



HAL
open science

Intelligence artificielle et prévision de l'impact de l'activité solaire sur l'environnement magnétique terrestre

Marina Gruet

► **To cite this version:**

Marina Gruet. Intelligence artificielle et prévision de l'impact de l'activité solaire sur l'environnement magnétique terrestre. Physique de l'espace [physics.space-ph]. UNIVERSITE DE TOULOUSE, 2018. Français. NNT: . tel-01987697

HAL Id: tel-01987697

<https://hal.science/tel-01987697>

Submitted on 21 Jan 2019

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.



ONERA

THE FRENCH AEROSPACE LAB

THÈSE

**En vue de l'obtention du
DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE**
Délivré par l'Institut Supérieur de l'Aéronautique et de l'Espace

Présentée et soutenue par

Marina GRUET

Le 28 septembre 2018

**Intelligence artificielle et prévision de l'impact de
l'activité solaire sur l'environnement magnétique
terrestre**

Ecole doctorale : **SDU2E - Sciences de l'Univers, de l'Environnement et de
l'Espace**

Spécialité : **Astrophysique, Sciences de l'Espace, Planétologie**

Unité de recherche :
ISAE-ONERA PSI Physique Spatiale et Instrumentation

Thèse dirigée par
Angelica SICARD et Sandrine ROCHEL

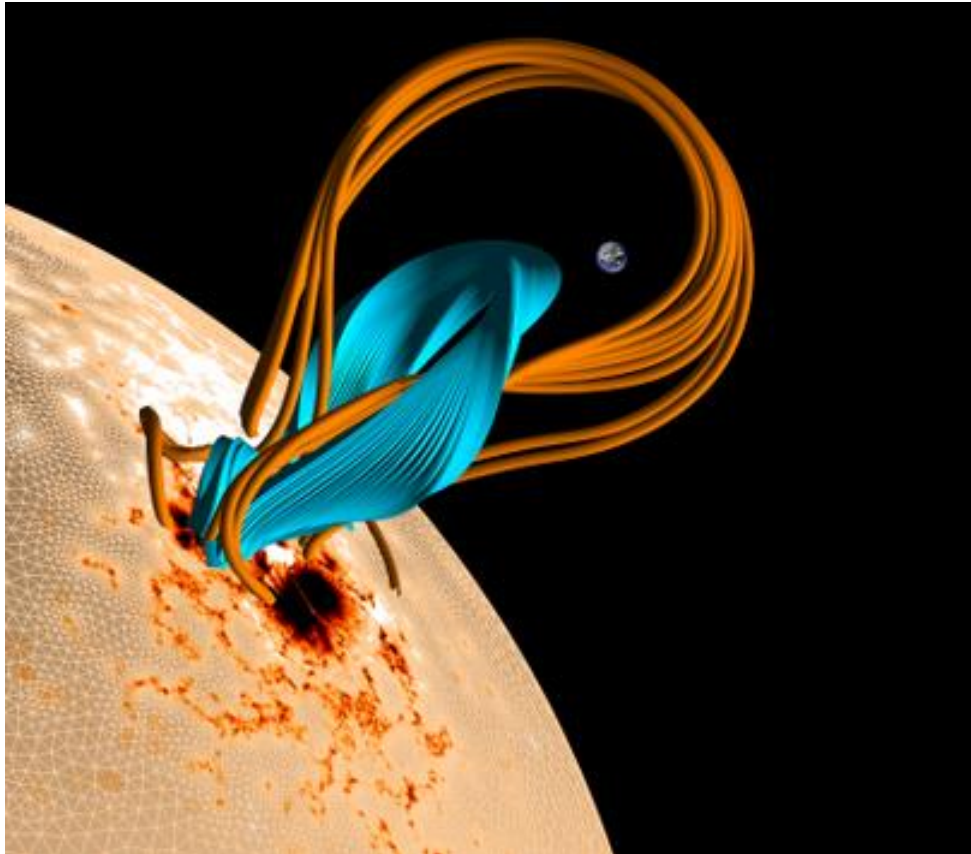
Jury

M. Thierry DUDOK DE WIT, Rapporteur
Mme Carine BRIAND, Rapporteur
Mme Aude CHAMBODUT, Examineur
Mme Angelica SICARD, Directeur de thèse
Mme Sandrine ROCHEL, Co-directeur de thèse
M. Jean LILENSTEN, Président

MEMBRE INVITÉE ET RÉFÉRENTE SUR L'ASPECT MATHÉMATIQUES DE CETTE

THÈSE :

NATHALIE BARTOLI, ONERA - DTIS



Modélisation de la cage magnétique solaire contenant une éruption. La Terre est représentée pour échelle [Amari et al., 2018].

**« AUCUN PESSIMISTE N’A JAMAIS DÉCOUVERT LE SECRET DES ÉTOILES,
NAVIGUÉ JUSQU’À DES TERRES INCONNUES, OU OUVERT UN NOUVEAU
CHEMIN POUR L’ESPRIT HUMAIN ».**

HELLEN KELLER.

REMERCIEMENTS

« A la naissance, on monte dans un train [...]. Au fur et à mesure que le temps passe, d'autres personnes montent dans le train. Et elles seront importantes : notre fratrie, nos amis, nos enfants, même l'amour de notre vie. Beaucoup démissionneront et laisseront un vide plus ou moins grand. D'autres seront si discrets qu'on ne réalisera pas qu'ils ont quitté leurs sièges. Ce voyage en train sera plein de joies, de peines, d'attentes, de bonjours, d'au-revoir et d'adieux. Le succès est d'avoir de bonnes relations avec tous les passagers pourvu qu'on donne le meilleur de nous-mêmes. On ne sait pas à quelle station nous descendrons, donc vivons heureux, aimons et pardonnons. Il est important de le faire car lorsque nous descendrons du train, nous ne devons laisser que de beaux souvenirs à ceux qui continueront leur voyage. Soyons heureux avec ce que nous avons. Aussi merci d'être un des passagers de mon train.

***Je veux dire à chaque personne qui lira ce texte
que je vous remercie d'être dans ma vie et de voyager dans mon train ».***

Jean d'Ormesson

Si certains trouvent que les remerciements sont parfois de trop dans un manuscrit de thèse, pour ma part il était nécessaire d'adresser noir sur blanc des remerciements à chacune des personnes qui a fait de cette thèse une expérience incroyable aussi bien professionnelle que personnelle. Cette thèse a été bien plus que trois années à programmer derrière un écran d'ordinateur. Ça a été trois de rencontres, que je ne suis pas prête d'oublier. Et si le temps me fait oublier, ces remerciements sauront me les rappeler.

Mes premiers remerciements dans le cadre de ce manuscrit de thèse vont à mes rapporteurs, Carine Briand et Thierry Dudok de Wit, ainsi qu'à mes examinateurs, Aude Chambodut et Jean Lilensten qui ont pris de leur temps durant l'été 2018 pour analyser mon travail et apporter un regard externe et expert, ayant amélioré sans aucun doute sa qualité.

Je tiens aussi à remercier les femmes qui ont encadré ma thèse, et oui dans ce milieu parfois trop masculin, il y a des « supers nanas ». Il y a Angélica, ma directrice, qui a repris ce sujet en plein chantier et m'a accordée le plus important que l'on puisse attendre d'un encadrant, sa confiance. Merci de m'avoir soutenue et défendue à chaque fois que j'avais une idée en tête, tu as été essentielle pour que je tiens en deux ans cette thèse. Il y a aussi Nathalie, toujours présente dès le début pour m'aider sur la moindre question mathématique, à n'importe quelle heure, surtout à la fin de la thèse ! On a toutes les deux complètement perdu la notion du temps au moment de la rédaction, je te remercie sincèrement pour ton soutien 24/24h, 7/7j. Et il y a bien sûr Sandrine, qui m'a choisie il y a 3 ans, m'a aidée à faire mes premiers pas dans ce milieu et a su me guider dans cette première année bien complexe.

J'ai eu la chance d'effectuer ces travaux dans différents laboratoires, et nulle part j'ai vu une ambiance aussi unique que celle de l'ONERA, plus spécifiquement celle du DPHY ! Un département encadré par Jean-Francois, je te remercie de m'y avoir accueillie. J'y ai côtoyé des scientifiques et des hommes incroyables surtout comme ceux de l'équipe ERS ! Merci à Didier, Sébastien, Daniel, Antoine pour tous les échanges que nous avons pu avoir, et à

Vincent pour avoir fait de l'ESWW un souvenir que je ne suis pas prête d'oublier ! Quentin je te cite à part car je tiens à te remercier comme Pauline pour avoir été des supers cobureaux, ça a été chouette d'avoir votre humour et votre franchise au quotidien pour avancer. Charlie et ses drôles de dames, une belle affaire ☺ Il y a bien sûr les anciens doctorants mais avant tout mes amis depuis la fac, Damien et Rémi, ça a été super d'avoir vos conseils dès mon arrivée et votre soutien permanent. Merci à vous deux les petits gars ! Et à MC qui essayait tant bien que mal de vous contrôler ! Et puis au DPHY, on ne se sent jamais vraiment seul. Il y a Christine, toujours présente pour nous aider dans nos galères, qu'aurais-je fais sans toi ! Il y a le duo infernal de l'humour, Monsieur Claude et le grand Stéphane (faudrait que tu me rendes mes places un de ces jours ! Ou que tu te décides à te joindre à nous ;)) Ce duo devient parfois un trio avec celui qui devrait enfin se présenter au concours du meilleur pâtissier, Gaël. Grâce à toi le DPHY n'a pas fini de se régaler à chaque repas de Noël ! Il y a aussi un autre duo, plus jeune, Tic et Tac, ou Marc qui se fracasse le crâne sur des tables et Pacaud qui fait des cris d'oiseaux dans le couloir. Duo qui devient parfois un trio encore avec Charles, toujours partant pour tout ! Merci pour le rire quotidien, indispensable, merci aussi à tous ceux qui font que le DPHY est, au-delà d'un lieu de travail, un endroit où on grandit humainement. J'espère que tous les doctorants qui auront la chance de passer par là se rendront compte de la chance qu'ils ont. A l'ONERA, je tiens aussi à remercier ceux en dehors du DPHY qui ont été présents à différents moments de ma thèse, et apporter leurs contributions scientifiques et humaines, notamment Sébastien et Sylvain du DTIS et Nadine de la communication.

Cette thèse a été cofinancée par le CNES, et je souhaite remercier Guy, Denis, Joëlle, Magali et Myriana pour tous ces beaux moments passés dans ce centre spatial, je n'aurai jamais cru passer de si chouettes moments grâce à une thèse. Avant cela, j'ai eu la chance de faire un stage au CNES, sous la direction de Bruno, un grand monsieur qui m'a beaucoup appris et m'a donné confiance en moi, et l'aide de Rémi, un super opérateur de satellites mais surtout un chouette homme avec qui je continue de discuter de mes doutes pour avancer dans le milieu professionnel. Merci à vous.

A l'IRAP, il y a une personne que je ne remercierai jamais assez, celui qui m'a un jour parlé du sujet qui a rythmé 3 ans de ma vie, Frédéric. Merci beaucoup !

Au nord, bien plus au nord, il y a Enrico, expert en processus gaussien au CWI mais surtout booster de recherches. Je t'ai rencontré durant ma première grande conférence, et j'ai tant appris à tes côtés. Ces quelques mois à Amsterdam m'ont apporté énormément humainement et scientifiquement, alors merci de m'avoir fait confiance pour travailler sur ce projet avec toi.

Plus personnellement, j'aurai de nombreuses personnes à remercier. Car ces 3 ans de thèse ont été des années intenses de vulgarisation, à la rencontre du public, notamment des plus jeunes.

Il y a eu le travail fait avec la Cité de l'Espace, comme le congrès scientifique des enfants. Christophe, Florence, vous faites un travail incroyable, merci, continuez à avoir cette motivation pour faire rêver les personnes qui viennent dans ce lieu.

Il y a eu les clubs d'astro avec le collègue Léon Blum avec Isabelle et toute la belle équipe. Merci pour tout ce que vous faites pour stimuler ces collégiens !

Il y a eu Maths en Scène, avec Houria et son super pouvoir pour montrer que oui, les maths, c'est super cool ! Merci !

Il y a eu Elise avec la super team de Délires d'encre, c'est toujours un vrai bonheur de vous retrouver pour Scientilivre, merci de faire confiance à UniverSCiel depuis 3 ans !

Ah oui UniverSCiel... là on parle de l'association qui me tient à cœur, que j'ai eu un immense plaisir à présider, et qui au-delà de moments inoubliables, m'a fait rencontrer des personnes folles, barrées, tarées, des jeunes chercheurs qui sont partants pour mettre leur temps libre au service de la transmission de connaissances durant une multitude de festivals, notamment durant le festival Astro-Jeunes. Il n'y a qu'avec une bande de vrais fous qu'on tient une semaine intense de festival avec plus de mille enfants de 4 à 17 ans. Merci à chacun de vous, merci à Minus et Cortex ou Gab et Jas pour m'avoir lancée dans cette aventure, merci à mon petit wilou, ma Lulu d'amour qui me manque toujours et super Turpin, Ppeille (désolé pour les licornes), Ines, la duchesse Anne Dory de Bretagne, Wrapin, Arnaud, Morgane, Ilane l'inimitable tata Bernard et toute la nouvelle génération : Dodo, Jeff, Loulou, Babak et ta bonne humeur inégalable, Hadrien, Adrien, Michael, Simon, Gabi... Bien sûr la team Rallye du Gers avec Babouche, Momo et Michmich, on se retrouvera bientôt avec Lulu au pour aller manger des gaufrettes chez pépé et Lili ☺ ! Et merci Sachatte pour avoir repris cette asso, j'ai toute confiance en toi, UniverSCiel a un bel avenir devant soi avec toi aux commandes ! Je vous dis à tous un IMMENSE merci, mais aussi mes excuses pour les surnoms pourris que je vous ai donné et qui (mal)heureusement sont restés !

En travaillant pour UniverSCiel, j'ai aussi rencontré des personnes qui font un travail dingue au quotidien pour la vulgarisation, bravo Thierry et Michael pour ce que vous faites avec la Ferme des Etoiles. Merci Bruno de m'avoir permis de présenter mon sujet de thèse au grand public au festival d'astronomie de Fleurance avant le grand jour de la soutenance, c'était une belle expérience.

Je n'étais pas du tout caillou de l'espace, et pourtant, notre super marraine Brigitte Zanda et son acolyte Sylvain avec Asma ont réussi à me passionner, vous êtes trop forts, bien joué à la team Vigie Ciel ! Et merci Sylvain pour m'avoir fait confiance pour que les petits cueilleurs d'étoiles prennent vie sur Toulouse, c'est un beau challenge que tu m'as offert.

Dès les premiers moments dans ce monde du spatial, il y a bien sûr Peter, un prof en or qui m'a suivi durant tout mon cursus et que je considère comme mon père scientifique, merci pour ton soutien !

Avant la thèse, il y a les amis qui me soutiennent depuis longtemps, ceux qui ont vécu mes meilleurs et pires moments. Mon petit pote, Thomas, rencontré dès les premiers instants de la fac, on s'était dit qu'on deviendrait docteur et nous y voilà, on l'a fait ! Il y a ma poulette, Céline, ma siamoise de la fac, une amie en or, une écoute à toute heure, je sais que je pourrai toujours compter sur toi et ton futur mari, le tonton Pierre ! De ce master, il y a aussi Armel

que j'aurai toujours plaisir à revoir, merci pour ton rire et et Simon le belge, thank you guy for being so belge ! Il y a aussi toute l'équipe du tennis de Frouzins, 11 ans que je vous connais, 11 ans que je suis vos péripéties et que vous suivez les miennes, 11 ans que les moments passés avec vous sont de vrais bols d'airs. Merci les biches, les Sausseureau, et les 3 Laclau-Bloomfield ! Je souhaite aussi remercier la team Mermoud qui m'a encouragé à partir dans cette voie alors que j'étais bien installée avec eux à travailler sur la mécanique des structures, Flo, Anthony et Séverin, vous êtes au top !

Pendant cette thèse j'ai eu la chance de rencontrer une super belgo-vietnamienne, une demoiselle petite par la taille mais grande par l'âme, je suis très fière de t'avoir comme amie, merci d'être dans ma vie Margaux.

Après la thèse, il y a ceux qui m'ont permis de transformer l'essai de la thèse. Ceux qui m'ont fait confiance dès que je leur ai dit que je réfléchissais à ce que je voulais faire, ceux qui sont devenus mes patrons, Starsky et Hutch ou plutôt Loïc et Frédéric. Merci à vous deux, mais aussi à toute l'équipe de Settis. Merci aussi à mes collègues qui m'ont soutenue quand c'était parfois compliqué de jongler entre fin de thèse et nouvel emploi. Merci à ma responsable Marie qui a toujours le sourire et toujours eu le mot en tant que docteure pour me rassurer. Merci à mon RT Fred pour sa bonne humeur permanente, merci à Eugenio, Captain Benoit, archi Alexandre et Pierre d'avoir suivi jusqu'à la soutenance !

Il y a bien évidemment la famille, ma Ohana, merci d'avoir toujours cru que j'en étais capable, mes petits frères et sœurs Lulu et Juju, vous voyez que tout est possible, même une blonde docteure en réseau de neurones ! Alors croyez en vous mes petits loups ☺ Ma mamy, je suis tellement heureuse de t'avoir eu dans le public pour ma soutenance ! Quand je parle de famille je pense aussi bien sûr à ma belle-famille, vous êtes géniaux, merci pour tout ce que vous m'apportez depuis tant d'années, vous êtes indispensables à ma réussite.

Enfin, il y a le seul qui résiste à mes projets les plus fous, un soutien permanent depuis plus de 11 ans, celui pour lequel je ferai tout pour qu'il soit fier de moi, je parle bien sûr de mon chat Jeanmi ! Plus sérieusement, je parle de minou, ou Simon pour les rares qui connaissent son prénom. On a tous besoin d'un repère dans ce joyeux bazar qu'est la vie, merci d'être le mien. Tu es un homme incroyable d'une compréhension qui dépasse l'entendement, j'ai hâte de voir les aventures qui nous attendent car je sais qu'avec toi tout finira bien. Merci infiniment !

ABSTRACT

In this thesis, we present models which belongs to the field of artificial intelligence to predict the geomagnetic index am based on solar wind parameters. This is done in terms to provide operational models based on data recorded by the ACE satellite located at the Lagrangian point L1. Currently, there is no model providing predictions of the geomagnetic index am . To predict this index, we have relied on nonlinear models called neural networks, allowing to model the complex and nonlinear dynamic of the Earth's magnetosphere. First, we have worked on the development and the optimisation of basics neural networks like the multilayer perceptron. These models have proven in space weather to predict geomagnetic index specific to current systems like the Dst index, characteristic of the ring current, as well as the global geomagnetic index Kp . In particular, we have studied a temporal network, called the Time Delay Neural Network (TDNN) and we assessed its ability to predict the geomagnetic index am within one hour, base only on solar wind parameters. We have analysed the sensitivity of neural network performance when considering on one hand data from the OMNI database at the bow shock, and on the other hand data from the ACE satellite at the L1 point. After studying the ability of neural networks to predict the geomagnetic index am , we have developped a neural network which has never been used before in Space Weather, the Long Short Term Memory or LSTM. Like the TDNN, this network provides am prediction based only on solar wind parameters. We have optimised this network to model at best the magnetosphere behaviour and obtained better performance than the one obtained with the TDNN. We continued the development and the optimisation of the LSTM network by using coupling functions as neural network features, and by developing multioutput networks to predict the sectorial am also called $a\sigma$, specific to each Magnetical Local Time sector. Finally, we developped a brand new technique combining the LSTM network and gaussian process, to provide probabilistic predictions up to six hours ahead of geomagnetic index Dst and am . This method has been first developped to predict Dst to be able to compare the performance of this model with reference models, and then applied to the geomagnetic index am .

RÉSUMÉ

Dans cette thèse, nous présentons des modèles appartenant au domaine de l'intelligence artificielle afin de prédire l'indice magnétique global am à partir des paramètres du vent solaire. Ceci est fait dans l'optique de fournir des modèles opérationnels basés sur les données enregistrées par le satellite ACE situé au point de Lagrange L1. L'indice am ne possède pas à l'heure actuelle de modèles de prédiction. Pour prédire cet indice, nous avons fait appel à des modèles non-linéaires que sont les réseaux de neurones, permettant de modéliser le comportement complexe et non-linéaire de la magnétosphère terrestre. Nous avons dans un premier temps travaillé sur le développement et l'optimisation des modèles de réseaux classiques comme le perceptron multi-couche. Ces modèles ont fait leurs preuves en météorologie de l'espace pour prédire aussi bien des indices magnétiques spécifiques à des systèmes de courant comme l'indice Dst , caractéristique du courant annulaire, que des indices globaux comme l'indice Kp . Nous avons en particulier étudié un réseau temporel appelé Time Delay Neural Network (TDNN) et évalué sa capacité à prédire l'indice magnétique am à une heure, uniquement à partir des paramètres du vent solaire. Nous avons analysé la sensibilité des performances des réseaux de neurones en considérant d'une part les données fournies par la base OMNI au niveau de l'onde de choc, et d'autre part des données obtenues par le satellite ACE en L1. Après avoir étudié la capacité de ces réseaux à prédire am , nous avons développé un réseau de neurones encore jamais utilisé en météorologie de l'espace, le réseau Long Short Term Memory ou LSTM. Ce réseau possède une mémoire à court et à long terme, et comme le TDNN, fournit des prédictions de l'indice am uniquement à partir des paramètres du vent solaire. Nous l'avons optimisé afin de modéliser au mieux le comportement de la magnétosphère et avons ainsi obtenu de meilleures performances de prédiction de l'indice am par rapport à celles obtenues avec le TDNN. Nous avons souhaité continuer le développement et l'optimisation du LSTM en travaillant sur l'utilisation de fonctions de couplage en entrée de ce réseau de neurones, et sur le développement de réseaux multisorties pour prédire les indices magnétiques am sectoriels ou $a\sigma$, spécifiques à chaque secteur Temps Magnétique Local. Enfin, nous avons développé une nouvelle technique combinant réseau LSTM et processus gaussiens, afin de fournir une prédiction probabiliste jusqu'à six heures des indices magnétiques Dst et am . Cette méthode a été dans un premier temps développée pour l'indice magnétique Dst afin de pouvoir comparer les performances du modèle hybride à des modèles de référence, puis appliquée à l'indice magnétique am .

AVANT-PROPOS

Si ce manuscrit présente les travaux produits durant trois années de doctorat, il est fondamental de rappeler avant toute chose les travaux qui ont été à l'origine de ce sujet. Tout au long de ce manuscrit, le lecteur trouvera des références sur des travaux pionniers dans le domaine de l'application de l'intelligence artificielle en météorologie de l'espace. Parmi toutes ces références, nous tenions à en présenter une en particulier, qui a été le point de départ de cette recherche. Certes, toutes les notions seront expliquées dans les chapitres à venir, et nous invitons le lecteur à revenir sur cet avant-propos après s'être imprégné de ces connaissances si nécessaire. Il était important de donner une place juste au travail fourni par Farida Mazouz [Mazouz et al., 2013] dans le cadre du projet ATMOP, travail à partir duquel nous avons construit cette thèse.

Le réseau développé dans le cadre du projet ATMOP est un réseau de neurones de type « feedforward backpropagation » comme défini au Chapitre 2 section 3.1.2.1. Le but de ce réseau était de calculer l'indice am à une heure et à trois heures à partir de paramètres du vent solaire et de l'indice am calculé à l'instant précédent (défini comme le am « nowcast » dans le rapport ATMOP). Les paramètres du vent solaire proviennent de la base OMNI et sont considérés entre 1995 et 2012, soit sur environ deux cycles solaires. Un historique de temps associé aux paramètres du vent solaire est également pris en compte, afin de mettre en place dans ce type de réseau non dynamique, un semblant de dynamique temporelle, et d'approximer au mieux le comportement de la magnétosphère. Deux matrices d'étude ont alors été considérées. Ces matrices contiennent les paramètres du vent solaire : la densité, la vitesse, le champ magnétique interplanétaire $|IMF B|$, et la composante B_z de l' $|IMF B|$ en coordonnées GSM à l'instant t , et jusqu'à 24 heures auparavant. Elles contiennent également l'indice magnétique am (en nT) correspondant à ces paramètres du vent solaire, ainsi que l'indice am « nowcast ». Pour analyser les résultats obtenus, le lecteur est invité à étudier le projet de [F. Mazouz et al., 2013]. Nous souhaitons souligner que la configuration qui a fourni les meilleurs résultats est celle où l'on considère en entrée tous les paramètres du vent solaire, avec un historique de temps de douze heures. On considère également l'indice am « nowcast ». Avec cette configuration, pour une prédiction à une heure, le réseau est meilleur lorsque l'activité est basse, c'est-à-dire lorsque l'indice am est inférieur à 20 nT. Globalement, les performances de prédiction à une heure sont meilleures que celles obtenues à trois heures. Cette première étude faite sur la prédiction de l'indice magnétique am a permis de lancer un sujet riche sur le thème de l'intelligence artificielle appliquée au développement d'outils opérationnels en météorologie de l'espace, et nous invitons le lecteur à en découvrir davantage au fil des pages de ce manuscrit.

SOMMAIRE

INTRODUCTION	31
CHAPITRE I	35
1. Les protagonistes de l'interaction Soleil-Terre	37
1.1. Le Soleil et le vent solaire	37
1.2. La magnétosphère.....	40
1.2.1. La magnétopause	40
1.2.2. Les cornets polaires	41
1.2.3. La queue magnétosphérique	42
1.2.4. La magnétosphère interne.....	43
1.2.5. L'ionosphère.....	44
2. La physique de l'interaction soleil-Terre	45
2.1. L'onde de choc terrestre	45
2.2. Les processus d'entrée de particules dans la magnétosphère	46
2.3. Les mécanismes de piégeage des particules	48
2.3.1. La théorie du piégeage.....	48
2.3.2. Les trois invariants adiabatiques	50
2.4. Les courants magnétosphériques.....	51
2.5. Les secteurs MLT.....	52
3. L'étude de l'interaction Soleil-Terre au travers des indices magnétiques	54
3.1. Les indices d'électrojets auroraux <i>AE, AU, AL</i>	54
3.2. Les indices d'activité polaire <i>PCN</i> et <i>PCS</i>	55
3.3. L'indice d'activité à l'équateur <i>Dst</i>	55
3.4. Les indices d'activité globaux	56
4. La complexité de la réponse de la magnétosphère à l'activité solaire ...	57
4.1. L'impact de la magnétogaine sur les paramètres du vent solaire	57
4.2. L'analyse de la réponse de la magnétosphère à l'activité solaire à l'aide des fonctions de couplage	57
4.3. La magnétosphère : un filtre non linéaire.....	59

5. L' utilisation des réseaux de neurones pour établir le lien entre le vent solaire et les indices magnétiques.....	61
5.1. Les premiers réseaux utilisés en météorologie de l'espace	61
5.2. Vers des réseaux plus complexes pour modéliser la dynamique magnétosphérique	62
6. Bilan sur l'état de l'art.....	63
CHAPITRE II.....	65
1. Les données utilisées et les séries temporelles	67
1.1. La nature des séries temporelles.....	67
1.1.1. Données discrètes et continues	67
1.1.2. Objectifs de l'analyse des séries temporelles	67
1.2. Données utilisées pour étudier la relation entre vent solaire et les indices magnétiques (<i>am</i> ou <i>Dst</i>).....	68
1.3. Préparation des données	69
2. Les prévisions de séries temporelles.....	70
2.1. Variables dépendantes et indépendantes	71
2.2. Variables continues et catégorielles	71
2.3. Régression versus Classification	71
3. Les modèles de prévision	72
3.1. Les réseaux de neurones.....	72
3.1.1. Eléments de base des réseaux de neurones.....	72
3.1.2. Topologie des réseaux de neurones utilisés dans le cadre de notre étude	78
3.2. Les processus gaussiens	87
3.2.1. Le théorème de Bayes	88
3.2.2. L'inférence Bayésienne	88
3.2.3. La régression au moyen des processus gaussiens.....	90
3.2.4. Prédire à partir des processus gaussiens	92
3.3. Les méthodes d'évaluation des performances d'un modèle de prédiction	93
3.3.1. L'erreur quadratique moyenne	93
3.3.2. Le coefficient de corrélation.....	93
3.3.3. La matrice de confusion	94
4. Bilan sur méthodes et matériels.....	96

CHAPITRE III	97
1. Analyse de la relation entre paramètres du vent solaire et indices magnétiques	99
1.1. Le coefficient de Kendall pour analyser les liens entre les paramètres du vent solaire et l'indice <i>am</i>	99
1.1.1. Le coefficient de Kendall	99
1.1.2. Comparaison des résultats obtenus en fonction des données considérées.....	101
1.2. Etude de l'historique de temps à considérer en entrée pour optimiser les performances de prédiction.....	104
2. Evaluation des réseaux de neurones à partir des données OMNI et ACE	108
2.1. Mise en évidence de la capacité du Time Delay Neural Network à prédire les effets de l'activité solaire sur l'environnement magnétique terrestre à partir des données OMNI.....	108
2.2. L'impact de l'utilisation des données en temps réel sur les performances des réseaux de neurones	110
2.3. L'analyse au travers d'un événement extrême : l'événement de Juillet 2004.....	113
3. Bilan sur l'étude de la capacité des réseaux de neurones à prédire l'impact de l'activité solaire sur l'environnement magnétique terrestre ..	119
CHAPITRE IV	121
1. Application du réseau Long Short Term memory à notre problématique	123
1.1. Développement du LSTM en Python et première mise en évidence de l'apport de ce réseau en comparaison au réseau de référence	123
1.2. Evaluation des performances de prédiction en fonction des données considérées.....	125
2. Analyse des effets de l'utilisation de fonction de couplage en entrée du LSTM.....	129
2.1. Point sur les fonctions de couplage et leurs applications en entrée des réseaux de neurones	129
2.2. L'apport des fonctions de couplage sur les performances des réseaux avec les données ACE.	134
3. Evaluation de la capacité du LSTM à fournir une prédiction multi-sortie dans le cadre de la prédiction de l'indice $a\sigma$	136
3.1. Le rôle de l' <i>am</i> sectoriel ou $a\sigma$ en météorologie de l'espace	136
3.2. Analyse des performances du LSTM pour la prédiction de l'activité magnétique associée à chaque secteur MLT.....	139

4. Bilan sur le Développement et l'analyse d'un nouveau réseau de neurones pour optimiser les prédictions de l'indice <i>am</i> à partir des paramètres du vent solaire	147
CHAPITRE V	149
1. Développement d'une nouvelle méthode de prédiction associant réseaux de neurones et processus gaussiens.....	151
1.1. L'intérêt des processus gaussiens en météorologie de l'espace	151
1.2. Description de la technique dite GPNN appliquée à la prédiction d'indices magnétiques de une heure à six heures	153
2. Développement et Evaluation de la capacité du GPNN à prédire l'indice magnétique <i>Dst</i> jusqu'à six heures en avance	157
2.1. Développement et optimisation du réseau LSTM pour définir la moyenne du GPNN	157
2.1.1. L'indice magnétique <i>Dst</i> , caractéristique des orages et sous-orages magnétiques.....	157
2.1.2. Définition du réseau LSTM pour la prédiction de l'indice magnétique <i>Dst</i>	159
2.1.3. Mise en évidence des atouts et faiblesses du LSTM à fournir des prédictions de l'indice magnétique <i>Dst</i> à partir de comparaison avec des modèles de référence	161
2.2. Analyse de la prédiction probabiliste fournie par le GPNN	164
2.2.1. Les courbes ROC pour évaluer les performances du GPNN en fonction de seuils d'activité	165
2.2.2. Analyse des diagrammes de fiabilité	169
2.2.3. Apport du GPNN pour la prédiction d'un événement extrême	170
3. Application de la technique hybride associant réseaux de neurones et processus gaussiens pour prédire l'indice magnétique <i>am</i>	171
4. Bilan sur l'optimisation des prédictions d'indices magnétiques à plus long termes au moyen d'une nouvelle technique hybride	177
CONCLUSION ET PERSPECTIVES	179
ANNEXES	183
1. Analyse de l'apport des paramètres <i>Vx</i> et <i>Bz</i> pour prédire l'indice magnétique <i>am</i> avec le TDNN	183
2. De Matlab vers Python	186
2.1. Définition des bibliothèques utilisées	187

2.2. Architecture globale du code.....	187
3. Liste des orages utilisés pour le test du processus gaussien combiné au réseau LSTM.....	189
4. Courbes ROC obtenues avec la méthode GPNN dans le cas de la prédiction de l'indice magnétique Dst.....	191
BIBLIOGRAPHIE	195
PUBLICATIONS.....	203

DATA ACKNOWLEDGEMENT

Tout au long de ce manuscrit, nous avons utilisé des bases de données fournies par des centres fournissant un travail sans lequel cette étude n'aurait pas été possible. Nous tenons donc à remercier les organismes suivants pour le travail conséquent effectué afin que ces modèles aient pu voir le jour :

- Les indices géomagnétiques ont été calculés et rendus disponibles par ISGI, à partir de données collectées par différents observatoires magnétiques. Ces indices proviennent d'instituts nationaux impliqués dans le réseau INTERMAGNET (<http://isgi.unistra.fr/>).
- Les données du vent solaire ont été obtenues à partir de la base OMNI, développée et entretenue par le National Space Science Data Center (NNSDC) de la NASA (<https://omniweb.gsfc.nasa.gov/ow.html>). Elles ont également été obtenues directement au niveau du satellite ACE au point de Lagrange L1 grâce à l'équipe du Caltech (<http://www.srl.caltech.edu/ACE/ASC/level2/index.html/>).
- Les données GPS ont été rendues publiques depuis peu par l'équipe CXD du Los Alamos National Laboratory (www.ngdc.noaa.gov/stp/space-weather/satellite-data/satellite-systems/gps).

NOMENCLATURE

* - Produit matriciel de Hadamard

ACE – Advanced Composition Explorer (satellite d’observation solaire localisé au point de Lagrange L1)

AE, AU, AL – Indices d’électrojets auroraux

am - Indice d’activité globale

aσ- Indice d’activité sectoriel spécifique à chaque secteur MLT

ASY-H - Indice caractéristique du courant annulaire (composante asymétrique)

b – Biais

B – Champ magnétique

B_T – Composante transverse du champ magnétique

B_z – Composante z du champ magnétique interplanétaire

CME – Coronal Mass Ejection (éjection de masse coronale)

C- Fonction de coût

CC – Coefficient de corrélation

C_t- Cell state

c_s – Vitesse du son

dΦ_{MP}/dt -Taux de variation du flux magnétique à la magnétopause

∇ - Gradient d’information

Dst – Indice magnétique caractéristique du courant annulaire

E – Espérance mathématique

ϵ - Energie sortante d’après [Perreault and Akasofu, 1978]

E_{in} – Energie entrante d’après [Wang et al, 2014]

η – Taux d’apprentissage

F10.7 Flux radio solaire à 10.7 cm

FAR – Taux de fausses alarmes

FPR – False Positive Rate

fs – Flow speed

GP – Gaussian Process (processus gaussien)

$GPNN$ – Gaussian Process-Neural Network

IMF – Interplanetary Magnetic Field

j - Invariant adiabatique

K – Flux d'énergie totale

$k(x, x')$ – Kernel ou fonction de covariance

Kp – Indice d'activité globale

L – Paramètre de Mc Illwain

L^* - Paramètre de Roederer ou de coquille

$LSTM$ - Long Short Term Memory (réseau de neurones avec mémoire à court et long termes)

$m(x)$ - Moyenne du processus gaussien

M – Moment magnétique relativiste

M_A – Nombre de Mach d'Alfvén

M_E – Moment du dipôle magnétique

MLP – MultiLayer Perceptron (perceptron multicouche)

MLT – Magnetic Local Time (Temps Magnétique Local)

n – Densité du vent solaire

\mathcal{N} - loi normale

$NARX$ - Non linear AutoRegressive with eXogenous inputs

\mathcal{NN} – Neural Network

μ_0 - perméabilité magnétique

ν – Viscosité cinématique

PCN et PCS – Indices d'activité polaire

POD – Probabilité de détection

R_\odot - Rayon solaire

R_e – Rayon terrestre

RMSE – Root mean square error

ROC – Receiver Operating Characteristic

SSN – Sunspot Number

SYM-H – Indice caractéristique du courant annulaire (composante symétrique)

tanh – Fonction tangente hyperbolique

τ – Coefficient de Kendall

τ_c – Période cyclotron

τ_{dd} – Période de rebond

τ_d – Période de dérive

TDNN – Time Delay Neural Network (réseau à retard de temps)

θ – IMF clock angle

TPR – True Positive Rate

U – Flux d'énergie entrant

V_A – Vitesse d'Alfvén

$V_{r,l}$ – Vitesse magnéto-sonore rapide et lente

W – Poids de l'information pour le réseau de neurones

X – Paramètres d'entrée ou feature du réseau de neurones ou du processus gaussien

LISTE DES FIGURES

Figure 1- Illustration de la structure du Soleil.....	37
Figure 2- Evolution du Sunspot Number (SSN) à gauche, et du flux radio F10.7 à droite depuis 1967.	38
Figure 3- Spirale de Parker [Parker, 1958].....	38
Figure 4- Evénements solaires extrêmes. a) Eruption Solaire, b) CME, c) Trou coronal s'ouvrant depuis l'un des pôles. Crédit : SOHO/EIT/ESA/NASA.	39
Figure 5- Vue d'artiste de la magnétosphère Les lignes de champ magnétique deviennent, dans cette représentation en trois dimensions, des surfaces qu' on appelle des coquilles magnétiques [Lilensten and Bornarel 2001]	40
Figure 6- En bleu on observe la magnétopause, surface créée par le contournement de la Terre par le vent solaire. Les flèches rouges représentent le courant de magnétopause [Lilensten and Bornarel 2001].....	41
Figure 7- Les flèches montrent les trajectoires de quatre particules du vent solaire piégées le long d'une ligne de champ associée aux cornets polaires. Elles peuvent alors être directement précipitées vers l'atmosphère polaire, ou être expulsées vers l'Espace [Lilensten and Bornarel 2001]	42
Figure 8 - Représentation schématique de la magnétosphère dans son ensemble.	43
Figure 9 - La magnétosphère interne. a) plasmasphère, b) ceintures de radiation interne et externe, c) courant annulaire. Crédit : Aurora Explorer.....	43
Figure 10 - En amont de la Terre, l' onde de choc crée un échauffement du vent solaire [Lilensten and Bornarel 2001].....	46
Figure 11- Entrées de particules dans la magnétosphère, [Lilensten and Bornarel 2001]	47
Figure 12 - Coquille de dérive définie par les mouvements de giration, rebond et dérive d'une particule.	49
Figure 13- Exemples de courants créés suite à l'interaction Soleil-Terre.....	52
Figure 14- Les secteurs MLT.	53
Figure 15 - Couvertures temporelles associées aux données considérées pour la prédiction de l'indice magnétique am et de l'indice Dst.	69
Figure 16- Le neurone biologique.....	73
Figure 17- Le neurone formel.	73
Figure 18- Du perceptron au perceptron multicouche [Nielsen, 2015]].	74
Figure 19- Répartition des données entre sous-ensemble d'entraînement, test et validation ...	76
Figure 20- Evolution des courbes d'erreur en fonction des itérations (ou epoch). La courbe rouge représente l'évolution durant l'apprentissage, la bleue celle durant les phases de test ou de validation. La zone verte représente les itérations ou epoch à partir desquelles on constate un surapprentissage avec une augmentation de l'erreur de test.	77
Figure 21 –Perceptron multicouche.	79
Figure 22- Perceptron Multicouche adapté à notre problématique. L'indice prédit par le réseau ou « nowcast » est renvoyé en entrée.	79
Figure 23- Le réseau à retard de temps ou TDNN.	81

Figure 24- Réseau récurrent en déroulé, X représente l'entrée, A le neurone, et H la sortie. (http://colah.github.io).....	82
Figure 25- Le réseau non linéaire autorégressif à entrées exogènes.	82
Figure 26- Problème de la dépendance dans le temps entre une information entrante X0 et une sortie à l'instant $ht + 1$ (http://colah.github.io).	83
Figure 27- Réseau LSTM.	84
Figure 28- Schéma réseau récurrent. (http://colah.github.io).....	84
Figure 29 - Schéma réseau LSTM. (http://colah.github.io).	84
Figure 30- Schéma du convoyeur du réseau avec les ponts sigmoïdes. (http://colah.github.io).	85
Figure 31- Schéma du "forget gate" et équation associée. (http://colah.github.io).....	85
Figure 32- Schéma du pont servant à garder les nouvelles informations et équations associées. (http://colah.github.io).....	86
Figure 33- Schéma de création de la nouvelle "cell state" et équation associée. (http://colah.github.io).....	86
Figure 34 - Schéma de création de la nouvelle sortie et équations associées. (http://colah.github.io).....	87
Figure 35- Schéma des "peephole connection"et équations associées. (http://colah.github.io).	87
Figure 36- Régression bayésienne pour un exemple hypothétique. La figure de gauche montre la distribution de fonctions a priori. La figure de droite montre la distribution a posteriori après avoir ajouté deux points de données observées. La ligne continue montre la prédiction moyenne, moyennée sur les différentes fonctions possibles. L'aire grisée représente l'écart type (standard deviation) à chaque point d'entrée [Rasmussen and Williams, 2006]	90
Figure 37 - Variations des résultats fournis par les processus gaussiens sur quatre points d'observations en fonction du kernel considéré : squared exponential, matern52, rational quadratic, periodic. (https://pythonhosted.org).	92
Figure 38- Répartition des am (nombre de points pour un seuil donné) et définition des seuils d'activité.	94
Figure 39- a) Distribution des valeurs de la vitesse (fs-flow speed) entre 1995 et 2012, b) Tracé de la vitesse en fonction de l'indice magnétique am pour des valeurs entre 1995 et 2012.	100
Figure 40- Réseaux considérés pour l'analyse de la capacité de réseaux de neurones à prédire l'indice magnétique am, a) le « feedforward backpropagation », b) le TDNN, c) le NARX. Les cadres rouges soulignent les paramètres temporels à optimiser pour la prédiction.....	105
Figure 41- POD et FAR du réseau « feedforward » en fonction de l'historique de temps du vent solaire considéré.	106
Figure 42-POD et FAR de chaque réseau avec les données OMNI en considérant un historique de temps de 6h et 12h. Le réseau « feedforward » est en vert, le TDNN en rouge et le NARX en bleu.	110
Figure 43- POD et FAR de chaque réseau avec les données ACE au point L1 en considérant un historique de temps de 6h et 12h. Le réseau « feedforward » est en vert, le TDNN en rouge et le NARX en bleu.	112

Figure 44- Exemple de données fournies par la base OMNI, de gauche à droite : année, jour de l'année (doy), heure, IMF B , densité et vitesse.	113
Figure 45- Description de l'événement de Juillet 2004. De haut en bas : vitesse du vent solaire, densité de proton, IMF B , Kp, am et le flux intégré d'électron enregistré par NPOES 15. .	114
Figure 46- Comparaison entre l'am réel et le am prédit pour l'événement de juillet 2004 avec les données OMNI en utilisant a) le réseau « feedforward », b) le réseau TDNN, c) le réseau NARX.....	116
Figure 47- Comparaison entre l'am réel et le am prédit pour l'événement de juillet 2004 avec les données ACE en utilisant a) le réseau « feedforward », b) le réseau TDNN, c) le réseau NARX.....	117
Figure 48- Evolution de l'erreur en fonction des epoch en phase d'entraînement et de validation pour le réseau LSTM à gauche et « feedforward » à droite.	124
Figure 49- POD et FAR du réseau LSTM en considérant la densité, la vitesse et l'IMF B . Le LSTM prenant en compte les données OMNI est en bleu, celui considérant les données ACE est en rose en fonction du niveau d'activité.	126
Figure 50- Événement de juillet 2004 avec les données OMNI en utilisant le réseau LSTM. Les données réelles sont en bleu, les données prédites en orange pointillé.	127
Figure 51- Événement de juillet 2004 avec les données ACE en utilisant le réseau LSTM. Les données réelles sont en bleu, les données prédites en orange pointillé.....	128
Figure 52- POD et FAR du réseau LSTM (rose foncé) et du TDNN (rose clair) en considérant la densité, la vitesse et l'IMF B de ACE. La fenêtre de spécialisation du TDNN est de six heures.	129
Figure 53- Entrée en Energie et impact sur les systèmes de courant d'après [Akasofu, 1981].	130
Figure 54- Estimation de l'Energie entrante au niveau de Mariner 5 et Explorer 34 d'après [Akasofu, 1981].....	131
Figure 55- Prédiction de Kp à partir de l'indice de Boyle à une heure et trois heures [Bala and Reiff, 2012].	133
Figure 56- POD et FAR du réseau LSTM en utilisant les paramètres du vent solaire à partir de ACE en rose, et utilisant la fonction de couplage définie par [Wang et al., 2014] à partir des données ACE en violet en fonction du seuil d'activité.	135
Figure 57- Événement de juillet 2004 avec l'équation de [Wang et al., 2014] à partir des données ACE en utilisant le réseau LSTM. Les données réelles sont en bleu, les données prédites en orange pointillé.	136
Figure 58 –Représentation UT/DOY des 4 indices associés aux secteurs MLT. La valeur moyenne sur les UT de ces indices en fonction du DOY est tracée dans l'encadré de droite [Chambodut et al., 2013].....	137
Figure 59 – Paramètres du vent solaire et indices géomagnétiques durant l'orage de Mai 2003. La pression du vent solaire est en vert, la vitesse en bleu, les composantes Z et Y de l'IMF B dans le repère GSM sont en rouge et bleu. L'indice am ainsi que les am sectoriels sont représentés à la suite ainsi que l'indice SYM-H [Chambodut et al., 2013].	138
Figure 60- Variation des indices am sectoriel pour l'événement de Juillet 2004.	139

Figure 61- Réseaux considérés pour la prédiction des indices sectoriels : un réseau multisortie vs. 4 réseaux monosortie. Dans tous les cas, le nombre de cellules au sein du LSTM est de 20.	140
Figure 62- POD et FAR du réseau LSTM multisortie en utilisant les paramètres du vent solaire à partir de ACE permettant de fournir des prédictions de σ_{dawn} , σ_{noon} , σ_{dusk} , $\sigma_{midnight}$ en fonction du niveau d'activité.	141
Figure 63- POD et FAR des 4 réseaux spécifiques à chaque indice sectoriel en utilisant les paramètres du vent solaire à partir de ACE permettant de fournir des prédictions de σ_{dawn} , σ_{noon} , σ_{dusk} , $\sigma_{midnight}$	142
Figure 64- Événement de juillet 2004 avec les données ACE en utilisant le réseau LSTM a) multisortie, b) monosortie pour la prédiction de σ_{dusk} . Les données réelles sont en trait continu, les données prédites sont en pointillé.	144
Figure 65- Événement de juillet 2004 avec les données ACE en utilisant le réseau LSTM a) multisortie, b) monosortie pour la prédiction de $\sigma_{midnight}$. Les données réelles sont en trait continu, les données prédites sont en pointillé.	146
Figure 66- Prédiction de l'indice Dst au moyen des processus gaussiens. La valeur réelle est en noire, la valeur prédite en rouge, les valeurs limites hautes et basses sont définies par les enveloppes vertes et bleues. [Chandorkar et al., 2017].	152
Figure 67- Etapes principales du développement du GPNN.	153
Figure 68- Événements considérés pour l'entraînement du processus gaussien.	155
Figure 69 - Intervalle de confiance pour évaluer la précision de l'estimation.	156
Figure 70- Fonctionnement opérationnel du GPNN.	156
Figure 71: Phases d'un orage magnétique en fonction du Dst.	158
Figure 72 - Couverture temporelle des études de référence et de notre étude.	161
Figure 73- Prédictions obtenues sans données GPS (en rouge) et avec les données GPS (en bleu) pour l'événement d'Halloween 2003. La valeur réelle est en gris.	164
Figure 74- Classification des orages magnétiques - site AER.	165
Figure 75 - Illustration de la notion de seuil © P. Calmant et E. Depiereux.	166
Figure 76 - Courbes ROC associées à une matrice de confusion. Chaque point définissant les courbes correspond à des couples TPR/FPR calculés pour différents seuils. La courbe de gauche représente une variation de TPR/FPR en fonction des valeurs seuils considérées. Celle du milieu représente le cas parfait pour lequel une discrimination totale est possible avec un FPR de 0 et un TPR de 1. la courbe de droite représente un cas pour lequel aucune discrimination n'est possible, le résultat est dû au hasard. © P. Calmant et E. Depiereux.	167
Figure 77 - Courbes ROC pour les prédictions à une heure obtenues avec le GPNN.	168
Figure 78 - TPR et FPR associés à chaque domaine d'activité, pour différentes seuils, en fonction du temps de prédiction considéré.	169
Figure 79 - Diagramme de fiabilité pour chaque temps de prédiction considéré. La diagonale en pointillé rouge représente une évolution parfaite de la fréquence observée en fonction de la probabilité de prédiction.	170
Figure 80- GPNN appliqué à la prédiction de l'orage d'Halloween 2003.	171
Figure 81 – Sous-ensembles d'entraînement du GPNN.	172
Figure 82- Courbes ROC associées aux sous-ensembles d'entraînement a) et b) de la Figure 81.	174

Figure 83- Diagrammes de fiabilité associés aux sous-ensembles d'entrainement a) et b) de la Figure 81.	175
Figure 84 - Résultats de la prédiction de l'événement de Juillet 2004 obtenus à partir des sous-ensembles d'entrainement a) et b) de la Figure 13. La ligne noire représente la valeur réelle, la ligne bleue la valeur moyenne prédite et la partie grisée représente la dispersion à $\pm 2\sigma$ autour de la valeur moyenne.....	176
Figure 85- POD et FAR de chaque réseau avec les données ACE au point L1 en considérant la densité, la vitesse, l'IMF B et B_z avec un historique de temps de 6h.. Le réseau « feedforward » est en vert, le TDNN en rouge et le NARX en bleu.	184
Figure 86- POD et FAR de chaque réseau avec les données ACE au point L1 en considérant la densité, la vitesse, l'IMF B et V_x avec un historique de temps de 6h. Le réseau « feedforward »est en vert, le TDNN en rouge et le NARX en bleu	185
Figure 87- Matlab vs Python.....	186
Figure 88- Architecture globale du code.....	188

INTRODUCTION

Notre Soleil, cette masse qui représente à elle seule environ 99.8% de la masse de notre système solaire et qui rend la vie possible sur Terre par l'apport de lumière et de chaleur est une étoile qui émet en permanence un flux de particules issu de la couronne solaire en expansion. L'activité de notre Soleil est complexe à évaluer, on parle de cycle solaire pour définir l'alternance de minima et de maxima d'activité solaire. Des événements extrêmes conduisent sporadiquement à une libération importante de particules, potentiellement dangereuses pour un environnement planétaire. Notre planète est protégée par un bouclier magnétique naturel qui nous protège des effets les plus nocifs. Cependant, il existe une interaction permanente entre le vent solaire et le champ magnétique terrestre, formant la magnétosphère terrestre. La première manifestation résultante de l'interaction Soleil-Terre est fascinante, de par la beauté qu'elle peut créer dans le ciel nocturne. Les aurores boréales ont depuis toujours émerveillés l'être humain. Mais si ces manifestations lumineuses semblent magiques pour bon nombre de personnes, derrière ces volutes se cachent une physique complexe, et de nombreux phénomènes problématiques pour nos technologies.

En effet, depuis la conquête spatiale à la fin des années 1950, le nombre de lancement de satellites dans notre environnement n'a cessé d'augmenter. Les technologies qui nous entourent sont pour la plupart connectées à nos satellites, comme la téléphonie, internet et les systèmes GPS. Comme il est expliqué dans l'ouvrage de [Lilensten and Bornarel 2001] notre technologie utilise les mêmes vecteurs que le Soleil pour transporter de l'énergie. En raison de cette similitude de nature physique, il existe un lien étroit entre activité solaire et activité technologique humaine. Par le passé, des événements solaires extrêmes ont fortement perturbé des satellites, conduisant parfois à leurs pertes, comme ce fut le cas pour le satellite ANIK en 1994 ou le satellite Telstar 401 en 1997. C'est ainsi qu'est née la météorologie de l'espace. Cette branche de l'astrophysique a pour but d'étudier les phénomènes causés par les rayonnements et les particules émises par le Soleil, ainsi que les interactions entre le Soleil, l'Espace interplanétaire et la Terre, dans le but de les comprendre et les anticiper.

Pour quantifier les effets du vent solaire sur l'environnement magnétique terrestre, les scientifiques font appel aux indices magnétiques. Ce sont des indicateurs des perturbations du champ magnétique terrestre, mesurés à partir de magnétomètres situés aux pieds des lignes de champ magnétique. En fonction de la position des magnétomètres sur le globe, ces indices peuvent être spécifiques à un système de courant, comme l'indice *Dst* caractéristique du courant annulaire avec des stations situées au niveau de l'équateur magnétique, ou l'indice *AE* caractéristique des électrojets auroraux avec des stations situées aux pôles. D'autres indices comme l'indice *Kp* ou *am* sont caractéristiques des perturbations magnétiques globales avec des stations situées à moyenne latitude.

Différentes méthodes détaillées dans le Chapitre 1 ont alors émergées pour modéliser la magnétosphère, et sa réponse à une perturbation associée à l'activité solaire. [Bargatze et al., 1985]. ont utilisé des modèles de filtres linéaires pour modéliser la magnétosphère, et ont ainsi analysé la réponse de l'activité magnétique mesurée au sol par les magnétomètres en fonction des perturbations externes provenant du vent solaire. Cette étude a permis de vérifier qu'il fallait considérer une composante de l'activité magnétique qui n'était pas directement connectée au vent solaire, et que la magnétosphère est impactée par d'autres mécanismes comme le chargement-déchargement de celle-ci. [Perreault and Akasofu, 1978] ont développé les premiers modèles de fonction de couplage permettant d'évaluer la dispersion en énergie depuis l'entrée de la magnétosphère au travers des différents systèmes de courants. Ces fonctions ont été les premières à fournir une évaluation des relations entre paramètres du vent solaire et indices magnétiques terrestres. Cependant, la magnétosphère reste un

système complexe à modéliser et à prévoir en ne faisant appel qu'à des modèles physiques. Son comportement non-linéaire, ainsi que la difficulté à anticiper l'activité solaire, a conduit les scientifiques à considérer des modèles appartenant aux techniques de l'intelligence artificielle, plus spécifiquement, les réseaux de neurones. Ces modèles peuvent établir des connexions non-linéaires entre des paramètres d'entrée comme les paramètres du vent solaire, et des paramètres de sortie comme les indices magnétiques. Les réseaux de neurones présentent l'intérêt de ne nécessiter aucune connaissance a priori sur les mécanismes d'interaction Soleil-Terre. Ce point est important car les relations entre vent solaire et indices magnétiques posent toujours question à l'heure actuelle. Ainsi, [Lundstedt and Wintoft, 1994], et [Boberg et al., 2000] ont montré qu'il était possible de prédire respectivement l'activité du courant annulaire avec l'indice *Dst*, l'activité spécifique aux électrojets auroraux avec l'indice *AE* et l'activité magnétique globale avec l'indice *Kp*, en utilisant des réseaux de neurones de type perceptron multicouche. Ces modèles, pionniers en météorologie de l'espace, sont statiques et ne sont pas suffisamment représentatifs de la dynamique de la magnétosphère terrestre. Par la suite, [Gleisner et al., 1996] ont développé des réseaux de neurones dynamiques de type Time Delay Neural Network pour prédire l'indice magnétique *Dst*. Ceci a permis de montrer que lorsque la dynamique est prise en compte dans le modèle de calcul non linéaire, les performances de prédiction en sont améliorées. L'évolution des orages magnétiques est ainsi mieux prédite, notamment la phase de recouvrement. Cependant, ces réseaux présentent des lacunes en cas d'activité extrême et ne sont pas suffisamment robustes pour faire de la prédiction en temps réel. Depuis, de multiples modèles ont été développés, notamment le modèle de [Wing et al. 2005], modèle opérationnel utilisé par la NOAA pour prédire l'indice *Kp* à une heure et à quatre heures.

L'ensemble de ces études a souligné la complexité de la réponse de la magnétosphère à l'activité solaire. Les réseaux de neurones présentent une solution optimale pour répondre à la problématique non linéaire de la dynamique magnétosphérique. Les prévisions fournies par ces méthodes sont un outil précieux en météorologie de l'espace pour anticiper un événement solaire extrême et permettre à un opérateur d'effectuer des manœuvres de protection en cas d'alerte. Cependant, ces modèles de prédictions manquent de précisions pour les épisodes d'activité solaire élevée. Ils présentent pour la plupart également l'inconvénient d'être développés à partir de données prétraitées fournies par la NASA, ce qui ajoute du temps de calcul pour obtenir une prédiction en temps réel. Ces réseaux ont été développés pour des indices spécifiques à un système de courant comme *AE* ou *Dst*, ou pour l'indice magnétique global *Kp*. Pour évaluer l'activité magnétique globale, l'indice magnétique *am* est un excellent indicateur, qui, contrairement à *Kp*, n'est pas défini sur une échelle logarithmique. Il est la traduction directe de la perturbation du champ magnétique terrestre, et n'est pas plafonné. A l'heure actuelle aucun modèle de prédiction n'existe pour cet indice. C'est ce qui justifie cette étude.

Cette étude consiste à développer des modèles de prédiction de l'indice magnétique *am* de type réseaux de neurones, en utilisant uniquement les paramètres du vent solaire enregistrés par un satellite d'observation solaire afin d'anticiper en temps réel l'impact de l'activité solaire sur l'environnement magnétique terrestre.

Après avoir présenté en détail les réseaux de neurones et processus utilisés dans notre étude dans le Chapitre 2, nous présentons dans le Chapitre 3 l'analyse faite sur la capacité du TDNN à prédire l'indice magnétique global *am* à court terme en utilisant uniquement les paramètres du vent solaire. Nous avons étudié ses performances en le comparant au réseau de référence, le perceptron multicouche, et à un réseau présentant d'excellentes performances de prévisions, le réseau non linéaire auto-régressif à entrées exogènes ou NARX. Ces deux derniers prennent en compte en entrée la valeur de l'indice magnétique prédite par le réseau. C'est une information dont nous souhaitons nous affranchir dans un cadre opérationnel car un indice n'est pas définitif avant un certain délai, délai

durant lequel un opérateur prend le risque de considérer en entrée une valeur erronée. Au moment où cette étude a été mise en place, le TDNN était le seul réseau de neurones permettant de réaliser une prédiction basée uniquement sur les paramètres du vent solaire. Nous avons étudié les performances des réseaux de neurones au travers de métriques classiques comme les coefficients de corrélation et l'erreur quadratique moyenne, mais aussi en utilisant des métriques statistiques basées sur une matrice de confusion que nous détaillons au Chapitre 2. Cette analyse a été faite à partir des données fournies par la base OMNI de la NASA, mais aussi à partir des données mesurées par le satellite ACE – Advanced Composition Explorer- situé au point de Lagrange L1. Nous présentons alors l'impact de l'utilisation de ces données sur les performances des modèles. Les travaux et résultats associés à cette étude ont été présentés dans un article en révision au Space Weather Space Climate (*Prediction of the geomagnetic index am based on the development and the performance comparisons of static and dynamic Neural Networks*, M. A. Gruet, N. Bartoli, S. Rochel, R. Benacquista, A. Sicard and G. Rolland).

Sachant que le TDNN est limité par une fenêtre temporelle fixe, nous avons souhaité programmer un réseau encore jamais utilisé en météorologie de l'espace, le réseau Long Short Term Memory (LSTM). Ce réseau possède une mémoire à court et long termes, ce qui est adapté pour modéliser la complexité de la dynamique magnétosphérique en réponse à une perturbation provenant du vent solaire. Ce réseau est, comme le TDNN, basé uniquement sur les paramètres du vent solaire. Nous avons programmé ce réseau, puis nous l'avons adapté aux données fournies en temps réel par le satellite ACE et nous avons étudié sa capacité à prédire des événements solaires extrêmes. Au Chapitre 4, nous présentons les performances de ce réseau de neurones pour prédire l'indice magnétique am , en utilisant en entrée dans un premier temps les paramètres du vent solaire pris séparément, puis une fonction de couplage définie par [Wang et al., 2014]. Nous détaillons également dans ce Chapitre 4 le développement du réseau LSTM pour prédire les indices am sectoriel ou $a\sigma$, spécifiques à chaque secteur Temps Magnétique Local (MLT) définis par [Chambodut et al., 2013].

Enfin, nous présentons au Chapitre 5 le fruit d'une collaboration avec le CWI (centrum voor wiskunde und informatica, le laboratoire de recherche appliquée en informatique et mathématique d'Amsterdam) sous la direction d'Enrico Camporeale, chercheur dans l'équipe MultiScale Dynamics. Ce projet a eu pour but de combiner la précision des réseaux de neurones à l'aspect probabiliste des processus gaussiens. En effet, ces processus permettent de répondre au besoin d'un opérateur qui souhaite évaluer l'erreur associée à une prédiction. Nous avons également développé cette technique afin de fournir une prédiction au-delà d'une heure d'indices magnétiques. Cette étude a été faite dans un premier temps pour prédire l'indice magnétique Dst , indice caractéristique des orages et sous-orages magnétiques, puis dans un second temps pour l'indice magnétique am . Les travaux et résultats associés à la prédiction de l'indice Dst au moyen de cette technique ont été acceptés dans le Journal Space Weather (*Multiple-Hour-Ahead forecast of the Dst index using a combination of Long Short-Term Memory Neural Network and Gaussian Process*, M.A. Gruet, M. Chandorkar, A. Sicard, E. Camporeale).

CHAPITRE I

ETAT DE L'ART : DU SOLEIL JUSQU'AUX MAGNÉTOMÈTRES, L'INTERACTION SOLEIL-TERRE

Cet état de l'art a pour but de fournir une photographie des connaissances existantes en amont du travail présenté dans la suite de ce manuscrit. Il permet d'aborder les relations Soleil-Terre depuis l'origine du vent solaire, jusqu'à son impact sur l'environnement terrestre, aussi bien au travers des processus physiques existants, de la présentation des indices magnétiques, indicateurs essentiels de la réponse de la magnétosphère, que des analyses physiques et mathématiques associées à la météorologie de l'espace. Comme le montre cette partie, ces analyses ont pour but de mieux comprendre l'interaction Soleil-Terre et de fournir des outils opérationnels pour prévoir les effets de l'activité solaire sur l'environnement terrestre.

1. Les protagonistes de l'interaction Soleil-Terre.....	37
1.1. Le Soleil et le vent solaire	37
1.2. La magnétosphère.....	40
1.2.1. La magnétopause	40
1.2.2. Les cornets polaires	41
1.2.3. La queue magnétosphérique	42
1.2.4. La magnétosphère interne.....	43
1.2.5. L'ionosphère.....	44
2. La physique de l'interaction soleil-Terre	45
2.1. L'onde de choc terrestre	45
2.2. Les processus d'entrée de particules dans la magnétosphère	46
2.3. Les mécanismes de piégeage des particules	48
2.3.1. La théorie du piégeage.....	48
2.3.2. Les 3 invariants adiabatiques.....	50
2.4. Les courants magnétosphériques.....	51
2.5. Les secteurs MLT	52
3. L'étude de l'interaction Soleil-Terre au travers des indices magnétiques	54

3.1.	Les indices d'électrojets auroraux <i>AE, AU, AL</i>	54
3.2.	Les indices d'activité polaire <i>PCN</i> et <i>PCS</i>	55
3.3.	L'indice d'activité à l'équateur <i>Dst</i>	55
3.4.	Les indices d'activité globaux	56
4.	La complexité de la réponse de la magnétosphère à l'activité solaire	57
4.1.	L'impact de la magnétogaine sur les paramètres du vent solaire	57
4.2.	L'analyse de la réponse de la magnétosphère à l'activité solaire au travers des fonctions de couplage	57
4.3.	La magnétosphère, un filtre non linéaire	59
5.	L'utilisation des réseaux de neurones pour établir le lien entre le vent solaire et les indices magnétiques	61
5.1.	Les premiers réseaux utilisés en météorologie de l'espace	61
5.2.	Vers des réseaux plus complexes pour modéliser la dynamique magnétosphérique	62
6.	Bilan sur l'état de l'art	63

1. LES PROTAGONISTES DE L'INTERACTION SOLEIL-TERRE

1.1. Le Soleil et le vent solaire

Notre Soleil est une étoile active, âgée d'environ 4,6 milliards d'années. Comme le montre la Figure 1 présentant la structure de notre astre, il est constitué principalement d'un cœur à $0,25 R_{\odot}$ au sein duquel se forme de l'hélium (5,92%) à partir de l'hydrogène (93,96%), d'une zone radiative en rotation uniforme ($0,45 R_{\odot}$) et d'une zone convective en rotation différentielle ($0,3 R_{\odot}$). Les réactions de fusion permettant la création d'hélium à partir d'hydrogène sont visibles en surface, grâce à la présence d'un champ magnétique complexe. Ce champ magnétique est produit dans la zone de transition entre la zone radiative et la zone convective, appelée tacholine. Il est soumis à des processus de dynamo chaotiques conduisant à deux effets :

-une activité permanente et turbulente visible à la surface. On observe alors des cellules de convection et une éjection permanente de matière dans l'Espace. Ceci conduit à des phénomènes violents et sporadiques sur l'échelle du jour,

-une activité cyclique d'environ 11 ans, associée au cycle solaire et définie suite au comptage des taches solaires comme illustré sur la Figure 2. Cette figure met en évidence les différents cycles solaires depuis 1967, en traçant respectivement deux indicateurs, le nombre de taches solaires (le sunspot number ou SSN en noir sur le graphique de gauche), et le flux radio solaire à 10,7 cm (ou F10.7 en jaune sur le graphique de droite).

