



**HAL**  
open science

# Solar wind / magnetosphere coupling inferred from machine-learning methods

Gautier Nguyen

► **To cite this version:**

Gautier Nguyen. Solar wind / magnetosphere coupling inferred from machine-learning methods. Earth and Planetary Astrophysics [astro-ph.EP]. Université Paris-Saclay, 2021. English. NNT : 2021UP-ASP012 . tel-03198435

**HAL Id: tel-03198435**

**<https://theses.hal.science/tel-03198435>**

Submitted on 14 Apr 2021

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

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

# Solar wind/magnetosphere coupling inferred from machine-learning methods

**Thèse de doctorat de l'Université Paris-Saclay**

École doctorale n° 127 Astronomie et Astrophysique d'Ile de France  
Spécialité de doctorat: Astronomie et astrophysique  
Unité de recherche: Université Paris-Saclay, CNRS, Ecole polytechnique, LPP,  
91128, Palaiseau, France.  
Réfèrent: : Faculté des Sciences d'Orsay

**Thèse présentée et soutenue à Palaiseau, le 1 Février 2021, par**

**Gautier Nguyen**

## Composition du jury:

<b>Karine Bocchialini</b> Professeur, Université Paris-Saclay	Présidente
<b>Philippe Louarn</b> Directeur de recherche , Université Paul Sabatier, Toulouse	Rapporteur et examinateur
<b>Zdenek Nemecek</b> Professeur, Faculty of Mathematics and Physics, Charles University, Prague	Rapporteur et examinateur
<b>Jonathan Eastwood</b> Maître de conférences, Imperial College, Londres	Examineur
<b>Vincent Genot</b> Astronome, Université Paul Sabatier, Toulouse	Examineur

## Direction de la thèse

<b>Dominique Fontaine</b> Directrice de recherche, Laboratoire de Physique des Plasmas	Directrice de thèse
<b>Nicolas Aunai</b> Chargé de recherche, Laboratoire de Physique des Plas- mas	Co-directeur de thèse

## Membre invité

<b>Maud Pastel</b> Referent Technique, Direction Générale de l'Armement	
--	--

---

Aujourd'hui, la seule chose que je trouvais à la fin de cette quête était une perte. Rien de plus normal: une quête ne s'achève pas par une trouvaille, car une quête n'a jamais de fin. En réalité, un quêteur n'a pas besoin de trouver: seule la quête est importante. Elle donne un sens à nos vies. Alors je compris que j'étais un quêteur, que je l'avais toujours été, que j'étais fier de l'être.

---

Luis Montero Manglano  
(La Table du roi Salomon)

---

## Remerciements

Au lecteur qui s'apprête à lire le présent manuscrit, tu trouveras là le fruit de trois années d'une quête tant scientifique qu'émotionnelle où l'extase du résultat s'oppose à la monotonie des péripéties qui y ont mené. Et si tu n'as probablement pas envie de lire le détail de ce que fut ma vie ces trois années durant, permet moi au moins, ô lecteur, de commencer mon récit en rendant les honneurs à l'ensemble des personnes qui m'ont accompagné, de près ou de loin, pendant tout ce temps.

Ma famille tout d'abord, ma mère, la première à m'inciter à faire une thèse, mon père, toujours soucieux de ma sécurité, ma grand-mère, partenaire d'infortune attirée, mon beau-père, mes soeurs, mon frère, ma tante et mes cousins pour leur soutien sans faille depuis 25 années maintenant.

La Direction Générale de l'Armement ensuite, pour m'avoir autorisé à faire cette thèse et grâce à qui j'ai pu découvrir les enjeux qui m'attendaient, au travers de la Formation Administrative et Militaire des Ingénieurs de l'Armement et d'un mois de mer entre Méditerranée et Océan Indien. Je porte à ce titre une attention toute particulière à Maud Pastel qui a accepté de suivre ma thèse quand bien même mon domaine lui était parfaitement inconnu et à Gwladys Theuillon grâce à qui je prends aujourd'hui un poste s'inscrivant dans la continuité de mes travaux de ces 3 dernières années.

Mes directeurs de thèse enfin, Dominique Fontaine et Nicolas Aunai, puisqu'une quête digne de ce nom ne peut se faire sans bons mentors. Dominique, pour sa disponibilité malgré le poids de ses responsabilités et l'engouement dont elle a fait preuve devant chacun de mes résultats aussi médiocres soient-ils m'incitant par la même occasion à les approfondir. Merci d'avoir donné un sens physique à ma première année et qui sait, peut-être un jour ferai-je l'étude statistique des ICMEs détectées. Nicolas, pour sa rigueur, sa curiosité et sa dévotion. Merci d'avoir accompagné ma progression numérique d'abord, observationnelle et physique ensuite ces trois dernières années durant en étant présent tout en me laissant la marge d'autonomie dont j'avais besoin pour pleinement m'épanouir. Merci aussi pour la patience et la pédagogie dont tu as fait preuve jusqu'à ce qu'enfin, à force d'acharnement je puisse mettre un sens en complexité sur l'ensemble des résultats que nous avons produits, quand bien même ceux-ci se sont finalement avérés être plutôt *"not that good"*. Merci finalement pour avoir insisté sur ces détails que je jugeais souvent futiles mais dont l'attention finissait souvent par nous révéler son lot d'intrigues et d'inattendus dont l'existence même justifiait ton intuition première.

Je remercie également l'ensemble des personnes avec qui j'ai pu travailler au cours de ces trois dernières années. Alexis Jeandet, mon chef d'équipe, pour sa maîtrise numérique inégalée et pour la disponibilité dont il a fait preuve que ce soit pour m'aider à installer Tensorflow ou remettre DSK puis Juno à flot quand je leur en demandais un petit peu trop. Erwan Le Pennec pour son expertise en Machine Learning et dont l'intuition a grandement facilité le développement de mes méthodes de détection. L'équipe du CDS Paris-Saclay, Balazs Kegl, Joris Vandebosche et Guillaume Lemaire pour l'organisation du RAMP et pour avoir utilisé le problème de la détection des ICMEs le temps d'un challenge data science. Bayane Michotte de Welle pour sa contribution à l'étiquetage sisyphéen des données et à la détection des traversées de choc, je lui souhaite désormais bon courage pour la thèse qui l'attend. Philippe Garnier, David Ecoffet, Thibaud Noel et Quentin Lenouvel pour les discussions sur la détection des chocs martiens et des EDRs en espérant avoir pu apporter ma pierre à l'édifice sur ces deux problèmes. J'accorde enfin une attention toute particulière à Benoît Lavraud pour l'aide apportée sur la détection des jets de reconnexion et sans qui ma conclusion finale sur l'indentation aurait été toute autre.

Sur le long chemin menant à la fin de la thèse, la moindre tâche anodine peut parfois s'avérer étonnamment ardue. Aussi, après une centaine de mails et refus administratifs, je remercie du fond du coeur l'ensemble des personnes qui ont accepté de faire partie de mon jury: mes deux rapporteurs, Philippe Louarn et Zdenek Nemecek auprès de qui je m'excuse par avance pour la moindre erreur physique, technique ou grammaticale passée entre les mailles des filets de la relecture, Karine Bocchialini pour en avoir accepté la présidence et Jonathan Eastwood pour l'intérêt porté à mon sujet. Je remercie encore plus grandement Vincent Génot qui, après avoir été le premier à

---

me parler de magnétopause quand je n'étais encore qu'en master et après de fructueuses réunions sur la détection d'événements par machine learning, a accepté de suivre mon évolution jusqu'à la fin de la thèse en devenant un des examinateurs.

Une quête c'est également des déceptions et des infortunes qui font prendre à l'aventure des détours, inattendus mais revêtant tout leur sens au moment de les replacer dans leur contexte. Pour ces raisons, je tiens à remercier dans un premier temps le referee de mon premier article sans qui nous n'aurions jamais envisagé de montrer que les performances de nos algorithmes étaient limitées par l'interprétation des données par un observateur externe. Dans un deuxième temps, j'exprime toute ma gratitude envers les referees 1 et 3 de mon deuxième article pour en avoir refusé la publication. Sans ce rejet, je ne me serai très probablement pas lancé dans l'étude statistique de la position de la magnétopause et je ne me serai très probablement pas posé toutes les questions que j'ai pu me poser ces 4 derniers mois.

Je n'oublie pas non plus le Laboratoire de Physique des Plasmas pour m'avoir accueilli et supporté pendant 3 ans. Merci à mon co-bureau, Clément Moissard pour les discussions sur tes résultats bizarres de simulation auxquels j'ai très souvent eu du mal à te proposer des solutions. Merci à Vitaliy en espérant ne pas avoir été un co-bureau trop bruyant. Merci à Alexandra Alexandrova, Olivier Le Contel et Mojtaba Akhavan-Tafti pour les discussions autour de mes résultats et de la détection des jets. Merci à l'équipe de gestion, Maryline, Cherifa et Edouard pour leur efficacité et leur patience vis à vis de mon incompréhension administrative. Merci à Olivier Guaitella pour avoir accepté de faire partie de mon comité de suivi et à Anne Bourdon pour le soutien moral des articles rejetés. Merci à l'ensemble des stagiaires, doctorants et post-doctorants du LPP pour les soirées films, les cafés, les apéros et les journées doctorants ou enfin j'en apprenais plus sur vos recherches respectives: Alejandro, Victor, Abhyuday, Ana Sofia, Antoine et Florian que j'ai suivi de l'X à la fin de la thèse, Benjamin, Edmond, Constance, Giulia, Annemarie, Clarence et Pierre. Merci finalement à Renaud Ferrand pour les magnans 11h30 passés à essayer de comprendre le pourquoi du comment d'IP Paris, l'expertise jeu de société et le soutien administratif quand il nous a fallu nous inscrire à Paris-Saclay.

De manière plus large, je remercie les doctorants de l'Ecole polytechnique pour ces moments qui nous faisaient oublier le dur quotidien de nos recherches respectives. Giuliano, Olga, Pierre, Clothilde, Pauline, Arnaud, Ambra, Svetlana, Nikolaï, Marco et, bien sûr, mon acolyte, Théo Courtois, petit ange parti trop tôt que j'espère vite revoir dans le cockpit d'un avion. Merci pour les réunions Doc'Union où nous entretenions l'espoir d'une interaction inter-laboratoires et inter-cursus globale. Merci pour les soirées du mardi soir, les barbecues et autres sessions poster. Merci pour les week-ends à Dammarie-les-Lys. Je n'oublie pas non plus les personnes de l'administration de l'X; Alexandra Belus, Emmanuel Fullenwarth, Audrey Lemaréchal et, évidemment, Elodie Lelaidier que je remercie tout particulièrement pour son efficacité, sa dévotion et sa bonne humeur à chacune de nos réunions.

La thèse, c'est aussi des conférences, des voyages, l'occasion de faire connaître sa recherche en cours mais aussi et surtout l'occasion de rencontrer du monde et d'élargir son horizon de connaissances. Je remercie donc les personnes rencontrées à l'AGU, à l'EGU ou encore au PNST pour les discussions animées sur nos recherches respectives, pour les soirées jeux et karaoké ou pour les simples visites de Vienne, Washington ou encore Biarritz. Matti, Erika, Jennimari, Léa, Eleanna, Jérémy, Naïs, et mon ami, Michael Lavarra, avec qui je brûle d'impatience de célébrer nos fins de thèse respectives d'ici quelques semaines. Je remercie par la même occasion mes compagnons d'embarquement à bord du BPC Dixmude, qu'ils soient de ma promotion ou administrateurs des affaires maritimes pour la découverte du milieu opérationnel, que je serai forcément amené à côtoyer, et pour les visites inattendues à Haïfa, Jerusalem ou Djibouti.

Ces trois années durant, le sport s'est avéré être un excellent moyen d'évacuer les frustrations des résultats non satisfaisants et de prendre le temps de s'évader pour mieux ensuite hiérarchiser sa pensée. Je remercie l'ensemble Supaero Handball pour les matchs à Bercy et la médaille de bronze au TSGED. J'ai également une pensée pour l'équipe de foot du LPP/LOA/Exotrail pour les matchs et entraînements de la pause midi, promis, un jour vous gagnerez le tournoi inter-labo et

---

d'ici là, j'apprendrai à cadrer mes frappes. J'en remercie tout particulièrement le capitaine, mon sosie et ami Thomas Charoy. Merci pour ton sens hors pair de la tactique, merci pour m'avoir aidé à maintenir Doc'Union à flots quand tout semblait désespéré et j'ai hâte de pouvoir lire la version éditée de ton livre à défaut de t'avoir vraiment aidé dans la relecture. Finalement, je remercie l'équipe de quidditch des Olympiens de Paris pour m'avoir fait découvrir la communauté quidditch grâce à laquelle j'ai pu pleinement m'épanouir émotionnellement, pour les voyages, les coupes d'Europe, la médaille de bronze et la mirabelle cup. J'ai hâte de vous retrouver sur les terrains après plusieurs mois d'absence. Un merci tout particulier à Hugo, Tess, Daniela, Solene, Sylvie et Clara pour le soutien moral des tous derniers mois et en encore plus grand merci à Denis, mon presque voisin de Massy, qui a eu la patience et la motivation de lire l'intégralité de ce manuscrit à la recherche de la moindre faute d'anglais.

Je remercie enfin mes amis d'école pour avoir suivi de près ou de loin l'évolution de ma thèse ces trois années durant. Merci à la Technical Hotline qui a résolu nombre de mes problèmes git, Python et autre LaTeX. Merci au phlyse pour les soirées NI et le fromage. Merci la RHDS pour la découverte des frangines. Merci à Paul pour m'avoir supporté dans sa coloc pendant toute ma première année. Merci aux Lazos pour les soirées jeux et age of, pour les vacances en Grèce et pour votre soutien dans les moments difficiles, hâte de célébrer mon doctorat en votre compagnie dans quelques semaines dans les Alpes.

Last but not least, ma toute dernière pensée, et non des moindres, est pour Naomi, pour sa patience, sa douceur, son intelligence et auprès de qui je m'excuse sincèrement d'avoir commencé notre relation par le confinement et la fin de thèse et que je ne remercierai jamais assez pour son soutien sans faille aux moments critiques de la fin de la rédaction et de la préparation de la soutenance.

---

## Abstract

The solar wind and the Earth magnetosphere form a complex duet which dynamics is ruled by a multitude of physical processes at every steps of this coupling. Upstream, large scale solar events such as [Interplanetary Coronal Mass Ejections \(ICMEs\)](#) transport important quantities of plasma and magnetic field which entry in the magnetosphere generates geomagnetic storms with a high impact on human activities. At the Earth proximity, varying solar wind conditions generate small-scale physical processes such as magnetic reconnection that rule the entire dynamics of the system.

Although the different elements of this coupling have been studied from an observational point of view for decades and by an important number of missions, the manual collection of the in-situ signature of the events of interest in the data is still a subjective, fastidious and hardly reproducible task and thus affects the quality of their associated statistical studies.

In this thesis, we take a step further in the direction of a global, statistically representative vision of the different actors of the solar wind-magnetosphere coupling by applying supervised machine learning algorithms to the automatic detection of their in-situ signatures.

In particular, we apply an ensemble of [Convolutional Neural Networks \(CNNs\)](#) to the automatic detection of [ICMEs](#), we use a gradient boosting algorithm to provide an automatic classification of the near-Earth regions and we combine the latter with a second gradient boosting classifier to detect plasma jets issued from magnetic reconnection at the magnetopause.

In the three cases, the method we develop outperforms state of the art automatic detection methods based on manual empirical thresholds on a reduced number of physical parameters. We also show that these methods are adaptable from a mission to another provided the regions visited by the concerned spacecraft share the same physical nature and offer the advantage of improving their detection performance with the simple increasing amount of data with time. This paves the way to the elaboration of additional detection methods, inspired from the one we develop, applied to the other actors of this coupling. However, we also show that the errors made by these algorithms is actually comparable to the difference that exists in the interpretation of in-situ data by two different human observers and that the quality of the predictions made by these methods is thus limited by the vision we have on the data and the events they measure .

These methods allow the rapid and reproducible elaboration of extensive multi-mission event catalogs that contain a reduced proportion of [False Positives \(FPs\)](#). These catalogs can then be used for further statistical analysis of in-situ measured events with an important number of samples. For instance, we use the magnetopause crossings catalogs obtained with the region classifier on the data of near-Earth missions with equatorial ([THEMIS](#), [MMS](#), [Double Star](#)), polar ([Cluster](#)) and lunar ([ARTEMIS](#)) to perform a statistical analysis of the position and shape of the magnetopause for different solar wind and seasonal conditions.

In the first place, this study confirms long-proved characteristics of the magnetopause such as the influence of the solar wind dynamic pressure or the azimuthal asymmetry induced by seasonal variations. In the second place, we bring answer elements to still open questions such as the influence of the [Interplanetary Magnetic Field \(IMF\)](#) radial component or the actual existence of a dawn-dusk asymmetry. In particular, we evidence the influence of the [IMF](#)  $B_y$  by showing that a varying clock angle affects the shape of the magnetopause because of the displacement of the reconnection sites it induces. These results are condensed into an analytical non-indented magnetopause model that offers a more precise description of this boundary on the night side of the magnetosphere.

Finally, we come back on the question of the near-cusp indentation of the magnetopause. We show that the crossings identified by both other researchers and by our method actually correspond to the crossings of the cusp inner boundary and result in overestimating the supposed magnetopause cusp indentation in models. We show that accounting for actual magnetopause current sheet crossings instead drastically changes the result, increasing the radial distance of the magnetopause boundary, although still showing an apparent depletion in comparison with a non-indented model.

---

## Résumé

La dynamique des différentes étapes du couplage entre le vent solaire et la magnétosphère terrestre est régie par une multitude de processus physiques aux échelles spatio-temporelles significativement différentes. En amont de la terre, des évènements solaires de grande envergure comme les Ejections de Masse Coronales Interplanétaires (ICMEs) transportent d'importantes quantités de plasma et de champ magnétique dont l'entrée dans la magnétosphère est à l'origine de tempêtes géomagnétiques aux effets dévastateurs sur les activités humaines. Au voisinage de la Terre, les variations des paramètres physiques du vent solaire sont à l'origine de processus physiques à petite échelle, tels que la reconnection magnétique, régissant l'ensemble de la dynamique du système.

Si les différents acteurs de ce couplage sont étudiés de manière observationnelle par de nombreuses missions depuis des décennies, la sélection manuelle d'évènements d'intérêt par reconnaissance de leur signature in-situ reste une tâche subjective, chronophage et difficilement reproductible affectant nécessairement la vision globale offerte par les études statistiques qui en découlent.

Dans ce travail de thèse, nous effectuons un pas supplémentaire vers l'acquisition d'une vision globale, statistiquement représentative des différents acteurs du couplage vent solaire - magnétosphère en appliquant des algorithmes d'apprentissage supervisé à la détection automatique de leurs signatures in-situ respectives.

En particulier, nous appliquons un ensemble de réseaux de neurones convolutionnels (CNNs) à la détection automatique des Ejections de Masse Coronale Interplanétaires (ICMEs), nous utilisons deux algorithmes à boosting de gradient, l'un pour classifier automatiquement les différentes régions de l'environnement terrestre proche, l'autre pour détecter les jets de plasma produits par la reconnexion magnétique à la magnétopause.

Dans les trois cas d'utilisation, les méthodes développées s'avèrent être plus performantes que les méthodes de détection automatiques basées sur l'utilisation manuelle de seuils établis sur un nombre réduits de paramètres physiques. Nous montrons également que ces méthodes sont adaptables d'une mission à une autre pourvu que les régions visitées par les différentes sondes ont la même nature physique. Ces méthodes ont également l'avantage de bénéficier d'une augmentation de leurs performances par simple augmentation de la quantité de données traitées. Les résultats obtenus sont le préambule à l'élaboration de méthodes de détection supplémentaires appliquées aux autres acteurs du couplage. Nous montrons cependant que les erreurs faites par ces algorithmes sont comparables aux divergences d'interprétation de données in-situ existantes entre deux observateurs humains différents. La qualité des prédictions de ces méthodes étant par conséquent limitées par l'interprétabilité que nous avons des données et des évènements qu'elles dérivent.

Ces méthodes permettent l'élaboration rapide et reproductible des catalogues multi-missions d'évènements qui contiennent une proportion réduite de Faux Positifs (FPs) tout en étant parmi les plus exhaustifs existants. Ces catalogues pouvant par la suite être exploités dans le cadre d'études statistiques d'évènements observés in-situ. En l'occurrence, nous utilisons le catalogue de traversées de magnétopause obtenu avec le classifieur de régions appliqué aux données des missions ayant des orbites équatoriales (THEMIS, MMS, Double Star), polaires (Cluster) et lunaires (ARTEMIS) pour réaliser une étude statistique de la position et de la forme de la magnétopause en fonction des conditions physiques du vent solaire et saisonales.

En premier lieu, cette étude confirme des caractéristiques de la magnétopause suggérées par de nombreuses études existantes telles que l'influence de la pression dynamique ou la non-axisymétrie azimuthale induite par les variations saisonales. En deuxième lieu, nous apportons des éléments de réponse aux questions encore ouvertes comme l'influence de la composante radiale du Champ Magnétique Interplanétaire (IMF) ou l'existence de l'asymétrie aube-crêpuscule. En particulier, nous mettons en évidence l'influence de la composante  $y$  de l'IMF  $B_y$  en montrant qu'une variation de l'angle horaire affecte le coefficient d'explosion de la magnétopause. L'ensemble de ces résultats sont condensés dans un modèle analytique, non indenté de la posi-

---

tion de la magnétopause qui offre une meilleure description de cette frontière sur le côté nuit de la magnétosphère.

Finalement, nous revenons sur la question de l'indentation de la magnétopause dans les cornets polaires. Nous suggérons que les traversées détectées par notre algorithme à boosting de gradient ainsi que les événements pris en compte par les précédentes études sur le sujet correspondent en fait aux frontières internes du cornet polaire et on tendance à sur-estimer la profondeur de l'indentation supposée dans cette région. Nous montrons, en prenant en compte des traversées de la couche de courant caractéristique de la magnétopause que cette dernière se situe en fait à des distances radiales plus élevées bien que les positions prises en compte continuent de suggérer la présence d'une déplétion vis-à-vis d'un modèle non-indenté.

# Contents

<b>Contents</b>	<b>ix</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Solar wind and the near-Earth environment	2
1.2 Large-scale solar events	7
1.3 The magnetopause, boundary between the solar wind and the magnetosphere	10
1.4 Small-scale physical processes of the near-Earth environment	15
1.5 Summary	19
1.6 Bibliography	20
<b>2 Machine learning as an automatic selection tool</b>	<b>27</b>
2.1 Introduction	28
2.2 Different types of supervised classification	29
2.3 Evaluating the performances of an algorithm	36
2.4 Implementing Machine Learning	39
2.5 Machine Learning in space physics	39
2.6 Bibliography	40
<b>3 An example of ambiguously labeled problem: automatic detection of ICMEs</b>	<b>43</b>
3.1 Introduction	44
3.2 Interplanetary Coronal Mass Ejections	44
3.3 Data	47
3.4 Algorithm	49
3.5 Results	54
3.6 Robustness	63
3.7 Global quality of the prediction	66
3.8 Conclusion	67
3.9 Bibliography	68
<b>4 Automatic classification of the three near-Earth regions</b>	<b>71</b>
4.1 Introduction	72
4.2 Data	73
4.3 Labeling THEMIS data	75
4.4 Algorithm selection	78
4.5 Algorithm performance	79
4.6 Adaptability of the model: from a mission to the other	80
4.7 Comparison with manually set thresholds	89
4.8 Massive detection of boundary crossings	90
4.9 Conclusion	94
4.10 Bibliography	95

<b>5</b>	<b>Statistical analysis of the magnetopause shape and location</b>	<b>101</b>
5.1	Introduction . . . . .	102
5.2	A brief insight on the magnetopause shape and location models . . . . .	103
5.3	Statistical analysis of the magnetopause crossing . . . . .	106
5.4	Fitting a new magnetopause model . . . . .	118
5.5	Nature of the near-cusp magnetopause . . . . .	123
5.6	Conclusion . . . . .	133
5.7	Bibliography . . . . .	134
<b>6</b>	<b>Automatic detection of magnetopause plasma flow</b>	<b>141</b>
6.1	Introduction . . . . .	142
6.2	Construction of the dataset . . . . .	145
6.3	Jet detection pipeline . . . . .	148
6.4	Method performance . . . . .	150
6.5	Adaptability of the method . . . . .	153
6.6	Massive detection of magnetopause plasma jets . . . . .	161
6.7	Conclusion . . . . .	162
6.8	Bibliography . . . . .	163
<b>7</b>	<b>Conclusions and perspectives</b>	<b>169</b>
7.1	Overview . . . . .	170
7.2	Application of supervised machine learning algorithms . . . . .	171
7.3	Position and shape of the magnetopause . . . . .	173
7.4	Potential of machine learning algorithms and larger perspectives . . . . .	175
<b>A</b>	<b>Coordinates systems</b>	<b>I</b>
A.1	Geocentric Solar Ecliptic (GSE) . . . . .	I
A.2	Geocentric Solar Magnetospheric (GSM) . . . . .	I
A.3	Local Magnetopause Normal (LMN) . . . . .	II
A.4	Magnetic Local Time (MLT) . . . . .	II
<b>B</b>	<b>Additional prediction examples</b>	<b>III</b>
B.1	ICMEs . . . . .	IV
B.2	Near-Earth regions . . . . .	VI
B.3	Cusp external boundary . . . . .	XI
B.4	Reconnection jets . . . . .	XII
<b>C</b>	<b>List of Acronyms</b>	<b>XV</b>

# Chapter 1

## Introduction

*C'est une planète  
Indomptable et secrète  
Où se jouent les rouages  
De nos humeurs sauvages*

---

Les Frangines

### Contents

---

<b>1.1 Solar wind and the near-Earth environment</b> . . . . .	<b>2</b>
1.1.1 Solar wind . . . . .	2
1.1.2 The near-Earth environment . . . . .	3
1.1.3 In-situ observations of the solar wind-magnetosphere coupling . . . . .	6
<b>1.2 Large-scale solar events</b> . . . . .	<b>7</b>
<b>1.3 The magnetopause, boundary between the solar wind and the magnetosphere</b> .	<b>10</b>
1.3.1 MHD discontinuities . . . . .	10
1.3.2 Location and shape of the magnetopause . . . . .	11
<b>1.4 Small-scale physical processes of the near-Earth environment</b> . . . . .	<b>15</b>
1.4.1 Magnetic reconnection . . . . .	15
1.4.2 Magnetic reconnection at the Earth magnetopause . . . . .	16
<b>1.5 Summary</b> . . . . .	<b>19</b>
<b>1.6 Bibliography</b> . . . . .	<b>20</b>

---

## 1.1 Solar wind and the near-Earth environment

### 1.1.1 Solar wind

The solar wind is the permanent stream of plasma in the interplanetary medium originating from the solar corona. Its existence was predicted by the theoretical work of Parker [1958] and first observed by spacecraft in 1959 with the mission Luna 1 [Beatty, 2007].

Mainly composed of protons and  $\alpha$  particles ( $\text{He}^{2+}$ ), it propagates in the interplanetary medium at supersonic and super-Alfvénic speed that is to say  $V > V_s = \sqrt{\gamma P_t / \rho}$  and  $V > V_a = B / \sqrt{\mu_0 \rho}$  where  $V$  is the velocity,  $P_t$  is the thermal pressure,  $\gamma$  is the ratio of specific heats,  $B$  is the amplitude of the magnetic field transported by the solar wind, the so-called **Interplanetary Magnetic Field (IMF)**,  $\rho$  is the particle density and  $\mu_0$  is the magnetic constant.

The **IMF** is frozen in the plasma. That is to say, the solar wind can be considered as a perfectly conducting plasma and the **IMF** field lines are transported in the interplanetary medium at the velocity of the solar wind. Because of the solar rotation, the **IMF** field lines are arranged into a spiral shape known as the Parker spiral [Parker, 1963]. A schematic representation of such arrangement in the ecliptic plane is shown in the Figure 1.1. Despite of this nominal, spiral configuration, the orientation of the **IMF** is still highly variable. This variability, transported at 1 **Astronomical Unit (AU)**<sup>1</sup> is a major actor of the dynamics of the interaction between the solar wind and the near-Earth environment. This importance will be discussed later-on. In particular, we will discuss the importance of the so-called clock angle  $\Omega$  defined as the angle of the projection of **IMF** in the plane perpendicular to the Sun-Earth direction. In the rest of this thesis, this direction will be defined as the X-axis. The convention adopted during the whole manuscript to define  $\Omega$ , in the **Geocentric Solar Magnetospheric (GSM)** coordinate system detailed in the Appendix A, is shown in the Figure 1.2.

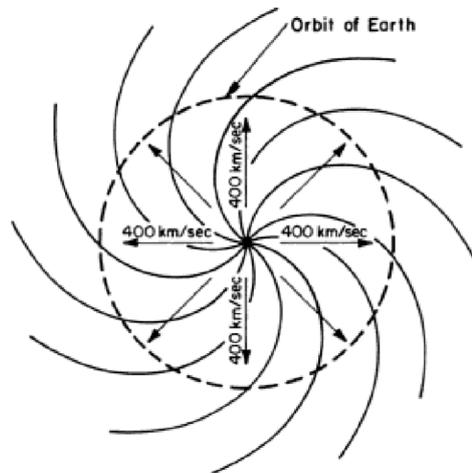


Figure 1.1: Representation of the Parker spiral (from Parker [1963])

The first observation of the solar wind by Luna 1 in 1959 was followed by a multitude of missions dedicated to Sun observation, from the ecliptic observatories SOHO and STEREO to the recently launched Parker Solar Probe and Solar Orbiter without forgetting the polar orbit of Ulysses. The permanent monitoring of the solar wind arriving at the Earth orbit is possible since the launch and the positioning of the missions WIND and ACE at the Earth Lagrange point L1 in the late 1990s. The combination of the data of these two missions and their time-shift to the nose of the Earth bow shock, the point on the Sun-Earth axis at which the solar wind becomes subsonic, forms the so-called NASA/GSFC's OMNI database [King and Papitashvili, 2005].

<sup>1</sup>Sun-Earth distance =  $150 \times 10^6$  km

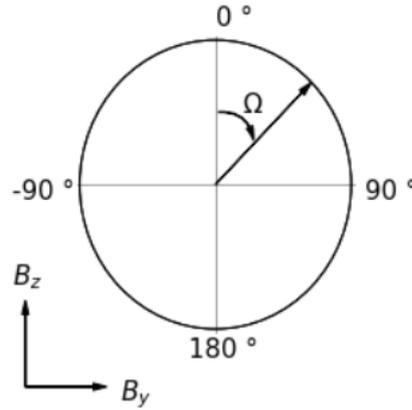


Figure 1.2: Schematic representation of the definition of the IMF clock angle  $\Omega$ .

Figure 1.3 represents the distribution of several solar wind parameters between 2001 and 2019. At the Earth orbit, the density of the plasma is in the orders of  $5 \text{ cm}^{-3}$  for an average velocity of approximately 400 km/s resulting in an average dynamic pressure of approximately 2 nPa. As a direct consequence of the Parker spiral, the IMF, that has an average amplitude in the orders of 5 nT, is essentially westward or eastward (e.g  $|\Omega| \sim 90^\circ$ ) with a fairly important radial component  $B_x$  which sign depends on the orientation of the Heliospheric current sheet, the surface where the polarity of the Sun's magnetic field switches from North to South.

### 1.1.2 The near-Earth environment

Since the IMF is frozen in the solar wind, the interaction between the solar wind and the dipole-like magnetic field of the Earth creates a cavity around our planet known as the *magnetosphere* [Chapman and Ferraro, 1931].

On the dayside part of the near-Earth environment, the geomagnetic field lines resulting from this interaction are compressed and bent earthward. This compression limits the extension of the dayside geomagnetic field lines to an average 8-9  $\text{Re}^2$  on the Sun-Earth axis. On the night side, the geomagnetic field lines are stretched out anti-sunward for more than 200  $\text{Re}$  forming the so-called *magnetotail*.

The magnetosphere constitutes an obstacle along the solar wind propagation in the interplanetary medium. Consequently, the solar wind that arrives in the near-Earth environment is slowed down and deflected forming in the process a collisionless *bow shock* downstream of which the solar wind becomes subsonic. The bow shock constitutes a first boundary that delimits the near-Earth environment from the upstream solar wind and one of the crucial surface of interest when studying the magnetosphere-solar wind coupling.

Downstream of the bow shock is a region where the plasma is compressed, decelerated and heated. In this region, the so-called *magnetosheath*, the IMF is also compressed and the associated field lines are draped around the magnetosphere [Kobel and Flückiger, 1994]. Downstream of the bow shock, the typical ion density of the shocked solar wind in the magnetosheath is about  $20 \text{ cm}^{-3}$  for an average velocity of approximately 200 km/s and a magnetic field amplitude in the orders of 20 nT.

<sup>2</sup>Earth radii, 1  $\text{Re} = 6371 \text{ km}$

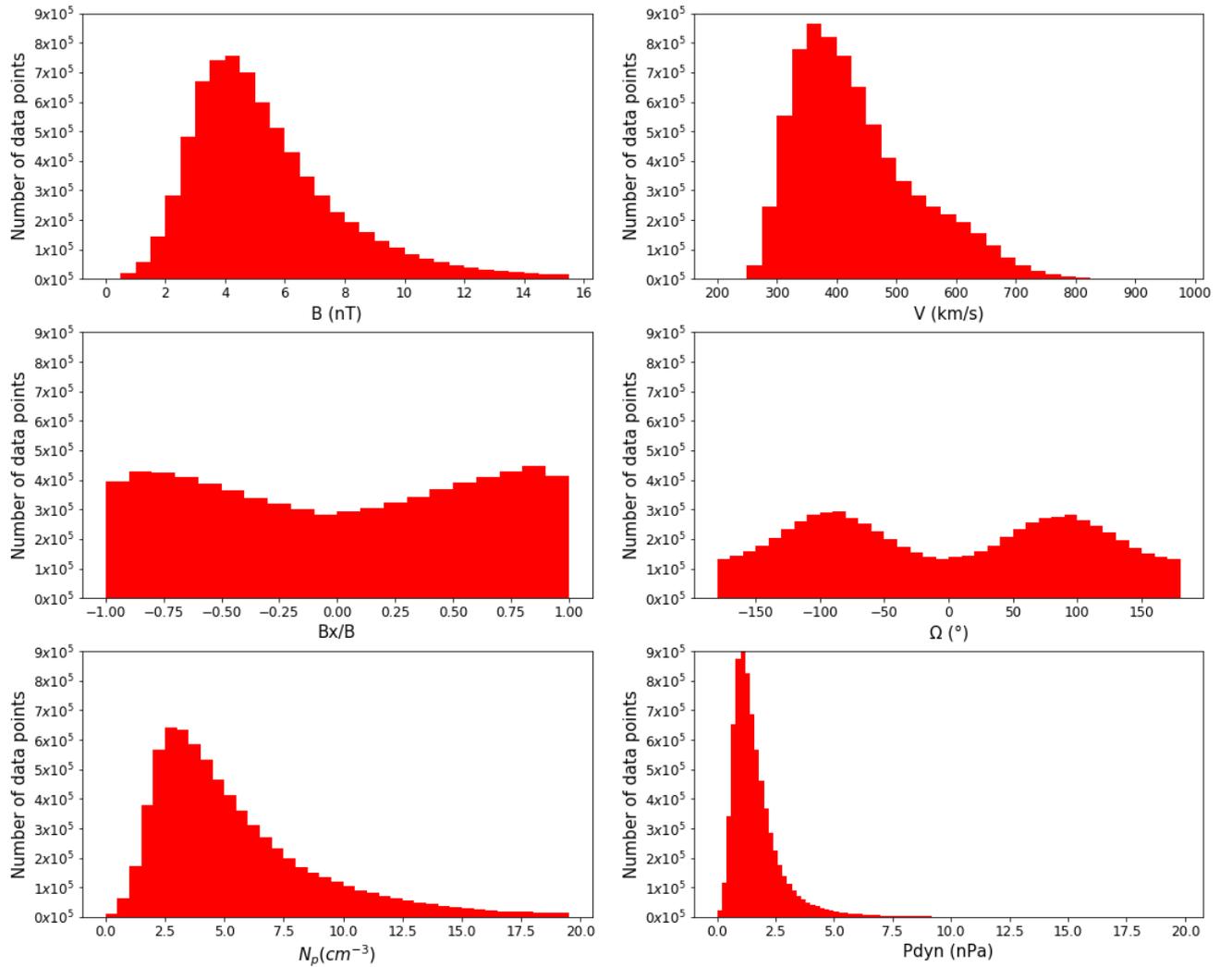


Figure 1.3: Distribution of the solar wind parameters of the OMNI database between 2001 and 2019: the amplitude of the IMF  $B$  (top left), the velocity  $V$  (top right), the IMF radial orientation  $B_x/B$  (middle left), the IMF clock angle  $\Omega$  (middle right), the proton density  $N_p$  (bottom left) and the dynamic pressure  $P_{dyn}$  (bottom right).

The boundary that delimits the shocked solar wind of the magnetosheath from the interior of the magnetosphere is called the *magnetopause*. At first order, this is the surface at which the solar wind and the magnetosphere total pressure balance each other and the current layer that, under ideal [Magnetohydrodynamics \(MHD\)](#) assumptions, prevents any transport of the shocked solar wind into the magnetosphere. In practice, the magnetopause is a dynamic boundary, strongly influenced by the physical parameters of the upstream solar wind, and affected by a multitude of small-scale physical processes occurring in its vicinity. Among them, magnetic reconnection, the merging and reconfiguration of the non-parallel geomagnetic and [IMF](#) field lines, is of uttermost importance as it allows the penetration of solar wind plasma and magnetic flux in the magnetosphere. This boundary will be one of the main topic of this thesis, whether it concerns its detection from satellite in-situ measurements, the identification of the small-scale physical processes occurring in its vicinity, or the statistical analysis of its location and shape. A more detailed description of both the magnetopause and magnetic reconnection will be given in the next section.

All regions of the near-Earth environment that we have presented so far are shown on the [Figure 1.4](#). In addition to the three "main" regions and the two boundaries that were previously discussed, the magnetosphere can be divided in additional regions hereafter described:

- The *polar cusps*, where the geomagnetic field lines fan out from the magnetic poles. These regions of particularly weak magnetic field are a privileged place of entry for the solar wind particles. The topography of this region and the associated shape of the magnetopause is particularly affected by magnetic reconnection as this will be seen in the chapter 5.
- The *plasma sheet*, the region located around the tail mid plane where most of the magnetotail plasma is concentrated. It is characterized by an average density of  $0.5 \text{ cm}^{-3}$  and a low magnetic field. In this region, the shear between the geomagnetic field lines of the two hemispheres is high and almost antiparallel. Therefore, the plasma sheet is centered around a current sheet called the *neutral sheet*.
- The *lobes*, a region almost empty of plasma (with typical densities lower than  $0.001 \text{ cm}^{-3}$ ) located between the plasma sheet and the magnetopause for which the geomagnetic field lines are open and stretched anti-sunward.
- The *plasmasphere*, much denser than the magnetospheric and solar wind plasmas located just outside of the ionosphere, the layer that separates the Earth magnetosphere from the atmosphere.
- The two *Van Allen radiation belts* made of highly energetic particles trapped on closed geomagnetic field lines at 2 and 6 *Re*.

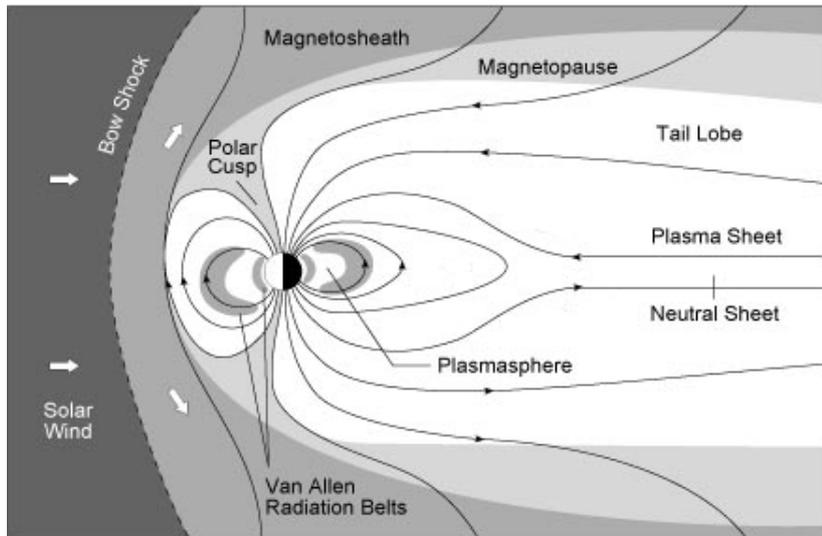


Figure 1.4: Meridional view of the different regions of the near-Earth environment. (Source: [https://ase.tufts.edu/cosmos/print\\_images.asp?id=29](https://ase.tufts.edu/cosmos/print_images.asp?id=29))

### 1.1.3 In-situ observations of the solar wind-magnetosphere coupling

The magnetosphere and the solar wind constitute a complex duet in permanent interaction at every scale. The variations of the solar wind modify the properties of the near-Earth regions, their boundaries, and generate small-scale physical processes such as magnetic reconnection that strongly affects the dynamics of this interaction. On the other hand, large-scale solar events such as [Coronal Mass Ejections \(CMEs\)](#), perturb the solar wind during their propagation in interplanetary space and strongly affect the magnetosphere by the formation of geomagnetic storm that can have huge consequences on the human activity.

With the easy access to this specific region of the interplanetary medium for a spacecraft launched from Earth, numerous are the missions focused on the study of the solar wind and its relation with the magnetosphere whether they are solar wind monitors at the Lagrange point L1 (Wind and ACE in particular) or explorers of the different regions of the near-Earth environment (Cluster, Double Star, [THEMIS](#) and [MMS](#) to mention just a few of them).

The increasing number of these missions and the associated analysis of their in-situ data measurements led to the multiplication of case studies that increased our knowledge on the different physical processes at stake.

Nowadays, the accumulation of decades of spacecraft in-situ data measurement allows the elaboration of massive, global statistical studies of the different actors of the solar wind-magnetosphere interaction that use the data of several different missions at the same time. In particular, we can cite the studies that have been made on the near-Earth regions ([Zhang et al. \[2019\]](#), [Lavraud et al. \[2004a\]](#) and references therein), their boundaries ([Hasegawa \[2012\]](#), [Paschmann et al. \[2018\]](#), [Němeček et al. \[2020\]](#) and references therein) and the physical processes at small ([Lewis and Fuselier \[2011\]](#), [Hoshi et al. \[2018\]](#), and references therein) and large scales ([Kilpua et al. \[2017\]](#), [Richardson \[2018\]](#), [Chi et al. \[2016\]](#) and references therein) that rule the interaction of the magnetosphere with the solar wind.

As they concatenate an important number of samples, such statistical studies contribute to the elaboration of a global, statistical vision of the different physical processes that affect the Sun-Earth relation. Nevertheless, they often rely on the manual selection of events of interest in the streaming in-situ time series data provided by spacecraft. This, in addition to being time-consuming, is an ambiguous task, strongly linked to the interpretation of an external observer and poorly reproducible<sup>3</sup>. This necessarily limits the information one can extract from the associated statistical studies. With the increasing number of spacecraft dedicated to the study of the

<sup>3</sup>Even by the authors themselves!

near-Earth environment and the ever-growing amount of data provided by the totality of these spacecraft, the proportion of selected data will represent an even smaller proportion of the total accessible data, which spoil the potential of their overall consideration.

From now on, the elaboration of automatic event detection methods in streaming in-situ time series data provided by spacecraft appears as an interesting option to accelerate the collection of data and improve the reproducibility of statistical studies. For this purpose, manually set thresholds on the values of physical quantities appear as the most intuitive and fastest solution one can think of to improve the detection [Jelínek et al., 2012; Lepping et al., 2005]. Nevertheless, these methods are limited by the high variability of in-situ data and the manual setting of optimal thresholds is another time-consuming, ambiguous and hardly reproducible task, limited by the visual inspection of huge quantities of data. Moreover, these methods are often tested on the dataset on which they have been developed and there are generally no clues on how well they do on unknown sets of data and how easy it is to apply them to the data of different missions. Additionally, the thresholds are often based on a reduced number of parameters and lose in efficiency when several physical features must be considered at the same time.

An interesting option we have to overcome these constraints then stands in using supervised machine learning algorithms. These algorithms, that have the ability to learn to perform a certain task after being trained on a given dataset, represent a promising tool to tackle already large and ever-growing bases of reliable data accumulated for decades, and their use in space physics is therefore progressing [Camporeale, 2019].

To what extent can these algorithms help us improve the automatic detection of the signatures in streaming in-situ data of the different physical processes that rule the interaction between the solar wind and the Earth magnetosphere? How can they help constructing a global, statistically representative vision of these processes?

The objectives of this thesis are then as follows:

- Investigate the potential of different machine learning algorithms in the optics of the fast, automatic and reproducible identification of the in-situ signatures of the different processes that intervene in the Sun-Earth interaction.
- Use these algorithms to perform a massive detection of the events of interest in the accumulated decades of in-situ data and generate some of the most exhaustive events catalogs.
- Exploit these catalogs to take a step further in the improvement of our global vision of the magnetosphere through the realization of massive, global statistical studies of the phenomena at stake.

Naturally, the question can be asked at every level of the interaction, and the application of machine learning algorithms thus has potential all along the chain of events from the Sun to the magnetosphere as this will be discussed in the next sections.

## 1.2 Large-scale solar events

In addition to the production of the solar wind, the physical processes that occur in the solar corona can generate large-scale solar events which, through the transport of important quantities of plasma and magnetic field, induce large-amplitude perturbations of the nominal solar wind with possible serious consequences regarding the dynamics of the near-Earth environment.

Among them, CMEs are the spectacular expulsion of large quantities of plasma and magnetic field in the interplanetary medium from the solar corona.

Suggested by the theoretical work of Chapman and Ferraro [1929] to explain geomagnetic disturbances in the Earth magnetosphere, their existence was confirmed by the ground observations of Hansen et al. [1971] and the on-board observations of Tousey [1973].

Produced in the solar corona, the ejecta propagates and expands in the interplanetary medium. This propagation of the ejecta, the so-called ICME, was first observed in by the measurements

of the interplanetary proton density and temperature by the Vela 3 spacecraft [Gosling et al., 1973] and linked to the production of CMEs by the combination of the data of 5 different spacecraft by Klein and Burlaga [1981].

The observation of a CME with the coronagraph of the Solar and Heliospheric Observatory (SOHO) spacecraft along with the schematic representation of an ICME arriving at Earth orbit are shown on the two panels of Figure 1.5.

The propagation of the ejecta in the interplanetary medium is accompanied by a three-dimensional expansion. At 1 AU, the average estimated width of an ICME is in the order of 0.3 AU [Wang et al., 2005]. The transported magnetic field is commonly described as having helical field lines forming a so-called flux rope [Goldstein, 1983]. By definition, the amplitude of the magnetic field is stronger than the average IMF and the typical values of the magnetic field of ICMEs found at 1 AU is in the order of 10 nT [Kilpua et al., 2017]. Following the ejection at large velocities in the solar corona, the transported plasma is usually faster than the preceding nominal solar wind, at 1 AU, the average velocity found for ICMEs is in the orders of 450 km/s. When the velocity of the ejecta is fast enough, the ICME pushes the downstream solar wind at supersonic speeds and is thus drives a shock. The propagation of this shock wave is at the origin of a turbulent plasma region of low anisotropy between the shock and the main body of the ICME, the so-called sheath [Klein and Burlaga, 1981; Moissard et al., 2019].

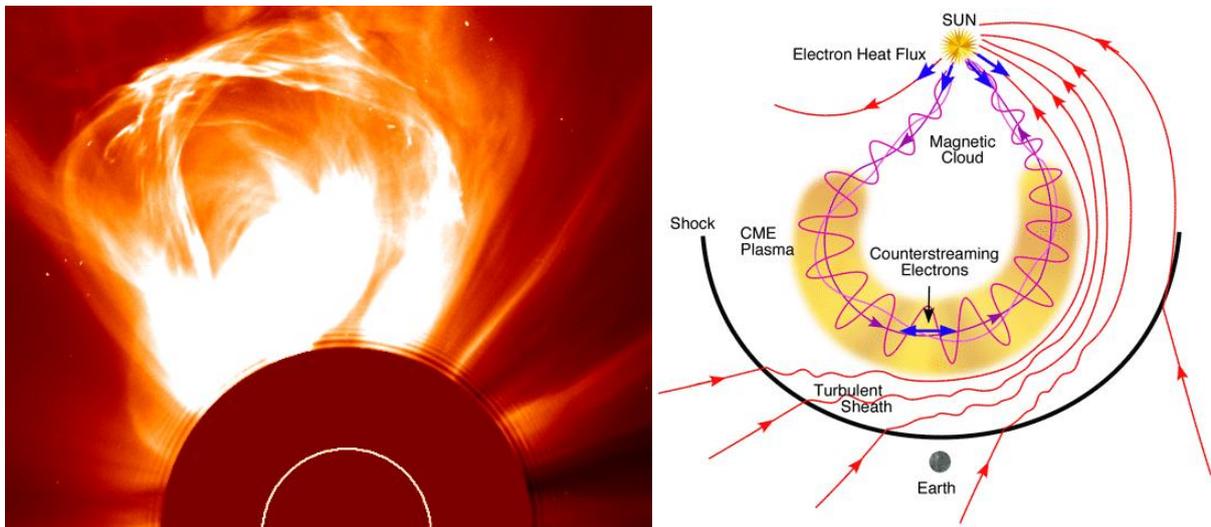


Figure 1.5: CME observed by the SOHO spacecraft on the 27th of February 2000 (right) and schematic representation of an ICME (right) (Adapted from Zurbuchen and Richardson [2006]).

These are the general average properties of ICMEs and these properties are also visible in solar wind observations during such events as shown in the Figure 1.6. If the different characteristics we just described define the commonly agreed in-situ signatures of these events, their identification is actually much more ambiguous and strongly related to the interpretation of an external observer [Shinde and Russell, 2003]. In Chapter 3, we will give a particular focus on how the application of machine learning highlights the difficulties inherent to the identification of ICMEs from in-situ data measurement.

The link between the CMEs and the geomagnetic disturbances recorded in the Earth magnetosphere was confirmed by Wilson [1987] who noticed the particular importance of the southward component of the IMF on the triggering of geomagnetic storms. From then on, storms were also observed for northward IMF [Du et al., 2008] and the CMEs are considered as the most geoeffective solar events [Yermolaev et al., 2012] at the origin of the greatest part of geomagnetic storms [Echer et al., 2005]. Among the various hazards already caused by these events on the human activity, one can typically cite the Bastille Day event [Webber et al., 2002], one of the largest geomagnetic storm ever recorded in space, or the March 1989 geomagnetic storm that led to an electrical power blackout in the entire Quebec [Boteler, 2019].

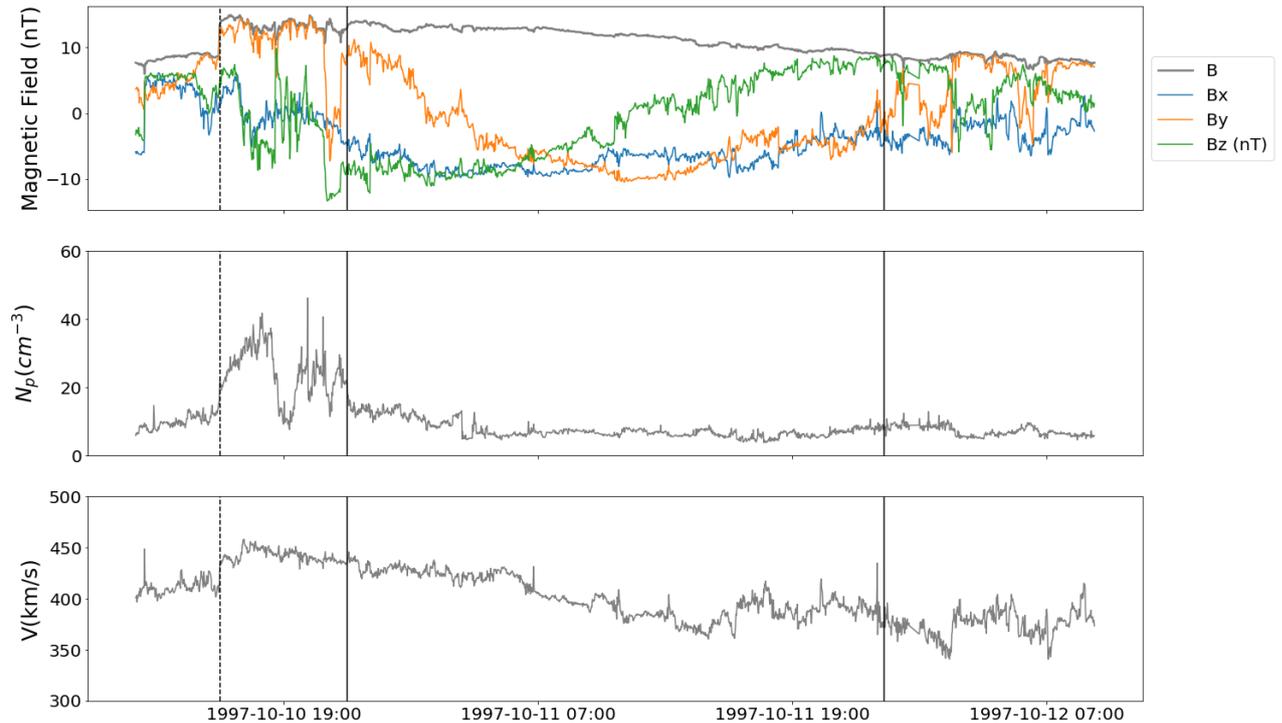


Figure 1.6: Solar wind observation during an ICME from the WIND spacecraft located at the Lagrangian Point L1. The solid vertical lines delimitate the ICME while the dashed vertical line indicate the beginning of the sheath. From the top to the bottom are represented : the magnetic field amplitude and components, the proton density and the solar wind velocity.

Eventhough the geoeffectiveness is expected to be related to the large quantities of plasma and magnetic field transported by the ejecta [Turc et al., 2014], the physical properties of ICMEs that are the most likely to affect the magnetospheric activity are still under debate [Kilpua et al., 2017].

In the development of space weather, a global, statistically representative vision of CMEs would be the opportunity to better understand the nature of these events and how they interact with the magnetosphere and affect the human activity. For this purpose, the increasing number of solar wind oriented missions (SOHO, STEREO, WIND, ACE, Solar Orbiter, Parker Solar Probe just to mention a few of them) led to the multiplication of the existing ICMEs catalogs and associated statistical studies of their different physical parameters [Lepping et al., 2006; Nieves-Chinchilla et al., 2018; Richardson and Cane, 2010]. Nevertheless, the lack of consensus on the typical in-situ signature of ICMEs associated to the fact these catalogs were elaborated after a manual selection of events resulted in incomplete, ambiguous and hardly reproducible lists which using masks the statistical vision we can have on such events.

In this context, elaborating automatic detection methods would allow the rapid and reproducible collection of such events for their further statistical analysis. Moreover, the analysis of the events detected by one of these methods could bring interesting information on how visual identification is made and how we interpret in-situ data measurements.

We will focus on those questions in the Chapter 3 that will entirely be dedicated to the automatic detection of ICMEs.

If CMEs are known so, they are far from being alone in the zoology of the large-scale solar events produced in the solar corona which transport of plasma and magnetic field strongly affects the Earth magnetosphere. Among this zoology, we can particularly cite the Corotating Interaction Regions (CIRs), the interaction of a stream of high speed solar wind emanating from coronal holes with the preceding slower nominal solar wind (Richardson [2018] and references therein), and the interplanetary shocks, direct consequence of the propagation of CMEs and CIRs in the interplanetary medium and responsible for the acceleration of particles to very high energies (Oliveira and Samsonov [2018] and references therein).

## 1.3 The magnetopause, boundary between the solar wind and the magnetosphere

### 1.3.1 MHD discontinuities

The different regions of the near-Earth environment are characterised by different physical properties. Consequently, the boundaries that delimit the three regions, the magnetopause and the bow shock, are discontinuities that evolve with the interaction of the plasma of the different medias.

In the frame of ideal [Magnetohydrodynamics \(MHD\)](#), when the plasma is considered as perfectly conducting, these discontinuities can be described by the Rankine-Hugoniot equations:

$$[\rho V_n] = 0 \quad (1.1)$$

$$[B_n] = 0 \quad (1.2)$$

$$[\rho V_n^2 + P + \frac{B^2}{2\mu_0}] = 0 \quad (1.3)$$

$$[\rho V_n \vec{V}_t - \frac{B_n \vec{B}_t}{\mu_0}] = \vec{0} \quad (1.4)$$

$$[B_n \vec{V}_t - V_n \vec{B}_t] = \vec{0} \quad (1.5)$$

Where  $\rho$  is the density,  $P$  is the thermal and kinetic pressure,  $[X]$  the jump of a parameter  $X$  across the discontinuity,  $X_n$  and  $X_t$  denotes respectively the normal and the tangential components of  $\vec{X}$ .

When  $[V_n] \neq 0$ , this is particularly what happens at the interface between the magnetosheath and the solar wind, the discontinuity is a shock.

Otherwise, we can distinguish three different configurations schematically represented in the [Figure 1.7](#) :

- If  $V_n = 0$  and  $B_n \neq 0$ , all the physical parameters but the density are continuous. No mass flow across the discontinuity is allowed and the jump in density is compensated by a jump in thermal pressure that rapidly disperse this so-called *contact discontinuity*.
- If  $V_n = 0$  and  $B_n = 0$ , the discontinuity is said *tangential (TD)*. The flow and the magnetic field are tangential to the discontinuity, the total pressure on the two sides balance and no mass or magnetic flux crossing is allowed. This is what happens at the magnetopause when no penetration of solar wind plasma is allowed.
- If  $V_n \neq 0$ , the discontinuity is said *rotational (RD)*. Mass flow crossing is allowed, the tangential velocity and magnetic field rotate but keep their magnitudes constant and equal to the Alfvén velocity across the discontinuity. This is for instance what happens when the [IMF](#) reconnects with the geomagnetic field as this will be detailed in the next section.

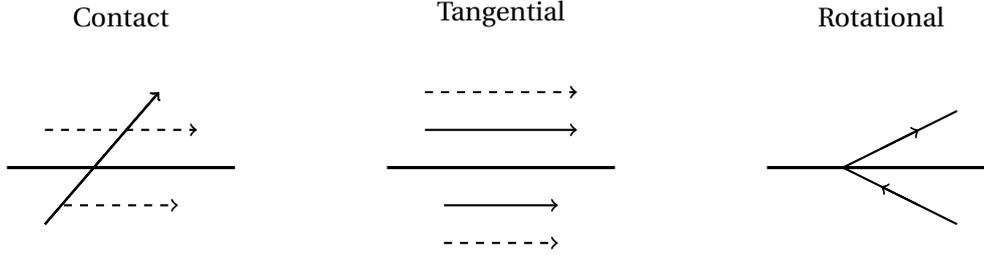


Figure 1.7: Schematic representation of a contact discontinuity (*left*), a **Tangential discontinuity (TD)** (*middle*) and a **Rotational discontinuity (RD)** (*right*), the solid arrows represent the magnetic field while the dashed arrows represent the velocity

### 1.3.2 Location and shape of the magnetopause

At first sight, magnetopause can locally be approximated by a **TD** and prevents any transport of the shocked solar wind into the magnetosphere. This boundary can be defined by the surface where the magnetosphere and the solar wind total pressure balance each other. In the solar wind, the total pressure can easily be approximated by the lone dynamic pressure  $P_{dyn} = \rho_{sw} V_{sw}^2$  because of the weak magnitude of the **IMF** and thermal pressure. In the magnetosphere, the thermal and dynamic pressures can be neglected in comparison to the magnetic pressure.

Thus, the pressure balance can be written as follows:

$$P_{dyn} \cos(\xi)^2 = \frac{B_{MSP}^2}{2\mu_0} \quad (1.6)$$

Where  $\xi$  is the incidence angle between the solar wind flow and the local magnetopause normal direction as shown in the Figure 1.8.

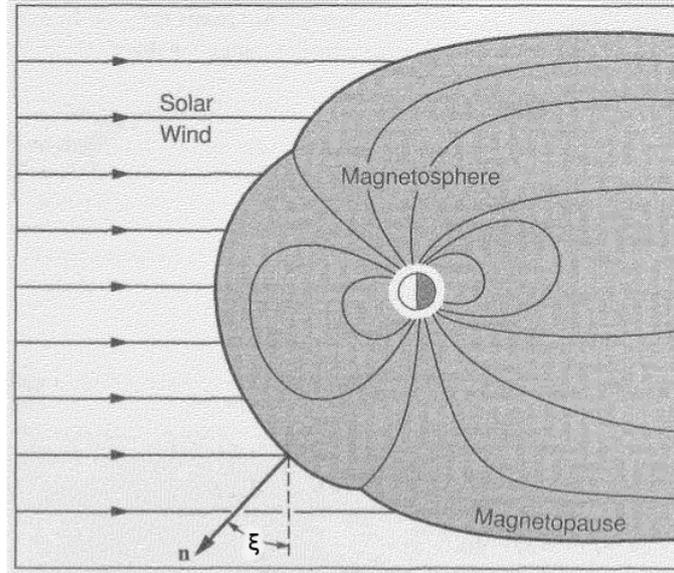


Figure 1.8: Meridional representation of the magnetopause obtained by balancing the solar wind dynamic pressure and the magnetosphere magnetic pressure (adapted from **Baumjohann and Treumann [1996]**)

When the incidence is normal (e.g when  $\xi = 0^\circ$ ), the flow in the magnetosheath reduces to 0 at the encounter with the magnetopause. Assuming a dipolar expression of the geomagnetic field, the radial distance of the magnetopause at this so-called *stagnation point*, the so-called *magnetopause stand-off distance* denoted  $r_0$  can be estimated by:

$$r_0 = \left( \frac{\kappa B_E^2}{2\mu_0 P_{dyn}} \right)^{1/6} \quad (1.7)$$

Where  $B_E$  is the magnetic field measured at the surface of the Earth and  $\kappa$  accounts for the deviation of the magnetic field from its dipolar value and the field generated at the boundary surface. Assuming  $\kappa = 2$ ,  $B_E = 3.1 \times 10^4$  nT, and  $P_{dyn} = 2$  nPa, we obtain a typical value expected for the stand-off distance under standard solar wind conditions:  $r_0 = 9.9$  Re.

An expression of the position and shape of the magnetopause can then be obtained by solving the equation (1.6). This is what was done analytically by Spreiter and Briggs [1962] and numerically by Sotirelis and Meng [1999].

A typical representation in the meridional plane of the magnetopause obtained from pressure balance is shown in the Figure 1.8. In addition to the stand-off distance, the *level of flaring* that describes the expansion of the surface in both equatorial and meridional planes is a key factor that characterises the position and shape of the magnetopause for changing solar wind conditions.

By solving the equation 1.6, Spreiter and Briggs [1962] noticed that the topological change of the geomagnetic field line at the polar cusps induced discontinuities between the dayside and the nightside magnetopause in these regions. These discontinuities are represented by the singular points on the magnetopause in the two hemispheres in the Figure 1.8. In practice, the magnetopause can be continuously extended in these regions through the introduction of an *indentation* that consider the geometry of the polar cusps with varying solar wind and seasonal conditions.

Following this theoretical definition, the first observation of the magnetopause was made with the measurements of Explorer 12 by Cahill and Amazeen [1963]. From then on, the accumulation of the missions that came across the different boundaries of the near-Earth environment allowed the multiplication of the studies focused on the magnetopause, whether they concern its location and shape (Němeček et al. [2020] and references therein) or its global dynamics (Hasegawa [2012]; Paschmann et al. [2018] and references therein). The collection of several observed magnetopause crossings allowed the establishment of empirical analytical models of the magnetopause shape and location that kept improving with the evidences of the influences of the different solar wind and seasonal parameters (Fairfield [1971]; Jelínek et al. [2012]; Lin et al. [2010]; Liu et al. [2015]; Shue et al. [1997] just to mention a few).

These observations also proved that the magnetopause is the theater of small-scale plasma processes that contribute to the dynamics of the boundary. Among them, magnetic reconnection fundamentally affects the location and shape of the magnetopause through the convection of the geomagnetic field lines it rearranges. Naturally, the evidence of this phenomenon questions the first definition of the magnetopause we gave. How do this process affect the position and shape of the magnetopause? How can we consider it in the frame of an analytical model fitted from in-situ data?

The generation and the characteristics of these processes is strongly dependent on the associated upstream solar wind conditions. Consequently, their effects on the magnetopause location and shape are seen through the variations of the boundary with changing solar wind and seasonal conditions.

With the important number of studies dedicated to the subject, the influence of some of these parameters has been showed for long: the dynamic pressure pushes the magnetopause earthward when increasing and the IMF  $B_z$  component increases the azimuthal flaring while reducing the equatorial one when negatively decreasing. The importance of some other parameters, the two other components of the IMF  $B_x$  and  $B_y$  for instance, is however unclear and still under debate. Additionally, the previous existing studies are limited in the night side and there is no indication if what we know about the magnetopause holds in the far night side where the identification of the magnetopause becomes even more ambiguous. Last but not least, the existence of magnetic reconnection fundamentally affects the topography of the polar cusp and blurs the nature of the magnetopause in this region. There is thus no clue about the reality of the theoretically predicted indentation.

In the wake of the existing studies, answering these open questions requires the collection of as many magnetopause crossings as possible.

A typical in-situ signature of a magnetopause crossing by the THEMIS E spacecraft is repre-

sented in the Figure 1.9. At first, the high density above  $20 \text{ cm}^{-3}$  and the omnidirectional differential energy ion fluxes indicate the spacecraft is in the magnetosheath, as the velocity, shown on the third panel, is low, the crossing happens near the stagnation point. Past 12 : 00, one can notice a drop in density followed by a jump in  $B_z$ , the spacecraft has crossed the magnetopause and is now in the magnetosphere. The interpretation of the high velocity peaks highlighted with the green intervals will be detailed in the next section.

In practice, the in-situ measurement of magnetopause crossings are not as clear in every region of the near-Earth environment and for every upstream solar wind condition and the identification of such events much less obvious<sup>4</sup>. The motion of the magnetopause with changing conditions result in partial crossings and the nature of the data measured by spacecraft with polar orbit is quite different from the nature of the data measured on an equatorial orbit. Additionally, the manual selection of such events is necessarily ambiguous and time-consuming. The automation of this task would then be an important improvement in the elaboration of statistical studies of the different properties of the magnetopause.

Because of the drawbacks of methods based on manually-set thresholds, one can see the potential of machine learning algorithms in the realisation of this task. The application of such algorithms in the frame of the automatic detection of the near-Earth regions and boundaries will be discussed in the Chapter 4 and the magnetopause crossings detected with these methods will be exploited through the statistical study of the magnetopause shape and location in the Chapter 5.

---

<sup>4</sup>Additional observational examples of magnetopause crossings are shown in the Appendix B.

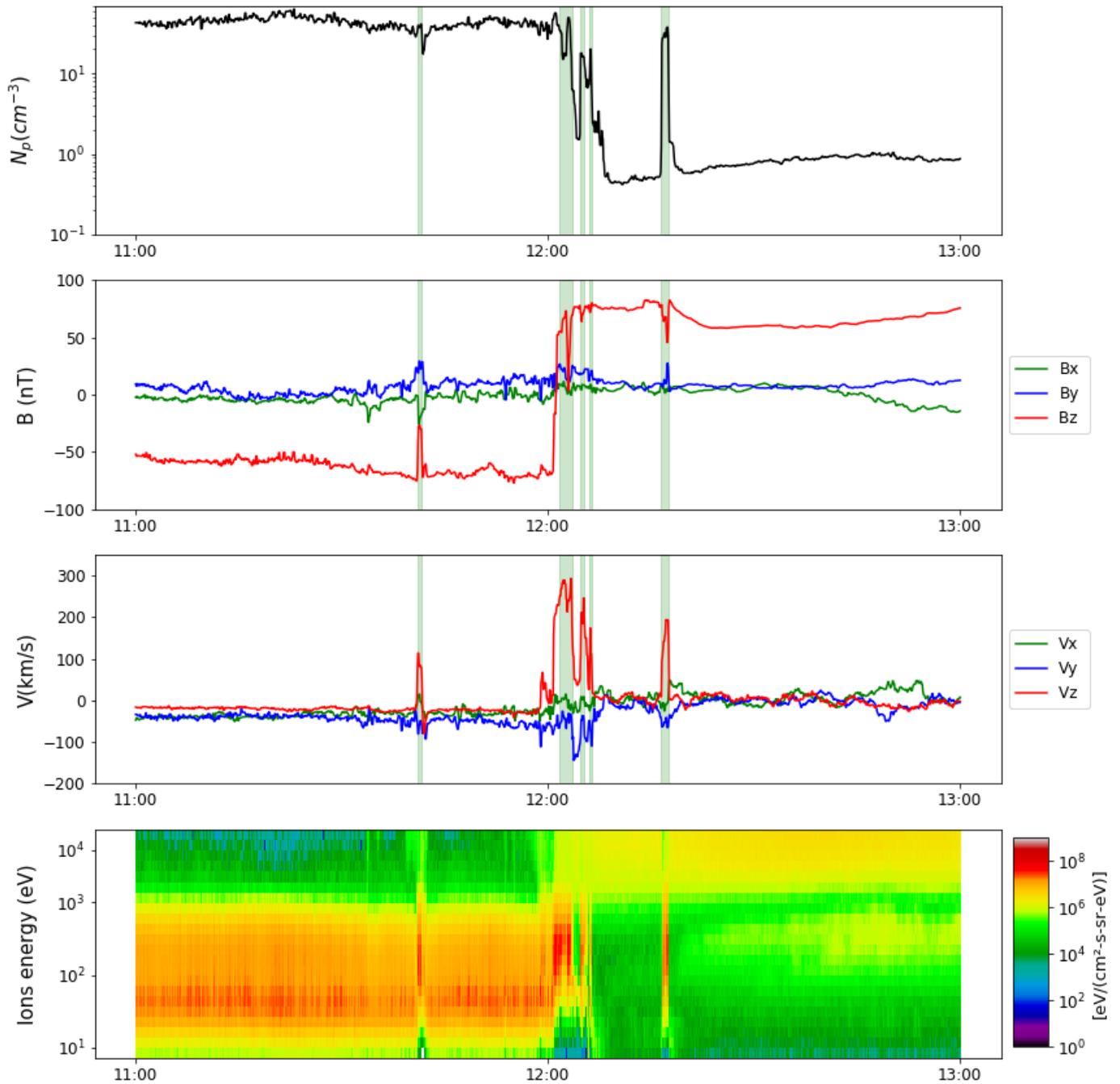


Figure 1.9: In-situ spacecraft measurement provided by the THEMIS E spacecraft during a magnetopause crossing on the 30<sup>th</sup> of April 2014. From top to bottom are represented the proton density, the three components of the magnetic field, the three components of the velocity and the omnidirectional differential energy fluxes of ions. The green intervals highlight magnetic reconnection plasma jets.

## 1.4 Small-scale physical processes of the near-Earth environment

Because the solar wind and the magnetosphere are two plasmas of different nature, their interaction is likely to create small-scale physical processes that strongly affects the dynamics of the system

Among them, we can for example mention Kelvin-Helmholtz instability that results in the propagation of surface waves along the magnetopause (Kivelson and Zu-Yin [1984] and references therein) or magnetic reconnection that occurs when two non-parallel field lines are merged and topologically rearranged.

The latter was evidenced as the dominant process when it comes to the transfer of momentum between the solar wind and the magnetosphere [Sibeck et al., 1999]. For this reason, the part of this thesis dedicated to the small-scale physical processes of the near-Earth environment focuses on magnetic reconnection.

### 1.4.1 Magnetic reconnection

Magnetic reconnection is likely to occur when two conductive plasma with non-parallel magnetic field interact with each other. The interacting field lines are merged resulting in a topological rearrangement of the system, often characterised by the conversion of magnetic energy into kinetic and thermal energy.

The term magnetic reconnection was first used by Dungey [1953] who showed that a breakdown of the frozen-in law could result in the rearrangement of the field lines connectivity and associated particle acceleration. From then on, the decades of studies dedicated to the comprehension of the process evidenced magnetic reconnection as a key factor of a wide range of phenomena. For instance, it plays a fundamental role in the formation of geomagnetic storms and auroras, in the relativistic jets emitted by active galactic nuclei or sawtooth oscillations observed in tokamak fusion plasmas (Yamada et al. [2010] and references therein). Magnetic reconnection is also believed to be the main actor at the origin of the formation of solar flares and CMEs that will constitute the main topic of the Chapter 3 [Shibata et al., 1995].

A schematic representation of magnetic reconnection is shown in the Figure 1.10. The incoming plasma flows are designated as the *inflows* while the flows of accelerated particles are denominated the *outflows*. When the frozen-in law is respected, the boundary between the two media is closed and no transfer of mass and momentum between the two plasmas is allowed. This is the case we described in the previous section when the plasma of the magnetosheath flows around the magnetopause.

With the violation of the frozen-in condition, the ions and the electrons decouple from the magnetic field in the proximity of the reconnection site. These regions of demagnetization are often called the *diffusion regions* and usually have a thickness in the order of the particles inertial length or thermal larmor radius, depending on the amplitude of the guide field. The *Ion Diffusion Region (IDR)* and the *Electron Diffusion Region (EDR)* are indicated in the Figure 1.10 with the grey and the blue rectangles respectively. The thick black lines of Figure 1.10 represent the field lines that are just being reconnected. As these lines separate the regions of different magnetic topology, they are called the *separatrices*. Because of the X shape they adopt, the point in the center of interest where they intersect is called the *X-point*. This is the point where the field lines are reconnected and from which they are convected with the outflow. In a 3D configuration, the counterpart of this reconnection site is called the *X-line*.

When all the physical parameters of the two interacting plasmas are equal but the orientation of their magnetic field, the reconnection is said to be *symmetric*. This well approximates what happens in the magnetotail as this will be described in the next subsection. Otherwise, reconnection is said to be *asymmetric*. This is especially what happens at the magnetopause where the dense and cold magnetosheath interacts with the hot, tenuous magnetosphere.

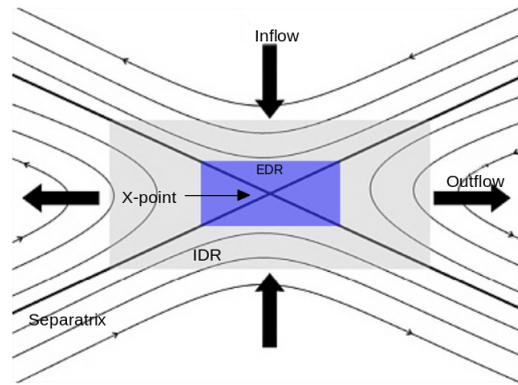


Figure 1.10: Schematic representation of magnetic reconnection, the black lines represent the magnetic field lines of the two media, the thick black lines are the separatrix, the black arrows indicate the plasma inflow and outflow, the grey rectangle represents the IDR and the blue rectangle represents the EDR.

### 1.4.2 Magnetic reconnection at the Earth magnetopause

Reconnection of the IMF with the geomagnetic field was first detailed by [Dungey \[1961\]](#) and represented in the Figure 1.11 in the case of a southward IMF.

The interplanetary field lines, represented in blue, and the geomagnetic field lines, represented in green, reconnect in the dayside of the magnetosphere at the reconnection site indicated by the grey rectangle (step ① of the Figure). It is worth noting that reconnection in this region is asymmetric. The reconnected field lines (represented in red), open on one side and attached to the Earth pole on the other side, are convected tailward<sup>5</sup> by the flow of solar wind and pile up in the magnetotail (step ② of the Figure). As the convected field lines point sunward in the northern hemisphere, and antisunward in the southern hemisphere, they reconnect in this region (step ③), interrupting the accumulation of flux in the process. On the nightward side of this symmetric reconnection site, which location is represented by the second grey rectangle, a bubble of plasma, known as a *plasmoid*, is expelled in the interplanetary medium. On the other side, the reconnected field lines are attached to the Earth and convected earthward carrying energetic accelerated particles which precipitation is at the origin of the formation of auroras (step ④). These closed field lines are then brought back to the dayside where they can reconnect again with the interplanetary field lines ensuring the continuity of the so-called *Dungey cycle* (step ⑤).

Following the description of [Dungey \[1961\]](#), when the IMF is northward and in a null dipole tilt condition, the interplanetary field lines and the closed geomagnetic field lines are parallel around the equatorial plane, indicating the absence of reconnection in this region. With the solar wind flowing around the magnetosphere, the interplanetary field lines are draped around the magnetopause and are likely to reconnect at high latitude where they are quasi anti-parallel to the geomagnetic field lines. The sunward convection of the newly reconnected field lines that appears in this case is opposed the flow of the solar wind.

In both cases, reconnection fundamentally affects the position and shape of the magnetopause by eroding the magnetosphere in a direction that depends on the IMF orientation, on the dayside equatorial plane when it is southward and at high-latitudes when it is northward [[Aubry et al., 1970](#)]. In the previous observational studies of the magnetopause, the effects of this erosion are considered through the study of the influence of the lone IMF  $B_z$  component that appears to be the component with the greatest impact on the magnetic topology of the magnetosphere. However, this consideration is reductive regarding the effect of the two other components of the IMF,  $B_x$  and  $B_y$ , on magnetic reconnection, which has been evidenced for long [[Gonzalez and Gonzalez, 1980](#); [Russell and Atkinson, 1973](#)]. Consequently, the influence of these two other components of the IMF is still unclear and open to further investigations. This open question will be one of the

<sup>5</sup>This convection process fills with plasma a layer between the magnetosheath and the lobes commonly known as the *plasma mantle*.

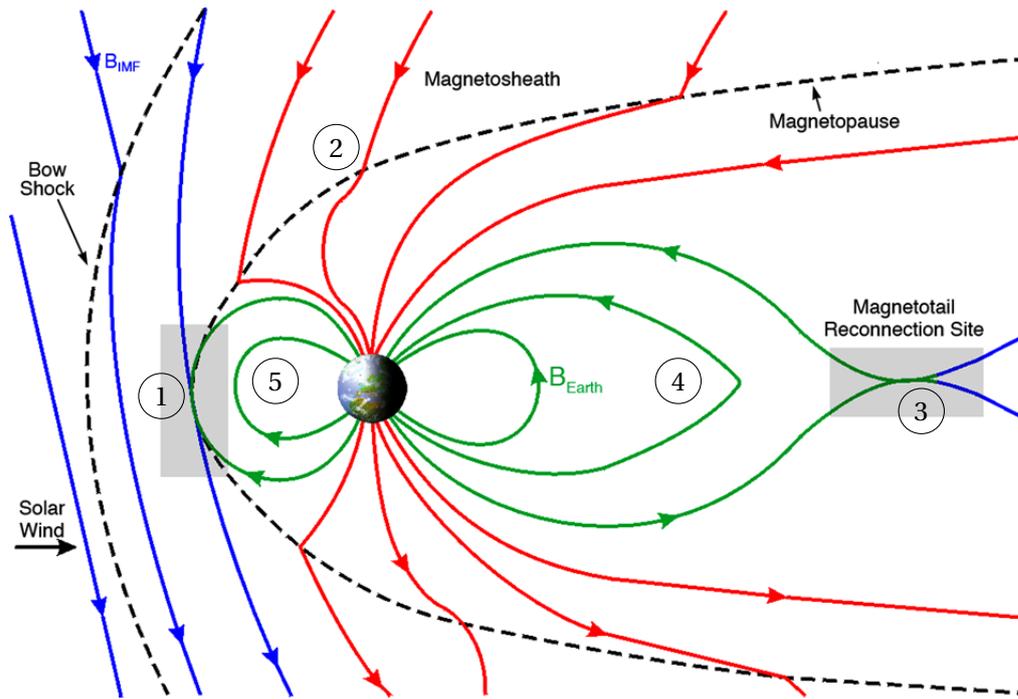


Figure 1.11: Schematic representation in the meridional plane of the dynamics of the magnetosphere when the IMF is southward. The blue lines represent the interplanetary field lines. The closed geomagnetic field lines are shown in green while the lines opened by reconnection are shown in red. The two grey boxes indicate the location of the dayside and the nightside reconnection sites and the black dashed line represent the bow shock and the magnetopause. The circled number depict the different steps of the Dungey cycle (see text) (adapted from Hughes [1995]).

main central topic of the chapter 5.

By reconfiguring the magnetic topology of the magnetosphere, reconnection modifies the nature of the near-cusp magnetopause. Indeed, the convection of the newly reconnected field lines creates boundaries that separate the plasma in the polar cusp from the magnetosheath called the *cusp external boundaries* [Lavraud et al., 2004b] by opposition with the so-called *cusp inner boundaries* that separate the cusp exterior from the magnetosphere. Without reconnection, the latter is the logical continuous extension between the dayside and the nightside magnetopauses. In the sense of reconnection, one of these boundaries actually corresponds to the separatrix of the geomagnetic and the interplanetary field lines. Consequently, the former appears as a more appropriate continuous extension of the magnetopause in the near-cusp region in the optics of reconnection that occurs whatever the orientation of the IMF might be. Although observed by various missions [Lavraud et al., 2004a; Zhou and Russell, 1997], the topology of this boundary for various solar wind conditions is still unclear and there are no existing clues on its actual indentation. We will come back on this still-open question in the Chapter 5.

The first in-situ evidence of magnetic reconnection was brought by Sonnerup and Cahill Jr. [1967] who observed non-zero normal magnetic field component at the crossing of the magnetopause by Explorer 12, indicating an interface between the magnetosheath and the geomagnetic fields that was not limited to the lone tangential discontinuity. This first observation was followed by the first evidence of accelerated plasma at the magnetopause found by Paschmann et al. [1979] using the ISEE satellites and by the evidence of reconnection in the magnetotail observed in IMP data by Hones Jr. et al. [1976] that were consistent with the predictions of Dungey. From then on, the evidences of reconnection occurring at the magnetopause multiplied with the accumulation of missions. In particular, we can cite Cluster, THEMIS and Double Star, some of the missions that will focus our attention in this thesis. Nowadays, the technological advances allow an in-situ measurement of the plasma properties with an even finer time resolution permitting a deeper investigation of the complexity of magnetic reconnection. For instance, the high resolution of the

measurements of the recently launched *MMS* allowed an observational insight on the *EDRs* for the first time.

For now, the most solid evidence we can collect happens when a spacecraft goes through a reconnection outflow during the crossing of the magnetopause<sup>6</sup>. Because of reconnection, the plasma of the outflow is accelerated, a spacecraft going through this region would then see a so-called *plasma jet* faster than the surrounding magnetosheath flow with a peaking component roughly corresponding to the component of the magnetic field that reverses during the crossing. This is especially the case for the jets represented by the green rectangles in the Figure 1.9. Here, the 5 jets we identified are oriented in the +Z direction indicating a spacecraft actually located north of the X-line, when the observed jets peak in both +Z and -Z directions, we usually talk about *reversal jets* that indicate a passage in the close proximity of the X-line by the spacecraft.

