Capon, C., B. Smith, M. Brown, R. Abay, and R. Boyce, 2019. Effect of ionospheric drag on atmospheric density estimation and orbit prediction. *Advances in Space Research*, **63**(8), 2495–2505. <https://doi.org/10.1016/j.asr.2019.01.013>, URL <https://www.sciencedirect.com/science/article/pii/S0273117719300213>.
- Carreau, C., 2013. Earth’s plasmasphere and the Van Allen belts. [Online; last accessed June 2023], URL <https://sci.esa.int/web/cluster/-/52831-earth-plasmasphere-and-the-van-allen-belts>.
- Carrington, R. C., 1859. Description of a singular appearance seen in the Sun on September 1, 1859. *Monthly Notices of the Royal Astronomical Society*, Vol. 20, p. 13-15, **20**, 13–15.
- Carvalho, D. V., E. M. Pereira, and J. S. Cardoso, 2019. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics*, **8**(8). 10.3390/electronics8080832, URL <https://www.mdpi.com/2079-9292/8/8/832>.
- Cellan-Jones, R., 2014. Stephen Hawking warns artificial intelligence could end mankind. *BBC news*, **2**(10), 2014.
- Cho, K., B. van Merriënboer, D. Bahdanau, and Y. Bengio, 2014a. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. *arXiv e-prints*, arXiv:1409.1259. 10.48550/arXiv.1409.1259, [1409.1259](https://arxiv.org/abs/1409.1259).
- Cho, K., B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, 2014b. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv e-prints*, arXiv:1406.1078. 10.48550/arXiv.1406.1078, [1406.1078](https://arxiv.org/abs/1406.1078).
- Choi, D., C. J. Shallue, Z. Nado, J. Lee, C. J. Maddison, and G. E. Dahl, 2019. On Empirical Comparisons of Optimizers for Deep Learning. *arXiv e-prints*, arXiv:1910.05446. 10.48550/arXiv.1910.05446, [1910.05446](https://arxiv.org/abs/1910.05446).
- Clerc, S., S. Brosse, and M. Chane-Yook, 2003. Sparcs: an advanced software for spacecraft charging analysis. 20–24.
- Coifman, R. R., S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, 2005. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the national academy of sciences*, **102**(21), 7426–7431.
- Cranmer, S. R., 2002. Coronal Holes and the High-Speed Solar Wind. *Space Science Reviews*, **101**(3), 229–294. 10.1023/A:1020840004535.
- Cranmer, S. R., 2009. Coronal Holes. *Living Reviews in Solar Physics*, **6**(1), 3. 10.12942/lrsp-2009-3, URL <https://doi.org/10.12942/lrsp-2009-3>.
- Cravens, T. E., 1997. Physics of Solar System Plasmas. Cambridge Atmospheric and Space Science Series. Cambridge University Press. 10.1017/CBO9780511529467.
- Cummings, W. D., and A. J. Dessler, 1967. Field-aligned currents in the magnetosphere. *Journal of Geophysical Research (1896-1977)*, **72**(3), 1007–1013. [Doi.org/10.1029/JZ072i003p01007](https://doi.org/10.1029/JZ072i003p01007), <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/JZ072i003p01007>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/JZ072i003p01007>.

-
- Cunningham, P., M. Cord, and S. J. Delany. Supervised Learning, 21–49. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008. ISBN 978-3-540-75171-7. 10.1007/978-3-540-75171-7_2, URL https://doi.org/10.1007/978-3-540-75171-7_2.
- Daglis, I. A., L. C. Chang, S. Dasso, N. Gopalswamy, O. V. Khabarova, et al., 2021. Predictability of variable solar–terrestrial coupling. *Annales Geophysicae*, **39**(6), 1013–1035. 10.5194/angeo-39-1013-2021, URL <https://angeo.copernicus.org/articles/39/1013/2021/>.
- Daglis, I. A., R. M. Thorne, W. Baumjohann, and S. Orsini, 1999. The terrestrial ring current: Origin, formation, and decay. *Reviews of Geophysics*, **37**(4), 407–438. <https://doi.org/10.1029/1999RG900009>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/1999RG900009>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/1999RG900009>.
- Delcroix, J.-L., and A. Bers, 1994. Physique des plasmas (Vol. I). EDP Sciences, Les Ulis. ISBN 9782759802876. Doi:10.1051/978-2-7598-0287-6, URL <https://doi.org/10.1051/978-2-7598-0287-6>.
- Dickinson, L. G., 1974. Defense meteorological satellite program (DMSP) user’s guide, vol. 74. AWS.
- Dombeck, J., C. Cattell, N. Prasad, E. Meeker, E. Hanson, and J. McFadden, 2018. Identification of Auroral Electron Precipitation Mechanism Combinations and Their Relationships to Net Downgoing Energy and Number Flux. *Journal of Geophysical Research: Space Physics*, **123**(12), 10,064–10,089. <https://doi.org/10.1029/2018JA025749>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2018JA025749>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018JA025749>.
- Doornbos, E., 2012. Thermospheric density and wind determination from satellite dynamics. Springer Science & Business Media.
- Drinkwater, M. R., R. Floberghagen, R. Haagmans, D. Muzi, and A. Popescu. GOCE: ESA’s First Earth Explorer Core Mission, 419–432. Springer Netherlands, Dordrecht, 2003. ISBN 978-94-017-1333-7. 10.1007/978-94-017-1333-7_36, URL https://doi.org/10.1007/978-94-017-1333-7_36.
- Dubey, S. R., S. K. Singh, and B. B. Chaudhuri, 2022. Activation functions in deep learning: A comprehensive survey and benchmark. *Neurocomputing*, **503**, 92–108. <https://doi.org/10.1016/j.neucom.2022.06.111>, URL <https://www.sciencedirect.com/science/article/pii/S0925231222008426>.
- Duchi, J., E. Hazan, and Y. Singer, 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, **12**(7).
- Dudok de Wit, T., and S. Bruinsma, 2017. The 30 cm radio flux as a solar proxy for thermosphere density modelling. *Journal of Space Weather and Space Climate*, **7**, A9. 10.1051/swsc/2017008.
- Dudok de Wit, T., S. Bruinsma, and K. Shibasaki, 2014. Synoptic radio observations as proxies for upper atmosphere modelling. *Journal of Space Weather and Space Climate*, **4**, A06. 10.1051/swsc/2014003, [1402.3946](https://doi.org/10.1051/swsc/2014003).
- Dungey, J. W., 1961. Interplanetary Magnetic Field and the Auroral Zones. , **6**(2), 47–48. 10.1103/PhysRevLett.6.47.
-

- Ebihara, Y., and T. Tanaka, 2022. Where Is Region 1 Field-Aligned Current Generated? *Journal of Geophysical Research: Space Physics*, **127**(3), e2021JA029,991. E2021JA029991 2021JA029991, <https://doi.org/10.1029/2021JA029991>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2021JA029991>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021JA029991>.
- Elphic, R. C., J. W. Bonnell, R. J. Strangeway, L. Kepko, R. E. Ergun, et al., 1998. The auroral current circuit and field-aligned currents observed by FAST. *Geophysical Research Letters*, **25**(12), 2033–2036. <https://doi.org/10.1029/98GL01158>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/98GL01158>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/98GL01158>.
- Elsken, T., J. H. Metzen, and F. Hutter, 2019. Neural architecture search: A survey. *The Journal of Machine Learning Research*, **20**(1), 1997–2017.
- Fairfield, D. H., and J. D. Scudder, 1985. Polar rain: Solar coronal electrons in the Earth's magnetosphere. *Journal of Geophysical Research: Space Physics*, **90**(A5), 4055–4068. <https://doi.org/10.1029/JA090iA05p04055>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/JA090iA05p04055>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/JA090iA05p04055>.
- Fang, T.-W., A. Kubaryk, D. Goldstein, Z. Li, T. Fuller-Rowell, G. Millward, H. J. Singer, R. Steenburgh, S. Westerman, and E. Babcock, 2022. Space Weather Environment During the SpaceX Starlink Satellite Loss in February 2022. *Space Weather*, **20**(11), e2022SW003,193. E2022SW003193 2022SW003193, <https://doi.org/10.1029/2022SW003193>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2022SW003193>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2022SW003193>.
- Farquhar, R. W., 1969. The control and use of libration-point satellites. Stanford University.
- Feldstein, Y. I., and G. V. Starkov, 1967. Dynamics of auroral belt and polar geomagnetic disturbances. *Planetary and Space Science*, **15**(2), 209–229. URL <https://www.sciencedirect.com/science/article/pii/0032063367901900>.
- Fessant, F., S. Bengio, and D. Collobert, 1996. On the prediction of solar activity using different neural network models. *Annales Geophysicae*, **14**(1), 20–26. 10.1007/s00585-996-0020-z.
- Fitzpatrick, R., 2011. Second Adiabatic Invariant. [Online; last accessed September 2023], URL <https://farside.ph.utexas.edu/teaching/plasma/lectures/node24.html>.
- Friis-Christensen, E., H. Lühr, D. Knudsen, and R. Haagmans, 2008. Swarm – An Earth Observation Mission investigating Geospace. *Advances in Space Research*, **41**(1), 210–216. <https://doi.org/10.1016/j.asr.2006.10.008>, URL <https://www.sciencedirect.com/science/article/pii/S0273117706005497>.
- Fuller-Rowell, T. J., R. A. Akmaev, F. Wu, A. Anghel, N. Maruyama, et al., 2008. Impact of terrestrial weather on the upper atmosphere. *Geophysical Research Letters*, **35**(9). <https://doi.org/10.1029/2007GL032911>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2007GL032911>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2007GL032911>.
- Fuller-Rowell, T. J., and D. S. Evans, 1987. Height-integrated Pedersen and Hall conductivity patterns inferred from the TIROS-NOAA satellite data. *Journal of Geophysical Research*, **92**(A7), 7606–7618. 10.1029/JA092iA07p07606.

-
- Geisser, S., 1975. The predictive sample reuse method with applications. *Journal of the American statistical Association*, **70**(350), 320–328.
- Ghahramani, Z. Unsupervised Learning, 72–112. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004. ISBN 978-3-540-28650-9. 10.1007/978-3-540-28650-9_5, URL https://doi.org/10.1007/978-3-540-28650-9_5.
- Gleisner, H., H. Lundstedt, and P. Wintoft, 1996. Predicting geomagnetic storms from solar-wind data using time-delay neural networks. *Annales Geophysicae*, **14**(7), 679–686. 10.1007/s00585-996-0679-1.
- Golub, L., and J. M. Pasachoff, 1997. The Solar Corona. Cambridge University Press.
- Gomez Toro, D., M. Arzel, F. Seguin, and M. Jézéquel, 2014. Soft Error Detection and Correction Technique for Radiation Hardening Based on C-element and BICS. *IEEE Transactions on Circuits and Systems II: Express Briefs*, **61**(12), 952–956. 10.1109/TCSII.2014.2356911.
- Gonzalez, W., and E. Parker, 2016. Magnetic reconnection. *Astrophysics and space science library*, **427**, 542.
- Goodfellow, I., Y. Bengio, and A. Courville, 2016. Deep Learning. MIT Press. <http://www.deeplearningbook.org>.
- Gopalswamy, N., L. Barbieri, E. W. Cliver, G. Lu, S. P. Plunkett, and R. M. Skoug, 2005. Introduction to violent Sun-Earth connection events of October–November 2003. *Journal of Geophysical Research: Space Physics*, **110**(A9). <https://doi.org/10.1029/2005JA011268>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2005JA011268>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2005JA011268>.
- Gosling, J. T., D. N. Baker, S. J. Bame, W. C. Feldman, R. D. Zwickl, and E. J. Smith, 1985. North-south and dawn-dusk plasma asymmetries in the distant tail lobes: ISEE 3. *Journal of Geophysical Research: Space Physics*, **90**(A7), 6354–6360. <https://doi.org/10.1029/JA090iA07p06354>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/JA090iA07p06354>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/JA090iA07p06354>.
- Gosling, J. T., D. N. Baker, S. J. Bame, and R. D. Zwickl, 1986. Bidirectional solar wind electron heat flux and hemispherically symmetric polar rain. *Journal of Geophysical Research: Space Physics*, **91**(A10), 11,352–11,358. <https://doi.org/10.1029/JA091iA10p11352>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/JA091iA10p11352>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/JA091iA10p11352>.
- Gruet, M., 2018. Intelligence artificielle et prévision de l’impact de l’activité solaire sur l’environnement magnétique terrestre. Ph.D. thesis. Thèse de doctorat dirigée par Sicard, Angélica et Rochel, Sandrine Astrophysique, sciences de l’espace, planétologie Toulouse, ISAE 2018, URL <http://www.theses.fr/2018ESAE0014>.
- Güdel, M., 2007. The Sun in Time: Activity and Environment. *Living Reviews in Solar Physics*, **4**(1), 3. 10.12942/lrsp-2007-3, URL <https://doi.org/10.12942/lrsp-2007-3>.
- Gupta, A., 2023. A comprehensive guide on Optimizers in deep learning. [Online; last accessed on July 2023], URL <https://www.analyticsvidhya.com/blog/2021/10/a-comprehensive-guide-on-deep-learning-optimizers/>.
- Hairy, P., 2021. Les Réseaux de Neurones récurrents pour les Séries Temporelles. [Online; last accessed July 2023], URL <https://metalblog.ctif.com/2021/09/06/les-reseaux-de-neurones-recurrents-pour-les-series-temporelles/>.
-

