Identification by machine learning and analysis of electron diffusion regions at the Earth's magnetopause observed by MMS

Quentin Lenouvel

To cite this version:

Quentin Lenouvel. Identification by machine learning and analysis of electron diffusion regions at the Earth's magnetopause observed by MMS. Earth and Planetary Astrophysics [astro-ph.EP]. Université Paul Sabatier - Toulouse III, 2022. English. NNT : 2022TOU30241 . tel-04075287

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)*

Présentée et soutenue le *10/11/2022* **par :**

Quentin LENOUVEL

Identification par apprentissage machine et analyse de régions de diffusion électronique à la magnétopause terrestre observées par la mission MMS

JURY

Christian Jacquey	Astronome	Président du Jury
Karine Issautier	Directrice de Recherche	Examinatrice
Vincent Génot	Astronome	Directeur de thèse
Philippe Garnier	Maître de conférences	Co-directeur de thèse
Roch Smet	Professeur	Rapporteur
Matthieu Kretzschmar	Maître de conférences	Rapporteur

École doctorale et spécialité :
SDU2E : Astrophysique, Sciences de l'Espace, Planétologie
Unité de Recherche :
Institut de Recherche en Astrophysique et Planétologie (UMR 5277)
Directeur(s) de Thèse :
Vincent GENOT et *Philippe GARNIER*
Rapporteurs :
Roch SMET et *Matthieu KRETZSCHMAR*

Remerciements

Mes premiers remerciements vont a mes deux directeurs de thèse Vincent Génot et Philippe Garnier, pour leur gentillesse, leur bienveillance et leur disponibilité durant ma thèse, notamment durant les multiples phases de confinement dues à la fameuse pandémie de COVID. Ces trois années de thèse ont pu être difficiles par moments, et je les remercie de m'avoir toujours conseillé, soutenu et encouragé pour que ma thèse se passe dans les meilleures conditions possibles.

Je remercie Christian Jacquey, Roch Smet, Matthieu Kretzschmar et Karine Issautier d'avoir accepté de faire partie de mon jury de thèse. Je souhaite également remercier Sergio Toledo-Redondo et Benoît Lavraud qui ont été très impliqué dans mon travail depuis le début de ma thèse et dont l'expertise scientifique a été d'une grande aide pour mener à bien un certain nombre d'études présentes dans ce manuscrit. De même, je tiens à remercier Nicolas Aunai et Gautier Nguyen pour leurs conseils ainsi que les discussions que l'on a eu lors de mes deux séjours au LPP.

Mon temps au labo n'aurait pas été le même sans tous ces midis passés à prendre le café dans le bureau J014. J'en profite donc pour remercier tous les camarades du J014, et plus particulièrement Xan Astiasarain et Erwan Quintin pour m'avoir soutenu et encouragé tout au long de la thèse. Et finalement, je remercie mes co-bureaux Naïs Fargette et Victor Réville pour leur gentillesse et la bonne humeur qu'ils ont pu apporter dans le bureau que nous avons partagé pendant 3 ans. Si vous lisez ceci, j'espère que mes bruits de clavier et la musique un peu forte qui pouvait sortir de mon casque ne vous ont pas trop dérangés !

Résumé

La reconnexion magnétique est un processus physique qui intervient quand des lignes de champ magnétique opposées interagissent ensemble, résultant en un changement de la topologie magnétique et un transfert de l'énergie magnétique vers l'énergie des particules environnantes. Ce phénomène peut être retrouvé dans de nombreux contextes astrophysiques dont l'un est à la magnétopause des planètes où le vent solaire et la magnétosphère interagissent ensemble. De nombreuses études ont été réalisées à ce sujet depuis les années 80 en utilisant des simulations et des données d'observations de missions comme THEMIS ou CLUSTER. Cependant, le coeur du processus de reconnexion, la région de diffusion électronique (ou Electron Diffusion Region), restait largement non étudié à cause de de la difficulté de son observation jusqu'au lancement de la mission Magnetospheric Multiscale (MMS) en 2015. Depuis lors, de plus en plus d'études sur l'EDR ont vu le jour grâce à la résolution sans précédent des données produites par MMS mais manquent toujours d'un nombre significatif de cas afin de généraliser les résultats obtenus, le nombre d'EDR identifiées n'était en effet que d'environ 32 en 2018 et n'a pas beaucoup changé dans les années suivantes.

Dans cette thèse, je détaille le travail que j'ai mené dans le but d'identifier de nouvelles EDR dans les données de la mission MMS en utilisant des techniques d'apprentissage profond et pour étudier certains des aspects de l'EDR qui ne sont pas encore complètement compris par la communauté scientifique comme la structure de l'EDR ou encore la conversion d'énergie au sein de cette région.

J'ai tout d'abord développé un Perceptron Multicouche pour identifier automatiquement de nouveaux cas d'EDR en utilisant 32 cas précédemment rapportés dans la littérature comme base de données d'entraînement. Ce modèle utilise plusieurs paramètres physiques classiques comme le champ magnétique, le champ électrique ainsi qu'un scalaire dérivé que j'ai créé appelé le "MeanRL". Ce paramètre a été conçu pour caractériser l'une des signatures clés de l'EDR à savoir la présence de formes de croissants dans les fonctions de distribution des vitesses des électrons prises dans cette région. L'algorithme entraîné a été appliqué aux données de la phase 1a de MMS et a permis l'identification de 18 nouveaux candidats d'EDR. Quatre fois plus d'EDR présentant de la conversion d'énergie positive (inner EDR) que d'EDR présentant de la conversion d'énergie négative (outer EDR) ont été trouvés durant la phase 1a, et il a été trouvé que le terme inertiel de la loi d'Ohm généralisée est responsable de la décélération des électrons dans l'outer EDR, causant la conversion d'énergie négative observée dans cette région.

J'ai ensuite utilisé les anciens et nouveaux cas d'EDR rapportés afin d'effectuer plusieurs études portant sur : le lien entre les paramètres du vent solaire et la conversion d'énergie à l'EDR, de potentiels nouveaux paramètres de petite échelle pour identifier des EDR, la position de plusieurs points d'intérêts au sein de l'EDR ainsi que sur l'intensité et les composantes parallèles/perpendiculaires de la conversion d'énergie à l'EDR.

Pour finir, j'ai conçu un réseau de neurones convolutionel (CNN) comme moyen alternatif pour rechercher des EDR. Ce modèle analyse les fonctions de distribution des vitesses des électrons pour directement chercher des croissants électroniques et essaie d'être une amélioration du paramètre "MeanRL". Les cas montrant de tels croissants sont alors analysés "à la main" pour confirmer ou non l'identification d'une EDR. Ce CNN a été appliqué sur les données de la phase 1b de MMS qui n'avait pas été traitée par le premier algorithme et a permis la détection de 17 cas additionnels d'EDR, amenant le nombre total d'EDR identifiés à environ 70 lors de l'écriture de cette thèse.

Abstract

Magnetic reconnection is a physical process that can occur when oppositely directed magnetic field lines interact with each other and change of the magnetic field topology, transferring the magnetic energy to the energy of the surrounding particles. This phenomenon can be found in numerous astrophysical contexts, one of which is at the magnetopause of the planets where the solar wind and the magnetosphere interact with each other. Many studies have been made on this matter since the 1980's using simulations and observations from missions like THEMIS or CLUSTER. However, the heart of the reconnection process, the electron diffusion region (or EDR), remained largely unstudied due to the difficulty to observe it until the launch of the Magnetospheric Multiscale (MMS) mission in 2015. Since then, more and more studies on the EDR emerge thanks to the unprecedented resolution of the data provided by MMS but still lack significant numbers of EDR cases to generalize the results, the number of identified EDRs was only of about 32 in 2018 and had not changed much in the next years.

In this thesis, I detail the work I have done to identify new EDR events in the data of the MMS mission using deep learning techniques and to study some of the physical aspects of the EDR that are not yet fully understood by the scientific community such as the structure of the EDR or the energy conversion process at the heart of this region.

I first developed a Multilayer Perceptron to automatically identify new EDR events using 32 cases previously reported in the literature as a training database. This model uses several classical physical parameters such as the magnetic field, the electric field or the magnetic current as well as one derived scalar that I created called the "MeanRL". This parameter was designed to characterize one of the key signatures of EDRs which is the presence of a crescent shape in the electron velocity distribution functions taken in this region. The trained algorithm was applied to the data of phase 1a of MMS and allowed us to identify 18 new EDR candidates. Four times as many EDRs with positive energy conversion (inner EDRs) as EDRs with negative energy conversion (outer EDR) were found in phase 1a, and the inertial term of the generalized Ohm's law was found responsible to the deceleration of the electrons in the outer EDR, leading to the negative energy conversion observed there.

Then, I used the previously and newly reported EDRs to perform several studies about : the link between the solar wind parameters and the energy conversion at the EDR, potential new small scale parameters to identify EDRs, the position of several points of interest within

the EDR as well as the intensity and the parallel/perpendicular components of the energy conversion at the EDR.

Finally, I designed a Convolutional Neural Network (CNN) as an alternative way to look for EDRs. This model analyzes the electron velocity distribution functions to directly search for electron crescents and tries to be an improvement of the ”MeanRL” parameter. Events showing such crescents are then ”manually” investigated to confirm or not the EDR detection. This CNN was applied on the data of phase 1b of MMS that had not been treated by the first algorithm and allowed for the detection of 17 additional EDR candidates, bringing the total number of identified EDRs to about 70 as of the writing of this thesis.

List of abbreviations

ADP	Axial Double Probe
AI	Artificial Intelligence
AFG	Analog Fluxgate Magnetometer
CNN	Convolutional Neural Network
CVA	Constant Velocity Approach
DFG	Digital Fluxgate Magnetometer
EDI	Electron Drift Instrument
EDP	Electric Double Probe
EDR	Electron Diffusion Region
eVDF	Electron Velocity Distribution Function
FGM	Fluxgate Magnetometer
FPI	Fast Plasma Investigation
GSE	Geocentric Solar Ecliptic
GSM	Geocentric Solar Magnetospheric
IDR	Ion Diffusion Region
IMF	Interplanetary Magnetic Field
LSTM	Long Short-Term Memory
MHD	Magnetohydrodynamics
MLP	Multilayer Perceptron
MMS	Magnetospheric MultiScale mission
MSp	Magnetosphere
MSh	Magnetosheath
MVA	Minimum Variance Analysis
SITL	Scientists In The Loop
SDP	Spin-plane Double Probe

Contents

Chapter 1

Introduction abrégée en Français

1.1 L'âge de la météo de l'espace

Un intérêt croissant de la communauté scientifique pour l'interaction entre le Soleil et la Terre a été observé ces dernières années, un intérêt qui inclut la reconnexion magnétique depuis sa découverte dans les années 80. Cette interaction, même si elle se produit loin au-dessus de nos têtes, peut avoir un impact très réel sur notre vie sur Terre. L'exemple le plus célèbre est celui des lumières polaires (également appelées ”aurores boréales”) que l'on peut observer dans les régions de hautes latitudes. Bien qu'étant des événements magnifiques à observer, ces produits de l'interaction entre le Soleil et la Terre peuvent avoir des conséquences plus gênantes, comme le 2 septembre 1859, lorsqu'un orage géomagnétique intense a provoqué l'aurore la plus brillante de l'histoire, visible jusqu'à Boston. En Angleterre, cette aurore a provoqué l'interruption d'environ 200000 km de lignes télégraphiques pendant plusieurs heures. C'est également la première fois que le lien entre les aurores et l'électricité a pu être établi sans ambiguïté.

L'influence de l'activité solaire sur les infrastructures terrestres et spatiales est ensuite une question d'importance croissante. La nécessité d'arrêter les instruments des satellites en cas de fortes tempêtes géomagnétiques est un exemple de précaution à prendre contre ce que l'on appelle désormais la ”météo de l'espace”. Ce terme est apparu dans les années 80 pour désigner un domaine qui devient de plus en plus pertinent de nos jours grâce aux nombreux outils de surveillance dont nous disposons. Les percées scientifiques telles que l'invention des ordinateurs et plus récemment l'avènement de l'intelligence artificielle nous permettent de prédire à distance ces événements afin de protéger nos infrastructures.

Ces prédictions sont intrinsèquement liées à notre compréhension actuelle des mécanismes d'interaction entre le Soleil et la Terre. Plus précisément, la compréhension des mécanismes régissant les échanges d'énergie entre ces deux corps célestes est cruciale pour continuer à améliorer les outils et techniques de prédiction, la reconnexion magnétique étant l'un de ces mécanismes. C'est pourquoi l'étude globale de la reconnexion magnétique est un sujet d'intérêt pertinent pour la communauté scientifique. Ce phénomène étant un vaste sujet qui ne peut être traité dans sa globalité, et surtout pas au cours d'un travail de thèse de 3 ans, mes études se concentrent sur une petite région au cœur du processus de reconnexion magnétique appelée la région de diffusion électronique.

1.2 Introduction à la reconnexion magnétique

Cette section sera axée sur l'introduction des fondements théoriques sur lesquels repose le concept de région de diffusion électronique en physique des plasmas. Je présenterai d'abord l'environnement plasma de la Terre afin de comprendre le contexte dans lequel se déroule la reconnexion magnétique présentée dans cette thèse. Ensuite, j'expliquerai comment le concept de congélation, et plus particulièrement sa rupture, est à l'origine des premiers modèles de reconnexion magnétique. Enfin, je parlerai de l'image communément admise de la reconnexion et de la façon dont elle change en fonction de l'endroit où se produit la reconnexion.

1.2.1 Environnement plasma de la Terre

Avant de décrire la reconnexion magnétique, je vais présenter l'environnement plasma de la Terre et présenter les différentes régions de plasma qui interagissent entre elles. La figure 1.1 montre une vue de cette interaction avec les lignes de champ magnétique provenant du soleil en jaune et celles provenant de la Terre en vert et violet. Les lignes vertes sont des lignes de champ fermées et les lignes violettes sont des lignes de champ ouvertes, reliant directement la Terre et le vent solaire.

Dans cette image, le premier acteur majeur de l'environnement plasma terrestre est le Soleil. Il libère constamment des particules chargées de sa couronne dans l'espace interplanétaire, et ce flux de plasma est appelé "vent solaire". Sa vitesse est observée entre 300 et 800 km/s et près de l'orbite de la Terre, sa vitesse moyenne est de 400 km/s et sa densité moyenne est de 5 cm^{-3}. Les particules composant le vent solaire sont principalement des électrons, des protons (H^+), des particules alpha α (He^{2+}) peuvent également être trouvées mais leur densité est plus faible que celle des protons. Le vent solaire est responsable de la présence loin du soleil et dans le milieu interplanétaire d'un champ magnétique appelé "champ magnétique interplanétaire" (ou IMF).

Le deuxième acteur majeur de la physique des plasmas autour de la Terre est la magnétosphère terrestre. Il s'agit d'une cavité magnétique créée par le confinement des lignes de champ magnétique de la Terre par la pression du vent solaire, les densités de particules à l'intérieur sont donc très faibles car elle agit comme un bouclier contre le vent solaire. Ce bouclier n'est cependant pas parfait aux endroits appelés "cuspides polaires" où le champ magnétique dipolaire de la Terre se déploie du dipôle magnétique, ce sont les points d'entrée privilégiés des particules du vent solaire dans la magnétosphère car le champ magnétique y est plus faible. Cette magnétosphère agit comme un obstacle pour le vent

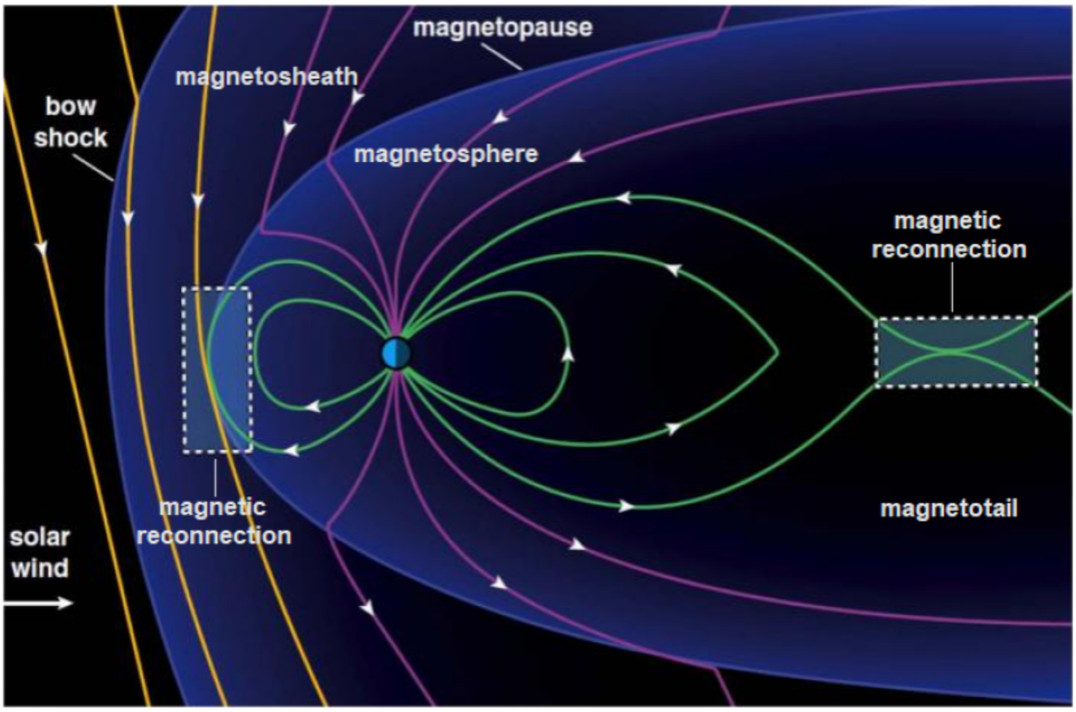


Figure 1.1: Figure montrant l'environment plasma de la Terre.

solaire supersonique qui est ralenti et dévié autour d'elle, créant une limite appelée "choc d'arc" où le plasma est ralenti et condensé, ce qui entraîne une augmentation de la densité, de la température et de l'intensité du champ magnétique.

La partie en aval du choc est appelée la magnétogaine, c'est une région dynamique de plasma turbulent qui sert d'état transitoire entre la région du vent solaire et la magnétosphère. Les valeurs typiques pour la densité de particules et l'intensité du champ magnétique dans cette région sont d'environ 20 cm^{-3} et 20 nT alors que les valeurs typiques pour la densité de particules et l'intensité du champ magnétique dans la magnétosphère sont d'environ 0,1 cm^{-3} et 50 nT. La magnétogaine est donc très facilement reconnaissable et se distingue de la magnétosphère dans les données.

La discontinuité qui sépare le champ géomagnétique du vent solaire choqué est une couche de courant appelée magnétopause. Il s'agit de la surface à laquelle la pression du vent solaire choqué et celle de la magnétosphère s'équilibrent. Sa position n'est donc pas constante et peut se déplacer vers l'intérieur ou l'extérieur en réponse à une augmentation ou une diminution de la pression du vent solaire. La magnétopause est la région autour de laquelle se concentre le travail présenté dans cette thèse, elle présente donc un intérêt particulier pour nous.

La dernière région de plasma d'intérêt qu'il nous reste à présenter est la "queue magnétique", qui est une extension de la magnétosphère du côté nuit, longue de plusieurs centaines de rayons terrestres. Cette extension bien au-delà de la Terre est la réponse du côté nuit de la magnétosphère à la pression du vent solaire du côté jour. Elle est composée de deux lobes presque vides de particules séparés par une nappe de plasma contenant la plupart des particules de la queue. Le cisaillement entre les lignes de champ ouvertes de direction opposée provenant des deux hémisphères crée une nappe de courant neutre située au milieu de la nappe de plasma.

C'est dans ce contexte que la reconnexion magnétique a lieu, la magnétopause côté jour et la magnétotail côté nuit étant les deux sites les plus proches de la Terre où la reconnexion magnétique peut avoir lieu en raison de la présence de lignes de champ magnétique de directions opposées. La reconnexion est toujours possible du côté jour car si le champ magnétique interplanétaire (FMI) transporté par le vent solaire est purement orienté vers le sud, la reconnexion peut avoir lieu près du ”point subsolaire”, qui est le point le plus externe de la magnétopause où le vent solaire rencontre d'abord la magnétosphère. Si l'IMF est purement orienté vers le nord, la reconnexion reste possible près des cuspides polaires, au-dessus des pôles, car le champ magnétique terrestre y est orienté dans la direction opposée à celle du champ du point subsolaire. Du côté nuit, la configuration des lignes de champ ouvertes fait que les lignes de champ provenant de l'hémisphère nord sont dans la direction opposée des lignes de champ provenant de l'hémisphère sud, ce qui permet également la reconnexion.

La reconnexion magnétique se retrouve également dans de nombreux autres contextes astrophysiques où de forts champs magnétiques sont impliqués et est à l'origine de certains des événements astrophysiques les plus puissants du système solaire tels que les éruptions solaires. La reconnexion magnétique est en effet un processus majeur de transfert d'énergie qui peut se produire lorsque des lignes de champ magnétique de direction opposée interagissent entre elles. Cette interaction se traduit par un changement de connectivité où la topologie des lignes de champ est réorganisée, et par une conversion de l'énergie magnétique en énergie cinétique, en énergie thermique et en accélération de particules.

1.2.2 Origine de la reconnexion magnétique et premiers modèles

Le terme ”reconnexion magnétique” a été utilisé pour la première fois par James Dungey en 1953 dans sa thèse de doctorat consacrée à l'étude de l'accélération des particules dans la magnétosphère terrestre (*Dungey* [1953]). La tâche qui lui a été confiée par son superviseur était d'expliquer ce qui se passe à un point neutre magnétique et quelles pourraient en être les implications pour les aurores. Dans le modèle qu'il a développé, la condition de gel de la MHD idéale se rompt pour permettre la dissipation de l'énergie et la présence d'un champ électrique. Le mouvement des particules crée une fine feuille de courant dans laquelle la diffusion du champ magnétique domine. Les lignes de champ traversant cette nappe de courant se brisent et se croisent, d'où les termes ”déconnexion” et ”reconnexion”. Enfin, le plasma reconnecté est éjecté perpendiculairement aux nouveaux flux de plasma entrants.

Trois ans après l'introduction du concept de reconnexion magnétique par Dungey en 1953, deux physiciens américains, Peter Sweet et Eugene Parker, ont développé le premier modèle théorique de la reconnexion. Leur modèle décrit la reconnexion bidimensionnelle en

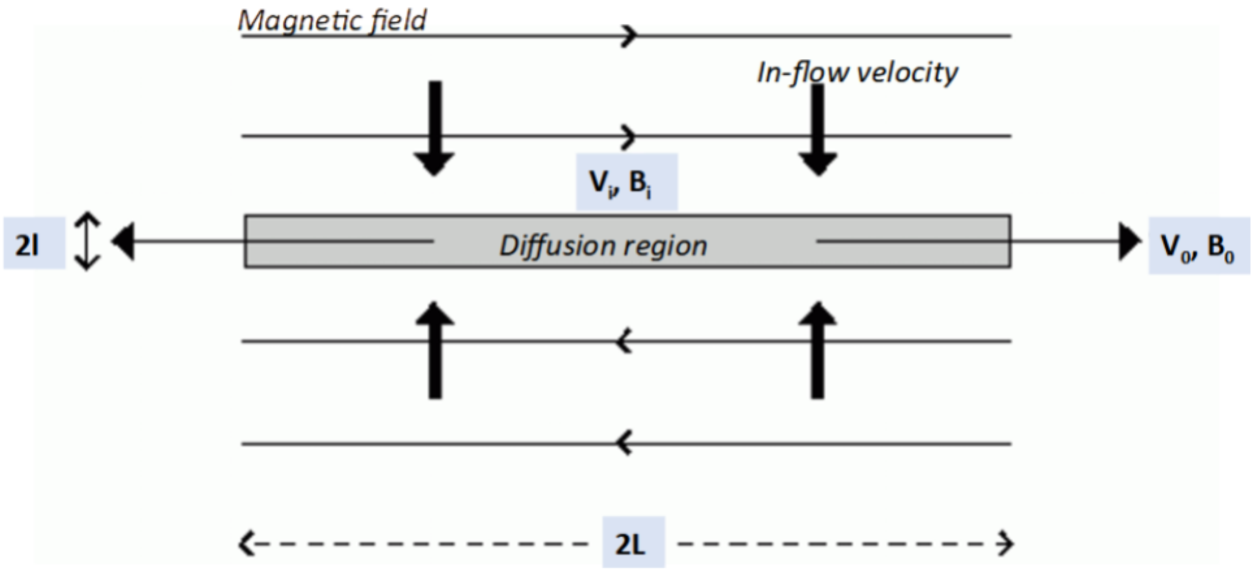


Figure 1.2: Figure montrant la géométrie du modèle de reconnexion de Sweet-Parker. X est selon la direction horizontale, Y selon la direction verticale et Z est perpendiculaire au plan XY dans lequel l'écoulement et le champ magnétique se trouvent. Image tirée de *Priest and Forbes* [2007].

régime permanent dans un plasma incompressible, sur la base du cadre de la MHD résistive. La figure 1.2 montre la géométrie de leur modèle, il considère une feuille de courant longue et mince de longueur $2L$ et de largeur $2l$. Pour comprendre pourquoi ce premier modèle était incomplet, il faut s'intéresser au taux de reconnexion. La démonstration suivante, qui suit celle réalisée dans *Priest and Forbes* [2007], montrera comment le trouver avec cette théorie ainsi que ce qu'il implique.

Nous commençons par le théorème de conservation de la masse qui nous dit que le taux auquel la masse entre dans la couche des deux côtés de la feuille de courant de la figure 1.2 est égal au taux auquel elle en sort, nous pouvons donc écrire :

$$2 \cdot \rho \cdot 2L \cdot v_{in} = 2 \cdot \rho \cdot 2l \cdot v_{out} \iff v_{in} L = v_{out} l \tag{1.1}$$

Nous pouvons obtenir une autre expression pour v_{in} en utilisant la loi d'Ohm pour un état stationnaire 2D qui s'écrit :

$$\mathbf{E} + \mathbf{v} \times \mathbf{B} = \mathbf{j}/\sigma \tag{1.2}$$

où $\mathbf{E} = E_z$ est un champ uniforme constant normal au plan XY. A l'entrée de la région d'afflux, le courant magnétique disparaît donc on a $||\mathbf{E}|| = v_{in} B_{in}$, et au centre de la région de diffusion, c'est le champ magnétique qui disparaît, ce qui donne $||\mathbf{E}|| = j/\sigma$. En utilisant la relation de Maxwell-Ampère ($\mathbf{j} = \nabla \times \mathbf{B}/\mu_0$) on obtient $||\mathbf{j}| = B_{in}/(\mu_0 l)$. On peut donc déduire de ces relations que $v_{in} = 1/(\mu_0 \sigma l) = \eta / l$ avec $\eta = 1/(\mu_0 \sigma)$.

En revenant à l'équation 1.1, nous pouvons maintenant remplacer la demi-épaisseur l en utilisant la nouvelle relation pour v_{in} pour obtenir $v_{in}^2 = \eta v_{out}/L$. Nous définissons le nombre de Mach d'Alfvén dans l'écoulement comme $M_{in} = v_{in}/v_{A,in}$ qui est également le taux de reconnexion sans dimension. Si nous utilisons la définition du nombre de Reynolds magnétique basée sur la vitesse d'Alfvén dans l'écoulement $R_{m_{in}} = Lv_{A,in}/\eta$, nous pouvons exprimer la vitesse v_{in} sous forme adimensionnelle : $M_{in}^2 = (1/R_{m,in} \cdot (v_{out}/v_{A,in})$.

Nous devons maintenant exprimer v_{out} et pour ce faire, nous allons partir de l'équation du mouvement dans un état stationnaire :

$$\rho(\mathbf{v} \cdot \nabla)\mathbf{v} = -\nabla p + \mathbf{j} \times \mathbf{B} \tag{1.3}$$

Si l'on néglige les effets de pression, la composante x de cette équation s'écrit :

$$(\mathbf{j} \times \mathbf{B})_x = jB_{out} = B_{in}B_{out}/(\mu_0 l) \tag{1.4}$$

en utilisant la définition de **j** écrite ci-dessus, qui est aussi la force de Lorentz le long de la feuille de courant. Cette force accélère le plasma au repos à la feuille neutre jusqu'à v_{out} sur une distance L. Ce terme est égal au terme d'inertie :

$$\rho(\mathbf{v} \cdot \nabla)v_x = \frac{\rho v_{out}^2}{L} \tag{1.5}$$

Donc, en revenant à l'équation 1.3 mais en regardant uniquement la composante x, nous pouvons écrire :

$$\frac{\rho v_{out}^2}{L} = \frac{B_{in}B_{out}}{\mu_0 l} \tag{1.6}$$

et en utilisant l'équation 1.1 ainsi que la conservation du flux magnétique qui s'écrit $v_{in}B_{in} = v_{out}B_{out}$, nous obtenons l'expression suivante pour v_{out} :

$$v_{out} = \frac{B_{in}}{\sqrt{\mu_0 \rho}} = v_{A,in} \tag{1.7}$$

Le taux de reconnexion que nous avons introduit précédemment s'écrit alors :

$$M_{in} = \frac{v_{in}}{v_{A,in}} = \frac{1}{\sqrt{R_{m,in}}} \tag{1.8}$$

Nous pouvons maintenant estimer le temps caractéristique de reconnexion à partir d'une longueur caractéristique L et de la "vitesse de reconnexion" v_{in} : $\tau = L/v_{in}$, et en utilisant l'équation 1.8 nous obtenons $\tau = \sqrt{R_{m,in}}L/v_{A,in}$. Les valeurs typiques dans une éruption solaire sont $R_{m,in} = 10^8$, $v_{A,in} = 100\ km/s$ et $L = 10^4\ km$ qui donnent un temps de reconnexion de 11 jours. Les valeurs typiques pour le vent solaire sont $R_{m,in} = 10^{11}$, $v_{A,in} = 100\ km/s$ et $L = 10^3\ km$, ce qui donne un temps de reconnexion de 37 jours. Le problème est que ces valeurs sont beaucoup trop grandes car le temps de reconnexion observé pour les éruptions solaires et le vent solaire se situe entre la minute et l'heure.

Le modèle de Sweet-Parker est un modèle de "reconnexion lente" car il n'était pas capable de décrire la reconnexion observée, c'est pourquoi en 1964, le physicien Harry Petschek a développé un modèle de "reconnexion rapide" comme une amélioration du modèle de Sweet-Parker. L'un des changements nécessaires pour obtenir des taux de reconnexion plus rapides était le rétrécissement de la taille de la région de diffusion L, ce qui entraîne la diminution de l'épaisseur de la nappe de courant et l'augmentation de la vitesse du processus de reconnexion. Par rapport au modèle de Sweet-Parker, la largeur de la région de diffusion est égale à sa hauteur comme on peut le voir sur la figure 1.3. Le modèle de Petschek donne un taux de reconnexion en accord avec les observations mais reste très controversé. Le problème vient du fait que le modèle nécessite deux ondes de choc sans collision et en mode lent qui peuvent être maintenues pendant la totalité du temps nécessaire au processus de reconnexion en régime permanent.

Topologie de la reconnexion magnétique Jusqu'à présent, j'ai détaillé le cas général de la reconnexion, mais je vais maintenant traiter plus spécifiquement de la reconnexion à la Terre. La magnétopause côté jour et la queue magnétique côté nuit sont les deux sites les plus proches de la Terre où la reconnexion magnétique peut avoir lieu, mais elle peut également se produire dans de nombreux autres contextes astrophysiques où de forts champs magnétiques sont impliqués et où la topologie des lignes de champ magnétique le permet.

L'image communément acceptée de la reconnexion magnétique (voir figure 1.4) est la suivante : nous avons d'abord deux régions de plasma différentes avec des lignes de champ magnétique de direction opposée situées en haut et en bas, séparées par une feuille de courant. Le plasma reconnecté est éjecté sur les côtés dans la "région d'échappement" et les délimitations entre l'échappement et les deux premières régions de plasma sont appelées les "séparatrices". Ces séparatrices se croisent au milieu de la feuille de courant en formant une forme de X, dont le point central est appelé le "point X". Le point X est l'endroit physique où il n'y a pas de champ magnétique, car c'est l'endroit précis où les lignes de champ se brisent et se reconnectent, et la ligne X est le prolongement du point X qui coupe la région d'échappement en deux là où les lignes de champ magnétique changent de direction.

À la longueur d'échelle inférieure à la longueur d'inertie des ions, la région est appelée région de diffusion des ions (ou IDR). C'est l'endroit où les ions sont démagnétisés des lignes de champ magnétique mais où les électrons sont toujours gelés dans le plasma. À l'intérieur de l'IDR, à une longueur d'échelle inférieure à la longueur d'inertie des électrons, se trouve la région de diffusion des électrons (ou EDR), où se produisent les échanges d'énergie. L'un des points d'intérêt à l'intérieur de l'EDR est le point de stagnation, c'est-à-dire l'endroit où la vitesse des électrons dans le plan de reconnexion est nulle. L'emplacement de ce point de stagnation est déterminé par l'asymétrie du système.

La reconnexion à la magnétopause de la Terre est appelée "reconnexion asymétrique" alors que la reconnexion à la queue magnétique est appelée "reconnexion symétrique". Cela est dû au fait qu'au niveau de la queue magnétique, le plasma des deux côtés de la feuille de courant est le même, de sorte que l'image de la reconnexion dont je parlais est symétrique. Cependant, à la magnétopause, le plasma provenant de la magnétosphère est très différent du plasma de la magnétogaine, les lignes de champ magnétosphérique étant comprimées par le vent solaire, l'intensité du champ magnétique y est beaucoup plus forte et la densité de particules plus faible. L'image de la reconnexion est alors asymétrique, une moitié de l'échappement étant plus petite que l'autre. Cela a également pour effet de déplacer le point de stagnation, normalement situé au point X dans une reconnexion symétrique, vers la magnétosphère. Les figures 1.4 et 1.5, qui résument les explications ci-dessus, montrent la différence entre la reconnexion symétrique et asymétrique.

La région de diffusion des électrons est le sujet principal de cette thèse, une description

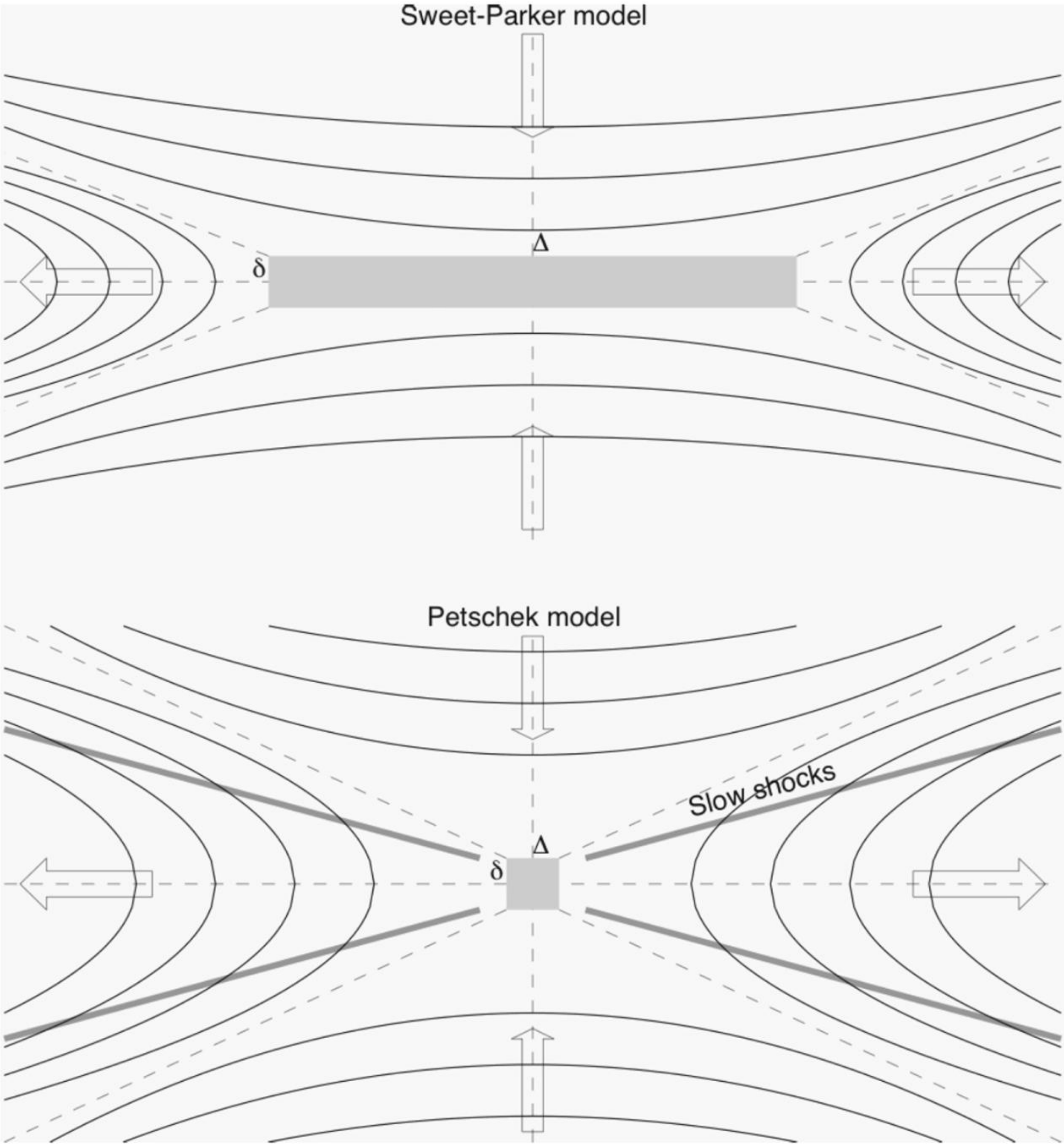


Figure 1.3: Figure comparant la géométrie du modèle de Sweet-Parker avec celle du modèle de Petschek. Image tirée de *Aschwanden* [2020]).

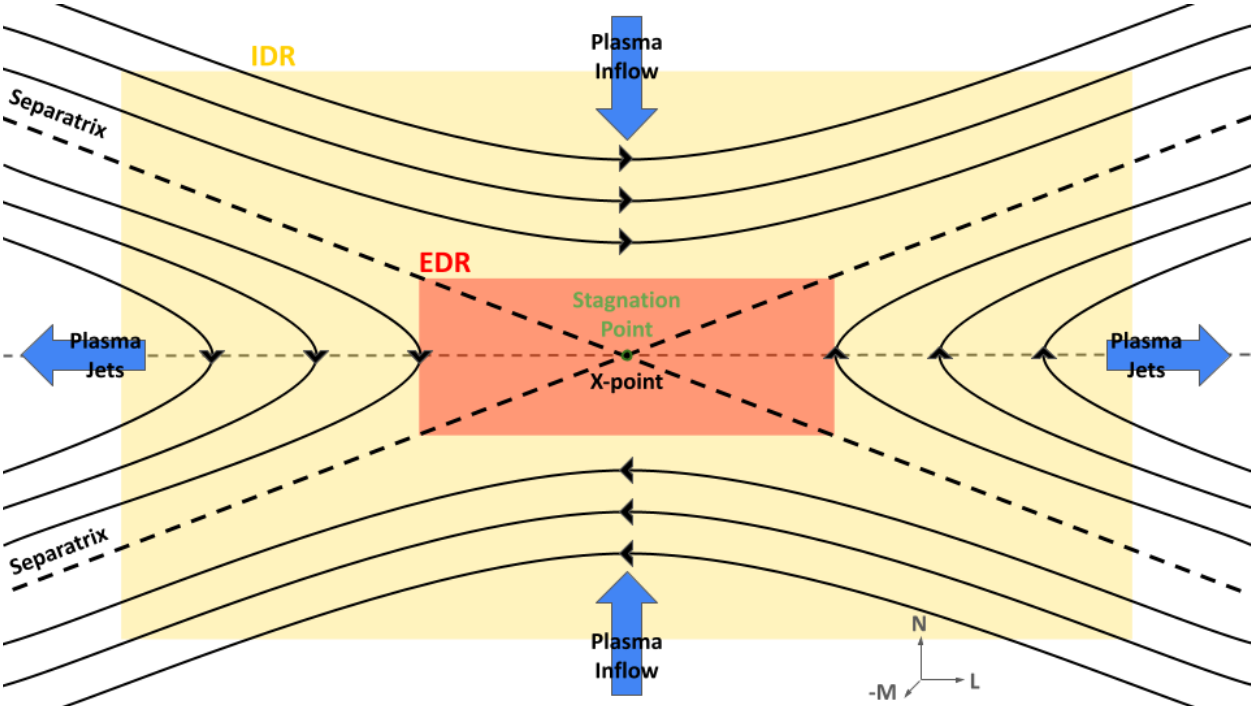


Figure 1.4: Figure montrant la topologie de la reconnexion magnétique symmétrique

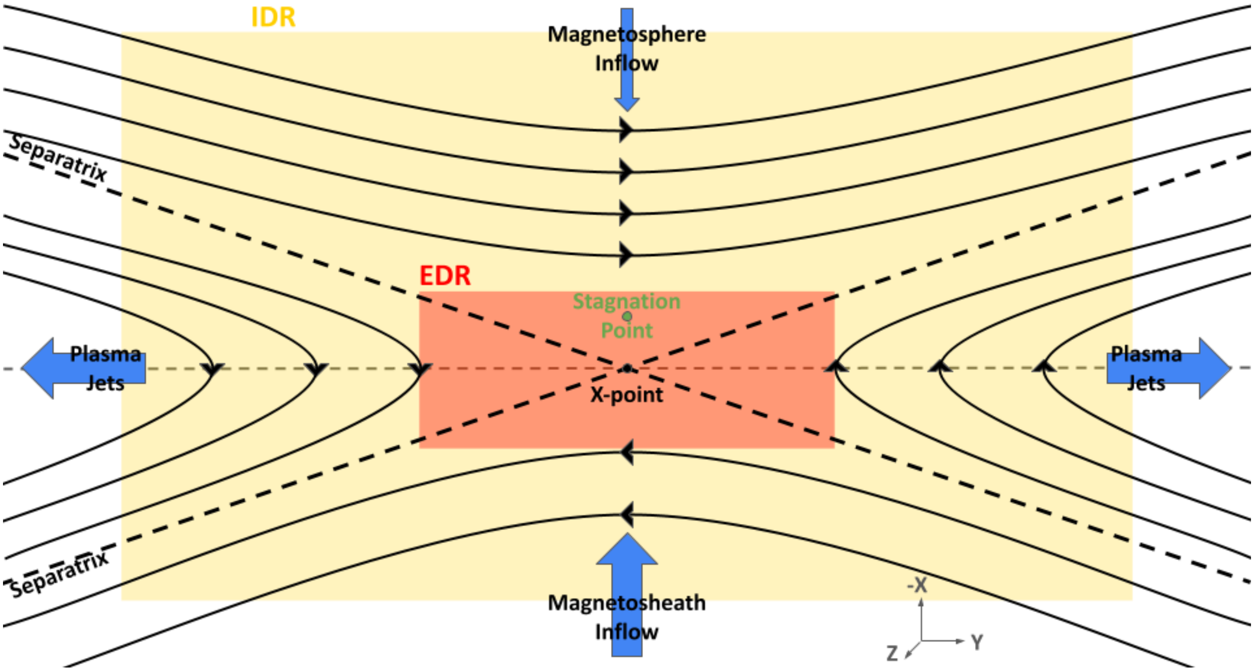


Figure 1.5: Figure montrant la topologie de la reconnexion magnétique asymmétrique

plus détaillée de la physique qui s'y déroule peut être trouvée dans la section 1.4. Le fait que sa taille soit beaucoup plus petite que celle de la région de diffusion des ions est l'une des raisons de la difficulté de sa détection dans les données. En raison de la vitesse de l'engin spatial, il n'est possible de rester que moins d'une seconde dans cette région. Pour surmonter ce problème, une très haute résolution était nécessaire et a été fournie par la mission Magnetospheric Multiscale.

1.3 La mission "Magnetospheric Multiscale" (MMS)

La mission Magnetospheric Multiscale (ou MMS) est une mission lancée le 13 mars 2015, et toujours active au moment de la rédaction de ce manuscrit (2022). L'objectif de la mission est d'étudier l'interaction entre le vent solaire du Soleil et la magnétosphère terrestre et plus précisément le phénomène de reconnexion magnétique. Elle est composée de quatre satellites identiques qui sont équipés d'instruments capables de mesurer les plasmas, les champs et les particules. Le plan de mission nominal était de 2 ans pendant lesquels les satellites voyageraient sur deux orbites terrestres différentes, hautement elliptiques. Les trois premiers mois de la mission sont consacrés à l'étude de la magnétopause côté jour, et les six derniers mois à l'étude de la reconnexion au niveau de la queue magnétique côté nuit.

MMS a été conçue pour succéder à la mission CLUSTER-II lancée par l'ESA en 2000, qui a permis d'étudier plusieurs aspects du processus de reconnexion magnétique (par exemple, *Hamrin et al.* [2011] a étudié les régions de conversion d'énergie avec Cluster, *Retinò et al.* [2006] a étudié la structure de la région de la séparatrice...). Plus précisément, l'un des principaux objectifs de la nouvelle mission MMS était d'approfondir la compréhension actuelle de la région de diffusion des électrons (EDR), ce qui n'était pas possible auparavant avec la résolution d'une minute fournie par CLUSTER. Il est possible de mieux comprendre cela en faisant ce simple calcul : l'orbite elliptique de CLUSTER a un périgée de 19 000 km et un apogée de 119 000 km, ce qui représente une distance d'environ 369 186 km que CLUSTER parcourt en 57 heures. Pour simplifier les choses, si nous supposons que la vitesse est constante tout au long de l'orbite, le calcul donne une vitesse de l'engin spatial de 1,8 km/s. Avec une résolution d'une minute, CLUSTER ne pourrait effectuer une mesure que tous les 108 km environ. C'est beaucoup plus que la taille de la région de diffusion des électrons qui n'est que de quelques kilomètres. Avec une résolution électronique de 30 ms, pour la même vitesse du vaisseau spatial, le MMS pourrait effectuer une mesure tous les 54 m, ce qui améliorerait considérablement les chances de fournir des échantillons prélevés à l'intérieur de la région de diffusion électronique.

1.4 La physique de la région de diffusion électronique (EDR)

L'un des objectifs de ma thèse était de trouver de nouveaux candidats pour la région de diffusion électronique, mais on peut légitimement se demander pourquoi il est nécessaire de chercher de nouveaux candidats. Il y a plusieurs réponses à cette question. Tout d'abord, j'ai dit précédemment que la mission MMS est la seule à permettre l'étude de l'EDR et qu'il n'y aura probablement pas d'autre mission capable de le faire dans les prochaines décennies, c'est donc le moment de faire ce genre d'étude. La deuxième réponse est que l'EDR est très petite, ses dimensions sont de l'ordre de quelques kilomètres. En raison de la vitesse des satellites de MMS, seuls quelques points de données peuvent être enregistrés dans cette région. L'identification de l'EDR dans les données du MMS n'est donc pas aussi facile qu'il n'y paraît. Depuis le lancement du MMS en 2015, seuls environ 35 événements EDR côté jour ont été signalés à ma connaissance, la plupart d'entre eux étant compilés dans *Webster et al.* [2018]. Avoir si peu d'événements rend les études statistiques sur les EDR plus difficiles, trouver de nouveaux candidats est alors crucial pour permettre des études plus poussées sur cette région. Plus généralement, la physique interne de l'EDR est une question très compliquée, il est donc nécessaire de disposer d'autant de cas que possible pour explorer la grande diversité des conditions que l'on peut trouver dans cette région.

Dans les prochaines sous-sections, je détaillerai des différentes signatures clés de l'EDR ainsi que de certains autres aspects physiques liés à l'EDR.

1.4.1 Conversion d'énergie

Lors de la reconnexion magnétique, la topologie des lignes de champ magnétique du plasma est modifiée, couplée à l'échange d'énergie des champs magnétiques vers les particules. Le changement topologique se produit au cœur de la région de diffusion des électrons, au point X. Cependant, l'échange d'énergie ne se produit pas au point X mais plutôt dans ce que l'on appelle la "région de dissipation", située du côté de la magnétosphère, loin du point X. Le processus de dissipation de l'énergie est régi par la loi d'Ohm généralisée, dérivée de l'équation du mouvement du fluide électronique, qui est la suivante :

$$\mathbf{E} + \mathbf{v}_e \times \mathbf{B} = -\frac{1}{ne}\nabla \cdot \mathbf{P}_e - \frac{m_e}{e}\left(\frac{\partial \mathbf{v}_e}{\partial t} + \mathbf{v}_e \cdot \nabla \mathbf{v}_e\right) \tag{1.9}$$

où $\mathbf{E}$ est le champ électrique, $\mathbf{v}_e$ est la vitesse de l'électron, $\mathbf{B}$ est le champ magnétique, $\mathbf{P}_e$ est le tenseur de pression des électrons.

Dans la région de diffusion électronique, les électrons sont démagnétisés, ce qui signifie qu'ils n'obéissent pas à la condition MHD idéale $\mathbf{E'} = \mathbf{E} + \mathbf{v}_e \times \mathbf{B} = 0$ où $\mathbf{E'}$ est le champ électrique dans le référentiel de repos du plasma. Ceci s'explique par le fait qu'aux grandes échelles sur de grandes périodes de temps, les gradients et la dérivée partielle par rapport au temps tendent tous vers zéro, mais ce n'est pas le cas dans l'EDR car nous considérons de très petites échelles sur de courtes périodes de temps. Vérifier si $\mathbf{E'} \neq 0$ est donc l'une des caractéristiques permettant d'identifier la région de diffusion électronique dans les données.

Une autre façon d'identifier la région de diffusion électronique, ou plutôt la région de dissipation à l'intérieur, est de regarder la dissipation Joule, qui est le taux de travail par volume effectué par le champ électrique sur le plasma exprimé par $\mathbf{J} \cdot \mathbf{E}$, avec $\mathbf{J}$ étant la densité de courant. On l'exprime aussi habituellement dans le cadre de repos de l'électron pour voir le travail effectué par les champs électriques non idéaux, il s'écrit alors $\mathbf{J} \cdot \mathbf{E'}$. La dissipation Joule est directement liée à la variation de l'énergie électromagnétique avec l'équation suivante :

$$\partial_t \varepsilon^m + \frac{c}{4\pi} \nabla \cdot (\mathbf{E} \times \mathbf{B}) = -\mathbf{J} \cdot \mathbf{E} \tag{1.10}$$

obtenue en utilisant une combinaison de l'équation de Vlasov avec les équations de Maxwell. Nous voyons que la dissipation Joule peut être positive ($\mathbf{J} \cdot \mathbf{E} > 0$) ou négative ($\mathbf{J} \cdot \mathbf{E} < 0$), ce qui affectera directement le signe de la variation de l'énergie électromagnétique $\partial_t \varepsilon^m = \frac{1}{8\pi}(\mathbf{B}^2 + \mathbf{E}^2)$. Si $\mathbf{J}.\mathbf{E}$ est positif, la variation de l'énergie électromagnétique est négative, ce qui signifie que le transfert d'énergie se fait des champs vers les particules et si $\mathbf{J} \cdot \mathbf{E}$ est négatif, le transfert d'énergie se fait des particules vers les champs, ce que l'on appelle respectivement l'effet de charge et l'effet dynamo.

À l'intérieur de l'EDR, l'effet de charge ($\mathbf{J} \cdot \mathbf{E} > 0$) est attendu avec un transfert d'énergie des champs vers les particules. Cependant, de nombreuses études ont montré qu'il peut également y avoir des valeurs négatives ou fluctuantes de $\mathbf{J} \cdot \mathbf{E}$ à l'EDR (voir la section 3.5).

1.4.2 Croissants d'électrons

L'un des aspects les plus importants de la mission MMS est de pouvoir visualiser les fonctions de distribution de la vitesse des électrons pour la première fois toutes les 30 ms. Cette cadence rapide permet l'étude de quelques fonctions de distribution prises à l'intérieur de l'EDR, ce qui n'était pas possible auparavant en raison des résolutions inférieures des missions précédentes.

Une fonction de distribution de la vitesse des électrons (eVDF) est un objet tridimensionnel représentant la distribution des vitesses des électrons dans l'espace des vitesses. Elle peut être représentée dans le système de coordonnées GSE (Vx, Vy, Vz) ou dans le système de coordonnées aligné sur le champ ($V_{||}$, $V_{\perp_1}$, $V_{\perp_2}$). L'eVDF complète étant un objet 3D, son étude sous cette forme n'est pas pratique, c'est pourquoi on tend à utiliser la fonction de distribution réduite à la place. La fonction de distribution réduite est le résultat de l'intégration de la fonction de distribution complète sur un certain angle autour d'un plan de l'espace. Cette appellation "fonction de distribution de la vitesse des électrons" est souvent utilisée de manière inexacte pour désigner la "fonction de distribution réduite de la vitesse des électrons" et c'est pourquoi toutes les "fonctions de distribution réduites" présentées tout au long de cette thèse seront souvent simplement appelées "fonctions de distribution". On s'attend à ce que les fonctions de distribution soient gyrotropiques, ce qui signifie qu'elles sont symétriques circulairement.

Avant le lancement de la mission MMS, seules quelques études utilisant des simulations ont été menées sur les fonctions de distribution des électrons pendant la reconnexion asymétrique. *Egedal et al.* [2011] a mis en évidence une anisotropie (aussi appelée agyrotropie) significative des fonctions de distribution des électrons dans la région d'afflux de la magnétosphère, considérée comme importante pour la génération de couches de courant allongées dans l'échappement. 3 ans plus tard, *Hesse et al.* [2014] a été le premier à montrer, à l'aide de simulations de type Particle-In-Cell (PIC), que la distribution des électrons au point de stagnation des électrons était constituée de deux populations d'électrons différentes, une distribution gyrotrope aux vitesses inférieures et une distribution en forme de croissant fortement non gyrotrope sur un seul côté aux vitesses supérieures. La fonction de distribution obtenue à partir de la simulation est visible sur la figure 1.6.

.

La présence de cette agyrotropie est due aux électrons énergétiques de la magnétosphère qui sont capables de se mélanger aux populations d'électrons de la magnétosphère en traversant la feuille de courant. Les électrons énergétiques sont ensuite réfléchis sur leur chemin de retour vers la magnétosphère, les piégeant du côté magnétosphérique de la feuille de courant. Pour aller un peu plus loin, cette forme particulière dans l'espace des vitesses peut être expliquée comme dans *Bessho et al.* [2016], en suivant la trajectoire d'un électron unique provenant de la magnétosphère ($z < 0$) et entrant dans l'EDR. Considérons un cas simplifié où le champ magnétique s'écrit ($B_x = bz, B_y = 0, B_z = 0$) avec b une constante positive et $z = z_X = 0$ le point neutre magnétique, et le champ électrique s'écrit ($E_x = 0, E_y = 0, E_z = -kzU(z)$) avec $k(> 0)$ la grandeur de la pente de E_z en $z > 0$ et U(z) la fonction de pas. L'équation de mouvement non relativiste dans le plan (y, z), en supposant $v_x = 0$, pour un électron interagissant avec de tels champs est la suivante :

$$m_e \frac{dv_y}{dt} = -\frac{e}{c} v_z bz, \tag{1.11}$$

$$m_e \frac{dv_z}{dt} = ekzU(z) + \frac{e}{c} v_y bz, \tag{1.12}$$

où m_e est la masse de l'électron. Dans la magnétosphère ($z \leq 0$), le terme de champ électrique disparaît et l'énergie de l'électron est conservée. Dans la magnétosphère ($z > 0$), E_z et B_x sont proportionnels à z, et si nous étudions le mouvement de l'électron dans le référentiel se déplaçant dans la direction y avec une vitesse d'impulsion $cE_z/B_x = -ck/b$, E_z disparaît tant que $k < b$ et l'énergie de l'électron est également conservée. L'hypothèse selon laquelle $cE_z/B_x = -ck/b = cst$ est basée sur des simulations. Si nous appliquons la formule de conservation de l'énergie, nous obtenons les deux cas suivants :

$$\frac{1}{2} m_e v^2 = cst \iff (v_y^2 + v_z^2) = cst \; quand \; z \leq 0 \tag{1.13}$$

$$\frac{1}{2} m_e v^2 = cst \iff (v_y - v_{y0})^2 + v_z^2 = cst \iff (v_y + ck/b)^2 + v_z^2 = cst \; lorsque \; z > 0 \tag{1.14}$$

Pour trouver la valeur de v_{y0} dans le cas où $z > 0$, il faut voir que lorsque $\frac{dv_z}{dt} = 0$, la valeur de v_y est $-ck/b$. Les deux cercles visibles sur la figure 1.7a représentent les deux seules trajectoires que peut prendre l'électron dans l'espace des vitesses en fonction de la valeur de z, formant un croissant dans le plan (v_y, v_z). Sur la figure 1.7b, nous pouvons voir la trajectoire réelle de l'électron dans l'espace physique, le champ magnétique étant le long de la direction x, l'électron tourne autour de celui-ci dans un sens ou dans l'autre et change de direction lorsqu'il traverse la magnétopause.

Quelques éléments à garder à l'esprit pour les prochains chapitres de cette thèse est tout d'abord que toutes les explications données ici à l'aide de simulations ont été faites en utilisant les coordonnées GSE. Dans ce système de coordonnées et en utilisant les conventions standard pour les conditions de départ de la simulation, le croissant sera situé dans le plan

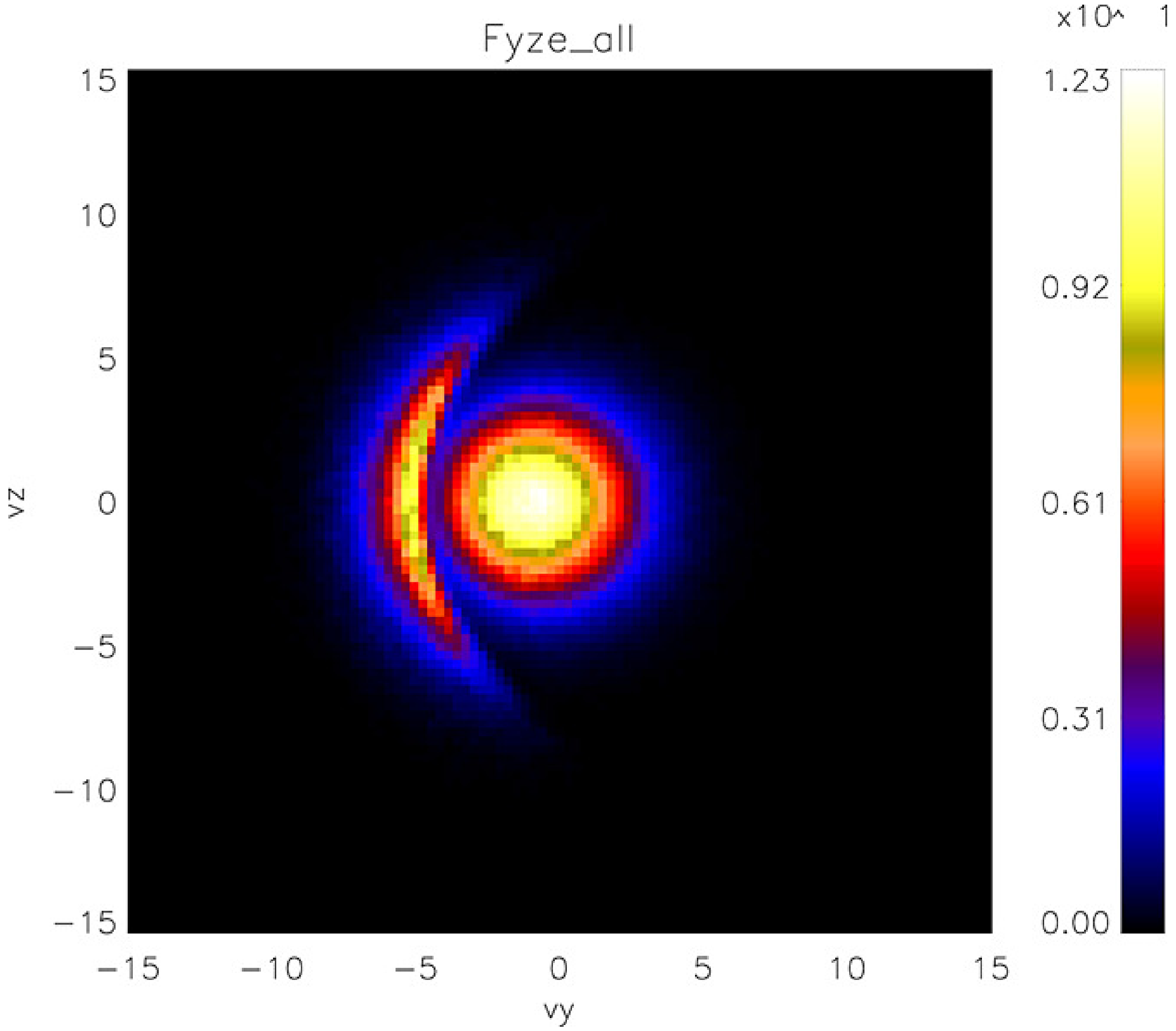


Figure 1.6: Figure montrant une fonction de distribution réduite obtenue par intégration d'une fonction de distribution complète prise au point de stagnation du flux dans une simulation de type Particle-In-Cell. La présence de deux populations d'électrons différentes est clairement visible avec une distribution gyrotrope au centre et une distribution en forme de croissant fortement non-gyrotrope située à environ -5 V_y. Image tirée de *Hesse et al.* [2014].

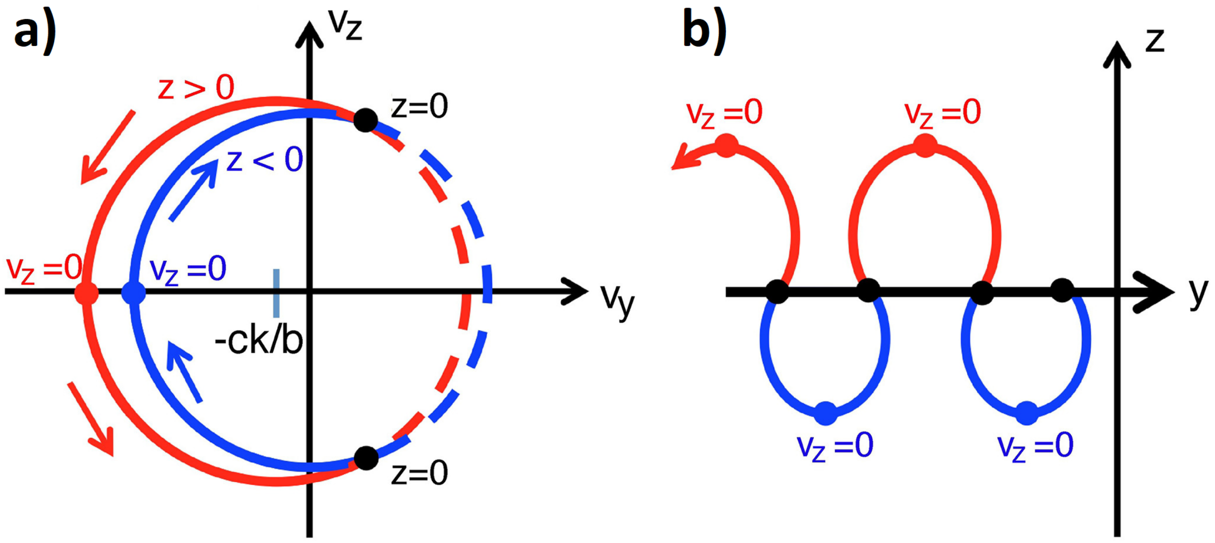


Figure 1.7: Figure montrant les trajectoires des électrons expliquant l'origine des croissants d'électrons au niveau de la fonction de distribution des électrons. Image tirée de *Bessho et al.* [2016].

(V_y, V_z), et plus précisément du côté négatif de V_y. Cependant, lorsqu'on utilise les données réelles d'un vaisseau spatial, on préfère généralement visualiser les fonctions de distribution des électrons dans le système de coordonnées aligné sur le champ, ce qui donne des croissants d'électrons dans le plan $(V_{\perp 1}, V_{\perp 2})$, et plus précisément du côté positif de $V_{\perp 1}$. De plus, les explications données ici indiquent que ces "croissants d'électrons" ne devraient être visibles que du côté magnétosphérique de la nappe de courant, alors que je montrerai plus tard que ce n'est pas toujours le cas et que des croissants d'électrons peuvent aussi être visibles du côté magnétosphérique.

La première observation de croissants d'électrons dans les données a été présentée dans *Burch et al.* [2016] où un événement d'EDR a été identifié dans les données du MMS un mois seulement après son lancement. La figure **??** illustre cette détection où l'on peut clairement voir l'évolution de la structure en forme de croissant lorsque le MMS traverse l'EDR. Dans le plan $(V_{\perp 1}, V_{\perp 2})$, le croissant est situé dans les valeurs positives de $V_{\perp 1}$, puis le croissant se transforme lentement en cercle, ce qui signifie que les électrons sont remagnétisés lorsque le MMS s'éloigne de l'EDR.

1.4.3 Spécificités de la reconnexion avec champ guide

Lorsque les lignes de champ magnétique qui se reconnectent ne sont pas complètement antiparallèles (ce qui signifie que l'angle de cisaillement entre les champs magnétiques magnétosphérique et magnétosphérique est inférieur à 180°), les composants hors plan ne se reconnectent pas. Le champ magnétique hors plan restant est appelé "champ guide". Pour le visualiser, on se place généralement dans le système de coordonnées "LMN", il s'agit d'une rotation du système de coordonnées GSE ou GSM de sorte que la composante L soit celle qui présente la plus forte variation de champ magnétique et que la N soit celle qui présente la plus faible variation de champ magnétique. Ce système de coordonnées est généralement trouvé à l'aide de l'analyse de variance minimale (MVA). Les différents systèmes de coordonnées ainsi que l'explication de la MVA se trouvent dans les sections annexes A et B. Dans le système de coordonnées "LMN", l'intensité du champ guide est mesurée par le rapport des composantes du champ magnétique B_M/B_L. Les événements dont les valeurs sont inférieures à 0,1 ou 0,2 sont considérés comme des événements à faible champ guide (ou à fort cisaillement). Pour les valeurs supérieures, les classifications de l'intensité du champ guide peuvent varier selon les auteurs, mais les événements dont les valeurs sont supérieures à 1 sont toujours considérés comme des événements à champ guide élevé (ou à faible cisaillement).