40 years of in-situ observation allowed an intensive study of the different aspects of magnetic reconnection that completed the theoretical and the numerical works dedicated to the subject. If these studies already confirmed a wide range of properties of magnetic reconnection, the clear influence of the conditions in the inflow regions [Cassak and Shay, 2007], the role of the different *IMF* components (Lavraud et al. [2005], Sonnerup [1974]) or its suppression by diamagnetic effects [Swisdak et al., 2003] just to mention a few, the secrets of this process are far from all being unlocked. Among the remaining unknowns, one can especially cite the conditions that initiate the process, the structure of the different diffusion regions or the parametric dependence of the X-line location.

The latter has been under debate since the very first premises of studies dedicated to magnetopause magnetic reconnection. The question to know if reconnection occurs where the magnetic field of both sides are anti-parallel (*anti-parallel reconnection* [Crooker, 1979]) or if only a component of the *IMF* actually reconnects (*component reconnection* [Sonnerup, 1974]) has been under debate for years. De facto, both scenarios have been observed in spacecraft data and the question is then not to know which one prevails but how does the actual location of reconnection line, that should probably be a combination of these two configurations, varies with changing solar wind and seasonal conditions. Using Polar observations, Trattner et al. [2007] inferred that reconnection occurs where the shear angle between the magnetosheath and the geomagnetic field lines is maximized and developed the so-called *maximum shear angle model*. Although the capacity of this model to predict the local orientation of the X-line has been indicated in an important number of observational studies (Cassak and Fuselier [2016] and references therein), one does still not know how to link this empirical predictions to the different parameters we believe to affect reconnection dynamics such as the density, the field amplitude or bulk velocity jumps across the magnetopause [Cassak and Shay, 2007]. The numerical and theoretical investigations of the X-line orientation also led to the development of numerous models among which we can cite the maximisation of the outflow speed by Swisdak and Drake [2007], the bisection between the magnetospheric and the magnetosheath fields by Aunai et al. [2016] or the maximisation of the current density of the magnetopause by Gonzalez and Mozer [1974]. Nevertheless, the comparison of all of these models, including the maximum shear angle model, to observational data has only been done through the investigation of the local orientation of the X-line [Souza et al., 2017] or on a small number of events [Trattner et al., 2012]. We then lack both of a global vision on how the reconnection sites extend over the whole magnetopause and of a comparison of these models with an important number of samples.

An interesting alternative we could bring to clarify the situation would be to exploit the accumulated data of spacecraft that observed the process, collect as much in-situ evidences of magnetic reconnection in the form of magnetopause plasma jets and perform a statistical analysis of their position relatively to the X-line for various solar wind and seasonal conditions. Collecting these jets in the data of various missions that have been crossing the magnetopause at various orbits and studying their position and velocity for different upstream conditions would result in a representation of the magnetopause plasma flow induced by reconnection which analysis would

<sup>6</sup>This outflow will also sometimes be called the *reconnection exhaust*.

give clue on the actual position of the X-line. Nevertheless, it implies the collection of an important number of magnetopause plasma jets in the data of different missions throughout their whole time period, which are in practice, much harder to locate than in the textbook case of Figure 1.9<sup>7</sup>. First because it requires the identification of magnetopause crossings which can already be an ambiguous task. Second because the in-situ signature of the exhaust can be obliterated by the presence of strong plasma fluctuations in the magnetosheath, this is particularly likely to happen at the flanks or when multiple crossings occur in a short time interval. Third because, depending on its trajectory, the apogee of a spacecraft can perfectly be located at the interface between the exhaust and one of the inflows. The measured in-situ signature will then be weaker necessarily and less distinguishable from the magnetosheath than what would have been measured if the crossing of the exhaust was complete.

From then on, elaborating an automated magnetopause plasma jets routine could constitute an important milestone in the statistical representation of the reconnection induced magnetopause plasma flow and the subsequent global investigation of the location of the X-line.

The first step of this improvement stands in the massive collection of magnetopause crossings as this is where reconnection is observed. Along with the possibility of investigating the large-scale influence of reconnection on the location and shape of the magnetopause from a statistical point of view, the interest of performing such massive detection is doubled and confirms all the potential such method would have. We will focus on this topic in the Chapter 4.

Concerning the jets detection properly speaking, and provided we have enough magnetopause crossings, machine learning algorithms appears again as good candidates in the frame of improving the automatic detection of magnetopause plasma jets. This will be the main focus of Chapter 6.

## 1.5 Summary

In this thesis, we apply different machine learning algorithms to provide a fast, automatic and reproducible identification of large-scale, medium-scale and small-scale events of interest in the in-situ data provided by spacecraft of multiple missions. In particular, we focus on the automatic detection of ICMEs at 1 AU, the classification of the three main near-Earth regions and the detection of magnetic reconnection plasma jets. In the three cases, the automatic detection methods allow us to rapidly generate reproducible catalogs of events with an important number of sample that can be used for additional statistical studies. Finally, we use the magnetopause crossings catalog generated with the region classifier to perform a statistical study of the position and shape of the magnetopause as a function of the upstream solar wind conditions. In particular, we benefit from the important number of samples detected by our method to investigate the influence of the IMF clock angle and to reconsider the issue of the near-cusp indentation of the magnetopause.

The outcome of the thesis is as follows, in Chapter 2, we present the principle of machine learning algorithms and describe the different algorithms that will be in use in the next chapters. In Chapter 3, we apply CNNs to the automatic detection of ICMEs. In chapter 4, we train a gradient boosting algorithm to distinguish the magnetosphere from the magnetosheath and the solar wind and use this method to provide a massive detection of the Earth magnetopause and bow shock crossings. Chapter 5 focuses on the exploitation of the obtained crossings catalog through the statistical study of the position and shape of the magnetopause as a function of the upstream solar wind conditions. Finally, we use the region classifier and the magnetopause crossings catalog as a basis for the elaboration of an automatic detection method of magnetic reconnection plasma jets in the Chapter 6.

---

<sup>7</sup>See chapter 6 and appendix B for additional observational examples.

## 1.6 Bibliography

- Aubry, M. P., Russell, C. T., and Kivelson, M. G.: Inward motion of the magnetopause before a sub-storm, *Journal of Geophysical Research* (1896-1977), 75, 7018–7031, <https://doi.org/10.1029/JA075i034p07018>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/JA075i034p07018>, 1970. 16
- Aunai, N., Hesse, M., Lavraud, B., Dargent, J., and Smets, R.: Orientation of the X-line in asymmetric magnetic reconnection, *Journal of Plasma Physics*, 82, 535820401, <https://doi.org/10.1017/S0022377816000647>, 2016. 18
- Baumjohann, W. and Treumann, R. A.: Basic space plasma physics, <https://doi.org/10.1142/p015>, 1996. 11
- Beatty, J. K.: Book Review: *Russian Planetary Exploration : History, Development, Legacy, and Prospects* / Springer/Praxis, 2007, *Sky & Telescope*, 113, 72, 2007. 2
- Boteler, D. H.: A 21st Century View of the March 1989 Magnetic Storm, *Space Weather*, 17, 1427–1441, <https://doi.org/10.1029/2019SW002278>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019SW002278>, 2019. 8
- Cahill, L. J. and Amazeen, P. G.: The boundary of the geomagnetic field, *Journal of Geophysical Research* (1896-1977), 68, 1835–1843, <https://doi.org/10.1029/JZ068i007p01835>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/JZ068i007p01835>, 1963. 12
- Camporeale, E.: The Challenge of Machine Learning in Space Weather: Nowcasting and Forecasting, *Space Weather*, 17, 1166–1207, <https://doi.org/10.1029/2018SW002061>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018SW002061>, 2019. 7
- Cassak, P. A. and Fuselier, S. A.: Reconnection at Earth's Dayside Magnetopause, vol. 427 of *Astrophysics and Space Science Library*, p. 213, [https://doi.org/10.1007/978-3-319-26432-5\\_6](https://doi.org/10.1007/978-3-319-26432-5_6), 2016. 18
- Cassak, P. A. and Shay, M. A.: Scaling of asymmetric magnetic reconnection: General theory and collisional simulations, *Physics of Plasmas*, 14, 102114, <https://doi.org/10.1063/1.2795630>, 2007. 18
- Chapman, S. and Ferraro, V. C. A.: The Electrical State of Solar Streams of Corpuscles, *Monthly Notices of the Royal Astronomical Society*, 89, 470–479, <https://doi.org/10.1093/mnras/89.5.470>, URL <https://doi.org/10.1093/mnras/89.5.470>, 1929. 7
- Chapman, S. and Ferraro, V. C. A.: A new theory of magnetic storms, *Terrestrial Magnetism and Atmospheric Electricity*, 36, 77–97, <https://doi.org/10.1029/TE036i002p00077>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/TE036i002p00077>, 1931. 3
- Chi, Y., Shen, C., Wang, Y., Xu, M., Ye, P., and Wang, S.: Statistical Study of the Interplanetary Coronal Mass Ejections from 1995 to 2015, *solphys*, 291, 2419–2439, <https://doi.org/10.1007/s11207-016-0971-5>, 2016. 6
- Crooker, N. U.: Dayside merging and cusp geometry, *Journal of Geophysical Research*, 84, 951–959, <https://doi.org/10.1029/JA084iA03p00951>, 1979. 18
- Du, A. M., Tsurutani, B. T., and Sun, W.: Anomalous geomagnetic storm of 21–22 January 2005: A storm main phase during northward IMFs, *Journal of Geophysical Research: Space Physics*, 113, <https://doi.org/10.1029/2008JA013284>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2008JA013284>, 2008. 8

- 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, 725–738, <https://doi.org/10.1080/14786440708521050>, URL <https://doi.org/10.1080/14786440708521050>, 1953. 15
- Dungey, J. W.: Interplanetary Magnetic Field and the Auroral Zones, *Physical Review Letters*, 6, 47–48, <https://doi.org/10.1103/PhysRevLett.6.47>, 1961. 16
- Echer, E., Alves, M., and Gonzalez, W.: A statistical study of magnetic cloud parameters and geoeffectiveness, *Journal of Atmospheric and Solar-Terrestrial Physics*, 67, 839 – 852, <https://doi.org/https://doi.org/10.1016/j.jastp.2005.02.010>, URL <http://www.sciencedirect.com/science/article/pii/S136468260500060X>, 2005. 8
- Fairfield, D. H.: Average and unusual locations of the Earth’s magnetopause and bow shock, *Journal of Geophysical Research*, 76, 6700, <https://doi.org/10.1029/JA076i028p06700>, 1971. 12
- Goldstein, H.: On the field configuration in magnetic clouds., in: *NASA Conference Publication*, vol. 228 of *NASA Conference Publication*, p. 0.731, 1983. 8
- Gonzalez, W. D. and Gonzalez, A. L. C.: Influence of the  $B_x$  component of the interplanetary magnetic field on magnetopause reconnection, *Geophysical Research Letters*, 7, 773–776, <https://doi.org/10.1029/GL007i010p00773>, 1980. 16
- Gonzalez, W. D. and Mozer, F. S.: A quantitative model for the potential resulting from reconnection with an arbitrary interplanetary magnetic field, *Journal of Geophysical Research*, 79, 4186, <https://doi.org/10.1029/JA079i028p04186>, 1974. 18
- Gosling, J. T., Pizzo, V., and Bame, S. J.: Anomalously low proton temperatures in the solar wind following interplanetary shock waves—evidence for magnetic bottles?, *Journal of Geophysical Research*, 78, 2001, <https://doi.org/10.1029/JA078i013p02001>, 1973. 8
- Hansen, R. T., Garcia, C. J., Groganard, R. J.-M., and Sheridan, K.: A Coronal Disturbance Observed Simultaneously with a White-Light Coronameter and the 80 MHz Culgoora Radioheliograph, 1971. 7
- Hasegawa, H.: Structure and Dynamics of the Magnetopause, 1, 71–119, <https://doi.org/10.5047/meep.2012.00102.0071>, 2012. 6, 12
- Hones Jr., E. W., Bame, S. J., and Asbridge, J. R.: Proton flow measurements in the magnetotail plasma sheet made with Imp 6, *Journal of Geophysical Research (1896-1977)*, 81, 227–234, <https://doi.org/10.1029/JA081i001p00227>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/JA081i001p00227>, 1976. 17
- Hoshi, Y., Hasegawa, H., Kitamura, N., Saito, Y., and Angelopoulos, V.: Seasonal and Solar Wind Control of the Reconnection Line Location on the Earth’s Dayside Magnetopause, *Journal of Geophysical Research (Space Physics)*, 123, 7498–7512, <https://doi.org/10.1029/2018JA025305>, 2018. 6
- Hughes, W. J.: The Magnetopause, Magnetotail, and Magnetic Reconnection, p. 227–287, Cambridge University Press, <https://doi.org/10.1017/9781139878296.010>, 1995. 17
- Jelínek, K., Němeček, Z., and Šafránková, J.: A new approach to magnetopause and bow shock modeling based on automated region identification, *Journal of Geophysical Research (Space Physics)*, 117, A05208, <https://doi.org/10.1029/2011JA017252>, 2012. 7, 12
- Kilpua, E., Koskinen, H. E. J., and Pulkkinen, T. I.: Coronal mass ejections and their sheath regions in interplanetary space, *Living Reviews in Solar Physics*, 14, 5, <https://doi.org/10.1007/s41116-017-0009-6>, 2017. 6, 8, 9

- King, J. H. and Papitashvili, N. E.: Solar wind spatial scales in and comparisons of hourly Wind and ACE plasma and magnetic field data, *Journal of Geophysical Research: Space Physics*, 110, <https://doi.org/10.1029/2004JA010649>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2004JA010649>, 2005. 2
- Kivelson, M. G. and Zu-Yin, P.: The Kelvin-Helmholtz instability on the magnetopause, *Planetary and Space Science*, 32, 1335 – 1341, [https://doi.org/https://doi.org/10.1016/0032-0633\(84\)90077-1](https://doi.org/https://doi.org/10.1016/0032-0633(84)90077-1), URL <http://www.sciencedirect.com/science/article/pii/0032063384900771>, 1984. 15
- Klein, L. W. and Burlaga, L. F.: Interplanetary magnetic clouds at 1 AU, Tech. rep., 1981. 8
- Kobel, E. and Flückiger, E. O.: A model of the steady state magnetic field in the magnetosheath, *Journal of Geophysical Research: Space Physics*, 99, 23 617–23 622, <https://doi.org/10.1029/94JA01778>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/94JA01778>, 1994. 3
- Lavraud, B., Fedorov, A., Budnik, E., Grigoriev, A., Cargill, P., Dunlop, M., Rème, H., Dandouras, I., and Balogh, A.: Cluster survey of the high-altitude cusp properties: a three-year statistical study, *Annales Geophysicae*, 22, 3009–3019, <https://doi.org/10.5194/angeo-22-3009-2004>, 2004a. 6, 17
- Lavraud, B., Phan, T., Dunlop, M., Taylor, M., Cargill, P., Bosqued, J., Dandouras, I., Rème, H., Sauvaud, J., Escoubet, C., Balogh, A., and Fazakerley, A.: The exterior cusp and its boundary with the magnetosheath: Cluster multi-event analysis, *Annales Geophysicae*, 22, 3039–3054, <https://doi.org/10.5194/angeo-22-3039-2004>, 2004b. 17
- Lavraud, B., Thomsen, M. F., Taylor, M. G. G. T., Wang, Y. L., Phan, T. D., Schwartz, S. J., Elphic, R. C., Fazakerley, A., Rème, H., and Balogh, A.: Characteristics of the magnetosheath electron boundary layer under northward interplanetary magnetic field: Implications for high-latitude reconnection, *Journal of Geophysical Research (Space Physics)*, 110, A06209, <https://doi.org/10.1029/2004JA010808>, 2005. 18
- Lepping, R. P., Wu, C.-C., and Berdichevsky, D. B.: Automatic identification of magnetic clouds and cloud-like regions at 1 AU: occurrence rate and other properties, *Annales Geophysicae*, 23, 2687–2704, <https://doi.org/10.5194/angeo-23-2687-2005>, 2005. 7
- Lepping, R. P., Berdichevsky, D. B., Wu, C. C., Szabo, A., Narock, T., Mariani, F., Lazarus, A. J., and Quivers, A. J.: A summary of WIND magnetic clouds for years 1995-2003: model-fitted parameters, associated errors and classifications, *Annales Geophysicae*, 24, 215–245, <https://doi.org/10.5194/angeo-24-215-2006>, 2006. 9
- Lewis, S. A. and Fuselier, W. S.: Properties of Near-Earth Magnetic Reconnection from In-Situ Observations, pp. 95–121, <https://doi.org/10.1007/s11214-011-9820-x>, 2011. 6
- Lin, R. L., Zhang, X. X., Liu, S. Q., Wang, Y. L., and Gong, J. C.: A three-dimensional asymmetric magnetopause model, *Journal of Geophysical Research (Space Physics)*, 115, A04207, <https://doi.org/10.1029/2009JA014235>, 2010. 12
- Liu, Z., Lu, J. Y., Wang, C., Kabin, K., Zhao, J. S., Wang, M., Han, J. P., Wang, J. Y., and Zhao, M. X.: Journal of Geophysical Research : Space Physics A three-dimensional high Mach number asymmetric magnetopause model from global MHD simulation, pp. 5645–5666, <https://doi.org/10.1002/2014JA020961>.Received, 2015. 12
- Moissard, C., Fontaine, D., and Savoini, P.: A Study of Fluctuations in Magnetic Cloud-Driven Sheaths, *Journal of Geophysical Research (Space Physics)*, 124, 8208–8226, <https://doi.org/10.1029/2019JA026952>, 2019. 8

- Nieves-Chinchilla, T., Vourlidas, A., Raymond, J. C., Linton, M. G., Al-haddad, N., Savani, N. P., Szabo, A., and Hidalgo, M. A.: Understanding the Internal Magnetic Field Configurations of ICMEs Using More than 20 Years of Wind Observations, *solphys*, 293, 25, <https://doi.org/10.1007/s11207-018-1247-z>, 2018. 9
- Němeček, Z., Šafránková, J., and Šimůnek, J.: An Examination of the Magnetopause Position and Shape Based Upon New Observations, chap. 8, pp. 135–151, American Geophysical Union (AGU), <https://doi.org/10.1002/9781119509592.ch8>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/9781119509592.ch8>, 2020. 6, 12
- Oliveira, D. and Samsonov, A.: Geoeffectiveness of interplanetary shocks controlled by impact angles: A review, *Advances in Space Research*, 61, 1 – 44, <https://doi.org/https://doi.org/10.1016/j.asr.2017.10.006>, URL <http://www.sciencedirect.com/science/article/pii/S0273117717307275>, 2018. 9
- Parker, E. N.: Dynamics of the Interplanetary Gas and Magnetic Fields., *The Astrophysical Journal*, 128, 664, <https://doi.org/10.1086/146579>, 1958. 2
- Parker, E. N.: Interplanetary dynamical processes., 1963. 2
- Paschmann, G., Papamastorakis, I., Sckopke, N., Haerendel, G., Sonnerup, B. U. O., Bame, S. J., Asbridge, J. R., Gosling, J. T., Russel, C. T., and Elphic, R. C.: Plasma acceleration at the earth's magnetopause - Evidence for reconnection, *Nature*, 282, 243–246, <https://doi.org/10.1038/282243a0>, 1979. 17
- Paschmann, G., Haaland, S. E., Phan, T. D., Sonnerup, B. U. Ö., Burch, J. L., Torbert, R. B., Gershman, D. J., Dorelli, J. C., Giles, B. L., Pollock, C., Saito, Y., Lavraud, B., Russell, C. T., Strangeway, R. J., Baumjohann, W., and Fuselier, S. A.: Large-Scale Survey of the Structure of the Dayside Magnetopause by MMS, *Journal of Geophysical Research (Space Physics)*, 123, 2018–2033, <https://doi.org/10.1002/2017JA025121>, 2018. 6, 12
- Richardson, I. G.: Solar wind stream interaction regions throughout the heliosphere, *Living Reviews in Solar Physics*, 15, 1–95, <https://doi.org/10.1007/s41116-017-0011-z>, URL <https://doi.org/10.1007/s41116-017-0011-z>, 2018. 6, 9
- Richardson, I. G. and Cane, H. V.: Near-Earth Interplanetary Coronal Mass Ejections During Solar Cycle 23 (1996 - 2009): Catalog and Summary of Properties, *solphys*, 264, 189–237, <https://doi.org/10.1007/s11207-010-9568-6>, 2010. 9
- Russell, C. T. and Atkinson, G.: Comments on a paper by J. P. Heppner, ‘Polar cap electric field distributions related to interplanetary magnetic field direction’, *Journal of Geophysical Research*, 78, 4001–4002, <https://doi.org/10.1029/JA078i019p04001>, 1973. 16
- Shibata, K., Masuda, S., Shimojo, M., Hara, H., Yokoyama, T., Tsuneta, S., Kosugi, T., and Ogawara, Y.: Hot-Plasma Ejections Associated with Compact-Loop Solar Flares, *The Astrophysical Journal*, 451, L83, <https://doi.org/10.1086/309688>, 1995. 15
- Shinde, A. A. and Russell, C. T.: What Defines an Interplanetary Coronal Mass Ejection?, in: AGU Fall Meeting Abstracts, vol. 2003, pp. SH21B–0133, 2003. 8
- Shue, J. H., Chao, J. K., Fu, H. C., Russell, C. T., Song, P., Khurana, K. K., and Singer, H. J.: A new functional form to study the solar wind control of the magnetopause size and shape, *Journal of Geophysical Research*, 102, 9497–9512, <https://doi.org/10.1029/97JA00196>, 1997. 12
- Sibeck, D. G., Paschmann, G., Treumann, R. A., Fuselier, S. A., Lennartsson, W., Lockwood, M., Lundin, R., Ogilvie, K. W., Onsager, T. G., Phan, T. D., Roth, M., Scholer, M., Sckopke, N., Stasiewicz, K., and Yamauchi, M.: Chapter 5-Plasma Transfer Processes at the Magnetopause, *Scientific Studies of Reading*, 88, 207–283, <https://doi.org/10.1023/A:1005255801425>, 1999. 15

- Sonnerup, B. U. Ö.: Magnetopause reconnection rate, *Journal of Geophysical Research*, 79, 1546–1549, <https://doi.org/10.1029/JA079i010p01546>, 1974. 18
- Sonnerup, B. U. and Cahill Jr., L. J.: Magnetopause structure and attitude from Explorer 12 observations, *Journal of Geophysical Research (1896-1977)*, 72, 171–183, <https://doi.org/10.1029/JZ072i001p00171>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/JZ072i001p00171>, 1967. 17
- Sotirelis, T. and Meng, C.-I.: Magnetopause from pressure balance, *Journal of Geophysical Research*, 104, 6889–6898, <https://doi.org/10.1029/1998JA900119>, 1999. 12
- Souza, V. M., Gonzalez, W. D., Sibeck, D. G., Koga, D., Walsh, B. M., and Mendes, O.: Comparative study of three reconnection X line models at the Earth's dayside magnetopause using in situ observations, *Journal of Geophysical Research (Space Physics)*, 122, 4228–4250, <https://doi.org/10.1002/2016JA023790>, 2017. 18
- Spreiter, J. R. and Briggs, B. R.: Theoretical determination of the form of the boundary of the solar corpuscular stream produced by interaction with the magnetic dipole field of the Earth, *Journal of Geophysical Research (1896-1977)*, 67, 37–51, <https://doi.org/10.1029/JZ067i001p00037>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/JZ067i001p00037>, 1962. 12
- Swisdak, M. and Drake, J. F.: Orientation of the reconnection X-line, *Geophysical research Letters*, 34, L11106, <https://doi.org/10.1029/2007GL029815>, 2007. 18
- Swisdak, M., Rogers, B. N., Drake, J. F., and Shay, M. A.: Diamagnetic suppression of component magnetic reconnection at the magnetopause, *Journal of Geophysical Research (Space Physics)*, 108, 1218, <https://doi.org/10.1029/2002JA009726>, 2003. 18
- Tousey, R.: The solar corona., in: *Space Research Conference*, vol. 2, pp. 713–730, 1973. 7
- Trattner, K. J., Mulcock, J. S., Petrinec, S. M., and Fuselier, S. A.: Probing the boundary between antiparallel and component reconnection during southward interplanetary magnetic field conditions, *Journal of Geophysical Research (Space Physics)*, 112, A08210, <https://doi.org/10.1029/2007JA012270>, 2007. 18
- Trattner, K. J., Petrinec, S. M., Fuselier, S. A., and Phan, T. D.: The location of reconnection at the magnetopause: Testing the maximum magnetic shear model with THEMIS observations, *Journal of Geophysical Research (Space Physics)*, 117, A01201, <https://doi.org/10.1029/2011JA016959>, 2012. 18
- Turc, L., Fontaine, D., Savoini, P., and Kilpua, E. K. J.: A model of the magnetosheath magnetic field during magnetic clouds, *Annales Geophysicae*, 32, 157–173, <https://doi.org/10.5194/angeo-32-157-2014>, 2014. 9
- Wang, C., Du, D., and Richardson, J.: Characteristics of the interplanetary coronal mass ejections in the heliosphere between 0.3 and 5.4 AU, *Journal of Geophysical Research*, 110, <https://doi.org/10.1029/2005JA011198>, 2005. 8
- Webber, W. R., McDonald, F. B., Lockwood, J. A., and Heikkila, B.: The effect of the July 14, 2000 “Bastille Day” solar flare event on >70 MeV galactic cosmic rays observed at V1 and V2 in the distant heliosphere, *Geophysical Research Letters*, 29, 15–1–15–3, <https://doi.org/10.1029/2002GL014729>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2002GL014729>, 2002. 8
- Wilson, R. M.: Geomagnetic response to magnetic clouds, *Planetary and Space Science*, 35, 329 – 335, [https://doi.org/https://doi.org/10.1016/0032-0633\(87\)90159-0](https://doi.org/https://doi.org/10.1016/0032-0633(87)90159-0), URL <http://www.sciencedirect.com/science/article/pii/0032063387901590>, 1987. 8

- Yamada, M., Kulsrud, R., and Ji, H.: Magnetic reconnection, *Rev. Mod. Phys.*, 82, 603–664, <https://doi.org/10.1103/RevModPhys.82.603>, URL <https://link.aps.org/doi/10.1103/RevModPhys.82.603>, 2010. 15
- Yermolaev, Y. I., Nikolaeva, N. S., Lodkina, I. G., and Yermolaev, M. Y.: Geoeffectiveness and efficiency of CIR, sheath, and ICME in generation of magnetic storms, *Journal of Geophysical Research: Space Physics*, 117, <https://doi.org/10.1029/2011JA017139>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2011JA017139>, 2012. 8
- Zhang, H., Fu, S., Pu, Z., Lu, J., Zhong, J., Zhu, C., Wan, W., and Liu, L.: Statistics on the Magnetosheath Properties Related to Magnetopause Magnetic Reconnection, *The Astrophysical Journal*, 880, 122, <https://doi.org/10.3847/1538-4357/ab290e>, URL <http://dx.doi.org/10.3847/1538-4357/ab290e>, 2019. 6
- Zhou, X.-W. and Russell, C. T.: The location of the high-latitude polar cusp and the shape of the surrounding magnetopause, *Journal of Geophysical Research: Space Physics*, 102, 105–110, <https://doi.org/10.1029/96JA02702>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/96JA02702>, 1997. 17
- Zurbuchen, T. H. and Richardson, I. G.: In-Situ Solar Wind and Magnetic Field Signatures of Interplanetary Coronal Mass Ejections, p. 31, [https://doi.org/10.1007/978-0-387-45088-9\\_3](https://doi.org/10.1007/978-0-387-45088-9_3), 2006. 8



# Chapter 2

## Machine learning as an automatic selection tool

*What we want is a machine that can learn from experience.*

Alan Turing

### Contents

---

<b>2.1 Introduction</b> . . . . .	<b>28</b>
<b>2.2 Different types of supervised classification</b> . . . . .	<b>29</b>
2.2.1 Optimization of loss functions: Gradient Descent . . . . .	30
2.2.2 Logistic regression . . . . .	30
2.2.3 Decision trees . . . . .	32
2.2.4 Gradient Boosting . . . . .	33
2.2.5 Artificial Neural Network (ANN) . . . . .	34
<b>2.3 Evaluating the performances of an algorithm</b> . . . . .	<b>36</b>
2.3.1 Calibrating the probabilistic output . . . . .	36
2.3.2 Metrics . . . . .	37
<b>2.4 Implementing Machine Learning</b> . . . . .	<b>39</b>
<b>2.5 Machine Learning in space physics</b> . . . . .	<b>39</b>
<b>2.6 Bibliography</b> . . . . .	<b>40</b>

---

## 2.1 Introduction

Machine learning designates a specific type of algorithms that have the specificity to learn from a given dataset to perform a certain task by exhibiting trends and patterns that are generalizable to unseen data. This concept is not recent in the field of computer science and the first uses of the term came with the first theoretical principles of such algorithms that appeared no later than the middle of the 20<sup>th</sup> century [Berkson, 1944; Fisher, 1936; Samuel, 1959].

With the recent advances in the field of artificial intelligence, with the emergence of the requisite technologies and for the potential they have to deal with already large and ever growing bases of reliable data accumulated for decades, the use of such methods is flourishing in various fields for a wide range of tasks from face recognition in social networks to weather prediction. Their contribution to science also does not have to be neglected and such algorithms are already used when it comes to disease diagnosis [Sajda, 2006] or galaxy classification [Nolte et al., 2019] to name just two examples.

The diversity of the tasks these algorithms are best suited for can be grouped in two main purposes: regression and classification.

The former should sound familiar to any reader who already attempted to fit a model to the data and a schematic representation of such a problem is shown on the left panel of Figure 2.1.

Given a dataset  $(X_{i,j})_{n \times m}$  made of  $n$  observations of  $m$  different features, represented in the left panel of Figure 2.1 by the blue points, (let us say the solar wind magnetic field, velocity and density), the goal here stands in fitting a function  $h$  called *hypothesis* that best describes the evolution of a given label  $(Y_i)_n$  (the position of the nose of the Earth bow shock for example), expected to be continuous and also accessible for each of the  $n$  samples of data.

This best description aspect is reached by finding the parameters of the function  $h$  that minimize the errors made by the hypothesis in comparison to the label, the so-called *loss function*. The expression for this function are numerous, the most frequently used of them being the **Mean Square Error (MSE)** defined as follows:

$$\text{MSE}(X) = \frac{1}{n} \sum_{i=1}^n (h(X_i) - Y_i)^2 \quad (2.1)$$

Where  $X_i$  is a  $m$  dimensional vector that represents the  $i_{th}$  row of  $X$ . The fitted hypothesis function is represented in the left panel of Figure 2.1 by the red dashed line.

The schematic representation of a classification problem is shown in the middle panel of Figure 2.1. In this case, the label  $Y$  has discrete values that represent the number of elements (or *classes*) we wish the dataset to be classified into (represented by the crosses and the points in Figure 2.1). The objective here is to fit a hypothesis function that best separates the elements of the two classes in the features space as shown with the dashed red line in Figure 2.1. The output of the algorithm can, depending on the requirements of the user, either be discrete, giving the predicted class of a given point, either continuous, between 0 and 1, and can thus be interpreted as the probability a given sample of data actually belongs to a given class.

For the two first panels of Figure 2.1, the algorithm is aware of the label it is supposed to predict for each observation of the dataset and has been trained iteratively to minimize the loss function by comparing predictions to labels. This kind of training is defined as *supervised learning*.

Another option we have in the fitting or *training* phase of the algorithms stands in not attributing any label to the data and let the algorithm find trends within the dataset by itself. This technique is called *unsupervised learning* and is schematized in the right panel of Figure 2.1. Without supposition on the class each datapoint belongs to, the idea here is to find *clusters* of data points in the feature space and use this *clustering* process to define the classes used for classification purpose.

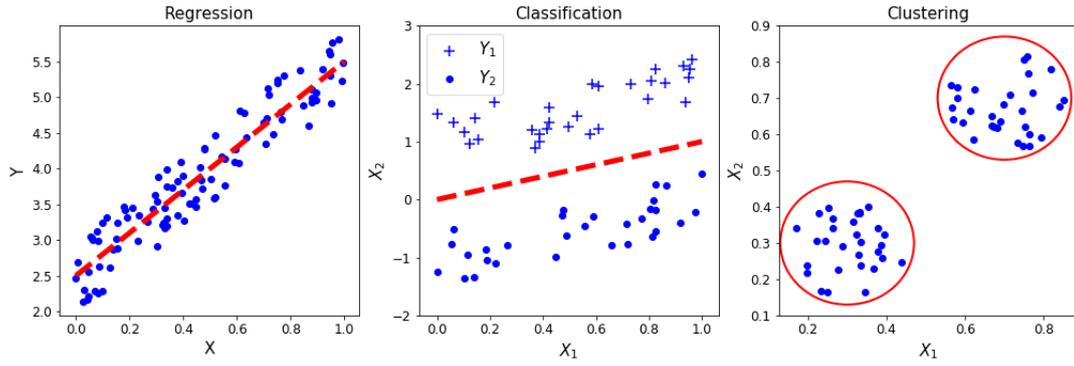


Figure 2.1: Schematic representation of a regression problem (*left*), a classification problem (*middle*) and a clustering problem (*right*). The blue markers represent the dataset points, the red figures represent the fitted hypothesis functions (*left*), the boundary between the two classes (*middle*) or the fitted clusters (*right*)

In the frame of our work, we intend to use machine learning algorithms to rapidly detect in-situ events and provide catalogs of them that would be consistent with what human experts would have manually labeled, but in a much faster and reproducible way. For this purpose, supervised learning algorithms used for classification tasks then appear as the most appropriate method. We will focus on the way they are trained and how they make predictions, in the next section.

Section 3 will present the metrics that can be used to give an insight on how the prediction of such algorithms can be trusted.

Section 4 will finally present how machine learning is currently used in space physics.

## 2.2 Different types of supervised classification

Having presented the main families of machine learning algorithms, we will here focus on 4 of the methods we have use in this thesis:

- Logistic regression
- Decision trees
- Gradient Boosting
- Artificial Neural Networks ([Artificial Neural Network \(ANN\)](#))

In the whole section, we will consider  $n$  observations of  $m$  different features  $(X_{i,j})_{n \times m}$  and the label  $(Y_i)_n$  associated for each observation.

Having trained an algorithm, one has to prove its capacity to generalize to unseen data. It is then necessary, when labeling a dataset, to keep a significant part of the data that will not intervene in the training phase but that will be used for the evaluation of the algorithm. This part of the dataset will be referred as the *test set* and usually represents 1/3 of the total dataset. The other 2/3 of the dataset are used for the training phase of the algorithm and will be referred as the *training set*.

For a given dataset, the performances of a trained algorithm can depend on how the training set and the test set are defined. This can easily lead to an algorithm that sticks too well to the training data without being able to generalize what it has learnt another set of data. Such algorithms are said to be *overfitting*. A typical solution we have to ensure this independence and avoid overfitting the data consists in removing a small subset of the training set and using this so-called *validation set* to test an algorithm trained with the remaining part of the training set. If an algorithm is independent from the separation made between the training and the validation set, its performances should be similar whatever the separation is. This independence can thus be verified by, in a so-called *cross-validation* process, repeating the operation for different separations between the training and the validation set and checking if the obtained performances are similar.

### 2.2.1 Optimization of loss functions: Gradient Descent

Each hypothesis function  $h$  comes with a set of parameters  $(\Theta_i)$  one can modify in order to make it fit to the training data and to the label. These parameters can for example be the coefficients of a polynome in the case of a polynomial fit.

The training phase then consists in finding the best values of  $(\Theta_i)$  in the parameter space that minimise the Loss function  $J(\Theta, X)$ .

Gradient Descent is one of the most useful algorithms used to perform this minimization task. The principle of the algorithm is then to come as close as possible to the Loss function minimum by taking iterative steps in the parameter space following the negative gradients of the loss functions.

A schematic representation of Gradient Descent is shown in Figure 2.2

Starting with initial conditions  $(\Theta_i)_0$ , each step of the algorithm then consists in updating the values of each parameters  $(\Theta_i)_l$  as follows:

$$\Theta_{i,l} = \Theta_{i,l-1} - \alpha \frac{\partial J(\Theta, X)}{\partial \Theta_i} \quad (2.2)$$

Where  $\Theta_{i,l}$  represents the  $i^{th}$  component of  $(\Theta_i)$  after the  $l^{th}$  iteration. The parameter  $\alpha$  is called the *learning rate* and describes how wide the steps are taken in the parameter space. A small learning rate will require more steps until convergence while large learning rates are the exposed to the risk of skipping the minimum.

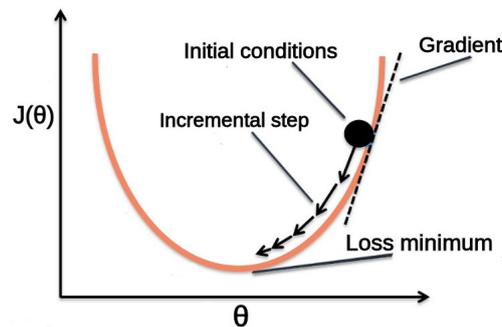


Figure 2.2: Schematic representation of Gradient Descent (Adapted from Lanham [2018])

Gradient descent is particularly efficient when the loss function is convex but is limited when the minimum is not unique as it could easily get stuck around a local minimum. Additionally, the consideration of the whole training set at each iteration makes it computationally heavier than an algorithm that consider small subsets of the training set at each iteration.

Starting from the principle we presented, the existing strategies to overcome these difficulties are numerous from the use of reduced portions of the dataset to compute the gradients at each iteration [Kiefer and Wolfowitz, 1952; Li et al., 2014], the acceleration of the process through the introduction of momentum [Botev et al., 2017] or algorithms with adaptative learning rates [Kingma and Ba, 2014].

### 2.2.2 Logistic regression

Logistic regression [Berkson, 1944] consists in defining the boundary between two classes (associated to a label being either equal to 0 or 1) as a hyperplane in the feature space. This concept is especially illustrated by the dashed red line of the middle panel of Figure 2.1.

Given a set of  $m + 1$  parameters  $(\Theta_i)$ , a given data point  $X_i$  will then belong to one or another class if it fall on one or the other side of the hyperplane defined by the vector  $(\Theta_i)$ . That is to say, if:

$$S(X_i) = \Theta_0 + \Theta_1 \times X_{i,1} + \Theta_2 \times X_{i,2} + \dots + \Theta_m \times X_{i,m} > 0 \quad (2.3)$$

This condition can be expressed as a probability by the application of the sigmoid function defined as:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (2.4)$$

With a resulting hypothesis function being equal to:

$$h_{\Theta}(X_i) = \frac{1}{1 + e^{-S(X_i)}} \quad (2.5)$$

The evolution of  $h_{\Theta}$  for varying values of  $S(X_i)$  is shown in Figure 2.3.

With the application of the sigmoid function, the condition 2.3 becomes  $h_{\Theta}(X_i) > 0.5$ , which indicates this hypothesis function can easily be understood as the probability a sample of the dataset has to belong to a given class.

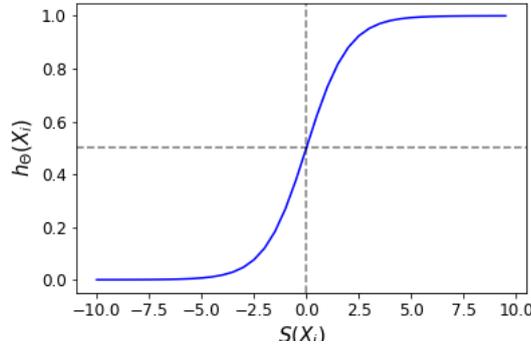


Figure 2.3: Evolution of the hypothesis function  $h_{\Theta}$  as a function of the different possible outputs of  $S(X_i)$

With this expression of the hypothesis, using the **MSE** as a loss function to minimize would bring to a non-convex problem with the presence of an important number of local minima resulting in an enhanced difficulty to find the global minimum. To cope with it, we define the logistic loss by:

$$J(\Theta) = -\frac{1}{n} \sum_{i=1}^n Y_i \log(h_{\Theta}(X_i)) + (1 - Y_i) \log(1 - (h_{\Theta}(X_i))) \quad (2.6)$$

Such a function has the advantage of attributing a low contribution to the extremal values of the hypothesis while giving a higher attention on the data points for which  $h_{\Theta}$  is close to 0.5. This is the function that will be minimized during the training phase of a logistic regression.

Logistic regression is also adaptable to classification problem where the label contains more than two classes. In this case, a common approach stands in training one logistic regression for each class and to compare the predictions made by these algorithms altogether. In this approach, called "One versus All", the final prediction for a given data point will then be the class for which the specifically trained algorithm will output the highest probability. Another possible approach stands in the generalization of the sigmoid function through the consideration of the so-called softmax function and the associated softmax regression [Ren et al., 2017].

Because of its simple design, a logistic regression algorithm is the one of the easiest algorithm to interpret as an insight on the found parameter,  $\Theta$ , directly gives the importance each feature has on the selection process of the algorithm, in addition to providing a linear relationship to discriminate one class from another. Nevertheless, its efficiency supposes a perfect linear separation between the different features of the dataset, which is far from being the case in the majority of machine learning problems.

### 2.2.3 Decision trees

Decision trees [Fisher, 1936] are a specific type of algorithm which final decision, the so-called *leaves* are reached through the application of a hierarchical sequence of thresholds-based tests on the different features of the dataset, the so-called *nodes*. Such techniques can be used for both and classification and regression purposes.

A schematic representation of such trees is shown in Figure 2.4. Here, the definition of the leaves and nodes have been set arbitrarily following the statistical a priori we have on the magnetosheath, the magnetosphere and the solar wind.

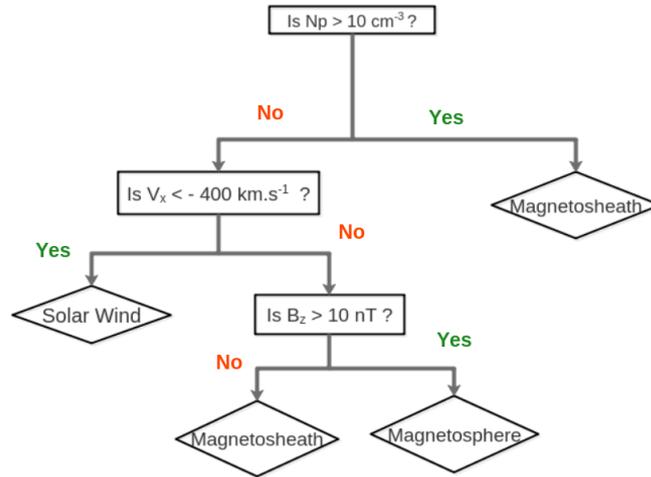


Figure 2.4: Schematic representation of a decision tree, each rectangle is a node of the tree, each diamond is a leaf

The machine learning adaptation of these trees, the so-called **Classification and Regression Trees (CART)** [Breiman et al., 1984], generalizes this sequential application of arbitrary threshold-based tests by defining the set of nodes, splits within the dataset and leaves that best suits the training set.

To do so, we start with the training set  $X_{i,j}$  and find the feature  $j$  and the observation  $\theta_j$  of this feature that minimize the defined cost-function:

$$J(\theta_j) = \frac{n_{left}}{n_{total}} G_{left} + \frac{n_{right}}{n_{total}} G_{right} \quad (2.7)$$

Where *left* and *right* denote the two dataset subsets obtained with the split,  $n_{total}$  represents the total number of samples in the node and  $n_{left}$  (resp.  $n_{right}$ ) represent the number of samples in the left (resp. right) split data subset.

G represents the so-called *Gini* index defined by:

$$G = 1 - \sum_{k=i}^{N_{classes}} p_k \quad (2.8)$$

Where  $p_k$  is the proportion of elements belonging to the class  $k$  in the considered data subset. This index measures the *impurity* and a null value of the Gini indicates a subset of data in which all the elements belong to the same class.

This splitting process is repeated for the newly-created nodes until one of these stopping conditions is reached:

1. All of the elements of the subsets created after a split belong to the same class, this is the natural stopping condition as no further split is then needed.
2. The number of nodes between the initial and the current one reaches a value called *maximum depth*. All the nodes created that reach this maximum depth become leaves.

3. The amount of data to split is below the minimum number of samples required to split an internal node. The concerned node then becomes a leaf.

For a given unknown data point, the predicted class will then correspond to the class that is in majority in the leaf of the tree this data point falls in. This output can also be expressed from a probabilistic point of view. The probability will then correspond to the proportion of this majority in the leaf in which the data point falls in.

The maximum depth and the minimal number of samples required to split are two hyper-parameters an external user can adjust to tune the decision tree. High values of these parameters allow a finer insight on the dataset but increase the risk the model will have to overfit the training data. Low values help reducing the training time but expose the model to the risk of not having seen enough training samples to exhibit any particular trend in the data and thus to *underfit* the data.

The concept of decision trees used for regression purpose is pretty similar with the noticeable differences that the *Gini* will be in this case substituted by a cost-function usually used in Regression such as the Root Mean Square Error and that the output of each leaf will here correspond to the average of the data points that fall into that leaf.

Because of the simplicity of their concept, decision trees are among the simplest machine learning algorithm to interpret. They also offer the advantage of requiring few preprocessing of the data and can even deal with missing features measurements. Nevertheless, these algorithms are very sensitive to small changes in the dataset and are thus very likely to overfit the training data. A common way to overcome these drawbacks then stands in considering an ensemble of several trees and considering a global prediction made of each of their individual guess. This is for example the principle of Random Forests [Breiman, 2001].

## 2.2.4 Gradient Boosting

Another approach we can have to overcome the disadvantages of decision trees we just mentioned stands in considering a sequential ensemble of decision trees and having each tree to correct the errors made by the previous one. This is the principle of Gradient Boosting algorithms [Friedman, 2001] that are among the most popular machine learning algorithms for their ability to rapidly deal with complex problems, that include missing data or an unbalanced dataset [Brown and Mues, 2012].

The principle of the construction of the algorithm is as follows:

1. For each element of the training set, define a preliminary score  $F_0(X_i)$  as the log of the odds of each element of the training set and define the associated preliminary probability  $P_0(X_i) = \sigma(F_0(X_i))$  where  $\sigma$  is the sigmoid function defined in the equation 2.4. This value  $F_0$  is uniform whatever the input is.
2. Fit a decision tree to the probabilistic residual error made by this preliminary prediction  $Y_i - P_0(X_i)$ , define the score  $h_1(\text{leaf}) = \frac{\sum_{i \in \text{leaf}} Y_i - P_0(X_i)}{P_0(X_i)(1 - P_0(X_i))}$  and predict the value of  $h_1$  for each element of the training set.
3. Define the updated score  $F_1 = F_0 + \alpha h_1$  where  $\alpha$  is defined as the *learning rate* of the algorithm and allow the reduction of the importance of the prediction of each trees in order to avoid an eventual overfit.
4. Compute the updated probabilities  $P_1(X_i) = \sigma(F_1(X_i))$ .
5. Repeat the operation for the number of trees used for the algorithm.

After the fitting phase, running instances of each of the trees that compose the algorithm then directly give the predicted score  $F$  and the associated probability of an unknown datapoint.

Even if the idea of fitting regression trees for classification purposes may seem counter intuitive, this approach offers an enhanced opportunity while considering a given classification problem and another example of this concept will be detailed in the next chapter.

In addition to the hyper-parameters that are used to fit the Decision Trees, the number of these trees and the learning rate  $\alpha$  are the two main parameters an external user can adjust to tune the decision tree. Decreasing the learning rate implies less importance granted to the updates provided by each tree. Increasing it exposes to the risk of taking into account the overfit of each tree and then ending up with an algorithm unable to generalize what it has learnt on the data. On the contrary, increasing the number of trees increases the chance of overfit, decreasing it increases the chance of underfit.

The same principle is followed when Gradient boosting is used for regression purposes, in this case, the probability and the score being replaced by the raw predictions made by the algorithm.

Despite of their efficiency, the implication of several trees at the same time and the increased number of hyper parameters make these algorithms harder to tune and much less interpretable than the Decision Trees or the Logistic Regression.

### 2.2.5 Artificial Neural Network (ANN)

Neural networks [Samuel, 1959] consist in a set of fully interconnected nodes, called *neurons* or simply *units*, arranged in a certain number of layers so that the output of each neuron of a layer to each neuron of the next one.

These algorithms, firstly conceptualized by analogy with the way the human brain processes the information, gained in popularity with the increasing amount of data, the technological advances that allowed a reduction of their training time, the utilisation of smarter optimization algorithms and thanks to their ability to successfully deal with massive datasets.

A schematic representation of a neural network is shown in Figure 2.5. For a given datapoint, the inputs of the first layer correspond to the measurement of each of the features for this point. These inputs are processed by the intermediate layers (unique in the case of Figure 2.5) called the *hidden layers* and result in a final output that can either be the probability of a class in a classification problem or the prediction of the quantity estimated in a regression problem.

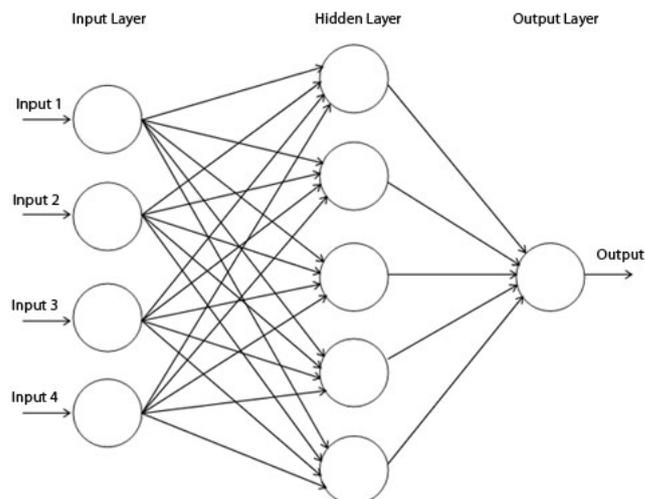


Figure 2.5: Schematic representation of a neural network (taken from <https://www.nicolamanzini.com/single-hidden-layer-neural-network/>)

For a given layer, each neuron independently processes the output of the previous layer and turns it into an output through the application of a so-called *activation function* that will be processed by the neurons of the following layer.

The process done by each neuron is of various nature depending on the nature of the neural network we intend to use. For the simplest one, the so-called *Multilayer Perceptrons* [Murtagh, 1991], this process consists in the weighted sum of the different neuron inputs. For more sophisticated networks such as CNN [Lecun et al., 1998] this process consists in the application of filters detecting the presence of a certain pattern in the data.

There are also various ways to define the activation function of each unit. The sigmoid function we presented in 2.4 can for example be an interesting choice of activation function, especially when comes the moment to return the probabilistic output expected at the last layer of the network. Another commonly used activation function is the *Rectified Linear Unit (ReLU)* defined as:

$$\text{ReLU}(x) = \max(0, x) \tag{2.9}$$

Because of its performances and its computation easiness [Hahnloser and Seung, 2001], this activation function is commonly used for the hidden layers and this is what will be done in the neural network used in Chapter 3.

Along with the activation function, the number of neurons per layer and the number of hidden layers are usual hyper-parameters modified when tuning the architecture of the algorithm. Neural networks that contain more than 2 hidden layers are usually designated as *deep neural networks*. In this case, we generally talk about *deep learning* rather than machine learning.

Having set the general architecture of the algorithms, neural networks are trained following the backpropagation algorithm [Rumelhart et al., 1988] which principle can be detailed as follows:

1. Perform a prediction on a subset of the training set (also called *batch*).
2. Compute the error made by this prediction.
3. Go through each layer starting from the final one and adjust the parameters of each neuron in order to reduce the error made by the prediction by applying a minimization process such as Gradient Descent on a defined cost function such as the Root Mean Square Error.
4. Repeat the operation until each sample of the training set has been predicted.

Following this backpropagation phase, we evaluate the performance of the algorithm on the validation set by computing the total cost function used in the training phase. We then train the neural network again starting this time with the previously found parameters. This operation is repeated for a certain number of iterations or *epoch* or until the cost function reaches its minimum. This ensures the capacity the algorithm has to generalize on the data and makes sure the cost function is at its minimum.

Following this description, the number of epochs, the number of samples in the batch, the cost function and the minimization algorithm all appear as important hyper-parameters an external user can modify to tune the algorithm and better the quality of the prediction.

Despite of their efficiency, the complexity of these algorithms go with their lack of interpretability. This is currently a hot topic in the field of Artificial Intelligence and is currently being addressed on simple cases such as digits recognition [Liu et al., 2018]. Thus, obtaining physical information from what these algorithms have learnt appears as coming out of the scope of this thesis. Nevertheless, we will figure out in the next chapter a couple of studies that can be made to investigate the influence each feature has on the quality of the prediction of these algorithms.

## 2.3 Evaluating the performances of an algorithm

The performances of an algorithm trained on the training set are evaluated by comparing the predictions performed on the test set to the associated ground truth label. In the case of classification, this prediction often comes as the probability a data point has to belong to the different accessible classes. The final decision of the algorithm can then be obtained by setting a probabilistic decision threshold and assigning an element to a class if its probability is above this threshold.

Without surprise, 0.5 is the most commonly used threshold but the performances reached by the algorithm of the constraint imposed by the problem we are focusing on might imply the choice of a different value.

### 2.3.1 Calibrating the probabilistic output

The probabilistic output gives a certain confidence in the prediction made by the algorithm. Nevertheless, this assertion is only true if it provides a correct representation of the data seen by the algorithm. That is to say, if  $x\%$  of the elements predicted with a probability of  $x\%$  for a given label are actual elements of this label. Such algorithm is defined as being *well-calibrated* and this verification must be done each time the probabilistic output of an algorithm is at stake.

To do so, we evaluate the fraction of the so-called *positive* elements on different probability intervals. A typical representation of such calibration curve is shown in the left panel of Figure 2.6 for a logistic regression, a decision tree and a gradient boosting. For a perfectly calibrated classifier, the calibration should stick to the gray dashed line. Seeing an almost perfect calibration for the blue curve of Logistic regression is not surprising as the main principle of the algorithm stand in minimizing the log of the odds of the prediction. On the opposite, decision trees are well known to show calibration issues [Niculescu-Mizil and Caruana, 2005] and often need a correction of their probabilistic output.

A solution we can apply to calibrate these probabilities stands in fitting the predicted probabilities to the associated training set label and keeping this final output as the final predicted probability [Niculescu-Mizil and Caruana, 2005]. This fit can either be a Logistic Regression, this is the principle of the so-called Platt scaling, or be a stair-shaped monotonous function, this is the principle of the so-called Isotonic Regression. The effect of the latter on the probabilities of a Decision Tree and a gradient boosting is shown in the right panel of Figure 2.6. Although not perfect, the calibration process allowed a more consistent distribution of the predicted probabilities especially for the highest probabilities for which we expect to have the most confidence.

Here again, the choice between Platt scaling and Isotonic regression is left free for an external user and the efficiency of one method compared to the other depends on the considered problem.

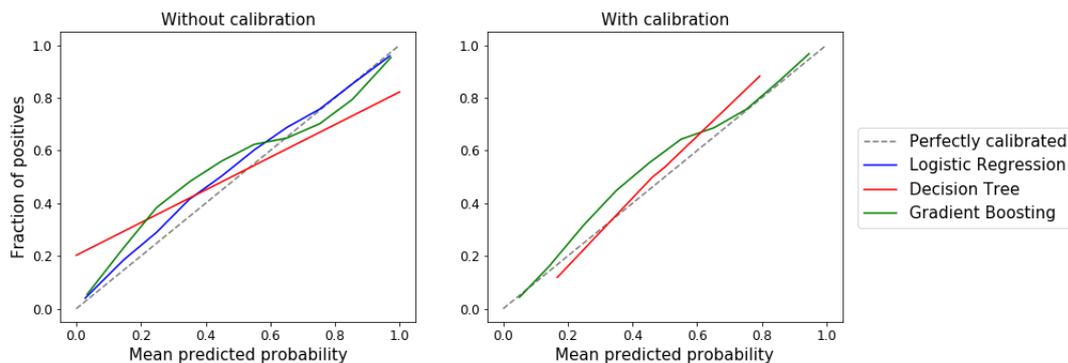


Figure 2.6: Typical representation of the calibration curve of a Logistic regression (blue), a Decision tree (red) and a Gradient Boosting (green) before (left) and after (right) the application of an isotonic regression. The gray dashed line represents the perfect calibration curve. As logistic regression is already a calibrated algorithm (see text), it is only represented in the left panel.

### 2.3.2 Metrics

Once the probabilities are calibrated, the choice of a particular decision threshold directly gives access to the final class predictions made by the trained algorithm. From now on, the predicted classification can be sorted into four categories:

- A **True Positive (TP)** is a point of a class that has been predicted correctly.
- A **True Negative (TN)** is a point not belonging to the concerned class that has been predicted as such.
- A **False Negative (FN)** is a point of a class that has not been correctly predicted.
- A **FP** is a point not belonging to the concerned class that has been predicted as belonging to the class.

With these definitions, the two types of errors likely to be made by an algorithm are the **FN** and the **FP**. From then on, the performance of a classifier can be summarized by two quantities, the recall or **True Positive Rate (TPR)**, defined as one minus the ratio of **FN** over the number of points in the associated class, and the precision defined as one minus the ratio of **FP** over the total number of points predicted in this class :

$$\text{Recall} = 1 - \frac{N_{\text{FN}}}{N_{\text{FN}} + N_{\text{TP}}} \quad (2.10)$$

$$\text{Precision} = 1 - \frac{N_{\text{FP}}}{N_{\text{FP}} + N_{\text{TP}}} \quad (2.11)$$

The value of these metrics comes with the chosen decision threshold and the evolution of the precision and the recall for varying decision thresholds can be represented in the precision-recall curve for which we have a typical representation on the left panel of Figure 2.7. The black dashed line represent the performances that would be reached by a random classifier and we expect the precision-recall curve to always be above this line. Logically, low values allow the prediction of more elements in a given class, which allows a decreasing number of **FN** but an increasing number of **FP**, the recall improves while the precision drops. On the opposite, high decision thresholds will drastically reduce the number of **FP** while augmenting the number of **FN**. The precision soars, the recall drops. All the interest in choosing properly a decision threshold then stands in finding the best compromise between recall and precision that best fulfills the user requirement. The elbow point of this curve is an interesting working point in the precision-recall curve as it offers the best compromise between a high precision and a high recall and we want this point to be as close to the right corner of the left panel of Figure 2.7 as possible. Nevertheless, for the purpose of physical studies, one may set a higher decision threshold in order to have a few number of **FP** provided the method still detects a fair number of events.

The quality of the precision-recall curve can be quantified by the computation of the so-called *average precision* defined by the area under the curve normalized by the area of the zone delimited by the minimal precision reached for the decision threshold of 0. We then expect this average precision to be as close to 1 as possible.

Another method we can use to express the performance of the algorithm is the **Receiving Operator Curve (ROC)** curve that represents the evolution of the recall as a function of the **False Positive Rate (FPR)** defined as:

$$\text{FPR} = \frac{N_{\text{FP}}}{N_{\text{FP}} + N_{\text{TN}}} \quad (2.12)$$

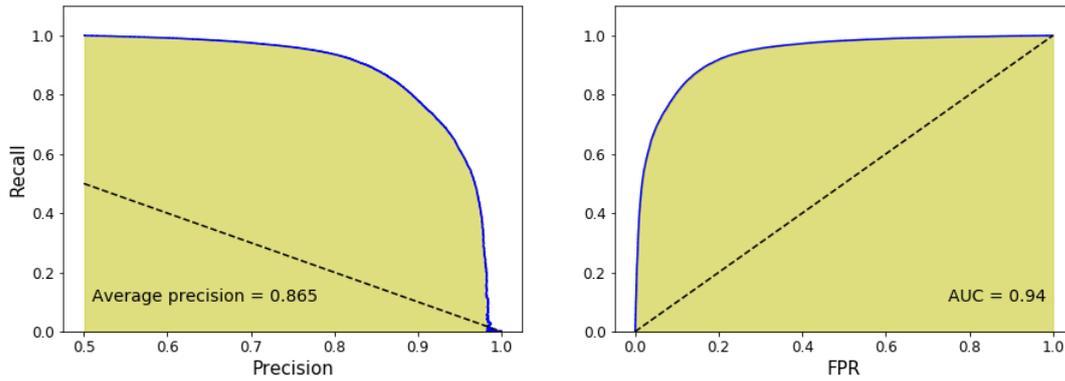


Figure 2.7: Typical representation of the precision-recall (left) and the ROC (right) curves for a trained algorithm. The yellow area represents the average precision and the AUC of the algorithm, the black dashed line represents the performance of a random classifier

A typical representation of the ROC curve is shown in the right panel of Figure 2.7. Similarly to the precision recall curve, the idea here is to find the best compromise between a high recall and a high FPR that here corresponds to the elbow point that we want to be as close to the upper left corner as possible. Once again, seeing this ROC curve above the black dashed line is reassuring as it means a classifier better than random.

The quality of the ROC curve can be assessed by the computation of the Area Under Curve (AUC) that is expected again to be as close to 1 as possible.

The definition of the FPR relies on the total number of TN. However, depending on the problem, it can be impossible to properly define a TN. This for instance the case of object detection in images or the exhibition of patterns in time series data. Thus, exploiting the ROC curve should only be considered when the notion of TN exists.

In addition to the ROC curve and the AUC, we can define the Heidke Skill Score (HSS) that compares the performance of the algorithm to what would come out of a random classifier:

$$HSS = \frac{\frac{N_{TP_s} + N_{TN_s}}{N} - \frac{(N_{TP_s} + N_{FN_s}) * (N_{TP_s} + N_{FP_s}) + (N_{FN} + N_{TN_s}) * (N_{FP} + N_{TN})}{N^2}}{1 - \frac{(N_{TP_s} + N_{FN_s}) * (N_{TP_s} + N_{FP_s}) + (N_{FN} + N_{TN_s}) * (N_{FP} + N_{TN})}{N^2}} \quad (2.13)$$

A negative HSS indicates randomness performs better than the classifier while a perfect forecast would be associated to a HSS of 1. The evolution of the HSS as a function of the decision threshold is shown in the Figure 2.8. Finding decreasing HSS for high decision threshold is not surprising as the number of FN will slightly increase in this case. The main interest in this curve then stands in the value of the HSS we do find for the final decision threshold we chose and the objective we have to bring this score as close to 1 as possible.

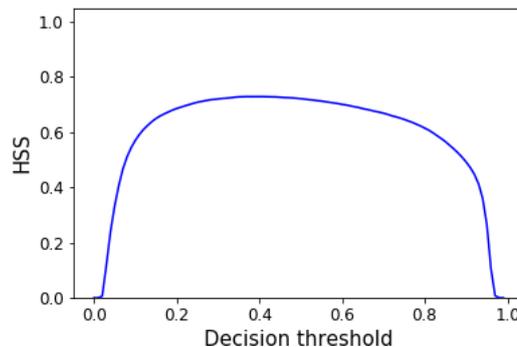


Figure 2.8: Typical representation of the evolution of the HSS as a function of the decision threshold

In the following, we will use machine learning algorithms to automatically elaborate event catalogs from in-situ data measurements. In addition to the mistakes we mentioned and their associated characterization, the beginning and ending times of the detected events also become a source of difference between the predicted and the labeled event list. To take into account this additional source of error, we define the Jaccard index between two events lists as:

$$\text{Jaccard}(A, B) = \frac{\text{duration}(A \cap B)}{\text{duration}(A \cup B)} \quad (2.14)$$

Once again, the better the prediction, the closest to 1 the Jaccard. With this definition, a low Jaccard can be induced by both events exclusive to one of the two lists and the differences on the boundaries of events. The Jaccard index gives clues on how well a predicted events list is similar to a labeled lists and an example of computation of this metrics will be detailed in the next chapter.

## 2.4 Implementing Machine Learning

All of the algorithms we presented are easily usable in Python with the associated packages. We used the machine learning algorithms implemented in the package *Scikit-learn* [Pedregosa et al., 2011] and the deep learning algorithms implemented in the package *Tensorflow* [Abadi et al., 2015] used as a backend of the library *Keras* [Chollet et al., 2015].

The machine learning algorithms have been trained on an AMD ryzen™threadripper™2990wx processor while the deep learning algorithms have been trained on two NVIDIA GeForce GTX 1080 TI™Graphical Processing Units.

## 2.5 Machine Learning in space physics

Space physics is not an exception in the diversity of the domains in which machine learning algorithms have a huge potential. Camporeale [2019] recently identified the future of their utilization to be mainly focused around three objectives: the use of solar imagery, the estimation of geomagnetic indices and the detection and classification of time-series patterns.

They have especially been widely used on solar images for tasks such as the prediction of solar flares (Colak and Qahwaji [2009] and references therein), the detection of sunspots [Yang et al., 2018] or even the classification of solar active regions that produced a solar flare with or without a CME [Bobra and Ionidis, 2016].

These algorithms also prove their worth in estimating the geomagnetic indices based on on-ground measurement (Zhelavskaya et al. [2019] and references therein) showing their usefulness in the field of space weather.

Concerning the detection and classification of streaming time series patterns from in-situ data, Miniere [1999] used neural networks to identify and classify electron and proton whistlers. Karimabadi et al. [2009] developed a data mining method called MineTool-TS they used to provide a classification of data intervals that contained Flux Transfer Event (FTE) or not as well as an extension to apply data mining to 3D simulation data [Sipes and Karimabadi, 2012]. Using a support vector machine on magnetopause crossings measured by 23 different spacecrafts, Wang et al. [2013] provided an empirical three dimensional model for the Earth magnetopause. Finally, Camporeale et al. [2017] provided an accurate method of solar wind classification into 4 classes using a Gaussian Process.

The emergence of these techniques in this specific type of problem even got further with the elaboration of methods to detect beginning and ending dates of events [Nguyen et al., 2019] or algorithms with the ability to classify specific regions in the Earth plasma environment [Olshevsky et al., 2019]. Nevertheless, despite of fair performances of such methods that will be discussed later-on, the efficiency of such methods on the streaming in-situ data provided by spacecraft is highly influenced by the representation we make of these objects or processes, and the definition

we have of the observational signatures we are supposed to see when crossing them. The underlying reasons standing behind this ambiguity and the tools we can use to deal with it will be detailed in the next chapter.

Chapters 3, 4 and 6, will focus on three examples of utilization of machine learning in the field of space physics in the frame of automatic detection and classification of specific patterns in streaming in-situ data.

## 2.6 Bibliography

- Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X.: TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, URL <https://www.tensorflow.org/>, software available from tensorflow.org, 2015. 39
- Berkson, J.: Application of the Logistic Function to Bio-Assay, *Journal of the American Statistical Association*, 39, 357–365, URL <http://www.jstor.org/stable/2280041>, 1944. 28, 30
- Bobra, M. G. and Ilonidis, S.: Predicting Coronal Mass Ejections Using Machine Learning Methods, *Astrophysical Journal*, 821, 127, <https://doi.org/10.3847/0004-637X/821/2/127>, 2016. 39
- Botev, A., Lever, G., and Barber, D.: Nesterov’s accelerated gradient and momentum as approximations to regularised update descent, in: 2017 International Joint Conference on Neural Networks (IJCNN), pp. 1899–1903, <https://doi.org/10.1109/IJCNN.2017.7966082>, 2017. 30
- Breiman, L.: Random Forests, *Mach. Learn.*, 45, 5–32, <https://doi.org/10.1023/A:1010933404324>, URL <https://doi.org/10.1023/A:1010933404324>, 2001. 33
- Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J.: *Classification and Regression Trees*, Wadsworth and Brooks, Monterey, CA, 1984. 32
- Brown, I. and Mues, C.: An experimental comparison of classification algorithms for imbalanced credit scoring data sets, *Expert Syst. Appl.*, 39, 3446–3453, <https://doi.org/10.1016/j.eswa.2011.09.033>, 2012. 33
- Camporeale, E.: The Challenge of Machine Learning in Space Weather: Nowcasting and Forecasting, *Space Weather*, 17, 1166–1207, <https://doi.org/10.1029/2018SW002061>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018SW002061>, 2019. 39
- Camporeale, E., Carè, A., and Borovsky, J. E.: Classification of Solar Wind With Machine Learning, *Journal of Geophysical Research (Space Physics)*, 122, 10,910–10,920, <https://doi.org/10.1002/2017JA024383>, 2017. 39
- Chollet, F. et al.: Keras, <https://keras.io>, 2015. 39
- Colak, T. and Qahwaji, R.: Automated Solar Activity Prediction: A hybrid computer platform using machine learning and solar imaging for automated prediction of solar flares, *Space Weather*, 7, S06001, <https://doi.org/10.1029/2008SW000401>, 2009. 39
- Fisher, R. A.: The use of multiple measurements in taxonomic problems, *Annals of Eugenics*, 7, 179–188, <https://doi.org/10.1111/j.1469-1809.1936.tb02137.x>, URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-1809.1936.tb02137.x>, 1936. 28, 32

- Friedman, J. H.: Greedy function approximation: A gradient boosting machine., *Ann. Statist.*, 29, 1189–1232, <https://doi.org/10.1214/aos/1013203451>, URL <https://doi.org/10.1214/aos/1013203451>, 2001. 33
- Hahnloser, R. H. R. and Seung, H. S.: Permitted and Forbidden Sets in Symmetric Threshold-Linear Networks, in: *Advances in Neural Information Processing Systems 13*, edited by Leen, T. K., Dietterich, T. G., and Tresp, V., pp. 217–223, MIT Press, URL <http://papers.nips.cc/paper/1793-permitted-and-forbidden-sets-in-symmetric-threshold-linear-networks.pdf>, 2001. 35
- Karimabadi, H., Sipes, T. B., Wang, Y., Lavraud, B., and Roberts, A.: A new multivariate time series data analysis technique: Automated detection of flux transfer events using Cluster data, *Journal of Geophysical Research (Space Physics)*, 114, A06216, <https://doi.org/10.1029/2009JA014202>, 2009. 39
- Kiefer, J. and Wolfowitz, J.: Stochastic Estimation of the Maximum of a Regression Function, *Ann. Math. Statist.*, 23, 462–466, <https://doi.org/10.1214/aoms/1177729392>, URL <https://doi.org/10.1214/aoms/1177729392>, 1952. 30
- Kingma, D. and Ba, J.: Adam: A Method for Stochastic Optimization, *International Conference on Learning Representations*, 2014. 30
- Lanham, M.: *Learn ARCore - Fundamentals of Google ARCore*, 2018. 30
- Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P.: Gradient-based learning applied to document recognition, in: *Proceedings of the IEEE*, pp. 2278–2324, 1998. 35
- Li, M., Zhang, T., Chen, Y., and Smola, A. J.: Efficient Mini-Batch Training for Stochastic Optimization, in: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14*, p. 661–670, Association for Computing Machinery, New York, NY, USA, <https://doi.org/10.1145/2623330.2623612>, URL <https://doi.org/10.1145/2623330.2623612>, 2014. 30
- Liu, X., Wang, X., and Matwin, S.: Interpretable Deep Convolutional Neural Networks via Meta-learning, *CoRR*, abs/1802.00560, URL <http://arxiv.org/abs/1802.00560>, 2018. 35
- Miniere, X.; Pincon, J. L. F.: A neural network approach to the classification of electron and proton whistlers , *ournal of Atmospheric and Terrestrial Physics*, 58, 911–924, [https://doi.org/10.1016/0021-9169\(95\)00077-1](https://doi.org/10.1016/0021-9169(95)00077-1), 1999. 39
- Murtagh, E.: Multilayer perceptrons for classification and regression, *Neurocomputing*, 2, 183 – 197, [https://doi.org/https://doi.org/10.1016/0925-2312\(91\)90023-5](https://doi.org/https://doi.org/10.1016/0925-2312(91)90023-5), URL <http://www.sciencedirect.com/science/article/pii/0925231291900235>, 1991. 35
- Nguyen, G., Aunai, N., Fontaine, D., Pennec, E. L., den Bossche, J. V., Jeandet, A., Bakkali, B., Vignoli, L., and Blancard, B. R.-S.: Automatic Detection of Interplanetary Coronal Mass Ejections from In Situ Data: A Deep Learning Approach, *The Astrophysical Journal*, 874, 145, <https://doi.org/10.3847/1538-4357/ab0d24>, 2019. 39
- Niculescu-Mizil, A. and Caruana, R.: Obtaining Calibrated Probabilities from Boosting, in: *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence, UAI'05*, p. 413–420, AUAI Press, Arlington, Virginia, USA, 2005. 36
- Nolte, A., Wang, L., Bilicki, M., Holwerda, B., and Biehl, M.: Galaxy classification: A machine learning analysis of GAMA catalogue data, *Neurocomputing*, 342, 172 – 190, <https://doi.org/https://doi.org/10.1016/j.neucom.2018.12.076>, URL <http://www.sciencedirect.com/science/article/pii/S0925231219301353>, *advances in artificial neural networks, machine learning and computational intelligence*, 2019. 28

- Olshevsky, V., Khotyaintsev, Y. V., Divin, A., Delzanno, G. L., Anderzen, S., Herman, P., Chien, S. W. D., Avanov, L., and Markidis, S.: Automated classification of plasma regions using 3D particle energy distribution, arXiv e-prints, arXiv:1908.05715, 2019. 39
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research*, 12, 2825–2830, 2011. 39
- Ren, Y., Zhao, P., Sheng, Y., Yao, D., and Xu, Z.: Robust Softmax Regression for Multi-class Classification with Self-Paced Learning, in: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pp. 2641–2647, <https://doi.org/10.24963/ijcai.2017/368>, URL <https://doi.org/10.24963/ijcai.2017/368>, 2017. 31
- Rumelhart, D. E., Hinton, G. E., and Williams, R. J.: *Learning Representations by Back-Propagating Errors*, p. 696–699, MIT Press, Cambridge, MA, USA, 1988. 35
- Sajda, P.: *Machine Learning for Detection and Diagnosis of Disease*, <https://doi.org/10.1146/annurev.bioeng.8.061505.095802>, 2006. 28
- Samuel, A. L.: Some Studies in Machine Learning Using the Game of Checkers, *IBM Journal of Research and Development*, 3, 210–229, <https://doi.org/10.1147/rd.33.0210>, 1959. 28, 34
- Sipes, T. B. and Karimabadi, H.: *MineTool-M 2 : An Algorithm for Data Mining of 2D Simulation Data*, 2012. 39
- Wang, Y., Sibeck, D. G., Merka, J., Boardsen, S. A., Karimabadi, H., Sipes, T. B., Šafránková, J., Jelínek, K., and Lin, R.: A new three-dimensional magnetopause model with a support vector regression machine and a large database of multiple spacecraft observations, *Journal of Geophysical Research (Space Physics)*, 118, 2173–2184, <https://doi.org/10.1002/jgra.50226>, 2013. 39
- Yang, Y., Yang, H., Bai, X., Zhou, H., Feng, S., and Liang, B.: Automatic Detection of Sunspots on Full-disk Solar Images using the Simulated Annealing Genetic Method, *Publications of the Astronomical Society of the Pacific*, 130, 104 503, <https://doi.org/10.1088/1538-3873/aadbfa>, 2018. 39
- Zhelavskaya, I. S., Vasile, R., Shprits, Y. Y., Stolle, C., and Matzka, J.: Systematic Analysis of Machine Learning and Feature Selection Techniques for Prediction of the Kp Index, *Space Weather*, 17, 1461–1486, <https://doi.org/10.1029/2019SW002271>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019SW002271>, 2019. 39

# Chapter 3

## An example of ambiguously labeled problem: automatic detection of ICMEs

*La conscience n'est jamais assurée de surmonter l'ambiguïté et l'incertitude.*

Edgar Morin  
(Le paradigme perdu)

### Contents

---

<b>3.1 Introduction</b> . . . . .	<b>44</b>
<b>3.2 Interplanetary Coronal Mass Ejections</b> . . . . .	<b>44</b>
<b>3.3 Data</b> . . . . .	<b>47</b>
3.3.1 WIND . . . . .	47
3.3.2 ICME catalog . . . . .	48
<b>3.4 Algorithm</b> . . . . .	<b>49</b>
3.4.1 Scaling . . . . .	49
3.4.2 Windowing . . . . .	50
3.4.3 Convolutional Neural Network (CNN) . . . . .	50
3.4.4 Post-processing . . . . .	52
3.4.5 Automatization . . . . .	53
<b>3.5 Results</b> . . . . .	<b>54</b>
3.5.1 Precision and recall . . . . .	54
3.5.2 High-recall region . . . . .	56
3.5.3 High-precision region . . . . .	60
<b>3.6 Robustness</b> . . . . .	<b>63</b>
3.6.1 Importance of the various features as ICME indicators . . . . .	63
3.6.2 Influence of the number of ICMEs in the training period . . . . .	65
3.6.3 Influence of the training, validating and testing period . . . . .	66
<b>3.7 Global quality of the prediction</b> . . . . .	<b>66</b>
<b>3.8 Conclusion</b> . . . . .	<b>67</b>
<b>3.9 Bibliography</b> . . . . .	<b>68</b>

---

### 3.1 Introduction

We mentioned in the previous chapter that the application of machine learning algorithms to the detection of specific physical signatures of events in streaming in-situ time series data provided by spacecraft found a bottleneck in the manually defined event catalogs given as an input. The underlying reasons of this ambiguity are numerous.

First of all, these events are measured whenever they are crossed by a spacecraft, the in-situ signature we obtain is then the 1D slice of a 3D much bigger structure, which bias the global overview we have on them.

Second, those structures are surrounded by a continuum of other events likely to happen in the solar wind or in the near-Earth environment which makes it hard to define properly where does an event begins or ends and if it really is the event we were expecting.

These two reasons can explain by themselves the difference of the criteria applied by different experts to label the same kind of event [Shinde and Russell, 2003]. Coupled to the different tolerance of each experts and the psychological factors <sup>1</sup> that can influence the experts criteria throughout the identification process, the existing event lists are assumed to be incomplete and ambiguous from a person to another.

Consequently, we expect an automatic event detection method to bring a considerable gain in time, objectivity and reproducibility in these catalogs. In this chapter, adapted from Nguyen et al. [2019], we use the problem of the automatic detection of ICMEs to show how the quality of the prediction of supervised machine learning is limited by the ambiguity of the events lists we use to label our dataset.

In Section 2, we will give a brief presentation of ICMEs and the problematics linked to their automatic detection. In section 3, we will present the data and event lists we used for the study. Section 4 will present in details the principle of our pipeline and how did we designed it. The obtained performances will be shown in Section 5. We will then discuss the robustness of the prediction regarding the different dataset features, the size of the event list used in the training set and the period considered for training, validating and testing our method. Finally, Section 6 will discuss the global quality of our prediction in comparison to ICME lists manually identified by experts.

### 3.2 Interplanetary Coronal Mass Ejections

CMEs are spectacular manifestations of the solar activity which are responsible for the expulsion at large velocities of large quantities of solar plasma and magnetic field. Their interplanetary counterpart, the so-called ICMEs, interact with the planetary environments. Their most famous subclass, the Magnetic Clouds (MCs), being well-known for their strong geoeffectiveness and their capacity to trigger magnetic storms that severely impact the Earth magnetosphere, ionosphere and even human activities.

After initial studies [Burlaga et al., 1981; Gosling et al., 1973; Klein and Burlaga, 1981], these events have been extensively investigated from in-situ measurements (Kilpua et al. [2017] and references therein).

Observationnally speaking, we observe the typical signature of these events whenever a spacecraft goes through the ICME following the dashed arrow represented on the schematic representation of an ICME shown in the Figure 3.1.

The associated typical-measurement of such events by the WIND spacecraft is shown in Figure 3.2. The top panel shows an enhanced magnetic field compared to the surrounding ambient solar wind and a long (here about 2 days), smooth rotation of the magnetic field. It is associated with a low proton temperature (fourth panel) resulting in low values of the parameter  $\beta$  defined as the ratio between the thermal pressure and the magnetic pressure (second panel). The third panel shows an enhanced velocity compared to the preceding solar wind with a declining profile. The

---

<sup>1</sup>especially its time consumption

MC is featured by a preceding abrupt and simultaneous jump in the magnetic field and velocity (indicated by the dashed line in Figure 3.2), and by a turbulent sheath, between the shock and the MC.

These are the main criteria generally used for the identification of ICMEs and we consider an event that fulfill all of these criterias to be an actual MC [Chi et al., 2016; Kilpua et al., 2017; Zurbuchen and Richardson, 2006]. However, not all ICMEs feature all standard ICME signatures and there is no signature that would be present in all ICMEs [Gosling et al., 1973; Kilpua et al., 2017; Richardson and Cane, 2010]. For example, about half of the ICMEs drive a fast upstream shock and are preceded by a sheath (Chi et al. [2016] and references therein).

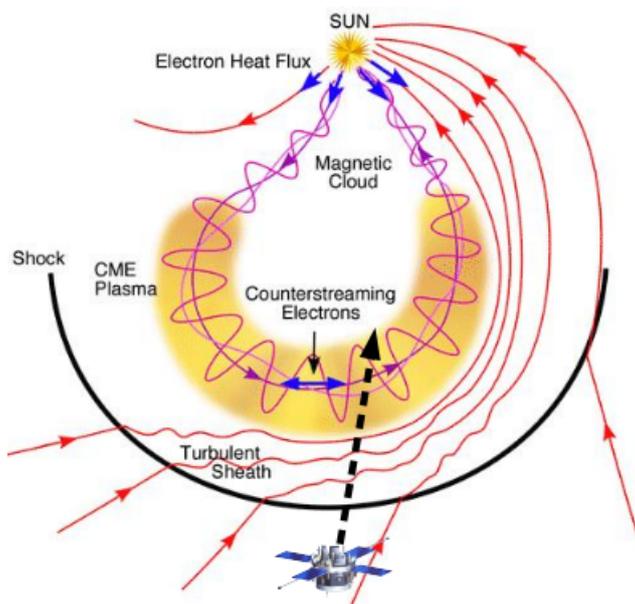


Figure 3.1: Same representation than the left panel of 1.5. The dashed arrow represents the 2D-slice realized by a spacecraft when providing the typical in-situ signature of an ICME represented in Figure 3.2. (Adapted from Zurbuchen and Richardson [2006])

With the increase of the number of solar wind oriented missions (WIND, ACE, STEREO and the more recent Parker Solar Probe and Solar Orbiter, ...), the number of ICMEs catalogs associated with these two missions flourished. Lepping et al. [2006] referenced 106 MCs between 1995 and 2008. Richardson and Cane [2010] listed 373 ICMEs between 1996 and 2015. Jian et al. [2006] listed 250 ICMEs between 1995 and 2009. Nieves-Chinchilla et al. [2018] listed 302 ICMEs between 1997 and 2016. And Chi et al. [2016] listed 465 ICMEs from 1995 to 2015. In average, 80% of the ICMEs of a given list are present in another list Chi et al. [2016].

The main difference we notice from one list to another stands in the criteria used by the authors to identify the ICME such as the presence of a front shock, the presence of a sheath, the fit to a flux rope model [Lepping et al., 2006] or the importance given by the authors to a specific physical parameter [Chi et al., 2016; Richardson and Cane, 2010]. Moreover, a given ICME might only fulfill a subset of these criteria or partially fulfill them which complicates their identification and consequently their automated detection.

The establishment of such catalogs allowed the study of ICMEs from a statistical point of view. These studies indicated the enhanced magnetic field and the low proton temperature as being typical characteristic of ICME in-situ signatures. They also indicated that the yearly occurrence of ICMEs is correlated to the solar cycle [Chi et al., 2016; Jian et al., 2006], and that ICMEs were considered to be long term events with an average duration being equal to 25 hours [Klein and Burlaga, 1981]. Complete conclusions of such studies can be found in Chi et al. [2016], Nieves-Chinchilla et al. [2018], Mitsakou and Moussas [2014], Kilpua et al. [2017] and references therein.