- Hardy, D. A., M. S. Gussenhoven, and E. Holeman, 1985. A statistical model of auroral electron precipitation. *Journal of Geophysical Research*, **90**(A5), 4229–4248. 10.1029/JA090iA05p04229.
- Hardy, D. A., M. S. Gussenhoven, R. Raistrick, and W. J. McNeil, 1987. Statistical and functional representations of the pattern of auroral energy flux, number flux, and conductivity. *Journal of Geophysical Research*, **92**(A11), 12,275–12,294. 10.1029/JA092iA11p12275.
- Hardy, D. A., E. G. Holeman, W. J. Burke, L. C. Gentile, and K. H. Bounar, 2008. Probability distributions of electron precipitation at high magnetic latitudes. *Journal of Geophysical Research (Space Physics)*, **113**(A6), A06305. 10.1029/2007JA012746.
- Harra, L., A. F. Battaglia, K. Barczynski, H. Collier, S. Krucker, K. K. Reeves, and G. Doschek, 2023. How Hot Can Small Solar Flares Get? *Solar Physics*, **298**(1), 13. 10.1007/s11207-022-02106-1, URL <https://doi.org/10.1007/s11207-022-02106-1>.
- Harvey, J. W., and J. Sheeley, N. R., 1979. Coronal Holes and Solar Magnetic Fields (Article published in the special issues: Proceedings of the Symposium on Solar Terrestrial Physics held in Innsbruck, May- June 1978. (pp. 137-538)). *Space Science Reviews*, **23**(2), 139–158. 10.1007/BF00173808.
- Hathaway, D. H., 2015. The Solar Cycle. *Living Reviews in Solar Physics*, **12**(1), 4. 10.1007/lrsp-2015-4, URL <https://doi.org/10.1007/lrsp-2015-4>.
- Heidrich-Meisner, V., and R. F. Wimmer-Schweingruber, 2018. Chapter 16 - Solar Wind Classification Via k-Means Clustering Algorithm. In E. Camporeale, S. Wing, and J. R. Johnson, eds., *Machine Learning Techniques for Space Weather*, 397–424. Elsevier. ISBN 978-0-12-811788-0. <https://doi.org/10.1016/B978-0-12-811788-0.00016-0>, URL <https://www.sciencedirect.com/science/article/pii/B9780128117880000160>.
- Helmboldt, J. F., 2020. The Properties and Origins of Corotating Plasmaspheric Irregularities: Part II—Tomography With Compact Arrays of GPS Receivers. *Journal of Geophysical Research: Space Physics*, **125**(6), e2020JA027,858. E2020JA027858 10.1029/2020JA027858, <https://doi.org/10.1029/2020JA027858>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2020JA027858>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020JA027858>.
- Heynderickx, D., B. Quaghebeur, E. Speelman, and E. Daly. ESA's Space Environment Information System (SPENVIS) - A WWW interface to models of the space environment and its effects, 2000. 10.2514/6.2000-371, <https://arc.aiaa.org/doi/pdf/10.2514/6.2000-371>, URL <https://arc.aiaa.org/doi/abs/10.2514/6.2000-371>.
- Hinton, G. E., and S. Roweis, 2002. Stochastic neighbor embedding. *Advances in neural information processing systems*, **15**.
- Hinton, G. E., and R. R. Salakhutdinov, 2006. Reducing the Dimensionality of Data with Neural Networks. *Science*, **313**(5786), 504–507. 10.1126/science.1127647, <https://www.science.org/doi/pdf/10.1126/science.1127647>, URL <https://www.science.org/doi/abs/10.1126/science.1127647>.
- Hochreiter, S., and J. Schmidhuber, 1997. Long Short-Term Memory. *Neural Computation*, **9**(8), 1735–1780. 10.1162/neco.1997.9.8.1735.
- Hocke, K., and N. Kämpfer, 2009. Gap filling and noise reduction of unevenly sampled data by means of the Lomb-Scargle periodogram. *Atmospheric Chemistry & Physics*, **9**(12), 4197–4206. 10.5194/acp-9-4197-200910.5194/acpd-8-4603-2008.

-
- Hoeksema, J. T., 1984. Structure and Evolution of the Large Scale Solar and Heliospheric Magnetic Fields. Ph.D. thesis, Stanford University, California.
- Hones, J., Edward W., 1973. Plasma flow in the plasma sheet and its relation to substorms. *Radio Science*, **8**(11), 979–990. 10.1029/RS008i011p00979.
- Horwitz, J. L., D. L. Gallagher, and W. K. Peterson, 1998. Geospace mass and energy flow : results from the International Solar-Terrestrial Physics Program. *Geophysical Monograph Series*, **104**. 10.1029/GM104.
- Hu, A., E. Camporeale, and B. Swiger, 2023. Multi-Hour-Ahead Dst Index Prediction Using Multi-Fidelity Boosted Neural Networks. *Space Weather*, **21**(4), e2022SW003286. E2022SW003286 2022SW003286, <https://doi.org/10.1029/2022SW003286>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2022SW003286>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2022SW003286>.
- Huang, M.-H., and R. T. Rust, 2018. Artificial intelligence in service. *Journal of service research*, **21**(2), 155–172.
- Hudson, H., 2002. Coronal holes as seen in soft X-rays by Yohkoh. *From Solar Min to Max: Half a Solar Cycle with SOHO*, **508**, 341–349.
- Hui, C. H., and C. E. Seyler, 1992. Electron acceleration by Alfvén waves in the magnetosphere. *Journal of Geophysical Research: Space Physics*, **97**(A4), 3953–3963. <https://doi.org/10.1029/91JA03101>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/91JA03101>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/91JA03101>.
- IBM, 2021. What is unsupervised learning? [Online; last accessed September 2023], URL <https://www.ibm.com/topics/unsupervised-learning>.
- Iijima, T., and T. A. Potemra, 1976. The amplitude distribution of field-aligned currents at northern high latitudes observed by Triad. *Journal of Geophysical Research (1896-1977)*, **81**(13), 2165–2174. <https://doi.org/10.1029/JA081i013p02165>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/JA081i013p02165>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/JA081i013p02165>.
- Inceoglu, F., Y. Y. Shprits, S. G. Heinemann, and S. Bianco, 2022. Identification of Coronal Holes on AIA/SDO Images Using Unsupervised Machine Learning. *The Astrophysical Journal*, **930**(2), 118. 10.3847/1538-4357/ac5f43, URL <https://dx.doi.org/10.3847/1538-4357/ac5f43>.
- James, G., D. Witten, T. Hastie, and R. Tibshirani, 2021. An Introduction to Statistical Learning. Springer Texts in Statistics. Springer New York, NY, 2 edn. 10.1007/978-1-0716-1418-1.
- Jin, B., D. A. Lorenz, and S. Schiffler, 2009. Elastic-net regularization: error estimates and active set methods. *Inverse Problems*, **25**(11), 115,022. 10.1088/0266-5611/25/11/115022, URL <https://dx.doi.org/10.1088/0266-5611/25/11/115022>.
- Johnson-Groh, M., 2019. In Solar System’s Symphony, Earth’s Magnetic Field Drops the Beat. NASA. [Online; last accessed September 2023], URL <https://www.nasa.gov/feature/goddard/2019/in-solar-system-s-symphony-earth-s-magnetic-field-drops-the-beat>.
- Jones, H. P., 2005. Magnetic Fields and Flows in Open Magnetic Structures. In Large-scale Structures and Their Role in Solar Activity, vol. 346, 229.

- Jones, T., 2017. Models for machine learning. URL <https://developer.ibm.com/articles/cc-models-machine-learning/>.
- Jordanova, V. K., 2017. SHIELDS Final Technical Report. 10.2172/1396150, URL <https://www.osti.gov/biblio/1396150>.
- Kallio, E., and G. Facskó, 2015. Properties of plasma near the moon in the magnetotail. *Planetary and Space Science*, **115**, 69–76. Solar wind interaction with the terrestrial planets, <https://doi.org/10.1016/j.pss.2014.11.007>, URL <https://www.sciencedirect.com/science/article/pii/S0032063314003432>.
- Kamide, Y., 1992. Is substorm occurrence a necessary condition for a magnetic storm? *Journal of geomagnetism and geoelectricity*, **44**(2), 109–117.
- Kaplan, A., and M. Haenlein, 2019. Siri, Siri, in my hand: Who’s the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, **62**(1), 15–25. <https://doi.org/10.1016/j.bushor.2018.08.004>, URL <https://www.sciencedirect.com/science/article/pii/S0007681318301393>.
- Karlsson, T., L. Andersson, D. M. Gillies, K. Lynch, O. Marghitsu, N. Partamies, N. Sivadas, and J. Wu, 2020. Quiet, Discrete Auroral Arcs—Observations. *Space Science Reviews*, **216**(1), 16. 10.1007/s11214-020-0641-7, URL <https://doi.org/10.1007/s11214-020-0641-7>.
- Kasapis, S., L. Zhao, Y. Chen, X. Wang, M. Bobra, and T. Gombosi, 2022. Interpretable Machine Learning to Forecast SEP Events for Solar Cycle 23. *Space Weather*, **20**(2), e2021SW002,842. E2021SW002842 2021SW002842, <https://doi.org/10.1029/2021SW002842>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2021SW002842>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021SW002842>.
- Katz, I., D. E. Parks, M. J. Mandell, J. M. Harvey, S. S. Wang, and J. C. Roche, 1977. NASCAP, a Three-Dimensional Charging Analyzer Program for Complex Spacecraft. *IEEE Transactions on Nuclear Science*, **24**(6), 2276–2280. 10.1109/TNS.1977.4329206.
- Kilcommons, L. M., R. J. Redmon, and D. J. Knipp, 2017. A new DMSP magnetometer and auroral boundary data set and estimates of field-aligned currents in dynamic auroral boundary coordinates. *Journal of Geophysical Research: Space Physics*, **122**(8), 9068–9079. <https://doi.org/10.1002/2016JA023342>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2016JA023342>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016JA023342>.
- King, J., and N. Papitashvili, 2006. One min and 5-min solar wind data sets at the Earth’s bow shock nose. *NASA Goddard Space Flight Cent., Greenbelt, Md.*
- Kingma, D. P., and J. Ba, 2014. Adam: A Method for Stochastic Optimization. *arXiv e-prints*, arXiv:1412.6980. 10.48550/arXiv.1412.6980, 1412.6980.
- Kivelson, M. G., and C. T. Russell, 1995. Introduction to Space Physics. Cambridge University Press. 10.1017/9781139878296.
- Kletzing, C. A., W. S. Kurth, M. Acuna, R. J. MacDowall, R. B. Torbert, et al., 2013. The Electric and Magnetic Field Instrument Suite and Integrated Science (EMFISIS) on RBSP. *Space Science Reviews*, **179**(1), 127–181. 10.1007/s11214-013-9993-6, URL <https://doi.org/10.1007/s11214-013-9993-6>.
- Knudsen, D. J., J. E. Borovsky, T. Karlsson, R. Kataoka, and N. Partamies, 2021. Editorial: Topical Collection on Auroral Physics. *Space Science Reviews*, **217**(1), 19. 10.1007/s11214-021-00798-8, URL <https://doi.org/10.1007/s11214-021-00798-8>.