Des études précédentes ont montré que la présence d'un champ guide affecte fortement les propriétés de la reconnexion ainsi que la dynamique des particules (voir *Pritchett and Coroniti* [2004]). Il a également été démontré que la présence d'un champ guide rend l'intensité du champ magnétique non nulle au point X, ce qui peut magnétiser les électrons près du point X, réduisant ainsi la taille du rayon de gyration des électrons par rapport à la taille de la couche de courant (voir *Swisdak et al.* [2005]). La remagnétisation des électrons au point X est également responsable de la diminution de l'intensité des croissants de la fonction de distribution de la vitesse des électrons observée dans l'EDR (voir *Hesse et al.* [2016]).

1.5 Aperçu de l'apprentissage automatique

L'apprentissage automatique fait partie de ce que l'on appelle "l'intelligence artificielle" et regroupe des techniques permettant aux algorithmes de s'améliorer automatiquement grâce à l'expérience et à l'utilisation de données. Cette section a pour but d'introduire les concepts de base de l'apprentissage automatique qui seront nécessaires à la compréhension des travaux abordés dans la suite de cette thèse. Nous commencerons par un bref historique des principales avancées de l'apprentissage automatique avant d'énumérer quelques-unes de ses applications en physique des plasmas.

1.5.1 Un bref historique de l'apprentissage automatique

La première utilisation du terme "apprentissage automatique" remonte à 1959, lorsqu'Alan Samuel, un informaticien américain d'IBM, l'a utilisé pour décrire le "programme de jeu de dames" qu'il avait conçu en 1952 (*Samuel* [1959]). La mémoire des ordinateurs étant très limitée à l'époque, il n'était pas possible pour l'algorithme de prédire tous les résultats possibles pour gagner. Samuel a donc créé ce que l'on appelle aujourd'hui le "prunning alpha-beta" : sur la base de la position des pièces sur le plateau, l'algorithme calcule une fonction de notation qui mesure la chance actuelle de gagner. Pour choisir le prochain coup, le programme recherche quelques coups à l'avance la configuration du plateau qui lui donne la meilleure valeur de score, et joue en conséquence.

La première méthode d'apprentissage qu'il a utilisée était appelée "apprentissage par cœur" et consistait simplement à enregistrer chaque configuration de plateau rencontrée associée à la valeur de la fonction de score calculée. De cette façon, lorsque l'on joue une partie avec une configuration de plateau déjà rencontrée, aucun calcul n'est nécessaire pour cette position, ce qui permet d'augmenter la profondeur de l'arbre de recherche pour le prochain coup. La deuxième méthode d'apprentissage, appelée "apprentissage par généralisation", était utilisée pour modifier les paramètres de la fonction de notation en faisant jouer le programme contre une autre version de lui-même.

Parallèlement aux travaux de Samuel, un autre informaticien américain, Franck Rosenblatt, a créé en 1957 le premier algorithme Perceptron (*Rosenblatt* [1958]) basé sur les travaux de Warren McCulloch et Walter Pitts sur la modélisation mathématique des neurones artificiels (appelés plus tard neurones MCP). Fondamentalement, le Perceptron est la combinaison de plusieurs neurones MCP où chaque neurone a un seuil d'activation et un poids de sorte qu'une sortie est produite en calculant la somme pondérée des entrées. La sortie de son algorithme était soit 0 soit 1, ce qui permettait une classification binaire des formes ou des lettres. Il a été le premier à montrer qu'un algorithme pouvait apprendre à partir de données. Au cours d'une phase d'apprentissage, l'algorithme reçoit des échantillons étiquetés et met à jour ses valeurs internes de sorte que la prochaine fois qu'il reçoit le même échantillon, sa réponse corresponde à l'étiquette, en se basant sur le fait que les échantillons ayant la même étiquette devraient produire des entrées similaires pour l'algorithme et donc une sortie similaire.

Dix ans plus tard, en 1967, Thomas Cover et Peter Hart publient l'algorithme de classification de motifs Nearest Neighbor (*Cover and Hart* [1967]). C'est la base de ce que l'on appelle aujourd'hui kNN (k-nearest neighbors) qui classe un point de données en fonction de la classification de ses voisins en supposant que des choses similaires existent à proximité les unes des autres dans l'espace des paramètres. En 1970, Seppo Linnainmaa a mis en œuvre le premier algorithme de rétropropagation permettant de traiter une erreur à la sortie d'un réseau neuronal, puis de la distribuer en amont dans les couches du réseau pour ajuster les couches cachées de neurones/nœuds (*Linnainmaa* [1976]). L'année suivante, Messenger et Mandell ont mis en œuvre pour la première fois un arbre de décision dans le cadre du projet THAID (*Messenger and Mandell* [1972]).

Mais dans le monde de l'IA, les années 70 sont surtout connues comme la "période d'hiver de l'IA". Il s'agissait d'une période de réduction des financements et de l'intérêt pour la recherche sur l'IA en raison du pessimisme de la communauté de l'IA et de la presse, ce qui a entraîné un manque de recherches sérieuses et de percées dans le domaine. L'hiver de l'IA a fait apparaître le schisme entre l'IA et l'apprentissage automatique. Jusque-là, l'apprentissage automatique avait été utilisé comme programme d'entraînement pour l'IA, mais à la fin des années 70, les chercheurs en informatique et en IA ont abandonné les réseaux neuronaux pour se tourner vers l'utilisation d'approches logiques et basées sur la connaissance. Un autre hiver de l'IA, bien que moins important, s'est produit à la fin des années 80. L'intérêt pour le domaine de l'IA avait recommencé à croître avec l'utilisation de programmes d'IA appelés "systèmes experts" par les grandes entreprises du monde entier. Cependant, les nombreux problèmes techniques de ces systèmes ont conduit à l'effondrement du marché du matériel spécialisé en IA en 1987.

Néanmoins, de nouveaux avancées dans le domaine de l'apprentissage automatique ont été faites par les chercheurs après l'hiver de l'IA, avec tout d'abord l'invention des algorithmes de "boosting". Le concept de boosting a été présenté pour la première fois en 1990 dans un article intitulé ”The Strength of Weak Learnability”, par Robert Schapire (*Schapire* [1990]). Ces algorithmes sont constitués de classificateurs faibles qui, seuls, donnent de mauvais résultats, mais si un grand nombre d’entre eux sont entraînés, la réponse moyenne est un classificateur fort. Il se base sur le phénomène connu appelé ”sagesse de la foule” où par exemple, il est possible de deviner le nombre de haricots dans un bocal en prenant la moyenne des réponses de chaque personne dans une foule, plus la foule est grande, plus le résultat sera proche de la vérité. En 1997, Jürgen Schmidhuber et Sepp Hochreiter ont décrit le modèle d’un réseau neuronal connu sous le nom d’algorithme de mémoire à long-court terme (ou LSTM) (*Hochreiter and Schmidhuber* [1997]). Les LSTM peuvent apprendre des tâches qui nécessitent la mémorisation d’événements qui ont eu lieu des milliers d’étapes discrètes auparavant, ce qui a permis d’énormes avancées dans les années suivantes dans le domaine de la reconnaissance vocale.

De l’an 2000 à aujourd’hui, l’apprentissage automatique est utilisé dans le monde entier grâce à des avancées significatives, permettant la création d’algorithmes de reconnaissance faciale, de véhicules à conduite autonome et de nombreuses autres applications. L’augmentation de la puissance de calcul au cours des deux dernières décennies a permis d’améliorer les architectures précédemment utilisées, comme les réseaux neuronaux convolutifs (CNN), et d’en créer de nouvelles, comme les réseaux adversariaux génératifs (GAN). Cela a permis à l’apprentissage automatique de faire partie intégrante de nombreux services et applications logiciels largement utilisés.

1.5.2 L’apprentissage automatique dans la physique des plasmas spatiaux

L’utilisation de techniques d’apprentissage automatique dans le domaine de la physique des plasmas spatiaux est liée au domaine de la météorologie spatiale. Ce n’est pas un sujet nouveau puisque des articles utilisant cette technologie ont déjà été publiés dans les années 1990, la prédiction de l’indice géomagnétique étant l’une des premières applications dans ce domaine (*Costello* [1998]). La prédiction de l’indice Kp, qui est une représentation globale de l’amplitude maximale de la variation géomagnétique sur des intervalles de 3 heures UT et qui peut être utilisé pour déterminer la probabilité d’observer des aurores le plus tôt possible, est toujours un sujet de recherche en cours comme le prouvent les publications récentes traitant de ce sujet (*Wintoft et al.* [2017]).

Les premières applications d’apprentissage automatique dans ce domaine étaient basées sur des architectures de réseaux neuronaux traditionnelles, mais de plus en plus d’architectures de pointe sont utilisées aujourd’hui grâce à l’augmentation de la puissance de calcul depuis les années 1990. Par exemple, l’utilisation de CNN pour analyser les images solaires afin de prédire les éruptions solaires n’était pas possible il y a quelques décennies, mais elle constitue aujourd’hui l’un des moyens possibles de s’attaquer à ce problème (*Huang et al.* [2018]). Il existe également de nombreuses autres applications d’apprentissage automatique en physique des plasmas spatiaux, allant de l’estimation de la densité électronique plasmasphérique à la prédiction du paramètre foF2 en physique ionosphérique. Nombre d’entre elles sont regroupées dans *Camporeale* [2019] qui est un article passant en revue l’utilisation de l’apprentissage automatique dans le domaine de la météorologie spatiale.

Pour terminer cette section sur l'apprentissage automatique, je vais maintenant examiner un peu plus en détail deux applications récentes de l'apprentissage automatique en physique des plasmas qui ont été publiées au cours des trois dernières années. L'idée est ici de montrer comment ces applications de plasma spatial passent des données aux résultats exploitables.

Détection automatique d'ICMEs

La première application dont nous allons parler est détaillée dans l'article "Automatic detection of Interplanetary Coronal Mass Ejections from in-situ data : a deep learning approach" publié dans Astrophysical Journal en 2019 *Nguyen et al.* [2019]. Les éjections de masse coronale (ou CME) sont des événements astrophysiques au cours desquels de grandes quantités de plasma solaire sont éjectées à grande vitesse. Les éjections de masse coronale interplanétaires (ICME) sont la contrepartie planétaire des éjections de masse coronale qui se déplacent hors de l'atmosphère solaire. Leur interaction avec les champs magnétiques planétaires peut déclencher des tempêtes géomagnétiques susceptibles d'avoir un impact important sur la magnétosphère de la Terre et donc sur l'activité humaine.

La mission WIND lancée en 1994 a fourni des années de données permettant de nombreuses études sur les ICME qui ont suggéré plusieurs critères pour les détecter. Parmi les caractéristiques les plus communes de ces ICME, on trouve un champ magnétique renforcé et en rotation régulière, une faible température des protons et un faible bêta du plasma. Cependant, en raison de la forte variabilité de ces événements, tous ces critères ne sont pas réunis pour chaque ICME, ce qui rend difficile une détection basée sur un seuil. La détection des ICME a été faite par inspection visuelle des données, ce qui conduit à des catalogues biaisés selon l'observateur. Cet article présente une méthode qui a été développée pour détecter automatiquement les ICME dans les données de la mission WIND en utilisant une combinaison de CNN et d'algorithme de détection des pics.

Dans la phase de prétraitement, ils ont utilisé une fusion de plusieurs listes d'ICME de WIND pour un total de 657 ICME. Les instruments WIND fournissent 30 caractéristiques d'entrée et ils ont calculé 3 variables supplémentaires, ce qui donne un total de 33 variables d'entrée. Les données ont été regroupées en fenêtres de différentes tailles allant de 1h à 100h glissant sur les ensembles de données à une période de 10 minutes. Pour chaque fenêtre, un paramètre de similarité est calculé, donnant 0 lorsqu'aucune ICME ne croise la fenêtre et 1 lorsqu'une fenêtre correspond parfaitement à une ICME.

Un CNN différent a été formé pour chaque taille de fenêtre afin d'apprendre et de prédire la similarité d'une fenêtre donnée associée aux données. L'architecture des CNN est présentée dans la figure 1.8.

La figure 1.9 montre une comparaison entre la prédiction attendue des ICME (en rouge) et la prédiction de l'algorithme (en vert) sur une période de 2 semaines après avoir appliqué un filtre gaussien sur les données. Un algorithme de détection des pics est appliqué sur les similarités calculées afin de créer les zones de prédiction vertes. Nous voyons que les prédictions de l'algorithme sont très proches de la classification ICME attendue.

La méthode présentée dans cet article montre que l'apprentissage automatique, et plus

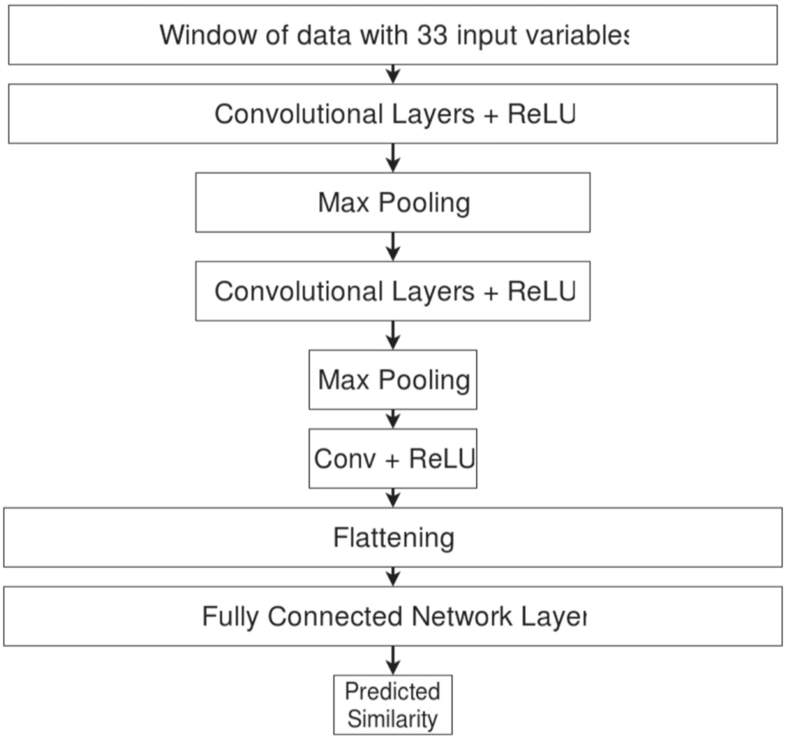


Figure 1.8: Figure montrant l'architecture du CNN utilisé pour détecter les ICMEs. Image tirée de *Nguyen et al.* [2019].

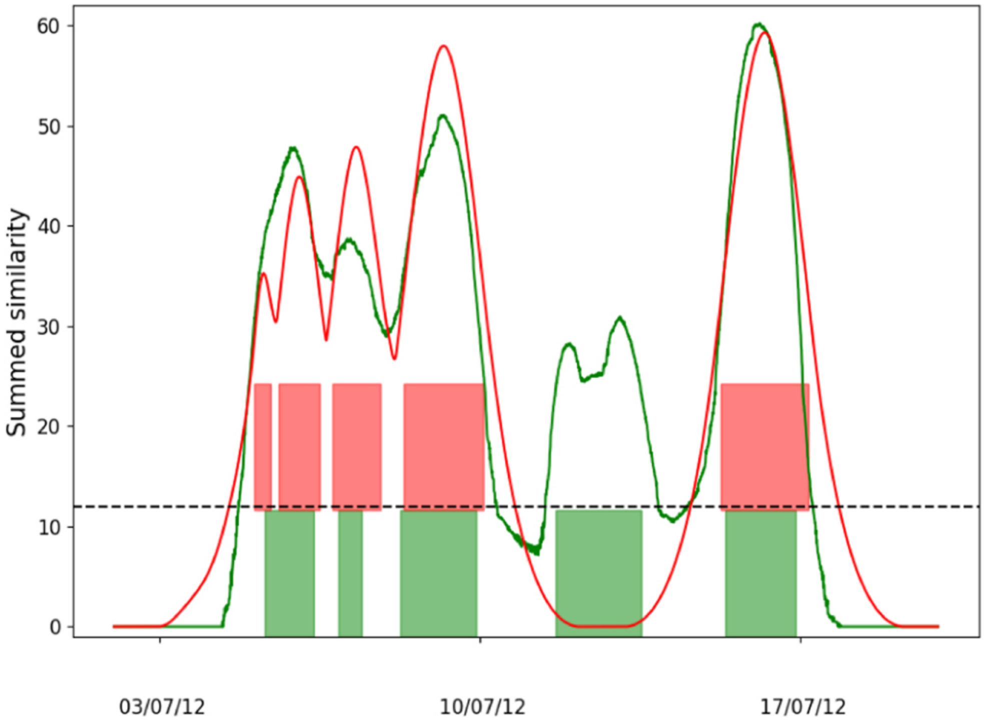


Figure 1.9: Figure montrant la comparaison entre les régions d'ICME attendues (en rouge) et les ICME prédites (en vert) par l'algorithme. Image tirée de *Nguyen et al.* [2019].

précisément les CNN, peut être utilisé pour créer des catalogues impartiaux et reproductibles d'événements astrophysiques.

Classification des régions de plasma à l'aide de l'apprentissage automatique

Une autre application récente de l'apprentissage automatique est détaillée dans l'article "Automatic Classification of Plasma Regions in Near-Earth Space With Supervised Machine Learning : Application to Magnetospheric Multi Scale 2016-2019 Observations" publié en 2020 (*Breuillard et al.* [2020]). Leur objectif était de classifier automatiquement 10 régions plasma clés à l'aide des données de séries temporelles de la mission MMS.

La classification des régions plasma se fait généralement par inspection visuelle en raison de la forte variabilité des paramètres du plasma qui ne permet pas aux méthodes basées sur des seuils de fonctionner parfaitement, une inspection humaine est toujours nécessaire pour valider les résultats. La classification est donc une tâche fastidieuse mais nécessaire car c'est le seul moyen de produire des études statistiques sur ces régions.

Pour former et construire un algorithme d'apprentissage supervisé, une grande quantité de données étiquetées est nécessaire, ce qui prend beaucoup de temps. Heureusement, un tel étiquetage manuel des régions du plasma est déjà effectué par le système SITL (Scientist-in-the-loop) qui implique un scientifique expert désigné pour sélectionner les données intéressantes qui seront transmises au sol sous forme de données à haute résolution. L'idée était d'utiliser un analyseur syntaxique pour convertir tous les commentaires en étiquettes standardisées pouvant être utilisées pour la phase d'apprentissage.

Leur modèle utilise un total de 12 variables, dont les composantes du champ magnétique, la densité ionique, la magnitude de la vitesse globale totale et ses composantes, les températures totale, parallèle et perpendiculaire. Les entrées devant être de la même taille, l'ensemble de données est regroupé sous forme de séries temporelles correspondant aux intervalles de temps étiquetés, divisés en tranches égales de 3 minutes.

L'architecture du modèle est appelée réseau entièrement convolutif (FCN), un sous-type d'apprentissage profond compétitif de CNN pour la classification des séries temporelles, composé d'une séquence de trois blocs de convolution temporelle suivis d'un bloc de mise "global average pooling". Chaque bloc de convolution temporelle est conçu de la même manière : une couche de convolution, une couche de normalisation de lot et une fonction d'activation ReLU (Rectified Linear Unit).

Une fois que le modèle est entraîné, il peut être utilisé sur des données MMS non vues pour classer les régions de plasma comme on peut le voir sur la figure 1.10. Le modèle classe les données par tranches de 3 minutes. La ligne noire montre la classification du modèle tandis que les sections colorées montrent la classification humaine qui aurait été faite sur cette série temporelle. Toutes les sections ne peuvent pas être facilement classifiés par une inspection humaine, d'où la quantité de sections de données qui ne sont pas colorés. Nous constatons que leur modèle se comporte bien par rapport aux prédictions humaines, malgré la grande variabilité des paramètres observables.

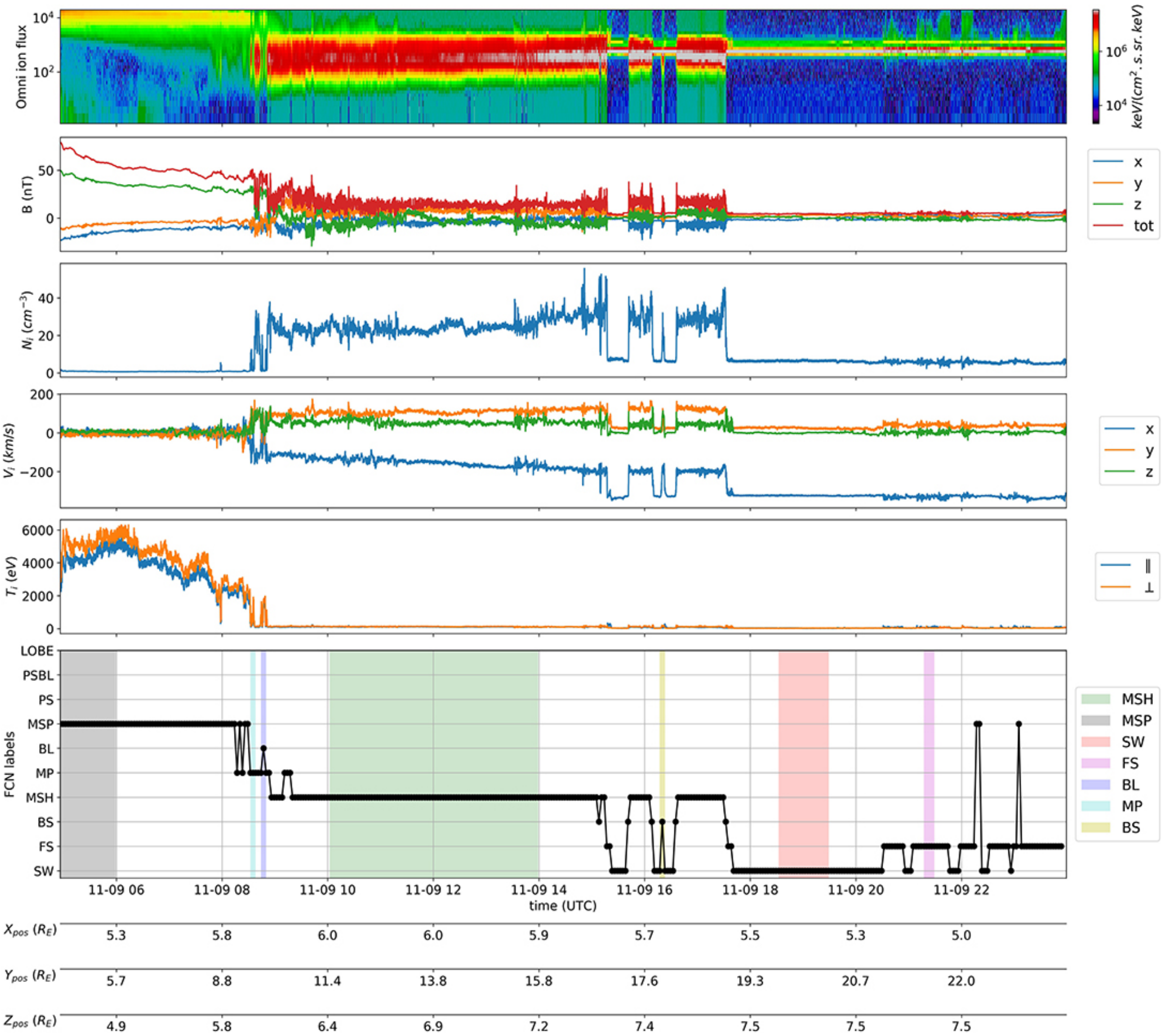


Figure 1.10: Figure montrant un exemple de classification de région de plasma à l'aide du modèle CNN entraîné sur les données MMS. La ligne noire montre la classification du modèle tandis que les sections colorées montrent la classification humaine qui aurait été effectuée sur cette série chronologique. Image tirée de *Breuillard et al.* [2020]

Conclusion

Cette application et la précédente montrent que même si l'utilisation de méthodes d'apprentissage automatique n'est pas encore un standard dans le domaine de la physique des plasmas, il existe un nombre toujours croissant de publications intégrant cette technologie et montrant de bons résultats. Dans le chapitre 3, je présenterai le travail que j'ai effectué sur la détection automatique des régions de diffusion des électrons en utilisant les éléments qui ont été présentés dans ce chapitre.

Chapter 2

Introduction

2.1 The age of space weather

A growing interest of the scientific community towards the interaction between the Sun and the Earth has been observed in the recent years, an interest that includes magnetic reconnection since its discovery in the 80's. This interaction, however occurring far above our heads, can indeed have a very real impact on our lives on Earth. The most famous example of this are the polar lights (also known as ”Auroras”) that can be observed in high-latitudes regions. Although being beautiful events to witness, these products of the interaction between the Sun and the Earth can have more bothering consequences such as on September 2, 1859 when an intense geomagnetic storm caused the brightest aurora in the history that could be seen as far as Boston. In England, this aurora caused the interruption of about 200 000 km of telegraph lines for many hours. It was also the first time where the link between auroras and electricity could unambiguously be established.

The influence of solar activity on terrestrial and spatial infrastructures is then a growing important matter. The necessity to shut down satellite instruments in case of strong geomagnetic storms is one example of precaution that has to be taken against what is now called ”space weather”. This term emerged in the 80's to designate a field becoming more and more relevant nowadays thanks to the many monitoring tools we have at our disposal. Scientific breakthroughs such as the invention of computers and more recently the advent of artificial intelligence allow us to remotely predict these events in order to protect our infrastructures.

These predictions are intrinsically linked to our current understanding of the interaction mechanisms between the Sun and the Earth. More precisely, the understanding of the mechanisms governing the energy exchanges between these two celestial bodies is crucial to keep enhancing the tools and techniques of prediction, magnetic reconnection being one of these mechanisms. This is why the global study of magnetic reconnection is a relevant topic of interest for the scientific community. Since this phenomenon is a vast subject that cannot be treated in its entirety, and especially not during a 3-year-long thesis work, my studies

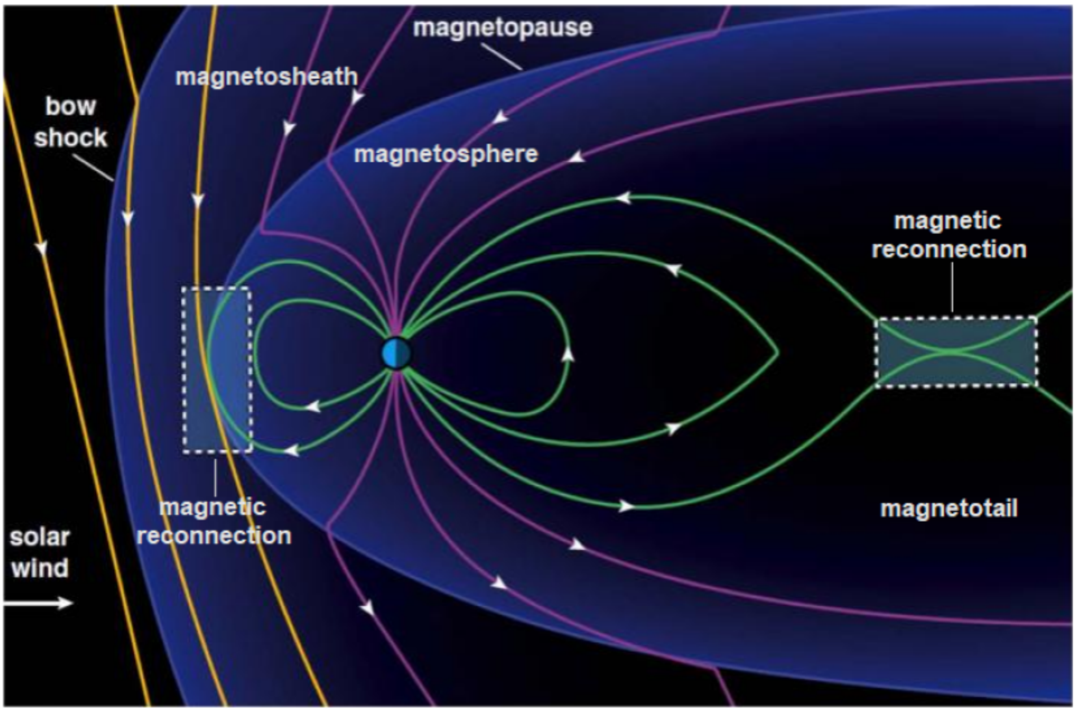


Figure 2.1: Figure showing the plasma environment of the Earth.

are focused on a small region at the heart of the magnetic reconnection process called the Electron Diffusion Region.

2.2 Introduction to magnetic reconnection

This section will be focused on the introduction of the theoretical grounds upon which the concept of electron diffusion region in plasma physics is built. I will first present Earth's plasma environment to understand the context in which the magnetic reconnection presented in this thesis takes place. Then I will explain how the frozen-in concept, and more specifically its breaking, is at the origin of the first magnetic reconnection models. Finally, I will discuss the commonly accepted picture of reconnection and how it changes depending on where the reconnection occurs.

2.2.1 Earth's plasma environment

Before describing magnetic reconnection, I will introduce the Earth's plasma environment and introduce the different plasma regions that interact together. Figure 2.1 shows a view of this interaction with the magnetic field lines coming from the Sun in yellow and that coming from the Earth in green and purple. The green lines are closed field lines and the purple ones are open field lines, directly interconnecting the Earth and the solar wind.

In this picture, the first major actor in the Earth's plasma environment is the Sun. It is constantly releasing charged particles from its corona into the interplanetary space, and this flux of plasma is referred to as ”Solar Wind”. Its speed is observed to be between 300 to 800 km/s and near the Earth's orbit, its average speed is 400 km/s and its average density is 5 cm^{-3}. The particles composing the solar wind are mainly electrons, protons (H^+), alpha particles α (He^{2+}) can also be found but their density is lower than that of the protons. The solar wind carries away from the Sun and into the interplanetary medium a magnetic field known as the ”Interplanetary Magnetic Field” (or IMF).

The second major actor of plasma physics around Earth is Earth's magnetosphere. It is a magnetic cavity created by the confinement of the Earth's magnetic field lines by the pressure of the solar wind, particle densities inside are thus very low as it acts as a shield against the solar wind. This shield is however not perfect at places called ”polar cusps” where the Earth's dipolar magnetic field fans out of the magnetic dipole, these are the privileged entry points for solar wind particles into the magnetosphere because of the lower magnetic field there. This magnetosphere acts as an obstacle for the supersonic solar wind which gets slowed down and deflected around it, creating a boundary called the ”bow shock” where plasma is slowed down and condensed, leading to an increase in plasma density, temperature and magnetic field strength.

The downstream part of the shock is called the magnetosheath, it is a dynamic region of turbulent plasma that serves as a transitory state between the solar wind region and the magnetosphere. Typical values for the particle density and the magnetic field intensity in this region are of about 20 cm^{-3} and 20 nT whereas typical values for the particle density and the magnetic field intensity in the magnetosphere are of about 0.1 cm^{-3} and 50 nT. The magnetosphere region is thus very easily recognizable and distinguishable from the magnetosheath region in the data.

The discontinuity separating the geomagnetic field from the shocked solar wind is a current layer called the magnetopause. It is the surface at which the pressure of the shocked solar wind and that of the magnetosphere balance each other. Its position is then not constant and can move inwards or outwards in response to a increase or a decrease in the solar wind's pressure. The magnetopause is the region around which the work presented in this thesis is focused on so it is of particular interest to us.

The last plasma region of interest left to introduce is the ”magnetotail”, which is an extension of the magnetosphere on the night side that is several hundreds of Earth's radii long. This extension far beyond the Earth is the night side response of the magnetosphere to the day side pressure of the solar wind. It is composed of two lobes almost empty of particles separated by a plasma sheet containing most of the tail's particles, and the shear between the oppositely directed open field lines coming from the two hemispheres creates a neutral current sheet lying at the middle of the plasma sheet.

It is in this context that magnetic reconnection takes place, the frontside magnetopause and the nightside magnetotail being the two closest sites to Earth where magnetic reconnection can take place due to the presence of oppositely directed magnetic field lines. Reconnection is always possible on the dayside because if the interplanetary magnetic field (IMF) carried by the solar wind has a purely southward orientation, reconnection can take place close to the "subsolar point", which is the outermost point of the magnetopause where the solar wind first encounters the magnetosphere. If the IMF has a purely northward orientation, then reconnection is still possible near the polar cusps above the poles because there, the Earth's magnetic field is in the opposite direction than that of the subsolar point field orientation. On the nightside, the open field lines' configuration makes it so field lines coming from the north hemisphere are in the opposite direction of the field lines coming from the south hemisphere which also allows reconnection.

Magnetic reconnection can also be found in many other astrophysical contexts where strong magnetic fields are involved and is at the origin of some of the most powerful astrophysical events in the solar system such as solar flares. Magnetic reconnection is indeed a major energy transfer process that can occur when oppositely directed magnetic field lines interact with each other. This interaction translates into a change of connectivity where the field lines' topology is rearranged, and a conversion of magnetic energy into kinetic energy, thermal energy, and particle acceleration.

2.2.2 The frozen-in flux concept

The physical framework mostly used to describe magnetic reconnection is called Magnetohydrodynamics (MHD), which is the combination of fluid mechanics and of Maxwell's equations. It describes the behavior of electrically conducting fluids in presence of magnetic fields. Plasma, the fourth state of matter which we will discuss throughout this thesis, can be considered as a perfect conductor with an infinite electrical conductivity (equivalent to having no electrical resistivity), meaning that a current flowing through a plasma should not lose power through the generation of heat. When considering ideal MHD (i.e. neglecting specific contributions to the electric field), one of the fundamental concepts is called the frozen-in flux theorem. Before addressing magnetic reconnection, I will show here the demonstration of this theorem as the whole concept of magnetic reconnection is based on the breaking of this theorem. Magnetic reconnection is indeed normally not possible according to ideal MHD.

The first step of the demonstration is to combine the Maxwell-Faraday, Maxwell-Ampere and the Ohm's laws which are respectfully written as follows :

$$\boldsymbol{\nabla} \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \tag{2.1}$$

$$\boldsymbol{\nabla} \times \mathbf{B} = \mu_0 \mathbf{J} \tag{2.2}$$

$$\mathbf{E} + \mathbf{v} \times \mathbf{B} = \mathbf{J}/\sigma \tag{2.3}$$

with $\mathbf{E}$ the electric field, $\mathbf{B}$ the magnetic field, $\mathbf{J}$ the electric current, $\mathbf{v}$ the velocity field and σ the electric conductivity of the fluid.

The combination of these equations yields the induction equation of MHD :

$$\frac{\partial \mathbf{B}}{\partial t} = \eta \nabla^2 B + \boldsymbol{\nabla} \times (\mathbf{v} \times \mathbf{B}) \tag{2.4}$$

where $\eta = 1/(\mu_0 \sigma)$ is the magnetic diffusivity. If we assume a typical speed V for the fluid and a typical length scale L, then we can rewrite $\eta \nabla^2 B$ as $\frac{\eta B}{L^2}$ and $\nabla \times (v \times B)$ as $\frac{VB}{L}$. The ratio of these two expressions is a dimensionless parameter called the magnetic Reynolds number : $R_m = \frac{LV}{\eta}$. It gives an estimate of the relative effect of induction (LV) to magnetic diffusion (η).

In a plasma at large magnetic Reynolds number (ideal plasma with an infinite electric conductivity), the magnetic diffusion term η tends towards 0, the induction equation then reduces to :

$$\frac{\partial \mathbf{B}}{\partial t} = \boldsymbol{\nabla} \times (\mathbf{v} \times \mathbf{B}) \tag{2.5}$$

If one now considers a curve C which enlaces a surface S moving with the plasma, during a time dt, one element dl of C enlaces a quantity of area $\mathbf{v} \times \mathbf{dl}\ dt$. The rate of variation of the magnetic flux through C is :

$$\frac{d}{dt} \iint_S \mathbf{B} \cdot \mathbf{dS} = \iint_S \frac{\partial \mathbf{B}}{\partial t} \cdot \mathbf{dS} + \oint_C \mathbf{B} \cdot \mathbf{v} \times \mathbf{dl} \tag{2.6}$$

When the curve C moves, the flux changes because of the change of magnetic field and the boundaries are moving in space. Using the vectorial identity $\mathbf{B} \cdot \mathbf{v} \times dl = -\mathbf{v} \times B \cdot dl$ and the Stokes theorem :

$$\oint_C \mathbf{A} \cdot \mathbf{dl} = \iint_S \boldsymbol{\nabla} \times \mathbf{A} \cdot \mathbf{dS} \tag{2.7}$$

we can replace the second right-hand side term of equation 2.6 to obtain the following equation :

$$\frac{d}{dt} \iint_S \mathbf{B} \cdot \mathbf{dS} = \iint_S \left(\frac{\partial \mathbf{B}}{\partial t} - \boldsymbol{\nabla} \times (\mathbf{v} \times \mathbf{B}) \right) \cdot \mathbf{dS} \tag{2.8}$$

We see that the right-hand side of equation 2.8 cancels out using equation 2.5. Therefore, the total magnetic flux through C remains constant as C moves with the plasma, or said differently, plasma parcels forming a flux tube remain the same over time. This is why in ideal MHD, we say that the magnetic field in frozen in the plasma. This frozen-in flux theorem is also known as Alfvén's theorem, and it holds even when the plasma is not perfectly conducting as long as the resistive and the electrical conductivity are large in spatial scale which still gives a large magnetic Reynolds number. The theorem however breaks when looking at small scales, which is one of the required conditions for magnetic reconnection to take place.

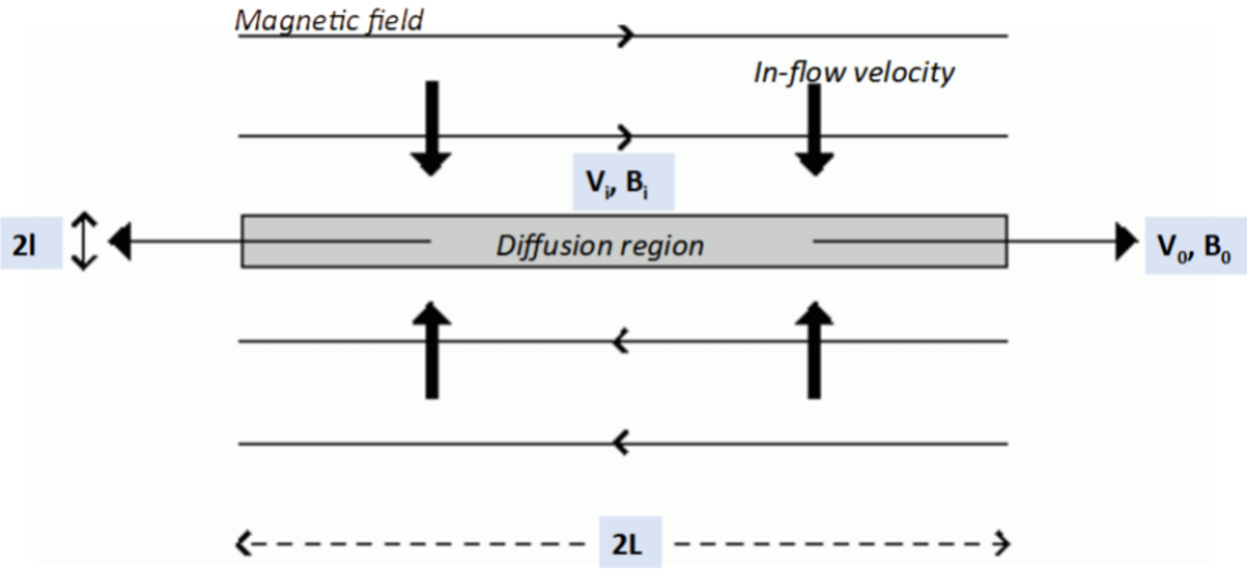


Figure 2.2: Figure showing the geometry of the Sweet-Parker model of reconnection. X is along the current sheet (horizontal direction), Y is perpendicular to the current sheet (vertical direction) and Z is normal to the XY plane in which the flow and the magnetic field lie. Image taken from *Priest and Forbes* [2007].

2.2.3 Origin of magnetic reconnection and early models

The term "magnetic reconnection" was first used by James Dungey in 1953 in his PhD thesis studying particle acceleration in the Earth's magnetosphere (*Dungey* [1953]). The task given to him by his supervisor was to explain what happens at a magnetic neutral point and what could be the implications for auroras. In the model he developed, the frozen-in condition of ideal MHD breaks down to allow energy dissipation and the presence of an electric field. The movement of particles creates a thin current sheet in which the diffusion of the magnetic field would dominate. Field lines passing through this current sheet would break and cross-connect with one another, thus the terms "disconnection" and "reconnection". Finally, the reconnected plasma is ejected perpendicularly to the incoming new plasma fluxes.

Three years after the introduction of the concept of magnetic reconnection by Dungey in 1953, two American physicists, Peter Sweet and Eugene Parker developed the first theoretical model of reconnection. Their model describes two-dimensional, steady-state reconnection in an incompressible plasma based on the resistive MHD framework. Figure 2.2 shows the geometry of their model, it considers a long and thin current sheet of length $2L$ and width $2l$. To understand why this first model was incomplete, we have to look at the reconnection rate. The following demonstration, which follows the demonstration done in *Priest and Forbes* [2007], will show how to find it with this theory as well as what it implies.

We start with the mass conservation theorem which tells us that the rate at which the mass enters the layer from both sides of the current sheet of figure 2.2 is equal to the rate at which it leaves it, we can thus write :

$$2 \cdot \rho \cdot 2L \cdot v_{in} = 2 \cdot \rho \cdot 2l \cdot v_{out} \Longleftrightarrow v_{in} L = v_{out} l \tag{2.9}$$

We can get another expression for v_{in} using the Ohm's law for a stationary 2D state which is written :

$$\mathbf{E} + \mathbf{v} \times \mathbf{B} = \mathbf{j}/\sigma \tag{2.10}$$

where $\mathbf{E} = E_z$ is a constant uniform field normal to the XY plane. At the entry of the inflow region, the current vanishes so we have $||\mathbf{E}|| = v_{in} B_{in}$, and at the center of the diffusion region, it is the magnetic field that vanishes, yielding $||\mathbf{E}|| = j/\sigma$. Using the Maxwell-Ampère relation ($\mathbf{j} = \nabla \times \mathbf{B}/\mu_0$) we get $||\mathbf{j}|| = B_{in}/(\mu_0 l)$. We can then deduce from these relations that $v_{in} = 1/(\mu_0 \sigma l) = \eta/l$ with $\eta = 1/(\mu_0 \sigma)$.

Going back to equation 2.9, we can now replace the half thickness l using the new relation for v_{in} to obtain $v_{in}^2 = \eta v_{out}/L$. We define the in-flow Alfvén Mach number as $M_{in} = v_{in}/v_{A,in}$ which is also the dimensionless reconnection rate. If we use the definition of the magnetic Reynolds number based on the in-flow Alfvén velocity $R_{m_{in}} = L v_{A,in}/\eta$, we can express the velocity v_{in} in dimensionless form : $M_{in}^2 = (1/R_{m,in}) \cdot (v_{out}/v_{A,in})$.

We now need to express v_{out} and to do so, we will start from the equation of motion in a stationary state :

$$\rho(\mathbf{v} \cdot \nabla)\mathbf{v} = -\nabla p + \mathbf{j} \times \mathbf{B} \tag{2.11}$$

If we neglect the pressure effects, the x component of this equation writes :

$$(\mathbf{j} \times \mathbf{B})_x = j B_{out} = B_{in} B_{out}/(\mu_0 l) \tag{2.12}$$

using the definition of $\mathbf{j}$ written above, which is also the Lorentz force along the current sheet. This force accelerates the plasma from rest at the neutral sheet and reaches v_{out} after a distance L. This term equals the inertia term :

$$\rho(\mathbf{v} \cdot \nabla) v_x = \frac{\rho v_{out}^2}{L} \tag{2.13}$$

So going back to equation 2.11 but looking at the x component only, we can write :

$$\frac{\rho v_{out}^2}{L} = \frac{B_{in} B_{out}}{\mu_0 l} \tag{2.14}$$

and using equation 2.9 as well as the magnetic flux conservation which writes $v_{in} B_{in} = v_{out} B_{out}$, we get the following expression for v_{out} :

$$v_{out} = \frac{B_{in}}{\sqrt{\mu_0 \rho}} = v_{A,in} \tag{2.15}$$

The reconnection rate we introduced earlier then writes :

$$M_{in} = \frac{v_{in}}{v_{A,in}} = \frac{1}{\sqrt{R_{m,in}}} \tag{2.16}$$

We can now estimate the characteristic time of reconnection from a characteristic length L and the "reconnection velocity" v_{in} : $\tau = L/v_{in}$, and using equation 2.16 we get $\tau = \sqrt{R_{m,in}} L/v_{A,in}$. Typical values in a solar flare are $R_{m,in} = 10^8$, $v_{A,in} = 100\ km/s$ and $L = 10^4\ km$ which yield a reconnection time of 11 days. Typical values in for the solar wind are $R_{m,in} = 10^{11}$, $v_{A,in} = 100\ km/s$ and $L = 10^3\ km$, giving a reconnection time of 37 days. The problem is that these values are way too big as the observed reconnection time for the solar flares and the solar wind is between the minute and the hour.

The Sweet-Parker model is a "slow reconnection" model as it was not able to describe the reconnection that is observed, and this is why in 1964, the physicist Harry Petschek developed a "fast reconnection" model as an improvement of the Sweet-Parker model. One of the changes needed to obtain faster reconnection rates was the shrinking of the size of the diffusion region L which leads to the decrease of the thickness of the current sheet and the increase of the speed of the reconnection process. Compared to the Sweet-Parker model, the width of the diffusion region is equal to its height as can be seen in figure 2.3. Petschek's model yields a reconnection rate in accord with the observations but remains highly controversial. The problem comes from the fact that the model requires the two collisionless, slow-mode shock waves that can be sustained for the entirety of the time needed for the steady state reconnection process to take place.

2.2.4 Topology of magnetic reconnection

Until now I discussed the general case of reconnection, but now I will discuss more specifically reconnection at the Earth. The frontside magnetopause and the nightside magnetotail are the two closest sites to Earth where magnetic reconnection can take place but it can also occur in many other astrophysical contexts where strong magnetic fields are involved and where the topology of the magnetic field lines allow it.

The commonly accepted picture of magnetic reconnection (see figure 2.4) is the following : first we have two different plasma regions with oppositely directed magnetic field lines located at the top and at the bottom, separated by a current sheet. The reconnected plasma is ejected on the sides into the "exhaust region" and the delimitations between the exhaust and the first two plasma regions are called the "separatrices". These separatrices cross at the middle of the current sheet forming an X shape, with the center point called the X-point. The X-point is the physical place where there is no magnetic field as it is the precise location where the field lines break and reconnect, and the X-line is the extension of the X-point that cuts the exhaust in two where the magnetic field lines change direction.

At the scale length below the ion inertial length, the region is called the Ion Diffusion Region (or IDR). It is where the ions are demagnetized from the magnetic field lines but the electrons are still frozen in the plasma. Embedded inside the IDR, at the scale length below the electron inertial length, there is the Electron Diffusion Region (or EDR), which is where the energy exchange occur. One of the points of interest inside the EDR is the stagnation

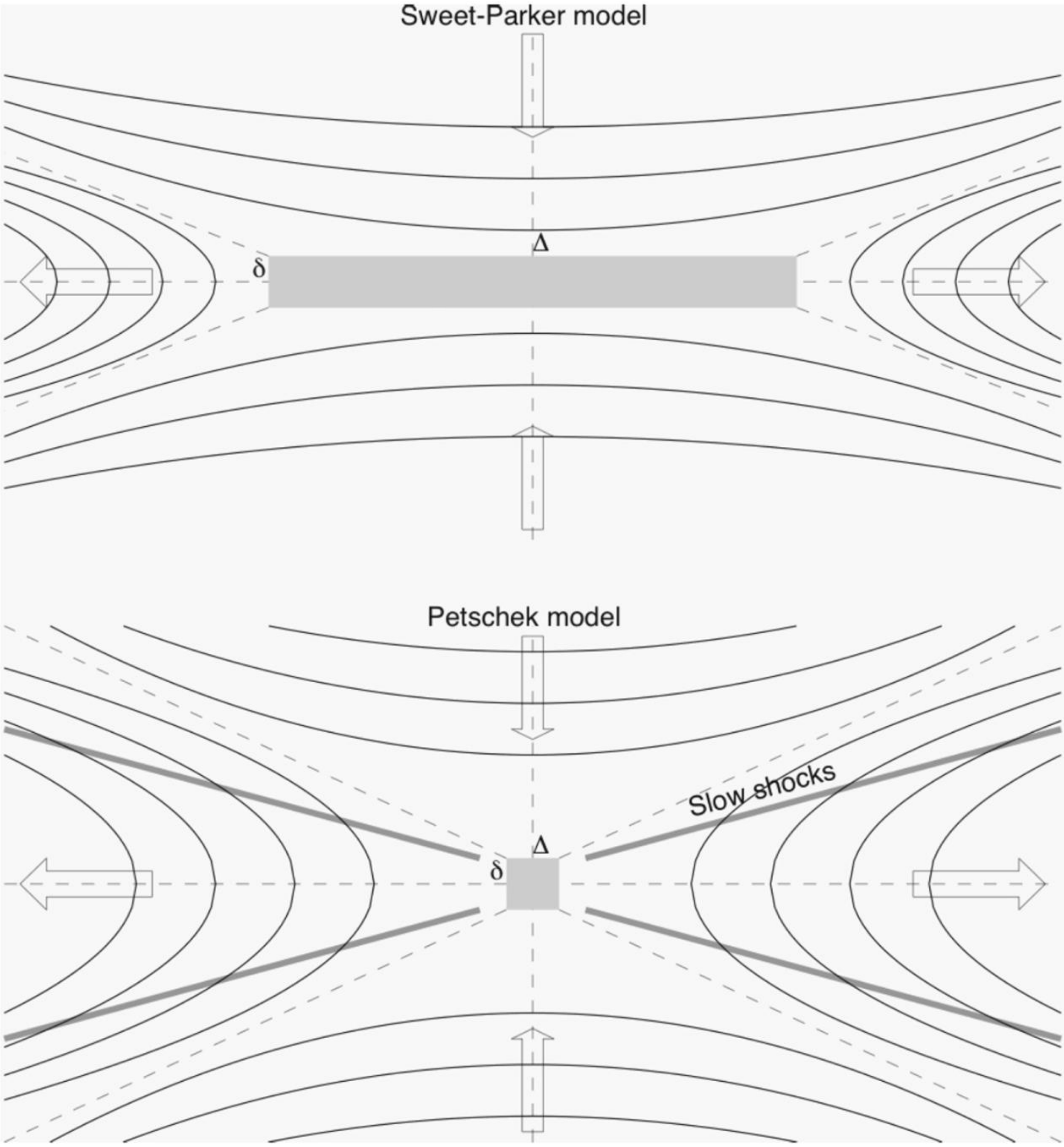


Figure 2.3: Figure comparing the geometry of the Sweet-Parker model and that of the Petschek model. Image taken from *Aschwanden* [2020]).

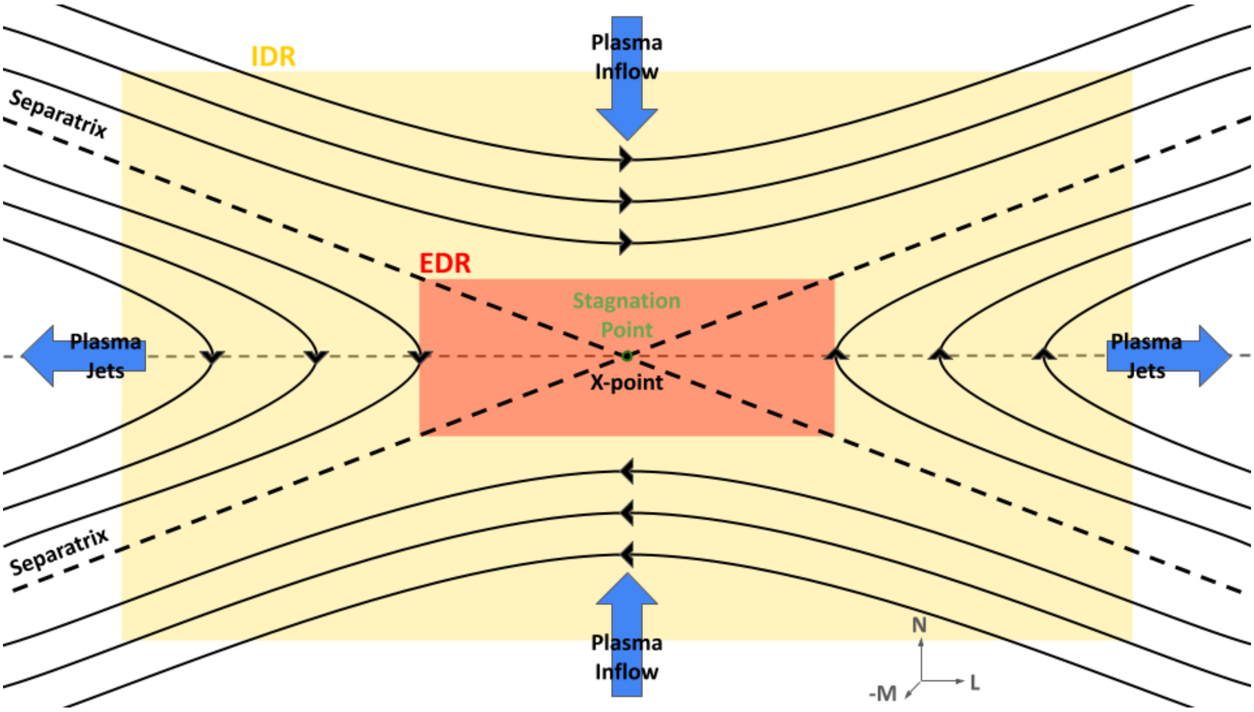


Figure 2.4: Figure showing the topology of symmetric magnetic reconnection. L,M and N are the current sheet coordinates.

point, it is the place where the velocity of the electrons in the reconnection plane is null. The location of this stagnation point is determined by the asymmetry of the system.

Reconnection at the Earth's magnetopause is referred to as "asymmetric reconnection" whereas reconnection at the magnetotail is called "symmetric reconnection". This is due to the fact that at the magnetotail, the plasma on both sides of the current sheet is the same, so the reconnection picture I was mentioning is symmetric. However at the magnetopause, the plasma coming from the magnetosheath is very different from the plasma in the magnetosphere, the magnetospheric field lines being compressed by the solar wind, the magnetic field intensity is much stronger there and the particle density lower. The reconnection picture is then asymmetric with one half of the exhaust being smaller than the other. This also has the effect of shifting the stagnation point, normally located at the X-point in symmetric reconnection, towards the magnetosphere. Figures 2.4 and 2.5, which summarize the explanations above, show the difference between the symmetric and asymmetric reconnection.

The Electron Diffusion Region is the main subject of this thesis, a more detailed description of the physics happening inside can be found in section 2.4. The fact that its size is much smaller than that of the Ion Diffusion Region is one of the reason for the difficulty of its detection in the data. Due to the speed of the spacecraft, it is only possible to stay less than a second inside this region. To overcome this problem, a very high resolution was needed and was indeed provided by the Magnetospheric Multiscale mission.

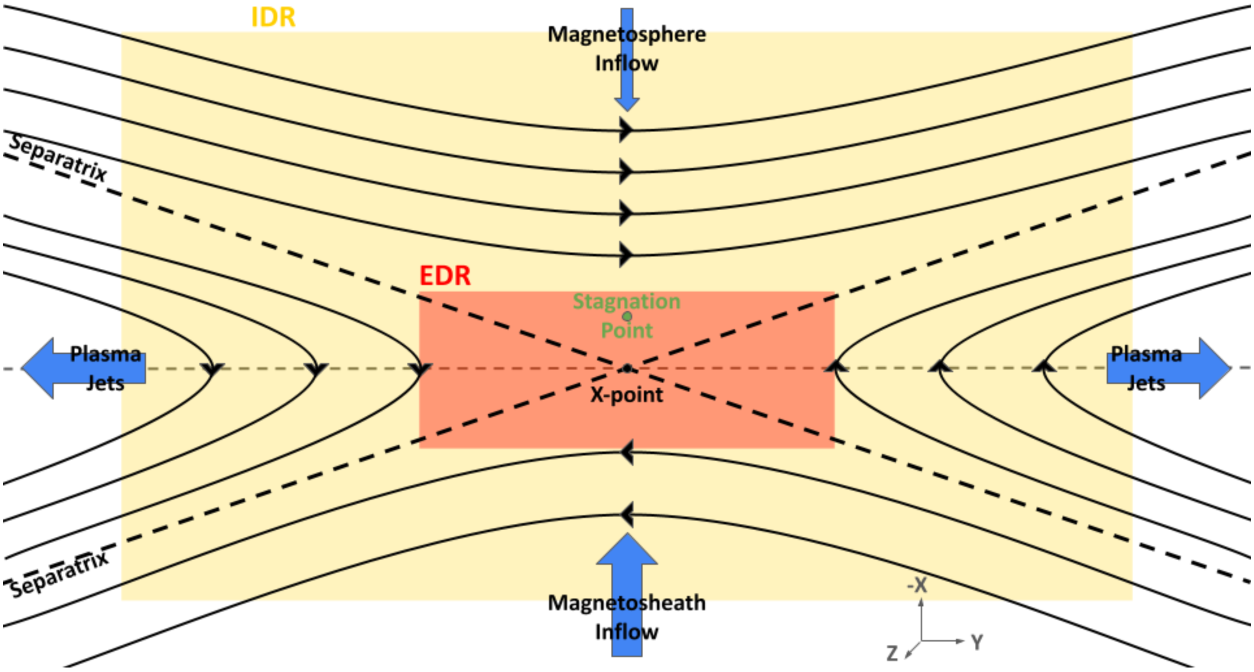


Figure 2.5: Figure showing the topology of asymmetric magnetic reconnection. X,Y and Z are the GSE coordinates.

2.3 The Magnetospheric Multiscale (MMS) Mission

2.3.1 Description of the mission

The Magnetospheric Multiscale Mission (or MMS) is a mission launched on March 13th 2015, and is still active at the time this manuscript is written (2022). The goal of the mission is to study the interaction between the Sun's solar wind and the Earth's magnetosphere and more precisely the phenomenon of magnetic reconnection. It is composed of four identical spacecraft that are equipped with instruments able to measure plasmas, fields and particles. The nominal mission plan was of 2 years during which the spacecraft would travel in two different highly elliptical Earth orbits. The first 3/4 of the mission are dedicated to the study of the day-side magnetopause, and the last 6 months are focused on the study of reconnection at the magnetotail on the night-side.

MMS was designed to succeed to the CLUSTER-II mission launched by the ESA in 2000 which helped study several aspects of the magnetic reconnection process (e.g. *Hamrin et al.* [2011] studied the energy conversion regions with Cluster, *Retinò et al.* [2006] investigated the structure of the separatrix region...). More precisely, one of the key objectives of the new MMS mission was to deepen the current understanding of the Electron Diffusion Region (EDR), which was not previously possible with the one minute resolution provided by CLUSTER. It is possible to better understand that by doing this simple computation : the elliptical orbit of CLUSTER has a perigee of 19 000 km and an apogee of 119 000 km, which represents a distance of roughly 369 186 km that CLUSTER travels in 57 hours. To simplify things, if we assume the velocity is constant throughout the orbit, the computation yields a spacecraft velocity of 1.8 km/s. With a resolution of one minute, CLUSTER could only do a measurement approximately every 108 km. It is far larger than the size of the Electron Diffusion Region which is only of a few kilometers. With an electronic resolution of 30 ms, for the same spacecraft velocity, MMS could make a measurement every 54 m, drastically improving the chance of providing samples taken inside the EDR.

2.3.2 Instruments

The instruments on board can be divided into three groups :

- Hot Plasma Suite : to study the charged gas or plasma present during reconnection, it includes the Fast Plasma Investigation and the Hot Plasma Composition Analyzer
- Energetic Particle Detector Suite : to study the fast moving, energetic particles, it includes the Fly's Eye Energetic Particle Sensor and the Energetic Ion Spectrometer
- Fields Suite : to study the electrical and magnetic fields and waves, it includes the Analog Fluxgate and Digital Fluxgate magnetometers, the Electron Drift Instrument, the Spin-plane Double Probe, the Axial Double Probe, and the Search Coil Magnetometer

The magnetic field **B** is measured with the Fluxgate Magnetometers (FGM) (*Russell et al.* [2016]) constituted of the Analog Fluxgate Magnetometer (AFG) and the Digital Fluxgate Magnetometer (DFG). It can measure the magnetic field in 3D with either a frequency of 16 Hz or 32 Hz by analysing the change of properties of a permeable material it carries in response to the presence of magnetic fields.

The electric field **E** is measured with the Electric field Double Probe (EDP) constituted of the The Spin-plane Double Probe (SDP) (*Lindqvist et al.* [2016]) and the Axial Double Probe (ADP) (*Ergun et al.* [2016]). They measure the voltage between two electrodes to determine the electric field, each residing at the ends of 60 meter-long booms for SDP and 9 meter-long booms for EDP that deploy away from the main body of each spacecraft after they are launched.

Another way of measuring the electric and magnetic fields is provided by the Electron Drift Instrument (EDI). It sends electron beams into space using two Gun Detector Units. By using the property of electrons to travel in circles around magnetic field lines, the instrument is able to measure the strength of the magnetic field by measuring the time taken by the sent electron beams to circle back to MMS. Moreover, by using the fact that electrons will not exhibit a perfect circular motion in the presence of an electric field due to what is called the "electron drift", the electric field can be measured by observing the sideways drift of the same returning electron beams.

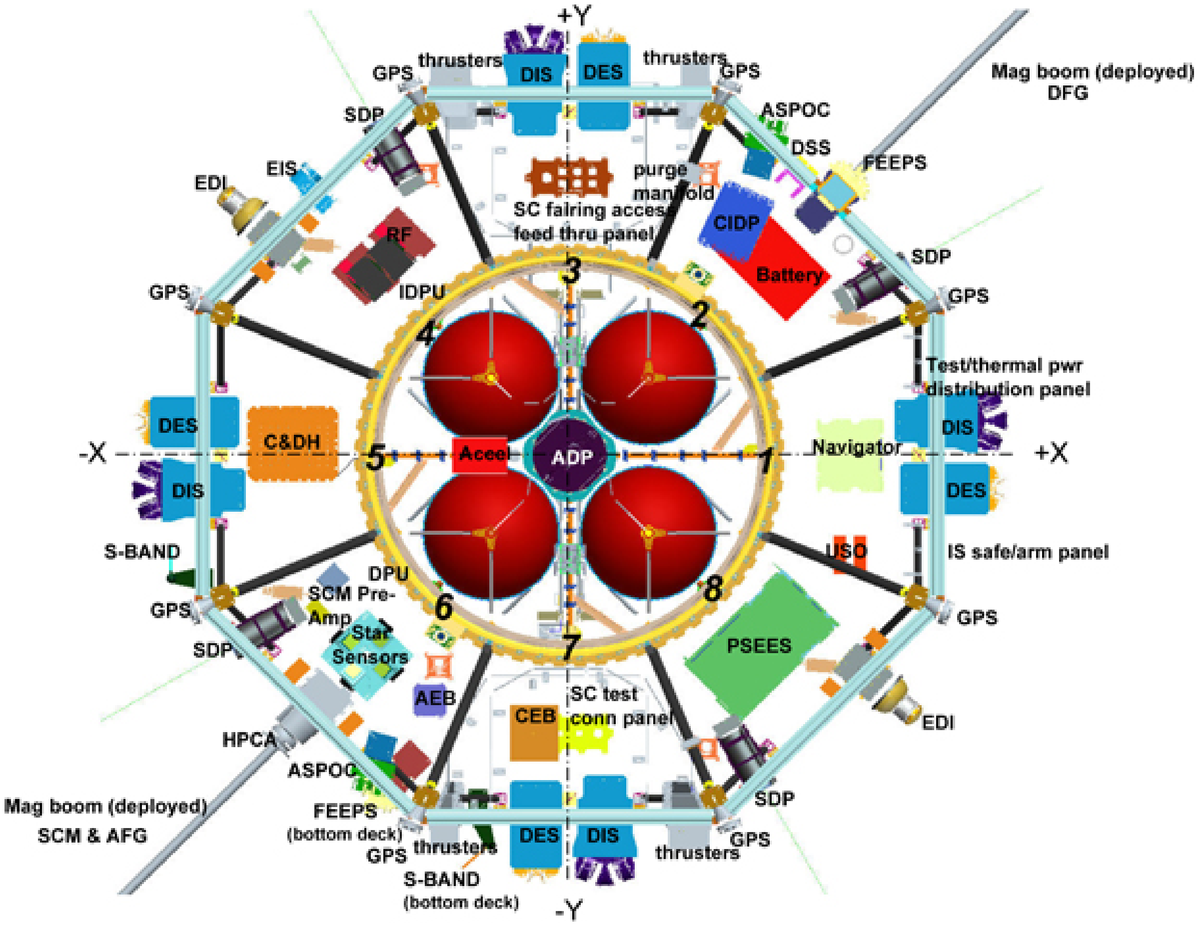


Figure 2.6: Figure showing the different instruments mounted on the MMS spacecraft. Image taken from mms.gsfc.nasa.gov

To measure the different particle quantities, the instrument used is the Fast Plasma Investigation (FPI) (*Pollock et al.* [2016]) instrument. Measurements of the particle fluxes are done every 150 ms for the ions and every 30 ms for the electrons. These measurements are then used to get full 3D distribution functions for the ions and for the electrons from which different moments can be computed such as the density n, the velocity u and the pressure tensor $\mathbb{P}$ with the following formulas :

$$n(\mathbf{r},t) = \int d^3\mathbf{v} f(\mathbf{r},\mathbf{v},t) \tag{2.17}$$

$$\mathbf{u}(\mathbf{r},t) = \frac{1}{n(\mathbf{r},t)} \int d^3\mathbf{v}\mathbf{v} f(\mathbf{r},\mathbf{v},t) \tag{2.18}$$

$$\mathbb{P}(\mathbf{r},t) = \int d^3\mathbf{v} m[\mathbf{v}-\mathbf{u}(\mathbf{r},t)][\mathbf{v}-\mathbf{u}(\mathbf{r},t)] f(\mathbf{r},\mathbf{v},t) \tag{2.19}$$

This allows for the computation of more complex parameters like the plasma current density J with the formula $\mathbf{J} = e \cdot (n_i\mathbf{v_i} - n_e\mathbf{v_e})$. All of these instruments make the MMS mission the best choice to study Electron Diffusion Regions. The next section will finally introduce in greater details this plasma region as well as its specificities.