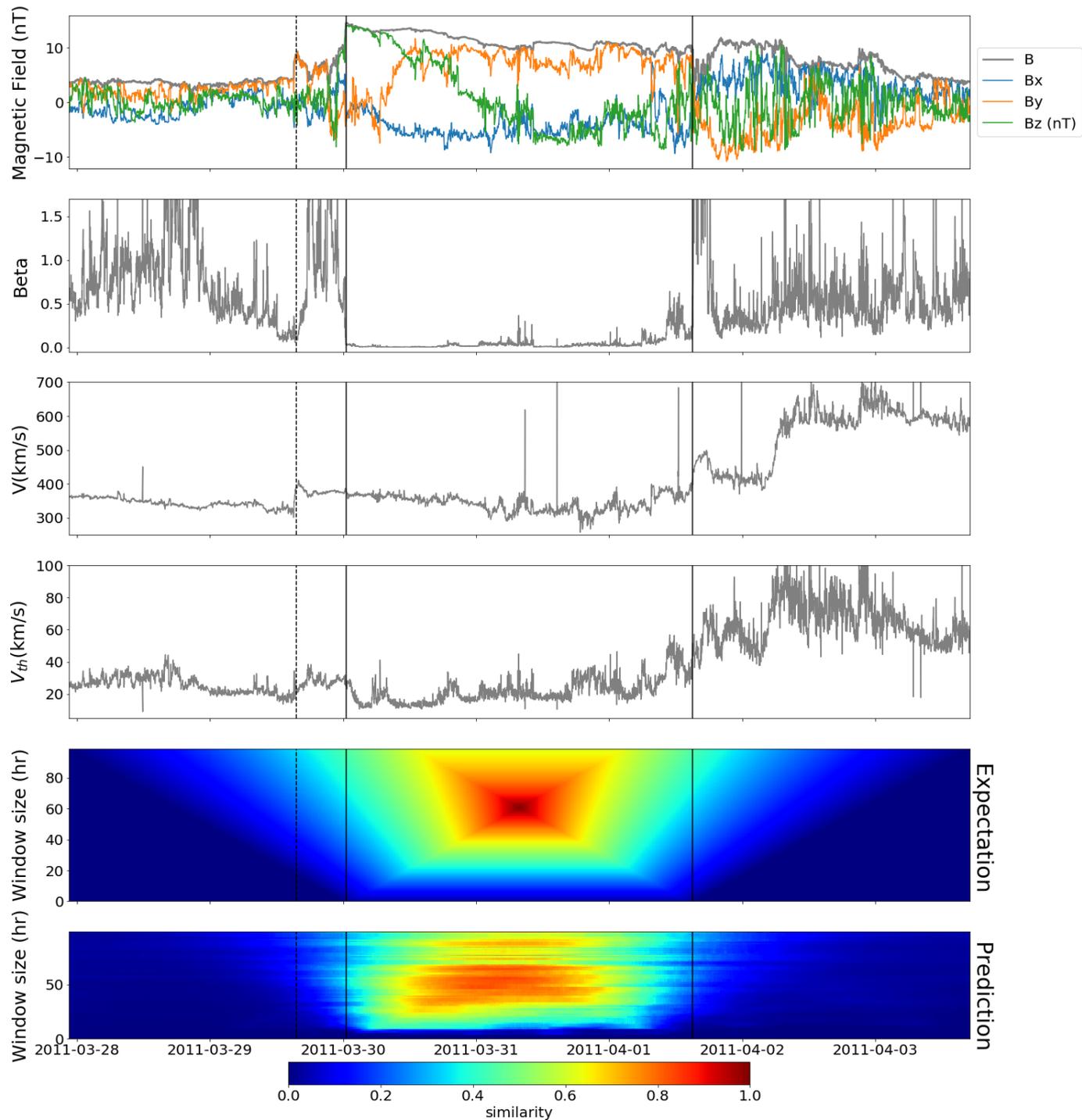


Figure 3.2: Solar wind observation during an ICME from the WIND spacecraft located at the Lagrangian Point L1. The solid vertical lines delimitate the ICME while the dashed vertical line indicate the beginning of the sheath. From the top to the bottom are represented : the magnetic field amplitude and components, the plasma parameter  $\beta$ , the solar wind velocity, the thermal velocity, the similarity the ICME have with sliding windows of various sizes (from 1 to 100 Hr) and the similarity predicted by our method.

All the beginning and ending dates of the ICMEs present in these catalogs have been identified by visual inspection. For the reason we mentioned in the introduction, this leads to incomplete, ambiguous and hardly reproducible catalogs which bias the statistical conclusions we can extract from them.

The development of an automatic identification method of ICMEs would then bring a considerable gain in time and objectivity in the elaboration of our catalogs.

Lepping et al. [2005] proposed an automatic detection method based on empirical thresholds. These thresholds are inferred from the expert knowledge of ICMEs properties and involve various physical and temporal parameters such as the duration, the plasma  $\beta$ , the magnetic field, the bulk velocity or the quality of the fit with a flux rope model.

Even though this method was able to recognize a fair quantity of identified events (45 on a total of 76 ICMEs in the period considered), the large number of found false positives (66 for a total of 111 predicted ICMEs) evidenced both the incompleteness of the list as well as the limits of using fixed thresholds for automatic identification.

Recently, Ojeda-Gonzalez et al. [2017] proposed an alternative automatic identification method based on the computation of a Spatio-Temporal Entropy. However, the method was tested on a very low number of ICMEs and its performance on long periods is not known.

Finally, the problem of identifying patterns in in-situ data measurement is at the root of many, if not all, observational studies. No matter how efficient previous automatic detection methods were, if some exist, they are based on expert and detailed knowledge of target event properties and thus are very specific to their detection. This methodology thus imposes to re-think the detection pipeline entirely for each kind of event, which constitutes a serious bottleneck that may be comparable or worse than doing the visual identification itself.

One way to overcome these constraints stands in the use of supervised machine learning algorithms that have proven their worth for various tasks in the field of space physics as explained in the previous chapter. Nevertheless, none of these methods was used to identify the starting and the ending times of a specific kind of event in streaming time series yet and this will be the challenge of the method we detail in the following sections.

## 3.3 Data

### 3.3.1 WIND

WIND is a NASA mission that was launched on the 1<sup>st</sup> of November 1994. After some time spent in the magnetosphere, the spacecraft went through a solar orbit at the Lagrangian point  $L_1$  where it continuously provided solar wind measurements from then on. We used the data provided by WIND between the 1st of October 1997 and the 1st of January 2016.

The magnetic field information were provided by the Magnetic Field Investigator (MFI) with a temporal resolution of 1 minute. The plasma moments were provided by the Solar Wind Experiment (SWE) with an approximate resolution of 90s and the particle distribution function between 0.3 and 10keV were provided by the 3-D Plasma and Energetic Particles Experiment (3DP) with a temporal resolution of 1 min.

The obtained dataset is therefore made of 30 primary input variables: the bulk velocity and its components  $V, V_x, V_y, V_z$ , the thermal velocity  $V_{th}$ , the magnetic field, its components and their Root Mean Square (RMS) :  $B, B_x, B_y, B_z, \sigma_{B_x}, \sigma_{B_y}, \sigma_{B_z}$ , the density of protons and  $\alpha$  particles obtained from both moment and non-linear analysis :  $N_p, N_{p,nl}$  and  $N_{\alpha,nl}$  as well as 15 canals of proton flux between 0.3 and 10 keV.

Due to instrumental constraints, holes are present within the whole dataset, the great majority of these holes have a duration between 2 and 10 minutes. On the other hand, the crossings of ICMEs with their sheath typically have durations of several hours. We therefore resample the data to a 10 minutes resolution, thereby eliminating the greatest majority of the holes while still remaining accurate in the determination of start and end times of labeled events.

In addition to this 30 input variables, we computed 3 additional features that will also serve as input variables : the plasma parameter  $\beta$ , the dynamic pressure  $P_{dyn} = N_p V^2$  and the normalized magnetic fluctuations :  $\sigma_B = \sqrt{(\sigma_{B_x}^2 + \sigma_{B_y}^2 + \sigma_{B_z}^2)}/B$ .

Apart from the sections 3.6.3 and 3.7, the period between 1998 and 2010 constitutes our *training set*, the period between 1997 and 1998 our *validation set* and the period between 2010 and 2016 our *test set*. This repartition has the advantage of considering a whole solar cycle (1997-2008) during the training phase and consequently giving to our algorithm the opportunity to notice the changes in the solar wind and in the frequency of ICMEs during a solar cycle.

### 3.3.2 ICME catalog

The ICME catalog we used consists in the union of the different WIND ICME lists [Chi et al., 2016; Jian et al., 2006; Lepping et al., 2006; Nieves-Chinchilla et al., 2018; Richardson and Cane, 2010]. During the various tests of our method, additional ICMEs that were not present in any of the existing lists were detected and have been progressively added to our catalog. Following these investigations, 148 new ICMEs have been discovered throughout our period and added to the dataset after a visual validation of the associated in-situ measurement. This represents 22% of our total dataset for a total of 657 ICMEs distributed as follows: 420 ICMEs in the *training set*, 13 ICMEs in the *validation set* and 232 ICMEs in the *test set*.

Our catalog can be found online<sup>2</sup> and will be designed, in the following, as the *Reference List (RL)*. We consider that this catalog is still not exhaustive and that events predicted by our pipeline but not being present in the catalog might be in fact actual ICME as it will be explained in 3.5.2.

Statistically speaking, we ensure the consistency of the *RL* by comparing it to the list established by Chi et al. [2016] that has the advantage of being extended over the same time period, as well as being the one containing the most events.

Figure 3.3 compares the yearly occurrence frequencies of the two catalogs. Even if the *RL* has more ICMEs, the trend observed in the annual variation of the number of ICMEs is conserved and confirms that a whole solar cycle is included in our *training set*.

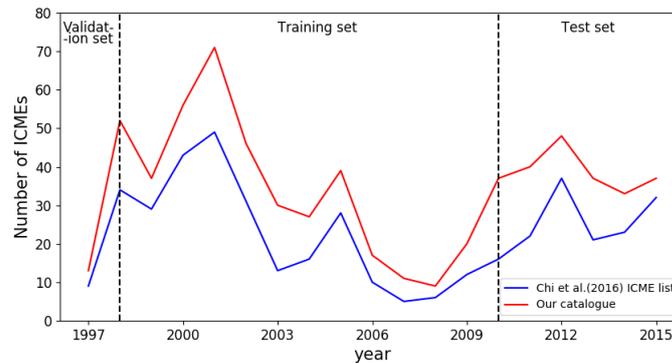


Figure 3.3: Yearly occurrence frequencies of ICMEs of the *RL* (red) and the list established by Chi et al. [2016] (blue). The vertical dashed line indicate the yearly disposition of our training, validation and test set.

As displayed by the bottom panel of Figure 3.4, the number of events in the *RL* (in pink) is larger than in the list of Chi et al. [2016] (in blue) (the overlap region appears in purple) but with a comparable distribution in duration. The first row of Figure 3.4 also shows consistent distributions of the magnetic field and the thermal velocity between both lists. To ensure this similarity as a proof of consistence, we compare the magnetic field and the thermal velocity of Chi et al. [2016] ICME list with random intervals of data in which no ICME was found, the duration of these intervals being distributed according to the duration distribution of our catalog (bottom panel). This comparison in the second row of Figure 3.4 shows distributions with larger magnetic fields and reduced thermal velocities in Chi et al. [2016] ICME list (in dark blue) than in the list without

<sup>2</sup><https://github.com/gautiernguyen/Automatic-detection-of-ICMEs-at-1-AU-a-deep-learning-approach>

ICMEs (in yellow) (overlap region in light blue). This is consistent with the expected ICMEs characteristics used by experts for their identification [Zurbuchen and Richardson, 2006]. On the other hand, this difference compared to the similarity we have for the two ICME lists (first row) ensures that the ICME catalog we used in our identification process is consistent with the previous existing ICME catalogs.

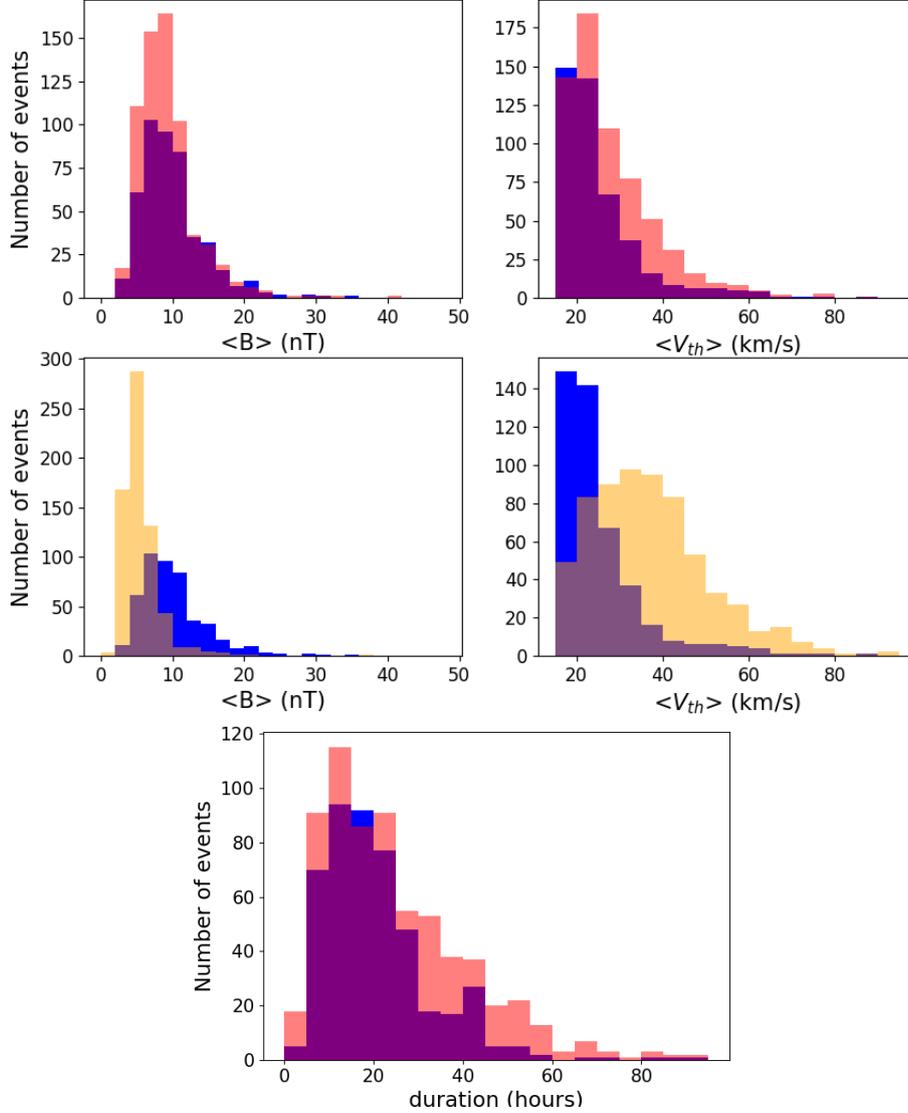


Figure 3.4: First row: distribution of the mean values over the whole ICME interval of the magnetic field and thermal velocity,  $\langle B \rangle$  and  $\langle V_{th} \rangle$  compared for the list of Chi et al. [2016] (blue) and our ICME catalog (pink) (overlap in purple). Second row: idem for the list of Chi et al. [2016] (blue) and random intervals of solar wind in which there is no ICME (yellow) (overlap in light blue). Bottom Panel : distribution of ICME duration for the list of Chi et al. [2016] and our ICME catalog (same color code as first row).

## 3.4 Algorithm

### 3.4.1 Scaling

We scale and normalize the data in order for each feature to have an average of 0 and a standard deviation of 1 in the training set. For the  $i^{th}$  of the  $j^{th}$  feature, this is done applying the following formula:

$$x_{ij,scaled} = \frac{x_{ij} - \mu_j}{\sigma_j} \quad (3.1)$$

Where  $\mu_j$  and  $\sigma_j$  are the average and the standard deviation of the  $j^{th}$  feature of the training set respectively.

This pre-process technique is particularly useful when used with a logistic regression or a neural network as it prevents an eventual feature preselection due to the differences in orders of magnitude that can exist among them [Juszczak et al., 2002].

### 3.4.2 Windowing

The data is grouped into windows of a hundred different sizes (from 1 to 100 hours) that are sliding on both our training and validation sets at a period corresponding to the global dataset sampling: 10 minutes. A window of data represents the values of the 33 input variables within this window that will be treated simultaneously. Our initial dataset is then converted into 100 datasets, each of them corresponding to a size of sliding window. Following this process, there are around 622000 windows of data in the training set, around 311000 in the test set and around 12960 in the validation set. In the following, we will refer to one of these datasets by calling it by its window size.

For each window size, the principle of the detection will stand in estimating a similarity parameter  $y_i$  for each window of data  $X_i$ , using regression methods for classification purpose. Logically, we would expect this parameter to be equal to 0 when no ICME intersects our window while it shall be equal to 1 when a window perfectly matches an ICME. The similarity  $s$  window  $X_i$  has with a given ICME could then easily be defined by :

$$s(\text{ICME}, X_i) = \frac{\text{duration}(X_i \cap \text{ICME})}{\text{duration}(X_i \cup \text{ICME})} \quad (3.2)$$

Given an ICME list and a window, we then define the expected similarity of the window  $X_i$  as:

$$s(X_i) = \max_{\text{ICME} \in \text{list}} s(\text{ICME}, X_i) \quad (3.3)$$

The aim of each regression would then stand in predicting a similarity  $y_i$  in order to make it as close to the expected similarity  $s(X_i)$  as possible.

Stacked together, similarities of many windows make a so-called *2D similarity map*. An example of such a map for a specific ICME is shown on Figure 3.2, fifth panel. The similarity is coded with the color bar, and the ordinate represents the window size from 1 to 100 hours. The maximum is reached in the middle of the event for the window corresponding to the ICME size. One can see that the similarity decreases faster in time for small windows than for large windows. Indeed, as they slide along time, small windows cease to see high similarities pretty quickly while large windows remain in range of ICME-like data - and thus high similarities - for quite longer times.

The orange bars in Figure 3.5 represent the distribution in similarity we have on the test set for a window size of 30 hours. Having the largest part of the computed similarity being equal to 0 is not surprising as the ICME are very seldom events in the solar wind .

### 3.4.3 Convolutional Neural Network (CNN)

#### Principle

For each window size, we fit a Convolutional Neural Network (CNN) to the associated set of similarities. CNNs are a specific kind of neural networks particularly good when it comes to images classification and object detection.

Here, each neuron is a learnable filter that, slided (convolved) on the data, produces a feature map that tries to recognize a specific pattern in the input data. The aim of additional layers is then here to go deeper in the details of the input data by fitting the characteristics of additional filters.

When moving from a layer to another, it can often be useful to reduce the dimensional of the produced feature maps by exhibiting their maximum. This operation, called *max pooling*, allows

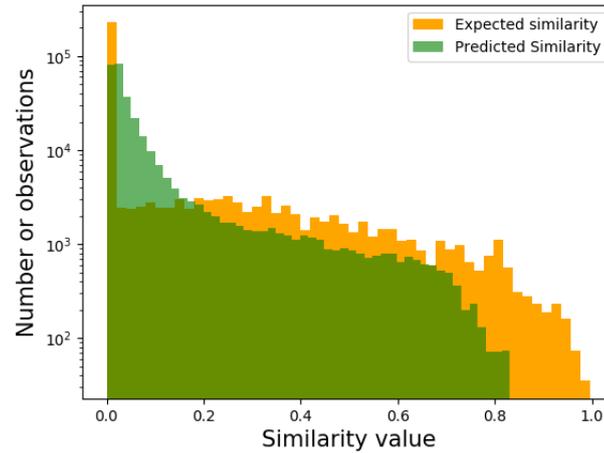


Figure 3.5: Distribution of the similarity values we have on the test set for a window size of 30 hours (orange) and the similarity repartition we have in our prediction for a window size of 30 hours (green).

both a computational gain in time and hampers the risk of overfitting by reducing the dimensionality of the input of the following neurons.

The final outputs of the last hidden layer are concatenated or *flattened* and used as the input of a fully interconnected neural network that produces the final prediction of the algorithm.

### Architecture

A schematic representation of the CNN trained in this chapter is shown in Figure 3.6.

The architecture of each CNN is widely adapted from the one found in Yang et al. [2015] that proved its efficiency for the recognition of patterns in time series data and is made as follows:

1. A first convolutional layer made of 80 filters activated with ReLU
2. Max pooling
3. A second convolutional layer made of 80 filters activated with ReLU
4. Max Pooling
5. A third convolutional layer made of 80 filters activated with ReLU
6. Flattening
7. A neural network made of one hidden layer of 10 units activated with ReLU
8. A final output cell activated with the sigmoid function.

In order to avoid the overfit of the training set, a fraction of the nodes that constitute the fully connected network layer are randomly dropped out at each step of the training phase. This operation ensures the robustness of the features learned by the algorithm and thus reduces the overfit.

The algorithm was trained for 100 epochs for a batch size of 128 samples, one of the commonly used batch sizes when it comes to the design of neural networks [Bengio, 2012], and we minimized the *log-cosh* cost function  $J(\mathbf{X}) = \frac{1}{n} \sum_{i=1}^n \log[\cosh(y_i - s(X_i))]$  that has the advantage of being less sensitive to falsely predicted value than the RMSE [Grover, 2019].

We minimize the cost function using the Stochastic Gradient Descent Algorithm [Kiefer and Wolfowitz, 1952] that has the advantage of being computationally cheaper than gradient descent while converging faster and being more likely to avoid local minima.

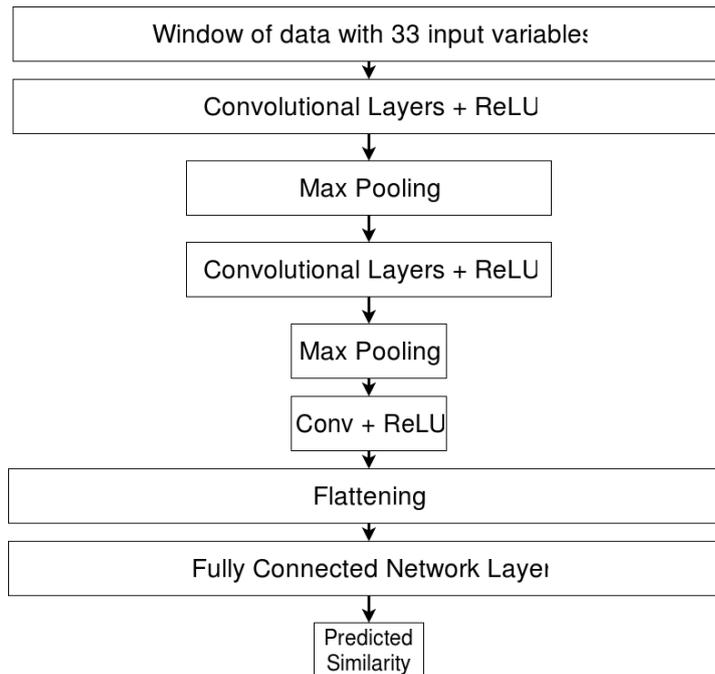


Figure 3.6: Illustration of the CNN architecture we used to estimate the similarity parameter for each window size.

### 3.4.4 Post-processing

The first panel of Figure 3.7 represents the expected similarity coded with the color bar computed for each of 100 CNNs (windows) during the period between the 8th of April 2012 and the 20th of July 2012. The raw predictions of the similarities (second panel) for each of these 100 CNNs are remarkably consistent with the expected similarities. However a small noise is present in this raw prediction, which is then smoothed using a median filter, and gives us the third panel of Figure 3.7.

It is worth noting already from this figure that the high predicted similarity values are very well localized in time on intervals very close to those of cataloged ICME, and there is a quite high contrast with the ambient solar wind background.

Additionally, the predicted similarity seems to be high for intermediate window sizes and lower for small and large windows, indicating the algorithm has well learned the typical length of the ICMEs from the training set. In addition, it is remarkable that ICMEs that are quite close from each other are not well separated by large windows, but often are separated by smaller ones. This is quite reasonable since, to the CNN, large windows features roughly look like a single ICME, whereas small windows actually have a chance to see data intervals that do not. The CNN here faces the same dilemma an observer would: "does an ICME cover the whole 100 hours, or are there two ICMEs, one closely following the other?". In these cases of true observational ambiguity, our 2D similarity map does not choose for the observer but rather provides a multi-scale suggestion.

The smoothed predicted similarity is also shown on the bottom panel of Figure 3.2 for an interval zoomed over a single ICME. Like the expected similarity, one can see a faster decrease of the similarity for small windows than for the large ones. It is worth noting that plotting the predicted similarity together with data is of great interest to very quickly and unambiguously identify visually ICMEs or periods of ICMEs, thereby shortening the usually long phases of data selection for observers. The first ICME of Figure 3.7 is a FN. Interestingly though, 2D maps still reveal a weak but coherent signature over the 100 - independent - CNN predictions, and that is visually detached from the ambient solar wind zero similarity. The penultimate predicted ICME is a FP even though the 2D maps reveal a strong signature for this event. The nature of these two types of events will be detailed in the next section.

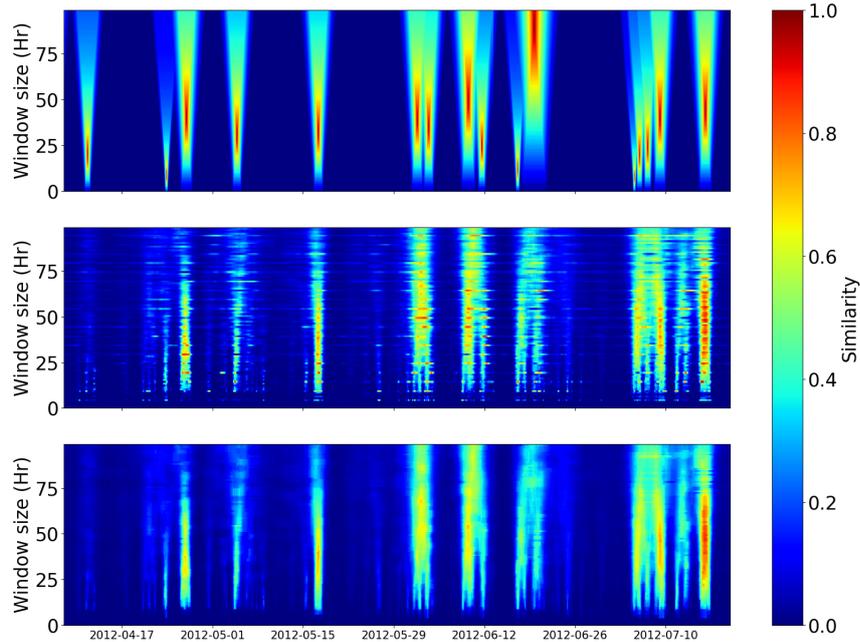


Figure 3.7: Three different representations of the similarity parameter during the period between 8 April 2012 and 20 July 2012. *Top* : Theoretical similarity computed from our prior ICME list, colored regions correspond to an actual ICME of our catalog. *Middle* : raw predictions obtained from each of our CNN. *Bottom*: Actual prediction of the pipeline after applying a median-filter to the raw predictions.

### 3.4.5 Automatization

The evaluation of the detection performance requires an automatization process that converts our predicted similarity into a list of predicted ICMEs. An important by-product of this process is the production of a reproducible and objective ICME catalog that can thus easily be updated incrementally with time without human intervention.

Because we only want to generate a list of start and end times, and because of noise in the window size dimension persisting after the application of the median filter, we reduce the predicted similarity to its time dependent integral along the window size axis, which defines the so-called *reduced similarity*.

We then regularize it by using a multiple-gaussian fit. An example of the fitted reduced similarity is shown as the green curve on Figure 3.8 for the period running from July 3rd 2012 to July 7th 2012.

A primary criterion is then applied to determine intervals within which ICME candidates may be searched. These intervals are defined as those for which the similarity exceeds a so-called *decision threshold*, shown as the dashed horizontal line in Figure 3.8.

Finally, a peak detection algorithm is applied to each of these intervals, from which the peak times and the half-height times will define our center, start and end times for predicted ICMEs.

To deal with the possible ambiguity that can exist for two close-by events, neighboring predicted ICMEs by less than two hours will be merged and considered as a single predicted event. This is the case for the fourth predicted ICME of Figure 3.8.

Finally, a predicted ICME having a duration of less than 2 hours is automatically considered as an inconsistent prediction since the CNN has never learned such a short duration event from our list, and is removed from the predicted list.

Predicted ICME intervals are represented on Figure 3.8 as green rectangles.

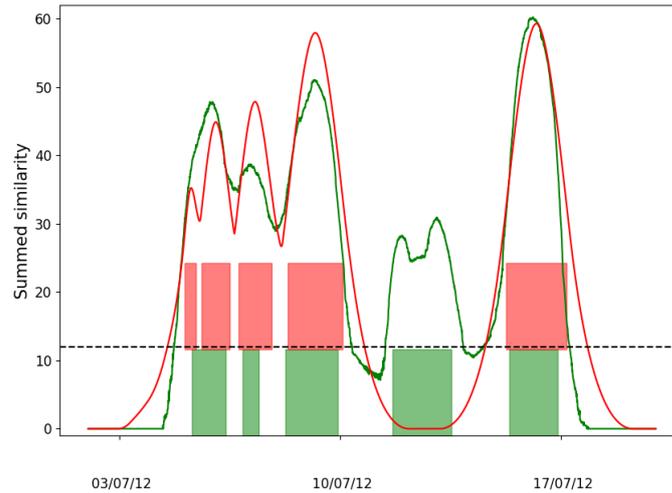


Figure 3.8: Expected (red) and predicted (green) reduced similarities for the period between the 3rd July 2012 and the 17th of July 2012. The red regions correspond to expected ICMEs from our catalog. Green regions correspond to predicted ICMEs after applying gaussian fitting and peak detection. The dashed line indicates the decision threshold (here equal to 12) we choose to make our prediction. The performances associated to this threshold correspond to the black dot of Figure 3.10.

### 3.5 Results

Having elaborated our pipeline, summarized on Figure 3.9, with the *training and validation set*, we can move on evaluating its performance by running it on the *test set*.

The performance of our method simply consists in comparing the predicted ICME list, obtained as explained above, with the ICMEs of the RL that are in the test period. It is important to remind that the RL, as any catalog, is not exhaustive and there are still ICME-like intervals never labeled in WIND data. Furthermore, time series represent a one dimensional slice into a non-stationary three-dimensional structure, therefore start and end times isolating events are based on an interpretation of the data. Labels do not represent an absolute truth, they vary from one expert to the other and one cannot expect any algorithm to outperform human in this subjective task. As a consequence, performance metrics are not perfect, their estimate cannot be expected to reach 100%.

#### 3.5.1 Precision and recall

Test ICME intervals are shown on Figure 3.8 as red rectangles. An existing (red) ICME is then considered as detected if more than 50% of its duration is overlapped by a predicted (green) ICME.

Due to the possible ambiguity that can exist in the transition from an ICME to a neighboring one, two ICMEs of our catalog are allowed to be detected by the same predicted ICME, which is the case for the first predicted ICME of Figure 3.8.

Computing the recall and precision for a continuously varying decision threshold gives us the evolution of the precision as a function of the recall, as represented in the precision-recall curve in Figure 3.10. The low values of this threshold allow weak similarity peaks to be seen as predicted ICMEs, thereby increasing FPs but decreasing FNs. Inversely, high decision thresholds result in less FPs but more FNs. The irregularities that are found on the curve, especially for high precision can be explained by the total number of predicted ICMEs that changes when we change our decision threshold. High recall and high precision regions are shown on Figure 3.10 as colored rectangles, the performances of our method in these regions will be detailed later-on. A perfect algorithm works at recall and precision 1. In practice it is rarely the case and observers will need to adjust the decision threshold so to maximize the recall, at the price of a smaller precision, or vice versa, depending on the objective.

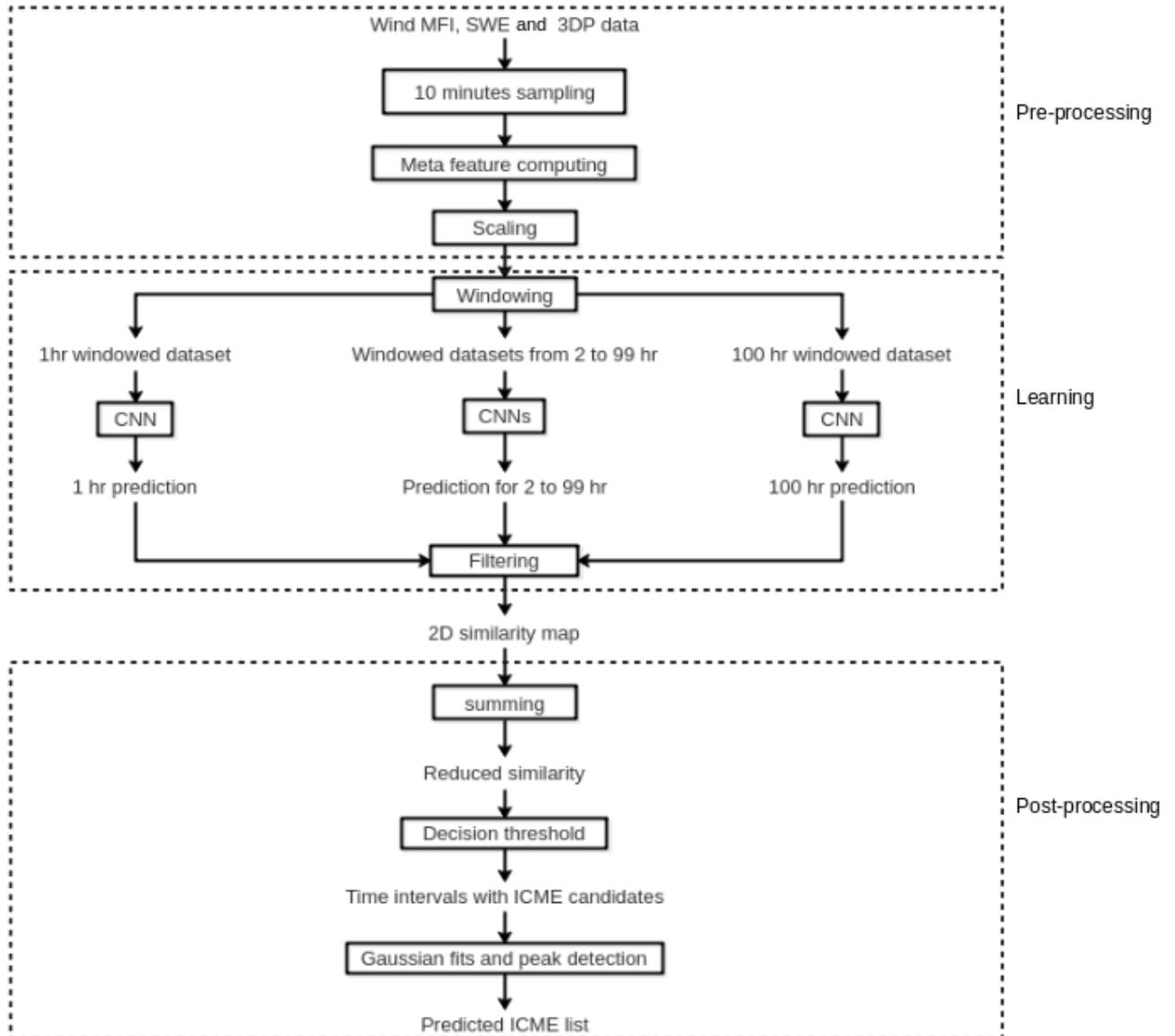


Figure 3.9: Scheme of our pipeline that converts raw data from WIND spacecraft into a generated ICME catalog, blocks with black contours represents operation lead on the data while blocks without contour represent the state of the dataset at the various steps of the pipeline. The dashed boxes indicate the different major steps of our pipeline.

Figure 3.10 shows recall and precision obtained in a previous work proposing automatic ICME detection [Lepping et al., 2005]. Because this method consists in fixing several arbitrary thresholds, it cannot easily produce precision recall curves, and only few points are to be compared with. Unlike it, our method comes with a single handle for the decision threshold, that moreover does not require prior knowledge of the physical nature of the events to be detected. Let us note here that at equivalent values of recall (resp. precision), our method leads to higher precision (resp. recall) and even reaches values of precision and recall that had not been reached by automatic identification methods yet.

To ensure the capacity our algorithm has to generalize on unknown data we also made a test of our whole pipeline on the training set and compared the performances of both predictions.

To do so, we compute the *average-precision* associated to the different precision-recall curves. Since our problem is limited by the ambiguity we have on the starting and ending times of the ICME signature combined with the non-exhaustivity of the RL, we do not expect such a high score. The average precision on our training set is 0.743 while it is 0.697 on the test set. As the training set has been used to set the characteristics of the filters of the CNN, it is not surprising to see a higher score on this prediction. The two scores are in the same order of value, which proves the capacity our algorithm has to generalize the knowledge about ICMEs it learned to unknown data.

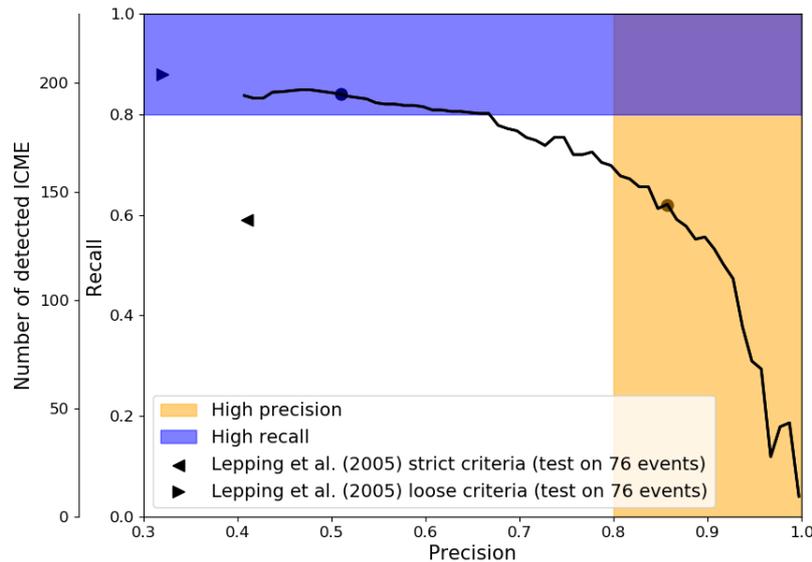


Figure 3.10: Precision-recall curve (black line) of our method. The region in blue indicates zones of high recall and fair precision where flux ropes and small ICME-like events can be detected. The region in yellow maps the high-precision zone for which the greatest part of the predicted ICMEs are in our catalog but do not represent it fully. The three markers indicate the performances of the previous attempts of ICME automatic identification : Lepping et al. [2005] strict (leftward triangle) and loose (rightward triangle) criteria. The black dot in the high precision region is associated with the decision threshold presented in Figure 3.8 and with the discussion of subsection 3.5.3. The black dot in the high recall region represents the working point used in the discussion of subsection 3.5.2.

### 3.5.2 High-recall region

The high recall region is represented by the blue rectangle in the Figure 3.10. In this region, our method detects the greatest part of the ICMEs present in the test set while generating a fair number of FPs.

To quantify the performances of our method, we selected the working point shown in Figure 3.10. At this point, 197 of the 232 ICMEs present in our test period are detected, which means a total recall of 84% and 160 of the 330 predicted ICMEs are considered as FPs, meaning a total precision of 51%.

The difference we notice between the number of detected ICMEs and the number of predicted ICMEs that are not FPs (170) comes from the fact that we allow neighboring ICMEs to be detected by the same predicted event.

For this decision threshold, 136 of the 150 ICMEs present in the list of Chi et al. [2016] were detected, which represents a recall of 91 % on this list. Likewise, we detected 100 of the 111 ICMEs present in the list of Nieves-Chinchilla et al. [2018] with a recall being equal to 90%. As our recall is close to its maximal value, the FNs we have correspond to ICMEs that will not be detected by this model whatever our decision threshold is. It is then in our interest to characterize these events. For the decision threshold we chose, we obtained 35 FNs, 10 were exclusive to the list of Chi et al. [2016], 7 came from the list of Nieves-Chinchilla et al. [2018], 3 were common to both lists and the 15 remaining came from the ICMEs we added after the different tests of our pipeline.

In the Figure 3.11, we represent the average reduced similarity of each ICME of the test set as a function of the time shift with the closest predicted ICME (that can also be a FP). The red points represent detected ICMEs while the blue ones represent the FNs. The size of each point is proportional to the associated ICME duration. The dashed line represents the decision threshold we chose to make our prediction. For a detected ICME, the closest predicted ICME is expected to be as close to the ICME as possible. This is why the time shift is usually low for detected ICMEs. The reason for which we can have more than 15 hours in the shift stands in the possibility we have to merge the predicted ICMEs. Additionally, it is not surprising to notice that the greatest majority of the detected ICMEs have an average reduced similarity above the decision threshold. Concerning the FNs, we can split them in two categories. On the one hand, 11 of them have an average reduced similarity below the decision threshold and can in fact be considered too weak in both their duration and pattern to be detected. On the other hand, all of the FNs with an average reduced similarity above the threshold are distant from a predicted ICME by less than 30 hours. Because of this proximity, the pipeline might not be able to distinguish the transition from an ICME to another and the prediction of the FN is then *absorbed* by its closest neighbor. Additionally, all of the FNs appear to have a short duration. Consequently, our FNs are short ICMEs that are either too weak to be detected, or close to another predicted ICME that entails their detection.

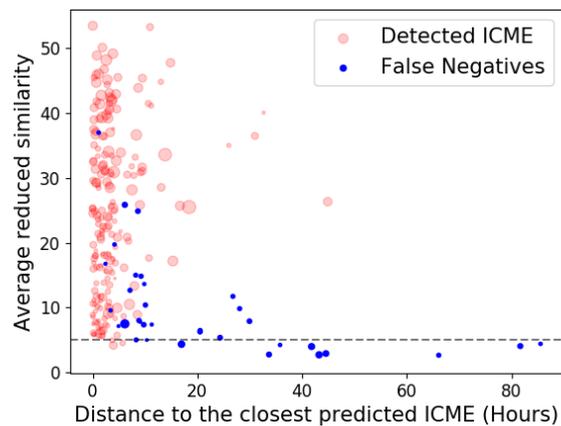


Figure 3.11: Average reduced similarity for each ICME of our test period as a function of the temporal shift to the closest predicted ICME. Red corresponds to detected ICME and blue corresponds to the FNs. The dashed line indicate the decision threshold we chose to make the prediction. The size of the circles corresponds to the duration of the event.

Similarly, a characterization of our FPs is necessary to understand the origin of the errors made by the model. Additionally, we remind that the RL cannot be considered as exhaustive, thus investigating FPs constitutes an opportunity to potentially discover ICMEs that had not been discovered yet.

Figure 3.12 represents the distribution of the ICMEs predicted by our pipeline according to their duration and the mean value of the reduced similarity during the event. As the FPs are a subset of these predicted ICMEs, the red bins then represent the predicted ICME that do contain one

or several ICMEs of our catalog. The possibility a predicted ICME has to cover several expected ICMEs is also at the origin of the difference we notice with the duration distribution shown in Figure 3.4. The left panel shows the great majority of the predicted ICMEs that have a duration below 20 hours are FPs. Similarly, the right panel shows that most of the ICMEs predicted with a low mean value of the reduced similarity are actually FPs. This confirms that augmenting the decision threshold is a good way to drastically reduce the number of FPs. Most of our FPs appear to have a short duration and a low mean value of the reduced similarity. An efficient way of ignoring them would then stand in the establishment of criteria according to these two parameters. However, as our catalog is assumed to be not exhaustive yet, some of these FPs might in fact be regions in the dataset where there could be one or several ICMEs that have not been discovered yet.

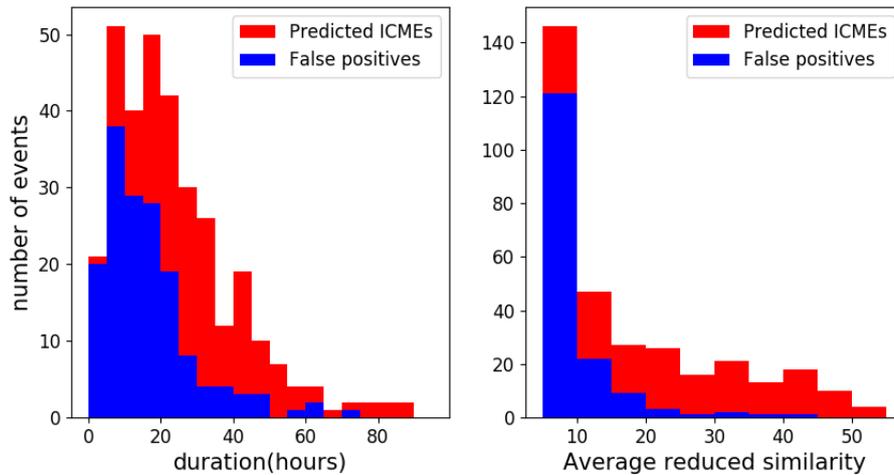


Figure 3.12: Distribution of the duration (left) and the mean value of the integral (right) for the predicted ICMEs (red) and the FPs (blue).

For these reasons, we inspected visually the 160 supposedly FPs and made a distinction between the ICME-like FPs, that did contained one or several time intervals that were susceptible of being actual ICME, and the non ICME-like FPs. A FP was considered as ICME-like if fitted at least three of the criteria used by Chi et al. [2016] to identify ICME manually. 102 of our 160 FPs were considered to be ICME-like while 58 others were considered to be non ICME-like.

Figure 3.13 represents the in-situ observation of the FP predicted by the models in the Figure 3.8 between the 11th of July 2012 and the 14th of July 2012 with the same panels than the one exposed in Figure 3.2. Between the two vertical solid lines that indicate the boundaries of the FP, one can see two ICME-like regions. This appears very clearly on the 2D maps for predicted similarity (bottom panel in Figure 3.13) with a single spot for high window sizes that splits into two for the low window sizes. In this case, the FP appear to have a large duration and a high mean value of the reduced similarity. One could then wonder if these two parameters could also serve as criteria to discriminate the ICME-like from the non ICME-like among the FP.

Figure 3.14 represents the distribution of the FPs predicted by our pipeline according to their duration and the mean value of the reduced similarity during the event. The blue bins represent the non ICME-like FPs while the green bins represent the ICME-like. Even if there is no real temporal discrimination visible on the left panel, FPs having a duration higher than 20 hr are likely to contain one or several ICME-like regions while the non-ICME like FPs usually have a short duration below 20 hr. Looking at the right panel, the reduced similarity then appears as a useful parameter to identify ICME-like FPs. Indeed, the great majority of the non ICME-like appear to have the lowest mean reduced similarity values while all but one FPs having the highest mean reduced similarity values are in fact ICME-like.

Setting our decision threshold in order to be in the high recall zone would then be useful to detect additional ICMEs in order to complete our current catalog. During the numerous trials of the pipeline, we then regularly checked the FPs predicted by our models in order to complete our

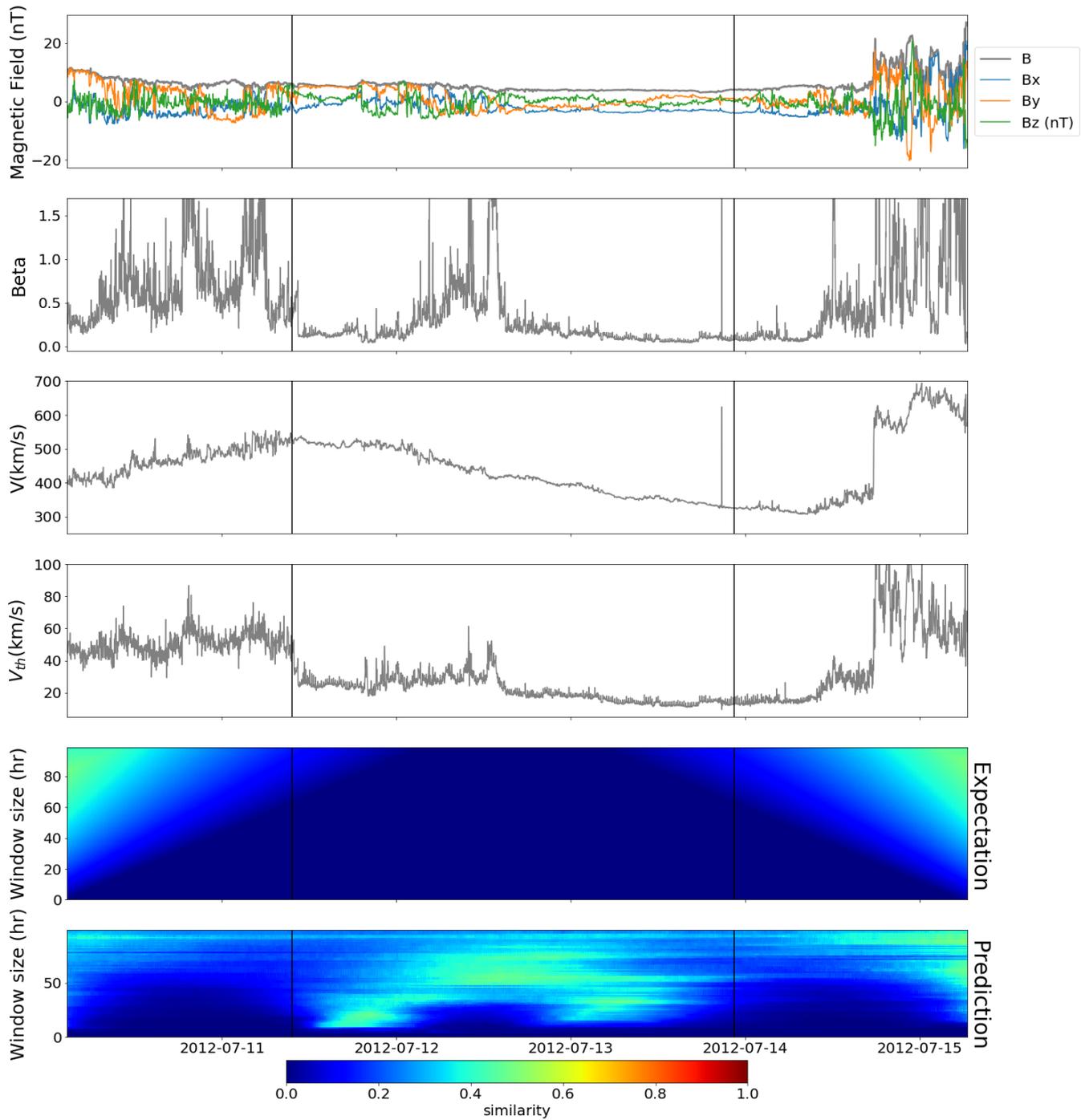


Figure 3.13: Solar wind observation of the FP predicted by our pipeline in the figure 3.8 that contains two ICME-like events. The solid vertical lines delimitate the boundaries of the predicted event. From the top to the bottom are represented : the magnetic field amplitude and components, the plasma  $\beta$ , the solar wind velocity, the thermal velocity, the similarity the ICME have with sliding windows of various sizes (from 1 to 100 Hr) and the similarity predicted by our method.

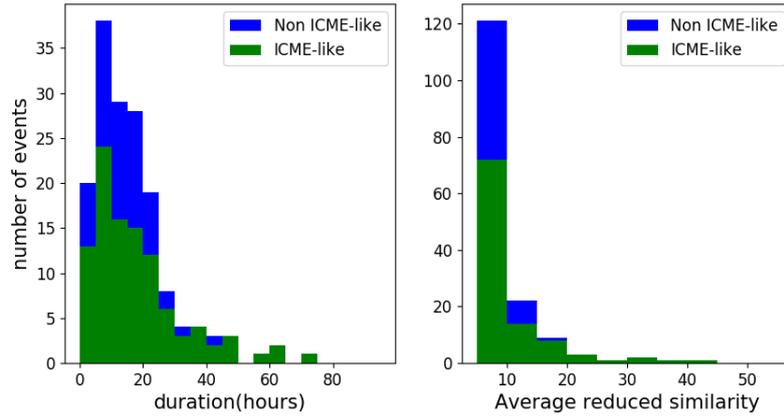


Figure 3.14: Distribution of the duration (left) and the mean value of the integral (right) for the acrsortFP depending if they are ICME-like (green) or not (blue).

catalog with potential new ICMEs. In order to find new ICMEs in the whole 1997-2015 period, predictions were also made on the 1997-2003 and 2004-2009 period by using the remaining period of the dataset for the training and the validation of our pipeline. This investigation led us to the ICME catalog that was presented previously.

### 3.5.3 High-precision region

The high precision region is represented by the yellow rectangle in the Figure 3.10.

In this region, the models generate an ICME list having a low number of FPs which ensures the consistency of the prediction. To quantify the performances of our method, we selected the working point shown in Figure 3.10 that corresponds to the decision threshold we used in the Figure 3.8.

At this point, 145 of the 232 ICMEs present in our test period are detected for a total recall of 62% and 25 of the 158 predicted ICMEs are considered as false positives for a total precision of 84%. For this decision threshold, 120 of the 150 ICMEs present in the list of Chi et al. [2016] were detected, which represents a recall of 80 % on this list. Likewise, we detected 82 of the 111 ICMEs present in the list of Nieves-Chinchilla et al. [2018] with a recall being equal to 74%. It is then worth noting that even with a high value of precision we still manage to detect the great majority of the previously detected ICMEs. The 87 FNs we obtained in this case were distributed as follows: 21 of them were exclusive to the list of Chi et al. [2016], 20 came from Nieves-Chinchilla et al. [2018], 9 were common to both list and 37 came from the ICMEs we discovered after different tests of our pipeline.

Following the distribution of the the average reduced similarity of the FNs in Figure 3.14, all but one of the 25 FPs we obtain, including the one represented in the Figure 3.13, will be considered as ICME-like and may contain one or several additional ICME we could add to our catalog.

By increasing our decision threshold, we ensure our pipeline will return ICMEs that have been predicted with high values of similarities just as the one shown in Figure 3.2. The generated predicted list is then supposed to contain easy-to-detect ICMEs that could be used for additional statistical study. To ensure it, we compared our predicted list to the ICMEs of our test period. Figure 3.15 shows the yearly occurrence frequencies of the two catalogs. As expected, the predicted list is shorter than the test list because the increased value of the decision threshold reduced the number of predictions. The two lists follow the same trend and both of them peak at the solar maximum in 2012 which is a first argument for the consistency of our predicted list.

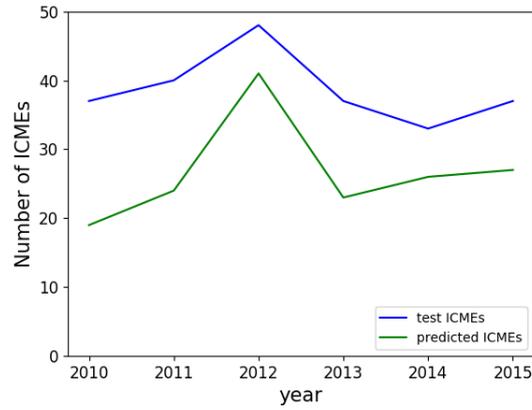


Figure 3.15: Yearly occurrence frequencies of ICMEs of our catalog during our test period (blue) and the list predicted by our pipeline (green) for the high precision point on Figure 3.10.

Figure 3.16 represents the distribution of the mean values of the magnetic field  $\langle B \rangle$ , the thermal velocity  $\langle V_{th} \rangle$  and the duration of our two catalogs, the left column corresponds to our test catalog while the right corresponds to our predicted list. Looking at the third row of subplots, the ICMEs we predict tend to be longer than the ICMEs of our catalog. This fact is partly due to the merge we did in our processing part as explained in the subsection 3.4.6. Nevertheless, the two first rows of subplots show similar distribution in magnetic field and thermal velocity. This confirms that our pipeline predicts consistent ICMEs that can be used by an external user for statistical studies.

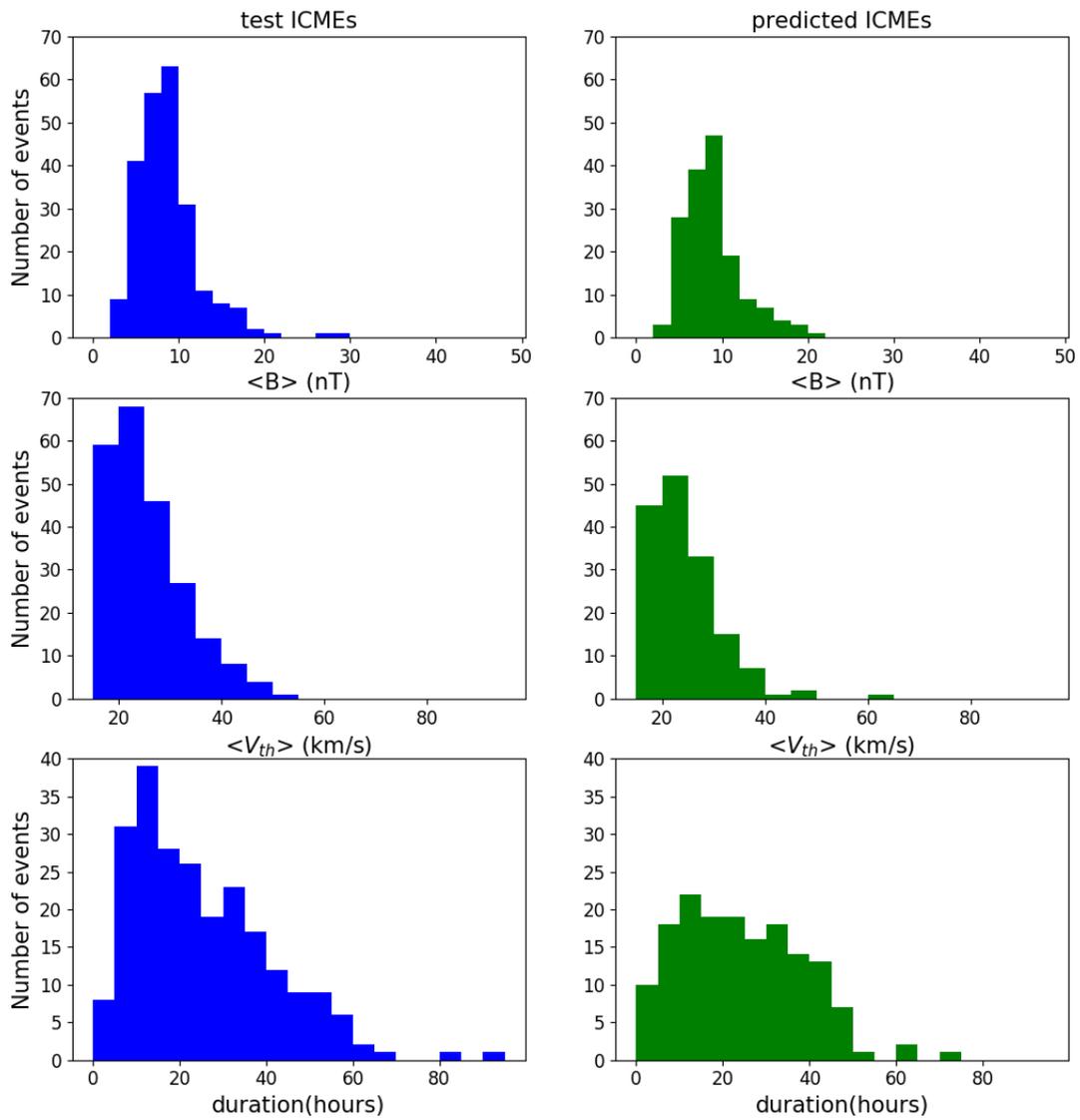


Figure 3.16: Distribution of the mean values of different parameters of the ICMEs in our test period (blue) and of the predicted ICMEs in our high precision region (green). From top to bottom are represented:  $\langle B \rangle$ ,  $\langle V_{th} \rangle$  and the duration.

## 3.6 Robustness

### 3.6.1 Importance of the various features as ICME indicators

To determine the relative importance of our different physical variables as important ICME indicators, we trained our pipeline on various configurations of our initial training set and compared the predictions that were made on our test set. The configurations we investigated are as follows :

- By considering solely the magnetic field magnitude and components data
- By considering the magnetic field data, the spectrogram and the  $\beta$
- By considering the proton fluxes only
- By considering the densities of protons and  $\alpha$  particles only

Figure 3.17 shows the similarity parameter which is expected (top panel) and predicted for the hundred windows from 1 to 100 hours in each of the above configurations on the same period between 8 April 2012 and 20 July 2012 as Figure 3.7. On the right part of the figure, the ICMEs that had been predicted with the strongest values of similarity are still detected with high values of similarity for the three first different arrangements of features. This proves the ability of our method to detect an ICME with missing parameters which is to say when data from an instrument are not available. Surprisingly, the lone measurements of densities can provide a fair detection of ICMEs despite the enhanced noise in this case. The prediction based on the densities values is even the only combination of features that detects the third ICME of Figure 3.17 apart from the detection based on our complete dataset. Similarly, the prediction on the lone magnetic field components and amplitude is the only arrangement of features that detects the fourth ICME.

As for visual detection of ICMEs, the magnetic field seems to play a key role in the CNN's learning. The possibility of detecting an ICME by using a specific set of features rather than an other is consistent with the possibility ICMEs have to partially fulfill the criteria generally used to detect them manually [Zurbuchen and Richardson, 2006]. The fact that no ICME during the period of Figure 3.17 detected by one of the subsets of features has not been also detected by our complete dataset indicates the importance each feature has in the characterization of the specific signatures of ICMEs as well as the importance of considering them altogether.

To understand the impact of removing features, we compute the precision-recall curves for each configuration that are shown in Figure 3.18. The average precision is then computed for each dataset configuration, these values are shown in Table 3.1. Unsurprisingly, the highest value of this area is obtained for the complete dataset while the predictions based on the proton fluxes and the one based on the densities have the lowest scores. This confirms the interest we have in considering the most complete set of features. The high values of the area obtained for the predictions based on the magnetic field,  $\beta$  and the proton fluxes indicate the major importance these features have on the automatic detection of ICMEs.

Dataset features	Average precision
All	0.697
$B, B_x, B_y, B_z$	0.593
$B, B_x, B_y, B_z, \beta$ and proton fluxes	0.621
Proton fluxes only	0.486
$N_p, N_{p,nl}, N_{\alpha,nl}$	0.334

Table 3.1: Areas under the precision recall curve for different dataset configurations, higher values indicate more efficient detection.

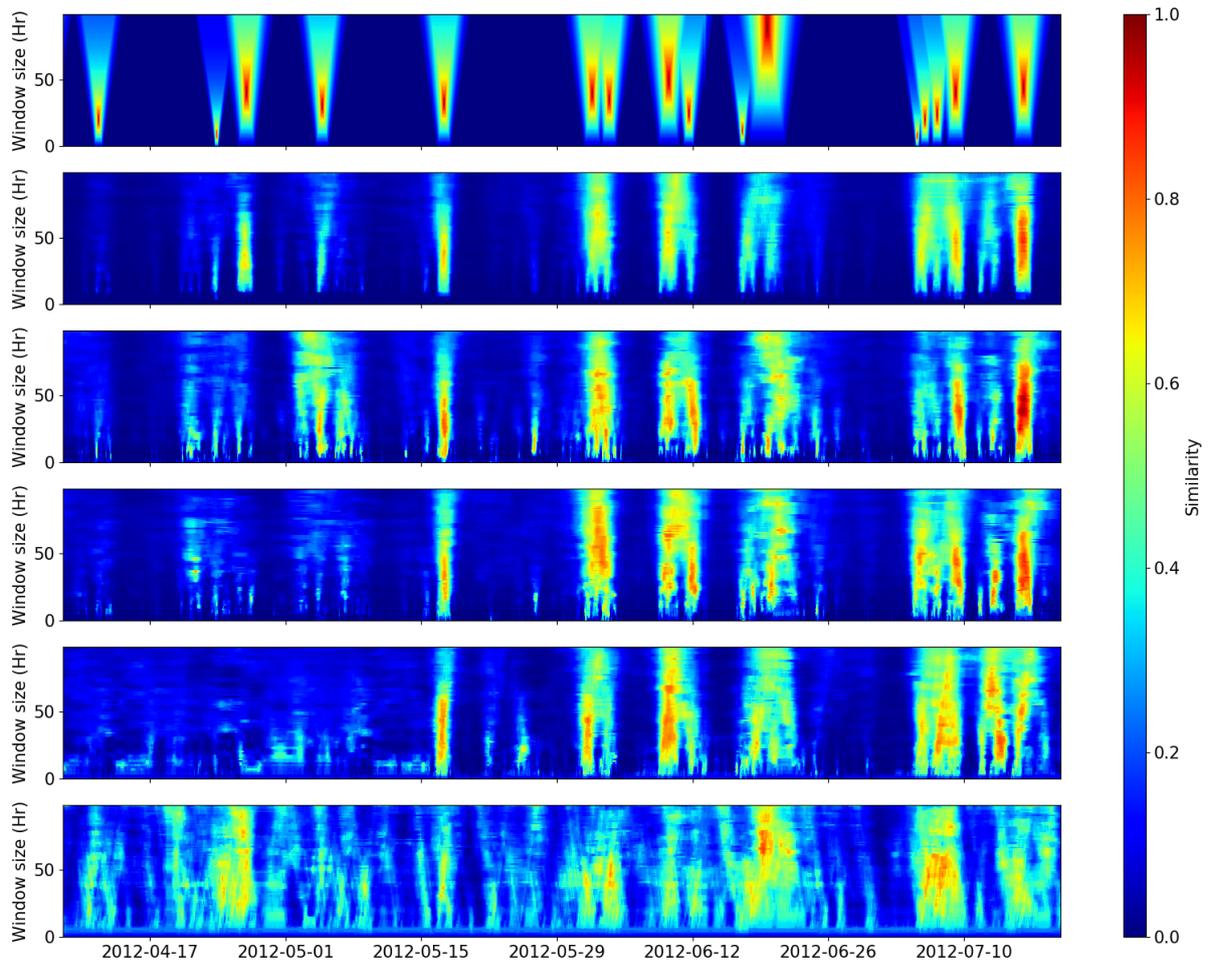


Figure 3.17: Estimation of the similarity parameter for different sets of features during the period between the 3rd of July 2012 and the 17th of July 2012. From top to bottom are represented : the expected similarity, the prediction with the complete dataset, the prediction based only on the magnetic field and its components, the prediction based on the magnetic field, the plasma parameter  $\beta$  and the proton fluxes, the prediction based on the lone proton fluxes and the prediction based on the lone measures of the protons and  $\alpha$  particles.

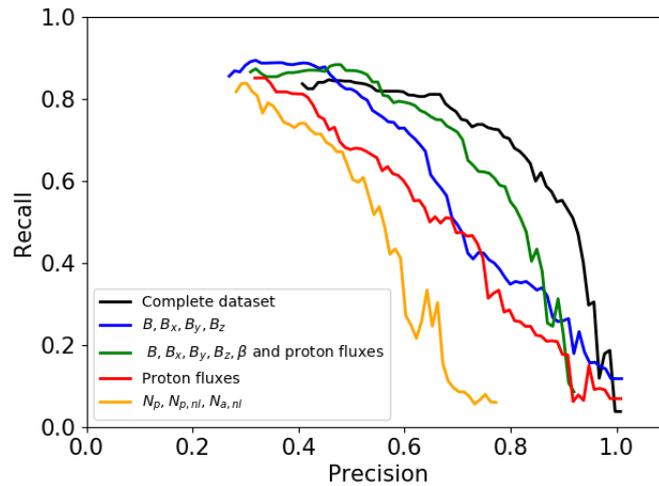


Figure 3.18: Precision recall curves of our method using different dataset features : Our complete dataset (black), the lone magnetic field information (blue), the magnetic field, the plasma  $\beta$  and the proton fluxes (green), the proton fluxes only (red) and the proton and  $\alpha$  particles densities only (yellow).

### 3.6.2 Influence of the number of ICMEs in the training period

Statistical learning methods, in particular deep learning algorithms, often need a large quantity of data to give meaningful predictions. In this section, we investigate the influence of the number of ICMEs in our training period and its impact on the general performances of the method. We thus progressively reduce the length of our training set by removing periods of data and thus ICMEs. This approach also allows us to investigate the real interest of considering a whole solar cycle of data during our training phase. Like in the previous subsection, we compute the average precision for each training period.

The evolution of this value as a function of the number of ICMEs in the training period is shown in Figure 3.19. The different tests we made are indicated with a cross x. The more we add ICMEs in our training period, the higher our average precision, which is consistent with the necessity of having a large quantity of data. However, the average precision starts with a sharp increase and evolves rapidly towards a weak inclination. The sharp increase indicates that a very low number of ICMEs are needed in order to reach fair performances.

Surprisingly, it is even possible to start detecting ICMEs with the knowledge given by a single event as shown by the second point of the Figure 3.19. It is very interesting to note that learning from few ICMEs with the complete set of features give similar performances than the whole training period with only particle densities or proton flux. Even if the progression is slower later-on, additional ICMEs keep improving the performances and it is then worth taking them into account.

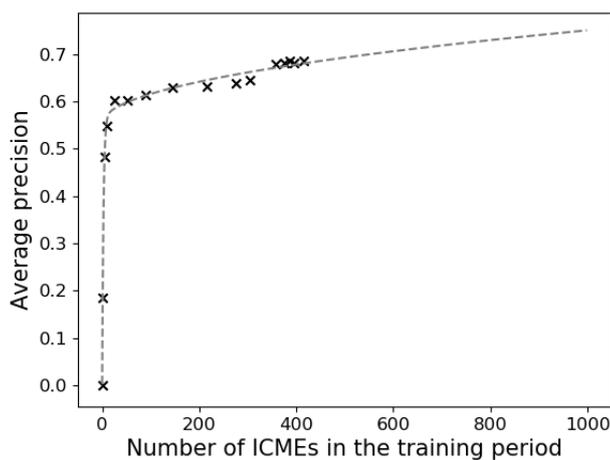


Figure 3.19: Average precision of our pipeline as a function of the number of ICMEs present in the training period we considered, the tests we made are represented by the crosses. The grey dashed line represent the estimation of what the average precision would be if additional ICMEs were added to our training period.

From then on, we could estimate the number of ICMEs we would need in order to reach a certain level of performances. This estimation is shown with the gray dashed line in the Figure 3.19. At first sight, as many ICMEs as what we currently have would be needed for an increase in average precision by 10 %. However, this expected number of ICMEs could be easily increased by completing our list with the FPs that appeared to be ICME-like and by extending our dataset period to the years before October 1997 and after 2015.

Additionally, we showed in previous subsections the capacity our pipeline had to predict ICMEs with one or several missing features. Coupled with the diversity of spacecraft that have been providing in-situ measurements of ICMEs for the past 22 years (STEREO, Helios, Ulyse, ACE,...), one could perfectly imagine a dataset composed of the in-situ measurements provided by various spacecraft standardized in order to have consistent features and sampling for each spacecraft.

This would increase drastically the number of given ICMEs and thus the performance of our pipeline provided a good compromise between the different instruments and products of each mission is found.

### 3.6.3 Influence of the training, validating and testing period

As mentioned in the subsection 3.5, we used our pipeline on three different training, validating and testing period. Changing the attribution of our three datasets period also has the interest of investigating the influence of the training period and the changes in WIND trajectory from 1997 to 2016 on our prediction. To do so, we changed the periods of our training, validation and test set as follows:

- Training period from the 1st of April 2004 to the 31st of December 2015, validation between the 1st of January 2004 and the 30th of March 2004 and test on the 1997-2003 period
- Training on the 1997-2003 period and from the 1st of April 2010 to the 31st of December 2015, validation between the 1st of January and the 30th of March 2010 and test on the 2004-2009 period

For each distribution, the average precision on the prediction made on the test set is shown on the Table 3.2. In the three cases we find similar values, which is consistent with the number of ICMEs (425, 530 and 353 respectively) contained in each training period and the average precision estimation provided by Figure 3.19. This confirms the importance of the diversity and of the number of ICMEs seen by the CNN during the training phase on the quality of the prediction made by our pipeline.

Considering these three values, the mean average precision for our pipeline is then  $0.694 \pm 0.003$ . Additionally, finding similar values for the three periods indicate that the changes in WIND trajectory especially during the 1997-2003 do not affect our results.

Testing period	Average precision
2010-2015	0.697
2004-2009	0.690
1997-2003	0.694

Table 3.2: Areas under the precision recall curve for different testing periods

## 3.7 Global quality of the prediction

The ambiguity that exists in the definition of the starting and ending times of an ICME combined of the non-exhaustivity of the different observers lists tends to limit the overview an event-based score would give on the quality of the detection made by our pipeline. It is thus interesting to also quantify to what extent the predicted list is globally similar to our list, and compare this global similarity to those of various independent expert lists covering the same time period.

For each prediction period, we computed the Jaccard index, defined in the chapter 2, for each decision threshold on our reduced similarity and represented the evolution of this index as a function of the temporal size of the list predicted by our pipeline in Figure 3.20.

High (resp. low) values of the total duration of the predicted list will correspond to a low (resp. high) decision threshold as the predicted list in this case will contain more (resp. less) events. The low value of the Jaccard in these cases is then mainly due to the important number of FPs (resp. FNs). For each of our prediction period, we notice a similar evolution of the Jaccard that peaks around the temporal size of the reference list represented by the vertical dashed line (that will vary with the considered period according to Figure 3.3). This proximity can be understood as the peak will correspond to the best compromise we can find between a high recall and a high precision.

To compare the quality of our prediction regarding the global similarity that exists between different expert lists, we computed the Jaccard between lists that had the same number of events in one of the three prediction periods that we considered (e.g Chi et al. [2016]; Nieves-Chinchilla et al. [2018]; Richardson and Cane [2010] and the RL for the three periods and Jian et al. [2006] for

the 1997-2003 and the 2004-2009 periods). The min-max interval for the different values of the Jaccard we found is represented on Figure 3.20 by the gray zone. The values we find that barely exceeds 50% find their origin in the non-exhaustivity of the lists and the ambiguity that exists in the definition of the temporal boundaries of the ICMEs. Below this interval, the generated list contains too many FPs or not enough events to give a significant insight about ICMEs.

Inside and above the interval, the list generated by our pipeline is as or more similar to our reference list than the human made lists between them. This is where the generated ICME lists can be used for further work such as the detection of additional events or statistical studies.

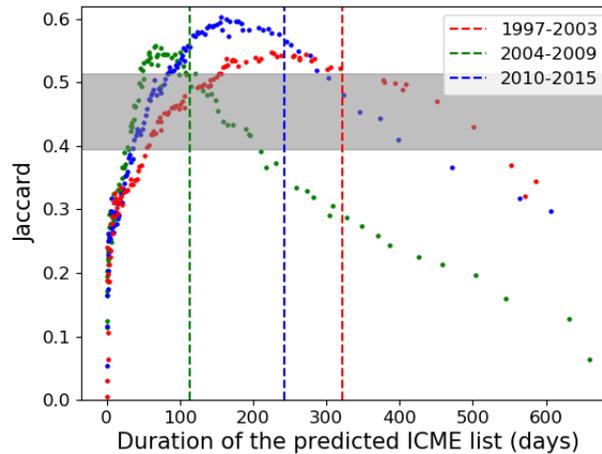


Figure 3.20: Jaccard between our ICME list and the generated ICME list as a function of the total duration of the generated ICME list (in days) for each prediction period: 1997–2003 (red), 2004–2009 (green), and 2010–2015 (blue). The vertical dashed lines represent the total duration of our list in each considered period. The gray line represents the confidence interval we have on the Jaccard between human-made lists.

In the three cases, a non negligible part of the Jaccards are inside of above the typical expert list Jaccards lying in the gray zone. This proves that the lists generated by our pipeline are as globally similar to the RL as experts lists are to one another, and can then be used, either for further detections or statistical studies.

### 3.8 Conclusion

Using Convolutional Neural Networks that estimated a similarity parameter for windows of data of various sizes, from 1 to 100hr, and a post processing method based on peak detection, we developed a pipeline that provides an automatic ICME detection from the WIND spacecraft in-situ measurements. The 2D-similarity map the pipeline returned and that is shown in Figure 3.7 provides an interesting visual indicator of zones of interest for an external user particularly in the case of neighbored ICMEs or multiple events with various duration<sup>3</sup>.

Our pipeline also has the ability to generate generic and reproducible ICME catalogs with a precision and a recall that has not been reached yet. From a Jaccard point of view, the list we predict are as comparable to our RL as two experts lists are together. Depending on the decision threshold we set on our detection, the pipeline offers the possibility to detect additional ICMEs (high-recall case) or to generate consistent and reproducible ICME catalogs that could be used for further statistical study (high-precision case).

From the insight we had on the FNs in the high recall region, we showed that the ICMEs that are never detected by our pipeline are either short and too weak to be detected or too close to another predicted ICME to be distinguished by the pipeline as a separate one. Up to now, a total of 148 additional ICMEs have been detected and were added to our WIND ICME catalog. Nevertheless, our catalog is not exhaustive yet and there are still ICMEs that have not been discovered yet.

<sup>3</sup>Additional prediction examples of our pipeline can be found in the appendix B.

Additional runs of the pipeline shall be needed in order to establish a consistent ICME catalog as much exhaustive as possible.

By testing our pipeline on datasets with missing features, we proved that even if the prediction was altered, our pipeline still has the capacity to detect ICMEs. On the first hand, this proves that our pipeline can be used even when one or several instruments of the spacecraft are defective in order to maintain a continuous prediction of ICMEs. On the other hand, this also proves that our pipeline can easily be adapted to the data of other spacecraft that have been measuring ICMEs over the past 22 years in different places of the solar system (Cluster, ACE, Stereo, Helios, Ulysse, Venus Express...) and for which measured features might be missing.

The influence of the number of ICMEs being present in the training set has been investigated by training our pipeline with reduced datasets. Even if a few number of ICMEs is enough to detect events properly, a large number of additional events is required if we want the quality of the prediction to improve significantly. These additional ICMEs could be added by considering additional training period such as the 2016-2018 period, by looking more precisely at the ICME-like FP that were found by our pipeline or even by considering the data provided by other spacecraft. Another way we would have to improve our performances would stand in the fine tuning of the parameters of the CNN we used to make our prediction.

The prediction of ICMEs has been established without giving to the algorithm any initial knowledge on ICMEs. The presence of an ICME in a given window of data being indicated to the algorithm through the similarity that only depends on the event temporal boundaries. This indicates the adaptability of our pipeline which could be used to detect other phenomena likely to be measured by spacecraft such as the sheaths of ICMEs, Co-rotating Interaction regions or Stream Interaction regions.

Finally, we established our method starting with an ambiguous input ICME list and proved their capability to generate events list containing the greatest part of events already detected by human observers while adding FPs that could be considered as actual events by an additional external observer. These studies made on the FPs and FNs of our pipeline therefore show the necessity, whenever the label is ambiguous, to go beyond the strict values of the used evaluation metrics. And the underlying importance of the insight given on the FPs and FNs of our method in the frame of massive event detection. A very similar insight on the errors made by our detection algorithms will be given in the next chapters where we will encounter the ambiguity issue again in both the problem of the automatic detection of magnetopause crossings and the automatic detection of magnetic reconnection jets.

### 3.9 Bibliography

Bengio, Y.: Practical recommendations for gradient-based training of deep architectures, CoRR, abs/1206.5533, URL <http://arxiv.org/abs/1206.5533>, 2012. 51

Burlaga, L., Sittler, E., Mariani, E., and Schwenn, R.: Magnetic loop behind an interplanetary shock - Voyager, Helios, and IMP 8 observations, *Journal of Geophysical Research*, 86, 6673–6684, <https://doi.org/10.1029/JA086iA08p06673>, 1981. 44

Chi, Y., Shen, C., Wang, Y., Xu, M., Ye, P., and Wang, S.: Statistical Study of the Interplanetary Coronal Mass Ejections from 1995 to 2015, *solphys*, 291, 2419–2439, <https://doi.org/10.1007/s11207-016-0971-5>, 2016. 45, 48, 49, 57, 58, 60, 66

Gosling, J. T., Pizzo, V., and Bame, S. J.: Anomalously low proton temperatures in the solar wind following interplanetary shock waves—evidence for magnetic bottles?, *Journal of Geophysical Research*, 78, 2001, <https://doi.org/10.1029/JA078i013p02001>, 1973. 44, 45

Grover, P.: 5 Regression Loss Functions All Machine Learners Should Know, <https://heartbeat.fritz.ai/5-regression-loss-functions-all-machine-learners-should-know-4fb140e9d4b0>, 2019. 51

- Jian, L., Russell, C. T., Luhmann, J. G., and Skoug, R. M.: Properties of Interplanetary Coronal Mass Ejections at One AU During 1995–2004, *solphys*, 239, 393–436, <https://doi.org/10.1007/s11207-006-0133-2>, 2006. 45, 48, 66
- Juszczak, P., Tax, D. M. J., and Duin, R. P. W.: Feature scaling in support vector data description, 2002. 50
- Kiefer, J. and Wolfowitz, J.: Stochastic Estimation of the Maximum of a Regression Function, *Ann. Math. Statist.*, 23, 462–466, <https://doi.org/10.1214/aoms/1177729392>, URL <https://doi.org/10.1214/aoms/1177729392>, 1952. 51
- Kilpua, E., Koskinen, H. E. J., and Pulkkinen, T. I.: Coronal mass ejections and their sheath regions in interplanetary space, *Living Reviews in Solar Physics*, 14, 5, <https://doi.org/10.1007/s41116-017-0009-6>, 2017. 44, 45
- Klein, L. W. and Burlaga, L. F.: Interplanetary magnetic clouds at 1 AU, Tech. rep., 1981. 44, 45
- Lepping, R. P., Wu, C.-C., and Berdichevsky, D. B.: Automatic identification of magnetic clouds and cloud-like regions at 1 AU: occurrence rate and other properties, *Annales Geophysicae*, 23, 2687–2704, <https://doi.org/10.5194/angeo-23-2687-2005>, 2005. 47, 56
- Lepping, R. P., Berdichevsky, D. B., Wu, C. C., Szabo, A., Narock, T., Mariani, E., Lazarus, A. J., and Quivers, A. J.: A summary of WIND magnetic clouds for years 1995–2003: model-fitted parameters, associated errors and classifications, *Annales Geophysicae*, 24, 215–245, <https://doi.org/10.5194/angeo-24-215-2006>, 2006. 45, 48
- Mitsakou, E. and Moussas, X.: Statistical Study of ICMEs and Their Sheaths During Solar Cycle 23 (1996 - 2008), *solphys*, 289, 3137–3157, <https://doi.org/10.1007/s11207-014-0505-y>, 2014. 45
- Nguyen, G., Aunai, N., Fontaine, D., Pennec, E. L., den Bossche, J. V., Jeandet, A., Bakkali, B., Vignoli, L., and Blancard, B. R.-S.: Automatic Detection of Interplanetary Coronal Mass Ejections from In Situ Data: A Deep Learning Approach, *The Astrophysical Journal*, 874, 145, <https://doi.org/10.3847/1538-4357/ab0d24>, 2019. 44
- Nieves-Chinchilla, T., Vourlidas, A., Raymond, J. C., Linton, M. G., Al-haddad, N., Savani, N. P., Szabo, A., and Hidalgo, M. A.: Understanding the Internal Magnetic Field Configurations of ICMEs Using More than 20 Years of Wind Observations, *solphys*, 293, 25, <https://doi.org/10.1007/s11207-018-1247-z>, 2018. 45, 48, 57, 60, 66
- Ojeda-Gonzalez, A., Mendes, O., Calzadilla, A., Domingues, M. O., Prestes, A., and Klausner, V.: An Alternative Method for Identifying Interplanetary Magnetic Cloud Regions, *Astrophysical Journal*, 837, 156, <https://doi.org/10.3847/1538-4357/aa6034>, 2017. 47
- Richardson, I. G. and Cane, H. V.: Near-Earth Interplanetary Coronal Mass Ejections During Solar Cycle 23 (1996 - 2009): Catalog and Summary of Properties, *solphys*, 264, 189–237, <https://doi.org/10.1007/s11207-010-9568-6>, 2010. 45, 48, 66
- Shinde, A. A. and Russell, C. T.: What Defines an Interplanetary Coronal Mass Ejection?, in: *AGU Fall Meeting Abstracts*, vol. 2003, pp. SH21B–0133, 2003. 44
- Yang, J., Nguyen, M. N., San, P. P., Li, X., and Krishnaswamy, S.: Deep Convolutional Neural Networks on Multichannel Time Series for Human Activity Recognition, in: *IJCAI International Joint Conference on Artificial Intelligence*, vol. 2015-January, pp. 3995–4001, 2015. 51
- Zurbuchen, T. H. and Richardson, I. G.: In-Situ Solar Wind and Magnetic Field Signatures of Interplanetary Coronal Mass Ejections, p. 31, [https://doi.org/10.1007/978-0-387-45088-9\\_3](https://doi.org/10.1007/978-0-387-45088-9_3), 2006. 45, 49, 63

## Chapter Summary

- In this chapter, we use CNNs on sliding windows and peak detection, to provide a fast, automatic and multi-scale detection of ICMEs in the in-situ data measurement provided by WIND.
- The pipeline we developed returns 2D-similarity maps that provide visual indications about the zones of interest in the data for an external observer labeling the data manually.
- Combining these maps to peak detection, our pipeline has the ability to generate generic and reproducible ICMEs catalogs with a precision and recall that passes the performance of the other existing detection method based on manual, empirical thresholds.
- Depending on the decision threshold we set for the production of the catalogs, they can be used either to detect additional ICMEs, either to study these events from a statistical point of view.
- Although less inaccurate, the method also works with one or several parameters and improves its performances by increasing the amount of input data.
- From a Jaccard point of view, the list we generate is as comparable to the list used to evaluate the algorithm than the lists manually elaborated by two different external observers.
- As the method uses no particular physical knowledge about ICMEs, the elaboration of our method paves the way for the automatic detection of the other large-scale events measured by solar monitors.