-
- Kodheli, O., E. Lagunas, N. Maturo, S. K. Sharma, B. Shankar, et al., 2021. Satellite Communications in the New Space Era: A Survey and Future Challenges. *IEEE Communications Surveys Tutorials*, **23**(1), 70–109. 10.1109/COMST.2020.3028247.
- Kohl, J., and S. Cranmer, 1999. Coronal Holes and Solar Wind Acceleration. Springer Science & Business Media.
- Kondrashov, D., Y. Shprits, and M. Ghil, 2010. Gap filling of solar wind data by singular spectrum analysis. *Geophysical Research Letters*, **37**(15), L15101. 10.1029/2010GL044138.
- Koons, H. C., and J. F. Fennell, 2006. Space weather effects on communications satellites. *URSI Radio Science Bulletin*, **2006**(316), 27–41. 10.23919/URSIRSB.2006.7909358.
- Kornfeld, R. P., B. W. Arnold, M. A. Gross, N. T. Dahya, W. M. Klipstein, P. F. Gath, and S. Bettadpur, 2019. GRACE-FO: The Gravity Recovery and Climate Experiment Follow-On Mission. *Journal of Spacecraft and Rockets*, **56**(3), 931–951. 10.2514/1.A34326, <https://doi.org/10.2514/1.A34326>, URL <https://doi.org/10.2514/1.A34326>.
- Koskinen, H. E., and E. K. Kilpua, 2022. Physics of Earth’s radiation belts: Theory and observations. Springer International Publishing. ISBN 978-3-030-82167-8. 10.1007/978-3-030-82167-8, URL <https://doi.org/10.1007/978-3-030-82167-8>.
- Kotsiantis, S., D. Kanellopoulos, and P. Pintelas, 2005. Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, **30**, 25–36.
- Krizhevsky, A., I. Sutskever, and G. E. Hinton, 2012. ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, eds., *Advances in Neural Information Processing Systems*, vol. 25. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.
- Kroll, A., 2000. Grey-box models: Concepts and application. *New frontiers in computational intelligence and its applications*, **57**, 42–51.
- Lai, S. T., 2011. Spacecraft charging. American Institute of Aeronautics and Astronautics.
- Lakhina, G. S., S. Alex, S. Mukherjee, and G. Vichare, 2006. On magnetic storms and substorms. In N. Gopalswamy and A. Bhattacharyya, eds., *Proceedings of the ILWS Workshop*, 320.
- LeCun, Y., Y. Bengio, and G. Hinton, 2015. Deep learning. *Nature*, **521**(7553), 436–444. 10.1038/nature14539, URL <https://doi.org/10.1038/nature14539>.
- Lecun, Y., L. Bottou, Y. Bengio, and P. Haffner, 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, **86**(11), 2278–2324. 10.1109/5.726791.
- Lee, S., E.-Y. Ji, Y.-J. Moon, and E. Park, 2021. One-Day Forecasting of Global TEC Using a Novel Deep Learning Model. *Space Weather*, **19**(1), 2020SW002,600. 2020SW002600 2020SW002600, <https://doi.org/10.1029/2020SW002600>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2020SW002600>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020SW002600>.
- Lilensten, J., and P.-L. Blelly, 2008. Du soleil à la Terre. EDP Sciences, Les Ulis. ISBN 9782868834676. Doi:10.1051/978-2-86883-467-6, URL <https://doi.org/10.1051/978-2-86883-467-6>.
- Lilensten, J., and J. Bornarel, 2001. Sous les feux du Soleil: Vers une météorologie de l’espace. EDP Sciences.
-

- Lilensten, J., J. P. C. Marques, M. Gruet, and F. Pitout, 2021. Météorologie de l'espace: vivre demain avec notre Soleil. De Boeck Supérieur.
- Lilensten, Jean, Dumbović, Mateja, Spogli, Luca, Belehaki, Anna, Van der Linden, Ronald, et al., 2021. Quo vadis, European Space Weather community? *J. Space Weather Space Clim.*, **11**, 26. 10.1051/swsc/2021009, URL <https://doi.org/10.1051/swsc/2021009>.
- Lilley Jr, J. R., D. L. Cooke, G. A. Jongeward, and I. Katz, 1989. POLAR User's Manual. *Tech. rep.*, S-CUBED LA JOLLA CA.
- Lin, J.-W., 2021. Geomagnetic Storm Related to Disturbance Storm Time Indices: Geomagnetic Storm. *European Journal of Environment and Earth Sciences*, **2**(6), 1–3.
- Liou, K., T. Sotirelis, and E. J. Mitchell, 2018. North-South Asymmetry in the Geographic Location of Auroral Substorms correlated with Ionospheric Effects. *Scientific Reports*, **8**(1), 17,230. 10.1038/s41598-018-35091-2, URL <https://doi.org/10.1038/s41598-018-35091-2>.
- Liu, H., K. Simonyan, O. Vinyals, C. Fernando, and K. Kavukcuoglu, 2017. Hierarchical Representations for Efficient Architecture Search. *arXiv e-prints*, arXiv:1711.00436. 10.48550/arXiv.1711.00436, [1711.00436](https://arxiv.org/abs/1711.00436).
- Liu, L., S. Zou, Y. Yao, and Z. Wang, 2020. Forecasting Global Ionospheric TEC Using Deep Learning Approach. *Space Weather*, **18**(11), e2020SW002,501. E2020SW002501 2020SW002501, <https://doi.org/10.1029/2020SW002501>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2020SW002501>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020SW002501>.
- Lui, A. T. Y., 1991. A synthesis of magnetospheric substorm models. *Journal of Geophysical Research*, **96**(A2), 1849–1856. 10.1029/90JA02430.
- Lundstedt, H., and P. Wintoft, 1994. Prediction of geomagnetic storms from solar wind data with the use of a neural network. *Annales Geophysicae*, **12**(1), 19–24. 10.1007/s00585-994-0019-2.
- Lütkepohl, H., and F. Xu, 2012. The role of the log transformation in forecasting economic variables. *Empirical Economics*, **42**(3), 619–638. 10.1007/s00181-010-0440-1, URL <https://doi.org/10.1007/s00181-010-0440-1>.
- Lyu, Y., H. Li, M. Sayagh, Z. M. J. Jiang, and A. E. Hassan, 2021. An Empirical Study of the Impact of Data Splitting Decisions on the Performance of AIOps Solutions. *ACM Trans. Softw. Eng. Methodol.*, **30**(4). 10.1145/3447876, URL <https://doi.org/10.1145/3447876>.
- Machol, J. L., J. C. Green, R. J. Redmon, R. A. Viereck, and P. T. Newell, 2012. Evaluation of OVATION Prime as a forecast model for visible aurorae. *Space Weather*, **10**(3). <https://doi.org/10.1029/2011SW000746>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2011SW000746>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2011SW000746>.
- Macpherson, K. P., A. J. Conway, and J. C. Brown, 1995. Prediction of solar and geomagnetic activity data using neural networks. *Journal of Geophysical Research*, **100**(A11), 21,735–21,744. 10.1029/95JA02283.
- Maehara, H., T. Shibayama, S. Notsu, Y. Notsu, T. Nagao, S. Kusaba, S. Honda, D. Nogami, and K. Shibata, 2012. Superflares on solar-type stars. *Nature*, **485**(7399), 478–481. 10.1038/nature11063, URL <https://doi.org/10.1038/nature11063>.

- Maharana, K., S. Mondal, and B. Nemade, 2022. A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings*, **3**(1), 91–99. International Conference on Intelligent Engineering Approach(ICIEA-2022), <https://doi.org/10.1016/j.gltp.2022.04.020>, URL <https://www.sciencedirect.com/science/article/pii/S2666285X22000565>.
- Makarov, G., 2022. GEOMAGNETIC INDICES ASY-H AND SYM-H AND THEIR RELATION TO INTERPLANETARY PARAMETERS. *Solar-Terrestrial Physics*, **8**(4), 36–43.
- Mandell, M., and V. Davis, 1990. User’s Guide to NASCAP/LEO. *Tech. rep.*, S-CUBED. Division of. Maxwell Laboratories inc.
- Mandell, M., V. Davis, B. Gardner, I. Mikellides, D. Cooke, and J. Minor, 2004. NASCAP-2K: An overview.
- Margolies, D. L., and T. von Roseninge, 1998. Advanced Composition Explorer (ACE): Lessons learned and final report. *NASA/Goddard (Greenbelt, MD)*.
- Marlowe, H. R., 2022. An unsupervised learning approach to superstorm signature identification in precipitating particle data. In *Proceedings of the 2nd Machine Learning in Heliophysics*, 10.
- Mathews, G., and S. S. Towheed, 1995. NSSDC OMNIWeb: The first space physics WWW-based data browsing and retrieval system. *Computer Networks and ISDN Systems*, **27**(6), 801–808. Proceedings of the Third International World-Wide Web Conference, [https://doi.org/10.1016/0169-7552\(95\)00033-4](https://doi.org/10.1016/0169-7552(95)00033-4), URL <https://www.sciencedirect.com/science/article/pii/0169755295000334>.
- Maynard, T., N. Smith, and S. Gonzalez, 2013. Solar storm risk to the North American electric grid. *Lloyd’s*, **1**(11).
- McGranaghan, R. M., 2016. Determining global ionospheric conductivity in the satellite and data assimilation age and assessing its influence on the magnetosphere-ionosphere-thermosphere system. Ph.D. thesis, University of Colorado at Boulder.
- McGranaghan, R. M., 2019. Eight lessons I learned leading a scientific “design sprint”. *Eos*, **100**. <https://doi.org/10.1029/2019EO136427>.
- McGranaghan, R. M., J. Ziegler, T. Bloch, S. Hatch, E. Camporeale, K. Lynch, M. Owens, J. Gjerloev, B. Zhang, and S. Skone, 2021. Toward a Next Generation Particle Precipitation Model: Mesoscale Prediction Through Machine Learning (a Case Study and Framework for Progress). *Space Weather*, **19**(6), e2020SW002,684. E2020SW002684 2020SW002684, <https://doi.org/10.1029/2020SW002684>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2020SW002684>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020SW002684>.
- McKay, A. J., 2004. Geoelectric fields and geomagnetically induced currents in the United Kingdom.
- Mikaelian, T., 2009. Spacecraft charging and hazards to electronics in space. *arXiv preprint arXiv:0906.3884*.
- Millward, G., R. Moffett, S. Quegan, and T. Fuller-Rowell, 1996. A coupled thermosphere-ionosphere—plasmasphere model, CTIP, STEP Handbook on Ionospheric Models. *RW Schunk*, 239–280.
- Miyake, F., K. Nagaya, K. Masuda, and T. Nakamura, 2012. A signature of cosmic-ray increase in ad 774–775 from tree rings in Japan. *Nature*, **486**(7402), 240–242. 10.1038/nature11123, URL <https://doi.org/10.1038/nature11123>.