En réalité, le champ magnétique solaire s'inversant tous les 11 ans environ, la véritable périodicité serait de 22 ans.

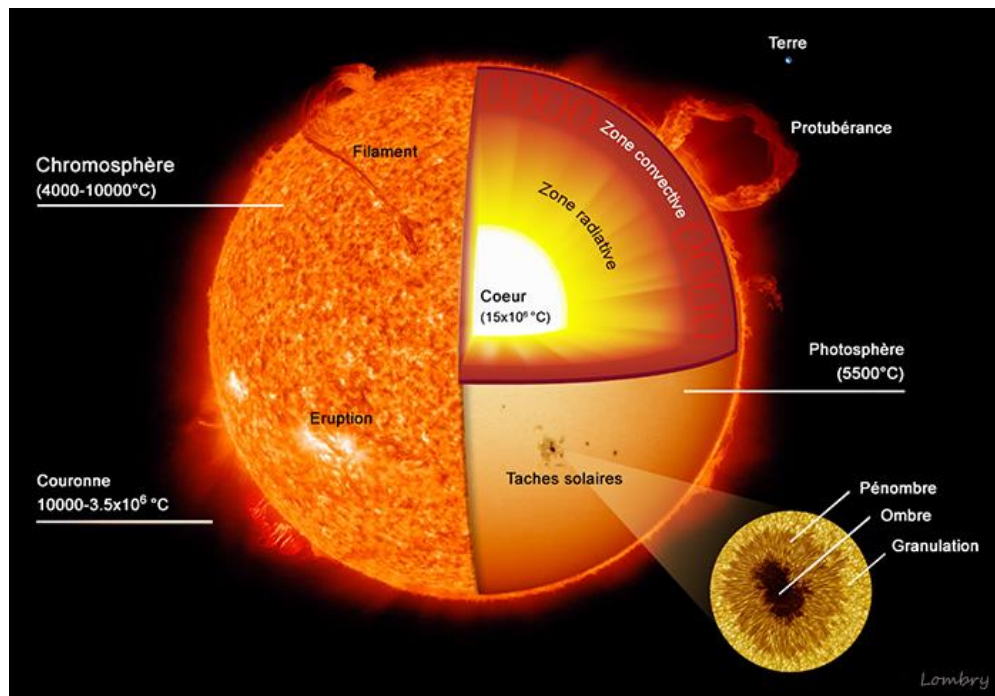


Figure 1- Illustration de la structure du Soleil.

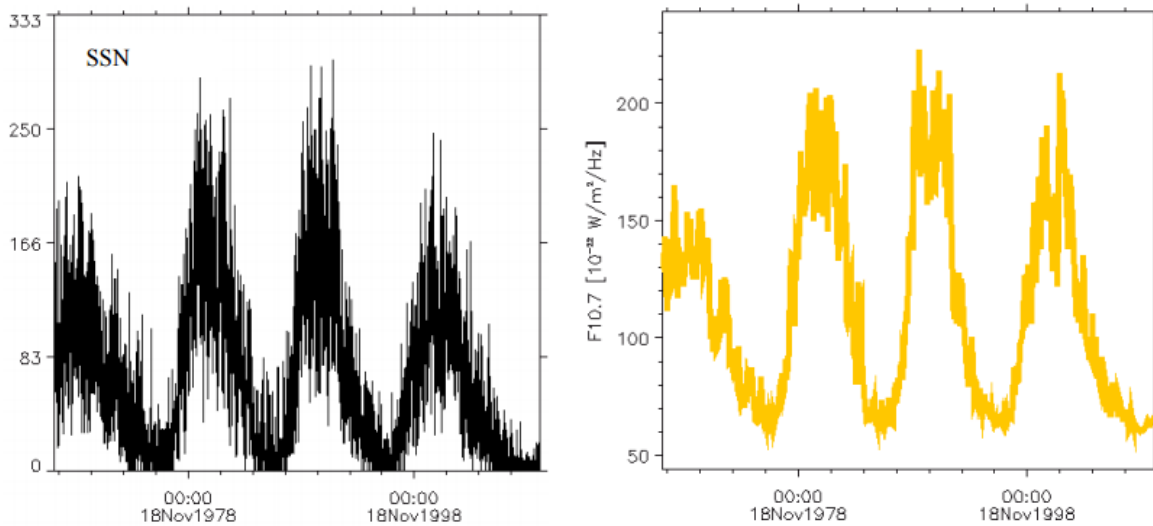


Figure 2- Evolution du Sunspot Number (SSN, nombre de tâches solaires) à gauche, et du flux radio F10.7 à droite depuis 1967.

L'atmosphère solaire est constituée de la chromosphère et de la couronne comme illustré sur la Figure 1, qui s'étend sur quelques rayons solaires. Cette dernière atteignant des températures de plusieurs milliers de degré Celsius, les particules chargées qui la composent peuvent échapper à l'attraction gravitationnelle du Soleil. Un flot de particules chargées en expansion radiale est ainsi créé et entraîne avec lui le champ magnétique solaire : le vent solaire. Par l'intermédiaire de son champ magnétique décrit précédemment et du plasma qu'il émet, l'influence du Soleil s'étend à tout l'espace interplanétaire, entraînant des variations périodiques caractéristiques du milieu.

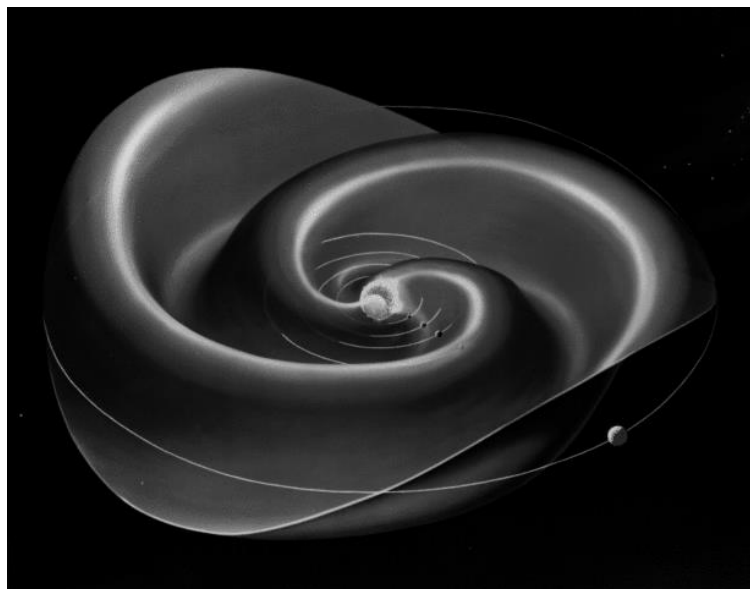


Figure 3- Spirale de Parker [Parker, 1958] .

La rotation du Soleil associée à l'expansion radiale du vent solaire donne au champ magnétique interplanétaire une forme de spirale : la spirale de [Parker, 1958] , illustrée sur la Figure 3. Ce vent solaire est composé principalement d'électrons et de protons, ainsi que d'ions Hélium (He^{2+}) et d'ions plus lourds. A une unité astronomique dans le plan de l'écliptique, le vent solaire est très peu dense, environ 7 cm^{-3} . Il est cependant rapide, entre 400 et 800 km.s^{-1} , bien supérieur à la vitesse du son dans le vent solaire (60 km.s^{-1}) ou la vitesse d'Alfvén (environ 40 km.s^{-1}) décrites à la section 2.1.

Des phénomènes localisés existent au cours desquels des flux élevés de particules (électrons, protons et ions lourds) sont émis. On distingue notamment trois catégories de phénomène: les éruptions solaires, les éjections de masse coronale, et les trous coronaux d'où s'échappe un vent solaire rapide. Ces phénomènes sont potentiellement géo-effectifs. La géo-effectivité définit la capacité d'un événement solaire à perturber l'environnement terrestre.

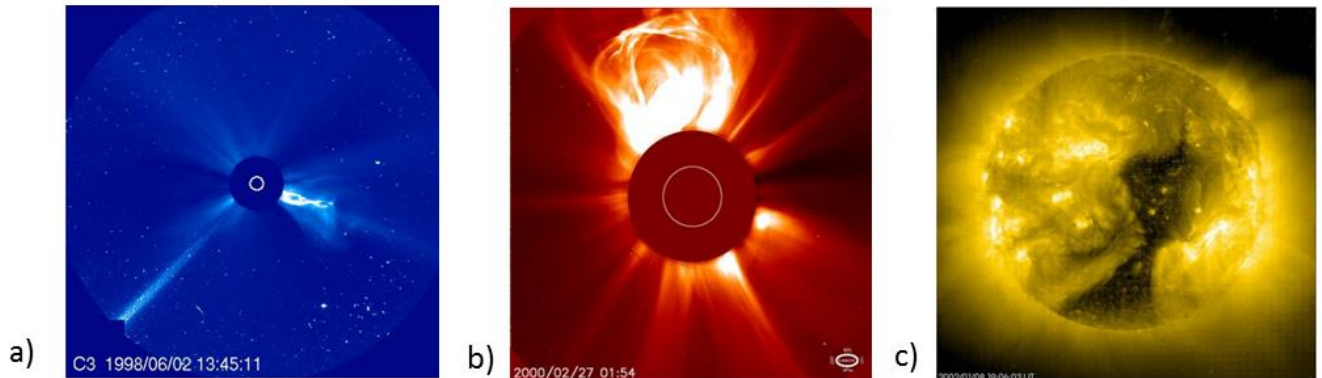


Figure 4- Événements solaires extrêmes. a) Eruption Solaire, b) CME, c) Trou coronal s'ouvrant depuis l'un des pôles. Crédit : SOHO/EIT/ESA/NASA.

Les éruptions solaires illustrées sur la Figure 4.a. sont des phénomènes locaux extrêmement violents de dissipation d'énergie magnétique et d'émission de photons et de particules jusqu'à une centaine de MeV. Elles proviennent du stockage d'énergie magnétique au niveau des taches solaires suivie par une brusque libération de cette énergie. L'énergie ainsi libérée est transférée à la matière environnante, permettant l'accélération des particules qui peuvent alors s'échapper vers le milieu interplanétaire. L'environnement électromagnétique terrestre peut être perturbé par ces éruptions, en raison des ondes de choc qu'elles génèrent dans le milieu interplanétaire (transmission au bout de quelques jours) et des particules énergétiques qu'elles injectent.

Une éjection de masse coronale (Coronal Mass Ejection, CME) représentée sur la Figure 4.b. est une explosion brutale où le plasma de la couronne va se trouver piégé dans une boucle magnétique. C'est un phénomène de grande ampleur dégageant une énergie colossale : la taille des CME en se propageant peut atteindre plusieurs dizaines de rayons solaires et ainsi potentiellement provoquer des orages magnétiques en interagissant avec le champ magnétique terrestre. La vitesse des éjections de matière varie de quelques centaines à quelques milliers de kilomètres par seconde. La fréquence des CME varie, jusqu'à 3 par jour durant les périodes de maximum d'activité solaire.

Lorsque les lignes de champ magnétique local à la surface du Soleil sont ouvertes vers l'espace, une grande quantité de matière sous forme de plasma est projetée au-delà de la couronne et forme ce qu'on appelle les vents solaires rapides (composés principalement de protons dont l'énergie n'excède pas les quelques keV et d'électrons, avec une vitesse supérieure à 700 km/s et une faible densité de 10 particules. cm^{-3} maximum). Les trous coronaux sont les zones où se produit ce phénomène. En raison de la perte de matière, la surface solaire est momentanément moins chaude et dense, ce qui explique que l'on visualise les trous coronaux sur la Figure 4.c. comme des tâches sombres sur des observations aux rayons X ou même UV.

1.2. La magnétosphère

La magnétosphère de la Terre a été découverte en 1958 par la sonde Explorer 1. Auparavant, les scientifiques savaient seulement que des courants électriques s'écoulaient dans l'Espace, suite aux perturbations magnétiques provoquées par les éruptions solaires.

[Gold, 1959] proposa le terme de magnétosphère, quand il écrivit :

« La région au-dessus de l'ionosphère, dans laquelle le flux magnétique de la Terre a un contrôle dominant sur les gaz et particules chargées rapides, est connue pour s'étendre sur une distance de 10 fois le rayon terrestre; son nom approprié pourrait être magnétosphère. »

Ainsi, la magnétosphère terrestre est une cavité modelée par le champ magnétique de la Terre, qui est lui-même engendré par les mouvements du noyau métallique liquide de celle-ci par effet dynamo. Le vent solaire modifie la forme de la magnétosphère comme on peut le voir sur la Figure 5. Il la comprime à une dizaine de rayons terrestres dans la direction du Soleil, et lui confère la forme d'une queue de comète du côté nuit où elle s'étend alors sur une distance 10 fois plus importante.

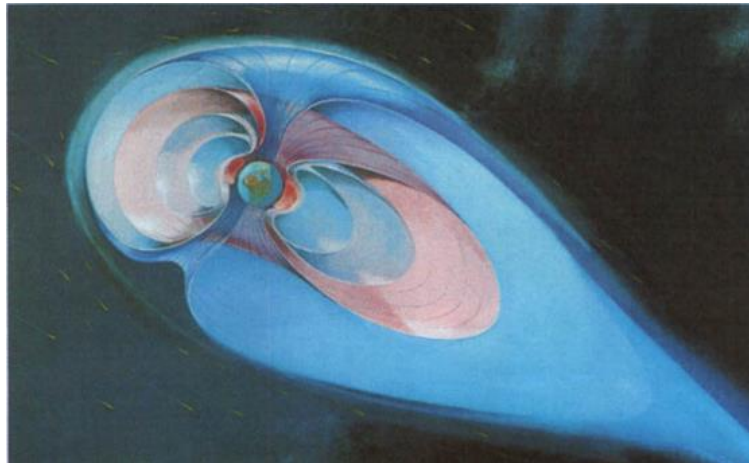


Figure 5- Vue d'artiste de la magnétosphère Les lignes de champ magnétique deviennent, dans cette représentation en trois dimensions, des surfaces qu'on appelle des coquilles magnétiques [Lilensten and Bornarel 2001].

1.2.1. La magnétopause

La magnétopause est le nom donné à la limite supérieure de la magnétosphère. Elle sépare le champ géomagnétique et le plasma principalement d'origine terrestre du plasma du vent solaire.

Cette limite a d'abord été proposée par [Chapman and Ferraro, 1931]. A l'époque, elle était définie comme un courant corpusculaire provenant du Soleil, présent uniquement durant les périodes d'activité solaire. Par conséquent la frontière était intermittente. Ils avançaient déjà à l'époque que la compression du champ magnétique par le plasma sortant causait des perturbations géomagnétiques observées à la surface terrestre, corrélées avec l'activité solaire.

Dans la plus simple des approximations, la magnétopause peut être définie comme une couche de courant se formant au point d'équilibre entre la pression dynamique du vent solaire et la pression magnétique du champ géomagnétique de la Terre.

A l'approche de la Terre, les particules chargées électriquement subissent l'action du champ géomagnétique croissant. Ainsi, comme l'illustre la Figure 6, une force perpendiculaire à la fois au vent et au champ (la force de Lorentz), dévie les ions vers le côté après-midi de la Terre (l'est) et les électrons vers le côté matin (l'ouest).

A l'extérieur, l'Espace est soumis au régime de vent solaire et au champ magnétique interplanétaire. A l'intérieur, c'est le champ géomagnétique de la Terre qui gouverne. La magnétopause se dresse comme une barrière entre l'un et l'autre.

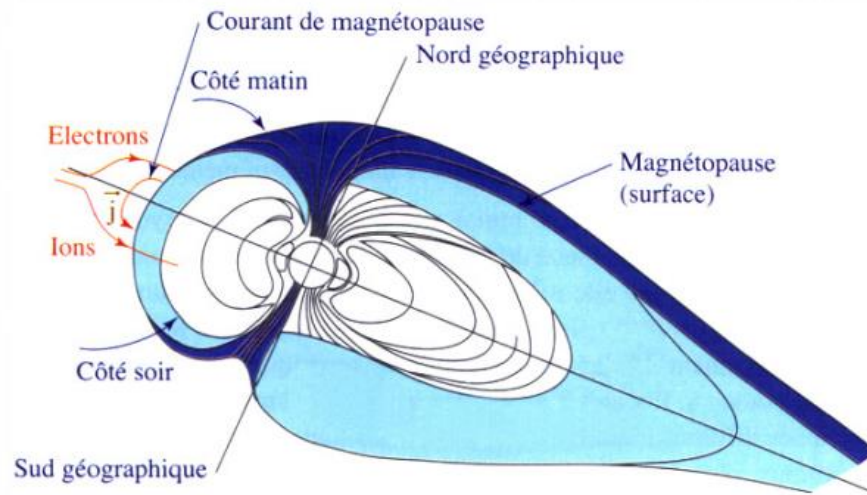


Figure 6- En bleu on observe la magnétopause, surface créée par le contournement de la Terre par le vent solaire. Les flèches rouges représentent le courant de magnétopause [Lilensten and Bornarel 2001]

1.2.2. Les cornets polaires

Les cornets polaires (cusps) marquent la séparation entre les lignes de champ magnétique situées côté jour et côté nuit de la magnétophère. Ces cornets polaires constituent une zone d'entrée directe du plasma. Comme l'illustre la Figure 7, les particules piégées suivent les lignes du champ magnétique terrestre le long des lignes rouges qui mènent aux pôles magnétiques avec un mouvement en spirale. Côté nuit, la couche frontière qui recouvre la partie à haute latitude de la magnétophère s'appelle le manteau de plasma. Ce terme a été introduit pour la première fois par [Rosenbauer et al. 1975]. Le manteau de plasma commence à la calotte polaire (polar cap) et s'étend vers la queue magnétophérique

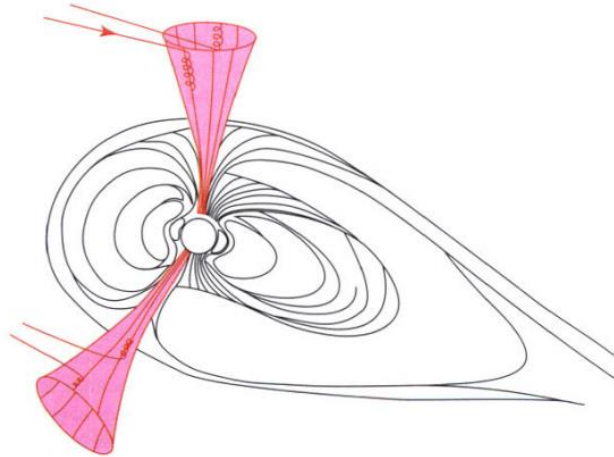


Figure 7- Les flèches montrent les trajectoires de quatre particules du vent solaire piégées le long d'une ligne de champ associée aux cornets polaires. Elles peuvent alors être directement précipitées vers l'atmosphère polaire, ou être expulsées vers l'Espace [Lilensten and Bornarel 2001].

Les cornets sont des zones de connexion entre le vent solaire et la magnétosphère. Les cornets polaires peuvent parfois être le siège d'aurores près de la Terre (bien que les aurores soient plus fréquentes et plus visibles dans les ovales auroraux), par précipitation de particules de hautes énergies issues de la queue magnétosphérique.

1.2.3. La queue magnétosphérique

La queue magnétosphérique est le nom donné à la région de la magnétosphère terrestre qui s'étire côté nuit, c'est à dire à l'opposé du Soleil.

C'est une région importante de la magnétosphère car elle agit comme un réservoir de plasma et d'énergie. L'énergie et le plasma sont relâchés dans la magnétosphère interne aperiodiquement durant des épisodes magnétiquement perturbés appelés sous-orages magnétosphériques que nous détaillons au Chapitre 5 section 2.1.1.

Dans cette région, on retrouve deux zones distinctes :

- Les lobes magnétosphériques (magnetic lobe), en bleu clair sur la Figure 8,
- La couche de plasma (plasma sheet), en jaune sur la Figure 8.

La couche de plasma s'étend dans le plan médian de la magnétosphère, entre les deux lobes. Le champ magnétique dans le lobe nord est dirigé vers la Terre, celui du sud en sens inverse. D'où le besoin d'un feuillet de courant pour séparer ces deux régions de champ magnétique opposé. Dans des conditions normales, cette couche de plasma est constituée des coquilles magnétiques fermées connectées aux régions aurorales. On distingue deux parties : la couche centrale (central plasma sheet, CPS), et la couche frontière du feuillet de plasma (plasma sheet boundary layer PSBL). Cette zone de transition est constituée d'un plasma chaud dont la densité moyenne est comprise entre 0.4 et 2 cm³. En comparaison, les lobes magnétosphériques sont constitués de plasma froid à très faible densité (10⁻² à 10⁻³ cm³).

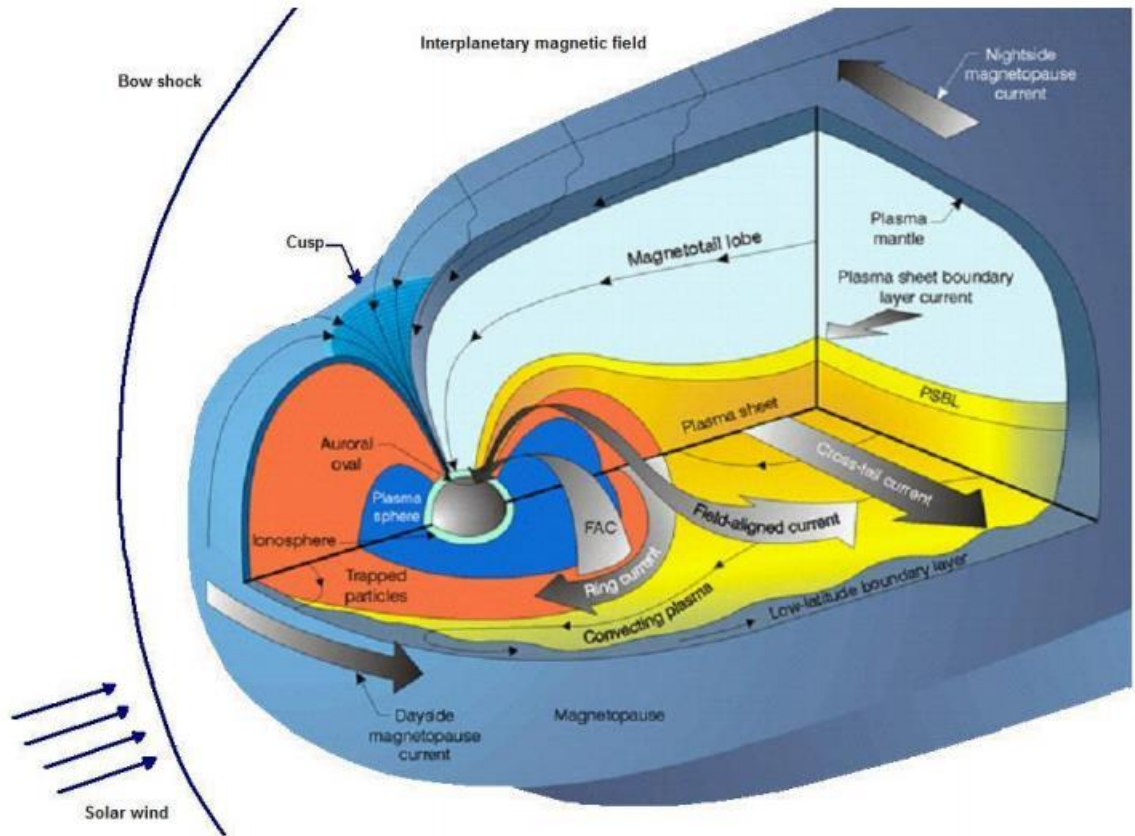
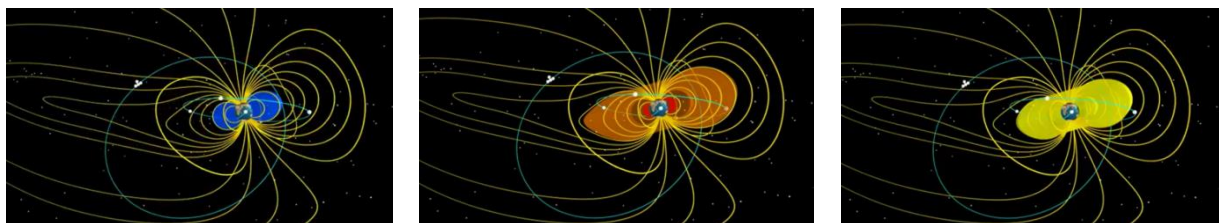


Figure 8 - Représentation schématique de la magnétosphère dans son ensemble.

1.2.4. La magnétosphère interne

La magnétosphère interne comporte la plasmasphère, les ceintures de radiation et le courant annulaire. Ces trois régions forment une région de haute densité avec trois populations de particules différentes.

La plasmasphère, en bleu sur la Figure 9.a., région toroïdale en corotation avec la Terre, est composée d'un plasma froid (électrons et protons d'énergie de l'ordre de l'électron-Volt) d'origine principalement ionosphérique. La limite de la plasmasphère est appelée magnétopause.



a)

b)

c)

Figure 9 - La magnétosphère interne. a) plasmasphère, b) ceintures de radiation interne et externe, c) distribution du courant annulaire. Crédit : Aurora Explorer.

Les ceintures de radiation ou « ceintures de Van Allen », correspondent aux régions annulaires, que l'on situe dans le plan de l'équateur magnétique. Dans ces régions, les particules chargées sont piégées par le champ magnétique et dérivent autour de la Terre. Elles forment des régions de plasma chaud, provenant de la queue magnétosphérique et de la magnétosphère interne. La première ceinture, en rouge sur la Figure 9.b. se trouve autour de 5000 kilomètres d'altitude et contient surtout des protons énergétiques, d'énergies supérieures à 10 MeV. La deuxième que l'on observe en orange sur la Figure 9.b. contient des électrons et des protons d'énergie moindre entre 0.1 et 10 MeV aux environs de 25 000 kilomètres.

Le courant annulaire, en jaune sur la Figure 9.c., que l'on décrira davantage à la section 3.3 se situe au bord de la ceinture externe. Il est composé d'un plasma tiède composé de particules énergétiques entre 10 et 200 keV, en dérive longitudinale, provenant de la queue magnétosphérique.

1.2.5. L'ionosphère

La couche d'atmosphère où prennent place les aurores polaires s'appelle l'ionosphère. Elle s'étend de 80 km jusqu'à plus de 600 km. Elle est principalement composée d'atomes et de molécules neutres provenant de l'atmosphère de la Terre. Elle peut être ionisée par deux sources d'énergie : le rayonnement solaire et les précipitations de particules en région aurorale. Le couplage entre la magnétosphère et l'ionosphère joue un rôle clef lors des sous-orages magnétosphériques autour de 150 km d'altitude. Ceux-ci engendrent des variations du champ magnétique mesurables au sol en région aurorale par des magnétomètres. Les mesures fournies par ces magnétomètres sont décrites à la section 3.3.

2. LA PHYSIQUE DE L'INTERACTION SOLEIL-TERRE

2.1. L'onde de choc terrestre

C'est au tout début des années soixante qu'émerge pour la première fois l'idée que l'interaction du vent solaire avec la magnétosphère de la Terre puisse générer une onde de choc en amont de l'environnement terrestre. Cette hypothèse est confirmée peu après avec les premières observations in situ de cette onde de choc par les sondes Mariner 2 et IMP 1.

Cette onde de choc est définie par différentes vitesses caractéristiques explicitées dans le Tableau 1. Elle résulte de l'interaction d'un fluide s'écoulant à une vitesse supérieure à l'une des vitesses caractéristiques du milieu avec un obstacle immobile. Ces vitesses sont celles des ondes générées par une perturbation (comme un obstacle) qui annoncent aux particules de ce fluide son approche, leur permettant ainsi de se réorganiser en fonction de cette perturbation.

Tableau 1 - Vitesses caractéristiques et nombre de Mach en milieu neutre et ionisé.

Fluide neutre	Milieu ionisé	Milieu ionisé
<i>Vitesse du son</i> [m.s ⁻¹]	<i>Vitesse d'Alfvén</i> [m.s ⁻¹]	<i>Vitesses magnétosonores rapide (Vr) et lente (Vl)</i> [m.s ⁻¹]
$c_s = \sqrt{\gamma k_B T / m}$, γ est l'indice polytropique du milieu, k_B la constante de Boltzmann, T la température et m la masse moyenne des particules	$V_A = B / \sqrt{\mu_0 \rho}$, B est l'amplitude du champ magnétique, μ_0 la perméabilité du vide et ρ la masse volumique du plasma	$V_{r,l} = \frac{1}{2} * (c_s^2 + V_A^2) \pm \frac{1}{2} \sqrt{(c_s^2 + V_A^2)^2 - 4c_s^2 V_A^2 \cos^2 \theta_{kB}}$, où le signe + correspond au mode rapide (Vr) et le signe - au mode lent (Vl). θ_{kB} est l'angle entre le champ magnétique et la direction de propagation des ondes.
Nombre de Mach sonique	Nombre de Mach alfvénique	Nombre de Mach magnétosonore
$M_S = V / c_s$	$M_A = V / V_A$	$M_{ms} = V / V_{r,l}$
<i>V vitesse caractéristique de l'écoulement dans le référentiel de l'obstacle</i>		

Le vent solaire s'approchant de la Terre avec une vitesse principalement super-alfvénique, une onde de choc doit alors se former pour permettre la déflexion de l'écoulement du plasma autour de l'obstacle que constitue la magnétosphère terrestre. On parle alors de formation de choc d'étrave.

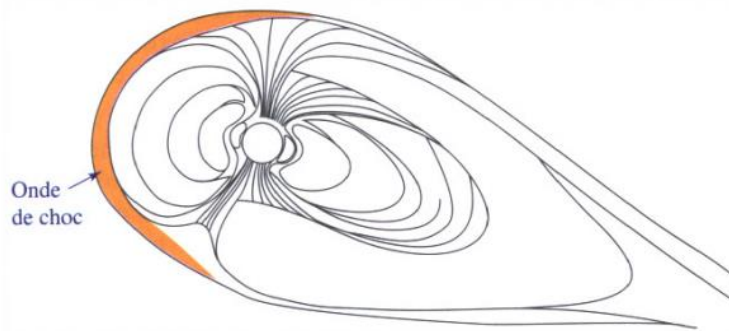


Figure 10 - En amont de la Terre, l'onde de choc crée un échauffement du vent solaire
[Lilensten and Bornarel 2001]

Cette zone représentée en orange sur la Figure 10 est une région de forts gradients, où les propriétés du milieu sont modifiées brutalement afin que la vitesse du vent solaire en aval devienne inférieure à la vitesse magnéto-sonore rapide locale, qui est la plus grande des trois vitesses caractéristiques dans le plasma.

Ainsi, en aval du choc, l'écoulement peut contourner l'obstacle formé par l'environnement planétaire car l'information de sa présence a le temps de se propager. Au passage du choc, il y a diminution de la vitesse, tandis que dans le même temps la densité, la température et l'intensité du champ magnétique augmentent. Cette transition s'effectue de façon irréversible, car il y a transformation d'énergie cinétique en d'autres formes d'énergie. Cette énergie est principalement de l'énergie thermique, désordonnée, ce qui conduit à une augmentation de l'entropie.

Un des paramètres clés de l'onde de choc est le nombre de Mach. Dans le plasma il existe trois nombres de Mach différents, définis dans le Tableau 1. Ce nombre de Mach nous renseigne sur la "force" du choc : plus il est élevé et plus l'écoulement incident est rapide, et donc plus grande est l'énergie qui doit être dissipée au niveau du choc.

Dans des conditions de vent solaire standard, les trois nombres de Mach, sonique, alfvénique et magnéto-sonore, sont du même ordre de grandeur, autour de 8 ou 10, pour l'onde de choc terrestre. Dans le cas d'évènements solaires transitoires, comme les éjections de masse coronales, ils peuvent varier dans une gamme allant de 1 à 30.

2.2. Les processus d'entrée de particules dans la magnétosphère

Le Soleil émet en continu des particules chargées dans le milieu interplanétaire. Le gaz ionisé ainsi formé constitue le vent solaire.

Le paramètre β d'un plasma définit le rapport de sa pression cinétique et de sa pression magnétique tel que

$$\beta = \frac{p}{p_m} = \frac{n k_B T}{B^2 / 2\mu_0} \quad (1)$$

avec n et T , la densité et la température du plasma en [m⁻³] et [K] respectivement, k_B la constante de Boltzmann, B l'induction magnétique [T] et μ_0 la perméabilité du vide [H.m⁻¹]. Deux régimes sont alors distinguables :

- $\beta < 1$, les effets du champ magnétique prédominent au sein du plasma

- $\beta > 1$, l'énergie cinétique prédomine : c'est le cas du vent solaire qui est un plasma supersonique à fort β , très conducteur et dont la densité décroît comme l'inverse du carré de la distance au Soleil.

Le diamètre du Soleil est 109 fois supérieur à celui de la Terre, donc le vent solaire ne fait pas que frapper la face avant de la magnétosphère, il longe aussi les flancs est-ouest et les frontières nord-sud.

Les particules vont alors pouvoir entrer dans la magnétosphère à différents endroits, comme le montre la Figure 11 :

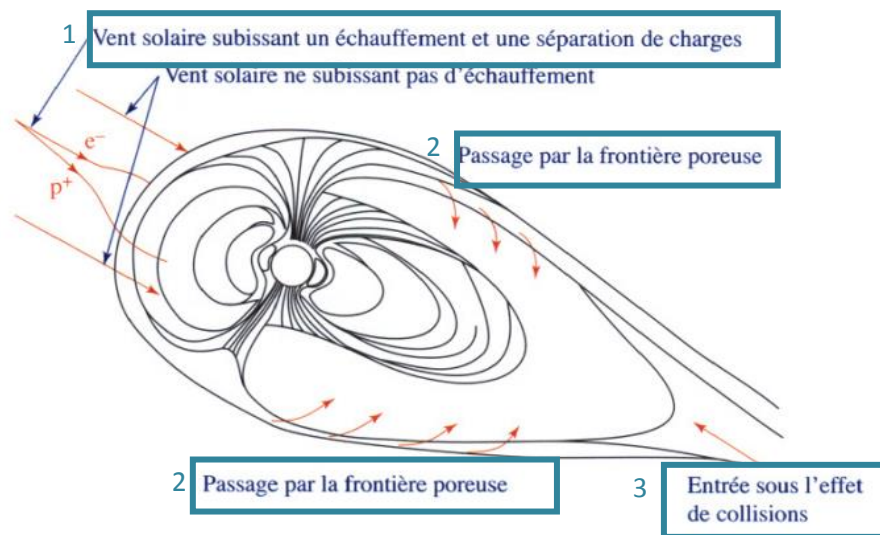


Figure 11- Entrées de particules dans la magnétosphère, [Lilensten and Bornarel 2001]

- « Zone 1 » : Lorsque les particules du vent solaire rencontrent la magnétopause dans l'axe Terre-Soleil, on observe deux phénomènes : un échauffement et une séparation de charges par déviation des ions vers le côté soir et des électrons vers le côté matin. Une fraction d'entre elles pénètre à l'intérieur de la magnétosphère (par exemple en suivant les systèmes de courant magnétosphériques ou encore en raison de la turbulence dans la magnétogaine). Cette fraction est associée aux caractéristiques du vent solaire à l'instant considéré.
- « Zone 2 » : Sur les flancs de la magnétosphère on retrouve un mélange de vent solaire globalement neutre en provenance directe du Soleil et de particules qui ont été accélérées sur la face avant de la magnétosphère. La magnétopause n'est pas totalement imperméable face à ces particules et 5 à 10 % d'entre elles arrivent à la traverser.
- « Zone 3 » : Lorsque le champ interplanétaire est orienté vers le Sud ($B_z < 0$), il est anti-parallèle au champ géomagnétique sur la face avant de la magnétosphère. Au niveau de la magnétopause, des processus agissant à l'échelle microscopique brisent localement les conditions de gel permettant ainsi l'interconnexion entre le champ interplanétaire et le champ terrestre.

Ces lignes de champ interconnectées sont ensuite convectées vers la queue de la magnétosphère, puis vers la couche de plasma. Des processus de reconnexion magnétique se développent dans la queue lointaine, permettant de recomposer la ligne de champ terrestre qui est injectée vers la Terre et la ligne de champ interplanétaire qui est redonnée au vent solaire. Une fraction de l'énergie transmise par le vent solaire est directement dissipée dans l'ionosphère par les précipitations de particules énergétiques. L'autre partie est accumulée dans la queue sous forme magnétique. L'énergie accumulée est ensuite

libérée lors de périodes agitées appelées sous-orages magnétosphériques dont nous parlons au Chapitre 5 section 2.1.1.

2.3. Les mécanismes de piégeage des particules

Nous avons vu précédemment que les particules chargées pénétrant dans la région interne de la magnétosphère subissent une accélération en raison du champ électrique, ainsi qu'une modification de la direction de leur vitesse due au champ magnétique terrestre pouvant les conduire au piégeage. Les ceintures de radiation en sont l'exemple le plus typique.

2.3.1. La théorie du piégeage

Une particule chargée soumise à un champ électromagnétique répond à la force de Lorentz \vec{F} définie par :

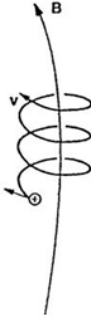
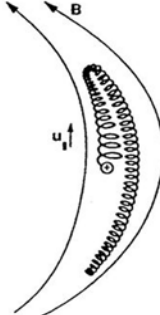
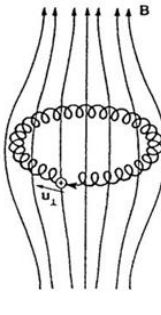
$$\vec{F} = q(\vec{E} + \vec{v} \wedge \vec{B}) \text{ [N]} \quad (2)$$

avec q la charge de la particule, v sa vitesse, E et B les champs électrique et magnétique. En présence d'un champ magnétique fort ou d'une vitesse importante, l'écriture de l'équation précédente se simplifie :

$$\vec{F} = \frac{d\vec{p}}{dt} = q(\vec{v} \wedge \vec{B}) \text{ [N]} \quad (3)$$

avec p la quantité de mouvement de la particule et t le temps. La solution de l'équation donne le mouvement d'une particule chargée dans un champ magnétique. Dans le cas d'un champ magnétique de type dipolaire, le mouvement est quasi-périodique. Grâce à l'approximation du centre-guide, ce dernier peut être décomposé en trois mouvements périodiques élémentaires que l'on décrit dans le Tableau 2, classés par ordre croissant de période.

Tableau 2- Mouvements périodiques élémentaires d'une particule piégée.

		
<p>Giration autour de la ligne de champ magnétique</p>	<p>Rebond entre deux points miroirs situés dans chaque hémisphère</p>	<p>Dérive autour de la Terre dans un plan perpendiculaire à l'axe du dipôle, vers l'est pour les électrons et vers l'ouest pour les protons. Dérive azimuthale perpendiculaire au champ magnétique</p>

La composition de ces trois mouvements forme une coquille de dérive comme illustré sur la Figure 12. Notons que l'approximation centre-guide ne peut être faite que si, et seulement si, le rayon de giration reste petit devant le rayon de courbure des lignes de champ.

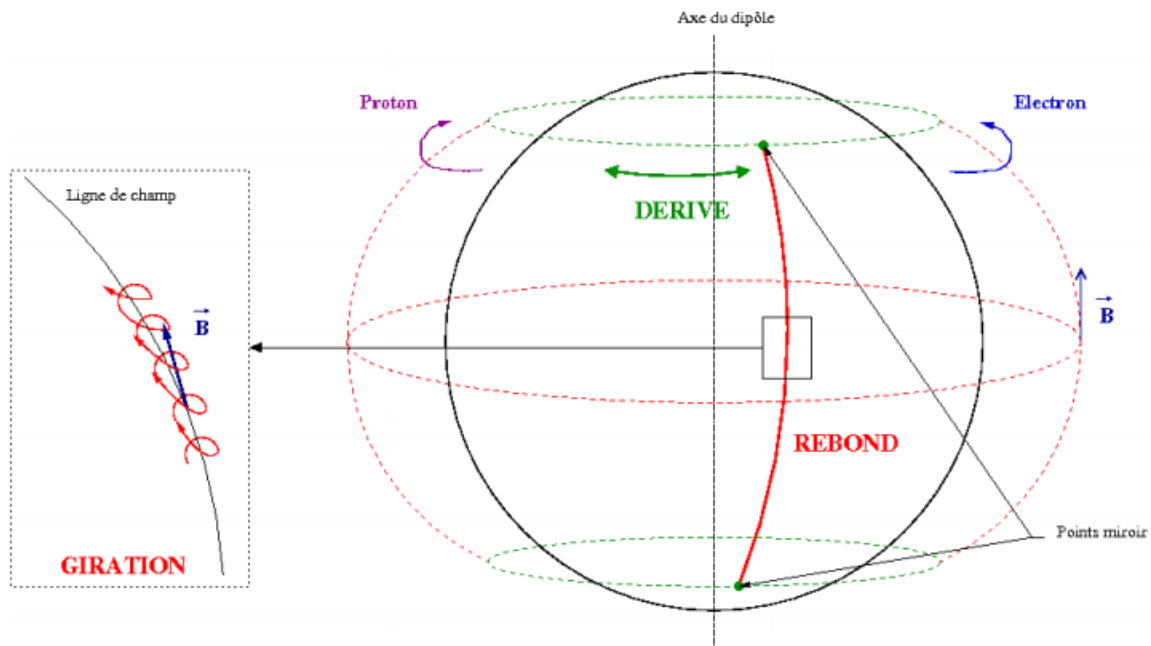


Figure 12 - Coquille de dérive définie par les mouvements de giration, rebond et dérive d'une particule.

La périodicité de chacun de ces mouvements élémentaires dépend de la nature de la particule et de son énergie. Des ordres de grandeur sont donnés dans le Tableau 3 pour des particules à $L=2$. Le paramètre adimensionné L , paramètre de McIlwain représente la distance, également exprimée en rayons terrestres, séparant le centre du dipôle à la ligne de champ à l'équateur magnétique.

Dans cette description, la particule peut être pseudo-piégée, en ce sens que le temps nécessaire pour quitter la sphère d'influence du champ magnétique terrestre est très supérieur au temps caractéristique de son mouvement pseudo-périodique le plus long. Ceci reste vrai tant que l'on peut négliger la contribution des champs électriques dans l'expression de la force de Lorentz.

Tableau 3- Périodes de giration, de rebond et de dérive pour des particules à $L=2$.

	Période cyclotron τ_c [s]	Période de rebond τ_{dd} [s]	Période de dérive τ_d [min]
Electron, 50 keV	1.10^{-5}	3.10^{-1}	460
Electron, 1 MeV	3.10^{-5}	1.10^{-1}	33
Proton, 50 keV	2.10^{-2}	1.10^{+1}	440
Proton, 1 MeV	2.10^{-2}	3	22

2.3.2. Les trois invariants adiabatiques

En considérant que les champs électromagnétiques régissant le mouvement global des particules chargées varient très lentement, les trois mouvements périodiques élémentaires peuvent être décrits, en première approximation, par des constantes de mouvement appelées invariants adiabatiques.

2.3.2.1. Le premier invariant $J1$

Le premier invariant s'exprime comme

$$J1 = \frac{1}{2\pi} \oint (\vec{p} + q\vec{A}) d\vec{l} = \frac{p_{\perp}^2}{2qB} \quad (4)$$

avec \vec{A} le potentiel vecteur magnétique, q sa charge, B le champ magnétique, $d\vec{l} = \sqrt{dr^2 + r^2 d\theta^2}$ l'élément infinitésimal de l'abscisse curviligne avec r et θ les coordonnées du système sphérique, \vec{p} le vecteur moment, et p_{\perp}^2 la composante perpendiculaire du vecteur moment.

Il est relié au moment magnétique relativiste M défini par :

$$M = J1 \frac{|q|}{m_0} = \frac{p_{\perp}^2}{2m_0 B} \quad (5)$$

avec m_0 la masse de la particule au repos

Cet invariant est conservé tant que les variations du champ magnétiques restent plus lentes que la période de giration.

2.3.2.2. Le second invariant $J2$

Le second invariant s'exprime comme

$$J2 = \frac{1}{2\pi} \oint \vec{p} d\vec{l} = \frac{1}{2\pi} \oint p_{\parallel} dl \quad (6)$$

avec p_{\parallel} la composante parallèle au champ magnétique de la quantité de mouvement.

Le deuxième invariant $J2$ peut être interprété comme une rigidité élastique qui définit la longueur de la corde magnétique entre les points miroirs. La position de ces deux points sur la ligne de champ dépend de l'angle d'attaque équatorial α_{eq} formé entre le champ magnétique et le vecteur vitesse de la particule à l'équateur magnétique.

2.3.2.3. Le troisième invariant $J3$

Le troisième invariant $J3$ est relié au flux magnétique Φ contenu dans la coquille de dérive que l'on définit par $\Phi = \int \vec{B} d\vec{S}$ avec $d\vec{S}$ correspondant au vecteur normal à l'élément infinitésimal de surface définie par l'intersection entre la coquille de dérive et le plan équatorial.

$$J3 = \frac{q}{2\pi} \oint \vec{A} d\vec{l} = \frac{q}{2\pi} \Phi \quad (7)$$

avec $d\vec{l}$ l'élément infinitésimal de circonférence équatoriale de la coquille de dérive.

Cet invariant peut être vu comme la contraction de la coquille de dérive afin de conserver le flux magnétique (d'après le théorème d'Ampère) lorsque le champ magnétique augmente à l'intérieur.

Il convient de noter que certaines interactions physiques peuvent entraîner jusqu'à la violation du premier et du second invariants adiabatiques en raison de leurs échelles de temps caractéristiques plus courtes que celles associées à ces invariants. Parmi elles, on peut notamment citer les interactions

coulombiennes et les interactions onde/particule qui provoquent toutes deux une diffusion en angle d'attaque des particules.

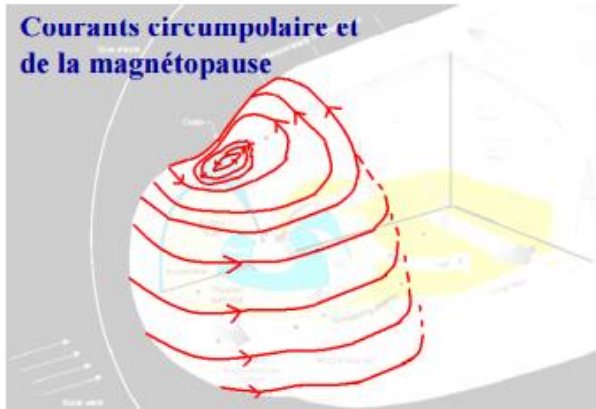
2.4. Les courants magnétosphériques

La magnétosphère est parcourue par un système complexe de courants de plusieurs types : courant de surface, courant annulaire, courants de la queue, courants alignés et électrojets ionosphériques. Ils jouent donc un rôle important dans la dynamique du système vent solaire-magnétosphère-ionosphère étant le siège des processus d'accélération et de conversion d'énergie.

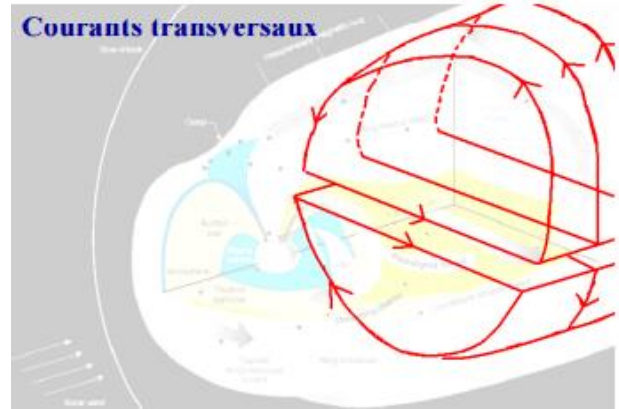
En particulier le courant à travers la queue s'écoulant dans la couche de plasma constitue la clef de voûte de la queue magnétosphérique et c'est la dynamique de ce courant qui permet d'accumuler l'énergie reçue par le vent solaire. C'est aussi en son sein que se développe le processus qui conduit à la dissipation explosive d'énergie lors des sous-orages.

Parmi tous les courants que créent les particules soumises au champ magnétique terrestre, on distingue les courants représentés sur la Figure 13.

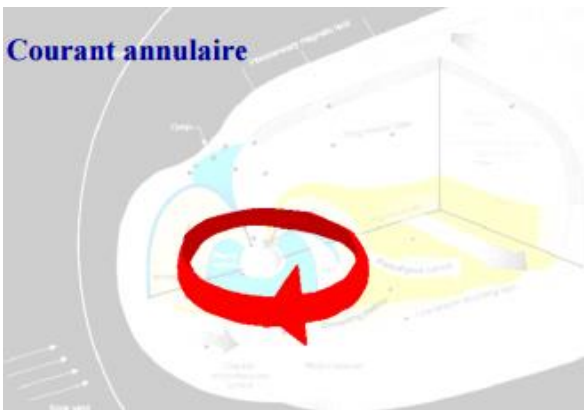
Les courants circumpolaires et de magnétopause représentés sur la Figure 13.a. sont des courants tournant vers l'Est à la surface de la magnétopause par interaction des particules du vent solaire avec le champ géomagnétique. Sur la Figure 13.b., les courants transversaux sont des courants liés à la circulation de particules chargées dans la queue de la magnétosphère, de part et d'autre du feuillet neutre. Ils sont gouvernés par le champ magnétique, dirigé vers la Terre dans le lobe Nord et dans le sens opposé pour le lobe Sud. Ceci implique donc la fermeture de ces boucles de courant dans le feuillet neutre, grâce à des courants perpendiculaires à la coupe transverse de la surface des lobes. Le courant annulaire représenté sur la Figure 13.c. est un courant directement lié à la dérive des particules piégées et aux particules magnétosphériques injectées. Il forme un espace annulaire de rayon intérieur de $2 R_E$ et de rayon extérieur de $6 R_E$ environ, où les électrons cheminent vers l'Est et des protons, vers l'Ouest. Enfin sur la Figure 13.d., les courants alignés sont des courants circulant le long des lignes de champ magnétique connectant la magnétosphère à l'ionosphère.



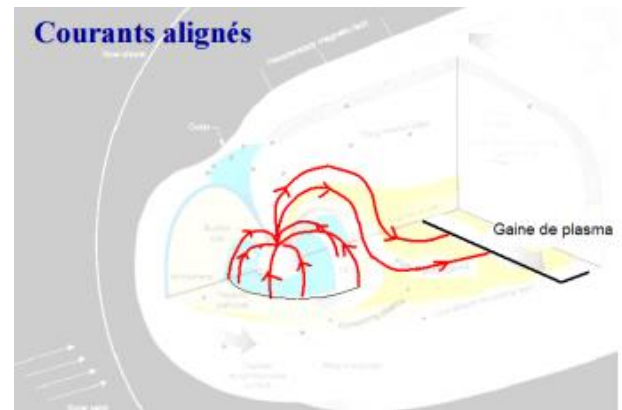
a)



b)



c)



d)

Figure 13- Exemples de courants créés suite à l'interaction Soleil-Terre.

2.5. Les secteurs MLT

Pour mieux décrire la dynamique interne de la magnétosphère, les scientifiques font appel à la notion de Temps Magnétique Local (ou MLT pour Magnetic Local Time). Cette notion permet de déterminer le positionnement d'un événement par rapport au Soleil et est défini en fonction de la longitude et de la latitude magnétique, ainsi que du Temps Universel Local.

La Figure 14 décrit la position des quatre secteurs définis : le méridien magnétique qui est face au Soleil, centré sur MLT=12 et correspond au côté jour. Le côté opposé est le côté nuit, centré sur MLT=0. Sur les flancs, on retrouve le côté matin ou aube centré sur MLT=6 et soir centré sur MLT=18.



Figure 14- Les secteurs MLT.

Dans une étude faite par [Chambodut et al., 2013], des indices dérivés de l'indice magnétique global *am* que nous définissons à la Section 3, les *am* sectoriels, ont montré l'importance d'étudier un événement en fonction du secteur MLT. En analysant un événement grâce à ces indices sectoriels, ils ont ainsi souligné que l'électrodynamique et les phénomènes physiques de type reconnexion côté nuit ou le couplage côté jour, conduisent à des perturbations bien différentes pour chaque secteur considéré.

3. L'ETUDE DE L'INTERACTION SOLEIL-TERRE AU TRAVERS DES INDICES MAGNETIQUES

Comme défini par [Mayaud, 1968] « le but d'un indice est de donner une information résumée de façon continue concernant un phénomène plus ou moins complexe qui varie avec le temps. Chaque indice est fait d'un ensemble de valeurs discrétisées et chacune d'elle caractérise le phénomène considéré sur une certaine plage de temps. » Un indice ne doit pas être trop sophistiqué car il ne substitue pas aux données originales mais a pour but d'en être un résumé. Il doit pouvoir être traçable et reconstruit, à la différence d'un proxy.

Les variations du champ magnétique terrestre peuvent être classées entre les variations séculaires dont les sources sont internes et les variations transitoires dont les sources sont externes et se trouvent dans l'environnement ionisé de la Terre.

Pour mieux comprendre les différents indices utilisés et la signification physique de chacun d'eux, nous allons les présenter au travers des caractéristiques les plus importantes : résolution spatiale, courants associés, position des magnétomètres. Toutes les figures présentées dans cette section, ainsi que les indices décrits sont obtenus grâce à ISGI (<http://isgi.unistra.fr/>).

3.1. Les indices d'électrojets auroraux *AE*, *AU*, *AL*

AE, *AU* et *AL*

Résolution temporelle

1 min

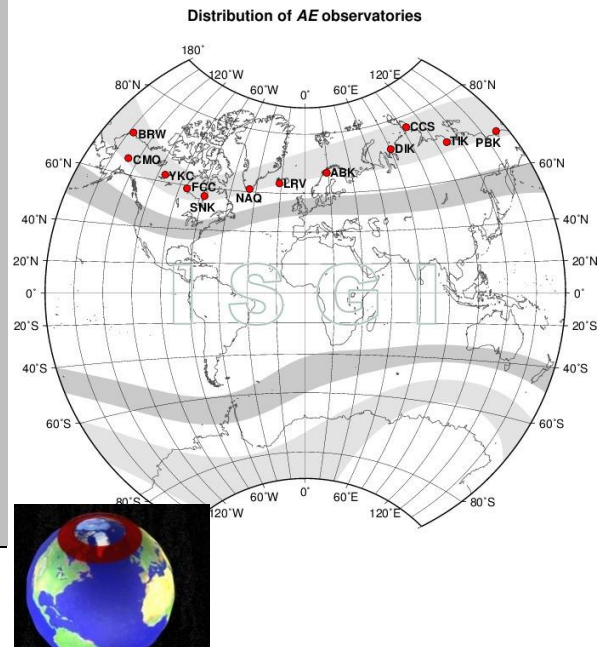
Mesures associées

AE, *AU*, *AL* donnent une estimation des électrojets dirigés vers l'est qui induisent une variation positive de la composante horizontale du champ magnétique (*AU*) ajouté à une mesure de l'électrojet vers l'ouest (*AL*)

$AE = (AU - AL)$, *AU* définissant l'enveloppe supérieure (upper) du magnétogramme de variation de la composante horizontale (*H*), et *AL* l'enveloppe inférieure (lower)

Réseau de stations

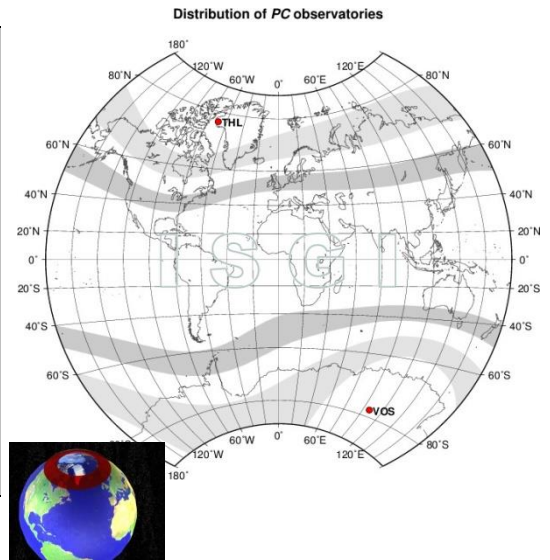
Douze stations situées entre 65 et 70° de latitude



3.2. Les indices d'activité polaire *PCN* et *PCS*

PCN et *PCS*

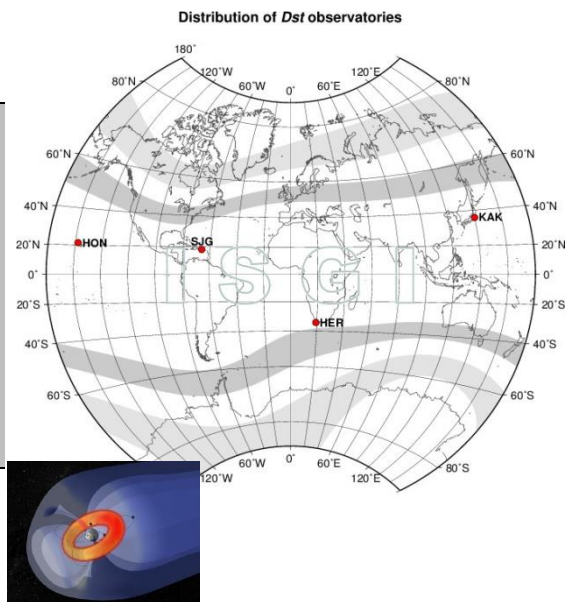
Résolution temporelle	1 min
Mesures associées	<p><i>PCN</i> et <i>PCS</i> permettent de contrôler l'activité géomagnétique au niveau de la calotte polaire, causé par des changements dans l'IMF et dans le vent solaire, conduit par le champ électrique interplanétaire géoeffectif, sans tenir compte du temps, des saisons et du cycle solaire.</p> <p><i>PCN</i> est au nord, <i>PCS</i> au sud</p>
Réseau de stations	Deux stations situées aux pôles



3.3. L'indice d'activité à l'équateur *Dst*

Dst

Résolution temporelle	1 h
Mesures associées	<p><i>Dst</i> donne une estimation horaire de l'intensité du courant annulaire. En temps calme, l'indice est proche de zéro et diminue fortement pendant la phase principale d'un orage. Le <i>Dst</i> est la valeur moyenne de la composante horizontale du champ magnétique.</p>
Réseau de stations	Quatre observatoires situés à des latitudes de 20 à 30° et à différentes longitudes



3.4. Les indices d'activité globaux

Kp et *ap*

Résolution temporelle

3h

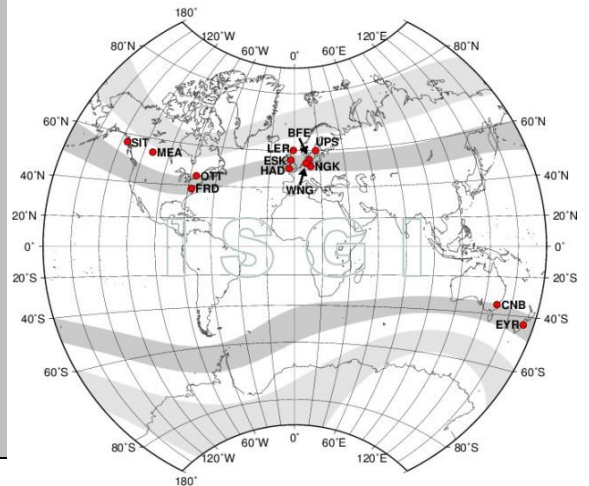
Mesures associées

L'indice *Kp* fournit le niveau de perturbation engendré par la composante horizontale du champ géomagnétique. Il s'agit de la moyenne des perturbations géomagnétiques, ramenée sur une échelle logarithmique entre 0 et 9. On obtient ensuite l'indice *ap* en convertissant l'échelle logarithmique *Kp* en échelle linéaire à l'aide d'une table de conversion.

Réseau de stations

13 stations (dont deux pour l'hémisphère Sud) à des latitudes comprises entre 46° et 63° Nord ou Sud

Distribution of *Kp* observatories



Kpm et *am*

Résolution temporelle

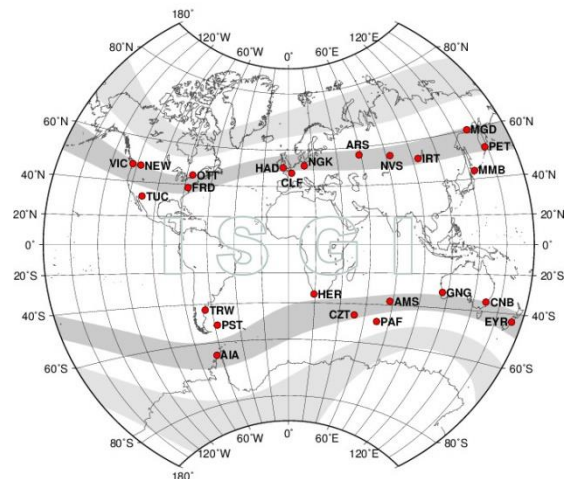
3h

Mesures associées

am fournit le niveau de perturbation engendré par la composante horizontale du champ géomagnétique en utilisant un nombre important de stations aussi bien au nord qu'au sud afin de mieux représenter l'activité magnétique suivant la longitude et les divergences hémisphériques possibles. On fait d'abord une moyenne des mesures au nord (*an*), puis au sud (*as*) afin d'obtenir $am = (an + as) / 2$.

De même que pour les indices *Kp* et *ap*, *Kpm* est le pendant en échelle logarithmique de *am*, que l'on peut obtenir grâce à une table de conversions.

Distribution of *am* observatories



Réseau de stations	23 stations situées autour de 50° de latitude géomagnétique
---------------------------	---

Dérivé de l'indice am , l'indice $a\sigma$ ou am sectoriel, est un indice défini pour chaque secteur MLT présenté à la section 2.5.

4. LA COMPLEXITE DE LA REPOSE DE LA MAGNETOSPHERE A L'ACTIVITE SOLAIRE

Pour mieux évaluer l'impact du vent solaire sur l'environnement terrestre, les scientifiques se sont interrogés sur la mise en place de relations empiriques entre le vent solaire et la magnétosphère. Les différentes analyses présentées dans cette partie montrent la complexité de la réponse de la magnétosphère à l'activité solaire.

4.1. L'impact de la magnétogaine sur les paramètres du vent solaire

Il a été montré qu'au passage de la magnétogaine, la zone située entre le choc et la magnétopause, les propriétés du vent solaire sont fortement modifiées.

Une étude faite par [Coleman, 2005] basée sur les données des satellites Geotail et Interball-Tail en aval du choc, ainsi que sur les mesures du satellite Wind, a montré que la direction du champ magnétique dans le plan perpendiculaire à l'axe Terre-Soleil n'est conservée que 30 % du temps à 10° près. La variation de cet angle reste en général inférieure à 30°, mais elle peut être bien plus importante dans un tiers des cas considérés. Ainsi l'hypothèse utilisée du "drapé parfait", c'est-à-dire que les lignes de champ s'enroulent autour de la magnétopause sans que leur direction dans le plan perpendiculaire à l'axe Terre-Soleil ne soit modifiée, est donc peu fiable.