# Chapter 4

## Automatic classification of the three near-Earth regions

*Le rêve de celles qui m'aimaient  
Est ma véritable amante*

---

Anonyme

### Contents

---

<b>4.1 Introduction</b>	<b>72</b>
<b>4.2 Data</b>	<b>73</b>
4.2.1 THEMIS	73
4.2.2 Double Star	74
4.2.3 MMS	74
4.2.4 Cluster	74
4.2.5 Multi-spacecraft datasets	75
<b>4.3 Labeling THEMIS data</b>	<b>75</b>
<b>4.4 Algorithm selection</b>	<b>78</b>
<b>4.5 Algorithm performance</b>	<b>79</b>
4.5.1 Temporal dependance	79
4.5.2 Influence of the manual labeling	79
<b>4.6 Adaptability of the model: from a mission to the other</b>	<b>80</b>
4.6.1 Double Star	80
4.6.2 MMS	82
4.6.3 Cluster	84
4.6.4 ARTEMIS	86
<b>4.7 Comparison with manually set thresholds</b>	<b>89</b>
<b>4.8 Massive detection of boundary crossings</b>	<b>90</b>
4.8.1 Magnetopause catalog	90
4.8.2 Bow shock catalog	93
<b>4.9 Conclusion</b>	<b>94</b>
<b>4.10 Bibliography</b>	<b>95</b>

---

## 4.1 Introduction

At first order, the magnetopause and the bow shock are the boundaries of three distinct regions of the near-Earth environment: the magnetosphere, the magnetosheath and the solar wind.

By definition, the shape and location of these boundaries do depend on the upstream solar wind conditions [Fairfield, 1971].

The ever-growing quantity of near-Earth in-situ data allowed the realisation of statistical studies dedicated to the position, shape and dynamics of both the magnetopause (Paschmann et al. [2018], Němeček et al. [2020] and references therein) and the bow shock (Kruparova et al. [2019] and references therein). Following these studies comes the task of their modelling and numerous are the existing magnetopause (Lin et al. [2010]; Shue et al. [1997]; Wang et al. [2013] and references therein) and bow shock (Farris and Russell [1994]; Jeřáb et al. [2005] and references therein) analytical or numerical models [Liu et al., 2015].

The first step of both empirical modelling and statistical studies is always the same and has already been mentioned in the chapter 1: establishing a consistent catalog of boundary crossings from the streaming in-situ data provided by missions of interest. Just like what was seen in the previous chapter, this appears to be a time-consuming, ambiguous and poorly reproducible task that should be automatized.

Nevertheless, the problem here is slightly different from detecting the beginning and ending date of an event, as seen with the ICMEs, as it consists here in finding the transition from a certain region in the data to another. In the former, an ICME is a spatial structure moving in a specific medium, the solar wind, that is measured when its trajectory encounters a spacecraft. The beginning and ending dates of the event then correspond to the entry and the exit of the structure by the spacecraft. In the latter, the before and the after of a crossing correspond to stable conditions reached on both sides of the boundary. In this context, no proper beginning or ending time can be defined objectively and the lone crossing of the boundary is not enough to provide information on the structure of the boundary. The method we presented in the previous chapter is consequently not adaptable here and the simplest way to achieve it then stands in training a model to classify the three near-Earth regions and finding the transitions intervals.

The in-situ measurements made by a spacecraft that went through these three regions are shown in Figure 4.1 where are represented from top to bottom the proton density, the magnetic field components, the velocity components and the omnidirectional energy fluxes of ions measured by THEMIS. The last panel will be explained in the following sections. This representation illustrates the typical values of these physical parameters in each of the region that we presented in chapter 1. The three regions are easily distinguishable by eye and the first method we could think about in this classification task would be to use manually set thresholds. Using the data provided by the five THEMIS spacecraft coupled with the solar wind conditions provided by WIND, Jelínek et al. [2012] established a method based on thresholds on the magnetic field amplitude  $B$  and the proton density  $N_p$  normalized by the IMF amplitude and proton density. They used this method to identify the three near-Earth regions and eventually build crossings lists from this classification. All the principle of the method then consists in manually setting the two straight lines that best separate the three regions in the  $(N_p, B)$  plane in a similar way than what is shown in Figure 4.12. Nevertheless, this still requires the manual setting of thresholds on a reduced number of parameters and there is additionally no guarantee on how well they will do on an unknown set of data and the separability of these two features presented here is not guaranteed on the whole magnetopause, especially in the case of nightside, flanks or high latitude boundary crossings. The method could thus be improved with additional features such as the amplitude of the ion bulk velocity or the ion temperature but this would lead to the establishment of manual thresholds in a  $N$ -dimensional space, which is usually a tricky task when done manually. Moreover, the best value found for these thresholds was not shared. The exploitation of this method by an external user thus requires to start the threshold setting all again from scratch.

Once again, machine learning appears as an interesting way to improve this automatization. The objective of this chapter is then to elaborate a machine learning based method that auto-

matically classifies the three near-Earth regions by looking at the magnetic field and the plasma moments provided by spacecraft of different missions. This method shall then be used to automatically elaborate massive and reproducible boundary crossings catalogs that will find their usefulness in the two next Chapters.

After presenting the data of the different missions we are concerned about, most of them being also used in the two next chapter, and the associated labels, we present the algorithm we use and why we choose it. We then evaluate its performances, investigate its adaptability to various missions: Double Star, [MMS](#), Cluster and [ARTEMIS](#). The performances of our method are then compared to a manually-set threshold method and is finally used to automatically elaborate boundary crossings catalogs.

## 4.2 Data

### 4.2.1 THEMIS

The [Time History of Events and Macroscale Interactions during Substorms \(THEMIS\)](#) mission is a NASA mission that was launched in February 2007. The mission consists of 5 identical spacecraft (from A to E) each measuring ions and electrons distributions between  $\sim 5$  eV and  $\sim 1$  MeV and computing the associated moments [[Angelopoulos, 2008](#)]. The five spacecraft have an equatorial orbit with a slowly rotating apogee that allows the mission to provide a complete sweeping of the dayside, the nightside, the dawnside and the duskside of the magnetosphere.

[THEMIS](#) B and C were the spacecraft with the largest apogee (respectively 30 and 20 Re) and frequently went through the bow shock while [THEMIS](#) A, D and E apogees (respectively 12, 12 and 10 Re) are more likely to provide an orbit tangential to the magnetopause. After February 2010, [THEMIS](#) B and C were inserted into the lunar orbit and became the [Acceleration Reconnection Turbulence & Electrodynamics of Moon's Interaction with the Sun \(ARTEMIS\)](#) mission.

We use the data between the 1<sup>st</sup> of March 2007 and the 1<sup>st</sup> of July 2019 and the data provided by [THEMIS](#) B and C until the 1<sup>st</sup> of January 2010. Past the latter period and until the 1<sup>st</sup> of July 2019, the data we have from the spacecraft B and C will constitute the [ARTEMIS](#) dataset we will especially use in the next section.

In both cases, the magnetic field measurements are provided by the Fluxgate Magnetometer (FGM, [Auster et al. \[2008\]](#)) while the plasma moments are provided by the Electrostatic Analyzer (ESA, [McFadden et al. \[2008\]](#)).

Concerning FGM, we used the spin-averaged data measurements of the magnetic field components in [GSM](#) coordinates for which we had a sample for each 3 seconds.

ESA provides three main modes of data: *Full* with an approximate 90s resolution, *Reduced* with an approximate resolution of 4s and *Burst* mode that provide high-resolution 3D distribution functions for disjoints 5 minutes intervals. The reduced mode can be separated into two distinct sub-modes. The Slow-Survey mode, where the particles distribution functions are composed of 32 omni directional energy channels and a singular solid-angle distribution, and the Fast-Survey mode, where the distribution functions are composed of 24 energy channels and 50 solid-angle distributions. Because of this singular solid-angle distribution, the computation of the ion bulk velocity is impossible in the Slow-Survey mode. We then use the plasma moments provided by the Fast-Survey reduced mode whenever they are available. We use the onboard moments to fill in the data gaps in the Slow-Survey mode. The remaining holes in the plasma moments are filled with the data measured in the full mode and linearly time interpolated in order to obtain streaming time series of the ion density, velocity and temperature with a uniform resolution of 4s.

The ESA and FGM are then synchronized to obtain a unique dataset with a common resolution of 5s.

### 4.2.2 Double Star

Double Star is a sino-european mission that consists in two spacecraft: TC-1 and TC-2. TC-1 was launched in December 2003 and introduced into an equatorial orbit while TC-2 was launched in July 2004 and had a polar orbit. The mission ended with the atmospheric re-entry of TC-1 in October 2007.

In this thesis, we will only consider the data provided by TC-1 during the whole mission period.

We used the magnetic field data provided by the *Fluxgate Magnetometer* (FGM, Carr et al. [2005]) and the plasma moments provided by the *Hot Ion Analyzer* instrument (CIS-HIA, Fazerley et al. [2005]) whenever these data were available. The data from the two instruments were synchronized in order to obtain a streaming physical features dataset with a uniform temporal resolution of 4s.

### 4.2.3 MMS

*Magnetospheric Multiscale* (MMS) is a NASA mission launched in March 2015. Specifically designed for the study of near-Earth magnetic reconnection, the mission is made of 4 identical spacecraft that orbit in the equatorial plane with a close interspacecraft proximity while forming a tetrahedron. This close interspacecraft proximity (at a maximum being equal to 160 km) combined with the high time resolution provided by the Fast Plasma Investigation (FPI, Pollock et al. [2016]) is designed for the measurement of electron-scale physics and the probing of EDR whenever the plasma and electromagnetic signatures suggest the spacecraft came to the close surroundings of an X-line region. At ion-scale, this small distance implies very similar interspacecraft moments measurements and there is then no need to consider the 4 spacecraft in the frame of our study.

For this reason, we solely used the data provided by MMS 1 between September 2015 and July 2019.

The magnetic field information is measured by the Fluxgate Magnetometer (FGM, Russell et al. [2016]) for which we use the survey mode that provides data with a temporal resolution of 4.5s.

The plasma moments and distribution functions are provided by FPI under three different modes: the slow survey with an approximate 60s resolution, the Fast Survey with an approximate resolution of 4.5s and the Burst mode switched on with a 150ms resolution whenever the spacecraft comes across a region of interest. The selection, analysis and labeling of such burst intervals is the specific task of a group of experts called the *Scientists In The Loop* (SITL) that continuously browse the data and choose the one of specific interest to be stored and transmitted to the ground for analysis. Implying manual and visual labeling of data, this task is obviously subjective and time-consuming and automatizing the data selection process is an interesting alternative to the SITL.

In order to obtain a streaming continuous dataset with a similar resolution than the one we use for the other missions, we use the measurements in the Fast Survey mode.

After removing the data gaps of each instrument, data from both FGM and FPI are synchronized to obtain a streaming continuous MMS dataset with a temporal resolution of 4s.

### 4.2.4 Cluster

Cluster is an european mission launched in July and August 2000. The mission consists in 4 spacecraft designed to study the near-Earth environment with a specific focus on the magnetospheric cusps thanks to a polar orbit that distinguishes this mission from the three we have been presenting so far.

Years before MMS, the orbits of the four spacecraft are the very first ones to be designed so that they formed a tetrahedron. The interspacecraft distance varying here between 10 and 20000 km.

Whenever they are available, we use the magnetic field measurement provided by the Fluxgate Magnetometer (FGM, Balogh et al. [2001]) and the plasma distribution functions and moments provided by the Hot Ion Analyzer of the Cluster Ion Spectrometry (CIS-HIA, Rème et al. [2001]). The data of both instruments being measured with a 4s resolution.

Depending on the mission phases and the spacecraft location, HIA works under different modes that correspond to different energy sweeping schemes and thus an eventual altered plasma moments computation:

- The magnetospheric and magnetosheath modes where the particle distribution functions come on the full 62 energy channels and 88 solid angles distribution.
- The solar wind modes where the energy channels are reduced to allow a higher energy resolution for the solar wind beam data which alters the associated computation of plasma moments.

Because they are limited to the energy ranges typically found in the solar wind, the HIA data measured under the solar wind is not well suited to compute properly the plasma moments in the magnetosphere and the magnetosheath for which low and high energy information is needed. Consequently, we removed the HIA data that are measured under the solar wind modes and we keep the data under the magnetospheric and magnetosheath modes only.

HIA being unavailable on Cluster 2, 4 and on Cluster 3 after 2009, we used the data provided by Cluster 1 between January 2001 and January 2013 as well as the data provided by Cluster 3 until November 2009.

#### 4.2.5 Multi-spacecraft datasets

Each of the different dataset then consists in 8 input variables; the ion bulk velocity components,  $V_x$ ,  $V_y$ ,  $V_z$ , the magnetic field components,  $B_x$ ,  $B_y$ ,  $B_z$ , the ion density  $N_p$  and the temperature  $T$ .

Due to the important differences existing between the different missions in the specificities of the distribution functions and particle energy or pitch angle spectrograms, we chose to focus on the plasma moments and magnetic field only.

For each dataset, the ion omnidirectional differential energy fluxes shown in 4.1 are then only be used for visual inspection of the data and to provide visual guidance in our labeling process.

In addition to the plasma and magnetic field measurements, we collect the position of each spacecraft that will be exploited, in the entire manuscript, in [GSM](#) coordinates.

### 4.3 Labeling [THEMIS](#) data

We start out this work with the data provided by [THEMIS](#) B on the whole available period for this spacecraft, that is to say between the 1<sup>st</sup> of March 2007 and the 1<sup>st</sup> of January 2010. To erase the noise due to very punctual partial crossings that will particularly hard to label and detect, we resample the data to a 1 minute resolution. A typical representation of such resampled data is shown in [Figure 4.1](#).

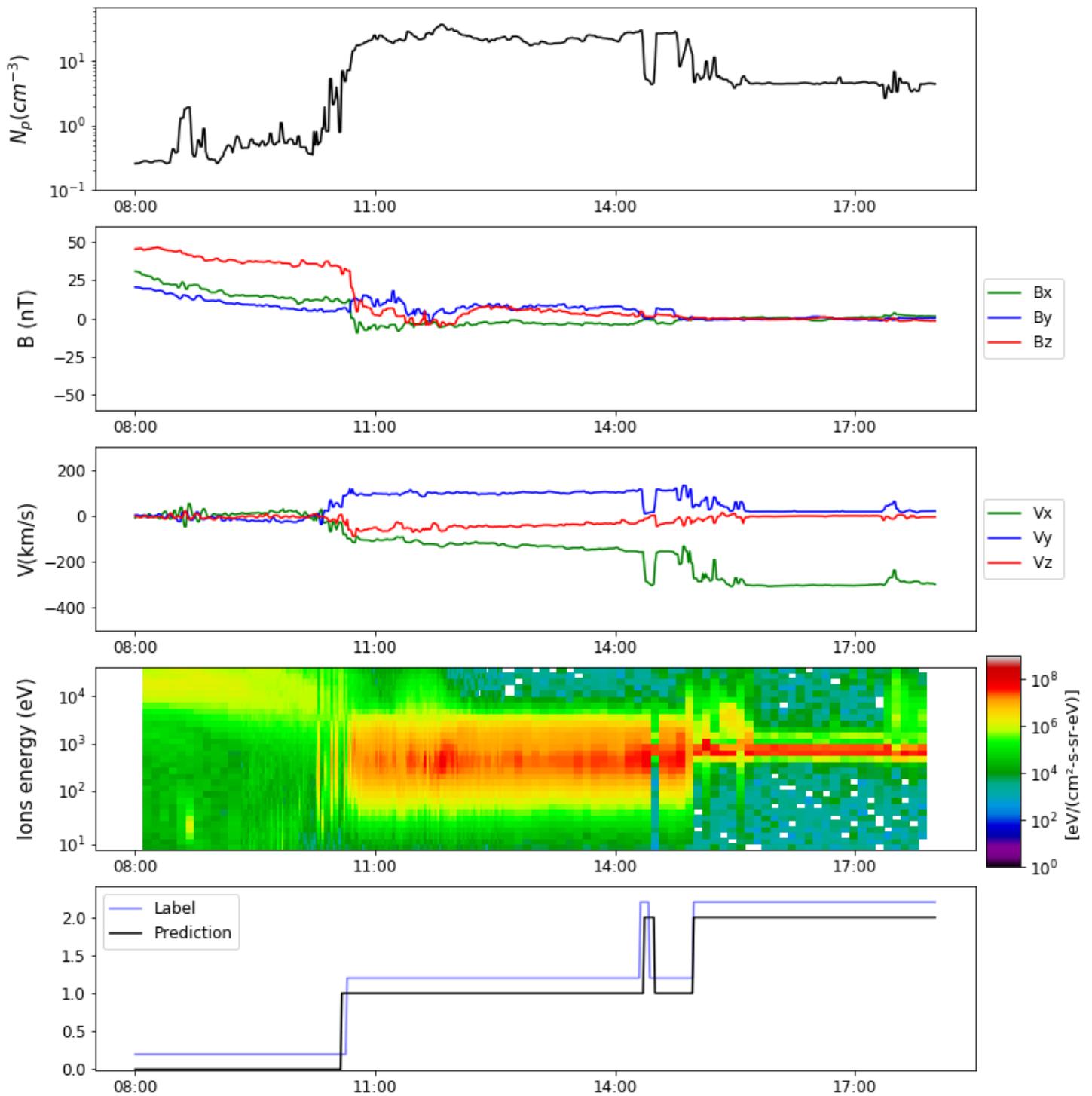


Figure 4.1: In-situ measurement provided by THEMIS B spacecraft on the 12<sup>th</sup> of May 2008. From the top to the bottom are represented: the ion density, the magnetic field components, the velocity components the omnidirectional differential energy fluxes of ions. The last bottom panel represents the evolution of the label (blue) , intentionally shifted for visual inspection and the prediction made by our algorithm (black).

Following the main characteristics of the near-Earth regions we presented in Chapter 1, we define the labeled regions as follows:

- Points in tenuous regions with almost no ion bulk flow and important magnetic field are identified as magnetosphere points and are associated to the label 0.
- Points in dense, fast regions for which we notice a monoenergetic beam in the ion spectrogram are identified as solar wind points and are associated to the label 2.
- Points that are not identified as solar wind or magnetosphere are identified as magnetosheath and are associated to the label 1. Those points correspond to the denser regions with an intermediate plasma velocity with a wide-spread ion spectrogram. With this definition, any region downstream of the bow shock that is not the magnetosphere is considered as the magnetosheath. This will be particularly the case of the regions of mixed plasmas such as the reconnection outflows or the near-cusp dense and hot plasma. In the optics of the detection of magnetopause crossings, this is as if we did not allowed any spontaneous mix of plasma at the magnetosphere-magnetosheath interface. Consequently, the magnetopause detected by this manner will actually correspond to the tangential definition of the magnetopause we gave in the Chapter 1. If this choice has negligible consequences regarding the magnetopause location at low-latitudes, it has a non-negligible impact on the representation we have on this boundary in the near-cusp region where the plasma mixing occurs on a much broader region. The consequences of this choice on the position and shape of the boundary in this specific region will be investigated in the next chapter.

We make those labels by inspecting the data visually and deciding, by selecting intervals, to which class their points belonged to. This requires to zoom in and out many intervals and is thus a long and fastidious process. To make it faster, in particular to zoom in regions of interest, we decide to guide our eyes with the preliminary predictions of algorithms trained on a dataset iteratively widened by our labels, plotted over the data.

The typical labeling of the three regions for a 1 minute resampled data interval is shown on the last subplot of Figure 4.1 where the theoretical label, shown in blue has been slightly shifted vertically for visual purpose. Following this process, our dataset is made of 59798 points of magnetosphere, 48056 points of magnetosheath and 150415 points of solar wind.

We selected data within dawn, dayside and dusk operation phases of [THEMIS](#) and thus expect a good [Magnetic Local Time \(MLT\)](#) coverage of both magnetopause and shock surfaces. This is confirmed by the actual spatial coverage of our labeled dataset shown in Figure 4.2. We then expect the method to be robust enough to the variability one can find in the data through the three different [THEMIS](#) operation phases.

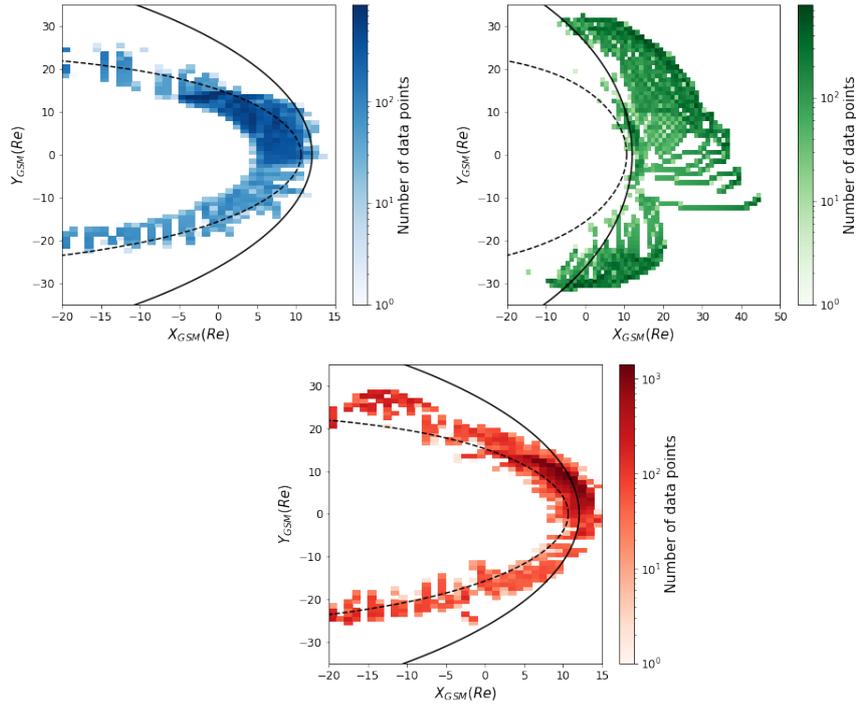


Figure 4.2: Spatial coverage of our labeled THEMIS dataset projected in the (X-Y) GSM plane, the solid black line represent a stand-off position the bow shock following Jeřáb et al. [2005] model while the dotted black line represent the magnetopause model of Lin et al. [2010]. Labels are spatially represented in a log-scale 2D histogram. Magnetosphere bins in blue vary between 1 and 901, Magnetosheath bins in red vary between 1 and 1421, solar wind bins in green vary between 1 and 788

#### 4.4 Algorithm selection

First of all, we randomly split our THEMIS dataset in 10 different ways in order to have 70% of the total dataset representing the training set and the remaining 30% constituting the test set. For each split, we train and test 3 different types of algorithms: logistic regression, decision tree and gradient boosting. From the AUC and HSS averaged over the three splits that are shown for each class and each algorithm in Table 4.1, gradient boosting appears as being the algorithm that performs best on differentiating the three regions although it should be noted here that even the simplest algorithm result in fair performances already. Gradient boosting will then be the algorithm we will be training and evaluating for the rest of the section.

	Logistic Regression	Decision Tree	Gradient Boosting
AUC magnetosphere	0.998	0.976	<b>0.999</b>
AUC magnetosheath	0.954	0.937	<b>0.997</b>
AUC solar wind	0.937	0.881	<b>0.999</b>
HSS magnetosphere	0.974	0.953	<b>0.987</b>
HSS magnetosheath	0.846	0.878	<b>0.975</b>
HSS solar wind	0.560	0.701	<b>0.992</b>

Table 4.1: AUC and HSS obtained for different algorithms for several train-test split.

The high AUC and HSS obtained for the three classes indicate how well this model performs in classifying the three regions. Moreover, the standard deviation obtained from the 10 different split cases is lower than  $10^{-3}$ , which shows that our method is independent from the split we make between our two sets. Seeing a lower score on the magnetosheath is not surprising as this is the class where we will find the most diversity in the physical nature of the data.

## 4.5 Algorithm performance

### 4.5.1 Temporal dependance

By performing a random split, we allow temporally close points, and thus almost identical points in the features space, to be in both training and test sets. Consequently, performing this split only partially shows the reproducibility of our method and how well it would perform on really unknown data. To ensure there are not any temporal overfit due to this split and that our method is truly reproducible, we train and evaluate our model 3 times by splitting our dataset temporally instead of randomly. For this, we consider our training set to be a time interval that represented 2/3 of our dataset and leave the remain to be the test set. The average [AUC](#) we obtain in this case are also shown in [Table 4.2](#) the very few variations we have compared to the random split ensures the temporal independence of our method as well as its reproducibility.

Mission	<a href="#">AUC Magnetosphere</a>	<a href="#">AUC Magnetosheath</a>	<a href="#">AUC Solar Wind</a>
<a href="#">THEMIS B</a> (w. Random split)	0.999	0.999	0.999
<a href="#">THEMIS B</a> (w. Temporal split)	0.999	0.997	0.999
Cluster 1 (without retraining)	0.988	0.983	0.996
Cluster 1 (with retraining)	0.999	0.998	0.999
Double Star TC1 (without retraining)	0.996	0.992	0.996
Double Star TC1 (with retraining)	0.999	0.998	0.999
<a href="#">MMS</a> (without retraining)	0.997	0.994	0.995
<a href="#">ARTEMIS</a>	0.999	0.999	0.999

Table 4.2: Comparison of the [AUC](#) of the [ROC](#) of our detection algorithms for different missions.

### 4.5.2 Influence of the manual labeling

The manual labeling process can be an important source of prediction errors. Thus, the label can eventually contain errors that could affect the quality of our prediction and high [AUC](#) would then not indicate the classification ability of our model but its ability to learn from an erroneous label. To figure this out, we perform trainings and evaluations of the algorithm by voluntarily mislabeling an ever-growing percentage of the dataset. If our model completely follows the indicated label in the training set, we expect a high [AUC](#) whatever this percentage might be. The mislabeling process is done as follows:

- We select a fraction of random points in the dataset
- The magnetosphere and the solar wind points are labeled as magnetosheath points
- Magnetosheath points are randomly mislabeled between the two other classes

The main reason that justifies this process stands in the fact that a human observer will never confuse magnetosphere and solar wind and there is of course some ambiguity in the labeling for classes concerned with a physical interface where data points don't strictly belong to either one or the other, but rather represent the finite transition region, omitted in our model. We repeat the operation for an ever growing percentage of the dataset until the proportion of the mislabeled points reaches 50% of the dataset. The random mislabeling and associated training and [AUC](#) computation are repeated 10 times at each step. The evolution of the [AUC](#) with the mislabeling proportion is shown in [Figure 4.3](#) for the three classes of the [THEMIS](#) dataset. The grey dashed lines represent the standard deviation we have between the different iterations of a given percentage of mislabeling.

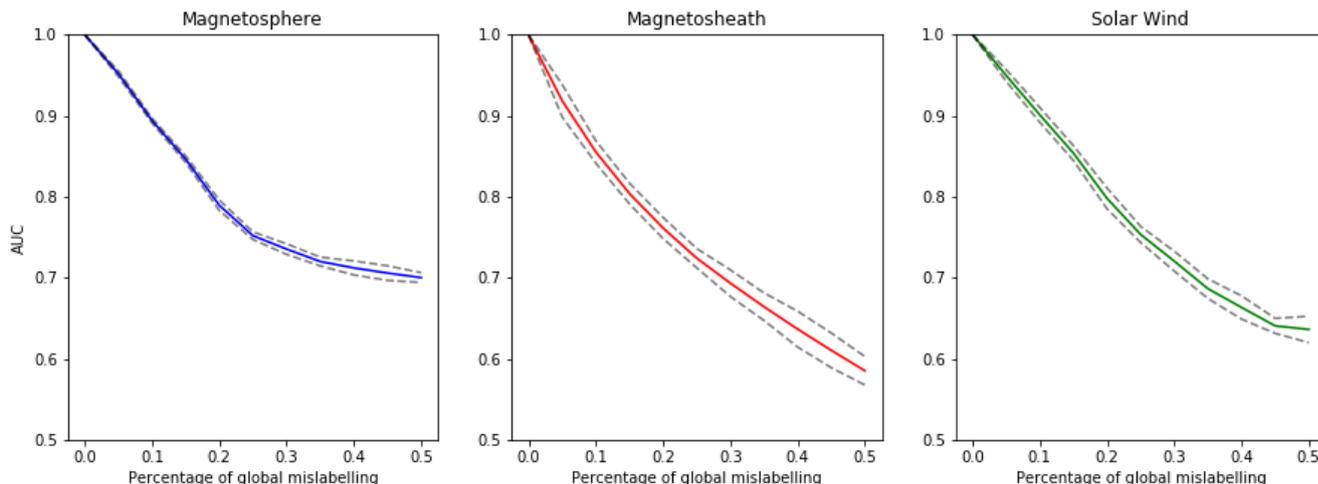


Figure 4.3: Evolution of the [AUC](#) as a function of the mislabelling percentage for the three different classes: magnetosphere (blue), magnetosheath (red) and solar wind (green). The gray dashed line represent the standard deviation we have between the different [AUC](#) scores of a same mislabelling percentage.

Having a more significant drop in the performances for the magnetosheath is not surprising as this is the class that will be most affected by our mislabelling process. Noticing that drop for the three different classes proves the model does not simply follow the indications provided from the labels but tries to find an intrinsic difference in the physical parameters of the three classes.

This shows the real capacity of our algorithm to classify the three near-Earth regions as well as the reliability of our label.

## 4.6 Adaptability of the model: from a mission to the other

Having developed an automatic detection method of the three near-Earth regions with high reliability, we should have few difficulties to adapt it to the data provided by additional spacecraft that goes through these regions. Even if a similar work can be adapted on the numerous past missions that went through the three near-Earth regions, we focus on this chapter on the most recent missions that offer the advantage of providing the data with the best quality, which removes an additional complexity that would appear with the oldest missions.

To do so, we label data points of each of the missions we are working on and compare this label to the predictions of our model trained with [THEMIS](#) data.

### 4.6.1 Double Star

A typical representation of the 1 minute resampled Double Star data data is shown in [Figure 4.4](#).

We label 20 671 magnetosphere points, 23 091 magnetosheath points and 4 944 solar wind points at the beginning of the year 2005 of our 1 minute resampled Double Star dataset. A third of these points constituting the Double Star test set, the other two thirds being kept in the case a different algorithm has to be trained to take into account the specificity of Double Star data. The spatial distribution of our labeled data is shown in the [Figure 4.5](#).

Since Double Star also has an equatorial orbit, we expect the model trained on [THEMIS](#) to perform well even without having to be retrained and this is the main reason why our label does not have to provide an entire coverage of the (X-Z) plane. And this is confirmed by the high [AUC](#) and [HSS](#) we have in [Tables 4.2](#) and the comparison of the [HSS](#) obtained for the different missions shown in the [Table 4.3](#).

Refitting the model would then allow a finer detection that would be specific to the quality of the data provided by Double Star in comparison to the [THEMIS](#) data but can be skipped as it does not bring a significant gain in [AUC](#) according to [Table 4.2](#).

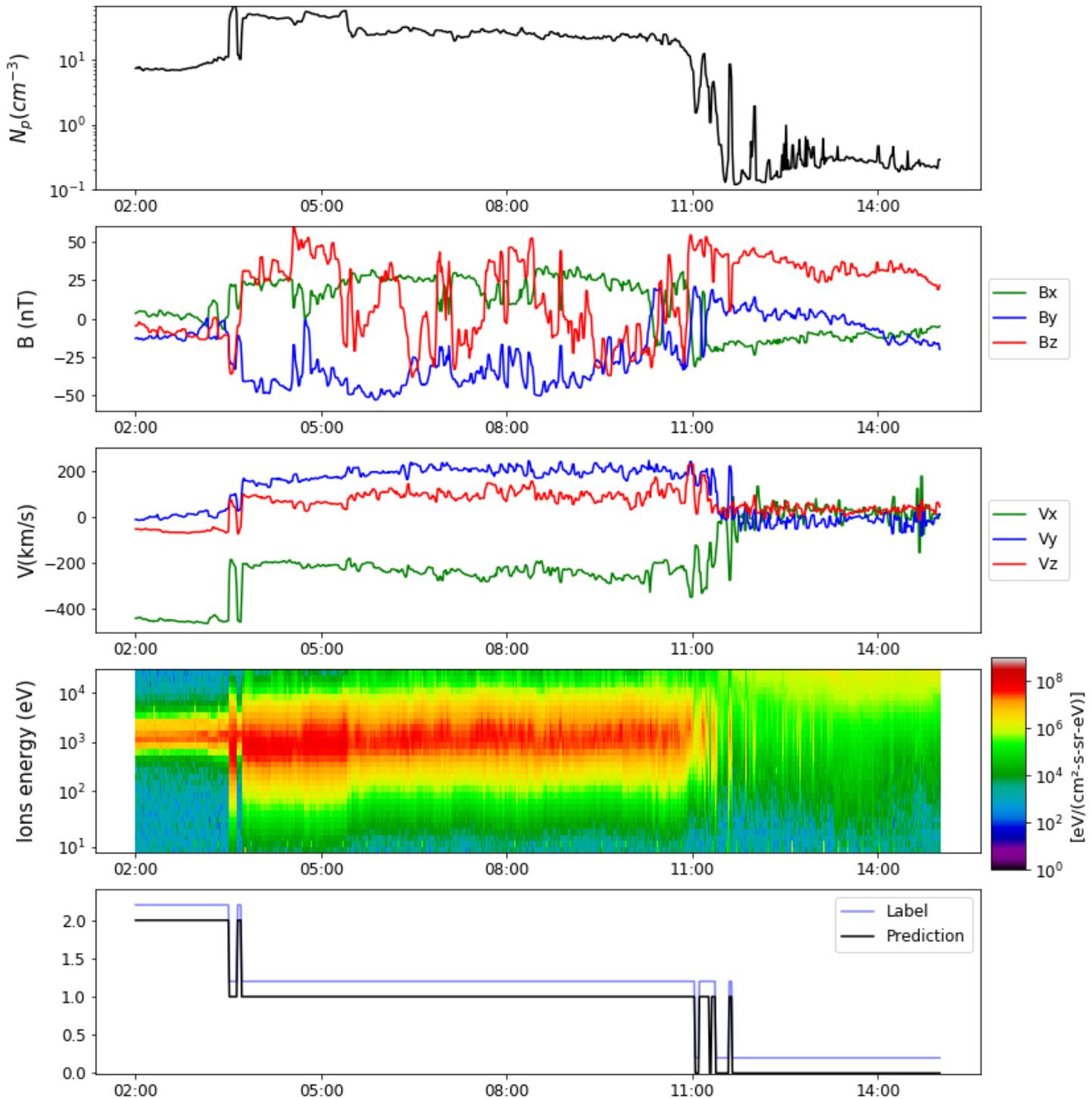


Figure 4.4: In-situ measurement provided by Double Star TC1 spacecraft on the 1<sup>st</sup> of January 2005. The legend is the same than in 4.1

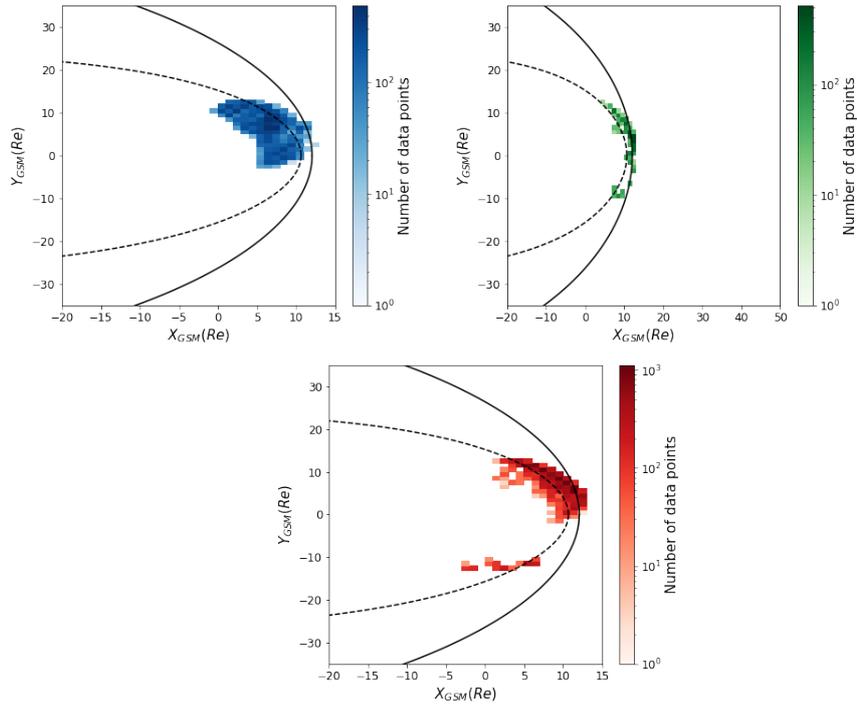


Figure 4.5: Spatial coverage of the Double Star labeled dataset. The legend is the same than in Figure 4.2

#### 4.6.2 MMS

A typical representation of the 1 minutes resampled MMS data is shown in Figure 4.6.

Since MMS also has an equatorial orbit, we once again expect the model trained on THEMIS to provide a very good classification of the three regions on MMS data as for the case of what has been shown for Double Star.

To figure it out, we label 7 612 magnetosphere points, 1 9272 magnetosheath points and 3 651 solar wind points during the first year of MMS and these labels the associated prediction of the classifier. The spatial coverage of these labeled points is shown in Figure 4.7

The high AUC and HSS shown in the Tables 4.2 and 4.3 confirms the adaptability of our classifier to equatorial missions without further additional fitting.

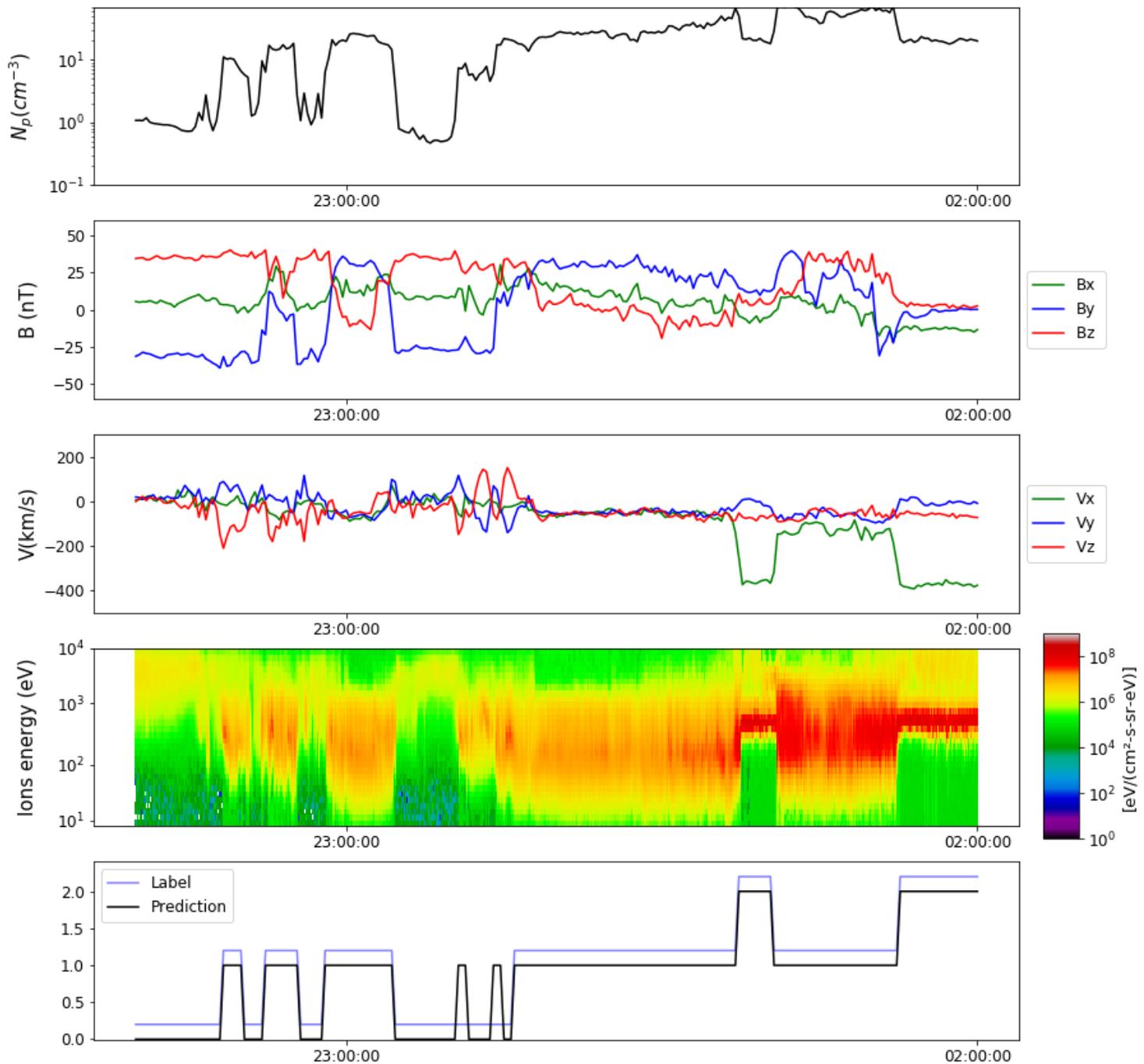


Figure 4.6: In-situ measurement provided by MMS spacecraft on the 31<sup>st</sup> of December 2015. The legend is the same than in 4.1

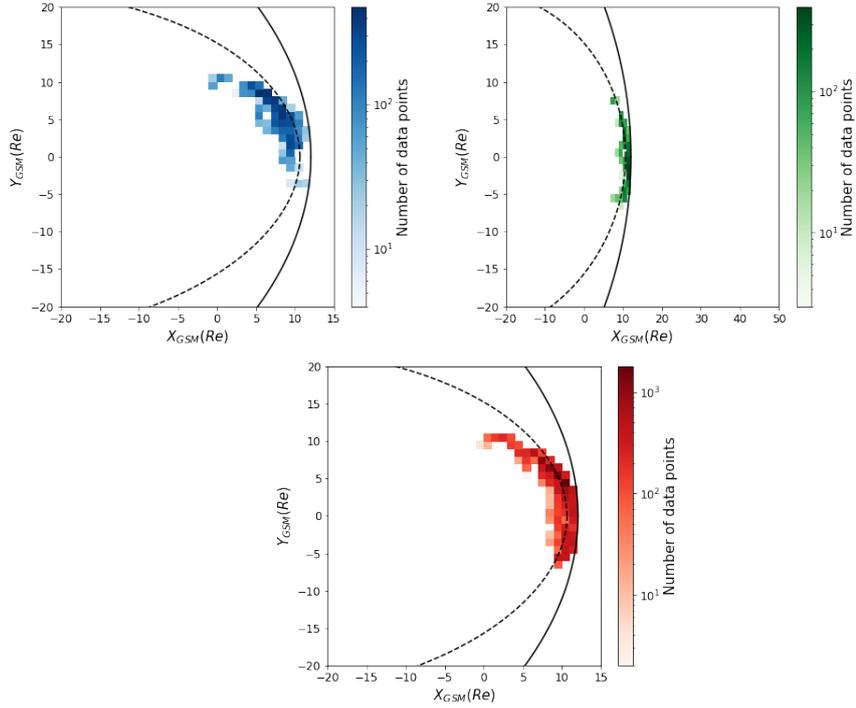


Figure 4.7: Spatial coverage of the MMS labeled dataset. The legend is the same than in Figure 4.2

### 4.6.3 Cluster

In comparison with the two previous cases, Cluster case might be more challenging because of the orbit, polar in this case, and the regions visited that have different physical properties than the one visited by equatorial missions. The data provided THEMIS and Cluster can therefore be substantially different and there is no real clue on how an algorithm trained on equatorial orbit data would perform on predicting on polar orbit data.

One minute sampled Cluster data are shown in Figure 4.8 and we here label 50 277 points of magnetosphere, 76 468 points of magnetosheath and 22 017 of solar wind between the years 2005 and 2006 which spatial distribution is shown in Figure 4.9. One third of these labeled points are used to evaluate the performances of the models while we kept the remaining two thirds in the case refitting the algorithm is needed. Applying our THEMIS-trained model, we notice a lower AUC for each of the three classes. This indicates the adaptability is not that obvious in this case.

We then adapt our classifier to the polar case by refitting the model trained on THEMIS with the Cluster labels. The increasing AUC we obtained, shown in Table 4.2 and the associated high HSS also shown in 4.3 proves the necessity we had to adapt our algorithm to the specificity of the Cluster data. It also shows our method can be easily adapted to the data of another mission, exploring regions with significant statistical deviations of the features, after a small labeling and refitting phase.

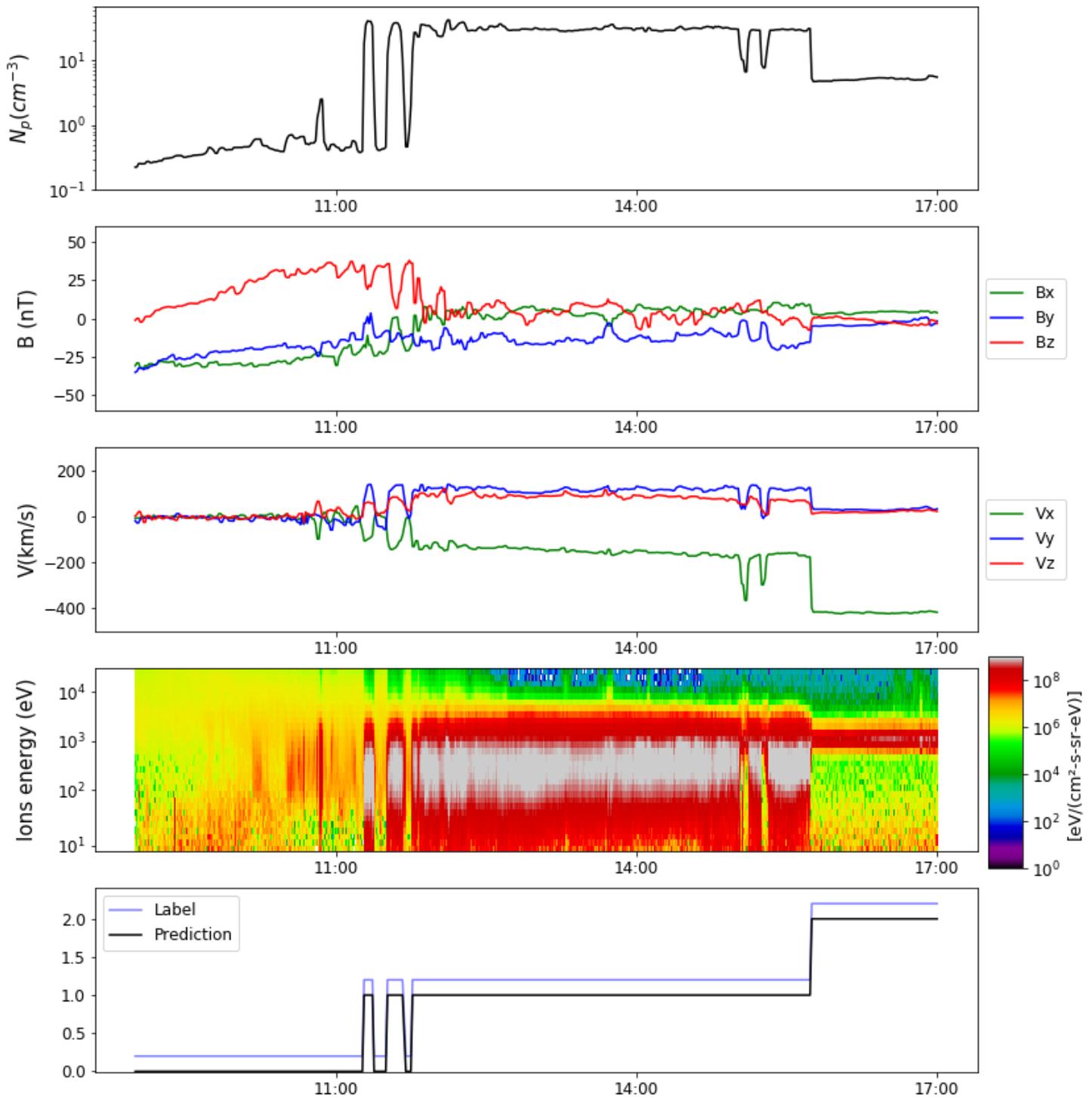


Figure 4.8: In-situ measurement provided by Cluster 1 spacecraft on the 6<sup>th</sup> of February 2005. The legend is the same than in 4.1

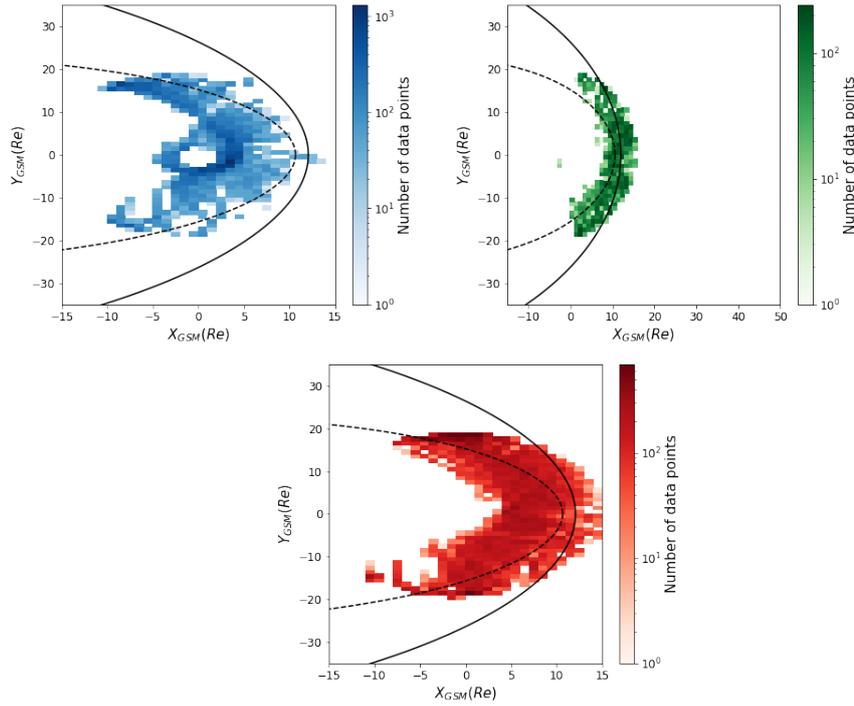


Figure 4.9: Spatial coverage of the Cluster labeled dataset. The legend is the same than in Figure 4.2

#### 4.6.4 ARTEMIS

The orbit of the **ARTEMIS** spacecraft is different from the orbit of the mission we have been investigating so far. This difference comes with a lot of change in the nature of the data measured by the spacecraft.

First of all, the spacecraft orbit the moon and are then much farther (around 40 Re) from the Earth than the spacecraft of the other missions we have studied. This implies the spacecraft does not explore the dayside regions and crosses the magnetopause and the bow shock in the night-side. At these distances, the magnetosheath plasma becomes almost as fast and as tenuous as the solar wind and small fluctuations on either one of the other side could easily be confused with a boundary crossing.

Second, the spacecraft spend most of their time in the solar wind, which make their measurement more sensitive to the data variability induced by the solar cycle that we neglected for the previous missions.

Finally, this specific type of orbit also introduces time intervals during which the data does not take values statistically close to any of our regions of interest. Indeed, once per orbit, **ARTEMIS** explores the lunar wake, characterized by an extremely low density and fluctuating velocity in every direction. These intervals, for which a typical representation of the data is shown in Figure 4.10, cannot be considered to belong to any of our existing region classes.

For this three reasons, the method we presented in the previous sections and successfully adapted to Double Star, **MMS** and Cluster cannot be used as is and the entire process from the labeling to the choice of the feature has to be designed from scratch.

To cope with the variability induced by the solar cycle we label a month per year and add the lunar wake as a fourth explored region (with an associated value of 3). The final labeled dataset is made of 26 560 magnetosphere points, 131 656 magnetosheath points, 429 283 solar wind points and 15 070 points of lunar wake which spatial distribution is shown in Figure 4.11.

We cope with the increasing difficulty to distinguish magnetosheath and solar wind by adding the spacecraft **GSM** coordinates as a feature of the dataset which will then consist in 11 input variables.

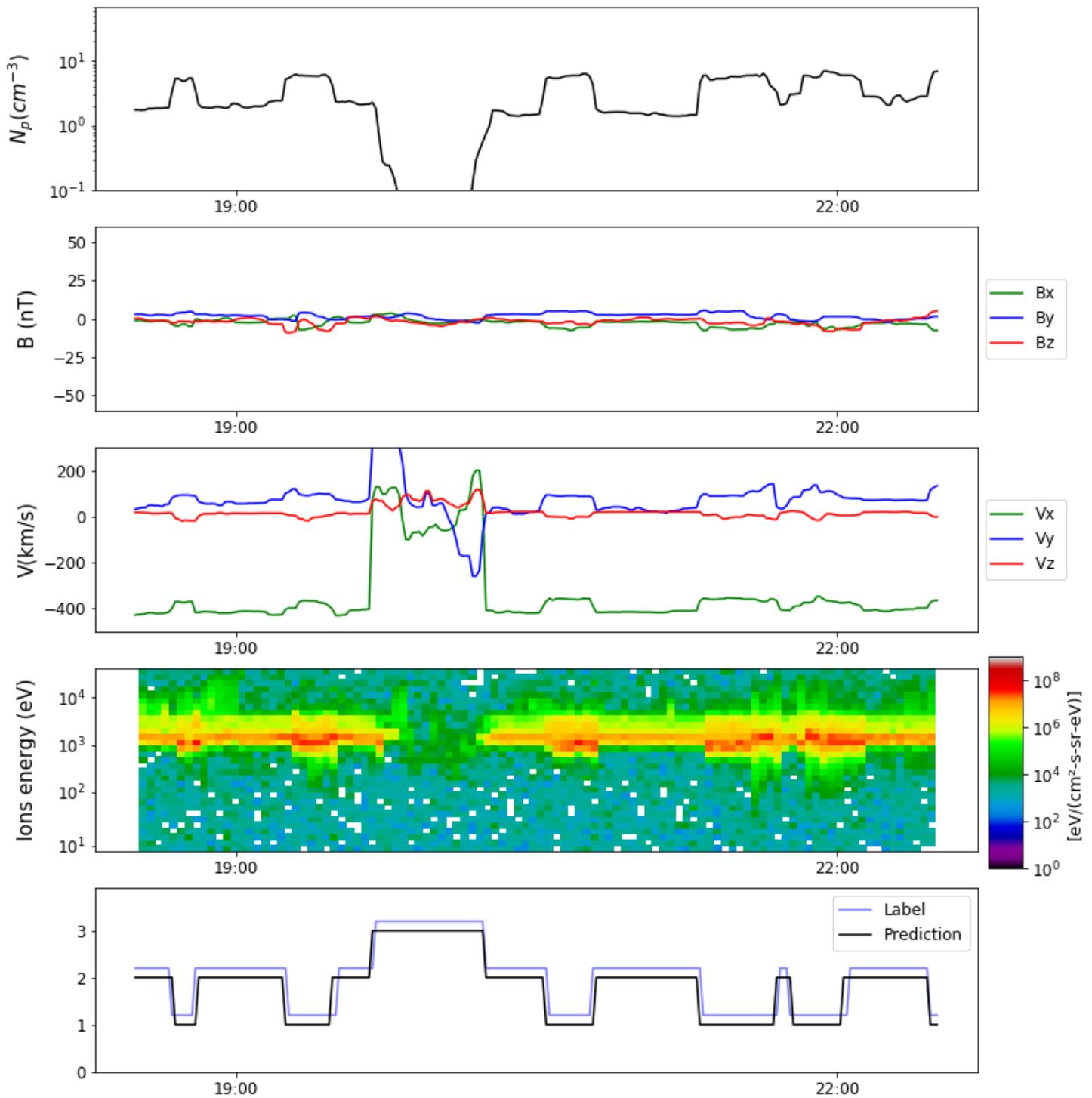


Figure 4.10: In-situ measurement provided by the [ARTEMIS B](#) spacecraft on the 13<sup>rd</sup> of August 2016. The legend is the same than in 4.1

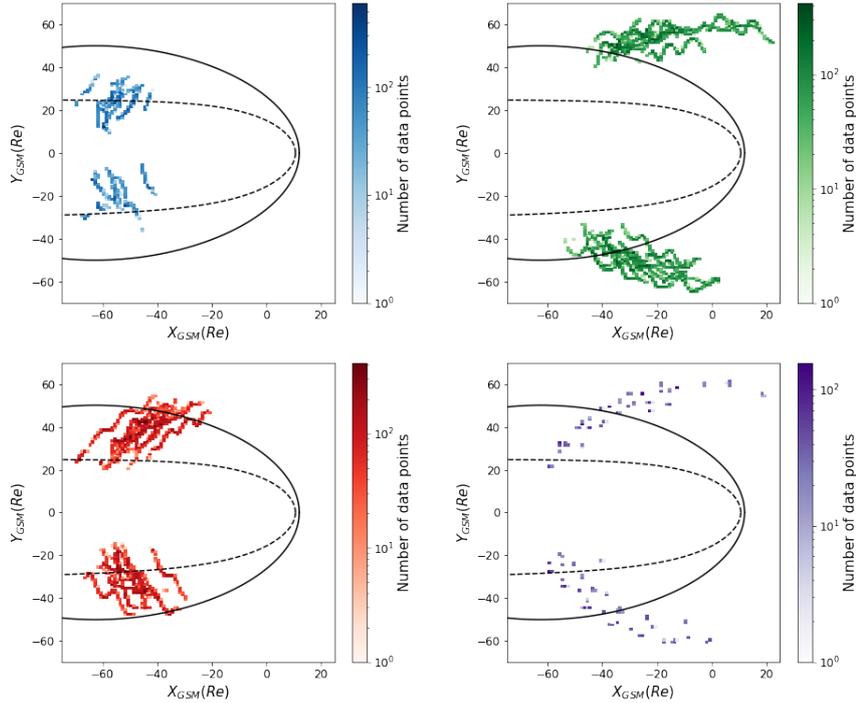


Figure 4.11: Spatial coverage of the **ARTEMIS** labeled dataset. The legend is the same than in Figure 4.2 with the addition of the Moon’s wake bins in purple which vary between 1 and 157

Having a different dataset and a different number of classes, we here cannot use the model trained in the previous section and we will then focus on the specific model we trained for this mission. The resulting high **AUC** shown in Table 4.2 proves the adaptability of our gradient boosting based method to another kind of orbit and and its flexibility and robustness regarding the addition of another region. This especially confirmed with the **AUC** and the **HSS** we found for the lunar wake region, that we respectively found equal to 0.97 and 0.947.

<b>Mission</b>	<b>HSS Magnetosphere</b>	<b>HSS Magnetosheath</b>	<b>HSS Solar Wind</b>
THEMIS B	0.987	0.975	0.993
Cluster 1	0.976	0.972	0.981
Double Star TC1	0.980	0.974	0.983
MMS	0.982	0.973	0.987
Artemis	0.976	0.962	0.974

Table 4.3: Comparison of the **HSS** of our detection algorithms for different missions.

## 4.7 Comparison with manually set thresholds

Having proved the efficiency of gradient boosting on different missions<sup>1</sup>, we want to compare it to the state of the art existing methods such as the one elaborated by Jelínek et al. [2012] that we described in the introduction.

Figure 4.12 represents the 2D histogram of  $B$  and  $N_p$  for THEMIS B, Double Star and Cluster 1 on the periods on which we labeled the different associated datasets. We divided these parameters by the corresponding solar wind density and the IMF amplitude that we obtained from the shifted OMNI data. At first sight, one can easily distinguish three main regions that are separated with the solid red lines for the three missions. Nevertheless, these linear boundaries have been set manually and we cannot ensure these could be the best choice for the three missions. To evaluate the quality of the classification, we compute the TPR and the FPR for the three missions and for varying boundary lines. We then use these values to compute the AUCs that are shown in the Table 4.4.

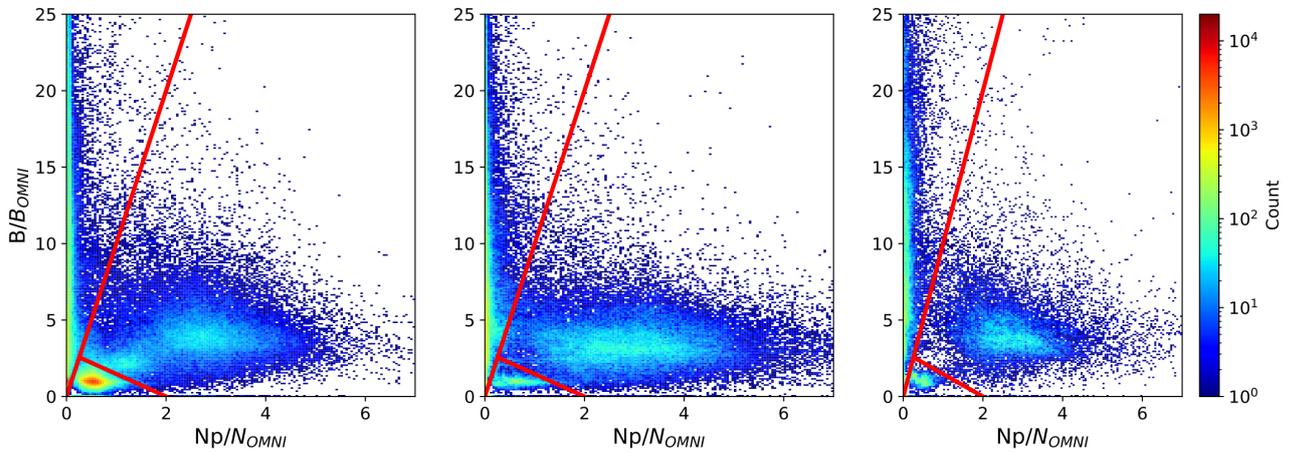


Figure 4.12: 2d histogram of  $B$  and  $N_p$  divided by the corresponding OMNI data for the three missions: THEMIS B (left), Cluster 1 (middle) and Double Star TC1 (right). The solid red lines indicate a possible set of linear boundaries we could define to separate the three regions

Mission	AUC Magnetosphere	AUC Magnetosheath	AUC Solar Wind
THEMIS B	0.915	0.908	0.859
Cluster 1	0.897	0.852	0.828
Double Star TC1	0.913	0.894	0.843

Table 4.4: AUC for the threshold-based method

Once again, we notice a lower AUC in the case of Cluster which is consistent with the difference we have between equatorial and polar orbits as explained in the previous section. Additionally, even if the boundaries plotted in the Figure 4.12 seem to provide a decent separation between the three regions, the AUC is lower than the one we obtained with the gradient boosting. This indicates our model performs better in classifying the three regions by setting more flexible boundaries on supplementary features while requiring less fitting time than the one required to manually set the thresholds used in the Figure 4.12.

The same kind of histogram gets messier with a much less obvious transition from the magnetosheath to the solar wind and the addition of the moon's wake as shown with the ARTEMIS data in Figure 4.13. This shows the difficulty manually set thresholds would have for a night side oriented mission and the interest of using machine learning in this case.

<sup>1</sup>Additional prediction examples can be found in the appendix B.

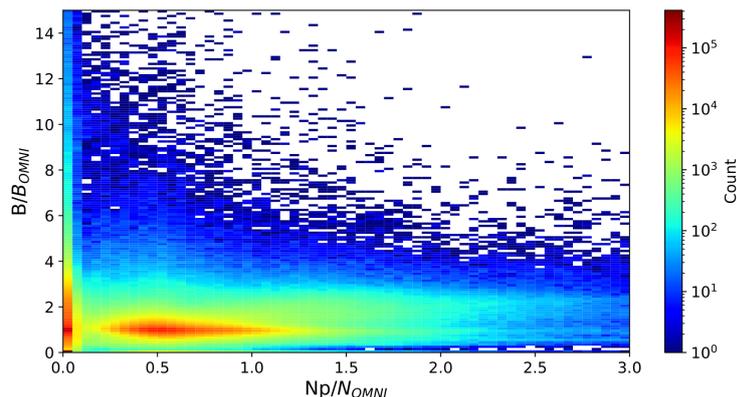


Figure 4.13: 2d histogram of  $B$  and  $N_p$  divided by the corresponding OMNI data for ARTEMIS B

## 4.8 Massive detection of boundary crossings

In the previous sections, we proved the efficiency, the reliability and the adaptability of our classifiers on data from several missions and spacecraft. From now on, these classifiers can be used to elaborate our own magnetopause and bow shock crossings catalogs by classifying the streaming in-situ data provided by any near-Earth spacecraft and by selecting time intervals enclosing two predicted regions. To do so, we train our 4 different models, THEMIS, Double Star, Cluster and ARTEMIS on their whole labeled datasets<sup>2</sup>.

### 4.8.1 Magnetopause catalog

We define a crossing as a 1 hour interval that contains as much magnetosheath points as magnetosphere points. With this definition, the magnetopause is defined as the region that separates the two labeled classes magnetosphere and magnetosheath. The consequence of such definition, especially in the polar cusps will be investigated in the next chapter.

We then elaborate a complete magnetopause crossing catalog by running our THEMIS model on the data provided by THEMIS A, B, C, D and E spacecraft. To gain time in the construction of the crossings and because we do not expect any magnetopause crossing in the nightside operation phase, we restrict ourselves to the dayside, dawn and dusk operation phase. As no crossing is expected far away in the solar wind ( $X_{GSM} > 15 \text{ Re}$ ) or close to the Earth dipole, we also remove these parts of the orbit.

The same model was used on the in-situ data provided by Double Star between 2004 and 2007 and MMS between 2015 and 2020.

We finally apply the same process on the in-situ data provided by Cluster 1 on the 2001-2016 period, by Cluster 3 on the 2001-2009 period and on ARTEMIS between 2010 and 2019 by using the corresponding trained model. The total number of crossings we obtained are summarized in the Table 4.5.<sup>3</sup>

<sup>2</sup>Those trained models can be found at [https://github.com/gautiernguyen/in-situ\\_Events\\_lists](https://github.com/gautiernguyen/in-situ_Events_lists)

<sup>3</sup>All of the magnetopause lists can be found at the same address.

<b>Mission</b>	<b>Magnetopause crossings</b>	<b>Bow shock crossings</b>
THEMIS A	2 824	1 590
THEMIS B	373	1 030
THEMIS C	658	1 238
THEMIS D	2 691	1 520
THEMIS E	2 726	1 511
Cluster 1	1 813	3 225
Cluster 3	1 534	2 004
Double Star TC1	931	846
MMS 1	810	1 035
ARTEMIS B	263	1 602
ARTEMIS C	373	1 626
Total	14 996	17 227

Table 4.5: Number of magnetopause and bow shock crossings we have for different missions

Given that our detection method has been evaluated on large parts of orbits, the high quality of the classification is made with regions where the spacecraft is not expected to cross a boundary. In these regions, the algorithm is less likely to hesitate on its prediction. On the other hand, it is more probable it hesitates on the predictions made close to the boundaries. Consequently, we have to ensure the classification is still of decent quality there.

Figure 4.14 represents the ROC we have on the classification between magnetosphere and magnetosheath points for THEMIS B, Cluster 1 and Double Star for the subset of our test set that lies in the proximity of a magnetopause or shock crossing. These predictions have been obtained with a model that has been trained with the complement part of the dataset, i.e. the subset that excludes the proximity of the crossings. Even if the AUC is lower than the one we obtained in the previous section, its still high value indicates the good quality of the classification when a spacecraft arrives close to the magnetopause and thus our capacity of building crossings from the prediction made by our model.

Another method we could use to ensure the consistency of the obtained crossings would stand in the certitude of the prediction made by the algorithm and their position in comparison to a theoretical magnetopause position. To do so, we computed the mean probability of each crossing by averaging the probabilities of belonging to the predicted class of each point present in the crossing.

As we explained it in Chapter 2, the use of the probabilistic output only makes sense if the probabilities are well-calibrated, which is not especially the case for ensemble algorithms such as gradient boosting Niculescu-Mizil and Caruana [2005]. The calibration curve of our THEMIS model is shown in Figure 4.15. Having a linear calibration curve close enough to the perfect calibration curve for the three regions, we consider the probabilistic output of our model to be decently well-calibrated and we can move on with using the probabilistic output of our models.

Events with high probability would correspond to undoubtful crossings while the events with the lowest probability would be the less likely to be actual crossings. The probability distribution of our 14996 is shown in Figure 4.16. Having a high probability for the greatest part of our events then ensures the consistency of our magnetopause list.

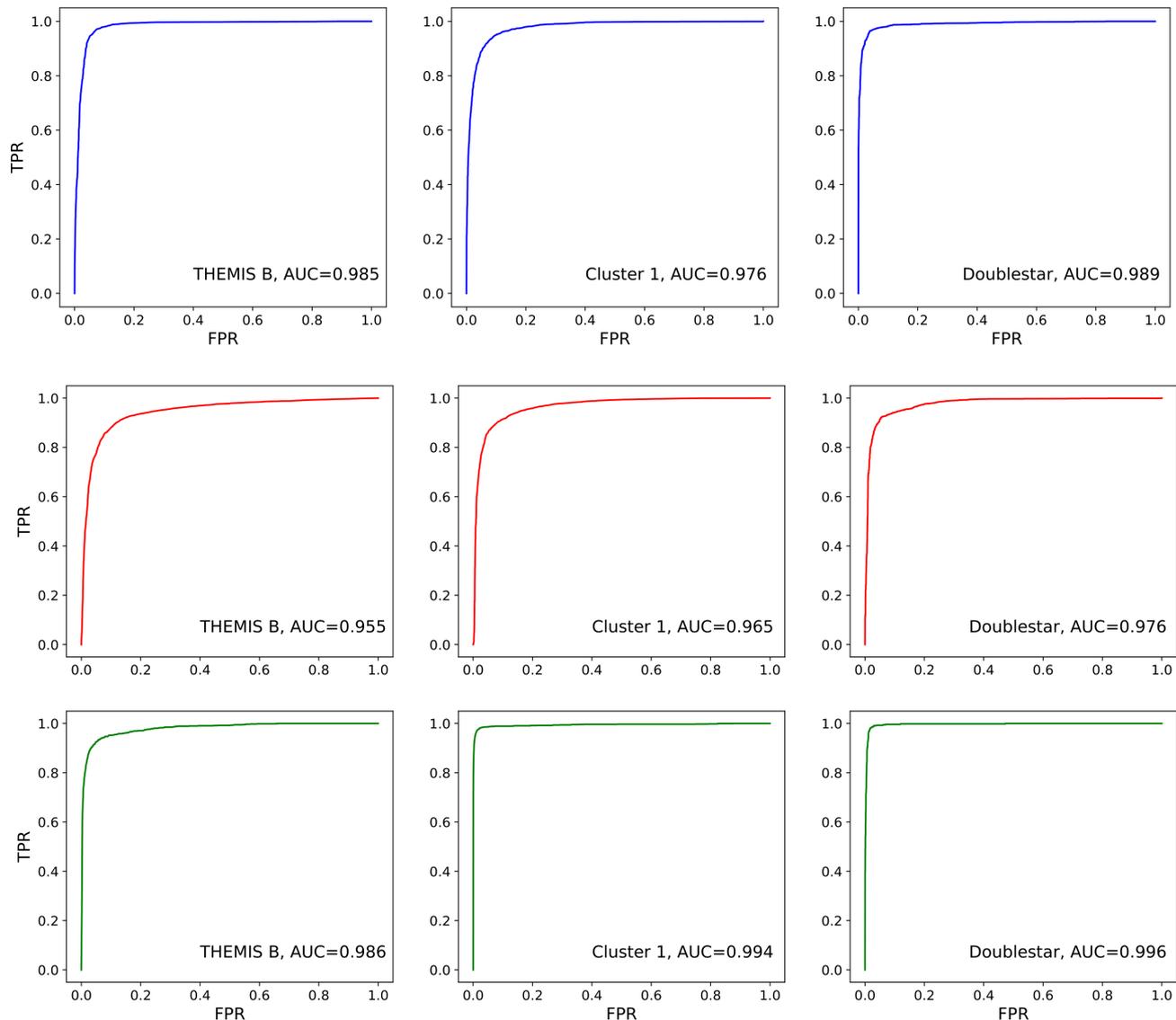


Figure 4.14: ROC curves evaluated on the labeled crossings for the three missions THEMIS B (left), Cluster 1 (middle) and Double Star (right) for the three classes: magnetosphere(top), magnetosheath(middle) and solar wind (bottom)

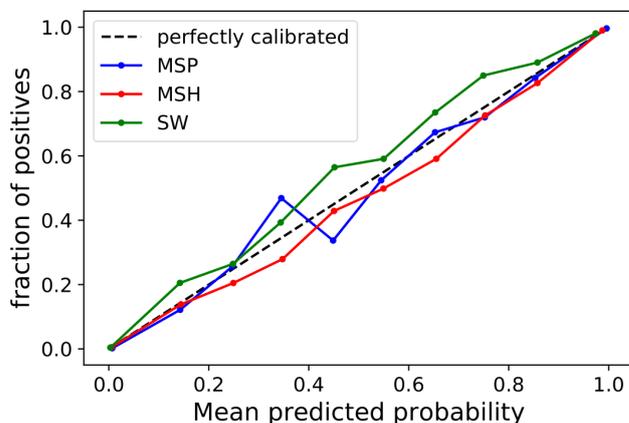


Figure 4.15: Calibration curve of our model trained on THEMIS data for the three regions. The black dashed line represent the calibration a perfectly-calibrated classifier would have.

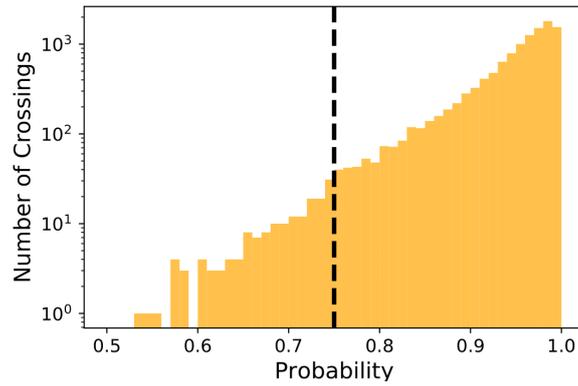


Figure 4.16: Distribution of the probability of the 14996 magnetopause crossings we built and summarized in Table 4.5. The solid dashed line represent the probability threshold we chose for the Figure 4.17

Finally, the spatial distributions of the crossings that have a probability higher than 75% in the GSM XZ, XY and YZ planes is shown in Figure 4.17. These crossings represent 98.5% of the crossings built with our models and are then expected to be the most likely to be actual magnetopause crossings. The solid black lines represent the stand off position of the magnetopause model established by [Lin et al., 2010] computed for a dynamic pressure of 2 nPa, a null  $B_z$  and assuming no dipole-tilt. The proximity between this distance and our actual crossings ends up proving the capacity our method has to elaborate a decent magnetopause crossings catalog with a decent coverage of the magnetopause at all latitudes and longitudes.

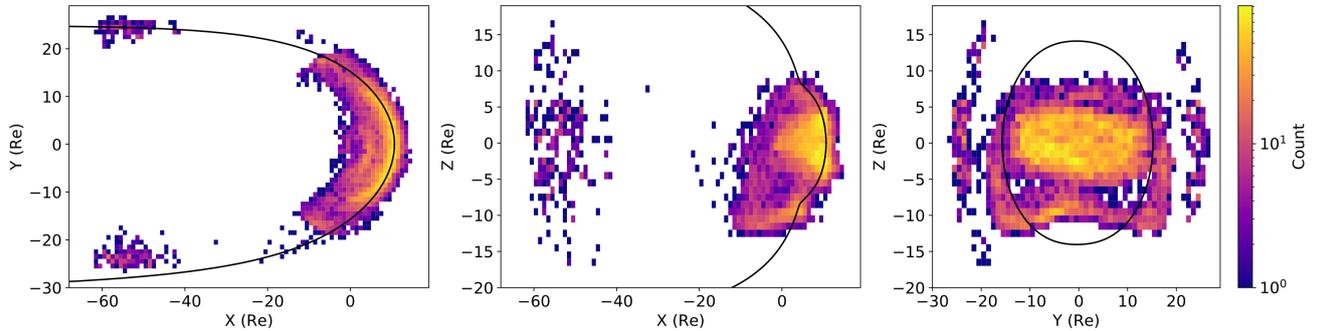


Figure 4.17: Spatial distribution of the crossings above the threshold in Figure 4.16 in the XY (left), XZ (middle), YZ (right) GSM planes. The solid black line indicate the Lin et al. [2010] magnetopause model with a dynamic pressure of 2 nPa and a null  $B_z$ .

## 4.8.2 Bow shock catalog

We define a bow shock crossing event as 10 minutes interval that contains as much magnetosheath points as solar wind points. We then run the models we trained for the different missions detailed in Section 3 on the same dataset we used in the case of the making of the magnetopause crossing catalog. The total number of obtained crossings is once again summarized in Table 4.5<sup>4</sup>.

The spatial distribution of the crossings with a probability higher than 75% in the GSM XZ, XY and YZ planes is shown in the Figure 4.19. The solid black line here represents the stand off position of the Jeřáb et al. [2005] bow shock model computed for a dynamic pressure of 2 nPa, a null  $B_z$  and an Alfvén Mach of 8.

<sup>4</sup>And the bow shock lists can once again be found at [https://github.com/gautiernguyen/in-situ\\_Events\\_lists](https://github.com/gautiernguyen/in-situ_Events_lists).

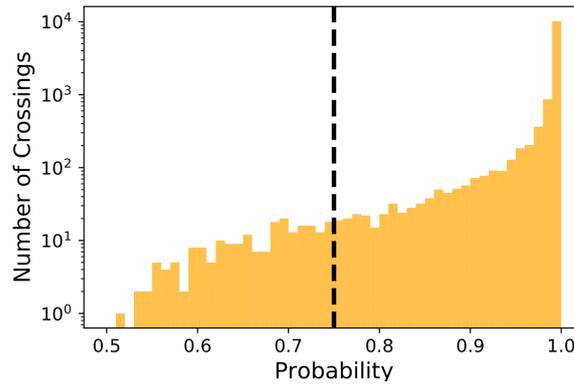


Figure 4.18: Distribution of the probability of the 17227 bow shock crossings we built and summarized in Table 4.5. The solid dashed line represent the probability threshold we chose for the Figure 4.19

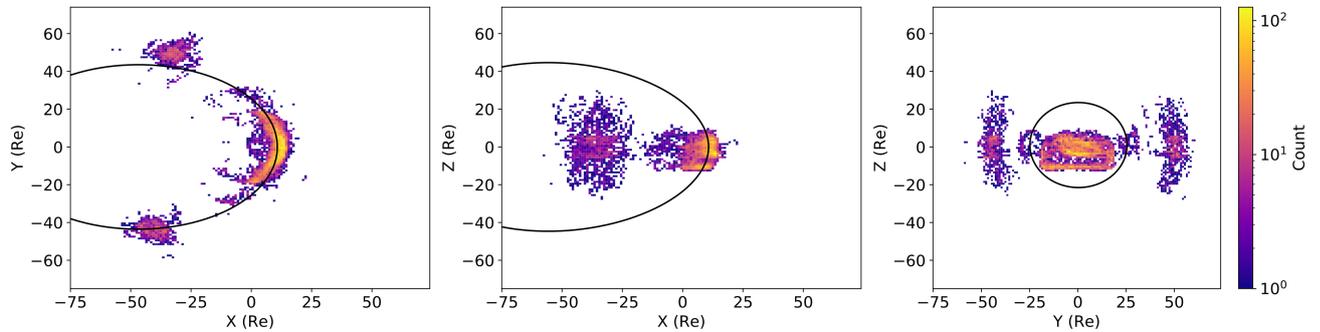


Figure 4.19: Spatial distribution of the crossings above the threshold in Figure 4.18 in the XY (left), XZ (middle), YZ (right) GSM planes. The solid black line indicate the Jeřáb et al. [2005] bow shock model with a dynamic pressure of 2 nPa, a null  $B_z$  and an Alfvén Mach of 8.