2.4 The physics of the Electron Diffusion Region (EDR)

One of the goals of my thesis was to find new electron diffusion region candidates but one could rightfully ask why is there a need to look for new candidates ? There are multiple answers to that question. First, I previously said that the MMS mission is the only one allowing the study of the EDR and there probably will not be another mission capable of doing so in the next decades, so now is the time for this kind of study. The second answer is that the EDR is very small, its dimensions are of the order of a few kilometers so due to the speed of the MMS spacecraft, only a few data points can be recorded inside that region. Identifying EDRs in the MMS data is thus not as easy as it may seem. Since the launch of MMS in 2015, only about 35 dayside EDR events have been reported to my knowledge, most of them are compiled in *Webster et al.* [2018]. Having so few events makes statistical studies on EDRs harder, finding new candidates is then crucial to allow further studies on that region. More generally, the internal physics of the EDR is a very complicated matter, so as many cases as possible are required to explore the vast diversity of conditions that can be found within this region.

In the next subsections, I will discuss the different key signatures of the EDR as well as some other physics aspects related to the EDR.

2.4.1 Conversion of energy

During magnetic reconnection, the topology of the magnetic field lines of the plasma is changed, coupled with the exchange of energy from the magnetic fields to the particles. The topological change takes place at the heart of the electron diffusion region, at the X point. However, the energy exchange does not occur at the X point but rather in what is called the "dissipation region", located on the magnetospheric side away from the X-point. The energy dissipation process is governed by the generalized Ohm's law, derived from the equation of motion of the electron fluid, which is the following :

$$\mathbf{E} + \mathbf{v}_e \times \mathbf{B} = -\frac{1}{ne}\nabla \cdot \mathbf{P}_e - \frac{m_e}{e}\left(\frac{\partial \mathbf{v}_e}{\partial t} + \mathbf{v}_e \cdot \nabla \mathbf{v}_e\right) \tag{2.20}$$

where $\mathbf{E}$ is the electric field, $\mathbf{v}_e$ is the electron velocity, $\mathbf{B}$ is the magnetic field, $\mathbf{P}_e$ is the electron pressure tensor.

In the electron diffusion region, electrons are demagnetized meaning that they do not obey the ideal MHD condition $\mathbf{E'} = \mathbf{E} + \mathbf{v}_e \times \mathbf{B} = 0$ where $\mathbf{E'}$ is the electric field in the plasma rest frame. This is explained by the fact that at large scales over large time periods, the gradients and the partial time derivative all tend towards zero but this is not the case in the EDR as we look at very small scales over short periods of time. Verifying if $\mathbf{E'} \neq 0$ is thus one of the features allowing to identify the electron diffusion region in the data.

Another way to identify the electron diffusion region, or rather the dissipation region within, is to look at the Joule dissipation, which is the per-volume rate of work done by the electric field on the plasma expressed $\mathbf{J} \cdot \mathbf{E}$, with $\mathbf{J}$ being the current density. It is also usually expressed in the electron rest frame to see the work done by the non ideal electric fields, it is then written as $\mathbf{J} \cdot \mathbf{E}$**'**. The Joule dissipation is directly linked to the variation of electromagnetic energy with the following equation :

$$\partial_t \varepsilon^m + \frac{c}{4\pi} \nabla \cdot (\mathbf{E} \times \mathbf{B}) = -\mathbf{J} \cdot \mathbf{E} \tag{2.21}$$

obtained using a combination of the Vlasov equation with the Maxwell's equations. We see that the Joule dissipation can either be positive ($\mathbf{J} \cdot \mathbf{E} > 0$) or negative ($\mathbf{J} \cdot \mathbf{E} < 0$), which will directly affect the sign of the variation of the electromagnetic energy $\partial_t \varepsilon^m = \frac{1}{8\pi}(\mathbf{B}^2 + \mathbf{E}^2)$. If $\mathbf{J}.\mathbf{E}$ is positive, the variation of the electromagnetic energy is negative meaning that the energy transfer is from the fields to the particles and if $\mathbf{J} \cdot \mathbf{E}$ is negative, the energy transfer is from the particles to the fields, which are respectively referred to as the load and dynamo effects.

Inside the EDR, the load effect ($\mathbf{J} \cdot \mathbf{E} > 0$) is expected with energy transfer from the fields to the particles. However, multiple studies have shown that there can also be either negative or fluctuating $\mathbf{J} \cdot \mathbf{E}$ values at the EDR (see section 3.5).

2.4.2 Electron Crescents

One of the greatest aspect of the MMS mission is to be able to visualize the electron velocity distribution functions for the first time every 30 ms. This fast rate allows for the study of a few distribution functions taken inside the EDR which was not previously possible due to the lower resolutions of the previous missions.

An electron velocity distribution function (eVDF) is a 3-dimensional object representing the distribution of the electron velocities in the velocity space. It can either be shown in the GSE coordinate system (Vx, Vy, Vz) or in the field-aligned coordinate system ($V_{||}$, $V_{\perp_1}$, $V_{\perp_2}$). The full eVDF being a 3D object, its study under this form is not practical, this is why the reduced distribution function tends to be used instead. The reduced distribution function is the result of the integration of the full distribution function over a certain angle around one plane of the space. This name "electron velocity distribution function" is often inaccurately used to designate the "reduced electron velocity distribution function" and this is why all the "reduced distribution functions" presented throughout this thesis will simply be called "distribution functions". The distribution functions are expected to be gyrotropic, meaning that they are circularly symmetric.

Before the launch of the MMS mission, only a few studies using simulations were conducted on the electron distribution functions during asymmetric reconnection. *Egedal et al.* [2011] showed significant anisotropy (also called agyrotropy) in the electron distribution functions in the inflow region in the magnetosphere, believed to be important for generating elongated current layers in the exhaust. 3 years later, *Hesse et al.* [2014] was the first to show using Particle-In-Cell simulations that the electron distribution at the electron stagnation point consisted of two different electron populations, a gyrotropic distribution at lower velocities and a highly non-gyrotropic crescent-shaped distribution on one side only at higher velocities. The distribution function obtained from the simulation can be seen in figure 2.7.

The presence of this agyrotropy is due to energetic electrons from the magnetosheath that are able to mix with electron populations from the magnetosphere by crossing the current sheet. The energetic electrons then get reflected on their way back to the magnetosheath, trapping them on the magnetospheric side of the current sheet. To go a little further, this peculiar shape in the velocity space can be explained as in *Bessho et al.* [2016], by following the trajectory of a single electron coming from the magnetosheath ($z < 0$) and entering the EDR. In this simulation, x_{simu}, y_{simu} and z_{simu} respectively correspond to Y_{GSE},$-Z_{GSE}$ and $-X_{GSE}$ from figure 2.5. Let us consider a simplified case where the magnetic field is written as $(B_x = bz, B_y = 0, B_z = 0)$ with b a positive constant and $z = z_X = 0$ the magnetic neutral point (the X-line), and the electric field is written $(E_x = 0, E_y = 0, E_z = -kzU(z))$ with $k(> 0)$ the magnitude of the slope of E_z in $z > 0$ and U(z) the step function. The nonrelativistic equation of motion in the (y, z) plane, assuming $v_x = 0$, for an electron interacting with such fields is the following :

$$m_e \frac{dv_y}{dt} = -\frac{e}{c} v_z bz, \tag{2.22}$$

$$m_e \frac{dv_z}{dt} = ekzU(z) + \frac{e}{c} v_y bz, \tag{2.23}$$

where m_e is the electron mass. In the magnetosheath ($z \leq 0$), the electric field term vanishes so the electron energy is conserved. In the magnetosphere ($z > 0$), E_z and B_x are proportional to z, and if we study the electron motion in the frame moving in the y direction with a boost velocity $cE_z/B_x = -ck/b$, E_z vanishes as long as $k < b$ and the electron energy is also conserved. Taking the assumption that $cE_z/B_x = -ck/b = cst$ is based on simulations. If we apply the energy conservation formula, we get the two following cases :

$$\frac{1}{2} m_e v^2 = cst \iff (v_y^2 + v_z^2) = cst \ when \ z \leq 0 \tag{2.24}$$

$$\frac{1}{2} m_e v^2 = cst \iff (v_y - v_{y0})^2 + v_z^2 = cst \iff (v_y + ck/b)^2 + v_z^2 = cst \ when \ z > 0 \tag{2.25}$$

To find the value of v_{y0} in the case when $z > 0$, one has to see that when $\frac{dv_z}{dt} = 0$, the value of v_y is $-ck/b$. The two circles visible in figure 2.8a represent the only two trajectories that the electron can take in the velocity space depending on the value of z, forming a crescent shape in the (v_y, v_z) plane. In figure 2.8b, we can see the actual trajectory of the electron in the physical space, with the magnetic field being along the x direction, the electron gyrates around it in one direction or the other and changes direction as it passes through the magnetopause. It is important to note that asymmetries in distribution functions can also be created by other filed configurations such as near separatrices, thus the presence of an asymmetry in the distribution function is not enough on its own to identify

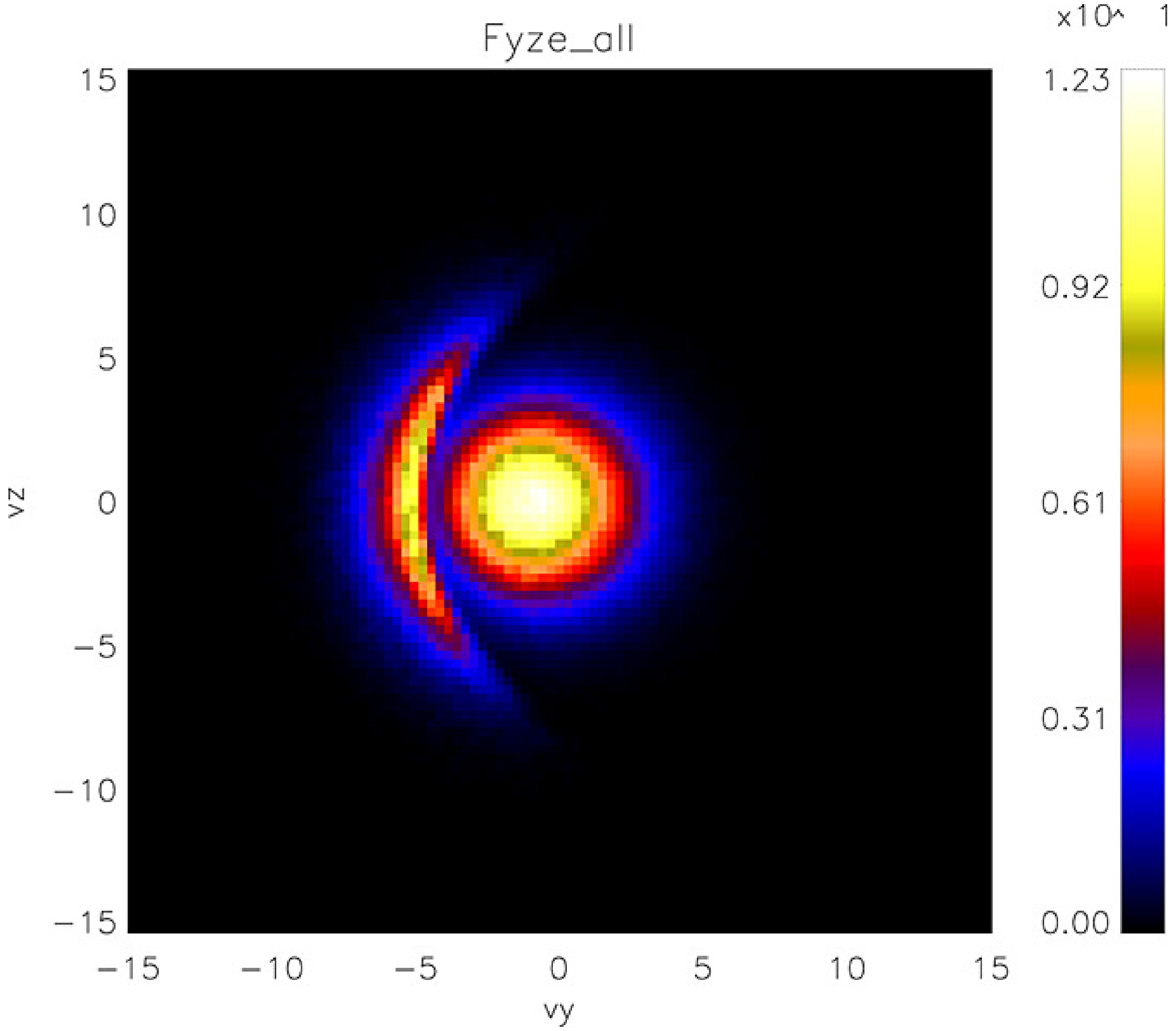


Figure 2.7: Figure showing a reduced distribution function obtained from integration of a full distribution function taken at the flow stagnation point in a Particle-In-Cell simulation. The presence of two different electron populations is clearly visible with a gyrotropic distribution at the center and a highly non-gyrotropic crescent-shaped distribution located at around -5 V_y. Image taken from *Hesse et al.* [2014].

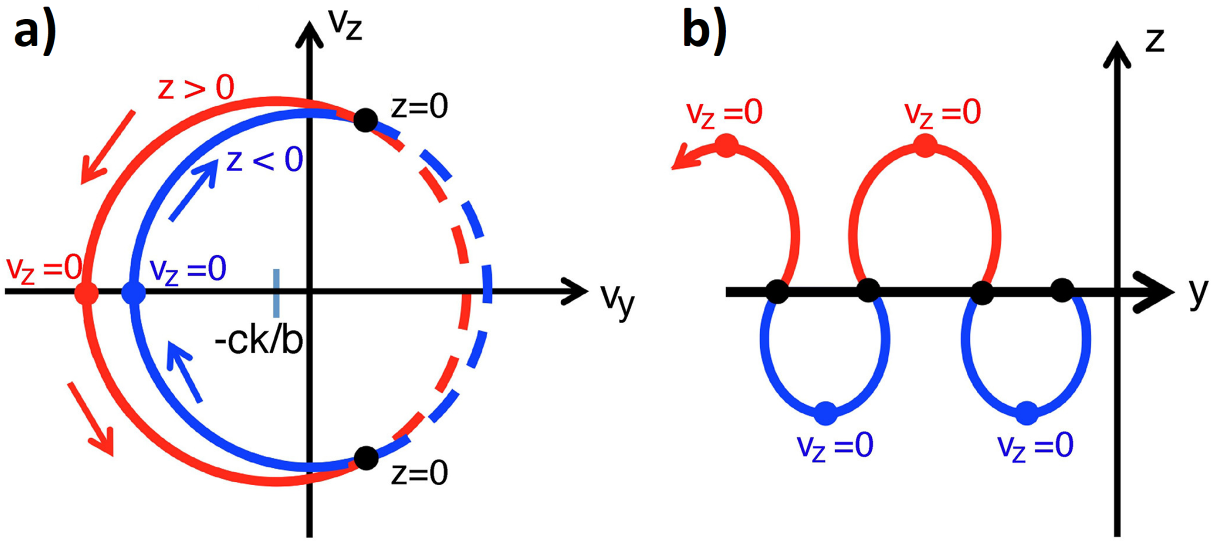


Figure 2.8: Figure showing electron trajectories explaining the origin of electron crescents at the electron distribution function. Image taken from *Bessho et al.* [2016].

an EDR.

A few things to bear in mind for the next chapters of this thesis is first that all of the explanations given here using simulations were done using the GSE coordinates. In that coordinate system and using standard conventions for the starting conditions of the simulation, the crescent will be located in the (V_y, V_z) plane, and more precisely on the negative V_y side. However, when using real spacecraft data, it is usually preferred to visualize the electron distribution functions in the field-aligned coordinate system, yielding electron crescents in the $(V_{\perp_1}, V_{\perp_2})$ plane, and more precisely on the positive $V_{\perp_1}$ side. Moreover, the explanations given here indicate that these "electron crescents" should only be visible on the magnetospheric side of the current sheet although I will show later that is not always the case and electron crescents can also be visible on the magnetosheath side.

The first observation of electron crescents in the data was presented in *Burch et al.* [2016] where an EDR event was identified in the MMS data only a month after its launch. Figure 2.9 shows this detection where we can clearly see the evolution of the crescent-shaped structure as MMS traverses the EDR. In the $(V_{\perp_1}, V_{\perp_2})$ plane, the crescent is located in the positive $V_{\perp_1}$ values and then the crescent slowly shifts into a circle, meaning that the electrons get remagnetized as MMS gets away from the EDR.

2.4.3 Specificities of guide field reconnection

When the magnetic field lines that reconnect are not completely anti-parallel (meaning that the shear angle between the magnetosheath and magnetospheric magnetic fields is less than 180°), the off-plan components do not reconnect. The remaining off-plan magnetic field is called a "guide field". To visualize it, one usually puts itself in the "LMN" coordinate

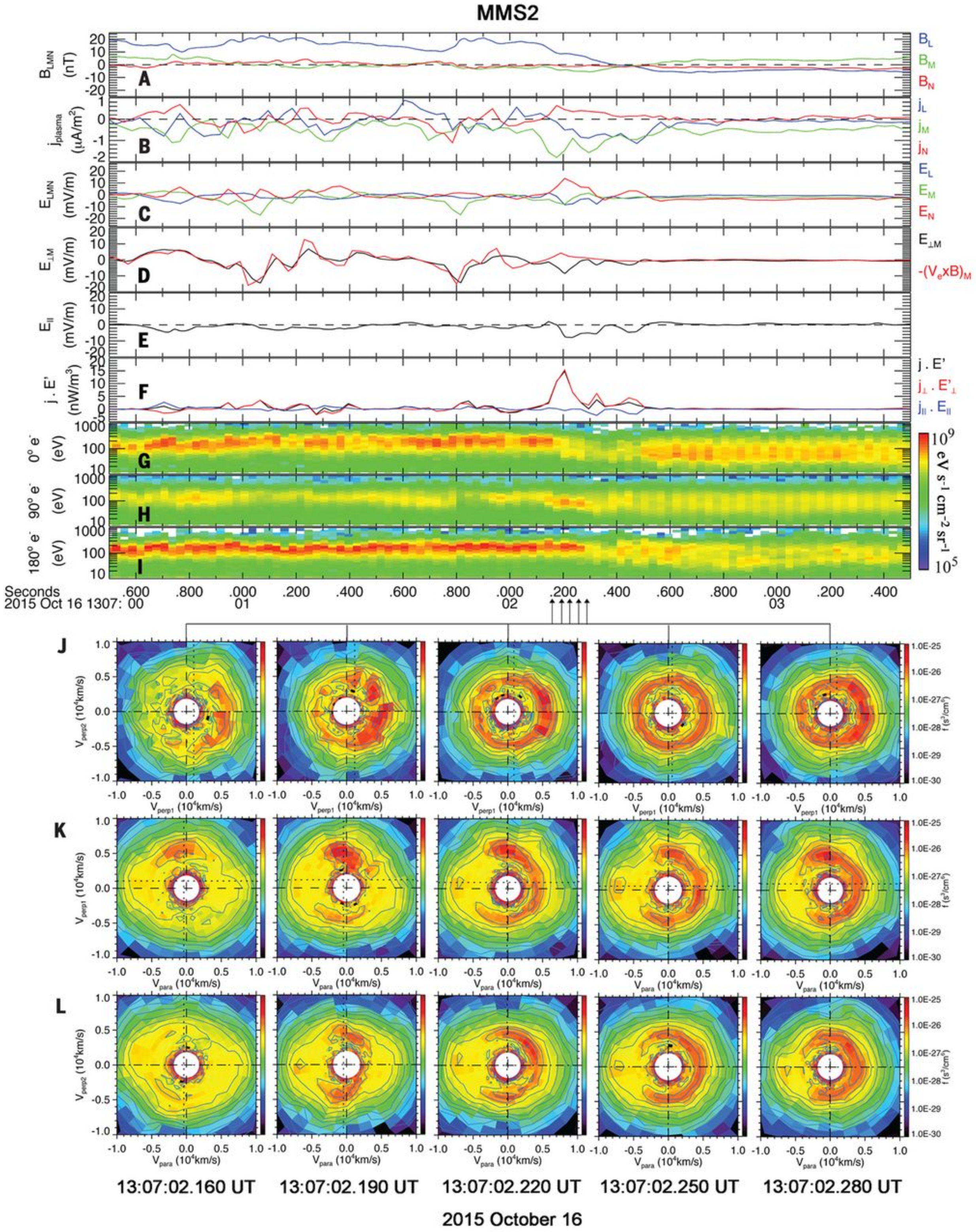


Figure 2.9: Figure showing the first electron velocity distribution function crescents observed in the MMS data. Image taken from *Burch et al.* [2016].

system, it is a rotation of the GSE or GSM coordinate system so that the L component is the one with the highest magnetic field variation and the N is the one with the smallest magnetic field variation. This coordinate system is usually found using the Minimum Variance Analysis (MVA), the different coordinate systems as well as the explanation of the MVA can be found in the Appendix sections A and B. In the "LMN" coordinate system, the intensity of the guide field is measured by the ratio of the magnetic field components B_M/B_L. Events with values below 0.1 or 0.2 are considered as low guide field (or high shear) events. For values above, the classifications of the intensity of the guide field can vary depending on authors but events with values above 1 are always considered as high guide field (or low shear) events.

Previous studies have shown that the presence of a guide field strongly affects the properties of reconnection as well as the particle dynamics (see *Pritchett and Coroniti* [2004]). It has also been shown that the presence of a guide field causes the magnetic field strength to be nonzero at the X-point which can magnetize electrons near the X-point, reducing the size of the electron gyroradius relative to the size of the current layer (see *Swisdak et al.* [2005]). The remagnetization of the electrons at the X-point is also responsible for the decrease of the intensity of the electron velocity distribution function crescents observed in the EDR (see *Hesse et al.* [2016]).

2.5 Machine Learning insights

Machine Learning is a part of what is called ”Artificial Intelligence” and regroups techniques enabling algorithms to improve automatically through experience and by the use of data. This section aims to introduce the basic concepts of machine learning that will be necessary to understand the work discussed in the rest of this thesis. We will first start with a brief history of the main breakthroughs in machine learning, then we will go on to explain the underlying mechanisms of machine learning before enumerating a few of its applications in plasma physics.

2.5.1 A brief history of machine learning

The first use of the term ”machine learning” can be traced back to 1959 where Alan Samuel, an American computer scientist from IBM, used it to describe the ”checkers-playing program” he designed in 1952 (*Samuel* [1959]). Since computer memory was very limited at the time, it was not possible for the algorithm to predict all possible outcomes in order to win, so instead Samuel created what is now called the ”alpha-beta prunning” : based on the position of the pieces on the board, the algorithm computes a scoring function that measures the current chance of winning. To choose the next move, the program would search a few moves ahead for the board configuration that would give it the best scoring value, and play accordingly.

The first learning method he used was called ”rote learning”, which simply consisted of saving each board configuration encountered paired with the computed score function value. In this way, when playing a game with a board configuration that had already been encountered, no computation was needed for this position, allowing for an increase in the depth of the search tree for the next move. The second learning method was called ”learning by generalization” and was used to modify the parameters of the scoring function by making the program play against another version of itself.

In parallel to Samuel’s work, another American computer scientist, Franck Rosenblatt, created in 1957 the first Perceptron algorithm (*Rosenblatt* [1958]) based on the work of Warren McCulloch and Walter Pitts on the mathematical modelization of artificial neurons (later called MCP neurons). Basically, the Perceptron is the combination of multiple MCP neurons where each neuron has an activation threshold and a weight so that an output is produced by computing the weighted sum of the inputs. The output of his algorithm was either 0 or 1 allowing for binary classification of shapes or letters. He was the first to show that an algorithm could learn from data. During a training phase, the algorithm is shown labeled samples and will update its internal values so that next time it is shown the same sample, its answer will match the label, based on the fact that samples with the same label should produce similar inputs for the algorithm and thus, a similar output.

10 years later, in 1967, Thomas Cover and Peter Hart published the Nearest Neighbor

pattern classification algorithm (*Cover and Hart* [1967]). It is the paving stone of what is nowadays known as kNN (k-nearest neighbors) which classifies a data point according to its neighbors' classification assuming that similar things exist close to each other in the parameter space. In 1970, Seppo Linnainmaa implemented the first backpropagation algorithm allowing for an error to be processed at the output of a neural network and then distributed backwards through the network's layers to adjust the hidden layers of neurons/nodes (*Linnainmaa* [1976]). The next year saw the first implementation of a decision tree by Messenger and Mandell in the THAID project (*Messenger and Mandell* [1972]).

But in the AI world, the 70's are mostly known as the "AI winter period". It was a period of reduced funding and interest in AI research due to pessimism in the AI community and in the press leading to a lack of serious research and breakthroughs in the field. The AI winter brought to the schism between AI and machine learning. Until then, machine learning had been used as a training program for AI, but in the late 70's, computer science and AI researchers abandoned neural networks and turned to the use of logical, knowledge-based approaches instead. Another, albeit smaller, AI winter occurred at the end of the 80's. Interest in the AI field had started to rise again with the use of AI programs called "expert systems" by large corporations around the world. However, the many technical issues with these systems lead to the collapse of the market for specialized AI hardware in 1987.

Nevertheless, new milestones in the machine learning field were still achieved by researchers after the AI winter with at first the invention of "boosting" algorithms. The concept of boosting was first presented in a 1990 paper titled “The Strength of Weak Learnability,” by Robert Schapire (*Schapire* [1990]). These algorithms are made up of weak classifiers which give poor results alone, but if a high number of them are trained, the average answer is a strong classifier. It is based on the known phenomenon called the "wisdom of the crowd" where for example, it is possible to guess the number of beans in a jar by taking the average of the answers of each person in a crowd, the bigger the crowd, the closer the result will be from the truth. In 1997, Jürgen Schmidhuber and Sepp Hochreiter described the model of a neural network known as the Long Short-term Memory algorithm (or LSTM) (*Hochreiter and Schmidhuber* [1997]). LSTM can learn tasks that require memory of events that took place thousands of discrete steps earlier, allowing for huge breakthroughs in the following years in the speech recognition field.

From the year 2000 to nowadays, machine learning is used worldwide thanks to significant breakthroughs, allowing for facial recognition algorithms, self-driving vehicles and many other applications. The increase of computational power in the last two decades lead to the enhancement of previously used architectures, such as Convolutional Neural Networks (CNN), and the creation of new ones, like Generative Adversarial Networks (GAN). This allowed machine learning to become integral to many widely used software services and applications.

2.5.2 The different available frameworks

We will now go a little bit further into how machine learning algorithms are conceived by first discussing the different frameworks that are available to solve problems. Choosing the right approach will usually result in better results in the end as well as allowing the user to better define what resources are needed in order to implement the algorithm.

The first learning approach in machine learning is called ”Supervised learning”, it is the most common type of framework in machine learning but it is the one that requires the largest amount of upstream work. The goal is to create a training dataset of labeled samples that the algorithm is going to learn from during a training phase. This phase enables the program to learn the underlying patterns and relationships between the input data and the output labels. Supervised learning algorithms are usually the most robust and trustworthy algorithms because they allow for the most control over the learning phase and thus the results.

Building the training dataset is the most crucial part as it will directly influence the quality of the predictions. The data used must be diverse enough to represent the complexity of the data that the algorithm will be running its predictions on and the data must be cleaned and balanced as missing and duplicate data will bias the algorithm leading to poor results when working on new data. Most of the machine learning algorithms used in our everyday lives are trained on labeled datasets. The range of applications can go from automatic image detection to price market predictions, and this is why personal data is so valuable nowadays to tech companies.

The second learning approach is called ”Unsupervised learning” and uses machine learning techniques to cluster and analyse unlabeled data. The aim of these models is to discover underlying patterns in the data and group samples accordingly, without the need for prior human definition of these groups. There is no need to provide labeled samples which can be very time efficient but since there is much less control over what the algorithm does, these algorithms are usually considered less trustworthy than supervised learning models.

The last learning approach is quite different than the previous two and is called ”Reinforced learning”. It works in a very different way as the objective is to train the algorithm to make a series of decisions based on its environment. For example, it was the approach used for Alan Samuel’s checkers-playing program. Usually, the algorithms decisions are chosen so that they will maximize a reward value. This reward value is assigned to a given configuration, it can be as simple as a board configuration for chess-playing programs or else as complex as reaching certain goals of safety or ride time reduction for autonomous vehicles.

This learning approach is used when a choice of actions is given to the algorithm and there is too many possible configuration to teach each one of them to the model. Instead, the model will train by ”playing” with the environment for long simulated times, sometimes even against itself for games or applications requiring an opponent, which can for example lead to having chess-playing algorithms like Google's AlphaZero with hundreds of years worth of experience compared to a human after having played several millions of games against itself. Since there is no training dataset, these types of algorithm usually are extremely bad at first and improve with time, usually outperforming humans to the task in the end.

Although machine learning algorithms are now used in many different fields and though the range of possible applications is unlimited, the core tasks that these algorithms can perform can be divided into 3 different groups. The first group includes ”Regression problems” which are problems where we are trying to predict a quantitative value based on the values of one or more variables. A famous study case of a regression problem in machine learning is the prediction of the market price of a house using a dataset containing the selling prices as well as the features of other houses in the same city. Then, ”Classification problems” are, on the other hand, problems where we are trying to predict a qualitative value. The classification of iris flowers among three species (Setosa, Versicolor, or Virginica) from sepals' and petals' length and width measurements is probably the most famous example of classification problem in the machine learning field. Finally, ”Clustering problems” are different than the two previous tasks as they aim to divide some data into homogeneous parts by minimising the distance in the parameter space between members of the same group. It falls into the unsupervised framework as it is possible for the user to control the number of clusters wanted but not to define these clusters.

2.5.3 Practical Machine Learning guide

The concept of learning explained with the Multilayer Perceptron (MLP)

As mentioned above, the first algorithm able to learn from data was Rosenblatt's Perceptron, which used the artificial neuron model from McCulloch and Pitts. At its core, a neuron is just the weighted sum of different input values to which is applied an activation function. Since then, applications in the machine learning field used architectures containing an ever growing number of neurons that can go from a few tens of neurons to several millions. To illustrate the concept of learning for algorithms, we will focus on a simple architecture that is an evolution of Rosenblatt's Perceptron : the Multilayer Perceptron (or MLP).

A Multilayer Perceptron is a supervised learning method which structure is composed of an input layer, one or several hidden layers and an output layer. Each layer in composed of one or several nodes and each node of the n'th layer is connected to every node of the n+1 and n-1 layers. This is why the Multilayer Perceptron can also be called ”Feed-forward Multilayer Perceptron”. The input layer receives only one sample from the dataset at the time and the output layer gives one answer per sample.

The number of nodes for the input layer is determined by the number of parameters, called ”features”, of a sample (such as the number of pixels if the samples are images) while the number of nodes of the output layer is determined by the task at hand. There is no rule to determine the number of nodes inside the hidden layers, this number is usually chosen after an optimization period of trial and error. For the input layer, each node will receive the value of a different feature of the same sample. An example of the architecture is shown in figure 2.10.

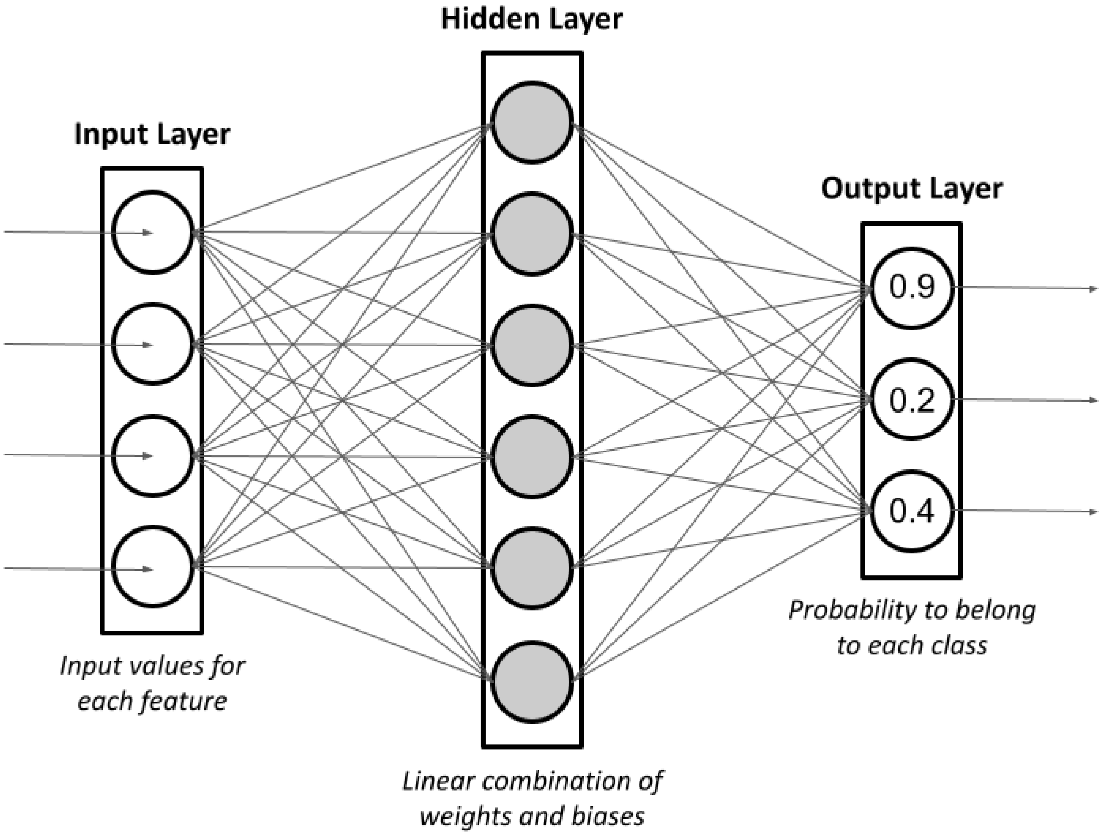


Figure 2.10: Figure showing an example of the Multilayer Perceptron (MLP) architecture in the case of a multiclass classification problem.

The number of nodes in the output layer will be one for a Binary Classification problem (the node giving a probability between 0 and 1) but for a Multiclass Classification problem, the number of nodes will be the number of different classes the sample can belong to (each node giving a probability between 0 and 1 to belong to each class). The number of hidden layers and of nodes inside them can however be variable and are free parameters called ”hyperparameters” over which the user has control. Other hyperparameters can be set by the user, as explained later.

This structure is designed to mimic the brain’s behavior where, if a neuron is activated, it will activate certain neurons and not others. To see how this occurs here, let us consider the activation of one neuron p from a hidden layer n. The activation of this neuron depends on the behavior of each neuron of the previous layer, the $n-1$ layer. Each neuron of the $n-1$ layer will be given a different weight ω with $\omega_{n-1,i}$ the weight of the i’th neuron of the $n-1$ layer. Let us call $a_{n-1,i}$ the activation value of the i’th neuron of the $n-1$ layer (the network converts every value from the input layer into a number between 0 and 1, making it an activation value). In a Multiclass Classification problem, the activation value of each node in the output layer will be the probability for the sample to belong to the class the node represents. Thus, the activation value of the p'th neuron of the n'th layer is computed with the following formula :

$$\sum_{i=1}^{m} f_{activation}(\omega_{n-1,i} \times a_{n-1,i} - b_{n,p}) = a_{n,p} \tag{2.26}$$

where m is the number of neurons in the the $n-1$ layer, $b_{n,p}$ is the bias of the p'th neuron of the n'th layer $f_{activation}$ is the chosen activation function, which will usually either be the sigmoid, softmax, hyperbolic tangent or ReLU function. For networks with many layers, the Rectified Linear Unit (ReLU) function is the standard for hidden layers because it overcomes the vanishing gradient problem, allowing models to learn faster and perform better. The ReLU function is defined as :

$$ReLU(a) = max(0, a) \tag{2.27}$$

returning 0 if a $\leq$ 0 and a otherwise. Due to the way it maps the values, ReLU is never used as the activation function of the output layer. For the output layer, the activation function will depend on the task and the problem at hand. The sigmoid function is usually used for Regression problems, it is defined as :

$$S(a) = \frac{1}{1+e^{a}} \tag{2.28}$$

In the case of Multiclass Classification problems, the activation function of the output layer will be the Softmax function, defined by the formula :

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}} \tag{2.29}$$

with z the activation value of the i'th neuron of the output layer and k the total number of neurons on the output layer. Applying the Softmax function on the activation values of the neurons on the output layer means that each component will be in the interval [0,1], and the components will add up to 1, so that they can be interpreted as probabilities, even though they are not probabilities in the strict mathematical sense. Thus the activation value of a neuron of the output layer gives the probability to belong to the class it represents, and the final answer of the algorithm can be defined either as the class represented by the neuron with the highest activation value on the output layer (Multi-class Classification) or by all the neurons with an activation value over 0.5 if the classes are not mutually exclusive (Multi-label Classification). The bias is a value telling how high the weighted sum needs to be before the neuron starts getting meaningfully active, making it possible to give a low activation value to a neuron if needed. The weights and the biases can then be positive or negative numbers between 0 and 1.

This same process is then applied to every node that is not in the input layer. At first, every weight and bias values are taken randomly so if the algorithm is given a sample, it is going to give a random answer at the end. But since we are in a supervised training environment, the correct answer that should have been given is known so the algorithm will change all the values of the weights and the biases so that they would give the right answer in the end. The algorithm is going to try to do that for every point in the training database, hence the term of "learning".

The way those weights and biases are tuned is by trying to minimize something called the "cost function", also called the "loss function". When the algorithm gives a wrong answer during the training part, which means that the activation value of the neuron representing the right class is not going to be 1 and the activation values of the other neurons is not going to be 0, it will compute a "cost function" telling how far from the right answer the algorithm was. This cost function is computed in the following way :

$$C = \sum_{i=0}^{m} (a_{output,i} - a^{d}_{output,i})^2 \tag{2.30}$$

where $a_{output,i}$ is the activation value of the i'th neuron of the output layer and $a^{d}_{output,i}$ is the desired activation value for the i'th neuron of the output layer (which is going to be either 1 or 0 for classification or any other number for regression). The activation value of a neuron of the output layer depends on the weights and biases of all the neurons that are connected to it so the total cost function can be written as $C(\omega_1, b_1, ..., \omega_n, b_n) = C(\mathbf{W})$ where $\mathbf{W}$ is a vector that contains every weights and biases of the system. The cost function is small when the network confidently classifies the sample correctly and is large when the network gives a random probability to belong to each class. Trying to minimize this cost function is trying to find its absolute minimum in the parameter space of C. During the process of "learning", the algorithm is going to move in the parameter space by following the negative gradient of the cost function $-\nabla C(\mathbf{W})$ which will tell how much the parameters should be changed in order to minimize the cost. Since the activation value of a neuron depends on the parameters of every neuron from the previous layer, the weights and biases of the system are changed going backwards, from the output layer to the input layer, hence the term of "Backpropagation".

The total cost function is the average of the cost function computed for every data point of the dataset. But instead of computing the total cost function, which requires a lot of computational work and memory, we are going to use fractions of the dataset called "batches" that are made after shuffling the whole dataset. The algorithm will run through every batches and will compute a cost function for each one of them. Each cost function will then be used to adjust the weights and biases of the system through backpropagation. The "batch size" is a hyperparameter over which the user has control. Another hyperparameter is the number of "epochs" which represents the number of times the algorithm will go through the whole dataset, creating new random batches after each completion.

Using batches instead of computing the total cost function of the whole dataset is called a "Stochastic Gradient Descent" and it is the easiest ways to optimize the performance of the system during the learning process even though there are several other ways to do it. The Stochastic Gradient Descent technique using fractions of the whole dataset to compute cost functions is less accurate that taking the gradient of the total cost function but it gives a good approximation of the direction of the gradient of the total cost function while being way more efficient in terms of computational work. The "Stochastic Gradient Descent" technique can however encounter many problems when there are for example local minima, the algorithm can get stuck in them, giving bad results in the end. That is why since 2014, the machine learning community uses a new standard called "ADAM" (*Kingma and Ba* [2014]). It is an improvement of the "Stochastic Gradient Descent" that uses momentum to avoid getting stuck in a local minimum and an adaptive learning rate.

The learning rate is the speed at which the algorithm is going to learn. If the learning rate is too low, it can lead to underfitting meaning that the algorithm will not have learned enough at the end of the training and if it is too high, it can lead to overfitting meaning that the algorithm learns by heart and does not really understand what it is doing. In more mathematical terms, underfitting is when the function used by the model to fit the training data is not complex enough resulting in overall bad performance and overfitting is when the function used for the fit is too complex, the algorithm will perform very well on the training samples but will fail to produce good results when generalizing the predictions on other samples than the ones it was trained on.

Figure 2.11 shows loss curves with a case of underfitting on the left and a case of overfitting on the right. Loss curves are a pair of curves representing the decrease of the loss function during the training as a function of the number of epochs for the training set and for the validation set. What we call the "validation set" is a part of the training dataset that the algorithm will not see and be trained on, it is used during the training phase to monitor the learning and is extremely useful notably to identify overfitting.

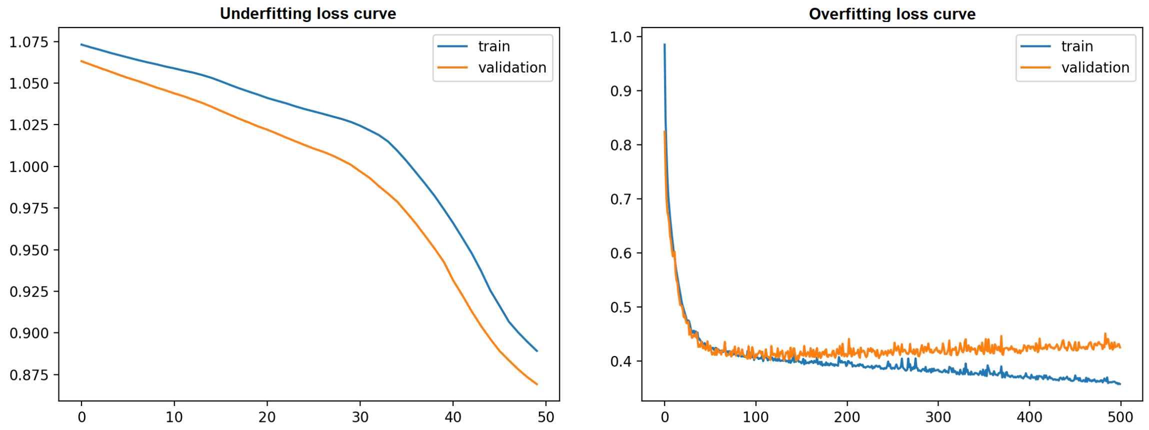


Figure 2.11: A case of Underfitting on a dataset (Left) and a case of Overfitting on a dataset (Right). The blue curve is the loss on the training set and the orange curve is the loss on the validation set (curves taken from machinelearningmastery.com)

With these two curves, it is easy to determine if the model is overfitting, underfitting or is behaving well. As can be seen on the left image of Figure 2.11, underfitting can be observed because the loss has not stopped decreasing at the end of the training so the training phase is incomplete. On the right, we can see that the training and validation curves both decrease at the start but the validation loss curve stops decreasing and starts increasing after around 100 epochs, meaning that the algorithm is starting to learn from noise in the training set and thus the performance of the algorithm on the validation set, which is unseen data attesting the generalization of the model, will decrease.

The need for and use of different metrics

Once our algorithm is trained on the training set, we want to know how it will perform before starting to properly use it. To do so, we test its predictions against a set of labeled samples that we had not used to train the algorithm, we call it the "test set". The method used to evaluate the algorithm is called a "metric" and its choice will vary depending on the type of problem and the application.

For regression tasks, the error is the difference between the target value and the predicted value. The most common metrics used are the Mean Squared Error (MSE), the Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE). MSE is computed as the mean of the squared differences between expected and predicted values in the dataset, it can be written as :

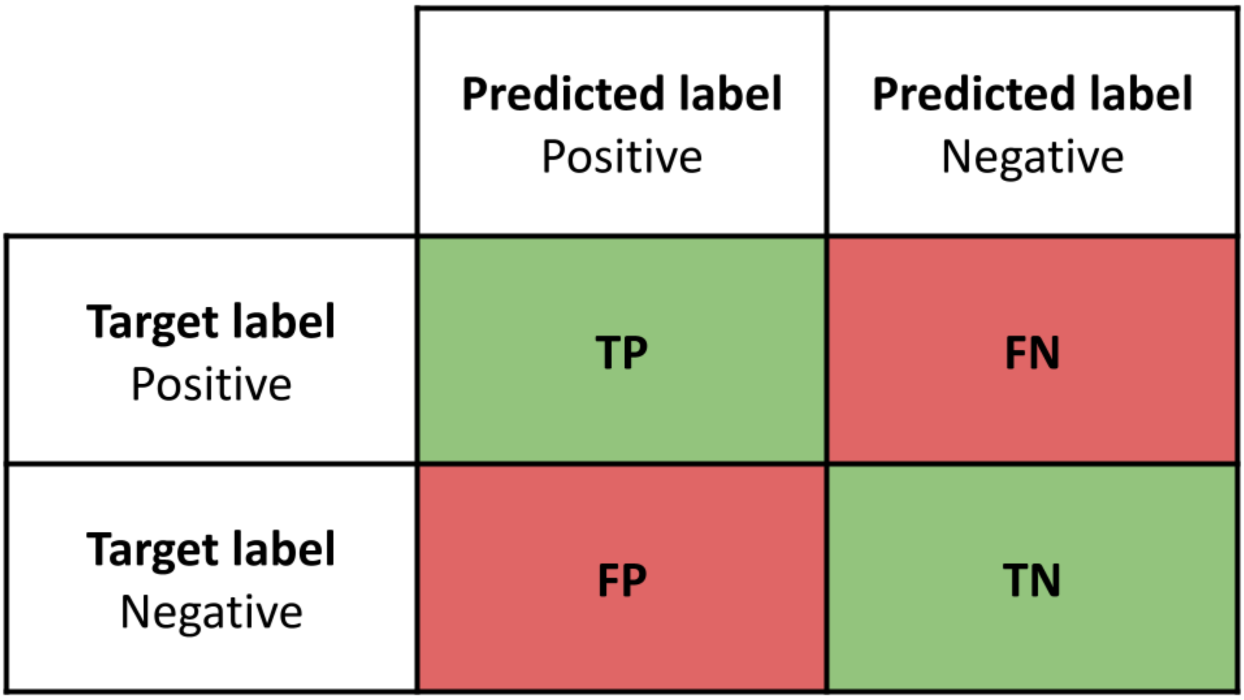
$$MSE = \frac{1}{N}\sum_{i}^{N}(y_i - \hat{y_i})^2 \tag{2.31}$$

where y_i is the i'th expected value, $\hat{y_i}$ is the i'th predicted value and N is the total number of samples in the test set. Since the difference is squared, the error is always positive. RMSE is computed by taking the square root of MSE so that the unit of RMSE is the same as that of the expected values, which is very convenient to quantify the real expected margin of error of the algorithm. Similarly, MAE is computed by taking the absolute value of the difference in MSE instead of the square of the difference. It means that the unit is the same as that of RMSE but MAE varies linearly contrary to RMSE.

For classification tasks, the metrics are computed for each class using a "confusion matrix" such as shown in figure 2.12. In the figure, we see how the predictions of the algorithm are classified into 4 categories when taking the "positive" label as reference. A true positive is when the predicted label of the algorithm matches the target label that was expected, a true negative is when the algorithm successfully classified a sample as not belonging to the reference class. A false negative is when a sample belonging to the reference class was mislabeled as belonging to another class by the algorithm, and a false positive is when a sample that is not from the reference class gets mislabeled as belonging to the reference class.

Using this matrix, it is possible to construct different metrics characterizing different types of mistakes. Some of the most common metrics used for classification are the Precision and the Recall. The formulas of these metrics are the following :

$$Precision = \frac{TP}{TP + FP} \tag{2.32}$$



	Predicted label Positive	**Predicted label** Negative
Target label Positive	**TP**	**FN**
Target label Negative	**FP**	**TN**

Figure 2.12: Figure showing an example of confusion matrix. The predicted label is the label given by the algorithm and the Target label is the expected label. The errors are taken into account according to the "Positive" label. TP = True positive, FP = False positive, FN = False negative, TN = True negative

$$Recall = \frac{TP}{TP + FN} \tag{2.33}$$

where TP, FP and FN respectively stand for number of True Positives, False Positives and False Negatives.

The precision tells which proportion of positive identifications were effectively correct while the recall tells which proportion of positive results were correctly identified. A low precision means that the algorithm will tend to overestimate the number of instances of the reference class and a low recall means that the algorithm will overlook a lot of instances of the reference class. Usually, increasing one of these metrics means the decrease of the other so depending on the application, one can prefer to either balance these two metrics or maximize one at the expense of the other. This is why another metric called the "F1-score" is commonly used in pair with the precision and the recall, it is the harmonic mean of the precision and of the recall and provides a balance between these two metrics.

$$F1\text{–}score = \frac{(2 \times Pre \times Rec)}{(Pre + Rec)} \tag{2.34}$$

where Pre and Rec respectively refer to Precision and Recall.

In the case of clustering, since it mostly is a collection of unsupervised learning techniques, there is no proper metric as defined for regression and classification. There are however ways to characterize the clustering errors with metrics such as the "silhouette score" or the "rand index" that respectively compute the distance between neighboring clusters and the similarity between clusters.

Another important architecture : the Convolutional Neural Network (CNN)

The previous sections referred to machine learning, which is a sub-domain of AI, but the architecture we will discuss here refers to a sub-domain of machine learning called "Deep Learning". Deep learning distinguishes itself from traditional "machine learning" by the fact that instead of having different generic layers containing neurons like in the Multilayer Perceptron, each layer will be a specific unit with a defined purpose, going further than traditional "machine learning" by creating more complex hierarchical models designed to mimic the human brain's operating mode.

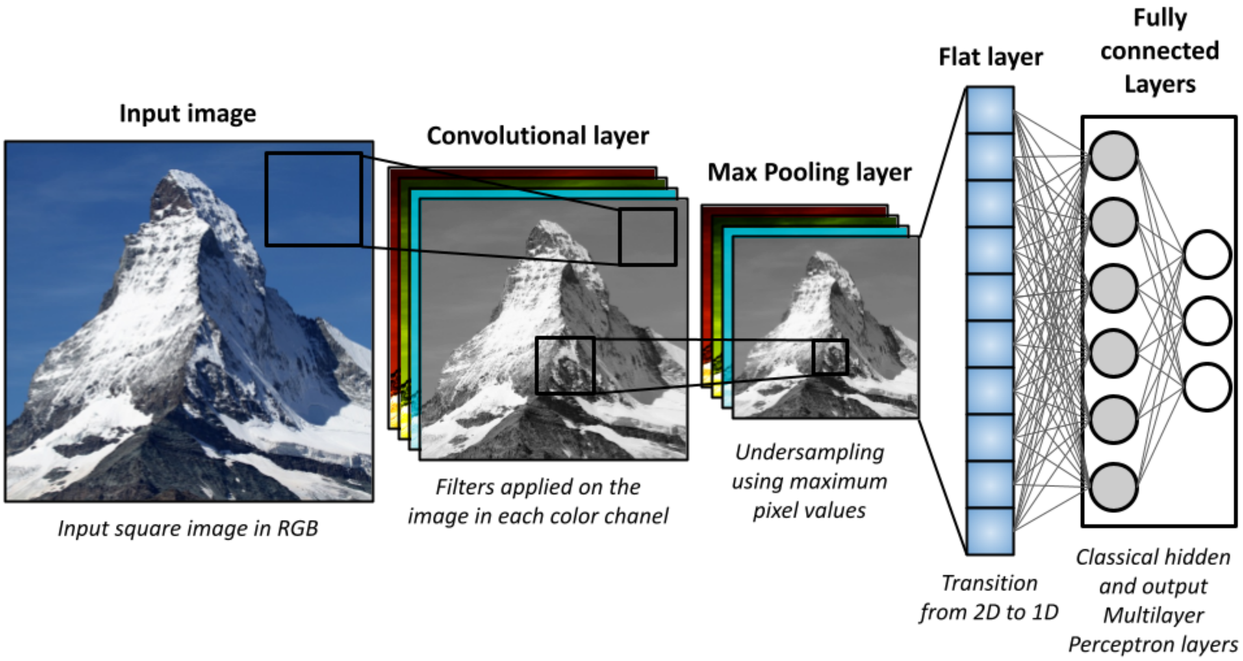
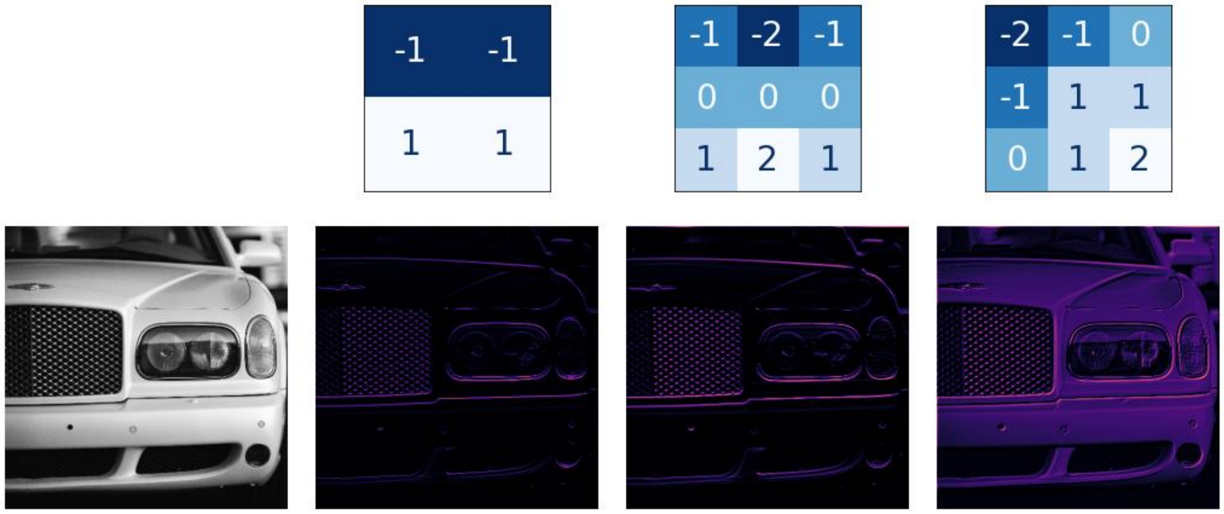


Figure 2.13: Figure showing an example of the Convolutional Neural Network (CNN) architecture in the case of a multiclass classification problem.

We will here explain the architecture known as "Convolutional Neural Network" (or CNN) that is the basis of the computer vision field. Its purpose is to deal with non 1D inputs, it can be images or audio files for example. The model, that can be seen in figure 2.13, is more complex than a simple machine learning model in that it is composed of several specialized layers that we will now discuss.

The first layer is called a ”Convolutional layer”, its role is to extract visual features from an image. During the training phase of the convolutional layers, the weights that are actually being adjusted are called ”kernels”. They can be represented as matrices and work by scanning over an image and producing a weighted sum of pixel values. The kernel will thus increase or decrease the importance of certain patterns of information in the image. During the training, a CNN tries to learn what features it needs to solve the classification problem which means finding the best values to fill its kernels with. Coupled with a ReLU activation function, the convolutional layer finishes the extraction according to the filter defined by the kernel.

Figure 2.14 shows how different kernels will affect the feature extraction. For example, the first kernel shown in the figure is symmetrical with negative values on the top and positive values on the bottom, showing a clear horizontal dichotomy. This kernel will thus extract horizontal features from the image, as can be seen on the extracted car image below the kernel where all the vertical lines are very shallow, almost invisible, whereas the horizontal lines are emphasised. The second kernel is similar to the first one but it emphasises even more horizontal lines. The third kernel, which is diagonally symmetric with positive values at the bottom right corner and negative values at the top left corner, emphasises bottom right corners on the image.



-1	-1
1	1

-1	-2	-1
0	0	0
1	2	1

-2	-1	0
-1	1	1
0	1	2

Figure 2.14: Figure showing different kernels and how they affect the feature extraction on an image. Image taken from kaggle.com

The next layer of the CNN is a ”Pooling layer” and its role is to condense the image to enhance the extracted features. Indeed, after the Convolutional layer with the ReLU activation function, a lot of zeros are present and keeping them would increase the size of the model so we condense the image while keeping the useful information. To compress the image, we move a sliding window on the image and apply the pooling. Two parameters control the movement of the window, the stride and the padding which are respectively the distance in pixels the window moves at each step and the method used to treat the boundaries of the input. If the padding is ”valid”, the window will stay entirely inside the input but we will lose some information at the boundaries because the image will shrink a little, if the padding is ”same”, zeros will be added around the borders but this can have the effect of diluting the influence of pixels at the borders. There are two different pooling methods which are ”average pooling” and ”max pooling”, they will either do the average of the pixels in a chunk or take only the maximum pixel value in the chunk, a chunk being an element of the grid over which the window will slide which depends of the padding and the stride. An example of how max pooling works can be observed in figure 2.15.

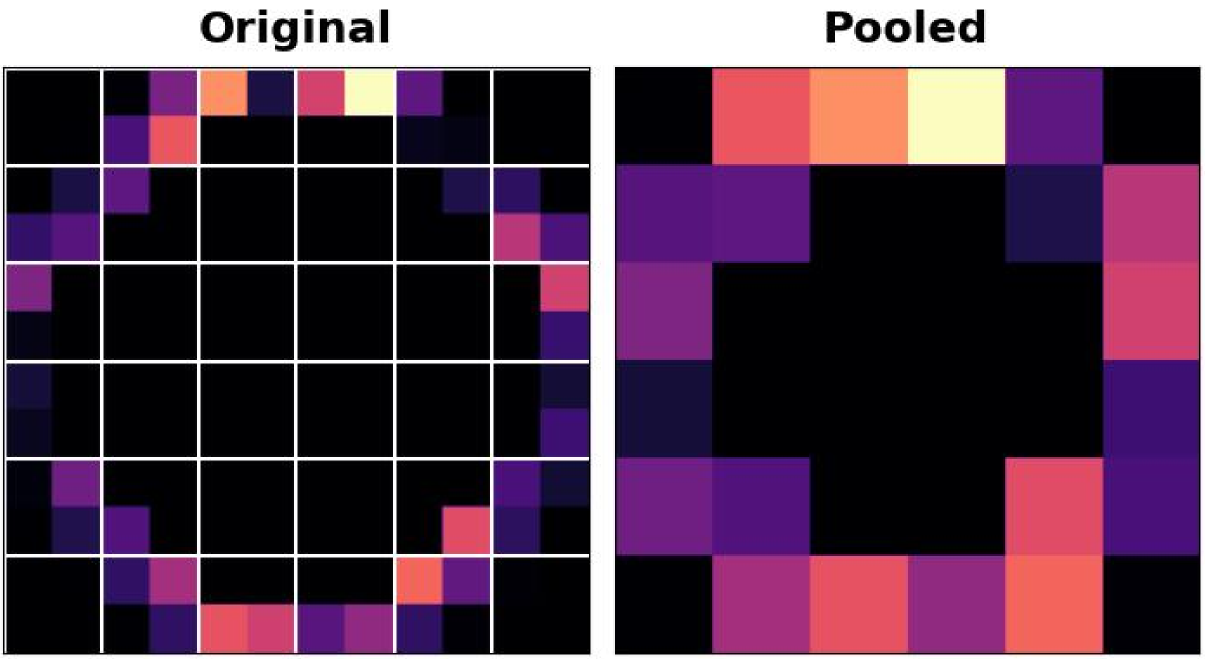


Figure 2.15: Figure showing the effect of max pooling on an image. Image taken from kaggle.com

The alternation of convolution and pooling layers can be repeated as many times as wanted, they constitute the basis of the model. However, the last layers of the CNN constitute the head and therefore cannot be repeated as it is the part used to do the classification. It is first constituted of a flattening layer to do the transition between 2D data and 1D data, and then this layer is itself fully connected to one or more hidden layers and an output layer working similarly to the ones in the Multilayer Perceptron. During the training, the weights and biases of these neurons are adjusted to treat the data recovered in the flattened layer and give the right predictions.

A quick word about other architectures

Many other deep learning architectures exist so we will not detail them all here but we will give some quick insights about some of them to complete the zoology of the different architectures.

We can first discuss the recurrent neural networks (or RNN) that is a fundamental architecture from which many other deep learning architecture are built. The difference with the classic Multilayer Perceptron lies in that instead of being completely feed-forward, meaning that each neuron of a layer is connected to each neuron of the next layer, a recurrent network might have connections that feed back into prior (or into the same) layers. This allows the RNN to remember some context during the processing of the data.

Some of the most advanced models to this day use the LSTM (Long-Short Term Memory) architecture. It uses the concept of memory cell which is based on the RNN architecture, it allows the network to retain its values for a short or a long time as a function of its inputs so that the cell can remember other things than just its last computed value. LSTM models are widely used in state-of-the-art applications like speech recognition or video processing. It is also of interest to the plasma physics field as it is a very powerful way to deal with pattern recognition using the physical context in time series data, which is what most of the spacecraft data look like.

Another state-of-the-art architecture is the Generative Adversarial Network (or GAN). It was presented in *Goodfellow et al.* [2014] to create fake images that look like real images. It uses two neural networks that work together, a generator and a discriminator. The discriminator is trained on real images and the generator's task is to create a fake image that will be mistaken for a real image by the discriminator. At first, the discriminator has no problem distinguishing the real images from the fake ones during the training but there will be a time when the output of the generator will be so good that it will reach a sort of equilibrium where the discriminator cannot tell the difference between the real and the fake images. This network can have scientific potential for space physics where someone could use this architecture to create images mimicking a physical phenomenon to train other algorithms when the amount of data is scarce to start with.

2.5.4 Machine Learning in space plasma physics

The use of machine learning techniques in the space plasma physics field is linked to the space weather field. It is not a new topic as papers using this technology were already published back in the 1990's, with the prediction of the geomagnetic index being among the earliest applications in the field (*Costello* [1998]). The prediction of the Kp index, which is a global representation of the maximum range of geomagnetic variation over 3-hour UT intervals and that can be used to determine the probability to observe auroras as early as possible, is still an ongoing research subject as proved by recent publications treating this subject (*Wintoft et al.* [2017]).

Early machine learning applications in this field were based on traditional neural network architectures but more and more state-of-the-art architectures are being used nowadays thanks to the increase of computational power since the 1990's. For example, the use of CNNs to analyse solar images to predict solar flares was not possible a few decades ago but it now is one of the possible ways to tackle this problem (*Huang et al.* [2018]). Many other machine learning applications in space plasma physics also exist going from plasmaspheric electron density estimation to the prediction of the foF2 parameter in ionospheric physics. A lot of them are regrouped in *Camporeale* [2019] which is a paper reviewing the use of machine learning in the space weather field.

To end this section about machine learning, I will now review in a bit more details two recent plasma physics applications of machine learning that were published in the last 3 years. The idea is here to show how these space plasma applications go from the data to the exploitable results.

Automatic detection of ICMEs

The first application we will discuss is detailed in the paper "Automatic detection of Interplanetary Coronal Mass Ejections from in-situ data: a deep learning approach" published in Astrophysical Journal in 2019 *Nguyen et al.* [2019]. Coronal mass ejections (or CMEs) are astrophysical events during which large quantities of solar plasma are ejected at high velocities. Interplanetary Coronal Mass Ejections (ICMEs) are the planetary counterpart of Coronal Mass Ejections that move out of the solar atmosphere. Their interaction with planetary magnetic fields can trigger geomagnetic storms that can heavily impact the Earth's magnetosphere and therefore human activity.

The WIND mission launched in 1994 has provided years of data allowing for many studies on ICMEs which have suggested several criteria to detect them. Some of the most common features of these ICMEs are an enhanced and smoothly rotating magnetic field, a low proton temperature and a low plasma beta. However, because of the strong variability of these events, not all these criteria are met for every ICMEs which makes an threshold based detection difficult to perform. Detection of ICMEs was done by visual inspection of the data which leads to biased catalogs depending on the observer. The paper shows a method that was developed to automatically detect ICMEs in the data of the WIND mission using a combination of CNN and peak detection algorithm.

In the Pre-processing phase, they used a fusion of multiple WIND ICME lists for a total of 657 ICMEs. The WIND instruments provide 30 input features and they computed 3 more variables leading to a total of 33 input variables. The data were grouped into windows of different sizes going from 1h to 100h sliding on the datasets at a period of 10 minutes. For each window, a similarity parameter is computed yielding 0 when no ICME intersects the window and 1 when a window perfectly matches an ICME.

A different CNN was trained for each window size to learn and predict the similarity

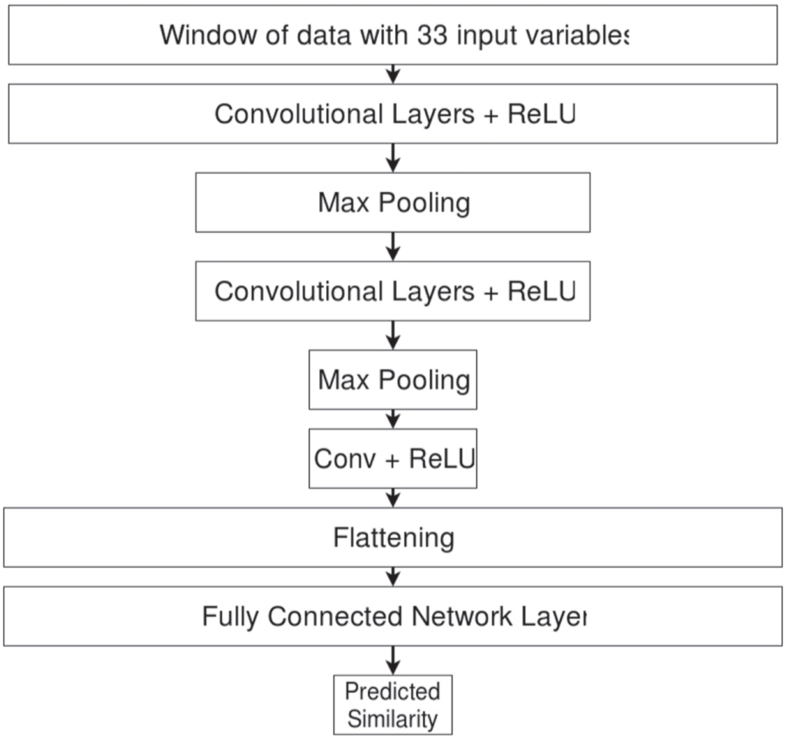


Figure 2.16: Figure showing the architecture of the CNN used to detect the ICMEs. Image taken from *Nguyen et al.* [2019].

of a given window associated with the data, the architecture of the CNNs can be found in figure 2.16.

Figure 2.17 shows a comparison between the expected ICME prediction (in red) and the prediction of the algorithm (in green) over a period of 2 weeks after applying a gaussian filter over the data. A peak detection algorithm is applied on the computed similarities in order to create the green prediction zones. We see that the predictions of the algorithm are very close to the expected ICME classification.

The method showed in this paper shows that machine learning, and more precisely CNNs, can be used to create unbiased, reproducible catalogs of astrophysical events.