- Molnar, C., 2023. Interpretable machine learning. [Online, last accessed June 2023], URL <https://christophm.github.io/interpretable-ml-book/>.
- Monigatti, L., 2022. A visual guide to learning rate schedulers in pytorch. [Online; last accessed on July 4, 2023], URL <https://towardsdatascience.com/a-visual-guide-to-learning-rate-schedulers-in-pytorch-24bbb262c863>.
- Muranaka, T., S. Hosoda, J.-H. Kim, S. Hatta, K. Ikeda, et al., 2008. Development of Multi-Utility Spacecraft Charging Analysis Tool (MUSCAT). *IEEE Transactions on Plasma Science*, **36**(5), 2336–2349. 10.1109/TPS.2008.2003974.
- Nakamura, M., A. Yoneda, M. Oda, and K. Tsubouchi, 2015. Statistical analysis of extreme auroral electrojet indices. *Earth, Planets and Space*, **67**(1), 153. 10.1186/s40623-015-0321-0, URL <https://doi.org/10.1186/s40623-015-0321-0>.
- NASA, 2013. Radiation Belts with Satellites. [Online; last accessed June 2023], URL https://www.nasa.gov/mission_pages/sunearth/news/gallery/20130228-radiationbelts.html.
- Newell, P. T., K. Liou, and G. R. Wilson, 2009. Polar cap particle precipitation and aurora: Review and commentary. *Journal of Atmospheric and Solar-Terrestrial Physics*, **71**(2), 199–215. <https://doi.org/10.1016/j.jastp.2008.11.004>, URL <https://www.sciencedirect.com/science/article/pii/S1364682608003738>.
- Newell, P. T., K. Liou, Y. Zhang, T. Sotirelis, L. J. Paxton, and E. J. Mitchell, 2014. OVA-TION Prime-2013: Extension of auroral precipitation model to higher disturbance levels. *Space Weather*, **12**(6), 368–379. <https://doi.org/10.1002/2014SW001056>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2014SW001056>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2014SW001056>.
- Newell, P. T., K. Liou, Y. Zhang, T. S. Sotirelis, L. J. Paxton, and E. J. Mitchell. Auroral Precipitation Models and Space Weather, chap. 18, 275–290. American Geophysical Union (AGU), 2015. ISBN 9781118978719. <https://doi.org/10.1002/9781118978719.ch18>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/9781118978719.ch18>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/9781118978719.ch18>.
- Newell, P. T., T. Sotirelis, J. M. Ruohoniemi, J. F. Carbary, K. Liou, J. P. Skura, C. I. Meng, C. Deehr, D. Wilkinson, and F. J. Rich, 2002. OVATION: Oval variation, assessment, tracking, intensity, and online nowcasting. *Annales Geophysicae*, **20**(7), 1039–1047. 10.5194/angeo-20-1039-2002.
- Newell, P. T., T. Sotirelis, and S. Wing, 2010. Seasonal variations in diffuse, monoenergetic, and broadband aurora. *Journal of Geophysical Research: Space Physics*, **115**(A3). <https://doi.org/10.1029/2009JA014805>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2009JA014805>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2009JA014805>.
- Ngwira, C. M., A. Pulkkinen, M. Leila Mays, M. M. Kuznetsova, A. B. Galvin, K. Simunac, D. N. Baker, X. Li, Y. Zheng, and A. Glocer, 2013. Simulation of the 23 July 2012 extreme space weather event: What if this extremely rare CME was Earth directed? *Space Weather*, **11**(12), 671–679. <https://doi.org/10.1002/2013SW000990>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2013SW000990>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2013SW000990>.

-
- Ni, B., R. M. Thorne, X. Zhang, J. Bortnik, Z. Pu, et al., 2016. Origins of the Earth's Diffuse Auroral Precipitation. *Space Science Reviews*, **200**(1), 205–259. 10.1007/s11214-016-0234-7, URL <https://doi.org/10.1007/s11214-016-0234-7>.
- Nichols, D. A., 1975. The Defense Meteorological Satellite Program. *Optical Engineering*, **14**(4), 144,273. 10.1117/12.7971832, URL <https://doi.org/10.1117/12.7971832>.
- Nishimura, Y., M. R. Lessard, Y. Katoh, Y. Miyoshi, E. Grono, et al., 2020. Diffuse and Pulsating Aurora. *Space Science Reviews*, **216**(1), 4. 10.1007/s11214-019-0629-3, URL <https://doi.org/10.1007/s11214-019-0629-3>.
- Nita, G., M. Georgoulis, I. Kitiashvili, V. Sadykov, E. Camporeale, et al., 2020. Machine learning in heliophysics and space weather forecasting: a white paper of findings and recommendations. *arXiv preprint arXiv:2006.12224*. 10.48550/arXiv.2006.12224.
- Northrop, T. G., and E. Teller, 1960. Stability of the Adiabatic Motion of Charged Particles in the Earth's Field. *Phys. Rev.*, **117**, 215–225. 10.1103/PhysRev.117.215, URL <https://link.aps.org/doi/10.1103/PhysRev.117.215>.
- Ofman, L., 2005. MHD Waves and Heating in Coronal Holes. *Space Science Reviews*, **120**(1-2), 67–94. 10.1007/s11214-005-5098-1.
- Okewu, E., P. Adewole, and O. Sennaiké, 2019. Experimental Comparison of Stochastic Optimizers in Deep Learning. In S. Misra, O. Gervasi, B. Murgante, E. Stankova, V. Korkhov, C. Torre, A. M. A. Rocha, D. Taniar, B. O. Apduhan, and E. Tarantino, eds., *Computational Science and Its Applications – ICCSA 2019*, 704–715. Springer International Publishing, Cham. ISBN 978-3-030-24308-1.
- Olshevsky, V., Y. V. Khotyaintsev, A. Lalti, A. Divin, G. L. Delzanno, et al., 2021. Automated Classification of Plasma Regions Using 3D Particle Energy Distributions. *Journal of Geophysical Research: Space Physics*, **126**(10), e2021JA029620. E2021JA029620 2021JA029620, <https://doi.org/10.1029/2021JA029620>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2021JA029620>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021JA029620>.
- Orcinha, M., N. Tomassetti, F. Barão, and B. Bertucci, 2019. Observation of a time lag in solar modulation of cosmic rays in the heliosphere. *Journal of Physics: Conference Series*, **1181**(1), 012,013. 10.1088/1742-6596/1181/1/012013, URL <https://dx.doi.org/10.1088/1742-6596/1181/1/012013>.
- Palmroth, M., M. Grandin, T. Sarris, E. Doornbos, S. Tourgaidis, et al., 2021. Lower-thermosphere–ionosphere (LTI) quantities: current status of measuring techniques and models. *Annales Geophysicae*, **39**(1), 189–237. 10.5194/angeo-39-189-2021, URL <https://angeo.copernicus.org/articles/39/189/2021/>.
- Papitashvili, N., D. Bilitza, and J. King, 2014. OMNI: A Description of Near-Earth Solar Wind Environment. In 40th COSPAR Scientific Assembly, vol. 40, C0.1–12–14.
- Park, W., J. Lee, K.-C. Kim, J. Lee, K. Park, et al., 2021. Operational Dst index prediction model based on combination of artificial neural network and empirical model. *Journal of Space Weather and Space Climate*, **11**, 38. 10.1051/swsc/2021021.
- Parker, E. N., 1959. Solar winds. Technical Information Division, Sandia Corporation; available from the Office of Technical Services, Department of Commerce, Washington.
-

- Parker, E. N., 1991. Heating Solar Coronal Holes. *Astrophysical Journal*, **372**, 719. 10.1086/170015.
- Parker, L. N., 2017. Surface Charging Overview. [Online; last accessed June 2023], URL <https://cpaess.ucar.edu/sites/default/files/meetings/2017/documents/Parker-Linda-Surface-Charging.pdf>.
- Partamies, N., M. Syrjäso, E. Donovan, M. Connors, D. Charrois, D. Knudsen, and Z. Kryzanowsky, 2010. Observations of the auroral width spectrum at kilometre-scale size. *Annales Geophysicae*, **28**(3), 711–718. 10.5194/angeo-28-711-2010, URL <https://angeo.copernicus.org/articles/28/711/2010/>.
- Pasco, X., 2017. Le nouvel âge spatial. De la Guerre froide au New Space. CNRS.
- Petschek, H. E., 1964. 50 MAGNETIC FIELD ANNIHILATION. In Proceedings of a Symposium Held at the Goddard Space Flight Center, Greenbelt, Maryland, October 28-30, 1963, vol. 50, 425.
- Pham, H., M. Guan, B. Zoph, Q. Le, and J. Dean, 2018. Efficient Neural Architecture Search via Parameters Sharing. In J. Dy and A. Krause, eds., Proceedings of the 35th International Conference on Machine Learning, vol. 80 of *Proceedings of Machine Learning Research*, 4095–4104. PMLR. URL <https://proceedings.mlr.press/v80/pham18a.html>.
- Picone, J. M., A. E. Hedin, D. P. Drob, and A. C. Aikin, 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. <https://doi.org/10.1029/2002JA009430>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2002JA009430>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2002JA009430>.
- Piper, L. G., B. D. Green, W. A. M. Blumberg, and S. J. Wolnik, 1986. Electron impact excitation of the N₂⁺ Meinel band. *Journal of Physics B: Atomic and Molecular Physics*, **19**(20), 3327. 10.1088/0022-3700/19/20/015, URL <https://dx.doi.org/10.1088/0022-3700/19/20/015>.
- Pisacane, V. L., 2008. The Space Environment and its Effects on Space Systems. American Institute of Aeronautics and Astronautics.
- Pneuman, G. W., and R. A. Kopp, 1971. Gas-Magnetic Field Interactions in the Solar Corona. *Solar Physics*, **18**(2), 258–270. 10.1007/BF00145940.
- Poivey, C., 2019. Radiation effects in space electronics. [Online; last accessed June 2023], URL https://indico.cern.ch/event/777129/contributions/3249529/attachments/1844695/3026130/6th_EIROforum_school_on_instrumentation_cpoivey.pdf.
- Priest, E., 2014. Magnetohydrodynamics of the Sun. Cambridge University Press. 10.1017/CBO9781139020732.
- Pulkkinen, A., E. Bernabeu, A. Thomson, A. Viljanen, R. Pirjola, et al., 2017. Geomagnetically induced currents: Science, engineering, and applications readiness. *Space Weather*, **15**(7), 828–856. <https://doi.org/10.1002/2016SW001501>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2016SW001501>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016SW001501>.
- Pulkkinen, T. I., E. I. Tanskanen, A. Viljanen, N. Partamies, and K. Kauristie, 2011. Auroral electrojets during deep solar minimum at the end of solar cycle 23. *Journal of Geophysical Research: Space Physics*, **116**(A4). <https://doi.org/10.1029/2010JA016098>, <https://agup>