## 4.9 Conclusion

Using a Gradient Boosting Classifier, we established an automatic detection method of the different near-Earth environment regions when they are traversed by the THEMIS spacecraft during the dawn, dusk and dayside mission phases. This method was successfully adapted on other equatorial dayside missions (Double Star and MMS) and, after a small retraining phase necessary to consider the orbital differences between different missions, its success on non-equatorial dayside missions such as Cluster. The adaptability of the method has even been tested on missions with a substantially different orbit such as ARTEMIS for which we provided a successful region classification after a small redesign of the observed features and the way the label was made. Having proved this adaptability, we could also think of using the method on the data of additional near-Earth missions, such as the 23 different spacecraft enumerated in Wang et al. [2013], provided enough information about the plasma moments with a sufficient resolution is provided.

For simplification, we only considered 3 classes and defined as magnetosheath any region where plasma differed from pristine solar wind and magnetospheric ones. The classification could then even be enhanced by the consideration of additional regions like the ion foreshock, the Plasma Depletion Layer (PDL), the exterior of the polar cusps, or the different boundary layers of the magnetopause.

At the same time of our study, [Olshevsky et al. \[2019\]](#) elaborated a [CNN](#) based method that analyzed the in-situ distribution functions provided by [MMS](#) to provide a near-Earth regions classification. Due to the important differences existing between different missions in the specificities of the distribution functions and particle energy or pitch angle spectrograms this method would require to start it all from scratch when it comes to its adaptability to other spacecraft. This re-training necessity comes with unavoidable extra hours of training due to the slow convergence of [CNNs](#) compared to the GB and the much heavier dataset. For these reasons, using the plasma moments paved the way to an easy adaptability from a specific type of mission to another and the production of light-weight algorithms that could eventually be taken onboard of upcoming missions to automatically select the data of interest and thus automatically decide of the data that should be stored and kept for further analysis. This would particularly bring a huge gain time in the data selection process that are either threshold triggered or human monitored like the [SITL](#). Moreover, the method does not use the specificity of being in the near-Earth and could then also be adapted to other planetary missions in the solar system.

We used this method to elaborate one of the most exhaustive existing magnetopause and bow shock crossing catalogs. A bonus to our method is that these catalogs can be readily and automatically grown as new data is made available. Having a large list of events also gives the opportunity to study these two near-Earth boundaries and physical processes occurring in their vicinity, from a statistical point of view. One could think, for instance, of the identification of the magnetic reconnection jets, which will be the topic of the Chapter 6 where the region classifier will even prove its utility to automatically extract the magnetosheath conditions associated to a given crossing and the eventual underlying subset of reconnection jets.

Last but not least, the high number of events we found is expected to be linked with a great variety in the associated solar wind conditions and it could then be interesting to link the position of the crossings with these upstream conditions through the construction of magnetopause and bow shock empirical analytical models, this will be the topic of the next chapter.

## 4.10 Bibliography

Angelopoulos, V.: The THEMIS Mission, *Scientific Studies of Reading*, 141, 5–34, <https://doi.org/10.1007/s11214-008-9336-1>, 2008. [73](#)

Auster, H. U., Glassmeier, K. H., Magnes, W., Aydogar, O., Baumjohann, W., Constantinescu, D., Fischer, D., Fornacon, K. H., Georgescu, E., Harvey, P., Hillenmaier, O., Kroth, R., Ludlam, M., Narita, Y., Nakamura, R., Okrafka, K., Plaschke, F., Richter, I., Schwarzl, H., Stoll, B., Valavanoglou, A., and Wiedemann, M.: The THEMIS Fluxgate Magnetometer, *Scientific Studies of Reading*, 141, 235–264, <https://doi.org/10.1007/s11214-008-9365-9>, 2008. [73](#)

Balogh, A., Carr, C., Acuña, M., Dunlop, M., Beek, T., Brown, P., Fornacon, K.-H., Georgescu, E., Glassmeier, K.-H., Harris, J., Musmann, G., Oddy, T., and Schwingenschuh, K.: The Cluster Magnetic Field Investigation: Overview of in-flight performance and initial results, *Annales Geophysicae*, 19, <https://doi.org/10.5194/angeo-19-1207-2001>, 2001. [74](#)

Carr, C., Brown, P., Zhang, T. L., Gloag, J., Horbury, T., Lucek, E., Magnes, W., O'Brien, H., Oddy, T., Auster, U., Austin, P., Aydogar, O., Balogh, A., Baumjohann, W., Beek, T., Eichelberger, H., Fornacon, K. H., Georgescu, E., Glassmeier, K. H., Ludlam, M., Nakamura, R., and Richter, I.: The Double Star magnetic field investigation: instrument design, performance and highlights of the first year's observations, *Annales Geophysicae*, 23, 2713–2732, <https://doi.org/10.5194/angeo-23-2713-2005>, 2005. [74](#)

Fairfield, D. H.: Average and unusual locations of the Earth's magnetopause and bow shock, *Journal of Geophysical Research*, 76, 6700, <https://doi.org/10.1029/JA076i028p06700>, 1971. [72](#)

- Farris, M. H. and Russell, C. T.: Determining the standoff distance of the bow shock: Mach number dependence and use of models, *Journal of Geophysical Research*, 99, 17 681–17 690, <https://doi.org/10.1029/94JA01020>, 1994. 72
- Fazakerley, A. N., Carter, P. J., Watson, G., Spencer, A., Sun, Y. Q., Coker, J., Coker, P., Kataria, D. O., Fontaine, D., Liu, Z. X., Gilbert, L., He, L., Lahiff, A. D., Mihalčić, B., Szita, S., Taylor, M. G. G. T., Wilson, R. J., Dedieu, M., and Schwartz, S. J.: The Double Star Plasma Electron and Current Experiment, *Annales Geophysicae*, 23, 2733–2756, <https://doi.org/10.5194/angeo-23-2733-2005>, 2005. 74
- Jelínek, K., Němeček, Z., and Šafránková, J.: A new approach to magnetopause and bow shock modeling based on automated region identification, *Journal of Geophysical Research (Space Physics)*, 117, A05208, <https://doi.org/10.1029/2011JA017252>, 2012. 72, 89
- Jeřáb, M., Němeček, Z., Šafránková, J., Jelínek, K., and Měrka, J.: Improved bow shock model with dependence on the IMF strength, *Planetary and Space Science*, 53, 85–93, <https://doi.org/10.1016/j.pss.2004.09.032>, 2005. 72, 78, 93, 94
- Kruparova, O., Krupar, V., Šafránková, J., Němeček, Z., Maksimovic, M., Santolik, O., Soucek, J., Němec, F., and Merka, J.: Statistical Survey of the Terrestrial Bow Shock Observed by the Cluster Spacecraft, *Journal of Geophysical Research (Space Physics)*, 124, 1539–1547, <https://doi.org/10.1029/2018JA026272>, 2019. 72
- Lin, R. L., Zhang, X. X., Liu, S. Q., Wang, Y. L., and Gong, J. C.: A three-dimensional asymmetric magnetopause model, *Journal of Geophysical Research (Space Physics)*, 115, A04207, <https://doi.org/10.1029/2009JA014235>, 2010. 72, 78, 93
- Liu, Z., Lu, J. Y., Wang, C., Kabin, K., Zhao, J. S., Wang, M., Han, J. P., Wang, J. Y., and Zhao, M. X.: *Journal of Geophysical Research : Space Physics* A three-dimensional high Mach number asymmetric magnetopause model from global MHD simulation, pp. 5645–5666, <https://doi.org/10.1002/2014JA020961>.Received, 2015. 72
- McFadden, J. P., Carlson, C. W., Larson, D., Ludlam, M., Abiad, R., Elliott, B., Turin, P., Marckwordt, M., and Angelopoulos, V.: The THEMIS ESA Plasma Instrument and In-flight Calibration, *Scientific Studies of Reading*, 141, 277–302, <https://doi.org/10.1007/s11214-008-9440-2>, 2008. 73
- Niculescu-Mizil, A. and Caruana, R.: Obtaining Calibrated Probabilities from Boosting, in: *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence, UAI'05*, p. 413–420, AUAI Press, Arlington, Virginia, USA, 2005. 91
- Němeček, Z., Šafránková, J., and Šimůnek, J.: An Examination of the Magnetopause Position and Shape Based Upon New Observations, chap. 8, pp. 135–151, American Geophysical Union (AGU), <https://doi.org/10.1002/9781119509592.ch8>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/9781119509592.ch8>, 2020. 72
- Olshevsky, V., Khotyaintsev, Y. V., Divin, A., Delzanno, G. L., Anderzen, S., Herman, P., Chien, S. W. D., Avanov, L., and Markidis, S.: Automated classification of plasma regions using 3D particle energy distribution, *arXiv e-prints*, arXiv:1908.05715, 2019. 95
- Paschmann, G., Haaland, S. E., Phan, T. D., Sonnerup, B. U. Ö., Burch, J. L., Torbert, R. B., Gershman, D. J., Dorelli, J. C., Giles, B. L., Pollock, C., Saito, Y., Lavraud, B., Russell, C. T., Strangeway, R. J., Baumjohann, W., and Fuselier, S. A.: Large-Scale Survey of the Structure of the Dayside Magnetopause by MMS, *Journal of Geophysical Research (Space Physics)*, 123, 2018–2033, <https://doi.org/10.1002/2017JA025121>, 2018. 72
- Pollock, C., Moore, T., Jacques, A., Burch, J., Gliese, U., Saito, Y., Omoto, T., Avanov, L., Barrie, A., Coffey, V., Dorelli, J., Gershman, D., Giles, B., Rosnack, T., Salo, C., Yokota, S., Adrian, M.,

Aoustin, C., Auletti, C., Aung, S., Bigio, V., Cao, N., Chandler, M., Chornay, D., Christian, K., Clark, G., Collinson, G., Corris, T., De Los Santos, A., Devlin, R., Diaz, T., Dickerson, T., Dickson, C., Diekmann, A., Diggs, F., Duncan, C., Figueroa-Vinas, A., Firman, C., Freeman, M., Galassi, N., Garcia, K., Goodhart, G., Guerro, D., Hageman, J., Hanley, J., Hemminger, E., Holland, M., Hutchins, M., James, T., Jones, W., Kreisler, S., Kujawski, J., Lavu, V., Lobell, J., LeCompte, E., Lukemire, A., MacDonald, E., Mariano, A., Mukai, T., Narayanan, K., Nguyen, Q., Onizuka, M., Paterson, W., Persyn, S., Piepgrass, B., Cheney, F., Rager, A., Raghuram, T., Ramil, A., Reichenthal, L., Rodriguez, H., Rouzaud, J., Rucker, A., Saito, Y., Samara, M., Sauvaud, J. A., Schuster, D., Shappirio, M., Shelton, K., Sher, D., Smith, D., Smith, K., Smith, S., Steinfeld, D., Szymkiewicz, R., Tanimoto, K., Taylor, J., Tucker, C., Tull, K., Uhl, A., Vloet, J., Walpole, P., Weidner, S., White, D., Winkert, G., Yeh, P. S., and Zeuch, M.: Fast Plasma Investigation for Magnetospheric Multiscale, *Scientific Studies of Reading*, 199, 331–406, <https://doi.org/10.1007/s11214-016-0245-4>, 2016.

74

Rème, H., Aoustin, C., Bosqued, J. M., Dandouras, I., Lavraud, B., Sauvaud, J. A., Barthe, A., Bouyssou, J., Camus, T., Coeur-Joly, O., Cros, A., Cuvilo, J., Ducay, F., Garbarowitz, Y., Medale, J. L., Penou, E., Perrier, H., Romefort, D., Rouzaud, J., Vallat, C., Alcaydé, D., Jacquy, C., Mazelle, C., D’Uston, C., Möbius, E., Kistler, L. M., Crocker, K., Granoff, M., Mouikis, C., Popecki, M., Vosbury, M., Klecker, B., Hovestadt, D., Kucharek, H., Kuenneth, E., Paschmann, G., Scholer, M., Sckopke, N., Seidenschwang, E., Carlson, C. W., Curtis, D. W., Ingraham, C., Lin, R. P., McFadden, J. P., Parks, G. K., Phan, T., Formisano, V., Amata, E., Bavassano-Cattaneo, M. B., Baldetti, P., Bruno, R., Chionchio, G., di Lellis, A., Marcucci, M. F., Pallochia, G., Korth, A., Daly, P. W., Graeve, B., Rosenbauer, H., Vasyliunas, V., McCarthy, M., Wilber, M., Eliasson, L., Lundin, R., Olsen, S., Shelley, E. G., Fuselier, S., Ghielmetti, A. G., Lennartsson, W., Escoubet, C. P., Balsiger, H., Friedel, R., Cao, J. B., Kovrazhkin, R. A., Papamastorakis, I., Pellat, R., Scudder, J., and Sonnerup, B.: First multispacecraft ion measurements in and near the Earth’s magnetosphere with the identical Cluster ion spectrometry (CIS) experiment, *Annales Geophysicae*, 19, 1303–1354, <https://doi.org/10.5194/angeo-19-1303-2001>, 2001. 74

Russell, C. T., Anderson, B. J., Baumjohann, W., Bromund, K. R., Dearborn, D., Fischer, D., Le, G., Leinweber, H. K., Leneman, D., Magnes, W., Means, J. D., Moldwin, M. B., Nakamura, R., Pierce, D., Plaschke, F., Rowe, K. M., Slavin, J. A., Strangeway, R. J., Torbert, R., Hagen, C., Jernej, I., Valavanoglou, A., and Richter, I.: The Magnetospheric Multiscale Magnetometers, *Scientific Studies of Reading*, 199, 189–256, <https://doi.org/10.1007/s11214-014-0057-3>, 2016. 74

Shue, J. H., Chao, J. K., Fu, H. C., Russell, C. T., Song, P., Khurana, K. K., and Singer, H. J.: A new functional form to study the solar wind control of the magnetopause size and shape, *Journal of Geophysical Research*, 102, 9497–9512, <https://doi.org/10.1029/97JA00196>, 1997. 72

Wang, Y., Sibeck, D. G., Merka, J., Boardsen, S. A., Karimabadi, H., Sipes, T. B., Šafránková, J., Jelínek, K., and Lin, R.: A new three-dimensional magnetopause model with a support vector regression machine and a large database of multiple spacecraft observations, *Journal of Geophysical Research (Space Physics)*, 118, 2173–2184, <https://doi.org/10.1002/jgra.50226>, 2013.

72, 94

## Chapter Summary

- In this chapter, we apply a gradient boosting algorithm to the in-situ data of the [THEMIS B](#) spacecraft to provide an automatic classification method of the 3 main near-Earth regions: the magnetosphere, the magnetosheath and the solar wind.
- This method outcomes the previous existing region classification methods based on the setting of manual, empirical thresholds on a reduced number of physical parameters.
- We test successfully the algorithm trained with [THEMIS B](#) data on the data provided by the other [THEMIS](#) spacecraft, by [MMS](#) and by Double Star TC1.
- After a retraining phase to take into account the differences induced by its polar orbit, we adapt the classifier to the data provided by Cluster.
- By reconsidering the set of features in the dataset and adding the lunar wake as a potential visitable region, we even manage to elaborate an equivalent of our method applied to the lunar orbit of [ARTEMIS](#).
- We use the region classifier to elaborate one of the most exhaustive and accessible magnetopause and bow shock crossings catalogs. These lists can be used for further additional studies of the properties of these two boundaries or as a basis for the elaboration of the automatic detection method of the in-situ signature of the small-scale physical processes of the near-Earth environment.





# Chapter 5

## Statistical analysis of the magnetopause shape and location

*There's much to do and many unknowns to the horizon...*

Herodotos

### Contents

---

<b>5.1 Introduction</b>	<b>102</b>
<b>5.2 A brief insight on the magnetopause shape and location models</b>	<b>103</b>
<b>5.3 Statistical analysis of the magnetopause crossing</b>	<b>106</b>
5.3.1 Dataset	106
5.3.2 The magnetopause stand-off distance	110
5.3.3 Evidencing the asymmetries	112
5.3.4 Dependencies of the flaring coefficients	114
<b>5.4 Fitting a new magnetopause model</b>	<b>118</b>
5.4.1 Fit	118
5.4.2 Evaluation and comparison with other models	119
5.4.3 Characteristics of the model	120
5.4.4 From a static to a dynamic model	122
<b>5.5 Nature of the near-cusp magnetopause</b>	<b>123</b>
5.5.1 Different boundaries of the polar cusps	123
5.5.2 Shape of the near-cusp magnetopause	128
<b>5.6 Conclusion</b>	<b>133</b>
<b>5.7 Bibliography</b>	<b>134</b>

---

## 5.1 Introduction

First observed by Cahill and Amazeen [1963], the Earth magnetopause has from then on been extensively studied from both in-situ measurements and numerical simulations (Hasegawa [2012] and references therein).

Using the observations of IMP and assuming a pressure balance between the solar wind and the magnetosphere without additional coupling, Fairfield [1971] and Formisano [1979] investigated the location and shape of the magnetopause and elaborated the very first empirical and analytical models of the magnetopause shape and location in the form of a quadric surface.

These observations also showed that reconnection eroded the magnetosphere in a location dependent on the orientation of the IMF resulting in an earthward motion and in the decrease of the level of flaring. Such finding was considered in the modeling of the magnetopause surface when Sibeck et al. [1991] and Petrinec et al. [1991] considered the solar wind dynamic pressure and the IMF  $B_z$  component as the main parameters of their models.

From then on, numerous analytical empirical models based on the solar wind dynamic pressure and the IMF  $B_z$  were developed [Petrinec and Russell, 1993; Roelof and Sibeck, 1993]. These models, that continued relying on the form of a quadric surface, were fitted using ISEE magnetopause crossings and progressively considered the extension of the magnetopause in the night-side [Petrinec and Russell, 1996]. With the measurements of the IMP8, ISEE 1 and ISEE 2 spacecraft used simultaneously, Shue et al. [1997] improved the accuracy of the magnetopause models by considering an inverse trigonometric function still in use in the most recent existing models.

All of the previously mentioned models used data from spacecraft that had an equatorial orbit and thus supposed a symmetry around the X axis. This symmetry was questioned by the investigation of Sotirelis and Meng [1999] that evidenced an influence of the Earth dipole tilt angle in accordance with the findings of Tsyganenko [1998]. The influence of the dipole tilt angle was later confirmed by the Hawkeye observation of Boardsen et al. [2000], Eastman et al. [2000] and the Interball observations of Šafránková et al. [2002]. They also observed a depletion of the selected boundary in the near-cusp region that was interpreted as the confirmation of an indented magnetopause. As already said in the chapter 1, this indentation can be seen as the logical continuous extension between the dayside and the nightside magnetopause in the case of a boundary locally tangential everywhere.

Using measurements of LANL and GOES, Kuznetsov and Suvorova [1998] evidenced the presence of a dawn-dusk asymmetry later confirmed by Dmitriev and Suvorova [1999] through the help of ANN. This asymmetry was later found to be linked to the aberration of the solar wind due to Earth orbital motion by Šafránková et al. [2002].

At the light of these new findings, Lin et al. [2010] fitted their magnetopause surface model that took into consideration the azimuthal asymmetry induced by the dipole tilt angle and also noticed a dawn-dusk asymmetry despite of the correction brought to the data in order to remove the aberration. Some years later, Wang et al. [2013] developed their own model by applying a support vector regression<sup>1</sup> to the combined crossings of 23 different spacecraft. Without assuming anything but symmetries on the magnetopause, they fell back on the dependencies on the dynamic pressure and the IMF  $B_z$ , proving in the process the potential of machine learning in the frame of such a regression problem. However, neither their data nor their model is shared, the study is consequently hardly reproducible and the results not usable.

Going further in our knowledge of the magnetopause, Dusik et al. [2010] and Grygorov et al. [2017] used THEMIS data and pointed out that an increasing IMF  $B_x$  pulled the magnetopause sunward while the MHD simulations of Liu et al. [2015] indicated this component would rather also contribute to the north-south asymmetry already induced by the dipole tilt angle. The latter also showed that increasing IMF  $B_y$  component twisted the magnetopause in the direction of the IMF.

<sup>1</sup>Although not presented in this thesis, *support vector machines* [Cortes and Vapnik, 1995] are another family of machine learning algorithms that can be used either for classification or regression purposes.

Despite of this important number of studies, the question of the magnetopause shape and location is not over yet and still numerous are the number of remaining questions:

1. First of all, the influence of the IMF cone angle is still unclear and the different conclusions drawn by [Dusik et al. \[2010\]](#) and [Liu et al. \[2015\]](#) calls for further investigation.
2. Second, [Liu et al. \[2015\]](#) evidenced the influence of the IMF clock angle numerically and [Dmitriev and Suvorova \[1999\]](#) inferred an influence of  $B_y$  on the level of flaring. Nevertheless this influence still lacks of observational evidences and has not yet been considered by any empirical analytical model.
3. Third, all of the magnetopause surface models considered data in the range  $X > -40 R_E$  and thus extrapolated to the far nightside whereas now spacecraft at lunar orbit such as [ARTEMIS](#) could provide useful information on the nature of the magnetopause at these distances.
4. Fourth, different observations of the polar cusps crossings led to different conclusions regarding the shape of the magnetopause in this region. If [[Boardsen et al., 2000](#); [Šafránková et al., 2002](#); [Šafránková et al., 2005](#)] suggested the existence of an indentation, [[Lavraud et al., 2004a](#); [Zhou and Russell, 1997](#)] suggested the opposite. [Eastman et al. \[2000\]](#) inferred that these divergences could be explained by the definition of the magnetopause considered by each study. Using Cluster data polar cusp events, [Lavraud et al. \[2004b\]](#) suggested that the indentation was observed because the inner boundary was considered instead of the actual magnetopause current sheet for which no specific depletion was noticed. Consequently, the question of the shape of the magnetopause in the near-cusp region is still open and requires additional investigations.

In this chapter, we exploit the 14 996 complete magnetopause crossings detected in the previous chapter to perform a statistical analysis of the magnetopause shape and location that result in the fit of an analytical empirical model.

After a brief presentation of three different existing magnetopause surface models, the two most recent [[Lin et al., 2010](#); [Liu et al., 2015](#)] and the most used [[Shue et al., 1997](#)], we will study the magnetopause stand-off distance, asymmetries and levels of flaring from a statistical point of view. The conclusions of the study will then be used to fit a new analytical empirical model of the magnetopause surface. The last section will investigate the nature of the polar cusps events of our dataset and consider the notion of indentation of the magnetopause in this region.

## 5.2 A brief insight on the magnetopause shape and location models

Intuitively, the two parameters we expect to influence the most the location and shape of the magnetopause are the dynamic pressure  $P_{dyn}$  and the  $B_z$  component of the solar wind and IMF. This is because of the definition of the magnetopause at first order and because of the influence of  $B_z$  on magnetic reconnection and the associated erosion as presented in the chapter 1.

Assuming an axial symmetry around the X axis, and after the statistical study of the influence of  $P_{dyn}$  and  $B_z$  on the crossings of the magnetopause by the IMP8, ISEE 1 and ISEE 2 spacecraft, [Shue et al. \[1997\]](#) developed their model defined as:

$$r = r_0 \left( \frac{2}{1 + \cos(\theta)} \right)^\alpha \quad (5.1)$$

$$r_0 = [a_0 + a_1 \tanh(a_2(B_z + a_3))] P_{dyn}^{a_4} \quad (5.2)$$

$$\alpha = (a_5 - a_6 B_z)(1 + a_7 P_{dyn}) \quad (5.3)$$

Where the value of  $\alpha$  can either describe an open ( $\alpha > 0.5$ ) or a closed ( $\alpha < 0.5$ ) magnetopause in the nightside.  $r_0$  represents the magnetopause stand-off distance at  $\theta = 0$  and the values of the

coefficients  $a_i$  in equations 5.3 and 5.2 were obtained through the fit of this model to their in-situ observations.

This model was questioned by the evidence of the possible dawn-dusk, the north-south and the azimuthal asymmetries [Boardsen et al., 2000; Kuznetsov and Suvorova, 1998; Sotirelis and Meng, 1999]. At the same time, the observations of Boardsen et al. [2000], Eastman et al. [2000] and Šafránková et al. [2002] indicated a possible indentation of the magnetopause in the near-cusp regions. These possible changes in the analytical expression of the magnetopause and shape were considered by Lin et al. [2010] who adapted the Shue et al. [1997] model by setting:

$$r = r_0 \left( \cos\left(\frac{\theta}{2}\right) + a_5 \sin(2\theta)(1 - e^{-\theta}) \right)^{\beta_0 + \beta_1 \sin(\phi) + \beta_2 \cos(\phi) + \beta_3 \cos(\phi)^2} + C(e^{d_n \psi_n^{a_{21}}} + e^{d_s \psi_s^{a_{21}}}) \quad (5.4)$$

$$r_0 = a_0 (P_{dyn} + P_m)^{a_1} \left( 1 + a_2 \frac{e^{a_3 B_z} - 1}{e^{a_4 B_z} + 1} \right) \quad (5.5)$$

$$\beta_0 = a_6 + a_7 \frac{e^{a_8 B_z} - 1}{e^{a_9 B_z} + 1} \quad (5.6)$$

$$\beta_1 = a_{10} \quad (5.7)$$

$$\beta_2 = a_{11} + a_{12} \gamma \quad (5.8)$$

$$\beta_3 = a_{13} \quad (5.9)$$

$$C = a_{14} (P_{dyn} + P_m)^{a_{15}} \quad (5.10)$$

$$d_{n,s} = a_{16} \pm a_{17} \gamma + a_{18} \gamma^2 \quad (5.11)$$

$$\psi_n = \arccos(\cos(\theta) \cos(\theta_n) + \sin(\theta) \sin(\theta_n) \cos(\phi)) \quad (5.12)$$

$$\psi_s = \arccos(\cos(\theta) \cos(\theta_s) + \sin(\theta) \sin(\theta_s) \cos(\phi - \pi)) \quad (5.13)$$

$$\theta_{n,s} = a_{19} \pm a_{20} \gamma \quad (5.14)$$

Where the values of the coefficients  $a_i$  are determined through fitting the model to in-situ measurements of magnetopause crossings. In the case of Lin et al. [2010], the value of the coefficients were determined using 2708 magnetopause crossings of 12 different spacecraft and are summarized in the Table 9 of the aforementioned paper<sup>2</sup>.

Here, the first term of the first equation is an extension of the model proposed by Shue et al. [1997] and setting  $a_5$  to 0 brings us back to the expression of  $r$  in equation 5.1.  $\beta_0$  controls the tail flaring and  $\beta_1$ ,  $\beta_2$  and  $\beta_3$  control the dawn-dusk, the north-south and the ( $Y-Z$ ) asymmetries respectively. The second term is the consideration of the cusps indentation and  $C$ ,  $d_n$  ( $d_s$ ),  $a_{21}$  and  $\theta_n$  ( $\theta_s$ ) control the depth, the scope, shape and location of the northern (southern cusp) indentation.

In addition to  $P_{dyn}$  and  $B_z$ , the model here depends on the solar wind magnetic pressure  $P_m$  and the dipole tilt angle  $\gamma$ . The consideration of  $P_m$  simply adds a term in the pressure balance that rules the magnetopause location and this is why its consideration in equations 5.5 and 5.10 appears as a sum with the dynamic pressure  $P_{dyn}$ . It is worth noting here that even if the most recent models take it into account, the influence of the magnetic pressure remains almost negligible in comparison to the effect of a changing  $P_{dyn}$ .

As we expect the North-South asymmetry and the position and shape of the cusps to be influenced by the dipole tilt, it is not surprising to notice its contribution in the expressions of  $\beta_2$ ,  $d_{n,s}$  and  $\theta_{n,s}$ .

This model provides a detailed prediction of the magnetopause location and shape for an important set of solar wind parameters. Nevertheless, they only considered the main solar wind parameters likely to affect the magnetopause: the magnetic and dynamic pressures and the IMF  $B_z$  component. They also developed their model without confirming or invalidating the influence

<sup>2</sup>Here again, the events used by the authors to develop their model is not shared and their study is consequently hardly reproducible in the same conditions.

of other possible parameters. For instance, the possible stretching of the magnetopause [Dmitriev and Suvorova \[1999\]](#) induced for different IMF clock angles and the eventual influence of the IMF cone angle is not accounted in the model.

Additionally, the cusp geometry expressed in the second term of equation 5.4 as an additive term has the drawback of being non-zero at the stand-off position of the magnetopause and in the nightside, biasing the interpretability we can have of  $r_0$  that cannot anymore be considered as the representation of the stand-off position.

Using the magnetopause detected in their MHD simulations, [Liu et al. \[2015\]](#) adapted the [Shue et al. \[1997\]](#) model by setting:

$$r = r_0 \left( \frac{2}{1 + \cos(\theta)} \right)^\alpha (1 - 0.1C \cos(\phi)^2) \quad (5.15)$$

$$\alpha = \alpha_0 + [\alpha_\phi + \delta_\alpha \text{sgn}(\cos(\phi))] \cos(2(\phi - \omega)) + \alpha_z \cos(\phi) \quad (5.16)$$

$$C = e^{-\frac{|\theta - l_n|}{w}} (1 + \text{sgn}(\cos(\phi))) + e^{-\frac{|\theta - l_s|}{w}} (1 + \text{sgn}(\cos(-\phi))) \quad (5.17)$$

$$r_0 = (a_0 + a_1 \tanh[a_2(B_z + a_3)])(P_{dyn} + P_m)^{a_4} \quad (5.18)$$

$$l_{n,s} = (a_5 + a_6 \tanh[a_7(B_z + a_8)])(1 \mp a_9 \gamma) \quad (5.19)$$

$$w = (a_{10} + a_{11} \log(P_{dyn}))(1 + a_{12} \gamma^2) \quad (5.20)$$

$$\alpha_0 = (a_{13} - a_{14} \tanh[a_{15}(B_z - a_{16})])(1 + a_{17} \log P_{dyn}) \quad (5.21)$$

$$\alpha_\phi = a_{18} + a_{19} \tanh[a_{20}(|B_z| - a_{21})](1 - a_{22} \log P_{dyn}) \quad (5.22)$$

$$\delta_\alpha = a_{23} \tanh(a_{24} B_x) \text{sgn}(B_x) \quad (5.23)$$

$$\omega = \arctan[a_{25} \left( \frac{B_y}{B_z} \right) (B_y^2 + B_z^2)^{a_{26}}] \quad (5.24)$$

$$\alpha_z = a_{27} \tanh(a_{28} \gamma) \quad (5.25)$$

In this case, the flaring is controlled by  $\alpha_0$ ,  $\alpha_\phi$  controls the azimuthal asymmetry between low and high latitudes,  $\delta_\alpha$  controls the north-south asymmetry induced by the IMF cone angle,  $\alpha_z$  describes the asymmetry induced by the tilt influence and  $\omega$  is the stretching direction of the magnetopause in the (Y-Z) plane induced by a changing clock angle. The consideration of the cusp indentation is here expressed by  $l_{n,s}$  and  $w$  that control the location and the angular width of the northern and southern cusps indentation respectively.

All of the coefficients  $a_i$ , expressed from equation 5.15 to 5.25, have been determined through the fitting of the model to MHD simulations of the magnetopause under various upstream solar wind conditions and the quality of the model has been assessed through the comparison of the predicted magnetopause to the 2168 in-situ observed crossings from the NASA magnetopause database<sup>3</sup>.

In comparison [Lin et al. \[2010\]](#), this model has the advantage to take into account the influence of the three components of the IMF while predicting the magnetopause with a slightly increased accuracy [[Liu et al., 2015](#)]. Nevertheless, this model assumes a dawn-dusk symmetry in opposition to what was assumed by [Lin et al. \[2010\]](#) and the cone angle is here set to only influence the north south asymmetry while having no influence on the stand-off distance  $r_0$  or the flaring  $\alpha_0$ , which is opposed to the observations made by [Dusik et al. \[2010\]](#) or [Grygorov et al. \[2017\]](#). Moreover, the crossings used for the observational comparison were by far mostly made in the north hemisphere and dayside region which thus gives poor evidence on how the model performs in the southern hemisphere or in the far nightside. Additionally, the perturbation induced by the consideration of the indentation depends on the azimuth angle and even if it vanishes in the far nightside (e.g  $\theta \sim \pi$ ), it is still non-zero on the X axis and thus results in a non-unicity of the stand-off distance.

<sup>3</sup><https://omniweb.gsfc.nasa.gov/ftpbrowser/magnetopause/Database.html>

The projections in the  $(X-Y)$ ,  $(X-Z)$  and  $(Y-Z)$  planes of the three models we presented in this section are represented in the Figure 5.1 for a dynamic pressure of 2 nPa and an IMF  $B_z$  component of  $-2$  nT. The advantages and disadvantages presented by these models indicate that the question of predicting accurately the position and shape of the magnetopause for a given set of solar wind and seasonal parameters is still open. For instance, the control parameters are still under debate and there is still uncertainties associated to the magnetopause behavior in the far nightside.

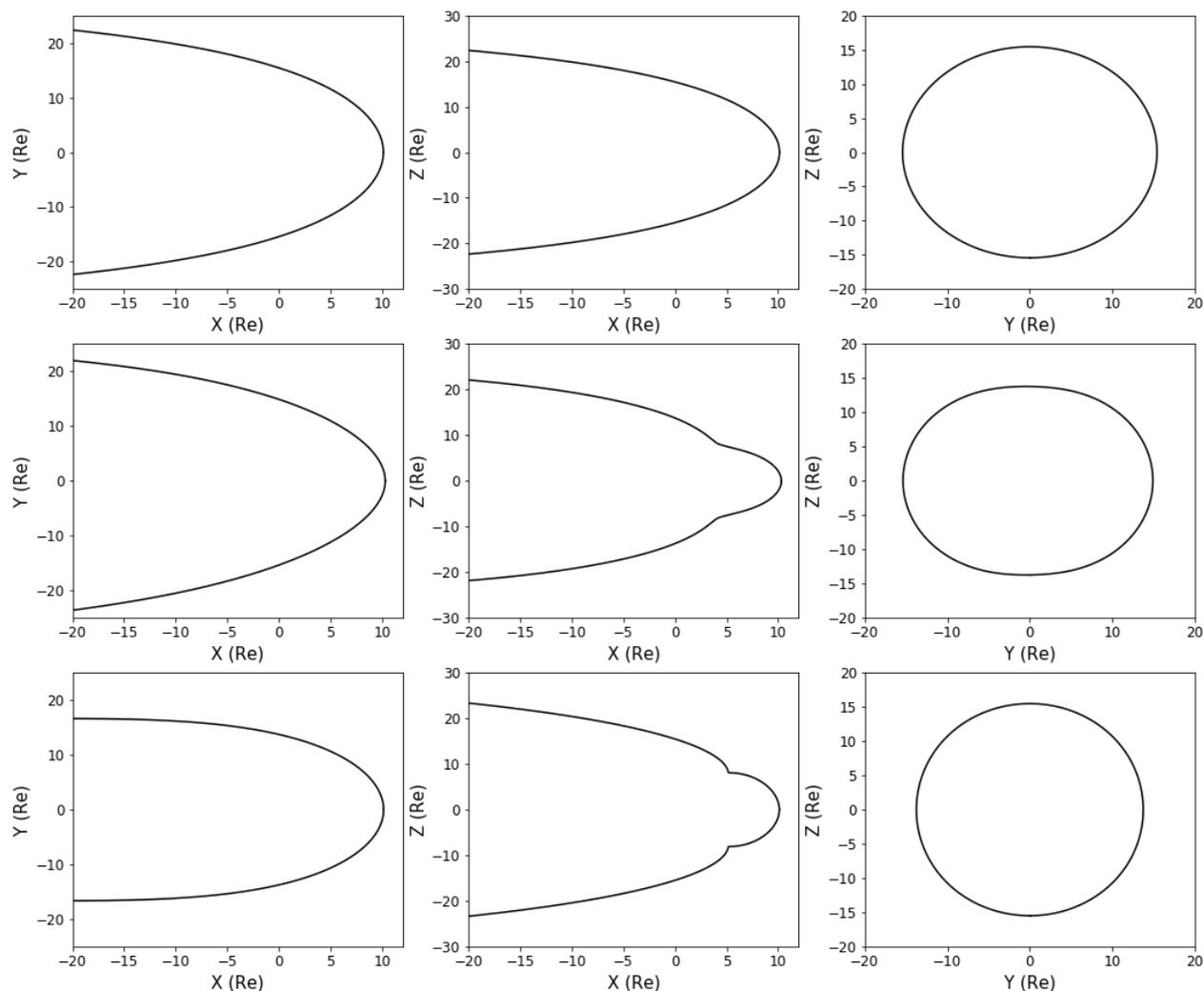


Figure 5.1: Projections in the  $(X-Y)$  (*left column*), in the  $(X-Z)$  (*middle column*) and in the  $(Y-Z)$  (*right column*) planes of the magnetopause models of Shue et al. [1997] (*top*), Lin et al. [2010] (*middle*) and Liu et al. [2015] (*bottom*) computed for a dynamic pressure of 2 nPa and an IMF  $B_z$  component of  $-2$  nPa.

## 5.3 Statistical analysis of the magnetopause crossing

### 5.3.1 Dataset

The detection of 14 996 complete magnetopause crossings in Chapter 4 offers a unique opportunity to provide a statistical insight on various geometrical properties of the magnetopause as functions of the upstream solar wind conditions. The 1 hour crossings we constructed in chapter 4 are here reduced to 10 minutes crossings by using the process described in 4.8.1. Each crossing is then associated to a set of solar wind upstream conditions obtained with a temporal shift of OMNI data.

We determine this shift time by applying the two-step propagation algorithm exposed in Šafránková et al. [2002]:

- for a given event at the GSM position  $X_0$  at the time  $t_0$ , we assume a solar wind velocity of  $V_0 = 400$  km/s and determine a first time lag  $\delta t$  from the position difference between OMNI and the spacecraft along the X axis:  $\delta t = (X_{\text{OMNI}} - X_0)/V_0$
- We measure  $V_1$ , the OMNI velocity at the shifted time  $t_0 - \delta t$  to determine the final time lag  $\delta t' = (X_{\text{OMNI}} - X_0)/V_1$
- We average OMNI data in the 5 minutes centered interval around  $t_0 - \delta t'$

Using this method, we removed 1 815 crossings that had no available upstream solar wind condition.

Such a dataset can be enhanced in size through the addition of the 2 168 online crossings of the missions IMP, ISEE, Geotail, Prognoz, Hawkeye, AMPTE, Explorer and OGO (<ftp://nssdcftp.gsfc.nasa.gov/spacecraftdata/magnetopausecrossings>) that were used in the comparison of Liu et al. [2015]’s model to observational data. The Hawkeye crossings were also used by Lin et al. [2010], especially when looking at the near-cusp magnetopause. The summary of such crossings and the mission they are associated with are shown in Table 5.1. When necessary, we will distinguish these events from the one detected by our region classifier by denominating them as *crossings from older missions*.

IMP	ISEE	Geotail	Prognoz	Hawkeye	AMPTE	Explorer	OGO	Total
75	333	76	91	1484	60	17	32	<b>2168</b>

Table 5.1: Summary of the 2168 crossings we added to the dataset

The combination of the two lists results in an ensemble of 15 349 magnetopause crossings distributed on 17 different spacecraft. If at first sight, we roughly have as much events as Wang et al. [2013], they considered an important number of partial crossings and we thus expect their dataset to offer a narrower range of solar wind and seasonal conditions at all altitudes and longitudes.

Having merged the two lists, we limit the dataset to the crossings for which  $X > -70$  Re, the minimal distance for which we detected ARTEMIS crossings. We correct the GSM position of each of the obtained 15 349 magnetopause crossings by removing the aberration due to the Earth’s revolution using a similar approach than what was done in Lin et al. [2010] and Boardsen et al. [2000] and assuming a revolution velocity of 30 km/s.

The histograms of the associated upstream solar wind parameters are shown in the Figure 5.2. For each panel, we notice a similar distribution to the one we show in the chapter 1 for OMNI. This indicates that the greatest part of the crossings occurred under normal solar wind conditions and these are the conditions under which we expect the statistics we are about to perform to be the most reliable.

As half of our crossings have been measured by spacecraft with relatively low apogees ( $\sim 12$  Re), it is important to make sure that our dataset is free from any orbital bias. Such limitations could indeed affect the importance of the different dependencies we will be focusing on in the next sections as discussed in Němeček et al. [2020].

To do so, we show in Figure 5.3 the projections in the  $(X - Y)$  and the  $(X - Z)$  plane of the crossings corrected position. The grey shading represents the time during which the different spacecraft were at given coordinates  $(X, Y)$  and  $(X, Z)$ . In both cases and for each value of X, the crossing with the highest Y or Z is located far from the maximal Y or Z reached by the spacecraft during the orbit. This suggests the crossings we selected are not limited by the orbit of the spacecraft we consider in this study.

Despite of having an even distribution in the  $(X - Y)$  plane, the majority of the high-latitude crossings are detected in the northern hemisphere. We balance this distribution by reverting the Z

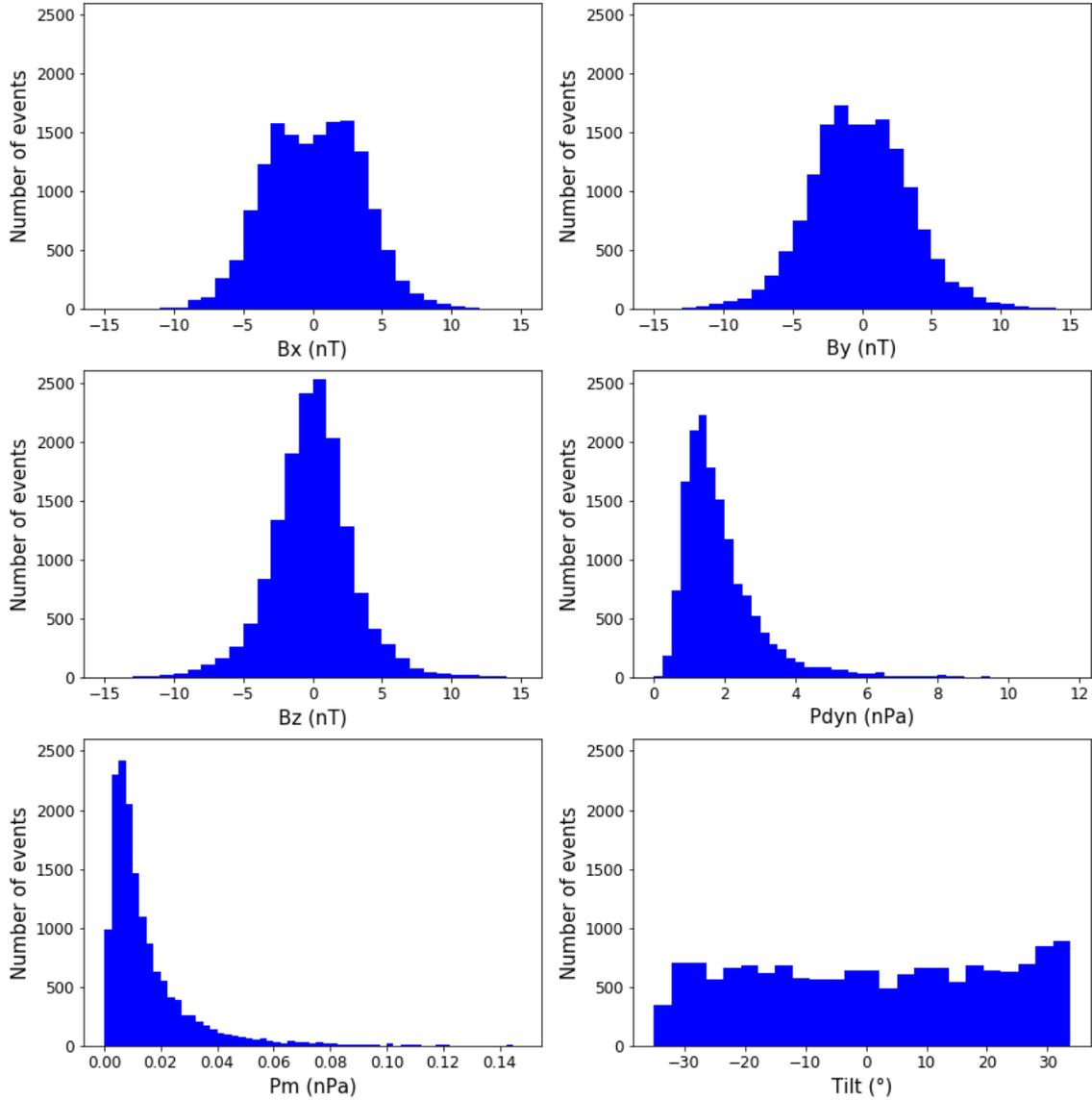


Figure 5.2: Histogram of the solar wind parameters of the 15349 magnetopause crossings: the three magnetic field components,  $B_x$  (top left),  $B_y$  (top right),  $B_z$  (middle left), the dynamic pressure  $P_{dyn}$  (middle right), the magnetic pressure  $P_m$  (bottom left) and the dipole tilt angle (bottom right).

coordinate and the tilt angle  $\gamma$  of every crossing, in a similar way than was done in Wang et al. [2013]. Assuming de facto that the northern summer hemisphere of the magnetopause is similar to the southern winter hemisphere:  $r(X, Y, Z, \gamma) = r(X, Y, -Z, -\gamma)$ .

The absence of bias is also confirmed by Figure 5.4 that represents the histograms of the solar wind parameters for both the entire dataset and the crossings measured by the spacecraft that have a high apogee (above 12  $R_e$ , corresponding to the crossings of any spacecraft but Double Star, MMS and THEMIS A, D and E.). Having similar blue and red histograms for each panel ensures no orbital bias is introduced whatever solar wind parameter is considered.

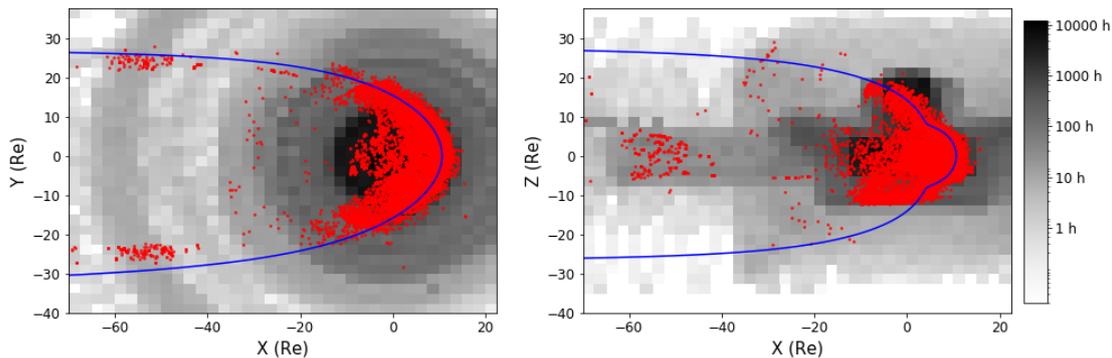


Figure 5.3: Projection in the  $(X - Y)$  (left) and in the  $(X - Z)$  (right) GSM plane of the 15349 magnetopause crossings (red dots). The gray shading represents the time spent by all of the spacecraft in a given region of the  $(X - Y)$  (resp.  $(X - Z)$ ) plane. The blue line represents Lin et al. [2010] magnetopause model with a dynamic pressure of 2 nPa and a null  $B_z$ .

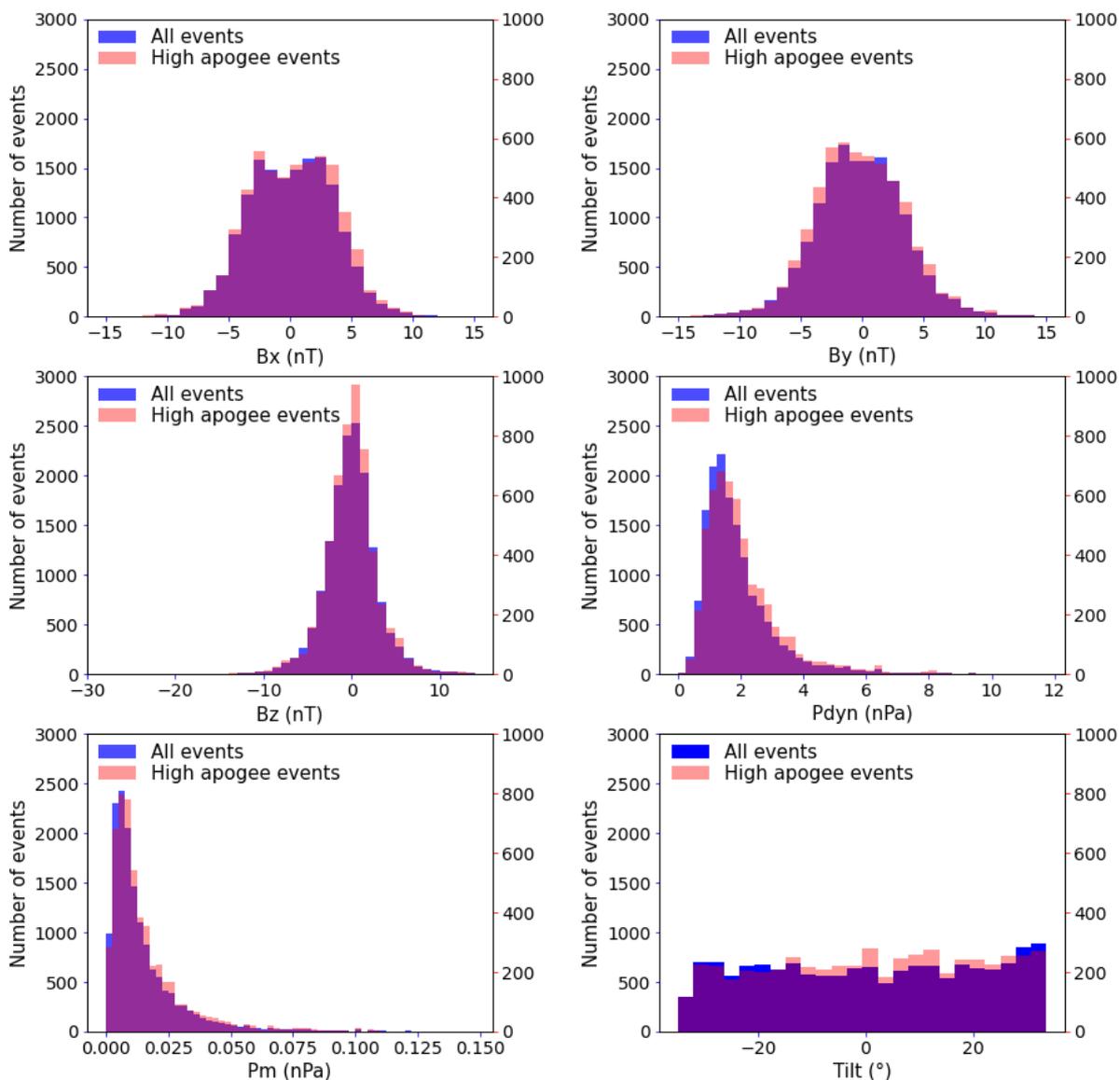


Figure 5.4: Histogram of the solar wind parameters of the 15 349 magnetopause crossings. Each panel is the same than in Figure 5.2. The red here bins show the same distribution for the events measured by high apogee spacecraft

In the introduction, we mentioned that the question of the magnetopause indentation is still open and a non-negligible part of the events of our dataset might be inner boundary crossings instead of being crossings of the actual magnetopause. Consequently, the crossings of our dataset that occurred in the near-cusp region will be studied separately from the other magnetopause crossings in the next section and the studies that we will show in the three following subsections have been performed without the consideration of the high latitude crossings that were likely to occur near the northern or the southern cusp.

Considering the expression of the indentation expressed by [Lin et al. \[2010\]](#) in equations 5.10 to 5.14, we define those so-called "out of cusp" crossings as the events for which the spherical coordinates  $\theta$  and  $\phi$  verify:

$$\begin{cases} (\theta - \theta_n)^2 + \phi^2 \geq \left(-\frac{1}{d_n}\right)^{\frac{2}{a_{21}}} & \text{if } Z \geq 0 \\ (\theta - \theta_s)^2 + \phi^2 \geq \left(-\frac{1}{d_s}\right)^{\frac{2}{a_{21}}} & \text{if } Z \leq 0 \end{cases} \quad (5.26)$$

The application of such criteria separates the dataset into 29 077 *out of cusp* crossings and 1 621 so-called *near-cusp* crossings.

### 5.3.2 The magnetopause stand-off distance

We study the magnetopause stand-off distance by selecting the 275 events for which  $\theta < 7.50^\circ$  and  $Z > 0$  and approximate the magnetopause stand-off distance  $r_0$  of these crossings by their actual radial distance  $r$ . The reason that explains the second selection criteria we make is that the crossings selected with the criteria on the zenith angle  $\theta$  correspond to a very narrow latitude and longitude band. It is then of no use here to consider the events in the southern hemisphere as they will just repeat the information brought by their northern hemisphere counterpart.

Naturally, we expect the total pressure  $P_{dyn} + P_m$  to be the feature that has the greatest influence on the stand-off distance. Figure 5.5 represents the radial distance of all of the events as a function of the total pressure. The stand-off distance  $r_0$  here appears to have a clear and consistent power-law dependency on the total pressure that was already exhibited in observations by [Shue et al. \[1997\]](#) and references therein and numerically by [Liu et al. \[2015\]](#). We exhibit this dependency by fitting  $r_0$  to the power law  $a_0(P_{dyn} + P_m)^{a_1}$  where  $a_0$  represents the stand-off distance of the magnetopause at 1 nPa and  $a_1$  the exponent of the power-law. The result of such fit is represented by the solid blue line and the grey interval that represents the 1-sigma confidence interval. The obtained values of  $a_0$  and  $a_1$  are shown in the top right corner of the Figure.  $a_0$  represents the stand-off distance at 1 nPa and the obtained value is thus consistent with the typical value we expect for the magnetopause nose. We also find an exponent value of  $-0.161$  for  $a_1$ , which is very close to the theoretical  $-\frac{1}{6}$  exposed in the chapter 1 and close to the values obtained by [Shue et al. \[1997\]](#), [Lin et al. \[2010\]](#) and [Liu et al. \[2015\]](#).

The dependency on the solar wind ram and magnetic pressure is so strong that studying now the dependency of the stand-off distance on the IMF components must be done with care. To cope with it, we separate the 275 events into 2 nT wide bins for each of the IMF components and fit  $r_0$  as a power law of the total pressure,  $r_0 = a_0(P_{dyn} + P_m)^{a_1}$  for each bin. We limit the bins for each component by looking at the total number of each event per defined bin. This still offers a wide range of upstream IMF conditions of the same order as the ones used in the studies led by [Roelof and Sibeck \[1993\]](#) and [Petrinec and Russell \[1993\]](#).

We represent the evolution of the fitted  $a_0$ , the stand-off distance at 1 nPa, as a function of  $B_x$  and  $B_y$  with the black circles in the two panels of Figure 5.6. In both cases, we notice an almost constant evolution of  $a_0$  that indicates no particular dependency of the stand-off distance for these two components. In the  $B_x$  case, this is different from the findings of [Dusik et al. \[2010\]](#) or [Grygorov et al. \[2017\]](#) who both exhibited a sunward motion of the magnetopause for a radial IMF but consistent with the study of [Liu et al. \[2015\]](#) for which the influence of the IMF  $B_x$  was almost negligible. The differences with the former might be explained by the spatial distribution

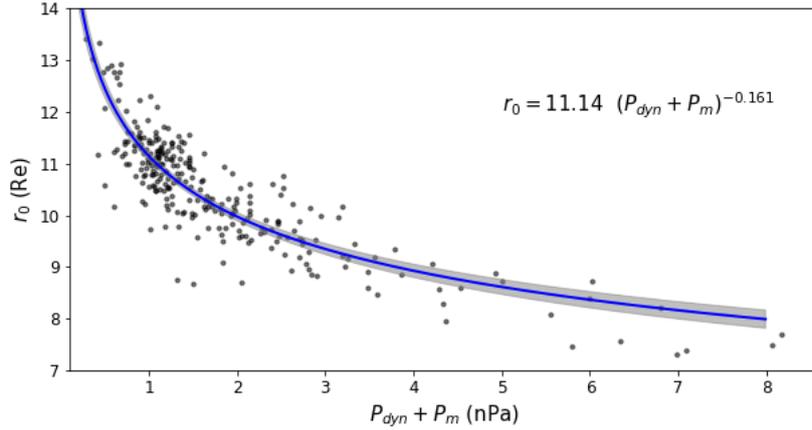


Figure 5.5: Variations of  $r_0$  with the dynamic pressure. The solid blue line represent the fit of  $r_0$  as a power-law of  $P_{dyn} + P_m$  and the grey interval is the 1-sigma confidence interval of such fit.

of our event selection that is much narrower. Our finding does not reflect the influence of the IMF cone angle on the whole magnetopause but is restricted to the lone stand-off distance for which the sunward motion due to a radial IMF was found in the order of magnitude of the errors usually made by the existing magnetopause models [Grygorov et al., 2017]. It would then be interesting in a further study to see if this apparent independence at the stand-off distance holds at higher latitudes and longitudes. For now, we will assume no particular  $B_x$  dependence.

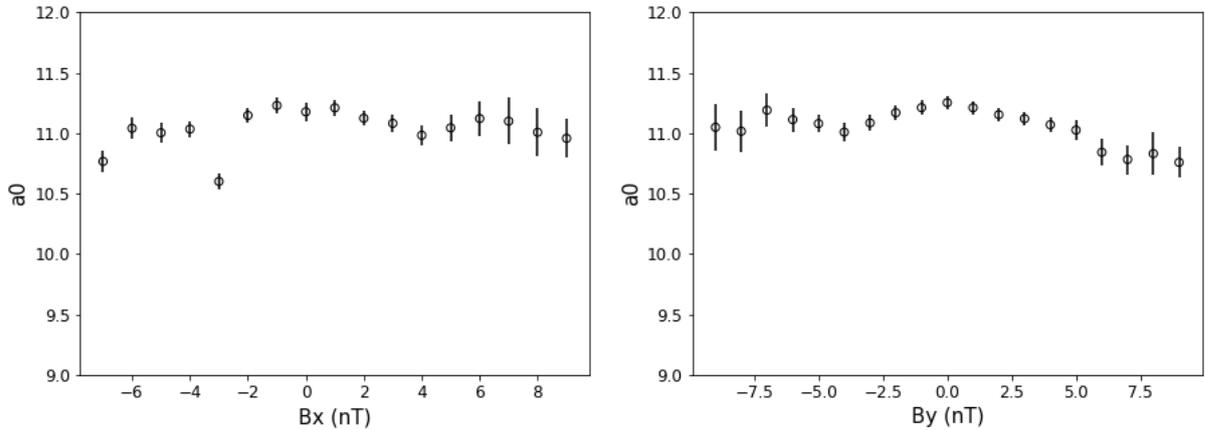


Figure 5.6: Evolution of the stand-off distance at 1 nPa,  $a_0$  as a function of the IMF  $B_x$  (left) and  $B_y$  (right). The black circles represent the value we obtain from fitting a power law to  $r_0$  for different  $B_z$  bins. The error bars represent the 1-sigma confidence interval of such fits.

The evolution of  $a_0$ , as a function of  $B_z$  is shown in Figure 5.7. Here, the noticed decrease of  $a_0$  for negative  $B_z$  is consistent with the erosion of the dayside magnetosphere when reconnection, favored by a southward IMF, occurs at low-latitude. This argument of the erosion of the dayside magnetosphere may also explain the saturation of  $a_0$  for positive  $B_z$  since magnetic reconnection is not thought to occur at low latitude for a northward IMF.

For comparative purposes, the three colored dashed lines represent the evolution of  $a_0$  with  $B_z$  previously obtained by Shue et al. [1997] (green line), Lin et al. [2010] (red line) and Liu et al. [2015] (blue line). For northward IMF, the values of  $a_0$  is consistent with the one obtained by both Shue et al. [1997] and Liu et al. [2015] while Lin et al. [2010]'s model over-estimate the stand-off distance by 2 Re further from the Earth. The three models also predict an erosion of the magnetopause for southward IMF and the slope of the decrease of  $a_0$  for negative  $B_z$  is close from the one exhibited by Shue et al. [1997].

The studies of Liu et al. [2015] and Shue et al. [1997] suggest a saturation of  $a_0$  for strong negative  $B_z$  (not shown in Figure 5.7) that confirms the observations made by Yang et al. [2002]. This saturation was also assumed by Lin et al. [2010] through the introduction of the third term of the equation 5.5. In our case, even if the decrease of our fitted values of  $a_0$  for negative  $B_z$  appears to be less important below -7 nT, the important error bars for this point and the restricted range of  $B_z$ , due to the restriction of our dataset to the stand-off distance, do not permit to draw of any conclusion on what happen for extreme values of  $B_z$ . Subsolar magnetopause crossings under extreme  $B_z$  are extremely scarce and it is then of no use to study them from a statistical point of view for now. They however constitute excellent samples for further case studies on the behavior of the magnetosphere under solar events, such as ICMEs, for which we expect such extreme conditions<sup>4</sup>. The saturation will be assumed when we will use the totality of the *out of cusp* crossings to fit an empirical magnetopause shape and location model.

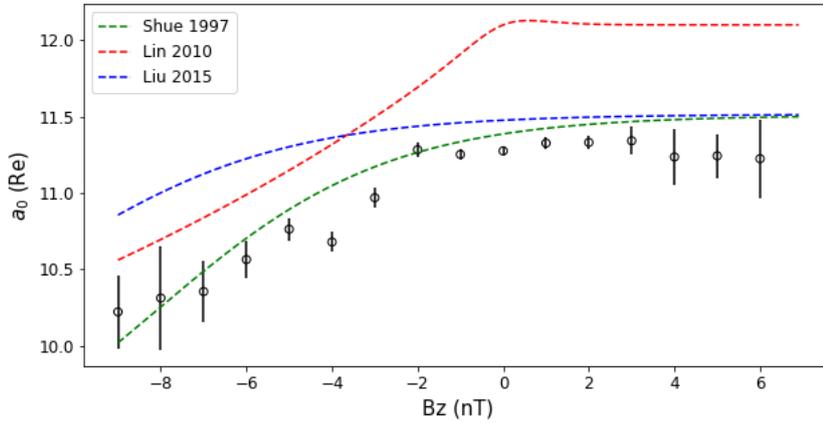


Figure 5.7: The same than Figure 5.6 but as a function of the IMF  $B_z$ . The three colored dashed line represent the evolution of  $a_0$  according three previous existing models: Shue et al. [1997] (green), Lin et al. [2010] (red) and Liu et al. [2015] (blue).

We can use the two evolutions we exhibited to establish a primary empirical expression of the magnetopause stand-off distance. Applying the Levenberg-Marquardt algorithm [Newville et al., 2014] to our 275 subsolar crossings, we then obtain:

$$r_0 = 10.75 (1 + 0.05 \tanh(0.35B_z + 1.6)) (P_{dyn} + P_m)^{-0.161} \quad (5.27)$$

### 5.3.3 Evidencing the asymmetries

The first empirical models [Fairfield, 1971; Formisano, 1979; Petrinec and Russell, 1993; Shue et al., 1997] of the magnetopause shape and location assumed axisymmetry around the GSM X axis. Nevertheless, the MHD simulations of Lu et al. [2011] evidenced an asymmetry between the magnetopause flaring in the (X – Y) plane and the flaring in the (X – Z) plane. They evidenced the IMF  $B_z$  component and the Earth dipole tilt angle as the main actors at the origin of such azimuthal asymmetry. This asymmetry was considered by the fits of Wang et al. [2013] and Lin et al. [2010] and already observed Šafránková et al. [2002] but never confirmed in the far nightside, below -20 Re.

We address this question by selecting on the one hand the 2154 out of cusp crossings for which  $|Y| < 2$  Re (the so-called X – Z plane events) and on the other hand the 5170 out of cusp crossings for which  $|Z| < 1$  Re (the so-called X – Y plane events). In both cases, we represent  $\log(r)$ , the radial distance of each crossing as a function of  $\log\left(\frac{2}{1+\cos(\theta)}\right)$ , the inverse trigonometric function used by Shue et al. [1997] and Liu et al. [2015] in equations 5.1 and 5.15, in the two panels of Figure 5.8. In

<sup>4</sup>This could especially been done through the study of the magnetopause crossings of our dataset at the date of which we also have a detected ICME.

both cases, we notice a clear linear dependency that legitimates the commonly used expression of the magnetopause shape and location as a power law of an inverse trigonometric function. It is also worth noting that the linear dependency holds for the far nightside crossings of ARTEMIS giving in the process another credit to the using of such analytical expression.

Following this linear dependency, the slope of a fitted linear expression should give a first estimate of the flaring coefficient  $\alpha$  in both the  $(X-Z)$  and the  $(X-Y)$  planes. The red curve and the associated grey shading confidence intervals are shown in the two panels of Figure 5.8. We find a value of  $\alpha$  that is lower in the  $(X-Z)$  plane than in the  $(X-Y)$  plane. Even if we lack observational evidences for the magnetopause at high latitude in the far nightside, this suggests the existence of an azimuthal asymmetry and a magnetopause that is more elongated in the  $Y$ -direction than in the  $Z$ -direction. Following 5.1, the intercept of the red curve of the two panels should correspond to an estimate of the average value of the stand-off distance  $r_0$ . This value is equal to  $10.5R_E$  in the  $(X-Z)$  plane and to  $10.38R_E$  in the  $(X-Y)$  plane, which is in the orders of magnitude of the average value of  $r_0$  we had in the previous subsection.

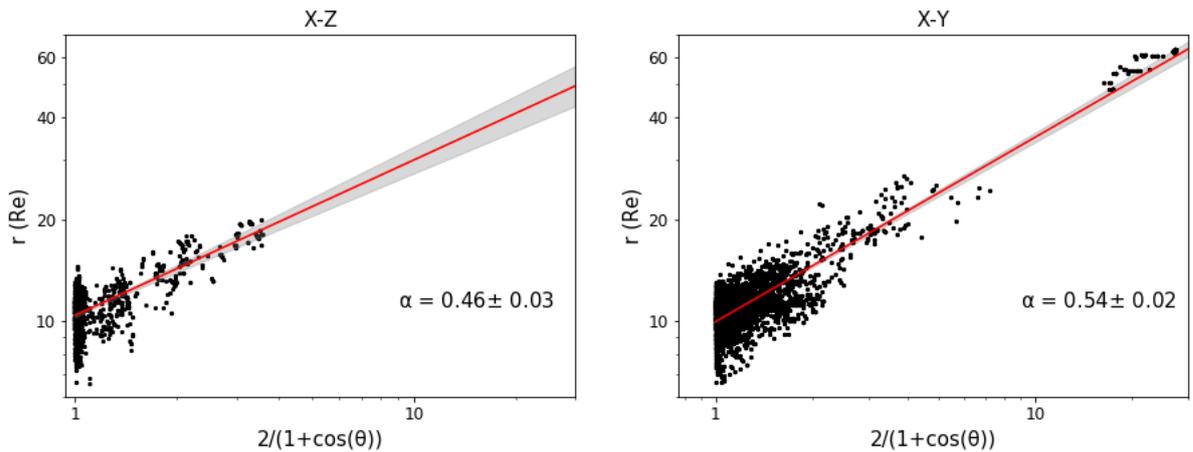


Figure 5.8: Evolution of the radial distance  $r$  of the crossings in the  $(X-Z)$  (*left*) plane and in the  $(X-Y)$  plane as a function of the inverted trigonometric function  $\frac{2}{1+\cos(\theta)}$  on a logarithmic scale. The solid red line represent the linear fit of  $\log r$  as a function of  $\frac{2}{1+\cos(\theta)}$ . The grey intervals represent the confidence intervals of such fits.

In addition to the azimuthal asymmetry, the MHD simulations of Liu et al. [2012] showed the dipole tilt angle induced a North-South asymmetry that was also observed by Boardsen et al. [2000] and considered by the fits of Lin et al. [2010] and Wang et al. [2013]. The symmetrization of the dataset we did in 5.3.1 already allows this asymmetry and the evolution of the northern (or the southern) magnetopause flaring with the dipole tilt angle will thus be properly evidenced and investigated in the next subsection.

Using Goes and LANL spacecraft, Kuznetsov and Suvorova [1998] evidenced a dawn-dusk asymmetry of the magnetopause, this asymmetry was also observed and linked to the solar wind aberration caused by the Earth orbital motion by Šafránková et al. [2002]. The latter noticed that the correction of this aberration was enough to erase the asymmetry. Having the aberration corrected, Lin et al. [2010] still noticed a dawn-dusk asymmetry in their fitted magnetopause model. Nevertheless, this was noticed in the MHD simulations of Liu et al. [2015] and references therein.

Following what we did to evidence the azimuthal asymmetry, we now separate the 5170 out of cusp crossings in the  $(X-Y)$  plane into 2832 dawnward ( $Y < 0$ ) and 2338 duskward ( $Y > 0$ ) and applied the same method than previously. The logarithm of the radial distance  $r$  of these two subsets of events as a function of the logarithm of the inverted trigonometric function  $\frac{2}{1+\cos(\theta)}$  is shown in the Figure 5.9.

In this case, we obtain almost equal values for  $\alpha$  in both sides that indicate no apparent dawn-dusk asymmetry and thus, a symmetric magnetopause regarding the  $Y = 0$  plane. In the light of those results, we decide to add another symmetry to the dataset by reverting the  $Y$  coordinate, increasing the size of the out of cusp crossings from 29077 to 58154.

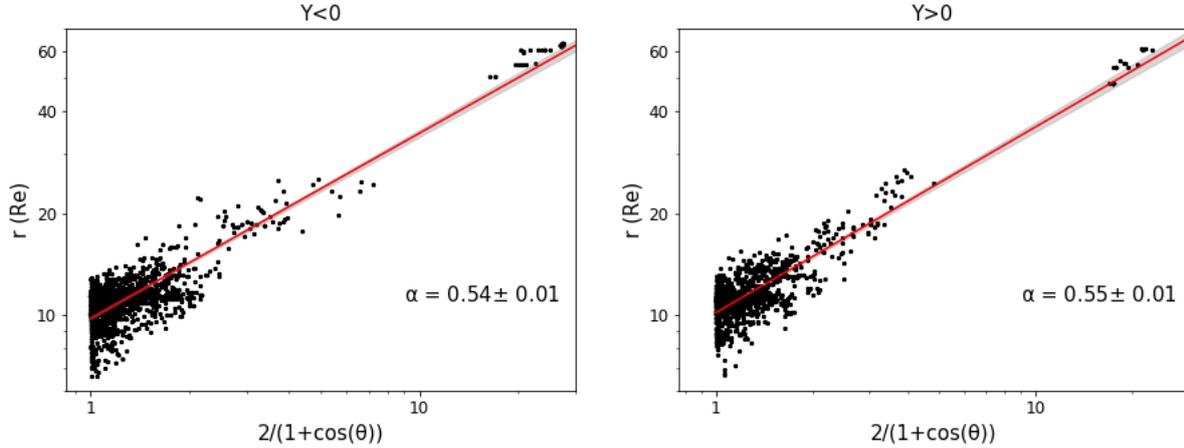


Figure 5.9: The same figure than 5.8 but with the crossings in the  $Y < 0$  (left) and in the  $Y > 0$  (right) half-spaces.

### 5.3.4 Dependencies of the flaring coefficients

The Figures 5.8 and 5.9 evidenced the magnetopause flaring as the power law of an inverse trigonometric function, consistently with (6.1). They also evidenced different flarings in the  $(X - Y)$  and in the  $(X - Z)$  directions indicating our necessity to treat the influence of the various solar wind parameters on the two flarings separately.

#### Influence of the dynamic pressure

Having symmetrized the dataset by reverting  $Y$ , we focus on the Equatorial flaring by selecting the 5170 duskward out of cusp crossings for which  $|Z| < 1$  Re and the 3882 northern out of cusp crossings for which  $|Y| < 2$  Re. We represent the averaged distribution of the total pressure in these two newly so-defined  $(X - Y)$  and  $(X - Z)$  planes in the two panels of Figure 5.10.

Although the observation is noisier in the left panel, one can see the appearance of clear parallel contours which intercept goes from 15 Re to 7 Re with an increasing pressure. This proves that the main effect of an increasing pressure stands in an earthward translation of the magnetopause along the  $X$  axis and thus, that the flaring coefficient of the magnetopause is independent from the upstream solar wind pressure. This finding is consistent with Lin et al. [2010] who found no particular pressure dependency and with the results of Liu et al. [2015] and Shue et al. [1997] that found a flaring coefficient that had very little variations with the pressure.

Another method we have to evidence this independence stands in separating the data of the two planes into sliding pressure bins between 0.5 and 6 nPa and estimating the flaring coefficient  $\alpha$  by fitting the radial position of the crossings to the equation 5.1 with  $r_0$  being defined by the equation 5.27. The result of such fits is shown in the two planes in Figure 5.11 and seeing very little variations of  $\alpha$  in the two planes confirms the independence.

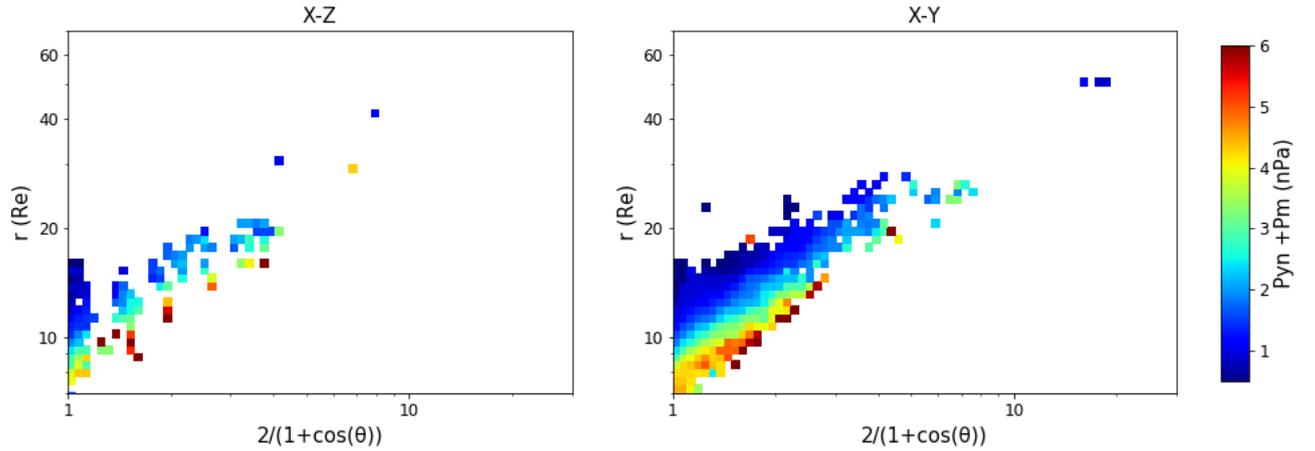


Figure 5.10: Averaged distribution of the solar wind total pressure ( $P_{dyn} + P_m$ ) associated to the crossings in the ( $X - Z$ ) (*left*) and in the ( $X - Y$ ) (*right*) planes.

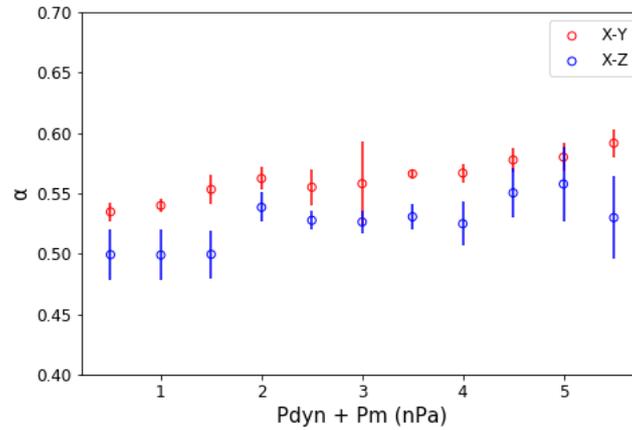


Figure 5.11: Evolution of the fitted flaring coefficient  $\alpha$  as a function of the total pressure for the out of cusp crossings in the ( $X - Y$ ) (*blue*) and in the ( $X - Z$ ) (*red*) planes. The error bars represent the 1-sigma confidence intervals of the different fits.

### Influence of the dipole tilt angle

The dipole tilt angle is expected to only influence the polar flaring of the magnetopause (Boardsen et al. [2000]; Lin et al. [2010] and references therein). Thus, we investigate the influence of the dipole tilt angle on the flaring by considering the ( $X - Z$ ) events previously defined. Working in a similar manner than for the pressure, we separate the crossings into sliding tilt angle bins and estimate  $\alpha$  for each group of events<sup>5</sup>.

The result of such fits are shown in the Figure 5.12 and show a clear linear increase of  $\alpha$  with an increasing  $\gamma$ . This indicates a northern hemisphere magnetopause that opens during summer and that tends to become more closed during winter.

This finding is consistent with the dependencies evidenced by Boardsen et al. [2000] and Lin et al. [2010] and very close from the hyperbolic tangent dependency chosen by Liu et al. [2015]. Having a symmetrized dataset, the observed flaring for the southern hemisphere will naturally also be a linear function of  $\gamma$  with the same intercept but an opposed slope.

<sup>5</sup>Although not shown, we also looked at the averaged spatial distribution of the Earth dipole tilt angle but did not notice any specific pattern with this method, proof of the absence of any translational effect  $\gamma$  on the magnetopause.

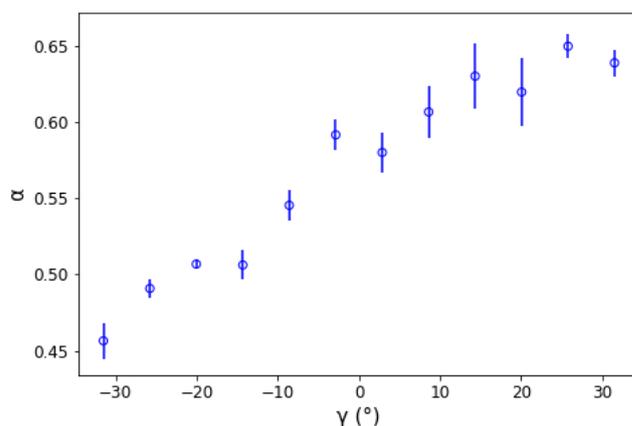


Figure 5.12: Evolution of the fitted flaring coefficient  $\alpha$  as a function of the dipole tilt angle  $\gamma$  for the out of cusp crossings in the  $(X-Z)$  plane. The error bars represent the 1-sigma confidence intervals of the different fits

### Influence of the IMF clock angle

The influence of the IMF orientation on the magnetopause shape was first noticed by Aubry et al. [1970] who noticed, using Ogo 5 measurements, an earthward motion of the boundary when the IMF was southward. The phenomena was later on frequently found in numerous statistical studies of the magnetopause location and shape [Petrinec and Russell, 1996; Sibeck et al., 1991] and considered in the most recent analytical models through the dependence on the IMF  $B_z$  component for both the subsolar stand-off distance and the level of flaring [Lin et al., 2010; Liu et al., 2015; Shue et al., 1997; Wang et al., 2013].

The aforementioned models all suggested an elliptic cross-section of the magnetopause which semi-major axis is located on the GSM Z (resp. Y) axis when the IMF is southward (resp. northward).

Such behaviour may be explained by the erosion mechanism triggered by magnetic reconnection and usually described by the so-called *onion peel model* described in Sibeck et al. [1991]. When the IMF turns southward, the X-line is believed to be located in the equatorial plane. Dayside magnetic flux is then convected by the magnetosheath flow to the magnetotail where its accumulation might result in a magnetopause surface that flares more in the azimuthal plane. The semi-major axis of the cross section is on the Z axis. On the opposite, when the IMF is northward, reconnection occurs in the lobes as the magnetic flux is here convected sunward, the azimuthal flaring decreases and the semi-major axis of the cross-section is now located on the Y axis.

Another reason that may explain this change of shape stands in the magnetic forces exerted on the boundary surface. A southward turning IMF results in a larger current flowing across the magnetopause surface and thus in an increased force exerted on the boundary in the equatorial plane.

It is also worth noting that the influence of the IMF orientation might depend on the value of the Alfvén Mach Number. Using MHD simulations, Lavraud and Borovsky [2008] suggested that, at low Mach numbers, both southward and northward IMF resulted in a semi-major axis on the Z axis while eastward (resp. orientation) rotated the semi-major axis anti-clockwise (resp. clockwise).

Although, the IMF  $B_z$  is believed to be the one with the greatest impact on the topology of the magnetosphere, its lone consideration is reductive regarding the totality of the parameters that have an influence on the shape of the magnetopause, one can especially think of the two other IMF components,  $B_x$  and  $B_y$  and their value in comparison with  $B_z$ . Consequently, instead of considering the influence of the lone value of  $B_z$ , we investigate the influence of the IMF clock angle  $\Omega$ . We leave the influence of the IMF cone angle for a further study.

Accordingly, we consider both the events in the  $(X-Y)$  (resp.  $(X-Z)$ ) plane and separate them

into  $30^\circ$  (resp.  $60^\circ$  wide clock angle bins between  $-180^\circ$  and  $180^\circ$  and estimate  $\alpha$  for each group of events. The obtained values are shown in the Figure 5.13.

For southward IMF orientations (green shaded intervals), the (X – Z) flaring, represented by the blue dots, is higher than the equatorial flaring represented with the red dots. The semi-major axis of the magnetopause cross-section is then oriented along the GSM Z axis. For northward orientations (yellow-shaded intervals), the equatorial flaring becomes higher than the (X – Z) flaring. The semi-major axis is now oriented along the GSM Y axis. This evolution is consistent with the suggestions of the other existing empirical magnetopause models [Lin et al., 2010; Liu et al., 2015; Shue et al., 1997; Wang et al., 2013] and could thus be explained either by the erosion mechanism triggered by magnetic reconnection, either by the magnetic forces exerted on the boundary surfaces. Nevertheless, these results are obtained with no distinction between the low and the high Alfvén Mach number crossings and are opposed to the suggestions previously Lavraud and Borovsky [2008]. It would thus be interesting, in a further study, to investigate if the result we evidenced still hold for low Alfvén Mach number values.

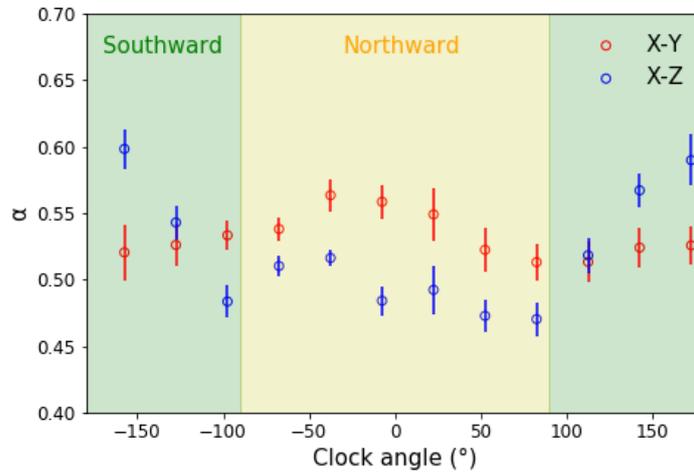


Figure 5.13: Evolution of the fitted flaring coefficient  $\alpha$  as a function of the IMF clock angle for the out of cusp crossings in the (X – Y) (blue) and in the (X – Z) (red) planes. The green intervals indicate the intervals where the IMF is southward and the yellow interval is the interval of northward IMF. The error bars represent the confidence intervals of the different fits.

The flaring coefficients evolving with changing values of the IMF clock angle also suggests a modification of the magnetopause shape for a changing IMF  $B_y$  component.