[Šafránková et al., 2009] ont montré que le signe de B_z était également impacté lors de la traversée de la magnétogaine. Les résultats montrent que si $|B_z|$ est faible ($|B_z| < 1$ nT), la probabilité de retrouver le même signe de B_z de part et d'autre du choc est de l'ordre de 0,5, ce qui revient à une coïncidence fortuite. Cette probabilité augmente toutefois avec $|B_z|$ et est autour de 0,95 pour $|B_z| > 10$ nT, mais de telles valeurs de $|B_z|$ ne sont rencontrées que 5% du temps. Au total, le signe de B_z dans le vent solaire n'est retrouvé dans la magnétogaine que dans 12% des cas. La majeure partie du temps, il est donc difficile de prédire le signe de B_z dans la magnétogaine.

La magnétogaine joue ainsi un rôle perturbateur dans l'interaction Soleil-Terre, et les études sur l'impact de celle-ci ont souligné l'impact sur la complexité de la modélisation de la réponse de la magnétosphère à l'activité solaire. Différentes techniques ont donc été utilisées pour analyser le comportement de la magnétosphère en réponse aux perturbations du vent solaire.

4.2. L'analyse de la réponse de la magnétosphère à l'activité solaire à l'aide des fonctions de couplage

Les fonctions de couplage ont été développées à la fin des années 70 dans le but d'évaluer la dispersion en énergie depuis l'entrée de la magnétosphère au travers des différents systèmes de courant.

La première a été développée par [Perreault and Akasofu, 1978] afin d'analyser les orages géomagnétiques en terme de flux d'énergie, et d'en extraire les paramètres du vent solaire contrôlant

les orages magnétiques et leurs évolutions. Pour ce faire, une analogie entre le flux d'énergie interplanétaire $U(t)$ et le flux de Poynting a été faite suivant l'équation 8. Le flux de Poynting représente le taux de variation par unité de surface de la quantité d'énergie d'un volume défini. Il est également interprété comme la différence entre le flux d'énergie entrant et le flux d'énergie sortant par unité de surface, ou la puissance véhiculée par une onde à travers une surface.

$$U(t) \cong \epsilon(t) = \frac{|E(t)| |B(t)|}{4\pi} (l_0 \sin^2 \frac{\theta'(t)}{2}) \text{ [W]} \quad (8)$$

avec $l_0 \cong 7R_e$ et θ' une mesure de l'angle entre le vecteur IMF et le vecteur du champ magnétosphérique en avant de la magnétosphère dans le plan équatorial. $\epsilon(t)$ représente alors l'énergie sortante, dissipée à travers différents systèmes de courant dans la magnétosphère.

[Perreault and Akasofu, 1978] ont ainsi étudié le taux de dissipation en énergie totale $U(t)$ en termes d'injection de particules dans le courant annulaire $U_i(t)$, de dissipation Joule dans l'ionosphère $U_j(t)$ et d'injections de particules aurorales $U_p(t)$. Suite à cette première étude, ils ont pu constater que la corrélation entre les paramètres géomagnétiques et interplanétaires était correcte, sauf durant les phases de récupération d'un orage.

Plus tard, [Akasofu, 1981] a repris cette étude et a constaté que la magnétosphère ne pouvait répondre pleinement à des flux d'énergie ayant une échelle temporelle inférieure à la constante de temps de la magnétosphère τ_m notamment à cause de l'inductance importante de la magnétosphère et de la résistance ionosphérique. En étudiant des événements extrêmes, il a pu souligner que pour déclencher un orage conséquent, c'est-à-dire engendrer un flux d'énergie important au niveau du courant annulaire, il fallait alors atteindre un niveau d'énergie entrante supérieure à $2,510^{19} \text{ erg.s}^{-1}$.

Suite à ces études, de nombreuses fonctions de couplage sont apparues. Celle de [Vasyliunas et al., 1982] représentée par le système d'équations (9) utilisait une analyse dimensionnelle basée sur la physique de la puissance extraite du vent solaire, en considérant l'importance relative du couplage électromagnétique, les effets de conductivité ionosphérique, et les couplages visqueux au bord de la magnétosphère.

$$P = \rho V^3 I_{CF}^2 F(M_A^2, H, R, \theta)$$

$$M_A^2 = \frac{\mu_0 \rho V^2}{B_T^2}$$

$$R = \frac{V L_{CF}}{\nu} \quad (9)$$

$$H = \mu_0 \sum_p \nu$$

$$I_{CF} = \left(\frac{M_E^2}{\mu_0 \rho V^2} \right)^{1/6}$$

avec ρ la densité du vent solaire, V la vitesse, I_{CF} la distance à la magnétopause définie par [Chapman and Ferraro, 1931], F est une fonction adimensionnelle des rapports adimensionnels (M_A^2, H, R) et de l'IMF clock angle θ . M_A est le nombre de Mach d'Alfvén basé sur la partie transverse du champ magnétique $B_T = (B_Y^2 + B_Z^2)^{1/2}$. H mesure l'importance relative de la conductivité ionosphérique comparée aux effets d'inertie en déterminant la force des courants de Birkeland [Hill and Rassbach, 1975]. R est le nombre de Reynolds, μ_0 est la perméabilité magnétique, ν la viscosité cinématique effective et M_E le moment du dipôle magnétique.

En accord avec cette expression, de nombreuses fonctions de couplage ont été introduites, en considérant différentes bases de données et des méthodes dérivées de celles-ci (voir [Murayama et al., 1982] ; [Bargatze et al., 1986]; [Xu and Shi, 1986]; [Stamper et al., 1999]; [Finch and Lockwood, 2007]). Cependant, la plupart de ces fonctions de couplage en énergie ne donnent pas quantitativement l'entrée en énergie depuis le vent solaire, en raison de coefficients toujours indéterminés.

Récemment, l'étude de [Wang et al., 2014] a proposé une fonction de couplage basée sur une étude en trois dimensions magnétohydrodynamiques. L'équation (10) donnée par cette étude montre l'importance de la vitesse dans l'interaction Soleil-Terre, ainsi que le rôle de l'IMF clock angle θ et fournit une relation qualitative entre l'entrée en énergie et les paramètres du vent solaire.

$$E_{in} = 3,78.10^7 n^{0.24} V^{1.47} B_T^{0.86} \left(\sin^{2.70} \left(\frac{\theta}{2} \right) + 0,25 \right) \text{ [W]} \quad (10)$$

avec E_{in} l'énergie entrante dans la magnétopause. En comparant les résultats fournis par cette équation et ceux obtenus avec l'équation de [Akasofu, 1981], [Wang et al., 2014] ont constaté qu'ils amélioreraient l'évaluation des pics d'activité d'un facteur 4 à 5, notamment en prenant en compte l'énergie dissipée par le chauffage de la plasmashet et l'écoulement plasmöide. Cette équation est similaire à celle fournie par [Newell et al., 2007] dans leur étude ayant eu pour but de fournir une fonction de couplage quasi-universelle en utilisant aussi bien des indices magnétiques que des données GOES ou SuperDARN, corrélées avec une vingtaine de paramètres du vent solaire. Cette étude a abouti à l'équation (11) qui comme l'équation (10), montre l'importance de la vitesse du vent solaire, ainsi que de la densité, et des paramètres associés aux champs magnétiques

$$d\Phi_{MP}/dt = V^{4/3} B_T^{2/3} \sin^{8/3}(\theta c/2) \quad (11)$$

avec $d\Phi_{MP}/dt$ le taux de variation du flux magnétique à la magnétopause.

Ces méthodes ont permis de mettre en relation les paramètres du vent solaire et les indices magnétiques mesurés au sol. D'autres techniques que nous présentons dans les sections suivantes, ont eu pour but d'étudier plus spécifiquement le comportement de la magnétosphère en réponse à la dynamique du vent solaire.

4.3. La magnétosphère : un filtre non linéaire

[Bargatze et al., 1985] ont travaillé sur des modèles de prédiction de l'indice AL associé à l'activité aurorale en utilisant une technique de filtrage linéaire. Ce filtrage était utilisé afin de modéliser la réponse de la magnétosphère à partir des paramètres du vent solaire. Dans cette analyse, les paramètres du vent solaire considérés étaient le produit VB_z . Cette analyse a souligné la complexité de la réponse de la magnétosphère, qui peut être associée à deux interprétations. La première est celle fournie par [Perreault and Akasofu, 1978] basée sur un modèle directement contrôlé par l'activité solaire, que nous avons décrit dans la section 4.2. La seconde provient de l'étude de [McPherron et al., 1970], qui ont montré que la magnétosphère se comporte comme une capacité qui se charge et se décharge en fonction de l'énergie accumulée.

Etant donné qu'aucune observation ne permettait de trancher vers un modèle ou l'autre, [Bargatze et al., 1985] ont alors utilisé le filtrage linéaire pour obtenir la relation linéaire généralisée entre des informations provenant du vent solaire et celles provenant de la magnétosphère. Cette étude a confirmé qu'on obtenait deux catégories de réponse. Une réponse rapide associée à des activités élevées, pour lesquelles l'activité magnétosphérique était directement liée au couplage avec le vent solaire, et une réponse plus lente associée cette fois à des mécanismes internes à la magnétosphère.

Ces analyses ont avant tout souligné le fait que la magnétosphère reste un système complexe à modéliser. Si les relations empiriques fournies par des analyses diverses (corrélation paramètres du vent solaire – indices magnétiques, magnétohydrodynamiques, flux de Poynting) permettent de souligner l'importance de certains paramètres du vent solaire et de mieux comprendre l'impact de l'activité solaire sur différents systèmes de courant magnétosphériques, la non-linéarité de la réponse de la magnétosphère à l'activité solaire est un élément clef de l'analyse de l'interaction Soleil-Terre. Pour étudier cette non-linéarité, les scientifiques se sont alors tournés vers des modèles prenant en compte cet aspect afin de fournir de nouveaux éléments de réponses dans la compréhension de cette interaction. Ceci a également été fait dans le but d'obtenir des outils opérationnels de prédiction de l'impact de l'activité solaire sur notre environnement magnétique, et donc sur nos technologies.

5. L' UTILISATION DES RESEAUX DE NEURONES POUR ETABLIR LE LIEN ENTRE LE VENT SOLAIRE ET LES INDICES MAGNETIQUES

Au cours des deux dernières décennies, nous avons pu constater un développement fulgurant des réseaux de neurones, et de leurs applications à la météorologie de l'espace. Les réseaux de neurones sont des modèles statistiques, inspirés des réseaux neuronaux humains. Nous décrivons leur fonctionnement et leur développement en détail dans le Chapitre 2.

Cet intérêt a démarré avec l'application réussie de cette technique puissante pour des problématiques très différentes, et dans des domaines aussi divers que la finance, la médecine, la production industrielle, la géologie ou encore la physique.

La possibilité d'apprendre sur la base d'exemples constitue l'une des nombreuses fonctionnalités des réseaux de neurones qui permettent à l'utilisateur de modéliser ses données et établir des règles précises qui vont guider les relations sous-jacentes entre différents attributs des données. L'utilisateur des réseaux de neurones collecte des données représentatives puis fait appel aux algorithmes d'apprentissage, qui vont apprendre automatiquement la structure des données.

Ici nous présentons les différents réseaux de neurones utilisés dans de précédentes études, montrant ainsi la capacité des réseaux de neurones à établir le lien entre vent solaire et indices magnétiques.

5.1. Les premiers réseaux utilisés en météorologie de l'espace

En météorologie de l'espace, les premiers réseaux utilisés ont été les « feedforward backpropagation ». Ces réseaux sont les plus simples à mettre en œuvre et à interpréter car le modèle est statique, allant uniquement de l'entrée vers la sortie. Son fonctionnement est décrit en détail dans le Chapitre 2.

[Lundstedt and Wintoft, 1994] ont fait appel à ces réseaux pour prédire les orages géomagnétiques avec l'indice *Dst*. Ils ont montré qu'avec ce modèle il était possible de prédire correctement les phases initiales et principales d'un orage, mais qu'il était plus complexe de prédire la phase de recouvrement.

[Gleisner and Lunsdtedt, 1997] ont utilisé ce modèle pour prédire la réponse des électrojets auroraux à l'activité solaire. Ils comparent notamment dans cette étude la différence sur les performances de prédiction en utilisant d'une part des fonctions d'activation linéaires au sein du réseau, et d'autre part des fonctions d'activation non linéaire. Ils ont ainsi montré l'apport des modèles non linéaires en obtenant des coefficients de corrélation supérieurs.

[Boberg et al., 2000] ont développé ce réseau afin de fournir des prédictions en temps réel de l'indice *Kp*. Plus précisément, un modèle hybride a été développé, afin d'avoir un réseau spécifique aux prédictions de *Kp* en période calme, et un réseau spécifique aux périodes agitées. En effet, il est complexe d'atteindre des performances optimales de prédiction avec ce type de réseau pour tous les niveaux d'activité, la dynamique de la magnétosphère étant hautement variable en fonction de ceux-ci.

Une première approche de la prédiction de l'indice magnétique *am* a été faite dans le cadre du projet ATMOP par [Mazouz et al., 2013] , comme défini en avant-propos. Le réseau feedforward backpropagation a alors été utilisé afin d'effectuer des prédictions à une heure et à trois heures de l'indice magnétique *am*.

Si le réseau « feedforward » permet de fournir une première approche des relations entre vent solaire et indices magnétiques, des modèles plus complexes ont été développés par la suite et appliqués à la

météorologie de l'espace pour fournir des prédictions plus précises de la réponse de la magnétosphère.

5.2. Vers des réseaux plus complexes pour modéliser la dynamique magnétosphérique

Les premières études fournies par [Lundstedt and Wintoft, 1994] ont souligné les faiblesses du réseau feedforward. Par la suite, des réseaux temporels comme le « Time Delay Neural Network » (TDNN) ont été utilisés. Ces réseaux sont basés sur des données du vent solaire décalées dans le temps par des processus internes au réseau, que nous décrivons dans le Chapitre 2. [Gleisner et al., 1996] ont fait appel à ce réseau pour prédire l'indice Dst et ont montré en comparaison à l'étude faite par [Lundstedt and Wintoft, 1994] qu'il était plus performant, en prédisant plus fidèlement la phase de recouvrement d'un orage.

Par la suite, des réseaux appartenant à une toute autre famille ont été utilisés en météorologie de l'espace. Il s'agit des réseaux récurrents. Ce sont des réseaux au sein desquels les neurones interagissent les uns avec les autres. Nous les décrivons plus en détail dans le Chapitre 2. Ils ont notamment été utilisés pour prédire l'activité au niveau du courant annulaire, courant spécifique aux orages magnétiques par [Wu and Lundstedt, 1996] et ont permis dans cette étude de mettre en place des fonctions basées sur les paramètres du vent solaire, utilisées par la suite en entrée de ce réseau pour fournir des prédictions en temps réel à partir du satellite WIND, situé au point de Lagrange L1.

[Wing et al. 2005] ont développé trois modèles basés sur le fonctionnement de ces réseaux récurrents afin de prédire Kp . Ces modèles opérationnels et disponibles sur le site de la NOAA (<https://www.swpc.noaa.gov/products/wing-kp>) sont basés sur les données fournies par le satellite ACE situé au point de Lagrange L1. Si ces modèles fournissent des prédictions en temps réel de Kp , ils ont surtout souligné encore une fois la complexité de l'étude de la réponse de la magnétosphère à cause de sa dynamique interne qui va dominer les effets des événements externes à la magnétosphère, ou être dominée par ceux-ci.

Plus récemment, un réseau récurrent spécifique a été utilisé en météorologie de l'espace, le réseau « Non linear AutoRegressive with eXogenous inputs » ou NARX. A l'origine, ce modèle est un modèle paramétrique polynomial. Par la suite, les mathématiciens ont intégré des fonctions sigmoïdes et introduits une non-linéarité. Ce réseau est basé sur une partie autorégressive ainsi que sur un historique des données exogènes (toutes les données autres que l'indice prédit). Il propose un modèle complexe à analyser mais fournissant des performances optimales de prédictions. [Cai et al., 2009] ont utilisé ce modèle pour prédire l'indice SYM-H spécifique du courant annulaire et ont montré grâce au fonctionnement du NARX, qu'en plus de l'impact direct du vent solaire sur l'évolution du courant associé aux orages magnétiques, l'état du courant en lui-même joue un rôle important sur son évolution, notamment dans la phase de recouvrement. [Bhaskar and Vichare, 2017] ont développé le NARX pour prédire SYM-H et ASY-H et ont confirmé les points soulignés par [Cai et al., 2009] suggérant l'importance de facteurs internes associés aux processus magnétosphériques. Dernièrement [Ayala Solares et al., 2016] ont utilisé le NARX pour prédire l'indice Kp , en comparant l'utilisation d'une fenêtre de données glissante et une approche directe. Dans les deux cas, ce modèle certes performant a montré que les prédictions obtenues étaient légèrement biaisées à cause de la partie autorégressive de celui-ci.

6. BILAN SUR L'ETAT DE L'ART

Cet état de l'art depuis l'origine du vent solaire jusqu'aux mesures faites au sol par les magnétomètres a permis de faire la photographie des connaissances existantes en amont du travail présenté dans ce manuscrit. Au travers de cette présentation des connaissances, différents éléments clefs au cœur des études passées ont été soulignés.

Tout d'abord, ces études ont montré qu'il existait un lien complexe entre la dynamique du vent solaire et les mesures faites au sol par les magnétomètres. En fonction de l'activité solaire, la réponse de la magnétosphère varie aussi bien sur une échelle temporelle que sur une échelle spatiale.

Les analyses physiques, orientées vers la mise en relation de paramètres du vent solaire avec les indices magnétiques ont souligné que certains paramètres du vent solaire jouaient un rôle prépondérant sur la dynamique magnétosphérique, notamment sa vitesse, sa densité ainsi que les composantes du champ magnétique.

Les analyses mathématiques faites par les réseaux de neurones permettent de fournir des outils opérationnels de prédiction, tout en confirmant la prépondérance des facteurs cités précédemment.

CHAPITRE II

MATÉRIELS ET MÉTHODES : MODÈLES ET DONNÉES POUR LA PRÉDICTION D'INDICES MAGNÉTIQUES

Dans cette partie Matériels et méthodes, nous présentons l'ensemble des données ainsi que les techniques utilisées. Le cœur des études effectuées par la suite étant la prédiction à court et long terme d'indices magnétiques, nous détaillons la notion de données temporelles et les précautions à prendre dans le cas de l'analyse de séries temporelles. Nous expliquons le traitement des données considérées effectué en amont dû à la complexité de celles-ci. Nous détaillons les méthodes utilisées pour effectuer des prévisions de séries temporelles. Ce sont des méthodes appartenant au domaine de l'apprentissage automatique supervisé, plus spécifiquement les réseaux de neurones et les processus gaussiens. Enfin, étant donné qu'il est important d'évaluer la qualité d'un modèle de prédiction, ainsi que sa précision et sa fiabilité, nous montrons les différentes métriques utilisées permettant cette évaluation.

1. Les données utilisées et les séries temporelles	67
1.1. La nature des séries temporelles	67
1.1.1. Données discrètes et continues	67
1.1.2. Objectifs de l'analyse des séries temporelles	67
1.2. Données utilisées pour étudier la relation entre vent solaire et les indices magnétiques (<i>am</i> ou <i>Dst</i>)	68
1.3. Préparation des données	69
2. Les prévisions de séries temporelles	70
2.1. Variables dépendantes et indépendantes	71
2.2. Variables continues et catégorielles	71
2.3. Régression versus Classification	71
3. Les modèles de prévision.....	72
3.1. Les réseaux de neurones.....	72
3.1.1. Eléments de base des réseaux de neurones.....	72
3.1.2. Topologie des réseaux de neurones utilisés dans le cadre de notre étude	78
3.2. Les processus gaussiens	87
3.2.1. Le théorème de Bayes	88

3.2.2.	L'inférence Bayésienne	88
3.2.3.	La régression au moyen des processus gaussiens.....	90
3.2.4.	Prédire à partir des processus gaussiens	92
3.3.	Les méthodes d'évaluation des performances d'un modèle de prédiction	93
3.3.1.	L'erreur quadratique moyenne	93
3.3.2.	Le coefficient de corrélation.....	93
3.3.3.	La matrice de confusion	94
4.	Bilan sur Matériels et méthodes	96

1. LES DONNEES UTILISEES ET LES SERIES TEMPORELLES

De nos jours, de nombreuses données définies sous formes de séries temporelles sont disponibles, dans un panel de domaines vaste allant des données médicales (e.g. la quantité de glucose dans le sang d'un patient [Deutsch et al, 1994]), aux biostatistiques (e.g. l'évolution d'une population [Bjørnstad and Grenfell, 2001]), à l'économie et à la finance (e.g. le taux de chômage, les indices boursiers [Franses, 1998]) en passant par le domaine clef de ce manuscrit, la météorologie de l'espace.

Le but de cette partie est de faire un point sur la définition d'une série temporelle, les données utilisées pour les études présentées dans ce manuscrit, ainsi que la préparation de ces données en amont de leurs utilisations pour en faire de la prédiction.

1.1. La nature des séries temporelles

Les séries temporelles font référence à un ensemble de points de données, s'étendant de façon séquentielle sur une période de temps. Ce sont des séries de valeurs ordonnées dans le temps. Les données possèdent alors un ordre naturel. C'est-à-dire que si on considère deux valeurs extraites d'un ensemble de données, il est possible de définir celle qui est arrivée avant l'autre.

1.1.1. Données discrètes et continues

Le mot « donnée » est un terme général pour un ensemble de variables et peut être catégorisé en donnée discrète ou donnée continue. Les données discrètes ne peuvent être définies que suivant des valeurs exactes spécifiques. Une variable, définie par des données discrètes, ne peut aller que d'une valeur spécifique à une autre. Les données continues consistent en des variables qui peuvent occuper n'importe quelle valeur sur un certain domaine défini. Elles ne sont pas restreintes suivant des valeurs spécifiques. Entre deux points, il peut y avoir un nombre infini de points de données.

Une série temporelle est considérée comme discrète quand les données sont échantillonnées suite à un processus, qui peut être lui-même discret ou continu, sur un certain intervalle de temps [Brown, 2004]. Dans ce contexte, échantillonner signifie approximer la réalité avec un sous-ensemble de données, mesurées à partir de procédés [Godambe, 1966]. Afin d'effectuer une analyse statistique, un ensemble fixé de variables (i.e. de données) est requis. Dans le cas de notre étude, les données du vent solaire et les indices magnétiques fournis chaque heure sont une représentation discrète d'un processus continu, l'interaction Soleil-Terre.

1.1.2. Objectifs de l'analyse des séries temporelles

La disponibilité des séries temporelles de données permet d'effectuer différentes formes d'analyses. Suivant [Chatfield, 2000], il existe quatre principaux objectifs associés à l'analyse de séries temporelles :

- Décrire les données en utilisant des graphiques et des mesures statistiques. C'est une première étape importante dans la plupart des analyses, notamment les tracés de séries temporelles pour analyser un événement.
- Trouver un modèle statistique décrivant le processus générant les données par approximation.
- Prédire, c'est-à-dire l'estimation des valeurs de futures observations. Nous précisons que prédictions et prévisions sont souvent utilisées de façon interchangeable dans la littérature. La prédiction sera le sujet principal de ce manuscrit.

- Contrôler et anticiper un processus en utilisant les prédictions de futures valeurs, de façon à prendre la main sur l'issue d'un processus : il s'agit souvent du but final de l'analyse de séries temporelles.

Dans les études décrites ici, le but principal est de fournir un modèle de prédiction qui permet d'anticiper de façon optimale l'impact de l'activité solaire sur l'environnement terrestre. Pour notre étude, nous nous sommes focalisés principalement sur le développement de modèles pour la prédiction de l'indice magnétique global *am*, à partir des paramètres du vent solaire. Dans un second temps, nous avons travaillé sur le développement de modèles pour la prédiction de l'indice magnétique *Dst*, caractéristique des orages et sous-orages magnétiques. Pour développer ces modèles, il est important de définir clairement les données utilisées pour ces études, ainsi que de les prétraiter. Dans la partie suivante, nous décrivons précisément ces éléments.

1.2. Données utilisées pour étudier la relation entre vent solaire et les indices magnétiques (*am* ou *Dst*)

Afin de développer des modèles capables de prédire l'indice magnétique *am* ou *Dst*, il est nécessaire dans un premier temps de définir les données d'entrées et de sorties de ces modèles. La Figure 15 présente les couvertures temporelles des différentes données considérées pour les modèles de prévision.

- *Les paramètres du vent solaire* : les données du vent solaire proviennent de deux bases de données. En premier lieu, nous avons considéré les données fournies par OMNI qui proviennent du National Space Science Data Center (NSSDC) de la NASA (<https://omniweb.gsfc.nasa.gov/ow.html>). Elles sont situées au niveau de l'onde de choc de la magnétosphère. Comme le montre la Figure 15, les données sont considérées entre le 1^{er} janvier 1995 et le 31 décembre 2012 pour les études faites sur la prédiction de l'indice magnétique *am*. Ensuite, nous avons utilisé les données fournies par le satellite Advanced Composition Explorer (ACE) [Stone et al., 1998]. Elles proviennent du ACE Science Center situé au Space Radiation Lab à Caltech (<http://www.srl.caltech.edu/ACE/ASC/level2/index.html>). Ces données sont localisées au niveau du point de Lagrange 1, en amont de la magnétopause. Ce sont des données de niveau 2, c'est-à-dire qu'elles ont été jugées adaptées pour des études scientifiques par l'équipe SPDF. Ces données ne sont disponibles qu'à partir du 5 février 1998, et nous les considérons jusqu'au 31 décembre 2012.

Les données du vent solaire considérées dans les études faites à la suite seront précisées (OMNI ou ACE).

- *Les données GPS* : Les données GPS sont obtenues à partir du site de la NOAA (<https://www.ngdc.noaa.gov/stp/space-weather/satellite-data/satellite-systems/gps/>). Ces données sont fournies par l'équipe travaillant sur le Combined X-ray dosimeter ou CXD du Los Alamos National Laboratory. Sur ce site est disponible depuis peu l'ensemble des données GPS, pour chaque satellite GPS. Pour notre étude, nous considérons le champ magnétique mesuré par le GPS ns41, car c'est celui qui possède la couverture temporelle la plus étendue [Morley et al. 2017].

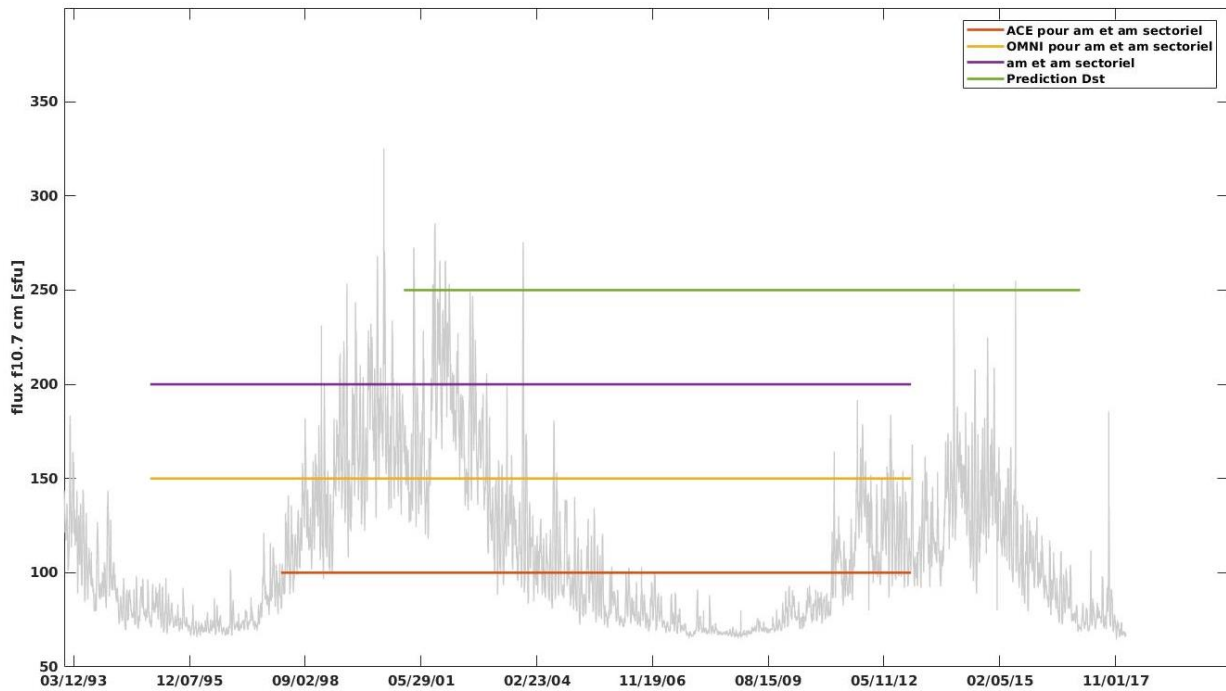


Figure 15 - Couvertures temporelles associées aux données considérées pour la prédiction de l'indice magnétique *am* et de l'indice *Dst*.

- *Les indices magnétiques :*

L'indice magnétique *am*, ainsi que les indices *am* sectoriels ou $a\sigma$ sont obtenus à partir de la base de l'International Service of Geomagnetic Index (ISGI) situé à l'Ecole des Sciences de la Terre (EOST), (<http://isgi.unistra.fr/>). Ces indices sont étudiés entre le 1^{er} janvier 1995 et le 31 décembre 2012 pour couvrir le cycle 23 comme le montre la Figure 15. L'indice magnétique *Dst* ainsi que les paramètres du vent solaire utilisés dans cette étude proviennent de la base OMNI décrites à la section 1.2. Ces données sont considérées entre le 1^{er} janvier 2001 et le 31 décembre 2016.

1.3. Préparation des données

- Interpolation de l'indice magnétique *am*

L'indice magnétique *am* est un indice défini sur une échelle temporelle de trois heures. Or, nous souhaitons prédire cet indice à une heure, nous effectuons alors une interpolation par spline à partir des données définies d'observations à trois heures pour obtenir des données d'observations à une heure. Ainsi, les prédictions faites à une heure sont faites à partir de données définies à une heure. L'interpolation par spline cubique est uniforme et définie par des polynômes de degré 3. Un polynôme de degré 3 s'écrivant $P(t) = a + bt + ct^2 + dt^3$, avec quatre équations (contraintes de continuité à chaque jonction pour la fonction, et pour ses dérivées premières et secondes) pour définir de manière unique les quatre coefficients (a,b,c,d) en résolvant le système linéaire de taille 4x4. [Andriambahoaka, Z., 2008] a montré que l'utilisation de splines pour combler les données manquantes dans le cadre d'analyses d'indices magnétiques était adaptée.

- Traitement des données manquantes

La question du traitement des données manquantes lors de l'analyse de séries temporelles est cruciale et se pose à deux niveaux. Le premier niveau concerne les données manquantes dans la base de données existantes. Si on considère les données du vent solaire, les données fournies par la base ACE et par la base OMNI présentent des trous de données, notamment en cas d'activités élevées. En effet, lors d'un événement solaire extrême, les détecteurs à bord des satellites sont saturés, et il n'est alors plus possible d'obtenir de mesures. Si les scientifiques à l'origine de la base OMNI tentent de palier la perte de ces données en combinant les données obtenues sur différents satellites, il n'empêche que des données manquent. Nous effectuons alors une interpolation par splines afin de conserver une base de données suffisante notamment à activités élevées pour entraîner au mieux nos modèles. Le deuxième niveau quant à lui concerne les données manquantes en temps réel. Cette question a été longuement posée lors du Workshop « Space Weather, a multidisciplinary approach » ayant eu lieu en Septembre 2017. Les modèles de prévisions ont besoin d'être alimentés en continu, et en temps réel, il n'est pas possible d'interpoler. Nous avons donc décidé de traiter ces données comme une fonction spécifique en fonction du langage de programmation considéré. Les langages considérés pour nos études sont spécifiés en annexe 2, et la technique utilisée est décrite lors des explications du développement des modèles aux Chapitres 3 et 4.

2. LES PREVISIONS DE SERIES TEMPORELLES

Etant donné un ensemble d'observations passées d'une certaine variable, la prédiction de série temporelle est utilisée pour développer des modèles permettant de décrire les relations sous-jacentes entre ces observations. Ces modèles doivent approximer la vraie fonction sous-jacente générant les données le plus fidèlement possible. Ils peuvent ensuite être utilisés pour extrapoler des séries dans le futur, en faisant des prédictions de valeurs futures d'une variable. L'extrapolation est un processus permettant d'estimer la valeur d'une variable, depuis les observations initiales, à partir de la relation obtenue grâce aux données observées [Armstrong, 1984].

Le processus permettant d'estimer une relation ou une fonction f à partir de données d'entrées X de façon à fournir une évaluation de la variable de sortie Y est appelé apprentissage statistique ou apprentissage automatique [James et al. 2013]. Cette relation peut être décrite par l'équation (12).

$$\hat{Y} = \hat{f}(X) \quad (12)$$

\hat{f} représentant une estimation de la fonction f , \hat{Y} le résultat de la prédiction associée à X .

La vraie relation entre X et Y peut être décrite par la relation (13)

$$Y = \hat{f}(X) + \epsilon \quad (13)$$

avec le terme d'erreur ϵ ou le bruit venant du processus tend à s'approcher de zéro, étant donné que l'erreur statistique n'est pas prédictible à partir des données.

Ce problème est un problème de prédiction de séries temporelles car une donnée passée est utilisée pour faire des prédictions futures. Une représentation mathématique est définie par l'équation (14), où le temps t joue un rôle central dans la prédiction

$$\hat{Y}_t = \hat{f}(X, t) \quad (14)$$

\hat{f} représente le modèle qui sera entraîné à partir des données, puis qui sera utilisé pour fournir des prédictions. Les fonctionnalités exactes de ce modèle dépendent du choix de l'algorithme. Ce choix est lié au contexte et aux données. Le théorème de Wolpert sur le « no free lunch theorem » statue qu'il n'existe pas d'algorithme meilleur qu'un autre sur l'ensemble des données [Wolpert and Macready, 1997].

2.1. Variables dépendantes et indépendantes

Les séries temporelles utilisées pour prédire le futur consistent en des variables dépendantes et indépendantes. Les variables indépendantes, appelées aussi prédicteurs ou « features », sont des variables autonomes qui ont des propriétés individuellement déterminées du phénomène observé. La variable dépendante, appelée également cible ou réponse (ou « target »), est ce qu'on souhaite prévoir ou prédire. Elle répond aux variables indépendantes. La tâche centrale de l'analyse prédictive est de comprendre et modéliser la relation entre les variables dépendantes et indépendantes. En d'autres mots, on souhaite approximer la fonction exacte qui décrit les données. Le concept central des séries temporelles est d'ailleurs le fait que le temps est la variable indépendante la plus importante.

2.2. Variables continues et catégorielles

Les données consistent généralement en des variables catégorielles et continues. Les variables qualitatives ou catégorielles sont des « features » qui prennent un nombre limité de valeurs, aussi appelées catégories ou niveaux. Elles ne représentent pas une quantité, par exemple un code postal ou un numéro de téléphone. Les variables continues ou quantitatives peuvent quant à elles prendre un nombre infini de valeurs numériques.

Ces concepts ne doivent pas être confondus avec les données discrètes et continues. Bien que les deux soient souvent utilisées de façon interchangeable, ils ne signifient pas exactement la même chose. Il est vrai qu'une donnée consiste en des variables continues ou discrètes. Cependant, cela ne définit pas que la donnée soit discrète ou continue. Les données proviennent de mesures provenant d'un processus sous-jacent qui peut être discret ou continu [Brown, 2004]. Les données discrètes peuvent par exemples provenir d'un échantillonnage périodique d'un processus continu. Il peut également être aussi inhérent à un processus discret. Par exemple, les données que nous utilisons en provenance du vent solaire sont des données rendues discrètes car elles sont échantillonnées sur des intervalles de temps fixes. Cependant, les variables considérées sont continues. Les données GPS sont également des données discrètes inhérentes à un processus continu, comme la variable cible que représente l'ensemble des données géomagnétiques. Le temps lui est une variable continue. Les données consistent donc en des variables continues, discrétisées sur des intervalles de temps fixes.

2.3. Régression versus Classification

En analyse statistique, la distinction est à faire entre régression ou bien interpolation et classification, en fonction du type de la variable de réponse (la cible). Les problèmes statistiques avec des données quantitatives sont souvent des problèmes de régression, tandis que ceux associés à des variables qualitatives sont des problèmes de classification [James et al. 2013]. Les méthodes d'apprentissage

statistique sont généralement choisies en fonction du type de problème et du type de la variable à prédire. Si les variables sont qualitatives ou quantitatives, les algorithmes à considérer sont différents.

L'analyse de série temporelle est souvent vue comme un problème de régression, où les valeurs de la variable cible continue sont approximées au fil du temps. Cependant, il peut également être un problème de classification où à chaque période de temps une variable est assignée à une catégorie spécifique. Par exemple, dans le cadre de prédictions d'indices magnétiques, on peut associer une variable à un niveau d'activité, calme ou agité.

Dans ce manuscrit, nous nous focalisons sur un problème de régression. Le but est d'effectuer une régression des indices magnétiques à partir des variables mesurées dans le vent solaire ou au niveau de l'orbite des satellites GPS en fonction de l'indice magnétique que l'on souhaite prédire. Les caractéristiques associées aux indices magnétiques ou « features », $X = (X_1, \dots, X_i)$ dites prédictives ont été observées sur un ensemble de n instances avec $d_1^n = \{(x_1, y_1), \dots, (x_n, y_n)\}$ avec $x_i \in \mathbb{R}^d$ s'il y a d paramètres en entrée.

3. LES MODELES DE PREVISION

Depuis quelques dizaines d'années, de nombreuses études ont été faites pour faire de la prévision de séries temporelles, tester ces modèles, découvrir de nouvelles propriétés et mettre de nouvelles applications en pratique. Souvent, si certains modèles offrent des performances prometteuses dans certains domaines d'applications, on s'attend à ce qu'il en soit de même lorsque ces modèles sont appliqués dans un autre domaine. Les modèles de prévision que nous présentons dans les sections suivantes ont initialement été développés pour des applications bien différentes de la météorologie de l'espace, mais les chercheurs s'y sont intéressés car les propriétés de ces modèles étaient potentiellement en accord avec les propriétés sous-jacentes à l'interaction Soleil-Terre.

3.1. Les réseaux de neurones

3.1.1. *Eléments de base des réseaux de neurones*

3.1.1.1. A l'origine, le fonctionnement de notre cerveau

Depuis longtemps les scientifiques sont inspirés par le cerveau humain. En 1943, Warren McCulloch et Walter Pitts ont développé le premier concept de réseau de neurones « une cellule seule vivant dans un réseau de cellules, recevant des entrées, calculant et générant des sorties en fonction de ces entrées. » (voir [McCulloch and Pitts, 1943]).

On parle de neuroscience computationnelle, la branche des neurosciences qui s'intéresse aux modèles mathématiques de la cognition, et donc en premier lieu à ceux de l'information mentale. On se base sur le neurone biologique comme défini sur la Figure 16 .

Les notions utilisées sont la plupart du temps assimilées à ces sciences :

- synapse : point de connexion avec les autres neurones,
- dendrites : entrées du neurone,
- axone : sortie du neurone,
- noyau qui active la sortie en fonction des stimulations d'entrée.

Un neurone ne va émettre une impulsion que si le signal transmis au corps cellulaire par les dendrites dépasse un certain seuil appelé seuil de déclenchement.

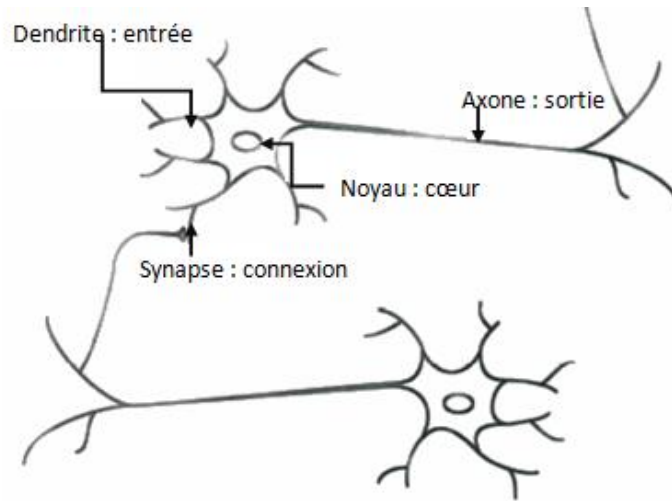


Figure 16- Le neurone biologique.

Afin de reproduire ce fonctionnement biologique au moyen de systèmes informatiques purement artificiels, il est nécessaire de transcrire les différents éléments cités précédemment sous formes de fonctions mathématiques.

Un réseau neuronal est l'association en un graphe plus ou moins complexe d'objets élémentaires : le neurone formel représenté sur la Figure 17.

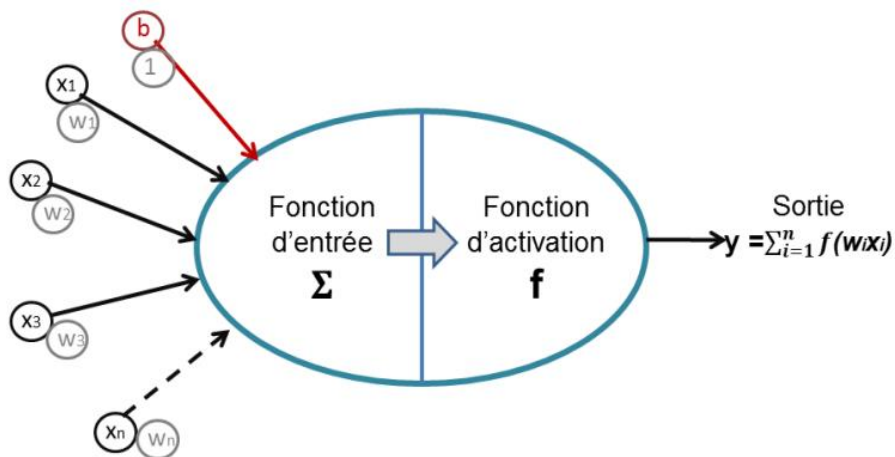


Figure 17- Le neurone formel.

Sur la Figure 17, les entrées X_1, \dots, X_n arrivent au neurone par l'intermédiaire d'une connexion avec une certaine force (respectivement W_1, \dots, W_n) connue sous le nom de poids. Plus la valeur d'un poids W_i est importante, plus l'intensité du signal entrant est forte, et donc, plus l'entrée correspondante est influente. Le biais représenté par le symbole b sur la Figure 17 est une constante et représente le seuil du neurone. Une fois que les signaux sont reçus, il faut effectuer la somme pondérée des entrées. Cela constitue l'entrée de la fonction d'activation f du neurone. L'activation du

neurone est une fonction mathématique qui convertit la somme pondérée des signaux afin de produire la sortie du neurone. La fonction d'activation peut profondément influencer sur la performance du réseau, nous définissons les fonctions principales dans la section 3.1.1.2. Ceci permet alors de produire la sortie y .

Pour illustrer le cas le plus simple du neurone formel, on peut utiliser l'exemple du perceptron [Rosenblatt, 1958]. Le perceptron peut être vu comme le réseau de neurones le plus simple, il s'agit d'un classifieur linéaire. Il prend des décisions basées sur des entrées pondérées et sur une certaine valeur seuil. En faisant varier les poids et le biais, différents modèles de décisions sont alors définis.

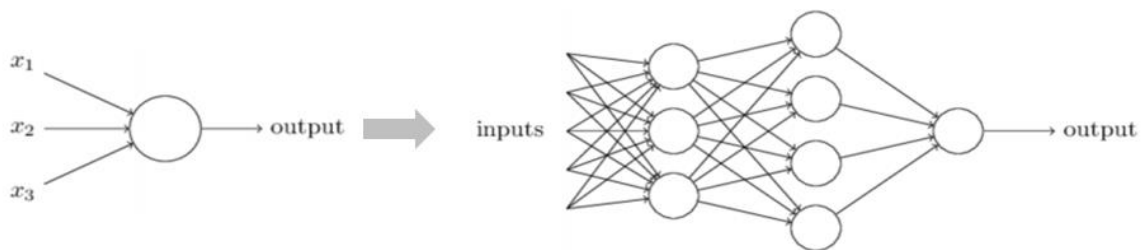


Figure 18- Du perceptron au perceptron multicouche [Nielsen, 2015] .

On peut alors définir un réseau plus complexe de neurones en combinant des perceptrons afin de prendre des décisions plus complexes, comme le montre la Figure 18. On appelle alors ces réseaux, les réseaux perceptron multicouche. Dans la première couche, trois décisions simples sont prises en pondérant les entrées. Ensuite ces décisions sont utilisées comme entrée de la seconde couche, chacune ayant son poids respectif, rendant alors la décision plus abstraite. On l'appelle la couche cachée car il ne s'agit ni d'une couche d'entrée, ni d'une couche de sortie. En termes d'équations, cela correspond au système défini par l'équation (15). Dans cette équation, on retrouve le produit scalaire entre l'information entrante X et son poids W , ainsi que le biais ou valeur seuil b .

$$output = \begin{cases} 0 & \text{si } W.X + b \leq 0 \\ 1 & \text{si } W.X + b > 0 \end{cases} \quad (15)$$

On applique dans cette équation (15) une fonction de Heaviside au potentiel post-synaptique. Cette fonction non-linéaire est une fonction d'activation, et nous en décrivons davantage à la section 3.1.1.2.

Ainsi, le neurone formel constitue la base du réseau de neurones. Les réseaux de neurones vont alors se distinguer par:

- L'organisation du graphe (architecture, niveau de complexité lié au nombre de neurones...),
- Les spécificités des neurones formels utilisés (fonction de transition ou d'activation),
- Par l'objectif visé (apprentissage supervisé ou non, optimisation ...).

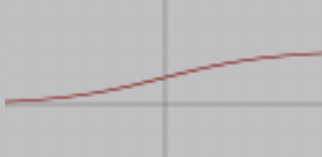
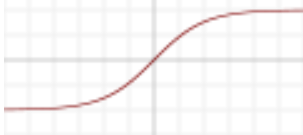
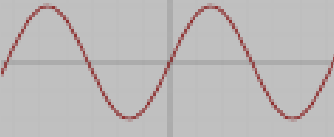
Ces paramètres sont définis dans les sections suivantes.

3.1.1.2. Les fonctions d'activation

En modifiant les poids et le biais, le modèle est capable d'apprendre à partir des données d'entraînement. Cependant, le moindre changement dans les poids ou biais peut changer

l'interprétation faite par le neurone et donc la sortie du réseau. Pour rendre le réseau plus sensible à ces transitions, de nouvelles fonctions ont été définies. Ainsi un changement subtil dans les paramètres clefs du réseau conduit à un changement similaire en sortie, et non à un changement radical. Ces fonctions sont présentées dans le Tableau 4. On les appelle fonctions d'activation non linéaires. Par exemple, dans le cas de la fonction sigmoïde, le neurone peut alors prendre n'importe quelle valeur entre 0 et 1, et fournir une valeur réelle entre ces deux valeurs en sortie. Dans notre problématique, nous serons amenés à considérer principalement la fonction tangente hyperbolique ou « tanh ».

Tableau 4- Fonctions d'activation.

Fonction	Définition	Description	Intervalle de définition
Sigmoïdale logistique	$\frac{1}{1 + e(-a)}$	Une courbe en « S » 	[0,1]
Tangente hyperbolique tanh	$\frac{e(a) - e(-a)}{e(a) + e(-a)}$	Une courbe sigmoïdale similaire à la fonction logistique. Produit généralement de meilleurs résultats que la fonction logistique en raison de sa symétrie. Idéale pour les perceptrons multicouches, en particulier pour les couches cachées 	(-1,+1)
Sinus	$\sin(a)$	N'est pas utilisé par défaut. 	[0,1]

3.1.1.3. L'apprentissage supervisé vs l'apprentissage non supervisé

Dans le cadre de l'apprentissage automatique, il existe différentes techniques d'apprentissage, notamment l'apprentissage supervisé, et le non supervisé. On distingue alors deux types de problèmes, en fonction de la présence ou non d'une variable à expliquer Y ou d'une forme à reconnaître qui a été, conjointement avec X , observée sur les mêmes objets. Dans le premier cas il s'agit bien d'un problème de modélisation ou apprentissage supervisé. On va trouver une fonction f susceptible, au mieux selon un critère à définir, de reproduire Y ayant observé X comme le décrit l'équation (13). Dans le cas contraire, en l'absence d'une variable à expliquer, il s'agit alors d'apprentissage dit non-supervisé. L'objectif généralement poursuivi est la recherche d'une typologie ou taxinomie des observations : comment regrouper celles-ci en classes homogènes mais les plus dissemblables entre elles. C'est un problème de classification (clustering).

Pour expliquer plus en détail l'apprentissage supervisé, qui est un des sujets majeurs lors du développement d'un réseau de neurones, il est important de préciser que l'un des objectifs principaux de l'apprentissage automatique est de trouver l'algorithme prédictif qui fournira les meilleures

performances grâce aux données mises à disposition. En fonction de l'ensemble des données considérées, la meilleure méthode à utiliser dépendra de la nature des données. De façon à mesurer les capacités de ces méthodes à fournir des prédictions optimales, différentes mesures de performance sont présentées dans la section 3.3. Ces techniques permettent d'évaluer dans quelle limite les prédictions obtenues grâce à un modèle approchent les données réelles. Etant donné qu'il est impossible d'utiliser des données nouvelles et non aperçues pour effectuer une comparaison, les données disponibles sont divisées en des sous-ensembles d'entraînement et de test, comme l'illustre la Figure 19. Le sous-ensemble d'entraînement ou d'apprentissage contient des données à partir duquel le modèle est construit ou entraîné. En général, on ne s'attache pas à la capacité du réseau à être optimal sur les données d'entraînement. On souhaite voir les performances du réseau sur des données inconnues. C'est là tout l'intérêt du sous-ensemble de test. Il ne sert pas à entraîner le réseau, mais à évaluer ces performances. La séparation des données en des sous-ensembles de données dépend de différents facteurs comme le type de données, le but de l'analyse, le nombre de données, ainsi que la répartition de ces données.

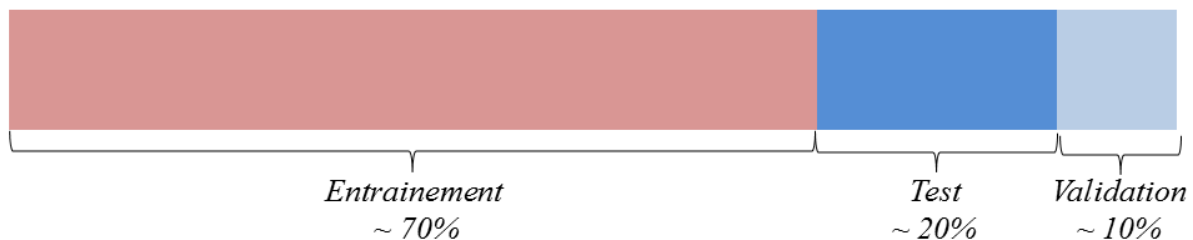


Figure 19- Répartition des données entre sous-ensemble d'entraînement, test et validation.

De nombreuses techniques d'apprentissage supervisé ne sont pas concernées uniquement par la comparaison de performances pour choisir le meilleur modèle. La plupart du temps, il y a également le besoin d'améliorer le modèle en réglant ses hyperparamètres (comme le nombre de couches cachées ou le nombre de neurones par couches). Si on fait un réglage de ces hyperparamètres sur le sous-ensemble de test, il y a un risque de surapprentissage ou overfitting du modèle. La Figure 20 représente l'évolution de l'erreur en fonction de l'époque (ou itérations) durant les phases d'apprentissage, test et validation, afin d'illustrer la notion de surapprentissage et sous-apprentissage. Le surapprentissage survient quand le modèle apprend des caractéristiques spécifiques à ces données, qui n'ont pas de relations causales avec la variable cible. Le modèle est alors trop flexible et apprend du bruit aléatoire au lieu d'apprendre les relations sous-jacentes. En général le surapprentissage peut être vu comme le fait que le modèle sera plus performant sur des données connues, et bien moins performants sur des nouvelles données [Hawkins, 2004].

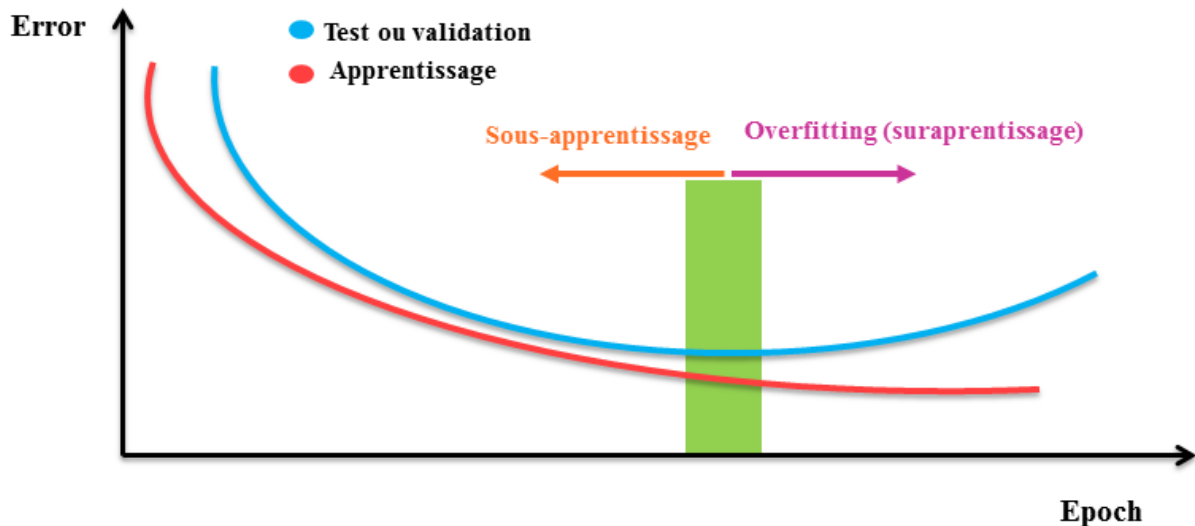


Figure 20- Evolution des courbes d'erreur en fonction des itérations (ou epoch). La courbe rouge représente l'évolution durant l'apprentissage, la bleue celle durant les phases de test ou de validation. La zone verte représente les itérations ou epoch à partir desquelles on constate un surapprentissage avec une augmentation de l'erreur de test.

Afin d'éviter le surapprentissage, un sous-ensemble de validation est utilisé lors du réglage des paramètres du modèle de prédiction. Les données sont alors initialement divisées en trois sous-ensembles : entraînement, validation et test. Durant la phase de réglage, le modèle est évalué en utilisant l'erreur de validation. Uniquement durant la dernière phase, les performances de prédiction sont évaluées en utilisant le sous-ensemble de test. On évite ainsi d'ajuster le modèle sur du bruit.

Cette évaluation du surapprentissage est également faite lors de la définition de la topologie de la structure, notamment pour définir le nombre de couches cachées, et le nombre de neurones dans chaque couche. C'est à force d'itérations successives ou epoch que l'on arrive à définir une topologie optimale en fonction d'une problématique donnée.

3.1.1.4. Les algorithmes d'optimisation

Maintenant que nous avons défini les notions clés de poids et de biais du réseau de neurones, ainsi que la notion d'apprentissage supervisé, nous définissons la méthode pour trouver ces hyperparamètres afin d'approximer le plus précisément possible $y(x)$ à partir des données d'entraînement x . Etant donné que l'on considère une méthode d'apprentissage supervisé, on a des données d'apprentissage à partir desquels on va chercher à minimiser l'écart entre la valeur réelle $y(x)$ et la valeur prédite par le réseau de neurones $\hat{y}(x)$. Pour ce faire, on introduit la notion de fonction de coût ou fonction de perte décrite par l'équation (16).

$$C(W, b) = \frac{1}{2n} \sum_{i=1}^n \| y(x_i) - \hat{y}(W, b) \|^2 \quad (16)$$

avec n le nombre de données d'entraînements, W les poids, b le biais et \hat{y} le vecteur de sortie du réseau qui dépend de W et de b . Cette fonction évalue les poids et les biais basés sur un coût quadratique ou erreur quadratique moyenne (MSE pour mean squared error). Pour des valeurs de W et de b fournissant une valeur optimale de $y(x)$, le coût sera faible, et vice-versa .

Pour un ensemble de poids et de biais donné, on souhaite déterminer la procédure à effectuer pour faire évoluer ces paramètres de façon à améliorer le modèle. Il faut minimiser la fonction de coût quadratique, et pour ce faire il est classique d'utiliser un algorithme d'optimisation de descente de gradient. Pour trouver un minimum local en utilisant une descente de gradient, les poids et les biais doivent être ajustés dans la direction de plus forte descente de la fonction au point considéré. Le gradient peut être vu comme la direction de plus forte pente de la fonction en ce point et c'est la direction opposée au gradient qui est utilisée. La descente de gradient est un processus itératif, où chaque variable est ajustée étape par étape. La fonction de coût est ainsi évaluée à chaque étape. La quantité à ajuster à chaque étape est appelée le pas ou encore le taux d'apprentissage de l'algorithme. Quand le gradient atteint une valeur nulle, un point stationnaire est atteint (un minimum, un maximum ou un point selle). Le gradient de la fonction de coût est donné par l'équation (17)

$$\nabla C = \left(\frac{\partial C}{\partial v_1}, \dots, \frac{\partial C}{\partial v_m} \right)^T = \left(\frac{\partial C}{\partial w_1}, \dots, \frac{\partial C}{\partial w_p}, \frac{\partial C}{\partial b_1}, \dots, \frac{\partial C}{\partial b_q} \right)^T \quad (17)$$

avec ∇C le vecteur du gradient et v une variable représentant les poids et les biais. Le vecteur du gradient consiste en la dérivée partielle de la fonction de coût par rapport à chaque variable. L'évolution de chaque variable peut être décrite par l'équation (18)

$$\nabla v = -\eta \nabla C \quad (18)$$

avec η le taux d'apprentissage qui est un réel positif. En pratique, on résout un problème d'optimisation mono-dimensionnelle (dans la direction du gradient) pour calculer le pas η . Appliqué aux poids et aux biais, la loi de descente de gradient pour les réseaux de neurones est établie par le système d'équations (19)

$$w_k \rightarrow w_{k+1} = w_k - \eta \frac{\partial C}{\partial w_k}(w_{k+1})$$

$$b_l \rightarrow b_{l+1} = b_l - \eta \frac{\partial C}{\partial b_l}(b_{l+1}) \quad (19)$$

Le processus permettant d'ajuster ces paramètres en évaluant les erreurs (différences entre la valeur réelle et la valeur prédite sur les points d'apprentissage), et de réitérer ce processus pas à pas pour minimiser la fonction de coût est appelé « backpropagation » ou rétropropagation en français.

Différentes descentes de gradients sont utilisées, la plus commune est l'algorithme de Levenberg Marquardt (voir [Levenberg, 1944], [Marquardt, 1963]) qui est adapté pour minimiser des fonctions non-linéaires.

3.1.2. Topologie des réseaux de neurones utilisés dans le cadre de notre étude

Nous présentons dans cette section les différents réseaux de neurones que nous avons développés, entraînés et optimisés dans le cadre de cette étude. Quatre réseaux ont été développés, en Matlab et en Python. Plus précisément, le perceptron multi-couche, le « Time Delay Neural network » et le NARX (« Non linear AutoRegressive with eXogenous inputs ») ont été programmés à partir de la toolbox Neural Network de Matlab Le réseau Long Short Term Memory a été programmé avec le langage Python, en faisant appel à la librairie Lasagne (<https://github.com/Lasagne/Lasagne>) et la surcouche Theano (<http://deeplearning.net/software/theano/>). Nous détaillons ceci dans l'annexe 2.

3.1.2.1. Le Perceptron multi-couche

Le perceptron multicouche, également appelé par abus de langage « feedforward backpropagation » suite à la technique d'entraînement utilisé est comme nous l'avons vu dans le Chapitre 2 section 5.1., le réseau le plus largement utilisé dans la communauté. Afin de comparer sa structure à celles de réseaux plus complexes, nous avons repris celle-ci avec un code couleur présenté sur la Figure 21. Nous limitons l'analyse à une seule couche cachée représentée en violet, alimentée par la couche d'entrée représentée en bleu. Cette limite à une seule couche est faite afin d'analyser au mieux la connexion entre entrées et sorties. Elle fournit à la couche de sortie la prédiction de l'indice magnétique am .

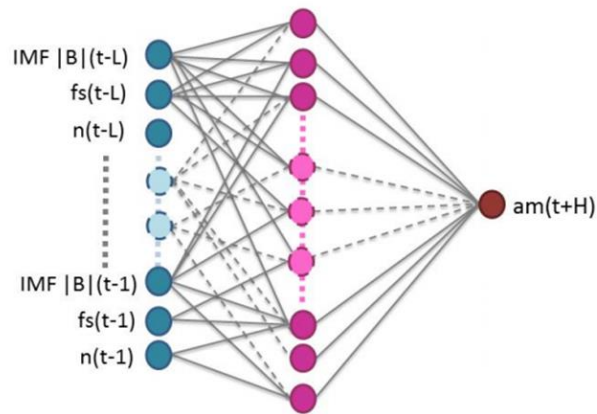


Figure 21 –Perceptron multicouche.

Cette structure par définition est une structure statique. Il n'y a pas de connexion entre les différentes couches du réseau, si ce n'est des connexions unidirectionnelles. Cependant, afin de représenter au mieux la complexité de la dynamique de la magnétosphère terrestre, nous considérons un historique d'entrée des paramètres du vent solaire (ici la densité n , la vitesse ou flow speed fs et l'IMF $|B|$ de taille L). Nous programmons également une connexion entre la sortie et l'entrée comme l'illustre la Figure 22 afin de donner au réseau l'information sur l'indice prédit par celui-ci, également appelé « nowcast index ». Ainsi, l'indice prédit est également utilisé en entrée du réseau de neurones.

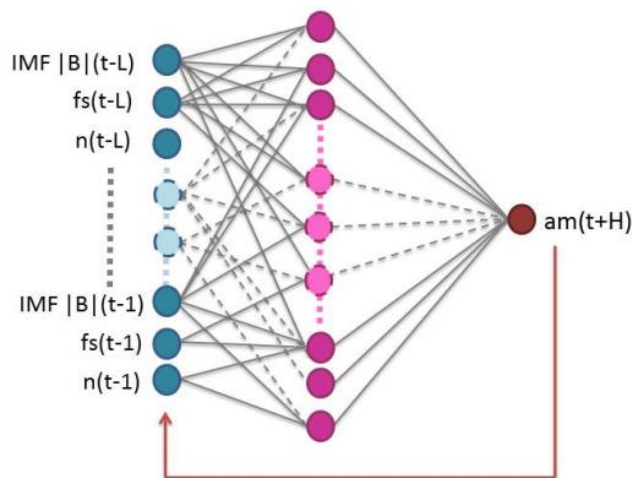


Figure 22- Perceptron Multicouche adapté à notre problématique. L'indice prédit par le réseau ou « nowcast » est renvoyé en entrée.

Ce réseau peut être défini par l'équation (20)

$$\hat{y}_k = \sum_{j=1}^{N_{hidden}} W_{jk} \tanh(\sum_{i=0}^{N_{input}} W_{ji} x_i) \quad (20)$$

avec l'indice i qui correspond au nombre de neurones N_{input} dans la couche d'entrée, l'indice j au nombre de neurones dans la couche cachée N_{hidden} et l'indice k à la couche de sortie. La fonction d'activation de la couche de sortie est une fonction linéaire, et celle de la couche cachée est une fonction tangente hyperbolique (\tanh) comme décrite dans le Tableau 4. Le poids W_{ji} connecte la couche d'entrée et la couche cachée, et W_{kj} est la connexion entre la couche cachée et la couche de sortie. N_{input} et N_{hidden} sont des hyperparamètres qui seront choisis pour minimiser l'erreur de validation comme expliqué dans la section 3.1.1.3.