-
- [ubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2010JA016098](https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2010JA016098), URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2010JA016098>.
- Qi, D., and A. J. Majda, 2020. Using machine learning to predict extreme events in complex systems. *Proceedings of the National Academy of Sciences*, **117**(1), 52–59.
- Quiñonero Candela, J., and Y. LeCun, 2016. Artificial Intelligence, revealed. URL <https://engineering.fb.com/2016/12/01/ml-applications/artificial-intelligence-revealed/>.
- Raschka, S., and V. Mirjalili, 2017. Python Machine Learning—Second Edition.
- Redmon, R. J., W. F. Denig, L. M. Kilcommons, and D. J. Knipp, 2017. New DMSP database of precipitating auroral electrons and ions. *Journal of Geophysical Research: Space Physics*, **122**(8), 9056–9067. <https://doi.org/10.1002/2016JA023339>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2016JA023339>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016JA023339>.
- Reep, J. W., and W. T. Barnes, 2021. Forecasting the Remaining Duration of an Ongoing Solar Flare. *Space Weather*, **19**(10), e2021SW002,754. 10.1029/2021SW002754.
- Reigber, C., P. Schwintzer, H. Lühr, et al., 1999. The CHAMP geopotential mission. *Boll. Geof. Teor. Appl.*, **40**, 285–289.
- Reiss, M. A., C. Möstl, R. L. Bailey, H. T. Rüdissler, U. V. Amerstorfer, T. Amerstorfer, A. J. Weiss, J. Hinterreiter, and A. Windisch, 2021. Machine Learning for Predicting the Bz Magnetic Field Component From Upstream in Situ Observations of Solar Coronal Mass Ejections. *Space Weather*, **19**(12), e2021SW002,859. E2021SW002859 2021SW002859, <https://doi.org/10.1029/2021SW002859>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2021SW002859>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021SW002859>.
- Ribeiro, M. T., S. Singh, and C. Guestrin, 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, 1135–1144. Association for Computing Machinery, New York, NY, USA. ISBN 9781450342322. 10.1145/2939672.2939778, URL <https://doi.org/10.1145/2939672.2939778>.
- Ridley, A. J., Y. Deng, and G. Tóth, 2006. The global ionosphere thermosphere model. *Journal of Atmospheric and Solar-Terrestrial Physics*, **68**(8), 839–864. 10.1016/j.jastp.2006.01.008.
- Riley, P., 2012. On the probability of occurrence of extreme space weather events. *Space Weather*, **10**(2). <https://doi.org/10.1029/2011SW000734>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2011SW000734>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2011SW000734>.
- Riley, P., J. A. Linker, J. Americo Gonzalez Esparza, L. Jian, C. Russell, and J. Luhmann, 2012. Interpreting some properties of CIRs and their associated shocks during the last two solar minima using global MHD simulations. *Journal of Atmospheric and Solar-Terrestrial Physics*, **83**, 11–21. Corotating Interaction Regions from Sun to Earth: Modeling their formation, evolution and geoeffectiveness, <https://doi.org/10.1016/j.jastp.2012.01.019>, URL <https://www.sciencedirect.com/science/article/pii/S1364682612000338>.
- Robert, E., 2023. Interpretation of satellites and ground-based data for the monitoring of low energies particles in auroral region in the frame of space weather. Ph.D. thesis, Université Grenoble-Alpes, UGA. Unpublished Thesis.
-

- Roble, R. G., E. C. Ridley, A. D. Richmond, and R. E. Dickinson, 1988. A coupled thermosphere/ionosphere general circulation model. *Geophysical Research Letters*, **15**(12), 1325–1328. 10.1029/GL015i012p01325.
- Roussel, J.-F., F. Rogier, G. Dufour, J.-C. Mateo-Velez, J. Forest, A. Hilgers, D. Rodgers, L. Girard, and D. Payan, 2008. SPIS Open-Source Code: Methods, Capabilities, Achievements, and Prospects. *IEEE Transactions on Plasma Science*, **36**(5), 2360–2368. 10.1109/TPS.2008.2002327.
- Roweis, S. T., and L. K. Saul, 2000. Nonlinear dimensionality reduction by locally linear embedding. *science*, **290**(5500), 2323–2326.
- Ruder, S., 2016. An overview of gradient descent optimization algorithms. *arXiv e-prints*, arXiv:1609.04747. 10.48550/arXiv.1609.04747, 1609.04747.
- Russell, C., 2000. The polar cusp. *Advances in Space Research*, **25**(7), 1413–1424. Proceedings of the DO.1 Symposium of COSPAR Scientific Commission D, [https://doi.org/10.1016/S0273-1177\(99\)00653-5](https://doi.org/10.1016/S0273-1177(99)00653-5), URL <https://www.sciencedirect.com/science/article/pii/S0273117799006535>.
- Russell, C. T., J. G. Luhmann, and R. J. Strangeway, 2016. *Space Physics: An Introduction*. Cambridge University Press. 10.1017/9781316162590.
- Samara, E., E. Chane, B. Laperre, C. Verbeke, M. Temmer, L. Rodriguez, J. Magdalenic, and S. Poedts, 2021. The Dynamic Time Warping as a means to assess modeled solar wind time series. In EGU General Assembly Conference Abstracts, EGU General Assembly Conference Abstracts, EGU21–12,459. 10.5194/egusphere-egu21-12459.
- Sandhu, J. K., M.-T. Walach, H. Allison, and C. Watt, 2019. A global view of storms and substorms. *Astronomy and Geophysics*, **60**(3), 3.13–3.19. 10.1093/astrogeo/atz144.
- Schatten, K. H., J. M. Wilcox, and N. F. Ness, 1969. A model of interplanetary and coronal magnetic fields. *Solar Physics*, **6**(3), 442–455. 10.1007/BF00146478, URL <https://doi.org/10.1007/BF00146478>.
- Schneider, F., L. Balles, and P. Hennig, 2019. DeepOBS: A Deep Learning Optimizer Benchmark Suite. *arXiv e-prints*, arXiv:1903.05499. 10.48550/arXiv.1903.05499, 1903.05499.
- Schrijver, C. J., and G. L. Siscoe, 2010. *Heliophysics 3 Volume Set*.
- Schroff, F., D. Kalenichenko, and J. Philbin, 2015. FaceNet: A Unified Embedding for Face Recognition and Clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
- Schumaker, T., 1988. Precipitating ion and electron detectors (SSJ/4) for the block 5D/flight 8 DMSP satellite, vol. 88. Air Force Geophysics Laboratory, United States Air Force.
- Selliez-Vandernotte, L., 2018. Optimisation de l’analyse de la matière organique à l’aide d’un nouveau spectromètre de masse basé sur le CosmOrbitrap dans un contexte de future mission spatiale. Ph.D. thesis, Université d’Orléans.
- Sexton, E. S., K. Nykyri, and X. Ma, 2019. Kp forecasting with a recurrent neural network. *Journal of Space Weather and Space Climate*, **9**, A19. 10.1051/swsc/2019020.

- Shepherd, S. G., 2014. Altitude-adjusted corrected geomagnetic coordinates: Definition and functional approximations. *Journal of Geophysical Research: Space Physics*, **119**(9), 7501–7521. <https://doi.org/10.1002/2014JA020264>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2014JA020264>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2014JA020264>.
- Shibata, K., and T. Yokoyama, 2002. A Hertzsprung-Russell-like Diagram for Solar/Stellar Flares and Corona: Emission Measure versus Temperature Diagram. *Astrophysical Journal*, **577**(1), 422–432. 10.1086/342141, [astro-ph/0206016](https://arxiv.org/abs/astro-ph/0206016).
- Shiokawa, K., W. Baumjohann, G. Haerendel, G. Paschmann, J. F. Fennell, et al., 1998. High-speed ion flow, substorm current wedge, and multiple Pi 2 pulsations. *Journal of Geophysical Research*, **103**(A3), 4491–4508. 10.1029/97JA01680.
- Shorten, C., and T. M. Khoshgoftaar, 2019. A survey on image data augmentation for deep learning. *Journal of big data*, **6**(1), 1–48.
- Shorten, C., T. M. Khoshgoftaar, and B. Furht, 2021. Text Data Augmentation for Deep Learning. *Journal of Big Data*, **8**(1), 101. 10.1186/s40537-021-00492-0, URL <https://doi.org/10.1186/s40537-021-00492-0>.
- Siciliano, F., G. Consolini, R. Tozzi, M. Gentili, F. Giannattasio, and P. De Michelis, 2021. Forecasting SYM-H Index: A Comparison Between Long Short-Term Memory and Convolutional Neural Networks. *Space Weather*, **19**(2), e2020SW002589. E2020SW002589 10.1029/2020SW002589, <https://doi.org/10.1029/2020SW002589>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2020SW002589>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020SW002589>.
- Sigernes, F., M. Dyrland, P. Brekke, S. Chernouss, D. A. Lorentzen, K. Oksavik, and C. S. Deehr, 2011. Two methods to forecast auroral displays. *J. Space Weather Space Clim.*, **1**(1). 10.1051/swsc/2011003, URL <https://doi.org/10.1051/swsc/2011003>.
- Sinaga, K. P., and M.-S. Yang, 2020. Unsupervised K-Means Clustering Algorithm. *IEEE Access*, **8**, 80,716–80,727. 10.1109/ACCESS.2020.2988796.
- Smith, A. W., C. Forsyth, I. J. Rae, T. M. Garton, T. Bloch, C. M. Jackman, and M. Bakrania, 2021. Forecasting the Probability of Large Rates of Change of the Geomagnetic Field in the UK: Timescales, Horizons, and Thresholds. *Space Weather*, **19**(9), e2021SW002788. E2021SW002788 2021SW002788, <https://doi.org/10.1029/2021SW002788>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2021SW002788>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021SW002788>.
- Smith, L. N., and N. Topin, 2019. Super-convergence: very fast training of neural networks using large learning rates. In T. Pham, ed., *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, vol. 11006, 1100612. International Society for Optics and Photonics, SPIE. 10.1117/12.2520589, URL <https://doi.org/10.1117/12.2520589>.
- Solanki, S. K., 2003. Sunspots: an overview. *The Astronomy and Astrophysics Review*, **11**, 153–286.
- Souza, V. M., C. Medeiros, D. Koga, L. R. Alves, L. E. Vieira, A. D. Lago, L. A. Da Silva, P. R. Jauer, and D. N. Baker, 2018. Chapter 13 - Classification of Magnetospheric Particle Distributions Via Neural Networks. In E. Camporeale, S. Wing, and J. R. Johnson, eds., *Machine Learning Techniques for Space Weather*, 329–353. Elsevier. ISBN 978-0-12-811788-0. [Doi.org/10.1016/B978-0-12-811788-0.00013-5](https://doi.org/10.1016/B978-0-12-811788-0.00013-5), URL <https://www.sciencedirect.com/science/article/pii/B978012811788000135>.

- Space Weather Prediction Center, S., 2023. F10.7 cm Radio Emissions. [Online; last accessed June 2023], URL <https://www.swpc.noaa.gov/phenomena/f107-cm-radio-emissions>.
- Spelman, V. S., and R. Porkodi, 2018. A Review on Handling Imbalanced Data. In 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), 1–11. 10.1109/ICCTCT.2018.8551020.
- Starkov, G., 1994. Mathematical model of the auroral boundaries. *Geomagnetism and Aeronomy*, **34**, 331–336.
- Stone, E. C., A. M. Frandsen, R. A. Mewaldt, E. R. Christian, D. Margolies, J. F. Ormes, and F. Snow, 1998. The Advanced Composition Explorer. *Space Science Reviews*, **86**(1), 1–22. 10.1023/A:1005082526237, URL <https://doi.org/10.1023/A:1005082526237>.
- Storz, M. F., B. R. Bowman, M. J. I. Branson, S. J. Casali, and W. K. Tobiska, 2005. High accuracy satellite drag model (HASDM). *Advances in Space Research*, **36**(12), 2497–2505. *Space Weather*, <https://doi.org/10.1016/j.asr.2004.02.020>, URL <https://www.sciencedirect.com/science/article/pii/S0273117705002048>.
- Stumpo, M., S. Benella, M. Laurenza, T. Alberti, G. Consolini, and M. F. Marcucci, 2021. Open Issues in Statistical Forecasting of Solar Proton Events: A Machine Learning Perspective. *Space Weather*, **19**(10), e2021SW002794. E2021SW002794 2021SW002794, <https://doi.org/10.1029/2021SW002794>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2021SW002794>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021SW002794>.
- Suess, S. T., 1979. Models of Coronal Hole Flows (Article published in the special issues: Proceedings of the Symposium on Solar Terrestrial Physics held in Innsbruck, May- June 1978. (pp. 137-538)). *Space Science Reviews*, **23**(2), 159–200. 10.1007/BF00173809.
- Sundararajan, M., A. Taly, and Q. Yan, 2017. Axiomatic attribution for deep networks. In International conference on machine learning, 3319–3328. PMLR.
- Sweet, P. A., 1958. The neutral point theory of solar flares. *Symposium - International Astronomical Union*, **6**, 123–134. 10.1017/S0074180900237704.
- Tang, R., Y. Tao, J. Li, Z. Chen, X. Deng, and H. Li, 2022. The Short-Time Prediction of the Energetic Electron Flux in the Planetary Radiation Belt Based on Stacking Ensemble-Learning Algorithm. *Space Weather*, **20**(2), e2021SW002969. E2021SW002969 2021SW002969, <https://doi.org/10.1029/2021SW002969>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2021SW002969>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021SW002969>.
- Tapley, B. D., S. Bettadpur, M. Watkins, and C. Reigber, 2004. The gravity recovery and climate experiment: Mission overview and early results. *Geophysical Research Letters*, **31**(9). <https://doi.org/10.1029/2004GL019920>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2004GL019920>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2004GL019920>.
- Teichmann, S., V. Heidrich-Meisner, L. Berger, and R. F. Wimmer-Schweingruber, 2023. Influence of solar wind parameters on unsupervised solar wind classification with k-means. 10.22541/essoar.168202115.54861110/v1, URL <https://doi.org/10.22541/2Fessoar.168202115.54861110%2Fv1>.
- Tenenbaum, J. B., V. d. Silva, and J. C. Langford, 2000. A global geometric framework for nonlinear dimensionality reduction. *science*, **290**(5500), 2319–2323.