To ensure it, we consider the 888 events in the (X – Z) plane for which  $1 \text{ nT} < B_z < 3 \text{ nT}$  and the 756 events in the (X – Z) plane for which  $-3 \text{ nT} < B_z < -1 \text{ nT}$ . In the two cases, we separate the events into 1 nT wide  $B_y$  bins and estimate  $\alpha$  for each group of events. The obtained values are shown in the Figure 5.14.

For a negative  $B_z$  (green dots), increasing absolute value of  $B_y$  results in a decreasing flaring coefficient. On the opposite, for a positive  $B_z$  (yellow dots), increasing absolute value of  $B_y$  suggests in a decreasing flaring coefficient. These evolutions are consistent with the evolution of  $\alpha$  for a changing IMF orientation in the (X – Z) plane, as shown in the Figure 5.13, and indicates that  $B_y$  modifies the magnetopause shape by displacing the reconnection sites.

It was also suggested in MHD simulations that a varying  $B_y$  induces a twisting of the magnetopause that followed the induced rotating IMF [Liu et al., 2015]. Nevertheless, such rotation results in an expected dawn dusk asymmetry that we neglected after the findings of Figure 5.9 and the symmetrization of the dataset that was there performed. The question of an eventual twisting of the magnetopause with changing IMF clock angle thus remains open and should be investigated in further studies.

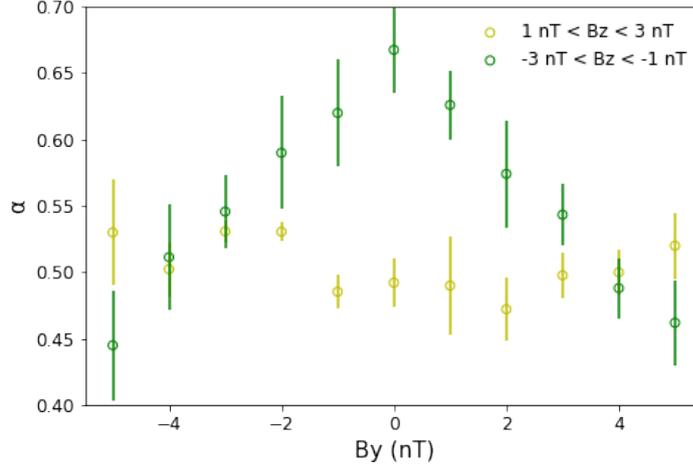


Figure 5.14: Evolution of the fitted flaring coefficient  $\alpha$  as a function of the IMF  $B_y$  component for the out of cusp crossings in the  $(X-Z)$  plane for which  $1 \text{ nT} < B_z < 3 \text{ nT}$  (yellow) and for which  $-3 \text{ nT} < B_z < -1 \text{ nT}$  (green). The error bars represent the confidence intervals of the different fits.

## 5.4 Fitting a new magnetopause model

### 5.4.1 Fit

We began this chapter by presenting several existing empirical models made from both observations and MHD simulations. In the light of the statistical studies that were just performed, we evidenced an expression of the magnetopause stand-off distance close from the one established by Shue et al. [1997] and Liu et al. [2015] and an azimuthal asymmetry that was mainly controlled by the dipole tilt angle and the IMF clock angle. If the tilt angle dependency is not new and was already taken into account by Lin et al. [2010], Liu et al. [2015] and Wang et al. [2013], the evolution of the flaring as a function of the clock angle is a novelty as previous studies considered the influence of  $B_z$  alone when taking into account the influence of magnetic reconnection.

Considering the results of the previous section, we can define an alternative expression of the non indented magnetopause surface as:

$$r = r_0 \left( \frac{2}{1 + \cos(\theta)} \right)^\alpha \quad (5.28)$$

$$r_0 = a_0 (P_{dyn} + P_m)^{a_1} (1 + a_2 \tanh(a_3 B_z + a_4)) \quad (5.29)$$

$$\alpha = \alpha_0 + \alpha_1 \cos(\phi) + \alpha_2 \sin(\phi)^2 + \alpha_3 \cos(\phi)^2 \quad (5.30)$$

$$\alpha_0 = a_5 \quad (5.31)$$

$$\alpha_1 = a_6 \gamma \quad (5.32)$$

$$\alpha_2 = a_7 \cos(\Omega) \quad (5.33)$$

$$\alpha_3 = a_8 \cos(\Omega) \quad (5.34)$$

Where  $\alpha_0$ ,  $\alpha_1$ ,  $\alpha_2$ , and  $\alpha_3$  control the different flarings we studied previously and  $\Omega$  is the IMF clock angle. The expressions of  $r_0$  is similar to the one exposed in 5.27 and the expression of  $\alpha$  has been modified from the previous models to consider the dependencies we evidenced in the previous subsections. It is worth noting that such an expression is unchanged by the substitution of  $\phi$  by  $-\phi$  and by the substitution of  $(\phi, \gamma)$  by  $(\pi - \phi, -\gamma)$  respecting consequently the two symmetries we supposed in the dataset. To ensure our model does not overfit the data, we randomly split the symmetrized dataset into a train set of 43664 events that will serve for the modeling and into a

test set of 14490 events that will be used to evaluate the accuracy of the fitted model on the next subsection.

We predetermined the values of  $a_0$ ,  $a_1$ ,  $a_2$ ,  $a_3$  and  $a_4$  in 5.27. The initial values of  $a_5$  and  $a_7$  are determined by fitting 5.28, 5.29, 5.30, 5.31 and 5.34 to the out of cusp crossings for which  $|Z| < 1Re$  and the initial values of  $a_5$ ,  $a_6$  and  $a_8$  by fitting 5.28, 5.29, 5.30, 5.31, 5.32 and 5.33 to the out of cusp crossings for which  $|Y| < 2Re$ . We show these initial fitting values in the Table 5.2 and adjust them by fitting the 7 equations, from 5.28 to 5.34, altogether to the 43664 out of cusp crossings all at once. We used the average initial fitted values of  $a_5$  as an initial guess for this coefficient.

$a_0$	$a_1$	$a_2$	$a_3$	$a_4$	$a_5$	$a_6$	$a_7$	$a_8$
10.75	-0.161	0.050	0.35	1.60	0.55, 0.51	0.026	0.015	-0.050

Table 5.2: Initial values of the coefficients of the equations (5.27) to (5.33) obtained from the initial fits around the subsolar point of  $a_0$ ,  $a_1$ ,  $a_2$ ,  $a_3$  and  $a_4$ , in the  $(X - Y)$  plane for  $a_5$  (*first value*) and  $a_7$  and in the  $(X - Z)$  plane for  $a_5$  (*second value*),  $a_6$  and  $a_8$

The final values obtained for each coefficients are presented in the Table 5.3 and result in an analytical empirical model of the non-indented magnetopause shape and location that depend on the solar wind total pressure  $P_{dyn} + P_m$ , the IMF  $B_z$  and clock angle  $\Omega$  and the dipole tilt angle  $\gamma$ .<sup>6</sup>

$a_0$	$a_1$	$a_2$	$a_3$	$a_4$	$a_5$	$a_6$	$a_7$	$a_8$
10.61	-0.150	0.027	0.207	1.62	0.558	0.135	0.015	-0.0839

Table 5.3: Final values of the coefficients of the equations 5.28 to 5.34 obtained after a total fit on the training set.

#### 5.4.2 Evaluation and comparison with other models

We evaluate the accuracy of the fitted model by computing the RMSE on the 14490 events of the test set and compare it to the RMSE of the models of Shue et al. [1997], Lin et al. [2010] and Liu et al. [2015] computed on the same test set. To keep a consistent comparison, we removed the indentation part of the models of Lin et al. [2010] and Liu et al. [2015] during the computation of the RMSE. The score we obtain for the different models for different spatial regions are shown in the Table 5.4 and visually represented in the Figure 5.15.

The obtained RMSE in the low-latitude, dayside regions, subsolar and flank, is almost similar for the four models, this indicates that our model continues providing an accurate description of the magnetopause shape and location in those regions and this is not surprising given the proximity of the expression and coefficients of the stand-off distance of the 4 models (equations 5.2, 5.5, 5.18 and 5.29). At high latitudes, we notice a more important error for the model of Shue et al. [1997] that is not surprising as no high latitude data was considered during the development of this model and this is particularly reflected by the similarity we find between our RMSE and the RMSE of Lin et al. [2010] and Liu et al. [2015].

Looking at the "close" nightside region, we notice this time a reduced error in comparison to the three others than can be explained by our consideration of the IMF clock angle in the expression of the flaring coefficient  $\alpha$  rather than the lone  $B_z$ .

Finally, the previous models were established without consideration of magnetopause crossings further than  $-30 Re$ , especially the one detected by ARTEMIS and it is thus not surprising to notice a lower RMSE for our model in the "far" nightside. Naturally, the error here is possibly substantially higher than in any other region we considered. This could be explained by the flapping of the magnetotail in the far nightside that could result in a much more variable boundary

<sup>6</sup>A numerical implementation of this model can be found at: [https://github.com/gautiernguyen/magnetopause\\_models](https://github.com/gautiernguyen/magnetopause_models) and will be the one used in the following paragraphs.

[Sergeev et al., 1998]. The underlying detection of magnetopause crossings will consequently be much more ambiguous and constitute a non-negligible source of error.

	Our model	Liu et al. [2015]	Lin et al. [2010]	Shue et al. [1997]
$X < -30$ (361)	$6.06 \pm 0.32$	$15.01 \pm 0.53$	$14.06 \pm 0.34$	$9.19 \pm 0.34$
$X > -30$ and $X < 0$ (2525)	$1.67 \pm 0.033$	$2.39 \pm 0.040$	$2.48 \pm 0.047$	$2.26 \pm 0.034$
$X > 0$ and $ Z  > 7.5$ (1188)	$1.84 \pm 0.042$	$1.78 \pm 0.042$	$1.72 \pm 0.044$	$2.24 \pm 0.043$
$X > 0$ and $ Z  < 7.5$ and $ Y  > 7.5$ (3992)	$1.02 \pm 0.016$	$0.99 \pm 0.017$	$1.22 \pm 0.016$	$1.19 \pm 0.016$
$X > 0$ and $ Z  < 7.5$ and $ Y  > 7.5$ (6424)	$0.93 \pm 0.011$	$0.86 \pm 0.010$	$0.96 \pm 0.010$	$1.00 \pm 0.012$
All regions (14490)	$1.53 \pm 0.013$	$2.73 \pm 0.021$	$2.65 \pm 0.023$	$2.06 \pm 0.015$

Table 5.4: RMSE of the different models in different region for the 14490 crossings of the test set, the uncertainty represents the standard error of mean of the error of each model. The number between brackets in the first column indicate the number of events per region.

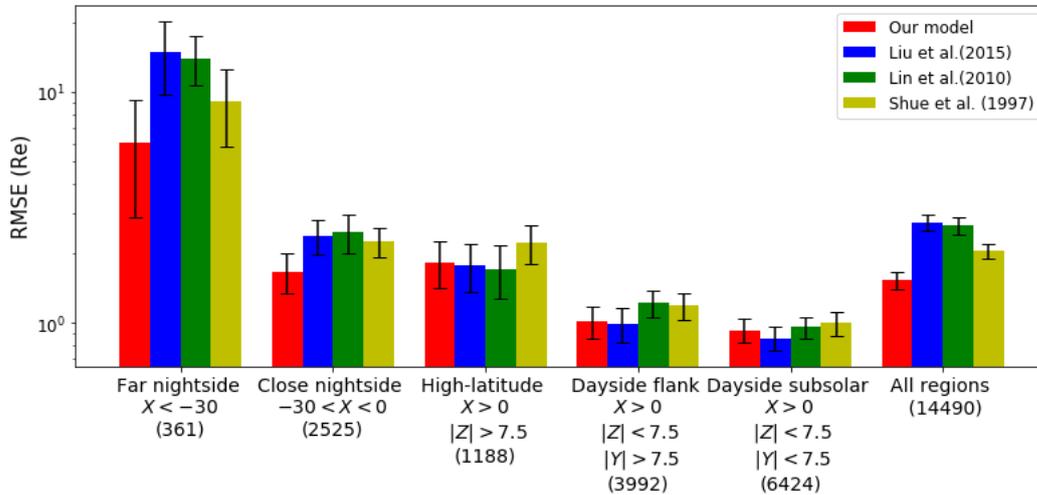


Figure 5.15: Visual representation of the RMSE of the different models exposed in the Table 5.4. The error bar represent 10 times the Standard Error of the Mean (SEM) of the error made by the models on the test set.

Combining the regions altogether (last group of bars of Figure 5.15), the precedent findings result in a global RMSE that is lower for our model in comparison to the others and thus ensures the reliability of our model and its legitimacy to be exploited in further magnetopause studies.

### 5.4.3 Characteristics of the model

We show the influence of the total pressure,  $P = P_{dyn} + P_m$  on our magnetopause model with the three panels of Figure 5.16. Following what we evidenced in the Figure 5.11, the total pressure pushes the magnetopause earthward along the X axis without influencing the flaring. This behavior confirms the findings of the previous models that showed very little pressure dependency. Additionally, we find a power law index  $a_1$  equal to -0.15 that is very close to the theoretical  $-1/6$  for a dipole in vacuum and in the same orders of magnitude than the values found by Shue et al. [1997], Lin et al. [2010] and Liu et al. [2015].

The influence of the IMF  $B_z$  is shown in the three panels of Figure 5.17. A decreasing  $B_z$  from a northward to a southward orientation will also translate the magnetopause Earthward along the X-axis, nevertheless, this effect, explained by the magnetopause erosion, is less important than the translation of the magnetopause induced by the variations of the total pressure. This is consistent with what we show in Figure 5.7 and the coefficients we obtain are very similar to what was obtained by Shue et al. [1997].

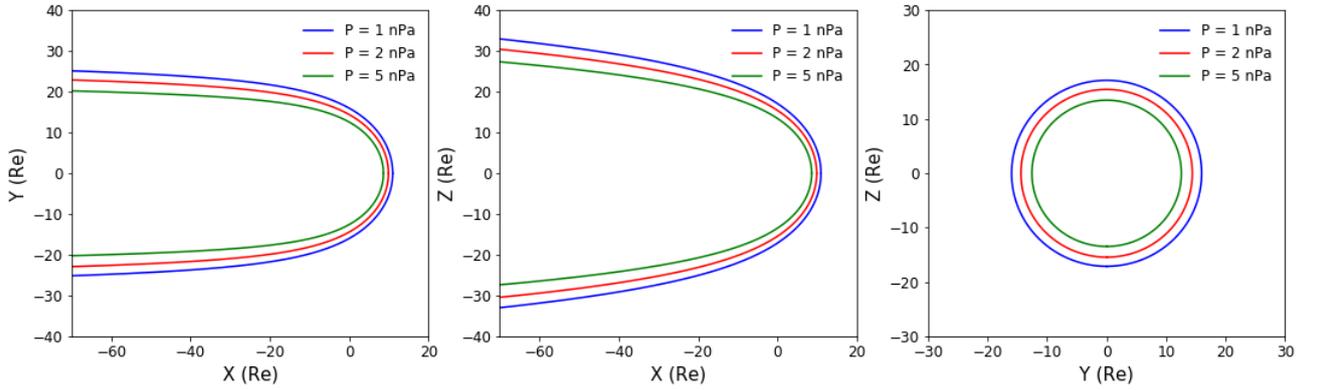


Figure 5.16: Projection in the  $(X-Y)$  (left),  $(X-Z)$  (middle), and  $(Y-Z)$  (right) planes of our model for varying total pressure  $P = P_{dyn} + P_m$ . The IMF is purely southward and  $B_z = -2$  nT.

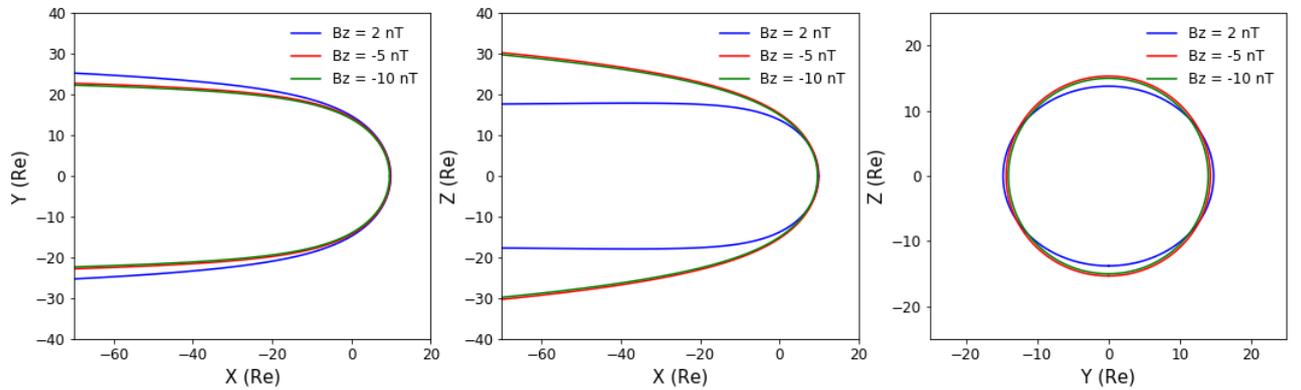


Figure 5.17: Projection in the  $(X-Y)$  (left),  $(X-Z)$  (middle), and  $(Y-Z)$  (right) planes of our model for varying  $B_z$ . The total pressure is equal to 2 nPa.

Additionally, a southward  $B_z$  induces an equatorial erosion and the opposite is observed for a northward IMF at high latitudes. This effect is represented with the three panels of Figure 5.18 that represent the influence of the clock angle. This is consistent with the change of flaring coefficients that was evidenced in the Figure 5.13 and results in an elliptic magnetopause cross section with a major axis on the  $Y$  axis for a positive  $B_z$  and on the  $Z$  axis for negative  $B_z$ .

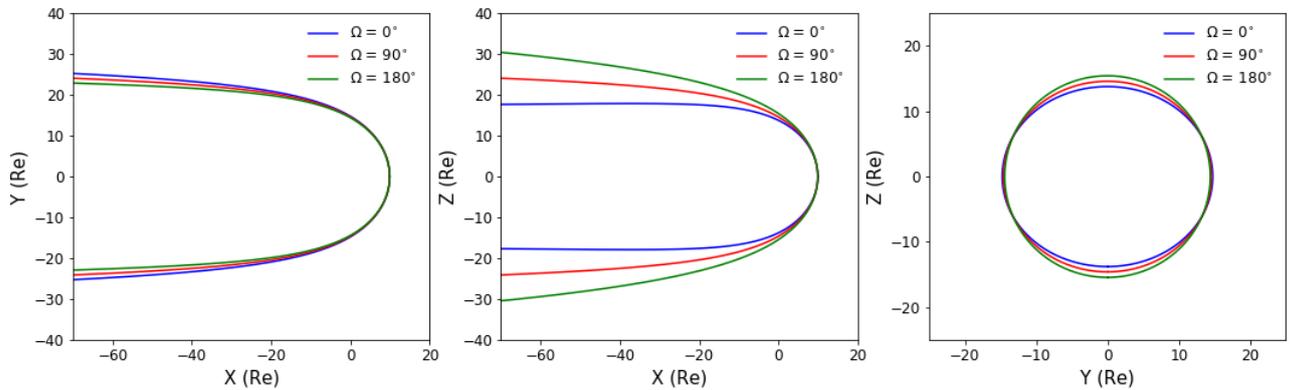


Figure 5.18: Projection in the  $(X-Y)$  (left),  $(X-Z)$  (middle), and  $(Y-Z)$  (right) planes of our model for varying clock angle  $\Omega$ . The total pressure is equal to 2 nPa and  $|B| = 2$  nT.

Finally, we show the influence of the dipole tilt angle  $\gamma$  with the three panels of Figure 5.19 that clearly indicate a magnetopause that rotates around the Y axis with a rotating dipole tilt angle resulting in a north hemisphere summer (respectively south hemisphere winter) shift of the magnetopause cross section.

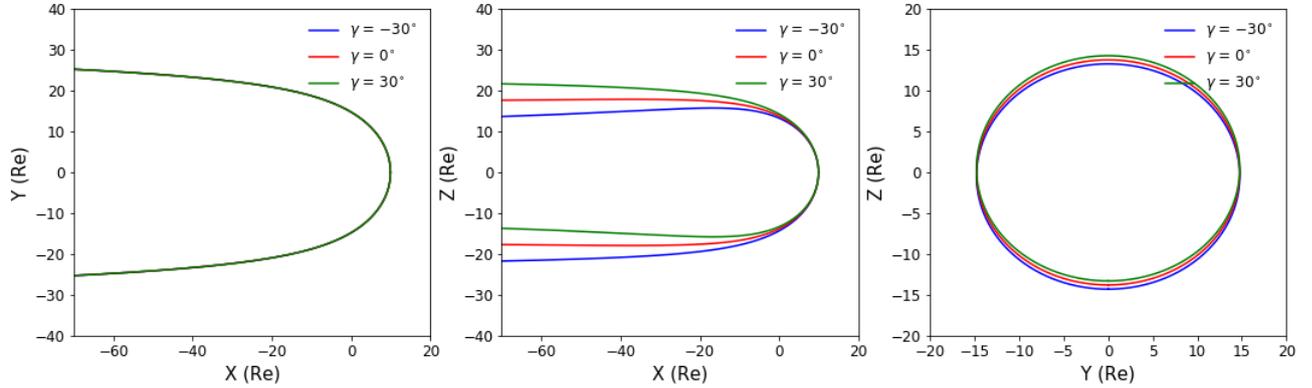


Figure 5.19: Projection in the  $(X-Y)$  (left),  $(X-Z)$  (middle), and  $(Y-Z)$  (right) planes of our model for varying dipole tilt angle  $\gamma$ . The total pressure is equal to 2 nPa, the IMF is purely northward and  $B_z = 2$  nT .

#### 5.4.4 From a static to a dynamic model

All of the magnetopause models we have been presenting so far, including our model, return a static representation of the magnetopause for a permanent upstream solar wind regime. Nevertheless, this regime is far from being the nominal case and the magnetopause is thus in a permanent evolution, following the variations of the upstream solar wind. This implies the necessity we have to adapt our static model into a dynamic magnetopause model that has the ability to provide radial distance of the magnetopause for every spherical coordinate,  $\theta$  and  $\phi$ , at any time.

To do so, we adapt the two-step propagation algorithm of Šafránková et al. [2002] to estimate the temporal shift to OMNI data needed for each zenith angle  $\theta$  :

1. At a given time  $t$  and a given zenith angle  $\theta$ , we estimate a first position of the magnetopause by computing our model for the averaged solar wind conditions in the interval between  $t$  and  $t - 30$  min. We chose 30 minutes as this is the typical shifting time we obtain for  $X \sim -70$  Re.
2. This first position serves to estimate a first value of the X coordinate of the magnetopause at this value of  $\theta$ .
3. We apply the two step propagation algorithm to estimate a first shifting time from this first X position and compute the associated radial position of the magnetopause.
4. We use this second radial position to re-estimate the X coordinate of the magnetopause at this value of  $\theta$ .
5. We apply the two-step propagation algorithm a second time to determine the final shifting time that will be used for this value  $\theta$  and we compute the associated final radial position of the magnetopause.

At a given time  $t$ , we apply this process for every zenithal position  $\theta$  and end up with a dynamical magnetopause model similar to the one shown in the two right panels of Figure 5.20 where we represented the projection of the magnetopause in the  $(X-Y)$  and in the  $(X-Z)$  planes at the time indicated by the black dashed line on the two left panels. Naturally, the shifting time increases with the zenith angle and the left boundary of the grey interval represented in the two left panels then corresponds to the magnetopause computed at the left border of the two right panels.

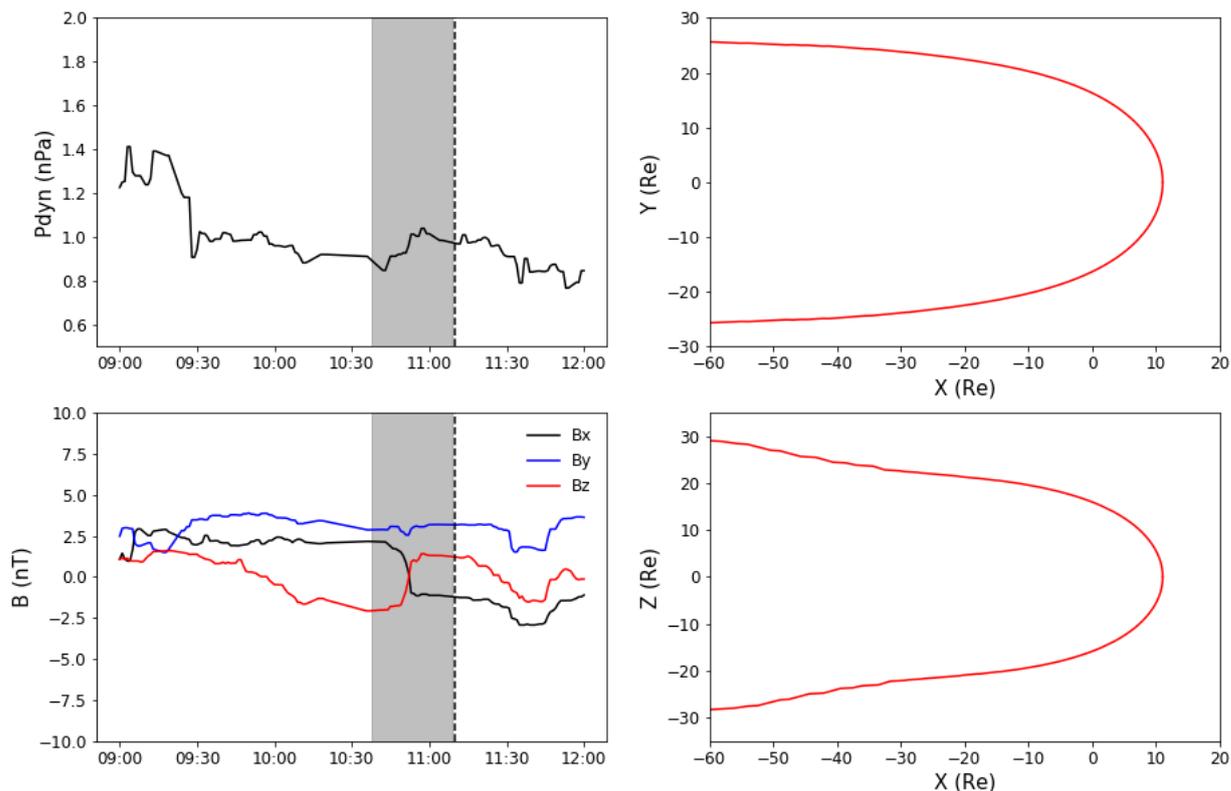


Figure 5.20: OMNI solar wind dynamic pressure (*top*) and magnetic field (*bottom*) measurement (*left column*) on the 3<sup>rd</sup> of March 2011 and projection of our dynamical magnetopause model (*right column*) in the  $(X-Y)$  (*top*) and in the  $(X-Z)$  (*bottom*) planes at the time corresponding to the black dashed line on the left column. The grey interval in the left panels represent the data interval propagated throughout the whole magnetopause.

The obtained magnetopause then considers an IMF shift from a negative to a positive  $B_z$  and the propagation of this transition is reproduced through the propagation of the erosion we notice in the  $(X-Z)$  plane.

Adapting the two step propagation algorithm, we then elaborated a process useful for providing a dynamical view of the magnetopause at any time. Additionally, this process is independent from the used static magnetopause model and thus easily adaptable to the models of Shue et al. [1997], Lin et al. [2010] and Liu et al. [2015].<sup>7</sup>

## 5.5 Nature of the near-cusp magnetopause

### 5.5.1 Different boundaries of the polar cusps

The polar cusps are defined in the two hemispheres as the regions where the geomagnetic field is vanishing. These regions, located at an average latitude of  $75^\circ$ , are the privileged entry place of solar particles in the magnetosphere as this was confirmed by the low-latitude observations of IMP5 data by Frank [1971] and the observations of ISIS data by Heikkila and Winningham [1971].

<sup>7</sup>The numerical implementation of this models can also be found at: [https://github.com/gautiernguyen/magnetopause\\_models](https://github.com/gautiernguyen/magnetopause_models).

Just like the different regions and boundaries of the near-Earth environment, the location and the geometry of the cusps are strongly affected by the solar wind conditions and the seasonal variations of the Earth dipole tilt angle. A southward (northward) IMF shifts the cusp equatorward (poleward) while a dawnward (duskward) orientation of the IMF shifts it dawnward (duskward) in the northern hemisphere, an increasing dynamic pressure widens it and a sunward tilt of the Earth dipole brings the cusp poleward (Russell [2000] and references therein).

As they are the privileged entry for the solar particles in the terrestrial magnetosphere, the polar cusps are particularly affected by magnetic reconnection and the orientation of the IMF as shown by the two panels of Figure 5.21:

- When the IMF is northward (right panel) reconnection occurs in the lobes and the merged field lines are convected sunward (green line). This convection, opposed to the tailward flow of the magnetosheath generates a region of dense, turbulent, and overall stagnant plasma, hotter than the magnetosheath and characterised by a low magnetic field defined as the *cusp exterior*. A passage of the Cluster 1 spacecraft in this region during northward IMF is shown between the two black lines in Figure 5.22.
- When the IMF is southward (left panel), reconnection occurs on the dayside, low latitude magnetopause. The merged field lines are convected tailward (green line). In this case, the direction of the two flows are similar and the associated cusp exterior, still dense, hot and associated to a low magnetic field, becomes a convective region where the plasma flow is oriented tailward. A passage of the Cluster 1 spacecraft in this region during southward IMF is shown between the two black lines in Figure 5.23.

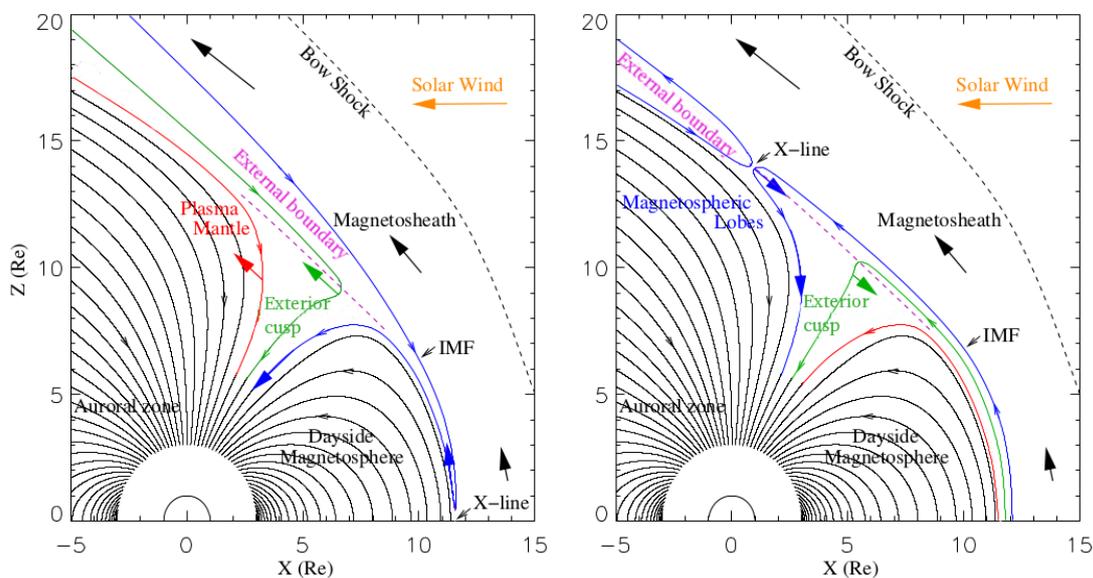


Figure 5.21: Schematic representation of the magnetic field topology and plasma flow in the near-cusp region for southward (*left*) and northward (*right*) IMF. The blue green lines show the time evolution of the reconnected field lines (see text). The red-line is the first non-convected field line and the dashed purple line represents the assumed location of the cusp external boundary. (adapted from Lavraud and Cargill [2005]).

In both cases, the convection of reconnected field lines are at the origin of the cusp exterior, a region of dense, hot plasma with a low magnetic field that can either be stagnant or convective depending on the orientation of the IMF. The differences of the physical parameters of this region with the magnetosheath imply the necessary existence of a discontinuity between the two regions known as the *cusp external boundary*, represented by the purple dashed line in Figure 5.21 and by the black dotted lines in Figures 5.22 and 5.23.

We define the boundary that delimits the exterior of the cusps from the magnetosphere as the *cusp inner boundary*. This boundary, represented by the black dotted lines in the 5.22 and 5.23, delimits the closed field lines of the magnetosphere<sup>8</sup>.

Both the inner and the external boundary can be seen as the continuous prolongation between the day side and the nightside magnetopause. Without reconnection, we expect the external boundary to vanish and the inner boundary should be TD that prevent any solar intrusion in the magnetosphere. It is then the logical continuous extension of the magnetopause in the near-cusp magnetopause. In the sense of reconnection, the inner boundary is actually the separatrix of the open and closed field lines on the dayside (resp. nightside) when the IMF is southward (resp. northward). The external boundary then appears as a more appropriate definition of the magnetopause in this region.

---

<sup>8</sup>Another observational example of the cusp external boundary is shown in the appendix B.

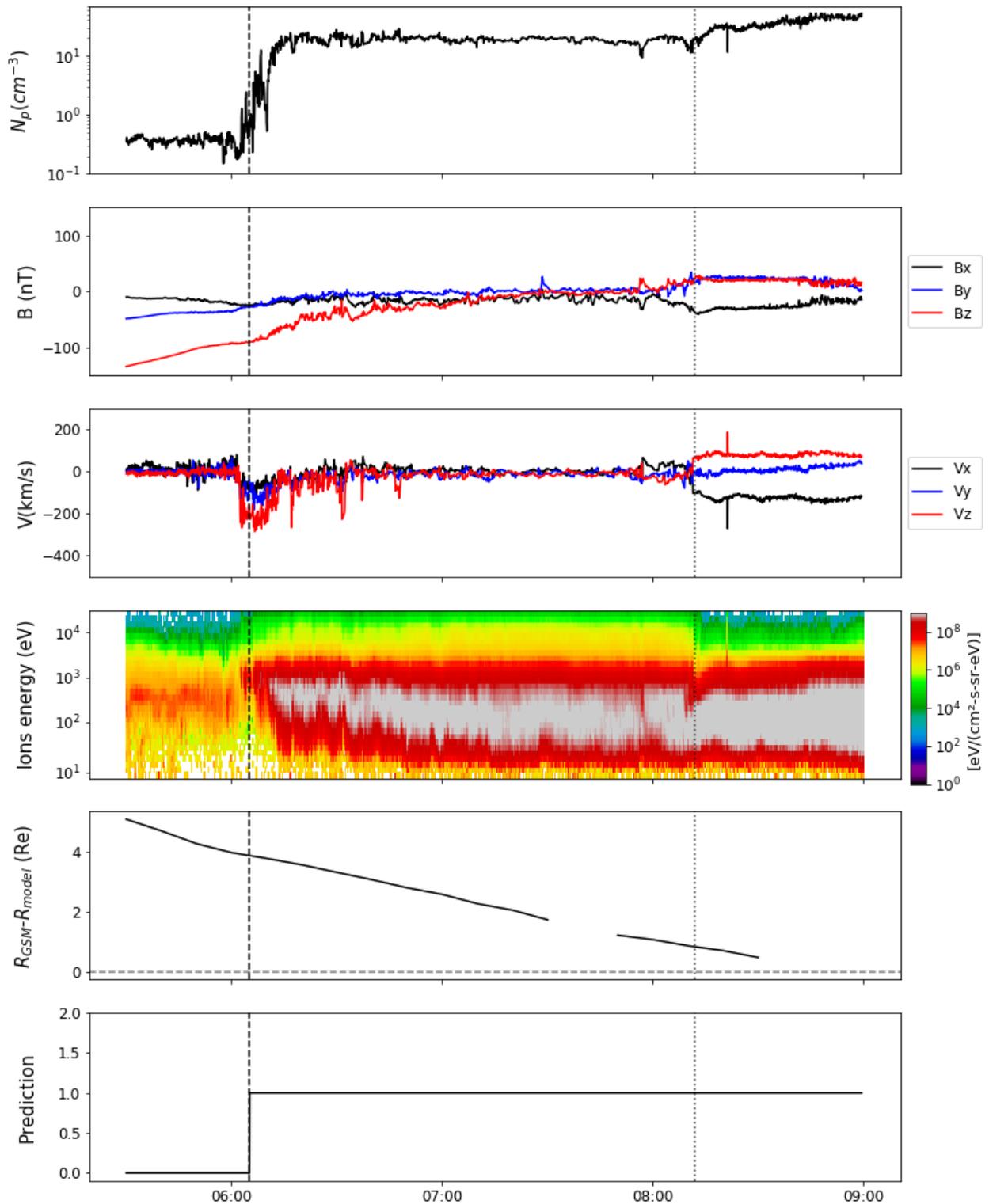


Figure 5.22: In-situ measurement provided by Cluster 1 spacecraft on the 16<sup>st</sup> of March 2002. From top to bottom are represented the ion density, the plasma magnetic field and velocity components, the omnidirectional energy fluxes of ions, The difference between the radial position of the spacecraft and the radial position predicted by our dynamic model and the prediction of the region classifier presented in Chapter 4. The grey dashed line indicate the magnetopause obtained from the gradient boosting prediction. The grey dotted line indicate the actual position of the magnetopause.

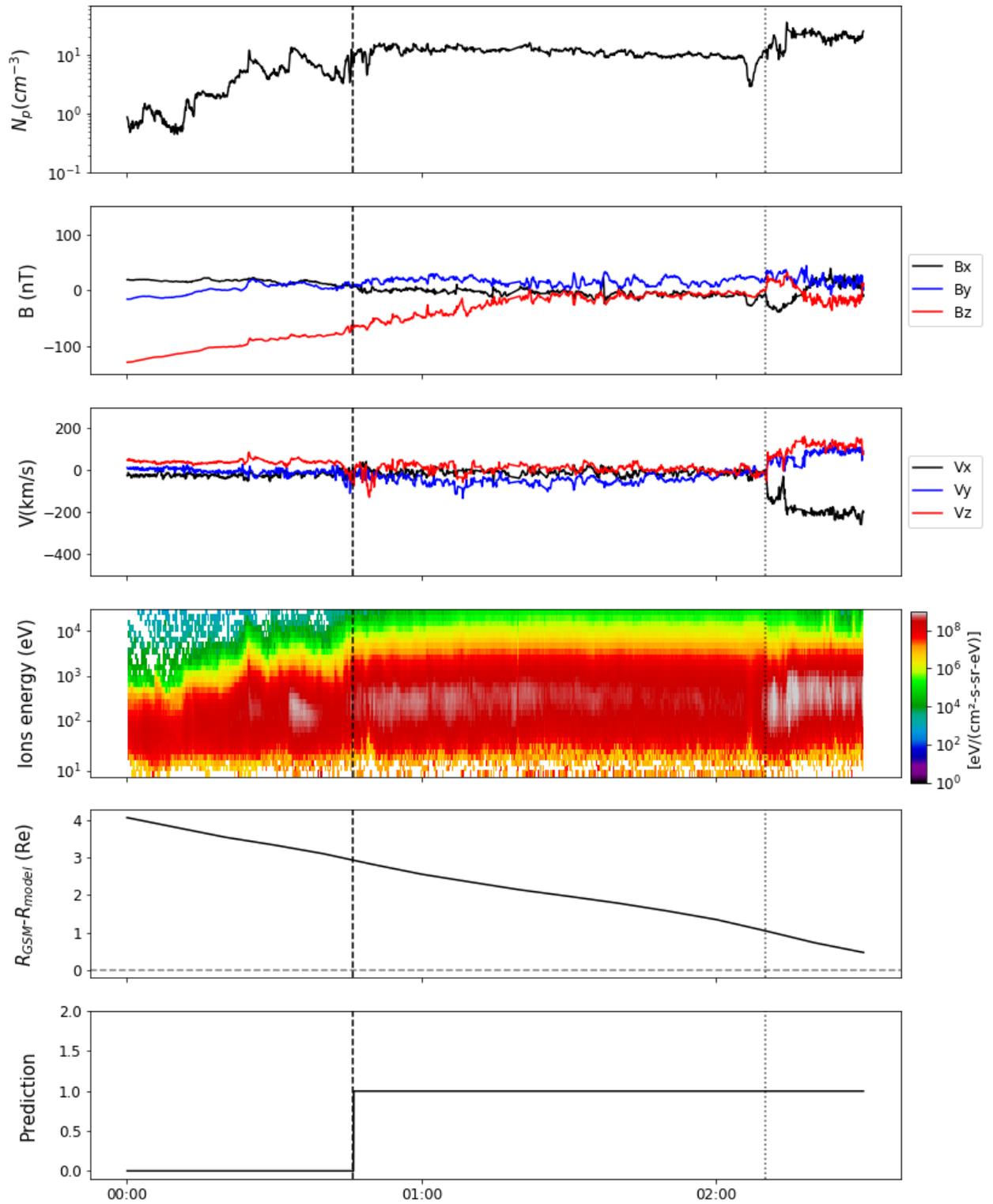


Figure 5.23: In-situ measurement provided by Cluster 1 spacecraft on the 21<sup>th</sup> of March 2002. The legend is the same than in Figure 5.22.

### 5.5.2 Shape of the near-cusp magnetopause

Based on pressure balance between the solar wind and the magnetosphere, the theoretical work of Spreiter and Briggs [1962] predicted a discontinuity between the dayside and the nightside magnetopause that could be extended by considering an indentation of the boundary in the near-cusp region. This indentation was later suggested by Haerendel et al. [1978] following their observations of HEOS data. These observations also proposed the existence of a dense, hot and stagnant region located on the assumed magnetosheath side of the indentation. From then on, the apparent near-cusp indentation was observed in a multitude of study based on the data of various spacecraft (Boardsen et al. [2000], Šafránková et al. [2002]) and considered by latest magnetopause models [Lin et al., 2010; Liu et al., 2015; Wang et al., 2013]. Although largely admitted, the existence of the indentation was questioned by the Hawkeye cusp observations of Zhou and Russell [1997]. Using the same set of data, Eastman et al. [2000] suggested that the absence of indentation they observed was linked to the definition they gave to the magnetopause in the near-cusp region. The Cluster observations of Lavraud et al. [2004a,b] indicated that the cusp external boundary, although a more appropriate continuous extension of the magnetopause in the near-cusp region, did not present any particular indentation while the inner boundary, an appropriate continuous extension of the magnetopause in the near-cusp region in the absence of reconnection, did so.

Nevertheless, their study was based on a very small number of samples and the actual existence of the near-cusp indentation still lacks of a large-scale, statistically relevant, confirmation. In this section, we take a step further in this direction by considering the 501 northern hemisphere in-cusp crossings<sup>9</sup> detected with our region classifier and the 205 in-cusp crossings from older missions that comply the same condition. In the following, the former events will be designated as the *automatically detected crossings* and the latter events will be designated as the *older crossings*. It is worth keeping in mind here that, given the orbit of the missions on which we applied our region classification routine, all of the automatically detected crossings are issued from our Cluster crossings catalogs.

In the previous chapter, we defined as magnetosheath any data point that was not defined either as magnetosphere or either at solar wind. Consequently, any region of mixed plasma found downstream of the bow shock, is classified as magnetosheath. In the near-cusp region, the cusp exterior, generated by the convection of the reconnected field lines, is then classified as magnetosheath and the detected boundary actually corresponds to the cusp inner boundary as shown for example by the last panel of the Figures 5.22 and 5.23. We thus expect the statistical insight on the position of the automatically detected crossings to predict an indented boundary, consistently with what has been said in the previous subsection.

To investigate the actual position of the near-cusp magnetopause, we visually collect 147 among the 501 automatically detected crossings for which Cluster 1 spacecraft goes through both the inner and the external cusp boundaries in a similar way than what is shown in the Figures 5.22 and 5.23.

Figure 5.24 represents the distribution of the solar wind parameters and of the Earth dipole tilt angle associated to the selected events. Having similar distributions of  $P_{dyn}$ ,  $B_y$  and  $B_z$  than what is shown in the Figure 5.2 and a balanced distribution of winter, summer and equinox events ensures our selection is representative enough and thus, not biased by any solar wind or seasonal parameter.

<sup>9</sup>The term *in-cusp* being defined with the equation 5.26

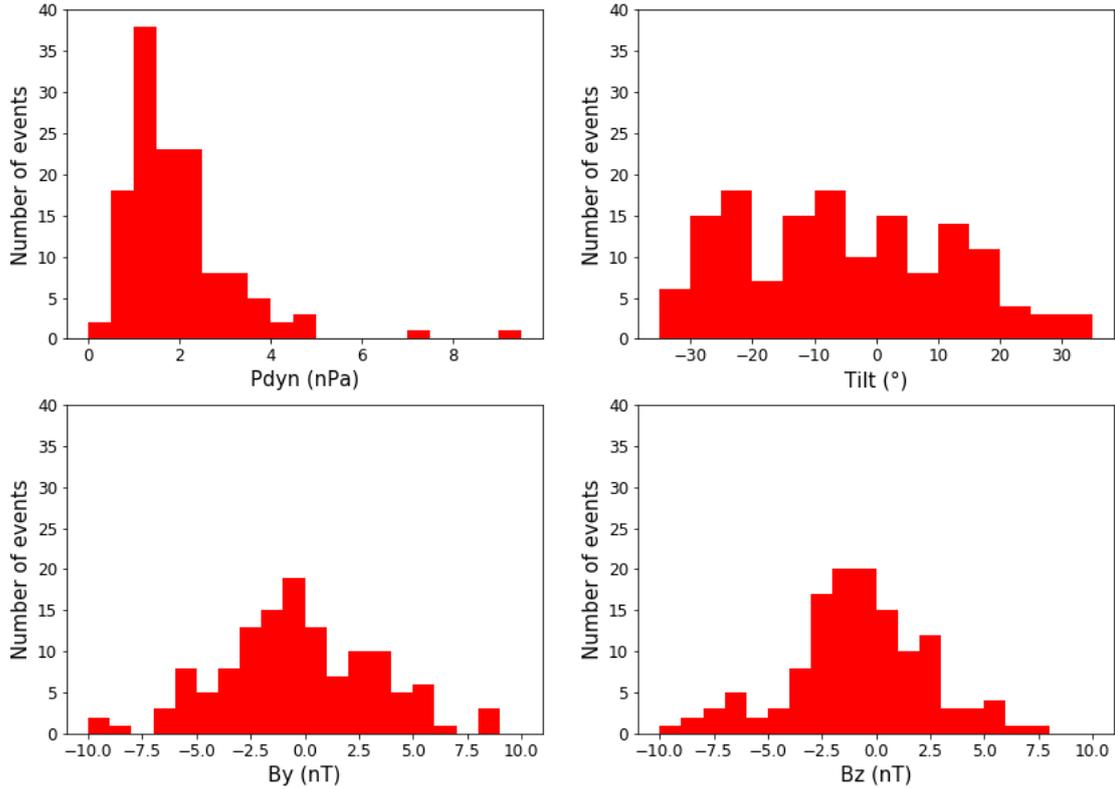


Figure 5.24: Histogram of the solar wind parameters and Earth dipole tilt angle of the 147 in-cusp magnetopause crossings manually identified:  $P_{dyn}$  (top left), the dipole tilt angle (top right),  $B_y$  (bottom left) and  $B_z$  (bottom right).

We characterise the events we selected by comparing the variations of the density, the velocity and the temperature at the crossing of the inner boundary to the same variations at the crossing of the external boundary. The results of such comparisons are shown on the three panels of Figure 5.25 :

- Looking at the left panel, the variations of density are much more important when crossing the inner boundary. This is consistent with the passage of a region almost empty of plasma (the lobe) to a dense region compared to the passage from a dense region to another. It is also worth noting that the density ratio between the magnetosheath and the cusp exterior is always above 1, indicating that the latter is on average more tenuous than the former.
- The middle panel indicates the jump in velocity is larger when crossing the external boundary. Additionally, the important standard deviation noticed in the distribution of this variation combined with the limitation of the jump the crossing of the inner boundary to  $\sim 150$  km/s is consistent with the flow of the cusp exterior we described in the previous subsection.
- The right panel indicates a temperature that increases when passing from the magnetosheath to the cusp exterior and no particular variations of temperature at the crossing of the inner boundary. This indicates the passage to a hotter region when crossing the cusp external boundary from the magnetosheath, consistently with the characteristics of the cusp exterior.

In the three cases, the variations induced by the crossing of the two boundaries are substantially different and the region they delimit is consistent with the characteristics of the cusp exterior presented previously.

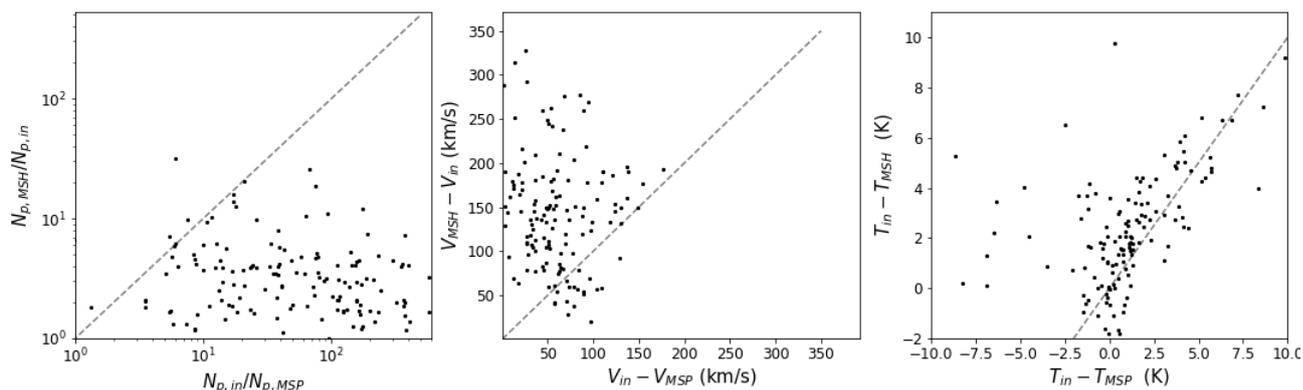


Figure 5.25: Variations of density (*left*), velocity (*middle*) and temperature (*right*) associated to the crossing of the cusp external boundary as a function to the variations of the same parameters associated to the crossing of the cusp inner boundary. The grey dashed line is the identity function.

Having characterised the crossing of the external boundary in comparison to the crossing of the inner boundary, we now focus on the position of the selected cusp external crossing. These events will be referred to as the *manually selected crossings*.

In order to draw a significant conclusions independent from the upstream solar wind conditions, we normalize the GSM position of the events of each list, manually selected, automatically detected and older, following the scheme presented in the Figure 5.26 and adapted from Lavraud et al. [2004a]:

- The GSM position of each event is projected in the ( $X-Z$ ) plane, in Figure 5.26, the projected event corresponds to the point A1.
- We define a reference magnetospheric field using the model of Tsyganenko and Sitnov [2005] computed for reference conditions defined:  $P_{dyn} = 2$  nPa,  $B_z = -2$  nT,  $B_y = B_x = 0$  nT,  $DST = -10$  nT and a null dipole tilt angle. At these conditions, we look for the last field line bent toward the dayside and define it as the reference last bent line represented by the solid red line.
- We compute the model of Tsyganenko and Sitnov [2005] for the solar wind and dipole conditions associated to our event and define the current last bent line as the last field line bent toward the dayside for these conditions, for instance, the green line.
- A1 is transformed into A2 through the rotation at the angle defined by the difference between the current last bent line and the reference last bent line.
- We use our magnetopause model to compute the position of the magnetopause in the ( $X-Z$ ) plane for the reference (red dashed line) and the current conditions (green dashed line). The radial distance of A2 is then scaled with the radial difference between the two computed magnetopause positions along the grey dotted line. This radial adjustment results in the final crossing with a normalized position represented by the point A3.

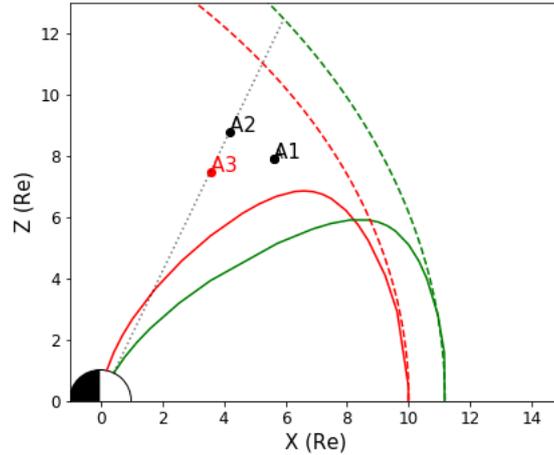


Figure 5.26: Illustration of the different steps of the coordinate transformation we apply to the in-cusp crossing. A1 is the projection of the GSM coordinate of an event in the  $(X-Z)$  plane. A2 is the image of the rotation of A1 with an angle defined by the difference between the current last bended line (green solid line) to the reference last bended line (red solid line). A3 is the image of the translation of A2 following the difference between the current (green dashed line) and the reference (red dashed line) magnetopause. in the direction defined by the zenith angle of A2 (grey dotted line).

The normalized positions of the automatically detected crossings are represented with the blue dots in the left panel of Figure 5.27. In comparison with our magnetopause model represented by the black solid line and despite of an important dispersion, we notice a clear depletion of these events location at the near-cusp latitudes. This is consistent with the nature of these crossings and the underlying expected indentation. It is worth noting that the position of the automatically detected crossings appear to be consistent with the near-cusp magnetopause predicted by Lin et al. [2010] and represented by the green line. This makes us infer that they fitted their model with a near-cusp magnetopause defined as the inner boundary. This assumption is confirmed by the normalized position of the older crossings represented by the blue dots of the right panel of Figure 5.27. Most of these crossings are the Hawkeye observations of Boardsen et al. [2000] and Eastman et al. [2000]<sup>10</sup> that are used in the fit of Lin et al. [2010] in this region and we notice the same depletion than in the left panel.

The average similar distribution of the normalized position noticed for the older and the automatically detected crossings is not surprising these events actually corresponds to the cusp inner boundary in both cases. The automatically detected crossings are thus consistent with magnetopause crossings manually detected by experts and this is not surprising as we had the same interpretation of the data. This proves, once again, that the quality of the prediction made by supervised learning algorithms is extremely linked to the interpretation of the data labeled by an external observer.

The normalized position of the manually selected crossings is represented by the red dots in the two panels of Figure 5.27. In this case, the detected crossings have a higher radial distance than the automatically detected and the older crossings, this is consistent with the definition of the cusp exterior external boundary for which we expect the location at higher radial distances than the inner boundary. Naturally, the agreement with the prediction of Lin et al. [2010] is much less obvious and this indicates that the indentation of the near-cusp magnetopause is not completely clear yet.

<sup>10</sup>Who defined the near-cusp magnetopause as the inner boundary.

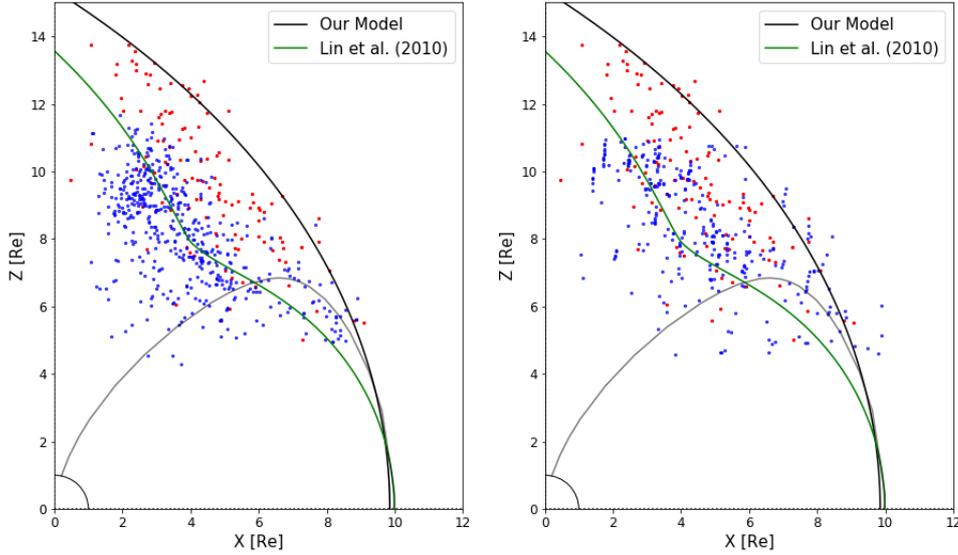


Figure 5.27: Comparison of the position in the normalized ( $X-Z$ ) plane of the in cusp events of our manually detected in cusp magnetopause crossings (*red dots*), the in cusp events detected in the Chapter 4 (*blue dots*, *left panel*) and the in cusp crossings from older mission (*blue dots*, *right panel*). The solid grey line represent the reference separatrix obtained from [Tsyganenko and Sitnov \[2005\]](#) magnetospheric magnetic field model. The black solid line is our magnetopause model computed for the reference conditions used in the normalization process (*see text*) and the solid green line is [Lin et al. \[2010\]](#) model computed for the same conditions.

Using Hawkeye and Cluster observations respectively, [Zhou and Russell \[1997\]](#) and [Lavraud et al. \[2004b\]](#) predicted a non-indented magnetopause. If this was actually the case we would intuitively expect the average normalized position of the manually selected crossing to be consistent with the projection in the ( $X-Z$ ) plane of our non-indented model. Nevertheless, almost all of these events are below our predicted magnetopause and keep suggesting a depletion in the near-cusp region.

We confirm this depletion by superimposing the normalised position of these crossings to the normalized position of the low and high latitude out of the cusp crossings for which  $|Y| < 2$  Re in the Figure 5.28. An indentation seems to appear in a way consistent with the near-cusp magnetopause predicted by [Liu et al. \[2015\]](#) and this can be explained by the fact that the authors identified the near-cusp magnetopause by looking at the variations of the thermal pressure. This gives an argument in favor of a near-cusp indentation that persists with the consideration of magnetopause magnetic reconnection. Nevertheless, this does not completely confirm its existence yet and further investigations would be needed.

From now on, an interesting option we could have to investigate this existence would be to perform a statistical study of the position and angular dispersion of crossing of the cusp external boundary as functions of the upstream solar wind and seasonal conditions. This would require the identification of additional events that could be detected with our region classifier applied on Cluster data provided we adapt the label to take the presence of the cusp exterior into consideration. A final argument in favor of the near-cusp indentation of the magnetopause would then be to observe a better fit to the data of a model that would account for this indentation in comparison with a model that would not. This would be one of the main focus of further studies on the shape of the near-cusp magnetopause.

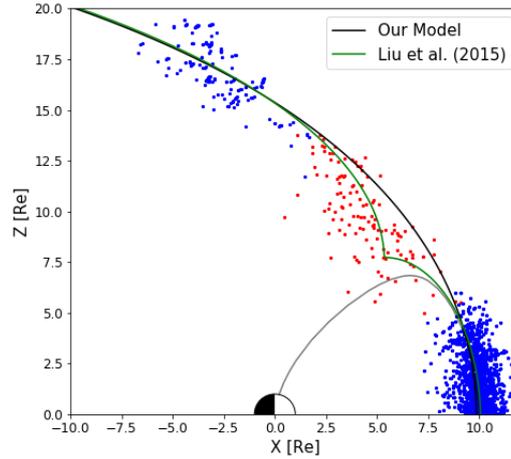


Figure 5.28: Position in the normalized  $(X-Z)$  plane of the manually detected in cusp magnetopause crossings (*red dots*) and the out of the cusp crossings for which  $|Y| < 2Re$ . The solid grey line represent the reference separatrix obtained from [Tsyanenko and Sitnov \[2005\]](#) magnetospheric magnetic field model. The black solid line is our magnetopause model computed for the reference conditions used in the normalization process (*see text*) and the solid green line is [Liu et al. \[2015\]](#) model computed for the same conditions.

## 5.6 Conclusion

Combining the magnetopause crossings we have detected in the Chapter 4 to online accessible magnetopause crossings, we provided a statistical analysis of the magnetopause shape and location through the study of the stand-off distance, the asymmetries and the level of flaring.

The findings of the statistical analysis can be summarized as follows:

1. The power-law that describes the evolution of the stand-off distance as a function of the solar wind dynamic pressure was found very close to the theoretical  $-1/6$ .
2. The influence of the IMF  $B_z$  on the stand-off distance was found similar to the one fitted by [Shue et al. \[1997\]](#).
3. We found no particular influence of the IMF  $B_x$  component on the stand-off distance. This finding appears to be in agreement with [Liu et al. \[2015\]](#) but is however restricted to the nose for which the sunward motion evidenced by [Grygorov et al. \[2017\]](#) appeared to be in the order of magnitude of the prediction errors usually made by the magnetopause models in this region. Thus, we gave no clue on the eventual influence of the IMF  $B_x$  component on the whole magnetopause and such focus would be the main topic of the future investigations.
4. We confirm the azimuthal asymmetry, and the influence of the dipole tilt angle on this asymmetry, as well as the lack of a dawn-dusk asymmetry.
5. The IMF  $B_y$  is found to affect the magnetopause shape and location through the influence of the clock angle. In comparison with all of the previous existing magnetopause models, this finding is more consistent with the role played of the IMF  $B_y$  component on the displacement of the reconnection site and the underlying erosion direction.

Following these results, we developed an empirical analytical asymmetric magnetopause shape and location model parameterized by the upstream solar wind dynamic and magnetic pressure, by the IMF  $B_z$  and  $B_y$  components and by the Earth dipole tilt angle. Comparing our model with the models of [Shue et al. \[1997\]](#), [Lin et al. \[2010\]](#) and [Liu et al. \[2015\]](#), we found the 4 models to predict the magnetopause location with equal accuracy on the dayside equatorial part of the magnetopause and the error made by our model is similar to the one made by [Lin et al. \[2010\]](#) and [Liu et al. \[2015\]](#) on the high-latitude dayside part of the magnetopause. On the nightside, the consideration of the clock angle instead of  $B_z$  and the consideration of crossings above  $-30 Re$  resulted in

a reduced error in comparison to the other existing model. Nevertheless, the lack of data in this region combined to the remnant ambiguity concerning the identification of the magnetopause at lunar distances indicate further studies are needed in this specific region of the near-Earth environment.

Assuming a dawn-dusk symmetry of the magnetopause, our model indicate a clock angle angle that squeezes the magnetopause in the Y direction and stretches it in the Z direction when the IMF turns from a northward to a southward orientation. This finding gives clues on the influence of the Y component of the IMF on the magnetopause but is not completely in agreement with Liu et al. [2015] that found  $B_y$  to twist the magnetopause instead. This shows the question of the dawn-dusk asymmetry and the nature of the influence of the clock angle on the magnetopause shape is still open and would need additional observational studies.

All of the existing magnetopause models, including ours, provide a static view of the magnetopause for a permanent upstream solar wind regime that is far from being the ground truth. For this reason, we adapted our static model to also provide a dynamic model of the magnetopause surface able to give an estimation of the magnetopause shape and location at any time.

Finally, we compared 147 manually selected cusp exterior boundary crossings to both the automatically detected in-cusp crossings and the older in-cusp crossings to give a global view of the shape of the two boundaries of the polar cusps.

The inner boundary represents a clear indentation. This boundary has been considered as the continuous extent of the magnetopause between the dayside and the nightside for long. This is the boundary considered by Lin et al. [2010] in the development of their model and the boundary we detect with our region classifier in the logic aftermath of the way we labeled the data.

Whenever reconnection occurs, the cusp external boundary appears as a more appropriate continuous extension between the dayside and the nightside magnetopauses. In this case, we noticed a depletion of the magnetopause location in the near-cusp region that appears in adequation with the numerical results of Liu et al. [2015]. Nevertheless, this finding is still preliminary and the confirmation or the invalidation of the actual existence of the indentation would requires further additional studies on the position of the cusp external boundary as a function of the upstream solar wind and seasonal condition.

Last but not least, the labeling of the magnetosheath we made in the Chapter 4 resulted in the detection of the crossing of the inner boundary in the near-cusp region. After the reproduction of the ambiguity of human made observation of ICMEs by the CNNs presented in the Chapter 3, this once again shows that the quality of detection method based on machine learning algorithms is highly linked to the interpretability of the external observer at the origin of the labeled dataset and the ambiguity, inherent to in-situ observations is also found in the prediction of such algorithms.

## 5.7 Bibliography

- Aubry, M. P., Russell, C. T., and Kivelson, M. G.: Inward motion of the magnetopause before a sub-storm, *Journal of Geophysical Research* (1896-1977), 75, 7018–7031, <https://doi.org/10.1029/JA075i034p07018>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/JA075i034p07018>, 1970. 116
- Boardsen, S. A., Eastman, T. E., Sotirelis, T., and Green, J. L.: An empirical model of the high-latitude magnetopause, *Journal of Geophysical Research: Space Physics*, 105, 23 193–23 219, <https://doi.org/10.1029/1998JA000143>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/1998JA000143>, 2000. 102, 103, 104, 107, 113, 115, 128, 131
- Cahill, L. J. and Amazeen, P. G.: The boundary of the geomagnetic field, *Journal of Geophysical Research* (1896-1977), 68, 1835–1843, <https://doi.org/10.1029/JZ068i007p01835>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/JZ068i007p01835>, 1963. 102
- Cortes, C. and Vapnik, V.: Support-vector networks, *Machine learning*, 20, 273–297, 1995. 102

- Dmitriev, A. V. and Suvorova, A. V.: Artificial neural network model of the dayside magnetopause: physical consequences, *Physics and Chemistry of the Earth C*, 25, 169–172, 1999. [102](#), [103](#), [105](#)
- Dusik, Š., Granko, G., Š, J., N, Z., and Jelínek, K.: IMF cone angle control of the magnetopause location : Statistical study, 37, 2–5, <https://doi.org/10.1029/2010GL044965>, 2010. [102](#), [103](#), [105](#), [110](#)
- Eastman, T. E., Sotirelis, T., and Green, J. L.: An Empirical Model of the High-latitude Magnetopause, 105, <https://doi.org/doi:10.1029/1998JA000143>, 2000. [102](#), [103](#), [104](#), [128](#), [131](#)
- Fairfield, D. H.: Average and unusual locations of the Earth's magnetopause and bow shock, *Journal of Geophysical Research*, 76, 6700, <https://doi.org/10.1029/JA076i028p06700>, 1971. [102](#), [112](#)
- Formisano, V.: The three-dimensional shape of the bow shock., *Planetary and Space Science*, 2C, 681–692, <https://doi.org/10.1007/BF02558125>, 1979. [102](#), [112](#)
- Frank, L. A.: Plasma in the Earth's polar magnetosphere, *Journal of Geophysical Research (1896-1977)*, 76, 5202–5219, <https://doi.org/10.1029/JA076i022p05202>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/JA076i022p05202>, 1971. [123](#)
- Grygorov, K., Šafránková, J., Němeček, Z., Pi, G., Přech, L., and Urbář, J.: Shape of the equatorial magnetopause affected by the radial interplanetary magnetic field, *Planetary and Space Science*, 148, 28 – 34, <https://doi.org/https://doi.org/10.1016/j.pss.2017.09.011>, URL <http://www.sciencedirect.com/science/article/pii/S0032063317302131>, 2017. [102](#), [105](#), [110](#), [111](#), [133](#)
- Haerendel, G., Paschmann, G., Sckopke, N., Rosenbauer, H., and Hedgecock, P. C.: The frontside boundary layer of the magnetosphere and the problem of reconnection, *Journal of Geophysical Research: Space Physics*, 83, 3195–3216, <https://doi.org/10.1029/JA083iA07p03195>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/JA083iA07p03195>, 1978. [128](#)
- Hasegawa, H.: Structure and Dynamics of the Magnetopause, 1, 71–119, <https://doi.org/10.5047/meep.2012.00102.0071>, 2012. [102](#)
- Heikkila, W. J. and Winningham, J. D.: Penetration of magnetosheath plasma to low altitudes through the dayside magnetospheric cusps, *Journal of Geophysical Research (1896-1977)*, 76, 883–891, <https://doi.org/10.1029/JA076i004p00883>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/JA076i004p00883>, 1971. [123](#)
- Kuznetsov, S. N. and Suvorova, A. V.: Solar wind magnetic field and pressure during magnetopause crossings at geosynchronous orbit, *Advances in Space Research*, 22, 63–66, [https://doi.org/10.1016/S0273-1177\(97\)01101-0](https://doi.org/10.1016/S0273-1177(97)01101-0), 1998. [102](#), [104](#), [113](#)
- Lavraud, B. and Borovsky, J. E.: Altered solar wind-magnetosphere interaction at low Mach numbers: Coronal mass ejections, *Journal of Geophysical Research: Space Physics*, 113, <https://doi.org/https://doi.org/10.1029/2008JA013192>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2008JA013192>, 2008. [116](#), [117](#)
- Lavraud, B. and Cargill, P. J.: Cluster reveals the magnetospheric cusps, *Astronomy & Geophysics*, 46, 1.32–1.35, <https://doi.org/10.1046/j.1468-4004.2003.46132.x>, URL <https://doi.org/10.1046/j.1468-4004.2003.46132.x>, 2005. [124](#)
- Lavraud, B., Fedorov, A., Budnik, E., Grigoriev, A., Cargill, P., Dunlop, M., Rème, H., Dandouras, I., and Balogh, A.: Cluster survey of the high-altitude cusp properties: a three-year statistical study, *Annales Geophysicae*, 22, 3009–3019, <https://doi.org/10.5194/angeo-22-3009-2004>, 2004a. [103](#), [128](#), [130](#)

- Lavraud, B., Phan, T., Dunlop, M., Taylor, M., Cargill, P., Bosqued, J., Dandouras, I., Rème, H., Sauvaud, J., Escoubet, C., Balogh, A., and Fazakerley, A.: The exterior cusp and its boundary with the magnetosheath: Cluster multi-event analysis, *Annales Geophysicae*, 22, 3039–3054, <https://doi.org/10.5194/angeo-22-3039-2004>, 2004b. 103, 128, 132
- Lin, R. L., Zhang, X. X., Liu, S. Q., Wang, Y. L., and Gong, J. C.: A three-dimensional asymmetric magnetopause model, *Journal of Geophysical Research (Space Physics)*, 115, A04207, <https://doi.org/10.1029/2009JA014235>, 2010. 102, 103, 104, 105, 106, 107, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 123, 128, 131, 132, 133, 134
- Liu, Z., Lu, J. Y., Wang, C., Kabin, K., Zhao, J. S., Wang, M., Han, J. P., Wang, J. Y., and Zhao, M. X.: Journal of Geophysical Research : Space Physics A three-dimensional high Mach number asymmetric magnetopause model from global MHD simulation, *Journal of Geophysical Research*, pp. 5645–5666, <https://doi.org/10.1002/2014JA020961>. Received, 2015. 102, 103, 105, 106, 107, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 123, 128, 132, 133, 134
- Liu, Z. Q., Lu, J. Y., Kabin, K., Yang, Y. F., Zhao, M. X., and Cao, X.: Dipole tilt control of the magnetopause for southward IMF from global magnetohydrodynamic simulations, *Journal of Geophysical Research (Space Physics)*, 117, A07207, <https://doi.org/10.1029/2011JA017441>, 2012. 113
- Lu, J. Y., Liu, Z. Q., Kabin, K., Zhao, M. X., Liu, D. D., Zhou, Q., and Xiao, Y.: Three dimensional shape of the magnetopause: Global MHD results, *Journal of Geophysical Research (Space Physics)*, 116, A09237, <https://doi.org/10.1029/2010JA016418>, 2011. 112
- Newville, M., Stensitzki, T., Allen, D., and Ingargiola, A.: LMFIT: Non-Linear Least-Square Minimization and Curve-Fitting for Python<sup>¶</sup>, <https://doi.org/10.5281/zenodo.11813>, 2014. 112
- Němeček, Z., Šafránková, J., and Šimůnek, J.: An Examination of the Magnetopause Position and Shape Based Upon New Observations, chap. 8, pp. 135–151, American Geophysical Union (AGU), <https://doi.org/10.1002/9781119509592.ch8>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/9781119509592.ch8>, 2020. 107
- Petrinec, S. M. and Russell, C. T.: An empirical model of the size and shape of the near-Earth magnetotail, *Geophysical Research Letters*, 20, 2695–2698, <https://doi.org/10.1029/93GL02847>, 1993. 102, 110, 112
- Petrinec, S. M. and Russell, C. T.: Near-Earth magnetotail shape and size as determined from the magnetopause flaring angle, *Journal of Geophysical Research: Space Physics*, 101, 137–152, <https://doi.org/10.1029/95JA02834>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/95JA02834>, 1996. 102, 116
- Petrinec, S. P., Song, P., and Russell, C. T.: Solar cycle variations in the size and shape of the magnetopause, *Journal of Geophysical Research*, 96, 7893–7896, <https://doi.org/10.1029/90JA02566>, 1991. 102
- Roelof, E. C. and Sibeck, D. G.: Magnetopause shape as a bivariate function of interplanetary magnetic field  $B_z$  and solar wind dynamic pressure, *Journal of Geophysical Research*, 98, 21 421–21 450, <https://doi.org/10.1029/93JA02362>, 1993. 102, 110
- Russell, C.: The polar cusp, *Advances in Space Research*, 25, 1413 – 1424, [https://doi.org/https://doi.org/10.1016/S0273-1177\(99\)00653-5](https://doi.org/https://doi.org/10.1016/S0273-1177(99)00653-5), URL <http://www.sciencedirect.com/science/article/pii/S0273117799006535>, proceedings of the DO.1 Symposium of COSPAR Scientific Commission D, 2000. 124

- Sergeev, V., Angelopoulos, V., Carlson, C., and Sutcliffe, P.: Current sheet measurements within a flapping plasma sheet, *Journal of Geophysical Research: Space Physics*, 103, 9177–9187, <https://doi.org/10.1029/97JA02093>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/97JA02093>, 1998. 120
- Shue, J. H., Chao, J. K., Fu, H. C., Russell, C. T., Song, P., Khurana, K. K., and Singer, H. J.: A new functional form to study the solar wind control of the magnetopause size and shape, *Journal of Geophysical Research*, 102, 9497–9512, <https://doi.org/10.1029/97JA00196>, 1997. 102, 103, 104, 105, 106, 110, 111, 112, 114, 116, 117, 118, 119, 120, 123, 133
- Sibeck, D. G., Lopez, R. E., and Roelof, E. C.: Solar wind control of the magnetopause shape, location, and motion, *Journal of Geophysical Research*, 96, 5489–5495, <https://doi.org/10.1029/90JA02464>, 1991. 102, 116
- Sotirelis, T. and Meng, C.-I.: Magnetopause from pressure balance, *Journal of Geophysical Research*, 104, 6889–6898, <https://doi.org/10.1029/1998JA900119>, 1999. 102, 104
- Spreiter, J. R. and Briggs, B. R.: Theoretical determination of the form of the boundary of the solar corpuscular stream produced by interaction with the magnetic dipole field of the Earth, *Journal of Geophysical Research (1896-1977)*, 67, 37–51, <https://doi.org/10.1029/JZ067i001p00037>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/JZ067i001p00037>, 1962. 128
- Tsyganenko, N. A.: Modeling of twisted/warped magnetospheric configurations using the general deformation method, *Journal of Geophysical Research*, 103, 23 551–23 564, <https://doi.org/10.1029/98JA02292>, 1998. 102
- Tsyganenko, N. A. and Sitnov, M. I.: Modeling the dynamics of the inner magnetosphere during strong geomagnetic storms, *Journal of Geophysical Research: Space Physics*, 110, <https://doi.org/10.1029/2004JA010798>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2004JA010798>, 2005. 130, 132, 133
- Šafránková, J., Němeček, Z., Dušík, v., Přeč, L., Sibeck, D. G., and Borodkova, N. N.: The magnetopause shape and location: a comparison of the Interball and Geotail observations with models, *Annales Geophysicae*, 20, 301–309, <https://doi.org/10.5194/angeo-20-301-2002>, URL <https://www.ann-geophys.net/20/301/2002/>, 2002. 102, 103, 104, 107, 112, 113, 122, 128
- Wang, Y., Sibeck, D. G., Merka, J., Boardsen, S. A., Karimabadi, H., Sipes, T. B., Šafránková, J., Jelínek, K., and Lin, R.: A new three-dimensional magnetopause model with a support vector regression machine and a large database of multiple spacecraft observations, *Journal of Geophysical Research (Space Physics)*, 118, 2173–2184, <https://doi.org/10.1002/jgra.50226>, 2013. 102, 107, 108, 112, 113, 116, 117, 118, 128
- Yang, Y.-H., Chao, J. K., Lin, C.-H., Shue, J.-H., Wang, X.-Y., Song, P., Russell, C. T., Lepping, R. P., and Lazarus, A. J.: Comparison of three magnetopause prediction models under extreme solar wind conditions, *Journal of Geophysical Research: Space Physics*, 107, SMP 3–1–SMP 3–9, <https://doi.org/10.1029/2001JA000079>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2001JA000079>, 2002. 112
- Zhou, X.-W. and Russell, C. T.: The location of the high-latitude polar cusp and the shape of the surrounding magnetopause, *Journal of Geophysical Research: Space Physics*, 102, 105–110, <https://doi.org/10.1029/96JA02702>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/96JA02702>, 1997. 103, 128, 132
- Šafránková, J., Dušík, , and Němeček, Z.: The shape and location of the high-latitude magnetopause, *Advances in Space Research*, 36, 1934 – 1939, <https://doi.org/https://doi.org/10.1016/j.asr.2004.05.009>, URL <http://www.sciencedirect.com/science/article/pii/>

[S0273117705004795](#), solar Wind-Magnetosphere-Ionosphere Dynamics and Radiation Models, 2005. [103](#)

## Chapter Summary

- In this chapter, we use the magnetopause crossings catalog generated in the Chapter 4 combined to an online accessible catalog of crossings by older missions to perform a statistical analysis of the position and shape of the magnetopause for various solar wind and seasonal conditions.
- The results of this study confirm a certain number of long-proven properties of the magnetopause such as the earthward pushing with an increasing dynamic pressure, the description of the flaring with the inverted trigonometric function, the influence of the IMF  $B_z$  on the standoff distance or the azimuthal asymmetry induced by the seasonal variations of the geomagnetic field.
- No particular dependence on the radial component of the IMF,  $B_x$ , on the stand-off distance was found, showing that the question of this influence is still open.
- We do not notice any dawn-dusk asymmetry once the aberration due to the Earth revolution is corrected.
- We investigate the role played by reconnection by investigating the influence of the IMF clock angle instead of considering the lone  $B_z$ . This allows us to evidence the influence of a changing  $B_y$ , which will affect the shape of the magnetopause by displacing the reconnection sites and the direction of the magnetosphere erosion.
- We condense our result in an analytical empirical magnetopause model that provides a more accurate prediction of the boundary location in the nightside of the magnetosphere. This model is also adapted into a dynamic magnetopause model that predicts the position and shape of the magnetopause at any given time.
- We compare the near-cusp crossings of our dataset to manually detected crossings of the cusp external boundary to prove that the representation we have on the magnetopause is strongly affected by its definition. Defined as the cusp inner boundary, the near-cusp magnetopause presents a clear and expected indentation. When it is defined as the cusp external boundary, the question is still open and requires further investigation although we still notice a depletion in comparison to a non-indented magnetopause model.



## Chapter 6

# Automatic detection of magnetopause plasma flow

*Pour soulever un poids si lourd,  
Sisyphé, il faudrait ton courage !  
Bien qu'on ait du coeur à l'ouvrage,  
L'Art est long et le Temps est court.*

---

Charles Baudelaire  
(Les Fleurs du mal)

### Contents

---

<b>6.1 Introduction</b>	<b>142</b>
6.1.1 Location of the reconnection sites	142
6.1.2 Detection of magnetopause plasma with manually-set thresholds	144
<b>6.2 Construction of the dataset</b>	<b>145</b>
6.2.1 Data	145
6.2.2 Magnetopause crossings	145
6.2.3 THEMIS C dataset	147
<b>6.3 Jet detection pipeline</b>	<b>148</b>
6.3.1 Peak detection	148
6.3.2 Peak classification	148
<b>6.4 Method performance</b>	<b>150</b>
6.4.1 Decision threshold	150
6.4.2 Insight on the pipeline's FNs and FPs	150
6.4.3 Summary	153
<b>6.5 Adaptability of the method</b>	<b>153</b>
6.5.1 Selection of test crossings	153
6.5.2 Decision threshold	154
6.5.3 Errors characterization	157
6.5.4 Global quality	160
<b>6.6 Massive detection of magnetopause plasma jets</b>	<b>161</b>
<b>6.7 Conclusion</b>	<b>162</b>
<b>6.8 Bibliography</b>	<b>163</b>

---

## 6.1 Introduction

### 6.1.1 Location of the reconnection sites

Magnetic reconnection is a key actor of the dynamics of the solar wind-magnetosphere system. The merging and convection of the field lines erodes the magnetosphere in a location that varies with the orientation of the IMF, on the dayside when it is southward, at higher latitudes when it is northward. This effect was especially confirmed by the statistical analysis of the magnetopause location and shape we performed in the last chapter and took into account in the magnetopause model we developed.

Locally, the question of the location of reconnection sites for varying solar wind and seasonal conditions is still an open question. Based on theoretical predictions, [Sonnerup \[1974\]](#) and [Gonzalez and Mozer \[1974\]](#) suggested that only a component of the magnetic fields is actually reconnecting. This so-called component reconnection model was opposed by the geometrical considerations of [Crooker \[1979\]](#) who suggested instead that reconnection occurred where the magnetic fields of both sides are anti-parallel. In practice, in-situ magnetopause observations revealed evidences of reconnection in favor of the anti-parallel model [[Gosling et al., 1991](#); [Petrinec and Fuselier, 2003](#); [Phan et al., 2003](#)] and evidences in favor of the component model [[Fuselier et al., 2005](#); [Gosling et al., 1990](#); [Trattner et al., 2007](#)] and the actual position of the reconnection line possibly lies in a combination of these two scenarii [[Fuselier et al., 2011](#)].

This being said, the orientation of the X-line, and the parameters likely to affect its location have been at the core of a myriad of observational and numerical studies. Simultaneous observations of reconnection evidence in the data of two different, distant spacecraft [[Dunlop et al., 2011](#); [Phan et al., 2006](#)] suggested the occurrence of reconnection on a globally extended line on the magnetopause surface. These findings seem in agreement with resistive MHD simulations who predicted reconnection to occur on a topological line, called the *separator*, discriminating the IMF from the domains of open and closed magnetospheric flux [[Dorelli et al., 2007](#); [Glocer et al., 2016](#); [Komar et al., 2013](#)]. Nevertheless, they do not give clues about whether reconnection occurs all along the separator or is restricted to a subset of the line.

Through the observation of ISEE 2 data, [Gosling et al. \[1990\]](#) inferred reconnection evidences were more likely to be found in the northern (resp. southern) dusk and the northern (resp. southern) dawn regions for a positive (resp. negative)  $B_y$  component. The following observational evidences of both anti-parallel and component reconnection at various latitudes prove a clear influence of the IMF clock angle on the location of reconnection sites. Using MHD simulations, [Hoilijoki et al. \[2014\]](#) and [Peng et al. \[2010\]](#) suggested a northward (resp. southward) motion of the reconnection line for a positive (resp. negative)  $B_x$  component and indicated that this shift was suppressed under high solar wind Alfvén Mach number. These findings were however observed on a very low number of events [[Lavraud et al., 2005](#)] or in the lone case of a southward IMF [[Hoshi et al., 2018](#)] and thus lack of observational confirmation.