Dans notre étude, nous n'avons qu'une seule sortie donc nous pouvons simplifier l'équation en remplaçant l'indice k par 1. On obtient alors l'équation (21)

$$\hat{y} = \sum_{j=1}^{N_{hidden}} W_j \tanh(\sum_{i=0}^{N_{input}} W_{ji} x_i) \quad (21)$$

3.1.2.2. Le réseau à retard de temps

Le « Time Delay Neural Network » (TDNN) ou réseau à retard de temps a tout d'abord été développé pour faire de la reconnaissance vocale par [Waibel et al., 1989]. Il a ensuite été utilisé en analyse de séquence par [Wohler and Anlauf, 1999]. Comme nous l'avons présenté dans le Chapitre 1 Section 5.2, ce réseau a également été utilisé pour prédire l'indice Dst . Il trouve donc sa place dans les études de séquence temporelle, c'est pourquoi nous avons souhaité voir son application dans le cadre de la prédiction de l'indice magnétique am .

Afin de garder une structure comparable à la structure du réseau précédent, nous conservons une couche cachée en violet sur la Figure 23. La couche d'entrée en bleu reçoit les informations du vent solaire à l'instant t , et est reliée avec un retard de temps à la couche en vert appelée fenêtre de spécialisation. Le nombre de délai τ_d ou retard de temps définit la longueur de la fenêtre de spécialisation. Cette fenêtre est également connectée à la couche cachée. Ainsi, cette couche est sensible aussi bien aux transitions rapides du signal d'entrée, qu'aux transitions plus lentes enregistrées par la fenêtre de spécialisation. On peut considérer que le TDNN a une structure de type perceptron multi-couche avec des poids partagés. Ceci correspond à des poids ayant la même valeur pour des connexions entre les neurones définies au même instant t .

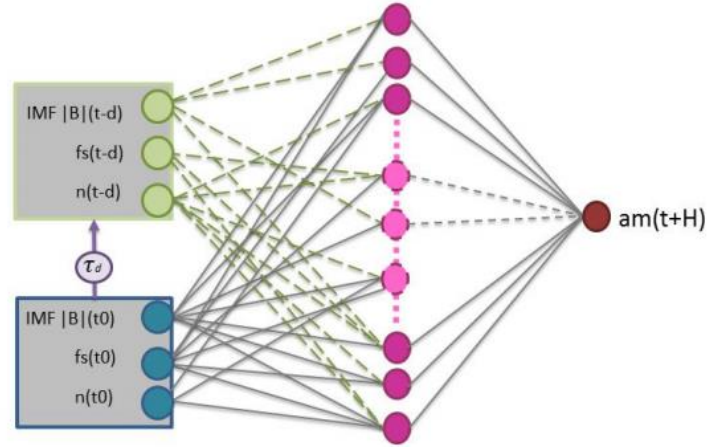


Figure 23- Le réseau à retard de temps ou TDNN.

Ce réseau ne considère en entrée que les paramètres du vent solaire, contrairement au réseau perceptron multicouche que nous avons défini précédemment qui considère également l'indice nowcast comme le montre la Figure 22 (avec la flèche de retour en rouge). Nous faisons ce choix car l'indice magnétique n'est pas définitif avant un certain délai pouvant aller jusqu'à plusieurs mois, il est alors intéressant d'avoir une structure basée uniquement sur les paramètres du vent solaire afin de ne pas ajouter en entrée une information que l'on ne peut vérifier rapidement. Le TDNN permet de voir comment un réseau se comporte en utilisant uniquement des informations externes à la magnétosphère, grâce aux poids partagés. Nous définissons le TDNN avec l'équation (22).

$$\hat{y} = \sum_{j=1}^{N_{hidden}} W_j \tanh\left[\left(\sum_{i=0}^{N_{input}} W_{ji} x_i + \sum_{l=0}^{N_{window}} W_{lji} x_l\right)\right] \quad (22)$$

avec l'indice l associé à la fenêtre de spécialisation. Le poids W_{lij} est la connexion entre la couche cachée et la couche d'entrée avec le retard de temps.

3.1.2.3. Les réseaux récurrents

Les réseaux récurrents possèdent au moins un cycle. Les neurones sont interconnectés et interagissent non linéairement. Ils créent un état interne au réseau, c'est comme si on avait des copies multiples d'un même réseau, chacun passant un message à son successeur comme illustré sur la Figure 24. Chaque unité est alors reliée par un arc ou synapse ayant un poids. Ils sont comparables à des réseaux de neurones classiques avec des contraintes d'égalité entre les poids du réseau. Les techniques d'entraînement du réseau sont les mêmes que pour les réseaux classiques, mais l'optimisation est plus complexe.

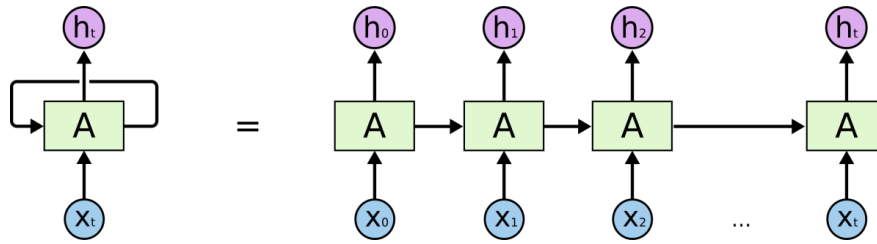


Figure 24- Réseau récurrent en déroulé, X représente l'entrée, A le neurone, et H la sortie. (<http://colah.github.io>).

Nous présentons ici deux réseaux utilisés pour notre étude, le réseau non linéaire autorégressif à entrées exogènes ou NARX, et un réseau à mémoire à court et long terme ou LSTM.

3.1.2.4. Le réseau non linéaire autorégressif à entrées exogènes

Le réseau non linéaire auto régressif à entrées exogènes (NARX) développé par [Leontaritis and Billings, 1985] dans les années 80 est un modèle puissant pour des applications à des systèmes non linéaires, et plus spécifiquement aux séries temporelles (voir [Haykin, 1998]; [Lin et al, 1996]; [Gao and Er, 2005]). Ces réseaux qui sont aussi puissants que des machines de Turing, voir même des super-Turing (voir [Kilian and Siegelmann, 1993], [Siegelmaan and Sontag, 1991], [Siegelmaan and Sontag, 1994], [Siegelmaan and Sontag, 1995], [Lin et al, 1996]). [Gao and Er, 2005], ont démontré que l'optimisation utilisant l'algorithme de gradient de descente type Levenberg Marquardt est plus efficace avec ce type de réseau, et qu'ils convergent et généralisent plus rapidement que les autres.

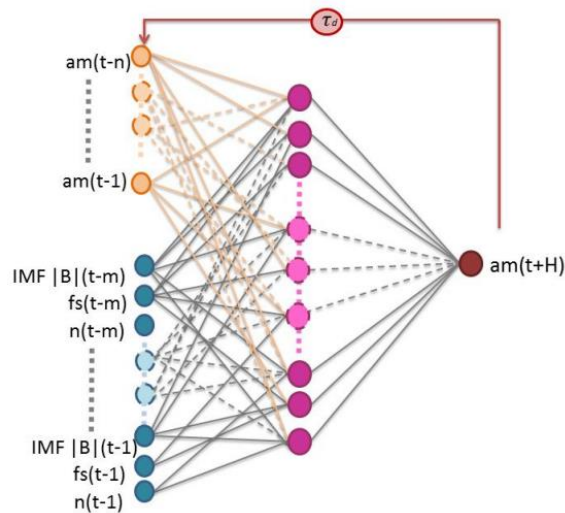


Figure 25- Le réseau non linéaire autorégressif à entrées exogènes.

Il existe différentes façons d'implémenter un NARX, dans notre cas nous souhaitons conserver une structure comparable à celle d'un réseau perceptron multicouche afin de mieux pouvoir les comparer par la suite, comme nous l'avons fait pour le TDNN. Nous gardons une seule couche cachée, alimentée par deux couches d'entrée différentes. Une des couches d'entrée correspond à une mémoire

intégrée contenant les entrées exogènes c'est-à-dire les paramètres du vent solaire, représentée en bleu sur la Figure 25. Une autre couche d'entrée correspond à la partie autorégressive avec une connexion à retard depuis la sortie du réseau, représentée en orange sur cette même figure. Cette couche va enregistrer les indices « nowcast » fournis par le réseau NARX et les injecter en entrée. Dans le cadre du réseau « feedforward backpropagation » (ou perceptron multicouche) illustré sur la Figure 22, on injecte les valeurs une par une, dans le cadre du réseau NARX, on construit un vecteur dynamique de taille n de ces valeurs avant de les injecter en entrée.

La sortie de ce réseau peut être ainsi décrite par l'équation (23)

$$\hat{y} = \sum_{j=1}^{N_{hidden}} W_j \tanh\left[\left(\sum_{i=0}^{N_{input_i}} W_{ji} x_i + \sum_{m=0}^{N_{input_m}} W_{kmj} f_m x_m\right)\right] \quad (23)$$

avec l'indice i correspondant à la couche contenant les paramètres exogènes (les paramètres du vent solaire), et l'indice m correspondant à la couche contenant la partie autorégressive (l'indice nowcast). W_{kmj} représente la connexion entre la couche autorégressive et la couche de sortie.

3.1.2.5. Le réseau à mémoire à court et long terme

Sur la Figure 24, on constate que ce réseau est comparable à une structure classique que l'on pourrait entraîner avec une rétropropagation de gradient, comme cela a été fait avec les précédents réseaux. Mais avec ce type d'entraînement, on fait face à un phénomène de « vanishing gradient » ou d'évanescence de gradient. Pour comprendre ce phénomène, il faut se rattacher à la définition de l'apprentissage à la section 3.1.1.3. Pour optimiser les poids, le réseau de neurones va essayer de minimiser au cours de l'entraînement une fonction d'erreur (équation (16)) dépendante de la sortie du réseau. Le gradient va décroître en général de manière exponentielle au fur et à mesure de l'éloignement dans le temps entre la donnée considérée à l'instant, et la donnée requise dans le passé. Plus l'événement nécessaire pour calculer la prédiction est lointain, moins la correction d'erreur entre l'événement et le présent ne sera efficace. Elle diminue exponentiellement avec l'intervalle de temps. Par conséquent, le réseau peut connecter les informations des instants précédents aux instants actuels, tant que l'information requise n'est pas trop éloignée dans le temps, comme illustré sur la Figure 26.

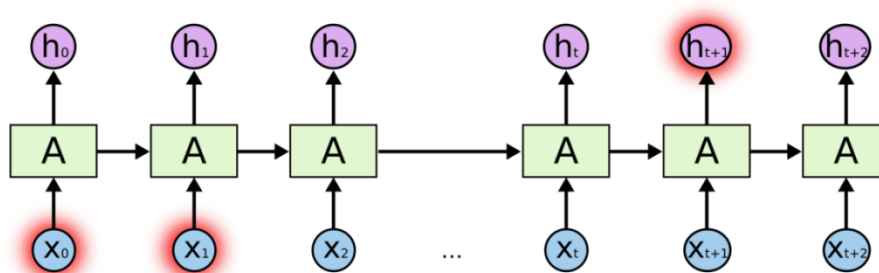


Figure 26- Problème de la dépendance dans le temps entre une information entrante X_0 et une sortie à l'instant h_{t+1} (<http://colah.github.io>).

Pour pallier à ce problème, des architectures particulières ont été développées, les réseaux LSTM. Pour comparer ce réseau aux réseaux précédents, nous l'avons représenté de façon schématique sur la Figure 27.

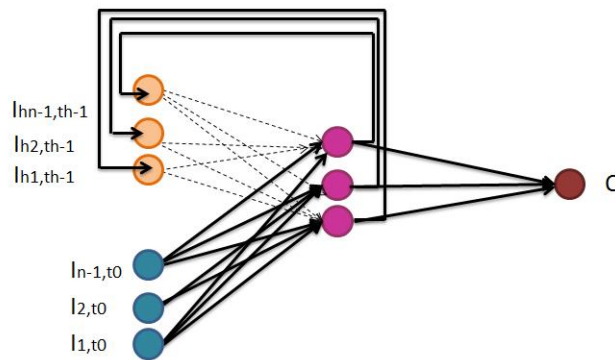


Figure 27- Réseau LSTM.

Ici nous ne précisons pas les données arrivant à la couche d'entrée en bleu car nous avons utilisé ce réseau aussi bien pour la prédiction de l'indice magnétique *am* que pour la prédiction de l'indice *Dst* caractérisé par le rond rouge sur la Figure 25. La couche cachée en violet interagit en permanence avec la couche orange afin de montrer qu'il y a une interaction interne au réseau, nous la détaillons plus précisément.

Le réseau LSTM a été développé en 1997 par [Hochreiter and Schmidhuber, 1997] et a pour but de répondre au problème de disparition de gradient. Pour comparaison, la Figure 28 représente une cellule type utilisée par un réseau récurrent, et la Figure 29 celle utilisée par un LSTM. Le cœur du LSTM est la « cell state », que l'on appelle également convoyeur, encadré en rouge sur la Figure 29. Dans le réseau récurrent classique, chaque unité de calcul ou neurone est liée à un état caché. Dans le réseau LSTM, chaque unité est aussi liée à la « cell state » qui joue le rôle de mémoire. Pour mieux comprendre le schéma de fonctionnement du LSTM, nous le détaillons pas à pas.

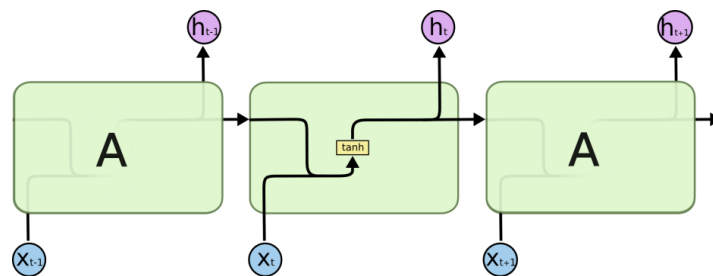


Figure 28- Schéma réseau récurrent. (<http://colah.github.io>).

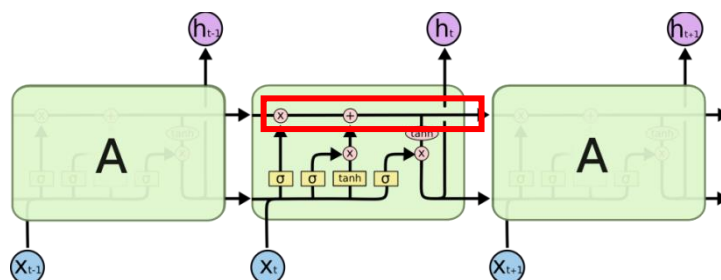


Figure 29 - Schéma réseau LSTM. (<http://colah.github.io>).

Sur la Figure 30, on représente le convoyeur qui fait passer la cellule de l'état C_{t-1} à C_t . Différents ponts ou « gates » sont programmés afin d'agir sur l'information, et de contrôler ce qui doit être retenu ou oublié ponctuellement en fonction de son importance dans le temps. Ces ponts sont des sigmoïdes, comme décrit dans le Tableau 4. Si elles valent 0, aucune information ne passe, en revanche si elles valent 1, toute l'information passe. Le programmeur peut alors contrôler le flux d'information grâce au contrôle opéré par la « cell state ».

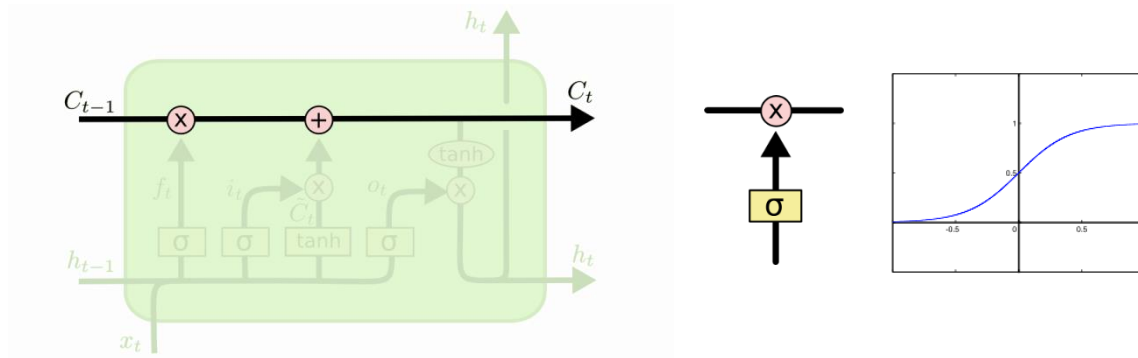


Figure 30- Schéma du convoyeur du réseau avec les ponts sigmoïdes. (<http://colah.github.io>).

La première étape consiste à définir les informations des instants précédents qu'il faut « garder ou jeter » par rapport à l'information entrante dans le neurone. On utilise un « forget gate » représenté sur la Figure 31.

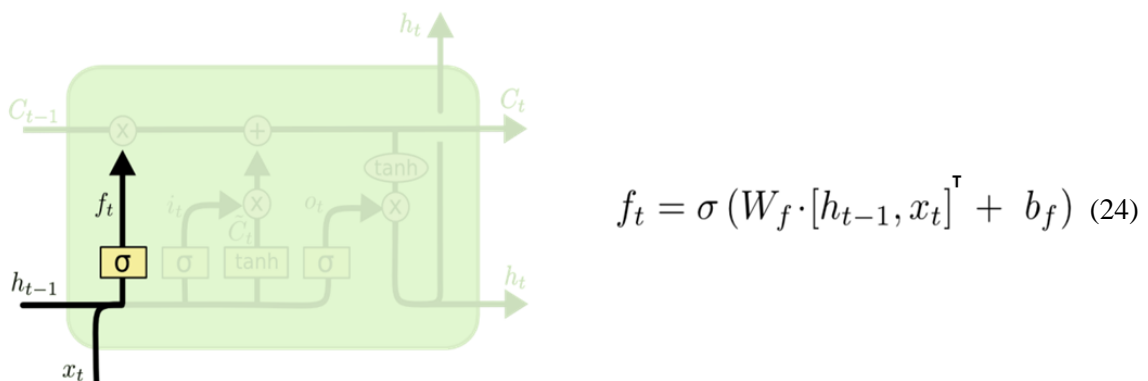
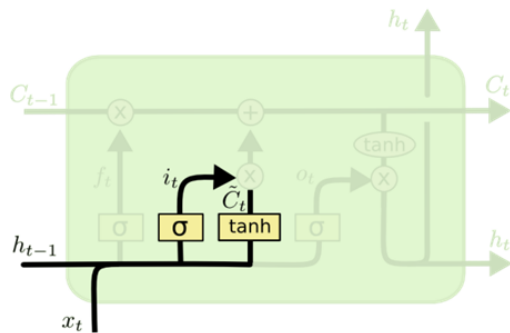


Figure 31- Schéma du "forget gate" et équation associée. (<http://colah.github.io>).

La fonction σ fournit un nombre compris entre 0 et 1 en fonction de l'importance de l'information, et le produit scalaire $W_f \cdot [h_{t-1}, x_t]^T$ nous donne la relation entre l'instant précédent et la nouvelle entrée arrivant dans l'unité de calcul.

Au cours de la seconde étape, il faut décider de l'information à stocker dans la « cell state » par rapport aux instants précédents. Donc sur la Figure 32, on peut voir qu'on a la même définition que dans le cas du « forget gate », mais cette fois le but n'est pas de savoir ce qu'il faut supprimer par rapport aux instants précédents, mais de garder ce qui est nécessaire. Donc on fait passer l'information par un pont i_t pour savoir ce qui est utile, et on définit un vecteur \tilde{C}_t contenant les nouvelles informations.

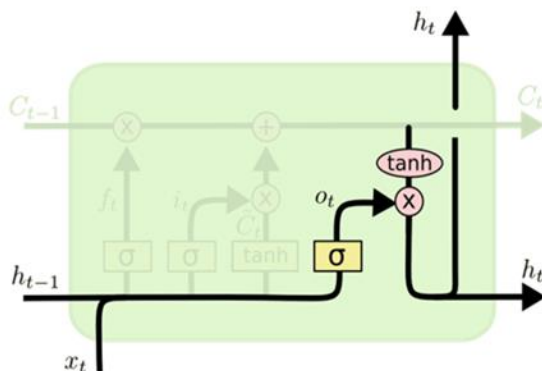


$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t]^T + b_i) \quad (25)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t]^T + b_C)$$

Figure 32- Schéma du pont servant à garder les nouvelles informations et équations associées. (<http://colah.github.io>).

En troisième étape, il faut passer de l'ancienne « cell state » à la nouvelle. On effectue alors une concaténation des étapes 1 et 2. Sur la Figure 33 on définit que la nouvelle cellule est la somme de ce qu'on a conservé des informations des instants précédents $f_t * C_{t-1}$ et ce qu'on ajoute comme nouvelle information $i_t * \tilde{C}_t$ en fonction de ce qui a été évalué comme nécessaire. L'opérateur * est associé au produit matriciel de Hadamard.

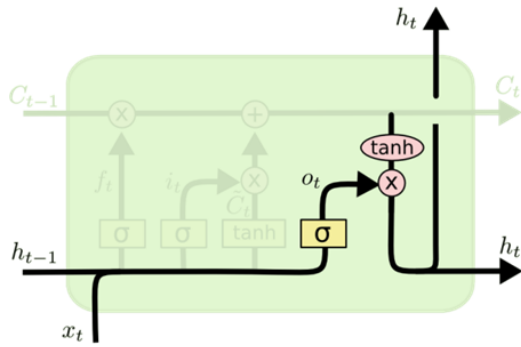


$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t]^T + b_o) \quad (26)$$

$$h_t = o_t * \tanh(C_t)$$

Figure 33- Schéma de création de la nouvelle "cell state" et équation associée. (<http://colah.github.io>).

Au final, il faut définir ce qu'on souhaite fournir en sortie h_t , en fonction de la « cell state ». On a alors une première étape comme illustrée par la Figure 34 qui va définir un premier calcul o_t en fonction de la sortie précédente, puis une seconde qui filtre cette information en fonction de la mémoire interne à la cellule représentée par la « cell state » C_t . C'est grâce à cette mémoire qu'on comble la perte d'information associée à l'effet de disparition de gradient et que la cellule conserve les informations essentielles pour faire des prédictions. Le LSTM possède une approche multiéchelle, ce qui est particulièrement intéressant pour traiter des systèmes auto-organisés dans lesquels la mémoire peut tendre vers l'infini.

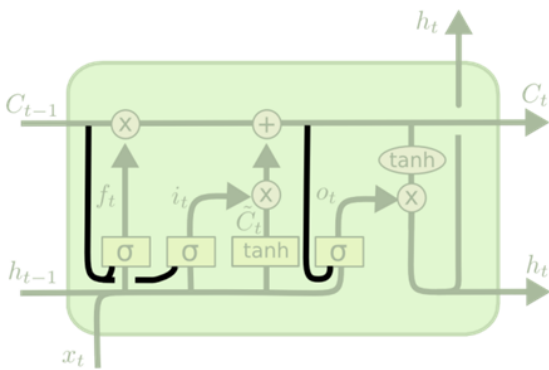


$$o_t = \sigma (W_o [h_{t-1}, x_t]^T + b_o) \quad (27)$$

$$h_t = o_t * \tanh (C_t)$$

Figure 34 - Schéma de création de la nouvelle sortie et équations associées. (<http://colah.github.io>).

Dans notre cas, nous avons souhaité ajouter des « peephole connection » entre les différentes étapes représentées sur la Figure 35. Chaque étape définie précédemment a une vue sur la « cell state » C_{t-1} et C_t . [Gers et al., 2002] ont montré que c'était mieux adapté pour résoudre des problèmes non linéaires, ce qui est notre cas avec l'étude de l'interaction Soleil –Terre.



$$f_t = \sigma (W_f \cdot [C_{t-1}, h_{t-1}, x_t]^T + b_f)$$

$$i_t = \sigma (W_i \cdot [C_{t-1}, h_{t-1}, x_t]^T + b_i) \quad (28)$$

$$o_t = \sigma (W_o \cdot [C_t, h_{t-1}, x_t]^T + b_o)$$

Figure 35- Schéma des "peephole connection" et équations associées. (<http://colah.github.io>).

Nous avons défini dans cette section les différents réseaux de neurones utilisés. Ces modèles fournissent des prévisions dites « single point », en donnant une information pas à pas dans le temps. Nous avons alors travaillé sur les processus gaussiens, afin de pouvoir à terme combiner les deux techniques et obtenir une information probabiliste sur la prédiction de l'indice considéré.

3.2. Les processus gaussiens

Bien qu'ils ne soient pas basés sur des modèles biologiques comme les réseaux de neurones, les processus gaussiens dans le domaine de la régression fournissent une mesure explicite de l'incertitude de prédiction. En incluant une incertitude dans le processus de définition du modèle, un ensemble de modèles peut être considéré dans un espace de solution ; en opposition à un seul modèle optimal dans le cas des réseaux de neurones. Comme nous l'avons vu dans la partie 2, le but des modèles de prévision de séries temporelles est d'approximer la fonction sous-jacente aussi près que possible, de façon à fournir des prédictions de qualité dans le futur. Au lieu d'avoir un seul modèle, chacun des modèles dans l'espace des solutions a une certaine probabilité d'expliquer les données. Cette façon de procéder est essentielle à l'inférence bayésienne [Frigola-Alcalde, 2015]. Avant d'expliquer le principe des processus gaussiens, il est important de définir le cadre d'étude lié à la statistique Bayésienne.

3.2.1. Le théorème de Bayes

La statistique Bayésienne est une théorie, contrairement aux statistiques classiques, où les probabilités sont appliquées à des problèmes statistiques. Autrement dit, il fournit aux chercheurs un outil pour mettre à jour leurs précédentes estimations en présence de nouvelles données. Ceci est généralement basé sur une probabilité conditionnelle et sur le théorème de Bayes. La probabilité conditionnelle est définie comme la probabilité P qu'un certain événement B ait lieu, étant donné l'occurrence d'un autre événement A . Ceci est écrit plus formellement avec la notation $P(B|A)$. Le théorème de Bayes est décrit par l'équation (29)

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \quad (29)$$

De plus $P(A)$ peut être écrite comme une marginalisation suivant l'équation (30)

$$P(A) = \sum_i P(A|B_i)P(B_i) \quad (30)$$

avec $\{B_i\}$ une partition de l'ensemble des possibles de B . On peut alors définir le théorème de Bayes par l'équation (31)

$$P(B|A) = \frac{P(A|B)P(B)}{\sum_i P(A|B_i)P(B_i)} \quad (31)$$

3.2.2. L'inférence Bayésienne

L'inférence Bayésienne permet d'appliquer les tâches de l'apprentissage supervisé dans le cadre bayésien. L'inférence signifie littéralement : une conclusion donnée sur la base de l'évidence. Dans ce contexte, une première tâche dans l'apprentissage supervisé est d'approximer $P(A|B)$ avec le modèle approprié, donnant un ensemble d'exemples correspondants de A et de B [Tipping, 2004]. Comme le décrit le principe de l'inférence Bayésienne, il faut mettre en place des modèles paramétrés pour définir notre probabilité conditionnelle comme décrit par l'équation (32)

$$P(B|A) = f(A, \theta) \quad (32)$$

avec θ un vecteur de paramètres, et f notre modèle. Etant donné un ensemble D de N données d'exemples de notre variable, une technique conventionnelle pour entraîner le modèle serait d'utiliser une fonction de perte pour optimiser les paramètres θ . Les prédictions de B sont ainsi faites, étant donné A , en évaluant $f(A, \theta)$ avec des paramètres θ fixés à leurs valeurs optimales.

Cependant, un élément clef de l'approche bayésienne est que les paramètres du modèle sont traités comme des variables aléatoires, comme A et B , donc la probabilité conditionnelle devient alors $P(B|A, \theta)$. Ainsi, la dépendance de B suivant θ , ainsi que suivant A est mise en évidence. Au lieu d'apprendre sur un ensemble exact de paramètres en utilisant des fonctions de pertes spécifiques, une distribution sur θ est déduite à partir de la loi de Bayes. Pour obtenir cette distribution postérieure, une distribution préalable $p(\theta)$ doit être spécifiée avant d'observer les données.

Pour redéfinir ceci plus simplement, en apprentissage supervisé classique, les problèmes sont approchés en définissant d'une part une classe restreinte de fonctions (e.g. les fonctions linéaires), et d'autre part un ensemble défini de paramètres obtenus suite à un processus d'optimisation. Avec l'approche Bayésienne, on requiert une probabilité de distribution a priori sur toutes les fonctions possibles, où les probabilités les plus élevées sont données aux fonctions qui sont les plus à même d'expliquer les données [Rasmussen and Williams, 2006]. Le besoin d'une distribution a priori est souvent vu comme un inconvénient de l'inférence Bayésienne. Cependant, en excluant toutes les variables sans intérêt, l'approche Bayésienne est capable de préférer des modèles simples qui suffisent à expliquer les données, sans inclure des complexités superflues [Tipping, 2004]. William of Ockham était un logicien du 14^{ème} siècle dont la théorie est interprétée comme « parmi toutes les hypothèses en compétition, celle avec le moins de complexité doit être choisie ». Cette théorie est souvent utilisée comme un guide heuristique en modélisation statistique.

L'inférence Bayésienne peut être résumée par l'équation (33), calculée par la loi de Bayes

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} \quad (33)$$

Si la donnée D est considérée comme un vecteur d'entrée X et la valeur cible correspondante y , alors l'équation (33) peut être reformulée suivant l'équation (34)

$$P(\theta|y, X) = \frac{P(y|X, \theta)P(\theta)}{P(y|X)} \quad (34)$$

avec $P(\theta)$ qui représente les connaissances a priori sur les paramètres avant d'observer les données. $P(y|X, \theta)$ est la vraisemblance d'observer nos résultats ou cibles, en fonction de données d'entrées définies, et une certaine distribution de θ . $P(y|X)$ est la constante de normalisation également appelée vraisemblance marginale. Il s'agit de la probabilité des données déterminée en intégrant sur toutes les valeurs possibles de θ . Après avoir observé les données, la connaissance a priori est ajustée avec la vraisemblance des données pour obtenir la distribution a posteriori $P(\theta|y, X)$. On pourrait alors réécrire l'équation (34) ainsi

$$\textit{distribution a posteriori} = \frac{\textit{vraisemblance} \times \textit{connaissance a priori}}{\textit{vraisemblance marginale}}.$$

La Figure 36 montre le fonctionnement de la méthode Bayésienne pour un problème simple de régression. A gauche on voit la distribution a priori à partir de laquelle quatre fonctions d'échantillonnages sont tracées. A droite, on montre la distribution a posteriori, après que deux points aient été ajoutés. On parle souvent de trajectoires, puis de trajectoires conditionnées par les points d'observation. Dans ces cas-là, la trajectoire est interpolante aux points d'observations. La ligne continue montre la prédiction moyenne, moyennée sur les différentes fonctions possibles. L'aire grisée représente l'intervalle de confiance associée à la variance en chaque point. Une façon de réduire la flexibilité dans la distribution des fonctions lorsque les données arrivent est de tracer différentes fonctions aléatoires à partir des connaissances a priori, et d'éliminer celles qui ne passent pas par les observations [Rasmussen and Williams, 2006]. On observe qu'une famille de courbes passent bien par les points, et ont des comportements divergents dans les régions où il n'y pas d'observations. Le concept de distribution des fonctions comme un espace de solution est central dans la modélisation Bayésienne [Roberts et al. 2012]. Plus récemment, [Chiplunkar, 2017] a également travaillé sur

l'application des processus gaussiens adaptées aux grandes bases de données ou pour les processus multi-sorties.

Quand on applique ce principe à l'analyse de séries temporelles, on parle d'ajustement de courbes. En ajustement de courbes, l'hypothèse est faite que la variable cible y est ordonnée par X , avec X représentant des données temporelles. L'inférence est alors faite en ajustant des courbes à un ensemble de points X, y . Par la suite, les prédictions peuvent être faites en extrapolant la courbe qui modélise les observations [Roberts et al. 2012].

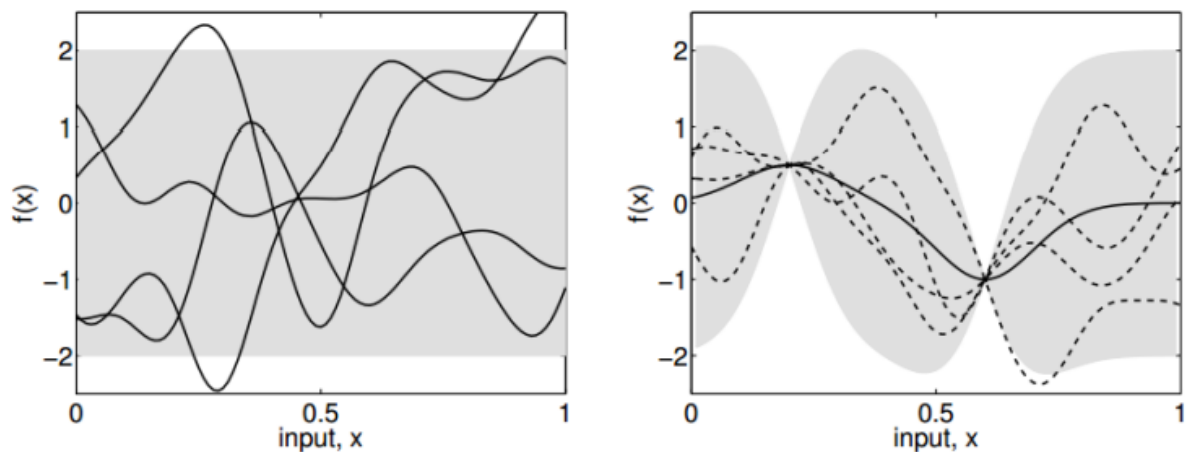


Figure 36- Régression bayésienne pour un exemple hypothétique. La figure de gauche montre la distribution de fonctions a priori. La figure de droite montre la distribution a posteriori après avoir ajouté deux points de données observées. La ligne continue montre la prédiction moyenne, moyennée sur les différentes fonctions possibles. L'aire grisée représente l'écart type (standard deviation) à chaque point d'entrée [Rasmussen and Williams, 2006].

3.2.3. La régression au moyen des processus gaussiens

Etant donné que la prévision est basée sur un ensemble infini de fonctions possibles, la question se pose de savoir comment un tel calcul peut être fait en un temps raisonnable, voir limité ? C'est ici que les processus gaussiens entrent en jeu. [Rasmussen and Williams, 2006] ont défini que les processus gaussiens sont une généralisation de la probabilité de distribution gaussienne. Alors que les probabilités de distribution sont utilisées pour décrire des données aléatoires, un processus (stochastique) gouverne les propriétés des fonctions. Au lieu de calculer une distribution complète sur les fonctions pour approximer un processus sous-jacent, nous allons nous limiter à la distribution des propriétés de ces fonctions sur un nombre fini de points.

Un processus gaussien est spécifié par une fonction moyenne $m(x)$ et une fonction de covariance ou noyau (« kernel ») $k(x, x')$. Il est utilisé pour modéliser des fonctions, en traitant chacun des points dans l'espace des fonctions $f(x)$ comme une variable aléatoire ayant une distribution gaussienne. De plus, ces points sont conjointement distribués gaussiennement, ce qui implique qu'il y a une covariance (ou encore corrélation) entre ces points. La fonction de covariance spécifie la façon dont chacun des points influence les valeurs que les points suivants sont susceptibles de prendre. L'idée principale est que si x et x' sont similaires d'après la fonction de covariance, on s'attend à ce que les

sorties $f(x)$ et $f(x')$ aient également un comportement similaire. Il existe différentes fonctions de covariance qui déterminent la forme (« smoothness », largeur temporelle etc...) du modèle. Le problème de l'apprentissage avec les processus gaussiens est de trouver les propriétés adaptées pour la fonction de covariance.

Un processus gaussien ou *GP* peut être décrit par le système d'équations (35)

$$\begin{aligned} f(x) &\sim GP(m(x), k(x, x')) \\ m(x) &= E[f(x)] \end{aligned} \tag{35}$$

$$k(x, x') = E[(f(x) - m(x))(f(x') - m(x'))]$$

avec $f(x)$ le processus réel que l'on souhaite modéliser, m la moyenne, k la fonction de covariance ou kernel, E l'espérance mathématique et x et x' deux points différents. Dans notre cas, les variables aléatoires à partir de la définition des processus gaussiens représentent la valeur de $f(x)$ en x . Plus spécifiquement, pour les séries temporelles, x sera les paramètres temporels du vent solaire décrits à la section 1.2.

Notons que le processus gaussien est un modèle non paramétrique, c'est-à-dire que tous les points de données doivent être mémorisés et sont utilisés pour prédire chaque point de données ultérieures. Dans le cadre des *GP*, on parle d'hyperparamètres à évaluer en fonction des données fournies en entrée. Ces hyperparamètres sont définis au cours d'un processus d'entraînement. La fonction de covariance permet de définir différents « prior » sur l'espace des fonctions. En fonction de la fonction choisie, les paramètres approximatés n'auront pas les mêmes valeurs. Comme le montre la Figure 37, en considérant la même base de données (ici 4 points), l'approximation varie beaucoup en fonction du noyau choisi.

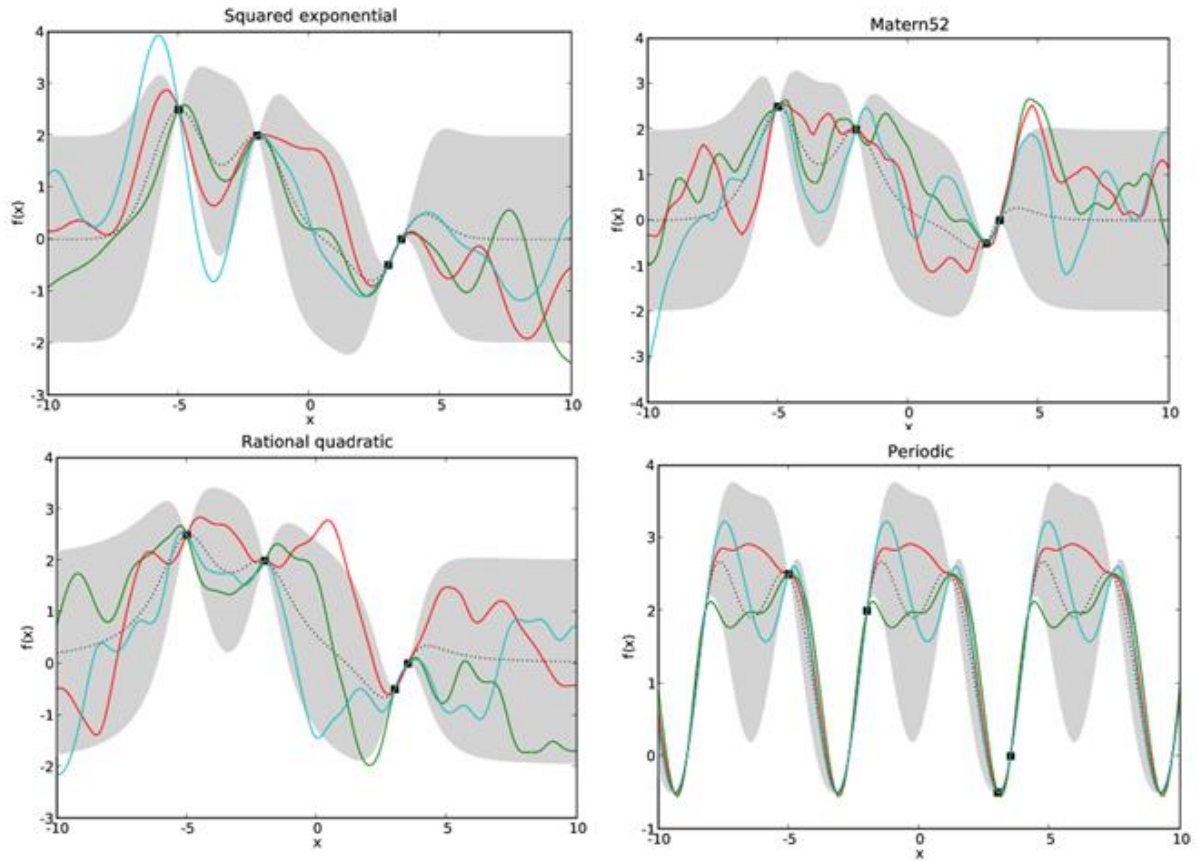


Figure 37 - Variations des résultats fournis par les processus gaussiens sur quatre points d'observations en fonction du kernel considéré : squared exponential, matern52, rational quadratic, periodic. (<https://pythonhosted.org>).

3.2.4. Prédire à partir des processus gaussiens

Etant donnée un ensemble de points de test X^* , la distribution de probabilité de la loi jointe entre les données d'entraînement $f(X)$ et les sorties de test $f(x^*)$ est donnée par une loi normale définie par l'équation (36).

$$\begin{bmatrix} f \\ f^* \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} m(x) \\ m(x^*) \end{bmatrix}, \begin{bmatrix} K(X, X) & K(X, X^*) \\ K(X^*, X) & K(X^*, X^*) \end{bmatrix} \right) \quad (36)$$

S'il y a n données d'entraînements et n^* données tests, alors $K(X, X^*)$ représente la matrice $n \times n^*$ de covariance entre les données d'entraînement X et les données de test X^* . \mathcal{N} définit le symbole de la loi normale.

Ainsi, lors d'une prédiction effectuée à partir d'un processus gaussien, la dépendance entre les données est prise en compte. Une fois que les hyperparamètres du processus ont été fixés lors de l'entraînement, c'est à dire si on réfère au principe de l'inférence bayésienne, que la connaissance a priori est définie, il est possible d'effectuer des prédictions au moyen des GP. Nous décrivons au Chapitre 5 plus en détails l'application du processus gaussien pour la prédiction d'indice magnétique, notamment en prenant en compte la sortie du réseau de neurones comme paramètre de moyenne du GP.

Pour optimiser les hyperparamètres, il est nécessaire de définir des ensembles d'entraînement. Nous montrons au Chapitre 5 l'importance de cette étape pour la prédiction de l'indice magnétique am à long terme.

Dans le cadre des réseaux de neurones, il est conseillé de considérer trois sous-ensembles de données entraînement –test – validation. Pour le développement de processus gaussien, nous considérons un sous-ensemble d'entraînement et un sous-ensemble de test.

3.3. Les méthodes d'évaluation des performances d'un modèle de prédiction

Lorsque différents modèles sont à comparer, il n'est pas évident de répondre à la question « lequel est le meilleur ? ». Cela dépend de la définition du « meilleur », du but spécifique et du contexte du problème [Chatfield, 2000]. D'autres facteurs importants qui doivent être considérés pour choisir un modèle de prédiction sont :

- La précision ou exactitude (accuracy),
- Le coût,
- Le temps de calcul,
- L'analyse experte.

Souvent, la précision du modèle est mise en avant pour comparer les performances du modèle, alors que les autres facteurs sont également à considérer. Si un algorithme propose de meilleures performances, mais qu'il est plus compliqué à implémenter et à interpréter, il serait risqué de se limiter à étudier un algorithme sur la base de ses performances.

Nous présentons alors les critères couramment utilisés pour évaluer les modèles en météorologie de l'espace, ainsi que des critères que nous avons souhaités mettre en place pour étudier plus spécifiquement les performances des modèles et apporter des informations supplémentaires à toute personne souhaitant les utiliser.

3.3.1. L'erreur quadratique moyenne

D'un point de vue statistique, les erreurs de prédiction sont les critères les plus évidents à considérer pour évaluer une prédiction. Une mesure classique est l'erreur quadratique moyenne ou mean squared error (MSE), plus spécifiquement ici sa racine carrée ou RMSE. On définit l'erreur de prédiction du modèle par ϵ_t , et le RMSE avec l'équation (37).

$$\epsilon_t = y_t - \hat{y}_t \tag{37}$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{t=1}^n \epsilon_t^2}$$

avec y la valeur réelle, \hat{y} la valeur prédite et n le nombre de points dans l'ensemble de validation dans le cas des réseaux de neurones, et dans l'ensemble de test dans le cas des processus gaussiens.

3.3.2. Le coefficient de corrélation

Une autre mesure classique est le coefficient de corrélation entre la valeur réelle y_t et la valeur prédite \hat{y}_t . Elle est définie le symbole CC et est décrite par l'équation (38)

$$CC = Cov(y_t, \hat{y}_t) / \sqrt{Var(y_t)Var(\hat{y}_t)}$$

$$Cov(y_t, \hat{y}_t) = E[(y_t - E[y_t])(\hat{y}_t - E[\hat{y}_t])] \quad (39)$$

$$Var(y_t) = E[(y_t - E[y_t])(y_t - E[y_t])^T]$$

avec E l'espérance mathématiques. Cependant, le coefficient de corrélation est fortement dominé par les variations lentes et tend à négliger les variations à court terme.

3.3.3. La matrice de confusion

D'une manière générale, une donnée est un élément parmi un ensemble délivré pour être étudié. L'analyse de données est l'art d'extraire les informations de cet ensemble de données. Lorsqu'on étudie une donnée, notamment la prédiction de données, il existe différents processus à mettre en œuvre afin d'en extraire une information.

Comme nous l'avons expliqué dans les sections précédentes, un modèle de prédiction doit être choisi, paramétré et validé. Parfois chercher le bon modèle est compliqué et lourd en calcul. Le paramètre le plus important est l'erreur, on va chercher à la minimiser au maximum. Nous rappelons alors qu'un bon modèle en apprentissage supervisé est un compromis entre qualité-complexité-efficacité et rapidité-simplicité-facilité. Pour optimiser ce traitement, nous avons décidé de définir des seuils d'activité afin d'évaluer notamment la capacité des réseaux de neurones à effectuer des prédictions en fonction de ces seuils. Les modèles d'apprentissage supervisé sont influencés par la répartition des données. Dans le domaine de la météorologie de l'espace, le souci récurrent est d'avoir un nombre bien plus conséquent de données à temps calme qu'à temps agité comme le montre la Figure 38 décrivant la répartition de l'indice am en fonction du seuil d'activité.

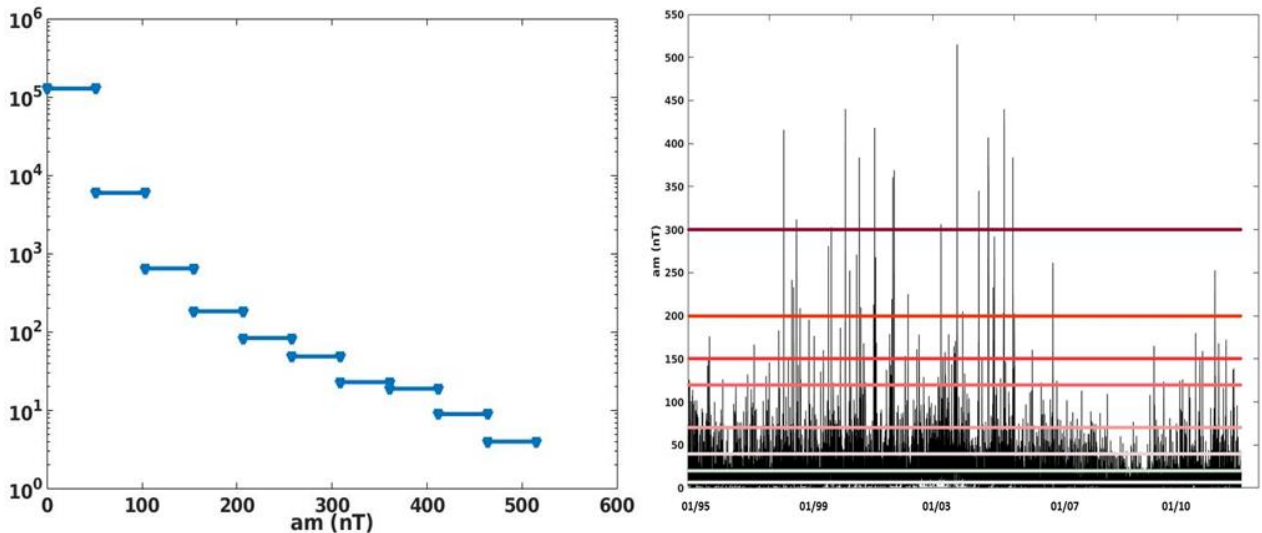


Figure 38- Répartition des am (nombre de points pour un seuil donné) et définition des seuils d'activité.

Nous avons alors défini les seuils décrits sur la Figure 38 et nous avons fait appel à une matrice de confusion pour évaluer la qualité du système de prédiction. Le choix des seuils a été fait principalement sur la base de la distribution des données existantes. La matrice de confusion est présentée sur le Tableau 5. Chaque colonne de la matrice de confusion représente le nombre

d'occurrences d'une classe estimée, tandis que chaque ligne représente le nombre d'occurrences d'une classe réelle. Un des intérêts de la matrice de confusion est qu'elle montre rapidement si le système parvient à prédire correctement. Sur la diagonale on trouve les éléments bien prédits dans un certain domaine d'activité, hors de la diagonale les éléments mal prédits.

Tableau 5- Matrice de confusion.

	Données positives	Données négatives	
Données estimées positives	Vraie positive (a)	Fausse positive (b)	Vraiment prédit
Données estimées négatives	Fausse négative(c)	Vraie négative (d)	Faussement prédit
	Vraiment observé	Faussement observé	Total

Un grand nombre de catégories statistiques est calculé à partir des éléments fournis par la matrice de confusion, afin de décrire certains aspects de performance de la prédiction. Par exemple :

- Précision (pourcentage correct, accuracy)

$$\text{Précision} = \frac{a+d}{total}$$

Cette fraction répond à la question : parmi tous les éléments, quelle fraction des événements prédits est correcte ? Les valeurs vont de 0 à 1, 1 étant la meilleure valeur possible.

C'est simple et intuitif mais ça peut être fortement induit en erreur car influencé par la catégorie la plus commune qui est souvent « pas d'événement ».

- Probabilité de détection POD ou True Positive Rate TPR:

$$POD = \frac{a}{a+c}$$

Avec ce rapport on peut savoir quelle fraction des événements « vrais » est correctement prédite. C'est sensible au vrai positif mais ignore les « faux négatifs » (fausses alarmes). Cependant ce calcul est très sensible à la périodicité des événements, mais il est bon pour les événements rares. Il est intéressant de le mettre en relation avec le taux de fausses alarmes. Les valeurs vont de 0 à 1, 1 étant la meilleure valeur possible.

Dans le cadre du développement de courbes ROC que nous explicitons au Chapitre 5 pour l'analyse de processus gaussien, nous parlons de True Positive Rate ou TPR. Le pendant du TPR est le FPR et est défini comme :

$$FPR = \frac{b}{b+d}$$

- False Alarm Ratio FAR (taux de fausses alarmes)

$$FAR = \frac{b}{a+b}$$

Cette fraction donne la fraction d'événement prédit qui en réalité ne s'est pas produit. C'est sensible aux fausses alarmes mais pas aux événements manqués (c). Les valeurs vont de 0 à 1, cette fois 0 est la meilleure valeur possible.

4. BILAN SUR METHODES ET MATERIELS

Dans cette partie nous avons présenté les différentes données et techniques utilisées dans les chapitres détaillés par la suite. Les données considérées proviennent de différentes bases. Pour le vent solaire nous considérons la base de données OMNI et les données fournies par le satellite ACE afin de voir l'importance de la localisation dans l'Espace de ces données sur les performances des méthodes de prédiction considérées. Nous utilisons également les données GPS rendues récemment publiques afin de voir leur application sur la prévision d'orage magnétique. Pour les données magnétiques, nous considérons les données fournies par la base ISGI pour les indices *am* et *am* sectoriels, ainsi que par OMNI pour l'indice *Dst*.

Pour mettre ces données en relation et fournir des modèles de prévision, nous avons présenté tout d'abord trois modèles ayant fait leurs preuves par le passé pour d'autres indices magnétiques comme expliqué dans le Chapitre 1 : le perceptron multicouche, le réseau à retard de temps ou TDNN et le réseau non linéaire autorégressif à valeurs exogènes ou NARX. Nous avons par la suite présenté un réseau encore jamais utilisé en météorologie de l'espace, le réseau à mémoire à court et long terme ou LSTM.

Afin de fournir des prévisions optimales sur du plus long terme et ajouter de l'information au-delà d'une information « single point » fournie par le réseau de neurones, nous avons alors présenté les processus gaussiens utilisés dans nos études. Le couplage de ces deux méthodes est étudié au Chapitre 5.

Les prochains chapitres présentent les résultats obtenus à partir de la programmation, l'optimisation, et la comparaison de ces techniques, dans le but de fournir des modèles de prévision toujours plus performants, dans l'intérêt d'un opérateur spatial.

CHAPITRE III

ÉTUDE DE LA CAPACITÉ DES RÉSEAUX DE NEURONES À PRÉDIRE LES EFFETS DU VENT SOLAIRE SUR L'ENVIRONNEMENT MAGNÉTIQUE TERRESTRE

Ce chapitre a pour but de présenter les performances de réseaux de neurones de référence dans le cadre de la prédiction à une heure de l'indice magnétique *am*. Ces réseaux sont le réseau statique « feedforward backpropagation » ou perceptron multicouche, le réseau temporel TDNN et le réseau récurrent NARX. Ils ont été appliqués par le passé à la prédiction d'autres indices magnétiques. Ici nous les avons développés, optimisés puis comparés pour étudier les relations entre les événements solaires et les perturbations magnétiques globales associées mesurées par *am*. Le but principal de cette étude a été de mettre en évidence la capacité du TDNN à prédire l'activité magnétique globale en ne considérant que les paramètres du vent solaire en entrée. En effet, parmi les réseaux de référence, le TDNN est le seul qui ne se base que sur les données fournies par la base OMNI de la NASA, ou celles fournies par le satellite ACE au point de Lagrange L1. Nous avons également défini un historique de temps optimal à considérer pour la prédiction de l'indice magnétique, assimilable au temps caractéristique de la réponse globale de la magnétosphère à une perturbation externe.

1. Analyse de la relation entre paramètres du vent solaire et indices magnétiques	99
1.1. Le coefficient de Kendall pour analyser les liens entre les paramètres du vent solaire et l'indice <i>am</i>	99
1.1.1. Le coefficient de Kendall	99
1.1.2. Comparaison des résultats obtenus en fonction des données considérées.....	101
1.2. Etude de l'historique de temps à considérer en entrée pour optimiser les performances de prédiction.....	104
2. Evaluation des réseaux de neurones à partir des données OMNI et ACE	108
2.1. Mise en évidence de la capacité du Time Delay Neural Network à prédire les effets de l'activité solaire sur l'environnement magnétique terrestre à partir des données OMNI.....	108
2.2. L'impact de l'utilisation des données en temps réel sur les performances des réseaux de neurones	110
2.3. L'analyse au travers d'un événement extrême : l'événement de Juillet 2004.....	113

3. Bilan sur l'étude de la capacité des réseaux de neurones à prédire l'impact de l'activité solaire sur l'environnement magnétique terrestre 119

1. ANALYSE DE LA RELATION ENTRE PARAMETRES DU VENT SOLAIRE ET INDICES MAGNETIQUES

La complexité de la réponse de la magnétosphère au vent solaire a été mise en évidence dans la première partie sur l'étude de l'interaction Soleil-Terre. En effet, afin de la modéliser, les scientifiques ont utilisé différentes approches comme les filtres linéaires, les filtres non linéaires, les fonctions de couplages et les réseaux de neurones. Ces approches ont souligné deux aspects importants de l'étude de cette interaction :

- Le premier aspect concerne les paramètres que l'on souhaite mettre en relation pour analyser le comportement de la magnétosphère. En fonction des études faites par le passé que nous avons présentées au Chapitre 2 sections 4 et 5, les paramètres considérés en entrée varient et n'offrent pas les mêmes informations sur la réponse de la magnétosphère au vent solaire.
- Le second aspect concerne le temps de réponse de la magnétosphère aux perturbations du vent solaire. En fonction de l'indice considéré, ce temps ne sera pas le même. Si l'on considère les indices d'activité aurorale comme AE, les mesures des magnétomètres sont associées aussi bien à des entrées directes de particules dans les cornets polaires, qu'à des injections de particules depuis le côté nuit suite à des phénomènes de reconnexion magnétique dans la queue magnétosphérique. L'indice magnétique global Kp est comme nous l'avons souligné au chapitre 1, section 3.4, basé sur des magnétomètres situés à moyenne latitude et rend compte de l'apport global d'énergie dans la magnétosphère. Lorsqu'un événement solaire perturbe l'environnement magnétosphérique terrestre, en fonction de l'importance de l'événement, le temps de réponse de la magnétosphère ne sera pas le même et peut varier sur des échelles temporelles allant de la minute à l'heure voir à plusieurs jours.

Dans cette étude, nous avons décidé d'effectuer une analyse de sensibilité basée sur le coefficient de Kendall pour définir l'existence d'une relation entre les différents paramètres clefs du vent solaire comme la densité, le champ magnétique, la vitesse avec l'indice magnétique am . Cette analyse est décrite dans la partie 1.1. Dans la partie 1.2, nous présentons l'étude de l'historique de temps des données du vent solaire à considérer en entrée afin d'analyser le temps de réponse de la magnétosphère à une perturbation. Cette analyse est faite à partir de paramètres extraits de la matrice de confusion que nous avons présentée dans le Chapitre 2 section 3.3.3.

1.1. Le coefficient de Kendall pour analyser les liens entre les paramètres du vent solaire et l'indice am

1.1.1. Le coefficient de Kendall

Dans la plupart des études, pour valider l'existence d'un lien entre deux variables comme un paramètre clef du vent solaire et un indice magnétique, une régression linéaire simple est effectuée. La qualité du lien supposé est mesurée par le coefficient de corrélation également appelé coefficient de Pearson [Pearson, 1895] qui varie entre -1 et 1, 1 indiquant une corrélation parfaite. Cependant, il existe des cas pour lesquels une mesure de corrélation sur les valeurs n'est pas la méthode la plus adaptée, notamment si les variables ne suivent pas une loi normale ou une loi gaussienne. La Figure 39.a. présente la distribution des valeurs de la vitesse ou fs pour flow speed, et on constate que cette distribution ne suit pas une loi normale. Ceci a un effet sur l'étude de la relation entre l'indice magnétique am et la vitesse présentée sur la Figure 39.b., avec un nuage de points complexe à analyser notamment dans le but d'en extraire un coefficient de corrélation.

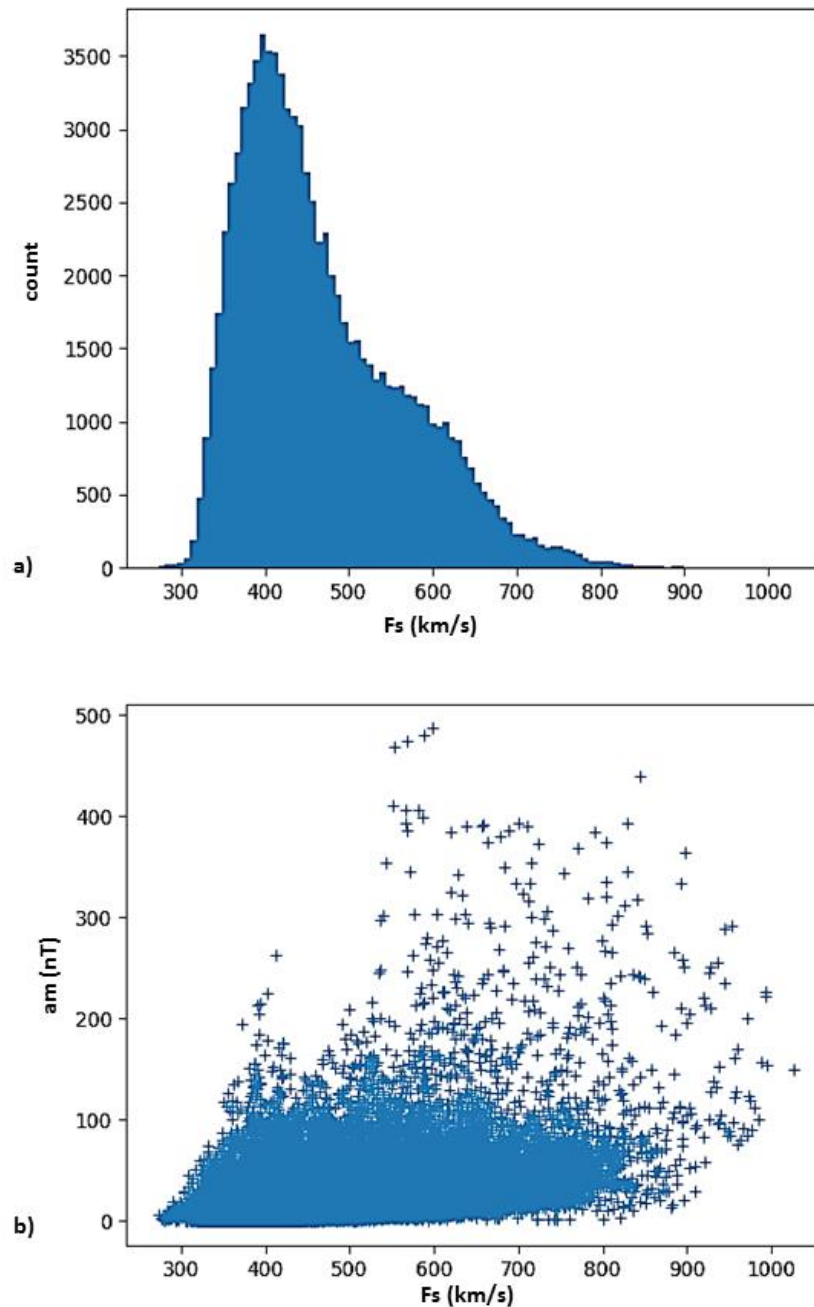


Figure 39- a) Distribution des valeurs de la vitesse (f_s -flow speed) entre 1995 et 2012, b) Tracé de la vitesse en fonction de l'indice magnétique a_m pour des valeurs entre 1995 et 2012.

Il existe d'autres méthodes permettant d'analyser les séries non paramétriques, c'est-à-dire lorsqu'on souhaite analyser un prédicteur ne prenant pas une forme prédéterminée mais qui est construit selon les informations provenant des données. Par exemple il existe la corrélation des rangs. On n'utilise alors pas les valeurs des observations dans les calculs, mais leur rang. Dans notre cas, il existe déjà un tri car nous étudions des séries temporelles. Les valeurs que prend la variable triée sont remplacées par leur numéro d'ordre croissant, et c'est sur ces rangs que la corrélation est effectuée. Avec ce type de méthode, de l'information est perdue. On cherche à savoir dans quelle mesure l'évolution est conjointe ou non. On ne peut pas observer la forme de la liaison, en revanche on perçoit l'existence d'une liaison, qu'elle soit monotone ou non-linéaire par exemple. Il existe plusieurs techniques concurrentes pour valider les résultats de cette étude sur les rangs, l'une d'elle est le coefficient de Kendall [Kendall, 1938]. Pour ce faire, on établit une statistique τ (ou « tau ») en triant une première série de

données (par exemple la vitesse) et en la comparant à une seconde série de données (par exemple l'indice magnétique am). Ensuite, pour chaque rang de la seconde colonne, on compte le nombre de valeurs suivantes qui lui sont supérieures (on attribue +1) et inférieures (-1). On obtient alors une nouvelle série de chiffres qui sont des soldes entre des nombres positifs et négatifs. La somme de ces soldes S est ensuite comparée au solde maximal qui serait obtenu avec une parfaite corrélation. Ainsi, plus la valeur du coefficient de Kendall décrit par l'équation (40) est proche de 1 ou -1, plus il existe une liaison entre les paramètres considérés

$$\tau = \frac{2S}{n(n-1)} \quad (40)$$

avec S la solde totale et n entier naturel (taille des séries).

Pour infirmer ou non l'hypothèse d'indépendance, on effectue ensuite un test paramétrique car ce coefficient est une variable aléatoire qui suit une loi normale sous des conditions très peu restrictives, d'espérance nulle pour des séries indépendantes.

1.1.2. Comparaison des résultats obtenus en fonction des données considérées

Le coefficient de Kendall décrit à la section 1.1.1 est appliqué pour étudier l'existence d'une relation entre des paramètres du vent solaire et l'indice magnétique am . Nous avons choisi les paramètres du vent solaire suivant : le champ magnétique interplanétaire (ou $IMF |B|$), les trois composantes associées B_x, B_y, B_z , la densité n , la vitesse fs (ou flow speed), les trois composantes associées

V_x, V_y, V_z , l'IMF clock angle θ et $B_t = \sqrt{B_x^2 + B_y^2}$.