- Thayer, J. P., X. Liu, J. Lei, M. Pilinski, and A. G. Burns, 2012. The impact of helium on thermosphere mass density response to geomagnetic activity during the recent solar minimum. *Journal of Geophysical Research: Space Physics*, **117**(A7). <https://doi.org/10.1029/2012JA017832>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2012JA017832>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2012JA017832>.
- Tieleman, T., and G. Hinton, 2012. Rmsprop: Divide the gradient by a running average of its recent magnitude. coursera: Neural networks for machine learning. *COURSERA Neural Networks Mach. Learn*, **17**.
- Toma, G., and C. Arge, 2005. Multi-wavelength Observations of Coronal Holes. In *Large-scale Structures and their Role in Solar Activity*, vol. 346, 251.
- Tsoi, A. C., 1997. Recurrent neural network architectures: an overview. *International School on Neural Networks, Initiated by IIASS and EMFCSC*, 1–26.
- Vallado, D., P. Crawford, R. Hujsak, and T. Kelso. Revisiting Spacetrack Report #3, 2006. 10.2514/6.2006-6753, <https://arc.aiaa.org/doi/pdf/10.2514/6.2006-6753>, URL <https://arc.aiaa.org/doi/abs/10.2514/6.2006-6753>.
- van den Oord, A., S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, 2016. WaveNet: A Generative Model for Raw Audio. *arXiv e-prints*, arXiv:1609.03499. 10.48550/arXiv.1609.03499, 1609.03499.
- van Dyk, D. A., and X.-L. Meng, 2001. The Art of Data Augmentation. *Journal of Computational and Graphical Statistics*, **10**(1), 1–50. 10.1198/10618600152418584, <https://doi.org/10.1198/10618600152418584>, URL <https://doi.org/10.1198/10618600152418584>.
- Vapnik, V., 1999. The nature of statistical learning theory. Springer science & business media.
- Vernile, A., 2018. The rise of private actors in the Space Sector. Springer.
- Vincent, P., H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, and L. Bottou, 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, **11**(12).
- Vokhmyanin, M. V., N. A. Stepanov, and V. A. Sergeev, 2019. On the Evaluation of Data Quality in the OMNI Interplanetary Magnetic Field Database. *Space Weather*, **17**(3), 476–486. <https://doi.org/10.1029/2018SW002113>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2018SW002113>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018SW002113>.
- Vorobjev, V., O. Yagodkina, and Y. Katkalov, 2013. Auroral Precipitation Model and its applications to ionospheric and magnetospheric studies. *Journal of Atmospheric and Solar-Terrestrial Physics*, **102**, 157–171. <https://doi.org/10.1016/j.jastp.2013.05.007>, URL <https://www.sciencedirect.com/science/article/pii/S1364682613001533>.
- Wallmark, J. T., and S. M. Marcus, 1962. Minimum Size and Maximum Packing Density of Nonredundant Semiconductor Devices. *Proceedings of the IRE*, **50**(3), 286–298. 10.1109/JR-PROC.1962.288321.
- Wang, Y. M., 2009. Coronal Holes and Open Magnetic Flux. *Space Science Reviews*, **144**(1-4), 383–399. 10.1007/s11214-008-9434-0.
- Wang, Y. M., and J. Sheeley, N. R., 1992. On Potential Field Models of the Solar Corona. *Astrophysical Journal*, **392**, 310. 10.1086/171430.

- Wanliss, J. A., and K. M. Showalter, 2006. High-resolution global storm index: Dst versus SYM-H. *Journal of Geophysical Research: Space Physics*, **111**(A2). <https://doi.org/10.1029/2005JA011034>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2005JA011034>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2005JA011034>.
- Wilson, A. C., R. Roelofs, M. Stern, N. Srebro, and B. Recht, 2017. The Marginal Value of Adaptive Gradient Methods in Machine Learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds., *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/81b3833e2504647f9d794f7d7b9bf341-Paper.pdf.
- Wu, J., D. J. Knudsen, D. M. Gillies, E. F. Donovan, and J. K. Burchill, 2017. Swarm Observation of Field-Aligned Currents Associated With Multiple Auroral Arc Systems. *Journal of Geophysical Research: Space Physics*, **122**(10), 10,145–10,156. <https://doi.org/10.1002/2017JA024439>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2017JA024439>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2017JA024439>.
- Yamazaki, Y., J. Matzka, C. Stolle, G. Kervalishvili, J. Rauberg, O. Bronkalla, A. Morschhauser, S. Bruinsma, Y. Y. Shprits, and D. R. Jackson, 2022. Geomagnetic Activity Index Hpo. *Geophysical Research Letters*, **49**(10), e2022GL098860. E2022GL098860 2022GL098860, <https://doi.org/10.1029/2022GL098860>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2022GL098860>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2022GL098860>.
- Yang, L., and Z. Huiyan, 1996. Shape preserving piecewise cubic interpolation. *Applied Mathematics*, **11**, 419–424.
- Ye, J., J. Zhao, K. Ye, and C. Xu, 2022. How to Build a Graph-Based Deep Learning Architecture in Traffic Domain: A Survey. *IEEE Transactions on Intelligent Transportation Systems*, **23**(5), 3904–3924. 10.1109/TITS.2020.3043250.
- Yeager, D. M., and L. A. Frank, 1976. Low-energy electron intensities at large distances over the Earth’s polar cap. *Journal of Geophysical Research (1896-1977)*, **81**(22), 3966–3976. <https://doi.org/10.1029/JA081i022p03966>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/JA081i022p03966>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/JA081i022p03966>.
- Yeakel, K. L., D. Turner, and I. Cohen, 2022. Automated detection and unsupervised classification of electron enhancement events in Earth’s magnetotail.
- Zeiler, M. D., and R. Fergus, 2013. Visualizing and Understanding Convolutional Networks. *arXiv e-prints*, arXiv:1311.2901. 10.48550/arXiv.1311.2901, 1311.2901.
- Zewdie, G. K., C. Valladares, M. B. Cohen, D. J. Lary, D. Ramani, and G. M. Tsidu, 2021. Data-Driven Forecasting of Low-Latitude Ionospheric Total Electron Content Using the Random Forest and LSTM Machine Learning Methods. *Space Weather*, **19**(6), e2020SW002,639. E2020SW002639 2020SW002639, 10.1029/2020SW002639, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2020SW002639>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020SW002639>.
- Zhang, C., S. Bengio, M. Hardt, B. Recht, and O. Vinyals, 2016. Understanding deep learning requires rethinking generalization. *arXiv e-prints*, arXiv:1611.03530. 10.48550/arXiv.1611.03530, 1611.03530.

- Zhang, G., Y. Liu, and X. Jin, 2020. A survey of autoencoder-based recommender systems. *Frontiers of Computer Science*, **14**(2), 430–450. 10.1007/s11704-018-8052-6, URL <https://doi.org/10.1007/s11704-018-8052-6>.
- Zhang, Y., and L. J. Paxton, 2008. An empirical Kp-dependent global auroral model based on TIMED/GUVI FUV data. *Journal of Atmospheric and Solar-Terrestrial Physics*, **70**(8), 1231–1242. URL <https://www.sciencedirect.com/science/article/pii/S1364682608000758>.
- Zhao, M.-X., J.-S. Wang, and X.-W. Zhao, 2022. A New Index to Describe the Response of Geomagnetic Disturbance to the Energy Injection from the Solar Wind. *Universe*, **8**(10). 10.3390/universe8100506, URL <https://www.mdpi.com/2218-1997/8/10/506>.
- Zheng, Y., N. Y. Ganushkina, P. Jiggins, I. Jun, M. Meier, et al., 2019. Space Radiation and Plasma Effects on Satellites and Aviation: Quantities and Metrics for Tracking Performance of Space Weather Environment Models. *Space Weather*, **17**(10), 1384–1403. <https://doi.org/10.1029/2018SW002042>, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2018SW002042>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018SW002042>.
- Ziegler, J., and R. M. Mcgranaghan, 2021. Harnessing expressive capacity of Machine Learning modeling to represent complex coupling of Earth’s auroral space weather regimes. In 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA), 1189–1196. 10.1109/ICMLA52953.2021.00193.
- Zirker, J. B., 1977. Coronal holes and high-speed wind streams. *Reviews of Geophysics*, **15**(3), 257–269.
- Zmuda, A. J., J. H. Martin, and F. T. Heuring, 1966. Transverse magnetic disturbances at 1100 kilometers in the auroral region. *Journal of Geophysical Research (1896-1977)*, **71**(21), 5033–5045. <https://doi.org/10.1029/JZ071i021p05033>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/JZ071i021p05033>.
- Zoph, B., and Q. V. Le, 2016. Neural Architecture Search with Reinforcement Learning. *arXiv e-prints*, arXiv:1611.01578. 10.48550/arXiv.1611.01578, [1611.01578](https://arxiv.org/abs/1611.01578).

List of Figures

1.1	Gyration motion of a particle along a magnetic field \mathbf{B}	11
1.2	Motion of a positive ion and an electron in a field (\mathbf{E}, \mathbf{B}) from Kivelson and Russell (1995). Accelerated by the \mathbf{E} field, the gyroradius is larger in the direction of \mathbf{E} , causing a drift \mathbf{u} perpendicular to both \mathbf{B} and \mathbf{E}	12
1.3	Movement of a positive ion and an electron in a non-uniform magnetic field \mathbf{B} in the plane, with a variation of the gyration radius resulting in a drift \mathbf{u} for positively charged particles. Inspired by Kivelson and Russell (1995).	13
1.4	Mirror Effect and Loss Cone Near Earth.	15
1.5	Summary of basic motions of trapped particles in the Earth's magnetic field.	17
1.6	Magnetic flux conservation in a field line from Lilensten and Bornarel (2001).	20
1.7	Plasma and frozen-in condition. In (a) and (b), we can see the deformation of field lines caused by the plasma motion. In (c) and (d), we can see the plasma pushing against a flux tube that it cannot cross (Brekke, 1997).	20
1.8	Sun parts from Steele Hill / NASA. Courtesy of SOHO consortium. SOHO is a project of international cooperation between ESA and NASA.	22
1.9	(a) International Sunspot Number (ISSN); (b) butterfly diagram : latitudes of sunspots as a function of time; (Russell et al., 2016).	24
1.10	Standard flare model from Russell et al. (2016).	25
1.11	Solar flare and CME from October 2003, during what is known as Halloween solar storms. (a) Michelson Doppler Imager (MDI) image of the Sun's disk, showing the large group of sunspots of active region 10486; (b) EUV Imaging Telescope (EIT) image of solar flare; (c) C2 camera of the Large Angle and Spectrometric Coronagraph (LASCO) showing the CME cloud; (d) LASCO C3 camera showing the CME cloud. Energetic particles show up in this image as bright points and streaks, when hitting the instrument's detectors. (Copyright: SOHO/MDI, SOHO/EIT, SOHO/LASCO (ESA & NASA)).	26
1.12	Overall solar emission spectrum from Russell et al. (2016) adapted from Golub and Pasachoff (1997).	27
1.13	Parker's model of Interplanetary Magnetic Field from Russell et al. (2016). (a) Motion of a single parcel of solar-wind fluid carrying the Sun's open field line (Russell et al., 2016). (b) Parker spiral field in the equatorial plane in case of a constant solar-wind speed of $400\text{km}\cdot\text{s}^{-1}$ from Kivelson and Russell (1995).	31
1.14	Isothermal coronal-expansion model from Pneuman and Kopp (1971) for a dipole magnetic field considered at the base of the corona. Dashed lines are field lines for classical dipole field. Yellow zone corresponds to a thin layer of high current density. Adapted from Kivelson and Russell (1995).	32
1.15	3D rendering of the interplanetary current sheet in a Parker solar wind model for a tilted dipole, from (Orcinha et al., 2019).	32
1.16	Complete solar-wind extension into interplanetary medium. The + (respect. -) symbol indicates positive outward (respect. negative inward) magnetic field. Region of slow solar wind are shaded (Russell et al., 2016).	33