Reconnection was also suggested to be strongly affected by the variations of the Earth dipole tilt angle. Both numerical simulations [[Park et al., 2006](#); [Russell et al., 2003](#)] and observations [[Hoshi et al., 2018](#); [Kitamura et al., 2016](#); [Trattner et al., 2012](#)] suggested a southward (resp. northward) shift of the reconnection line during summer (resp. winter) but once again, these findings lack of observational confirmation with an important number of samples.

Finally, additional theoretical and numerical investigations indicated the crucial role played by the variations of the physical parameters across the magnetopause on the efficiency of reconnection, one can especially cite the jump of both the magnetic field amplitude and the density [[Cassak and Shay, 2007](#)] or the shear flow [[Cassak and Otto, 2011](#)].

All of these parametrical considerations were completed by the numerical investigation of the local orientation of the X-line. Such studies led to the development of a multitude of models which allow the step by step construction of a global X-line over the whole magnetopause surface, often based on the maximisation of a given quantity. Among them, one can especially cite the orientation of the reconnection line following the bisection between the magnetosheath and the magne-

tospheric fields, [Aunai et al., 2016; Moore et al., 2002] or the maximisation of the outflow speed [Schreier et al., 2010; Swisdak and Drake, 2007].

If the numerical comparison of these models showed that all of them did reasonably well under a southward IMF [Komar et al., 2015], comparing these models to in-situ data is a tricky task as it requires knowledge of the concerned parameters that are not available for in-situ spacecraft measurement. The consideration of a single event at a time only gives indications on the local orientation of the X-line without providing any global answer elements. The Polar observations of Trattner et al. [2007] led to the development of the maximum shear angle model which states that reconnection occurs along the line where the shear between the magnetosheath and the magnetospheric field lines is maximized. From then on, numerous observational studies gave consistency to this model [Petrinec et al., 2011; Trattner et al., 2012; Vines et al., 2017]. It was even used in the MMS design effort to predict the encounters of reconnection sites [Fuselier et al., 2016]. Nevertheless, one does still not know how to link this lone reliance on the shear angle to our understanding of reconnection dynamics and its dependency on physical quantities such as the density, the field amplitude or bulk velocity jumps across the magnetopause. Additionally, the numerous observational tests of this model have only considered the local orientation of the reconnection line for each single event [Souza et al., 2017] or have been done on a small set of reconnection evitences [Trattner et al., 2017]. They also gave no guarantee about whether reconnection occurs all along the line predicted by the model or just on a subset of it.

From now on, an interesting approach we can have stands in collecting as many in-situ evidence of magnetopause plasma flow, defined in the following as *magnetopause plasma jets*, as we can in the data of both equatorial (THEMIS, Double Star, MMS, ...) and polar missions (Cluster, Geotail, ...), superimpose these events together in the form of magnetopause flow maps. Assuming reconnection as the main process at the origin of the magnetopause flow and if they are steady enough, a global pattern would appear in the different produced flow maps and would follow the expected evolution of the X-line as solar wind and seasonal condition change. This is especially what has been done with Double Star data by Trenchi et al. [2008] and with THEMIS data by Hoshi et al. [2018] and represented in the Figure 6.1 for different values of the clock angle  $\Omega$ . This approach proved its potential by showing consistency with the idea of an extended X-line on the magnetopause surface and the suggestions made about the influence of a changing  $B_y$ , a changing  $B_x$  or the seasonal variations. Nevertheless, the small number of events they selected limited the precision of their conclusion and their results concerning the influence of both the IMF clock angle and the Earth dipole tilt angle were limited to the case of a southward IMF.

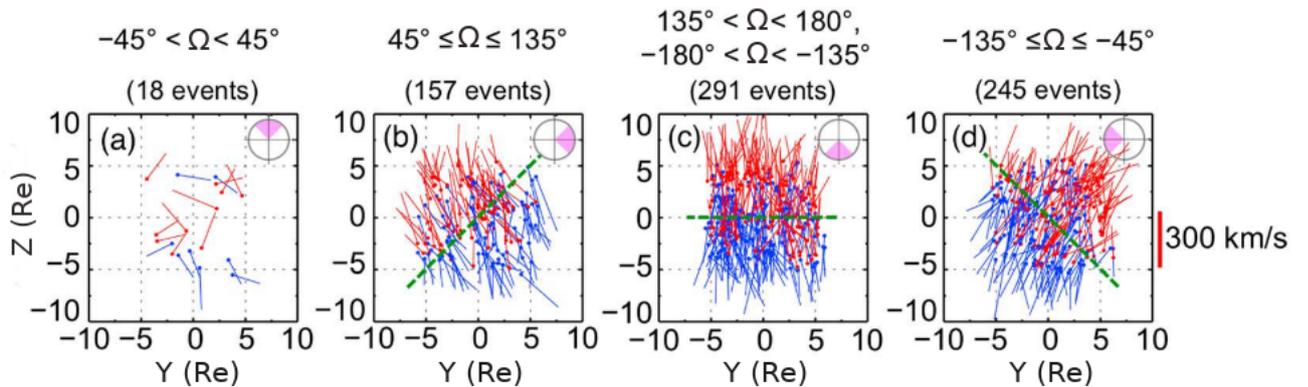


Figure 6.1: Spatial distributions of dayside reconnection jets and their velocities projected in the GSM ( $Y-Z$ ) plane under northward (a), westward (b), southward (c) and eastward (d) IMF. The blue and the red bars represent the southward and the northward jets respectively and their length indicate their relative velocity with the magnetosheath. The green dashed lines represent the rough location of the reconnection line expected by the bisection model [Moore et al., 2002]. Adapted from Hoshi et al. [2018].

A typical in-situ measurement of a magnetopause crossing with a spacecraft going through the reconnection exhaust has been shown in the Chapter 1 in Figure 1.9. There, we identified magnetopause plasma jets by considering the velocity peaks that are higher than the average of what is observed in the surrounding magnetosheath and for which the peaking direction was consistent with the reversed component of the magnetic field. This observation in the plasma moments and magnetic fields was then correlated with an enhanced ion flux at higher energy indicating the mixing of two ions populations. These are the main criteria when one attempts to collect magnetopause plasma jets with the plasma moments and the magnetic field, but the consideration of all of these criteria at once and the decision of their fulfillment is ambiguous because of its dependency on the external observer's decision and because of the variability that exists globally in the dataset<sup>1</sup>.

Here again, the complexity of the magnetopause plasma jets gathering task is a serious bottleneck to these studies that often leads to poorly reproducible catalogs limited to the few most obvious events. Automating this collection task and performing it on all of the available missions measurement then once again appears as a serious milestone in the elaboration of consistently reproducible massive statistical studies of events observed in-situ and would bring us a step further in the study of the location of the X-line through the analysis of the reconnection induced magnetopause plasma flow.

### 6.1.2 Detection of magnetopause plasma with manually-set thresholds

Following the theoretical work of [Levy et al. \[1964\]](#) who predicted that the acceleration could either be due to slow shocks or [Rotational discontinuities \(RDs\)](#) and the numerous observations of Alfvénic flows at the crossing of the magnetopause [[Paschmann et al., 1986](#)], a common option we have to avoid the fastidious manual detection of jets stand in selecting the plasma peak that are faster than the magnetosheath and comparing their relative velocity  $\Delta\vec{V}_{observed}$  to their theoretical expected Alfvén velocity  $\Delta\vec{V}_{expected}$  given by the so-called Walén relation:

$$\Delta\vec{V}_{expected} = \vec{V}_{outflow} - \vec{V}_{inflow} = \pm \left( \frac{1 - \alpha_{inflow}}{\mu_0 \rho_{inflow}} \right)^{1/2} \left[ \vec{B}_{outflow} \left( \frac{1 - \alpha_{outflow}}{1 - \alpha_{inflow}} \right) - \vec{B}_{inflow} \right] \quad (6.1)$$

Where  $\alpha = \mu_0 \frac{p_{\parallel} - p_{\perp}}{B^2}$  is an anisotropic factor with  $p_{\parallel}$  and  $p_{\perp}$  being the thermal pressure parallel and perpendicular to the magnetic field respectively.

If the selected peak corresponds to an Alfvénic flow, then  $\Delta\vec{V}_{observed}$  and  $\Delta\vec{V}_{expected}$  should be aligned or anti-aligned verifying the so-called Walén test [[Paschmann et al., 1986](#)]:

$$R_W = \left| \frac{\Delta\vec{V}_{observed}}{\Delta\vec{V}_{expected}} \right| \sim 1 \text{ and } |\cos(\Theta_W)| = \frac{\Delta\vec{V}_{observed} \cdot \Delta\vec{V}_{expected}}{|\Delta\vec{V}_{observed}| |\Delta\vec{V}_{expected}|} \sim 1 \quad (6.2)$$

Naturally, one should not expect perfect fulfillment of this relation because of the variability of in-situ measurement and because of the ideal assumptions that lead to 6.1. For this reason, the verification of the Walén test is usually done by introducing thresholds on both  $R_W$  and  $\Theta_W$ . Although this test allows the selection of an important number of actual magnetopause plasma jets, it also misses a lot of them while making an important number of false predictions as shown by [Paschmann et al. \[2018\]](#) and are often followed by a post processing step to eliminate the falsely predicted events after their visual inspection or the introduction of additional manually-set thresholds to a method already based on manually-set thresholds. Additionally, this relation implies the manual definition of a reference magnetosheath and the setting of arbitrarily set thresholds that can hardly generalize for different crossings and to different types of missions. Consequently, the automatic detection of reconnection jet still requires improvements.

In the aftermath of what was done in the chapters 3 and 4, the next step that could be taken in this automation process would be the application of machine learning algorithms. The problem

<sup>1</sup>Additional less obvious magnetopause crossings and associated jets are shown in the Appendix B.

here has a lot of common points with the problem we exposed in chapter 3: both problems are about defining beginning and ending dates of events from streaming in-situ data and both signatures are limited by the ambiguity that exists from an observer to another. A consistent attempt in this approach would then to adapt the CNN we used to detect ICME to magnetic magnetopause plasma. Nevertheless, the huge amount of required training data and associated manually labeled jets of different types and the required training time make this solution hardly adaptable in practice. Moreover, the pipeline we designed in the chapter 3 is adapted to the detection of events that have a wide duration dispersion, which is not particularly the case for magnetopause plasma which often have a weak duration in comparison to the resolution of the instrument.

In this chapter, we will elaborate another machine learning based method that automatically detects magnetopause plasma in the data provided by various spacecraft that went across the magnetopause. After presenting the datasets we used, we will detail the different steps of our method and evaluate its performance and its adaptability from a spacecraft to another. The method will finally be used to rapidly build the massive magnetopause plasma catalog fit for further statistical studies, that unfortunately comes out of the scope of this thesis.

## 6.2 Construction of the dataset

### 6.2.1 Data

We use the datasets of the equatorial missions, the five THEMIS spacecraft, Double Star TC1 and MMS 1 that we presented in the chapter 4. With the difficulties of creating a dataset made of particle distribution functions that would be homogeneous for all of the missions we are working on, we once again choose to focus on the plasma magnetic field and moments while keeping the spectrograms for the visual inspection and the easing of our labeling process.

In the following, we will have to identify the magnetospheric and the magnetosheath parts of the different crossings we consider. To do so, we use the THEMIS region classifier we presented in the chapter 4 that was proved also adaptable to Double Star and MMS data.

Even though a work similar to what we do in this chapter could be applied to non-equatorial missions such as Cluster, the observational differences in the typical signature of magnetopause plasma and the associated magnetopause crossings when moving from an equatorial to a polar orbit add another complexity to the problem that is skipped for now.

### 6.2.2 Magnetopause crossings

As our objective is to detect magnetopause plasma occurring at the dayside magnetopause, there is no use of considering each dataset in its globality.

Consequently, we restrict our 7 datasets to the magnetopause crossings we found with the region classifier presented in the previous chapter. Among the 11 634 accessible crossings, we keep those that are in the MLT range from 8 to 16 hr and enlarge the initial 1hr-crossings to ensure the crossing are complete and the spacecraft do returned in the magnetosheath or the magnetosphere after crossing the boundary layer.

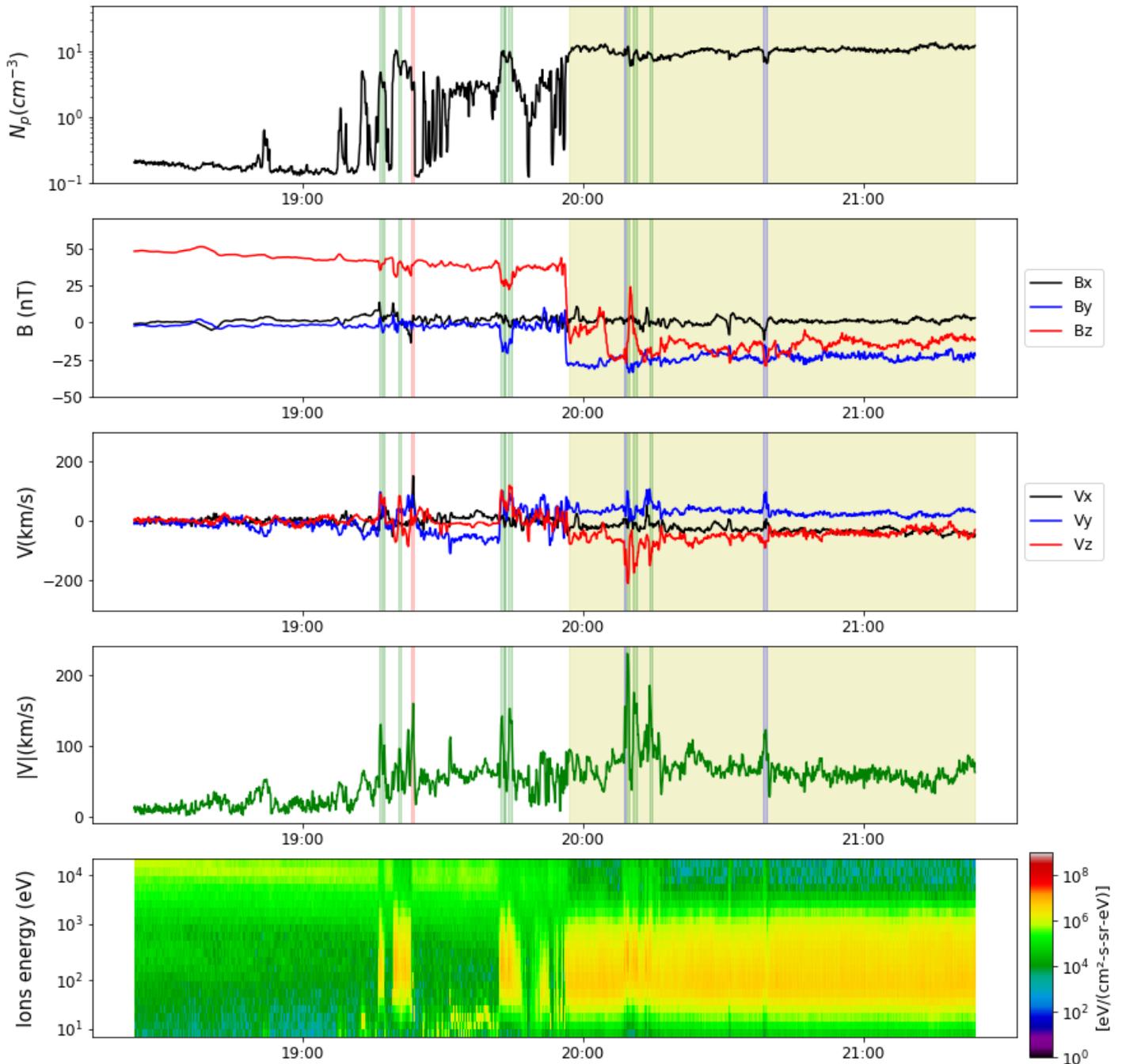


Figure 6.2: In-situ measurement provided by THEMIS C spacecraft during a magnetopause crossing on the 27<sup>th</sup> of July 2009. From the top to the bottom are represented the ion density, the magnetic field components in GSM coordinates, the velocity components and magnitude and the omnidirectional differential energy fluxes of ions. The yellow shading highlight the reference magnetosheath we define. The green intervals indicate the algorithm TP. The blue interval is a FN. The red interval is a FP.

The enlargement of the crossings is done automatically following this process:

1. Consider a 1-hr crossing
2. Iteratively define the beginning time (resp. the ending time) of the crossing 5 minutes earlier (resp. later) until one of the following conditions is fulfilled:
  - The spacecraft crosses the bow shock (i.e the region classifier detects solar wind).
  - The spacecraft comes back in the magnetosphere/magnetosheath (i.e the region classifier detects magnetosphere/magnetosheath in the 5 minutes added interval while it was in the magnetosheath/magnetosphere before resp. after).
  - The spacecraft comes in the vicinity of the Earth's dipole (i.e the magnetic field amplitude goes above 100 nT).
  - The new spacecraft beginning time is 90 minutes after the ending date of a preceding crossing (resp. before the beginning date of a following crossing) (this condition is added to make sure none of the enlarged crossings overlapped one with each other).
  - The total crossing duration is equal to 3 hours.

We finally keep the so-called *non-hesitating crossings* that we define either as crossings for which the region classifier detects less than 4 magnetosheath intervals within the crossing, either as crossings for which the largest detected magnetosheath interval represent more than 75% of the total detected magnetosheath beneath the crossing.

A typical representation of such crossing is shown for [THEMIS C](#) data in [Fig. 6.2](#) where from top to bottom are represented the ion density, the magnetic field [GSM](#) components, the velocity [GSM](#) components and module and the omnidirectional differential ion energy fluxes. The yellow shading depicts the magnetosheath interval detected by the region classifier while the meaning of the blue, green and red rectangles will be explained later-on.

The final list of observed magnetopause crossings is then composed of 7126 events distributed on the 7 spacecraft we consider in this chapter as shown in the [Table 6.3](#).

### 6.2.3 THEMIS C dataset

We use the [THEMIS C](#) dataset to fit and evaluate the algorithm before adapting it to the other spacecraft and missions. This set is made of 240 crossings that we randomly split into a train set made of 164 events and a test set in which there are 76.

The spatial distribution of the two sets is shown in [Figure 6.3](#). The presence of events for all [MLT](#) longitudes shown in the two panels indicate we will give our algorithm the opportunity to take into account the physical diversity of the plasma flow throughout the magnetopause.

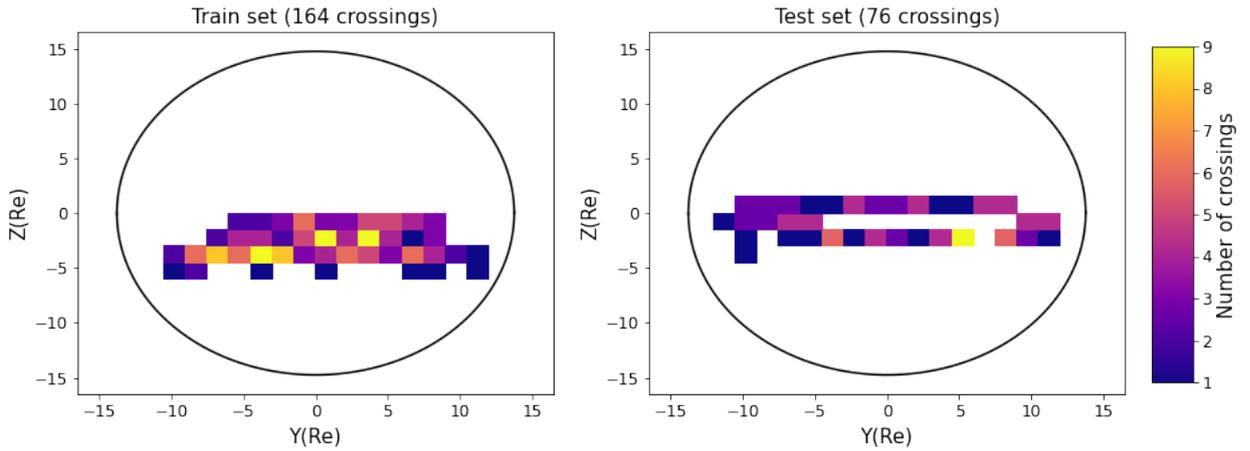


Figure 6.3: Spatial distribution in the **GSM** (Y-Z) plane of the **THEMIS C** crossings that constitute the training (*left*) and the test (*right*) set of our classifier. The solid black line indicate the (Y-Z) projection of the magnetopause model developed in the chapter 5 computed with a dynamic pressure of 2 nPa and a null  $B_z$

## 6.3 Jet detection pipeline

### 6.3.1 Peak detection

For a given magnetopause crossing, we expect such jets to be faster than the average surrounding magnetosheath flow.

Thus, the first step in the detection of magnetopause plasma jets stand in the proper definition of a reference magnetosheath for every considered crossing. We define this reference magnetosheath by the largest magnetosheath interval that has been detected by the region classifier. This reference magnetosheath is represented in the Figure 6.2 by the yellow interval.

For each crossing, we apply a **Minimum Variance Analysis (MVA)** to represent the vectorial quantities (velocity and magnetic field) in the **Local Magnetopause Normal (LMN)** coordinate system<sup>2</sup>. We then select the velocity module peaks that are above the median velocity module of this reference magnetosheath and consider as potential jets the time intervals that corresponds to the half-height width of each peak. Those detected peaks are the one that will be classified as being actual magnetopause plasma or not and some of them are represented by the colored blue, red and green intervals in Figure 6.2.

We then manually inspect every detected peaks of the 240 **THEMIS C** crossings and manually label those that are actual in-situ signature of magnetic magnetopause plasma. Following this process, the train (resp. test) is the made of 9358 (resp. 4534) peaks, 705 (resp. 313) of which being actual magnetopause plasma. The total number of considered crossings and detected jets for **THEMIS C** is reminded in Table 6.1.

### 6.3.2 Peak classification

For each detected peak, we compute the following set of features:

- The **LMN** components of the velocity difference between the peak and the reference magnetosheath,  $\Delta V_l$ ,  $\Delta V_m$  and  $\Delta V_n$ .
- The local variation of the L component of the magnetic field  $\delta B_l$ .
- The ion jet temperature T within the peak.
- The density difference between the peak and the reference magnetosheath,  $\Delta N_p$ .

<sup>2</sup>The definition of such coordinate system and the principle of **MVA** is presented in the Appendix A.

- The module of the velocity  $V$  of the peak.
- The difference between the jet velocity module and the reference magnetosheath median velocity module,  $V - V_{msh}$ .
- The ratio  $(V - V_{msh})/\sigma_{V_{msh}}$  where  $\sigma_{V_{msh}}$  represents the standard deviation of the median velocity module in the reference magnetosheath.

These are the 9 features we will use to classify the detected peaks and determine which of them are magnetopause plasma.

For the ability it has proved (see Chap. 4) to rapidly deal with complex and unbalanced datasets, we choose once again to train a Gradient Boosting Classifier here made of 500 base tree estimators.

Being interested in the massive detection of magnetopause plasma for statistical purpose, we want the algorithm prediction to provide as less FPs as possible even if this implies reducing the recall with an augmented decision threshold.

Naturally, this implies to exploit the probabilistic output of the trained algorithm. We then have to make sure it is well-calibrated. The calibration curve of the algorithm represented by the blue dotted curve in Figure 6.4 indicates the necessity we have to calibrate this probabilistic output. To do so, we apply an Isotonic regression [Niculescu-Mizil and Caruana, 2005] to the probabilistic output of the algorithm. The modified calibration curve is represented by the blue solid line in Figure 6.4. Seeing it better sticking to the dashed grey line that represents the case of a perfectly calibrated classifier, especially for the highest probabilities, then proves the efficiency of such calibration process and gives consistency to the probabilistic output of the algorithm that will be used in the upcoming sections.

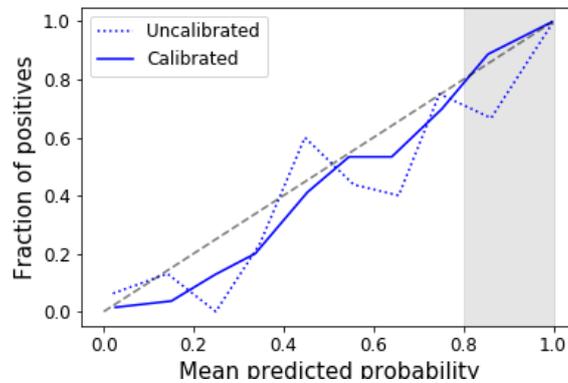


Figure 6.4: Calibration curve of the Gradient Boosting Classifier the peak classification. The gray dashed line represents the curve of a perfectly-calibrated classifier. The blue dotted (resp. solid) line represents the calibration curve of of the Gradient Boosting before (resp. after) applying the Isotonic regression to the probabilistic output of the algorithm. The grey interval represents the probability interval on which we will focus in the following sections.

## 6.4 Method performance

### 6.4.1 Decision threshold

We evaluate the performances of the calibrated gradient boosting classifier fitted to our train set, by comparing the peak classification performed on our test set to our manual jet labeling.

A typical prediction made by our algorithm is shown in Figure 6.2 where the green intervals represent our test set TP. The blue (resp. red) interval is a FN (resp. FP) of the test set. It is already worth noting that, just like the CNNs of chapter 3 and the gradient boosting of chapter 4, the prediction provides an interesting visual indicator in the form of colored shading intervals that can be used by an external observer as suggestions about the patterns in the data likely to be actual magnetopause plasma jets.

Figure 6.5 shows the precision-recall curve we obtain for a varying probabilistic decision threshold. The best compromise we find between a low FP and low FN, the elbow point of the curve, is reached for a decision threshold of 0.5 and leads to a precision and a recall of 0.86.

The leftward triangle indicates the precision and the recall we have when we apply the Walén test with the typical thresholds values :  $0.4 < R_W < 3$  and  $|\Theta_W - 90^\circ| > 60^\circ$ <sup>3</sup>.

In this case, we notice both a low precision ( $\sim 20\%$ ) and a low recall ( $\sim 20\%$ ), this is not surprising as we expect a significant part of the peaks in the test set to have a low  $V - V_{msh}$ , coming from the velocity fluctuations that can be found in the crossing reference magnetosheath. Such peaks are usually removed by the setting of a velocity threshold before applying the application of a Walén test [Hoshi et al., 2018; Paschmann et al., 2018].

Thus, we only apply the Walén test to the peaks of the test set for which  $V - V_{msh} > 150 \text{ km/s}$ , which is the criteria used in Hoshi et al. [2018] and Trenchi et al. [2008]. In this case, the obtained precision and recall is shown by the rightward triangle. Although the addition of this 5<sup>th</sup> threshold does reduce efficiently the total number of FPs (for a precision now equal to  $\sim 85\%$ ), it still leads to an important number of FNs ( Seen in the recall now equal to  $\sim 50\%$ ).

Similarly to what we noticed in Figure 3.10, equivalent values of recall (resp. precision), leads to higher precision (resp. recall) for our method. This indicates the efficiency of our method in comparison to the manual set of threshold to ensure the respect of the rotational discontinuity relation.

Additionally, the Gradient Boosting Classifier presents the advantage of coming with a single degree of freedom when it comes to the choice of the decision threshold, which is not the case with the Walén test for which we can count 4 or 5 of them depending on the setting of a velocity threshold or not.

### 6.4.2 Insight on the pipeline's FNs and FPs

For the purpose of physical studies, we want the massively detected event lists to contain as few FPs as possible even if this implies reducing the recall<sup>4</sup>. To do so, we select the working point indicated by the black dot on the Figure 6.5.

At this point, which corresponds to a decision threshold of 0.8, the algorithm correctly classifies 230 of the 313 jets of the test set (or in other words gives a recall of 73%) and only returns 9 FPs (which gives a precision of 96%). In order to understand the origin of the errors made by the algorithm, a characterization of these FNs and FPs is needed.

<sup>3</sup>These are the typical values used by Hoshi et al. [2018] and Trenchi et al. [2008] for instance.

<sup>4</sup>Without missing the greatest part of the existing jets of course.

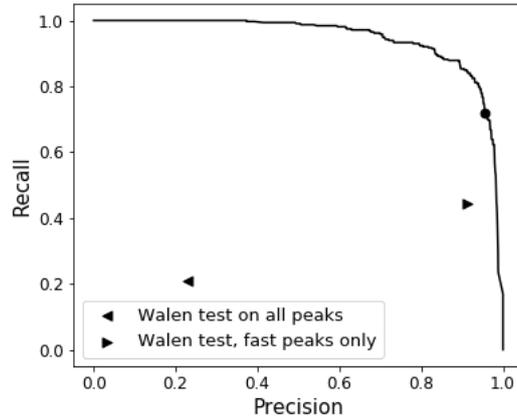


Figure 6.5: Precision-recall curve of the calibrated gradient boosting algorithm. The left pointing black triangle indicate the performances reached by the Walén Test on the test while the right pointing black triangle indicate the performances of the Walén Test on the fastest peaks only. The black dot indicate the value of the precision and the recall for the decision threshold we chose in the next sections.

A visual inspection reveals that the 9 obtained FPs can be divided into 3 categories:

- 4 of them are peaks that look like magnetopause plasma but are either too weak in comparison to the reference magnetosheath, either isolated in the magnetosheath and thus temporally far from the the other detected magnetopause plasma within the crossing.
- 2 of them have a velocity that peaks in a direction uncorrelated to the observed change in the magnetic field orientation
- 3 of them are actual magnetopause plasma that were forgotten in the labeling process

Even if their low number prevents us from extracting any statistical information about their nature, this distribution gives us trends on how the FP present in the upcoming massive jet detection can be characterised.

Due to their higher number, a more consistent statistical insight can be provided on the FNs.

Figure 6.6 shows the scatter plot of  $\Theta_W$  versus  $R_W$  computed for each of the TPs (left panel, green scatter) and each of the FNs (right panel, blue scatter). The color scale here represents the velocity difference the event has with its associated reference magnetosheath  $V - V_{msh}$  and the gray zones represents the zones of the plan ( $R_W$ ,  $\Theta_W$ ) a given event has to fall into to be considered as a reconnection jet by the Walén test. The boundaries of these zones have been set using the habitual thresholds.

The distribution of the TPs in the left panel is consistent to what is shown in [Paschmann et al., 2018] as the greatest part of the fastest jets are located either in the gray zones, either in their surroundings, most of them being located in the  $50^\circ < \Theta_W < 120^\circ$  and  $R_w < 1$  region. On the opposite, the greatest part of the FNs are outside of the gray regions and do fail the Walén test.

Thus, the majority of the FNs are jets for which the rotational discontinuity relation is not respected. Nevertheless also having TPs that fail the Walén test proves this is not the only parameter we can look at to fully describe those FPs. Additionally, this shows the difficulties faced by the Walén test to provide an exhaustive collection of magnetopause plasma jets as well as the observer bias introduced by the application of manual thresholds when performing this test and consequently confirms the interest of the construction of more elaborated methods such as ours.

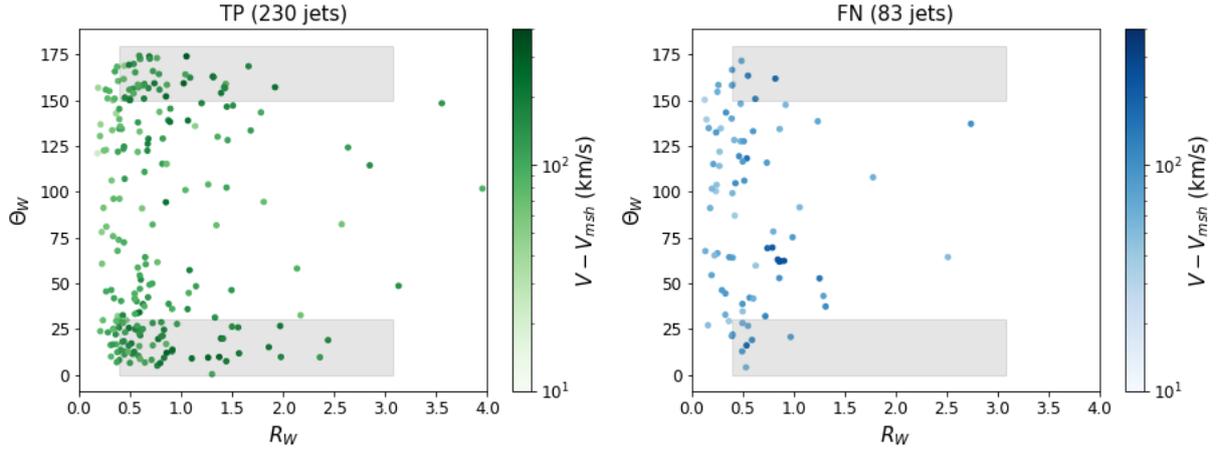


Figure 6.6: Walén test results of the **TPs** (left) and the **FNs** (right) of the test set for a decision threshold of 0.8. Each dot represent the Walén Angle  $\Theta_W$  of a given jet as a function of its Walén ratio  $R_W$ . The dots are colored according to their velocity in comparison the velocity of their associated reference magnetosheath. The grays intervals represents the criterias set the region in the  $(R_W, \Theta_W)$  space defined by the typical Walén test thresholds.

Another feature we could look at is the temporal distance of the **FNs** in comparison to the detected **TPs** of the same crossing and about how well these events are separated from the surrounding reference magnetosheath. Figure 6.7 shows the **Kernel Density Estimation (KDE)** we obtain for both **TPs** and **FNs** of the three different parameters:

- The distance to the closest **TP** within the same crossing (first panel).
- The ratio  $\frac{N_{10}}{N_{crossing}}$  where  $N_{10}$  is the number of **TPs** in a 10 minutes interval centered around the event and  $N_{crossing}$  is the total number of **TPs** in the associated magnetopause crossing (second panel).
- The difference  $V - V_{msh}$  between the jet velocity and the median velocity of the reference magnetosheath (third panel)

To obtain this **KDE**, we consider each value of the concerned parameter as a Gaussian density function centered on this observation and sum these densities altogether for each concerned events list.

Looking at the first panel, we notice similar distribution. This indicates that the **FNs** are not especially more isolated from the surrounding other jets than what it is for the **TPs**. Looking at the second panel now, the **FNs** appear to be less surrounded by **TPs** than the **TPs** themselves. Consequently, the **FNs** are more likely to occur at the beginning or at the ending of a time interval in which the proportion of jets is important e.g. when the spacecraft just enters or is about to leave the reconnection outflow, where the signatures of reconnection might be the hardest to distinguish from the reference magnetosheath. This trend is confirmed with the different observed distributions we have in the third panel that shows the **FNs** tend to be slower than the **TPs** and then harder to distinguish from the reference magnetosheath.

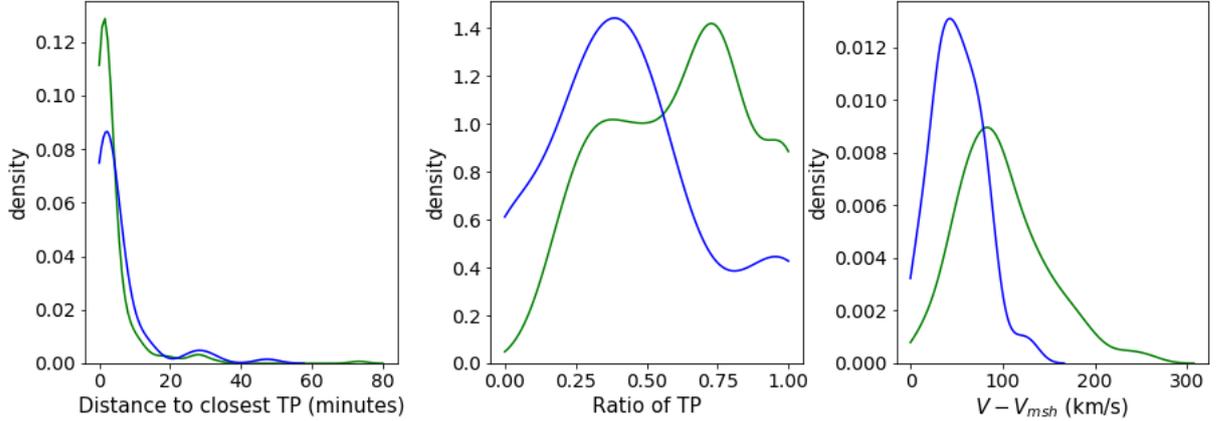


Figure 6.7: KDE of the distance of a given jet to the closest TP (*left*), of the proportion of TPs in a 10 minutes interval around a given jet (*middle*) and the difference of velocity between a given jet and its associated reference magnetosheath (*right*) for both the TPs (green curves) and the FNs (blue curves) of the THEMIS C test set

### 6.4.3 Summary

In this section, we showed the efficiency of a calibrated gradient boosting classifier to automatically detect the evidence of magnetopause plasma jets.

For the purpose of physical studies, setting a high decision threshold leads to a precision of 96% with FPs that are mostly jet-like but hardly distinguishable from the associated reference magnetosheath or that have been forgotten during the labeling process.

This decision threshold also leads to a recall of 73% with a great part of the missed events that corresponds to the beginning or the ending time of the crossing of the outflow by the spacecraft or slow events, hard to distinguish from the reference magnetosheath that are furthermore likely to fail the Walén test.

## 6.5 Adaptability of the method

### 6.5.1 Selection of test crossings

Having shown the efficiency of our method on the THEMIS C dataset, we expect it to be easily adaptable to the data of additional missions in a similar way to what was shown with the region classifier in the previous chapter. To do so, we randomly select 90 magnetopause crossings of THEMIS A, MMS and Double Star and manually inspect the detected velocity peaks of each crossing to label those that are actual magnetopause plasma in a similar way to what was done for THEMIS C.

We make sure our crossing selection is still representative of all of our accessible longitude by looking at their spatial distribution shown in the Figure 6.8 in the (Y-Z) GSM plane for the three spacecraft separately and considered altogether and indicate the total number of labeled events in Table 6.1.

Spacecraft	Concerned crossings	Number of Jets
THEMIS A	90	417
THEMIS C	240	1018
Double Star	90	722
MMS	90	583
Total	510	2740

Table 6.1: Number of manually labeled magnetopause plasma and associated reconnecting magnetopause crossings we have for different spacecrafts

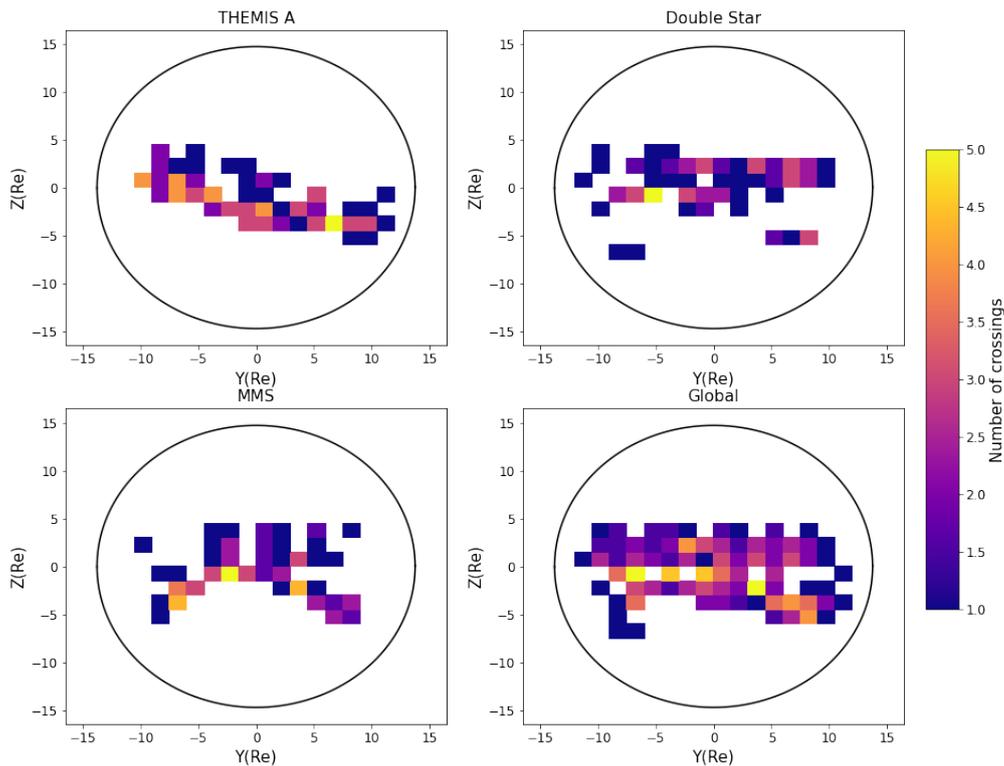


Figure 6.8: Spatial distribution in the GSM (Y-Z) plane of the test crossings we selected for different missions: THEMIS A (upper left), Double Star (upper right), MMS (bottom left) and the three missions altogether (bottom right). The solid black line is similar as the one shown in Fig. 6.3

### 6.5.2 Decision threshold

Typical predictions made by our peak detector are shown for Double Star and MMS in the Figures 6.9 and 6.10 with a disposition of panels and legends that are the same than in Figure 6.2. At first sight, the given prediction looks consistent with the one we had for THEMIS C and we perform a study similar to what is done in the previous section to ensure it.

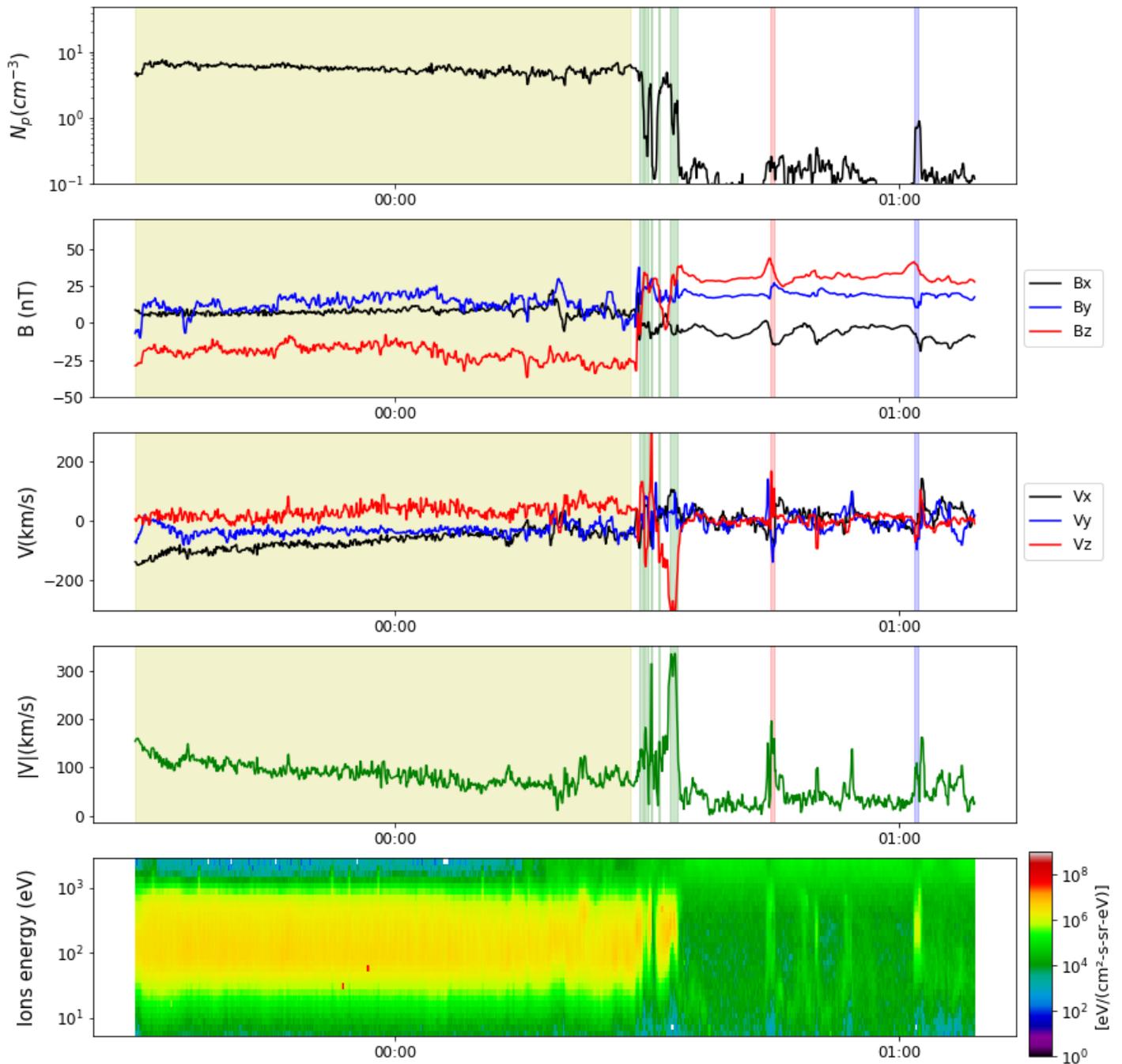


Figure 6.9: In-situ measurements provided by the Double Star TC1 spacecraft on the 16<sup>th</sup> of April 2005 that include jets detected by our model. The legends are the same than in Figure 6.2

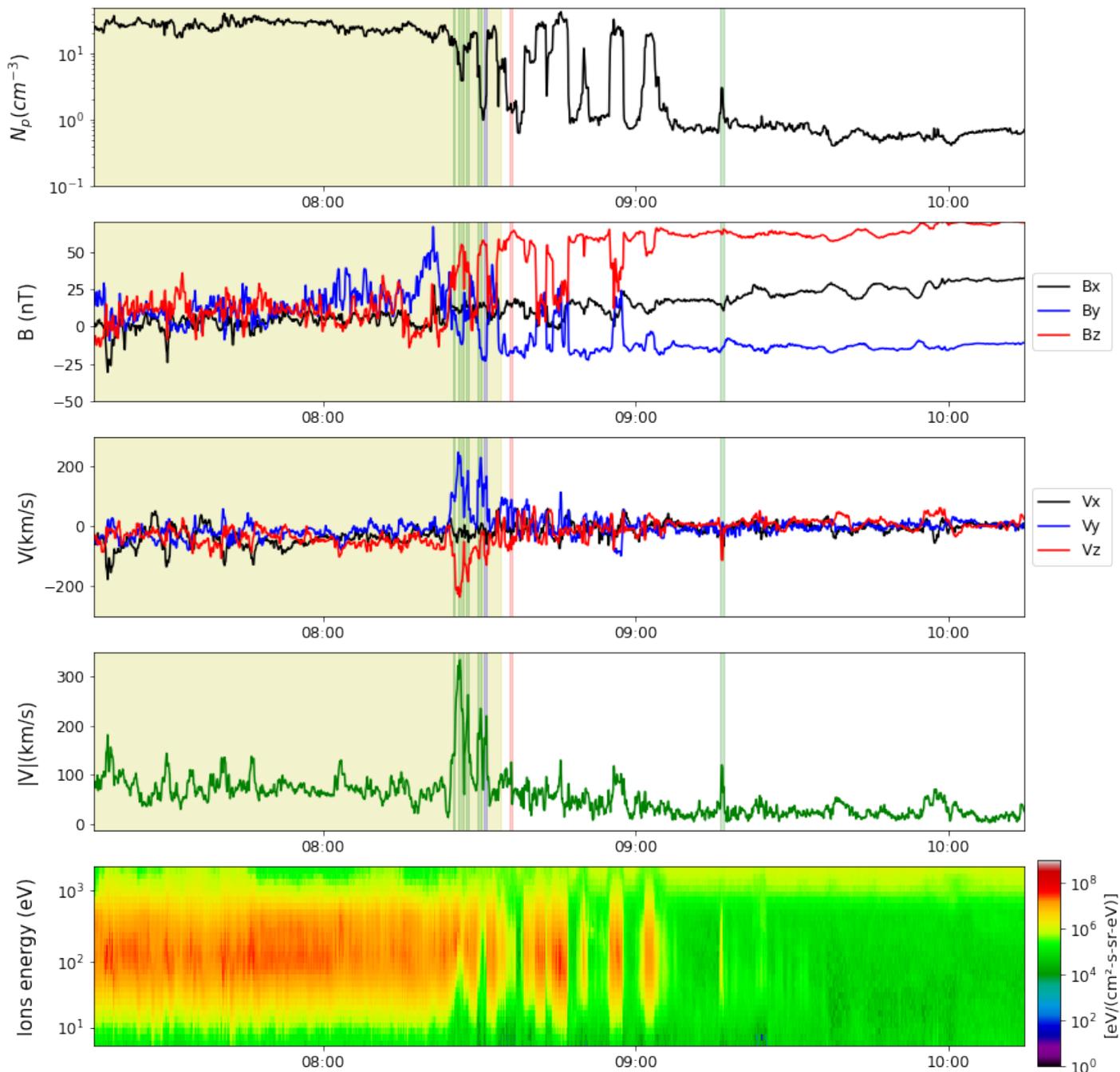


Figure 6.10: In-situ measurements provided by the MMS 1 spacecraft on the 4<sup>th</sup> of January 2016 that include jets detected by our model. The legends are the same than in Figure 6.2

Figure 6.11 shows the precision-recall curves we have for our three missions separately and altogether. Even if the recall drop is here more abrupt than the one we have for [THEMIS C](#), the best recall-precision compromise we can find is above 80% for each panel and is thus consistent with what we show in Figure 6.5. This proves the method can be easily extended to a wide range of equatorial missions provided they offer the same features.

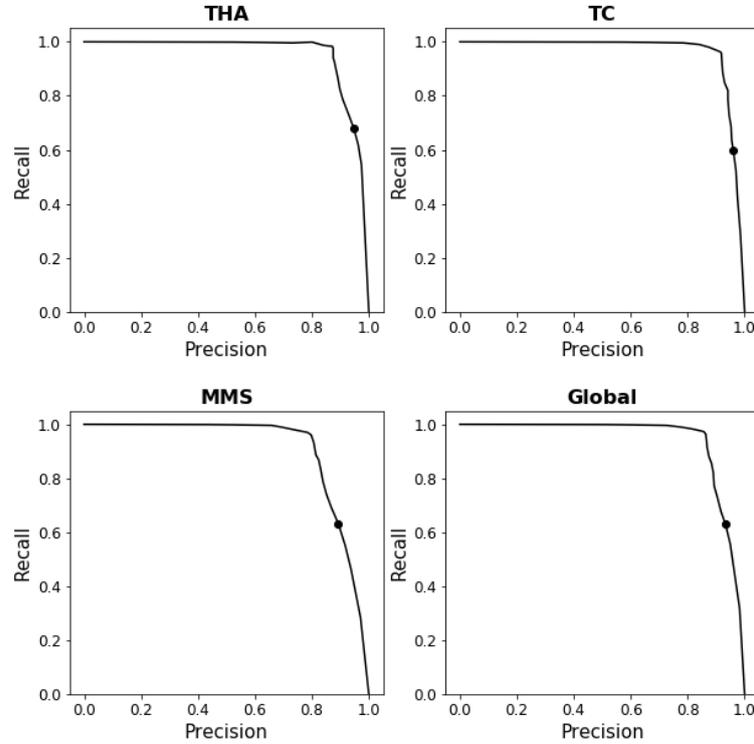


Figure 6.11: Precision-recall curves of our model on the subsets of the magnetopause crossings of [THEMIS A](#) (upper left), [Double Star](#) (upper right), [MMS](#) (bottom left) and for the three missions at once (bottom right). The black dot corresponds to the performances we have for a probability threshold of 0.8

### 6.5.3 Errors characterization

Being consistent to what is done with [THEMIS C](#), we characterize the errors made by the algorithm by selecting the working point that corresponds to a decision threshold of 0.8 for each of the three spacecraft. The associated precisions and recalls are represented by the black points in the four panels of Figure 6.11.

At this working point, the algorithm detects 1163 jets, 78 of which being [FPs](#) (for a precision of 93%) and misses 637 labeled jets (for a recall of 63%). The performances for the three spacecraft are detailed in Table 6.2.

At first sight, the lower precision we have for [MMS](#) could be a serious bottleneck to the massive detection as the proportion of [FPs](#) will be more important for this mission than the others at a given decision threshold. Nevertheless, this precision is linked to the value we chose for the decision threshold and we could perfectly increase it in this specific case in order to reach a precision equivalent to the one we obtain for [THEMIS](#) and [Double Star](#).

Spacecraft	Number of TPs	Number of FNs	Number of FPs	Recall	Precision
THEMIS A	284	133	16	0.68	0.95
Double Star	433	289	18	0.6	0.96
MMS	368	215	44	0.63	0.89
Total	1085	637	78	0.63	0.93

Table 6.2: Number of TPs, FNs, FPs and obtained precision and recall for the three different missions for a decision threshold of 0.8

The multi-mission equivalent of Figure 6.6 is shown in Figure 6.12 where we also added the same scatter plot for the FPs (middle panel, red scatter) that are here more numerous and can then be considered statistically speaking.

The distribution we observe for the TPs in the top left panel is once again consistent with the one observed in Paschmann et al. [2018] with a faster majority being present in the gray zones or in their vicinity. This distribution is once again much less localized in the case of the FNs even though, a larger proportion of events here respect the Walén test. This once again shows a significant part of the missed jets are far from succeeding the Walén test in addition to be statistically slower than the TPs and thus harder to distinguish from the surrounding reference magnetosheath.

The same distribution is more scattered for the FPs and no particular trend can be inferred from the bottom middle panel. With this in mind, we still notice some events, particularly the fastest that pass the Walén test or fall in its vicinity. And some of these events could perfectly be actual jets we forgot during the labeling process as this was a possibility mentioned with the FPs of THEMIS C.

Moving on with the characterization, similar KDEs to what we presented in Figure 6.7 are shown in Figure 6.13 where the computed KDEs of the FPs are represented with the red curves.

Here, the difference in the distributions we notice between TPs and FNs are similar to the one we had in the Figure 6.7. This confirms the supposition we made on how the majority of these FNs looked like and gives a rough idea on the nature of these events we will miss during our massive detection process.

Looking now at the density estimations of the FPs, the left panel shows they are more likely to be found far from the magnetopause crossing where we expect most of the detected jets to be found than the FNs. However, the three distribution in the Figure 6.13 that concern the FPs appear to have a similar evolution than the one observed for the FNs. This proves that, just like the latter, they correspond to events that are either far in the magnetosheath and temporally distant from the majority of the other detected jets, either at the beginning or ending time of the the outflow and could actually correspond to actual jets that were not labeled by omission or because of their weaker in-situ signature. The latter is confirmed by the distribution of the velocity of the FPs that proves these events are statistically slower than the other detected jets and thus present a less obvious signature in the data.

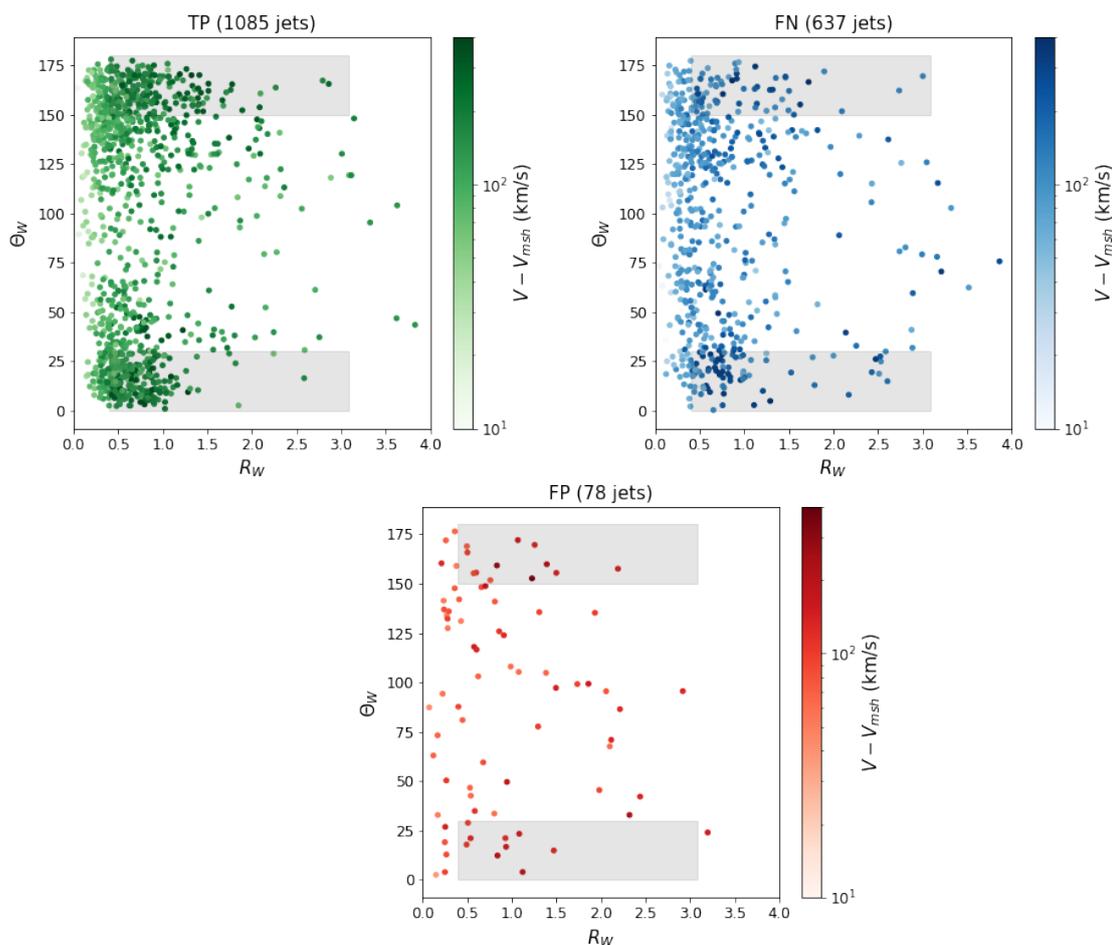


Figure 6.12: Walén tests results for the TPs (left), FNs (right) and FPs (middle). The legends are the same than in Fig. 6.6

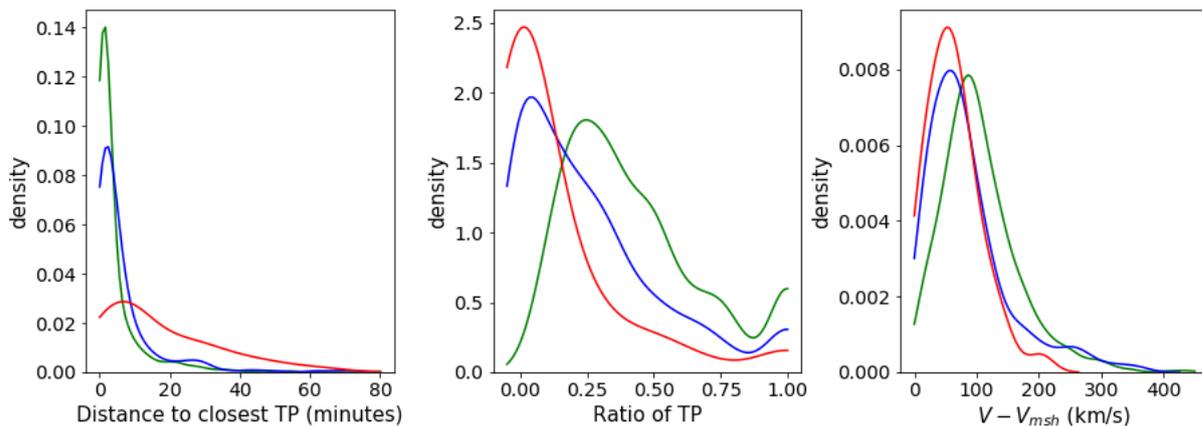


Figure 6.13: Kernel Density Estimation of the distance of a given jet to the closest TPs (top left), of the proportion of TPs in a 10 minutes interval around a given jet (bottom middle) and the difference of velocity between a given jet and its associated reference magnetosheath (top right) for both the TPs (green curves), the FNs (blue curves) and the FPs (red curves) of our multi mission test set

### 6.5.4 Global quality

The clues we evidenced on the nature of the **FNs** and on the **FPs** of the algorithm show that the boundary between a missed, a predicted and a falsely predicted jet is not as sharp as what we would like it to be. This is understandable as, if it existed, the knowledge of such a boundary would have brought us to a perfect detection free of any mistake and the problem of the automatic detection of magnetopause plasma would have been solved with the setting of a single threshold wisely chosen. This lack of sharpness is another proof of the ambiguity, due to the lack of consensus on the definition of a jet and to the difference of perception from an observer to another, that is hidden behind this problem.

In a similar way than what we did in chapter 3, it could then be interesting to compare the wrong predictions made by the algorithm to the differences that exist between the manual selection of magnetopause plasma jets done by two different external human observers. We do it by having two different manual label of the jets of 20 of the 90 crossings considered for each spacecraft in this chapter and we compute the precision associated with the comparison of one event list to another.

The min-max interval of these 6 different precisions is shown with the gray interval in Figure 6.14. The interval expands from 0.83 to 0.98 and never reaches 1. Coupled with the average precision of 0.86, represented by the black dashed line, this is the evidence of the difference of perception that exists from an observer to another and the ambiguity linked with this detection problem.

The green line represented in this Figure is the evolution of the precision obtained on the three spacecraft considered altogether as a function of the chosen decision threshold and the black point represent the precision we have for the value 0.8 we used in this section.

For a wide range of decision thresholds, and particularly the one we used, 0.8, the precision is in this min-max interval and the method then returns jets catalog that contain as much difference between the predicted and the labeled list than the inconsistencies that exist between in the event selection made by two different human experts compared to one another.

This statement is particularly strengthened by the characterization we made on the **FNs** and **FPs** for which we evidenced a significant part of them to be ambiguous events, either isolated from the other predicted **TPs**, either at the start or the end of a time interval that contained a lot of jets. Such events were harder to distinguish from the associated reference magnetosheath, which are typical labeling mistakes that are likely to be made by a human observer.

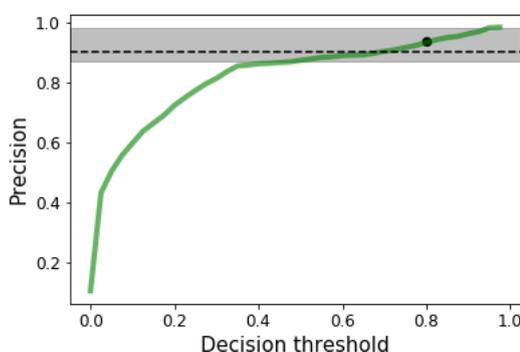


Figure 6.14: Evolution of the precision of the algorithm on the three missions at once as a function of the probabilistic decision threshold. The gray zone (resp. the black dashed line) represents the confidence interval (resp. the average precision) we have between two human made lists compared to one another. The black dot represents the precision we have for the decision threshold we chose in the study

## 6.6 Massive detection of magnetopause plasma jets

In the two previous sections, we showed the reliability and the adaptability of the jet detector, and after having given an insight on the mistakes most likely to be made by the algorithm, we can now perform the massive jet detection by running it on the totality of the crossings we selected for each of our 7 concerned spacecraft. In order to provide lists with as less FPs as possible while still seeing a fair number of jets, we set our decision threshold to 0.8 to decide if a peak had to be kept or not in our final events lists.

The total number of crossings analysed by the method and the total number of associated magnetopause plasma is summarized in the Table 6.3<sup>5</sup>.

Spacecraft	Concerned crossings	Number of Jets
THEMIS A	1924	5245
THEMIS B	171	455
THEMIS C	240	1018
THEMIS D	1385	3745
THEMIS E	1892	5131
Double Star	202	852
MMS 1	379	1511
Total	7126	17957

Table 6.3: Number of magnetopause plasma and associated reconnecting magnetopause crossings we have for different spacecrafts

Obviously, we are aware of the non-exhaustivity of these lists. Nevertheless, the assumed few proportion of FPs and the assumed important number of total detected jets make of these lists the most exhaustive that have been made so far. The spatial distribution of the projected position in the (Y–Z) plane of the jets we detected is shown in the Figure 6.15. As this distribution is balanced in the MLT range we defined, this opens the door for a future statistically representative analysis of the reconnection induced magnetopause plasma flow and the associated position of the X-line.

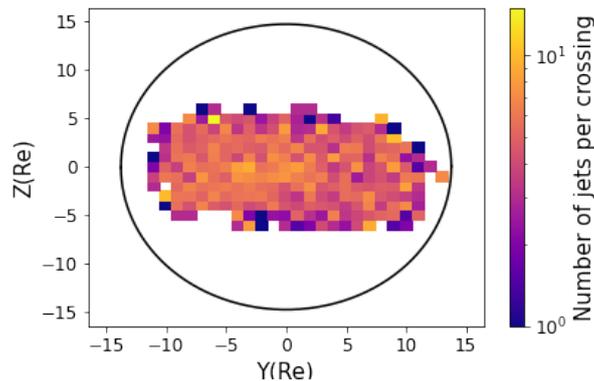


Figure 6.15: Spatial distribution in the (Y-Z) GSM plane of the 17 957 jets detected by the algorithm with a probability above 0.8 The solid black line indicate our magnetopause model with a dynamic pressure of 2 nPa and a null  $B_z$ .

<sup>5</sup>The events list are all accessible here: [https://github.com/gautiernguyen/in-situ\\_Events\\_lists](https://github.com/gautiernguyen/in-situ_Events_lists)

## 6.7 Conclusion

A summary of the process we applied to perform a fast, reproducible detection of magnetopause plasma on the 53 cumulated years of the equatorial data we used in this thesis is shown in Figure 6.16.

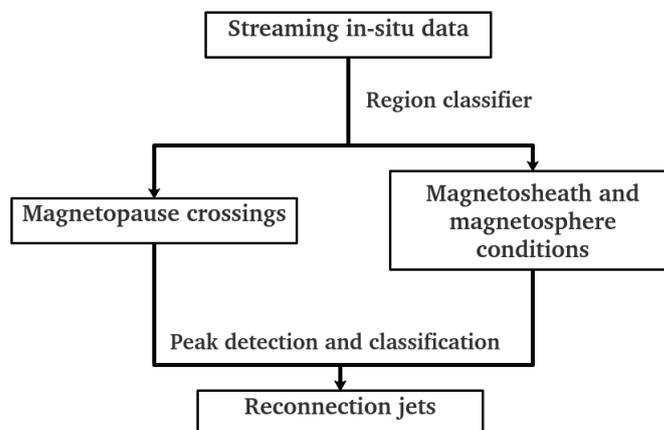


Figure 6.16: Scheme of the pipeline that takes as an input streaming in-situ data and returns a catalog of magnetopause plasma with their associated magnetosheath and magnetosphere conditions

A first gradient boosting classifier is applied to restrict the datasets to the regions in which the spacecraft crossed the magnetopause and retrieve the associated magnetosheath conditions. This is the classifier we developed in the chapter 4.

Using the obtained magnetosheath conditions, we detect the velocity peaks that are faster than the reference magnetosheath and classify them with a second gradient boosting classifier, this time calibrated with an isotonic regression, to determine which of these peaks are actual magnetopause plasma. Similarly to ICMEs, the obtained predictions, the green intervals of Figure 6.2, 6.9 and 6.10 provide an interesting visual indicator that could even be used as an assistant in the frame of the manual labeling of in-situ data.

Naturally, basing our definition of jets on fast flows and the application of a classifier with an imperfect recall necessarily leads to an incomplete and far from being exhaustive events lists. Nevertheless, for the probabilistic decision threshold we choose, the proportion of these FN is small compared to the number of detected jets and we showed that a significant part of them were slower events, harder to distinguish from the magnetosheath, that tended to be at the edges of a region with an important concentration of jets or that were more likely to fail the Walén test. Moreover, this method allows the fast detection of a great quantity of events. We could then easily get on statistically without their consideration.

On the other side, setting a high value of the probabilistic decision threshold results in an even smaller proportion of FPs. Following the conclusion on the statistical insight we gave on them, we show that a majority of them are likely to be either temporally far from the other detected magnetopause plasma within their associated crossing, either located at the beginning or at the ending time of the region where most of the detected jets of a given crossing were found. Moreover, they were on average found slower than the detected TPs and thus also harder to distinguish from the magnetosheath because of their weaker signature.

It is worth noting here that the characteristics of the FPs are pretty much similar to the characteristics of the FNs and that either one or the other side could perfectly have been labeled (respectively unlabeled) as an actual reconnection jet and that the two types of mistakes made by the peak classifier could perfectly have been made by an external observer in the labeling phase of the elaboration of the method. This is confirmed by the comparison of the performances of our algorithm to the difference between the manual classifications of two observers on the same set of peaks.

For the third time in three applications of supervised machine learning algorithms, we prove that the quality of the detection of the in-situ signature of a specific event in the streaming time-series measurement of a spacecraft is highly linked to the interpretation of the data and the definition given to those signatures by an observer. The ambiguity, inherent to the analysis of in-situ data, present from an observer to another is then also found in the prediction made by these algorithms and is consequently the main limiting factor in the utilisation of these methods for this purpose.

It is however not a sufficient reason to completely do without the potential of such algorithms in the frame of the study of the solar wind-magnetosphere coupling as they proved all along this thesis their capacity to provide a fast, reproducible detection of the events of interest while outperforming the quality of the prediction made by manually-set thresholds.

Coming back to magnetopause plasma, we used the pipeline described in the Figure 6.16 to elaborate the largest multi mission catalog of magnetopause plasma ever made and thus, opened the door to study of reconnection at the Earth magnetopause through the statistical analysis of the plasma flow they induce at the interface magnetosphere-magnetosheath and although it comes out of the temporal constraints of this thesis, this would be the logical aftermath of our work.

Concerning the detection strictly speaking, the work presented in this chapter was done for a *MLT* range from 8 to 16 hours and for the most obvious magnetopause crossings and a nice improvement for the jet detection pipeline would then stand in its adaptation to crossings that contain several magnetopause partial crossings and in its opening to a wider *MLT* range.

Having made the detection possible in the equatorial plane, this work could be adapted to the data of additional missions with different orbits. For instance, considering the high-altitude crossings by Cluster would be an interesting adaptation of the method described in this chapter. However, and at the light of the discussion we made on the nature of the near-cusp magnetopause in the last chapter, this would first imply a reconsideration of the region classifier in order to detect the actual magnetopause everywhere.

Finally, even though the features we choose to fit the algorithm are much less obvious than the one we used in the two previous chapters, it could be interesting to adapt it to the detection of additional reconnection signature such as the one that can occur in the solar wind or the detection of reconnection evidences in the magnetotail.

## 6.8 Bibliography

- Aunai, N., Hesse, M., Lavraud, B., Dargent, J., and Smets, R.: Orientation of the X-line in asymmetric magnetic reconnection, *Journal of Plasma Physics*, 82, 535820 401, <https://doi.org/10.1017/S0022377816000647>, 2016. 143
- Cassak, P. A. and Otto, A.: Scaling of the magnetic reconnection rate with symmetric shear flow, *Physics of Plasmas*, 18, 074501, <https://doi.org/10.1063/1.3609771>, 2011. 142
- Cassak, P. A. and Shay, M. A.: Scaling of asymmetric magnetic reconnection: General theory and collisional simulations, *Physics of Plasmas*, 14, 102114, <https://doi.org/10.1063/1.2795630>, 2007. 142
- Crooker, N. U.: Dayside merging and cusp geometry, *Journal of Geophysical Research*, 84, 951–959, <https://doi.org/10.1029/JA084iA03p00951>, 1979. 142
- Dorelli, J. C., Bhattacharjee, A., and Raeder, J.: Separator reconnection at Earth's dayside magnetopause under generic northward interplanetary magnetic field conditions, *Journal of Geophysical Research: Space Physics*, 112, <https://doi.org/10.1029/2006JA011877>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2006JA011877>, 2007. 142
- Dunlop, M. W., Zhang, Q. H., Bogdanova, Y. V., Trattner, K. J., Pu, Z., Hasegawa, H., Berchem, J., Taylor, M. G. G. T., Volwerk, M., Eastwood, J. P., Lavraud, B., Shen, C., Shi, J. K., Wang, J.,

- Constantinescu, D., Fazakerley, A. N., Frey, H., Sibeck, D., Escoubet, P., Wild, J. A., Liu, Z. X., and Carr, C.: Magnetopause reconnection across wide local time, *Annales Geophysicae*, 29, 1683–1697, <https://doi.org/10.5194/angeo-29-1683-2011>, 2011. 142
- Fuselier, S. A., Trattner, K. J., Petrinec, S. M., Owen, C. J., and Rème, H.: Computing the reconnection rate at the Earth’s magnetopause using two spacecraft observations, *Journal of Geophysical Research (Space Physics)*, 110, A06212, <https://doi.org/10.1029/2004JA010805>, 2005. 142
- Fuselier, S. A., Trattner, K. J., and Petrinec, S. M.: Antiparallel and component reconnection at the dayside magnetopause, *Journal of Geophysical Research (Space Physics)*, 116, A10227, <https://doi.org/10.1029/2011JA016888>, 2011. 142
- Fuselier, S. A., Lewis, W. S., Schiff, C., Ergun, R., Burch, J. L., Petrinec, S. M., and Trattner, K. J.: Magnetospheric Multiscale Science Mission Profile and Operations, *Space Science Review*, 199, 77–103, <https://doi.org/10.1007/s11214-014-0087-x>, 2016. 143
- Glocer, A., Dorelli, J., Toth, G., Komar, C. M., and Cassak, P. A.: Separator reconnection at the magnetopause for predominantly northward and southward IMF: Techniques and results, *Journal of Geophysical Research: Space Physics*, 121, 140–156, <https://doi.org/10.1002/2015JA021417>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015JA021417>, 2016. 142
- Gonzalez, W. D. and Mozer, F. S.: A quantitative model for the potential resulting from reconnection with an arbitrary interplanetary magnetic field, *Journal of Geophysical Research*, 79, 4186, <https://doi.org/10.1029/JA079i028p04186>, 1974. 142
- Gosling, J. T., Thomsen, M. F., Bame, S. J., Elphic, R. C., and Russell, C. T.: Plasma flow reversals at the dayside magnetopause and the origin of asymmetric polar cap convection, *Journal of Geophysical Research*, 95, 8073–8084, <https://doi.org/10.1029/JA095iA06p08073>, 1990. 142
- Gosling, J. T., Thomsen, M. E., Bame, S. J., Elphic, R. C., and Russell, C. T.: Observations of reconnection of interplanetary and lobe magnetic field lines at the high-latitude magnetopause, *Journal of Geophysical Research: Space Physics*, 96, 14 097–14 106, <https://doi.org/10.1029/91JA01139>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/91JA01139>, 1991. 142
- Hoilijoki, S., Souza, V. M., Walsh, B. M., Janhunen, P., and Palmroth, M.: Magnetopause reconnection and energy conversion as influenced by the dipole tilt and the IMF B<sub>x</sub>, *Journal of Geophysical Research: Space Physics*, 119, 4484–4494, <https://doi.org/10.1002/2013JA019693>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2013JA019693>, 2014. 142
- Hoshi, Y., Hasegawa, H., Kitamura, N., Saito, Y., and Angelopoulos, V.: Seasonal and Solar Wind Control of the Reconnection Line Location on the Earth’s Dayside Magnetopause, *Journal of Geophysical Research (Space Physics)*, 123, 7498–7512, <https://doi.org/10.1029/2018JA025305>, 2018. 142, 143, 150
- Kitamura, N., Hasegawa, H., Saito, Y., Shinohara, I., Yokota, S., Nagai, T., Pollock, C. J., Giles, B. L., Moore, T. E., Dorelli, J. C., Gershman, D. J., Avakov, L. A., Paterson, W. R., Coffey, V. N., Chandler, M. O., Sauvaud, J. A., Lavraud, B., Torbert, R. B., Russell, C. T., Strangeway, R. J., and Burch, J. L.: Shift of the magnetopause reconnection line to the winter hemisphere under southward IMF conditions: Geotail and MMS observations, *Geophysical Research Letters*, 43, 5581–5588, <https://doi.org/10.1002/2016GL069095>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016GL069095>, 2016. 142

- Komar, C. M., Cassak, P. A., Dorelli, J. C., Glocer, A., and Kuznetsova, M. M.: Tracing magnetic separators and their dependence on IMF clock angle in global magnetospheric simulations, *Journal of Geophysical Research: Space Physics*, 118, 4998–5007, <https://doi.org/10.1002/jgra.50479>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/jgra.50479>, 2013. 142
- Komar, C. M., Fermo, R. L., and Cassak, P. A.: Comparative analysis of dayside magnetic reconnection models in global magnetosphere simulations, *Journal of Geophysical Research: Space Physics*, 120, 276–294, <https://doi.org/10.1002/2014JA020587>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2014JA020587>, 2015. 143
- Lavraud, B., Thomsen, M. F., Taylor, M. G. G. T., Wang, Y. L., Phan, T. D., Schwartz, S. J., Elphic, R. C., Fazakerley, A., Rème, H., and Balogh, A.: Characteristics of the magnetosheath electron boundary layer under northward interplanetary magnetic field: Implications for high-latitude reconnection, *Journal of Geophysical Research (Space Physics)*, 110, A06209, <https://doi.org/10.1029/2004JA010808>, 2005. 142
- Levy, R. H., Petschek, H. E., and Siscoe, G. L.: Aerodynamic aspects of the magnetospheric flow, *AIAA Journal*, 2, 2065–2076, <https://doi.org/10.2514/3.2745>, 1964. 144
- Moore, T. E., Fok, M. C., and Chandler, M. O.: The dayside reconnection X line, *Journal of Geophysical Research (Space Physics)*, 107, 1332, <https://doi.org/10.1029/2002JA009381>, 2002. 143
- Niculescu-Mizil, A. and Caruana, R.: Obtaining Calibrated Probabilities from Boosting, in: *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence, UAI'05*, p. 413–420, AUAI Press, Arlington, Virginia, USA, 2005. 149
- Park, K. S., Ogino, T., and Walker, R. J.: On the importance of antiparallel reconnection when the dipole tilt and IMF  $B_y$  are nonzero, *Journal of Geophysical Research: Space Physics*, 111, <https://doi.org/10.1029/2004JA010972>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2004JA010972>, 2006. 142
- Paschmann, G., Papamastorakis, I., Baumjohann, W., Sckopke, N., Carlson, C. W., Sonnerup, B. U. Ö., and Lühr, H.: The magnetopause for large magnetic shear: AMPTE/IRM observations, *Journal of Geophysical Research*, 91, 11 099–11 115, <https://doi.org/10.1029/JA091iA10p11099>, 1986. 144
- Paschmann, G., Haaland, S. E., Phan, T. D., Sonnerup, B. U. Ö., Burch, J. L., Torbert, R. B., Gershman, D. J., Dorelli, J. C., Giles, B. L., Pollock, C., Saito, Y., Lavraud, B., Russell, C. T., Strangeway, R. J., Baumjohann, W., and Fuselier, S. A.: Large-Scale Survey of the Structure of the Dayside Magnetopause by MMS, *Journal of Geophysical Research (Space Physics)*, 123, 2018–2033, <https://doi.org/10.1002/2017JA025121>, 2018. 144, 150, 151, 158
- Peng, Z., Wang, C., and Hu, Y. Q.: Role of IMF  $B_x$  in the solar wind-magnetosphere-ionosphere coupling, *Journal of Geophysical Research: Space Physics*, 115, <https://doi.org/10.1029/2010JA015454>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2010JA015454>, 2010. 142
- Petrinec, S. M. and Fuselier, S. A.: On continuous versus discontinuous neutral lines at the dayside magnetopause for southward interplanetary magnetic field, *Geophysical Research Letters*, 30, <https://doi.org/10.1029/2002GL016565>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2002GL016565>, 2003. 142
- Petrinec, S. M., Dayeh, M. A., Funsten, H. O., Fuselier, S. A., Heirtzler, D., Janzen, P., Kucharek, H., McComas, D. J., Möbius, E., Moore, T. E., Reisenfeld, D. B., Schwadron, N. A., Trattner, K. J., and Wurz, P.: Neutral atom imaging of the magnetospheric cusps, *Journal of Geophysical Research (Space Physics)*, 116, A07203, <https://doi.org/10.1029/2010JA016357>, 2011. 143

- Phan, T., Frey, H. U., Frey, S., Peticolas, L., Fuselier, S., Carlson, C., Rème, H., Bosqued, J.-M., Balogh, A., Dunlop, M., Kistler, L., Mouikis, C., Dandouras, I., Sauvaud, J.-A., Mende, S., McFadden, J., Parks, G., Moebius, E., Klecker, B., Paschmann, G., Fujimoto, M., Petrinec, S., Marcucci, M. F., Korth, A., and Lundin, R.: Simultaneous Cluster and IMAGE observations of cusp reconnection and auroral proton spot for northward IMF, *Geophysical Research Letters*, 30, <https://doi.org/10.1029/2003GL016885>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2003GL016885>, 2003. 142
- Phan, T. D., Hasegawa, H., Fujimoto, M., Oieroset, M., Mukai, T., Lin, R. P., and Paterson, W.: Simultaneous Geotail and Wind observations of reconnection at the subsolar and tail flank magnetopause, *Geophysical Research Letters*, 33, <https://doi.org/10.1029/2006GL025756>, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2006GL025756>, 2006. 142
- Russell, C. T., Wang, Y. L., and Raeder, J.: Possible dipole tilt dependence of dayside magnetopause reconnection, *geophysical research Letters*, 30, 1937, <https://doi.org/10.1029/2003GL017725>, 2003. 142
- Schreier, R., Swisdak, M., Drake, J. F., and Cassak, P. A.: Three-dimensional simulations of the orientation and structure of reconnection X-lines, *Physics of Plasmas*, 17, 110704–110704, <https://doi.org/10.1063/1.3494218>, 2010. 143
- Sonnerup, B. U. Ö.: Magnetopause reconnection rate, *Journal of Geophysical Research*, 79, 1546–1549, <https://doi.org/10.1029/JA079i010p01546>, 1974. 142
- Souza, V. M., Gonzalez, W. D., Sibeck, D. G., Koga, D., Walsh, B. M., and Mendes, O.: Comparative study of three reconnection X line models at the Earth's dayside magnetopause using in situ observations, *Journal of Geophysical Research (Space Physics)*, 122, 4228–4250, <https://doi.org/10.1002/2016JA023790>, 2017. 143
- Swisdak, M. and Drake, J. F.: Orientation of the reconnection X-line, *Geophysical Research Letters*, 34, L11106, <https://doi.org/10.1029/2007GL029815>, 2007. 143
- Trattner, K. J., Mulcock, J. S., Petrinec, S. M., and Fuselier, S. A.: Probing the boundary between antiparallel and component reconnection during southward interplanetary magnetic field conditions, *Journal of Geophysical Research (Space Physics)*, 112, A08210, <https://doi.org/10.1029/2007JA012270>, 2007. 142, 143
- Trattner, K. J., Petrinec, S. M., Fuselier, S. A., and Phan, T. D.: The location of reconnection at the magnetopause: Testing the maximum magnetic shear model with THEMIS observations, *Journal of Geophysical Research (Space Physics)*, 117, A01201, <https://doi.org/10.1029/2011JA016959>, 2012. 142, 143
- Trattner, K. J., Burch, J. L., Ergun, R., Eriksson, S., Fuselier, S. A., Giles, B. L., Gomez, R. G., Grimes, E. W., Lewis, W. S., Mauk, B., Petrinec, S. M., Russell, C. T., Strangeway, R. J., Trenchi, L., and Wilder, F. D.: The MMS Dayside Magnetic Reconnection Locations During Phase 1 and Their Relation to the Predictions of the Maximum Magnetic Shear Model, *Journal of Geophysical Research (Space Physics)*, 122, 11,991–12,005, <https://doi.org/10.1002/2017JA024488>, 2017. 143
- Trenchi, L., Marcucci, M. F., Pallochia, G., Consolini, G., Bavassano Cattaneo, M. B., di Lellis, A. M., Rème, H., Kistler, L., Carr, C. M., and Cao, J. B.: Occurrence of reconnection jets at the dayside magnetopause: Double Star observations, *Journal of Geophysical Research (Space Physics)*, 113, A07S10, <https://doi.org/10.1029/2007JA012774>, 2008. 143, 150
- Vines, S. K., Fuselier, S. A., Petrinec, S. M., Trattner, K. J., and Allen, R. C.: Occurrence frequency and location of magnetic islands at the dayside magnetopause, *Journal of Geophysical Research (Space Physics)*, 122, 4138–4155, <https://doi.org/10.1002/2016JA023524>, 2017. 143

## Chapter Summary

- In this chapter, we combine the magnetopause crossings lists generated in the Chapter 4, the region classifier developed in the same chapter and a calibrated gradient boosting classifier to provide an automatic detection of the magnetopause plasma jets.
- The method is first developed and tested on the in-situ data measurement provided by the [THEMIS C](#) spacecraft. Before being adapted to the data of other spacecraft with an equatorial orbit, the other [THEMIS](#) spacecraft, [MMS](#) and Double Star.
- For every mission we consider, the method we developed performs better than the Walén test and comes with an interesting visual indicator that can be used as an assistant in the frame of manual event labeling.
- We use the jet detection pipeline to generate one of the most exhaustive, generic, multi-mission catalog of magnetopause magnetopause plasma. Using these catalogs in the frame of statistical study of the magnetopause plasma flow induced by reconnection opens the door for a novel investigation on the determination of the position of the reconnection sites for changing solar wind and seasonal conditions.
- By having an insight on both the [FNs](#) and [FP](#) of our test set and by comparing the performances of our algorithm to the manual label of different experts, we show that the detection inconsistencies made by our method could perfectly have been made by an external human observer. This brings another proof of the limitation of the application of supervised machine learning algorithms by our own interpretation of the data and the events they measure.