Les données utilisées sont les données fournies par la base OMNI. Pour faire cette étude, nous avons considéré deux cas, associés au traitement des données du vent solaire :

- un premier cas où nous n'avons conservé que les périodes contenant des données, nous avons supprimé de la base de données toutes les périodes pour lesquelles certains paramètres n'étaient pas définis. En effet, une des problématiques lorsqu'on effectue une analyse à partir des paramètres du vent solaire est qu'il existe des données manquantes (cette problématique est décrite dans la section 1.3 du Chapitre 2).
- Un deuxième cas où nous avons interpolé les données manquantes, afin de voir comment l'interpolation de données pour combler les trous de données pouvait avoir un impact sur l'existence d'un lien entre les paramètres du vent solaire et l'indice magnétique am . Ce cas est important à considérer car pour développer et analyser les réseaux de neurones dans cette étude, nous avons interpolé les données manquantes.

Ainsi, dans un premier temps, nous nous sommes d'abord affranchis des trous de données dans la base existante. Les résultats de l'analyse du coefficient de Kendall sont présentés dans le Tableau 6. Les trois paramètres qui ressortent principalement de cette étude sont la vitesse, la composante V_x de la vitesse et l' $IMF |B|$ avec des valeurs présentées dans le Tableau 6 respectivement égales à 0.420, 0.418 et 0.415. Le coefficient de Kendall entre l'indice magnétique am et la densité n dans ce cas-là est moins élevé, montrant ainsi qu'il existe une relation entre ces deux paramètres et l'indice magnétique am mais qu'elle n'est pas autant explicite qu'entre les paramètres cités précédemment.

Une fois que cette première analyse a été faite, nous avons relancé l'algorithme pour voir les résultats obtenus en considérant la base de données avec les données manquantes interpolées. Ce traitement

n 'est pas anodin et va principalement concerner les cas d'événements extrêmes. En effet, c'est lors de ces événements que les détecteurs à bord des satellites présenteront des faiblesses, et pour lesquels les scientifiques traitant les bases de données comme OMNI et ACE seront amenés à définir les paramètres du vent solaire associés comme inexistantes. Les résultats de l'analyse des effets de l'interpolation des valeurs manquantes sur l'évolution du coefficient de Kendall sont présentés dans le Tableau 7.

Tableau 6- Coefficient de Kendall calculé entre les séries temporelles d'indices magnétiques et des paramètres du vent solaire explicités dans la colonne de gauche en supprimant les trous de données.

Paramètre	Coefficient de Kendall
F_s	0.420
V_x	0.418
IMF B 	0.415
B_t	0.276
B_z	0.189
V_y	0.094
V_z	0.0301
N	0.0234
B_x	0.0193
B_y	0.0174
Theta θ	0.00420

Les trois paramètres qui ressortent principalement de cette analyse sont comme précédemment la vitesse f_s , la composante V_x de la vitesse et l' $IMF |B|$ avec pour coefficient respectivement 0.344, 0.346 et 0.371. Ainsi, en interpolant les données manquantes, l' $IMF |B|$ est le paramètre du vent solaire pour lequel l'existence d'une relation avec l'indice magnétique am est la plus explicite. On constate également que la densité joue un rôle plus important en comparaison avec le cas où les données manquantes sont supprimées, avec un coefficient égal ici à 0.122 contre 0.0234 précédemment. Grâce à cette comparaison, nous soulignons le rôle de la densité n dans l'analyse d'événement extrême, ce paramètre explicitant davantage une relation avec l'indice magnétique am lorsque les données associées à ces événements ne sont pas supprimées mais interpolées.

Tableau 7- Coefficient de Kendall calculé entre les séries temporelles d'indices magnétiques et des paramètres du vent solaire explicités dans la colonne de gauche en interpolant les trous de données.

Paramètre	Coefficient de Kendall
IMF B 	0.371
F_s	0.346
V_x	0.344
B_t	0.234
B_z	0.215
N	0.122
V_y	0.101
V_z	0.0484
B_x	0.0165
B_y	0.0141
Theta θ	0.00418

Dans les deux cas, les paramètres présentant une relation faible avec l'indice magnétique am sont les composantes B_x et B_y du champ magnétique interplanétaire, et l'angle horaire ou IMF clock angle θ . Ce dernier résultat est à souligner car dans les analyses faites sur l'entrée en énergie dans la magnétosphère associée aux événements solaires comme celle de [Akasofu, 1981] explicitée avec l'équation (41a) et celle de [Wang et al., 2014] avec l'équation (42b), l'angle θ est un paramètre clef.

$$\text{Equation d'Akasofu [1978]} \quad \epsilon(t) = \frac{|E(t)| |B(t)|}{4\pi} (L_0 \sin^2 \frac{\theta}{2})^{2|W|} \quad (41a)$$

$$\text{Equation de Wang [2014]} \quad E_{in} = 3,78.10^7 n^{0,24} f_s^{1,47} B_T^{0,86} (\sin^{2,7} \frac{\theta}{2} + 0,25) \quad (42b)$$

avec $l_0 \cong 7R_e$ (rayon terrestre), E le champ électrique en V/m, n la densité en cm^{-3} , f_s la vitesse du vent solaire en km/s, $B_T = \sqrt{B_x^2 + B_y^2}$ en nT.

Ces fonctions de couplage ont été développées en considérant des interactions spécifiques du vent solaire avec l'environnement terrestre comme l'injection de particules dans le courant annulaire, la dissipation Joule dans l'ionosphère et l'injection de particules aurorales. Ainsi, le but d'une fonction de couplage est de définir une fonction permettant d'évaluer l'entrée en énergie associée au vent solaire et sa répartition à travers les différents systèmes de courant. Ceci est fait en analysant l'évolution des phénomènes physiques explicités précédemment en fonction des paramètres du vent solaire, et non pas directement l'évolution des variations globales du champ magnétique en fonction des paramètres du vent solaire. Cette fonction peut par exemple être utilisée pour évaluer par la suite un indice magnétique, qui rend compte des perturbations du champ magnétique. Dans le cas de l'analyse de sensibilité faite ici à partir du coefficient de Kendall, nous étudions l'existence d'une relation directe entre les paramètres du vent solaire et l'indice magnétique am , afin de définir les paramètres d'entrée importants pour les réseaux de neurones. Ce point est à souligner afin d'éviter une confusion sur l'interprétation des résultats fournis par l'analyse du coefficient de Kendall.

Grâce à cette étude, nous avons pu mettre en évidence l'existence d'une relation entre l' $IMF |B|$ et am , ainsi qu'entre la vitesse f_s et am . Ces deux paramètres seront donc utilisés en entrée des réseaux de neurones développés par la suite. Afin d'éviter une éventuelle redondance des informations associées au fait d'utiliser par exemple la vitesse f_s et sa composante V_x , nous faisons le choix de ne conserver que f_s . Et ce même si V_x présente un coefficient de Kendall élevé. De même, B_t et B_z ne sont pas conservés étant donné que l' $IMF |B|$ est utilisé. En revanche, nous décidons de considérer la densité n car le coefficient de Kendall augmente lorsqu'on interpole les données manquantes, et que pour l'entraînement et l'optimisation des réseaux de neurones, nous considérons des bases pour lesquelles une interpolation est effectuée.

Au final, trois paramètres sont conservés pour une première étude : la vitesse f_s , l' $IMF |B|$ et la densité n . Si nous avons fait le choix de ne pas considérer un grand nombre de paramètres en entrée, c'est parce que le but des réseaux de neurones développés dans le cadre de cette étude est de fournir des prédictions de l'indice magnétique am , tout en étudiant la dépendance de am avec des paramètres du vent solaire, en fonction d'un historique de temps. Dans l'annexe 1, nous analysons l'apport des paramètres V_x et B_z sur les performances des réseaux.

1.2. Etude de l'historique de temps à considérer en entrée pour optimiser les performances de prédiction

A la section précédente, nous avons défini à partir du coefficient de Kendall, trois paramètres du vent solaire présentant une relation avec l'indice magnétique am , $\{n, fs, IMF|B|\}$. Maintenant que des paramètres d'entrée ont été fixés, pour optimiser les performances des réseaux de neurones, il faut définir un historique de temps à considérer pour ces paramètres d'entrée. Cette analyse est faite à partir des données OMNI.

Comme expliqué au Chapitre 2, pour évaluer les performances des réseaux, nous faisons le choix de ne pas nous baser sur des mesures largement utilisées par la communauté, comme sont le coefficient de corrélation CC et l'erreur quadratique moyenne ou $RMSE$ (root mean square error). Nous nous basons sur une matrice de confusion.

A partir de la matrice de confusion, nous extrayons deux paramètres statistiques, la probabilité de détection (POD) et le taux de fausses alarmes (FAR). Ces paramètres sont calculés en fonction du seuil d'activité défini au Chapitre 2 section 3.3.3. Grâce à la POD, nous pouvons définir le fait qu'un événement a eu lieu, et que le réseau a été capable de le définir dans le bon domaine d'activité magnétique. Son pendant est le FAR qui permet de rendre compte du fait que le réseau a prédit un événement dans le mauvais domaine d'activité. Ces mesures sont explicitées plus largement dans le Chapitre 2 section 3.3.3.

Comme nous l'avons expliqué précédemment, le but de ce chapitre est de définir la capacité des réseaux de neurones à prédire l'indice magnétique am . Nous développons et optimisons dans un premier temps trois réseaux ayant fait leurs preuves pour la prédiction d'autres indices magnétiques : le réseau multilayer perceptron (MLP) également appelé « feedforward backpropagation », le « time delay neural network » (TDNN) ou réseau à retard de temps, ainsi que le réseau récurrent autorégressif à entrées exogènes (NARX). Nous rappelons les structures de ces réseaux sur la Figure 40. Les cadres rouges représentent les paramètres temporels à optimiser. Dans le cas du réseau « feedforward backpropagation », il s'agit de la taille L du vecteur de paramètres du vent solaire. Pour le TDNN on définit la taille de la fenêtre de spécialisation d et pour le NARX la taille de la série exogène m .

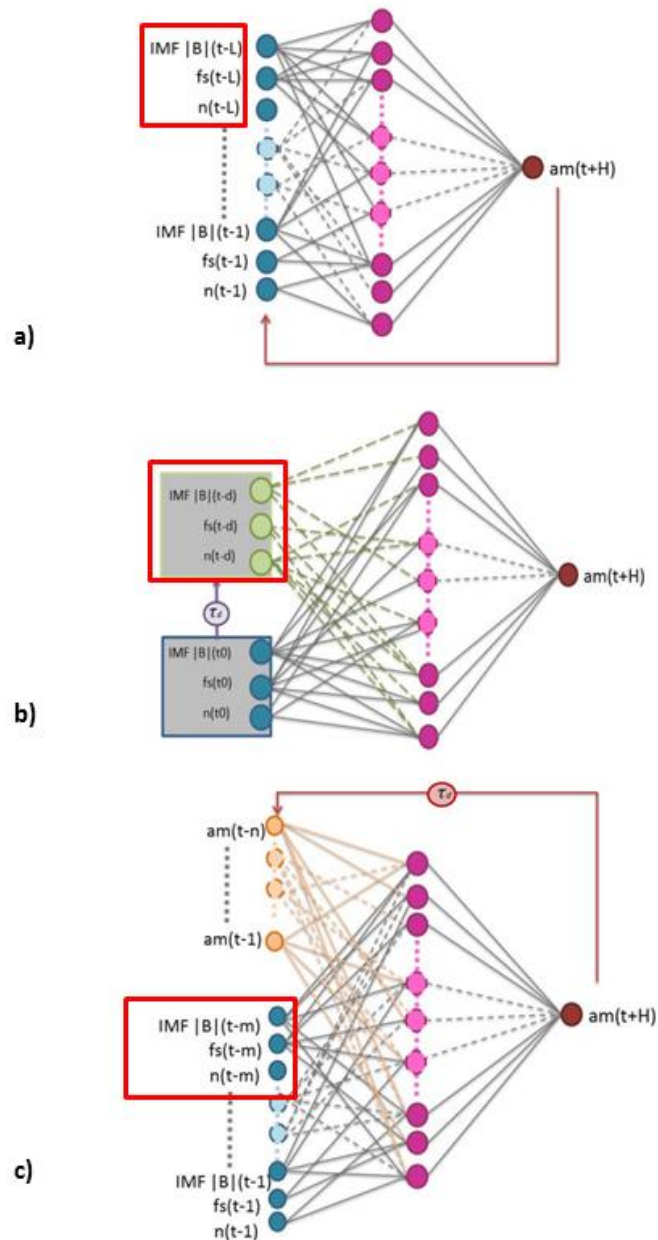
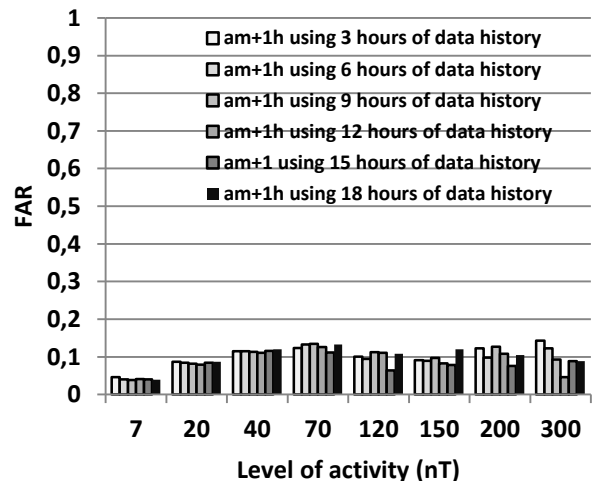
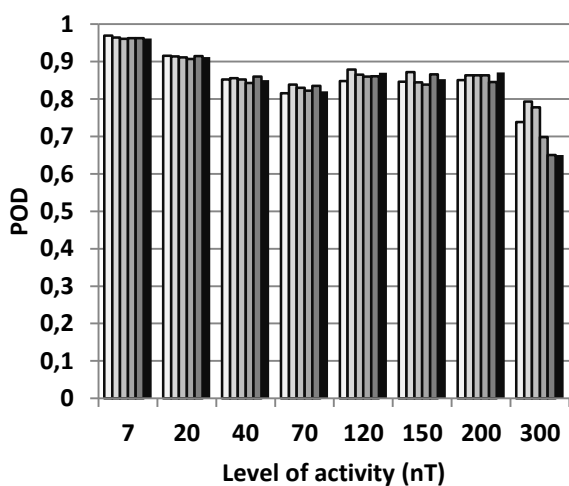


Figure 40- Réseaux considérés pour l'analyse de la capacité de réseaux de neurones à prédire l'indice magnétique am , a) le « feedforward backpropagation », b) le TDNN, c) le NARX. Les cadres rouges soulignent les paramètres temporels à optimiser pour la prédiction.

En développant le réseau TDNN appliqué à la prédiction de Dst , [Gleisner and Lunsdtedt, 1997] ont utilisé une méthode que nous allons également appliquer pour définir les paramètres temporels décrits précédemment (L , d et m) et qui est expliquée ci-après. Dans le but d'optimiser la prévision à une heure de l'indice magnétique am , on se base sur la structure la plus simple, le « feedforward backpropagation ». Ensuite, on fait varier l'historique de temps pour définir la taille L du vecteur des paramètres du vent solaire à considérer en entrée. En comparant les POD et FAR obtenus en fonction de L , on définit un L optimal, et on utilise cette valeur pour définir le paramètre d du TDNN ainsi que le paramètre m du NARX. Cette démarche est utilisée car il existe de nombreux paramètres à optimiser lors du développement de réseaux de neurones. Comme [Gleisner and Lunsdtedt, 1997], nous faisons donc le choix d'optimiser un paramètre sur la structure la plus simple, puis de reporter ce paramètre sur les autres structures.



	POD						FAR					
	3h	6h	9h	12h	15h	18h	3h	6h	9h	12h	15h	18h
7	0,970	0,964	0,961	0,963	0,963	0,962	0,046	0,040	0,039	0,041	0,040	0,039
20	0,916	0,914	0,912	0,907	0,915	0,912	0,087	0,084	0,082	0,080	0,085	0,087
40	0,852	0,856	0,852	0,843	0,860	0,850	0,115	0,115	0,113	0,111	0,116	0,121
70	0,815	0,839	0,830	0,823	0,835	0,820	0,123	0,133	0,135	0,126	0,112	0,133
120	0,848	0,879	0,865	0,860	0,861	0,870	0,100	0,095	0,113	0,111	0,064	0,108
150	0,846	0,872	0,845	0,839	0,866	0,853	0,091	0,089	0,097	0,082	0,078	0,120
200	0,851	0,864	0,864	0,864	0,846	0,871	0,122	0,098	0,127	0,108	0,076	0,105
300	0,738	0,794	0,778	0,698	0,651	0,651	0,143	0,123	0,093	0,047	0,089	0,089

Figure 41- POD et FAR du réseau « feedforward » en fonction de l'historique de temps du vent solaire considéré.

La Figure 41 nous montre qu'en fonction de l'historique de temps considéré en entrée (L), les performances ne sont pas les mêmes et varient également en fonction du seuil d'activité. Comme nous l'avons montré à la section 3.4.3. du Chapitre 2, la base de données contient davantage de données à temps calme, qu'à temps agité. La base de données joue un rôle fondamental dans l'entraînement d'un réseau de neurones. Ainsi, plus il y a de données pour entraîner le réseau, meilleures sont les performances. On constate cet effet sur la Figure 41 présentant les performances du réseau « feedforward » en fonction du niveau d'activité. Les POD pour des seuils d'activité faible sont bien supérieures à celles associées à des seuils d'activité élevée, notamment pour le seuil le plus élevé (>300 nT). Cet effet est visible également sur les FAR. Ils sont plus faibles à temps calme qu'à temps agité. Pour un opérateur, il est important d'évaluer les capacités du réseau pour les seuils d'activité élevé, cas pour lesquels les risques associés aux technologies et infrastructures humaines sont les plus importants. Nous nous sommes donc focalisés sur le seuil d'activité le plus élevé (>300 nT) pour définir l'historique de temps à considérer, L , dans le cas du réseau « feedforward backpropagation ».

A ce niveau d'activité, on constate que la POD la plus élevée vaut 0.794 (FAR de 0.123) et correspond à un historique de temps de six heures. [Boberg et al., 2000] en développant un réseau « feedforward » pour la prédiction de Kp ont fait un constat similaire. Dans leur étude, il était démontré que lors d'orages magnétiques importants, un historique de temps de six heures réduisait l'erreur entre valeur prédite et valeur réelle, et améliorait la corrélation.

Il est également approprié de considérer l'historique de temps qui minimise le taux de fausses alarmes. Et cet historique correspond à douze heures au lieu de six heures, avec un FAR de 0.047, soit trois fois moins élevé. En revanche, en considérant douze heures, la POD diminue de 0.794 à 0.698. Un utilisateur aurait alors à faire un compromis entre choisir un historique de temps qui offre le plus de chance de prédire de « vrais » événements et choisir celui qui donne le moins de risque de prédire de « faux » événements.

Dans la suite de ce chapitre, nous avons fait le choix de conserver deux valeurs pour le paramètre L égales à six heures et à douze heures pour le réseau « feedforward », et nous définissons la largeur d de la fenêtre de spécialisation du TDNN également à 6h et à 12h, de même pour la largeur m de la série exogène du NARX. Ainsi, dans les sections suivantes, nous comparons les performances des réseaux de neurones en considérant ces deux historiques de temps, caractéristiques du temps de réponse de la magnétosphère à une perturbation magnétique globale.

2. EVALUATION DES RESEAUX DE NEURONES A PARTIR DES DONNEES OMNI ET ACE

2.1. Mise en évidence de la capacité du Time Delay Neural Network à prédire les effets de l'activité solaire sur l'environnement magnétique terrestre à partir des données OMNI

Précédemment, nous avons défini un ensemble de paramètres du vent solaire (densité n , vitesse fs et champ magnétique interplanétaire $IMF |B|$) ainsi qu'une fenêtre temporelle de ces paramètres à considérer (six heures et douze heures) afin d'optimiser les prédictions de l'indice magnétique am à une heure. Cette étude a été faite en considérant la base OMNI.

Nous considérons trois réseaux représentés sur la Figure 40 : le réseau « feedforward », le TDNN et le NARX. Les trois réseaux ont en commun d'utiliser en entrée les paramètres du vent solaire. Le TDNN présente la particularité de n'être basé que sur ces paramètres. En effet, c'est le seul réseau parmi les trois qui ne va pas utiliser en entrée l'indice magnétique am prédit par le réseau. Cette spécificité présente un intérêt pour un réseau opérationnel car la valeur définitive de l'indice magnétique am n'est pas définitive immédiatement. Dans un cadre d'utilisation de ces réseaux en temps réel, l'opérateur prend le risque d'utiliser en entrée une valeur erronée avec le réseau « feedforward » ou le réseau NARX. En effet, ces deux réseaux, comme le montre la Figure 40, considèrent également l'indice magnétique prédit, également appelé indice « nowcast ».

Le défi est donc d'étudier la capacité d'un réseau de neurones basé uniquement sur les paramètres du vent solaire à prédire l'activité magnétique globale. C'est un défi car l'indice magnétique est un indicateur de l'état de perturbation de l'environnement magnétique terrestre, et l'état actuel de cet environnement est dépendant des événements passés. S'affranchir de cette information devrait avoir un effet sur la capacité du réseau à prévoir cet indicateur magnétique.

La Figure 42 nous montre les performances obtenues avec les différents réseaux en considérant les données OMNI pour chaque seuil d'activité.

A tout niveau d'activité, le « feedforward » en vert sur la Figure 42 est celui qui présente les moins bonnes performances, en considérant six heures et douze heures d'historique de temps. Si à bas niveau d'activité les performances sont correctes, pour le seuil d'activité le plus élevé (supérieur à 300 nT) les POD sont inférieures à 0.8 (0.794 en considérant un historique des paramètres du vent solaire de six heures, et 0.698 avec un historique de douze heures). Ce réseau a également le FAR le plus élevé à tout niveau d'activité, avec une valeur maximale de 0.133 pour des valeurs d' am comprises entre 70 nT et 120 nT avec six heures d'historique de temps du vent solaire.

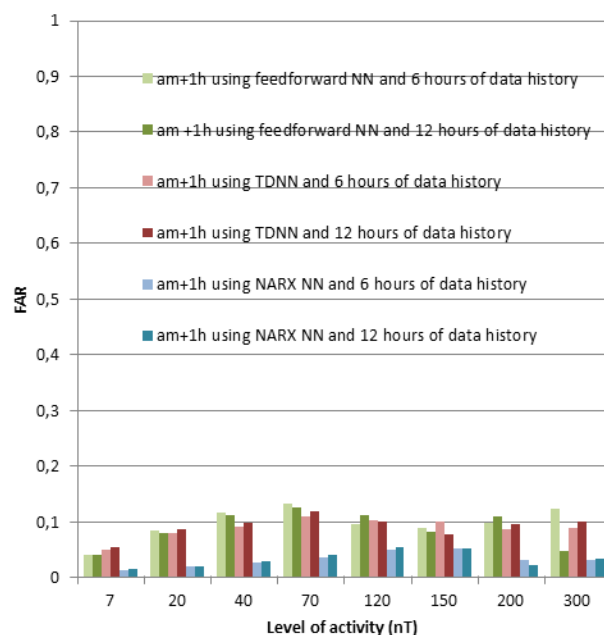
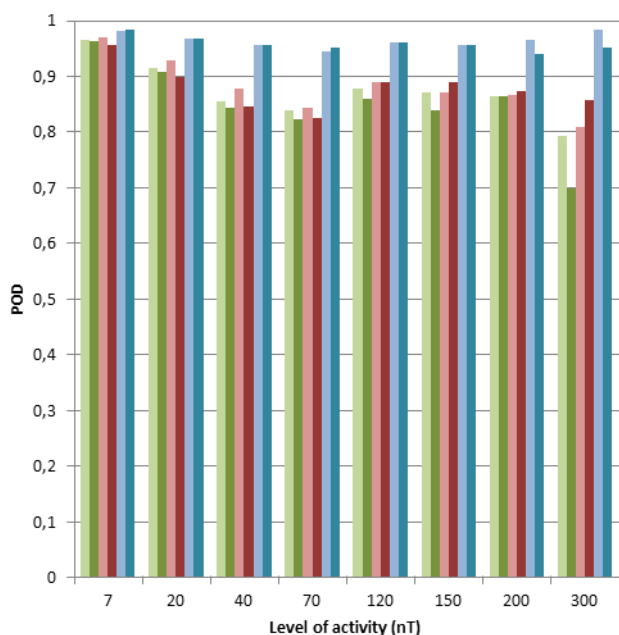
En revanche, le NARX est celui qui offre les prédictions les plus précises, à tout niveau d'activité et pour tout historique de temps, avec une POD comprise entre 0.939 et 0.984 et un FAR compris entre 0.013 et 0.054. Avec de telles performances, ce réseau démontre les forces d'un réseau récurrent, dont l'interaction dynamique entre les couches permet de s'adapter en fonction de l'information entrante.

Le TDNN quant à lui est plus complexe à analyser. Si on considère le seuil d'activité le plus élevé, avec en entrée un historique de temps des paramètres du vent solaire de douze heures, ce réseau présente une POD de 0.857 et un FAR de 0.100, contre une POD de 0.810 et un FAR égal à 0.089 avec un historique de six heures. Si à la section précédente, nous soulignons le compromis à faire entre un historique de six heures et douze heures des paramètres du vent solaire à considérer en entrée, ici nous soulignons les effets que cela peut avoir sur le comportement d'un réseau. Avec douze heures d'historique de temps, le réseau détecte mieux des événements qui vont vraiment avoir lieu, mais il y

a aussi plus de risque de prédire de faux événements. Pour la prédiction de l'indice magnétique am sur la base des paramètres explicités précédemment, le NARX est le réseau le plus performant. Pour mettre en évidence les capacités du TDNN à fournir des prédictions de am , nous le comparons au réseau « feedforward » qui est un réseau largement utilisé par la communauté. A activité faible et moyenne ($am < 120$ nT), le TDNN offre des performances similaires au « feedforward ». Pour les deux réseaux, en termes de POD, l'historique de temps de six heures est le mieux adapté. En termes de FAR, c'est également le cas pour le TDNN. En revanche pour le « feedforward », on observe davantage de variations ce qui rend impossible de définir un historique de temps des paramètres du vent solaire optimal. Lorsque l'activité augmente, le TDNN présente alors de meilleures performances en terme de POD que le réseau « feedforward », notamment pour des valeurs extrêmes d' am . Avec un historique des paramètres d'entrée de douze heures, la POD du TDNN est de 0.857 contre 0.698 pour le « feedforward ». En revanche, le réseau « feedforward » présente un FAR plus faible que celui du TDNN avec une valeur de 0.047 contre 0.100.

Le réseau TDNN, réseau par définition temporel et basé uniquement sur les paramètres du vent solaire, permet donc de prédire l'indice magnétique am avec des performances correctes en comparaison avec celles obtenues avec le réseau de référence « feedforward ». Comme nous l'avons souligné précédemment, le défi avec le TDNN est de s'affranchir de l'information concernant l'état de perturbation dans lequel est la magnétosphère, sachant que son état à l'instant $t + 1$ est dépendant de son état aux instants précédents. Définir un réseau basé uniquement sur les paramètres du vent solaire pour définir un indice magnétique global représente un avantage d'un point de vue opérationnel car on s'affranchit en entrée d'une information dont on ne peut vérifier la fiabilité en temps réel. Ceci permet également d'analyser la relation entre paramètres du vent solaire et indices magnétiques. Etant donné qu'il n'y a que la densité n , la vitesse fs et l' $IMF |B|$ en entrée, on peut alors évaluer la réponse de la magnétosphère directement en fonction de ces paramètres d'entrée. La relation n'est pas biaisée par la valeur de l'indice magnétique « nowcast » grâce à la structure temporel du TDNN.

D'un point de vue strictement opérationnel, le TDNN présente donc l'avantage de n'être basé que sur des paramètres d'entrée dont un opérateur n'aurait pas à se soucier de leurs exactitudes. Cependant, pour obtenir des performances optimales de prévision, le NARX reste le meilleur en termes de POD et de FAR. L'opérateur a donc un compromis à faire entre les paramètres qu'il souhaite utiliser, l'historique de temps à considérer et la complexité du réseau.



	POD						FAR					
	Feed forward NN 6h	Feed forward NN 12h	TDNN 6h	TDNN 12h	NARX NN 6h	NARX NN 12h	Feed forward NN 6h	Feed forward NN 12h	TDNN 6h	TDNN 12h	NARX NN 6h	NARX NN 12h
7	0,964	0,963	0,970	0,957	0,982	0,984	0,040	0,041	0,050	0,055	0,013	0,014
20	0,914	0,907	0,929	0,898	0,968	0,969	0,084	0,080	0,079	0,087	0,019	0,020
40	0,856	0,843	0,877	0,845	0,955	0,956	0,115	0,111	0,090	0,097	0,025	0,028
70	0,839	0,823	0,843	0,826	0,946	0,952	0,133	0,126	0,108	0,119	0,036	0,041
120	0,879	0,860	0,888	0,888	0,961	0,961	0,095	0,111	0,102	0,101	0,049	0,054
150	0,872	0,839	0,871	0,889	0,957	0,957	0,089	0,082	0,099	0,076	0,053	0,053
200	0,864	0,864	0,867	0,872	0,964	0,939	0,098	0,108	0,086	0,095	0,031	0,021
300	0,794	0,698	0,810	0,857	0,984	0,952	0,123	0,047	0,089	0,100	0,031	0,032

Figure 42-POD et FAR de chaque réseau avec les données OMNI en considérant un historique de temps de 6h et 12h. Le réseau « feedforward » est en vert, le TDNN en rouge et le NARX en bleu.

2.2. L'impact de l'utilisation des données en temps réel sur les performances des réseaux de neurones

Après avoir optimisé les réseaux en utilisant les données OMNI, nous avons effectué ce processus en utilisant les paramètres fournis par le satellite ACE situé au point de Lagrange L1. Ces données ont déjà été utilisées par le passé, par exemple [Costello, 1998] les a utilisées pour développer un réseau « feedforward » basé sur les paramètres du vent solaire pour prédire Kp une heure en avance. [Wing et al. 2005] ont également utilisé ces données en développant des modèles de prédiction de Kp pour la NOAA (<http://www.spc.noaa.gov/products/wing-kp>).

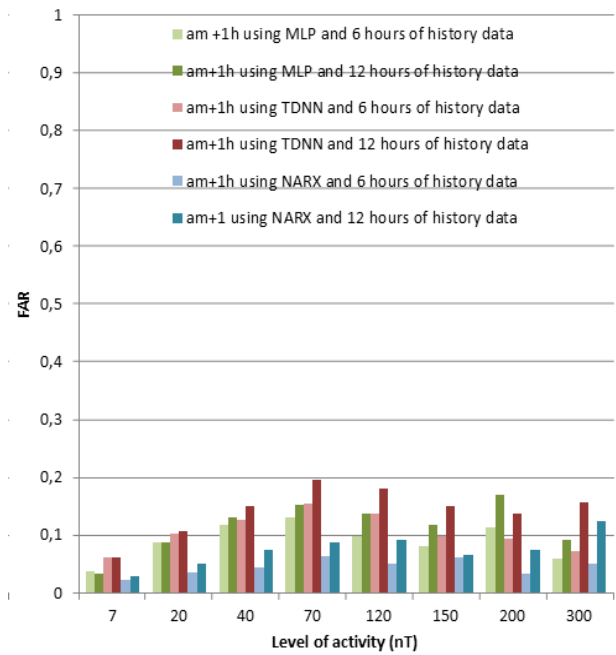
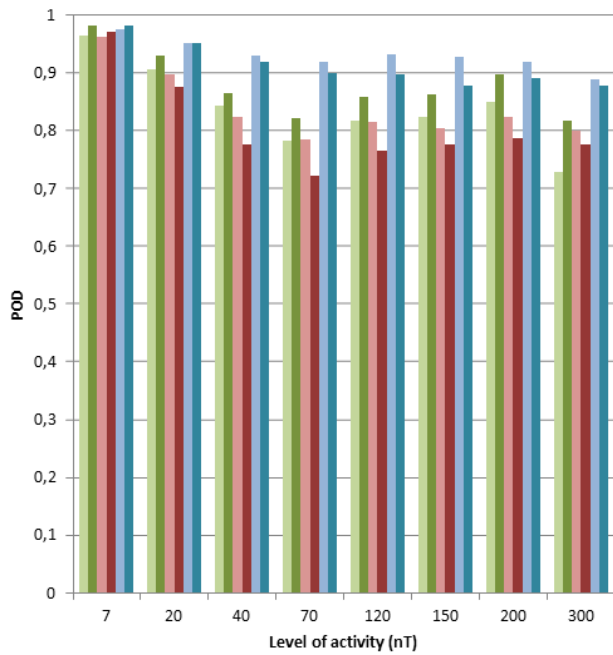
L'intérêt de ces données réside dans le fait qu'elles sont fournies en L1, c'est-à-dire bien en amont de la magnétosphère. Les données OMNI sont fournies au niveau de l'onde de choc, cette localisation étant définie par un prétraitement fait par l'équipe scientifique de SPDF de la NASA. Afin de développer un modèle indépendant des divers prétraitements à la base d'OMNI, et calculé à partir de

données enregistrées directement au niveau d'un satellite d'observation solaire, nous avons fait le choix de nous tourner vers le satellite ACE. Dans nos calculs nous avons pris en compte la différence dans l'horizon de prédiction entre les données ACE et les données OMNI.

Si en nous tournant vers des données fournies directement par le satellite ACE nous gagnons l'avantage de prétraiter celles-ci suivant un mode opératoire que nous définissons, nous perdons un avantage clef d'OMNI qui est le traitement des données manquantes. En effet, lorsqu'un événement extrême conduit à une libération importante de particules énergétiques, les détecteurs à bord des satellites vont être saturés et les données seront manquantes. Les scientifiques à l'origine de la base de données OMNI recourent alors les données manquantes avec des données provenant d'autres satellites comme WIND et IMP8 avant de les faire converger au niveau de l'onde de choc.

Pour notre étude, nous ne conservons que les données fournies par le satellite ACE. Les données sont comme nous l'avons décrit au Chapitre 2 section 1.2 considérées à partir du 5 février 1998. Ceci a déjà pour effet de réduire la base de données de deux ans, et comme nous l'avons vu précédemment, plus la base de données est conséquente, meilleur est l'apprentissage. Nous avons souhaité vérifier cette hypothèse, et voir comment les réseaux « feedforward backpropagation », TDNN et NARX se comportent lorsque les données ne sont plus fournies par OMNI au niveau de l'onde de choc, mais par ACE au niveau du point de Lagrange 1.

La Figure 43 nous montre les performances des différents réseaux pour chaque niveau d'activité en utilisant les données ACE. En comparant la Figure 42 et la Figure 43, on constate une différence de performances globales associées au changement de la base de données. Les performances de chaque réseaux sont ainsi moins élevées en utilisant les données ACE qu'en utilisant les données OMNI. Ce constat souligne l'importance de la base d'entraînement pour un réseau de neurones. En dehors de la réduction de deux ans de la base de données, on a également une perte de données notamment en cas d'événement extrême. La Figure 44 par exemple présente les données fournies par OMNI lors de l'orage d'octobre 2003 également appelé Halloween Storm. Les données associées à la densité et la vitesse sont remplacées respectivement par 999.9 et 9999, indiquant l'absence de données pour ces instants.



	POD						FAR					
	Feed forward NN 6h	Feed forward NN 12h	TDNN 6h	TDNN 12h	NARX NN 6h	NARX NN 12h	Feed forward NN 6h	Feed forward NN 12h	TDNN 6h	TDNN 12h	NARX NN 6h	NARX NN 12h
7	0,965	0,982	0,962	0,971	0,976	0,982	0,036	0,033	0,062	0,060	0,022	0,029
20	0,906	0,929	0,898	0,875	0,952	0,951	0,088	0,087	0,103	0,106	0,035	0,050
40	0,843	0,864	0,824	0,777	0,929	0,920	0,117	0,130	0,127	0,149	0,044	0,074
70	0,782	0,822	0,784	0,721	0,919	0,900	0,131	0,152	0,155	0,195	0,062	0,086
120	0,816	0,858	0,814	0,765	0,932	0,896	0,098	0,136	0,136	0,179	0,051	0,091
150	0,824	0,863	0,805	0,776	0,927	0,877	0,081	0,117	0,098	0,149	0,061	0,064
200	0,850	0,896	0,824	0,786	0,918	0,890	0,112	0,169	0,094	0,136	0,032	0,074
300	0,727	0,816	0,800	0,776	0,889	0,878	0,059	0,091	0,071	0,156	0,051	0,122

Figure 43- POD et FAR de chaque réseau avec les données ACE au point L1 en considérant un historique de temps de 6h et 12h. Le réseau « feedforward » est en vert, le TDNN en rouge et le NARX en bleu.

Pour cette étude, comme pour l'analyse effectuée avec les données OMNI, nous étudions les performances avec un historique de temps de six heures, et un historique de douze heures. On constate qu'en prenant les données au niveau L1, le réseau « feedforward » offre de meilleures performances que le TDNN à seuil d'activité élevé. En considérant un historique de temps de douze heures, la POD du « feedforward » est de 0.816 (FAR = 0.091) et celle du TDNN est de 0.776 (FAR=0.156).

Le réseau NARX est celui qui continue à offrir les meilleures performances de prédiction, mais à seuil d'activité élevé (>300 nT) la POD diminue de 0.984 à 0.889 et le FAR augmente de 0.031 à 0.051 en utilisant ACE au lieu d'OMNI.

2003	301	8	16.9	7.4	612.
2003	301	9	15.4	4.2	653.
2003	301	10	12.9	1.4	759.
2003	301	11	11.4	1.7	767.
2003	301	12	11.5	1.6	800.
2003	301	13	10.7	1.8	809.
2003	301	14	10.4	999.9	9999.
2003	301	15	9.0	999.9	9999.
2003	301	16	9.3	999.9	9999.
2003	301	17	9.3	999.9	9999.
2003	301	18	9.3	999.9	9999.
2003	301	19	9.7	999.9	9999.

Figure 44- Exemple de données fournies par la base OMNI, de gauche à droite : année, jour de l'année (day), heure, IMF|B|, densité et vitesse.

Lorsqu'on a utilisé les données OMNI en entrée, il était difficile de définir un seul historique de temps idéal à considérer en entrée des réseaux de neurones, et ce notamment dans le cas du TDNN. En utilisant les données ACE, l'analyse des résultats de ce réseau montre que l'historique de temps idéal à considérer en entrée est de six heures. En effet, à tout niveau d'activité, la POD est plus élevée, et le FAR plus bas que dans le cas où on considère douze heures d'historique de temps. Par exemple, pour les *am* supérieurs à 300 nT, la POD vaut 0.800 (et le FAR vaut 0.071) en considérant six heures d'historique, contre une POD de 0.776 (et un FAR de 0.156) pour un historique de douze heures.

Globalement, à tout seuil d'activité et tout historique de temps confondus, le TDNN est le moins performant des trois réseaux lorsqu'on considère les données fournies par le satellite ACE. Cette baisse de performance est principalement liée à la définition de ce réseau temporel basé uniquement sur les paramètres du vent solaire. Les autres réseaux ont une partie autorégressive, dans le cas du réseau « feedforward » c'est une partie que nous avons ajoutée, dans le cas du NARX c'est dans la définition même du réseau. Cet aspect autorégressif permet à ces réseaux de continuer à avoir une information pour prévoir l'indice magnétique *am*, même en cas d'absence de données du vent solaire. Cette faiblesse du TDNN est à prendre en compte pour définir des réseaux basés uniquement sur les paramètres du vent solaire qui soient plus robustes lors d'événements extrêmes.

Ces réseaux ont été comparés en utilisant des mesures statistiques, et ce pour chaque seuil d'activité. Afin d'évaluer les performances de ces réseaux concrètement, nous les avons testés sur différents événements extrêmes et ainsi mis en valeur leurs avantages et faiblesses.

2.3. L'analyse au travers d'un événement extrême : l'événement de Juillet 2004

Après avoir comparé les performances globales des réseaux, nous avons souhaité comparer leurs performances sur des cas d'événement extrêmes. Un cas qui nous a tout particulièrement intéressé est l'événement de Juillet 2004, présenté sur la Figure 45. Sur cette figure, on peut voir les caractéristiques du vent solaire que nous utilisons en entrée (*n*, *fs* et *IMF |B|*), les indices *Kp* et *am* fournis par ISGI et le flux intégré d'électrons d'énergie supérieure à 0.3 MeV enregistré par les satellites NOAA POES 15 entre le 1er juillet 2004 et le 1er septembre 2004.

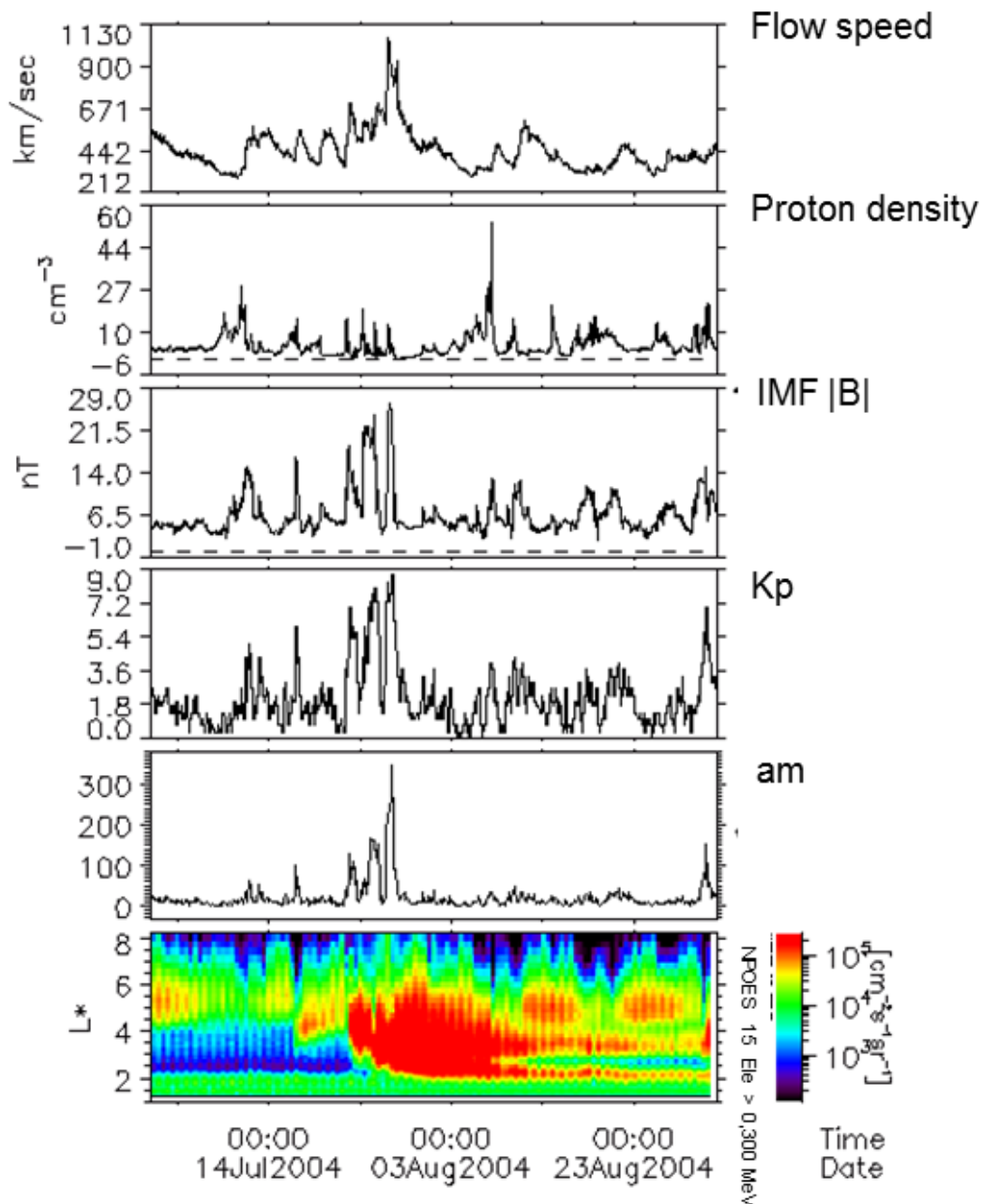


Figure 45- Description de l'événement de Juillet 2004 présenté du 1^{er} juillet au 31 août 2004. De haut en bas : vitesse du vent solaire, densité de proton, IMF |B|, Kp, am et le flux intégré d'électron enregistré par NPOES 15.

Le flux d'électron de hautes énergies en fonction du paramètre de coquille de dérive L^* a été représenté afin de voir comment l'environnement dans lequel évolue les satellites a pu être impacté lors de cette succession d'événements. Des flux importants ont été générés aussi bien à $L^*=6.6$, quand le satellite traverse la ligne de champ magnétique correspondant à l'orbite géostationnaire, jusqu'à $L^*=2$, c'est-à-dire dans la magnétosphère interne.

En observant les variations d'activité aussi bien au niveau du vent solaire qu'au niveau de l'environnement géomagnétique, on peut constater que cet événement est le résultat d'une succession d'événements, avec trois pics notables d'activité pour am jusqu'à atteindre 350 nT. Les orages

géomagnétiques résultats ont été associés à un évènement de type « coronal hole stream » suivi par une CME associée à un nuage magnétique [Kataoka and Miyoshi, 2008]. Cette variation rapide et violente est un excellent cas test pour les réseaux de neurones, afin de voir comment notamment le TDNN basé uniquement sur les paramètres du vent solaire est capable de s'adapter rapidement pour fournir des prédictions optimales.

Nous pouvons noter les différences entre les variations de Kp et celles d' am . Kp est défini sur une échelle logarithmique, mais grâce à la variation de am qui traduit directement les perturbations géomagnétiques, un opérateur peut mieux se représenter les variations de l'activité en fonction du vent solaire. Par exemple, dans le cas de l'évènement de Juillet 2004 présenté sur la Figure 45, quand Kp atteint une valeur de 7, cela correspond aussi bien à un am de 120 nT, 160 nT ou 230 nT. Les effets associés à ces évènements ne sont pas les mêmes du point de vue des flux dans les ceintures de radiation. Au moment du premier pic correspondant à un Kp de 7 et un am de 120 nT, le flux augmente jusqu'à des $L^*=3.5$, tandis qu'au second pic, pour un Kp de 7 et un am de 160 nT, le flux augmente en se rapprochant de $L^*=2$. Cette variation est importante du point de vue des flux de particules énergétiques, et pour cela le Kp comprime la dynamique de la magnétosphère contrairement à l' am .

Nous avons donc voulu observer les performances des différents réseaux, leurs capacités à prédire les variations d'activité, avec un premier pic de 5 nT à 127 nT le 22 juillet, un second de 6 nT à 169 nT le 25 juillet puis un troisième pic de 3 nT à 350 nT le 27 juillet.

La Figure 46 représente les performances des réseaux en utilisant la base de données OMNI avec un historique de temps de six heures. La Figure 46.a. nous montre que le réseau « feedforward » prédit correctement les variations de l'activité mais a parfois tendance à surestimer les pics d'activité. Au lieu de prédire un pic d'activité à 223 nT, il prédit un pic à 268 nT. Sur la Figure 46.b, on constate que le réseau TDNN est moins stable dans ses prédictions, avec davantage de fluctuations dans les variations des perturbations du champ magnétique. Il offre des performances moins élevées que le « feedforward » notamment pour les pics d'activité avec une prédiction à 162 nT au lieu de 226 nT. Ceci est probablement dû au fait que ce réseau ne prend en compte en entrée que les paramètres du vent solaire et n'a donc pas l'information sur l'état de la magnétosphère. Comme nous l'avons explicité auparavant, la magnétosphère se comporte comme une capacité qui se charge et se décharge avec le temps. Ainsi, en fonction des évènements son comportement est complexe à prévoir. Si les indices magnétiques posent problème dans leur utilisation en entrée d'un réseau de neurones à cause du fait qu'ils ne sont pas définitifs avant un certain temps, ils fournissent une information précieuse qui permet d'optimiser les prédictions. On constate cet effet avec le NARX sur la Figure 46.c, ce réseau offre les meilleures prédictions possibles en anticipant au mieux les variations du champ magnétique.

Sur la Figure 47, on observe les performances des différents réseaux en utilisant la base de données ACE et un historique de temps de six heures. Comme nous avons pu le constater en comparant les POD et FAR des réseaux en utilisant ces données à différents niveaux d'activité à la section 2.2, les performances de tous les réseaux chutent à cause du manque de données en cas d'évènements extrêmes. En effet, pour montrer l'absence de données, nous avons réalisé une interpolation par spline sur les données de base. On peut ainsi constater avec les prédictions en pointillé sur la Figure 47 que l'absence de données impacte le réseau. Il n'est alors plus capable de prédire l'activité et l'opérateur reste sans information jusqu'à ce que les données du vent solaire soient à nouveau disponibles.

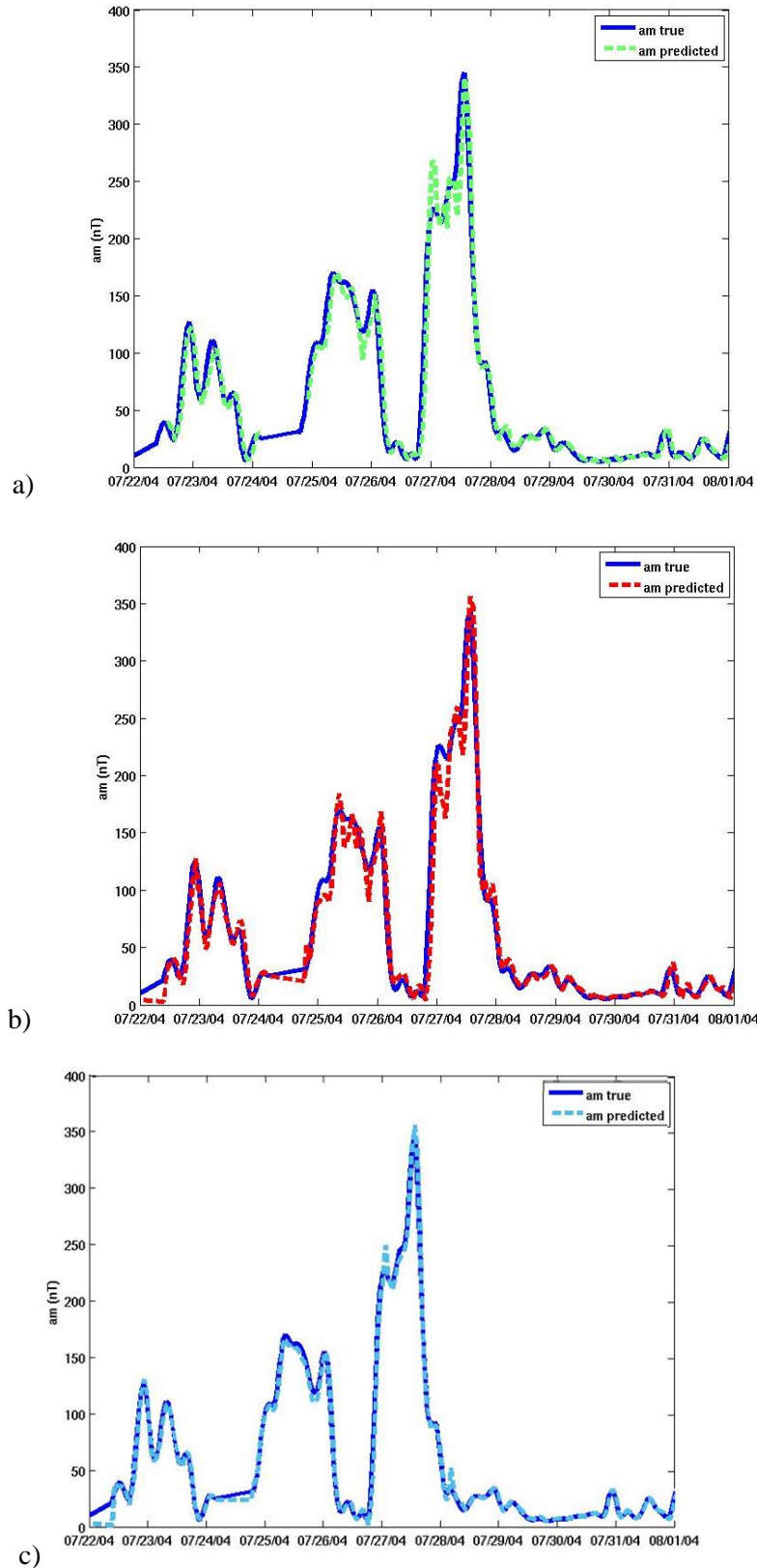


Figure 46- Comparaison entre l'am réel et le am prédit pour l'événement de juillet 2004 avec les données OMNI en utilisant a) le réseau « feedforward », b) le réseau TDNN, c) le réseau NARX. Pour mieux visualiser la prédiction, un décalage d'une heure est utilisée pour la représentation graphique.

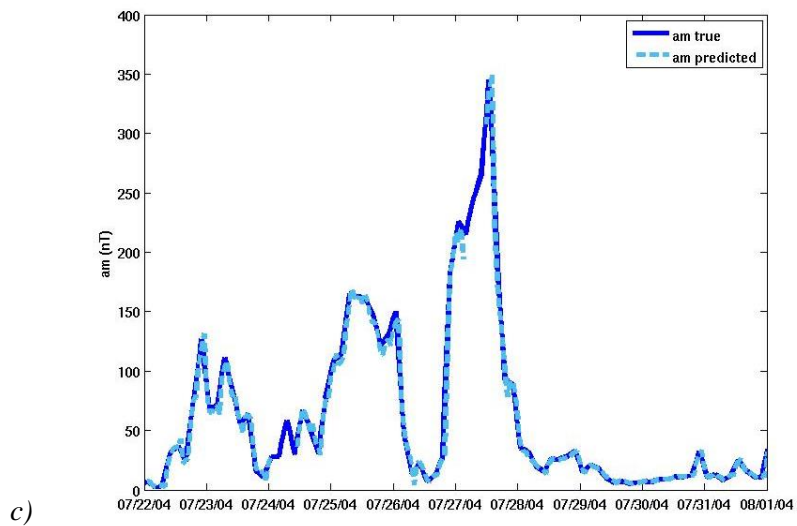
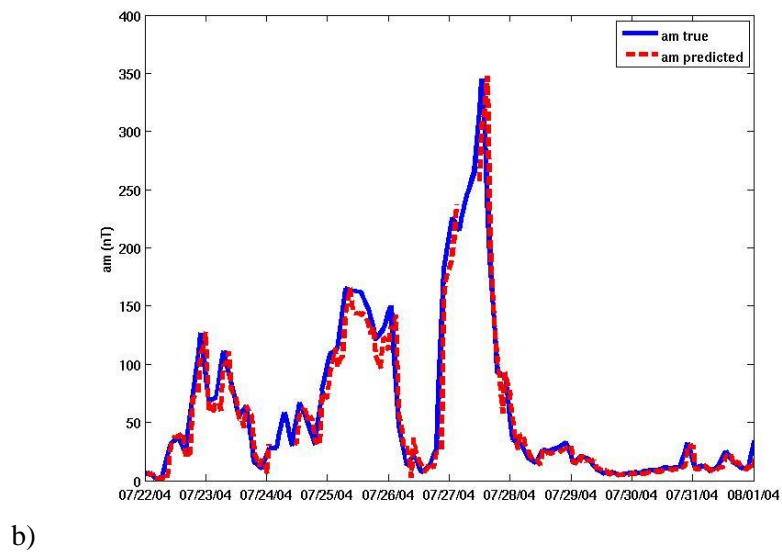
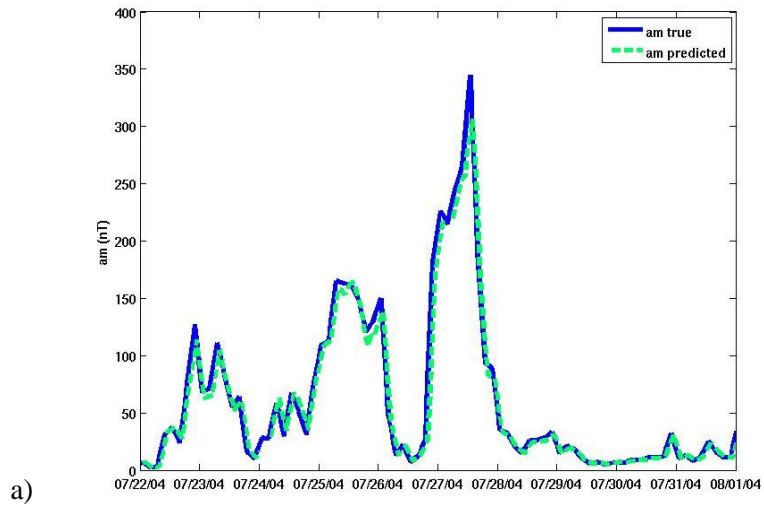


Figure 47- Comparaison entre l'am réel et le am prédit pour l'événement de juillet 2004 avec les données ACE en utilisant a) le réseau « feedforward », b) le réseau TDNN, c) le réseau NARX. Pour mieux visualiser la prédiction, un décalage d'une heure est utilisée pour la représentation graphique.

Comme nous l'avons étudié en faisant une analyse globale des réseaux, le NARX reste tout de même le réseau le plus performant en comparaison avec le « feedforward » et le TDNN. Ce résultat est valable aussi bien avec les données OMNI qu'avec les données fournies par le satellite ACE.

En utilisant les données ACE, on constate que sur un cas d'événement extrême comme celui de Juillet 2004, le TDNN fournit une prédiction similaire au « feedforward » pour le pic d'activité le plus élevé, mais que la prévision fournie par ces réseaux pour le pas de temps suivant n'est pas du tout la même. Le TDNN fournit une prévision de l'indice plus élevée que celle fournie par le « feedforward », ce qui est sûrement lié à la définition même des réseaux. En effet, si les paramètres du vent solaire considérés, ainsi que l'historique de temps jouent un rôle fondamental dans l'analyse des performances des réseaux de neurones, la structure de ceux-ci est également un élément clef. Comme expliqué au Chapitre 2, le réseau « feedforward » est un réseau par définition statique auquel nous appliquons une pseudo-dynamique en réinjectant l'indice prédit ou indice « nowcast » en entrée, ainsi qu'un historique de temps des paramètres du vent solaire en entrée. Le TDNN est un réseau temporel, utilisant la notion de poids partagé, qui rend la couche cachée sensible à la fois aux transitions rapides enregistrées par la couche d'entrée et aux transitions plus lentes enregistrées par la fenêtre de spécialisation. Pour rappel, lorsqu'on parle d'historique de temps dans le cadre du TDNN, on définit la taille de la fenêtre de spécialisation. Les performances du TDNN d'un point de vue global comme nous l'avons étudié dans la section précédente, et d'un point de vue spécifique à un événement extrême comme celui de Juillet 2004 nous montrent qu'il est possible de fournir des prédictions de l'indice magnétique *am* en ne considérant que les paramètres du vent solaire. En utilisant les données fournies directement au niveau du satellite ACE au point de Lagrange L1, en considérant un historique de temps de six heures, les prévisions sont optimisées, mais peuvent manquer de robustesse comparées à un modèle récurrent comme le NARX.

3. BILAN SUR L'ETUDE DE LA CAPACITE DES RESEAUX DE NEURONES A PREDIRE L'IMPACT DE L'ACTIVITE SOLAIRE SUR L'ENVIRONNEMENT MAGNETIQUE TERRESTRE

Cette analyse de la capacité des réseaux de neurones à prédire l'impact de l'activité solaire sur l'environnement magnétique terrestre a été basée sur le développement, l'optimisation et la comparaison de trois réseaux de neurones. Ces trois réseaux sont le réseau perceptron multicouche (MLP) également appelé « feedforward backpropagation », le réseau temporel TDNN et le réseau récurrent NARX. Ces réseaux ont déjà été utilisés par la communauté pour prédire d'autres indices magnétiques, nous avons alors souhaité reprendre ces modèles afin de voir s'ils étaient applicables à l'indice magnétique am . Nous avons également souhaité voir s'ils permettaient de mettre en évidence des relations entre un événement solaire caractérisé par des paramètres fournis par la base OMNI ou le satellite ACE, et l'indice magnétique am représentatif de l'état de perturbation global de l'environnement magnétique terrestre.

Pour étudier les paramètres du vent solaire à considérer en entrée, nous avons analysé le coefficient de Kendall. Ce coefficient basé sur une analyse des rangs permet de mettre en évidence l'existence d'une relation entre deux séries de données. Nous avons alors étudié l'existence d'une relation entre des paramètres du vent solaire et l'indice am . Trois paramètres ont ainsi été identifiés, la vitesse f_s , l' $IMF |B|$ et la densité n . Ces paramètres ont été utilisés en entrée des réseaux développés par la suite. Nous avons fait le choix de considérer les paramètres du vent solaire pris séparément, il serait possible d'analyser à l'avenir les résultats obtenus en utilisant des combinaisons des paramètres du vent solaire.

Cette analyse a été faite avec des données interpolées, une analyse détaillée sur les incertitudes associées aux données OMNI permettrait d'évaluer l'impact de celles-ci sur la composante stochastique générée par l'interpolation.

Nous avons ensuite comparé les performances des trois réseaux de neurones avec les paramètres fournis par la base OMNI. Cette étude a mis en évidence les performances du NARX, supérieures à celles des réseaux « feedforward » et TDNN. Elle a également souligné la capacité du TDNN à fournir des prédictions de l'indice magnétique am , ce qui est un véritable défi étant donné que ce réseau est basé uniquement sur les paramètres du vent solaire.

Afin d'étudier les performances de ces réseaux dans un cadre opérationnel, nous avons considéré les données directement fournies par le satellite ACE situé au point de Lagrange L1. Cette étude nous a principalement permis de comprendre l'impact associé à l'utilisation de ces données. En effet, les scientifiques qui développent la base OMNI fournissent un travail conséquent sur le traitement des données manquantes, ce qui arrive surtout en cas d'événement extrême. La base de données OMNI est plus complète que celle de ACE, et la base de données est un élément clef de l'entraînement d'un réseau de neurones. Plus la base de données est riche, plus les paramètres du réseau de neurones comme les poids et les biais ont la possibilité de s'adapter à différentes situations. Nous avons alors pu constater qu'en utilisant la base ACE, les performances des réseaux diminuaient. Le réseau NARX restait malgré tout le plus performant, et le TDNN a montré sa capacité à fournir des prédictions optimales notamment à seuil d'activité élevé, en comparaison avec le réseau « feedforward ». Nous avons également constaté que pour développer un réseau basé uniquement sur les paramètres du vent solaire comme le TDNN avec les données fournies par le satellite ACE, il faut considérer un historique de temps de six heures.

Cette comparaison de réseaux a été faite globalement avec l'analyse des POD et des FAR pour différents niveaux d'activité, puis a été appliquée à un cas d'événement extrême en Juillet 2004. Le

TDNN a alors montré sa capacité à prédire des variations rapides et importantes de pic d'activité, mais n'atteint pas la précision du réseau NARX.

Le choix des mesures a été fait par rapport à des seuils que nous avons défini, nous pouvons adapter par la suite ces mesures en fonction des besoins des opérateurs. La réponse attendue n'est pas la même en fonction d'un opérateur satellite qui serait intéressé par une erreur relative ou par un prévisionniste radio qui lui souhaiterait connaître l'erreur absolue.

Afin de donner une base de référence de comparaison avec des mesures classiques comme le coefficient de corrélation (CC) et l'erreur quadratique moyenne (RMSE), le Tableau 8 présente les valeurs de ces mesures pour chacun des réseaux présentés dans ce chapitre, en considérant un historique de temps de six heures. Nous présentons également les résultats associés à un réseau « feedforward » qui serait basé uniquement sur les paramètres du vent solaire, sans le « nowcast » index, afin de montrer l'apport du TDNN, pour des mesures globales offertes par la RMSE et le CC. Si on considère les données ACE, le réseau « feedforward » basé uniquement sur les paramètres du vent solaire présente un CC à 0.766 et une RMSE à 14.1, tandis que le TDNN présente un CC à 0.958 et une RMSE à 6.85. Ceci montre l'apport de la fenêtre de spécialisation et de la notion de poids partagé du TDNN pour mieux assimiler la réponse de la magnétosphère à une perturbation externe.

	RMSE		CC	
	OMNI	ACE	OMNI	ACE
Feedforward backpropagation	4.28	5.20	0.983	0.975
Feedforward backpropagation without nowcast index	10.6	14.1	0.912	0.766
TDNN	8.85	6.85	0.913	0.958
NARX	3.32	3.65	0.989	0.988

Tableau 8- Coefficient de corrélation (CC) et erreur quadratique moyenne (RMSE) des différents réseaux utilisés en considérant un historique de temps de six heures.