1.17	Visual representation of a corotating interaction region (CIR) forming compression and rarefaction zones. From Russell et al. (2016).	34
1.18	3D representation of the magnetosphere showing the major regions and currents, from Russell et al. (2016)	38
1.19	Cross section of the magnetosphere showing the major regions, including the outer magnetosphere and the magnetotail. From Russell et al. (2016).	38
1.20	(a) 3-D view from above of the magnetopause and the currents j_{tail} and j_{mp} from Dr. A. Marchaudon's class, found online https://www.slideserve.com/ion/le-champ-magn-tique-d-origine-externe-aur-lie-marchaudon-lpc2e and (b) identical perspective view of the northern portion of the magnetopause current, as seen from above the ecliptic plane. Charged particles in the solar wind are deflected in opposite directions by Earth's main field, creating a boundary current (from Encyclopædia Britannica, https://www.britannica.com/science/geomagnetic-field/The-ionospheric-dynamo , last accessed September 2023).	39
1.21	Schematic view of the reconnection geometry at the magnetopause. NIF = normal incidence frame. Adapted from Russell et al. (2016).	40
1.22	Dungey cycle from Kivelson and Russell (1995). Magnetospheric convection due to the magnetic reconnections. A field line labeled 1 connects with a field line labeled 1' causing a change of topology of the field line. Numbers show the successive configurations of the field line, in the antisunward direction, in the magnetosphere, polar cap and auroral ionosphere.	42
1.23	Cutaway model of the radiation belts with the 2 Van Allen Probes satellites flying through them. Credit: NASA (NASA, 2013).	45
1.24	Radiation belts (red) and plasmasphere (blue) under low (a) moderate (b) and high (c) geomagnetic activity. From ESA website - (Carreau, 2013).	46
1.25	Characteristic signature of a geomagnetic storm in the Dst index [nT] (Amory-Mazaudier et al., 2017).	47
1.26	Plasma waves and their region in the inner magnetosphere. Adapted from both NASA's Goddard Space Flight Center/Mary Pat Hrybyk-Keith image and Kletzing et al. (2013).	50
1.27	Layers of the ionosphere in our atmosphere by the Encyclopaedia Britannica, Inc (Britannica, 2012).	53
1.28	(a) Field-aligned currents system from Britannica (1994), (b) Currents towards and away from the ionosphere, from Russell et al. (2016).	54
1.29	Overview of the field-aligned current systems, showing Hall and Pedersen currents. Figure from Palmroth et al. (2021).	55
1.30	First observation of the cycle of auroral substorm in the pole from Akasofu (1968).	59
1.31	"Animated model aurora ovals as a function of Kp index for 24th December 2009 at 08:50 UT. The transparent polygons represent Feldstein-Starkov ovals (A). The faint white outer ring is the equatorward boundary of the diffuse aurora (C). The Zhang-Paxton ovals are displayed on top with green intensity values scaled according to the electron energy flux (B). The yellow scaled intensity areas are the intersection ($A \cap B$) between the two models." from Sigernes et al. (2011).	60
1.32	Short-term 30 to 90 minute forecast of the location and intensity of the aurora based on the OVATION model. Credits: NOAA Space Weather Prediction Center; https://www.swpc.noaa.gov/products/aurora-30-minute-forecast .	61

1.33	Most stations and observatories used for geomagnetic indices. ABK = Abisko, Sweden; AIA = Argentine Islands, Antarctica (Ukraine); AMS = Martin De Vivies-Amsterdam Island, French Southern and Antarctic Lands (France); ARS = Arti, Russia; BFE = Brorfelde, Denmark; BRW = Barrow, United States of America; CCS = Cape Chelyuskin, Russia; CLF = Chambon La Foret, France; CNB = Canberra, Australia; CMO = College, United States of America; CZT = Port Alfred, French Southern and Antarctic Lands (France); DIK = Dixon, Russia; ESK = Eskdalemuir, United Kingdom; EYR = Eyrewell, New Zealand; FCC = Fort Churchill, Canada; FRD = Fredericksburg, United States of America; GNG = Gingin, Australia; HAD = Hartland, United Kingdom; HER = Hermanus, South Africa; HON = Honolulu, United States of America; IRT = Irkutsk, Russia; KAK = Kakioka, Japan; KEP = King Edward Point, United Kingdom; LER = Lerwick, United Kingdom; LRV = Leirvogur, Iceland; MEA = Meanook, Canada; MGD = Magadan, Russia; MMB = Memambetsu, Japan; NAQ = Narsarsuaq, Greenland (Denmark); NEW = Newport, United States of America; NGK = Niemegek, Germany; NVS = Novosibirsk, Russia; OTT = Ottawa, Canada; PAF = Port-Aux-Francais, French Southern and Antarctic Lands (France); PBK = Pebek, Russia; PET = Paratunka/Petropavlovsk, Russia; PST = Port Stanley, Falkland Islands (United Kingdom); SIT = Sitka, United States of America; SJG = San Juan, United States of America; SNK = Sanikiluaq, Canada; THL = Qaanaaq (Thule), Greenland (Denmark); TIK = Tixie Bay, Russia; TUC = Tucson, United States of America; UPS = Uppsala, Sweden; VIC = Victoria, Canada; VOS = Vostok, Antarctica (Russia); WNG = Wingst, Germany; YKC = Yellowknife, Canada. From ISGI <code>isgi.unistra.fr/</code>	66
1.34	Table from the presentation of Dr. Linda Neergaard Parker (Parker, 2017) representing surface and internal charging based on particle energy.	69
1.35	Overview of the different processes during spacecraft charging. Figure from the presentation of Dr. Linda Neergaard Parker (Parker, 2017).	69
1.36	Formation of the induced electric field and the GICs as a consequence. Figure from McKay (2004).	74
1.37	Classifications of SEEs, from Gomez Toro et al. (2014).	76
1.38	All three categories of radiation effects, from Poivey (2019).	77
1.39	Space Weather Effects Overview. Credits: ESA.	80
2.1	(a) AlphaGo's victory over Go champion Lee Sedol in the Google DeepMind Challenge Match, featured in Nature (January 28th, 2016). (b) DALL-E's impressive image extension capability demonstrated on Johannes Vermeer's "Meisje met de parel." (c) DeepDream output example.	85
2.2	Different branches of AI. Note that "AI Ethics and Responsible AI" could be represented differently from other branches, as it is not exactly a branch in itself but rather a cross-cutting consideration across all branches.	87
2.3	Categories of Machine Learning algorithms.	87
2.4	Search results highlighting the occurrence of the words 'machine learning' in the AGU Space Weather Journal website's search bar https://agupubs.onlinelibrary.wiley.com/journal/15427390 . 183 results that can be research articles (158), technical articles (7), editorial (5), issue information (4), commentary (3), commissioned manuscript (1), feature (2), meeting report (1), news article (1). Words can appear in either the title, keywords, abstract, author affiliation or funding agency of the results.	88

2.5	Decision tree with artificial data. Instances with a value greater than 3 for feature x_1 end up in node 5. All other instances are assigned to node 3 or node 4, depending on whether values of feature x_2 exceed 1. Credits: Christophe Molnar, <i>Interpretable Machine Learning</i> , https://christophm.github.io/interpretable-ml-book/	92
2.6	Example of the U-k-means clustering ending in a 6-cluster dataset, from Sinaga and Yang (2020). Unlike the classical k-means algorithm, the amount of clusters is here an output.	93
2.7	Simplified structure of reinforcement learning algorithms, adapted from IBM's resources, https://developer.ibm.com/articles/cc-models-machine-learning/ (Jones, 2017).	95
2.8	Simplified and summarized structures for all main three types of machine learning algorithms: supervised, unsupervised and reinforcement learning, adapted from IBM's resources, https://developer.ibm.com/articles/cc-models-machine-learning/ , last accessed June 22, 2023 (Jones, 2017).	96
2.9	Sigmoid function.	100
2.10	Schematic diagram of the logistic regression classification, adapted from Raschka and Mirjalili (2017).	100
2.11	Schematic diagram of a 2-layer neural network.	101
2.12	Lineplots of learning rate evolution as a function of the epoch, from Monigatti (2022). Of course, the shape of the curve depends on the parameters specified by the user.	111
2.13	Non-exhaustive list of activation functions for artificial neural networks. Credits: Sebastian Raschka, 2016 https://sebastianraschka.com	113
2.14	Example of confusion matrix for 1000 training samples in a binary classification. Note: TP = True Positives; TN = True Negatives; FN = False Negatives; FP = False Positives.	115
2.15	Example of overfitting and underfitting fits on simple cases. Credits: MathWorks courses https://www.mathworks.com/discovery/overfitting.html , last accessed in June 2023.	117
2.16	Example of validation and training loss curves from Brownlee (2019).	118
2.17	Example of a convolutional layer applied to a 6x6 input with 3 channels, using 4x4 filters (3 channels). No padding and a stride of one are applied. Adapted from https://indoml.com	128
2.18	Example of the filter sliding over the three channels. Padding has been applied, and the stride is set to one.	128
2.19	Example of the a dilation factor of 2 where the filter skips one pixel in between each value, creating a sparse sampling pattern. Here, the filter is patched on the input (in blue) to create the first value of the output matrix (in cyan).	129
2.20	Example of a pooling layer using the max pooling method. Source: Computer Science Wiki https://computersciencewiki.org/index.php/File:MaxpoolSample2.png , last accessed in July 2023.	129
2.21	Example of a convolutional neural network with 5 convolutional layers and 5 pooling layers, culminating in three fully-connected layers. These final layers can be employed because the pixel values have been transformed into vectors through a "flatten" operation, allowing us to return to the familiar concept of classic neural networks. Image from https://learnopencv.com/understanding-convolutional-neural-networks-cnn/##Convolutional-Blocks-and-Pooling-Layers , last accessed in July 2023.	130