# Chapter 7

## Conclusions and perspectives

*Nos choix sont comme les rides sur l'eau. Au début, ils semblent minuscules et insignifiants, mais avec le temps, ils peuvent se muer en raz-de-marée dévastateurs.*

---

Socrate

### Contents

---

<b>7.1 Overview</b> . . . . .	<b>170</b>
<b>7.2 Application of supervised machine learning algorithms</b> . . . . .	<b>171</b>
7.2.1 Detection of ICMEs . . . . .	171
7.2.2 Classification of the Near-Earth Regions . . . . .	172
7.2.3 Detection of magnetopause plasma jets . . . . .	173
<b>7.3 Position and shape of the magnetopause</b> . . . . .	<b>173</b>
<b>7.4 Potential of machine learning algorithms and larger perspectives</b> . . . . .	<b>175</b>

---

## 7.1 Overview

The ensemble solar wind-magnetosphere is a complex system which interaction dynamics is ruled by a multitude of physical processes. In the upstream solar wind, large-scale events produced in the solar corona such as ICMEs transport important quantities of plasma, magnetic field and energy which propagation at Earth orbit can generate geomagnetic storms with high impact on the human activity. Around the Earth, the interaction of two different types of plasmas defines different regions of the near-Earth environment separated by two main boundaries, the magnetopause and the bow shock, and generates small-scale physical processes such as Kelvin-Helmholtz instability or magnetic reconnection that have a strong influence on the dynamics of the system by modifying the location and shape of the boundaries or allowing the transfer of mass and momentum between the two parts of the couple.

The study of the different actors of this coupling can be done through the statistical analysis of the in-situ data measurement by spacecraft orbiting the Earth or the Sun. Nevertheless, these studies often rely on a small number of samples usually selected after a time-consuming, ambiguous and poorly reproducible manual selection of events. This necessarily restricts the resulting statistical vision we can have on the different events of interest and spoils the potential of the decades of accumulation of spacecraft in-situ data measurement considered altogether. Improving the automatic event detection methods then appears as a necessity in the frame of the construction of a global, statistically representative vision of the different actors of the Sun-Earth relation.

In this thesis, we take a step further in this direction by providing automatic detection methods based on supervised machine learning algorithms. Although limited by our own interpretation of the data and definition of the events of interest, the predictions of these methods outperform the quality of the predictions made by the previous existing methods based on manually-set thresholds. This allows the exploitation of the greatest majority of the accessible spacecraft data at our disposal through the rapid and reproducible detection of an important number of events in more than 80 cumulated years of observation of the near-Earth environment and thus offers the opportunity to perform statistical studies of the different actors of the solar-wind magnetosphere coupling with an important number of samples. We apply such methods on three specific actors of the relation: ICMEs, the different near-Earth regions and the magnetic reconnection plasma jets at the magnetopause. We then use the massive detection of magnetopause crossings issued from the regions classifier to investigate how the location and shape of the magnetopause vary with different solar wind conditions.

The different results presented in this thesis are summarized<sup>1</sup> and put in the global context of the Sun-Earth interaction in the Figure 7.1. We come back more precisely on these different results and the precise perspectives they offer in the two next sections.

---

<sup>1</sup>An advised reader will recognize here the Figures 3.2, 4.1, 5.20, 5.28 and 6.2.

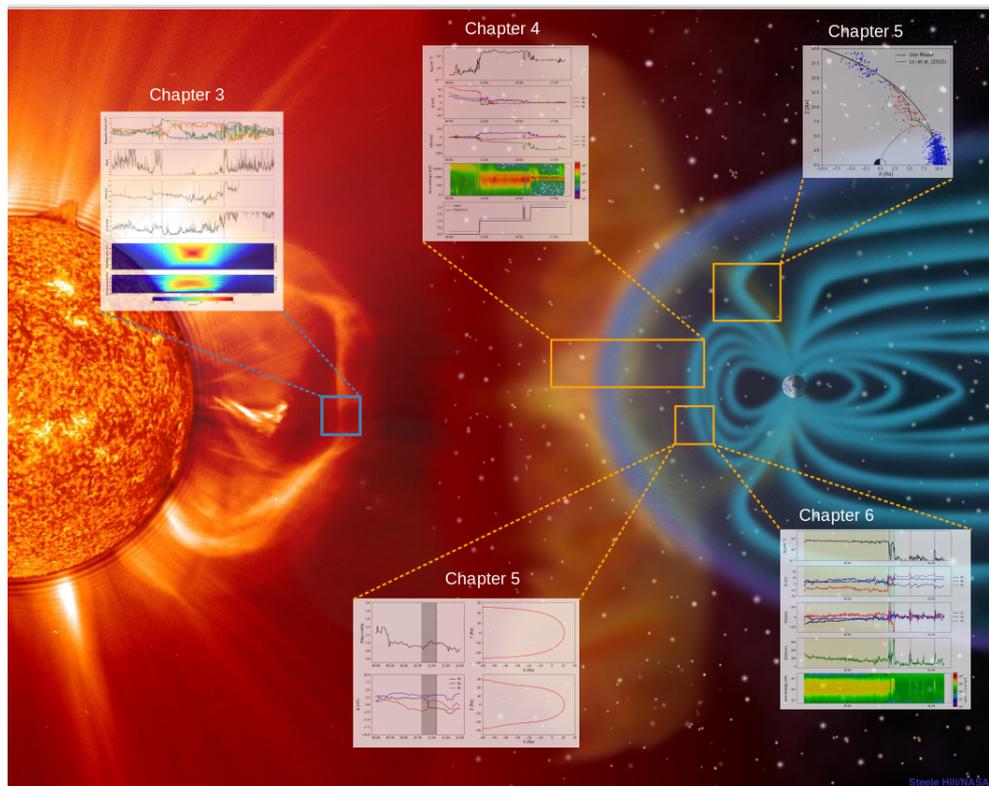


Figure 7.1: Non exhaustive summary of the results obtained in this thesis replaced in the entire context of the Sun-Earth interaction (Adapted from the artist view of Steel Hill: [https://www.nasa.gov/mission\\_pages/hinode/solar\\_004.html](https://www.nasa.gov/mission_pages/hinode/solar_004.html))

## 7.2 Application of supervised machine learning algorithms

### 7.2.1 Detection of ICMEs

In the Chapter 3, we used the data of the solar monitor WIND between 1997 and 2016 to elaborate an automatic detection method of ICMEs at their arrival at the Earth Lagrange point L1. To do so, we trained CNNs to estimate the extent at which a window of data was similar to the typical in-situ signature of an ICME through the prediction of a so-called similarity parameter. Due to the wide dispersion of the duration of these events, we considered 100 CNNs for 100 different windows sizes, from 1 to 100 hr.

Stacked together, the prediction of this ensemble of CNNs returned a 2D-similarity map where the intervals of data likely to correspond to the in-situ signature of one or several events and this constituted a first interesting multi-scale visual indicator in the frame of the manual selection of ICMEs<sup>2</sup>. After a small post-processing, we exploited these maps to rapidly generate reproducible catalogs of events.

The varying performances of this detection for a changing similarity decision threshold evidenced two regimes of detection of our pipeline. At high recall, a non negligible part of the FPs appear to have an ICME-like in-situ signature and the FNs always missed by the detection actually correspond to ambiguous events that would not have been labeled by every experts. This regime can be used to complete the existing catalogs through the selection of additional events and we used it to complete the list of recorded ICMEs between 1997 and 2016 through the detection of 148 additional events in this period. At high precision, the pipeline produces a small part of FPs while still seeing an important number of events. This regime can thus be used to perform the so-desired statistical analysis of the different properties of ICMEs and investigate, for instance, the

<sup>2</sup>The 2D-similarity maps established by this method for the 23 years of the WIND data in use in this thesis can be found here: [https://hephaistos.lpp.polytechnique.fr/data/machine\\_learning/ICME/index.html](https://hephaistos.lpp.polytechnique.fr/data/machine_learning/ICME/index.html)

link between the properties of such events and their geoeffectiveness.

Although less accurate, we showed that the pipeline still provides a decent detection of ICMEs when one or several input features is missing. Additionally, the method showed its capacity to improve itself with an increasing amount of data. These two results suggest the possibility and the interest we would have in adapting the pipeline to the data of other missions that have been monitoring the solar wind at Earth orbit or elsewhere. They would help increase the quality of the detection provided by the algorithm and would offer the unique opportunity to study given ICMEs for various points in the interplanetary medium and thus obtain more precise information on their spatial structure and propagation throughout the solar system.

Another interest of the pipeline we developed is the fact it does not require specific physical knowledge about the event it detects and it could thus perfectly be adapted to any large-scale spatial structure that propagates in the solar wind, one could think of the CIRs we briefly mentioned in the Chapter 1 or the sheath of the ICMEs for instance.

Finally, we proved, through the comparison of different human-made lists, that the difference between the list predicted by the pipeline and the RL used to train and evaluate it was comparable to the differences between the lists established by two different human experts on the same set of data. Consequently, the ambiguity inherent to our own interpretation of data appears as a limiting factor to the quality of the prediction made by supervised machine learning algorithms.

### 7.2.2 Classification of the Near-Earth Regions

In the Chapter 4, we applied a gradient boosting classifier to the plasma moments and magnetic field measurements of THEMIS B to provide an automatic classification of the 3 main near-Earth regions: the solar wind, the magnetosheath and the magnetosphere.

Outperforming the methods based on manually-set threshold, the method was tested on the data of two other missions with equatorial orbit, Double Star and MMS, and classified the different regions with the same quality. Following this success, the same technique was used on the data provided by the non-equatorial Cluster mission and the measurements at lunar orbit provided by ARTEMIS. In both cases, we reached the same prediction quality after a retraining phase mandatory for the consideration of the physical characteristics of the different regions visited by the concerned spacecraft. In the specific case of ARTEMIS, this adaptation implied the addition of the position vector as an input feature and the consideration of a fourth region visited by the spacecraft: the lunar wake.

The developed region classifier was then used to elaborate the most exhaustive and publicly accessible complete magnetopause and bow shock crossing catalogs to our knowledge, easily and rapidly updatable with the increasing amount of data. At first, these catalogs can be used to perform statistical studies of the different properties of the near-Earth boundaries. This is for instance what we did with the analysis of the position and shape of the magnetopause in the chapter 5. They also constitute the starting point of the identification, and the underlying statistical study, of the different physical processes likely to happen in their vicinity. This is for instance what we did with the magnetopause plasma jets in the chapter 6. Additionally, gradient boosting classifiers are light-weighted algorithms that could thus easily be taken onboard of the future upcoming missions and automatically select the data of interest that should be stored and kept for further analysis, transforming the SITL process introduced with the launch of MMS into a *Machine In The Loop* process.

However, the classifier was trained with a magnetosheath defined as any region that was not neither the solar wind, neither the magnetosphere. If this simplification was reasonable at low-latitudes to exploit the spacecraft crossing locations for modeling the position and shape of the magnetopause, it leads, in the near-cusp region, to the detection of the cusp inner boundaries and not to the detection of the actual magnetopause current sheet. This seriously impacts the shape of the magnetopause if accounted as is in model fits as shown in the Chapter 5.

From now on, an interesting solution we can have to enhance the classifier would be to consider the remaining near-Earth regions as additional labels of the algorithm. This improvement

would also permit the statistical analysis of the properties of the different regions of the near-Earth environment with an increased level of detail. Among these regions, one could for instance cite the Ion Foreshock or the different magnetopause boundary layers but given the conclusions of Chapter 5, the cusp exterior appear as the priority region we should add to the classifier adapted to high-altitude data.

### 7.2.3 Detection of magnetopause plasma jets

In the chapter 6, we combined the region classifier to a peak detection and a second gradient boosting classifier to provide an automatic detection of magnetopause plasma jets at the dayside, low-latitude magnetopause. First trained on THEMIS C data, this pipeline was tested on the whole THEMIS mission, on Double Star and on MMS.

Once again, the jet detection pipeline we developed performed better than state of the art method based on manually-set thresholds. Concerning reconnection this shows that looking for the flows that respect the RD relation is not enough to fully describe the characteristics of magnetopause plasma flow. This jet detection pipeline was then used to elaborate the most exhaustive and accessible multi-mission catalog of magnetopause plasma jets paving in the process the way for the elaboration of statistical studies of the plasma flow induced by magnetic reconnection at the dayside magnetopause. This study could be completed by the application of the jet detection pipeline to the non-equatorial data provided by Cluster. Once this done, we could use all of the detected jets to construct maps of the plasma flow induced by reconnection at the magnetopause that could then be used to investigate the position of the reconnection sites for varying solar wind and seasonal conditions consistently with the problematics exposed in the chapter 6.

Naturally, the lists we elaborated are far from being completely exhaustive and will probably never be. This was proved through the statistical insight we gave on the FNs and the FPs of the pipeline that often appeared as having an ambiguous in-situ signature and would not have been considered as reconnection by every experts. In addition, we showed that the proportion of FPs in a list generated by our pipeline was similar to the proportion of differences we can find in two lists made by two different experts on the same dataset. Consequently, the quality of the detection provided by our pipeline is once again limited by the ambiguity that resides in interpretation of the data, different from an observer to another and the choices made during the labeling phase of the elaboration of the method heavily impact the nature of the events of the detected algorithms.

Finally, the specificity of the features we used to elaborate the peak classifier makes it not as adaptable to other structures as the ICME detector presented in the Chapter 3. Nevertheless, it could be worth finding an equivalent adapted to the detection of reconnection evidences elsewhere in the near-Earth environment or the interplanetary medium. One could especially think of magnetotail reconnection or solar wind reconnection.

## 7.3 Position and shape of the magnetopause

The rapid and reproducible obtainment of events catalogs from decades of in-situ data measurement is the entrance pass for the world of massive statistical studies of the different actors of the solar wind-magnetosphere with an important number of samples. Following this perspective, allowed by the application of supervised machine learning algorithms to the automatic detection of in-situ event signature, we used the multi-mission magnetopause crossings catalog that was elaborated with the region classifier combined to online accessible crossings to perform, in the Chapter 5, a statistical study of the position and shape of the magnetopause.

From the important number of crossings at our disposal and from the variety of both their associated upstream conditions and their location, this study improved the global vision we could have on the magnetopause.

On the one hand, we confirmed characteristics of the magnetopause that have been proven for long, the earthward pushing with an increasing dynamic pressure, the influence of the IMF  $B_z$  on the stand-off distance and the azimuthal asymmetry induced by the seasonal variations of the geomagnetic field.

On the other hand, we brought answer elements to the questions concerning the location and shape of the magnetopause that were still open:

1. We did not notice any dawn-dusk asymmetry once the aberration due to the Earth's revolution was removed.
2. We found no particular influence of the IMF radial component  $B_x$  on the stand-off distance.
3. Through the evidence of the influence of the IMF clock angle, we showed that a changing IMF  $B_y$  could induce changes on the magnetopause shape, which is consistent with the expected effect of this component on the displacement of the reconnection sites on the day-side.

In the three cases, the results we obtained differed with some existing studies while agreeing with others. Naturally, this shows that our study has not brought the final answer to these still-open questions that are the main point on which the upcoming study of the magnetopause location and shape with additional data shall focus. An interesting option we have to perform further investigation would be to consider the data of the future upcoming near-Earth missions, in particular the missions that crossed the magnetopause in the nightside, or spacecraft, such as Cluster 4, that were not considered in this thesis. The other option would be to redefine the way we detect the magnetopause in order to collect even the smallest partial crossings in a similar way than what is done in the previous existing studies. These two options would help increase the size of the crossing catalogs on which the statistical analysis will be performed and would thus allow an even more detailed investigation of the different parameters for which the influence on the magnetopause is still uncertain.

In all of the previous existing observational magnetopause models and statistical studies, the effect of magnetic reconnection is considered through the lone investigation of the influence of the IMF  $B_z$ . The third point we mentioned is then particularly interesting as it credits the influence of a changing IMF  $B_y$  on the displacement of the reconnection sites. On a large scale, reconnection affects the shape of the magnetopause by eroding the magnetosphere along a specific direction defined by the reconnection line. For a changing IMF, the reconnection sites are displaced on the boundary surface, the underlying erosion of the magnetosphere is also necessarily displaced and reflected in the modification of the shape of the magnetopause. In the chapter 6, we emitted the perspective of using the detected jets to construct magnetopause plasma flow maps and exploit them to predict the location of the X-line for various solar wind and seasonal conditions. We could then use these predictions to infer the influence of the different solar wind and seasonal parameters, one could especially think of the IMF cone angle, on the shape of the magnetopause and confirm this supposition through the statistical analysis of additional crossings.

The different results of this statistical study were condensed in the development of an empirical, analytical, asymmetric, static, non-indented model of the magnetopause shape and location. This model was even turned into a dynamical model that offered the possibility to predict the magnetopause shape and location at a given time rather than predicting it for a given set of solar wind and seasonal parameters. If the prediction of the magnetopause of our model is as accurate as the predictions of the other previous existing models on the dayside, the prediction error made by this model was found reduced in the nightside and particularly in the far nightside further than -30 Re. This is mostly due to the consideration of an additional parameter, the IMF clock angle, and to the presence in the dataset of magnetopause crossings at lunar distances provided by ARTEMIS. Nevertheless, the prediction error, due to both the lack of data and the ambiguity of magnetopause identification, even more important at lunar distances, resulted in a still important prediction error that indicates the necessity of further study of the characteristics of this boundary

at these distances from the Earth. Provided we have enough data, this would be an interesting improvement of the study we made and the model we developed.

Finally, we evidenced another effect of reconnection on the shape of the magnetopause when focused on the position of the crossings, detected and from older missions, in the near-cusp region. For every orientation of the IMF, the convection of the reconnected field lines generates a cusp external boundary that separates the cusp exterior from the magnetosheath. Following the way we labeled data in the chapter 4, the boundary we detected at high latitudes actually corresponds to the cusp inner boundary. This boundary is consistent with the concept of magnetopause in absence of reconnection and results in an apparent discontinuity between the dayside and the nightside. As a consequence, the magnetopause is necessarily indented in this region and this is what we noticed when we looked at the position of our near-cusp detected crossings. We also showed that the crossings from older missions, identified by other researchers and used in analytical models in the literature dominantly show inner boundary crossings too and result in overestimating the supposed magnetopause cusp indentation.

In the sense of reconnection, the external boundary appear as a more appropriate continuous extension of the magnetopause in the near-cusp region. Through the analysis of manually labeled crossings of this boundary, and despite of an increased radial distance, we noticed a depletion, in comparison to non-indented magnetopause models, in the near-cusp region that still suggests the existence of the indentation. To make it clear in our mind, we would have to perform a more detailed statistical analysis of the position of this boundary for various solar wind and seasonal conditions. With those indications, the comparison of the accuracy of a refitted, non-indented, model and an all new indented model would then give the final argument in favor or against the actual existence of the near-cusp indentation. This is definitely one of the priorities of a future work lead on the subject and the first step of this future work would be to adapt the region classifier to the consideration of additional near-Earth regions in order to collect as many cusp external boundaries as possible and make this vision statistically representative in terms of number of samples.

## 7.4 Potential of machine learning algorithms and larger perspectives

In the three cases of study we presented in this thesis, we applied supervised machine learning algorithms on the streaming in-situ data measurement provided by the spacecraft of various missions in order to provide a fast, automatic, reproducible detection method of different actors of the solar wind-magnetosphere coupling. Each method presents a design adapted to the spatio-temporal scales of the structures they are detecting or classifying and to the nature of their in-situ signature. In the three cases, the method we developed outperforms the quality of the prediction of the state of the art existing methods based on thresholds manually set on a reduced number of physical parameters. This is consistent with the complex physical nature of the different structures we focused on which multiple characteristics cannot be reduced to a couple of features.

At first sight, their prediction, superposed to the associated data, constitute an interesting visual indicator that can perfectly be used by an external observer to ease the observation and the selection of data. An interesting application of these algorithms would thus be their implementation on visualisation tool where they would act as an artificial assistant to any external observer that has an insight on a given set of in-situ data. This is especially the objective of the upcoming SciQLOP<sup>3</sup> tool currently under development at LPP for the CDPP<sup>4</sup>. With their ability to rapidly analyze and provide an accurate classification of data, one can also consider their onboard adaptation where they could be used either to provide an automatic selection and labeling of the data, either to select the specific mode under which they are best appropriated to be sent back on Earth. The confirmed efficiency of such algorithms is then the first step of the elaboration of a *Machine*

---

<sup>3</sup><https://github.com/SciQLop/SciQLop>

<sup>4</sup>Plasma Physics Data Center

*In The Loop* process dedicated to the primary analysis and selection of the raw in-situ data that will be provided by the upcoming missions.

These algorithms can also be used to rapidly elaborate some of the most exhaustive event catalogs. On the one hand, the increase of the size of the existing event catalogs will help improve the quality of the detections made by these algorithms as this was proved in the Chapter 3. On the other hand, the rapid and reproducible obtainment of these lists paves the way to reproducible statistical analysis of the various properties of the different actors of the solar wind-magnetosphere coupling with an important number of samples. This is the incredible opportunity to condense the decades all of the missions at once and is thus a tremendous milestone in our way to the global statistically representative vision we can have on the different elements that rule the dynamics of the near-Earth environment and its interaction with the solar wind.

At short term, we can consider the realisation of such type of studies through the exploitation of the different catalogs constructed in this thesis and this perspective was already mentioned in the case of ICMEs and magnetopause plasma jets. Concerning the magnetopause, it was actually done in the Chapter 5 where we performed a statistical analysis of the position and shape of the magnetopause that resulted in the production of an empirical and analytical model of this boundary. From now on, it would be interesting to perform the same work in the case of the bow shock crossing in order to have a complete statistically representative view of the dynamics of the two main boundaries at stake in the coupling.

On a longer term, the adaptation of our detection methods to additional solar events, near-Earth regions or small-scale physical processes, mentioned as an interesting perspective for the three cases, would obviously lead to the detailed study of these elements from a statistical point of view. Following what was previously said, we could think of the CIRs, the cusp exterior boundary or the Kelvin-Helmholtz instability. We could also think of elaborating an automatic detection method adapted to small-scale structures which observation is only allowed by very high resolution measurement even though this would probably imply the definition of another detection concept. This could for example be tested on the EDRs, which observation is permitted with the burst modes of MMS. Going even further into the potential of such methods, we could even think of their adaptation to the data of the other missions of the solar system.

On an even longer term, the accumulation of these different automatic detection methods paves the way to their combination into a single, general, data analyzer that would have the ability to read and classify any plasma and magnetic field measurement made in the solar system. This would constitute a huge achievement in the field of in-situ data analysis.

Nevertheless, we proved for each of the three cases that the quality of the detection was limited by our own interpretation of the data. In the case of ICMEs and magnetopause plasma jets, this was shown by comparing the list generated by our detection methods to different lists obtained by the manual selection of events by several external human observers. In the case of the boundaries, this was shown by showing how the choice we made on the definition of the magnetosheath impacted the representation we had on the location and shape of the magnetopause. The ambiguity, characteristics of any interpretation in-situ data, will then also be found in the prediction of these automatic detection methods and this can be explained by two reasons closely linked one to another. First, the label of the data was done manually<sup>5</sup> and this task is obviously ambiguous and hardly reproducible even when the same observer labels the same set of data twice. Second, we used supervised algorithms, the introduction of a label with a forced ambiguity in the training phase will then necessarily result in a prediction that reflects this ambiguity.

<sup>5</sup>Either by ourselves or by external observers through the utilisation of already existing events catalogs.

From then on, looking for the increase of the recall, precision or AUC appears as the illusory purpose of a future work given the consequent hindsight one must have when interpreting the values of these metrics. It would be, for instance, much more interesting to have a direct insight on the how these algorithms have learnt to perform the task they were dedicated to. Indeed, interpreting the way these algorithms have learnt their knowledge from the data would bring significant information on the physical nature of the events they are trying to detect, surely, but more interestingly, they could also bring us clues on how we interpret the data and how we can reduce the ambiguity that underlies from an observation to another. However, the interpretability of machine learning, especially when it comes to neural networks, is still a hot topic in the field of artificial intelligence and comes far from the scope of this thesis. Another alternative approach we can have concerning the introduction of machine learning in the analysis of in-situ data would be the utilisation of unsupervised techniques where we could benefit from the clustering of data according a particular trend in the data to evidence new, unexpected properties of the main actors of the solar wind-magnetosphere coupling.

Despite of this so-called *ambiguous limitation*, the range of the possible application of supervised machine learning algorithms is extremely wide and the full potential of such technique in the field of in-situ data analysis is consequently only starting. There is however no doubt it will be in the upcoming years given the encouraging promises glimpsed with its first applications.



# Appendix A

## Coordinates systems

In this section, we will briefly detail the different coordinate systems in use for this thesis.

### A.1 Geocentric Solar Ecliptic (GSE)

In this coordinate system, we define the X axis as the Sun-Earth axis oriented sunward. The Z axis is perpendicular to the ecliptic plane and the Y axis completes the orthogonal set.

### A.2 Geocentric Solar Magnetospheric (GSM)

In [GSM](#) the X axis is also defined as the Sun-Earth axis oriented sunward. The Y axis is perpendicular to the Earth Dipole, such that the Z axis is in the plane defined by the X axis and the Earth Dipole and then completes the orthogonal set..

In this thesis, we will mostly use these [GSM](#) cartesian coordinates, X, Y and Z. We will also punctually consider their spherical equivalent  $r, \theta$  and  $\phi$  where  $r = \sqrt{X^2 + Y^2 + Z^2}$  is the radial distance from the center of the Earth,  $\theta$  is the zenith angle between the direction of  $r$  and the X axis and the azimuth  $\phi$  is the angle between the projection of  $r$  in the Y – Z plane and the positive direction of the Z axis:

$$X = r \cos(\theta) \tag{A.1}$$

$$Y = r \sin(\theta) \sin(\phi) \tag{A.2}$$

$$Z = r \sin(\theta) \cos(\phi) \tag{A.3}$$

The spherical system we adopted in [GSM](#) coordinates is represented in the [Figure A.1](#). It is worth noting here that the convention adopted to define the azimuth angle  $\phi$  is the same than the one adopted to define the [IMF](#) clock angle  $\Omega$  in the [Figure 1.2](#).

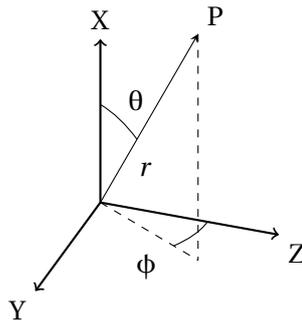


Figure A.1: Representation of the spherical coordinates used in [GSM](#) coordinates in this thesis.

### A.3 Local Magnetopause Normal (LMN)

#### A.3.1 Principle

LMN coordinates is a system adapted to the local study of the magnetopause and thus particularly useful when focusing on the small scale physical processes occurring in the vicinity of this boundary. The N direction is the normal outward vector to the magnetopause, the M direction perpendicular to the plan that contain the N direction and the GSM Z axis and the L direction completes the orthogonal set.

In practice, the LMN coordinates of a given point of space can be determined by two manners.

At first, the N direction can be estimated with the use of an analytical magnetopause such as the one we developed in the Chapter 5.

The other method, the so-called **Minimum Variance Analysis (MVA)** that was used in the Chapter 6, stands in finding the direction for which the magnetic field, measured during the boundary crossing, has the less variations. The three directions , L, M and N, are thus found by finding the eigenvalues of the so-called magnetic variance matrix defined as:

$$M_i^j = \langle B_i B_j \rangle - \langle B_i \rangle \langle B_j \rangle \quad (\text{A.4})$$

where  $i, j = 1, 2, 3$  are the three cartesian GSM components of the magnetic field.

### A.4 Magnetic Local Time (MLT)

At a given point in space, MLT is the hour angle formed by the meridional plane that contain the subsolar point with the meridional plane in which this point is.

By convention, the subsolar point is found at the so-called magnetic noon and is thus found at a MLT of 12 Hr.

The representation of the MLT in comparison with the GSM coordinates is shown in the Figure A.2

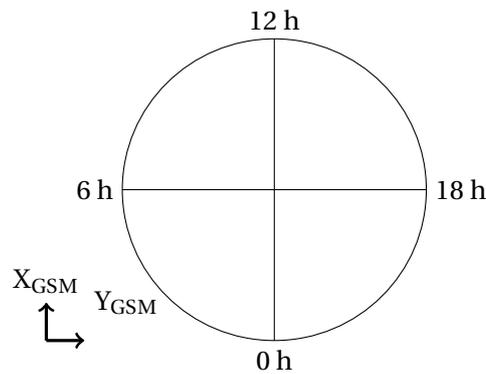


Figure A.2: Schematic definition of the MLT.

## **Appendix B**

# **Additional prediction examples**

## B.1 ICMEs

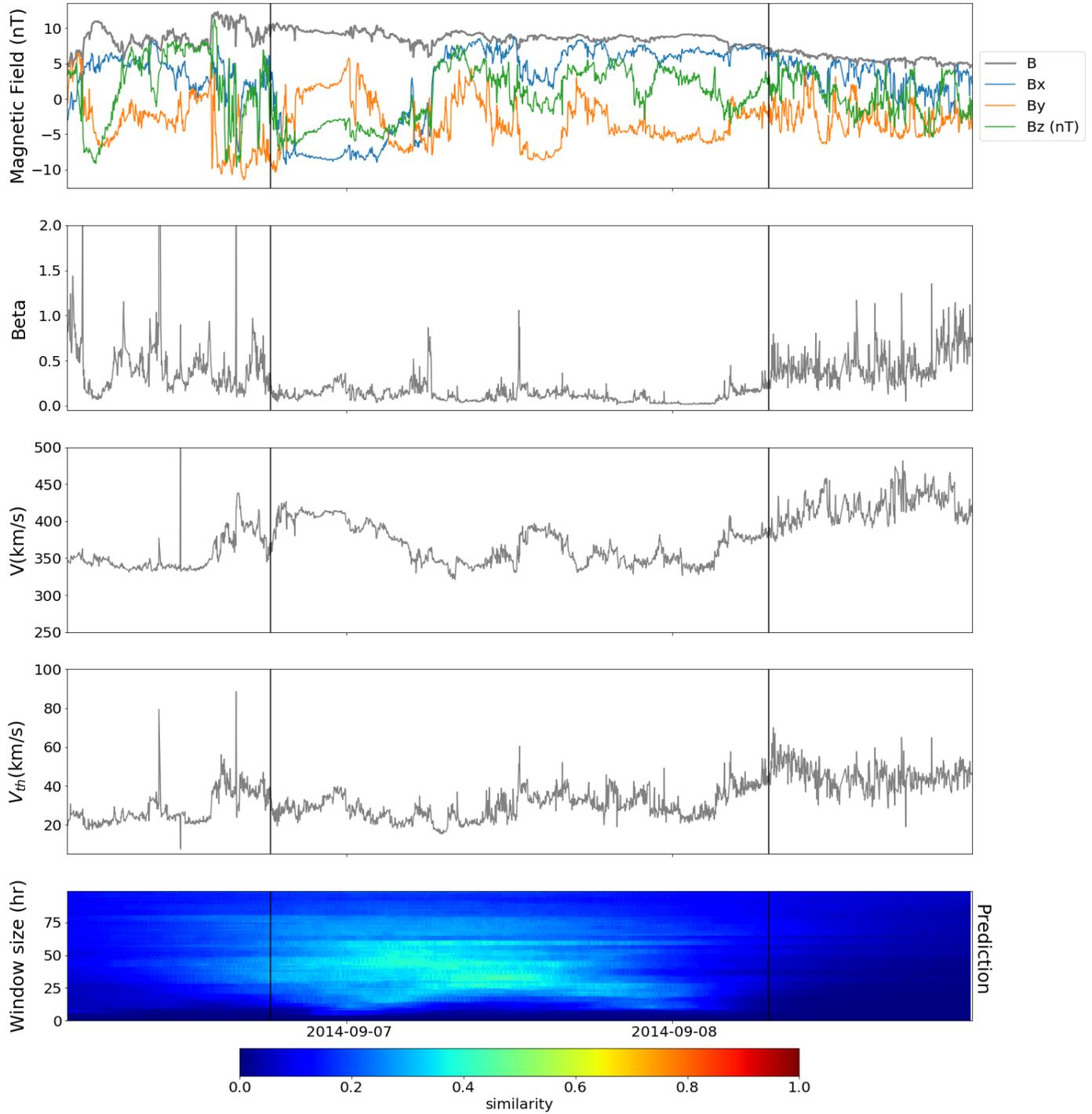


Figure B.1: Solar wind observation during an ICME from the WIND spacecraft. The legend is the same than in the Figure 3.2

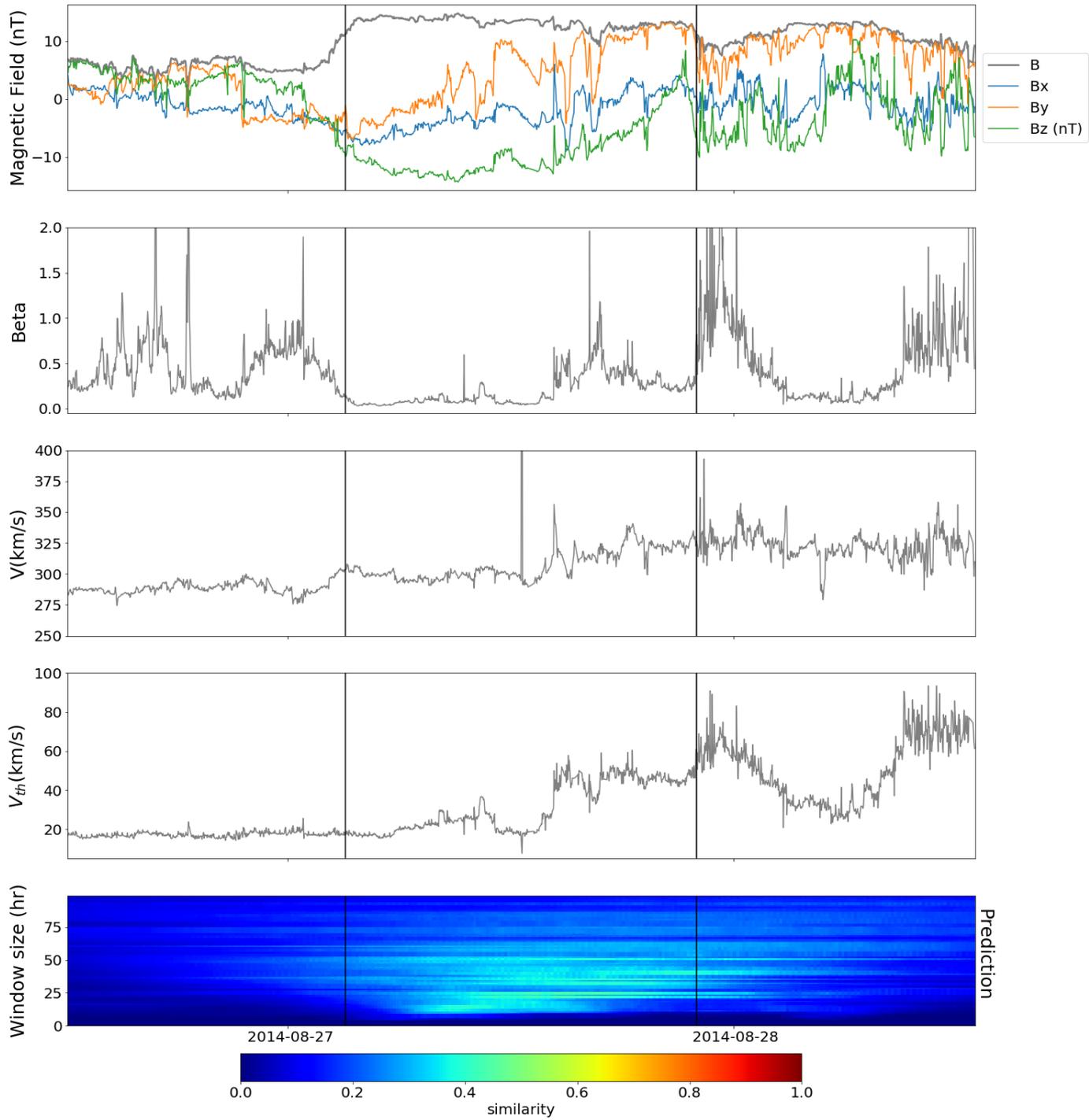


Figure B.2: Solar wind observation during an ICME from the WIND spacecraft. The legend is the same than in the Figure 3.2

## B.2 Near-Earth regions

### B.2.1 THEMIS

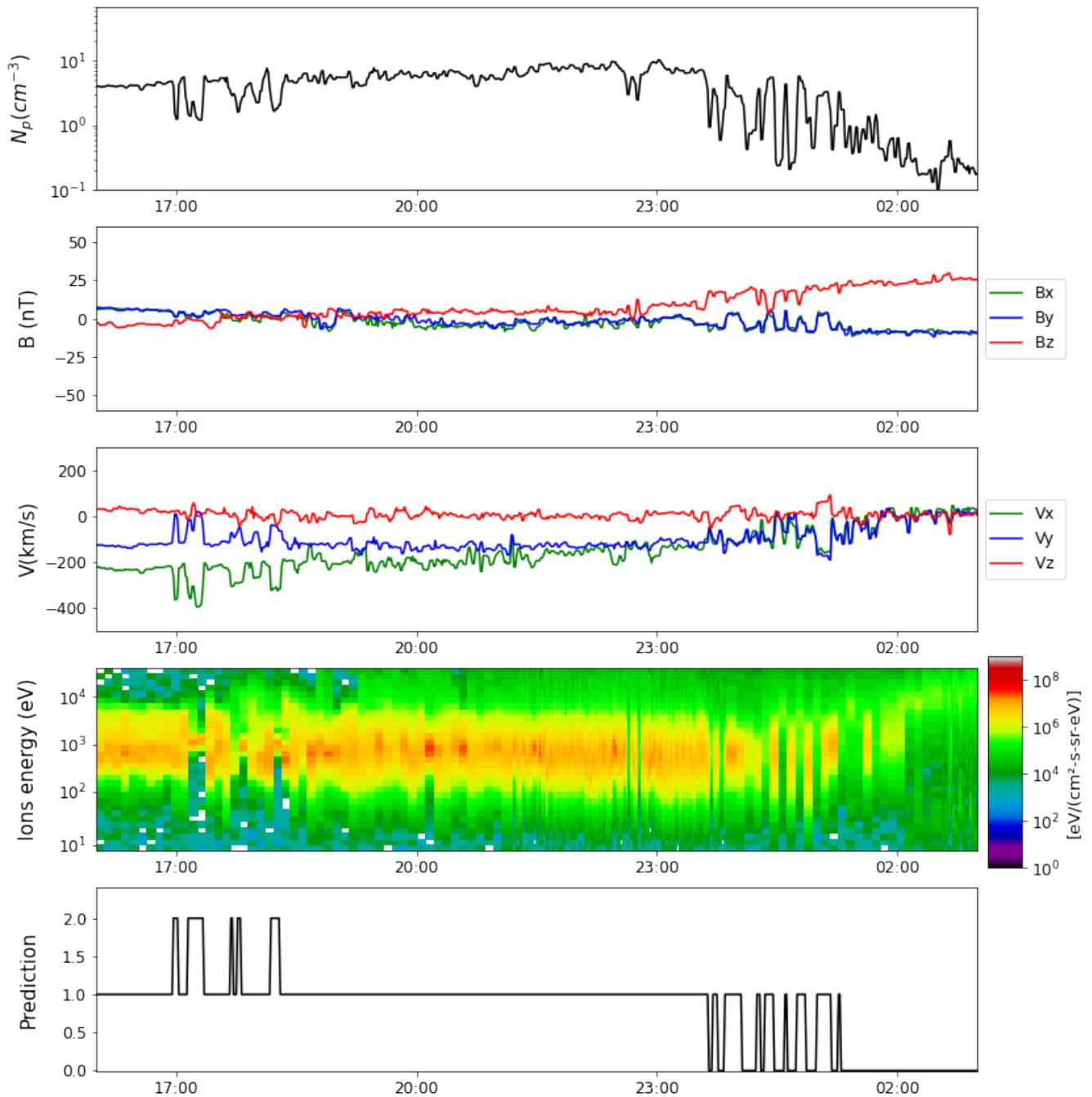


Figure B.3: In-situ measurement provided by THEMIS B spacecraft on the 10<sup>th</sup> of November 2008. The legend is the same than in 4.1.

**B.2.2 Double Star**

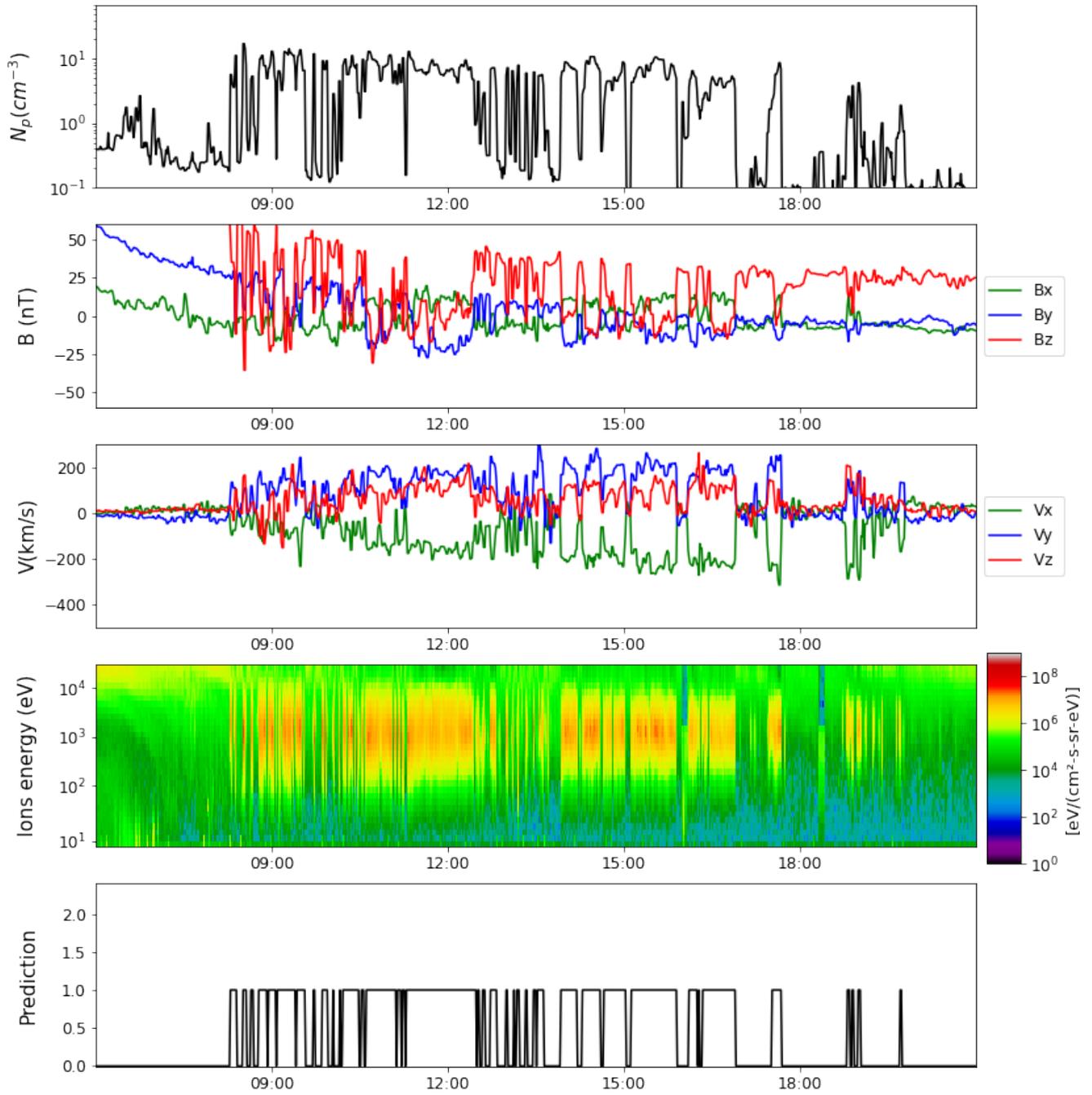


Figure B.4: In-situ measurement provided by Double Star TC 1 spacecraft on the 15<sup>th</sup> of January 2005. The legend is the same than in 4.1.

## B.2.3 MMS

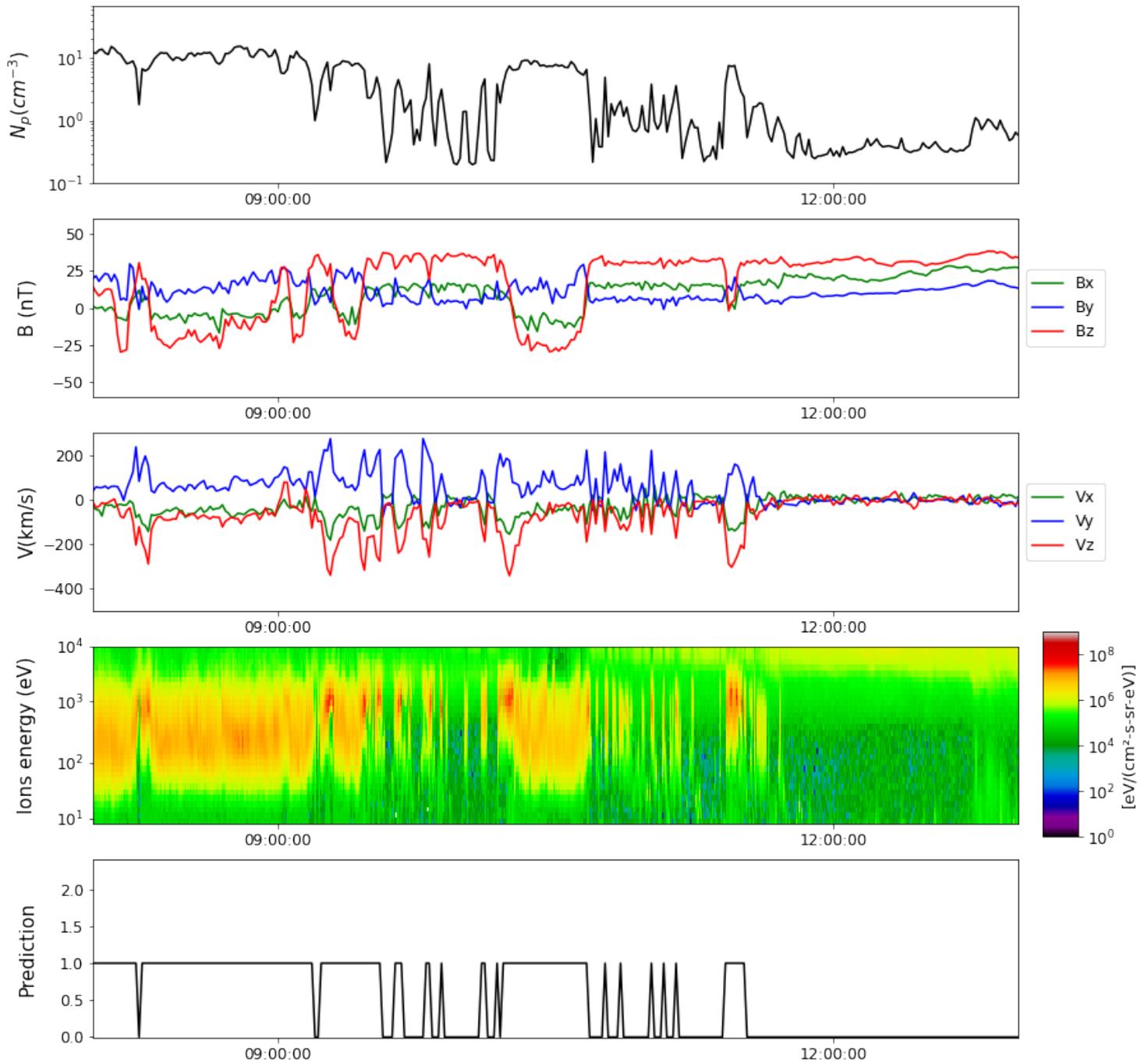


Figure B.5: In-situ measurement provided by MMS 1 spacecraft on the 2<sup>nd</sup> of December 2015. The legend is the same than in 4.1.

**B.2.4 Cluster**

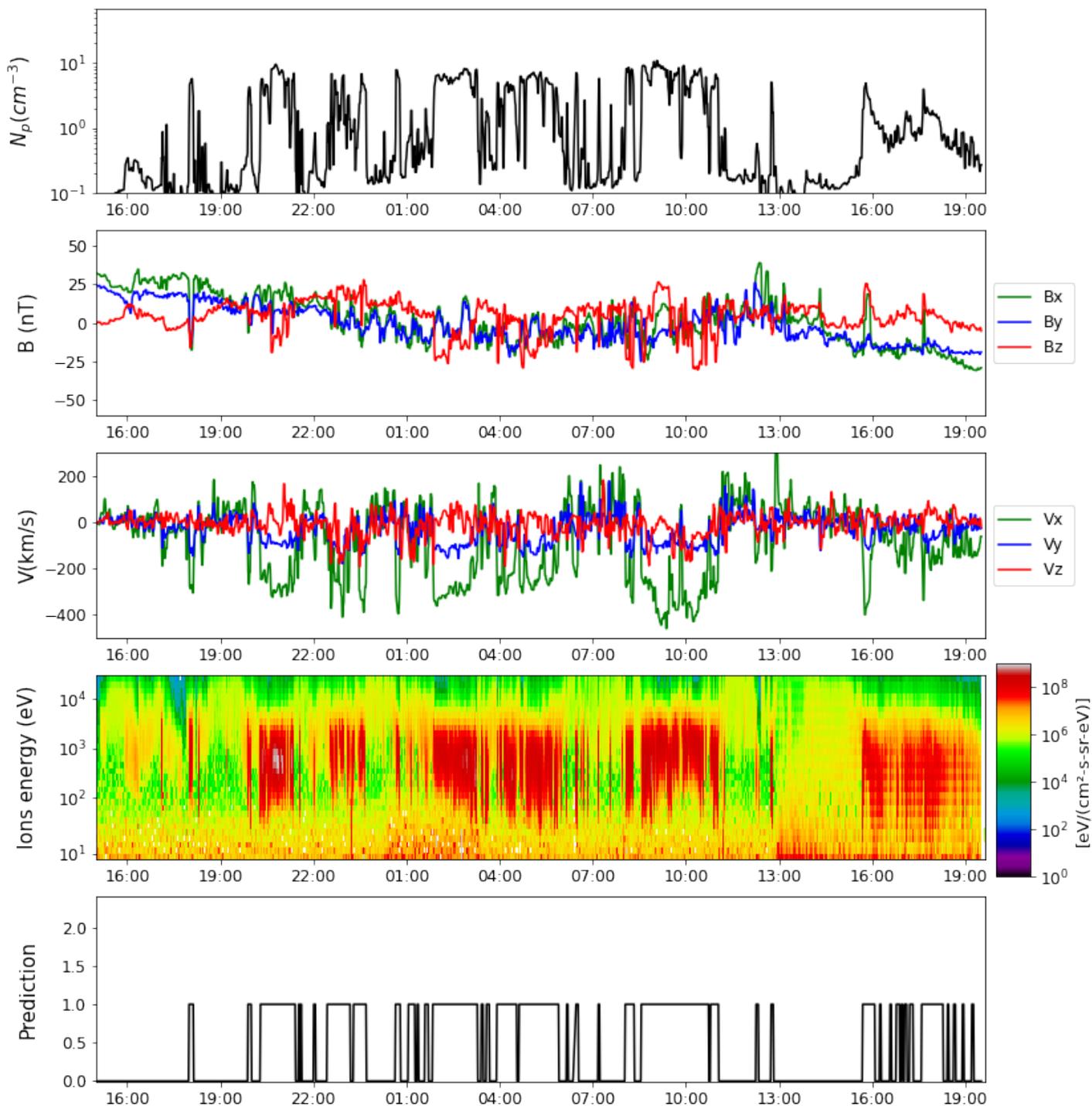


Figure B.6: In-situ measurement provided by Cluster 3 spacecraft on the 23<sup>rd</sup> of June 2003. The legend is the same than in 4.1.

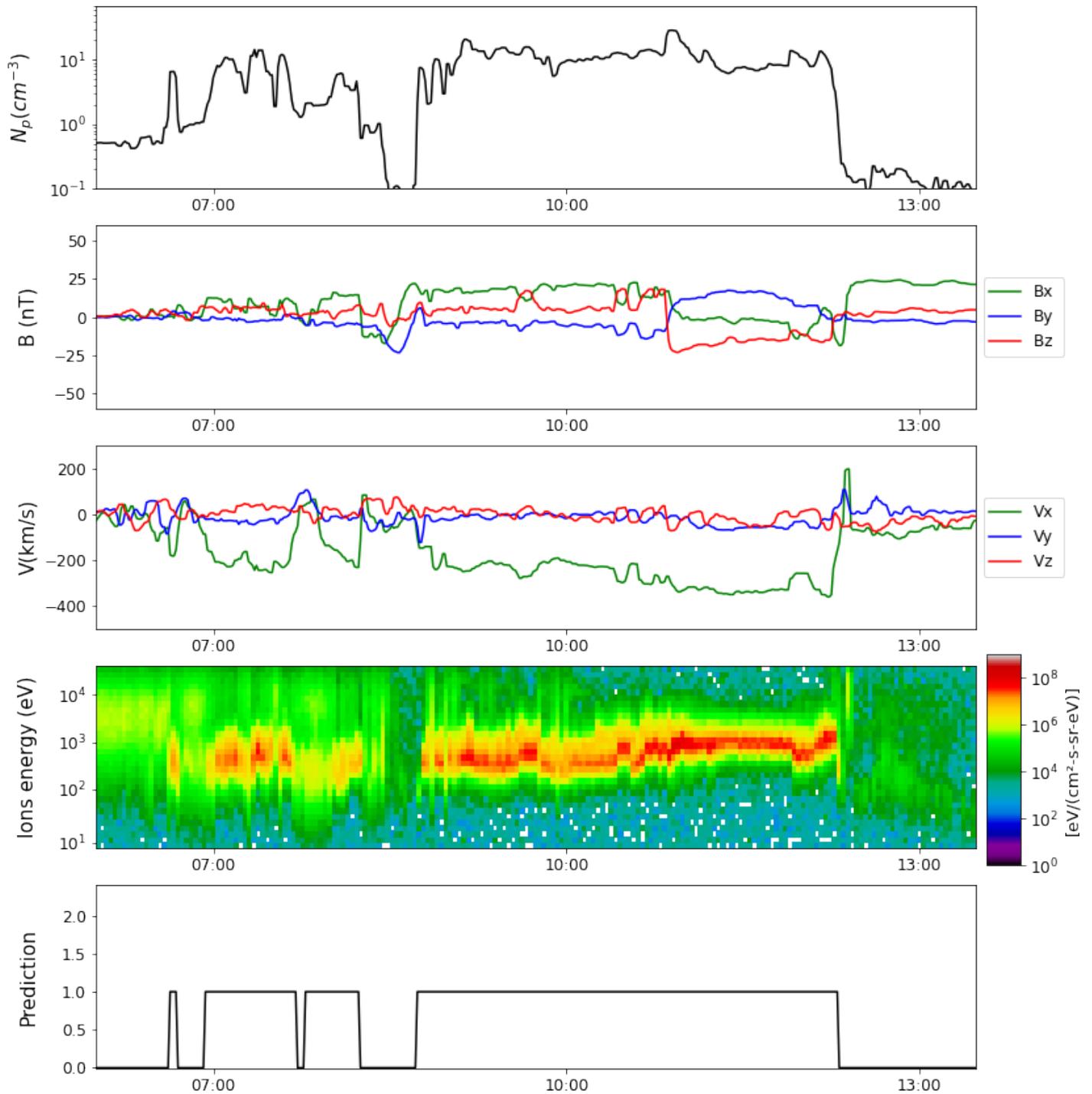
**B.2.5 ARTEMIS**

Figure B.7: In-situ measurement provided by ARTEMIS B spacecraft on the 24<sup>th</sup> of April 2013. The legend is the same than in 4.1.

### B.3 Cusp external boundary

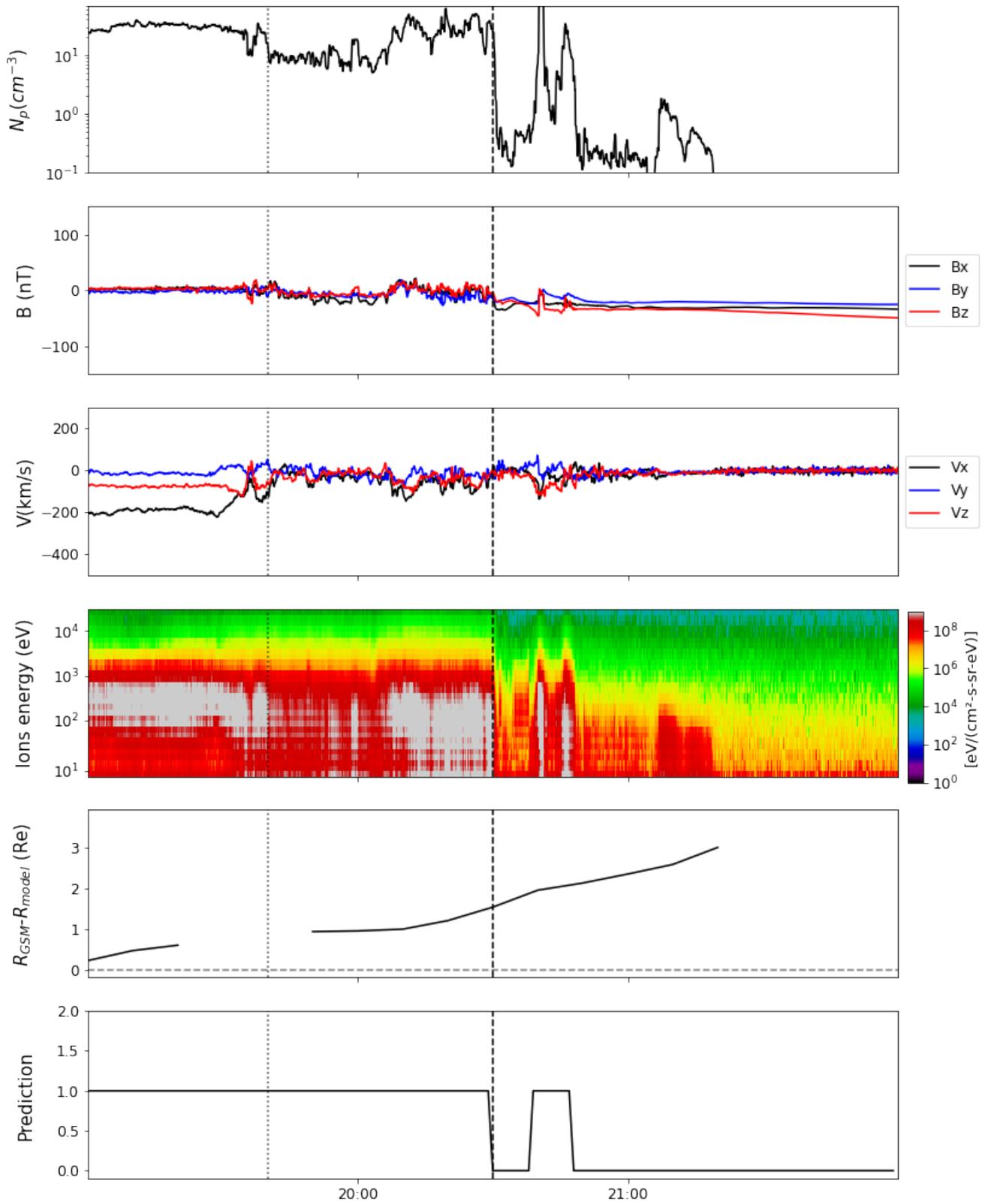


Figure B.8: In-situ measurement provided by Cluster 1 spacecraft on the 21<sup>st</sup> of April 2007. The legend is the same than in 5.22

## B.4 Reconnection jets

### B.4.1 THEMIS

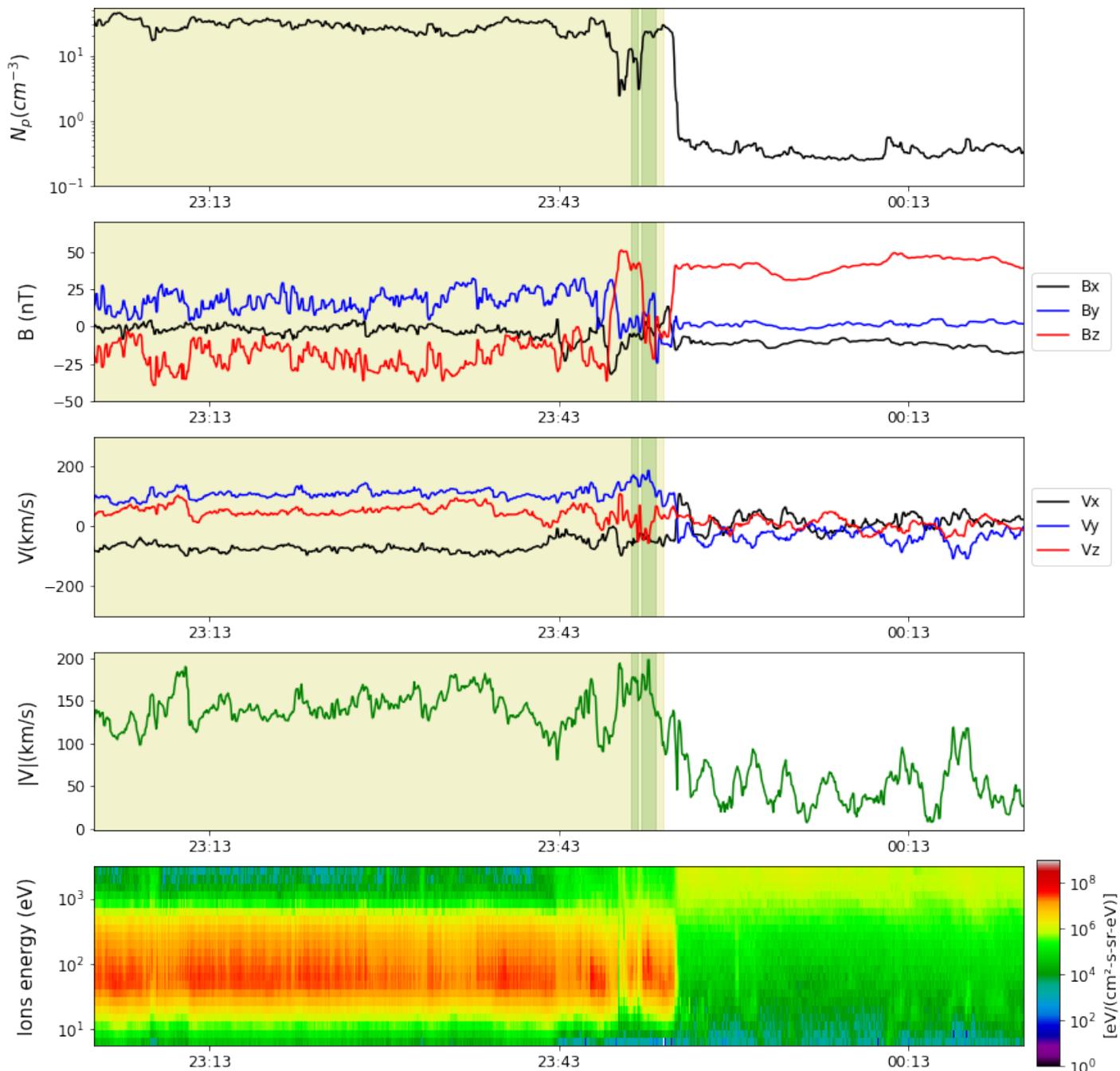


Figure B.9: In-situ measurement provided by THEMIS E spacecraft on the 1<sup>st</sup> of November 2011. The legend is the same than in 6.2 except for the green intervals that here represent the velocity peaks classified as reconnection jets by the peak classifier.

**B.4.2 Double Star**

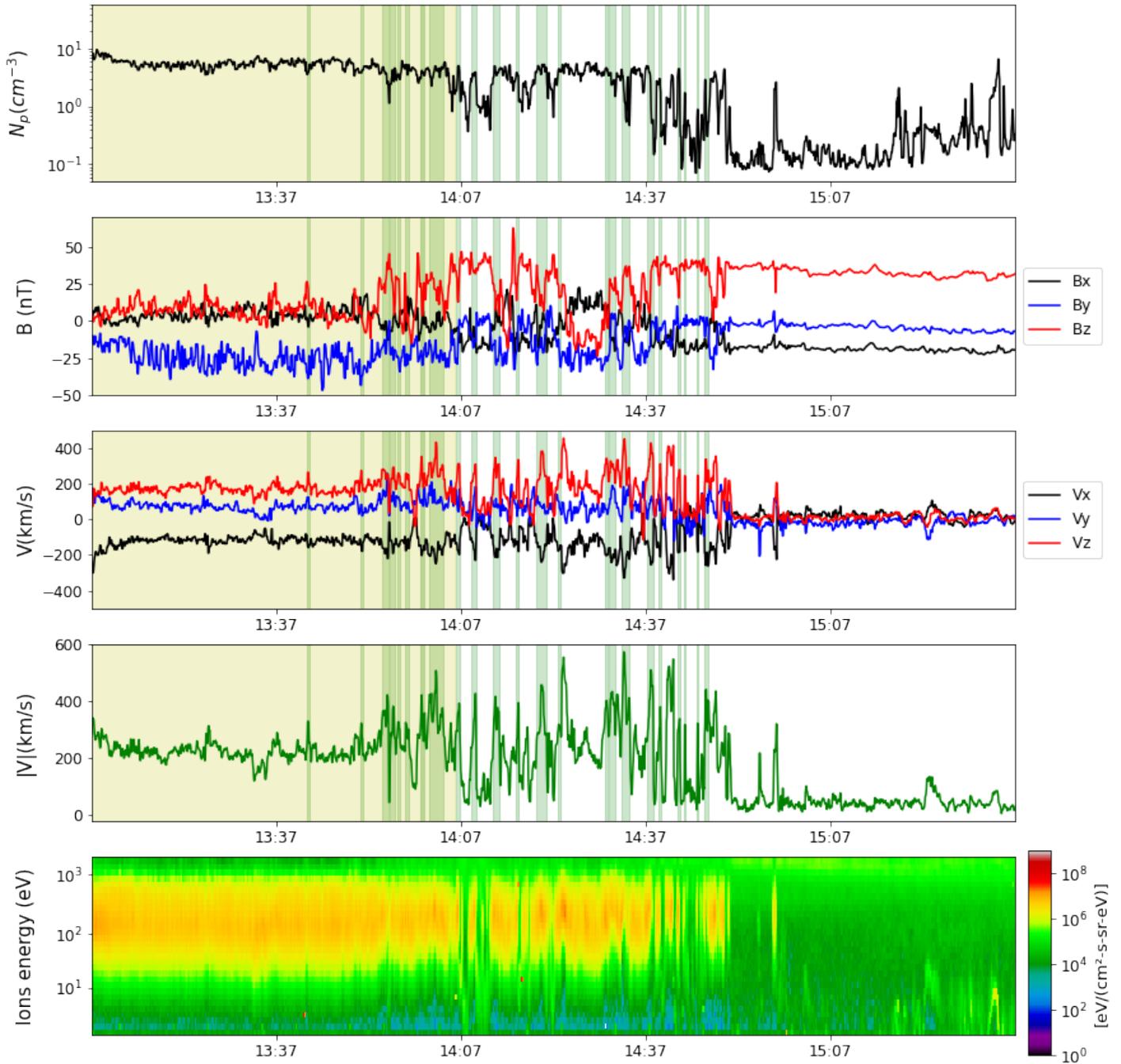


Figure B.10: In-situ measurement provided by the Double Star TC1 spacecraft on the 10<sup>th</sup> of March 2005. The legend is the same than in 6.2

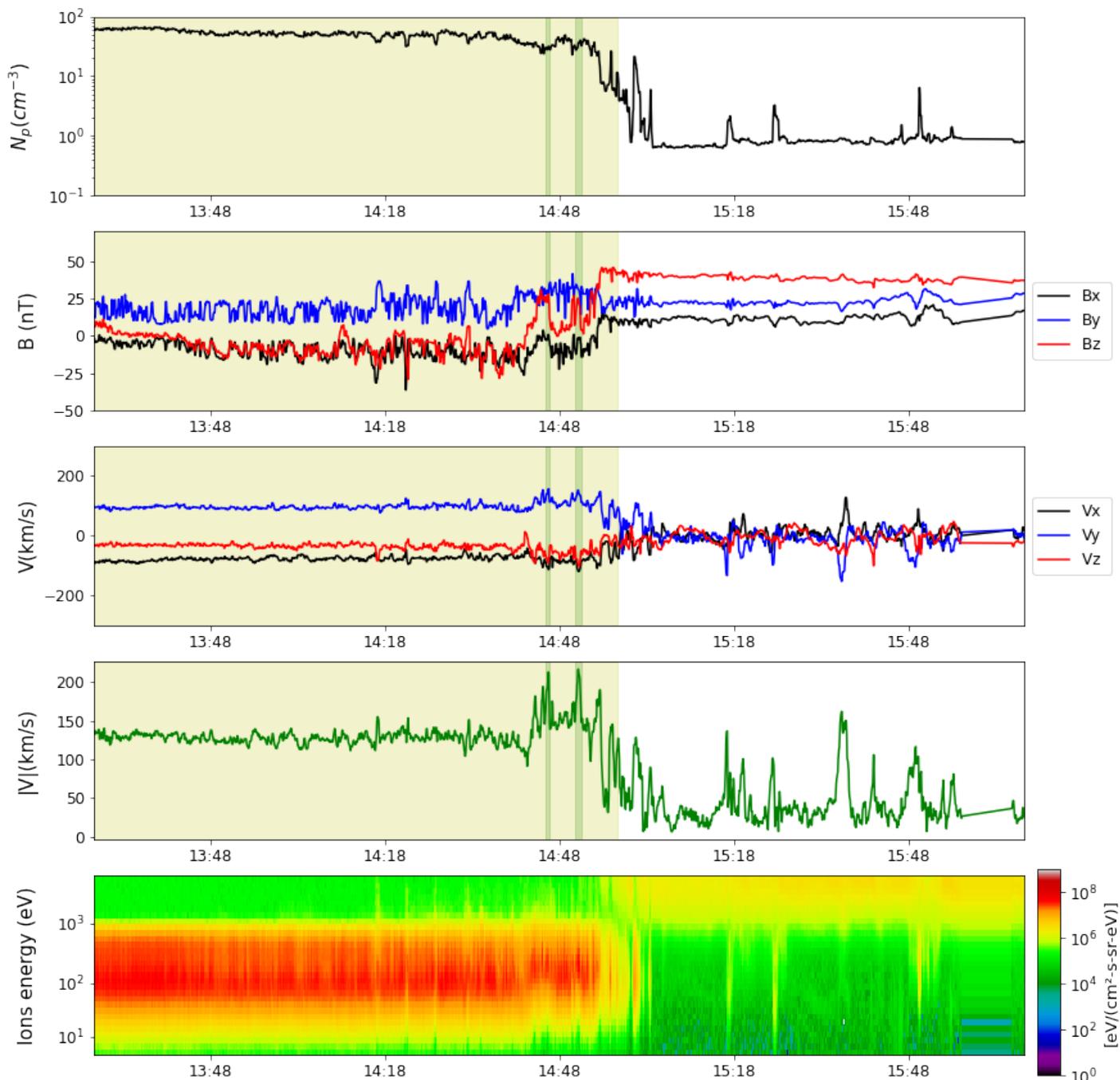
**B.4.3 MMS**

Figure B.11: In-situ measurement provided by MMS 1 spacecraft on the 29<sup>th</sup> of October 2015. The legend is the same than in B.9.

# Appendix C

## List of Acronyms

- ANN** Artificial Neural Network. [27](#), [29](#), [34](#), [102](#)
- ARTEMIS** Acceleration Reconnection Turbulence & Electrodynamics of Moon's Interaction with the Sun. [vi](#), [vii](#), [71](#), [73](#), [79](#), [86–91](#), [94](#), [98](#), [103](#), [107](#), [113](#), [119](#), [172](#), [174](#), [X](#)
- AU** Astronomical Unit. [2](#), [8](#), [19](#)
- AUC** Area Under Curve. [38](#), [78–80](#), [82](#), [84](#), [88](#), [89](#), [91](#), [177](#)
- CART** Classification and Regression Trees. [32](#)
- CIR** Corotating Interaction Region. [9](#), [172](#), [176](#)
- CME** Coronal Mass Ejection. [6–9](#), [15](#), [39](#), [44](#)
- CNN** Convolutional Neural Network. [vi](#), [vii](#), [19](#), [35](#), [43](#), [50–53](#), [56](#), [63](#), [66](#), [68](#), [70](#), [95](#), [134](#), [145](#), [150](#), [171](#)
- EDR** Electron Diffusion Region. [15](#), [16](#), [18](#), [74](#), [176](#)
- FN** False Negative. [37](#), [38](#), [52](#), [54](#), [57](#), [60](#), [66–68](#), [141](#), [146](#), [150–153](#), [158–160](#), [162](#), [167](#), [171](#), [173](#)
- FP** False Positive. [vi](#), [vii](#), [37](#), [52](#), [54](#), [56–60](#), [66–68](#), [141](#), [146](#), [149–151](#), [153](#), [157–162](#), [167](#), [171](#), [173](#)
- FPR** False Positive Rate. [37](#), [38](#), [89](#)
- FTE** Flux Transfer Event. [39](#)
- GSE** Geocentric Solar Ecliptic. [x](#), [I](#)
- GSM** Geocentric Solar Magnetospheric. [x](#), [2](#), [73](#), [75](#), [78](#), [86](#), [93](#), [94](#), [107](#), [109](#), [112](#), [116](#), [117](#), [130](#), [131](#), [143](#), [146–148](#), [153](#), [154](#), [161](#), [I](#), [II](#)
- HSS** Heidke Skill Score. [38](#), [78](#), [80](#), [82](#), [84](#), [88](#)
- ICME** Interplanetary Coronal Mass Ejection. [vi](#), [vii](#), [x](#), [7–9](#), [19](#), [43–50](#), [52–58](#), [60–63](#), [65–68](#), [70](#), [72](#), [112](#), [134](#), [145](#), [162](#), [169–173](#), [176](#), [IV](#), [V](#)
- IDR** Ion Diffusion Region. [15](#), [16](#)
- IMF** Interplanetary Magnetic Field. [vi](#), [vii](#), [2–5](#), [8](#), [10–12](#), [16–19](#), [72](#), [89](#), [102–106](#), [110–112](#), [116–121](#), [123–125](#), [133](#), [134](#), [139](#), [142](#), [143](#), [174](#), [175](#), [I](#)
- KDE** Kernel Density Estimation. [152](#), [153](#), [158](#)

**LMN** Local Magnetopause Normal. [x](#), [148](#), [II](#)

**MC** Magnetic Cloud. [44](#), [45](#)

**MHD** Magnetohydrodynamics. [1](#), [5](#), [10](#), [102](#), [105](#), [112](#), [113](#), [116–118](#), [142](#)

**MLT** Magnetic Local Time. [x](#), [77](#), [145](#), [147](#), [161](#), [163](#), [II](#)

**MMS** Magnetospheric Multiscale. [vi](#), [vii](#), [6](#), [18](#), [71](#), [73](#), [74](#), [79](#), [82–84](#), [86](#), [90](#), [91](#), [94](#), [95](#), [98](#), [108](#), [143](#), [145](#), [153](#), [154](#), [156–158](#), [161](#), [167](#), [172](#), [173](#), [176](#), [VIII](#), [XIV](#)

**MSE** Mean Square Error. [28](#), [31](#)

**MVA** Minimum Variance Analysis. [148](#), [II](#)

**PDL** Plasma Depletion Layer. [94](#)

**RD** Rotational discontinuity. [10](#), [11](#), [144](#), [173](#)

**ReLU** Rectified Linear Unit. [35](#), [51](#)

**RL** Reference List. [48](#), [54](#), [56](#), [57](#), [66](#), [67](#), [172](#)

**RMSE** Root Mean Square Error. [51](#), [119](#), [120](#)

**ROC** Receiving Operator Curve. [37](#), [38](#), [79](#), [91](#), [92](#)

**SEM** Standard Error of the Mean. [120](#)

**SITL** Scientists In The Loop. [74](#), [95](#), [172](#)

**SOHO** Solar and Heliospheric Observatory. [8](#)

**TD** Tangential discontinuity. [10](#), [11](#), [125](#)

**THEMIS** Time History of Events and Macroscale Interactions during Substorms. [vi](#), [vii](#), [ix](#), [6](#), [12](#), [14](#), [17](#), [71–73](#), [75–80](#), [82](#), [84](#), [89–92](#), [94](#), [98](#), [102](#), [108](#), [141](#), [143](#), [145–148](#), [153](#), [154](#), [157](#), [158](#), [161](#), [167](#), [172](#), [173](#), [VI](#), [XII](#)

**TN** True Negative. [37](#), [38](#)

**TP** True Positive. [37](#), [146](#), [150–153](#), [158–160](#), [162](#)

**TPR** True Positive Rate. [37](#), [89](#)

**Titre:** Etude du couplage magnétosphère/vent solaire par des méthodes de machine learning

**Mots clés:** Magnetosphere, vent solaire, magnetopause, machine learning

**Résumé:** Les décennies d'accumulation de données provenant de missions explorant le vent solaire ainsi que l'environnement terrestre proche permettent l'étude de la relation Soleil-Terre de manière statistique. Ces études sont toutefois limitées par la sélection manuelle des événements d'intérêt dans les données qui reste une tâche fastidieuse, subjective et difficilement reproductible.

En nous appuyant sur des outils d'apprentissage statistique, nous mettons au point des méthodes de détection automatique d'événements à partir de mesures de données in-situ. Qu'il s'agisse de détecter les éjections de masse coronale interplanétaires, de classifier les régions de l'environnement terrestre proche ou de détecter les jets de plasma issus de la reconnexion magnétique à la magnétopause, nos méthodes font moins d'erreurs que celles basées sur des seuils empiriques généralement utilisées pour sélectionner des événements. Elles sont de plus adaptables d'une mission à une autre pourvu que la nature des régions traversées par les sondes soient similaires. Nous montrons toutefois que l'interprétation des données par ces méthodes sont limitées par notre

propre interprétation physique des données et des événements qu'elles mesurent.

Ces méthodes ouvrent la porte aux études statistiques d'événements mesurés in-situ à grand nombre d'échantillons. La classification des différentes régions de l'environnement terrestre proche nous permet par exemple d'étudier statistiquement la position et la forme adoptée par la magnétopause en s'appuyant sur les données de missions d'orbites équatoriales (THEMIS, MMS, Double Star), polaires (Cluster) et lunaires (ARTEMIS). En plus de confirmer l'influence saisonale, le rôle joué par la pression dynamique, et l'asymétrie azimutale, nous montrons que l'orientation azimuthale du champ magnétique interplanétaire modifie la forme de la magnétopause par le biais de la reconnexion magnétique et discutons de la nature de cette frontière au niveau des cornets polaires. L'étude résulte en la production d'un modèle analytique de la position de la magnétopause offrant une description plus précise de cette frontière du côté nuit de la magnétosphère terrestre.

**Title:** Solar wind/magnetosphere coupling inferred from machine-learning methods

**Keywords:** Magnetosphere, solar wind, magnetopause, machine learning

**Abstract:** Decades of in-situ data measurement by missions focused on the study of the solar wind and its relation with the near-Earth environment allowed the study of the Sun-Earth coupling from a statistical point of view. Nevertheless, these studies are limited by the manual selection of the events of interest in the data that is still a subjective, fastidious and hardly reproducible task.

Using machine learning algorithms, we elaborate automatic detection methods of events from in-situ data measurement. Whether they are applied to the detection of interplanetary coronal mass ejections, to the classification of the near-Earth regions or to the identification of magnetopause magnetic reconnection jets, the developed methods are more accurate than those based on manual, empirical thresholds. They are also adaptable from a mission to another provided the regions visited by the spacecraft are of the same nature. We show that the interpretation of the data by these methods is limited by the vision

we have on the data and the events they measure.

These methods pave the way for statistical studies of in-situ measured events with an important number of samples. Thereby, we use the classification of the different regions of the near-Earth environment to statistically study the position and shape of the magnetopause using the data of missions with equatorial (THEMIS, MMS, Double Star), polar (Cluster) and lunar (ARTEMIS) orbits. In addition to confirming the seasonal dependence, the azimuthal asymmetry and the influence of the solar wind dynamic pressure, we show that the clock angle of the interplanetary magnetic field modifies the shape of the magnetopause through the process of magnetic reconnection and lead a discussion on the nature of this boundary around the polar cusps. We combine the results of the study into an analytical model of the magnetopause position and shape that offers a more precise description of this boundary on the night side of the Earth magnetosphere.