Classification of plasma regions using machine learning

Another recent application of machine learning is detailed in the paper ”Automatic Classification of Plasma Regions in Near-Earth Space With Supervised Machine Learning: Application to Magnetospheric Multi Scale 2016–2019 Observations” published in 2020 (*Breuillard et al.* [2020]). Their goal was to automatically classify 10 key plasma regions using time series data from the MMS mission.

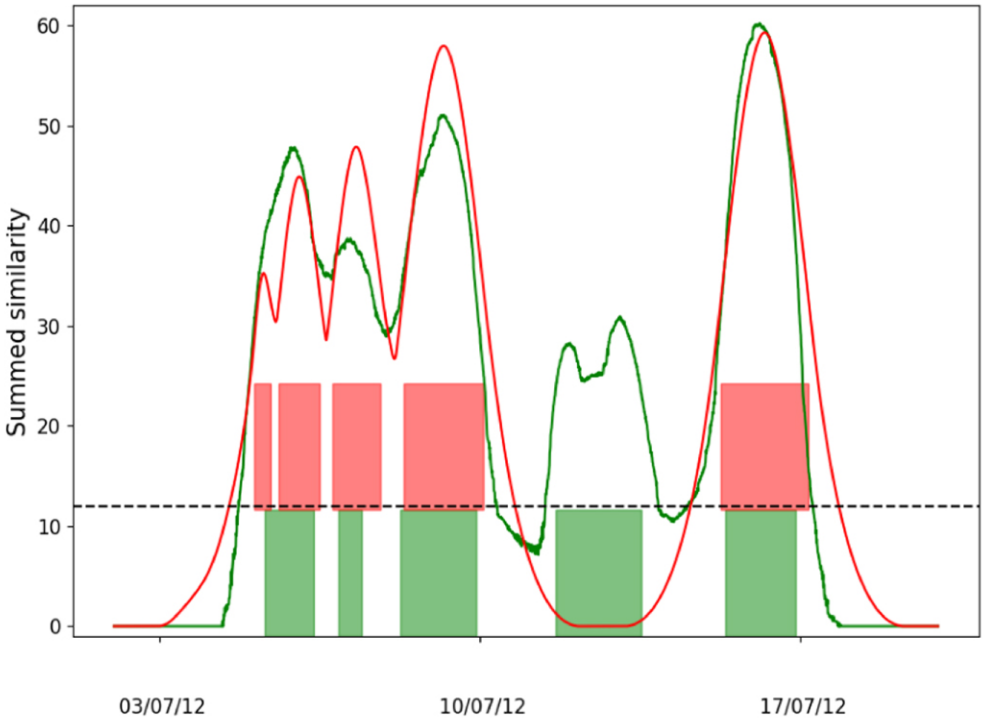


Figure 2.17: Figure showing the comparison between the expected ICMEs regions (in red) and the predicted ICMEs (in green) by the algorithm. Image taken from *Nguyen et al.* [2019].

Plasma region classification is usually done by visual inspection because of the strong variability of plasma parameters which does not allow for threshold based methods to perfectly work, human inspection is still needed to validate the results. The classification is thus a time consuming but necessary task as it is the only way to produce statistical studies on these plasma regions.

To train and build a supervised learning algorithm, a large amount of labeled data is required which is a very time consuming task. Fortunately, such a manual labeling of the plasma regions is already done by the Scientist-in-the-loop (SITL) system which involves a human expert scientist designated to select data of interest that will be transmitted to the ground as high-time resolution data. The idea was to use a parser to convert all the comments into standardized labels that can be used for the training phase.

Their model uses a total of 12 variables which include the magnetic field components, the ion density, the total bulk velocity magnitude and its components, the total, parallel and perpendicular temperatures. Since inputs must be of the same size, the dataset is grouped as time series corresponding to the labeled time intervals split into equal chunks of 3 minutes.

The architecture of the model is called a fully convolutional network (FCN) which is a competitive deep-learning sub-type of CNN for time series classification composed of a sequence of three time convolution blocks followed by a global average pooling block. Each time convolution block is designed the same way : a convolution layer, a batch normalization layer, and a Rectified Linear Unit (RELU) activation function.

Once the modeled is trained, it can be used on unseen MMS data to classify the plasma regions as can be seen in figure 2.18. The model classifies the data by chunks of 3 minutes. The black line shows the classification of the model while the colored sections show the human classification that would have been done on this time series. Not all the chunks can be easily classified by human inspection, hence the amount of chunks of data that are not colored. We see that their model performs well against the human predictions despite the high variability of the parameters that can be observed.

Conclusion

This application and the previous one show that even though the use of machine learning methods is not yet a standard in the plasma physics field, there is an ever growing number of publications incorporating this technology and showing strong results. In the next chapter, I will present the work I have done on the automatic detection of electron diffusion regions using the elements that were introduced in this chapter.

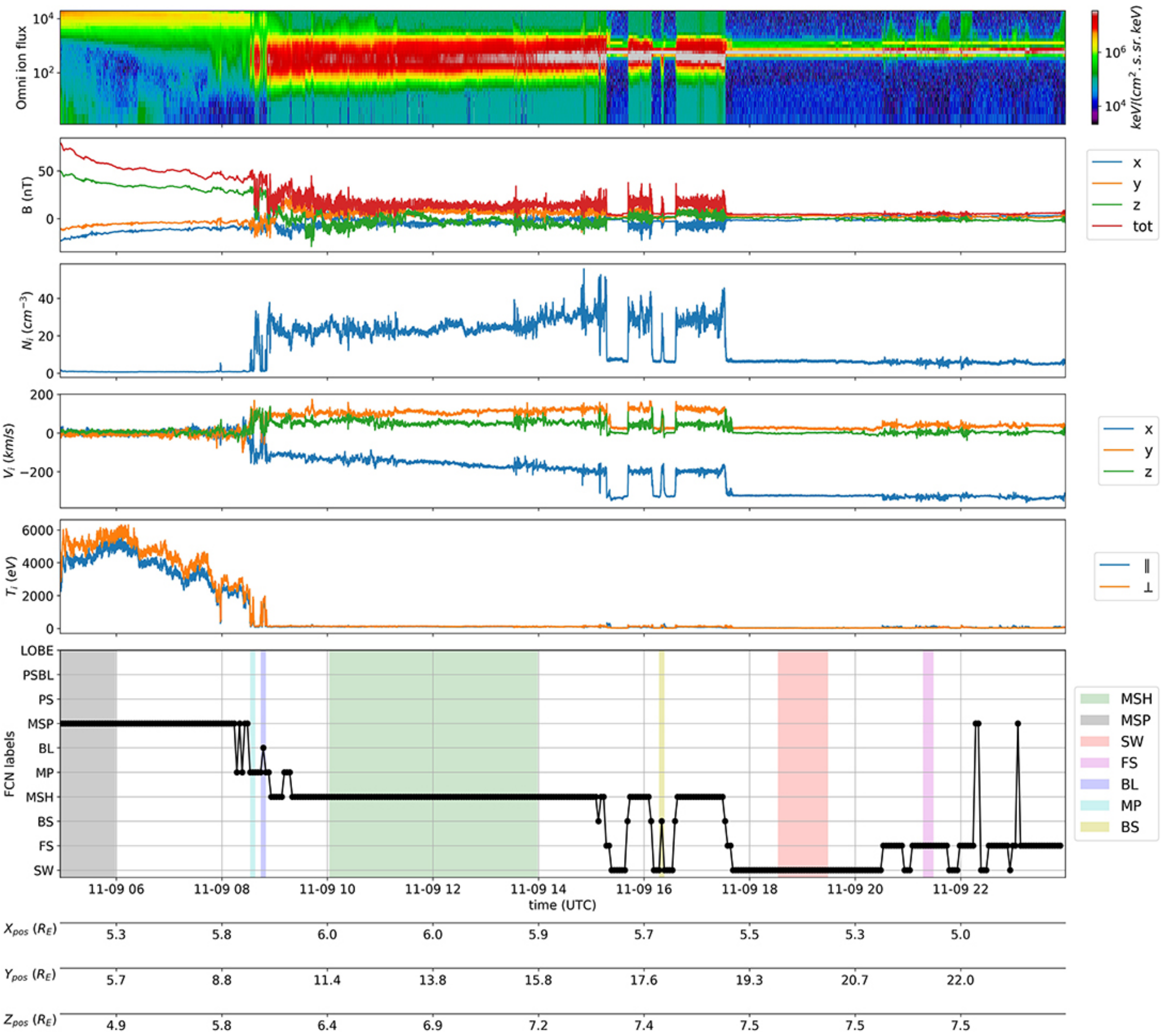


Figure 2.18: Figure showing an example of plasma region classification using the trained CNN model on the MMS data. The black line shows the classification of the model while the colored sections show the human classification that would have been done on this time series. Image taken from *Breuillard et al.* [2020]

Chapter 3

Automatic identification of EDRs at the Earth's magnetopause

3.1 Introduction

As previously mentioned in section 2.4, the electron diffusion region (EDR) is the heart of the magnetic reconnection process. The challenge of its detection lies in the fact that these regions are only a few kilometers wide, or to put it differently, of the order of the electron inertial length, which means that a spacecraft passing through an EDR will usually stay less than half a second in it. In the past, plasma physics missions like CLUSTER did not have the instrumental resolution to correctly identify EDRs. The electron diffusion region could only really be studied using plasma simulations until the launch of the MMS mission in 2015.

MMS was the first mission with the instrumental resolution necessary to observe and study these regions and a year after its launch the first detection of an EDR was reported in *Burch and Phan* [2016]. As of 2018, only 32 EDRs had been reported on the dayside in various papers and they were all compiled in *Webster et al.* [2018]. The scarce amount of EDR detections despite the instrumental resolution of MMS can be explained by the fact that each detection has to be made by human observation of the data which can be hard and biased. This was what lead to the idea of using machine learning to automatically detect new EDR cases in the MMS data. But before going into the details of the handling of the data and the conception of the algorithm, I will briefly review the physical characteristics and signatures of the EDR that were presented in section 2.4.

The EDR is the heart of the reconnection where the magnetic energy gets converted into kinetic and thermal energy. Combining Vlasov and Maxwell's equations, we get that the term $\mathbf{J} \cdot \mathbf{E}$ is equal to the variation of electromagnetic energy, meaning that one of the key signatures of the EDR is to have a high energy conversion term $\mathbf{J} \cdot \mathbf{E}$. Another EDR signature comes from the breaking of the ideal MHD condition $\mathbf{E'} = \mathbf{E} + \mathbf{v}_e \times \mathbf{B} = 0$. Since the measurement of the electric field with MMS is one of the most challenging, incertitude on the measurements make that $\mathbf{E'}$ is not always exactly null outside the EDR, this is why we like to look at places where **E'** is as high as possible to differentiate this parameter from ”noise”. Due to the complex magnetic configuration and the gyration of the electrons at the EDR, one can observe crescent-shaped electron velocity distribution functions when the guide field is low enough. These crescents are of great value to identify EDRs and thus constitute another key signature of EDRs. It is important to state that crescents can also form at other places than at the EDR so the more key signatures are observed, the more confident one can be about having crossed an EDR and only finding electron crescents is usually not enough.

3.2 Treatment of the MMS data

3.2.1 Handling of the data products

Every space mission is faced with telemetry constraints, and every mission has its own way of dealing with it. In the case of the MMS mission, data are divided in two categories, ”survey mode” and ”burst mode” for low-resolution data and high-resolution data. This distinction is necessary because the amount of high-resolution burst data recorded on board of the MMS spacecraft is such that only about 4% can be downlinked to Earth. Still in the case of the MMS mission, there is a burst management system that consists of the automated burst system (ABS) that selects burst intervals using trigger data numbers, of a human Scientist-in-the-Loop (SITL) who examines all of the low-resolution survey data, and who manually selects and classifies burst intervals, and of a ground loop system (GLS) which is a system of machine learning or empirical models that automate the event classification process using all of the data available to the SITL. These three systems allow for the selection of regions of interest that suit the MMS mission objectives with one of them being the potential identification of new EDRs.

For the following study, only burst data were used as it is the only way to have the necessary resolution to resolve electron dynamics inside the EDR. In burst data mode, MMS instruments have two different resolutions, one of 150ms for ion measurements and one of 30ms for electron measurements whereas the resolution is of 4.5s for electron measurements in survey mode for example. MMS also provides measurements with even higher resolutions for the electric and magnetic fields with respectively 0.01 ms and 0.96 ms. Since I planned to use both particles and fields measurements taken at different rates, the first part of the work was to upsample the ion measurements on the nominal FPI electron measurement cadence using linear interpolation and to downsample the magnetic field and electric field measurements also on the nominal FPI electron measurement cadence as well by taking the means of the measurements around the electron measurement times. These operations were done using the IRF matlab package [1] developed by a part of the MMS science team at Uppsala University in Sweden. The processed data was then written into .txt files in order to be read by the python algorithms used for the rest of the study.

---

[1] https://sites.google.com/site/irfumatlab/

The other very important part of the MMS data treatment is the handling of the electron velocity distribution functions provided by the FPI instrument on MMS. To produce the distribution function images used throughout this thesis, the integration of the full 3D distribution function is done over a +/-15^o angle above and below the $V_{\perp,1}$,$V_{\perp,2}$ plane. This integration angle was chosen as a good compromise between having enough data to spot EDR crescents that are expected to be seen in the $V_{\perp,1}$,$V_{\perp,2}$ plane and not drowning the crescent's information in too much noise when taking the average value of the energy bins of interest with that of bins located outside of the $V_{\perp,1}$,$V_{\perp,2}$ plane.

3.2.2 Creation of the MeanRL parameter

As said earlier in section 2.4, the electron velocity distribution function crescents are a big hint for the presence of an EDR. They are located in the $V_{\perp,1}$,$V_{\perp,2}$ plane in the magnetic field-aligned coordinate system, on the positive $V_{\perp,1}$ side of the electron velocity distribution function where $V_{||}$ is in the direction of the magnetic field, $V_{\perp,1}$ is in the direction of V_e and $V_{\perp,2}$ is along the direction of $V_{||} \times V_{\perp,1}$. The distribution function is a complex product provided by MMS so the idea was to summarize the asymmetry of the distribution function into a scalar parameter easier to handle.

This is why I came up with the idea of creating an adimensional scalar value that could be read as a time series which would characterize the presence of a crescent in the distribution functions. The parameter is called the "MeanRL" for "Mean Right/Left" because the idea is here to take the mean value of the bins from the right side of the image and to divide them by the mean values of the bins on the left side, yielding high values where there is an asymmetry in the image characterized by a crescent on the right.

Since the flux levels are very different from low to high energy electron bins (with higher fluxes in the core of the distribution), the first step is to remove the energy dependence of the phase space densities, each bin of the distribution function f is first normalized by the mean of the bins at the same energy in the phase space :

$$f'(\mathcal{E}_\perp) = \frac{f(\mathcal{E}_\perp)}{\left(\sum_i^n f_i(\mathcal{E}_i = \mathcal{E}_\perp)\right)/n} \tag{3.1}$$

where n is the total number of bins with the energy $\epsilon = \epsilon_\perp$, $\mathcal{E}_\perp = \frac{1}{2}m_e(V_{\perp,1}^2 + V_{\perp,2}^2)$ is the perpendicular energy with m_e the mass of the electron. This action also enhances the contrast between the two sides of the image in case of an asymmetry in the distribution function. Then, the MeanRL is computed with the following formula :

$$MeanRL = \frac{\sum\limits_i f_i'(V_{\perp,1} > 0, \mathcal{E}_1 < \mathcal{E}_\perp < \mathcal{E}_2)}{\sum\limits_i f_i'(V_{\perp,1} < 0, \mathcal{E}_1 < \mathcal{E}_\perp < \mathcal{E}_2)} \tag{3.2}$$

where $\mathcal{E}_1 = 40$ eV and $\mathcal{E}_2 = 275$ eV since EDR crescents are not expected outside of this energy range. The result of these transformations can be seen on Figure 3.1. The range of the MeanRL parameter is between 0 and 1 if there is a crescent shape on the left and above 1 if there is a crescent shape on the right.

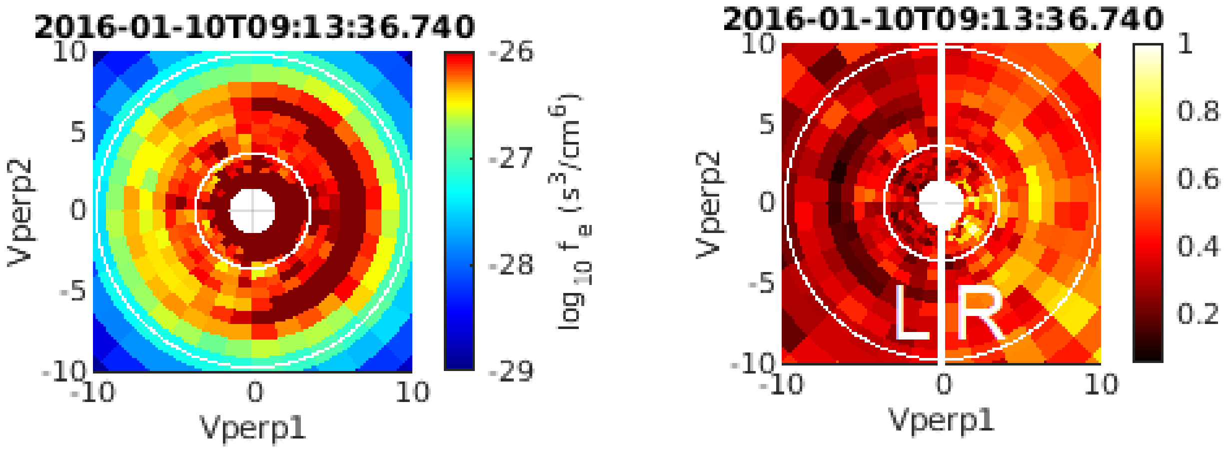


Figure 3.1: Left: Electron velocity distribution function (f) taken inside an EDR and presenting a crescent on the positive $V_{\perp,1}$ side in the $V_{\perp,1}$,$V_{\perp,2}$ plane. Right: Normalization for each energy ring of the electron velocity distribution function (f') on the left to compute the MeanRL scalar. $V_{\perp,1}$ and $V_{\perp,2}$ are in 10^3 km/s. The two white circles delimit the range of energies [40 eV, 275 eV].

In the context of looking for EDRs that show exclusively crescents on positive $V_{\perp,1}$ side of electron velocity distribution functions, values over 2 and beyond are usually expected based on detailed analysis of the EDR cases reported in the literature (see section 3.4). The highest observed values yet go up to around 8. Sometimes, when dealing with distribution functions displaying high fluxes, the MeanRL may be high even though no crescent is present but a small asymmetry. Indeed with this computation method, phase space elements with high phase space densities have more weight than elements with smaller phase space densities giving rise to some false positives. This is why an approach directly using the image of the distribution function coupled with the use of a CNN is a logical evolution of this idea and is presented in chapter 5

This parameter can be computed for each distribution function which are provided by the FPI instrument at the electron measurement cadence of 30ms. This means that no further resampling was needed and that this parameter could easily be included in the .txt files produced by matlab with the other parameters necessary to start the creation of a training dataset for the automatic identification of EDRs.

3.3 The automatic detection algorithm

3.3.1 Description of the algorithm

The problem of identifying EDRs with machine learning brings a number of challenges, the first one being the training of the algorithm due to the scarce amount of available data. The 32 previously reported cases from *Webster et al.* [2018] provide a small amount of data points (200 samples) that can be used to train the algorithm to recognize EDRs compared to the amount of data generally used in other applications in the same field (see *Nguyen et al.* [2019] and *Breuillard et al.* [2020]). The second challenge is the feature selection which is the selection of the best physical parameters to provide to the algorithm due to our lack of complete understanding of the microphysics inside EDRs.

For building the algorithm, I chose to identify Electron Diffusion Regions as well as 3 other classes: Separatrix and Ion Diffusion Regions (IDR), Magnetosphere regions (MSP), and Magnetosheath and Boundary Layer regions (MBL). I chose this approach to be able to better contextualize the EDR detections regarding the location of the other plasma regions. Although the detection of EDRs was our main objective, studying and trying to detect these other plasma regions was of interest to see how the physical parameters describing the EDR compared to those describing these other regions. Separatrices and Ion Diffusion Regions are grouped together because they exhibit similar features such as high $\mathbf{E} + \mathbf{v}_i \times \mathbf{B}$ values and non zero $\mathbf{J} \cdot \mathbf{E}$ values due to strong currents, they are thus often difficult to distinguish from each other. For similar reasons, I also grouped magnetosheath and boundary layer. Figure 3.2 shows a flowchart describing the whole process of the algorithm that I will now describe.

3.3.2 Creating a training database

I used the list of 32 EDR events from *Webster et al.* [2018] to build the training database. I manually labeled data points in 80s burst data intervals taken around the reported EDR events, and an example of how the labeling was done can be found in figure 3.3. This length of interval was chosen to be able to see the whole context of the event and possible transitions from the magnetosphere to the IDR and to the EDR. As can be seen on this figure, the labeling was done using the old labels before regrouping some of the classes together. Some of the reported events were not included into the training database, based on visual inspection of the electron distribution functions, namely events 13, 15, 22, 23, 24 and 30, since deciding which data points could be considered as actually belonging to the EDR and not to other classes was ambiguous for these events by looking at each electron distribution function around the reported times. Treating the 4 spacecraft independently, I gathered 200 individual EDR data points for the EDR training database. For the other classes (IDR, Magnetosphere and Boundary Layer) the number of data points per class for the training database are respectively 6809, 5334 and 18334, taken from the multiple 80s burst intervals.

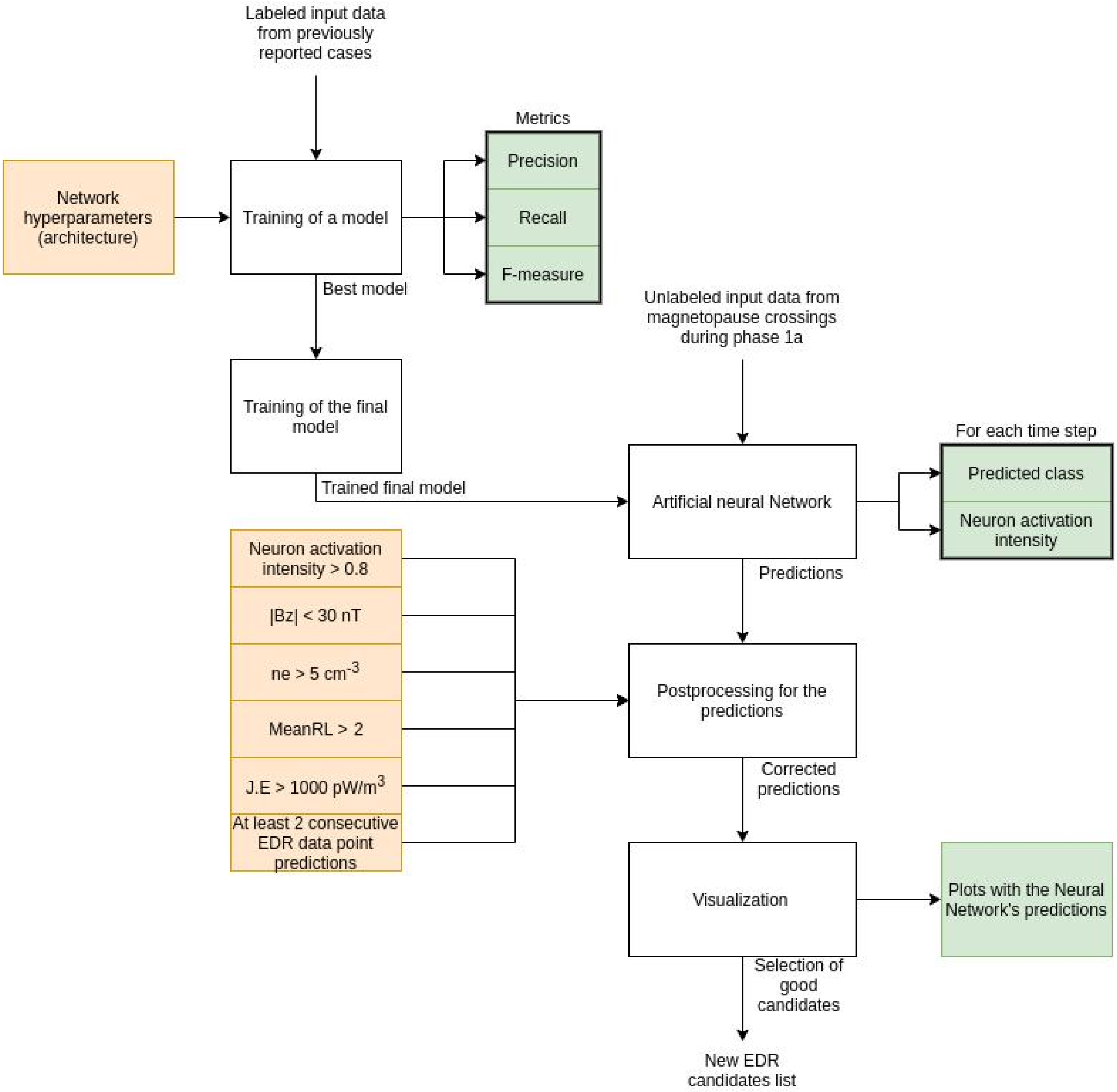


Figure 3.2: Flowchart of the processes of the algorithm. Orange rectangles represent parameters that could be tuned by the user and green rectangles represent outputs of the process it is linked to.

I then chose the relevant physical parameters to feed the algorithm with, which are referred to as "features". The selection of the features was made to reflect our current understanding of EDRs while still being able to identify the other plasma regions of interest. Thus I chose features characterizing the magnetic field, the generalized Ohm's law and the presence of electron physics including crescents in the velocity distribution functions. The features that were provided to the algorithm are the following :

- The magnetic field (FGM) : B_z, $|\mathbf{B}|$
- The ion velocity (FPI) : V_{ix}, V_{iy}, V_{iz}

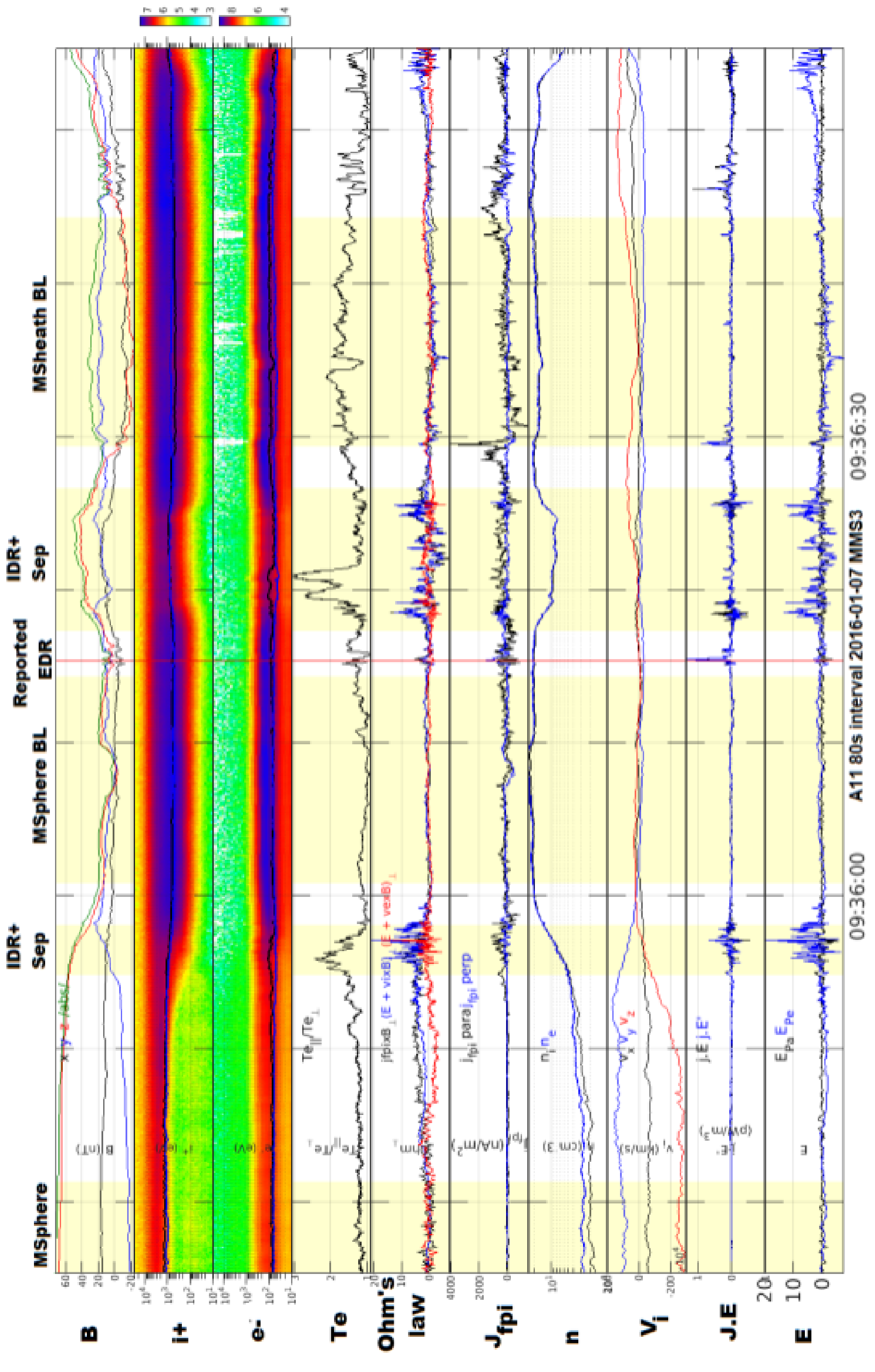


Figure 3.3: Figure showing the manual labeling process using an event where all the different classes were visible.

- The electron density (FPI) : n_e
- The electric field (EDP) : $E_{||}$, $E_\perp$
- The electric current (computed from FGM and FPI data) : $J_{||}$, $J_\perp$
- The terms of the generalized Ohm's law (computed from FPI data) : $(\frac{\mathbf{J}\times\mathbf{B}}{en})_\perp$, $(\mathbf{E}+\mathbf{V_e}\times\mathbf{B})_\perp$, $(\mathbf{E}+\mathbf{V_i}\times\mathbf{B})_\perp$
- The energy dissipation (computed from FPI and EDP data) : $\mathbf{J}\cdot\mathbf{E}$
- The electron temperature (FPI) : $Te_{||}$, $Te_\perp$
- The MeanRL parameter

The x,y,z coordinates correspond to GSE coordinates while the parallel and perpendicular coordinates correspond to field aligned coordinates $(||, \perp_1, \perp_2)$. Field aligned coordinates for a vector $\mathbf{V}$ are obtained by first taking the dot product of the vector $\mathbf{V}$ and the magnetic field vector $\mathbf{B}$ to get $V_{||}$. The $V_{\perp_1}$ coordinate can be obtained by taking the dot product of $\mathbf{V}$ and $\mathbf{B}\times[1,0,0]$. The final coordinate $V_{\perp_2}$ is obtained by taking the dot product of $\mathbf{V}$ and the vector $\mathbf{B}\times(\mathbf{B}\times[1,0,0])$. The perpendicular coordinate given here being the mean of the values of the 2 perpendicular coordinates $\perp_1$ and $\perp_2$. The electric current $\mathbf{J}$ is computed from particle measurements (FPI) using a single spacecraft.

Figure 3.4 shows boxplots of the distributions of different physical parameters for each plasma region I manually labeled during phase 1 which constitutes the training database. The boxplot graphical representation gives information on the dispersion of the data within a class. From these boxplots, we can see that most of the parameter distributions do not allow a simple threshold approach as for each parameter, at least 2 classes usually have close ranges of values. The absence of simple thresholds for each class (in particular for the IDR and EDR) is one of the main reasons for choosing a neural network approach, which by contrast allows for the building of a complex relationship between several physical parameters.

Nevertheless, the different regimes and characteristics expected for the different classes are found in the boxplots: IDRs exhibit a wide range of values for the terms in the Ohm's law, magnetosphere regions are characterized by low particle density (Figure 3.4b) and large magnetic field (Figure 3.4a), whereas Boundary Layer regions show a wide range of values for the intensity of the magnetic field (Bz, Figure 3.4a), and large electron density (Figure 3.4b), but small energy dissipation since it is not the place where the exchange of energy occurs ($\mathbf{J}\cdot\mathbf{E}$, Figure 3.4c).

The EDR class is characterized, as expected, by a strong dissipation ($\mathbf{J}\cdot\mathbf{E}$), non ideal Ohm's law components and relatively low magnetic field intensities. I initially considered the agyrotropic index $\sqrt{Q}$ introduced by *Swisdak* [2016] (see also *Aunai et al.* [2013]) which gives a measure of the nongyrotropy of a particle distribution based on the measured pressure tensor. However, we can see on figure 3.4 that it is inefficient to identify EDRs, their values for the EDR class are indeed very close to the values of other classes and in particular magnetospheric like regions (Figure 3.4g). This can be explained by the fact that when the

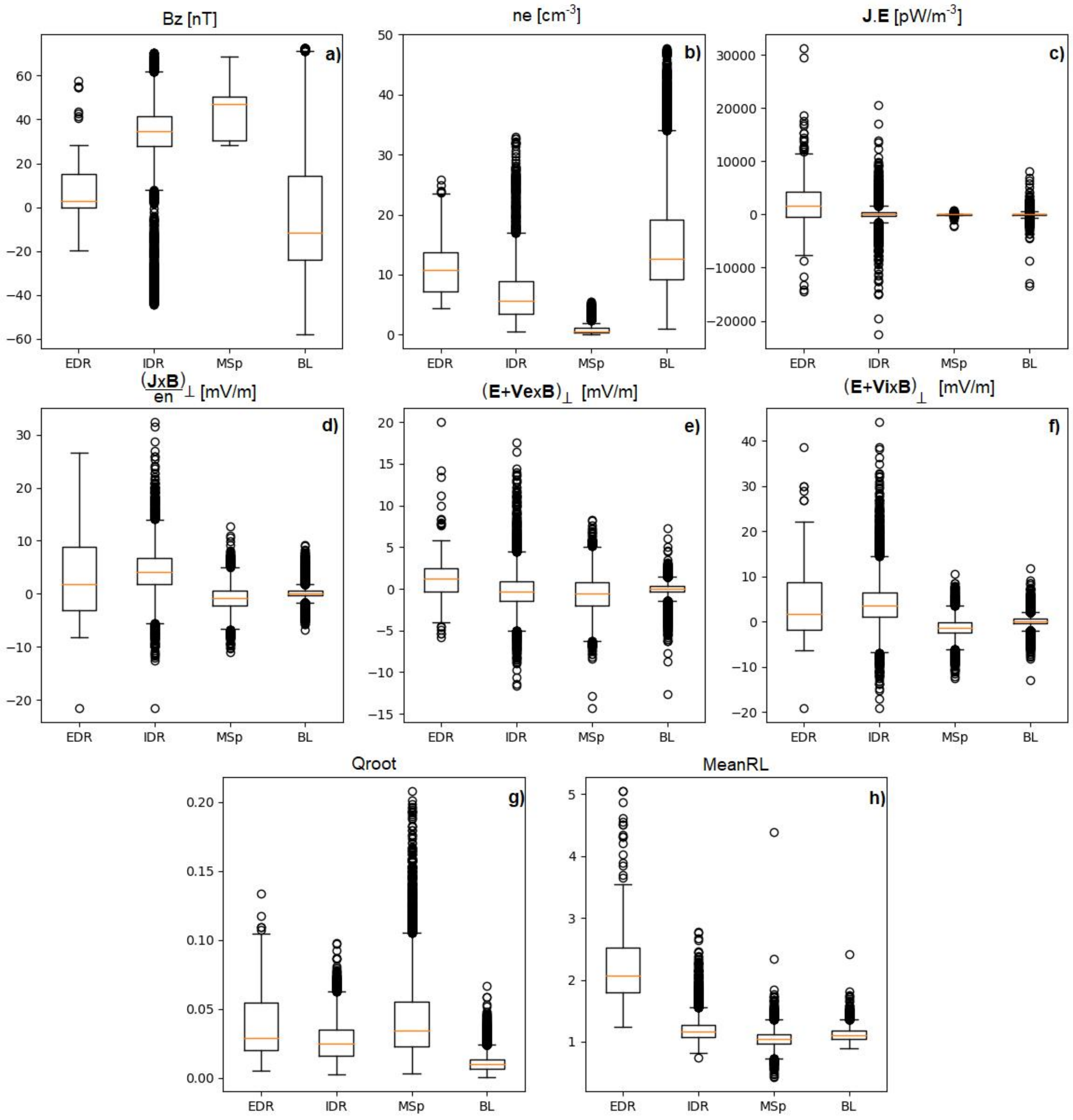


Figure 3.4: Boxplot distributions of the values of different features for manually labeled data points during phase 1 for each class. The 4 classes are: Electron Diffusion Region (EDR), Ion Diffusion Region + Separatrix (IDR), Magnetosphere region (MSp) and Boundary Layer region (BL). They respectively contain 200, 6809, 5334 and 18334 points. Each rectangle contains 50% of the values of the distribution and are delimited by Q_1 and Q_3 being the first and third quartiles, the yellow lines represent the median of the distribution. Black circles are outliers of the distribution, they are defined by being below the threshold $Q_1 - 1.5 \times (Q_3 - Q_1)$ or above the threshold $Q_3 + 1.5 \times (Q_3 - Q_1)$.

particle counts are low (below 5 cm^{-3}) like in the magnetosphere, the off-diagonal components of the pressure tensor in a field aligned basis are less reliable (close to noise level), which frequently results in large and nonphysical values for this parameter. That is the reason why I chose not to include the agyrotropy index as a feature for the algorithm. On the other hand, the MeanRL parameter that I defined is performing much better to separate the different classes (Figure 3.4h), the range of values for the EDR class being larger than that of all the other classes.

The last modification I performed to the database was to remove outlier points. Since the data points were all manually labeled and since it is usually hard to clearly identify which data points constitute the boundaries between the different plasma regions, it is possible that some of the manually labeled data points were mislabeled. So, based on boxplots of all features for each class, I chose to remove in all the classes except for the EDR class the data points that were outliers of the distribution, defined by being below the threshold $Q_1 - 1.5 \times (Q_3 - Q_1)$ or above the threshold $Q_3 + 1.5 \times (Q_3 - Q_1)$ for at least one feature, with Q_1 and Q_3 being the first and third quartiles. No EDR points were removed because it is already the smallest class of the dataset. The number of data points for each class after removing outliers is 5037 (initially 6809) for the IDR class, 4201 (initially 5334) for the Magnetosphere class, 12870 (initially 18334) for the Boundary Layer class, and 200 for the EDR class. All the resampled values for the features of the remaining manually labeled data constitutes the training database.

As discussed in the machine learning introduction, it is common practice to split the training database into a training, a validation and a test set. The first one is the set from which the algorithm is going to learn, the second one is used to control the and detect potential overfitting during the learning of the algorithm and is never used to train the algorithm, and the last one is used to assess the performances of the algorithm before using it on unseen data. Due to the great imbalance in the number of data points available for each class, I used a stratified splitting which keeps the proportion of the classes in each set. The repartition of the training database data points in each set is 60% for the training set, 20% for the validation set and 20% for the test set. Thus for the EDR class, 60% of all available data points go into the training set and the remaining 40% are equally split between the validation and the test sets. At this point, I standardized the data for each feature, which is a transformation yielding a mean of zero and a standard deviation of 1 for the distribution, in order to keep a priori the same importance for all features.

3.3.3 Training and evaluation of the model

The performance of the trained algorithm can be assessed by looking at different metrics that are chosen depending on the problem at hand and computed for each class. I chose to look at 3 metrics being the Precision, the Recall and the F1-score. The Precision tells which proportion of positive identifications were effectively correct while the Recall tells which proportion of positive results were correctly identified. A low Precision means that the algorithm will tend to overestimate the number of instances of the class and a low Recall means that the algorithm will overlook a lot of instances of the class. Usually, increasing one of these metrics means the decrease of the other so depending on the application, one can prefer to either balance these two metrics or maximize one at the expense of the other. For our goal of finding new EDRs, a high Recall was necessary in order not to miss too many cases so a lower Precision was tolerated. The F-measure metric is the harmonic mean of the Precision and the Recall and is used to see the balance between these two metrics. The formulas of these metrics are the following :

$$Precision = \frac{TP}{TP + FP} \tag{3.3}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.4}$$

$$F1{-}score = 2 \times \frac{(Pre \times Rec)}{Pre + Rec} \tag{3.5}$$

where TP, FP and FN respectively stand for number of True Positives, False Positives and False Negatives, and Pre and Rec respectively refer to Precision and Recall.

Using the training and validation sets, I then experimented to find the neural network architecture that would give the best metrics, being in our case the highest possible Recall with still a relatively high Precision to limit the number of false positives.

The algorithm used is a classical feed-forward Multilayer Perceptron. It was developed using Keras/Tensorflow in Python and trained using a ”categorical crossentropy” type loss function with the ”Adam” optimizer (*Kingma and Ba* [2014]). I also used a number of Epochs of 60 and a batch size of 128. The best combination of hyperparameters has been determined through trial and error combined with Bayesian optimization techniques (*Wu et al.* [2019]), though our tests suggested that the hyperparameter space of the algorithm was rather flat as different combinations of hyperparameters were giving results with similar metrics in the end.

The architecture of the final model (which is the model trained using all of the available data) is an input layer of 17 neurons, then a first hidden layer of 17 neurons with a ”ReLU” activation function followed by a second hidden layer of 33 neurons with a ”tanh” activation function, and finally an output layer of 4 neurons corresponding to the 4 classes to identify with a ”softmax” activation function which is the standard practice for classification tasks in machine learning.

As said in the beginning of this section, the scarce number of EDR data points was problematic for the training of the algorithm, as the number of data points between the four classes was very unbalanced. To tackle this problem, during the training I used weights to give more importance to the smallest classes. These weights are computed by attributing to each point the inverse of the apparition frequency of the class it belongs to. The results of the training can be found in Table 3.1. The algorithm performs very well for IDR, MSp and MBL (F1-score $> 97\%$). For EDR, the metrics also indicate satisfactory results (F1-score

Table 3.1: Results of the predictions of the algorithm on the test set after the training

Class	Test data points	Precision	Recall	F1-score
EDR	40	83%	95%	88%
IDR + Separatrix Region	1008	97%	97%	97%
Magnetosphere	840	98%	100%	99%
Boundary Layer	2574	99%	99%	99%

$> 88\%$).

3.3.4 Post-processing and prediction on unlabeled magnetopause crossings data

The final model of the neural network is trained using all available data regardless of the sets established before. The input data is then replaced by unlabeled data points taken from September 2015 to March 2016 corresponding to phase 1a of MMS. I used the magnetopause crossings database from *Paschmann et al.* [2018] to reduce the MMS database to magnetopause crossings of phase 1a and reduce the overall time of computation. For each data point, the algorithm yields a predicted class and a neuron activation intensity (the highest output value among those of the neurons representing each class) going from 0 to 1 that can be interpreted as a level of confidence of the algorithm, giving lower values when the data point was harder to classify, probably because it did not resemble enough data points included in the training database.

The scarcity of EDR training examples was a challenge for this study as explained previously. The highest precision achieved for the algorithm is relatively high but when it is applied to the number of data points that the predictions are done on, it is normal to find a lot of potential false positives. This is why I chose some criteria of post-processing in order to reduce the number of potential EDR cases, leaving the predictions of the other plasma regions untouched. I chose to only consider the EDR predictions of data points that corresponded to the following criteria :

- Neuron activation intensity > 0.8 : to discard ambiguous predictions.
- $|B_z| < 30$ nT : the algorithm tends to mislabel some IDR data points as EDR and I found that it was particularly true for "Separatrix IDR" data points so removing the EDR detections with higher B_z would remove some of these cases. EDRs at the dayside magnetopause should exhibit small Bz values, since they are supposed to be the center of the current sheet magnetic field reversal.
- Electron density $n_e > 5$ cm^{-3} : the regions with small number density typically correspond to the magnetosphere. In these regions, electron distribution functions become less reliable owing to low count detections by FPI and therefore MeanRL becomes unreliable. Furthermore, the dayside magnetosphere is often populated by

cold (a few eV) plasma, which affects **E** field measurements (*Toledo-Redondo et al.* [2019]). Therefore, we exclude EDR selections featuring low electron density because of these potential issues.

- MeanRL > 2 : to increase the likelihood of having well identifiable crescents in the electron velocity distribution functions.
- $\mathbf{J} \cdot \mathbf{E} > 1000$ pW/m^3 : to avoid false positives that could also be present in magnetosphere regions.
- We require a minimum of 2 consecutive EDR points : the identification of only one data point as EDR is more likely to be a false positive resulting from time aliasing or spurious data points, and the detection of too many consecutive EDR points should not be possible given the speed of the spacecraft and the size of the EDR.

After the post-processing, less than 10% of the detected EDR cases remain. These post-processing criteria may remove some real EDR detections in the end, but they significantly reduce the number of false positives and permit manual checking of each remaining candidate. For example, a new EDR event recently found during phase 1a and detailed in *Zhong et al.* [2020] was detected by the algorithm but removed from the pool of EDR cases to manually check after the post-processing. The reason is that I fixed the MeanRL threshold at 2 to limit false positive detections but for this case, the data points labeled as EDR by the algorithm had MeanRL values just below 1.8, thus not making it past the post-processing step.

3.4 Overview of the new EDR candidates

In the rest of this thesis, I will qualify as "EDR candidates" the EDRs that were manually selected after being identified by my algorithms, that will also be the case for EDR identified in chapter 5. After the post-processing, each remaining EDR predictions in the span of a single burst data CDF file was given an ID and manually inspected. Out of the 137 manually checked events found during the phase 1a of MMS, 18 were selected as new EDR candidates based on visual inspection of crescents and physical parameters, which are presented in table 3.2. To our knowledge, these EDR crossings had not been previously reported. Candidates are separated into 2 labels being EDR and Outer EDR (OEDR) based on the average sign of $\mathbf{J} \cdot \mathbf{E}$ over the event (see the detailed explanation in next section).

On Figure 3.5 we can see the spatial spread of the new EDR candidates along the MMS orbit, as well as that of the 32 previously reported dayside EDR cases. We can observe that the distribution of the new EDR candidates is similar to that of previously reported cases during phase 1a but the detections of the algorithm seem to be more located near the flanks. This could be explained by the fact that MMS spacecraft tend to do relatively faster cuts through the magnetopause around the subsolar point than near the flanks where MMS spacecraft are more often skimming the magnetopause, increasing the time spent in regions likely to contain EDR signatures.

Table 3.2: List of the selected new EDR candidates found by the algorithm. Two times are given if the time of the EDR signature is different from one spacecraft to another. Values separated by a slash indicate that the feature rapidly changes from one value to another during the EDR signature. The ID column shows the start time of the burst CDF file containing the EDR candidate.

ID	EDR Time	MMS	Mean SC Separation (km)	J.E (nW/m3)	MeanRL	Label
20150909_084324	08:43:58	3	200	5	3	EDR
20150909_125814	13:00:22	4	154	1	3.5	EDR
20150909_142734	14:28:51	3	145	1	2.6	EDR
20150911_101214	10:13:35	3	178	-10/+2	5	EDR
20150914_161634	16:17:50	2	139	-10	3.5 / 2.75	OEDR
20150919_092544	09:26:27	3	62	7.5/4	3/1.8 ; 2.5	EDR
20150922_134024	13:41:31	3	51	-6	3.5	OEDR
20150923_090914	09:09:38	4	62	-6	4	OEDR
20151001_065214	06:53:43	2	51	-5	5	EDR
20151006_141714	14:17:54	1	26	1.8	3	EDR
20151202_011514	01:17:02	1	18	-4	3.2	OEDR
20160102_234614	23:46:17	4	41	-2.1	2.75	OEDR
20160107_221104(2)	22:12:20 ; 22:11:37	3,4	44	5 ; -10/+20	2.5 ; 2.75	EDR
20160205_221924	22:19:46	1	14	3	2.25	EDR
20160211_015924(2)	02:01:04	1,3	14	1.2 ; 1	6 ; 4.8	EDR
20160214_204124(3)	20:41:56	2,3,4	15	-4/+5 ; -1/+1 ; 10	4.75 ; 4.3 ; 5	EDR
20160219_183904(2)	18:42:38-39 ; 18:42:38-40	1,3	15	1.2/2.75 ; 1.1/1.5	3.1/2.5 ; 3/2.5	EDR
20160228_010604(2)	01:07:33	1,3	16	4 ; 7.5	5 ; 2.4	EDR

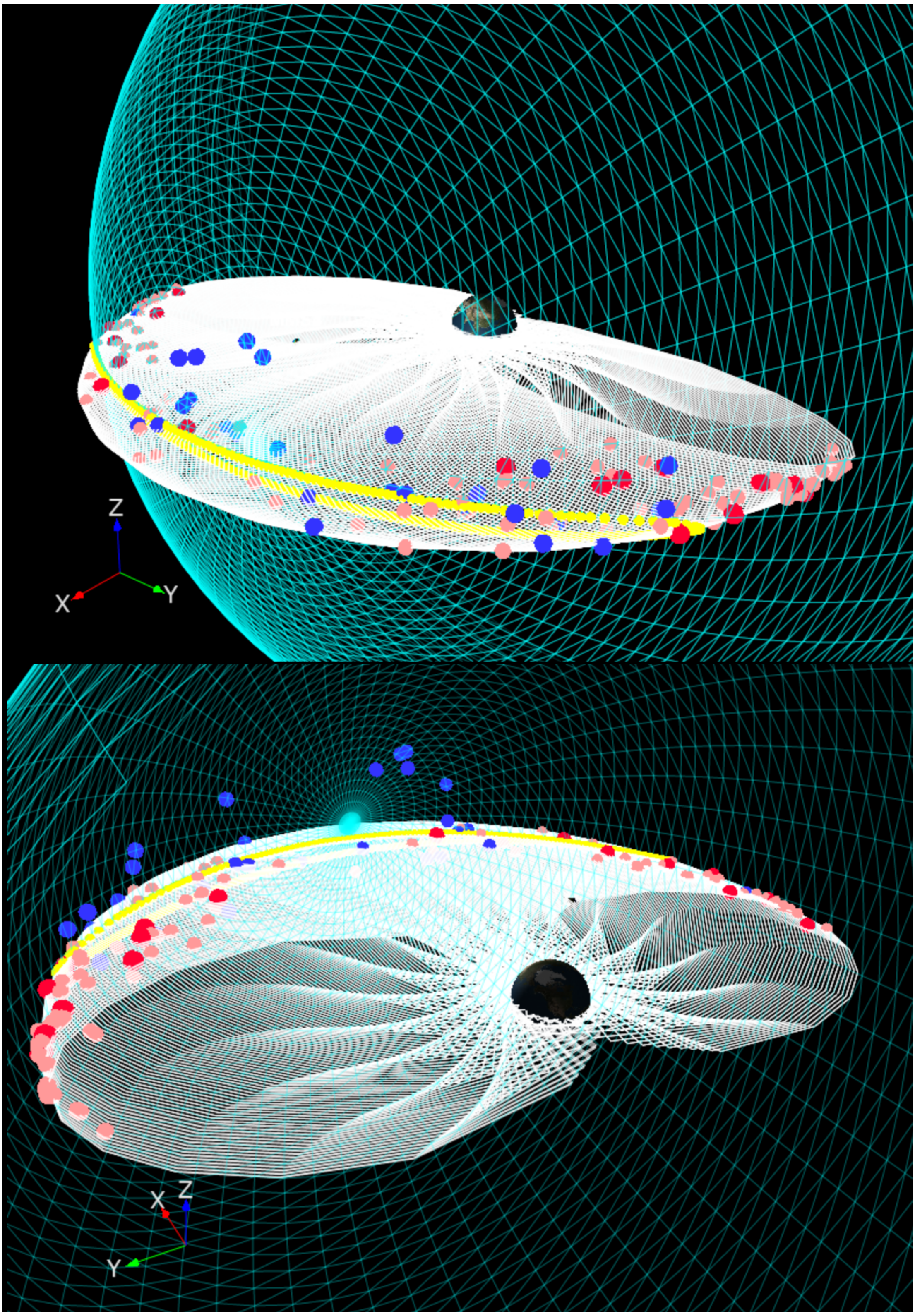


Figure 3.5: Spatial repartition of the selected new EDR candidates (red dots), of the other detections labeled as EDR by the algorithm during phase 1a (pink dots), as well as that of the 32 dayside EDR (blue dots) cases previously reported in *Webster et al.* [2018] using CDPP's 3Dview software (*Génot et al.* [2018]). Yellow points represent magnetopause crossings during phase 1a of MMS computed from the Shue model (*Shue et al.* [1997]) and represented by blue lines, while white lines show the trajectory of MMS during phase 1a.

Figure 3.6 shows a boxplot comparing the distribution of the values of the features of the previously reported EDR and of the new EDR candidates. We can observe that the values for the new EDR candidates are very different from that of other regions and very close to those for previously reported EDR cases, reinforcing the belief that the cases found by the algorithm and that I then manually selected are actual EDRs. Note that MeanRL values are all above 2 because of the post-processing which explains why the distribution is different from that of Webster's cases.

In addition to the 32 previously reported cases listed in *Webster et al.* [2018] and the more recent EDR candidate propositions (e.g. *Xu et al.* [2020]), the 18 new EDR candidates found during the phase 1a of MMS bring the number of identified EDR candidates during the phase 1 to over 50. This number is close to the 56 EDR encounters that were expected at the dayside magnetopause during the 2.5 years of the nominal mission of MMS (*Fuselier et al.* [2016]) based on the maximum shear reconnection model by *Trattner et al.* [2007a, b]. As new dayside EDR candidates found during phase 1 keep being reported, the number of EDRs may well exceed the number of expected EDRs. One possible explanation to this would be that the size of the EDR may be larger than expected, allowing MMS to do multiple crossings of the same EDR structures.

3.5 Discussion about the inner/outer EDR

The energy dissipation $\mathbf{J} \cdot \mathbf{E}$ is a parameter reflecting the energy conversion of the system, a positive value meaning that the magnetic energy is converted into kinetic and thermal energy and a negative value meaning that the energy is transferred from the particle to electromagnetic fields, with positive energy dissipation being the expected behavior at the EDR since the EDR is the place where the breaking of the magnetic field lines occurs, releasing their energy which accelerates surrounding particles. PIC simulations from *Cassak et al.* [2017] and *Swisdak et al.* [2018a] showed that regions of positive and negative energy dissipation both exist, showing the natural patchiness of the structure of the EDR. Following studies from *Burch et al.* [2018a] explained this structure by the combination of oscillating electric fields and oblique quasi-electrostatic whistler waves.

The classical picture of the electron diffusion region is that of an elongated box embedded inside the ion diffusion region. However, some studies (e.g. *Zenitani et al.* [2011], *Shay et al.* [2007], *Karimabadi et al.* [2007]) suggested that the EDR is rather a two-scale structure composed of an inner region where the usual signatures of the EDR are observed ($E' \neq 0$, $\mathbf{J} \cdot \mathbf{E} > 0$) and of an outer region where the energy conversion term $\mathbf{J} \cdot \mathbf{E}$ is negative. The outer EDR then acts as a generator region due the electron bulk flow slowing down there and which transfers the energy from the particles to the fields as opposed to in the inner EDR (see *Karimabadi et al.* [2007]) where kinetic energy dominates.

In the 32 previously reported EDR events, 24 events have $\mathbf{J} \cdot \mathbf{E} > 0$ (75%), 2 events

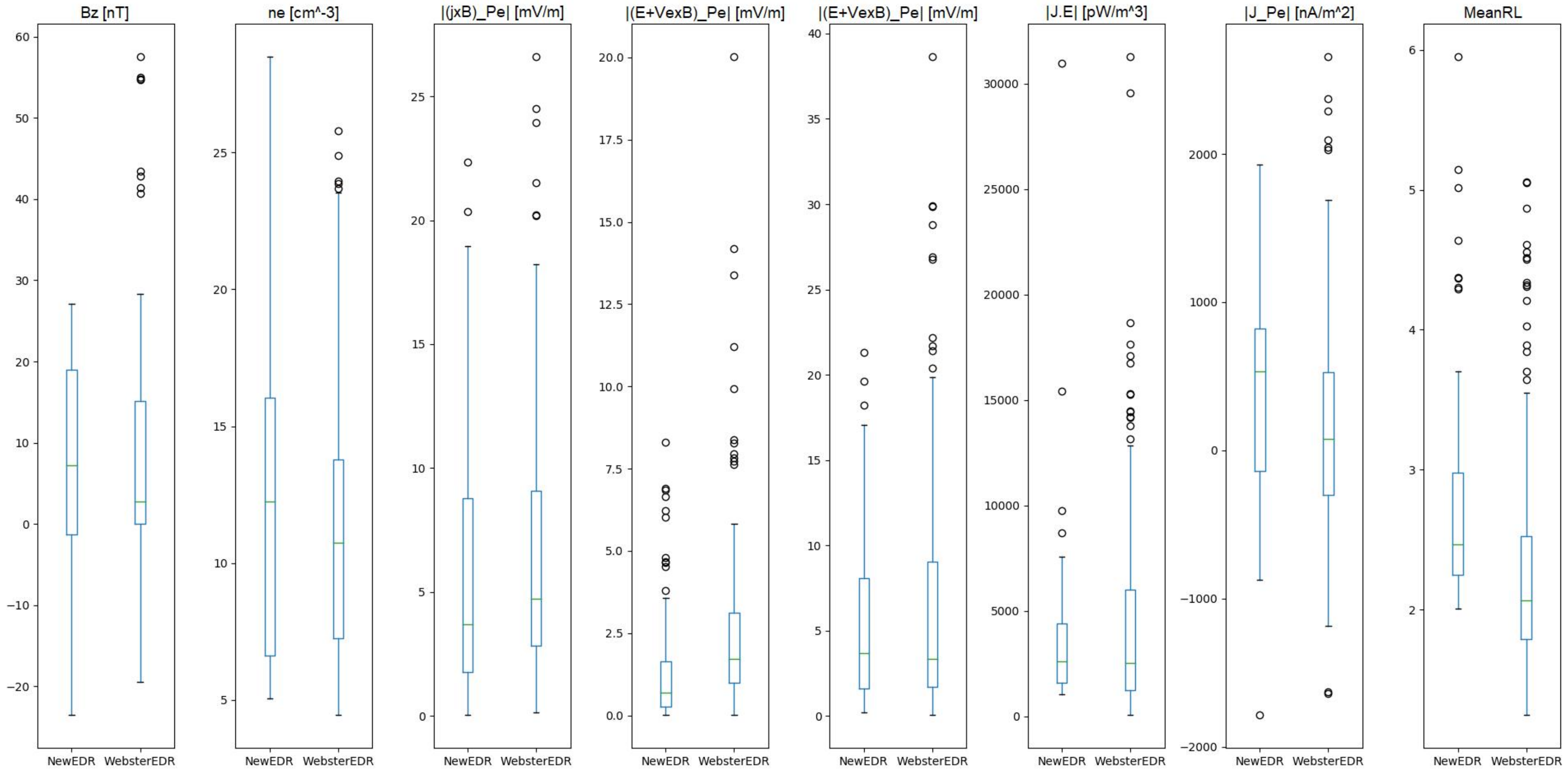


Figure 3.6: Boxplots showing for each physical parameter the distribution of the values for the 18 selected new EDR candidates and for the 32 dayside EDR events previously reported in *Webster et al.* [2018]. Each rectangle contains 50% of the values of the distribution, the line inside a rectangle represents the median of the distribution. Black circles are outliers of the distribution, they are defined by being below the threshold $Q_1 - 1.5 \times (Q_3 - Q_1)$ or above the threshold $Q_3 + 1.5 \times (Q_3 - Q_1)$ with Q_1 and Q_3 being the first and third quartiles.

(A08 and B32) have $\mathbf{J} \cdot \mathbf{E} < 0$ (6.25%) and 6 have an oscillating energy dissipation (18.75%). Treating each spacecraft separately, the complete list of 137 potential EDR candidates identified by the algorithm comprises 195 different magnetopause crossings. Among these, there are 112 crossings with $\mathbf{J} \cdot \mathbf{E} > 0$ (57.44%), 59 crossings with $\mathbf{J} \cdot \mathbf{E} < 0$ (30.25%) and 24 crossings with an oscillating $\mathbf{J} \cdot \mathbf{E}$ (12.31%), meaning the energy dissipation goes from positive to negative or from negative to positive during the same potential EDR crossing. If we only look at the 18 cases presented in this study which constitutes a number of 24 separate crossings, we get 14 crossings with $\mathbf{J} \cdot \mathbf{E} > 0$ (58.33%), 7 crossings with $\mathbf{J} \cdot \mathbf{E} < 0$ (29.17%) and 3 crossings with an oscillating energy dissipation (12.5%). These numbers can be compared to those presented by *Hamrin et al.* [2011] where they identified three times as many energy conversion load regions as energy conversion generator regions in the Cluster data at scales larger than the EDR. If we sum the previously reported events and the new events found by the algorithm that have $\mathbf{J} \cdot \mathbf{E} > 0$ and those that have $\mathbf{J} \cdot \mathbf{E} < 0$ (discarding the oscillating energy conversion cases), we get 4 times as many energy conversion load regions as energy conversion generator regions (38 load regions and 9 generator regions) which is still close to the results presented by *Hamrin et al.* [2011].

I attempted without success to link the positive and negative energy dissipations of our events to the outflow criterion for being in the inner EDR given by *Cozzani et al.* [2019] (which studies the B32 event from *Webster et al.* [2018]). The criterion is $(V_{A,i} < V_{e,L} < V_{A,e})$, with $V_{e,L}$ the L coordinate of the electron velocity, $V_{A,i}$ the Alfvén velocity of the ions and $V_{A,e}$ the Alfvén velocity of the electrons. The meeting of this condition means that the electron jet has not fully developed and thus that the spacecraft is located in the inner EDR. The main problem is that the electron speed in our cases is always far lower than the Alfvén velocity of the electrons and it is sometimes even lower than the Alfvén velocity of the ions, making it impossible to meet the criteria for the outer EDR. The determination of the inner/outer limit of the EDR is probably too complex to be assessed using a single criterion on the electron jet velocity. Geometric considerations (intrinsically 3D picture, guide field, ...) may be at work and cannot be captured efficiently by the localized nature of the electron jets.

In a recent paper, *Xiong et al.* [2022] suggested based on a 2.5D kinetic PIC simulation that the negative energy conversion is caused by the decrease of the electron bulk velocity in the outer EDR. According to this study, the newly reconnected magnetic field in the outer EDR induces an electric field that slows the electrons, the velocity gradient formed by this deceleration is thus responsible for the negative energy conversion there. They also show that the contribution of the electron inertial term from the generalized Ohm's law to the negative energy conversion dominates that of the pressure tensor term. However, they state that the contribution of the inertial term is connected to the mass ratio of the simulations and this may be the reason why *Genestreti et al.* [2018a] observed in the MMS data negative energy conversion in the inflow region with a dominance of the electron pressure tensor term. I investigated this question with three outer EDR cases with clear negative energy conversion : A08 from *Burch and Phan* [2016], B32 from *Webster et al.* [2018] and L08 found by the machine learning model (in table 5.6 from the end of chapter 5).

Figures 3.7, 3.8 and 3.9 are plots of three outer EDR events where one can see multiple

EDR key signatures such as a B_z gradient, non-zero E', large energy conversion (negative in these cases) as well as multiple very clear electron crescents (only one is shown in each figure). One can see on the last panel of figure 3.7 that the pressure term clearly dominates at the time of the negative energy conversion. On the last panel of figure 3.8, one can see that it is the inertial term that dominates at the time of the negative energy conversion, but at the time step before, it is the pressure term that dominates. On the last panel of figure 3.9, the first half of the time span during which **J** · **E'** is negative is dominated by the inertial term while the second half is dominated by the pressure term. This shows that reality is more complex than expected since these results agree with the results from both *Xiong et al.* [2022] and *Genestreti et al.* [2018a] depending on which event and time are analyzed. As mentioned in *Xiong et al.* [2022], there are thus possibly two different mechanisms to explain the formation of the negative energy conversion regions.

Conclusion

This first try at automatically identifying new EDR events by training a Multilayer Perceptron on the previously reported cases allowed for the discovery of 18 new EDR candidates, bringing the number of identified EDRs in the phase 1 of MMS from 36 to 54. The new identified EDRs are very similar to the 32 reported cases from *Webster et al.* [2018] which was expected as the model was taught what an EDR is using these previously reported cases. Despite the small number of training samples to identify EDRs, the ”simplicity” of this first algorithm allowed it to work with such a small training dataset. Many other EDR samples would indeed be necessary to implement bigger models such as LSTMs. One idea we had was to see if a link between real data and simulation data could be made with the goal to be able to train an algorithm using simulated EDR data and to make predictions on real data. As a first step, work was done by Ambre Ghisalberti during her internship at IRAP on the possible use of the model presented in this chapter to classify plasma regions in the data obtained from a magnetic reconnection simulation. Since it is not my work, as I only helped her, the work she achieved is presented in Appendix C. everal improvements could be made on the first model presented in this chapter obtain better results. Other approaches are possible to automatically identify new EDRs, one of them is by using a CNN which is presented in chapter 5. In the next chapter, I will present different statistical studies that make use of the high number of EDR cases that I now had at my disposal.