2.22	Architecture of a basic autoencoder with W representing the weights of the encoder part and W' representing the weights of the decoder part. Figure adapted from Zhang et al. (2020).	131
2.23	Example of a denoising autoencoder where the corrupted input image is encoded to a representation and then decoded. Figure adapted from Bank et al. (2020).	132
2.24	Schematic view of the functioning of a Recurrent Neural Network. From Ye et al. (2022).	132
2.25	Schematic view of a LSTM cell. Several cells are then put together to form the LSTM architecture. Figure adapted from Hairy (2021).	133
2.26	This figure illustrates 1D convolution for <code>nr_input_channels</code> , the number of input features, on a sequence of length <code>input_length</code> (for example, 30 minutes of data for two temporal variables).	134
2.27	Explanation of zero padding in the case of no dilation.	134
2.28	Schematic view of the receptive field. A full history coverage would mean no blue squared in the bottom row.	135
2.29	Example of dilation 2^i over several layers.	135
2.30	Architecture of a TCN.	136
2.31	Example of a very simple network with PyTorch, demonstrating the 7 essential steps to create a functioning machine learning algorithm. All 7 steps are explained in Section 2.4.2.1.	140
2.32	Comparison between PyTorch and PyTorch Lightning, highlighting key differences. Image sourced from the PyTorch-Lightning documentation https://pytorch-lightning.readthedocs.io/en/0.7.1/introduction_guide.html , last accessed in July 2023.	141
3.1	Details of the configuration of the aperture and curved plates of an SSJ/4 Electrostatic Analyzer (ESA), from Schumaker (1988).	148
3.2	Adjusted channel response characteristics of the SSJ/4 F8 electron detector, taken as an example to understand energy channels in the SSJ/4 instrument, from Schumaker (1988).	149
3.3	Details of the configuration of the SSJ/5 Electrostatic Analyzer (ESA) field of view. Taken from Boston College DMSP website, https://dmsp.bc.edu/html2/ssj5_inst.html , last accessed September 2023.	150
3.4	Typical color spectrogram of the SSJ4 data obtained in one polar pass of a DMSP F13 spacecraft during the crossing of the northern polar region on 97 May 15 during a geomagnetic storm. Credits: Boston College website, https://dmsp.bc.edu/html2/dmsspssj4_data.html , last accessed in July 2023.	151
3.5	Schematic details of the ACE satellite and its instruments (image credit: NASA, accessed through the ESA's eoportal, https://www.eoportal.org/satellite-missions/ace##rf-communications , last accessed September 2023).	153
3.6	Histogram of the DMSP SSJ Electron total energy flux, from F06 to F18, after applying a base-10 logarithm.	177
3.7	Solar cycle represented by $F_{10.7}$ measures in solar flux units. The green region represents the measures by F06 to F09 in 1987 and 1988. The red region represents the measures by F12 to F18.	178
3.8	2D 24x24 polar grid plotting DMSP observational density with (a) the combined north and south pole, (b) the south pole data only and (c) the north pole data.	178
3.9	Histogram of the Magnetic Local Time observations (in hours) for all satellites combined (black) and separately (lineplots). Differences arise from the slightly different orbits of the satellites, as well as their progressive shifts over time.	179
3.10	Histogram of the Magnetic Latitude observations (in degrees).	179

3.11	DMSP SSJ data points ordered per satellite. Data values in $\log_{10}(\text{eV}/\text{cm}^2/\text{sec}/\text{ster})$.	180
3.12	DMSP SSJ data points in $\log_{10}(\text{eV}/\text{cm}^2/\text{sec}/\text{ster})$ ordered per satellite, and in time as in Figure 3.11, with data points above the 99.995th quantile highlighted in red.	181
3.13	Histograms of SYM-H (nT), AL(nT) and AU (nT).	185
3.14	Diagram illustrating the preparation of the available datasets (inputs in blue and outputs in red). The referenced articles are King and Papitashvili (2006), Redmon et al. (2017), and the AI ready dataset (here https://zenodo.org/record/4281122) from McGranaghan (2019); McGranaghan et al. (2021).	210
3.15	Diagram illustrating the integration of inputs and outputs to produce the ultimate combined datasets employed in our study. Blue arrows delineate the transition of OMNI data, red arrows denote the total electron energy flux data, and black arrows represent the joint movement of both OMNI and DMSP data (often following synchronization of the two). The green arrow corresponds to the index list.	211
3.16	Operational schema of the TCN with presentation of selected inputs and outputs.	213
3.17	Schematic representation of the historical data retrieval process for each parameter. Adapted from McGranaghan et al. (2021).	213
3.18	Illustration of inputs and outputs within the context of a fully connected neural network using our dataset number 2 (we may selectively utilize data from the "INPUTS" category).	214
4.1	Organization of our code using the PyTorch Lightning framework. A solid arrow pointing from box X to box Y indicates that X instantiates Y. A dashed arrow pointing from box X to box Y indicates that X refers to Y (i.e., in X, we can perform the operation <i>self.Y</i>).	218
4.2	Model Evaluation Metrics for the machine Learning Model (PrecipNet) from McGranaghan et al. (2021).	220
4.3	Left: Histograms of precipitated electron flux values on the training set and the validation set where the validation set are data from the satellite F16 in the year 2010. Right: Histograms of data from all satellites separately, showing a clear difference between F16 and the others.	221
4.4	Training and Validation Loss from McGranaghan et al. (2021) on the left and from our reproduction on the right. The Loss function is a mean squared error (MSE).	222
4.5	Loss curves for the PrecipNet-R training, both for random sampling to produce the training (yellow curve) and validation sets (black curve), and the choice of the 2010 year of satellite F16 (blue and red curves for training and validation sets, respectively).	223
4.6	Training and Validation loss functions for PrecipNet-R during training over 1000 epochs without the early-stopping criterion.	223
4.7	PrecipNet-R training for various architectures without Dropout. Numbers found in the description of the architecture correspond to the amount of neurons in each layer (e.g., V0 is made of to 2 hidden layers containing respectively 7000 and 4 neurons).	224
4.8	Original PrecipNet-R (V1) vs. a simpler architecture (V0) loss curves.	224
4.9	Actual values compared with model predictions for the set-aside test set. The model is PrecipNet-R with position and date inputs only shown on the left and the original PrecipNet-R on the right.	225
4.10	A closer look at the curves observed in Figure 4.9.	225
4.11	On the left: training and validation loss functions for the 12:00 MLT cell (from 11h to 12h MLT) and 75° MLAT (magnetic latitude between 74° and 75°). On the right: a closer comparison between labels and predictions on the training set.	226

4.12	Top: Training and validation curves for models 1 and 2. Bottom: Training and validation curves for models 3 and 4.	230
4.13	Histogram comparison of the test dataset versus the predictions made by trained models 1 to 4.	231
4.14	Density displays of the test dataset data points versus predictions made by the three models. Density is shown in logarithmic scale for visualization clarity, with 1 added to the result to represent gaps in white. Thus, no point has a value between 0 and 1. The line $x = y$ is highlighted in black.	232
4.15	Histograms of OVATION, models 1 and 4 predictions over 66,882 samples taken randomly from the test set.	235
4.16	Density estimates between DMSP real measurements and OVATION, models 1 and 4 predictions over 66,882 samples taken randomly from the test set. The density is displayed in logarithmic scale for visualization purposes, and we added 1 to the results, making any gaps appear white. Thus, no point has a value between 0 and 1. The line $x = y$ appears in black.	236
4.17	Display of the results obtained by PrecipNet sourced from the article McGranaghan et al. (2021) compared with our two most relevant models: our model 1 for its simplicity and our model 4 for its results on extreme values. Note: some PrecipNet data that we do not have appear here. They may correspond to missing points obtained in PrecipNet by interpolation, which we therefore do not have. Some minor differences are observed between our datasets, the cause of which we have not identified.	237
4.18	Display of the histogram and density estimates obtained by forecast using model 1's architecture. The density is displayed in logarithmic form for visualization purposes, and we added 1 to the result, making any gaps appear white. Thus, no point has a value between 0 and 1. The line $x = y$ appears in black.	239
4.19	Training and validation loss curves over 50 epochs.	242
4.20	Evolution curves of the learning rate, following the exponential LR set at 0.9 (the learning rate is multiplied by 0.9 at each epoch).	243
4.21	Histograms of Model 1 and 2 results compared to the histogram of actual DMSP values on the test set.	243
4.22	Density of points when plotting predicted values from the TCN models against actual values on the test set. The density is displayed in logarithm for visualization purposes, and we added 1 to the result and made voids appear in white. Thus, no point has a value between 0 and 1. The line $x = y$ is shown in black.	244
4.23	Overview of the code used to display what is visible in Figures 4.24 and 4.25.	245
4.24	Output of the code plotting a certain number of points "amount_data_plot" (here 14) after a given date "start_datetime" (here March 18, 2010, at 02:01) as output from the trained TCN. Also displayed here are the true DMSP values as well as OVATION's forecasts. The color code of the points comes from Figure 4.25.	246
4.25	Same data as in Figure 4.24 displayed with a color code on a polar graph representing the poles. The passage points represent 15 values for the F16 satellite between 02:01 and 02:14 on March 18, 2010.	247
4.26	Display of the final product output, based on the TCN model, for two arbitrary dates (March 18, 2010, at 02:00 and November 22, 2003, at 03:21) showing a tumultuous period (left) and a calmer one (right). The precision in magnetic local time is 0.25h, and in magnetic latitude, it's 0.5°	248

4.27	Three displays showing the attribution values from the integrated gradients method for three distinct examples. The first example (top) corresponds to the sample from May 23, 2011, at 05:38. The second (middle) corresponds to the sample from September 20, 2004, at 07:08. The third (bottom) corresponds to March 18, 2010, at 01:59.	249
5.1	Idea of a stacking method to combine several AI algorithms to forecast DMSP values.	255

List of Tables

1.1	Fast and slow solar wind properties according to Russell et al. (2016)	29
1.2	Inner and outer belts characteristics	46
1.3	Summary of frequencies, regions, sources and main impacts of different wave modes in the inner magnetosphere. Table built using Koskinen and Kilpua (2022) and Kletzing et al. (2013)	51
1.4	Characteristics of the various ionospheric regions. From Pisacane (2008)	53
1.5	Summary of aurora types and their characteristics	58
1.6	Summary of main geomagnetic and solar indices	65
1.7	Summary of effects of GICs and overall disturbances in the geomagnetic field on power systems, from Beccutti (2013)	75
2.1	Examples of large datasets used in Space Weather, from Camporeale (2019). Note: ACE = Advanced Composition Explorer; DSCOVR = Deep Space Climate Observatory; SOHO = Solar and Heliospheric Observatory; STEREO = Solar Terrestrial Relations Observatory; SDO = Solar Dynamics Observatory; VAP = Van Allen Probes; GOES = Geostationary Operational Environmental Satellite system; POES = Polar Operational Environmental Satellites; DMSP = Defense Meteorological Satellite Program; GONG = Global Oscillation Network Group.	89
2.2	Examples of libraries (top) and AI-ready datasets (bottom) for applying machine learning in space weather.	90
2.3	Brief comparison between the gradient descent and normal equation methods	99
2.4	Example of loss functions for regression in the case of deterministic forecasts	103
2.5	All notation used to understand the training process of an algorithm. Some notations, such as the	105
2.6	Example of metrics	114
2.7	Non-exhaustive list of the main hyperparameters and factors which can be modified by the user.	122
4.1	Table presenting the hyperparameters of the four final models used. Note: FC = Fully-Connected. AE = Autoencoder. The hardware used was from a single MSI computer (GPU: nvidia Quadro RTX; CPU: Intel i7-10875H @3GHz).	229
4.2	Values of the selected metrics on the test dataset, which the algorithm has not previously encountered. "MSE Xth" refers to the MSE value for data outside the Xth percentile. "Time" indicates the training duration. It is for the hardware from a single MSI computer (GPU: nvidia Quadro RTX; CPU: Intel i7-10875H @3GHz).	233
4.3	MAE, RMSE and MSE for satellite F16 and year 2010 data	233
4.4	Comparison of metrics results between models 1, 4 and OVATION over 66,882 samples taken randomly from the test set.	235
4.5	Results when training the architecture of model 1 with only historical values of the inputs B_x , B_y , B_z in GSE coordinates, flow speed, proton density, and pressure (in logarithm) and AL, AU, and SYM-H. Time is for the hardware from a single MSI computer (GPU: nvidia Quadro RTX; CPU: Intel i7-10875H @3GHz).	238

4.6 Hyperparameters of the two versions of our Temporal Convolutional Network . . . 241
4.7 Results for our metrics on versions 1 and 2 of the TCN 242