Le réseau NARX est de tous les réseaux le plus performant, mais sa structure requiert d'avoir en entrée une partie autorégressive, c'est-à-dire de considérer l'indice « nowcast ». Le compromis serait donc de trouver une structure qui aurait les performances d'un réseau récurrent sans liaison autorégressive.

Dans le chapitre suivant, nous présentons le développement d'un nouveau réseau, encore jamais utilisé en météorologie de l'espace, le réseau Long Short Term Memory (LSTM). Ce réseau appartient à la famille des réseaux récurrents, et nous le développons en ne considérant que les paramètres du vent solaire en entrée.

CHAPITRE IV

DÉVELOPPEMENT ET ANALYSE D'UN NOUVEAU RÉSEAU DE NEURONES POUR OPTIMISER LES PRÉDICTIONS DE L'INDICE AM À PARTIR DES PARAMÈTRES DU VENT SOLAIRE

Dans ce chapitre sur le développement et l'analyse d'un nouveau réseau de neurones pour la prédiction de l'indice magnétique am à partir des paramètres du vent solaire, nous avons développé puis optimisé un nouveau réseau, le réseau Long Short Term Memory ou LSTM. Ce réseau n'a encore jamais été utilisé en météorologie de l'espace, et présente l'avantage de n'être basé que sur les paramètres du vent solaire considérés à l'instant présent. Dans cette étude, nous avons tout d'abord étudié le réseau LSTM et son apport concret à la prédiction de l'indice am . Puis nous avons optimisé davantage les prédictions en utilisant en entrée une fonction de couplage permettant de rendre compte de l'entrée en énergie associée à un événement solaire extrême. Enfin, nous avons développé des réseaux plus complexes basés sur le LSTM, afin de fournir une nouvelle information sur l'évolution spatiale d'une perturbation à un opérateur. En effet, nous avons travaillé sur la prédiction de l'indice am sectoriel ou $a\sigma$, spécifique à chaque secteur MLT.

1. Application du réseau Long Short Term memory à notre problématique	123
1.1. Développement du LSTM en Python et première mise en évidence de l'apport de ce réseau en comparaison au réseau de référence	123
1.2. Evaluation des performances de prédiction en fonction des données considérées	125
2. Analyse des effets de l'utilisation de fonction de couplage en entrée du LSTM	129
2.1. Point sur les fonctions de couplage et leurs applications en entrée des réseaux de neurones	129
2.2. L'apport des fonctions de couplage sur les performances des réseaux avec les données ACE.	134
3. Evaluation de la capacité du LSTM à fournir une prédiction multi-sortie dans le cadre de la prédiction de l'indice $a\sigma$	136
3.1. Le rôle de l' am sectoriel ou $a\sigma$ en météorologie de l'espace	136

3.2. Analyse des performances du LSTM pour la prédiction de l'activité magnétique associée à chaque secteur MLT	139
4. Bilan sur le Développement et l'analyse d'un nouveau réseau de neurones pour optimiser les prédictions de l'indice <i>am</i> à partir des paramètres du vent solaire.....	147

1. APPLICATION DU RESEAU LONG SHORT TERM MEMORY A NOTRE PROBLEMATIQUE

Au Chapitre 3, nous avons comparé trois réseaux de neurones de référence afin de déterminer la capacité de ces modèles à prédire l'indice magnétique am . Dans un premier temps, nous avons montré que le réseau récurrent NARX était le plus performant de tous. Il présentait la probabilité de détection la plus élevée et le taux de fausses alarmes le plus faible, et ce à tout niveau d'activité. Cependant, ce réseau possède une partie autorégressive. C'est-à-dire qu'il va considérer en entrée des valeurs passées de l'indice am prédites par le réseau. Ceci peut être problématique dans un cadre opérationnel car les valeurs de l'indice magnétique am ne sont pas définitives immédiatement. Un opérateur prendrait alors le risque, en utilisant ce réseau, de considérer une valeur erronée, le temps que la prévision fournie par celui-ci soit vérifiée. Dans un second temps nous avons mis en évidence la capacité du réseau TDNN à fournir des prédictions de l'indice magnétique am en ne considérant que les paramètres du vent solaire. Ce réseau temporel est moins performant que le réseau récurrent NARX, mais est plus efficace dans la majorité des cas que le réseau « feedforward » largement utilisé dans la communauté.

Après avoir mis en évidence ces deux résultats, nous avons conclu que pour optimiser les prévisions de l'indice magnétique am , en utilisant uniquement les paramètres du vent solaire, il est nécessaire de définir un réseau de neurones alliant la possibilité de ne considérer en entrée que les paramètres du vent solaire, et la performance d'un réseau récurrent. Nous avons alors programmé le réseau récurrent Long Short Term Memory (LSTM), que nous avons décrit dans le Chapitre 2 section 3.1.2. Nous décrivons dans cette partie l'analyse des performances de ce réseau en comparaison aux réseaux précédemment développés, afin de voir l'apport qu'il représente pour la prédiction de l'indice am à une heure.

1.1. Développement du LSTM en Python et première mise en évidence de l'apport de ce réseau en comparaison au réseau de référence

Précédemment, les réseaux « feedforward », TDNN et NARX ont été programmés en Matlab avec la toolbox Neural Network. Le réseau LSTM est programmé en Python, en utilisant la librairie Lasagne et la surcouche de calcul scientifique Theano décrite en annexe 2.1. Afin de comparer les performances du LSTM au réseau utilisé classiquement par la communauté, nous avons également reprogrammé le réseau « feedforward » en Python.

La question du langage de programmation est importante lorsque l'on souhaite développer des codes utilisables par des opérateurs. Au-delà de la question de l'analyse des relations entre les paramètres d'entrée et l'indice à prédire, se pose la question de l'utilisation concrète du programme. Si la première partie de nos travaux était orientée vers l'analyse de l'utilisation et de l'apport des réseaux de neurones pour la prédiction de l'indice am , maintenant que nous avons mis en évidence les propriétés du réseau idéal pour un modèle opérationnel, il est nécessaire de réfléchir à un langage utilisable de façon opérationnel. Dans l'annexe 2, nous explorons en détail les différences entre le langage Matlab et Python, et mettons en évidence l'apport de ce dernier pour fournir un modèle transparent, portable et utilisable par tous.

Il est important de noter que lorsqu'on définit un réseau récurrent, il est plus juste de parler en termes de nombre de cellules que de nombre de neurones. Pour définir le réseau récurrent NARX, nous avons programmé une structure comparable à une structure « feedforward », afin de pouvoir comparer des architectures similaires avec une égalité entre le nombre de neurones cachés pour le réseau « feedforward » et le nombre de cellules pour le LSTM. Ceci avait été fait dans le Chapitre 3 pour

comparer les réseaux « feedforward », TDNN et NARX afin de comprendre notamment comment l'apport de couches spécifiques, comme une couche autorégressive pour le NARX, peut apporter une précision sur la prévision fournie par le réseau. Ceci a également été fait afin d'évaluer l'impact de l'historique de temps sur les performances des réseaux. Dans le cadre du développement du réseau LSTM, nous revenons à la définition même du réseau récurrent qui est le développement en chaîne du réseau. L'information est transmise d'une cellule à l'autre, au moyen d'un convoyeur d'information, afin de pallier au défaut des réseaux récurrents classiques qui est l'évanescence de gradient ou « vanishing gradient » [Hochreiter, 1998] décrite au Chapitre 2 section 3.1.4.2.

Pour cette étude, le réseau « feedforward » possède vingt neurones dans la couche cachée de calculs, tandis que le réseau LSTM est défini sur un enchaînement de vingt cellules de calcul. Le LSTM est alimenté uniquement par les paramètres du vent solaire à l'instant t , tandis que le réseau « feedforward » est alimenté par un historique de temps des données du vent solaire, et par l'indice « nowcast ». Nous allons donc analyser l'apport de la structure du LSTM pour fournir une prédiction de l'indice magnétique am à une heure en ne considérant que les paramètres $\{n, fs, IMF|B|\}$ en entrée.

Pour entraîner ces réseaux, la technique dite de descente de gradient, explicitée dans le Chapitre 2 section 3.1.1.3, est utilisée. Pour entraîner un réseau en utilisant cette technique, il faut que l'approximation de la fonction fournie par ce réseau soit différentiable. Il est bien connu que le réseau « feedforward » est différentiable [Gégout et al., 1995]. Ce mécanisme est également applicable pour les réseaux récurrents. En effet, on peut « dérouler » dans le temps le réseau et obtenir un réseau de type « feedforward » qui partage une matrice de poids entre les couches cachées. Lorsqu'on « déroule » un réseau récurrent et qu'on applique un système d'entraînement de type « backpropagation » pour effectuer une descente de gradient, on applique un algorithme de « Backpropagation Through Time » ou BPTT [Werbos, 1990]. Les cellules à mémoire temporelle comme les LSTM ajoutent des ponts multiplicatifs aux réseaux récurrents classiques, ce qui ne change pas le fait de pouvoir appliquer ce processus d'entraînement.

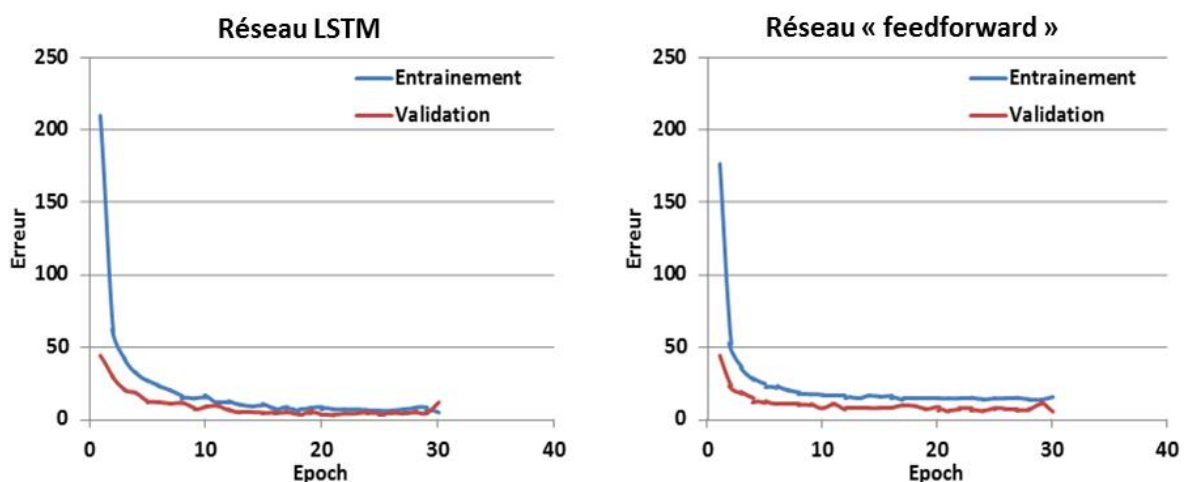


Figure 48- Evolution de l'erreur en fonction des epoch en phase d'entraînement et de validation pour le réseau LSTM à gauche et « feedforward » à droite.

Nous utilisons dans un premier temps les données OMNI pour l'entraînement. La Figure 48 présente l'évolution de l'erreur en fonction des « epoch », c'est-à-dire des périodes associées aux passages des sous-ensemble de données durant les phrases d'entraînement et de validation pour réajuster et valider les poids et biais. La répartition des données entre les sous-ensembles est similaire à celle utilisée au Chapitre 3 illustrée par la Figure 19, c'est-à-dire 70% pour l'entraînement et 10 % pour la validation (les 20 % restants étant associés au sous-ensemble de test). On constate que le réseau LSTM converge plus rapidement que le réseau feedforward et présente une erreur sur le test final moins élevée avec une valeur de 11.39 contre 15.39 pour le réseau « feedforward ». Ainsi, avec cette première analyse nous constatons que le développement en chaîne d'un réseau de neurones permet d'optimiser le processus d'entraînement et d'améliorer les performances finales du réseau. La prédiction de série temporelle est complexe à développer. En effet, les réseaux existants sont adaptés pour de la classification, ou de la régression. Les réseaux utilisés précédemment comme le « feedforward », le TDNN et le NARX permettent de faire de la prédiction en utilisant le principe de régression. Les informations entrantes sont combinées avec des poids puis transformées pour fournir une réponse qui est la valeur prédite par le modèle. Les coefficients du modèle sont calculés durant la phrase d'entraînement afin que les valeurs prédites soient les plus proches possibles des valeurs observées afin de minimiser l'erreur. Pour que ces modèles soient représentatifs du comportement complexe de la magnétosphère, nous avons utilisé des fonctions de transfert non linéaires. Nous avons ainsi dépassé les limites d'une régression. Cependant, contrairement aux modèles de prédiction de type régression, les prédictions de séries temporelles ajoutent le problème de la dépendance entre les variables entrantes. Cette dépendance est prise en compte sur une échelle temporelle restreinte grâce à la fenêtre de spécialisation dans le cas du TDNN. Les réseaux NARX et « feedforward » ont une prise en compte de cette dépendance limitée par l'historique de temps que nous considérons en entrée des données. Dans le cas du réseau LSTM, le développement en chaîne des cellules permet de conserver cette dépendance dans le temps, en alimentant le réseau uniquement avec les paramètres du vent solaire à l'instant présent. D'un point de vue physique, on se rapproche du comportement réel de la magnétosphère qui possède une dynamique interne sur différentes échelles temporelles. Avec le LSTM nous n'avons pas à optimiser cette échelle de temps, la structure du réseau permet de s'adapter en fonction des données que le réseau reçoit. D'un point de vue opérationnel, la plus-value du LSTM est importante car on réduit l'information entrante à traiter aux données arrivant à l'instant présent.

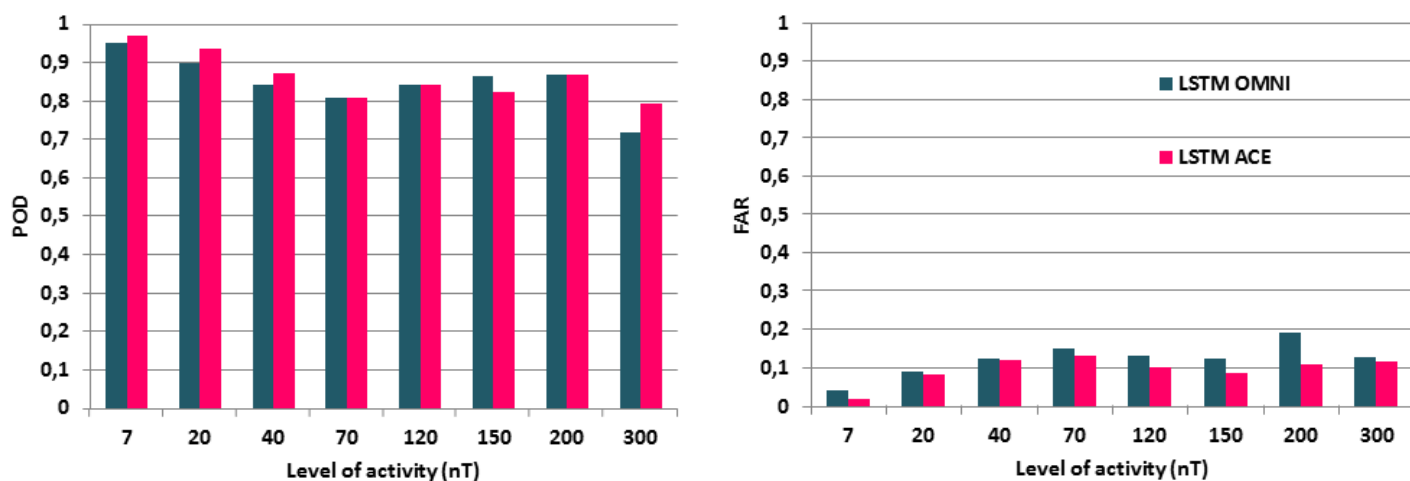
Nous allons alors étudier les performances du LSTM développé en Python en le comparant avec le réseau TDNN développé en Matlab, afin de comprendre comment le LSTM améliore la prédiction de l'indice magnétique am à une heure.

1.2. Evaluation des performances de prédiction en fonction des données considérées

Dans un premier temps, nous avons développé le réseau LSTM en considérant les données OMNI. Nous avons ensuite optimisé à nouveau le réseau en considérant les données ACE. Cette démarche a été effectuée afin de voir l'effet du traitement des données manquantes en Python sur les performances du réseau de neurones. Comme nous l'avons expliqué au Chapitre 2 section 1.3, pour traiter les données manquantes nous avons effectué une interpolation par splines. Cependant, en temps réel, lorsqu'il y aura des données manquantes, il sera impossible d'effectuer une interpolation par spline. Nous avons alors fait le choix de développer un module traitant les données entrantes, et les transformer en « not a number » afin que le réseau puisse continuer à tourner, même en cas de données manquantes. Dans le cadre du développement de réseau opérationnel, ce point est crucial.

La Figure 49 présente les POD et FAR du LSTM en considérant les données OMNI et ACE. En utilisant le traitement des données manquantes décrit précédemment, on constate que l'utilisation des

données ACE impacte peu les performances du réseau, bien au contraire. Si on se focalise sur le niveau d'activité le plus élevé, avec les données ACE, le LSTM présente une POD de 0.791 et un FAR de 0.116, tandis qu'avec les données OMNI on obtient une POD de 0.716 et un FAR de 0.127. Ainsi, avec notre module de prétraitement des données manquantes, on permet au réseau de ne plus être autant impacté que les précédents réseaux par le risque d'absence de données plus élevé avec la base ACE qu'avec OMNI. Le réseau peut être opérationnel avec les données du vent solaire considérées directement au point de Lagrange L1, et s'affranchit d'un prétraitement supplémentaire qui est celui fourni par l'équipe scientifique de SPDF à l'origine d'OMNI.



	POD		FAR	
	LSTM - OMNI	LSTM - ACE	LSTM - OMNI	LSTM - ACE
7	0,951	0,970	0,042	0,0178
20	0,896	0,934	0,0903	0,0812
40	0,842	0,869	0,124	0,119
70	0,807	0,809	0,149	0,131
120	0,841	0,839	0,131	0,101
150	0,863	0,823	0,125	0,0863
200	0,866	0,868	0,193	0,110
300	0,716	0,791	0,127	0,116

Figure 49- POD et FAR du réseau LSTM en considérant la densité, la vitesse et l'IMF|B|. Le LSTM prenant en compte les données OMNI est en bleu, celui considérant les données ACE est en rose en fonction du niveau d'activité.

Le Tableau 9 présente les coefficients de corrélation (CC) et l'erreur quadratique moyenne (RMSE) du LSTM, en considérant d'une part les données OMNI, et d'autre part les données ACE. Le constat est le même que celui obtenu avec les mesures de POD et FAR, en utilisant les données ACE, les performances du LSTM avec notre module de prétraitement ne sont pas fortement impactées, avec par exemple un CC de 0.974 et une RMSE de 4.73 avec les données OMNI contre un CC de 0.975 et une RMSE de 5.82 avec les données ACE.

Tableau 9- Coefficient de corrélation (CC) et erreur quadratique moyenne (RMSE) du LSTM avec les données OMNI et ACE.

	RMSE		CC	
	OMNI	ACE	OMNI	ACE
LSTM	4.73	5.82	0.974	0.975

Les Figure 50 et Figure 51 présentent les prédictions fournies par le LSTM pour l'événement de Juillet 2004 (présenté au Chapitre 3 section 2.3.) en utilisant respectivement les données OMNI et ACE. Cette comparaison confirme les performances présentées en Figure 49. En effet, si on se focalise sur le pic d'activité du 27 juillet présentant une valeur réelle de 246 nT, on constate que pour ce domaine d'activité compris entre 200 et 300 nT, les POD du LSTM en considérant les données OMNI et ACE sont équivalentes, égales respectivement à 0.866 et 0.868. Cependant, le FAR est plus élevé avec les données OMNI (0.193 contre 0.110 avec les données ACE). La Figure 50 illustre cet effet sur la prédiction du pic d'activité entre le 27 et le 28 juillet 2014. On observe une surestimation du pic d'activité avec une valeur prédite à 298 nT au lieu de 246 nT avec les données OMNI, tandis que le LSTM avec les données ACE offre une prédiction proche de la valeur réelle du pic d'activité. En ce qui concerne le pic d'activité le plus élevé à 350 nT, les deux réseaux proposent des prédictions similaires à 300 nT, et ce même si la POD et le FAR obtenus avec les données ACE sont meilleurs que ceux obtenus avec les données OMNI (POD = 0.716 et FAR = 0.127 avec OMNI, POD = 0.791 et FAR = 0.166 avec ACE). Ceci montre alors les limitations du modèle pour prédire l'indice magnétique am en considérant les paramètres du vent solaire $\{n, fs, IMF|B|\}$.

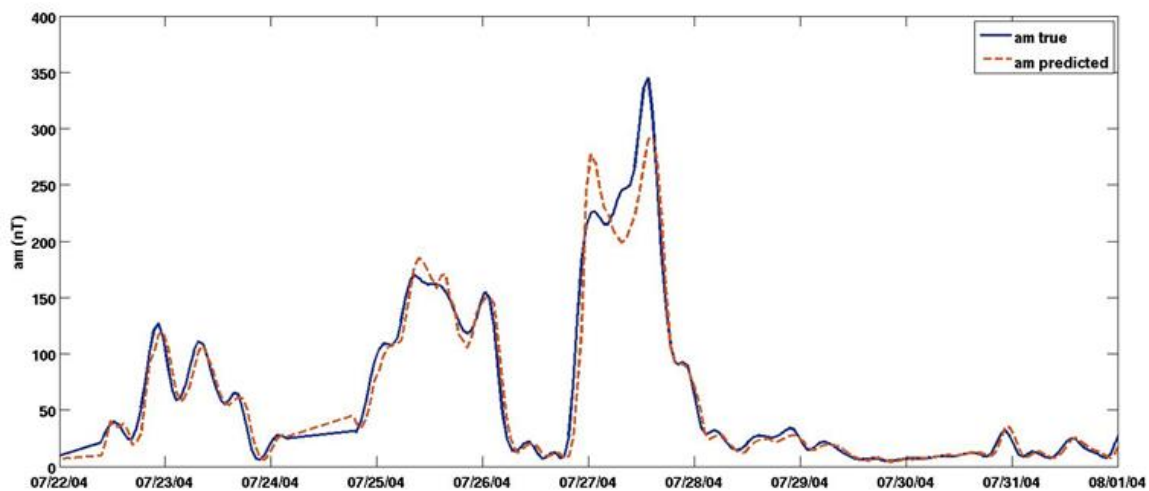


Figure 50- Événement de juillet 2004 avec les données OMNI en utilisant le réseau LSTM. Les données réelles sont en bleu, les données prédites en orange pointillé. Pour mieux visualiser la prédiction, un décalage d'une heure est utilisée pour la représentation graphique.

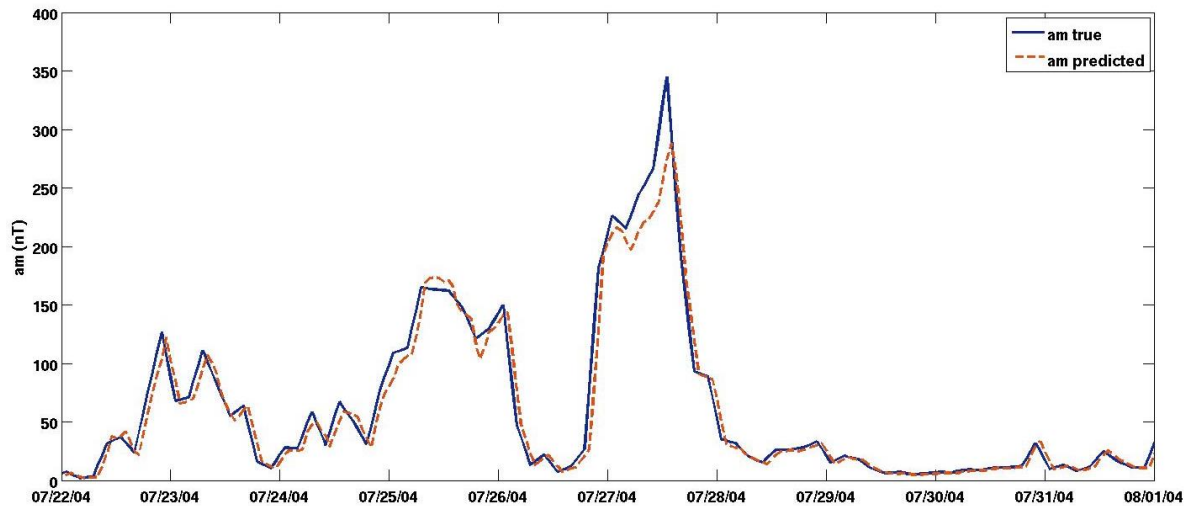
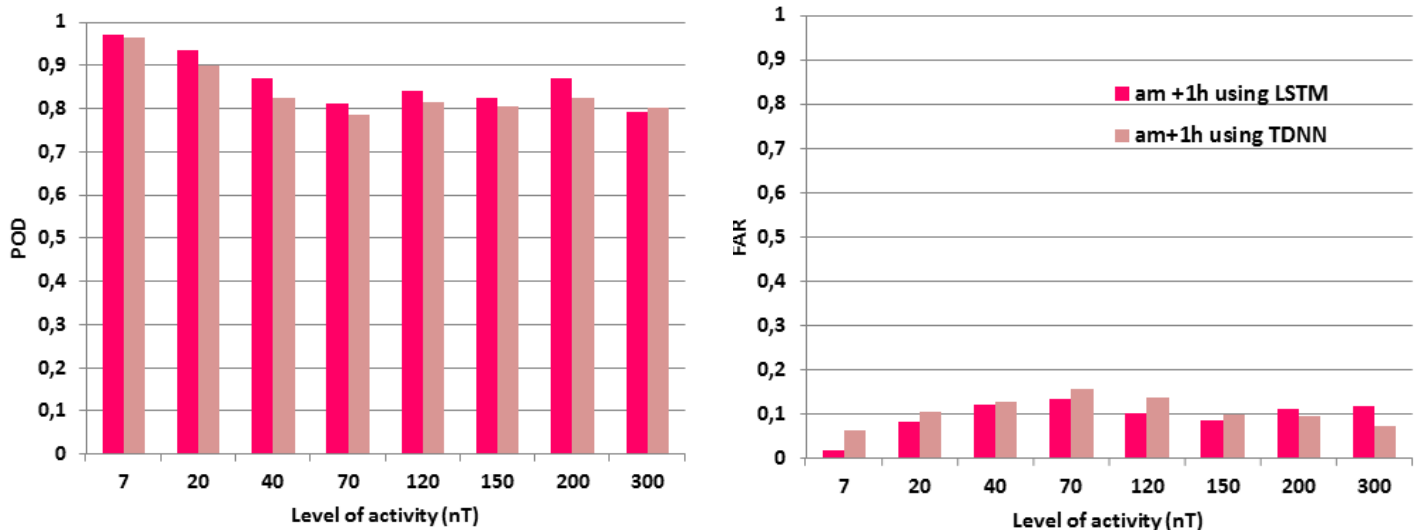


Figure 51- Evénement de juillet 2004 avec les données ACE en utilisant le réseau LSTM. Les données réelles sont en bleu, les données prédites en orange pointillé. Pour mieux visualiser la prédiction, un décalage d'une heure est utilisée pour la représentation graphique.

Par la suite, nous avons souhaité étudier l'apport du réseau LSTM par rapport au réseau TDNN. Cette comparaison est délicate car les deux réseaux ne sont pas programmés avec le même langage de programmation. Mais elle reste nécessaire car notre but tout au long de cette étude a été de développer un réseau basé uniquement sur les paramètres du vent solaire comme le TDNN, le plus performant possible et adapté aux besoins opérationnels. La Figure 52 compare les performances des réseaux LSTM et TDNN en utilisant les données ACE. Le TDNN est le même que celui optimisé au Chapitre 3 section 2.2. c'est-à-dire en considérant les données ACE, avec une fenêtre de spécialisation de six heures.

Nous constatons alors que le réseau LSTM est plus performant globalement que le TDNN avec une POD plus élevée et un FAR plus faible, sauf pour le seuil d'activité le plus élevé. On obtient alors pour le TDNN une POD de 0.800 et un FAR de 0.071 contre une POD de 0.791 et un FAR de 0.116 pour le LSTM. Le réseau récurrent LSTM basé uniquement sur les paramètres du vent solaire à l'instant présent permet alors d'améliorer les prévisions pour quasiment tous les niveaux d'activité. Ce résultat est sûrement lié au fait que le développement sous Matlab des réseaux de neurones est cadré par les outils fournis par la toolbox Neural Network. Ce cadrage est associé notamment à la détection automatique de la période à considérer pour cesser l'entraînement et éviter le surapprentissage comme nous l'avons décrit au Chapitre 2 section 3.1.1. En développant le réseau en Python, nous nous affranchissons de l'aspect boîte noire présenté par une toolbox Matlab, mais nous perdons un outil d'optimisation, le notre étant potentiellement moins robuste que celui développé par Matlab.

Ainsi, notre réseau en Python est mieux adapté aux besoins d'un opérateur que le TDNN, et présente globalement de meilleures performances que celui-ci, en étant légèrement moins performant pour le cas d'activité le plus élevé. Etant donné que le but est de fournir des prédictions optimales à tout niveau d'activité, et notamment pour les niveaux les plus élevés, nous avons réfléchi à des améliorations possibles du LSTM. Un des éléments possibles qui nous est venu en discutant avec différents scientifiques, notamment lors d'un séminaire à l'IASB à Bruxelles pour présenter les résultats obtenus en septembre 2017, a été l'utilisation de fonction de couplage en entrée du réseau de neurones.



	POD		FAR	
	LSTM - ACE	TDNN- ACE	LSTM - ACE	TDNN - ACE
7	0,970	0,962	0,0178	0,062
20	0,934	0,898	0,0812	0,103
40	0,869	0,824	0,119	0,127
70	0,809	0,784	0,131	0,155
120	0,839	0,814	0,101	0,136
150	0,823	0,805	0,0863	0,098
200	0,868	0,824	0,110	0,094
300	0,791	0,800	0,116	0,071

Figure 52- POD et FAR du réseau LSTM (rose foncé) et du TDNN (rose clair) en considérant la densité, la vitesse et l'IMF|B| de ACE. La fenêtre de spécialisation du TDNN est de six heures.

2. ANALYSE DES EFFETS DE L'UTILISATION DE FONCTION DE COUPLAGE EN ENTREE DU LSTM

Au Chapitre 1 section 4.2, nous avons introduit les fonctions de couplage. Ces travaux ont été initiés par [Perreault and Akasofu, 1978] dans le but d'analyser la répartition en énergie des perturbations du vent solaire au travers de différents systèmes de courant. Ces fonctions définissent une quantité d'énergie entrante dans la magnétosphère terrestre à partir de paramètres du vent solaire. Elles sont définies empiriquement, sur la base d'observations d'événements extrêmes. Depuis 1978, différentes fonctions de couplages ont été développées, nous souhaitons étudier l'apport de ces fonctions de couplage pour la prévision de l'indice magnétique global am .

2.1. Point sur les fonctions de couplage et leurs applications en entrée des réseaux de neurones

L'équation d'entrée en énergie permet de rendre compte de l'importance de la perturbation apportée par un événement solaire sur les différents systèmes de courant terrestres. La première équation a été développée par [Perreault and Akasofu, 1978] pour analyser les orages géomagnétiques en termes de flux d'énergie. Le but était de trouver les paramètres du vent solaire qui contrôlent les orages, et pour ce faire, l'estimation du flux d'énergie interplanétaire $\epsilon(t)$ a été faite en termes de flux de Poynting

(l'équation est décrite au Chapitre 1, section 4.2.). Le flux de Poynting représente le taux de variation par unité de surface de la quantité d'énergie d'un certain volume. On l'interprète comme la différence entre le flux d'énergie entrant et le flux d'énergie sortant par unité de surface, ou encore la puissance véhiculée par une onde à travers une surface. Sa variation temporelle est ensuite comparée au taux de dissipation en énergie en termes d'injection de particules dans le courant annulaire $U_R(t)$, de dissipation Joule dans l'ionosphère $U_J(t)$ et d'injection de particules aurorales $U_A(t)$. Il a alors été établi par [Perreault and Akasofu, 1978] que le taux de dissipation en énergie totale $U(t)$ est la somme des trois contributions précédentes et qu'on peut établir un lien direct entre l'énergie entrante et l'énergie dissipée au travers des différents courants magnétosphériques $U(t) \cong \epsilon(t)$. Cette relation a été établie après une étude basée sur 17 paramètres géomagnétiques et interplanétaires, et a montré qu'il y avait une bonne corrélation entre ces deux éléments, sauf durant les phases de recouvrement après un orage.

Par la suite, [Akasofu, 1981] a repris les études de couplage et a constaté que la magnétosphère ne pouvait répondre pleinement à des flux d'énergie d'échelle temporelle inférieure à la constante de temps de la magnétosphère τ_m à cause de l'inductance importante de la magnétosphère et de la résistance ionosphérique de 0.1Ω . La Figure 53 présente l'analyse faite dans cet article sur deux événements extrêmes, ceux de février et mars 1973, en présentant les variations du flux entrant ϵ , celles du flux total interne à la magnétosphère U_T , du flux d'énergie au niveau du courant annulaire U_R et celui associée à la dissipation joule U_J . Comme il est démontré par les équations mises en place par [Akasofu, 1981] illustrées sur la Figure 53, le courant annulaire U_R est directement dérivé de l'indice Dst , tandis que le courant de dissipation Joule dans l'ionosphère U_J et d'injection de particules aurorales U_A sont reliés à l'indice AE . En étudiant ces événements, différentes relations ont été constatées entre le flux d'énergie entrant $\epsilon(t)$ et les systèmes de courant U_R et U_J . Lorsque $\epsilon(t)$ augmente, on peut voir sur la Figure 53 que cela provoque une augmentation du flux interne à la magnétosphère, principalement au niveau du courant annulaire U_R . L'activité aurorale a également augmenté en conséquence, mais le flux d'énergie est bien plus important au niveau du courant annulaire, déclenchant par la suite un orage. La majorité de l'énergie entrante est alors dissipée suite à des phénomènes de reconnexion magnétique et de déclenchement de sous-orages et d'orages dans la magnétosphère interne.

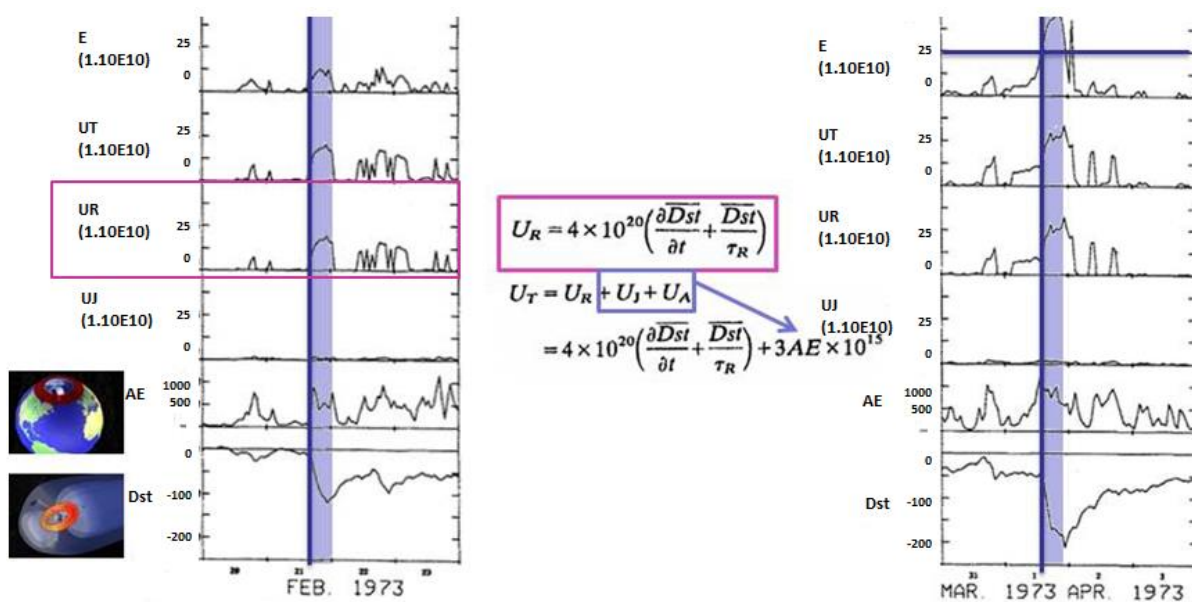


Figure 53- Entrée en Energie et impact sur les systèmes de courant d'après [Akasofu, 1981].

Les éléments mis en évidence par les études de [Perreault and Akasofu, 1978] et [Akasofu, 1981] soulèvent ainsi un élément clef dans le cadre de notre étude sur l'application d'une fonction de couplage pour monitorer la prédiction d'un indice magnétique. En effet, il existe une relation forte entre l'entrée en énergie associée à un événement solaire et les indices magnétiques.

[Akasofu, 1981] a également démontré grâce aux observations de Mariner 5, situé à 1.9 millions de km de la Terre, que le flux d'Énergie est conservé entre ce point et un point proche de la Terre au niveau d'Explorer 34 qui était alors dans la magnétosphère, comme illustré sur la Figure 54. Ceci est une estimation qui concerne directement notre étude car cela signifierait que le flux d'Énergie est conservé entre le point L1 qui est lui à 1.5 millions de km et la magnétosphère. On peut donc analyser le lien de causalité entre le flux d'énergie entrant ainsi que l'énergie stockée en se basant directement sur les paramètres du vent solaire enregistrés par le satellite ACE.

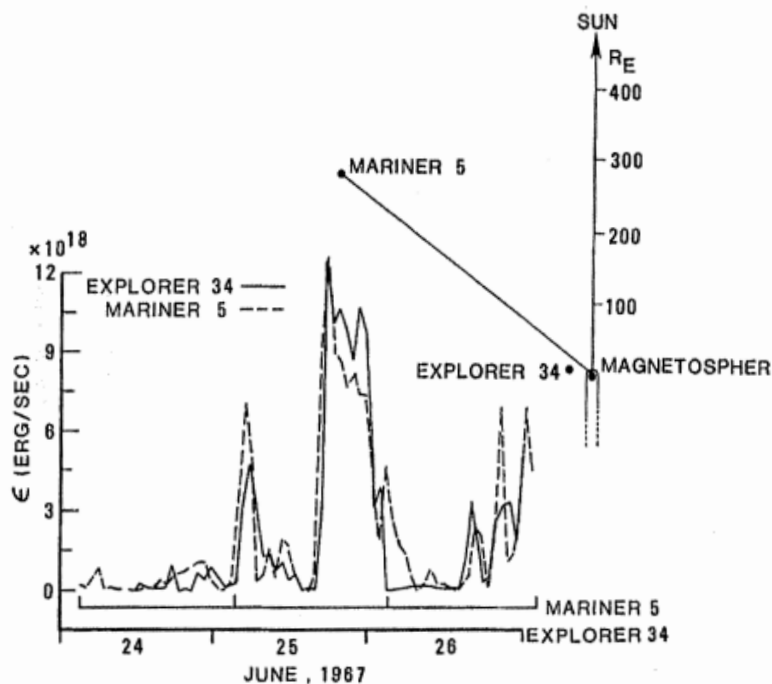


Figure 54- Estimation de l'Énergie entrante au niveau de Mariner 5 et Explorer 34 d'après [Akasofu, 1981].

Par la suite, [Vasyliunas et al., 1982] ont proposé une analyse dimensionnelle en partant du postulat suivant : « il existe un ensemble limité de variables d'entrées indépendantes (des propriétés du vent solaire et de la Terre) à partir desquelles on peut calculer les variables de sorties dépendantes ». Pour prendre en compte la magnétohydrodynamique du vent solaire, c'est-à-dire le fait que le vent solaire est un fluide conducteur du courant électrique en présence de champs électromagnétique, trois paramètres ont été considérés, la vitesse du vent solaire f_s , la densité du vent solaire n et le champ magnétique interplanétaire $IMF |B|$. Il a également été défini que si la conductivité ionosphérique joue un rôle (magnitude des courants de Birkeland), il fallait considérer le moment du dipôle magnétique terrestre M_E ainsi que la conductivité ionosphérique de Pedersen Σp . Enfin, la viscosité cinématique effective est prise en compte car il y a une interaction visqueuse entre la magnétosphère et le flux du vent solaire (mouvement de convection ou hypothèse de la magnétosphère fermée). Le système d'équations développé par [Vasyliunas et al., 1982] pour donner l'expression de la puissance entrante dans la magnétosphère est présenté au Chapitre 1 section 4.2.

Avec cette analyse dimensionnelle, [Vasyliunas et al., 1982] ont ainsi proposé une expression qualitative. De nombreuses fonctions de couplage ont été développées par la suite, basées sur différentes méthodes ou bases de données (voir [Murayama et al, 1982] [Bargatze et al, 1986], [Stamper et al., 1999], [Finch and Lockwood, 2007]). Cependant, la plupart de ces fonctions ne permettent pas d'évaluer quantitativement l'entrée en énergie associée au vent solaire à cause des coefficients qui restent indéterminés (voir [Murayama et al, 1982] [Bargatze et al, 1986], [Stamper et al., 1999], [Finch and Lockwood, 2007]). Les simulations MHD fournissent une approche efficace pour évaluer le flux d'énergie global entrant dans le système « SW-M-I » (vent solaire-magnétosphère-ionosphère) [Papadopoulos et al., 1999]. L'approche MHD a été utilisée dans de multiples études pour analyser la dissipation en énergie dans la magnétosphère, mais encore une fois d'un point de vue qualitatif (voir [Palmroth et al., 2005], [Palmroth et al., 2006], [Palmroth et al., 2010], [Palmroth et al., 2012], [Pulkkinen et al., 2002], [Pulkkinen et al., 2008]). [Wang et al., 2014] ont alors utilisé cette approche MHD pour se focaliser sur le transfert d'énergie depuis le vent solaire vers la magnétosphère, et en déduire une fonction de couplage en énergie à partir de simulations numériques. Ceci a principalement pour but de fournir la première analyse quantitative, contrairement aux précédentes qui n'étaient que qualitatives. Pour ce faire, une première étape illustrée consiste à identifier la surface de la magnétopause à partir de simulations MHD. Cette étape est effectuée à partir de la technique développée par [Palmroth et al., 2003]. Ensuite, [Wang et al., 2014] déterminent le flux d'énergie passant à travers la magnétopause en utilisant le système d'équations (43)

$$dEq = dAK \cdot \hat{n}$$

$$\mathbf{K} = \left(U + P - \frac{B^2}{2\mu_0} \right) \mathbf{V} + \frac{1}{\mu_0} \mathbf{E} \times \mathbf{B} \quad (43)$$

$$U = \frac{P}{\gamma-1} + \frac{1}{2} \rho V^2 + \frac{B^2}{2\mu_0}$$

$$(E_{thermique} + E_{cinétique} + E_{magnétique})$$

avec dA un élément de surface, \hat{n} vecteur normal à la surface, \mathbf{K} le flux d'énergie totale, $\gamma = 5/3$ l'exposant polytropique, P la pression thermique, \mathbf{B} le champ magnétique, V la vitesse et $\mathbf{E} = V \times \mathbf{B}$ le champ électrique de convection. Enfin, le flux d'énergie totale à travers la surface de la magnétopause est l'intégrale des flux d'énergie pour la surface totale.

La formule empirique représentée par l'équation (44) a ainsi été déduite de l'étude de [Wang et al., 2014] à partir de 37 tests à l'état quasi stable

$$E_{in}(W) = 3,78 \cdot 10^7 n^{0.24} V^{1.47} B_T^{0.86} \left(\sin^{2.70} \left(\frac{\theta}{2} \right) + 0,25 \right) \quad (44)$$

avec E_{in} l'énergie entrante dans la magnétosphère et B_T le champ magnétique transverse

$$B_T = \sqrt{B_x^2 + B_y^2}$$

Cette étude a montré alors une nette amélioration des corrélations entre les résultats donnés par la simulation et la fonction définie par [Wang et al., 2014] en comparaison avec la fonction de couplage d' [Akasofu, 1981]. Cette fonction sous-évaluait des entrées en énergie d'un facteur 4 à 5 d'après [Wang et al., 2014].

Lorsque nous avons étudié au Chapitre 3 section 1.1.2. le coefficient de Kendall analysant l'existence d'une relation directe entre les paramètres du vent solaire et l'indice magnétique am , l'IMF clock angle faisait partie des paramètres présentant le moins de relation directe avec la perturbation magnétique mesurée au sol. En analysant la relation d'un point de vue énergétique avec une fonction de couplage, une forte dépendance avec l'IMF clock angle ressort. Dans la logique que nous étudions

ici, la fonction de couplage permet d'évaluer dans un premier temps la quantité d'énergie entrante dans la magnétosphère, qui va être dissipée par la suite au travers de différents systèmes de courant. Il est complexe d'évaluer d'un point de vue quantitatif et qualitatif la répartition de cette énergie à travers les systèmes, et donc d'évaluer un indice magnétique résultant des perturbations engendrées par cette entrée en énergie. Le réseau de neurones trouve sa place pour connecter les deux informations, d'une part l'évaluation de l'énergie entrante dans la magnétosphère fournie par la fonction de couplage de [Wang et al., 2014], et la perturbation magnétique globale mesurée par l'indice magnétique *am*.

Une démarche similaire avait été étudiée par [Bala and Reiff, 2012], en utilisant en entrée d'un réseau de neurones classique de type « feedforward » l'indice de Boyle. Cet indice est une approximation empirique évaluant le potentiel au niveau de la calotte polaire en utilisant les paramètres du vent solaire. Ce modèle avait été utilisé notamment pour la prédiction de l'indice *Kp* jusqu'à six heures à partir des données ACE. La Figure 55 présente les résultats obtenus avec cette méthode pour la prédiction de *Kp* à une heure et trois heures avec l'événement de Juillet 2009. L'« ANN-Kp » est le *Kp* obtenu avec la technique développée par [Bala and Reiff, 2012] à partir de réseaux de neurones. Le « NOAA Kp » est le *Kp* prédit disponible sur le site de la NOAA. L'« official Kp » est le *Kp* définitif. Ce modèle opérationnel présente de bonnes performances pour prédire *Kp* à partir d'une fonction comme l'indice de Boyle, avec un coefficient de corrélation de 0.88 pour une prédiction à une heure.

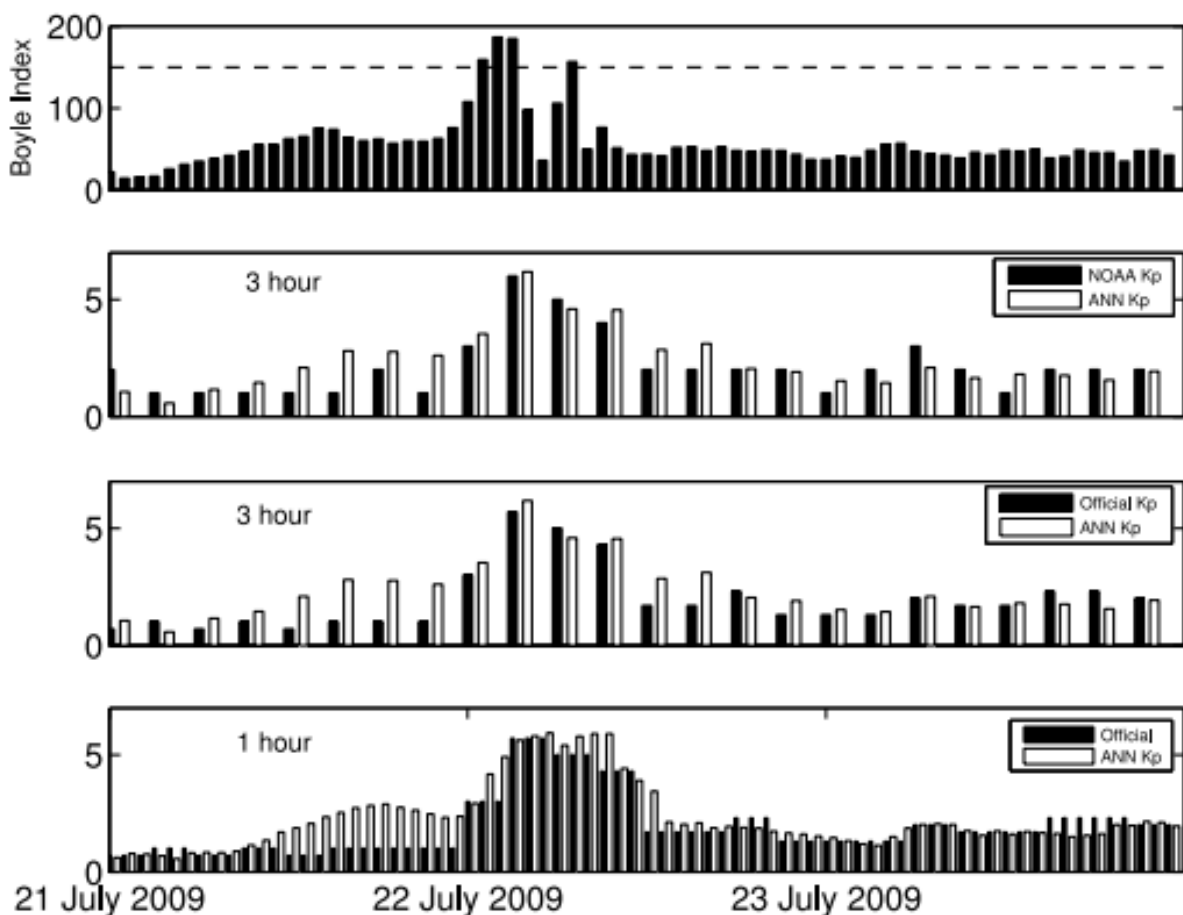


Figure 55- Prédiction de *Kp* à partir de l'indice de Boyle à une heure et trois heures [Bala and Reiff, 2012].

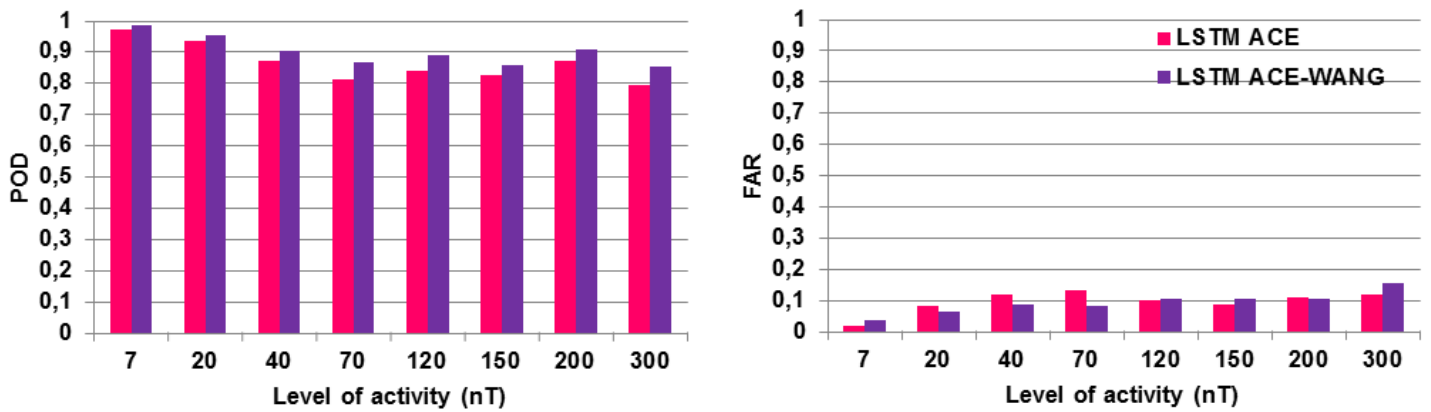
Nous allons donc évaluer l'apport de l'utilisation de la fonction de couplage définie par [Wang et al., 2014] en entrée du LSTM afin de prédire l'indice magnétique am à une heure, et son apport en comparaison de l'utilisation de paramètres du vent solaire considérés séparément en entrée du LSTM.

2.2. L'apport des fonctions de couplage sur les performances des réseaux avec les données ACE

Pour analyser l'apport de la fonction de couplage définie par [Wang et al., 2014], nous calculons à partir des données ACE l'énergie entrante afin de l'utiliser en entrée du réseau LSTM. L'entrée du réseau de neurones est donc uniquement la fonction de couplage de [Wang et al., 2014].

La Figure 56 présente les résultats obtenus en termes de POD et de FAR pour le réseau LSTM en considérant d'une part uniquement les données ACE $\{n, fs, IMF |B|\}$ (en rose sur la Figure 56) et d'autre part en considérant la fonction de couplage de [Wang et al., 2014] calculée à partir des données ACE (en violet sur la Figure 56). La POD est meilleure en utilisant la fonction de couplage à tout niveau d'activité. Au seuil d'activité le plus élevé, la POD s'élève à 0.853 en utilisant la fonction de couplage, contre 0.791 en ne considérant que les paramètres du vent solaire $\{n, fs, IMF |B|\}$ pris séparément. Cependant, le FAR est variable et lorsque la valeur d' am est extrême, le FAR est supérieur avec la fonction de couplage avec une valeur de 0.155 contre 0.116 en utilisant les données du vent solaire prises séparément.

Du point de vue des mesures globales, le coefficient de corrélation est de 0.986, et l'erreur quadratique moyenne est de 3.75 nT. En comparant ces performances à celles présentés dans le Tableau 9, on constate que ces résultats sont meilleurs. En effet, le coefficient de corrélation et l'erreur quadratique moyenne obtenus utilisant les paramètres du vent solaire pris séparément sont respectivement égaux à 0.975 et 5.82.



	POD		FAR	
	LSTM - ACE	LSTM - ACE/WANG	LSTM - ACE	LSTM - ACE/WANG
7	0,970	0,982	0,0178	0,0373
20	0,934	0,952	0,0812	0,0636
40	0,869	0,902	0,119	0,0878
70	0,809	0,864	0,131	0,0832
120	0,839	0,888	0,101	0,104
150	0,823	0,858	0,0863	0,104
200	0,868	0,907	0,110	0,103
300	0,791	0,853	0,116	0,155

Figure 56- POD et FAR du réseau LSTM en utilisant les paramètres du vent solaire à partir de ACE en rose, et utilisant la fonction de couplage définie par [Wang et al., 2014] à partir des données ACE en violet en fonction du seuil d'activité.

Pour évaluer l'impact que cela peut avoir sur la prévision d'événements extrêmes, nous avons évalué les performances du réseau sur l'événement de Juillet 2004. La Figure 57 présente les résultats obtenus sur cet événement en utilisant l'équation de [Wang et al., 2014] en entrée à partir des données ACE. Cette figure est à comparer avec la Figure 51 où nous avons également utilisé les données ACE mais en considérant chaque paramètre du vent solaire $\{n, fs, IMF |B|\}$ pris séparément. Pour le pic d'activité le plus extrême à 350 nT, le LSTM alimenté par l'équation de [Wang et al., 2014] propose une valeur prédite plus proche que celle proposée par le LSTM alimenté par les données ACE prises séparément. En effet, avec ce dernier, la valeur prédite était de 300 nT, tandis qu'avec l'équation d'énergie en entrée on obtient une valeur prédite de 345 nT. Et ce, même si comme le montre la Figure 56, le FAR est plus élevé en utilisant l'équation d'énergie en entrée, mais la POD est bien supérieure et compense cet effet.

Nous avons ainsi optimisé la prédiction de l'indice am avec le réseau LSTM, en utilisant les données fournies par le satellite ACE au point de Lagrange L1. Nous avons amélioré les performances de prédiction de ce réseau en utilisant en entrée une fonction de couplage, celle de [Wang et al., 2014], qui rend compte de l'entrée en énergie dans la magnétosphère. Afin de compléter cette étude, nous avons souhaité travailler sur l'indice am sectoriel ou $a\sigma$. Cet indice est dérivé de l'indice am . Nous

présentons cet indice, l'apport qu'il représente dans le cadre de la météorologie de l'espace ainsi que l'analyse de la prédiction de celui-ci dans la section suivante.

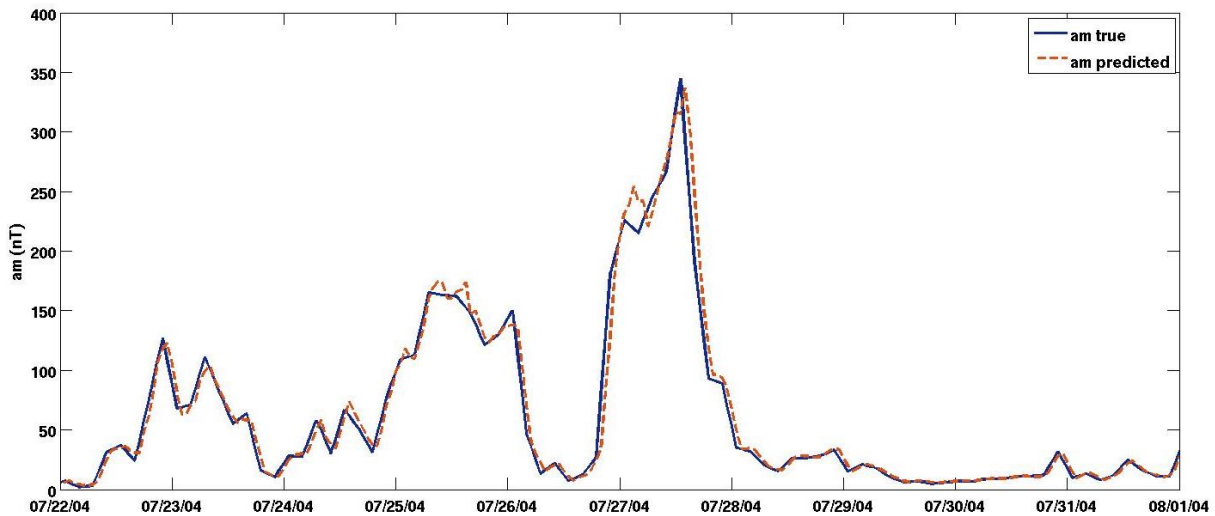


Figure 57- Evénement de juillet 2004 avec l'équation de [Wang et al., 2014] à partir des données ACE en utilisant le réseau LSTM. Les données réelles sont en bleu, les données prédites en orange pointillé. A comparer avec la Figure 50 pour les résultats obtenus avec les données du vent solaire prises séparément. Pour mieux visualiser la prédiction, un décalage d'une heure est utilisée pour la représentation graphique.

3. EVALUATION DE LA CAPACITE DU LSTM A FOURNIR UNE PREDICTION MULTI-SORTIE DANS LE CADRE DE LA PREDICTION DE L'INDICE $a\sigma$

L'indice $a\sigma$ a été publié par [Chambodut et al., 2013]. C'est un indice dérivé de l'indice am , basé sur le même réseau de station, également défini en nanoTesla et trihoraire. Il fournit une caractérisation de l'activité géomagnétique locale, associée aux secteurs MLT définis au Chapitre 1 section 2.5. Ainsi, 4 indices sont définis à chaque instant : $a\sigma_{dawn}$ (aube), $a\sigma_{noon}$ (midi) $a\sigma_{dusk}$ (crépuscule) $a\sigma_{midnight}$ (minuit). Ils couvrent respectivement les secteurs 03-09 MLT, 09- 15 MLT, 15-21 MLT, 21-03 MLT. Ceci présente un intérêt en météorologie de l'espace, car l'activité magnétique en fonction du secteur MLT n'est pas la même. Nous le présentons plus en détail dans la section suivante, avant d'effectuer une analyse sur la prédiction des am spécifiques à chaque secteur MLT.

3.1. Le rôle de l' am sectoriel ou $a\sigma$ en météorologie de l'espace

En proposant l'indice magnétique sectoriel $a\sigma$, [Chambodut et al., 2013] ont proposé une nouvelle fenêtre sur l'observation d'une perturbation magnétique et de son évolution en fonction du secteur MLT. Il a été observé qu'en fonction de l'indice sectoriel considéré, et donc de l'activité géomagnétique observée, différentes propriétés statistiques associées à des processus physiques étaient mises en évidence. Par exemple, la Figure 58 extraite de [Chambodut et al., 2013] présente les variations statistiques pour chaque indice sectoriel sur la période 1959-2011. On observe alors des phénomènes bien spécifiques aux équinoxes avec des maxima autour du solstice de Juin et de Décembre, notamment pour les secteurs dusk (aube) et midnight (nuit). Cet effet est aussi bien présent sur les 4 premiers encadrés où l'intensité est fonction du UT (Temps Universel) et du DOY (day of

year), que dans l'encadré de droite où la valeur moyenne de l'indice sectoriel est représentée en fonction du DOY.

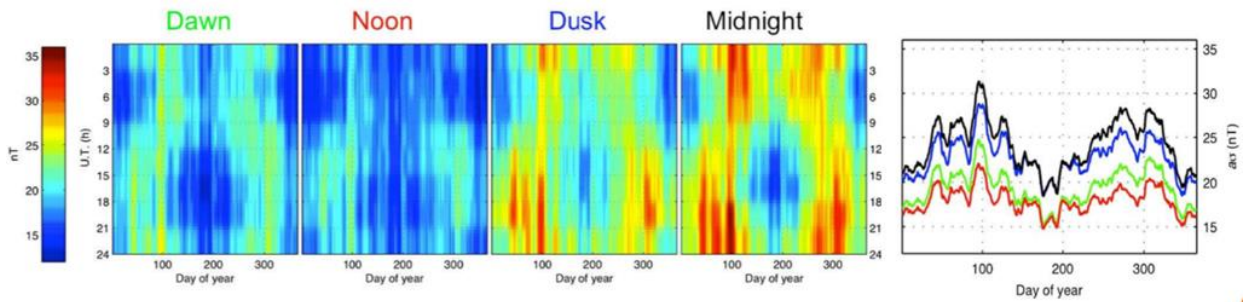


Figure 58 – Représentation UT/DOY des 4 indices associés aux secteurs MLT. La valeur moyenne sur les UT de ces indices en fonction du DOY est tracée dans l'encadré de droite [Chambodut et al., 2013].

L'observation de l'évolution d'une perturbation magnétique associée à un événement solaire extrême au travers des indices sectoriels permet de mettre en avant des processus physiques bien spécifiques. [Chambodut et al., 2013] en analysant l'événement de Mai 2003 illustré sur la Figure 59 ont pu analyser différents effets en fonction des variations des indices sectoriels. Par exemple, une croissance soudaine du $a\sigma_{Noon}$ (midi) est associée à une augmentation de la pression du vent solaire et à un $IMF |B|$ orienté vers le sud, ainsi qu'à une augmentation de la reconnexion côté jour de la magnétopause ce qui améliore le couplage vent solaire-magnétosphère côté jour. La décroissance du $a\sigma_{Noon}$ (midi) peut par la suite être expliquée par le changement dans l'orientation de l' $IMF |B|$. On peut alors suivre l'évolution de l'orage associé en fonction des différents secteurs. Dans cette même analyse, il a été souligné que suite à l'augmentation des phénomènes de reconnexion, on peut observer une forte asymétrie dawn-dusk (aube-crépuscule), probablement liée au maximum de l'orage mis en évidence par les variations de l'indice SYM-H, et donc de l'intensification du courant annulaire partiel (voir [Shi et al. 2005], [Shi et al. 2008]) ou de l'asymétrie des électrojets auroraux. Enfin, une augmentation importante des indices sectoriels côté midnight et dusk (minuit et aube) est la conséquence de sous-orages intensifiant le courant annulaire partiel. Ainsi en suivant les variations des indices sectoriels, on peut identifier différents effets, leurs variations, et les dissocier pour mieux anticiper leurs conséquences.