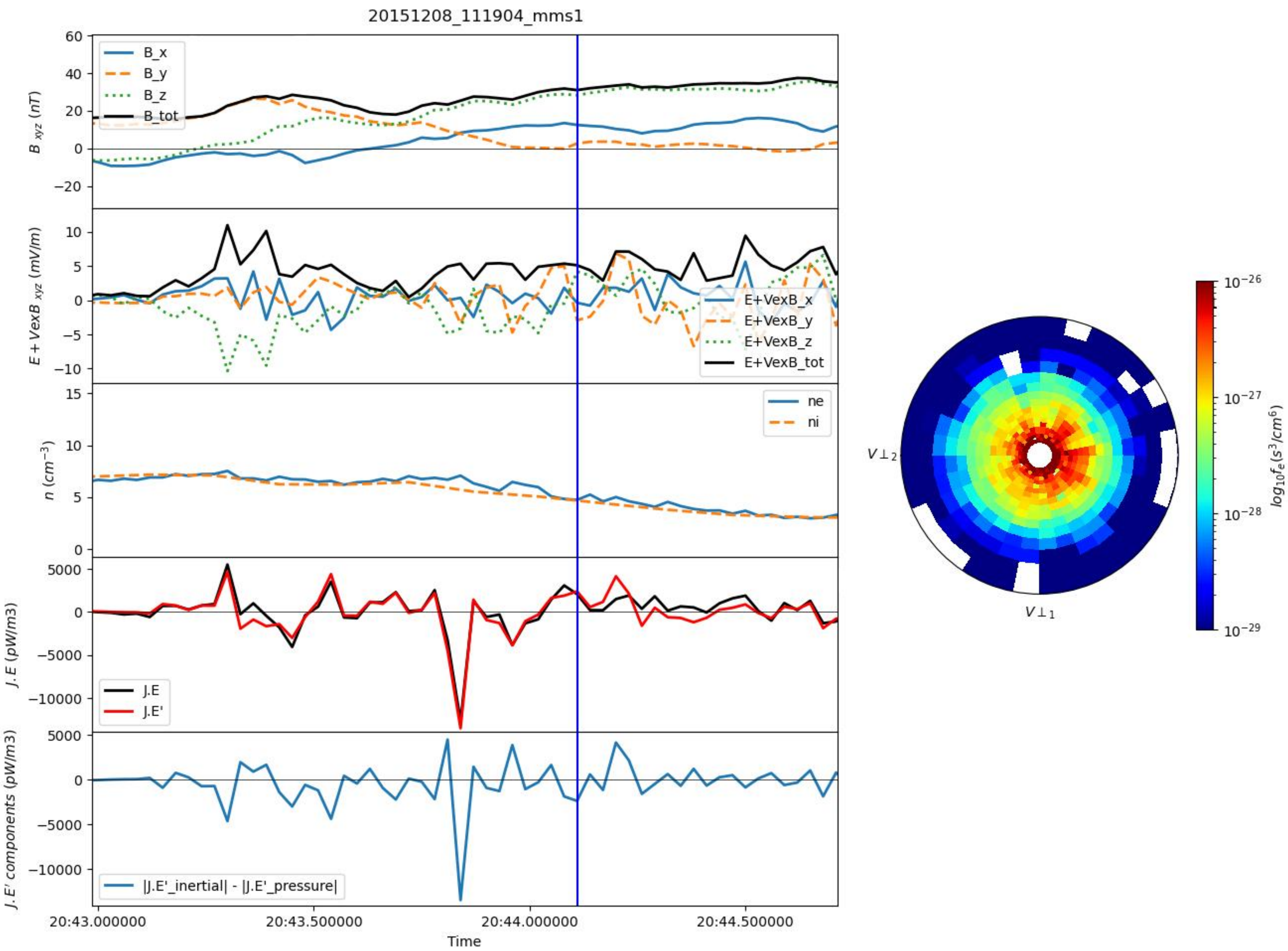


Figure 3.7: Figure showing the dominance of the pressure component over the pressure component of the energy conversion $\mathbf{J} \cdot \mathbf{E}'$ (last panel) for the EDR event A08 reported by *Burch and Phan* [2016].

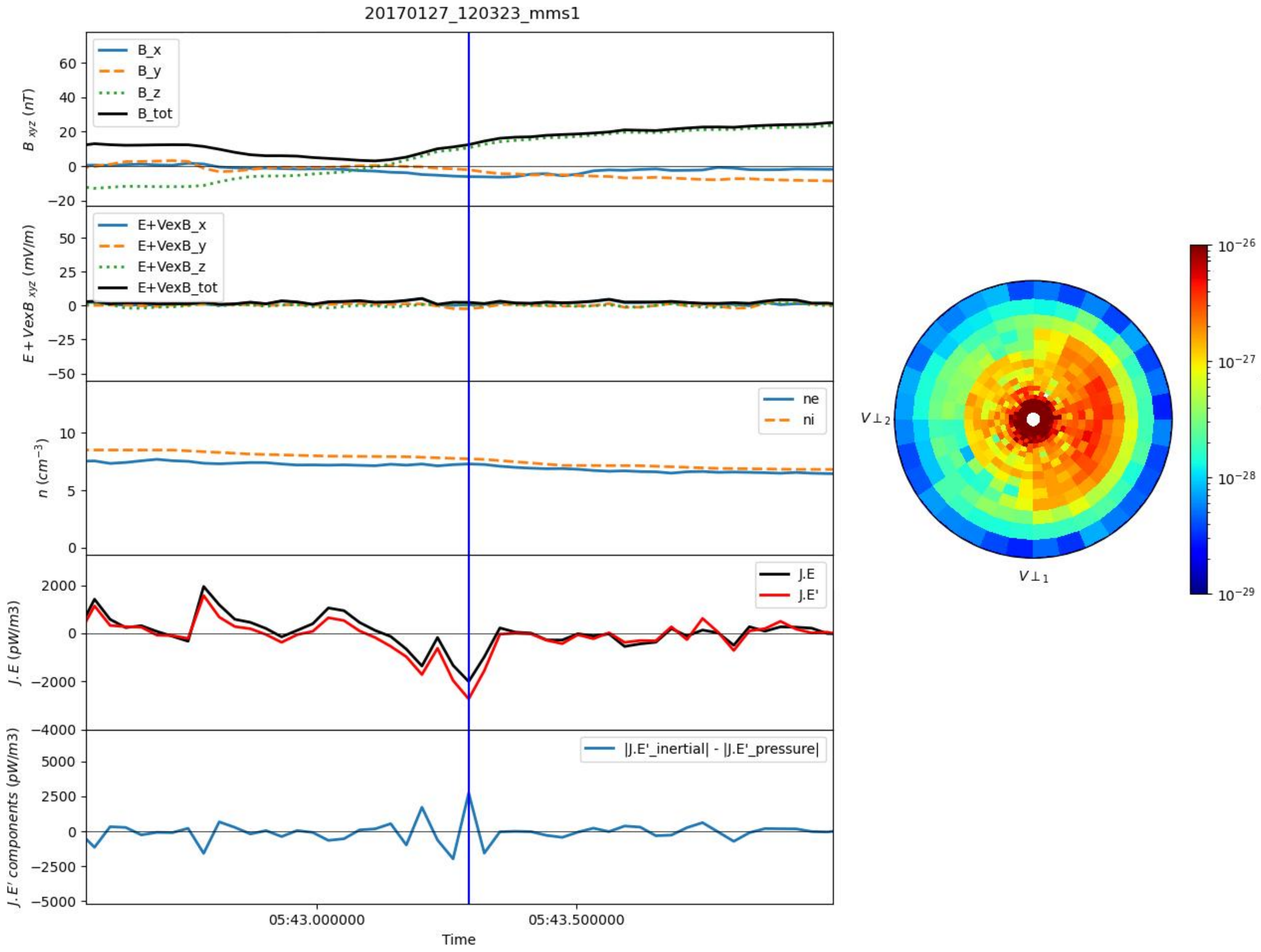


Figure 3.8: Figure showing the dominance of the inertial component over the pressure component of the energy conversion $\mathbf{J}\cdot\mathbf{E}'$ (last panel) for the EDR event B32 reported by *Webster et al.* [2018].

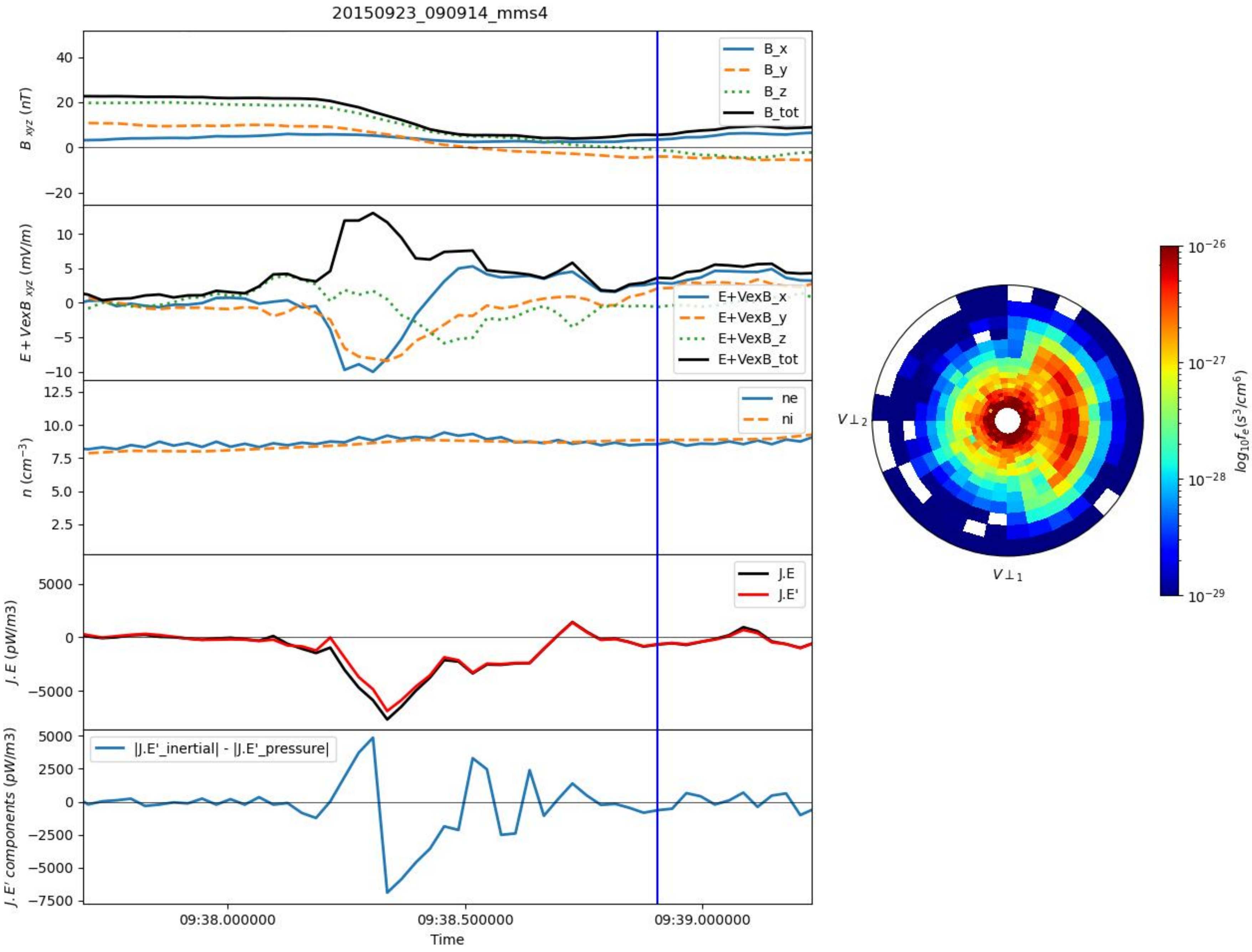


Figure 3.9: Figure showing the dominance of the pressure component over the pressure component of the energy conversion $\mathbf{J} \cdot \mathbf{E}'$ (last panel) for the EDR event L08 found by the machine learning model.

Chapter 4

Analysis of new EDR candidates

In this section, I present a series of detailed studies I have done after having found new EDRs with the help of the algorithm described in the previous section. In the literature, very few papers investigate the properties of the EDR region based on more than one specific EDR event (e.g. *Zhou et al.* [2021] and *Wang et al.* [2022]), and draw conclusions based on single event analysis. My aim in this section is to better understand the EDR based on a statistical approach, by analyzing various properties, such as : investigating a potential dependence with respect to solar wind parameters, characterizing the small scale parameters of the EDR, characterizing the structure of the EDR (and in particular the stagnation line) and studying the link between energy conversion and guide field. If no strong conclusion arose from these analysis, this clearly shows that the EDR structure and dynamics are more complex that expected from simple models and that considering a multiple event approach is a fruitful method to avoid drawing biased conclusions. It also shows that these results should be compared with advanced simulation results (e.g. *Malakit* [2012]) that can provide the full picture with however specific assumptions that do not necessarily cover the large range of conditions observed.

4.1 Potential link between EDRs and solar wind parameters

The solar wind and IMF parameters are known to influence the solar wind interaction at Earth and the magnetopause reconnection process in particular. Indeed, *Scurry et al.* [1994a, b] showed using a geomagnetic activity index as a proxy measure of the efficiency of dayside reconnection that the solar wind magnetosonic Mach number, upstream solar wind beta, and the cone angle of the interplanetary magnetic field control the plasma beta in the magnetosheath downstream of the Earth's bow shock. More recently, *Koga et al.* [2019] showed that the outflow speed and the reconnection electric field are affected by the solar wind-Alfvén speed and the solar wind-Alfvén Mach number. *Trattner et al.* [2016] found that the location of the magnetopause reconnection adjusts with changes in the solar wind, leading to the proposal of the Maximum Magnetic Shear Model (*Trattner et al.* [2021]).

It is thus of interest to investigate, based on our list of EDR events composed of previously reported EDRs and the new ones found in the phase 1a MMS data, to which extent some of the properties of the EDR depend on the solar wind parameters. The energy conversion ($\mathbf{J} \cdot \mathbf{E}$) that takes place at the EDR may in particular depend on the solar wind parameters, through a modification of the magnetosheath inflow conditions that are known e.g. to influence the total integrated energy conversion rate in Petscheck-like reconnection (see *Goodbred et al.* [2021]). I thus tried to link the solar wind parameters and the energy conversion intensity and sign. The solar wind parameters that were studied are the following :

- The IMF magnetic field components in GSE coordinates B_x IMF, B_y IMF, B_z IMF
- The Alfvénic Mach number $M_a = |V_r|/V_A$, with V_r being the radial velocity of the solar wind and V_A the Alfven velocity
- The solar wind density n_{SW}
- The dynamic pressure P_{dyn}
- The solar wind velocity V_{SW}
- The solar wind temperature T_{SW}

All these parameters were obtained for each EDR event reported in *Webster et al.* [2018] and *Lenouvel et al.* [2021] using data from the OMNI mission retrieved using AMDA[1] (see *Génot et al.* [2021]).

I first looked at the correlations between these parameters and the sign of the energy conversion. I split the EDRs of the list in three groups depending on whether the sign of the energy conversion is positive, negative or fluctuating. Figure 4.1 shows one of the density plots that I made for each solar wind parameter combination. No trend was observed for any solar wind parameter combination as can be seen on figure 4.1 hinting that there is no influence of the solar wind (both IMF intensity and solar wind velocity) on the energy conversion's sign at the EDR.

Then I wanted to look at the influence of the IMF clock angle on the intensity of the energy conversion. The clock angle determines the orientation of the IMF in the terminator plane of the planet, it is defined as the angle between the geomagnetic north and the projection of the IMF vector onto the GSM Y-Z plane ($= arctan(B_y, B_z)$). The IMF clock angle is known to play a key role in the efficiency of the reconnection process at the magnetopause, since it determines the orientation of the IMF and thus how and where the largest magnetic shear appears (*Trattner et al.* [2007a, b]). The clock angle has for example a significant impact on the tilt of the reconnection line at the magnetopause (see *Pulkkinen et al.* [2010]), as well as on the reconnection voltage (see *Hu et al.* [2009]) or on the type of reconnection and reconnection rate according to *Ouellette et al.* [2010]. It is thus justified to investigate the influence of the clock angle of the IMF on the EDR properties.

---

[1] http://amda.cdpp.eu/

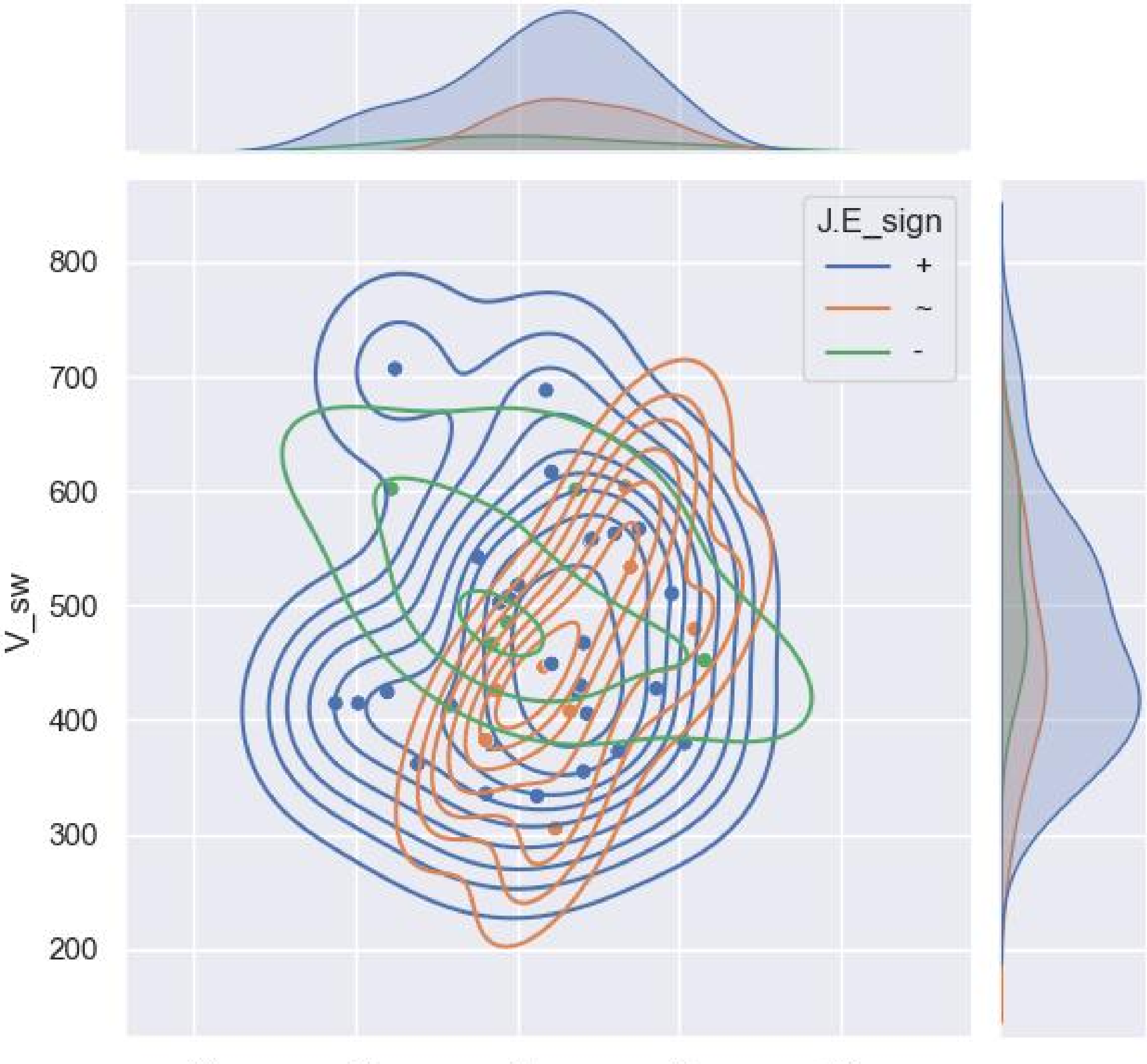


Figure 4.1: Example of 2D density map showing the distributions of EDR cases depending on their values of solar wind velocity V_{SW} and IMF B_y. The colors blue, orange and green represent respectively the positive, oscillating or negative sign of the energy conversion $\mathbf{J} \cdot \mathbf{E}$ for the related EDR case.

Figure 4.2 shows two angular plots : on the left there is the mean of the maximum $J \cdot E$ values for the EDR cases depending on the IMF clock angle, and on the right there is the distribution of IMF clock angle values among the EDRs. We see on the left plot that there is no particular tendency regarding the influence of the IMF clock angle on the intensity of the energy conversion. On the right plot, we can however see that 69% of the EDRs were found with an IMF clock angle in the two lower quadrants. This repartition tells us that EDRs are favored by southward pointing IMF clock angles, which most probably shows the preferred southward orientation of the IMF to induce a reconnection event with the northward magnetospheric magnetic field. This repartition is also the reason why looking at the mean of the maximum $J \cdot E$ values for the EDR cases depending on the IMF clock angle is inconclusive as the distribution of clock angle values is clearly not uniform.

Although we cannot link the intensity of the energy conversion and the IMF clock angle, there is an interesting trend regarding the value of B_y IMF and the intensity of the energy conversion. Figure 4.3 shows two 2D histograms of B_y IMF and the logarithm of $J \cdot E$ for the data points classified as EDR and IDR. What is interesting is that for the EDR class, the distribution of data points is not symmetric regarding the B_y IMF values as there is a clear skew towards positive B_y IMF values that is not observed for IDR class data points. While the IDR distribution is representative of the data coverage with a number of magnetopause crossings with a significant B_y component of the IMF (seen in the clock angle distribution), the EDR distribution shows few cases with negative B_y values.

This interesting behaviour, if confirmed in the future by a larger number of events (here we have only few cases per histogram bin), could be linked to the findings detailed in *Reistad et al.* [2020] in which they showed that there is an explicit dependence of the combination of the IMF clock angle and the IMF B_y on the reconnection rate. These authors indeed showed that when the Earth dipole tilt is negative, positive IMF B_y conditions are associated with a larger radius of the polar current systems than during negative IMF B_y conditions. Since the polar cap size is an indicator of the levels of dayside reconnection (see *Clausen et al.* [2013]), this suggests that the previously reported influence of IMF B_y on the ionospheric currents (see *Laundal et al.* [2018]) could be related to an external cause such as an asymmetry in the solar wind magnetosphere coupling affecting the dayside reconnection rate. According to *Reistad et al.* [2020], the underlying reason for the B_y dependence is unclear but could be the modification of the location of the X-line by a combination of the dipole tilt and IMF B_y towards the subsolar point where the reconnection is expected to be more efficient. In the future, we will however need to confirm the IMF B_y dependence of the energy conversion in the EDR with more events, and then investigate in more details the underlying mechanism.

4.2 Small scale parameters' analysis

For this study, I used the list of previously reported EDRs and the list of new EDRs found by the algorithm during phase 1a to create another database still containing the four classes from the original identification which are EDR, IDR, Magnetosphere side and Boundary Layer. The number of data samples for each class is respectfully 5 033, 131 014, 128

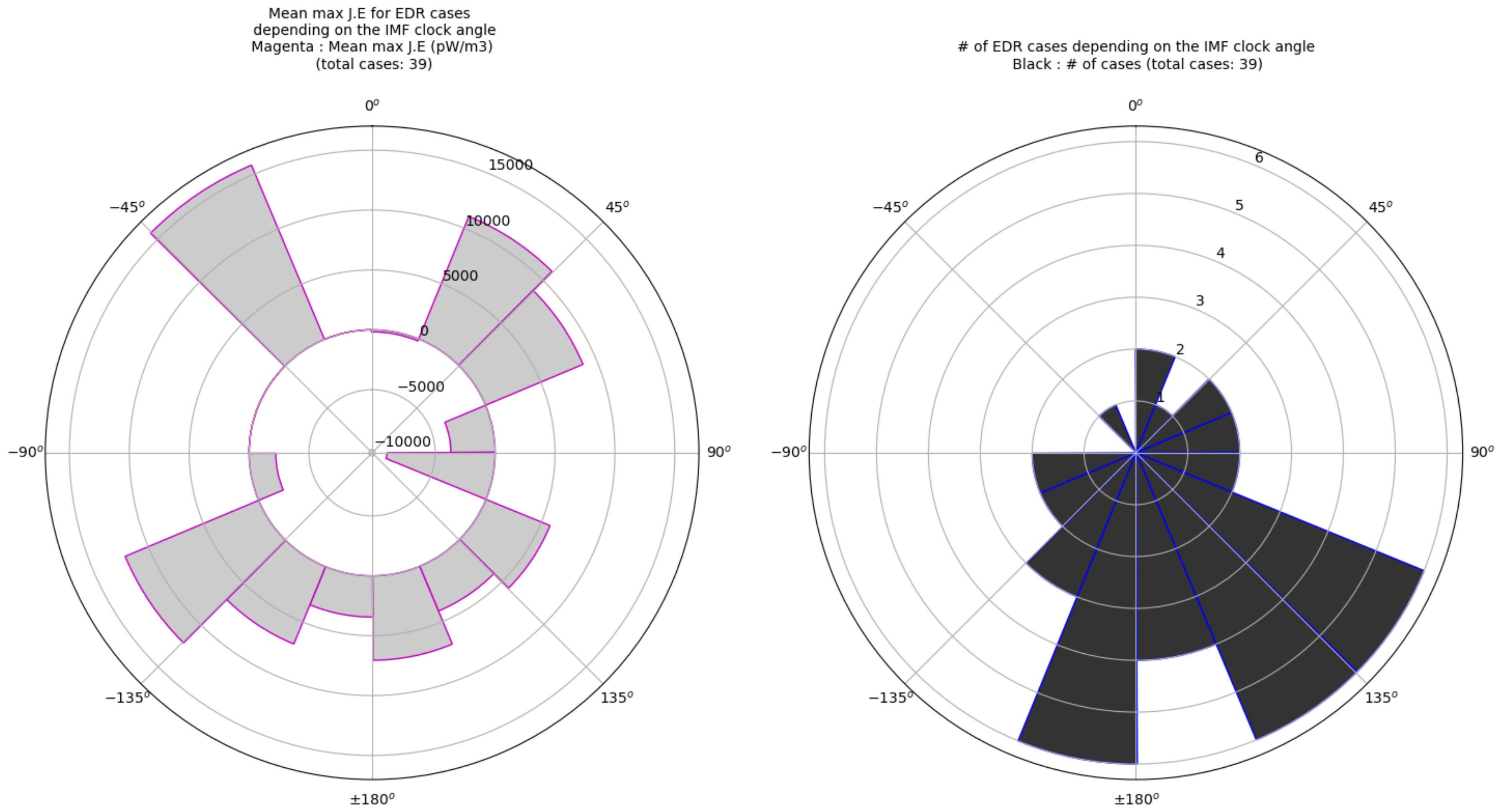


Figure 4.2: Figure showing two angular histograms representing the IMF clock angle. The plot on the left shows the mean of the maximum $J \cdot E$ values for the EDR cases depending on the IMF clock angle, and the plot on the right shows the distribution of IMF clock angle values among the EDRs.

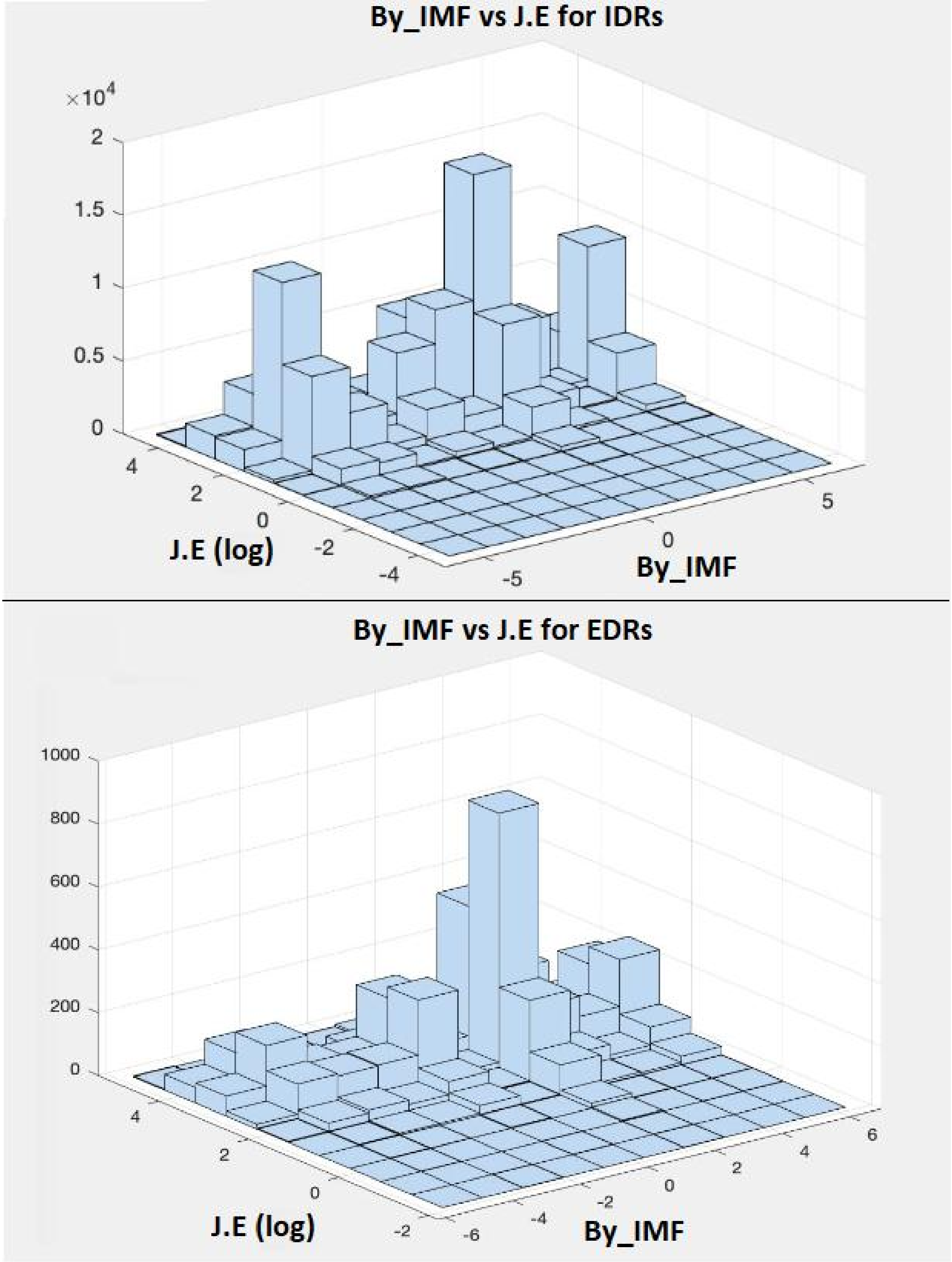


Figure 4.3: Figure showing two 2D histograms of the IMF B_y values versus the logarithm of the energy conversion $J \cdot E$ for the IDR (top) and for the EDR (bottom) classes.

970 and 386 219. The total number of EDR cases that were used in this statistical study is 46.

Once new EDR candidates were found, I thought it would be interesting to look for parameters that were not included in the first algorithm but that could be used to better identify EDRs in the time series. The EDR detections by previous authors or using my algorithm are based on few specific characteristics expected from earlier models and simulations, but the discovery of a significant number of EDR events allows to revisit the parameters used to detect this region. Moreover, during my thesis, several works based on a single case study proposed that specific parameters could be used to detect the EDR. We are thus able with our enlarged dataset to investigate this.

One of the reasons that made us want to do this study was the paper by *Hwang et al.* [2019] in which they propose to look at the electron vorticity $\boldsymbol{\Omega}_e = \boldsymbol{\nabla} \times V_e$ as a proxy to identify the edges of the EDR. The vorticity is a parameter that is classically used in hydrodynamics to investigate the vortices appearing in a fluid at small scale or in climate models. Plasma vortices also observed with Cluster (see *Alexandrova et al.* [2021]) may also appear in non-collisional magnetized astrophysical plasmas, in particular in planetary environments such as the plasma wake behind the planet Venus (see *Lundin et al.* [2013]). The vorticity may be used to detect such plasma vortices, that may be induced by the mutual interaction between the electric current and the magnetic field that produces a helical field configuration within a vortex (see *Durand-Manterola and Flandes* [2022]).

More precisely, *Hwang et al.* [2019] present an EDR in the magnetotail found on July 17^{th}, 2017 where the electron vorticity $\boldsymbol{\Omega}_e$ gets twice as high as the electron gyrofrequency $\omega_{ce} = \frac{eB}{m_e c}$. In theory, if the characteristic frequency of a certain plasma region (which would here be the electron vorticity $\boldsymbol{\Omega}_e$) is below the ion gyrofrequency $\boldsymbol{\Omega}_e < \omega_{ci}$, the system can be described by magnetohydrodynamics. If the characteristic frequency is larger than the ion gyrofrequency but lower than the electron gyrofrequency $\omega_{ci} < \boldsymbol{\Omega}_e < \omega_{ce}$, then the Hall physics present in the IDR governs the behavior of the plasma. And finally if the characteristic frequency is larger than the electron gyrofrequency $\boldsymbol{\Omega}_e > \omega_{ce}$, then the system enters into the kinetic regime that is dominant inside the EDR. Investigating when this condition is fulfilled could be a new EDR signature.

I managed to reproduce the results of *Hwang et al.* [2019] but I did not observe $\boldsymbol{\Omega}_e > \omega_{ce}$ in any of the 46 EDR events present in the list for the statistical study. In each case, the electron vorticity is at least two orders of magnitude lower than the electron gyrofrequency. Nevertheless, I made a boxplot of the ratio between the electron vorticity and the electron gyrofrequency for each of the 4 plasma regions to see if we could see some noticeable enhancement for the EDR class. This boxplot can be seen in the top left panel of figure 4.4. We can indeed see that the mean value of the $\boldsymbol{\Omega}_e / \omega_{ce}$ ratio is almost twice as high for the EDR class than for the other classes. The values are however quite small as the highest value observed is of about 0.025. The vorticity inequality proposed by *Hwang et al.* [2019] as a method to detect the EDR edges is interesting but only looking at when the $\boldsymbol{\Omega}_e > \omega_{ce}$ condition is met appears inappropriate based on a statistical basis. The event featured in their study is located in the magnetotail where the electron gyrofrequency is lower than at

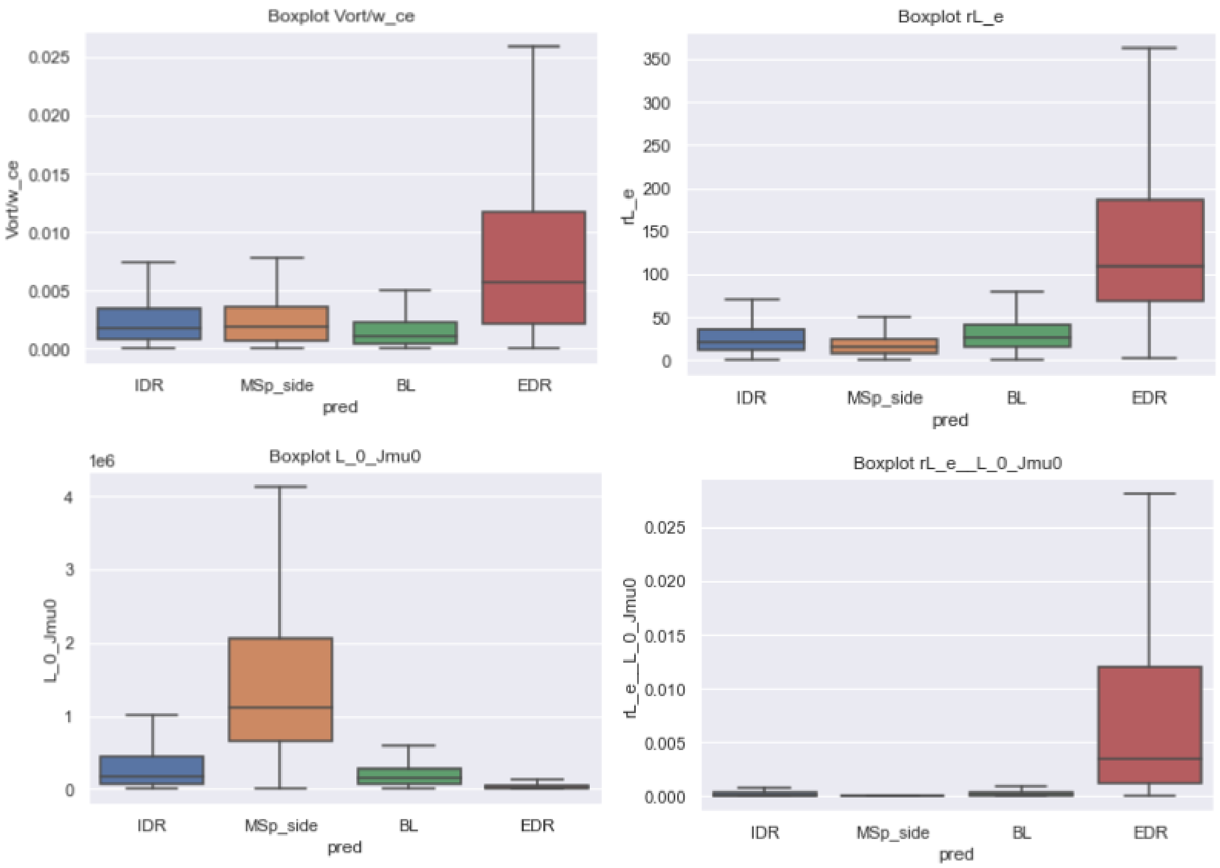


Figure 4.4: Boxplots of 4 different small scale parameters for 4 different plasma regions.

the magnetopause due to the lower magnetic field there so it is unlikely to meet the $\Omega_e > \omega_{ce}$ condition at the magnetopause. The Ω_e/ω_{ce} parameter could be used with precaution to strengthen EDR classifications but it cannot be considered as a key signature of the EDR.

Another parameter I looked at is the electron Larmor radius $r_L = \frac{m_e V_{e_\perp}}{eB}$. This parameter is supposed to be large when being close to the X-point and thus in the EDR because as the magnetic field intensity gets lower, the electrons will orbit further from the magnetic field lines meaning that their gyroradius will increase. A boxplot can be seen in the top right panel of figure 4.4. The mean value for the EDR class is above the highest value of any other plasma region so this parameter is quite interesting to look for EDRs. However, high values for this parameter for the EDR class are biased by the fact that most of the EDRs that have been found are low guide field cases with a low magnetic field. Events with high guide fields will produce low electron Larmor radius values as the electrons orbit closer to stronger magnetic field lines so looking for EDRs with high electron Larmor radius biases us to search for low guide field EDRs only.

I also looked at the magnetic gradient scale $L_0 = \frac{B}{\mu_0 J}$ with $\mu_0 J$ coming from the Maxwell-Ampère equation $\nabla \times \mathbf{B} = \mu_0 \mathbf{J} + \frac{1}{c^2}\frac{\partial \mathbf{E}}{\partial t}$ with no electric field, which tells us the scale of variation of the magnetic field. This parameter is supposed to be small when crossing a structure with strong magnetic gradient, typically at the electron scale like the EDR, due to the presence of a reduced magnetic field with strong currents, so it could be a good indicator of the presence of an EDR. This boxplot can be seen in the bottom left panel of figure 4.4. We see that the parameter is very low for the EDR class and is the largest for the Magnetospheric side class.

The last boxplot at the bottom right of figure 4.4 is the ratio of the Larmor radius and the magnetic gradient scale. This ratio does not have a physical meaning beyond the presence of currents and of reduced magnetic fields due to the proximity to the X-line, but has the advantage of giving very high values for the EDR class and very low values for the other classes. 67% of the EDR cases chosen for the study show a significant peak of the Larmor radius and of the magnetic gradient scale right at the reported time of the EDR, and the ratio of these parameters shows an even clearer peak at this time. This combination of the two parameters can thus be a good parameter to look at for future wide EDR search in time series data, but it still has to be combined with other parameters as it is not a decisive parameter. Moreover, this parameter is strongly biased towards EDR events without guide field, which limits the ability of such a parameter to search for events with an existing guide field.

There is one last parameter that was not included in the boxplot which is the adiabatic parameter κ introduced by *Lavraud et al.* [2016] and defined as $\kappa^2 = R_c/R_g$ with R_c the local magnetic field curvature and R_g the particle gyroradius. The local magnetic field curvature can be written as follows :

$$\frac{1}{R_c} = \left|\frac{\mathbf{B}}{|B|} \cdot \nabla \frac{\mathbf{B}}{|B|}\right| \tag{4.1}$$

Theory predicts that when κ^2 approaches 25, particle scattering occurs and particle dynamics becomes chaotic for values below 10. At the EDR, low values around 10 are expected as the motion of the electrons is supposed to be chaotic due to their demagnetization. After watching the behavior of this parameter for each EDR case studied here, only 4 EDRs exhibit κ^2 values below or near 10 and 6 EDRs have values close to 25. In about 34% of the EDR cases, the κ^2 parameter is the lowest of the whole burst mode duration at the EDR and in 26% of the cases, this parameter is just locally low but lowest values exist at other times. It is an interesting parameter regarding the physical insight it provides but it is not reliable enough to be a key EDR signature, although it can be used as a confirmation element if other key EDR signatures are associated with a low κ^2 value.

The analysis of the boxplots of specific magnetic and velocity parameters in the different regions show that none of them can be used to detect EDRs via a simple threshold approach, which underlines the importance to analyze several parameters by eye or through an automatic approach such as machine learning. Moreover, the magnetic derived parameters would further bias the detection towards non (or low) guide field reconnection events.

4.3 Geometry of the EDR structure

4.3.1 Location of some points of interest of the EDR relative to the energy conversion region

The spatial distribution of the energy conversion during a reconnection event and its dependence on guide field is poorly understood, and was debated by several authors. Previous works mentioned that energy conversion can happen not only at the X-line but also in the outflow region where the magnetic flux pile up occurs (see *Huang et al.* [2015]). *Yi et al.* [2019] found that the energy conversion rate peaks significantly later than the reconnection rate does, because the energy conversion primarily occurs at the reconnection fronts (the leading boundary of the reconnection outflow characterized by the enhancement of the reconnected magnetic field component) rather than at the X-line. *Xu et al.* [2020] suggested that the guide field may play an important role in modifying the location where the energy conversion occurs together with the electric field near the X-line, since the magnetized electrons along the guide field could be drifted by the force of $E_N \times B_g$ and accelerated by the different electric field systems generated by local waves. On the other side, the conjoined study by *Genestreti et al.* [2017] and *Cassak et al.* [2017] revealed a specific behavior of the energy conversion depending on the guide field based on few EDR events, as also discussed in *Burch et al.* [2018b]. *Genestreti et al.* [2017] indeed determined the location where the energy conversion is positive for 11 asymmetric EDR cases and found that as the guide field increases, the location of the energy conversion region gets closer to the X-line. The presence of a guide field allows for the presence of electrons at the X-line which will enhance the energy conversion at this location and decrease the energy near the electron crescent point which is the location of the electron velocity distribution function crescents. *Cassak et al.* [2017] confirmed this using 2D PIC simulations of three MMS events.

Based on our larger list of EDR events, I thus attempted to confirm or infirm the conclusions of these studies, i.e the location of the maximum energy conversion rate with respect to structural points of the reconnection region, at least for the low guide field regime where most of our EDR cases happen. Out of the 50 events from *Webster et al.* [2018] and *Lenouvel et al.* [2021], 11 were discarded of the study due to the impossibility to provide a reliable LMN coordinate system, giving a final list of 39 EDR events listed in Table 4.1. The LMN coordinates were all manually found using Minimum Variance Analysis (MVA) except for events A02 *Norgren et al.* [2016], A04 *Phan et al.* [2016], A07 *Khotyaintsev et al.* [2016], A08 *Burch and Phan* [2016] and A10 *Chen et al.* [2017] were the LMN coordinates were taken directly from the corresponding papers. The MVA technique is explained in Appendix B.

Figures 4.5 and 4.6 show scatterplots of the temporal and spatial separation between the spot of maximum energy conversion and multiple spots of interest of the reconnection respectively as a function of the guide field and of the density asymmetry. The spots of interests are the $B_L = 0$ point, the electron crescent point and the estimated stagnation point (see figure 4.7 which is a reproduction of figure 2.5 from chapter 2 here for the context). The time separations "DeltaT" are also normalized in ion inertial lengths to avoid

Table 4.1: List of EDR events composing the database used for the study and the associated LMN coordinates computed using MVA in the GSE coordinates. The ”Event ID” column shows the date and the UTC time at which the EDR crossing can be found, the time being given to the nearest second. The ”LMN coordinates” column shows the LMN vectors found using MVA in the GSE coordinates and used through the rest of the study for each event.

Event ID	MMS	AKA	LMN coordinates		
2015-09-09T13:00:22	4	L02	L=[0.13,0.25,0.96]	M=[0.93,-0.37,-0.03]	N=[0.35,0.89,-0.28]
2015-09-09T14:28:51	3	L03	L=[0.25,0.33,0.91]	M=[0.67,-0.73,0.09]	N=[0.70,0.59,-0.41]
2015-09-11T10:13:35	3	L04	L=[0.47,-0.09,0.88]	M=[0.81,-0.35,-0.47]	N=[0.35,0.93,-0.09]
2015-09-19T09:26:27	3	L06	L=[0.01,0.07,1]	M=[0.86,-0.51,0.03]	N=[0.51,0.86,-0.07]
2015-09-23T09:09:38	4	L08	L=[-0.12,0.70,0.70]	M=[0.64,-0.49,0.60]	N=[0.76,0.52,-0.39]
2015-10-01T06:53:43	2	L09	L=[0.53,-0.10,0.93]	M=[-0.11,0.98,0.14]	N=[-0.93,-0.15,0.34]
2015-10-06T14:17:54	1	L10	L=[0.54,-0.03,0.84]	M=[0.30,0.94,-0.16]	N=[-0.79,0.33,0.52]
2015-12-02T01:17:02	1	L11	L=[-0.11,-0.76,0.64]	M=[-0.43,0.62,0.66]	N=[-0.89,-0.21,-0.40]
2016-01-02T23:46:17	4	L12	L=[0.38,-0.24,0.89]	M=[0.33,0.94,0.11]	N=[-0.86,0.25,0.44]
2016-01-07T22:11:37	4	L13	L=[-0.31,-0.59,0.74]	M=[0.66,0.43,0.62]	N=[-0.68,0.68,0.26]
2016-02-05T22:19:46	1	L14	L=[0.32,-0.21,0.92]	M=[0.83,0.53,-0.17]	N=[-0.45,0.82,0.34]
2016-02-11T02:01:04	1	L15	L=[-0.07,-0.62,0.78]	M=[1,-0.04,0.06]	N=[0,0.78,0.62]
2016-02-14T20:41:56	2	L16	L=[0.06,-0.01,1]	M=[0.29,0.96,-0.01]	N=[-0.95,0.29,0.06]
2016-02-19T18:42:38	1	L17	L=[-0.60,-0.62,0.51]	M=[0.75,-0.21,0.63]	N=[-0.28,0.76,0.59]
2016-02-28T01:07:33	3	L18	L=[-0.1,-0.51,0.86]	M=[0.61,0.65,0.46]	N=[-0.78,0.57,0.25]
2015-10-16T10:33:30	1	A02	L=[0.14,0.17,0.98]	M=[0.33,-0.94,0.12]	N=[0.94,0.3,-0.18]
2015-10-16T13:07:02	3	A03	L=[0.35,0.15,0.93]	M=[0.69,-0.71,-0.15]	N=[0.64,0.69,-0.35]
2015-10-22T06:05:21	1	A04	L=[0.29,0.11,0.95]	M=[0.17,-0.98,0.07]	N=[0.94,0.14,-0.3]
2015-11-12T07:19:21	1	A06	L=[0.39,-0.52,0.76]	M=[-0.91,-0.1,0.4]	N=[-0.13,-0.85,-0.51]
2015-12-06T23:38:31	2	A07	L=[0.03,-0.5,0.86]	M=[0.56,0.72,0.4]	N=[-0.83,0.47,0.3]
2015-12-08T11:20:44	1	A08	L=[0.46,-0.41,0.76]	M=[-0.14,0.84,0.52]	N=[-0.88,-0.35,0.32]
2015-12-09T01:06:11	4	A09	L=[-0.01,-0.51,0.86]	M=[0.89,0.38,0.24]	N=[-0.45,0.77,0.45]
2015-12-14T01:17:39	1	A10	L=[0.27,-0.54,0.79]	M=[-0.41,-0.82,-0.41]	N=[0.87,-0.21,-0.44]
2016-01-10T09:13:37	4	A12	L=[-0.27,-0.65,0.71]	M=[0.74,0.33,0.58]	N=[-0.61,0.69,0.39]
2016-02-07T20:23:34	1	A13	L=[0.09,-0.60,0.79]	M=[-0.87,0.35,0.36]	N=[-0.49,-0.72,-0.49]
2016-10-22T12:58:40	2	B14	L=[0.1,0.28,0.95]	M=[0.63,-0.76,0.16]	N=[0.77,0.59,-0.25]
2016-11-02T14:46:18	4	B15	L=[-0.18,0.5,0.85]	M=[-0.73,0.51,-0.45]	N=[-0.66,-0.7,0.28]
2016-11-06T08:40:58	2	B16	L=[0.39,0.03,0.92]	M=[0.38,-0.92,-0.13]	N=[0.84,0.4,-0.37]
2016-11-12T17:48:46	2	B17	L=[-0.24,0.19,0.95]	M=[-0.86,0.42,-0.30]	N=[-0.46,-0.89,0.06]
2016-11-13T09:10:40	4	B18	L=[0.27,0.48,0.83]	M=[0.04,0.86,-0.51]	N=[-0.96,0.17,0.21]
2016-11-23T07:49:52	2	B20	L=[0.14,0.29,0.95]	M=[0.4,-0.89,0.21]	N=[0.91,0.35,-0.23]
2016-11-28T15:46:59	2	B23	L=[0.01,-0.53,0.85]	M=[0.96,-0.22,-0.15]	N=[0.27,0.82,0.5]
2016-12-11T04:41:50	4	B24	L=[0.37,-0.17,0.91]	M=[0.24,0.97,0.08]	N=[-0.90,0.19,0.4]
2016-12-19T14:15:01	2	B25	L=[0.12,0.31,0.94]	M=[-0.89,-0.38,0.23]	N=[0.43,-0.87,0.23]
2017-01-11T04:22:43	4	B27	L=[-0.18,0.73,0.66]	M=[0.29,0.68,-0.68]	N=[-0.94,0.07,-0.33]
2017-01-20T12:32:07	2	B28	L=[-0.03,-0.17,0.99]	M=[0.69,0.71,0.14]	N=[-0.73,0.68,0.09]
2017-01-22T10:15:46	2	B29	L=[0.08,-0.04,1]	M=[0.40,0.91,0]	N=[-0.91,0.4,0.09]
2017-01-22T10:47:33	2	B31	L=[-0.04,-0.61,0.79]	M=[-0.85,-0.4,-0.34]	N=[0.52,-0.69,-0.5]
2017-01-27T12:05:43	1	B32	L=[-0.15,-0.20,0.97]	M=[0.72,-0.7,-0.03]	N=[0.68,0.69,0.25]

possible biases due to the variability of the upstream magnetic field conditions among the events included. The $B_L = 0$ point was determined as the closest point of the maximum $J.E$ value where the sign of B_L changes. The electron crescent point was determined as the point with the highest MeanRL value as defined in *Lenouvel et al.* [2021] and in section 3.2.2, which is the point where the electron velocity distribution function crescent shape is the most distinguishable. For each event, the stagnation line (that will be assimilatyed to the electron stagnation point) was estimated by taking the closest points to the maximum $J.E$ value among the minimum electron velocity values for the event. If a minimum value for V_e was too high (over 50 km/s) to be close to the real stagnation point or if it was too far away of the maximum $J.E$ point, the point was not included which explains why the scatterplots involving the stagnation point have less points than the other scatterplots.

We observe a large dispersion of the relative positions of the points of interest, with distances that are of the order of a fraction of ion inertial length. Some points are located at distances that should not be physical since the time separation is above the second which is larger than the typical scale of the EDR. In this case, it shows that the determination of the points of interest is complex and their definition can not be done in a simple manner. In figure 4.5, we observe faint trends as a function of the guide field that do not reverse when transforming the time separation into normalized distances but become even fainter. It indeed seems that the $B_L = 0$ point and the electron stagnation point get closer to the maximum energy conversion point as the guide field increases, and the electron crescent point gets further away from the maximum energy conversion point as the guide field increases. However, looking at the green area which represents the 95% confidence interval for the linear regression, it is clear that the observed trends are not robust enough and thus do not allow to draw definitive conclusions. Regarding the electron density ratio influence, in figure 4.6 we see that the 3 points of interest seem to get further away from the maximum energy conversion point when looking at the time separations only. However, the trend reverses for the electron stagnation point and the electron crescent point when transforming the time separations into normalized distances. The confidence intervals tell us that the results are however not robust enough to be certain of the conclusions observed.

4.3.2 Distance between the X-line and the stagnation point

Another study done on the EDRs reported in *Webster et al.* [2018] and *Lenouvel et al.* [2021] is focused on the computation of the distance between the X-line and the stagnation point. The starting point of this study is the paper from *Cassak et al.* [2017] in which they study the location where the energy conversion is maximum. They observed that for weak guide fields, the maximum of the energy conversion is located halfway between the X-line and electron stagnation point on the magnetospheric side. As the guide field increases, the electron stagnation point gets closer to the X-line and the maximum of the energy conversion moves closer to the electron stagnation point.

Other studies have been made on the subject of the electron stagnation point, for example *Cassak and Shay* [2007] first showed that the X-line and the stagnation point are not collocated in asymmetric reconnection, the stagnation point being on the side of the

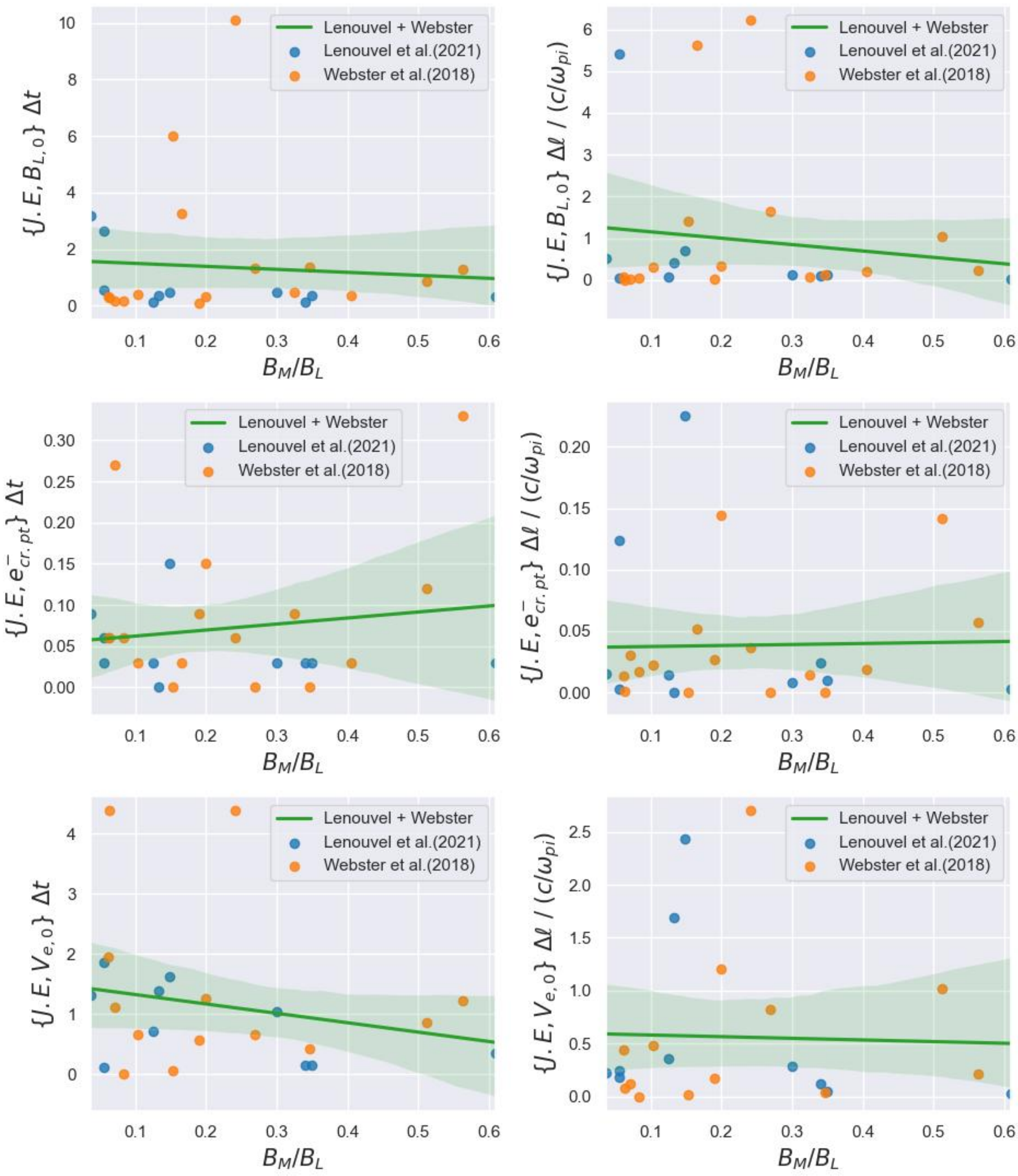


Figure 4.5: Scatterplot of the time delta (left) in seconds between the maximum energy conversion point and various points of interest located in the EDR and scatterplot of the distance in ion inertial length (right) between the maximum energy conversion point and various points of interest located in the EDR, as a function of the guide field B_M/B_L. $B_L = 0$ refers to the crossing of the X-line, $e^-_{cr,pt}$ refers to the electron crescent point which is the closest point where an electron crescent is visible in the electron velocity distribution function, and $V_{e,0}$ is the electron stagnation point.

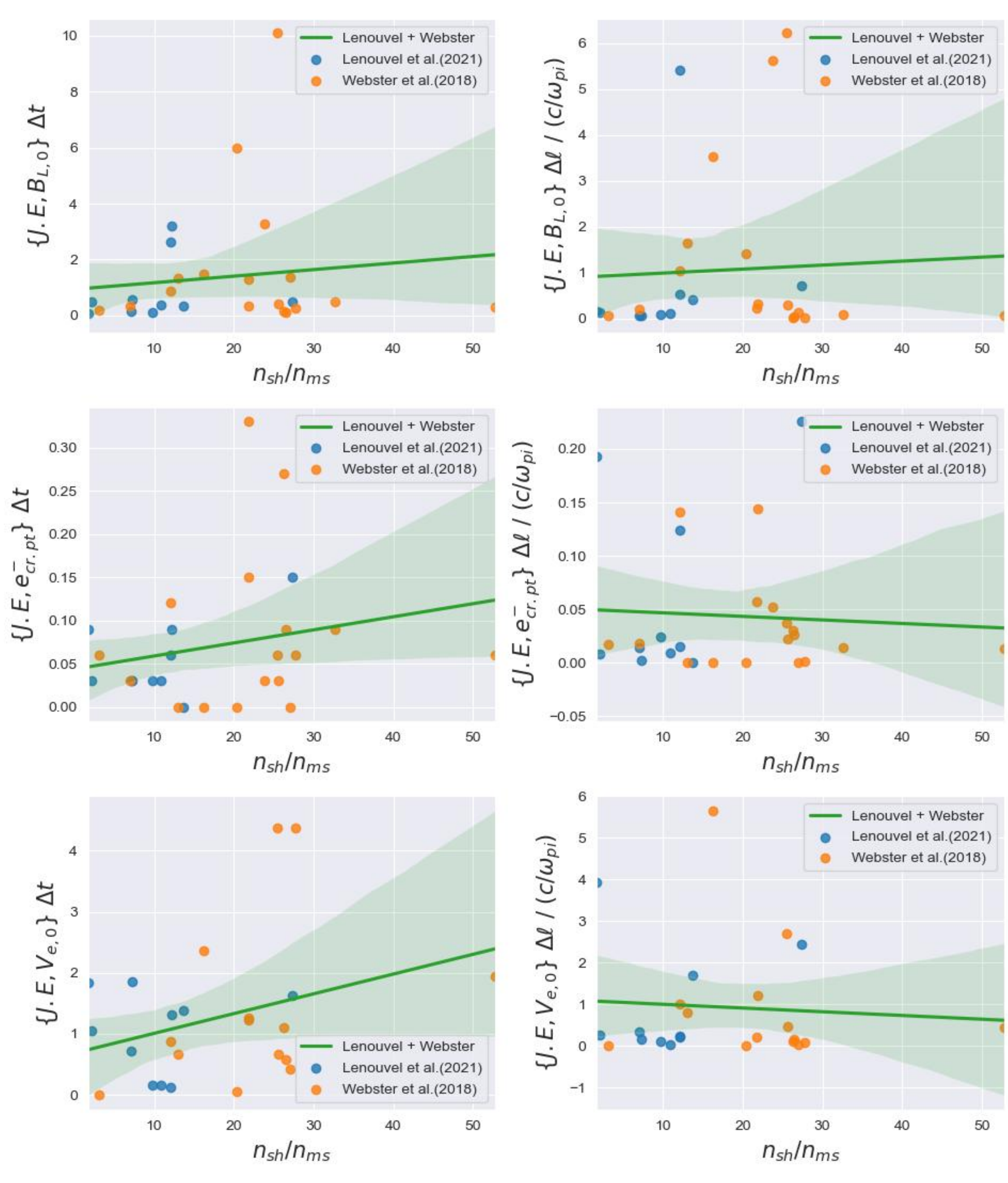


Figure 4.6: Scatterplot of the time delta (left) in seconds between the maximum energy conversion point and various points of interest located in the EDR and scatterplot of the distance in ion inertial length (right) between the maximum energy conversion point and various points of interest located in the EDR, as a function of the electron density ratio n_{sh}/n_{ms}. $B_L = 0$ refers to the crossing of the X-line, $e^-_{cr,pt}$ refers to the electron crescent point which is the closest point where an electron crescent is visible in the electron velocity distribution function, and $V_{e,0}$ is the electron stagnation point.

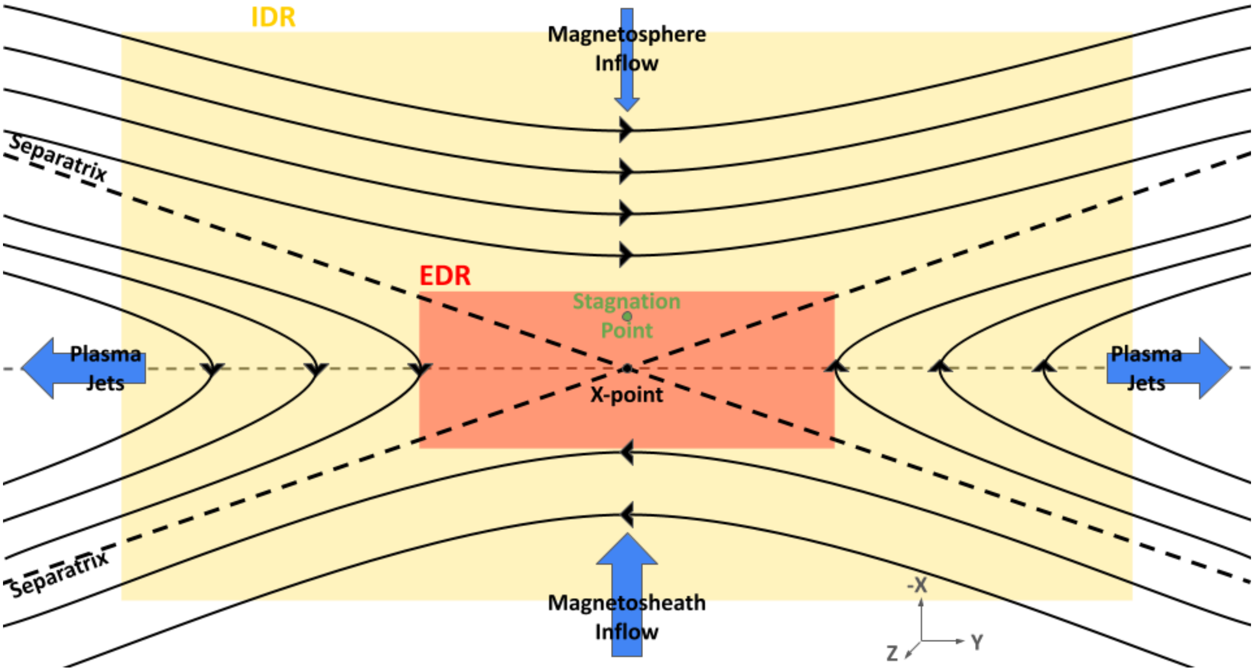


Figure 4.7: Figure showing the topology of asymmetric magnetic reconnection

X-line where the Alfvén velocity is the highest. In the case of symmetric reconnection, the X-line, the stagnation point and the maximum energy conversion point are supposed to be all collocated.

My objective was to use the analytical formula for the distance between the X-line and the stagnation point in the N direction proposed in *Cassak et al.* [2017], and to compare the results with a manual measurement of the distance between the X-line and the stagnation point estimated in the N direction from a time separation method applied on the EDR cases that have been reported. The goal was also to see the evolution of that distance regarding the magnetic field ratio $B_{L_{sheath}}/B_{L_{sphere}}$ and the electron density ratio $n_{e_{sheath}}/n_{e_{sphere}}$.

The analytical formula from *Cassak et al.* [2017], based on a Sweet-Parker-type scaling analysis for asymmetric antiparallel reconnection by *Cassak and Shay* [2007], is the following :

$$\frac{\delta_{(X,Se)_N}}{2\delta_e} \simeq \frac{\rho B^2_{L_{e,MSp}} - B^2_{L_{e,MSh}}}{(B_{L_{e,MSh}} + B_{L_{e,MSp}})(\rho B_{L_{e,MSp}} + B_{L_{e,MSh}})} \tag{4.2}$$

$$\rho = \frac{n_{e,MSh}}{n_{e,MSp}} \tag{4.3}$$

with $\delta_{(X,Se)_N}$ the distance between the X-line and the electron stagnation point in the N direction, δ_e the width of the current sheet, $B_{L_{e,MSh}}$ the L component in the LMN coordinate system of the magnetic field on the magnetosheath side, $B_{L_{e,MSp}}$ the L component in the

LMN coordinate system of the magnetic field on the magnetospheric side, $n_{e,MSh}$ the electron density on the magnetosheath side and $n_{e,MSp}$ the electron density on the magnetospheric side. This formula does not expressively depend on the guide field, the expression used to compute the width of the current sheet does however. *Cassak et al.* [2017] state that there is no hard threshold but if the guide field is below 0.1 or 0.2, one can consider to be in the zero guide field limit and can use the following expression for the width of the current sheet :

$$\delta_e \sim \frac{1}{2}\left(\sqrt{\frac{B_{L^{e,MSh}}}{B_{L^{e,MSp}}}} + \sqrt{\frac{B_{L^{e,MSp}}}{B_{L^{e,MSh}}}}\right) d_{e,out} \tag{4.4}$$

$$d_{e,out} = \sqrt{\frac{\epsilon_0 m_e c^2}{q_e^2\, n_{e,out}}} \tag{4.5}$$

$$n_{e,out} \sim \frac{n_{e,MSh} B_{L^{e,MSp}} + n_{e,MSp} B_{L^{e,MSh}}}{B_{L^{e,MSh}} + B_{L^{e,MSp}}} \tag{4.6}$$

with $d_{e,out}$ the asymmetric inertial scale, ϵ_0 the vacuum permittivity, m_e the electron mass, c the speed of light, q_e the electron charge and $n_{e,out}$ the outflow electron density.