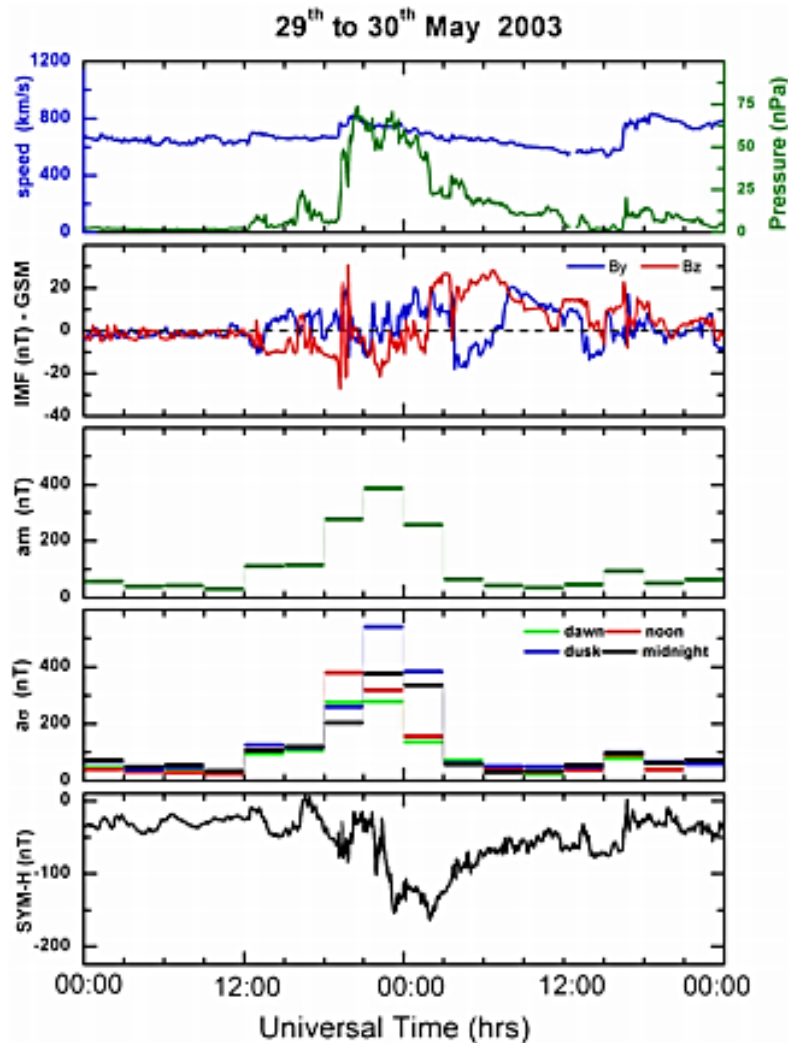


Figure 59 – Paramètres du vent solaire et indices géomagnétiques durant l'orage de Mai 2003. La pression du vent solaire est en vert, la vitesse en bleu, les composantes Z et Y de l'IMF $|B|$ dans le repère GSM sont en rouge et bleu. L'indice am ainsi que les am sectoriels sont représentés à la suite ainsi que l'indice SYM-H [Chambodut et al., 2013].

L'événement que nous avons analysé tout au long de ce manuscrit est l'événement de Juillet 2004. Une analyse détaillée de cet événement est également faite par [Chambodut et al., 2013] et nous présentons les variations des indices sectoriels sur la Figure 60. Les phénomènes à l'origine des différents pics d'activité sont comme pour l'événement de Mai 2003, associés à des perturbations physiques bien spécifiques comme l'augmentation du courant annulaire partiel qui impacte le secteur dawn-dusk (aube-crêpuscule), ou une augmentation forte côté midnight (minuit) du pic d'activité impliquant des sous-orages forts contrôlés par l'IMF $|B|$ orienté vers le sud dans le nuage magnétique.

Ainsi, grâce à ces indices sectoriels, il est possible de fournir une analyse spécifique d'une perturbation magnétique et de son évolution au travers des différents systèmes de courant. C'est une information pouvant jouer un rôle clef à l'avenir pour un opérateur dans le spatial car en fonction de la position du satellite, l'impact observé n'est pas le même en fonction du secteur comme nous avons pu le voir aussi bien pour l'événement de Mai 2003 que pour celui de Juillet 2004. Les dispositions à prendre en conséquence ne seraient donc pas les mêmes. Nous avons donc travaillé sur le

développement de modèles de prévision de ces différents indices sectoriels dans le but d'apporter de nouveaux outils et méthodes en météorologie de l'espace.

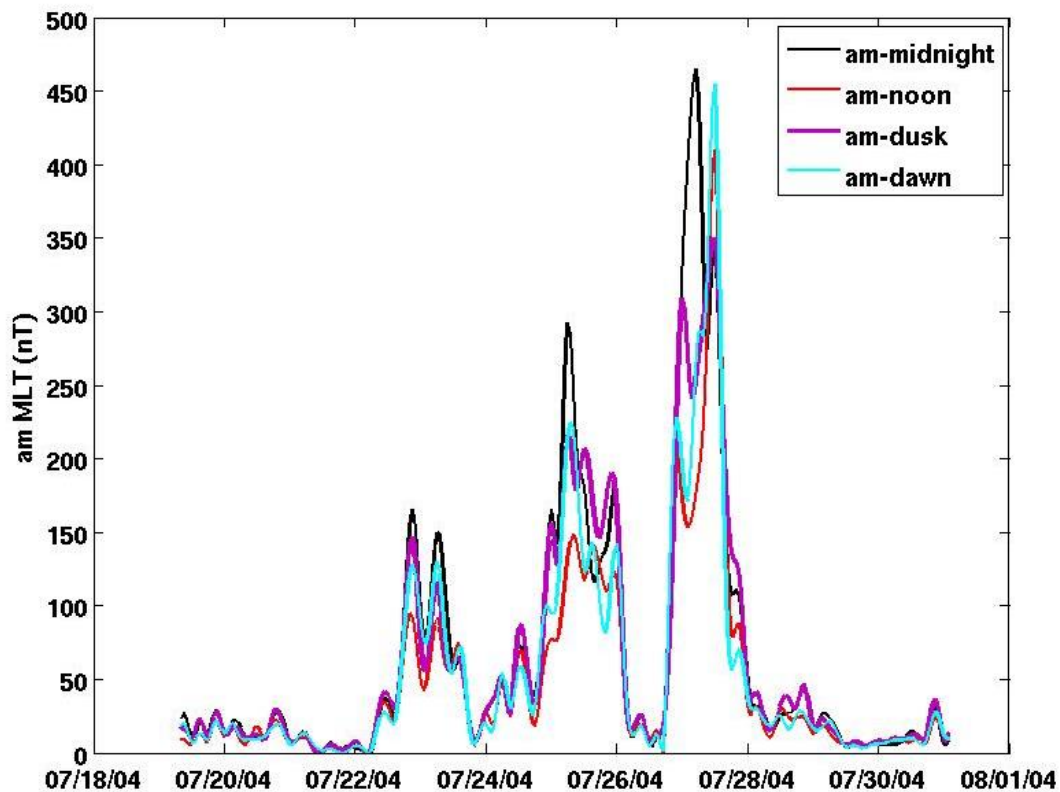


Figure 60- Variation des indices am sectoriel pour l'événement de Juillet 2004.

3.2. Analyse des performances du LSTM pour la prédiction de l'activité magnétique associée à chaque secteur MLT

Afin de prédire l'indice $a\sigma$, c'est-à-dire fournir une prédiction pour chacun des secteurs et donc des valeurs associées à $a\sigma_{dawn}$, $a\sigma_{noon}$, $a\sigma_{dusk}$, $a\sigma_{midnight}$, il existe deux possibilités. Dans un premier temps, nous avons développé un réseau mult sortie représenté à gauche sur la Figure 61, c'est-à-dire que le réseau fournit en sortie un vecteur contenant les valeurs prédites des quatre am sectoriels. Dans un second temps, nous avons développé un réseau spécifique à chaque secteur MLT représenté à droite sur la Figure 61, soit quatre réseaux. Il est en effet possible de développer des réseaux multi sortie, cette méthode est utilisée classiquement par les réseaux de classification qui utilisent une couche supplémentaire en sortie appelée softmax [Sutton and Barto, 1998]. Cette technique permet de représenter une loi catégorielle, c'est-à-dire une loi de probabilité sur un ensemble K de différents résultats possibles. On l'utilise ainsi dans différentes méthodes de classifications multiples faisant appel à des réseaux de neurones. Les méthodes de prédictions que nous développons ici font appel au concept de régression et non de classification. Nous ne pouvons donc pas utiliser cette technique directement. Nous avons donc défini un réseau dont l'entraînement est fait à partir d'un vecteur des données du vent solaire $\{n, fs, IMF |B|\}$, et pour lequel la sortie est un vecteur contenant les prédictions des indices $\{a\sigma_{dawn}, a\sigma_{noon}, a\sigma_{dusk}, a\sigma_{midnight}\}$. Cette méthode n'a pas encore été utilisée en météorologie de l'espace, et n'a été que très peu développée dans d'autres domaines, mais elle permet de prendre en compte les dépendances entre sorties [An et al. 2012].

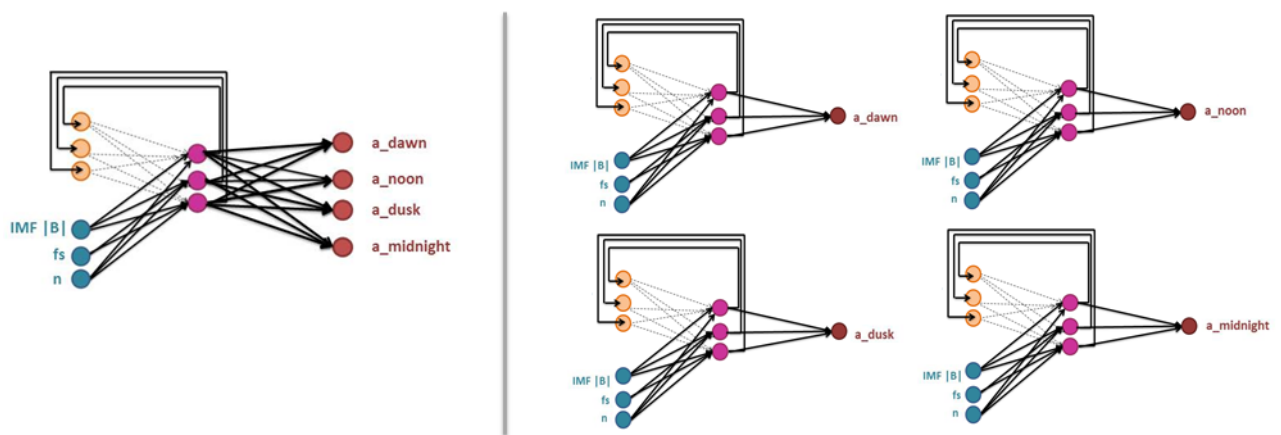
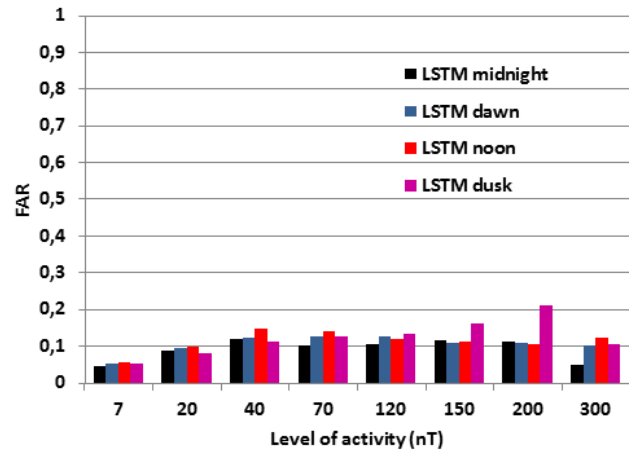
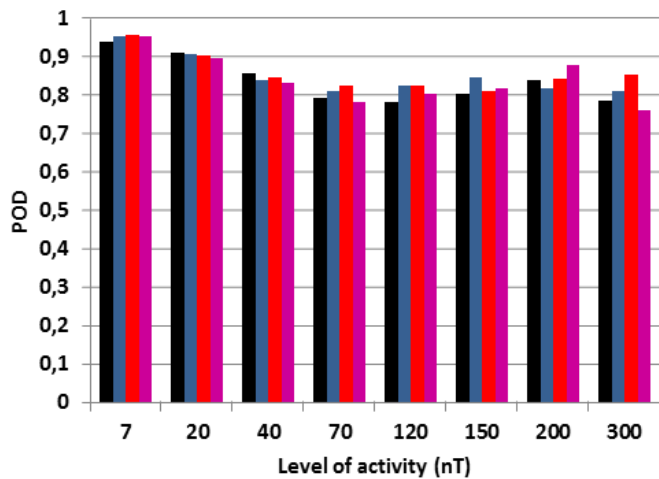


Figure 61- Réseaux considérés pour la prédiction des indices sectoriels : un réseau multisortie vs. 4 réseaux monosortie. Dans tous les cas, le nombre de cellules au sein du LSTM est de 20.

Nous avons donc travaillé sur ces deux techniques afin de voir si, à l'heure actuelle, avec les techniques que nous avons, il est plus judicieux de considérer un réseau multisortie, ou quatre réseaux monosortie. La Figure 62 montre les performances du réseau multisortie en utilisant les données ACE pour fournir les prédictions des indices sectoriels. La Figure 63 montre les performances associées aux quatre réseaux spécifiques à chaque indice sectoriel.

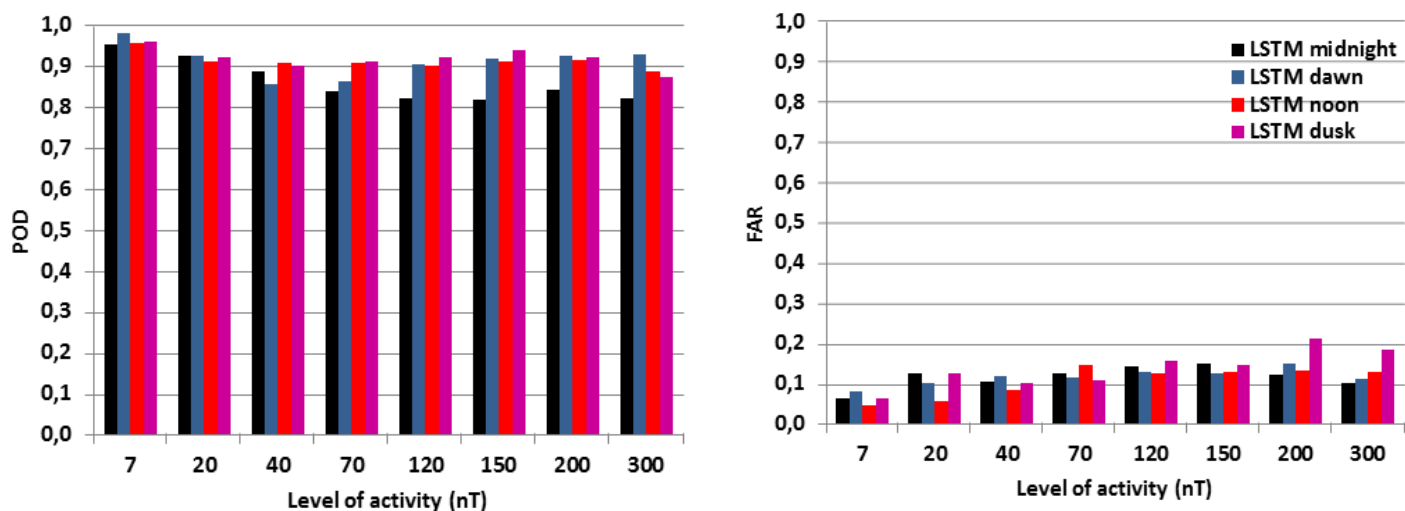
Le premier constat que l'on peut faire est qu'en considérant le seuil d'activité le plus élevé ($a\sigma > 300nT$) les POD obtenues en utilisant les réseaux monosortie sont plus élevées que celles obtenues par le réseau multisortie. En effet, quel que soit l'indice sectoriel considéré, la POD est meilleure. Par exemple pour le cas du $a\sigma_{Dawn}$ (aube) la POD du réseau multisortie est de 0.812 tandis que celle du réseau monosortie associée à cet indice est de 0.930. En revanche, si la POD est meilleure, le FAR est plus élevé avec les réseaux monosortie, et ce à tout niveau d'activité et indice considéré. Pour comparaison, si on considère l'indice $a\sigma_{Dawn}$ (aube dans le domaine d'activité le plus élevé, le FAR est à 0.101 dans le cadre du multisortie, et de 0.112 dans le cadre du monosortie. Il faut donc peser le pour et le contre de ces techniques. L'avantage d'un réseau spécifique à une sortie est que l'optimisation des paramètres du réseau comme les poids et biais est faite pour une sortie bien spécifique. En revanche, cette méthode ne prend pas du tout en compte les dépendances existantes entre les sorties, c'est-à-dire ici entre les indices magnétiques sectoriels.



	POD				FAR			
	Midnight	Dawn	Dusk	Noon	Midnight	Dawn	Dusk	Noon
7	0,941	0,956	0,956	0,958	0,0439	0,0518	0,0514	0,0548
20	0,912	0,907	0,898	0,903	0,0857	0,092	0,0802	0,097
40	0,858	0,841	0,834	0,847	0,118	0,121	0,109	0,147
70	0,795	0,811	0,784	0,826	0,101	0,123	0,124	0,139
120	0,782	0,826	0,804	0,826	0,105	0,125	0,132	0,117
150	0,804	0,847	0,817	0,810	0,116	0,107	0,158	0,111
200	0,842	0,819	0,879	0,845	0,111	0,109	0,209	0,103
300	0,788	0,812	0,761	0,856	0,048	0,101	0,104	0,12

Figure 62- POD et FAR du réseau LSTM multisortie en utilisant les paramètres du vent solaire à partir de ACE permettant de fournir des prédictions de $\alpha\sigma_{dawn}$, $\alpha\sigma_{noon}$, $\alpha\sigma_{dusk}$, $\alpha\sigma_{midnight}$ en fonction du niveau d'activité.

Physiquement, il existe une dépendance entre les secteurs MLT car si certains phénomènes sont spécifiques à un secteur MLT comme les phénomènes de reconnexion coté minuit, ou les phénomènes d'amplification du courant annulaire partiel côté dawn-dusk (aube-crépuscule) [Chambodut et al., 2013], ces phénomènes sont liés. En effet, comme on a pu le constater sur la Figure 60 avec l'événement de Juillet 2004, un premier pic d'activité est observé côté midnight (minuit), c'est-à-dire du côté où les particules sont réinjectées lors d'orages magnétiques, puis cette perturbation s'est propagée sur les autres secteurs MLT avec l'augmentation du courant annulaire partiel. Avec les méthodes existantes actuellement, le développement de réseau multisortie dans le cadre de la régression logistique n'est pas optimal. Pour ce faire, il faudrait réadapter les réseaux de neurones en définissant des classes, et développer des classifieurs multisortie, avec une sortie spécifique à un niveau d'activité. .



	POD				FAR			
	Midnight	Dawn	Dusk	Noon	Midnight	Dawn	Dusk	Noon
7	0,952	0,983	0,961	0,958	0,063	0,080	0,063	0,046
20	0,927	0,924	0,923	0,912	0,124	0,101	0,127	0,056
40	0,888	0,858	0,901	0,909	0,103	0,118	0,101	0,085
70	0,838	0,863	0,913	0,907	0,125	0,114	0,108	0,145
120	0,822	0,907	0,921	0,901	0,142	0,128	0,156	0,127
150	0,819	0,920	0,941	0,913	0,151	0,125	0,147	0,129
200	0,844	0,925	0,921	0,915	0,123	0,151	0,213	0,133
300	0,822	0,930	0,875	0,889	0,101	0,112	0,185	0,128

Figure 63- POD et FAR des 4 réseaux spécifiques à chaque indice sectoriel en utilisant les paramètres du vent solaire à partir de ACE permettant de fournir des prédictions de $a\sigma_{dawn}$, $a\sigma_{noon}$, $a\sigma_{dusk}$, $a\sigma_{midnight}$.

Le Tableau 10 présente les RMSE et CC associés aux réseaux multisorties et monosortie, pour chaque indice sectoriel. Le réseau multisortie est moins performant que les réseaux monosortie, avec des RMSE plus élevées et des CC plus faibles. Cependant, nous tenons à souligner qu'avec des développements futurs de techniques mieux adaptées à la prise en compte des dépendances au sein du réseau, il sera possible d'améliorer ces performances.

Tableau 10- Coefficient de corrélation (CC) et erreur quadratique moyenne (RMSE) du LSTM multisortie et des 4 LSTM monosortie avec les données ACE.

	RMSE				CC			
	Midnight	Dawn	Dusk	Noon	Midnight	Dawn	Dusk	Noon
LSTM multisortie	8,043	7,025	6,080	7,624	0,958	0,954	0,954	0,955
4 LSTM monosortie	7,19	6,45	5,32	5,42	0,963	0,966	0,967	0,963

Nous avons étudié les performances de ces réseaux sur l'événement de Juillet 2004, les résultats sont présentés sur la Figure 64 pour l'indice $a\sigma_{dusk}$ (crépuscule) et Figure 65 pour l'indice $a\sigma_{midnight}$ (minuit). Pour l'indice $a\sigma_{dusk}$ (crépuscule), les prédictions fournies par le réseau a) multisortie ou b) monosortie sont comparables à activité calme ou modérée, mais lorsque l'activité augmente et varie fortement entre le 25 et le 28 juillet, les prédictions fournies par les deux réseaux sont bien différentes. Par exemple, entre le 25 et le 26 juillet 2004, on observe trois pics d'activité dans le secteur dusk (crépuscule). Un premier à 226 nT, un second à 215 nT et un troisième plus faible à 196 nT. Le réseau multisortie fournit de meilleures prédictions que le réseau monosortie, spécifique à la prédiction de l'indice $a\sigma_{dusk}$ (crépuscule). En effet, le réseau multisortie prédit pour les trois pics présentés précédemment des valeurs respectivement égales à 232 nT, 234 nT et 198 nT. Le réseau monosortie surestime les valeurs à prédire avec des valeurs respectivement égales à 251 nT, 251 nT et 221 nT. Ceci rejoint les résultats observés sur les Figure 62 et Figure 63. Dans le domaine d'activité compris entre 200 et 300 nT, si la POD est meilleure avec le réseau monosortie (0.921 pour le monosortie contre 0.879 pour le réseau multisortie), le FAR est plus élevé (0.213 pour le monosortie contre 0.209 pour le multisortie). Une meilleure POD ne suffit pas à compenser le fait que le FAR soit plus élevé dans ce cas-là. Cependant, entre le 27 et le 28 juillet 2004, on observe deux pics d'activité plus intenses que ceux observés durant la période du 25 au 26 juillet 2004. Il y a un premier pic à 301 nT, puis un second à 350 nT. Dans ce cas-là, le réseau monosortie propose de meilleures prédictions avec un premier pic prédit à 303 nT et un second à 353 nT, tandis que le réseau multisortie prédit un premier pic à 325 nT puis un second à 275 nT. Ceci rejoint également les résultats présentés par les Figure 62 et Figure 63. En effet, dans le domaine d'activité supérieur à 300 nT, la POD pour le multisortie est plus faible que celle du monosortie (0.761 contre 0.875). Et bien que le FAR soit plus élevé avec le monosortie qu'avec le multisortie (0.195 contre 0.104), le fait que la POD soit bien supérieure pour le monosortie permet de fournir une meilleure prédiction en limitant la surévaluation de la valeur supposée par le FAR plus élevé.

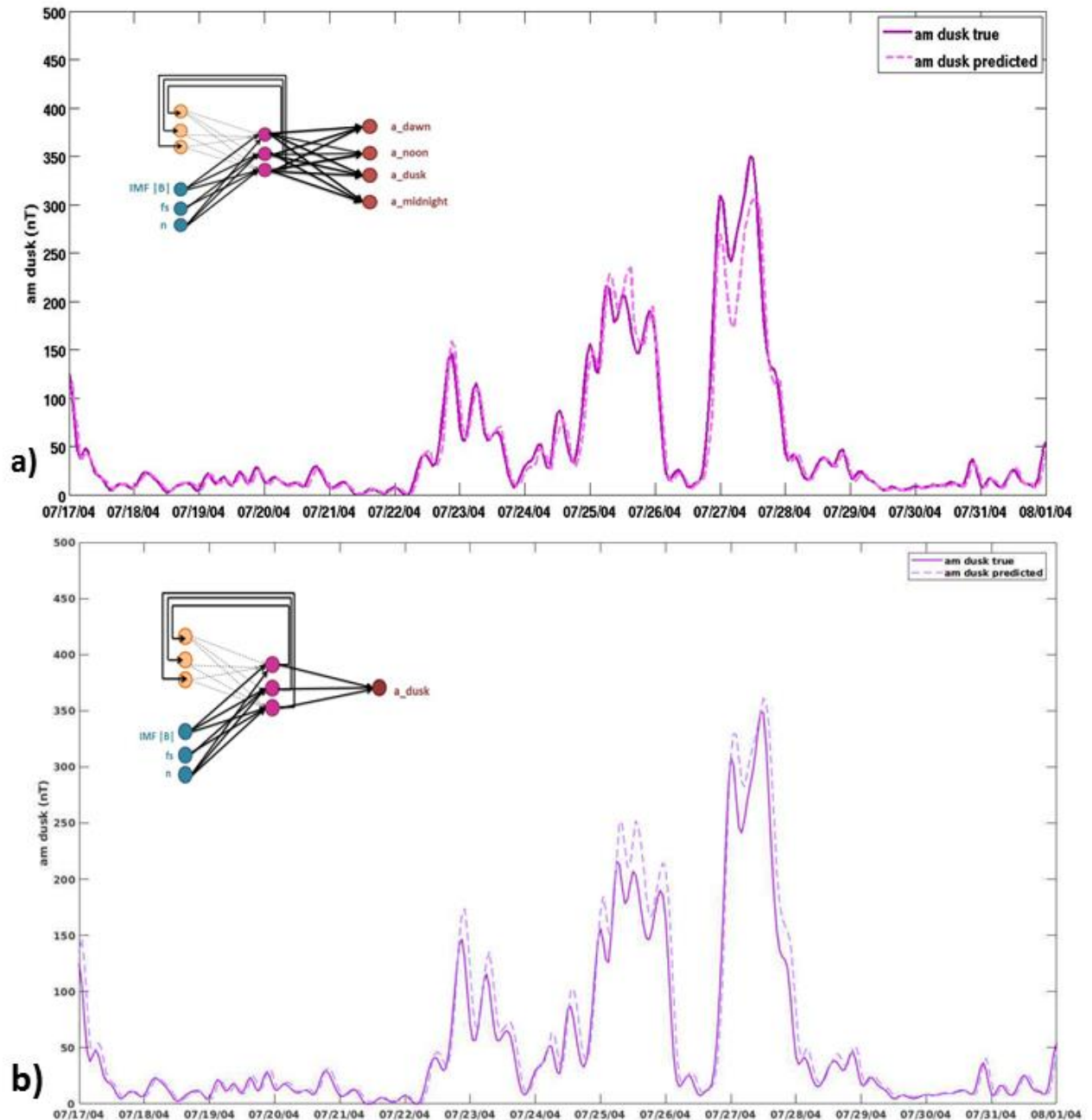


Figure 64- Evénement de juillet 2004 avec les données ACE en utilisant le réseau LSTM a) multisortie, b) monosortie pour la prédiction de $\alpha\sigma_{dusk}$. Les données réelles sont en trait continu, les données prédites sont en pointillé. Pour mieux visualiser la prédiction, un décalage d'une heure est utilisée pour la représentation graphique.

Sur la Figure 65, on constate que les valeurs prédites par le réseau multisortie et le réseau monosortie, spécifique à la prédiction de l'indice $\alpha\sigma_{midnight}$ (minuit) sont comparables à activités faibles, modérées, voir intenses comme c'est le cas pour le pic d'activité à 300 nT entre le 25 et le 26 juillet 2004. Cependant, pour le pic d'activité extrême à 475 nT entre le 27 et 28 juillet 2004, les deux réseaux ne présentent pas les mêmes performances. Le réseau multisortie prédit une valeur à 325 nT tandis que le réseau monosortie prédit une valeur à 398 nT. La valeur obtenue avec ce dernier se rapproche de la valeur réelle, mais ne correspond pas exactement. On observe également que suite à un pic aussi élevé, le réseau présente de moins bonnes performances pour prédire un pic d'activité

intense que précédemment lors des pics d'activités intenses observés entre le 25 et le 26 juillet 2004. En effet, le pic d'activité à 350 nT est prédit à 275 nT avec le multisortie, et 303 nT avec le monosortie. Ceci est probablement lié à la structure en chaîne du LSTM qui enregistre les informations obtenues par le passé afin d'évaluer les valeurs à prédire. Si l'activité à prédire a été sous-estimée, cela impacte la prédiction à l'instant suivant.

Avec cette étude de l'événement de Juillet 2004, nous avons souligné la complexité de l'analyse de réseaux multisortie en comparaison avec le réseau monosortie. Pour des activités faibles à moyennes, les deux réseaux sont globalement comparables, en revanche, pour les pics d'activité extrême, le réseau monosortie offre de meilleures prédictions. Ceci est lié au fait que les paramètres du réseau s'adaptent spécifiquement à la sortie à prédire, et non à un ensemble de sortie. Pour prédire de multiples sorties, présentant une dépendance entre elles, il sera important pour la suite de développer des techniques adaptées de réseaux de neurones ayant des sorties calculant les dépendances entre celles-ci.

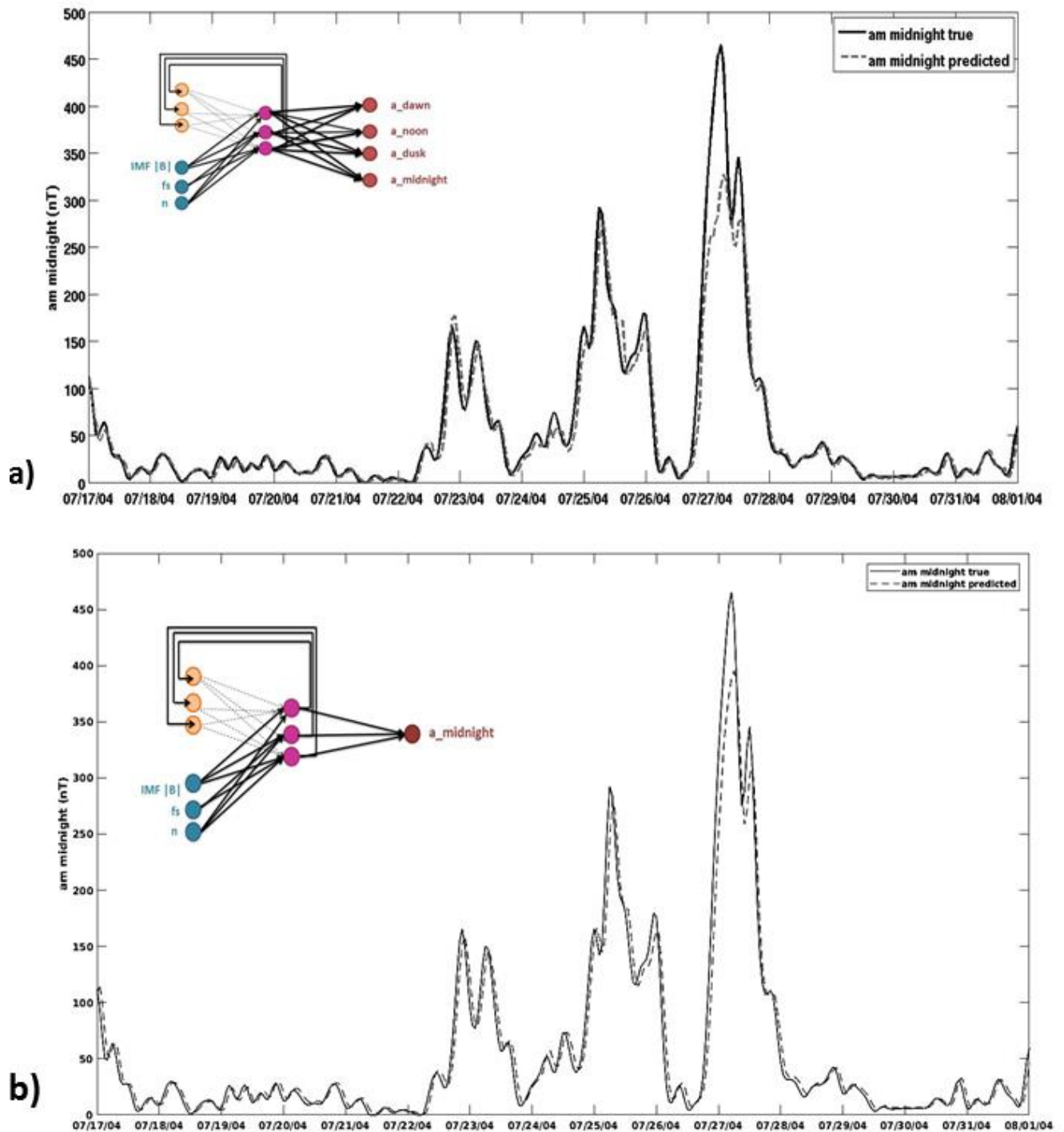


Figure 65- Evénement de juillet 2004 avec les données ACE en utilisant le réseau LSTM a) multisortie, b) monosortie pour la prédiction de $a_{\sigma_midnight}$. Les données réelles sont en trait continu, les données prédites sont en pointillé. Pour mieux visualiser la prédiction, un décalage d'une heure est utilisé pour la représentation graphique.

4. BILAN SUR LE DEVELOPPEMENT ET L'ANALYSE D'UN NOUVEAU RESEAU DE NEURONES POUR OPTIMISER LES PREDICTIONS DE L'INDICE am A PARTIR DES PARAMETRES DU VENT SOLAIRE

Dans ce chapitre sur le développement et l'analyse d'un nouveau réseau de neurones pour optimiser les prédictions de l'indice am à partir des paramètres du vent solaire, nous avons présenté le développement et les performances d'un réseau encore jamais utilisé en météorologie de l'espace, le réseau Long Short Term Memory ou LSTM. Ce réseau récurrent présente l'avantage de n'être basé que sur les paramètres du vent solaire à l'instant présent. Le TDNN que nous avons étudié au Chapitre 3 avait la capacité de fournir des prédictions également à partir des paramètres du vent solaire uniquement, mais présentait le défaut de devoir optimiser une fenêtre temporelle fixe, et présentait des performances plus faibles que le réseau récurrent NARX. Ce dernier ayant une partie autorégressive dont nous avons souhaité nous affranchir pour développer des réseaux opérationnels.

Dans un premier temps, nous avons présenté le réseau LSTM, évalué ses avantages comme une capacité d'optimisation plus rapide que les réseaux classique de type feedforward. Nous avons ensuite comparé les performances de ce réseau récurrent au réseau TDNN, et mis en évidence le fait que grâce au LSTM, nous avons réussi à améliorer les prédictions fournies par un réseau basé uniquement sur les paramètres du vent solaire.

Pour continuer à améliorer les performances du LSTM, nous avons considéré une nouvelle entrée pour celui-ci, une fonction de couplage. Plus spécifiquement, celle développée par [Wang et al., 2014] obtenue suite à une analyse MHD. Cette fonction de couplage permet d'évaluer l'entrée en énergie dans la magnétosphère associée à un événement solaire. Cette entrée en énergie est la nouvelle entrée du réseau de neurones. On peut alors connecter cette information à l'indice magnétique qui mesure la perturbation associée à cette entrée en énergie, ce qui rejoint l'essence de notre étude, modéliser le comportement de la magnétosphère au moyen des réseaux de neurones. Nous avons constaté que l'utilisation de la fonction de couplage permet d'améliorer les performances des réseaux de neurones pour l'indice am .

Enfin, afin de fournir une nouvelle information, spécifique à chaque secteur MLT, nous avons travaillé sur la prédiction de l'indice am sectoriel ou $a\sigma$. Cet indice développé par [Chambodut et al., 2013] permet de fournir à un opérateur une mesure de la perturbation magnétique en fonction du Temps Magnétique Local, et de voir l'évolution d'une perturbation en fonction de ceux-ci. Nous avons alors développé deux types de réseaux, un réseau multisortie fournissant quatre prédictions associés à $\{a\sigma_{dawn}, a\sigma_{noon}, a\sigma_{dusk}, a\sigma_{midnight}\}$, et quatre réseaux monosortie spécifiques à chaque indice sectoriel. Suite au développement et à l'optimisation de ces réseaux, nous avons constaté qu'avec les techniques utilisées actuellement pour faire de la régression, le réseau multisortie est moins performant que le réseau monosortie pour les cas d'activités les plus extrêmes. Mais à l'avenir, il sera nécessaire d'optimiser le réseau multisortie avec de nouvelles techniques, car ces réseaux sont plus en accord avec la physique observée étant donné qu'ils prennent en compte la dépendance entre les sorties. Il faudrait alors soit développer une technique de régression logistique multinomiale, méthode qui généralise les techniques de régression à des problèmes multiclassés, c'est-à-dire avec un nombre multiple de valeurs discrètes, soit transformer le problème actuel de prédiction fait à partir d'un modèle de régression, en problème de classification.

Maintenant que les prédictions à une heure de l'indice magnétique am ont été optimisées, nous avons souhaité nous tourner vers de nouvelles techniques, afin de fournir des prédictions à plus long terme. Ces travaux sont au cœur du chapitre suivant.

CHAPITRE V

OPTIMISATION DES PRÉDICTIONS D'INDICES MAGNÉTIQUES JUSQU'À SIX HEURES AU MOYEN D'UNE NOUVELLE TECHNIQUE COMBINANT RÉSEAUX DE NEURONES ET PROCESSUS GAUSSIENS

Dans ce chapitre sur l'optimisation des prédictions d'indices magnétiques jusqu'à six heures, nous avons développé une technique combinant la précision d'une prédiction fournie par un réseau de neurones, à l'aspect probabiliste d'un processus gaussien. Nous analysons d'une part l'apport de cet aspect probabiliste pour répondre au besoin d'un opérateur qui souhaite évaluer l'erreur associée à une prédiction. D'autre part, nous étudions l'apport de cette méthode hybride pour fournir une prédiction d'indices magnétiques au-delà d'une heure. Cette étude a été faite dans un premier temps pour prédire l'indice magnétique *Dst*, indice caractéristique des orages et sous-orages magnétiques. Ceci nous a permis de comparer les performances de notre nouveau modèle à des modèles de référence. Ensuite, nous avons appliqué ce processus que nous appelons GPNN (Gaussian Process – Neural Network) à la prédiction de l'indice magnétique *am*.

1. Développement d'une nouvelle méthode de prédictions associant réseaux de neurones et processus gaussiens.....	151
1.1. L'intérêt des processus gaussiens en météorologie de l'espace	151
1.2. Description de la technique dite GPNN appliquée à la prédiction d'indices magnétiques de une heure à six heures	153
2. Développement et Evaluation de la capacité du GPNN à prédire l'indice magnétique <i>Dst</i> jusqu'à six heures en avance	157
2.1. Développement et optimisation du réseau LSTM pour définir la moyenne du GPNN	157
2.1.1. L'indice magnétique <i>Dst</i> , caractéristique des orages et sous orages magnétiques	157
2.1.2. Définition du réseau LSTM pour la prédiction de l'indice magnétique <i>Dst</i>	159
2.1.3. Mise en évidence des atouts et faiblesses du LSTM à fournir des prédictions de l'indice magnétique <i>Dst</i> à partir de comparaison avec des modèles de référence	161
2.2. Analyse de la prédiction probabiliste fournie par le GPNN	164

2.2.1.	Les courbes ROC pour évaluer les performances du GPNN en fonction de seuils d'activité	165
2.2.2.	Analyse des diagrammes de fiabilité.....	169
2.2.3.	Apport du GPNN pour la prédiction d'un événement extrême	170
3.	Application de la technique combinatoire associant réseaux de neurones et processus gaussiens pour prédire l'indice magnétique <i>am</i>	171
4.	Bilan sur l'optimisation des prédictions d'indices magnétiques à plus long termes au moyen d'une nouvelle technique combinatoire.....	177

1. DEVELOPPEMENT D'UNE NOUVELLE METHODE DE PREDICTION ASSOCIANT RESEAUX DE NEURONES ET PROCESSUS GAUSSIENS

Dans les chapitres précédents, nous avons étudié la capacité de différents réseaux de neurones à fournir une prédiction à une heure de l'indice magnétique am . Au travers de ces analyses, nous avons démontré qu'il était possible de fournir des prédictions basées uniquement sur les paramètres du vent solaire, à partir des données fournies par le satellite ACE au point de Lagrange L1. Ces modèles de prévisions sont importants pour un opérateur en météorologie de l'espace car ils permettent d'aider à prendre des décisions en cas d'événements solaires pouvant impacter fortement l'environnement magnétique terrestre. Afin que ces modèles soient plus utiles dans un cadre opérationnel, il est nécessaire de fournir des prédictions à plus long terme. Pour ce faire, nous avons développé un nouvel algorithme, en faisant notamment appel à la technique des processus gaussiens que nous avons présenté au Chapitre 2, section 3.2. Ces travaux ont été développés dans le cadre d'une collaboration de trois mois avec le Centre pour les mathématiques et l'informatique d'Amsterdam, le CWI (Centrum voor Wiskunde en Informatica), sous la direction d'Enrico Camporeale, chercheur dans l'équipe Multiscale Dynamics. Enrico Camporeale travaillait avec son doctorant Mandar Chandorkar sur l'application des processus gaussiens à la prédiction de l'indice Dst , nous avons alors souhaité travailler sur le développement d'une méthode combinant leur expertise à celle que nous avons développé sur les réseaux de neurones afin de proposer un nouveau modèle.

1.1. L'intérêt des processus gaussiens en météorologie de l'espace

Les processus gaussiens sont des modèles permettant de fournir une mesure explicite de l'incertitude sur la donnée prédite. Ceci est fait en définissant un ensemble de modèles associé à un espace de solutions. Nous avons décrit les processus gaussiens en détail dans le Chapitre 2, section 3.2. Cette mesure de l'incertitude est une information clef pour un opérateur, afin d'évaluer la fiabilité de la prédiction fournie par le modèle. Précédemment, nous avons effectué un travail sur l'évaluation de la fiabilité des réseaux de neurones lors des phases d'optimisation en appliquant les mesures de probabilité de détection et les taux de fausses alarmes. Désormais nous souhaitons fournir une évaluation de l'erreur sur la mesure, à chaque prédiction fournie par le modèle.

Les processus gaussiens ont fait leur première apparition en apprentissage machine avec l'étude de [Neal, 1996], comme un cas limitant de l'inférence bayésienne appliquée à des réseaux de neurones avec un nombre infiniment grand de neurones dans la couche cachée. Bien que leurs applications en apprentissage automatique soient récentes, leurs origines remontent au domaine de la recherche en géo-statistique où ils étaient connus sous le nom de méthode de Krigeage [Krige, 1951]. Dans le domaine purement mathématique, les processus gaussiens ont été largement étudiés, et leur existence a été prouvée par l'extension du théorème de Kolmogorov [Tao, 2011]. L'application de ces processus à l'apprentissage automatique a été analysée en profondeur par [Rasmussen and Williams, 2006]. Ce dernier a été un ouvrage de référence pour le développement de notre modèle.

Les processus gaussiens n'ont été que très peu utilisés en météorologie de l'espace, afin de fournir des prédictions d'indices magnétiques. La première application a été faite par [Chandorkar et al., 2017], dans le but de fournir des prédictions à une heure de l'indice magnétique Dst , spécifique des orages et sous-orages magnétiques. Ces prédictions sont effectuées à partir de la vitesse fs et de la composante B_z du champ magnétique interplanétaire. La Figure 66 présente les résultats associés à la prédiction d'un événement extrême. On observe que la valeur prédite est proche de la valeur réelle, mais on note surtout l'existence d'un intervalle de confiance entre une valeur limite haute définie par la courbe bleue, et une valeur limite basse définie par la courbe verte. Cela signifie que grâce aux processus

gaussiens, un opérateur a accès à un ensemble de valeurs prédites, et non à un seul point de prédiction. Ce modèle rejoint alors une nécessité grandissante en météorologie de l'espace qui est l'accès aux barres d'erreur sur une prédiction en temps réel. Cette approximation de l'erreur est d'autant plus nécessaire à mesure que l'on souhaite fournir une prédiction éloignée dans le temps. En effet, la dépendance entre paramètres du vent solaire et indices magnétiques diminue avec l'augmentation du temps de prédiction.

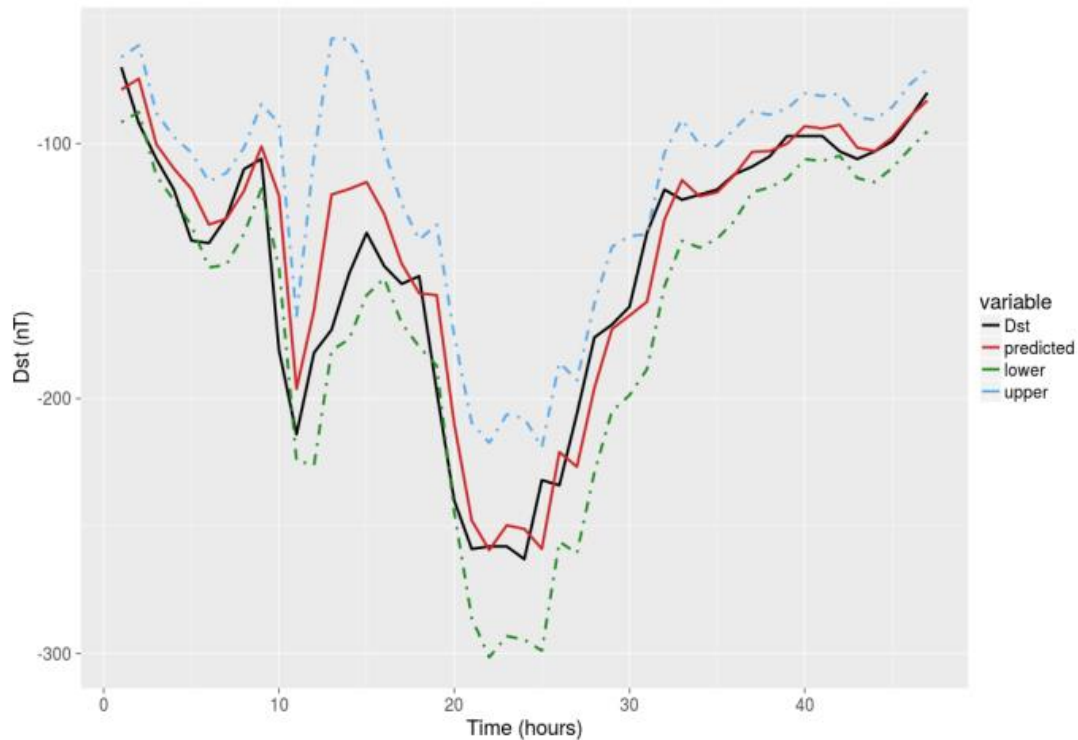


Figure 66- Prédiction de l'indice Dst au moyen des processus gaussiens. La valeur réelle est en noire, la valeur prédite en rouge, les valeurs limites hautes et basses sont définies par les enveloppes vertes et bleues [Chandorkar et al., 2017].

Les processus gaussiens ont alors trouvé leur place dans ce domaine, en démontrant de très bonnes performances pour prédire l'indice magnétique *Dst* à une heure. Cependant, les mesures globales de performance démontrées par les processus gaussiens ne sont pas aussi optimales que celles fournies par les réseaux de neurones. C'est pourquoi nous avons souhaité développer un modèle combinant la précision des réseaux de neurones, et l'aspect probabiliste des processus gaussiens, afin de produire des prédictions d'indices magnétiques de une heure à six heures.

Nous avons appliqué ce modèle dans un premier temps à la prédiction de l'indice *Dst* puis dans un second temps à l'indice *am*. Nous avons fait ce choix car durant tout ce travail de thèse, nous avons développé des techniques afin de prédire l'indice magnétique *am* pour lequel il n'existait aucun autre modèle de prédiction. Après avoir optimisé le développement du LSTM pour la prédiction de l'indice *am*, il était donc judicieux d'utiliser ce réseau de neurones pour prédire un autre indice pour lequel il existe des modèles de prédiction jusqu'à six heures (voir [Wu and Lundstedt, 1997], [Bala and Reiff, 2012] et [Lazzús et al., 2017]). Ceci nous permet alors de comparer les performances du LSTM à ces modèles. Le choix de l'indice *Dst* a également été soutenu par le fait que ce projet a été fait en

collaboration avec l'équipe Multiscale dynamics du CWI, au sein de laquelle il existe une expertise sur le développement de modèles pour la prédiction de l'indice Dst .

1.2. Description de la technique dite GPNN appliquée à la prédiction d'indices magnétiques de une heure à six heures

Afin de développer la technique combinant les réseaux de neurones aux processus gaussiens, technique que nous appelons par la suite GPNN (pour Gaussian Process - Neural Network), nous avons effectué les différentes étapes décrites sur la Figure 67.

Dans un premier temps, nous avons défini un réseau LSTM pour la prédiction de l'indice Dst , afin d'obtenir un vecteur contenant des valeurs prédites de une heure à six heures de l'indice Dst , que l'on note $\widehat{Dst}(t+p)_{NN}$, p représentant le délai auquel on souhaite prédire l'indice.

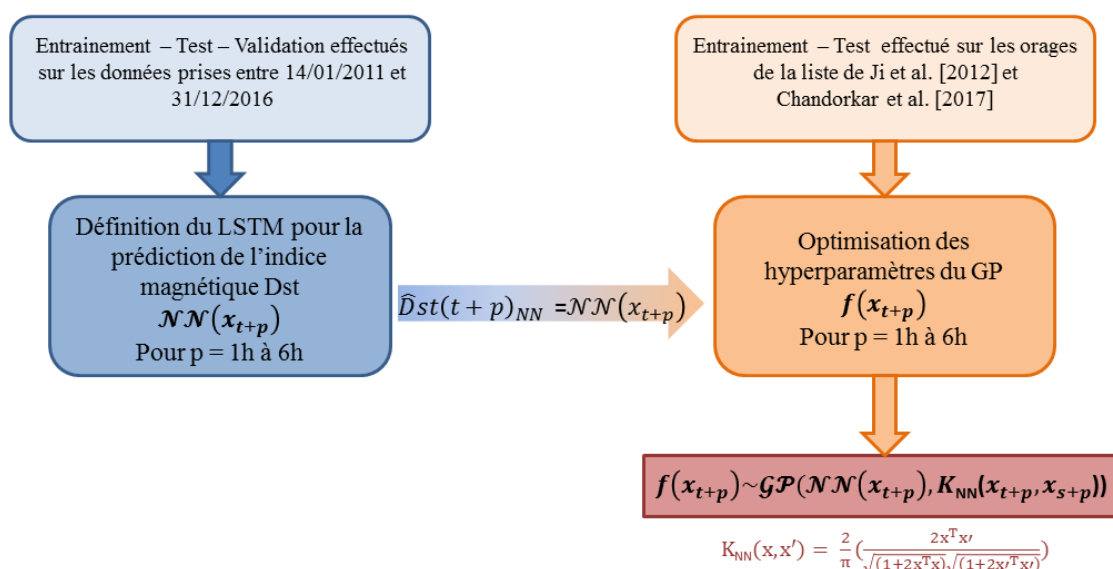


Figure 67- Etapes principales du développement du GPNN.

Pour définir ce réseau, nous avons utilisé des données décrites au Chapitre 2 section 1.3. Ces données correspondent aux paramètres du vent solaire, ainsi qu'aux données GPS. Nous entrons plus en détail sur les données utilisées à la section suivante.

Ces données sont réparties en trois sous-ensembles afin d'entraîner, tester et valider les réseaux LSTM. Dans le cadre de la prédiction de une heure à six heures, deux choix se présentent à nous, comme lorsque nous avons travaillé sur la prédiction des indices magnétiques sectoriels. Il est possible de développer un réseau multisortie fournissant un vecteur contenant les six valeurs prédites de l'indice magnétique Dst , ou de développer six réseaux, chacun étant spécifique à une prédiction de une heure jusqu'à six heures. Suite à l'étude faite sur la prédiction des indices sectoriels au Chapitre 4 section 3, nous avons conclu qu'avec les techniques développées actuellement, la prise en compte de la dépendance entre les sorties n'est pas correctement faite. Nous avons également déduit que pour obtenir des réseaux plus performants, il était plus judicieux d'utiliser un réseau spécifique à une sortie

et non à plusieurs afin d'optimiser les poids et biais sur chaque indice. Nous développons donc six réseaux de neurones, un pour chaque temps de prédiction, et nous concaténons les valeurs prédites par chacun des réseaux pour obtenir un vecteur $\widehat{Dst}(t+p)_{NN}$, utilisés en entrée du processus gaussien. Ceci définit la partie gauche du schéma représentée en bleu sur la Figure 67.

Ainsi, une fois que les réseaux sont optimisés, nous pouvons utiliser les données prédites par les réseaux, soit $\widehat{Dst}(t+p)_{NN}$, pour définir la valeur moyenne du processus gaussien. Cette étape est signalée par la flèche reliant la partie réseau de neurones en bleu, à la partie processus gaussien en orange sur la Figure 67. Au Chapitre 2 section 3.2.3, nous avons exprimé un processus gaussien comme étant décrit par une moyenne $m(x)$ et une fonction de covariance, $k(x, x')$ (équation (45)).

$$f(x) \sim GP(m(x), k(x, x')) \quad (45)$$

où x et x' représentent deux points dans l'espace des variables d'entrée.

La valeur moyenne est ainsi décrite par la sortie du LSTM. Pour la fonction de covariance, nous avons fait le choix d'utiliser la fonction de types réseaux de neurones. [Chandorkar et al., 2017] avaient analysé de multiples fonctions de covariance, et la fonction de type réseaux de neurones décrite par l'équation (46) paraît adaptée pour spécifier la façon dont chaque point influence la valeur que les autres points sont susceptibles de prendre dans le cadre de la prédiction d'indices magnétiques.

$$k_{NN}(x, x') = \frac{2}{\pi} \left(\frac{2x^T x'}{\sqrt{(1+2x^T x)}\sqrt{(1+2x'^T x')}} \right) \quad (46)$$

Afin de pouvoir effectuer des prédictions à partir des processus gaussiens, il est nécessaire, comme le décrit le principe de l'inférence bayésienne sur lequel repose les fondements du processus gaussien, d'avoir une distribution a priori sur les fonctions à évaluer. On restreint cette distribution aux fonctions passant par des points de données observés. Ce processus est décrit en détail dans le Chapitre 2 section 3.2.3. On définit alors un ensemble d'événements d'entraînement (correspondant à la distribution a priori) pour ajuster les hyperparamètres du processus gaussien (comme les paramètres de la fonction de covariance). Les événements utilisés pour cet entraînement sont présentés sur la Figure 68. On utilise une période de temps calme avec des valeurs de Dst variant entre 12 et -37 nT, et trois périodes d'orages magnétiques, extraits de la liste de [Ji et al, 2012] : un orage modéré avec un pic à -147 nT, et deux orages intenses avec un pic à -395 nT et un pic à -272 nT. Ces événements ont été déterminés suite à des itérations sur des sous-ensembles d'entraînement-test que nous décrivons à la suite.

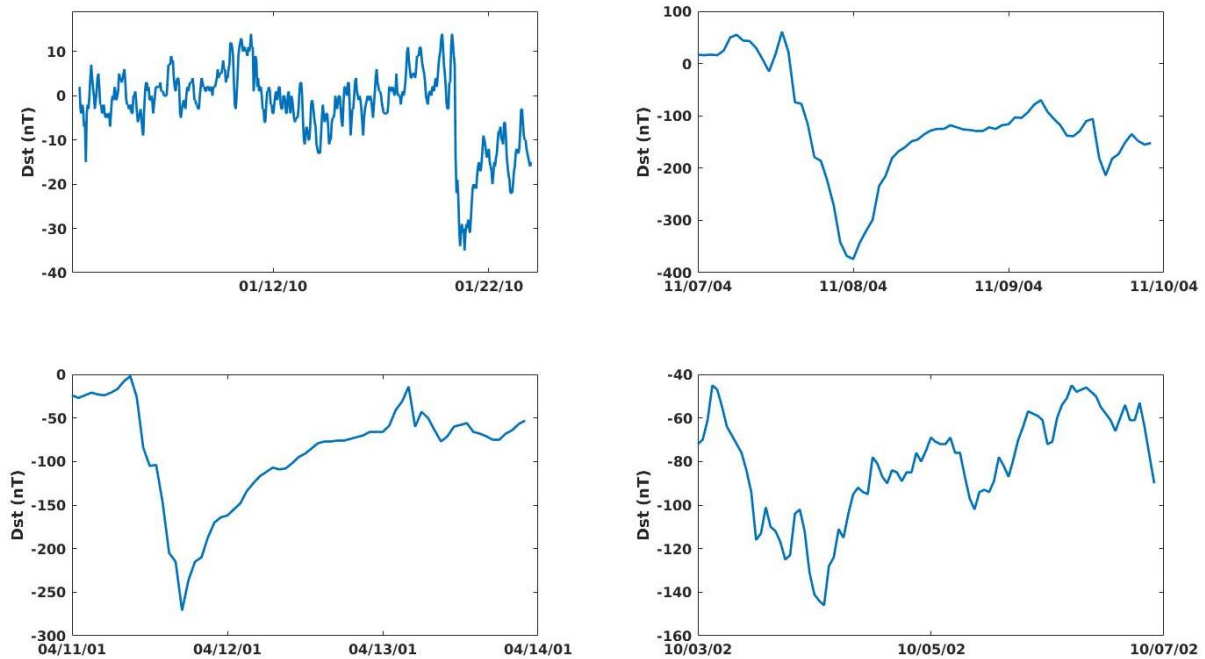


Figure 68- Événements considérés pour l'entraînement du processus gaussien.

Dans un premier temps, nous faisons une première évaluation des hyperparamètres du processus gaussien sur le sous-ensemble d'entraînement en évaluant l'erreur quadratique moyenne et le coefficient de corrélation. Une fois que cette optimisation est faite sur le sous-ensemble d'entraînement, nous évaluons le processus gaussien sur un sous-ensemble de tests, plus précisément une liste d'événements correspondant à des orages magnétiques modérés à intenses. Nous pouvons alors évaluer les performances du processus gaussien pour prédire l'indice magnétique Dst avec les erreurs quadratiques moyennes et coefficient de corrélation sur le sous-ensemble de test. Le processus d'entraînement - test est effectué de façon itérative pour définir un sous-ensemble d'entraînement optimal pour fixer les hyperparamètres de façon définitive. Cette analyse nous a conduit à définir les sous-ensemble présentés sur la Figure 68, et les orages utilisés pour tester le processus gaussien sont présentés en annexe 3. Une fois que les hyperparamètres sont fixés, le processus gaussien basé sur une valeur moyenne fournie par le vecteur de sortie des réseaux LSTM est alors capable de fournir en quelques secondes un vecteur contenant les valeurs prédites de une heure à six heures de l'indice magnétique Dst , avec un intervalle de confiance correspondant à la distribution autour de la valeur moyenne. L'intervalle de confiance permet d'évaluer la précision de l'estimation d'un paramètre statistique sur un ensemble de données. Dans notre cas, nous considérons un intervalle de confiance à 95 %, illustré par la Figure 69. Cela signifie qu'on définit un encadrement correct 95 fois sur 100 en moyenne, donc que si on pouvait répéter des estimations similaires, en affirmant à chaque fois que la prédiction se trouve dans cet intervalle, on se tromperait en moyenne 5 fois sur 100.

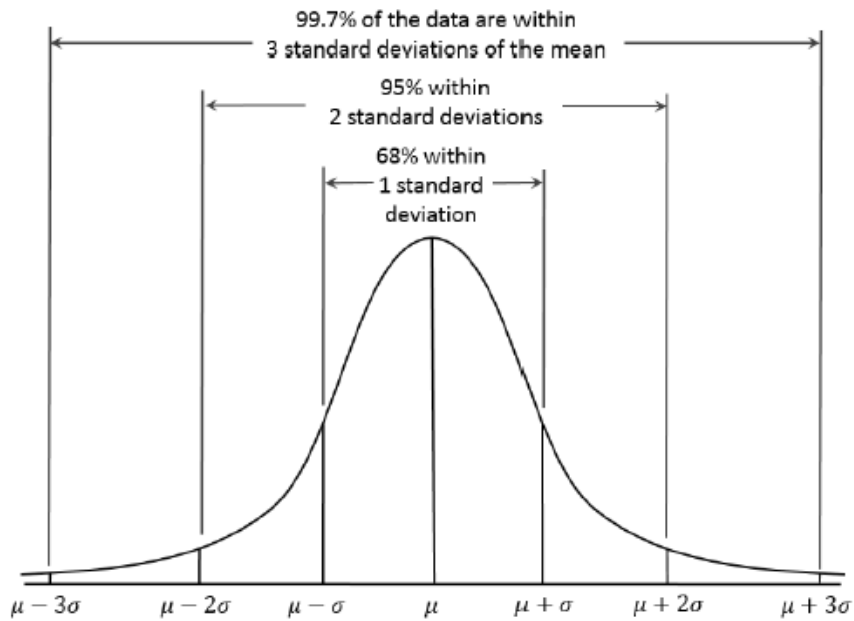


Figure 69 - Intervalle de confiance pour évaluer la précision de l'estimation.

Le mode opératoire utilisable en temps réel est décrit par la Figure 70. Les paramètres du vent solaire alimentent les entrées des réseaux LSTM, qui fournissent un vecteur de valeurs prédites de l'indice *Dst* de une heure à six heures, utilisés comme moyenne du processus gaussien.

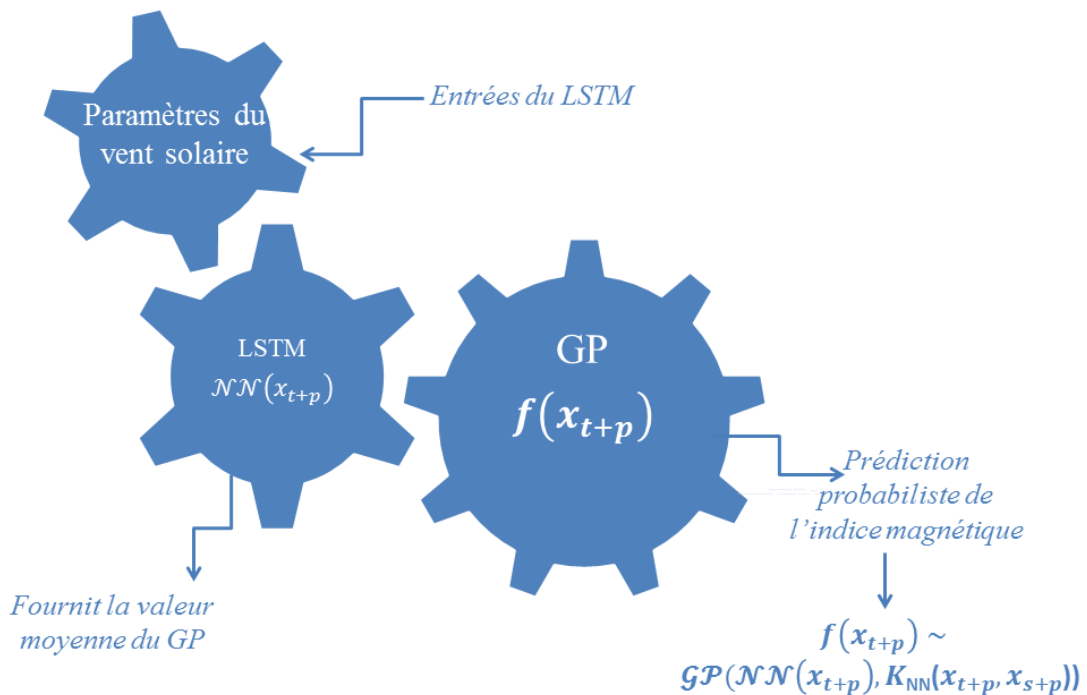


Figure 70- Fonctionnement opérationnel du GPNN.

Dans les sections suivantes, nous décrivons les données utilisées en entrée du processus combinant réseaux de neurones et processus gaussien, l'entraînement et la validation des réseaux de neurones, ainsi que l'optimisation et l'évaluation du GPNN au moyen de métriques spécifiques à l'analyse de prédiction probabiliste.

2. DEVELOPPEMENT ET EVALUATION DE LA CAPACITE DU GPNN A PREDIRE L'INDICE MAGNETIQUE *DST* JUSQU'A SIX HEURES EN AVANCE

Afin de développer le modèle GPNN, nous devons dans un premier temps optimiser les réseaux LSTM. Pour ce faire, nous définissons d'abord un ensemble de données d'entrées à partir des paramètres du vent solaire et des données GPS. Pour évaluer les performances des réseaux LSTM, nous comparons celles-ci à des réseaux de référence développés par le passé pour la prédiction de l'indice magnétique *Dst* de une heure à six heures.