When the guide field is not negligible anymore, one should replace the asymmetric inertial scale by the electron Larmor radius ρ_e when computing the width of the current sheet, yielding the following formula :

$$\delta_{e,GF} \sim \frac{1}{2}\left(\sqrt{\frac{B_{L^{e,MSh}}}{B_{L^{e,MSp}}}} + \sqrt{\frac{B_{L^{e,MSp}}}{B_{L^{e,MSh}}}}\right) \rho_e \tag{4.7}$$

$$\rho_e = \frac{v_{out}}{\Omega_{ce}(B_{tot})} \tag{4.8}$$

$$\Omega_{ce}(B_{tot}) = \frac{|q_e| B_{tot}}{m_e} \tag{4.9}$$

$$v_{out} = c_{e,asym} = \sqrt{\frac{k_B T_{e,asym}}{m_e}} \tag{4.10}$$

$$T_{e,asym} = \frac{n_{e,MSh} T_{e,MSh} B_{L^{e,MSp}} + n_{e,MSp} T_{e,MSp} B_{L^{e,MSh}}}{n_{e,MSp} B_{L^{e,MSh}} + n_{e,MSh} B_{L^{e,MSp}}} \tag{4.11}$$

with v_{out} the outflow velocity given by $c_{e,asym}$ the electron sound speed, B_{tot} is the average of the total magnetic field $(B_L^2 + B_M^2)^{1/2}$ on the two sides, Ω_{ce} the electron cyclotron frequency, k_B the Boltzman constant, $T_{e,asym}$ the electron asymmetric temperature, $T_{e,MSh}$ the electron temperature on the magnetosheath side and $T_{e,MSp}$ the electron temperature

Events	$n_{e,MSh}$	$n_{e,MSp}$	$B_{L,MSh}$	$B_{L,MSp}$	$\delta_{(X,Se)_N}/2\delta_e$	$\delta_{(X,Se)_N}/(c/\omega_{pe})$
Eriksson et al. [2016]	15	6	15	30	0.509	1.214
Burch et al. [2016]	11.3	0.7	23	39	0.594	1.523
Burch and Phan [2016]	7.5	3	15	35	0.554	1.335

Table 4.2: Table showing the computed distance between the X-line and the stagnation point either normalized by the width of the current sheet or by the electron inertial length, as well as some of the values needed to compute them for 3 different simulations of MMS events from *Cassak et al.* [2017]. The density and magnetic field intensity values are taken directly from this paper.

on the magnetospheric side.

From looking at equation 4.2, we see that as the guide field increases, the right hand side of equation is not going to change. However, the width of the current sheet δ_e depends on the electron Larmor radius in the presence of a guide field. An increase of the guide field will result in the electron being more magnetized, leading to a decrease of the gyroradius and thus of the current sheet width. To keep the right hand side of equation 4.2 unchanged, the distance between the X-line and the electron stagnation point must decrease as the guide field increases. *Cassak et al.* [2017] consequently predict that the distance between the X-line and the location of the energy conversion site decreases with a guide field.

In their study, they make PIC simulations of three observed EDR events to study the location of the stagnation point : the 8 September 2015 event reported in *Eriksson et al.* [2016], the 16 October 2015 event reported in *Burch et al.* [2016] and the 8 December 2015 event reported in *Burch and Phan* [2016]. The simulation of these events is useful as they allow to study things that were outside of the MMS spacecraft trajectory during these events and thus to study the "full picture". Table 4.2 shows the computed distance between the X-line and the stagnation point either normalized by the width of the current sheet or by the electron inertial length for these 3 PIC simulations. The computation of $\delta_{(X,Se)_N}/2\delta_e$ is done using equation 4.2, which does not depend on the guide field, the guide field dependence being embedded in the computation of the width of the current sheet δ_e.

In order to compare the distance between the X-line and the stagnation point in the N direction computed from the theoretical formula and the distance computed using the manual time separation method, I needed to be able to apply both methods to the same cases. I started with the 50 cases list from *Webster et al.* [2018] and *Lenouvel et al.* [2021]. LMN coordinates could be computed successfully for 39 of them, and among these 39 cases, only 23 showed a clean magnetosphere/magnetosheath distinction allowing the retrieving of the required density and magnetic field intensity values on both sides of the magnetopause. The analytical formula could thus be applied to these 23 cases, which results can be found in table 4.3

To apply the time separation method, I first needed to compute the magnetopause velocity in the N direction, which could only be done for 15 cases. The magnetopause velocity in

Table 4.3: List of EDR events with the relevant values to apply the analytical formula. The "Event time" column shows the date and the UTC time at which the EDR crossing can be found, the time being given to the nearest second. The "SC" column shows in the data of which spacecraft the event was observed. The "AKA" column shows the nickname used to identify more easily each event, the events which nickname starts with an L come from *Lenouvel et al.* [2021] and the others come from *Webster et al.* [2018]. The "$n_{e,MSh}$" and "$n_{e,MSp}$" columns show the electron density in cm^{-3} for the magnetosheath and for the the magnetosphere side and the "$B_{L,MSh}$" and "$B_{L,MSp}$" columns show the magnetic field intensity in nT. The last column shows the distance between the X-line and the stagnation line normalized by the width of the current sheet, computed with the analytical formula from *Cassak et al.* [2017]

Event time	AKA	SC	$n_{e,MSh}$	$n_{e,MSp}$	$B_{L,MSh}$	$B_{L,MSp}$	$\delta_{(X,Se)_N}/2\delta_e$
2015-09-11T10:13:35	L04	3	21.59	1.43	56.99	50.79	0.401
2015-09-19T09:26:27	L06	3	26.65	1.14	29.98	55.06	0.624
2015-10-06T14:17:54	L10	1	20.09	1.28	21.63	32.11	0.556
2016-01-02T23:46:17	L12	4	21.12	2.10	16.39	48.60	0.715
2016-02-19T18:42:38	L17	1	7.714	0.80	9.13	40.19	0.791
2016-02-28T01:07:33	L18	3	19.16	1.70	39.71	48.00	0.478
2015-10-16T10:33:30	A02	1	13.85	0.61	29.45	40.81	0.549
2015-10-16T13:07:02	A03	3	10.19	0.36	21.42	38.48	0.623
2015-10-22T06:05:21	A04	1	18.69	3.83	28.93	56.43	0.565
2015-11-12T07:19:21	A06	1	7.043	0.23	15.67	29.14	0.632
2015-12-06T23:38:31	A07	2	17.98	0.43	48.95	64.84	0.551
2015-12-08T11:20:44	A08	1	10.82	0.65	27.10	50.50	0.619
2015-12-09T01:06:11	A09	4	7.011	0.36	30.92	48.21	0.576
2015-12-14T01:17:39	A10	1	10.21	0.74	10.51	43.86	0.789
2016-01-10T09:13:37	A12	4	22.57	0.99	36.02	60.05	0.599
2016-02-07T20:23:34	A13	1	11.05	0.36	8.36	39.87	0.819
2016-10-22T12:58:40	B14	2	29.05	1.83	29.09	49.89	0.596
2016-11-02T14:46:18	B15	4	15.81	1.25	28.30	34.52	0.488
2016-11-06T08:40:58	B16	2	30.89	2.26	28.26	60.18	0.647
2016-11-12T17:48:46	B17	2	8.593	0.54	27.78	46.80	0.591
2016-11-13T09:10:40	B18	4	13.93	0.42	36.30	43.84	0.522
2016-12-19T14:15:01	B25	2	16.29	1.24	12.72	43.50	0.751
2017-01-22T10:47:33	B31	2	11.92	0.19	8.471	42.40	0.830

the N direction is computed using the Constant Velocity Approach (CVA) which is detailed in the Appendix B. The computed magnetopause velocity was subtracted from the electron velocity to get the electron velocity in the plasma frame moving with the magnetopause as suggested in *Malakit* [2012] to find the stagnation point. Then, I needed to get the position of the electron stagnation line, which is where the electron velocity in the N direction is null ($V_{e,N} = 0$). In only 12 cases could an approximate stagnation point be located near the $B_L = 0$ point. An example of the localisation of these points can be found in figure 4.8. The last step of the time separation method is to measure the time separation between the $B_L = 0$ point and the $V_{e,N} = 0$ point and to multiply it by the previously computed magnetopause velocity to get a distance. To compare this distance with the results obtained with the analytical formula, the distance must be normalized by the width of the current sheet $2\delta_e$, which can either be computed in the zero guide field limit or using the guide field dependent formula. This was successfully done for 10 EDR cases that are presented in table 4.4

Figure 4.9 shows a 3D scatterplot of the distances from tables 4.3 and 4.4 as a function of the magnetic field ratio $B_{L_{sheath}}/B_{L_{sphere}}$ and the electron density ratio $n_{e_{sheath}}/n_{e_{sphere}}$. The 2D surface that is visible represents all the possible values for the analytical formula from *Cassak et al.* [2017]. The blue points, all located on the surface, are the values obtained using the analytical formula from *Cassak et al.* [2017]. The red points are the values obtained using the time separation method, using the zero guide field limit for the computation of the width of the current sheet. Finally, the green points are the values obtained using the time separation method, using the guide field dependent formula for the computation of the width of the current sheet. Some of the values are not visible on this 3D plot for visualisation purposes, but they are present on figure 4.10 and 4.11, which are the projections on the two axes of the bottom plane of figure 4.9. These figures show error bars computed using the uncertainty of measure on the location of the X-line, the stagnation point and the uncertainty in the measure of the magnetopause velocity used to get in the plasma frame and to transform the time separation measured into a distance.

We see that as the magnetic field ratio increases, the theoretical distance computed with equation 4.2 decreases. This is because as the magnetic field ratio gets closer to 1, we get closer to the symmetrical reconnection scenario where the stagnation point and the X-line are collocated. When the electron density ratio increases, we also get closer to the symmetrical reconnection scenario, but the density ratio does not have the same weight as that of the magnetic field ratio in the computation of the distance $\delta_{(X,Se)_N}/2\delta_e$. After 10, for the same magnetic field ratio, the distance $\delta_{(X,Se)_N}/2\delta_e$ only increases very slightly to become almost constant.

If we now look at the results from the time separation method, we see a big difference between the red points that use the zero guide field limit to compute the width of the current sheet, and the green points that use the guide field dependant formula to compute the width of the current sheet. When the guide field increases, the width of the current sheet decreases a lot, increasing the ratio of the distance from the X-line to the stagnation point with the width of the current sheet, as can be seen in table 4.4. Thus, the green points are all above the blue and red points. The red points are above and under the blue points meaning that

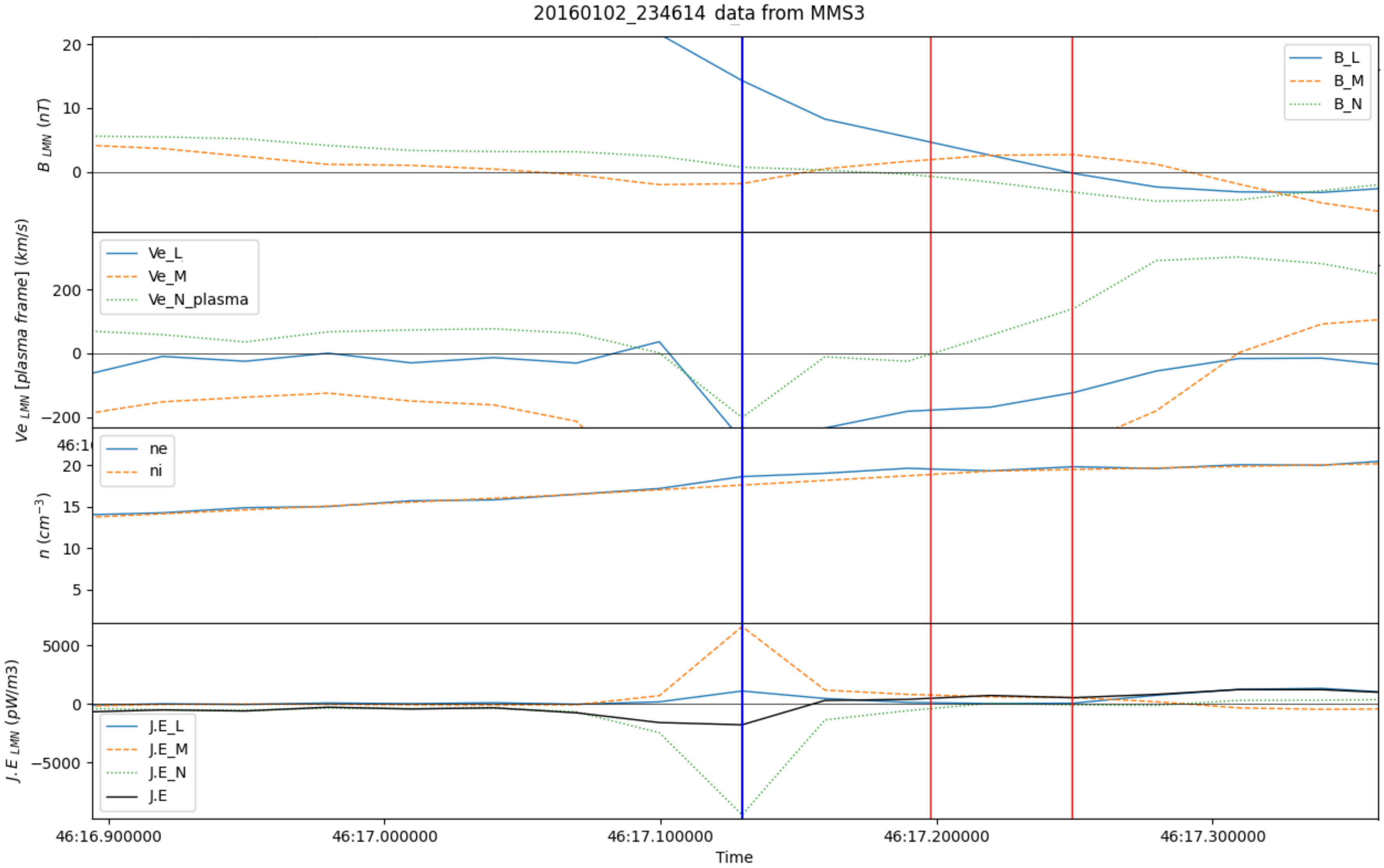


Figure 4.8: Figure showing for the event 20160102_234614 the locations of the max J.E point (blue line), of the electron stagnation line (left red line) and of the X-line (right red line). The magnetosphere is located on the left side of the plot and the magnetosheath-like region is located on the right after the second vertical red line.

Table 4.4: List of EDR events with the relevant values for the time separation method. The "Event time" column shows the date and the UTC time at which the EDR crossing can be found, the time being given to the nearest second. The "SC" column shows in the data of which spacecraft the event was observed. The "AKA" column shows the nickname used to identify more easily each event, the events which nickname starts with an L come from *Lenouvel et al.* [2021] and the others come from *Webster et al.* [2018]. The "MP velocity" column shows the magnetopause velocity computed in (in km/s). The "δ_e" and "$\delta_{e,GF}$" columns show the half-width of the current sheet in (in m) computed either in the zero guide field limit or with the guide field dependent formula. The "$\Delta_t(B_L = 0, V_{e,N} = 0)$" column shows the time separation between the magnetopause crossing and the stagnation point (in s). The last two columns show the distance between the X-line and the stagnation line, normalized by the width of the current sheet in the zero guide field limit and with the guide field dependent formula.

Event time	SC	AKA	MP velocity	δ_e	$\delta_{e,GF}$	$\Delta_t(B_L = 0, V_{e,N} = 0)$	$\delta_{(X,Se)_N}/2\delta_e$	$\delta_{(X,Se)_N}/2\delta_{e,GF}$
2015-09-19T09:26:27	3	L06	17.765	1323.232695	323.3161502	0.0925	0.6209273	2.5412626
2016-01-02T23:46:17	4	L12	-63.942	1514.211992	555.6116326	0.027318	0.57679224	1.5719357
2016-02-19T18:42:38	1	L17	52.871	2696.093923	968.1631051	0.21	2.0590732	5.734008
2015-12-06T23:38:31	2	A07	36.624	1661.167981	344.1384602	0.03	0.33070347	1.5963168
2015-12-08T11:20:44	1	A08	37.534	2066.534584	532.1497605	0.03	0.272438	1.0579776
2015-12-09T01:06:11	4	A09	-49.609	2591.642491	464.6634679	0.2	1.9139909	10.67521
2015-12-14T01:17:39	1	A10	26.573	2323.176928	945.4868103	0.17	0.97225755	2.3889558
2016-01-10T09:13:37	4	A12	24.722	1442.282106	326.3365899	0.049	0.41994333	1.8559879
2016-11-02T14:46:18	4	B15	78.912	1755.421332	451.6490832	0.93	20.90329	81.24467
2016-11-13T09:10:40	4	B18	-101.857	1909.433596	524.3010248	0.29	7.734863	28.169327

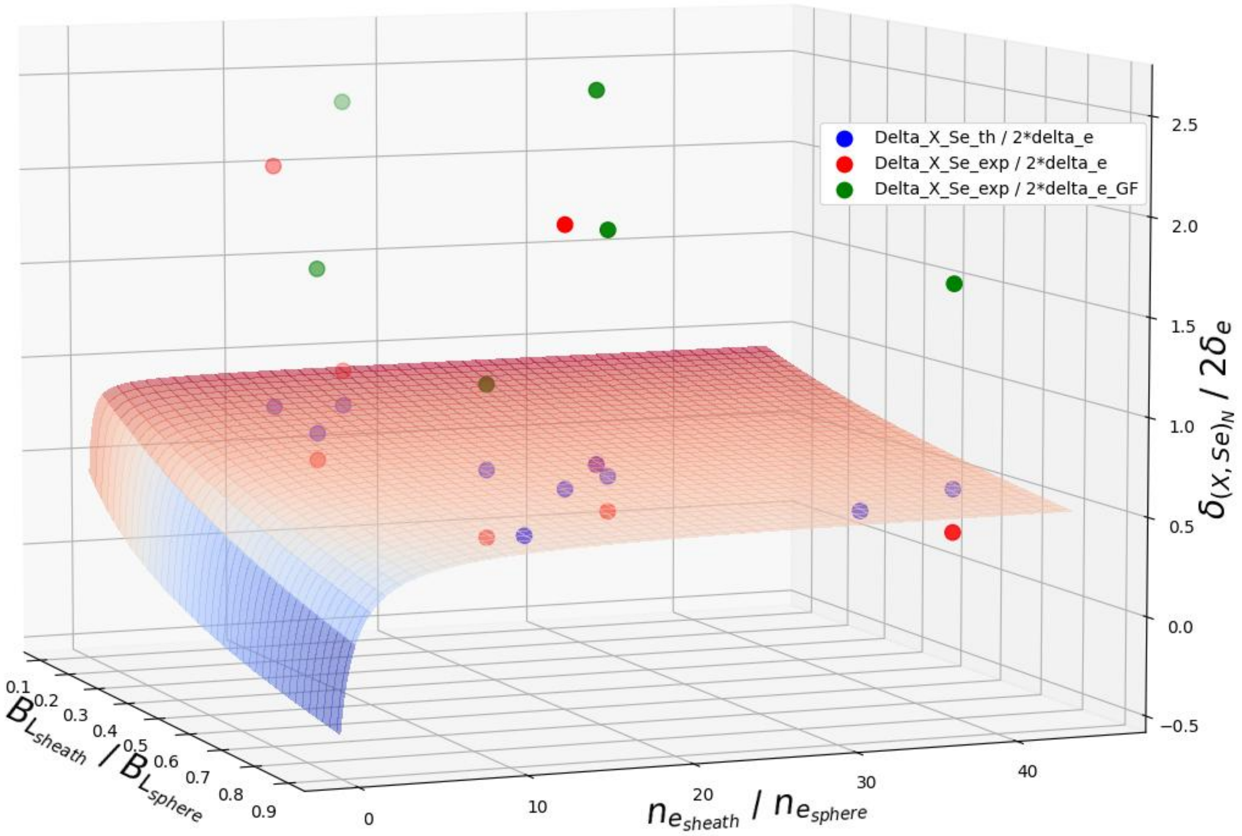


Figure 4.9: 3D scatterplot of the distance in the N direction between the X-line and the stagnation line normalized by the width of the current sheet, versus the ratio between the L component of the magnetic field in the magnetosheath and the L component of the magnetic field in the magnetosphere, and versus the electron density ratio between the magnetosheath and the magnetosphere. The 2D surface represents all the possible values for the analytical formula from *Cassak et al.* [2017]. The blue points on the surface are the values obtained using the analytical formula from *Cassak et al.* [2017]. The red points are the values obtained using the time separation method using the zero guide field limit for the computation of the width of the current sheet. The green points are the values obtained using the time separation method, using the guide field dependent formula for the computation of the width of the current sheet.

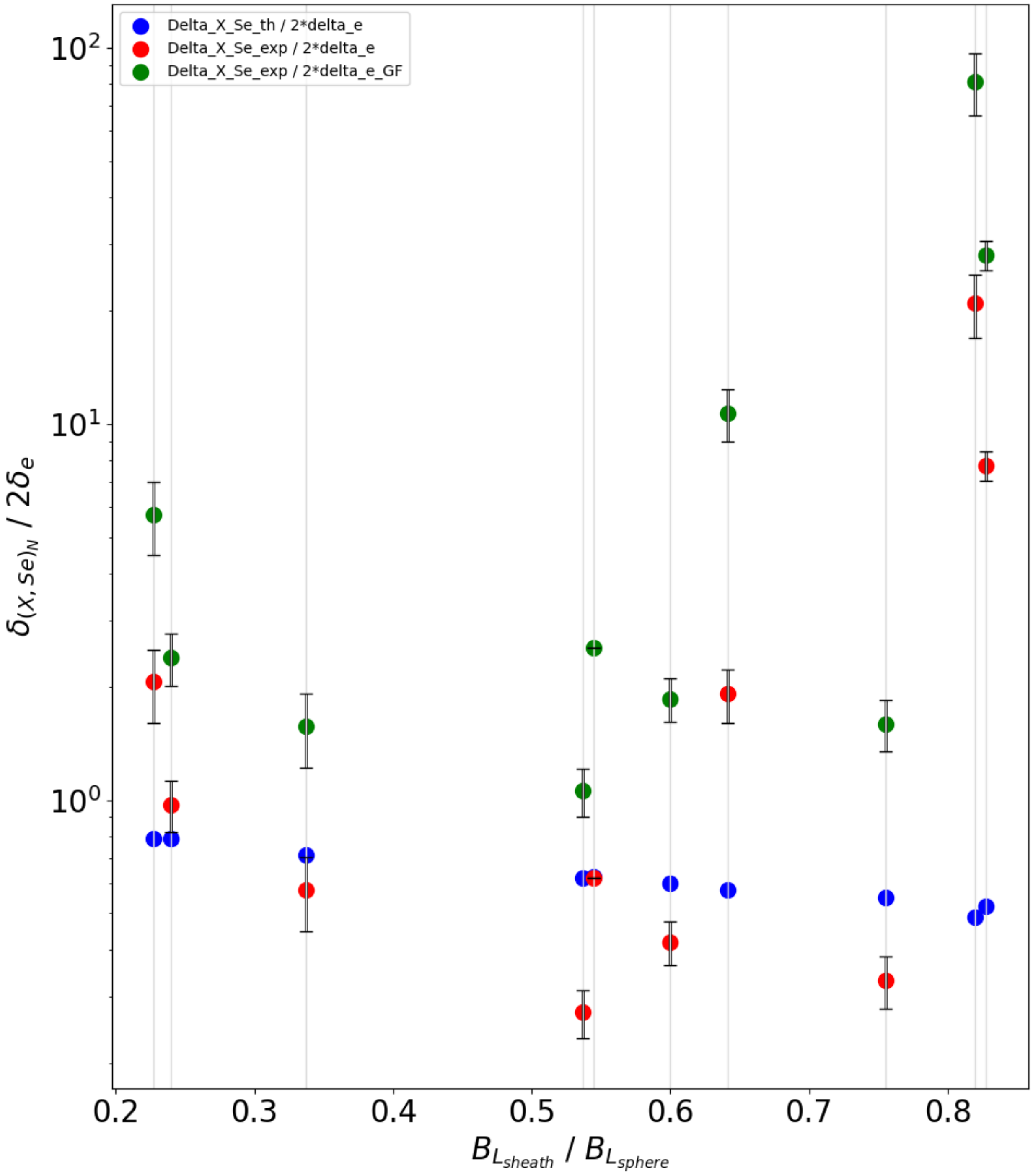


Figure 4.10: Scatterplot of the distance in the N direction between the X-line and the stagnation line normalized by the width of the current sheet versus the ratio between the L component of the magnetic field in the magnetosheath and the L component of the magnetic field in the magnetosphere

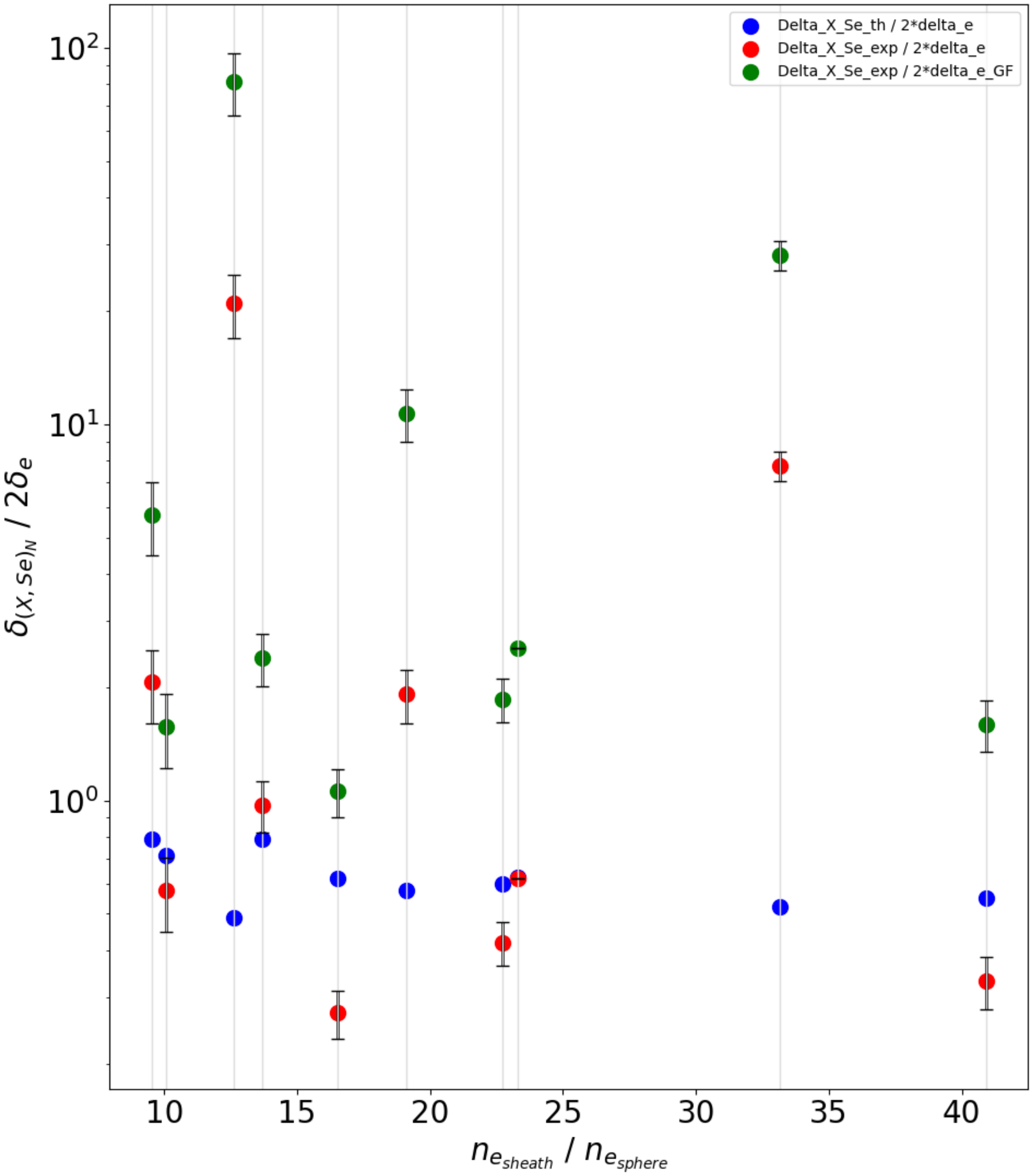


Figure 4.11: Scatterplot of the distance in the N direction between the X-line and the stagnation line normalized by the width of the current sheet versus the electron density ratio between the magnetosheath and the magnetosphere

the distance is either larger or smaller than the theoretical one but always quite close to it. The two events with a magnetic field ratio above 0.8 show the biggest difference between the theoretical distances and the measured ones. I have no physical explanation for this behavior, these events have very different electron density ratios but the magnetopause velocity was quite high for these two events (79 km/s and -102 km/s) which probably led to a bad measurement of the time separation between the X-line and the stagnation point, the two time deltas being probably too high to be physically meaningful.

Since the paper at the origin of this study was focused on the location of the maximum energy conversion point, I decided to look at the distance between the X-line and the maximum energy conversion point and how it relates to the location of the stagnation point. Figure 4.12 shows the results for the 10 EDR cases studied before. *Cassak et al.* [2017] showed that for their cases, the max energy conversion point is always located between the X-line and the electron stagnation point and the maximum energy conversion point gets closer to the stagnation point (and further away from the X-line) as the guide field increases. However, we see in figure 4.12 that the maximum energy conversion point is beyond the estimated electron stagnation point for 8 of the 10 EDR cases, and the proximity between the maximum energy conversion point and the estimated electron stagnation point does not follow any pattern as the guide field increases. This may also be the consequence of the significant uncertainties regarding the determination of the electron stagnation point due to the trajectory of the spacecraft combined with the temporal variability of the electron velocities measured. On the other side, the figure suggests a possible trend regarding the distance between the max $\mathbf{J} \cdot \mathbf{E}$ point and the X-line, which seems to increase as the guide field increases, with the lowest distances only at low guide field ratios (below 0.15), which is in agreement with the expectations by *Cassak et al.* [2017].

4.4 Nature of the energy conversion at the EDR

Understanding the energy conversion mechanisms inside the reconnection sites is among the key topics in the community. A number of studies focused in particular on the influence of the guide field on the energy conversion close to the X-line. First it is expected to reduce the reconnection efficiency by getting away from the ideal situation with opposite fields from both sides with a null magnetic field at the center of the reconnection. *Yi et al.* [2019] showed indeed that a guide field leads to the reduction of both the reconnection rate and the energy conversion rate. *Goodbred et al.* [2021] also investigated the active role of the guide field in Petschek-type reconnection, where it changes the exhaust structure and can significantly reduce both the energy conversion rate and the reconnection rate.

The extension of the region where energy conversion is localized has also been debated. *Yamada et al.* [2016] reported the observation of energy conversion across a region significantly larger than the narrow electron diffusion region due to non classical heating based on MRX experimental studies. Later, *Burch et al.* [2018b] showed strong energy conversion is highly localized within the electron diffusion region (EDR) and can occur at both the X-line

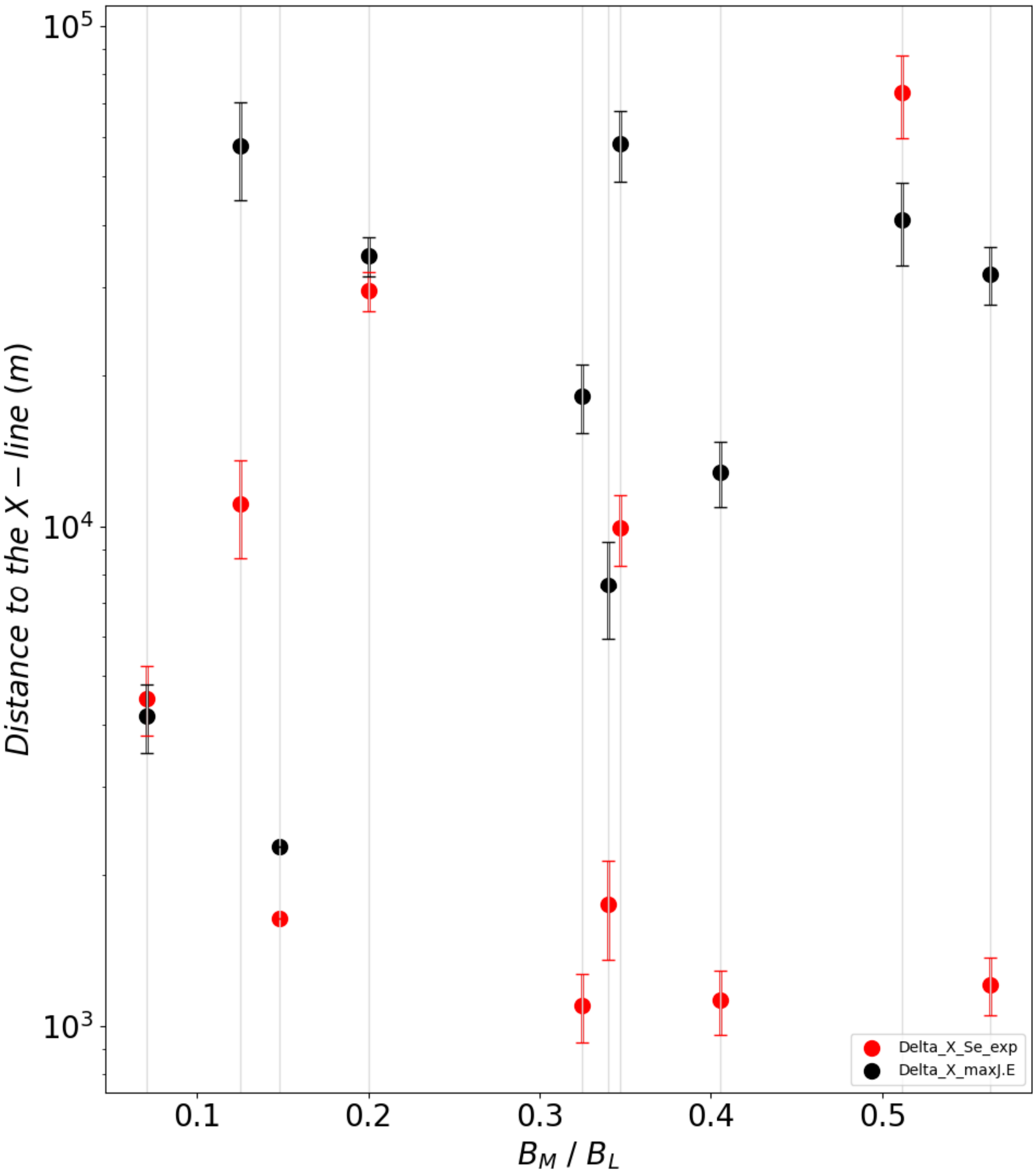


Figure 4.12: Scatterplot of the distance in the N direction between the X-line and the stagnation line (red points) and the distance in the N direction between the X-line and the max energy conversion (J.E) point (black points)

and the electron stagnation region, depending on the guide field. They also discussed the existence of oscillatory J.E structures with alternative positive/negative energy conversion rates (as discussed in section 3) induced presumably by large electric field amplitude standing waves. *Swisdak et al.* [2018b] suggested that these oscillatory J.E structures should not exist in strong guide field reconnection.

Several studies discussed in particular how the guide field can modify the structure of the energy conversion rate, driven either by parallel or perpendicular J.E, such as *Yamada et al.* [2016] or *Wilder et al.* [2018], or where the maximum energy conversion rate occurs with respect the X-line (*Genestreti et al.* [2017]; *Cassak et al.* [2017]; *Burch et al.* [2018b]; *Yi et al.* [2019]). I chose to focus on the papers by *Wilder et al.* [2018] and the conjoined work by *Genestreti et al.* [2017] and *Cassak et al.* [2017], since they proposed frameworks that allow for a statistical study based on our list of EDR events, at a period where the authors lacked enough cases.

4.4.1 Influence of the guide field on the parallel/perpendicular energy conversion ratio

Investigating the characteristics of the energy conversion rate is crucial to understand the mechanisms at work in the reconnection region. In particular, previous studies showed that the parallel and the perpendicular contributions to the total energy conversion rate for electrons ($J_e \cdot E_e = J_{e,\perp} \cdot E_{e,\perp} + J_{e,||} \cdot E_{e,||}$) should behave differently in the core of the reconnection. *Yamada et al.* [2016] and several other authors (see also *Fox et al.* [2018]) concluded that the perpendicular rate is significantly larger than the parallel one near the X point in the absence of guide field, since electrons gain their energy essentially from the out-of-plane reconnection electric field which is by definition perpendicular to the magnetic field near the X point in the absence of guide field. The dominance of the perpendicular contribution should remain until moderate guide field conditions (30% of the reconnection magnetic field), but above this threshold the parallel reconnection electric field should increase and the parallel contribution of energy conversion should dominate. *Wilder et al.* [2018] verified these predictions based on a small number of IDR events and current sheets, but no such study was performed on EDR events. I aimed at extending this paper results to our list of EDR events and investigate the relative importance of the parallel and perpendicular contributions to the energy conversion rate inside the core of the reconnection process.

Figure 4.13 is taken from *Wilder et al.* [2018]. It shows the evolution of the $(\mathbf{J} \cdot \mathbf{E'})_{ratio}$ (defined below) as a function of the guide field B_g. The vertical dashed line at $B_g = 0.5$ represents the threshold value after which the reconnection electric field will become parallel dominated. The conclusion of their study on multiple IDRs and current sheets taken close to EDRs is that the dissipation due to the parallel electric field becomes more significant when the guide field increases. They define the $(\mathbf{J} \cdot \mathbf{E'})_{ratio}$ as :

$$(\mathbf{J} \cdot \mathbf{E'})_{ratio} = \frac{max[(\mathbf{J} \cdot \mathbf{E'})_{||})] - max[(\mathbf{J} \cdot \mathbf{E'})_{\perp}]}{max(\mathbf{J} \cdot \mathbf{E'}))} \tag{4.12}$$

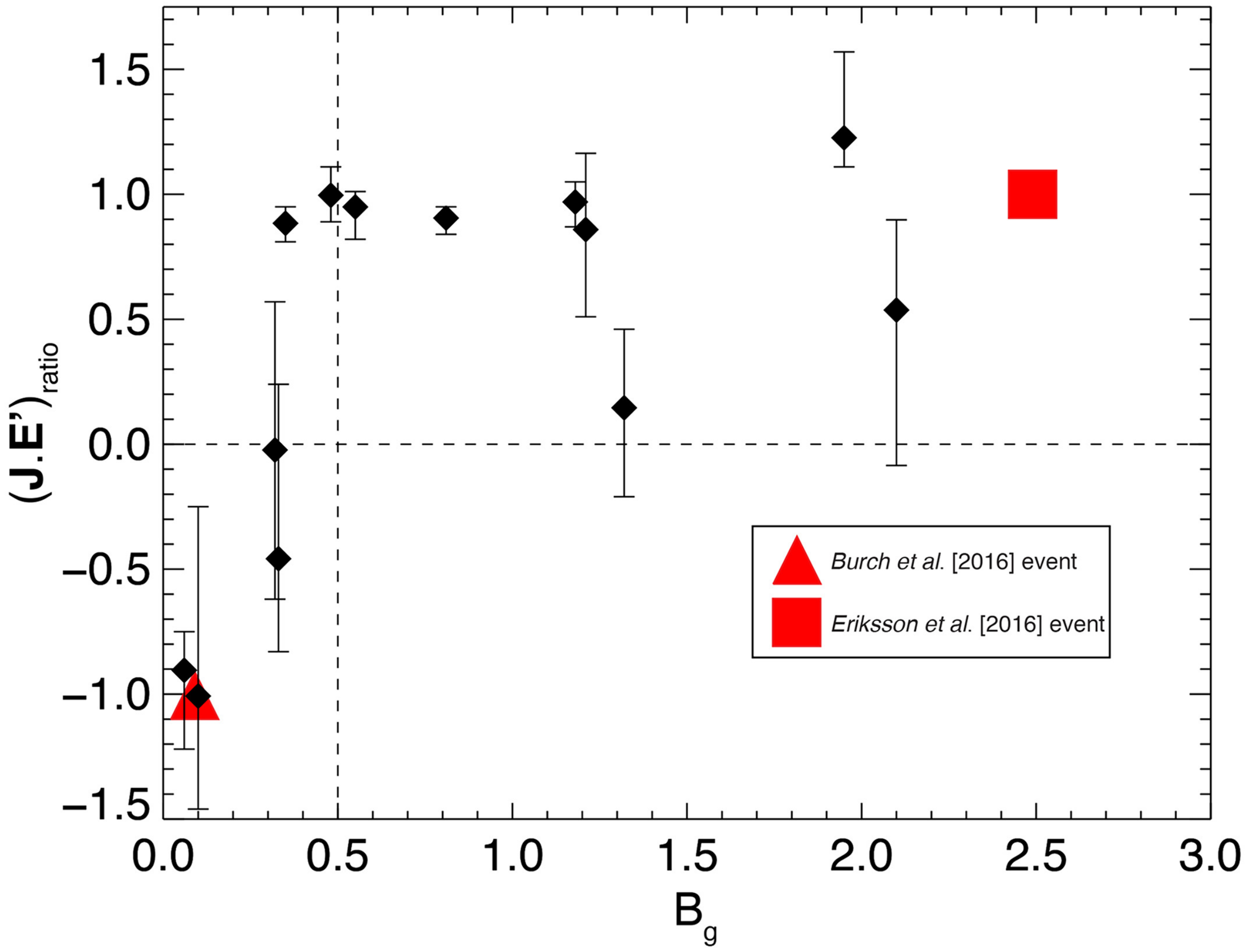


Figure 4.13: Figure showing the $(\mathbf{J} \cdot \mathbf{E'})_{ratio}$ as a function of the guide field B_g for 15 EDR cases. Figure taken from *Wilder et al.* [2018].

where the "max" is the maximum value of the quantity measured within the current sheet, implying that the two maxima can be taken at different places. The guide field B_g is defined as :

$$B_g = tan(90 - \frac{\theta_b}{2}) \tag{4.13}$$

$$\theta_b = \frac{arccos(B_1 \cdot B_2)}{||B_1|| \cdot ||B_2||} \tag{4.14}$$

with θ_b being the shear angle between the magnetic field vectors on each side of the reconnecting current sheet.

I computed the shear angle and the guide field B_g for 25 EDR cases out of the 50 cases from *Webster et al.* [2018] and *Lenouvel et al.* [2021]. I used the $\mathbf{J} \cdot \mathbf{E}$ ratio instead of the $J \cdot E'$ ratio but the qualitative results remain unchanged between $\mathbf{J} \cdot \mathbf{E}$ and $\mathbf{J} \cdot \mathbf{E'}$. The results are presented in table 4.5.

Figure 4.14 shows a scatterplot of the evolution of the $(\mathbf{J} \cdot \mathbf{E})_{ratio}$ as a function of the

Table 4.5: Energy conversion ratio and guide field values computed for multiple EDRs.

Event	**MMS**	$(\mathbf{J}\cdot\mathbf{E})_{ratio}$	B_g
20150909_125814	4	0.57	0.08
20150911_101214	3	1.85	0.25
20150919_092544	3	-0.64	0.34
20151001_065214	2	-0.51	0.10
20151006_141714	1	0.91	0.08
20160102_234614	4	-2.15	0.30
20160107_221104	4	-0.44	0.14
20160228_010604	3	-0.63	0.18
20151016_103254	1	-1.50	0.07
20151016_130524	3	-0.83	0.15
20151022_060324	1	2.30	0.35
20151112_071854	1	0.06	0.15
20151206_233734	2	-0.10	0.26
20151208_111904	1	1.24	0.49
20151209_010344	4	-0.25	0.04
20151214_011634	1	-0.64	0.34
20160110_091124	4	-0.86	0.32
20161022_125714	2	-0.71	0.08
20161102_144544	4	-0.40	0.30
20161112_174804	2	-0.68	0.39
20161113_090924	4	0.42	0.33
20161128_154544	2	1.69	0.13
20161219_141424	2	-0.61	0.28
20170111_042223	4	-0.92	0.27
20170122_104513	2	-0.41	0.18

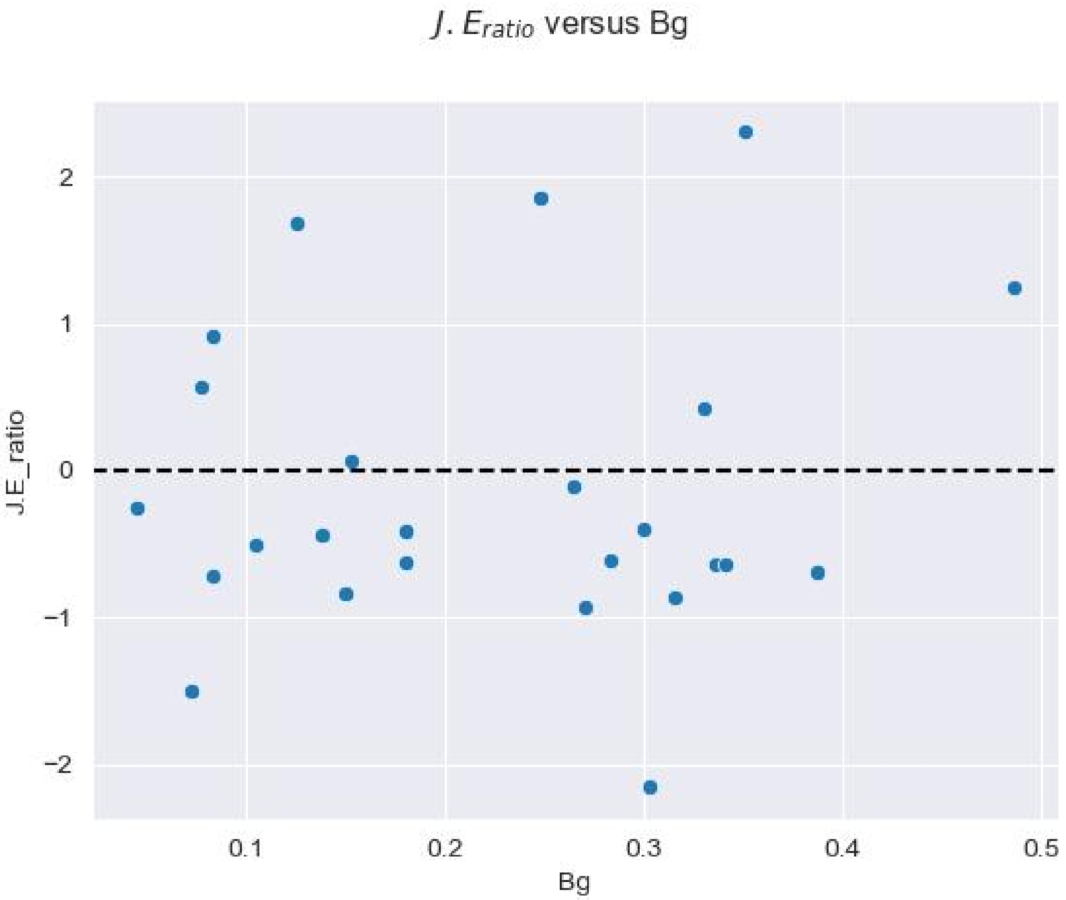


Figure 4.14: Scatterplot showing the $(\mathbf{J}\cdot\mathbf{E})_{ratio}$ as a function of the guide field B_g for 25 EDR cases.

guide field B_g, made using the values shown in table 4.5. This figure is to be compared with figure 4.13 from *Wilder et al.* [2018]. We see that there is no particular trend to be observed as the guide field increases. However, there are two things that can be said. The first is that the B_g range of our EDR cases only goes up to 0.5 whereas the B_g range from *Wilder et al.* [2018] goes up to 2.5 with the high guide field EDR event reported in *Eriksson et al.* [2016]. Since one of the strong criteria for identifying EDRs is the presence of electron crescents and the fact that the magnetic configuration allowing the formation of crescents gets destroyed with large guide fields, it is only logical that most EDR cases found so far are the low guide field events like the first EDR case reported by MMS from *Burch et al.* [2016]. The second interesting thing is that, even though no trend is visible on figure 4.14, we see that the ratio of events above and under the $(\mathbf{J}\cdot\mathbf{E})_{ratio} = 0$ is the same as that of figure 4.13 from *Wilder et al.* [2018] before the vertical dashed line at $B_g = 0.5$, with a majority of negative ratios and thus of larger perpendicular rates.

This result confirms that below $B_g = 0.5$, the dissipation seems to be driven by the perpendicular electric field more than the parallel electric field, and the transition may indeed occur at $B_g = 0.5$. More high guide field EDR cases would be needed to fully confirm the

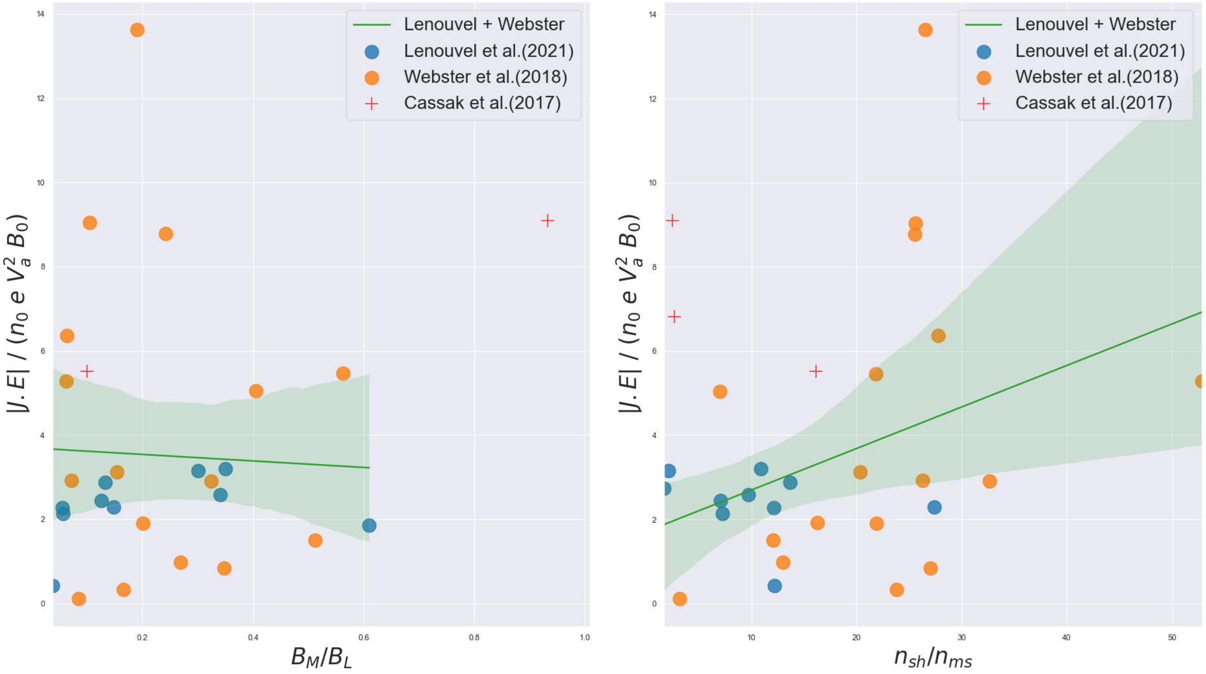


Figure 4.15: Scatterplots of the normalized absolute value of the energy conversion J.E as a function of the guide field B_M/B_L (left) and of the density asymmetry n_{sh}/n_{ms} (right)

hypothesis but finding high guide field EDRs is more difficult so it would require the specific search of such events. Nevertheless, this provides a statistical confirmation that the parallel conversion rate dominates in the EDRs for large guide field.

4.4.2 Variability of the intensity of the energy conversion

The EDR database presented earlier in table 4.1 from section 4.3 which was used to compare results with those of the conjoined study by *Genestreti et al.* [2017] and *Cassak et al.* [2017] was also used to observe the influence of the guide field and of the density asymmetry on the intensity of the energy conversion at the EDR.

Figure 4.15 shows two scatterplots of the normalized absolute value of the energy conversion $J.E$ as a function of the guide field B_M/B_L (Left) and of the density asymmetry n_{sh}/n_{ms} (Right). On the left plot, the event from *Cassak et al.* [2017] with $B_M/B_L = 5$ was not included on the plot for a better readability but it is still taken into account for the computation of the linear regression. The absolute value of $J.E$ is normalized by $(n_0\ e\ V_a^2\ B_0)$ to yield a dimensionless value where n_0 is the average ion density, e is the elementary charge, V_a is the Alfvén velocity and B_0 is the average magnetic field intensity taken on the magnetosheath side of the current sheet. By developing the terms, one can see that the normalisation $(n_0\ e\ V_a^2\ B_0)$ is equivalent to $(e\ B_0^3)/(\mu_0\ m_p)$.

We see that the intensity of the energy conversion seems to slightly decrease as the guide field increases but the confidence interval makes this trend doubtful given the large dispersion. This is in agreement with previous works based on few events, which suggested a change of the location of the maximum energy conversion due to the guide field rather than a modification of its intensity. On the other hand, the intensity of the energy conversion increases significantly as the density asymmetry between the magnetosheath and the magnetospheric side of the current sheet increases, and the shape of the confidence interval makes this trend believable. This is understandable because if the plasma is denser then the current J which depends on the density and the velocity of the particles will increase if the other parameters do not change.

Conclusion

In this section, I first looked at a potential link between EDRs and solar wind parameters. Although the intensity of the energy conversion could not be linked to the IMF clock angle, an interesting trend regarding the value of B_y IMF and the intensity of the energy conversion was found, the distribution of this parameter for the IDR labeled data points is homogeneous and representative of the data coverage, while the distribution of IMF B_y for EDR cases is highly inhomogeneous with mostly positive B_y cases. The origin of this trend is not completely clear yet but could be due to the modification of the location of the X-line by a combination of the dipole tilt and IMF B_y towards the subsolar point where the reconnection is expected to be more efficient according to *Reistad et al.* [2020].

The electron Larmor radius $r_L = \frac{m_e V_{e_\perp}}{eB}$, the magnetic gradient scale $L_0 = \frac{B}{\mu_0 J}$ and the electron vorticity $\Omega_e = \boldsymbol{\nabla} \times V_e$ (proposed as a new proxy parameter to identify the borders of the EDR in *Hwang et al.* [2019]) were investigated as potential new small scale parameters to identify EDRs in the future and showed promising results, even though the latter one could not be considered as a key parameter of the EDR due to its inconsistency to allow the identification of the EDR often enough.

I further investigated the energy conversion characteristics, its intensity, the parallel versus perpendicular contributions, its spatial location, as well as the location of key structural points of the reconnection (stagnation line, electron crescent point, X-line) based for the first time on a relatively large number of EDR events. I sought to compare with specific studies that provided conclusions based on simulations/theory or few case studies. It appears that the characteristics derived show similar orders of magnitude and sometimes the same behavior as expected, but overall the results show a situation that is more complex than could be expected at first. This may reveal a complex interplay of various phenomena (3D effects, waves), of various initial conditions (magnetosheath and magnetospheric conditions, guide field). This can also be induced by our reduced capability to investigate such problems due to the still too low number of cases (only 10 for some analysis), or the difficulty to analyze the characteristics of the EDR structure such as the electron stagnation point that appears difficult to determine from the MMS data. Further work is needed to better understand the energy conversion mechanisms at work in the EDR and the dependence of its structure with respect to key parameters like the guide field. One of the first step is clearly to enhance the number of new EDR cases and thus augment the statistics. This is the reason why the next chapter proposes a modified approach to automatically detect EDRs, and provide a larger list of events.

Chapter 5

Automatic detection of Electron velocity distribution function (eVDF) crescents

5.1 Introduction

After the first automatic EDR detection algorithm and the statistical analysis of the new EDR candidates found, I chose to improve the method of detection of the EDR crescents. Until then, I was using a scalar we created called the MeanRL (see section 3.2.2) which attempted to characterize the presence of an asymmetry in the electron velocity distribution functions as detailed in section 3.2.2. However, this first algorithm lead to a large number of false positives, including a number of cases with large MeanRL values despite the absence of crescents. My idea was to create and train a Convolutional Neural Network making use of the full complexity of the distribution function instead of a reduced scalar, using crescent-shaped distributions from all EDRs found so far to automatically identify crescents in the distribution functions of phase 1b that had not been treated yet. The identification of crescents being a first strong signature that will indicate that other EDR signatures should be looked for at this time.

5.2 Creation of a training dataset

To create a training set for the model, I used the crescents of the EDR identified during phase 1 and reported in *Webster et al.* [2018] and *Lenouvel et al.* [2021], yielding a total of 214 crescents using data from the 4 MMS spacecraft. As for the first machine learning application of chapter 3, we have access to an almost unlimited amount of random distribution functions belonging to various plasma regions other than EDRs so getting data for the negative class is not a problem.

5.2.1 Transformation of the distribution function

The first step is to transform the data into a convenient form for the algorithm. The first obstacle is indeed that the distribution functions that we look at to identify crescent shapes are actually polar histograms in logarithmic scale whereas CNNs usually take a 2D matrix of pixel intensity as an input. Some work is necessary in order to make the data usable by the model.

The CDF file that we get from MMS to make the distribution function is a heavy file containing for each time step : 32 values of electron fluxes, 33 values defining the energy bins and 33 values defining the angle bins. However, not all of these information are useful to look for crescent shapes, only the flux values really matter because the energy bins and the angle bins only slightly change from a time step to another. Thus, assuming these values to be constant for every distribution function is acceptable if we prioritize the geometry, which is what the CNN does anyway.

The 32 bins of energy and angle are very convenient as it allows us to straightforwardly transform our polar histogram into a Cartesian histogram of 32 pixels by 32 pixels where each pixel holds the value of the electron flux for a given angle and energy range. To visually identify a crescent shape, we look for colored crescent patterns so the question of the transformation of flux values has to carefully be addressed as well.

When working with CNNs, it is convenient to work with black and white images because colored RGB images have 3 channels so they take longer to process and it also means that the model will be more complex. Since the logarithm of the phase space density values I use are always capped between -29 and -26, I mapped the lowest flux value to a pixel intensity value of 0, corresponding to a black pixel, and the highest flux value to a pixel intensity value of 255, corresponding to a white pixel. The white bins of the polar distribution functions are ”not a number (NaN)” values which needed to be dealt with so I decided to replace them by the lowest flux value corresponding to a black pixel in the end.

Figure 5.1 shows the transformation of the polar distribution function into the Cartesian image for distribution functions with and without a crescent shape. Going from the top to the bottom of the cartesian image means going to higher energies and going from left to right means going in a circular motion on the polar plot so black and white pixels on the left or right edge are actually neighbors in the polar plot. The cut between left and right was chosen so that crescents on the right of the polar image would be in the center of the cartesian image and we see that the crescent shape in the bottom polar image can be well identified in the cartesian image. We can thus be confident that identifying crescents in the black and white image is doable by the CNN.

It is important to mention that the polar distribution functions shown in figure 5.1 and in the rest of this thesis only show the first 18 bins of energy for convenience of visualization (corresponding roughly to an energy range of [0 eV, 700 eV]) because the energy bins are

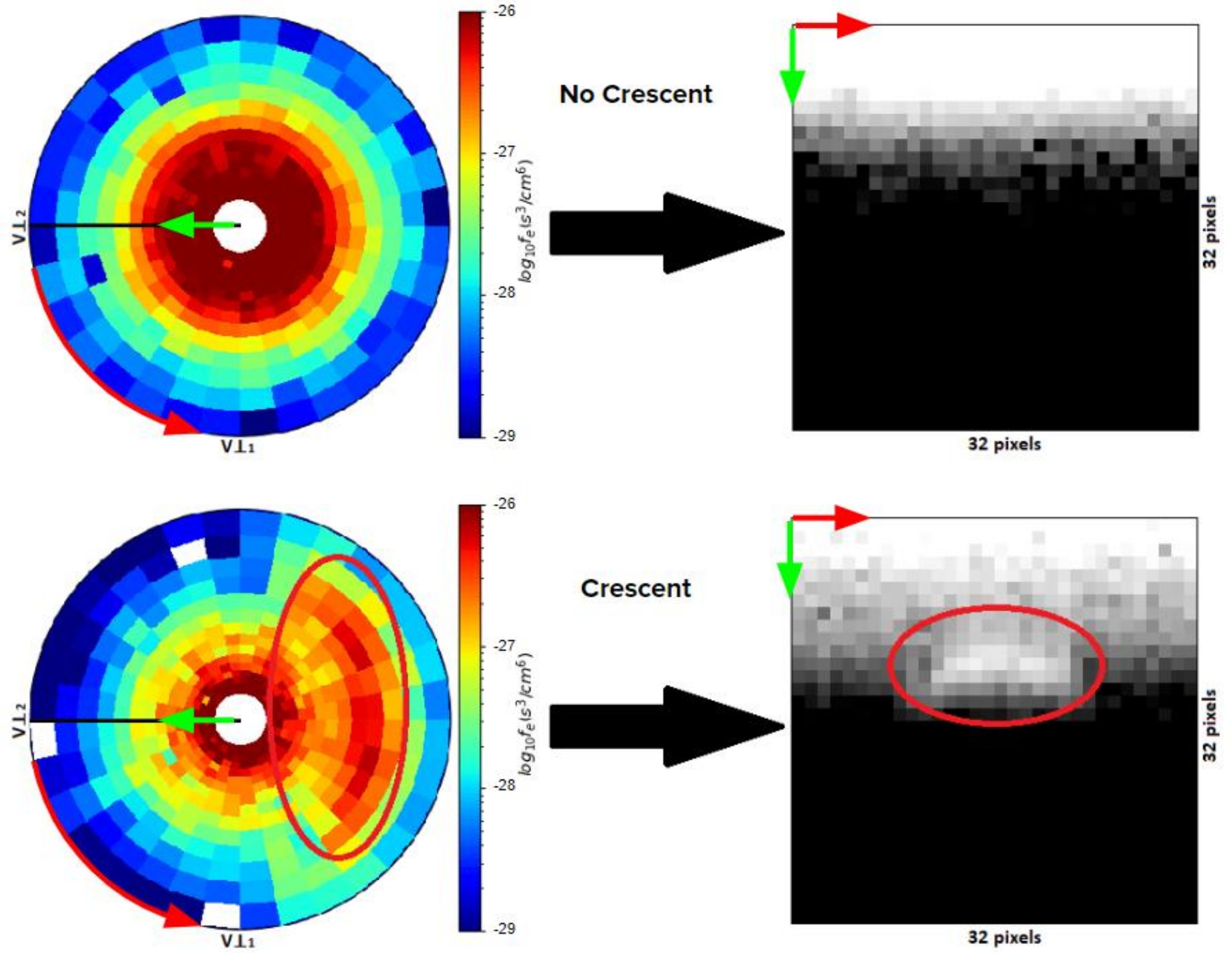


Figure 5.1: Transformation of the polar distribution function into the cartesian image for distribution functions with and without a crescent shape. The green arrow shows the correspondence between a movement to higher energies in the polar plot and in the cartesian plot, the red arrow shows the correspondence between a circular movement in the polar plot and a horizontal movement in the cartesian plot.

logarithmic so as we get further away from the center, we look at higher energy bins that get larger and are either empty or almost empty. Showing all the energy bins would not be useful because the area containing the crescent would be very small and crescents have never been observed at these higher energies.

Using this transformation, I was able to transform the 214 crescents in the polar plots into 214 black and white cartesian images. However, 214 images for the class we are trying to identify is absolutely not enough to train a CNN so data augmentation was needed in order to build a training dataset.

5.2.2 Data augmentation

Artificially augmenting a dataset composed of images is usually not a complicated task. If we take a picture of a dog as an example, many operations can be done on the image such as rotation, translation, cropping, mirroring and resizing to produce a very high number of new images that will still be representing a dog. In our case, data augmentation is more complicated because we are looking for crescent shapes in the cartesian black and white image. Resizing or cropping the crescent in the cartesian image is not really doable, rotating it in another direction or translate it on the image is not something we want because it would break the correspondence between the polar and Cartesian representations, making the training of the model potentially harder for no reason. This could also become contradictory with the physics of the EDR, since the location and orientation of the crescent is important and related to the structure of the EDR itself. Only horizontally mirroring the crescent is possible which would only double our initial 214 images and would still be insufficient.

I tried a different approach which is to create artificial crescents by making the assumption that combining 2 real observed crescent shapes should yield a realistic crescent shape. This is a reasonable assumption, since this will create crescents whose orientation and energy limits are in the range of the observed crescents, and not create artificial crescents whose orientation and energy range are unrealistic and outside the ranges observed. This data augmentation method is not possible with traditional CNN images of dogs for example and has the advantage of allowing the generation of many new images while being an interesting take on data augmentation techniques.

To achieve the creation of synthetic crescents, I first made a selection of 112 crescents out of the 214 that I found to be the most ”beautiful” to ensure that the combination of two good crescents of the selection would always make a good new crescent. The quantitative selection criterion was to have a very clear left/right asymmetry with a high contrast of color so that they would be unarguably classified as exhibiting a crescent by a human observer. This of course introduces a strong selection bias for ”strong” crescents but this is acceptable since we want to limit the false positives. Each of these 112 crescents was to be mixed with every other crescent yielding a total of $C_2^{112} = 6126$ new synthetic crescents. Each crescent undergoes a series of transformation before the mixing to break the linearity and ensure that each new image is a new information for the model. The transformations and mixing of the crescents are done the following way :

- The matrices taken to apply the transformations on and the mixing contain the logarithms of the electron phase space densities.
- A random small angular rotation of 0 or 1 bin is applied on each matrix corresponding to either a horizontal movement on the left of one pixel in the cartesian image, a horizontal movement on the right, or no movement at all. Such a small movement ensures that the original information is not changed too much.
- Logarithmic noise is added to each matrix. Each original flux value gets multiplied by a random value taken from a logarithmic normal distribution of mean = 0.5 and of

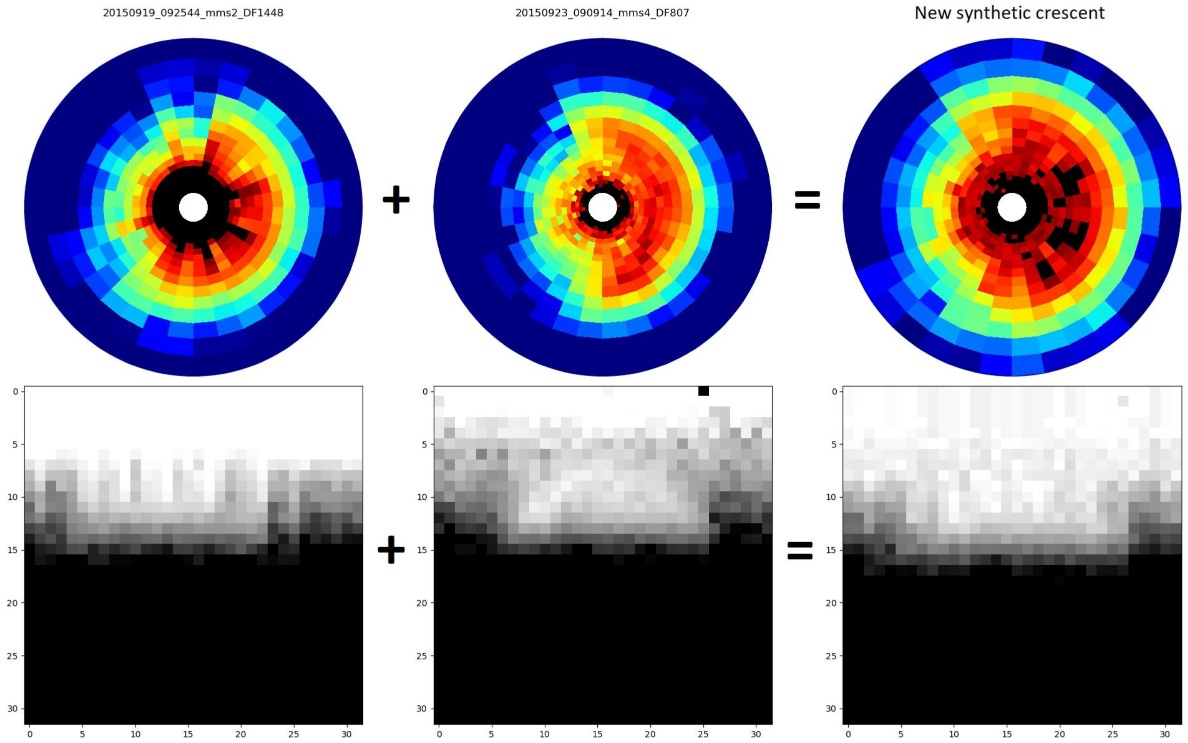


Figure 5.2: Figure showing the result of the combination of 2 electron velocity distribution functions exhibiting crescent-shaped distributions to create new synthetic crescents. The top row shows the polar representations of the distribution functions while the bottom row shows the corresponding cartesian representation of each polar distribution function.

sigma = 0.25. These values were chosen because they yielded good results after some trial and error.

- Both matrices are combined by averaging the values element-wise.
- The corresponding polar representation of the distribution function is also produced as a way to control if the result of the combination is satisfying, meaning that the crescent shape could indeed be potentially observed in real data and that it would be unarguably classified as a crescent by a human observer.

Figure 5.2 shows the result of one of the 6126 crescent combinations. We can see that thanks to the transformations applied before the combination, the new crescent produced can be considered as new data as it is not a simple averaging of the original crescents.

5.3 Training and evaluation of the model

Thanks to the data augmentation step, there was finally enough data to train a CNN. The first step was to divide the data into a training, validation and a test set. Since one wants to avoid data leakage between the training set and the test set so that the results are not biased, I chose to put the 102 crescents that were not used during the data augmentation step into the test set and added 102 random distribution functions (without a crescent) to get a number of 204 test samples equally representing both classes. These 102 crescents were not showing clear and unarguable asymmetry so one could be expecting the model to struggle on this test set. The total number of samples in the training set was 10125 and the total number of samples in the validation set was 2531 which corresponds to a 80%/20% split between the training and the validation.

The choice of the architecture of the model was challenging. In most standard CNN applications, the architecture is a succession of convolution and pooling layers with an increasing number of filters. This way, the model is able to extract more and more small scale features as the pooling layers decrease the resolution of the image. In our case, the structures we are looking for in the image are almost of the size of the image so the number of filters has to be higher at the beginning and must decrease as the resolution of the image decreases.

The final architecture is the following :

- Convolution layer with 128 filters of 3x3 kernels and with a ”relu” activation function
- 2D max pooling layer with a 2x2 kernel and the padding parameter ”same”
- Convolution layer with 64 filters of 3x3 kernels and with a ”relu” activation function
- 2D max pooling layer with a 2x2 kernel and the padding parameter ”same”
- Convolution layer with 32 filters of 3x3 kernels and with a ”relu” activation function
- Global 2D max pooling layer (making the transition between the 2D and the 1D architectures)
- Dropout layer with a 0.1 probability (to avoid overfitting)
- Dense layer of 25 neurons with the ”relu” activation function
- Dense layer of 25 neurons with the ”relu” activation function
- Dense layer of 1 neuron with the ”sigmoid” activation function (to make the binary classification)

The total number of trainable parameters is 95037 which is very small compared to the millions of parameters of complex CNNs.

The optimizer chosen is ”adam”, the loss is keras’ ”binary crossentropy”, the metric is ”binary accuracy” and the batch size chosen for the training is 128. To avoid overfitting during the training, I used an early stopping method based on the monitoring of the validation loss. The training lasted only 4 epochs with the early stopping enabled. Figure 5.3 shows the training curves of the model. We see that the model would overfit quite rapidly without early stopping. Such a low number of epochs for the training is not a problem and can be explained by the low number of images to train the CNN (compared to the millions of images sometimes required to train larger CNNs) and the low complexity of these images (only one color channel and the pattern searched by the CNN is not as complex as searching

Table 5.1: Confusion matrix of predictions of the model on the test set

	Predicted ”Crescent”	Predicted ”No Crescent”
True ”Crescent”	90	12
True ”No Crescent”	1	101

Table 5.2: Metrics corresponding to the predictions of the model on the test set

Class	Number of samples	Precision	Recall	F1-score
Crescent	102	98.90%	88.24%	93.26%
No Crescent	102	89.38%	99.02%	93.95%

for an animal shape for example).

Then the evaluation of the model is performed by looking at the predictions on the test set that had never been seen by the algorithm before. Tables 5.1 and 5.2 show the results of the model on the test set. We see that the recall is lower than the precision hinting that the model may miss a few crescents (which is coherent with the data augmentation strategy) without showing too many false positives.

To evaluate the model, we look at the predictions of the algorithm and compare them to the true labels to compute a score. However, the model does not give a class but a number between 0 and 1, and the standard thing is to assume the delimitation between the 2 classes to be at 0.5, which is what was done for the results shown in tables 5.1 and 5.2. But there is another thing that can be done to enhance the results of the model which is to choose the prediction threshold that gives the best results. Indeed, depending on the data we have, 0.5 may not be the best delimitation for our problem.

Figure 5.4 shows the precision and recall curves as a function of the decision threshold. We see that as the threshold gets closer to 1, the precision drops to 0 and the recall goes to 1 because a threshold of 1 means every image gets classified as a crescent. The best decision threshold is chosen to be the one that gives the smallest difference between precision and recall. That value was found to be 0.147 and is located at the dashed vertical black line on the plot. At this decision threshold, the recall and precision are of about 98%, while keeping the decision threshold at 0.5 would reduce the precision and the recall leading to much more false positives.

Using this value as decision threshold, tables 5.3 and 5.4 show the results on the training set with the custom decision threshold values of 0.147. We see that the results are much better and this is why this decision threshold is used for the predictions on the new MMS data.

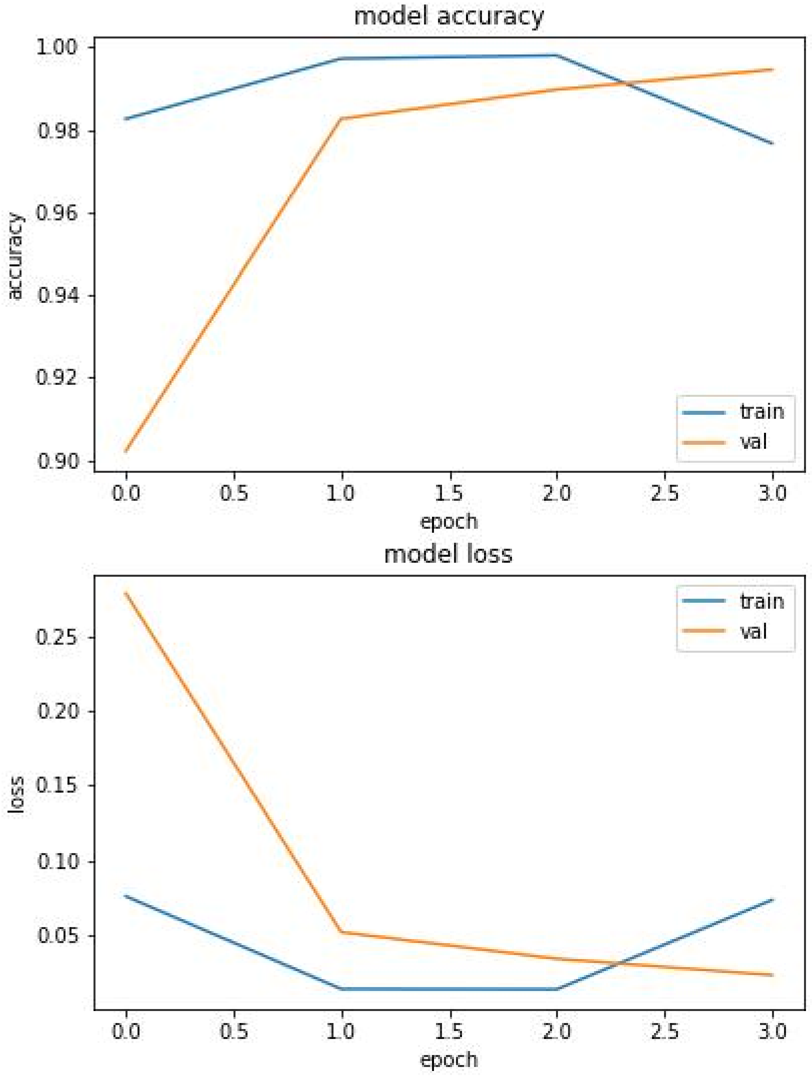


Figure 5.3: Figure showing the training curves of the crescent recognition CNN model. The top panel shows the accuracy of the model during the training and the bottom panel shows the evolution of the loss during the training. The accuracy and the loss are also shown for the validation set that the algorithm never trains on, which is computed after each epoch.

Table 5.3: Confusion matrix of predictions of the model on the test set with the custom decision threshold

	Predicted ”Crescent”	Predicted ”No Crescent”
True ”Crescent”	100	2
True ”No Crescent”	2	100

Table 5.4: Metrics corresponding to the predictions of the model on the test set with the custom decision threshold

Class	Number of samples	Precision	Recall	F1-score
Crescent	102	98.04%	98.04%	98.04%
No Crescent	102	98.04%	98.04%	98.04%

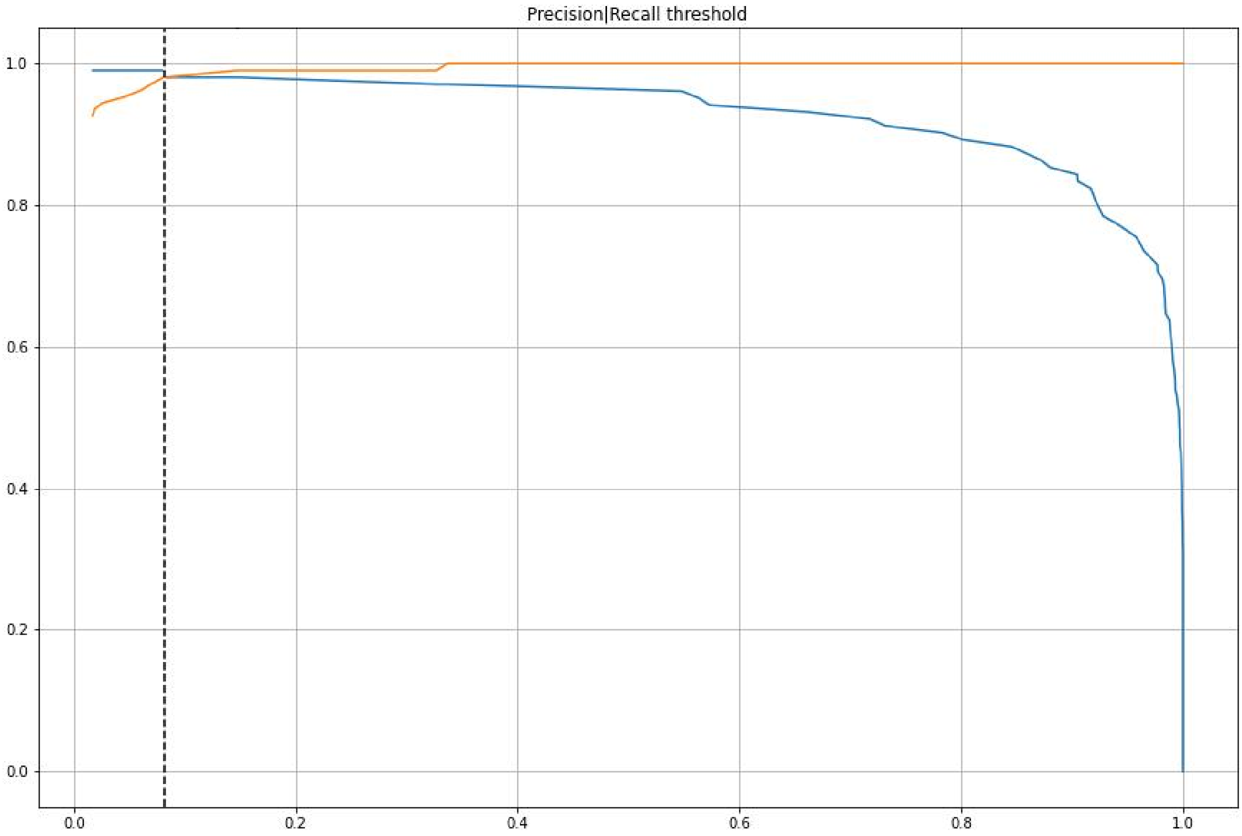


Figure 5.4: Figure showing the precision and recall curves as a function of the decision threshold chosen. The blue line is the precision and the orange line is the recall. The dotted vertical line is the location of the chosen threshold value.

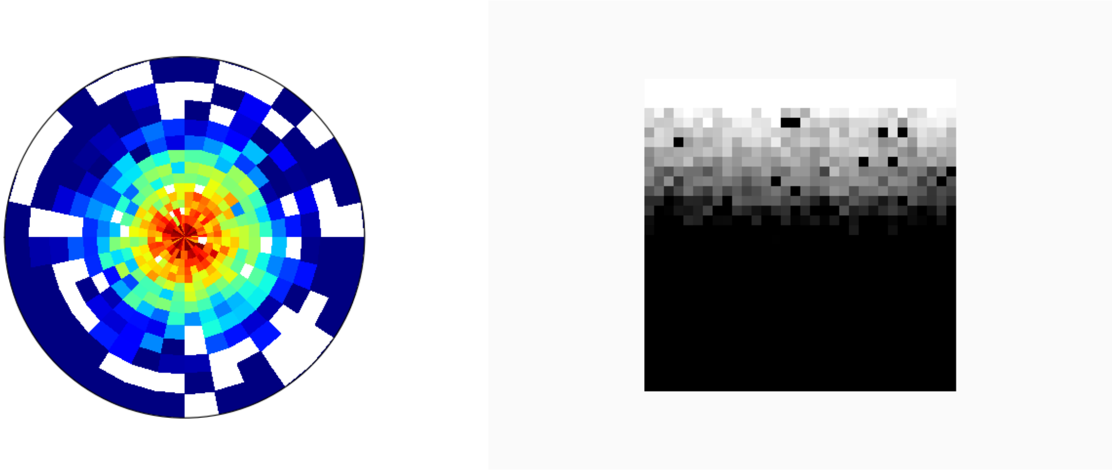


Figure 5.5: Figure showing an example of false positive crescent in the eVDF (Cartesian coordinates on the left, polar on the right) detection by the CNN that can not be easily explained.

5.4 Predictions on new MMS data of phase 1b

The CNN model being trained, the goal was then to identify potential new EDRs in the data of phase 1b of MMS which goes from November 26th 2016 to January 29th 2017. I used the list of magnetopause crossings from the ISSI team (*Paschmann et al.* [2018]) to only retrieve data from time intervals labeled as ”magnetopause crossing”. For each time step of these time intervals, a black and white cartesian distribution function image was produced for the model to classify resulting in 31 038 394 images. The method used to produce these images is the same as detailed previously.

Predictions were only done on MMS 1 and 2 because of the size of the data. The model classified 7487 images as ”Crescents” which makes a detection rate of 0.024%. Based on the number of EDRs expected to be observed by with MMS (*Fuselier et al.* [2016]), it is impossible to have so many real crescents in the data and many detections are indeed false positives. I had to manually review each image to sort the good and the bad predictions which was easy because most of the bad predictions can be discarded in less than a second by visual analysis. Some false positives such as shown in figure 5.5 can not be easily explained and may be due to a proximity to the decision threshold. Some other false positives can very easily be understood such as when a lot of data is missing, making a lot of black pixels on the cartesian image that can produce the same pattern as if there were a crescent. An example of that is shown in figure 5.6. A post-treatment of the detections was considered at some point with for example the exclusion of the positive crescent detections when the polar eVDF had too many NaN values. However, no satisfying threshold could be found so no post-treatment was applied as the number of positive crescent detections could still easily be manually checked in a short time and removing the post treatment removed the chances of discarding a few good detections that could lead to new EDR identifications.

After visual inspection of the images classified as crescents by the model, I labeled 335 out of the 7487 images as ”true positives” interesting enough to be investigated to look

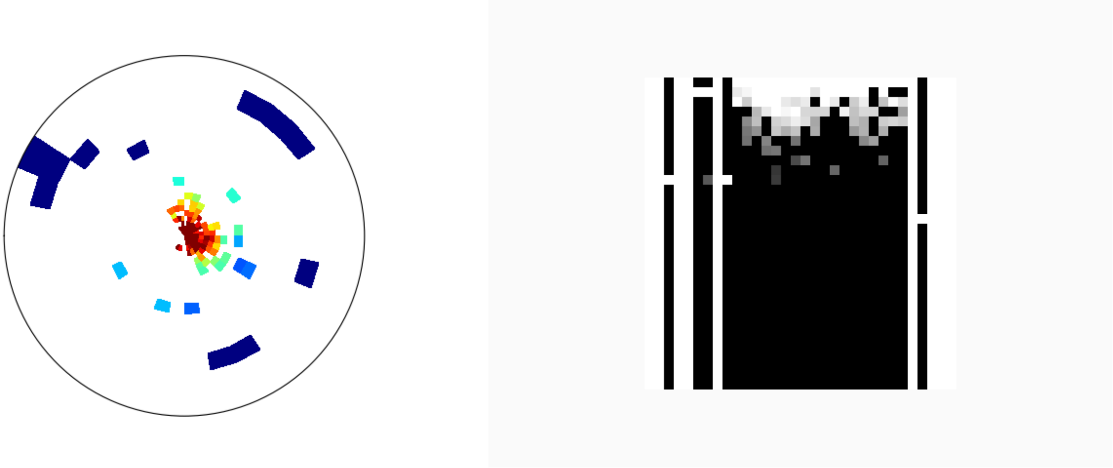


Figure 5.6: Figure showing an example of false positive crescent detection by the CNN that can be easily explained by high amount of missing data, producing the same pattern as if there were a crescent.

for new EDRs because, as mentionned in section 2.4, the presence of an asymmetry in the distribution function does not necessarily mean the presence of an EDR as these can also be found near separatrices for example. These 335 distribution functions came from MMS 1 and 2 which represented 66 individual events with possible multi-spacecraft detections. I inspected each event to look for EDR signatures other than the presence of electron crescents such as :

- Low magnetic field B
- High energy dissipation $J \cdot E$
- High deviation from the frozen-in condition : $E + V_e \times B \neq 0$

The final list of selected EDR candidates contains 17 events based on these criteria and is presented in table 5.5. These new events have not been published yet and are presented for the first time in this thesis.

The fact that I searched for crescents first and then confirmed or not the presence of an EDR makes it so it is almost certain that there are other EDRs in the phase 1b of MMS that have not been found by this algorithm. There are two reasons for that : first, the model is not perfect so it is certainly possible that some crescents were missed by the model, and the second reason is that by looking for crescents, I inevitably discard potential high guide field EDR cases that would not exhibit the crescent-shaped distributions.

5.5 Case study of one new EDR candidate

In this section, I will present in detail one of the new EDR candidates found using the electron crescent detection CNN and explain the reasons that make me say that this event

Table 5.5: Table showing the new EDR candidates found during phase 1b of MMS using the electron crescent detection CNN.

Event	**Date**	**Time (UTC)**	**MMS**	**Author**
L19	2016/09/28	14:25:47	MMS2	Lenouvel (unpublished)
L20	2016/09/28	18:39:27	MMS1 MMS2	Lenouvel (unpublished)
L21	2016/09/28	18:43:44	MMS2	Lenouvel (unpublished)
L22	2016/10/13	11:24:28	MMS2	Lenouvel (unpublished)
L23	2016/10/13	11:29:53	MMS1	Lenouvel (unpublished)
L24	2016/10/16	17:37:13	MMS1	Lenouvel (unpublished)
L25	2016/10/22	13:01:31	MMS2	Lenouvel (unpublished)
L26	2016/11/13	15:25:34	MMS1 MMS2	Lenouvel (unpublished)
L27	2016/11/14	13:13:20	MMS1 MMS2	Lenouvel (unpublished)
L28	2016/11/27	06:45:40	MMS1 MMS2	Lenouvel (unpublished)
L29	2016/11/30	13:05:18	MMS2	Lenouvel (unpublished)
L30	2016/12/23	02:53:10	MMS1 MMS2	Lenouvel (unpublished)
L31	2016/12/26	11:20:29	MMS2	Lenouvel (unpublished)
L32	2016/12/29	11:56:11	MMS2	Lenouvel (unpublished)
L33	2017/01/07	08:09:55	MMS1 MMS2	Lenouvel (unpublished)
L34	2017/01/09	02:27:31	MMS2	Lenouvel (unpublished)
L35	2017/01/27	00:57:07	MMS1	Lenouvel (unpublished)

is indeed an EDR.

The EDR candidate chosen for this case study is L30. On December 23rd 2016, MMS 1 and 2 (and potentially MMS3 and MMS4) encountered a current sheet at 02:53:10 UTC while crossing the magnetopause going from the magnetosheath to the magnetosphere, with a reversal of the B_z component of the magnetic field. Figure 5.7 shows a large overall plot of the event seen by MMS1 produced by the MMS science data center. We can see that the particle density is of about 15 cm^{-3} in the magnetosheath and drops below 5 cm^{-3} upon entering the magnetosphere a first time after crossing the current sheet. During the crossing, the magnetic field intensity rapidly rises from about 10 nT to about 30 nT. MMS then briefly reenters the magnetosheath with the particle density reaching 10 cm^{-3} before going back into the magnetosphere with the density finally dropping to about 1 cm^{-3}. Electron and ion spectrograms also show the increased presence of high energy particles at that time correlated with a temporary drop in the magnetic field intensity, confirming MMS going back and forth between the magnetosheath and the magnetosphere.

Figures 5.8 and 5.9 (zoom) show, in addition to figure 5.7, the $E + V_e \times B$ and $J \cdot E$ parameters, respectively describing the departure from the frozen-in condition and the energy conversion. The MeanRL parameter described in section 3.2.2 is also shown, and we see that the value reach a peak at about 1.75, meaning that this case would have been filtered out by the post-processing of the first EDR detection algorithm from chapter 3. We see that at the time where the crescent-shaped distribution is shown, there is a significant rise of the $E + V_e \times B$ parameter, reaching about 10 mV/m. Higher values can be seen on the large scale plot in the magnetosphere but these values are probably wrong due to the large uncertainty on the electric field measurement when the particle counts are low, producing these very noisy curves.

At the same time than the $E + V_e \times B$ quantity rises, we see a large positive spike in the $J \cdot E$ parameter at about $5000pW/m^3$ indicating a high transfer of energy from the magnetic field to the particles. No negative energy conversion is observed indicating that this event is a classic EDR and not an outer EDR case (see the discussion about inner/outer EDRs in section 3.4). Only one distribution function is shown here but 2 other crescents were observed on MMS1 $30ms$ before and $30ms$ after the presented distribution function.

At the time of the crossing, the magnetic field is of about $23nT$ which can seem a bit high for an EDR but the previous study on the first reported EDRs presented in section 3.3.2 show that these magnetic field intensity values are possible for EDRs. These are all the reasons hinting that this event is a good new EDR candidate.

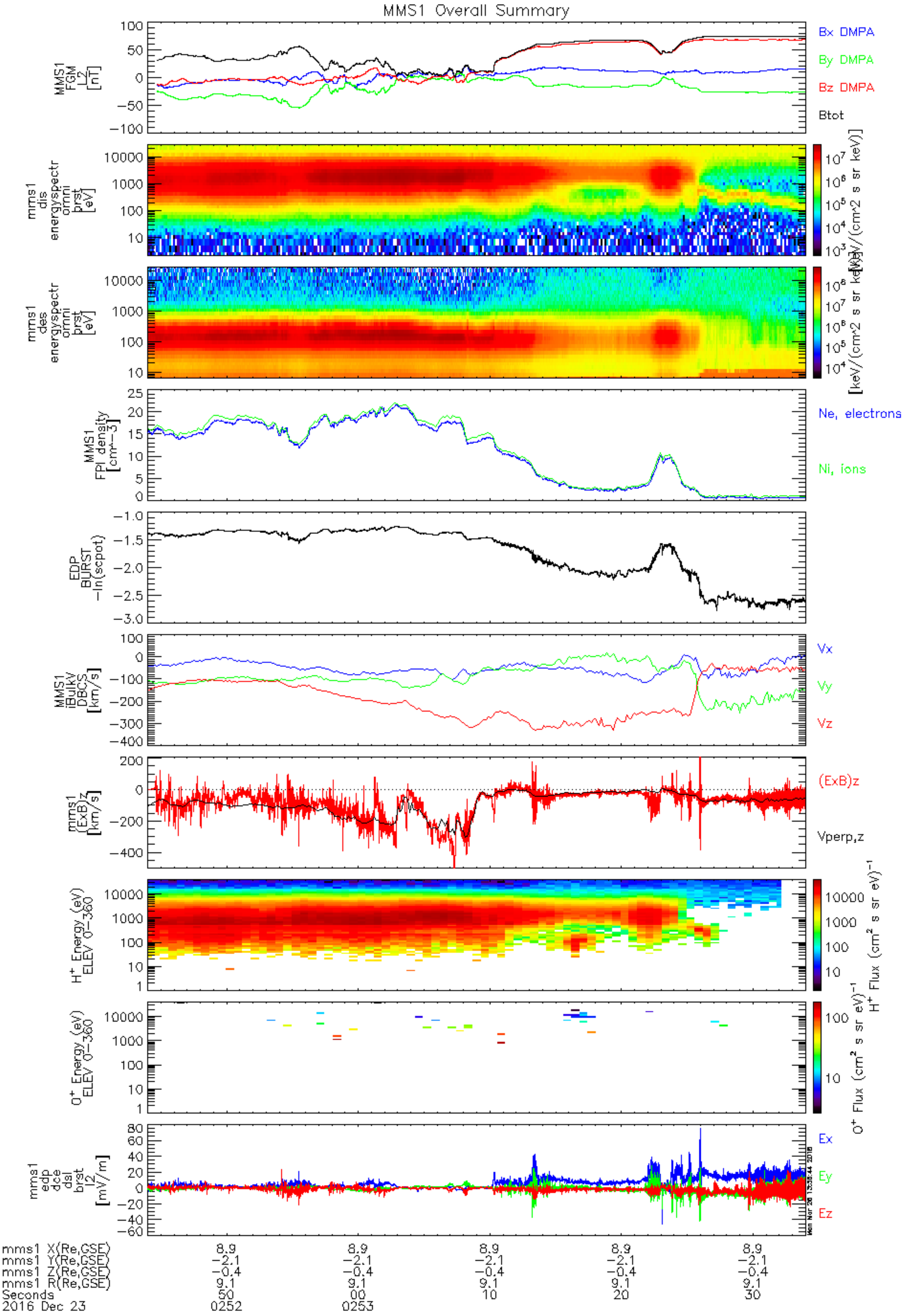


Figure 5.7: Overall plot produced by the MMS science data center of the EDR candidate event from December 23rd 2016 seen by MMS1.

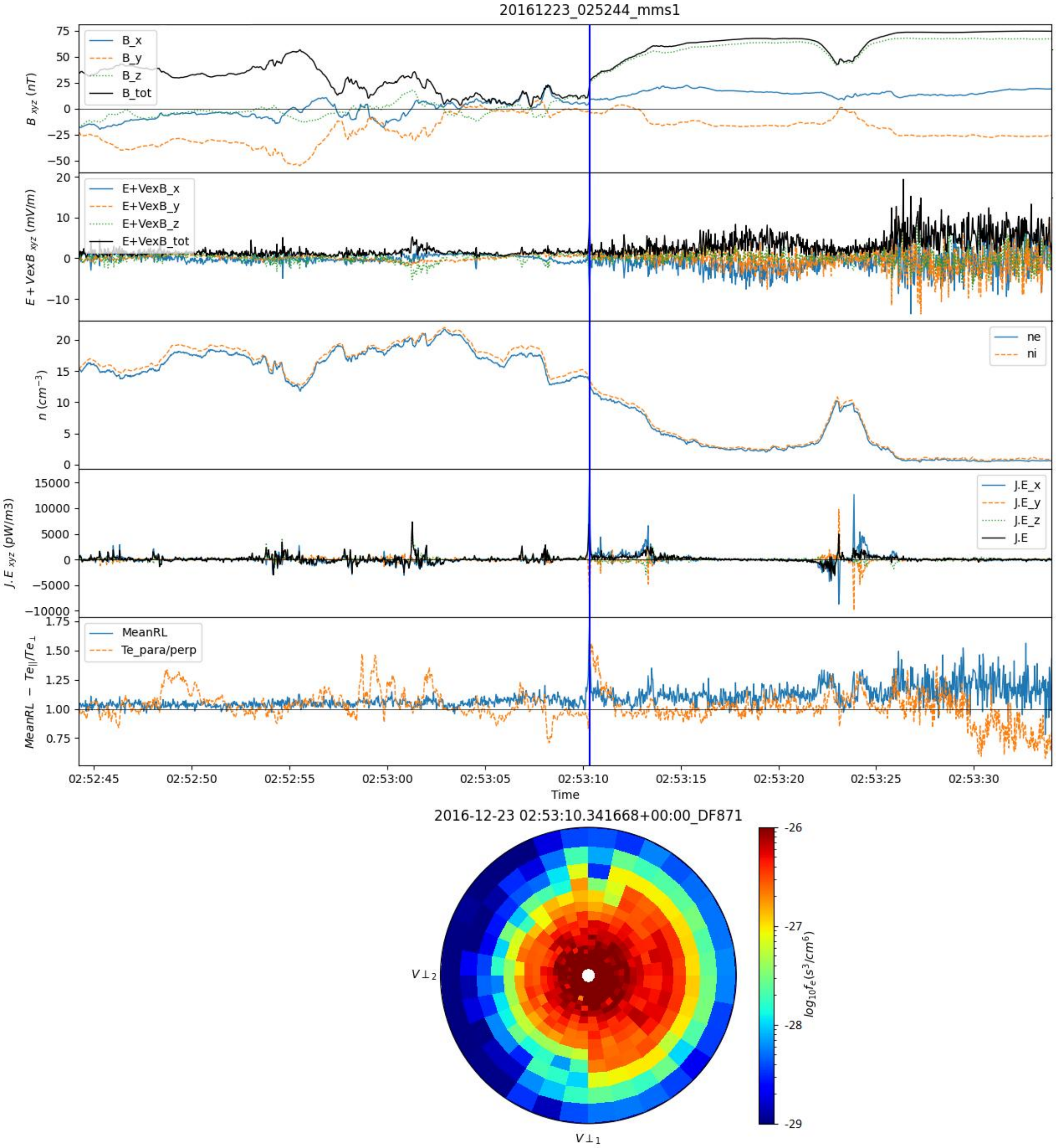


Figure 5.8: Large scale plot of the the EDR candidate event from December 23rd 2016 seen by MMS1 showing the location of the detected crescent-shaped electron velocity distribution function.

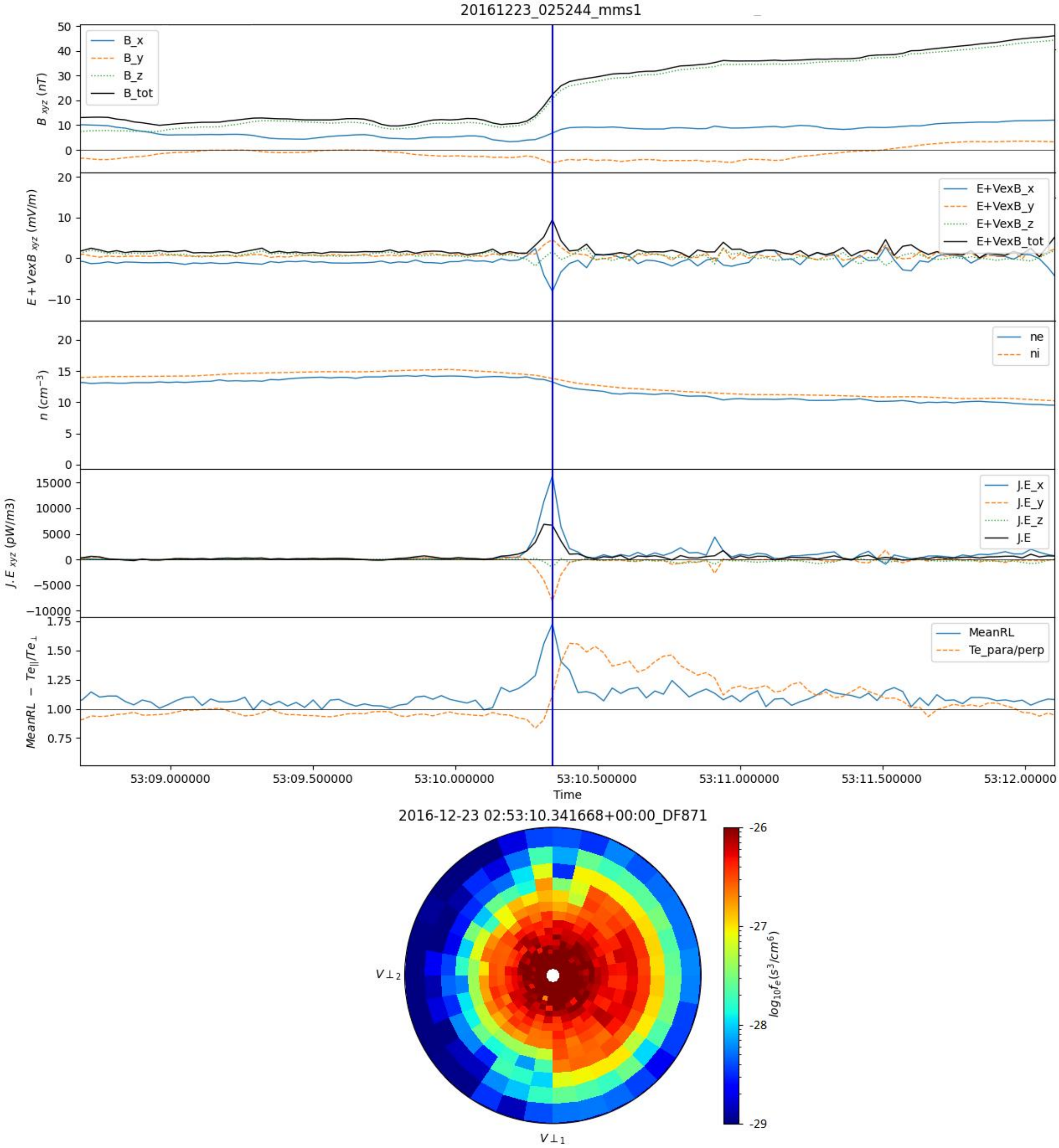


Figure 5.9: Small scale plot of the the EDR candidate event from December 23rd 2016 seen by MMS1 showing the location of the detected crescent-shaped electron velocity distribution function.

5.6 Complete list of identified dayside EDR during phase 1 of MMS

Tables 5.6 and 5.7 show the list of all the identified dayside EDR candidates during the phase 1 of MMS. This list includes previously reported EDRs in the literature as well as the new candidates found by my two algorithms, all sorted in chronological order. Events labeled with the letter A or B followed by two digits are events regrouped in *Webster et al.* [2018], events with the letter A or B followed by roman numerals are events that were not listed in *Webster et al.* [2018] and events labeled with the letter L were either identified by the first algorithm and presented in *Lenouvel et al.* [2021] or found by the second algorithm and marked as ”unpublished”.

Conclusion

In this section, I presented a way to identify electron crescents using a convolutional neural network as an enhancement of the previous ”MeanRL” parameter. This new model led to the discovery of 17 new EDR candidates during phase 1b of MMS. The final list shows 71 EDRs with 35 of them being ”EDR candidates” identified with my two different AI approaches. This number represents a proportion of almost 50% of the total number of EDRs identified during phase 1 of MMS. Enhancing the number of identified EDRs to be studied by the community was one of the goal of my thesis work as it would allow future new discoveries to be made about the complex physics occurring at the EDR.

Table 5.6: Dayside EDRs reported during Phase 1a of MMS.

Event	**Date**	**Time (UTC)**	**MMS**	**Author**	**Status**
AI	2015/09/08	11:01:20	MMS3	*Eriksson et al.* [2016]	Reported
L01	2015/09/09	08:43:58	MMS3	*Lenouvel et al.* [2021]	Candidate
L02	2015/09/09	13:00:22	MMS4	*Lenouvel et al.* [2021]	Candidate
L03	2015/09/09	14:28:51	MMS3	*Lenouvel et al.* [2021]	Candidate
L04	2015/09/11	10:13:35	MMS3	*Lenouvel et al.* [2021]	Candidate
L05	2015/09/14	16:16:34	MMS2	*Lenouvel et al.* [2021]	Candidate
A01	2015/09/19	07:43:29	MMS4	*Chen et al.* [2016]	Reported
L06	2015/09/19	09:25:44	MMS3	*Lenouvel et al.* [2021]	Candidate
L07	2015/09/22	13:41:31	MMS3	*Lenouvel et al.* [2021]	Candidate
L08	2015/09/23	09:09:38	MMS4	*Lenouvel et al.* [2021]	Candidate
L09	2015/10/01	06:53:43	MMS2	*Lenouvel et al.* [2021]	Candidate
L10	2015/10/06	14:17:54	MMS1	*Lenouvel et al.* [2021]	Candidate
A02	2015/10/16	10:33:30	MMS1	*Norgren et al.* [2016]	Reported
A03	2015/10/16	13:07:02	MMS1	*Burch et al.* [2016]	Reported
A04	2015/10/22	06:05:21	MMS1	*Phan et al.* [2016]	Reported
A05	2015/11/01	15:08:06	MMS3	*Fuselier et al.* [2017]	Reported
A06	2015/11/12	07:19:21	MMS1	*Fuselier et al.* [2017]	Reported
L11	2015/12/02	01:17:02	MMS1	*Lenouvel et al.* [2021]	Candidate
A07	2015/12/06	23:38:31	MMS2	*Khotyaintsev et al.* [2016]	Reported
A08	2015/12/08	11:20:44	MMS1	*Burch and Phan* [2016]	Reported
A09	2015/12/09	01:06:11	MMS4	*Webster et al.* [2018]	Reported
AII	2015/12/13	09:43:18	MMS1	*Dong et al.* [2021]	Reported
A10	2015/12/14	01:17:39	MMS1	*Chen et al.* [2017], *Ergun et al.* [2017], *Graham et al.* [2017]	Reported
L12	2016/01/02	23:46:17	MMS4	*Lenouvel et al.* [2021]	Candidate
A11	2016/01/07	09:36:15	MMS3	*Webster et al.* [2018]	Reported
L13	2016/01/07	22:13:20 22:11:37	MMS3 MMS4	*Lenouvel et al.* [2021]	Candidate
A12	2016/01/10	09:13:37	MMS4	*Fuselier et al.* [2017]	Reported
L14	2016/02/05	22:19:45	MMS1	*Lenouvel et al.* [2021]	Candidate
L15	2016/02/11	02:01:04	MMS1 MMS3	*Lenouvel et al.* [2021]	Candidate
L16	2016/02/14	20:41:56	MMS2 MMS3 MMS4	*Lenouvel et al.* [2021]	Candidate
L17	2016/02/19	18:42:38	MMS1 MMS3	*Lenouvel et al.* [2021]	Candidate
A13	2016/02/27	20:23:34	MMS1	*Fuselier et al.* [2017]	Reported
L18	2016/02/28	01:07:33	MMS1 MMS3	*Lenouvel et al.* [2021]	Candidate

Table 5.7: Dayside EDRs reported during Phase 1b of MMS.

Event	**Date**	**Time (UTC)**	**MMS**	**Author**	**Status**
L19	2016/09/28	14:25:47	MMS2	Lenouvel (unpublished)	Candidate
L20	2016/09/28	18:39:27	MMS1 MMS2	Lenouvel (unpublished)	Candidate
L21	2016/09/28	18:43:44	MMS2	Lenouvel (unpublished)	Candidate
L22	2016/10/13	11:24:28	MMS2	Lenouvel (unpublished)	Candidate
L23	2016/10/13	11:29:53	MMS1	Lenouvel (unpublished)	Candidate
L24	2016/10/16	17:37:13	MMS1	Lenouvel (unpublished)	Candidate
B14	2016/10/22	12:58:40	MMS2	*Webster et al.* [2018]	Reported
L25	2016/10/22	13:01:31	MMS2	Lenouvel (unpublished)	Candidate
B15	2016/11/02	14:46:18	MMS4	*Webster et al.* [2018]	Reported
B16	2016/11/06	08:40:58	MMS2	*Webster et al.* [2018]	Reported
B17	2016/11/12	17:48:46	MMS2	*Webster et al.* [2018]	Reported
B18	2016/11/13	09:10:40	MMS4	*Webster et al.* [2018]	Reported
L26	2016/11/13	15:25:34	MMS1 MMS2	Lenouvel (unpublished)	Candidate
L27	2016/11/14	13:13:20	MMS1 MMS2	Lenouvel (unpublished)	Candidate
B19	2016/11/18	12:08:11	MMS2	*Webster et al.* [2018]	Reported
B20	2016/11/23	07:49:33	MMS2	*Webster et al.* [2018]	Reported
B21	2016/11/23	07:49:52	MMS4	*Webster et al.* [2018]	Reported
B22	2016/11/23	07:50:30	MMS1	*Webster et al.* [2018]	Reported
L28	2016/11/27	06:45:40	MMS1 MMS2	Lenouvel (unpublished)	Candidate
BI	2016/11/28	07:36:55	MMS1	*Genestreti et al.* [2018b]	Reported
B23	2016/11/28	15:46:59	MMS2	*Webster et al.* [2018]	Reported
L29	2016/11/30	13:05:18	MMS2	Lenouvel (unpublished)	Candidate
B24	2016/12/11	04:41:49	MMS4	*Webster et al.* [2018]	Reported
B25	2016/12/19	14:15:01	MMS2	*Webster et al.* [2018]	Reported
L30	2016/12/23	02:53:10	MMS1 MMS2	Lenouvel (unpublished)	Candidate
BII	2016/12/24	15:03:32	MMS1	*Li et al.* [2020]	Reported
L31	2016/12/26	11:20:29	MMS2	Lenouvel (unpublished)	Candidate
L32	2016/12/29	11:56:11	MMS2	Lenouvel (unpublished)	Candidate
B26	2017/01/02	02:58:12	MMS2	*Webster et al.* [2018]	Reported
L33	2017/01/07	08:09:55	MMS1 MMS2	Lenouvel (unpublished)	Candidate
L34	2017/01/09	02:27:31	MMS2	Lenouvel (unpublished)	Candidate
B27	2017/01/11	04:22:43	MMS4	*Webster et al.* [2018]	Reported
B28	2017/01/20	12:32:07	MMS2	*Webster et al.* [2018]	Reported
B29	2017/01/22	10:15:46	MMS2	*Webster et al.* [2018]	Reported
B30	2017/01/22	10:15:58	MMS2	*Webster et al.* [2018]	Reported
B31	2017/01/22	10:47:33	MMS2	*Webster et al.* [2018]	Reported
L35	2017/01/27	00:57:07	MMS1	Lenouvel (unpublished)	Candidate
B32	2017/01/27	12:05:43	MMS1	*Webster et al.* [2018]	Reported

Chapter 6

Conclusions

The Magnetospheric Multiscale (MMS) mission has allowed for a deeper understanding of the underlying physical mechanisms of magnetic reconnection after the success of the CLUSTER mission. The unprecedented temporal resolution of the instruments on board the four spacecraft composing the MMS tetrahedron led to the first observation of an Electron Diffusion Region (EDR) in 2016, a region that could only be studied through simulations of magnetic reconnection until then. One of the instruments of the MMS mission named the Fast Plasma Investigation (FPI) allowed for the first time the observation of crescent-shaped electron velocity distribution functions (eVDF), theorized several years before and supposed to be a key signature of these regions based on simulations. This observation was one of the main discoveries of that first detection in 2016. Since then, other EDRs located at the magnetopause and at the magnetotail have been reported by the MMS community. The number of reported EDRs on the dayside was of about 30 in 2018, most of them were regrouped in *Webster et al.* [2018] which served as a base catalogue for many studies done by the MMS community. However, such a small number of events is far too small to understand the complexity of the core of the reconnection process, and in addition most of these studies only took into account a small portion of these EDR cases leading to a lack of generalization for some of the conclusions of these studies.

The first task of my thesis work has been to find new EDR candidates in order to enhance the number of available cases to study. The identification of these regions was done manually which was very time consuming due to the large amount of data made available by MMS. The idea of my thesis was to try to bring machine learning solutions to the search of EDRs. Artificial intelligence (AI) is indeed a tool that is getting more and more popular in the astrophysical community, as demonstrated by the growing number of studies involving AI methods in our community (see *Nguyen et al.* [2019] and *Breuillard et al.* [2020]), even though its use is still marginal compared to the other fields of physics and the other scientific domains. I thus developed a first algorithm to identify EDRs in the data of Phase 1a of MMS, the architecture used is that of a Multilayer Perceptron (MLP) that does point by point multiclass classification using several physical parameters such as the magnetic field, the electric field, the particle densities and the current as inputs. This algorithm used these inputs to predict if at the time, one given MMS spacecraft was in the magnetosphere, the magnetosheath, the ion diffusion region or in the electron diffusion region. One of the input features of this MLP called the ”MeanRL” parameter was created to identify inside the electron velocity distribution functions the presence of crescent-shaped distributions that are a key signature of the EDR. The ”MeanRL” parameter characterizes the presence of an asymmetry in the eVDF in the form of a scalar that can be plotted and viewed as a time series along with the other physical parameters given by MMS. This first algorithm allowed for the identification of 18 new EDR candidates that were reported in *Lenouvel et al.* [2021], bringing the number of known dayside EDRs to about 50. This number is therefore close to the 56 EDR encounters that was estimated in *Fuselier et al.* [2016] to be found at the dayside magnetopause during the 2.5 years of the nominal mission of MMS. The matter of the inner/outer EDR was also discussed, four times as many inner EDRs (positive energy conversion) as outer EDRs (negative energy conversion) were found in phase 1a, meaning that the outer EDR still represents a significant portion of the reconnection zone. The inertial and pressure terms of the generalized Ohm’s law were found to alternatively dominate each other depending on the event and the time analyzed, supposing the existence of two different mechanisms to explain the creation of the negative energy conversion regions.

This large EDR database was the starting point of several statistical studies I made in order to deepen our understanding of the electron diffusion region. I first studied a potential link between the solar wind parameters and the energy conversion at the EDR. After seeing that there were no correlations between the solar wind parameters and the sign of the energy conversion at the EDR, I looked at the influence of the IMF clock angle on the intensity of the energy conversion as well as the repartition of the IMF clock angles among the reported EDR cases. I observed that 69% of the EDRs were found with an IMF clock angle between -112.5 and +112.5 degrees showing the preferred southward orientation of the IMF to induce a reconnection event with the northward magnetospheric magnetic field. An interesting behavior of the EDR energy conversion versus IMF B_y was also found : the distribution of this parameter for the IDR labeled data points is homogeneous and representative of the data coverage, while the distribution of IMF B_y for EDR cases is highly inhomogeneous with mostly positive B_y cases. The origin of this trend is not completely clear yet but could be due to the modification of the location of the X-line by a combination of the dipole tilt and IMF B_y towards the subsolar point where the reconnection is expected to be more efficient according to *Reistad et al.* [2020]. The IMF B_y dependence of the energy conversion in the EDR will have to be confirmed with a broader EDR events list.

I then tried to find new small scale parameters that characterize EDRs and could be used to better identify them in time series data. The electron vorticity Ω_e was proposed as a new proxy parameter to identify the borders of the EDR in *Hwang et al.* [2019]. More precisely, they proposed to look at the condition that the electron vorticity exceeds the electron gyrofrequency ($\Omega_e > \omega_{ce}$) which theoretically makes the system transition into the EDR regime. My analysis of 46 of the reported dayside EDRs revealed that only the unique conditions of the event presented in *Hwang et al.* [2019] allowed this condition to be met as this condition was never met for any of the events I studied. However, a boxplot analysis of the parameter Ω_e/ω_{ce} showed that this parameter was generally higher for the data points of phase 1a classified in the EDR class by the first MLP algorithm than for data points classified in the 3 other classes. This parameter could thus be used with precaution to strengthen EDR classifications but it cannot be considered as a key signature of the EDR. The other parameters that were analyzed as potential new small scale parameters to identify the EDR are the electron Larmor radius $r_L = \frac{m_e V_{e_\perp}}{eB}$ and the magnetic gradient scale $L_0 = \frac{B}{\mu_0 J}$. At the EDR in the absence of significant guide field, the Larmor radius is supposed to be larger and the magnetic gradient scale smaller than in the other plasma regions (due to the expected small magnetic field) which is what was observed using boxplot visualization. These parameters thus seems relevant to identify EDRs with small guide fields. The last parameter studied is the adiabatic parameter κ introduced by *Lavraud et al.* [2016]. It indicates that the particle dynamics becomes chaotic when the values are below 10 which is expected at the EDR due to the demagnetization of the electrons. Only 4 of the EDR cases studied showed values below 10 but in one third of the cases, κ was the lowest at the EDR. This parameter may thus be used as a confirmation element if other key EDR signatures are associated with a low κ value but it is not reliable enough to be a key EDR signature.

The structure of the EDR and how it changes based on the physical configuration is a complex matter which I tried to study by making use of the large EDR database at our disposal. Based on a conjoined study by *Genestreti et al.* [2017] and *Cassak et al.* [2017] where they found that the location of the energy conversion region gets closer to the X-line as the guide field increases, I determined the location of the maximum energy conversion rate with respect to structural points of the reconnection region : the $B_L = 0$ point, the electron crescent point and the electron stagnation point. A large dispersion of the relative positions of the points of interest was observed, with distances that are of the order of a fraction of ion inertial length. No clear trends were found regarding the evolution of the distance between these structural points and the location of the maximum energy conversion rate as the guide field or the electron density asymmetry increased. In addition, the confidence intervals of the linear regressions performed told us that the results are not robust enough to allow for definitive conclusions to be drawn. We thus found that the determination of the points of interest is complex and their definition can not be done in a simple manner.

The distance between the X-line and the electron stagnation point was also investigated. The study by *Cassak et al.* [2017] on the matter showed that the maximum of the energy conversion is located halfway between the X-line and electron stagnation point on the magnetospheric side. As the guide field increases, the electron stagnation point gets closer to the X-line and the maximum of the energy conversion moves closer to the electron stagnation point. I used the analytical formula for the distance between the X-line and the electron stagnation point in the N direction proposed by *Cassak et al.* [2017] and compared the results with a manual measurement of the distance between the X-line and the electron stagnation point in the N direction from a time separation method applied on the EDR cases that have been reported. One should bear in mind that the determination of the stagnation point is hard and prone to important uncertainties. To be able to compare both methods, the distances obtained from the time separation method are normalized by the width of the current sheet which can be computed using a formula taking the presence of a guide field into account or not. We observed that the distances computed with the theoretical formula and the distances measured with the time separation method and normalized in the zero guide field limit are close. However, computing the current sheet width with the formula including the guide field gives distances that are always bigger than in the two other cases.

Cassak et al. [2017] also predicted that the max energy conversion point is always located between the X-line and the electron stagnation point and the maximum energy conversion point gets closer to the stagnation point (and further away from the X line) as the guide field increases. I thus decided to look at the distance between the X-line and the maximum energy conversion point and how it relates to the location of the stagnation point. We observed that with our method, the maximum energy conversion point is beyond the electron stagnation point for 8 of the 10 EDR cases studied, and the proximity between the maximum energy conversion point and the electron stagnation point does not follow any pattern as the guide field increases. A possible trend regarding the distance between the max $\mathbf{J}\cdot\mathbf{E}$ point and the X-line was however observed, which seems to increase as the guide field increases, with the lowest distances only at low guide field ratios (below 0.15), which is in agreement with the expectations by *Cassak et al.* [2017].

The nature of the energy conversion was also the subject of two studies. The energy conversion at the EDR was the first of the two to be investigated using the EDR database. Previous studies showed that the perpendicular contributions to the energy conversion should dominate until moderate guide field conditions but above this threshold the parallel reconnection electric field should increase and the parallel contribution of energy conversion should dominate. This was verified by *Wilder et al.* [2018] on a small number of IDR events and current sheets but I aimed at extending this paper results to our list of EDR events. I found that below $B_g = 0.5$, which is the threshold above which the parallel contribution is supposed to dominate over the perpendicular one, the ratio of events with an energy conversion dominated by a perpendicular contribution and those with a parallel contribution dominance is the same as that of the figure from *Wilder et al.* [2018] with a majority of negative ratios and thus of larger perpendicular rates. This result confirms that below $B_g = 0.5$, the dissipation seems to be driven by the perpendicular electric field more than the parallel electric field, and the transition may indeed occur at $B_g = 0.5$. More high guide field EDR cases would be needed to fully confirm the hypothesis. On a second time, I looked at the variability of the intensity of the energy conversion using the same EDR database. We observed that the intensity of the energy conversion seems to slightly decrease as the guide field increases but the large dispersion of the interval confidence did not allow for robust conclusions to be drawn. However, the intensity of the energy conversion increases significantly as the density asymmetry between the magnetosheath and the magnetospheric side of the current sheet increases. This result was expected as the current J, which depends on the density and the velocity of the particles, will increase if the other parameters do not change.

One of the conclusions of the several statistical studies I made was that the number of EDRs found, even with the addition of the 18 EDR candidates found with the first algorithm, was still too low to perform certain studies. EDRs are complex regions whose observed characteristics can vary depending on the conditions : solar wind, magnetosheath, geographical location, guide field intensity, spacecraft trajectory... Enhancing the number of EDR cases is thus essential to deepen our understanding of the underlying physical mechanisms with further studies. To do so, I proposed a modified approach to automatically detect EDRs in the Phase 1b of MMS that had not been treated with the first algorithm. This new approach acts as an enhancement of the previous ”MeanRL” parameter used to characterize the asymmetry in the electron velocity distribution functions and consists in the use of a Convolutional Neural Network (CNN) to directly analyse and classify the distribution functions. Particular care was taken in the transformation of the data to ensure it could be easily used by the CNN. The eVDF used to visualize electron crescents are indeed polar 2D histograms of the logarithm of phase space densities but the CNN needs a matrix of values between 0 and 1 as an input. The transformation of the eVDF into a normalized Cartesian image was thus crucial to ensure that the signal of the crescent would not be destroyed in the process. A large amount of images are usually required to train a CNN, but only 214 crescent-shaped distributions characteristic of the EDR have been identified in the MMS data which is too few, data augmentation was then necessary to provide a reasonable number of images for the model to be trained on. However, regular image augmentation techniques used on real images could not be applied in this case due to the physical meaning of the distribution function images. I circumvented this problem by creating artificial crescents made by the combination of some of the 214 observed EDR crescents, bringing the total number of EDR crescents from 214 to 6126 and allowing for the proper training of the model. The trained CNN was then applied on distribution functions taken from phase 1b of MMS. The time series containing distribution functions classified as exhibiting a crescent were manually analyzed to look for additional EDR key signatures. This led to the discovery of 17 new EDR candidates using only data from spacecraft 1 and 2 due to the difficulty to manage the large volume of data needed to produce all the electron velocity distribution functions. Since some EDRs are visible on multiple spacecraft and others are not depending on the way the MMS tetrahedron crosses the EDR, about 5 additional EDRs could potentially be identified after running the CNN on the data from MMS 3 and 4.

It is clear based on the several studies presented in this thesis that future work needs to be done to further enhance the number of EDR cases at the disposal of the community and to address unanswered questions regarding the physics underlying the EDR. The first MLP model that was presented allowed for the discovery of new EDR candidates but lacked the use of the context of the time series in order to reach its full potential. Using an LSTM type model could be one way to enhance this algorithm even though the current small number of identified events could still be a challenge to choose such an approach. Creating a model mixing the use of physical context from the first model as well as EDR crescent detection using the CNN approach of the second model that was presented may be the most promising attempt at creating a powerful and reliable automatic EDR identification algorithm that could be used by the whole community. This model could moreover make use of some of the small scale parameters analysed in this thesis such as the Larmor radius or the magnetic scale gradient in order to strengthen the EDR detections, but further studies on new key EDR parameters should be pursued and could be eased by looking at the new current list of identified EDR presented in section 5.6. However, a number of criteria correspond only to low guide field type of reconnection (including the crescents), which underlines the need to enlarge the search of EDRs to strong guide field cases that are now rare.

One of the important aspects of the EDR that should be studied in the future is the location of the different points of interest within the EDR to understand how they are affected by reconnection parameters such as the density asymmetry or the guide field. Another important aspect that should be investigated is the study of the mechanisms governing the energy conversion at the EDR, especially those responsible for some of the observed negative energy conversion cases that are not yet fully understood by the community.

Chapter 7

Conclusion (Français)

La mission Magnetospheric Multiscale (MMS) a permis d'approfondir la compréhension des mécanismes physiques sous-jacents de la reconnexion magnétique après le succès de la mission CLUSTER. La résolution temporelle sans précédent des instruments embarqués à bord des satellites composant le tétraèdre MMS a conduit à la première observation d'une région de diffusion électronique (EDR) en 2016, une région qui ne pouvait jusqu'alors être étudiée que par des simulations de la reconnexion magnétique. L'un des instruments de la mission MMS nommé "Fast Plasma Investigation" (FPI) a permis pour la première fois d'observer des fonctions de distribution de vitesse des électrons (eVDF) en forme de croissant, théorisées plusieurs années auparavant et supposées être une signature clé de ces régions sur la base de simulations. Cette observation a été l'une des principales découvertes de cette première détection en 2016. Depuis, d'autres EDR situées à la magnétopause et à la queue magnétique ont été signalées par la communauté MMS. Le nombre d'EDR signalées du côté jour était d'environ 30 en 2018, la plupart d'entre elles ont été regroupées dans *Webster et al.* [2018] qui a servi de catalogue de base pour de nombreuses études réalisées par la communauté MMS. Cependant, un si petit nombre d'événements est bien trop faible pour comprendre la complexité du cœur du processus de reconnexion, la plupart de ces études n'ayant pris en compte qu'une petite partie de ces cas d'EDR, conduisant à un manque de généralisation pour certaines des conclusions de ces études.

La première tâche de mon travail de thèse a été de trouver de nouveaux candidats d'EDR afin d'augmenter le nombre de cas disponibles à étudier. L'identification de ces régions a été faite manuellement, ce qui a pris beaucoup de temps en raison de la grande quantité de données mises à disposition par MMS. L'idée de ma thèse était d'essayer d'apporter des solutions d'apprentissage automatique à la recherche des EDR. L'intelligence artificielle (IA) est en effet un outil de plus en plus populaire dans la communauté astrophysique, comme en témoigne le nombre croissant d'études impliquant des méthodes d'IA dans notre communauté (voir *Nguyen et al.* [2019] et *Breuillard et al.* [2020]), même si son utilisation reste marginale par rapport aux autres domaines de la physique et aux autres domaines scientifiques. J'ai donc développé un premier algorithme pour identifier les EDR dans les données de la phase 1a de MMS, l'architecture utilisée est celle d'un Perceptron multicouche (MLP) qui fait une classification multiclasse point par point en utilisant comme entrées plusieurs paramètres physiques tels que le champ magnétique, le champ électrique, les densités de particules et le courant magnétique. Cet algorithme a utilisé ces données pour prédire si, à un moment donné, un vaisseau spatial MMS se trouvait dans la magnétosphère, la magnétosphère, la région de diffusion ionique ou la région de diffusion électronique. L'une des caractéristiques d'entrée de ce MLP, appelée paramètre "MeanRL", a été créée pour identifier, à l'intérieur des fonctions de distribution de la vitesse des électrons, la présence de distributions en forme de croissant qui sont une signature clé de l'EDR. Le paramètre "MeanRL" caractérise la présence d'une asymétrie dans la eVDF sous la forme d'un scalaire qui peut être tracé et visualisé comme une série temporelle avec les autres paramètres physiques donnés par MMS. Ce premier algorithme a permis d'identifier 18 nouveaux candidats EDR qui ont été rapportés dans *Lenouvel et al.* [2021], portant le nombre d'EDR connus côté jour à environ 50. Ce nombre est donc proche des 56 rencontres d'EDR qui ont été estimées dans *Fuselier et al.* [2016] à la magnétopause côté jour pendant les 2,5 années de la mission nominale du MMS. La question de l'EDR interne/externe a également été discutée, quatre fois plus d'EDR internes (conversion d'énergie positive) que d'EDR externes (conversion d'énergie négative) ont été trouvés dans la phase 1a, ce qui signifie que l'EDR externe représente encore une partie importante de la zone de reconnexion. On a constaté que les termes d'inertie et de pression de la loi d'Ohm généralisée dominent alternativement l'un et l'autre selon l'événement et le temps analysés, ce qui suppose l'existence de deux mécanismes différents pour expliquer la création des régions de conversion d'énergie négative.

Cette grande base de données d'EDR a été le point de départ de plusieurs études statistiques que j'ai réalisées afin d'approfondir notre compréhension de la région de diffusion électronique. J'ai d'abord étudié un lien potentiel entre les paramètres du vent solaire et la conversion d'énergie au niveau de l'EDR. Après avoir constaté qu'il n'y avait pas de corrélation entre les paramètres du vent solaire et le signe de la conversion d'énergie à l'EDR, j'ai examiné l'influence de l'angle d'horloge du champ magnétique interplanétaire (IMF) sur l'intensité de la conversion d'énergie ainsi que la répartition des angles d'horloge de l'IMF parmi les cas d'EDR rapportés. J'ai observé que 69% des EDR ont été trouvées avec un angle d'horloge de l'IMF entre -112.5 et +112.5 degrés, montrant l'orientation préférée vers le sud de l'IMF pour induire un événement de reconnexion avec le champ magnétique magnétosphérique vers le nord. Un comportement intéressant de la conversion d'énergie à l'EDR en fonction du B_y de l'IMF a également été trouvé : la distribution de ce paramètre pour les points de données marqués IDR est homogène et représentative de la couverture des données, alors que la distribution du B_y de l'IMF pour les cas EDR est fortement inhomogène avec la plupart des cas B_y positifs. L'origine de cette tendance n'est pas encore complètement claire mais pourrait être due à la modification de l'emplacement de la ligne X par une combinaison de l'inclinaison du dipôle et du B_y de l'IMF vers le point subsolaire où la reconnexion devrait être plus efficace selon *Reistad et al.* [2020]. La dépendance de l'IMF B_y sur la conversion d'énergie dans l'EDR devra être confirmée par une liste plus large d'événements d'EDR.

J'ai ensuite essayé de trouver de nouveaux paramètres de petite échelle qui caractérisent les EDR et pourraient être utilisés pour mieux les identifier dans les données de séries temporelles. La vorticité électronique Ω_e a été proposée comme nouveau paramètre de substitution pour identifier les limites de l'EDR dans *Hwang et al.* [2019]. Plus précisément, ils ont proposé d'examiner la condition selon laquelle la vorticité des électrons dépasse la gyrofréquence des électrons ($\Omega_e > \omega_{ce}$) qui, théoriquement, fait passer le système dans le régime de l'EDR. Mon analyse de 46 des EDR raportés côté jour a révélé que seules les conditions uniques de l'événement présenté dans *Hwang et al.* [2019] permettaient de remplir cette condition, car cette condition n'a jamais été remplie pour aucun des événements que j'ai étudiés. Cependant, une analyse par boxplot du paramètre Ω_e/ω_{ce} a montré que ce paramètre était généralement plus élevé pour les points de données de la phase 1a classés dans la classe EDR par le premier algorithme MLP que pour les points de données classés dans les 3 autres classes. Ce paramètre pourrait donc être utilisé avec précaution pour renforcer les classifications d'EDR mais il ne peut être considéré comme une signature clé de l'EDR. Les autres paramètres qui ont été analysés comme de nouveaux paramètres potentiels à petite échelle pour identifier l'EDR sont le rayon de Larmor électronique $r_L = \frac{m_e V_{e_\perp}}{eB}$ et l'échelle de gradient magnétique $L_0 = \frac{B}{\mu_0 J}$. Au niveau de l'EDR, en l'absence d'un champ guide significatif, le rayon de Larmor est supposé être plus grand et l'échelle du gradient magnétique plus petite que dans les autres régions du plasma (en raison du petit champ magnétique attendu), ce qui a été observé en utilisant la visualisation par boxplot. Ces paramètres semblent donc pertinents pour identifier les EDRs avec de petits champs guides. Le dernier paramètre étudié est le paramètre adiabatique κ introduit par *Lavraud et al.* [2016]. Il indique que la dynamique des particules devient chaotique lorsque les valeurs sont inférieures à 10, ce qui est attendu au niveau de l'EDR en raison de la démagnétisation des électrons. Seuls 4 des cas d'EDR étudiés présentaient des valeurs inférieures à 10 mais dans un tiers des cas, κ était le plus faible à l'EDR. Ce paramètre peut donc être utilisé comme un élément de confirmation si d'autres signatures clés de l'EDR sont associées à une faible valeur de κ, mais il n'est pas suffisamment fiable pour être une signature clé de l'EDR.

La structure de l'EDR et la façon dont elle change en fonction de la configuration physique est une question complexe que j'ai essayé d'étudier en utilisant la grande base de données d'EDR à notre disposition. En me basant sur une étude conjointe de *Genestreti et al.* [2017] et *Cassak et al.* [2017] où ils ont trouvé que l'emplacement de la région de conversion d'énergie se rapproche de la ligne X lorsque le champ guide augmente, j'ai déterminé l'emplacement du taux de conversion d'énergie maximum par rapport aux points structurels de la région de reconnexion : le point $B_L = 0$, le point de croissant d'électron et le point de stagnation électronique. Une grande dispersion des positions relatives des points d'intérêt a été observée, avec des distances qui sont de l'ordre d'une fraction de la longueur inertielle ionique. Aucune tendance claire n'a été trouvée concernant l'évolution de la distance entre ces points structurels et l'emplacement du taux de conversion d'énergie maximal lorsque le champ guide ou l'asymétrie de la densité électronique augmente. De plus, les intervalles de confiance des régressions linéaires effectuées nous ont indiqué que les résultats ne sont pas assez robustes pour permettre de tirer des conclusions définitives. Nous avons donc constaté que la détermination des points d'intérêt est complexe et que leur définition ne peut se faire de manière simple.

La distance entre la ligne X et le point de stagnation électronique a également été étudiée. L'étude de *Cassak et al.* [2017] sur la question a montré que le maximum de la conversion d'énergie est situé à mi-chemin entre la ligne X et le point de stagnation électronique du côté de la magnétosphère. Lorsque le champ guide augmente, le point de stagnation électronique se rapproche de la ligne X et le maximum de la conversion d'énergie se rapproche du point de stagnation électronique. J'ai utilisé la formule analytique de la distance entre la ligne X et le point de stagnation électronique dans la direction N proposée par *Cassak et al.* [2017] et j'ai comparé les résultats avec une mesure manuelle de la distance entre la ligne X et le point de stagnation électronique dans la direction N à partir d'une méthode de séparation temporelle appliquée sur les cas d'EDR qui ont été rapportés. Il faut garder à l'esprit que la détermination du point de stagnation est difficile et sujette à d'importantes incertitudes. Pour pouvoir comparer les deux méthodes, les distances obtenues par la méthode de séparation temporelle sont normalisées par la largeur de la nappe de courant qui peut être calculée à l'aide d'une formule prenant en compte ou non la présence d'un champ guide. Nous avons observé que les distances calculées avec la formule théorique et les distances mesurées avec la méthode de séparation temporelle et normalisées dans la limite du champ guide nul sont proches. Cependant, le calcul de la largeur de la feuille de courant avec la formule incluant le champ guide donne des distances qui sont toujours plus grandes que dans les deux autres cas.

Cassak et al. [2017] a également prédit que le point de conversion énergétique maximale est toujours situé entre la ligne X et le point de stagnation des électrons et que le point de conversion énergétique maximale se rapproche du point de stagnation (et s'éloigne de la ligne X) à mesure que le champ guide augmente. J'ai donc décidé d'examiner la distance entre la ligne X et le point de conversion d'énergie maximale et sa relation avec l'emplacement du point de stagnation. Nous avons observé qu'avec notre méthode, le point de conversion énergétique maximale se situe au-delà du point de stagnation électronique pour 8 des 10 cas d'EDR étudiés, et que la proximité entre le point de conversion énergétique maximale et le point de stagnation électronique ne suit aucune tendance lorsque le champ guide augmente. Une tendance possible concernant la distance entre le point du maximum de $\mathbf{J} \cdot \mathbf{E}$ et la ligne X a cependant été observée, qui semble augmenter avec l'augmentation du champ guide, les distances les plus faibles n'étant observées que pour les faibles rapports de champ guide (inférieurs à 0,15), ce qui est en accord avec les attentes de *Cassak et al.* [2017].

La nature de la conversion d'énergie a également fait l'objet de deux études. La conversion énergétique au niveau de l'EDR a été la première des deux à être étudiée en utilisant la base de données de l'EDR. Des études précédentes ont montré que les contributions perpendiculaires à la conversion d'énergie devraient dominer jusqu'à des conditions de champ de guidage modérées, mais qu'au-delà de ce seuil, le champ électrique de reconnexion parallèle devrait augmenter et la contribution parallèle de la conversion d'énergie devrait dominer. Ceci a été vérifié par *Wilder et al.* [2018] sur un petit nombre d'événements IDR et de feuilles de courant mais j'ai cherché à étendre les résultats de cet article à notre liste d'événements EDR. J'ai trouvé qu'en dessous de $B_g = 0.5$, qui est le seuil à partir duquel la contribution parallèle est supposée dominer sur la contribution perpendiculaire, le ratio des événements avec une conversion d'énergie dominée par une contribution perpendiculaire et ceux avec une contribution parallèle dominante est le même que celui de la figure de *Wilder et al.* [2018] avec une majorité de ratios négatifs et donc de taux perpendiculaires plus importants. Ce résultat confirme qu'en dessous de $B_g = 0,5$, la dissipation semble être dominée par le champ électrique perpendiculaire plus que par le champ électrique parallèle, et la transition peut effectivement se produire à $B_g = 0,5$. D'autres cas d'EDR à champ guide élevé seraient nécessaires pour confirmer pleinement cette hypothèse. Dans un deuxième temps, j'ai examiné la variabilité de l'intensité de la conversion d'énergie en utilisant la même base de données d'EDR. Nous avons observé que l'intensité de la conversion d'énergie semble diminuer légèrement lorsque le champ de guidage augmente, mais la grande dispersion de l'intervalle de confiance n'a pas permis de tirer des conclusions solides. Cependant, l'intensité de la conversion d'énergie augmente de manière significative lorsque l'asymétrie de densité entre la magnétosphère et le côté magnétosphérique de la feuille de courant augmente. Ce résultat était attendu, car le courant magnétique J, qui dépend de la densité et de la vitesse des particules, augmente si les autres paramètres ne changent pas.

Une des conclusions des différentes études statistiques que j'ai réalisées était que le nombre d'EDR trouvées, même en ajoutant les 18 candidats d'EDR trouvés avec le premier algorithme, était encore trop faible pour réaliser certaines études. Les EDR sont des régions complexes dont les caractéristiques observées peuvent varier en fonction des conditions : vent solaire, magnétosheath, localisation géographique, intensité du champ guide, trajectoire du satellite... Il est donc essentiel d'augmenter le nombre de cas d'EDR pour approfondir notre compréhension des mécanismes physiques sous-jacents par des études complémentaires. Pour ce faire, j'ai proposé une approche modifiée pour détecter automatiquement les EDR dans la phase 1b de MMS qui n'avaient pas été traités avec le premier algorithme. Cette nouvelle approche agit comme une amélioration du précédent paramètre "MeanRL" utilisé pour caractériser l'asymétrie dans les fonctions de distribution de la vitesse des électrons et consiste en l'utilisation d'un réseau de neuronnes convolutif (CNN) pour analyser et classer directement les fonctions de distribution. Un soin particulier a été apporté à la transformation des données afin qu'elles puissent être facilement utilisées par le CNN. Les eVDF utilisées pour visualiser les croissants d'électrons sont en effet des histogrammes polaires 2D du logarithme des densités dans l'espace des phases, mais le CNN a besoin d'une matrice de valeurs entre 0 et 1 comme entrée. La transformation de l'eVDF en une image cartésienne normalisée était donc cruciale pour s'assurer que le signal du croissant ne serait pas détruit dans le processus. Une grande quantité d'images est généralement nécessaire pour entraîner un CNN, mais seulement 214 distributions en forme de croissant caractéristiques de l'EDR ont été identifiées dans les données de MMS, ce qui est trop peu, l'augmentation des données était donc nécessaire pour fournir un nombre raisonnable d'images pour l'entraînement du modèle. Cependant, les techniques habituelles d'augmentation des images utilisées sur des images réelles ne pouvaient pas être appliquées dans ce cas en raison de la signification physique des images de la fonction de distribution. J'ai contourné ce problème en créant des croissants artificiels faits par la combinaison de certains des 214 croissants EDR observés, portant le nombre total de croissants EDR de 214 à 6126 et permettant l'entraînement adéquat du modèle. Le CNN entraîné a ensuite été appliqué aux fonctions de distribution extraites de la phase 1b de MMS. Les séries temporelles contenant des fonctions de distribution classées comme présentant un croissant ont été analysées manuellement afin de rechercher d'autres signatures clés de l'EDR. Cela a conduit à la découverte de 17 nouveaux candidats EDR en utilisant uniquement les données des satellites 1 et 2 en raison de la difficulté à gérer le grand volume de données nécessaire pour produire toutes les fonctions de distribution de la vitesse des électrons. Étant donné que certaines EDR sont visibles sur plusieurs satellites et que d'autres ne le sont pas, en fonction de la façon dont le tétraèdre de MMS croise l'EDR, environ 5 EDR supplémentaires pourraient potentiellement être identifiées après avoir appliqué le CNN sur les données de MMS 3 et 4.

Il est clair que, sur la base de plusieurs études présentées dans cette thèse, des travaux futurs doivent être effectués pour augmenter encore le nombre de cas d'EDR à la disposition de la communauté et pour répondre aux questions restées sans réponse concernant la physique sous-jacente à l'EDR. Le premier modèle de MLP qui a été présenté a permis de découvrir de nouveaux candidats d'EDR mais il lui manquait l'utilisation du contexte de la série temporelle pour atteindre son plein potentiel. L'utilisation d'un modèle de type LSTM pourrait être une façon d'améliorer cet algorithme, même si le petit nombre actuel d'événements identifiés pourrait encore constituer un défi pour le choix d'une telle approche. La création d'un modèle combinant l'utilisation du contexte physique du premier modèle ainsi que la détection du croissant de l'EDR à l'aide de l'approche CNN du second modèle présenté pourrait être la tentative la plus prometteuse pour créer un algorithme d'identification automatique des EDR puissant et fiable qui pourrait être utilisé par l'ensemble de la communauté. Ce modèle pourrait en outre utiliser certains des paramètres à petite échelle analysés dans cette thèse tels que le rayon de Larmor ou le gradient d'échelle magnétique afin de renforcer les détections d'EDR, mais des études supplémentaires sur de nouveaux paramètres clés de l'EDR devraient être effectuées et pourraient être facilitées par l'examen de la nouvelle liste actuelle d'EDR identifiés présentée dans la section 5.6. Cependant, un certain nombre de critères ne correspondent qu'à des types de reconnexion à faible champ guide (y compris les croissants), ce qui souligne la nécessité d'élargir la recherche d'EDR aux cas à fort champ guide qui sont pour le moment rares.

L'un des aspects importants de l'EDR qui devrait être étudié à l'avenir est la localisation des différents points d'intérêt dans l'EDR pour comprendre comment ils sont affectés par les paramètres de reconnexion tels que l'asymétrie de densité ou le champ guide. Un autre aspect important qui devrait être étudié est l'étude des mécanismes régissant la conversion d'énergie au niveau de l'EDR, en particulier ceux responsables de certains des cas de conversion d'énergie négative observés qui ne sont pas encore totalement compris par la communauté.

Appendix A

Useful coordinate systems for reconnection

In plasma physics, different coordinate systems can be used depending on what is studied and how. We will here only review the ones that are useful for the comprehension of this thesis.

The most basic coordinate system used in plasma physics is the ”Geocentric Solar Ecliptic” system (GSE). The X-axis is towards the Sun, the Z-axis is perpendicular to the plane of the Earth’s orbit around the Sun, being positive in the north direction, and the Y-axis completes the right-handed orthogonal triad. This system is usually the reference is which most of the space physics data is given as it is useful to specify magnetic boundaries and spacecraft positions when they are orbiting the Earth. There is another coordinate system close to GSE which is the ”Geocentric Solar Magnetospheric” (GSM) system. The X-axis is still oriented in the Sun’s direction but the Y-axis is perpendicular to the Earth’s magnetic dipole, towards the dusk and include in the magnetic equator plane so that the X-Z axis contains the dipole axis. The difference between GSE and GSM is therefore a simple rotation around the X axis. An image summarizing these 2 coordinate systems can be found in figure A.1.

Next, another important coordinate system while studying the effect of magnetic fields on particles is the ”Field-Aligned Coordinate” (FAC) system. The 3 field-aligned components are $(||, \perp_1, \perp_2)$and for a vector **A**, they are obtained by first taking the dot product of the vector **A** and the magnetic field vector **B** to get $A_{||}$. The $A_{\perp_1}$ coordinate can be obtained by taking the dot product of **A** and the direction of the electron velocity vector V_e. The final coordinate $V_{\perp_2}$ is along the direction of $V_{||} \times V_{\perp,1}$.

The last important coordinate system we will talk about is the ”LMN” coordinate system, used when studying events related to the magnetopause. At the magnetopause, the L axis is contained in the reconnection plane along the direction of the component of the magnetic field **B**, the N axis is perpendicular to the current sheet, positive away from the

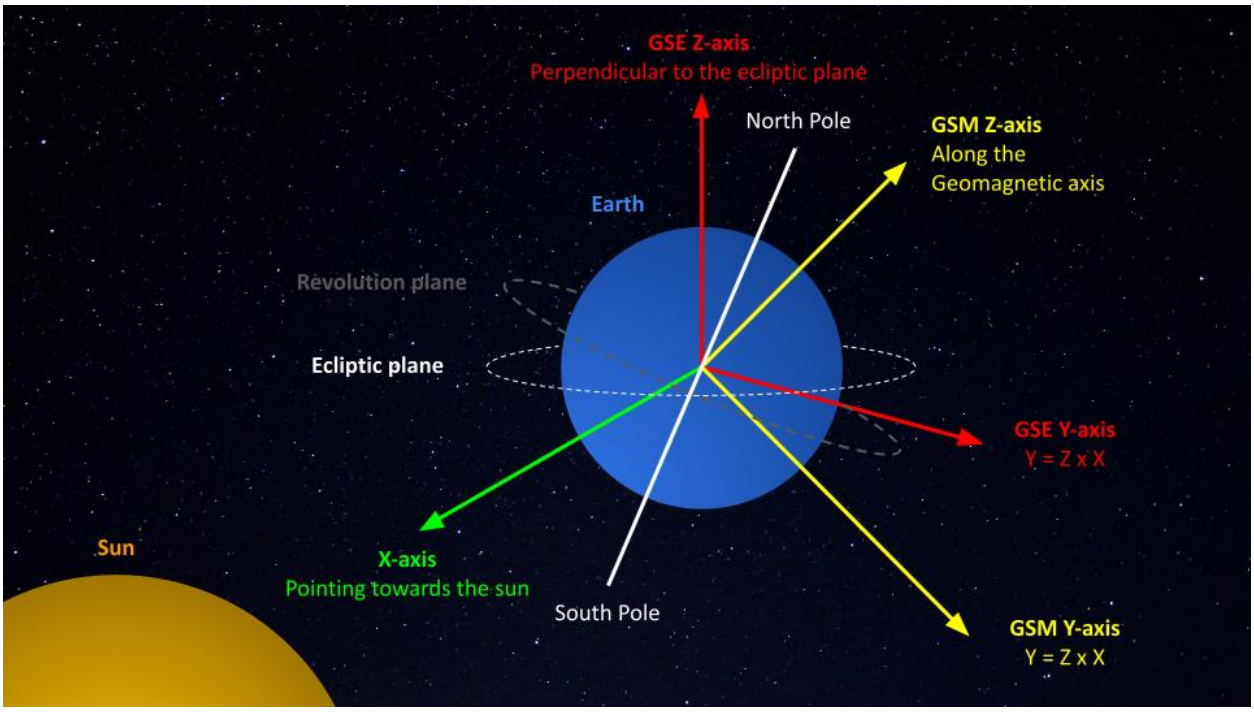


Figure A.1: Figure showing the difference between the GSE and GSM coordinate systems.

Earth, and the M axis, positive westward, completes the right-handed orthogonal triad. The determination of this system is more complex than for the previous coordinate systems and can be done using different techniques. One of them, which is the Minimum Variance Analysis of B (MVAB) is detailed in Appendix B.

Appendix B

A brief overview of some single and multi-spacecraft techniques

B.1 Minimum Variance Analysis of B (MVAB)

The main purpose of the minimum variance analysis (MVA) is to estimate the coordinates of the unit normal vector $\hat{\mathbf{n}}$ to an idealised one-dimensional transition layer in a plasma from single-spacecraft data, this was first developed by *Sonnerup and Cahill Jr.* [1967]. The aim of the method is to determine the $\hat{\mathbf{n}}$ vector so that the field-component set $\{\mathbf{B}^{(m)} \cdot \hat{\mathbf{n}}\}$ (m=1,2,3,...,M) has minimum variance. $\hat{\mathbf{n}}$ is thus determined by the minimization of σ^2 defined as :

$$\sigma^2 = \frac{1}{M} \sum_{m=1}^{M} |(\mathbf{B}^{(m)} - \langle \mathbf{B} \rangle) \cdot \hat{\mathbf{n}}|^2 \tag{B.1}$$

where $\langle \mathbf{B} \rangle$ is the average of the magnetic field defined as :

$$\langle \mathbf{B} \rangle \equiv \frac{1}{M} \sum_{m=1}^{M} \mathbf{B}^{(m)} \tag{B.2}$$

and with the normalisation constraint $|\hat{\mathbf{n}}|^2 = 1$. This translates by seeking the solution of a set of three homogeneous linear equations which are the following :

$$\frac{\partial}{\partial n_X}(\sigma^2 - \lambda(|\hat{n}|^2 - 1)) = 0 \frac{\partial}{\partial n_Y}(\sigma^2 - \lambda(|\hat{n}|^2 - 1)) = 0 \frac{\partial}{\partial n_Z}(\sigma^2 - \lambda(|\hat{n}|^2 - 1)) = 0 \tag{B.3}$$

with λ being a Lagrange multiplier. The resulting set of three equations can be written in matrix form the following way :

$$\sum_{\nu=1}^{3} M_{\mu\nu}^{B} n_\nu = \lambda n_\mu \tag{B.4}$$

where $\mu, \nu = 1, 2, 3$ represent the cartesian components X, Y, Z in the GSM or GSE coordinate system and $M_{\mu\nu}^{B}$ is the magnetic variance matrix written as :

$$M_{\mu\nu}^{B} \equiv \langle B_\mu B_\nu \rangle - \langle B_\mu \rangle \langle B_\nu \rangle \tag{B.5}$$

The λ values from equation B.4 are the eigenvalues $\lambda_1, \lambda_2, \lambda_3$ of $M^B_{\mu\nu}$ which represent the directions of maximum, intermediate and minimum variance of the field component along each vector. The eigenvalue with the lowest absolute value is λ_N and the associated eigenvector N defines the boundary normal direction. This vector is chosen to point outwards of the magnetopause by definition. The eigenvalue with the highest absolute value is λ_L and the associated eigenvector is L, and the third vector M can be found knowing $M = N \times L$. *Sonnerup and Cahill Jr.* [1967] defined that the method can be considered valid only if the ratio λ_M/λ_N is greater than 1.5. Any vector in a given coordinate system can then be converted into the LMN coordinate system using the transformation matrix :

$$T = \begin{pmatrix} L_X & L_Y & L_Z \\ M_X & M_N & M_Y \\ N_X & N_Y & N_Z \end{pmatrix} \tag{B.6}$$

B.2 Constant Velocity Approach (CVA)

The constant velocity approach (CVA) was first developed by *Russell et al.* [1983] and uses a multi-spacecraft timing method to determine the velocity of a plasma discontinuity. It assumes a constant velocity of the discontinuity during the crossing of the spacecraft. The method allows for the determination of a unit vector $\hat{\mathbf{n}}$ normal to the discontinuity as well as a velocity V_0 along this same vector. The assumption of a constant velocity at the magnetopause does not hold for a long period of time due to the rapid changing motions so the results are only valid locally. The general case when $N \geq 4$ is called Polynomial velocity approach (PVA) and was developed by *Haaland et al.* [2004], the velocity $V(t)$ is written as :

$$V(t) = A_0 + A_1 t + A_2 t^2 + A_3 t^3 + \ldots + A_{N-4} t^{N-4} \tag{B.7}$$

In the case of $N = 4$, PVA reduces to CVA and the velocity thus reduces to $V(t) = A_0$. The spacecraft traversals are ordered according to increasing time, the first crossing is CR0 and occurs at $t = t_0 = 0$ while the last crossing is CR(N-1) and occurs at $t = t_{N-1}$. The spacecraft separations are given relative to the order of the spacecraft crossing the discontinuity, so by defining $\mathbf{SR}_i$ as the spacecraft position vector of the i^{th} spacecraft to cross the boundary, the separation vector $\mathbf{R}_i$ $(i = 1, 2, 3, \ldots, N-1)$ is :

$$\mathbf{R}_i = (\mathbf{SR}_i - \mathbf{SR}_0) \tag{B.8}$$

The component of that vector along $\hat{\mathbf{n}}$ can then be written as :

$$\mathbf{R}_i \cdot \hat{\mathbf{n}} = \int_{t=0}^{t=t_i} V(t)dt = A_0 t_i \tag{B.9}$$

Using 4 spacecraft, there are (N-1) equations of this form and the system can be written in matrix from as follows :

$$\begin{pmatrix} R_{1x} & R_{1y} & R_{1z} \\ R_{2x} & R_{2y} & R_{2z} \\ R_{3x} & R_{3y} & R_{3z} \end{pmatrix} \begin{pmatrix} m_x \\ m_y \\ m_z \end{pmatrix} = \begin{pmatrix} t_1 \\ t_2 \\ t_3 \end{pmatrix} \tag{B.10}$$

with $\mathbf{m} \equiv \hat{\mathbf{n}}/A_0$. After solving the system, the coefficient A_0 is obtained from the normalisation condition $\hat{\mathbf{n}}^2 = 1$, which leads to $A_0 = 1/|\mathbf{m}|$ and $\hat{\mathbf{n}} = \mathbf{m}/|\mathbf{m}|$

Appendix C

Application of the first Multilayer Perceptron model on EDR simulation data

In this section, I will report some of the work done by Ambre Ghisalberti during her internship at IRAP from March to July 2021. The work that will be presented is an attempt at using the automatic plasma region detection algorithm presented in chapter 3 on data taken from simulations. This work is important because it asks the question of the feasibility to train a machine learning model on real data to use it on simulations, even though the real goal would be to train a model on simulation data and see if it could be applied to do classification on real data. A number of applications show the ability of machine learning techniques to better classify real data by training an algorithm on simulated data. A large number of simulations can indeed help to augment training data to detect rare events in real data. In our case, we cannot perform a large number of different EDR simulations, but we can try to compare data and simulations by trying to detect EDRs in simulations from our algorithm trained on data, and identify the key issues. As we will see in this section, going from simulated to real data and vice-versa is a major challenge due to how simulations are made, all the parameters are indeed normalized so if a model was trained on real data, one needs to denormalize the simulation data so that the values fall in the range of what the model was trained on.

C.1 Description of the simulation data

The simulation data come from a two-dimensional fully kinetic simulation of magnetic reconnection using the particle-in-cell (PIC) code SMILEI. It is a special simulation as it features a large density plasmaspheric plume modeled by a large amount of cold plasma. All the specificities of this simulation are presented in *Dargent et al.* [2020] so I will only summarize the important parts to understand the work Ambre has done.

In the simulation, the data are normalized using ion scale quantities, the magnetic field

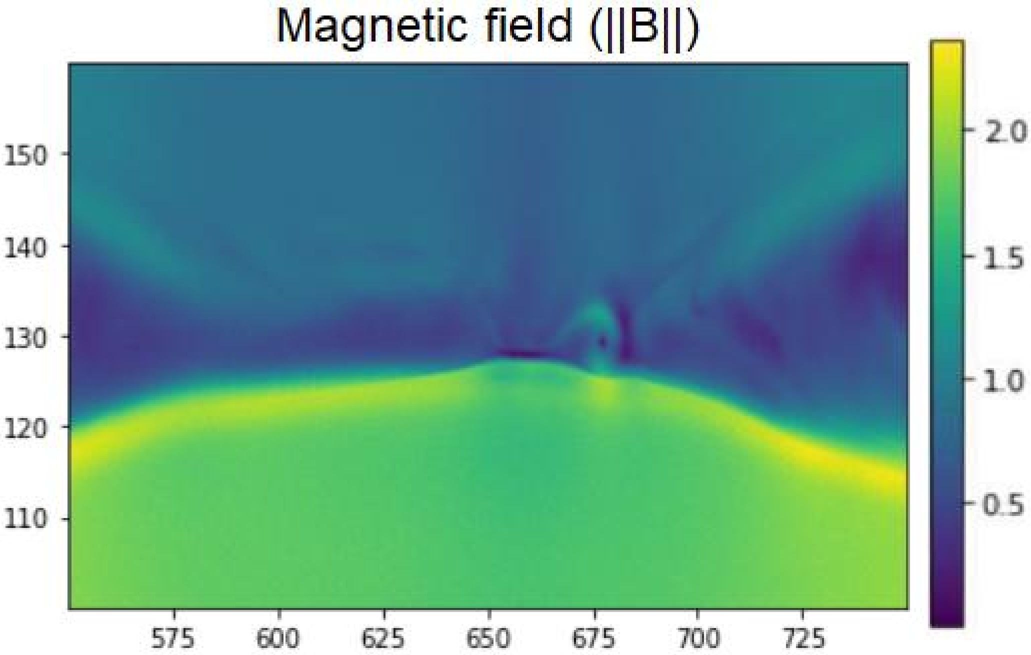


Figure C.1: Figure showing the absolute value of the magnetic field for the simulation in the range $x = [550, 750]$ and $y = [100, 160]$. The values are normalized by an arbitrary B_0

is normalized to an arbitrary value B_0 and the density is normalized to an arbitrary quantity n_0 as well. The masses and charges are normalized to the proton mass m_p and charge e, lengths are normalized to the inverse of the proton inertial length $\delta_i = c/\omega_{pi}$ ans time to the inverse of the proton gyrofrequency $\omega_{ci}^{-1} = m_p/(eB_0)$ in the simulation. There are 25 600 cells in the x direction, 10 240 cells in the y direction, and initially 50 particles per cell and per population. It is good to note that the x and y axes in the simulation respectively correspond to the z and x GSE axes, the out of plane axis of the simulation then corresponds to the GSE y axis. The chosen mass ratio m_i/m_e is 25 (instead of 1836) in order to accelerate the computation, which may induce biases for some of the physical parameters (in particular the energy conversion $\mathbf{J} \cdot \mathbf{E}$ where masses appear).

The system has periodic boundary conditions in the x direction and reflective boundary conditions in the y direction. Figures C.1 and C.2 respectively show the absolute value of the magnetic field and the ion density for a small part of the simulation centered around the EDR in the middle. In this small area, one time step was chosen to run the classification model on is such that this area is not affected by the periodic and reflective boundary conditions yet. The plasmaspheric plume is visible at the bottom of figure C.2 with very high ion density values just below the area with the lowest ion density values.

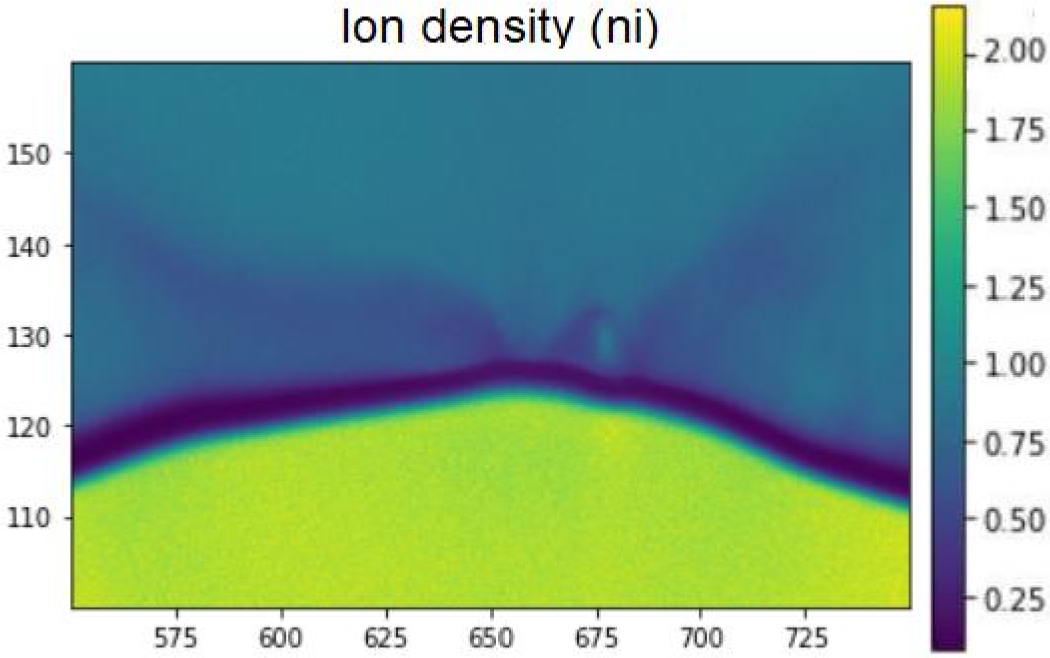


Figure C.2: Figure showing the ion density for the simulation in the range $x = [550, 750]$ and $y = [100, 160]$. The values are normalized by an arbitrary n_0

C.2 Linking the simulation data and the real data

The two main problems that arose when Ambre tried to adapt the algorithm on the simulations data were the denormalization of the simulation data and the computation of the MeanRL parameter. The first problem comes from the fact that the simulation data are normalized by arbitrary values B_0 and n_0 which will determine all the other parameters in the simulation. Since the algorithm was trained on real data, the first step is to "denormalize" the simulation data by choosing a B_0 and an n_0 that will fix all the other values of the simulation. This is a very important task because choosing inappropriate values for B_0 and n_0 can result in other variables being out of the range of the training data of the algorithm. The goal is to use the same scaler that was used to scale (with a standardization scaler in the python algorithm) the training data of the algorithm as described in chapter 3 and apply it to the denormalized simulation data so that the algorithm can run predictions. If the data are not properly scaled then the algorithm will fail at making the right region predictions. To check if the data are properly scaled, Ambre used boxplots to compare the distribution of values among the different classes (Magnetosphere, Magnetosheath, IDR and EDR) for the simulation data and for the real data using several combinations of B_0 and n_0 values.

The second problem regarding the computation of the MeanRL comes from the fact that the way the electron velocity distribution functions are computed in the simulation is not the same as with the real data. The MeanRL computation method was specially design to work with the polar histogram representation meaning that some work was necessary to adapt the MeanRL computation to the simulation data. Figure C.3 shows an example of raw electron velocity distribution functions taken at two different places of the simulation.

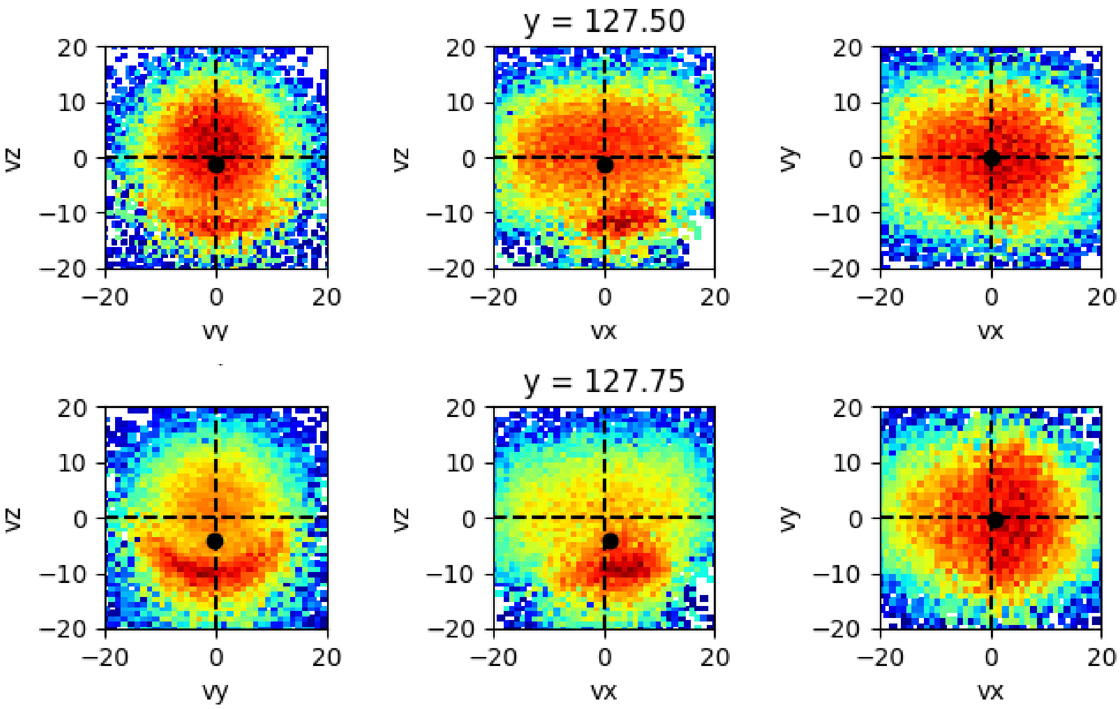


Figure C.3: Electron velocity distribution functions examples taken from the simulation in the region $x = [654, 657]$ and $y = [127.5, 127.75]$.

One of the things that are different from the eVDF from MMS is that in the simulation we have access to the low velocity part of the distribution.

Since the data are originally in GSE coordinates, the first thing to do was to use the field aligned coordinates to visualize the crescents in the ($V_{\perp,1}$,$V_{\perp,2}$) plane. To avoid problems that could be due to the denormalization of the data, the computation of the MeanRL was not done using the energy bins and was rather simplified to making the ratio of the number of particles with a positive $V_{\perp,1}$ velocity with the number of particles with a negative $V_{\perp,1}$ velocity. A zoomed map of the MeanRL adapted to the simulation data at the X-point can be seen in figure C.4. We see that the high values are mostly located close to the X-line as expected but on the magnetosheath side (top of the picture), there are values that are higher than expected in this region producing a patchy color map that could be due to the method of computation of this adapted MeanRL. The classic MeanRL value also gave high values sometimes when there were no real asymmetry observed so the source of the patchiness observed could be the imprecisions of the concept of the MeanRL itself.

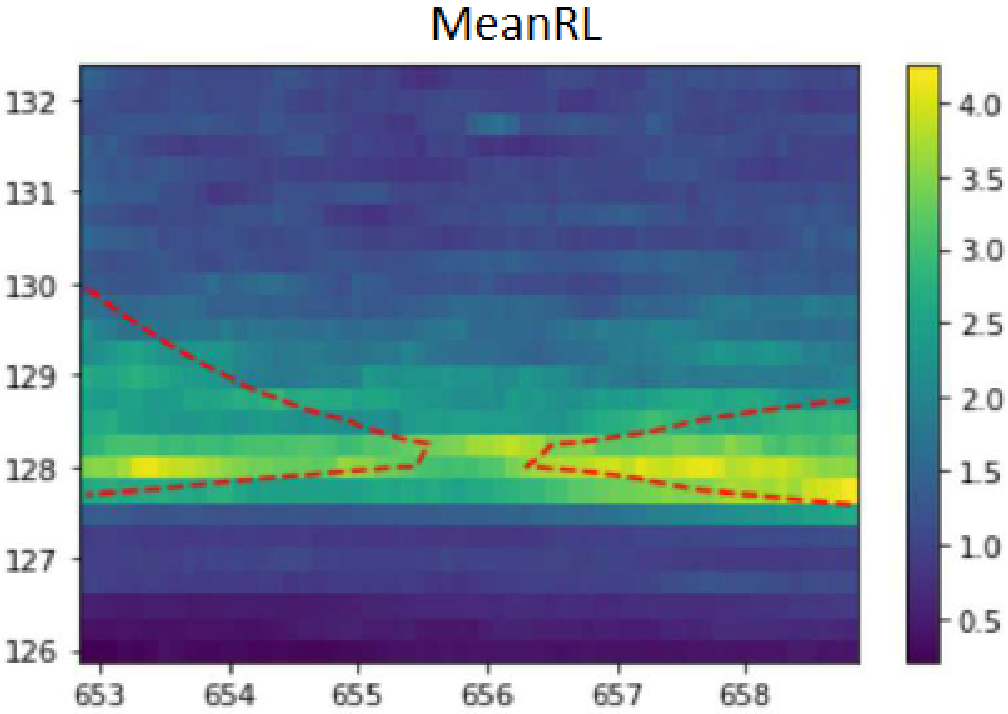


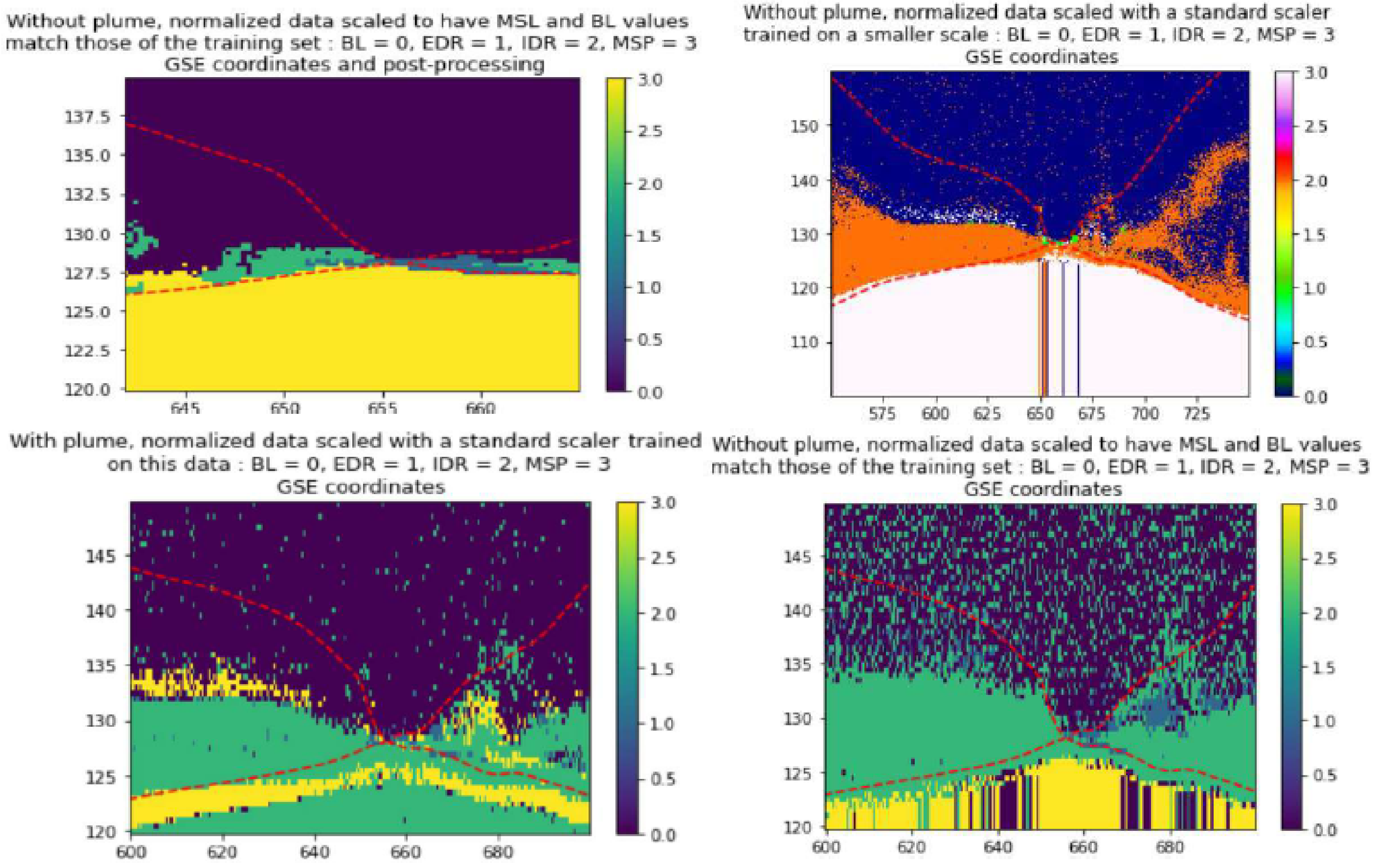
Figure C.4: Figure showing the adapted MeanRL parameter for the simulation in the range $x = [653, 659]$ and $y = [126, 132]$ (at the X-point).

C.3 Plasma region predictions of the model on the simulation data

The value finally chosen for B_0 was 18 nT and the value for n_0 was 10 cm^{-3} which are typical values at the magnetopause. The problem was however that using the original scaler for the algorithm resulted in the distribution of some of the simulation data parameters being very different from that of the real data. The algorithms predictions cannot be accurate if the data it is given look nothing like what it was trained on. Therefore, the decision was made not to use the original scaler but to fine tune a new scaler manually better adapted to the training set. Electron velocity distribution functions were only available in the $x = [600, 700]$ and $y = [120, 150]$ area of the simulation so the EDR predictions, which relies on the MeanRL to work properly, could only be made in this area.

Figure C.5 shows four different attempts at using the plasma region prediction algorithm on the simulation data. The presence of a plume appeared as an issue since it strongly modifies the range of physical parameters inside the simulation which impacts the results of the automatic classification. We thus tried to apply modifications, by keeping or removing the plume. Only the bottom left classification model keeps the plume in the data whereas for the other attempts, each column of values inside the plume were replaced by the last value of the same column that is inside the magnetosphere layer, and this is why on the two right plots we can see some vertical lines with the same predictions as inside the plume. The bottom right plot (with a scaler adapted to the real data training set in the inner simulation region) shows a distinction between the magnetosheath and the magnetosphere regions but the prediction of the location of the IDR is extremely noisy and a large EDR is predicted in the exhaust region and not near the X-point which is not convincing. The top right plot (with a scaler trained based only on the inner simulation region) is less noisy than the previous plot and the EDR is better located around the X-point with only a few EDR predictions in the exhaust. The bottom left plot shows that without knowledge of what the plume is, the algorithm interestingly classifies it as an IDR but we do not know why this is the case. The top left plot (with a scaler adapted to the training set) uses post-processing to smooth the results, the view is also closer to better visualize the EDR prediction in the middle. We see that the EDR is very well located on the X-line, with IDR predictions inside the separatrices and rather on the magnetospheric side.

The results obtained from these attempts are promising results even though one has to bear in mind that a heavy human implication was needed to obtain this result during the data scaling part that is very tricky and yet crucial. Future work is needed to obtain satisfying results in a more automatic way. Linking simulation data and real data would be extremely beneficial as simulations can yield an almost infinite amount of training data for events such as the EDR that is easier to simulate than to observe and this application is a encouraging first step towards that direction. The work done by Ambre Ghisalberti shows that even though the physics in the simulation is not an exact replica of the real physics (notably due to the mass ratios used in the simulations), an algorithm trained on real data can still make satisfying predictions on simulated data. The next step would be to try other simulation runs with various conditions and evaluate how robust the needed scaling is, and then to reversely use simulation data to train an algorithm that would be used on real data test sets. Moreover, as mentioned in the manuscript, reconnection signatures are different depending on the region considered (dayside magnetopause or magnetotail), which may need the use of several different algorithms or ensemble models.


Without plume, normalized data scaled to have MSL and BL values match those of the training set : BL = 0, EDR = 1, IDR = 2, MSP = 3 GSE coordinates and post-processing

Without plume, normalized data scaled with a standard scaler trained on a smaller scale : BL = 0, EDR = 1, IDR = 2, MSP = 3 GSE coordinates

With plume, normalized data scaled with a standard scaler trained on this data : BL = 0, EDR = 1, IDR = 2, MSP = 3 GSE coordinates

Without plume, normalized data scaled to have MSL and BL values match those of the training set : BL = 0, EDR = 1, IDR = 2, MSP = 3 GSE coordinates

Figure C.5: Figure showing four different attempts at using the plasma region prediction algorithm on the simulation data. The red dashed lines represent the separatrices.

Bibliography

Alexandrova, O., C. Rossi, M. Maksimovic, P. Hellinger, C. Lacombe, and A. Mangeney, Coherent vortices in Solar Wind Turbulence at Sub-Ion and Electron Scales, in *AGU Fall Meeting Abstracts*, vol. 2021, pp. SH44B–02, 2021.

Aschwanden, M. J., Global energetics of solar flares. x. petschek reconnection rate and alfvén mach number of magnetic reconnection outflows, *The Astrophysical Journal*, *895*(2), 134, doi:10.3847/1538-4357/ab8aec, 2020.

Aunai, N., M. Hesse, and M. Kuznetsova, Electron nongyrotropy in the context of collisionless magnetic reconnection, *Physics of Plasmas*, *20*(9), 092903, doi:10.1063/1.4820953, 2013.

Bessho, N., L.-J. Chen, and M. Hesse, Electron distribution functions in the diffusion region of asymmetric magnetic reconnection, *Geophysical Research Letters*, *43*(5), 1828–1836, doi:10.1002/2016GL067886, 2016.

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

Burch, J. L., and T. D. Phan, Magnetic reconnection at the dayside magnetopause: Advances with mms, *Geophysical Research Letters*, *43*(16), 8327–8338, doi:10.1002/2016GL069787, 2016.

Burch, J. L., et al., Electron-scale measurements of magnetic reconnection in space, *Science*, *352*(6290), aaf2939, doi:10.1126/science.aaf2939, 2016.

Burch, J. L., et al., Localized oscillatory energy conversion in magnetopause reconnection, *Geophysical Research Letters*, *45*(3), 1237–1245, doi:10.1002/2017GL076809, 2018a.

Burch, J. L., et al., Localized oscillatory energy conversion in magnetopause reconnection, *Geophysical Research Letters*, *45*(3), 1237–1245, doi:https://doi.org/10.1002/2017GL076809, 2018b.

Camporeale, E., The challenge of machine learning in space weather: Nowcasting and forecasting, *Space Weather*, *17*(8), 1166–1207, doi:https://doi.org/10.1029/2018SW002061, 2019.

Cassak, P. A., and M. A. Shay, Scaling of asymmetric magnetic reconnection: General theory and collisional simulations, *Physics of Plasmas*, *14*(10), 102,114, doi:10.1063/1.2795630, 2007.

Cassak, P. A., et al., The effect of a guide field on local energy conversion during asymmetric magnetic reconnection: Particle-in-cell simulations, *Journal of Geophysical Research: Space Physics*, *122*(11), 11,523–11,542, doi:10.1002/2017JA024555, 2017.

Chen, L.-J., et al., Electron energization and mixing observed by mms in the vicinity of an electron diffusion region during magnetopause reconnection, *Geophysical Research Letters*, *43*(12), 6036–6043, doi:https://doi.org/10.1002/2016GL069215, 2016.

Chen, L.-J., et al., Electron diffusion region during magnetopause reconnection with an intermediate guide field: Magnetospheric multiscale observations, *Journal of Geophysical Research: Space Physics*, *122*(5), 5235–5246, doi:10.1002/2017JA024004, 2017.

Clausen, L. B. N., S. E. Milan, J. B. H. Baker, J. M. Ruohoniemi, K.-H. Glassmeier, J. C. Coxon, and B. J. Anderson, On the influence of open magnetic flux on substorm intensity: Ground- and space-based observations, *Journal of Geophysical Research: Space Physics*, *118*(6), 2958–2969, doi:https://doi.org/10.1002/jgra.50308, 2013.

Costello, K. A., *Moving the Rice MSFM into a real-time forecast mode using solar wind driven forecast modules*, Rice University, 1998.

Cover, T. M., and P. E. Hart, Nearest neighbor pattern classification, *IEEE Trans. Inf. Theory*, *13*, 21–27, 1967.

Cozzani, G., et al., In situ spacecraft observations of a structured electron diffusion region during magnetopause reconnection, *pre*, *99*(4), 043204, doi:10.1103/PhysRevE.99.043204, 2019.

Dargent, J., N. Aunai, B. Lavraud, S. Toledo-Redondo, and F. Califano, Simulation of plasmaspheric plume impact on dayside magnetic reconnection, *Geophysical Research Letters*, *47*(4), e2019GL086,546, doi:https://doi.org/10.1029/2019GL086546, e2019GL086546 10.1029/2019GL086546, 2020.

Dong, X.-C., et al., Observation of nonuniform energy dissipation in the electron diffusion region of magnetopause reconnection, *Geophysical Research Letters*, *48*(13), e2020GL091,928, doi:https://doi.org/10.1029/2020GL091928, e2020GL091928 2020GL091928, 2021.

Dungey, J., Lxxvi. conditions for the occurrence of electrical discharges in astrophysical systems, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *44*(354), 725–738, doi:10.1080/14786440708521050, 1953.

Durand-Manterola, H. J., and A. Flandes, Plasma vortices driven by magnetic torsion generated by electric currents in non-magnetic planetary wakes, *Advances in Space Research*, *69*(10), 3902–3908, doi:https://doi.org/10.1016/j.asr.2022.03.002, 2022.

Egedal, J., A. Le, P. L. Pritchett, and W. Daughton, Electron dynamics in two-dimensional asymmetric anti-parallel reconnection, *Physics of Plasmas*, *18*(10), 102,901, doi:10.1063/1.3646316, 2011.

Ergun, R. E., et al., The axial double probe and fields signal processing for the mms mission, *Space Science Reviews*, *199*(1), 167–188, doi:10.1007/s11214-014-0115-x, 2016.

Ergun, R. E., et al., Drift waves, intense parallel electric fields, and turbulence associated with asymmetric magnetic reconnection at the magnetopause, *Geophysical Research Letters*, *44*(7), 2978–2986, doi:https://doi.org/10.1002/2016GL072493, 2017.

Eriksson, S., et al., Magnetospheric multiscale observations of the electron diffusion region of large guide field magnetic reconnection, *Phys. Rev. Lett.*, *117*, 015,001, doi:10.1103/PhysRevLett.117.015001, 2016.

Fox, W., et al., Energy conversion by parallel electric fields during guide field reconnection in scaled laboratory and space experiments, *Geophysical Research Letters*, *45*(23), 12,677–12,684, doi:https://doi.org/10.1029/2018GL079883, 2018.

Fuselier, S. A., W. S. Lewis, C. Schiff, R. Ergun, J. L. Burch, S. M. Petrinec, and K. J. Trattner, Magnetospheric Multiscale Science Mission Profile and Operations, *ssr*, *199*(1-4), 77–103, doi:10.1007/s11214-014-0087-x, 2016.

Fuselier, S. A., et al., Large-scale characteristics of reconnection diffusion regions and associated magnetopause crossings observed by mms, *Journal of Geophysical Research: Space Physics*, *122*(5), 5466–5486, doi:10.1002/2017JA024024, 2017.

Genestreti, K. J., et al., The Effect of a Guide Field on Local Energy Conversion During Asymmetric Magnetic Reconnection: MMS Observations, *Journal of Geophysical Research (Space Physics)*, *122*(11), 11,342–11,353, doi:10.1002/2017JA024247, 2017.

Genestreti, K. J., et al., Mms observation of asymmetric reconnection supported by 3-d electron pressure divergence, *Journal of Geophysical Research: Space Physics*, *123*(3), 1806–1821, doi:10.1002/2017JA025019, 2018a.

Genestreti, K. J., et al., Mms observation of asymmetric reconnection supported by 3-d electron pressure divergence, *Journal of Geophysical Research: Space Physics*, *123*(3), 1806–1821, doi:https://doi.org/10.1002/2017JA025019, 2018b.

Goodbred, M., Y.-H. Liu, B. Chen, and X. Li, The relation between the energy conversion rate and reconnection rate in petschek-type reconnection—implications for solar flares, *Physics of Plasmas*, *28*(8), 082,103, doi:10.1063/5.0050557, 2021.

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, Generative adversarial nets, *Advances in neural information processing systems*, *27*, 2014.

Graham, D. B., et al., Instability of agyrotropic electron beams near the electron diffusion region, *Phys. Rev. Lett.*, *119*, 025,101, doi:10.1103/PhysRevLett.119.025101, 2017.

Génot, V., et al., Science data visualization in planetary and heliospheric contexts with 3dview, *Planetary and Space Science*, *150*, 111 – 130, doi:https://doi.org/10.1016/j.pss.2017.07.007, enabling Open and Interoperable Access to Planetary Science and Heliophysics Databases and Tools, 2018.

Génot, V., et al., Automated multi-dataset analysis (amda): An on-line database and analysis tool for heliospheric and planetary plasma data, *Planetary and Space Science*, *201*, 105,214, doi:https://doi.org/10.1016/j.pss.2021.105214, 2021.

Haaland, S. E., et al., Four-spacecraft determination of magnetopause orientation, motion and thickness: comparison with results from single-spacecraft methods, *Annales Geophysicae*, *22*(4), 1347–1365, doi:10.5194/angeo-22-1347-2004, 2004.

Hamrin, M., O. Marghitu, P. Norqvist, S. Buchert, M. André, B. Klecker, L. M. Kistler, and I. Dandouras, Energy conversion regions as observed by cluster in the plasma sheet, *Journal of Geophysical Research: Space Physics*, *116*(A1), doi:https://doi.org/10.1029/2010JA016383, 2011.

Hesse, M., N. Aunai, D. G. Sibeck, and J. Birn, On the electron diffusion region in planar, asymmetric, systems, *Geophysical Research Letters*, *41*, 8673–8680, doi:10.1002/2014GL061586, 2014.

Hesse, M., Y.-H. Liu, L.-J. Chen, N. Bessho, M. Kuznetsova, J. Birn, and J. L. Burch, On the electron diffusion region in asymmetric reconnection with a guide magnetic field, *Geophysical Research Letters*, *43*(6), 2359–2364, doi:https://doi.org/10.1002/2016GL068373, 2016.

Hochreiter, S., and J. Schmidhuber, Long short-term memory, *Neural Comput.*, *9*(8), 1735–1780, doi:10.1162/neco.1997.9.8.1735, 1997.

Hu, Y. Q., Z. Peng, C. Wang, and J. R. Kan, Magnetic merging line and reconnection voltage versus imf clock angle: Results from global mhd simulations, *Journal of Geophysical Research: Space Physics*, *114*(A8), doi:https://doi.org/10.1029/2009JA014118, 2009.

Huang, S. Y., et al., Electromagnetic energy conversion at dipolarization fronts: Multispacecraft results, *Journal of Geophysical Research: Space Physics*, *120*(6), 4496–4502, doi:https://doi.org/10.1002/2015JA021083, 2015.

Huang, X., H. Wang, L. Xu, J. Liu, R. Li, and X. Dai, Deep learning based solar flare forecasting model. i. results for line-of-sight magnetograms, *The Astrophysical Journal*, *856*(1), 7, 2018.

Hwang, K. J., et al., Electron Vorticity Indicative of the Electron Diffusion Region of Magnetic Reconnection, *grl*, *46*(12), 6287–6296, doi:10.1029/2019GL082710, 2019.

Karimabadi, H., W. Daughton, and J. Scudder, Multi-scale structure of the electron diffusion region, *Geophysical Research Letters*, *34*(13), doi:https://doi.org/10.1029/2007GL030306, 2007.

Khotyaintsev, Y. V., et al., Electron jet of asymmetric reconnection, *Geophysical Research Letters*, *43*(11), 5571–5580, doi:https://doi.org/10.1002/2016GL069064, 2016.

Kingma, D. P., and J. Ba, Adam: A Method for Stochastic Optimization, *arXiv e-prints*, arXiv:1412.6980, 2014.

Koga, D., W. D. Gonzalez, V. M. Souza, F. R. Cardoso, C. Wang, and Z. K. Liu, Dayside magnetopause reconnection: Its dependence on solar wind and magnetosheath conditions, *Journal of Geophysical Research: Space Physics*, *124*(11), 8778–8787, doi:https://doi.org/10.1029/2019JA026889, 2019.

Laundal, K. M., C. C. Finlay, N. Olsen, and J. P. Reistad, Solar wind and seasonal influence on ionospheric currents from swarm and champ measurements, *Journal of Geophysical Research: Space Physics*, *123*(5), 4402–4429, doi:https://doi.org/10.1029/2018JA025387, 2018.

Lavraud, B., et al., Currents and associated electron scattering and bouncing near the diffusion region at Earth's magnetopause, *grl*, *43*(7), 3042–3050, doi:10.1002/2016GL068359, 2016.

Lenouvel, Q., et al., Identification of electron diffusion regions with a machine learning approach on mms data at the earth's magnetopause, *Earth and Space Science*, *8*(5), e2020EA001,530, doi:https://doi.org/10.1029/2020EA001530, e2020EA001530 2020EA001530, 2021.

Li, W. Y., et al., Electron bernstein waves driven by electron crescents near the electron diffusion region, *Nature Communications*, *11*(1), 141, doi:10.1038/s41467-019-13920-w, 2020.

Lindqvist, P.-A., et al., The spin-plane double probe electric field instrument for mms, *Space Science Reviews*, *199*(1), 137–165, doi:10.1007/s11214-014-0116-9, 2016.

Linnainmaa, S., Taylor expansion of the accumulated rounding error, *BIT Numerical Mathematics*, *16*, 146–160, 1976.

Lundin, R., S. Barabash, Y. Futaana, M. Holmström, H. Perez-de Tejada, and J.-A. Sauvaud, A large-scale flow vortex in the venus plasma tail and its fluid dynamic interpretation, *Geophysical Research Letters*, *40*(7), 1273–1278, doi:https://doi.org/10.1002/grl.50309, 2013.

Malakit, K., Asymmetric magnetic reconnection: A particle-in-cell study, Ph.D. thesis, University of Delaware, 2012.

Messenger, R., and L. Mandell, A modal search technique for predictive nominal scale multivariate analysis, *Journal of the American Statistical Association*, *67*(340), 768–772, doi: 10.1080/01621459.1972.10481290, 1972.

Nguyen, G., N. Aunai, D. Fontaine, E. L. Pennec, J. V. den Bossche, A. Jeandet, B. Bakkali, L. Vignoli, and B. R.-S. Blancard, Automatic detection of interplanetary coronal mass ejections from in situ data: A deep learning approach, *The Astrophysical Journal*, *874*(2), 145, doi:10.3847/1538-4357/ab0d24, 2019.

Norgren, C., et al., Finite gyroradius effects in the electron outflow of asymmetric magnetic reconnection, *Geophysical Research Letters*, *43*(13), 6724–6733, doi:https://doi.org/10.1002/2016GL069205, 2016.

Ouellette, J. E., B. N. Rogers, M. Wiltberger, and J. G. Lyon, Magnetic reconnection at the dayside magnetopause in global lyon-fedder-mobarry simulations, *Journal of Geophysical Research: Space Physics*, *115*(A8), doi:https://doi.org/10.1029/2009JA014886, 2010.

Paschmann, G., et al., Large-Scale Survey of the Structure of the Dayside Magnetopause by MMS, *Journal of Geophysical Research (Space Physics)*, *123*(3), 2018–2033, doi:10.1002/2017JA025121, 2018.

Phan, T. D., et al., Mms observations of electron-scale filamentary currents in the reconnection exhaust and near the x line, *Geophysical Research Letters*, *43*(12), 6060–6069, doi:10.1002/2016GL069212, 2016.

Pollock, C., et al., Fast plasma investigation for magnetospheric multiscale, *Space Science Reviews*, *199*(1), 331–406, doi:10.1007/s11214-016-0245-4, 2016.

Priest, E., and T. Forbes, *Magnetic Reconnection: MHD Theory and Applications*, Cambridge University Press, 2007.

Pritchett, P. L., and F. V. Coroniti, Three-dimensional collisionless magnetic reconnection in the presence of a guide field, *Journal of Geophysical Research: Space Physics*, *109*(A1), doi:https://doi.org/10.1029/2003JA009999, 2004.

Pulkkinen, T. I., M. Palmroth, H. E. J. Koskinen, T. V. Laitinen, C. C. Goodrich, V. G. Merkin, and J. G. Lyon, Magnetospheric modes and solar wind energy coupling efficiency, *Journal of Geophysical Research: Space Physics*, *115*(A3), doi:https://doi.org/10.1029/2009JA014737, 2010.

Reistad, J. P., K. M. Laundal, A. Ohma, T. Moretto, and S. E. Milan, An explicit imf b dependence on solar wind-magnetosphere coupling, *Geophysical Research Letters*, *47*(1), e2019GL086,062, doi:https://doi.org/10.1029/2019GL086062, e2019GL086062 10.1029/2019GL086062, 2020.

Retinò, A., et al., Structure of the separatrix region close to a magnetic reconnection x-line: Cluster observations, *Geophysical Research Letters*, *33*(6), doi:https://doi.org/10.1029/2005GL024650, 2006.

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

Russell, C. T., M. M. Mellott, E. J. Smith, and J. H. King, Multiple spacecraft observations of interplanetary shocks: Four spacecraft determination of shock normals, *Journal of Geophysical Research: Space Physics*, *88*(A6), 4739–4748, doi:https://doi.org/10.1029/JA088iA06p04739, 1983.

Russell, C. T., et al., The magnetospheric multiscale magnetometers, *Space Science Reviews*, *199*(1), 189–256, doi:10.1007/s11214-014-0057-3, 2016.

Samuel, A. L., Some studies in machine learning using the game of checkers, *IBM Journal of Research and Development*, *3*(3), 210–229, doi:10.1147/rd.33.0210, 1959.

Schapire, R. E., The strength of weak learnability, *5*(2), doi:10.1023/A:1022648800760, 1990.

Scurry, L., C. T. Russell, and J. T. Gosling, Geomagnetic activity and the beta dependence of the dayside reconnection rate, *Journal of Geophysical Research: Space Physics*, *99*(A8), 14,811–14,814, doi:https://doi.org/10.1029/94JA00794, 1994a.

Scurry, L., C. T. Russell, and J. T. Gosling, A statistical study of accelerated flow events at the dayside magnetopause, *Journal of Geophysical Research: Space Physics*, *99*(A8), 14,815–14,829, doi:https://doi.org/10.1029/94JA00793, 1994b.

Shay, M. A., J. F. Drake, and M. Swisdak, Two-scale structure of the electron dissipation region during collisionless magnetic reconnection, *Physical Review Letters*, *99*(15), doi: 10.1103/physrevlett.99.155002, 2007.

Shue, J. H., J. K. Chao, H. C. Fu, C. T. Russell, P. Song, K. K. Khurana, and H. J. Singer, A new functional form to study the solar wind control of the magnetopause size and shape, *jgr*, *102*(A5), 9497–9512, doi:10.1029/97JA00196, 1997.

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

Swisdak, M., Quantifying gyrotropy in magnetic reconnection, *grl*, *43*(1), 43–49, doi:10.1002/2015GL066980, 2016.

Swisdak, M., J. F. Drake, M. A. Shay, and J. G. McIlhargey, Transition from antiparallel to component magnetic reconnection, *Journal of Geophysical Research: Space Physics*, *110*(A5), doi:https://doi.org/10.1029/2004JA010748, 2005.

Swisdak, M., J. F. Drake, L. Price, J. L. Burch, P. A. Cassak, and T.-D. Phan, Localized and intense energy conversion in the diffusion region of asymmetric magnetic reconnection, *Geophysical Research Letters*, *45*(11), 5260–5267, doi:10.1029/2017GL076862, 2018a.

Swisdak, M., J. F. Drake, L. Price, J. L. Burch, P. A. Cassak, and T.-D. Phan, Localized and intense energy conversion in the diffusion region of asymmetric magnetic reconnection, *Geophysical Research Letters*, *45*(11), 5260–5267, doi:https://doi.org/10.1029/2017GL076862, 2018b.

Toledo-Redondo, S., et al., Electrostatic spacecraft potential structure and wake formation effects for characterization of cold ion beams in the earth's magnetosphere, *Journal of Geophysical Research: Space Physics*, *124*(12), 10,048–10,062, doi:10.1029/2019JA027145, 2019.

Trattner, K. J., J. S. Mulcock, S. M. Petrinec, and S. A. Fuselier, Location of the reconnection line at the magnetopause during southward imf conditions, *Geophysical Research Letters*, *34*(3), doi:10.1029/2006GL028397, 2007a.

Trattner, K. J., J. S. Mulcock, S. M. Petrinec, and S. A. Fuselier, Probing the boundary between antiparallel and component reconnection during southward interplanetary magnetic field conditions, *Journal of Geophysical Research: Space Physics*, *112*(A8), doi:10.1029/2007JA012270, 2007b.

Trattner, K. J., S. M. Petrinec, and S. A. Fuselier, The location of magnetic reconnection at earth's magnetopause, *Space Science Reviews*, *217*(3), 41, doi:10.1007/s11214-021-00817-8, 2021.

Trattner, K. J., et al., The response time of the magnetopause reconnection location to changes in the solar wind: Mms case study, *Geophysical Research Letters*, *43*(10), 4673–4682, doi:https://doi.org/10.1002/2016GL068554, 2016.

Wang, S., et al., Whistler waves associated with electron beams in magnetopause reconnection diffusion regions, *Earth and Space Science Open Archive*, p. 40, doi:10.1002/essoar.10511586.1, 2022.

Webster, J. M., et al., Magnetospheric Multiscale Dayside Reconnection Electron Diffusion Region Events, *Journal of Geophysical Research (Space Physics)*, *123*(6), 4858–4878, doi:10.1029/2018JA025245, 2018.

Wilder, F. D., et al., The role of the parallel electric field in electron-scale dissipation at reconnecting currents in the magnetosheath, *Journal of Geophysical Research: Space Physics*, *123*(8), 6533–6547, doi:https://doi.org/10.1029/2018JA025529, 2018.

Wintoft, P., M. Wik, J. Matzka, and Y. Shprits, Forecasting kp from solar wind data: input parameter study using 3-hour averages and 3-hour range values, *Journal of Space Weather and Space Climate*, *7*, A29, 2017.

Wu, J., X.-Y. Chen, H. Zhang, L.-D. Xiong, H. Lei, and S.-H. Deng, Hyperparameter optimization for machine learning models based on bayesian optimizationb, *Journal of Electronic Science and Technology*, *17*(1), 26 – 40, doi:https://doi.org/10.11989/JEST.1674-862X.80904120, 2019.

Xiong, Q. Y., et al., Formation of negative j e in the outer electron diffusion region during magnetic reconnection, *Journal of Geophysical Research: Space Physics*, *127*(2), e2022JA030,264, doi:https://doi.org/10.1029/2022JA030264, e2022JA030264 2022JA030264, 2022.

Xu, J., F. Z. Peng, C. T. Russell, B. Giles, P.-A. Lindqvist, R. B. Torbert, R. E. Ergun, and J. L. Burch, Observation of energy conversion near the x-line in asymmetric guide-field reconnection, *The Astrophysical Journal*, *895*(1), L10, doi:10.3847/2041-8213/ab8e31, 2020.

Yamada, M., J. Yoo, and S. Zenitani, *Energy Conversion and Inventory of a Prototypical Magnetic Reconnection layer*, pp. 143–179, Springer International Publishing, Cham, doi:10.1007/978-3-319-26432-5_4, 2016.

Yi, Y., M. Zhou, L. Song, and X. Deng, On the energy conversion rate during collisionless magnetic reconnection, *The Astrophysical Journal*, *883*(1), L22, doi:10.3847/2041-8213/ab40c1, 2019.

Zenitani, S., M. Hesse, A. Klimas, and M. Kuznetsova, New measure of the dissipation region in collisionless magnetic reconnection, *Physical Review Letters*, *106*(19), doi:10.1103/physrevlett.106.195003, 2011.

Zhong, Z. H., et al., Extension of the electron diffusion region in a guide field magnetic reconnection at magnetopause, *The Astrophysical Journal*, *892*(1), L5, doi:10.3847/2041-8213/ab7b7c, 2020.

Zhou, M., H. Man, Y. Yang, Z. Zhong, and X. Deng, Measurements of energy dissipation in the electron diffusion region, *Geophysical Research Letters*, *48*(24), e2021GL096,372, doi:https://doi.org/10.1029/2021GL096372, e2021GL096372 2021GL096372, 2021.