Une fois que les réseaux sont optimisés, nous pouvons travailler sur le développement du processus gaussien utilisant en valeur moyenne les valeurs prédites des LSTM. Pour les évaluer, nous utilisons des courbes ROC et des diagrammes de fiabilité.

2.1. Développement et optimisation du réseau LSTM pour définir la moyenne du GPNN

2.1.1. L'indice magnétique *Dst*, caractéristique des orages et sous-orages magnétiques

Les différents mécanismes décrits au Chapitre 1, section 2 nous permettent de comprendre que la magnétosphère terrestre est en permanence soumise à une dynamique, d'intensité variable. Nous avons vu au Chapitre 1 section 2.2 que lorsque le champ magnétique est orienté vers le Sud ($B_z < 0$), on observe des mécanismes de reconnexion magnétique au niveau de la queue magnétosphérique. La circulation du vent solaire à travers le champ interconnecté se traduit aussi par l'apparition d'une différence de potentiel d'un flanc à l'autre de la magnétosphère et donc par la présence d'un champ électrique dirigé de l'est vers l'ouest. Ce champ électrique va être responsable (via la dérive électrique) de la circulation à grande échelle du plasma dans la magnétosphère, la convection magnétosphérique. Il va aussi avoir pour effet d'intensifier le système de courant de la queue et donc d'y accumuler de l'énergie magnétique. Lorsque de l'énergie est relâchée depuis la queue magnétosphérique vers la magnétosphère interne, ceci a pour effet de créer de violentes perturbations magnétiques.

Au 19^{ème} siècle et dans la première moitié du 20^{ème}, les perturbations magnétiques les plus étudiées furent les "orages magnétiques", ainsi que les avait dénommés Alexandre von Humboldt. Ces perturbations du champ magnétique sont mondiales et aisément observées un peu partout. Typiquement, un orage magnétique se développe sur une demi-journée, et disparaît progressivement en quelques jours.

Les orages magnétiques sont relativement rares. Par contre, de plus petits "sous-orages" sont beaucoup plus fréquents, à quelques heures d'intervalle. Ces deux variétés de perturbations sont naturellement en relation, et, au cours des orages magnétiques, on observe généralement dans les régions polaires d'intenses sous-orages. Les orages sont classés d'après le nombre d'ions et d'électrons injecté de la queue dans la ceinture externe de radiation, et par la perturbation magnétique qu'ils entraînent reflétant une croissance rapide du courant annulaire. On peut considérer les orages magnétiques comme des séries de sous-orages très intenses, mais il y a des facteurs additionnels. En

particulier il faut aux orages magnétiques des stimuli externes comme l'arrivée d'une onde de choc ou d'un jet rapide de vent solaire.

D'un point de vue indice magnétique, des signatures sont caractéristiques des sous-orages et des orages. Dans les deux cas, l'indice AE qui est l'indice caractéristique des électrojets auroraux montre des signatures dans le cas des sous-orages et des orages. L'indice Dst qui est lui caractéristique du courant annulaire, et donc d'événements plus importants en terme d'énergie, ne montre des signatures que dans le cas d'orages magnétiques.

Un orage magnétique peut être décomposé en trois phases distinctes illustrées à la Figure 71.

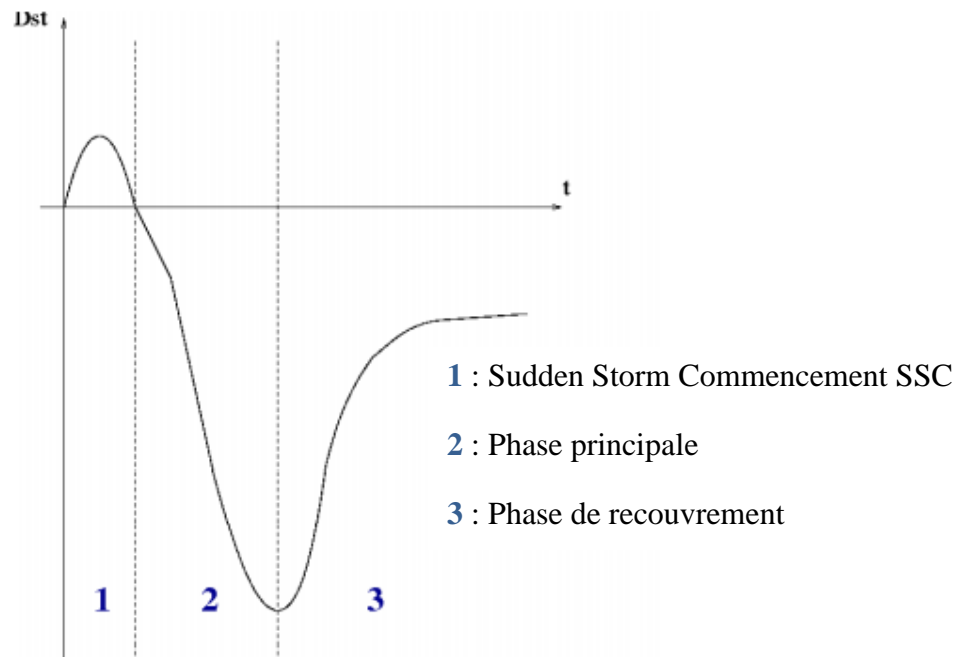


Figure 71: Phases d'un orage magnétique en fonction du Dst .

- Le SSC :

Il s'agit d'une phase transitoire de courte durée (de l'ordre de quelques heures) durant laquelle des particules à haute vitesse, émises par le Soleil et transportées par le vent solaire, ainsi que le champ magnétique interplanétaire, viennent comprimer le côté jour de la magnétosphère entraînant un changement soudain du champ magnétique observé à la surface de la Terre. Cette phase s'accompagne d'une augmentation du Dst en réponse au resserrement des lignes de champ et à la perte de particule dans le courant annulaire.

- La phase principale :

Cette phase se déroule sur plusieurs heures. Lors de l'ouverture des lignes de champ du côté jour de la magnétosphère, une partie de l'énergie du vent solaire est directement dissipée dans l'ionosphère aurorale et une autre partie est emmagasinée dans la queue. Cette énergie est par la suite convertie en énergie cinétique entraînant l'accélération de particules du feuillet neutre (de quelques keV à quelques centaines de keV) en direction de la Terre et leur injection dans la magnétosphère interne ou leur précipitation dans les zones aurorales. Une autre partie peut néanmoins être éjectée dans la direction

antisolaire (plasmoïdes, jets antisolaires). Une chute du Dst est alors observable, correspondant à une intensification du courant annulaire. Par ailleurs, la plasmopause se rapproche de la Terre. Au cours de cette phase, le niveau de flux dans la magnétosphère interne peut facilement être multiplié jusqu'à trois ordres de grandeur en l'espace de quelques heures.

- La phase de relaxation

La phase de diminution des perturbations électromagnétiques peut s'étaler sur plusieurs jours. Elle est marquée par une augmentation du Dst , et une diminution de l'indice Kp , plus ou moins rapide. La plasmopause reprend alors sa configuration initiale.

2.1.2. Définition du réseau LSTM pour la prédiction de l'indice magnétique Dst

Pour prédire l'indice magnétique Dst , nous avons utilisé en entrée du GPNN un ensemble de paramètres du vent solaire et de données GPS. Le choix des données du vent solaire a été guidé par les précédentes études faites sur la prédiction de l'indice magnétique Dst de une heure à six heures. Ces données sont résumées dans le Tableau 11. Nous avons alors retenu comme paramètres du vent solaire la vitesse fs , la densité n , l' $IMF |B|$ et la composante B_z du champ magnétique. Nous utilisons également en entrée les données GPS fournies par le satellite ns41, car celui-ci possède la couverture temporelle la plus importante (comme nous l'avons présenté au Chapitre 2 section 1.3). Plus précisément, nous considérons le champ magnétique mesuré au niveau du satellite $Bsat_{GPS}$. Ces données étant rendues publiques depuis peu [Morley et al. 2017], nous souhaitons évaluer l'apport de celles-ci sur les prédictions d'un indice spécifique aux orages magnétiques. A l'heure actuelle, il n'y a pas d'études qui ait mis en évidence des relations spécifiques entre variation de Dst et perturbations sur satellites GPS. Une étude de [Kumar et al., 2012] a constaté des chutes rapides du Dst lors de variation au niveau du Total Electron Content (TEC) mesuré par les satellites GPS, mais pas de corrélation permettant de souligner un lien direct entre variation du Dst et perturbations sur mesures GPS.

Comme il a été mis en évidence qu'il existe une dépendance forte entre la valeur actuelle de l'indice Dst , et ses valeurs passées, nous considérons également en entrée les valeurs précédentes de cet indice magnétique, jusqu'à six heures, $Dst(t - 1h), Dst(t - 2h), \dots, Dst(t - 6h)$.

On définit alors l'équation (47) pour fournir des prédictions de une heure à six heures de l'indice magnétique Dst avec les réseaux LSTM, le $Bsat_{GPS}(t)$ étant optionnel car nous souhaitons évaluer dans la section suivante l'apport de cette donnée.

$$\begin{aligned} \widehat{Dst}(t + p)_{NN} = \mathcal{NN}(n(t), V(t), IMF|B|(t), Bz(t), Bsat_{GPS}(t), \\ Dst(t - 1h), Dst(t - 2h), \dots, Dst(t - 6h)) \end{aligned} \quad (47)$$

Tableau 11- Etudes sur la prédiction de l'indice magnétique Dst de une heure à six heures.

Article de référence	Données utilisées en entrée
Bala, R., and P. Reiff (2012),	Indice de Boyle : $\Phi(kV) = v^2 \cdot 10^{-4} + 11,7B\sin^3\left(\frac{\theta}{2}\right)$ Avec v le flow speed, B le champ magnétique interplanétaire (IMF B) et θ l'IMF clock angle
Lazzús, J. A., Vega, P., Rojas, P., and I. Salfate. (2017)	Purement autorégressif
Wu, J.-G., and H. Lundstedt (1997),	$IMF B $ B_s , composante dirigée vers le sud de l'IMF B $Si B_z < 0, B_s = -B_z, si B_z > 0, B_s = 0$ n la densité du vent solaire f_s le flow speed $P = nv^2$ la pression dynamique du vent solaire $\varepsilon = vB^2L^2 \sin^4\left(\frac{\theta}{2}\right)$ avec $L = 7$ rayon terrestre l'équation d'Akasofu (1981)

Dans les chapitres précédents, pour entraîner les réseaux de neurones nous utilisons un gradient de descente de type Levenberg Marquardt, décrit au Chapitre 2 section 3.1.1.3. Ce gradient est largement utilisé dans le cadre du développement de réseaux de neurones. Depuis peu, un nouvel algorithme d'optimisation est sorti, le *RMSprop*. C'est une méthode d'apprentissage avec un taux d'apprentissage adaptatif proposé par Geoff Hinton. Elle n'est pas publiée à l'heure actuelle¹. Les paramètres comme les poids et les biais du réseau sont décrits en utilisant la notation θ_i . On peut alors définir avec l'équation (48) le gradient de la fonction J noté $g_{t,i}$ associée aux paramètres θ_i au temps t .

$$g_{t,i} = \nabla_{\theta} J(\theta_{t,i}) \quad (48)$$

La mise à jour des paramètres en utilisant la méthode *RMSprop* est décrite par l'équation (49). Tout d'abord on calcule la moyenne glissante $E(g^2)$ au temps t , puis on l'applique au calcul des paramètres θ_i

$$E(g^2)_{t,i} = 0.9E(g^2)_{t-1,i} + 0.1 g_{t,i}^2 \quad (49)$$

¹ La description de la méthode *RMSprop* n'est pas publiée mais est décrite sur ce site http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf

$$\theta_{t+1,i} = \theta_{t,i} - \frac{\eta}{E[g^2]_{t,i} + \epsilon} g_{t,i}$$

avec η le taux d'apprentissage et ϵ un terme permettant d'éviter la division par zéro.

Pour développer les LSTM spécifiques à chaque temps de prédiction, la base de données est divisée en 3 sous-ensembles : 70% pour l'entraînement, 20 % pour le test, et 10 % pour la validation, comme illustré sur la Figure 19.

2.1.3. Mise en évidence des atouts et faiblesses du LSTM à fournir des prédictions de l'indice magnétique *Dst* à partir de comparaison avec des modèles de référence

Pour évaluer les performances des réseaux LSTM, nous comparons les coefficients de corrélation et erreurs quadratiques moyennes obtenus avec et sans les données GPS, avec ceux obtenus par les modèles de [Wu and Lundstedt, 1997], [Bala and Reiff, 2012] et [Lazzús et al., 2017]. Les couvertures temporelles de chacune de ces études sont représentées sur la Figure 72.

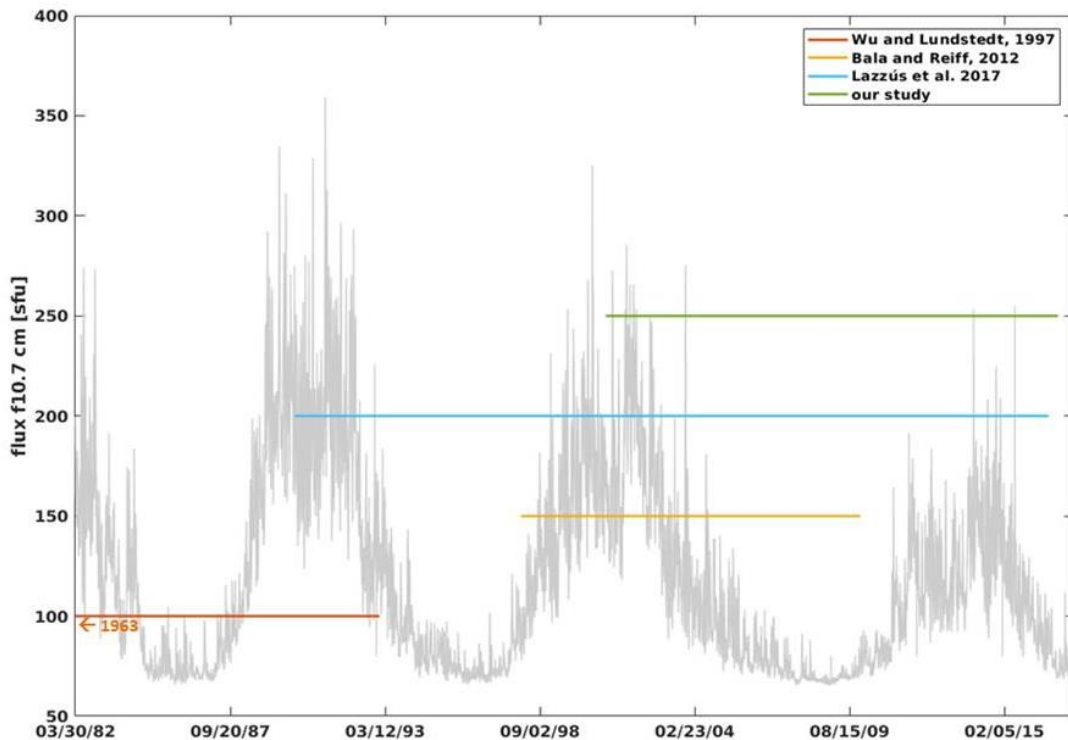


Figure 72 - Couverture temporelle des études de référence et de notre étude.

Les Tableau 12 et Tableau 13 présentent les résultats de cette comparaison. Sur ces tableaux sont également représentés la persistance. La persistance = utilise la valeur précédente de *Dst* pour prédire le *Dst* de l'instant suivant, soit $Dst = Dst(t - 1)$. C'est un modèle simple qui peut être utilisé comme « baseline » ou référence. On observe que ce modèle simple présente cependant des performances correctes dues à la corrélation forte entre les valeurs des *Dst* sur un délai d'une heure. Cette corrélation diminue avec le temps. La persistance à une heure a un CC de 0.925 et une RMSE de 9.32 nT, et à six heures un CC de 0.859 et une RMSE de 19.8 nT.

Les réseaux LSTM, avec ou sans les données GPS, fournissent des performances proches de celles obtenues par [Lazzús et al., 2017] de une heure à trois heures. Pour des temps de prédiction supérieures, nos modèles avec ou sans les données GPS présentent de meilleures performances. Par exemple, en considérant une prédiction à six heures, notre modèle avec les données GPS a un CC de 0.873 et un RMSE de 9.86 nT, tandis que le modèle de [Lazzús et al., 2017] a un CC de 0.826 et un RMSE de 13.09 nT. Etant donné que le modèle de [Lazzús et al., 2017] est basé uniquement sur les valeurs passées de l'indice *Dst*, cela montre l'apport de l'utilisation de données exogènes comme les paramètres du vent solaire ou le champ magnétique mesuré par le GPS.

[Bala and Reiff, 2012] ont utilisé en entrée du réseau de neurones l'indice de Boyle, décrit dans le Tableau 11 et ont obtenu des performances similaires aux nôtres. Si on considère à nouveau une prédiction à six heures, leur modèle présente un CC de 0.77 et une RMSE de 11.09 nT, ce qui est un peu moins élevé que nos prédictions obtenues avec ou sans les données GPS.

Nous avons également comparé nos modèles avec ceux développés par [Wu and Lundstedt, 1997] car c'est le premier modèle utilisant des réseaux récurrents. Nous souhaitons comparer les performances d'un réseau récurrent classique à celles obtenues avec le LSTM, et voir comment la complexité de la structure du LSTM permet d'améliorer les prédictions. Le modèle développé par [Wu and Lundstedt, 1997] présente à six heures un CC de 0.82 et une RMSE de 20.8 nT. Ainsi, la structure spécifique du LSTM permet d'obtenir davantage de précision sur les prédictions qu'un réseau récurrent classique.

Enfin, nous constatons qu'en utilisant les données GPS, on obtient de meilleures performances pour prédire l'indice magnétique *Dst* de une heure à six heures. En effet, à tout temps de prédiction considéré, le coefficient de corrélation est plus élevé, et l'erreur quadratique moyenne est plus faible.

Tableau 12- Coefficient de corrélation des modèles de prédictions de l'indice *Dst* de une heure à six heures. Les couvertures temporelles utilisées pour l'évaluation de ces modèles sont présentées Figure 72.

<i>Correlation Coefficient (CC)</i>						
	<i>Persistence</i>	<i>Our model</i>	<i>Our model using GPS data</i>	<i>Lazzús et al. 2017</i>	<i>Bala and Reiff 2012</i>	<i>Wu and Lundstedt 1997</i>
<i>t+1h</i>	<i>0,925</i>	<i>0,966</i>	<i>0,966</i>	<i>0,982</i>	<i>0,86</i>	<i>0,91</i>
<i>t+2h</i>	<i>0,918</i>	<i>0,946</i>	<i>0,946</i>	<i>0,949</i>	-	<i>0,89</i>
<i>t+3h</i>	<i>0,916</i>	<i>0,923</i>	<i>0,928</i>	<i>0,918</i>	<i>0,84</i>	<i>0,86</i>
<i>t+4h</i>	<i>0,885</i>	<i>0,902</i>	<i>0,910</i>	<i>0,887</i>	-	<i>0,83</i>
<i>t+5h</i>	<i>0,875</i>	<i>0,882</i>	<i>0,892</i>	<i>0,858</i>	-	<i>0,82</i>
<i>t+6h</i>	<i>0,859</i>	<i>0,865</i>	<i>0,873</i>	<i>0,826</i>	<i>0,77</i>	<i>0,82</i>

La Figure 73 présente les résultats obtenus avec les réseaux LSTM pour prédire l'indice magnétique *Dst* durant l'événement d'Halloween 2003 de une heure à six heures, avec les données GPS en bleu, et sans les données GPS en rouge. On constate que les prédictions obtenues à une heure et deux heures

sont similaires. En revanche, si on considère les prédictions faites à trois heures, le modèle sans les données GPS fournit une prédiction du pic à -348 nT tandis que celui avec les données GPS prédit un pic à -405 nT, la valeur réelle étant à -422 nT. Pour une prédiction faite à quatre heures, le modèle sans les données GPS prédit ce pic à -355 nT et celui avec les données GPS prédit ce pic à -380 nT. A cinq heures, les valeurs prédites pour ce pic à -422 nT sont quasiment similaires, autour de -365 nT. A six heures, on constate que le réseau n'est plus suffisant pour établir une connexion entre les différents paramètres considérés en entrée du réseau de neurones, et l'indice magnétique *Dst*. Nous avons pu voir notamment en analysant la persistance que la corrélation entre l'indice *Dst* à six heures et celui des instants précédents décroît. Il peut aussi y avoir un effet lié à la réponse de la magnétosphère à un événement extrême de type super orage. La magnétosphère se comporte comme une capacité qui se charge et se décharge en fonction du temps et des événements qui l'impactent [Rochel et al. 2016]. Ainsi, il est probable qu'à six heures, les effets associés à l'événement solaire considéré enregistré au niveau de l'onde de choc ne soient plus associables aux perturbations magnétosphériques mesurées par les indices magnétiques. Il est important en effet de rappeler que les données que nous utilisons sont fournies par la base de données OMNI, donc que les résultats obtenus analysent les relations mises en place par le réseau de neurones LSTM entre des paramètres mesurées à l'onde de choc, ainsi que les données mesurées par un GPS, donc dans l'environnement proche terrestre, à l'indice magnétique *Dst*.

Tableau 13- Root mean square error des modèles de prédictions de l'indice *Dst* de une heure à six heures.

<i>Root Mean Square Error (RMSE)</i>						
	<i>Persistence</i>	<i>Our model</i>	<i>Our model using GPS data</i>	<i>Lazzús et al. 2017</i>	<i>Bala and Reiff 2012</i>	<i>Wu and Lundstedt 1997</i>
<i>t+1h</i>	9,32	5,34	5,25	4,24	8,83	14,5
<i>t+2h</i>	10,2	6,65	6,55	7,05	-	16,3
<i>t+3h</i>	13,3	7,86	7,59	8,87	9,6	18,2
<i>t+4h</i>	14,5	8,86	8,53	10,44	-	19,9
<i>t+5h</i>	17,5	9,59	9,18	11,65	-	20
<i>t+6h</i>	19,8	10,24	9,86	13,09	11,09	20,8

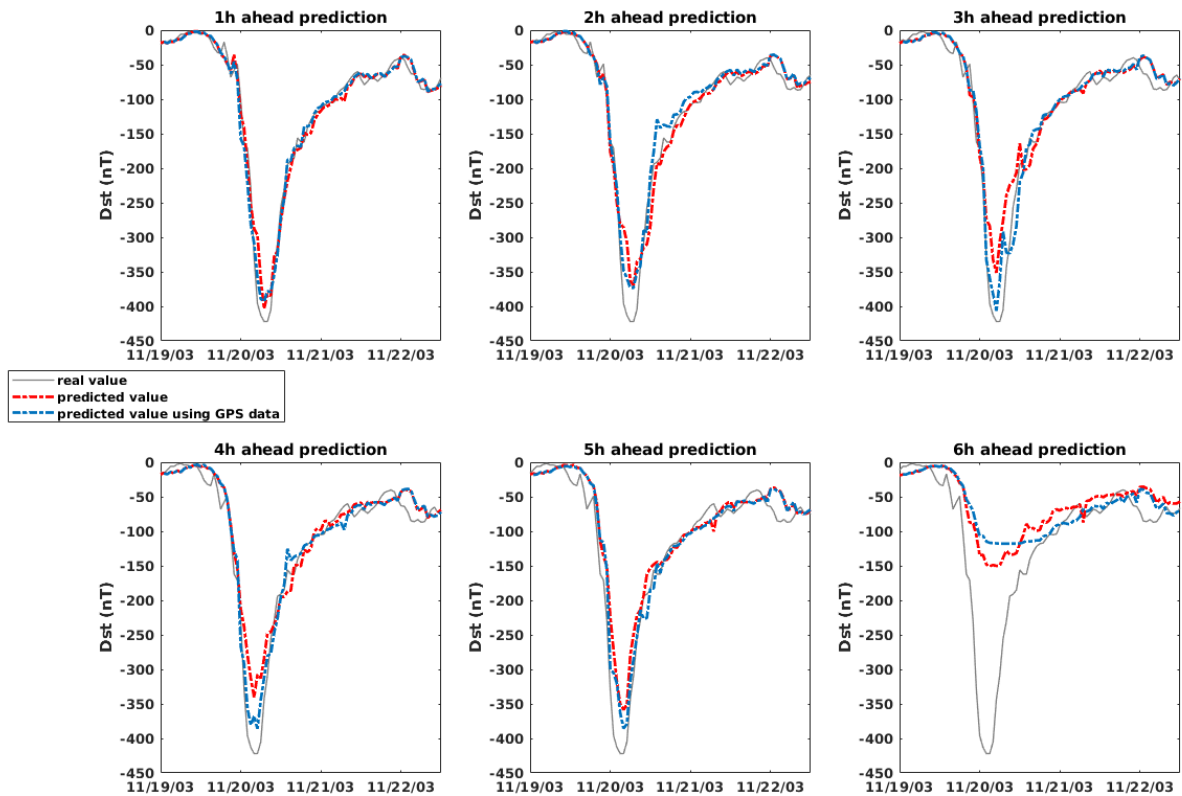


Figure 73- Prédications obtenues sans données GPS (en rouge) et avec les données GPS (en bleu) pour l'événement d'Halloween 2003. La valeur réelle est en gris.

Ainsi, nous avons développé six réseaux LSTM qui fournissent des prédictions de une heure jusqu'à six heures de l'indice magnétique Dst . Nous avons constaté que les données GPS peuvent apporter un gain de précision en fonction du temps de prédiction. Nous décidons de conserver la mesure GPS du champ magnétique $Bsat_{GPS}$ dans le développement à suivre. Les prédictions de l'indice magnétique Dst obtenues à partir des données GPS ainsi que des autres données décrites par l'équation (47) sont alors utilisées comme moyenne du processus gaussien, afin de fournir une prédiction probabiliste de une heure à six heures avec une barre d'erreur. Dans la section suivante, nous analysons la prédiction fournie par le GPNN, au travers de mesures spécifiques comme la courbe ROC ou courbe Receiver Operating Characteristic et le diagramme de fiabilité, ainsi qu'en étudiant l'apport du GPNN pour prédire un événement extrême comme l'Halloween storm de 2003.

2.2. Analyse de la prédiction probabiliste fournie par le GPNN

Un processus gaussien a pour but de fournir non seulement une valeur moyenne, assimilable à une prédiction « single point », mais aussi une incertitude associée. Des mesures comme les CC et RMSE définies pour une prédiction « single point » ne sont pas adaptées pour évaluer une prévision probabiliste.

Nous avons alors fait appel à de nouvelles mesures, la courbe Receiver Operating Characteristic ou ROC, et le diagramme de fiabilité pour analyser les prédictions fournies par le GPNN.

2.2.1. Les courbes ROC pour évaluer les performances du GPNN en fonction de seuils d'activité

Dans un premier temps, nous avons étudié la fiabilité du GPNN au moyen des courbes ROC. Ces courbes font appel à deux métriques, le True Positive Rate (TPR) et le False Positive Rate (FPR). Ces métriques sont extraites d'une matrice de confusion. Les matrices de confusion sont définies pour analyser une catégorie par rapport à une autre. Nous décrivons plus en détail le mode opératoire pour obtenir les TPR et FPR dans le Chapitre 2 section 3.3.3.

Dans notre cas nous définissons trois catégories, spécifiques à trois « types » d'orages, et nous analysons l'évolution des TPR et FPR d'une catégorie par rapport aux deux autres. C'est le principe du « one versus all the others » décrit par [Camporeale et al., 2016] pour classer les vent solaires entre ejecta trou coronal, « sector reversal » et « streamer belt ». Nous faisons le choix de distinguer ces orages car en fonction de leur intensité, l'impact sur les technologies humaines n'est pas le même.

La Figure 74 présente une classification des orages magnétiques en fonction des valeurs de *Dst* fournies par le site AER²

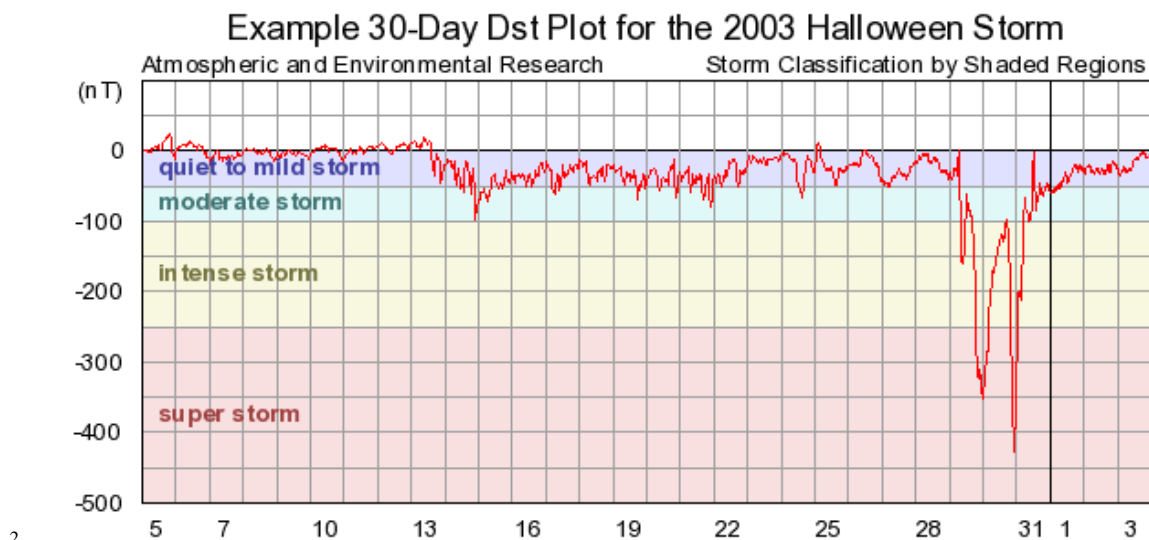


Figure 74- Classification des orages magnétiques - site AER.

Nous nous basons sur cette classification pour définir trois catégories d'orages. Comme le présente le Tableau 14, on définit des orages modérés pour des valeurs de *Dst* supérieures à -50 nT, des orages intenses pour des valeurs de *Dst* comprises entre -50 et -250 nT, et des supers orages pour des *Dst* inférieurs à -250 nT.

² <https://www.aer.com/science-research/space/space-weather/space-weather-index/>

Tableau 14- Classification des orages en fonction du niveau d'activité.

Niveau d'activité	Classification de l'orage
$Dst > -50 \text{ nT}$	Modéré
$-250 \text{ nT} < Dst < -50 \text{ nT}$	Intense
$Dst < -250 \text{ nT}$	Super orage

Le modèle GPNN fournit à un opérateur une prévision probabiliste, qui peut être utilisée pour prendre une décision en fonction de l'événement impactant l'environnement magnétique terrestre. Par exemple, un opérateur peut décider d'éteindre un système en fonction du niveau de sévérité de l'orage, quand la probabilité que cet orage arrive excède un seuil de déclenchement prédéfini. Ainsi, pour chaque orage, on peut construire une courbe ROC, qui est le tracé du FPR en fonction du TPR, pour différents seuils donnés.

Pour définir la notion de seuil, nous faisons appel à un exemple utilisé dans le domaine de la médecine. Les courbes ROC sont souvent utilisés dans ce domaine, afin d'effectuer des analyses sur les résultats d'un patient. Les médecins effectuent des tests diagnostiques quantitatifs pour analyser une caractéristique observée, par exemple un taux d'hormone dans le sang. La précision du test dépend alors de la valeur seuil choisie (arbitrairement) pour distinguer les personnes malades des personnes saines. La Figure 75 représente la distribution des résultats possibles d'un test chez les personnes "non-malades" (en vert) et chez les personnes "malades" (en rouge).

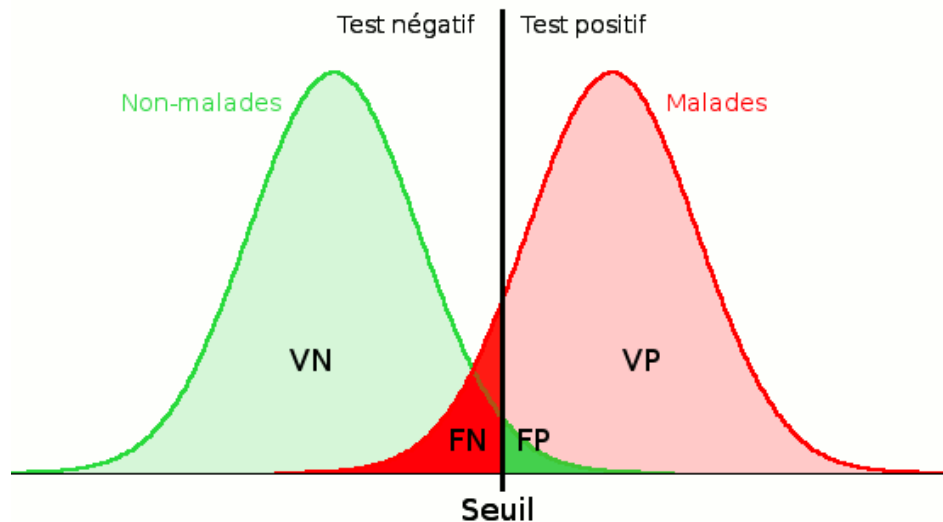


Figure 75 - Illustration de la notion de seuil © P. Calmant et E. Depiereux

On peut alors définir une valeur seuil, à partir de laquelle un médecin pourra définir qu'une patiente est malade ou saine. Les personnes présentant un test supérieur à une valeur seuil sont considérées comme positives au test, et donc supposées malades, tandis que celles présentant un résultat inférieur au seuil sont considérées comme négatives, et donc supposées non-malades. On constate que certaines

personnes sont considérées comme malades (car positives au test) alors qu'elles ne le sont pas en réalité; ce sont les faux positifs (FP, en vert foncé sur la Figure 75). De même, certaines personnes sont considérées comme non-malades (car négatives au test) alors qu'elles sont malades; ce sont les faux négatifs (FN, en rouge foncé sur la Figure 75). Lorsque la valeur seuil choisie change, les nombres de vrais positifs, vrais négatifs, faux positifs et faux négatifs s'en trouvent modifiés, modifiant par conséquent les valeurs de sensibilité et de spécificité de ce test. Ainsi, pour chaque valeur seuil, il est possible de déterminer les valeurs de sensibilité et de spécificité correspondantes. La relation entre la sensibilité et la spécificité du test, ou TPR et FPR pour chacune des valeurs seuils possibles, peut être représentée sous forme d'un graphique: la courbe ROC.

Dans notre cas, au lieu d'un cas malade/non malade à étudier, nous considérons trois niveaux d'activité à distinguer. Nous allons évaluer les TPR et FPR pour différents seuils et dans ce cas ces seuils permettant d'évaluer la probabilité qu'un événement appartenant à ce domaine d'activité ait lieu ou non.

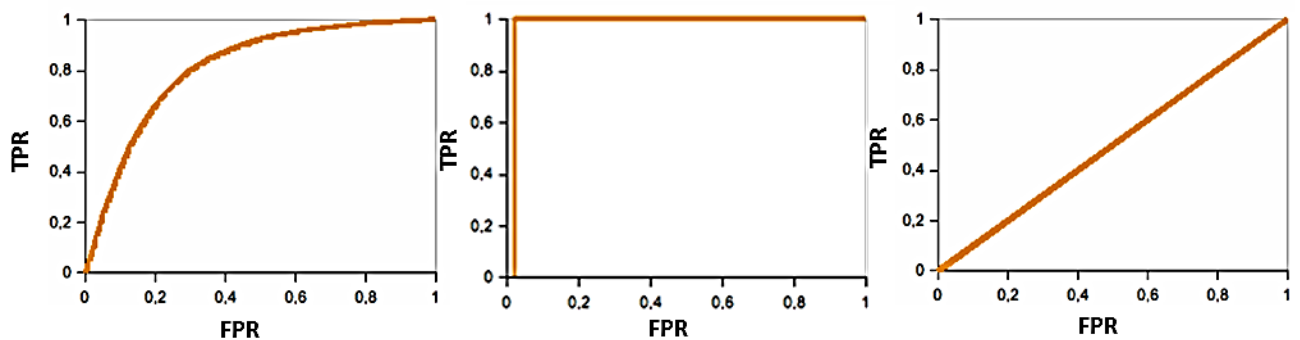


Figure 76 - Courbes ROC associées à une matrice de confusion. Chaque point définissant les courbes correspond à des couples TPR/FPR calculés pour différents seuils. La courbe de gauche représente une variation de TPR/FPR en fonction des valeurs seuils considérées. Celle du milieu représente le cas parfait pour lequel une discrimination totale est possible avec un FPR de 0 et un TPR de 1. La courbe de droite représente un cas pour lequel aucune discrimination n'est possible, le résultat est dû au hasard. © P. Calmant et E. Depiereux

La Figure 76 illustre des courbes ROC, montrant avec la figure du milieu qu'une classification parfaite correspond à un FPR de 0 et un TPR de 1. Ainsi, la valeur du seuil qui produit le point le plus proche de ces valeurs correspond au seuil optimal.

Nous avons donc défini les courbes ROC associées aux prédictions fournies par le GPNN. Afin de donner une représentation graphique, la Figure 77 présente les courbes ROC obtenues pour une prédiction à une heure. Mais pour des questions pratiques et pour faciliter l'analyse nous travaillons à partir de la Figure 78 qui présente les résultats obtenus pour chaque temps de prédiction et chaque catégorie d'orages, en fonction du seuil considéré. Le seuil optimal est en rouge traitillé et est calculé pour minimiser la distance Euclidienne entre le point de calcul et le couple (FPR=0, TPR=1). Si le lecteur souhaite avoir une visualisation graphique des résultats, les courbes ROC obtenues pour les prédictions jusqu'à six heures sont présentées en annexe 4.

Pour l'analyse du FPR, on constate qu'à tout niveau d'activité et tout temps de prédiction, on obtient des valeurs très faibles pour chaque seuil (la valeur la plus élevée étant égale à $2.7 \cdot 10^{-3}$ pour un seuil

de 10% en considérant un temps de prédiction d'une heure). Les variations du TPR sont plus complexes à généraliser. Dans un premier temps, on se focalise sur les valeurs obtenues pour le niveau d'activité le plus élevé, c'est-à-dire pour des valeurs de Dst inférieures à -250 nT. Pour des prédictions faites de une heure à cinq heures, les valeurs sont toujours supérieures à 0.719 pour des seuils de 10 à 40%, puis il y a une baisse des valeurs. Si on se focalise ensuite sur les prédictions faites à six heures, le meilleur TPR vaut 0.5 pour un seuil de 10 %. Cela signifie que plus il y a de risques qu'un super orage se produise, moins le modèle GPNN est capable de le prédire sans le sous-estimer six heures en avance. Cependant, pour les cas d'orages intenses ($-250 \text{ nT} < Dst < -50 \text{ nT}$), le GPNN fournit des TPR supérieurs à 0.670 pour des seuils compris entre 10% et 80%, et pour des orages modérés, ce modèle présente un TPR supérieure à 0.649 pour chaque seuil de une heure à six heures.

Ainsi, nous constatons en étudiant les courbes ROC que le GPNN trouve des limites pour prédire des supers orages à six heures. Ceci est probablement lié au fait, comme nous l'avions mentionné précédemment, que la perturbation ne peut être anticipée six heures en avance avec des mesures prises au niveau de l'onde de choc, comme c'est le cas avec les données OMNI. L'intérêt du processus gaussien trouve alors sa place pour analyser ce processus, car à mesure que la prédiction est éloignée dans le temps, l'erreur augmente. Grâce à la distribution sur la prédiction fournie par le GPNN, l'opérateur est capable d'évaluer la fiabilité de cette prédiction, et de l'analyser avec une courbe ROC.

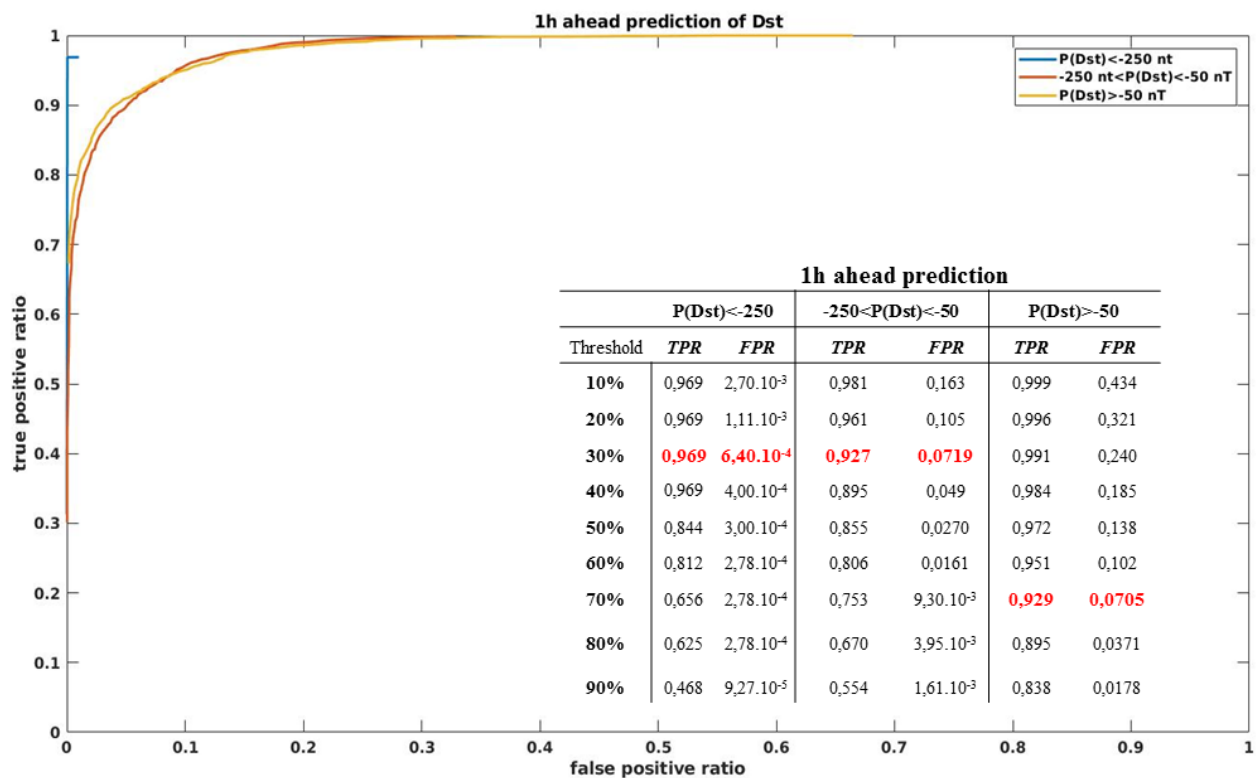


Figure 77 - Courbes ROC pour les prédictions à une heure obtenues avec le GPNN.

1h ahead prediction							2h ahead prediction						
Threshold	P(Dst)<-250		-250<P(Dst)<-50		P(Dst)>-50		Threshold	P(Dst)<-250		-250<P(Dst)<-50		P(Dst)>-50	
	TPR	FPR	TPR	FPR	TPR	FPR		TPR	FPR	TPR	FPR	TPR	FPR
10%	0,969	2,70.10 ⁻³	0,981	0,163	0,999	0,434	10%	0,969	3,15.10 ⁻³	0,963	0,199	0,999	0,388
20%	0,969	1,11.10 ⁻³	0,961	0,105	0,996	0,321	20%	0,937	9,27.10 ⁻⁴	0,934	0,142	0,984	0,273
30%	0,969	6,40.10⁻⁴	0,927	0,0719	0,991	0,240	30%	0,937	3,71.10⁻⁴	0,914	0,105	0,973	0,211
40%	0,969	4,00.10 ⁻⁴	0,895	0,049	0,984	0,185	40%	0,906	1,85.10 ⁻⁴	0,891	0,0834	0,961	0,167
50%	0,844	3,00.10 ⁻⁴	0,855	0,0270	0,972	0,138	50%	0,781	1,85.10 ⁻⁴	0,863	0,0565	0,943	0,134
60%	0,812	2,78.10 ⁻⁴	0,806	0,0161	0,951	0,102	60%	0,6875	9,27.10 ⁻⁵	0,824	0,0390	0,917	0,107
70%	0,656	2,78.10 ⁻⁴	0,753	9,30.10 ⁻³	0,929	0,0705	70%	0,656	9,27.10 ⁻⁵	0,783	0,0268	0,895	0,0845
80%	0,625	2,78.10 ⁻⁴	0,670	3,95.10 ⁻³	0,895	0,0371	80%	0,500	9,27.10 ⁻⁵	0,720	0,0156	0,858	0,0646
90%	0,468	9,27.10 ⁻⁵	0,554	1,61.10 ⁻³	0,838	0,0178	90%	0,437	0	0,601	5,6810 ⁻³	0,802	0,0363

3h ahead prediction							4h ahead prediction						
Threshold	P(Dst)<-250		-250<P(Dst)<-50		P(Dst)>-50		Threshold	P(Dst)<-250		-250<P(Dst)<-50		P(Dst)>-50	
	TPR	FPR	TPR	FPR	TPR	FPR		TPR	FPR	TPR	FPR	TPR	FPR
10%	0,875	3,24.10 ⁻³	0,958	0,254	0,984	0,373	10%	0,906	3,24.10 ⁻³	0,968	0,311	0,970	0,339
20%	0,843	9,27.10⁻⁴	0,939	0,186	0,971	0,278	20%	0,875	1,29.10⁻³	0,953	0,252	0,949	0,243
30%	0,813	4,64.10 ⁻⁴	0,912	0,139	0,955	0,228	30%	0,813	7,42.10 ⁻⁴	0,933	0,208	0,931	0,192
40%	0,750	1,86.10 ⁻⁴	0,890	0,106	0,940	0,182	40%	0,813	6,49.10 ⁻⁴	0,916	0,169	0,906	0,144
50%	0,625	9,27.10 ⁻⁵	0,880	0,0819	0,919	0,146	50%	0,781	9,27.10 ⁻⁵	0,895	0,138	0,874	0,104
60%	0,593	0	0,809	0,0606	0,893	0,1058	60%	0,687	9,27.10 ⁻⁵	0,843	0,106	0,841	0,0803
70%	0,593	0	0,766	0,0451	0,826	0,0865	70%	0,562	9,27.10 ⁻⁵	0,795	0,0812	0,802	0,0636
80%	0,437	0	0,714	0,0291	0,814	0,0594	80%	0,468	9,27.10 ⁻⁵	0,742	0,0621	0,76	0,0449
90%	0,406	0	0,614	0,0164	0,747	0,0413	90%	0,437	9,27.10 ⁻⁵	0,640	0,0403	0,699	0,0300

5h ahead prediction							6h ahead prediction						
Threshold	P(Dst)<-250		-250<P(Dst)<-50		P(Dst)>-50		Threshold	P(Dst)<-250		-250<P(Dst)<-50		P(Dst)>-50	
	TPR	FPR	TPR	FPR	TPR	FPR		TPR	FPR	TPR	FPR	TPR	FPR
10%	0,812	3,06.10 ⁻³	0,956	0,316	0,962	0,346	10%	0,500	8,34.10⁻³	0,953	0,352	0,932	0,307
20%	0,812	1,02.10⁻³	0,934	0,246	0,945	0,265	20%	0,437	4,92.10 ⁻³	0,928	0,289	0,909	0,241
30%	0,750	4,63.10 ⁻⁴	0,917	0,189	0,926	0,215	30%	0,437	3,24.10 ⁻³	0,904	0,244	0,886	0,186
40%	0,719	9,27.10 ⁻⁵	0,891	0,148	0,906	0,171	40%	0,406	2,78.10 ⁻³	0,890	0,202	0,862	0,161
50%	0,625	9,27.10 ⁻⁵	0,856	0,120	0,881	0,139	50%	0,375	1,76.10 ⁻³	0,859	0,167	0,834	0,130
60%	0,562	9,27.10 ⁻⁵	0,824	0,0942	0,853	0,107	60%	0,375	1,39.10 ⁻³	0,821	0,138	0,798	0,113
70%	0,468	0	0,779	0,0740	0,810	0,081	70%	0,281	7,47.10 ⁻⁴	0,788	0,115	0,757	0,0914
80%	0,468	0	0,725	0,055	0,754	0,0654	80%	0,281	3,70.10 ⁻⁴	0,735	0,0926	0,712	0,0693
90%	0,468	0	0,639	0,0381	0,685	0,0430	90%	0,281	2,78.10 ⁻⁴	0,661	0,0691	0,649	0,0455

Figure 78 - TPR et FPR associés à chaque domaine d'activité, pour différentes seuils, en fonction du temps de prédiction considéré.

2.2.2. Analyse des diagrammes de fiabilité

La courbe ROC présentée dans la section précédente permet d'indiquer à un opérateur la capacité du GPNN à détecter l'occurrence d'un événement géomagnétique pour un seuil donné, en termes de FPR et TPR. Les diagrammes de fiabilité permettent de mesurer la proximité entre la probabilité qu'un événement prédit ait lieu, avec la fréquence réelle de l'événement observé. Une méthode de prédiction parfaite fournit une probabilité p d'un événement égale en moyenne à la fréquence f à laquelle il est observé. Si la courbe de fiabilité est en dessous ou au-dessus de la diagonale parfaite, la prévision est

alors respectivement « under confident » ou « over confident », c'est-à-dire qu'elle présente des probabilités plus petites ou plus élevées que la fréquence réelle observée.

La Figure 79 présente les diagrammes de fiabilité obtenus pour des prédictions de l'indice magnétique *Dst* de une heure à six heures. La diagonale en pointillé rouge représente le cas parfait entre probabilité de prédiction et fréquence observée. A une heure, on constate que la prédiction sous-estime légèrement un orage, quand il y a plus de 35 % de probabilité pour une valeur donnée de *Dst*. Par exemple, quand il y a 80% de risque pour un orage prédit, la fréquence réelle observée pour cet orage est de 90%. Le GPNN fournit des prédictions fiables à deux heures, étant donné que la fréquence observée d'un orage définit une diagonale presque parfaite avec la probabilité de prédiction de cet orage. Pour des prédictions faites au-delà de trois heures, plus on va loin dans le temps, plus le GPNN surestime la probabilité d'un orage. Si on se focalise sur des prédictions faites à six heures, lorsque le GPNN fournit une probabilité de prédiction à 90%, la fréquence réelle observée de l'orage est de 65%. Ce diagramme peut être utilisé à l'avenir pour optimiser les valeurs du GPNN, c'est-à-dire en réajustant par exemple la dispersion obtenue en sortie (la valeur du σ).

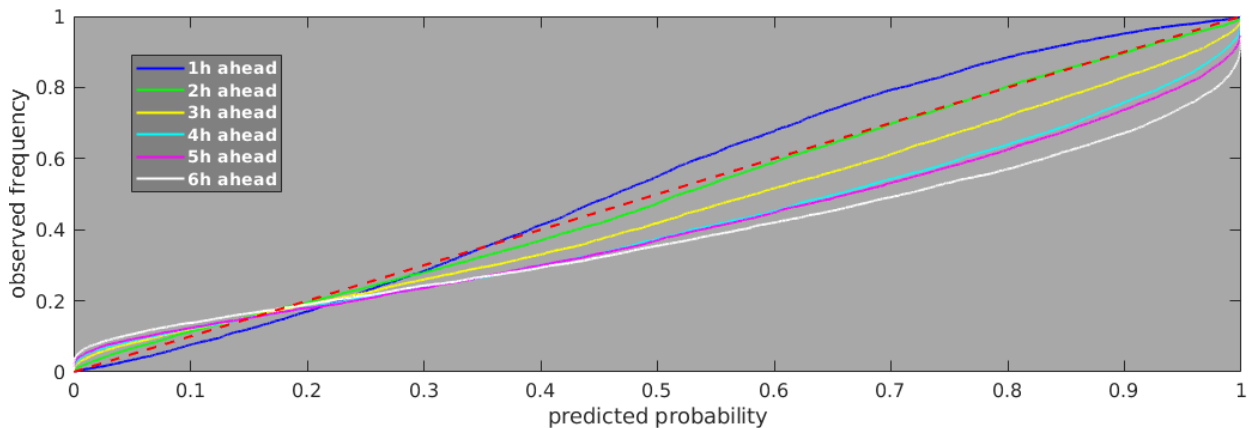


Figure 79 - Diagramme de fiabilité pour chaque temps de prédiction considéré. La diagonale en pointillé rouge représente une évolution parfaite de la fréquence observée en fonction de la probabilité de prédiction.

Ainsi, avec le diagramme de fiabilité et les courbes ROC, on peut évaluer les prédictions probabilistes fournies par le GPNN. Avec ces deux méthodes, nous constatons qu'à six heures, il est possible de fournir des prédictions de l'indice magnétique *Dst*, mais qu'il faut être prudent sur l'utilisation de la valeur considérée. Cette limite à six heures est sûrement donnée par le fait qu'au-delà d'un tel délai, même si l'algorithme est optimisé, il faut considérer les données directement au niveau du Soleil, et non au niveau de l'environnement proche terrestre.

2.2.3. Apport du GPNN pour la prédiction d'un événement extrême

Afin d'évaluer l'apport du GPNN par rapport à une prédiction « single point » (à un seul point) fournie par un réseau de neurones comme le LSTM que nous avons étudié à la section 2.1, nous analysons les prédictions fournies par le GPNN pour l'événement d'Halloween 2003. La Figure 80 présente les résultats du GPNN pour cet événement, il est alors possible de le comparer à la Figure 73 qui représente les résultats obtenus avec le LSTM. Pour des prédictions faites de une heure à cinq

heures, grâce au GPNN, la valeur prédite est proche de la valeur réelle. Par exemple, pour une prédiction faite à cinq heures, le pic d'activité prédit est à -391 nT, pour une valeur réelle de -422 nT. Si la valeur prédite par le GPNN est comparable à celle obtenue avec le LSTM, l'intervalle de confiance fourni par le GPNN contient la valeur maximale du pic d'activité et permet à un opérateur d'évaluer une limite haute du pic qui peut être atteinte. Cet apport est également visible lors de la prédiction à six heures, où on peut observer d'une part une amélioration de la prédiction de la valeur moyenne du pic d'activité et d'autre part, les barres d'erreur permettent d'englober la valeur maximale du pic. Cependant, la dispersion autour de la valeur moyenne est importante à six heures, ce qui montre encore les limites de ce modèle basé sur les mesures fournies par la base OMNI et les données GPS.

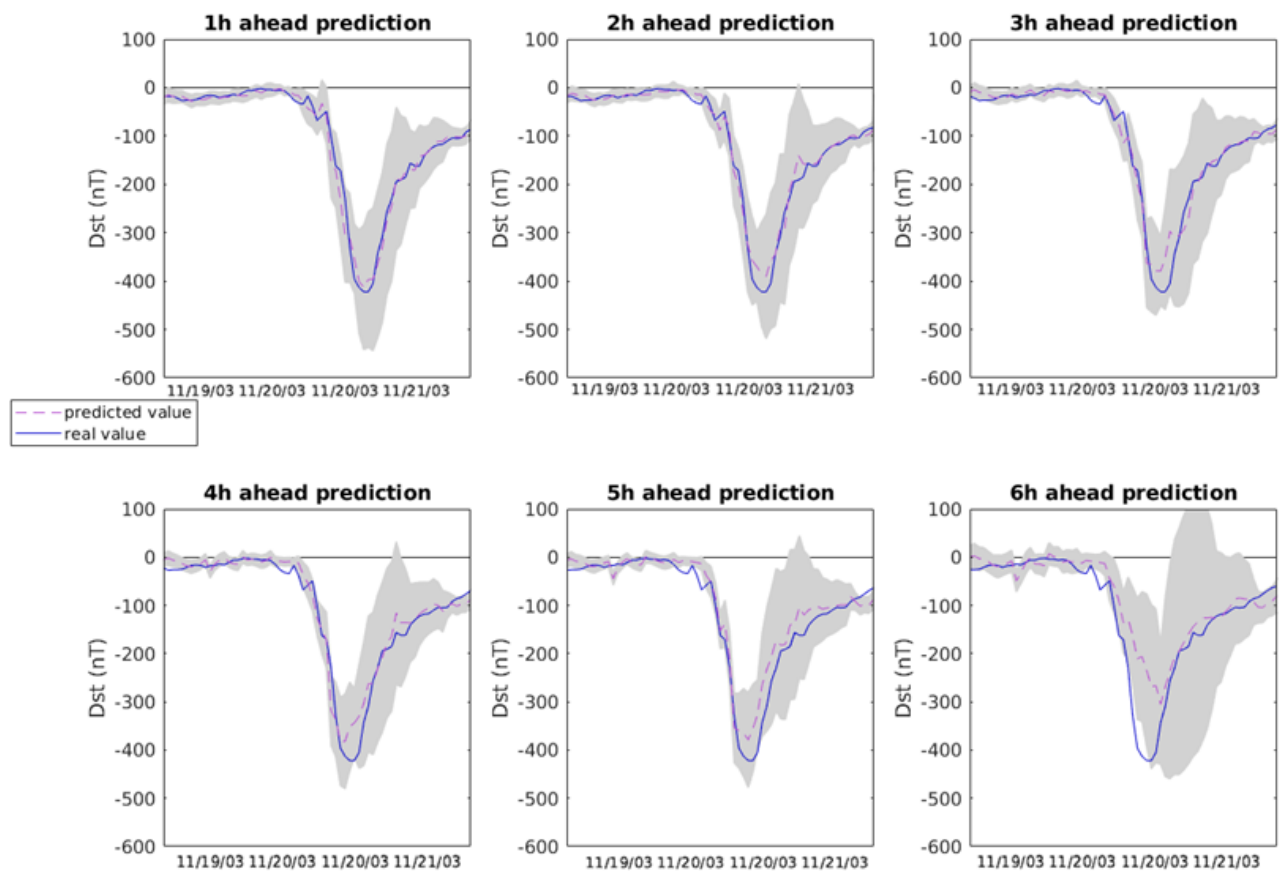


Figure 80- GPNN appliqué à la prédiction de l'orage d'Halloween 2003.

3. APPLICATION DE LA TECHNIQUE HYBRIDE ASSOCIANT RESEAUX DE NEURONES ET PROCESSUS GAUSSIENS POUR PREDIRE L'INDICE MAGNETIQUE AM

Après avoir développé et appliqué la technique du GPNN à la prédiction de l'indice magnétique *Dst*, nous avons souhaité étudier son application à la prédiction de l'indice magnétique *am*.

Nous avons donc dans un premier temps repris le modèle du LSTM que nous avons décrit au Chapitre 4 section 1, c'est-à-dire celui défini par l'équation (50). Ce modèle a donc pour entrées la densité *n*, la

vitesse fs et $IMF|B|$. Nous l'avons entraîné puis optimisé pour obtenir six réseaux LSTM pour la prédiction de l'indice magnétique am de une heure jusqu'à six heures.

$$\widehat{am}(t + p)_{NN} = \mathcal{NN}(n(t), fs(t), IMF|B|(t)) \quad (50)$$

Le même schéma que celui développé dans le cadre de la prédiction de l'indice magnétique Dst , représenté sur la Figure 67 est utilisé. Une fois les six réseaux LSTM optimisés, nous avons travaillé sur l'entraînement du GPNN pour la prédiction de l'indice magnétique am .

Cette dernière étape dans notre travail nous a montré la complexité de l'entraînement du GPNN qui est associée à la définition des hyperparamètres du processus gaussien. Pour définir des sous-ensembles d'entraînement, nous nous sommes basés sur la liste de « stream interaction region » développée par [Jian et al., 2006]. La Figure 81 présente des exemples de sous-ensembles d'entraînement utilisés pour cette étude.

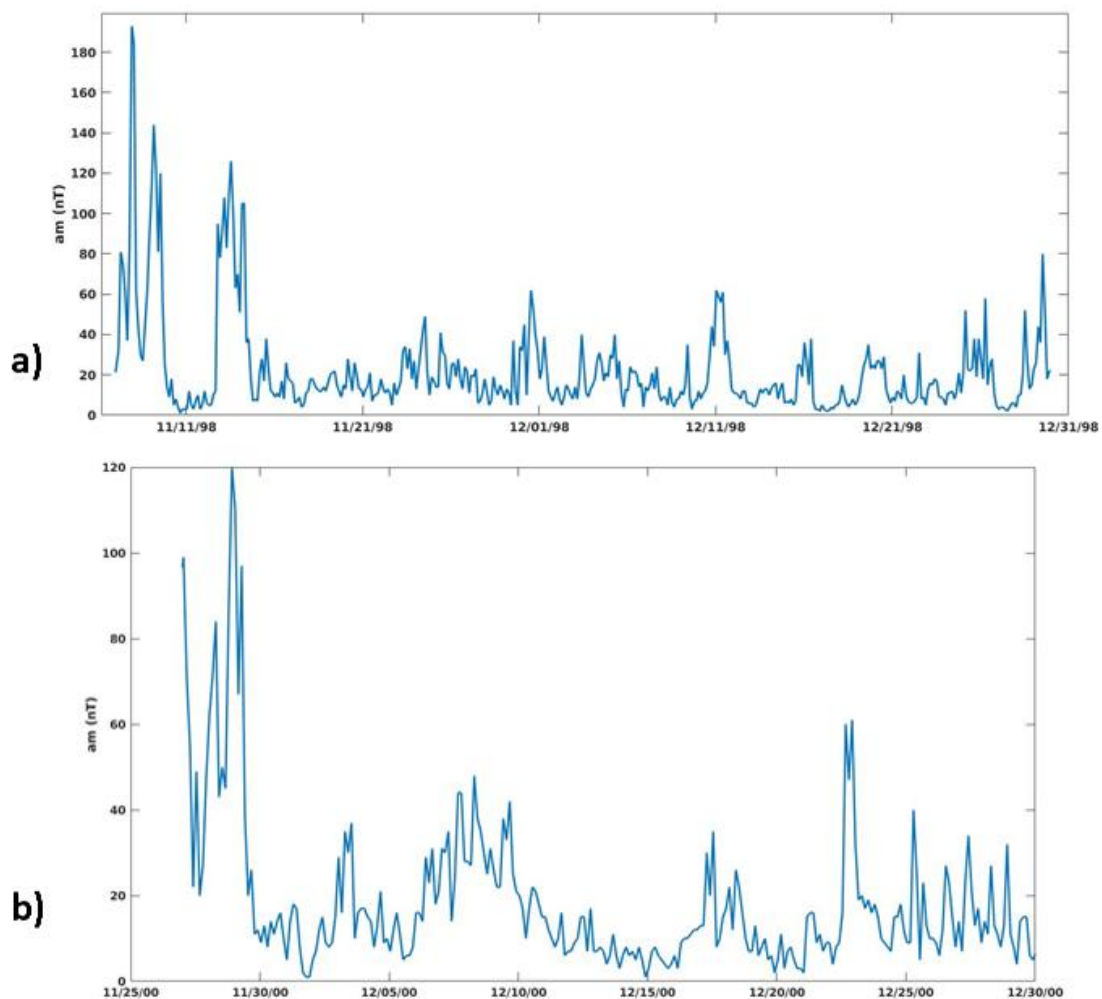


Figure 81 – Sous-ensembles d'entraînement du GPNN.

Nous présentons alors les résultats obtenus à partir de ces deux sous-ensembles d'entraînement, afin d'évaluer la capacité actuelle du GPNN à prédire l'indice magnétique am . D'autres sous-ensembles pourront être choisis par la suite, nous avons fait le choix de ceux-ci car ils présentent des variations importantes sur des échelles de temps variables. Une analyse plus complète sur les sous-ensembles seraient nécessaires pour optimiser les sous-ensembles d'entraînement.

La Figure 82 représente les courbes ROC associées aux sous-ensembles d'entraînement a) et b) définis sur la Figure 81. Afin d'effectuer une première évaluation au moyen des courbes ROC, les niveaux d'activité pour cette étude ont été restreints à trois niveaux et non à huit comme nous le faisons précédemment pour l'indice magnétique am . Ces niveaux ont été définis à partir de la table K_{pm}^3 pour effectuer une comparaison avec l'indice K_p en terme de sévérité. Nous avons alors défini un premier niveau pour des $am < 50$ nT, correspondant à un K_{pm} compris entre 0 et 4, un second niveau pour des am compris entre 50 et 150 nT, ce qui correspond à des K_{pm} de 4 et 6, et un dernier niveau pour des am supérieurs à 150 nT, donc des K_{pm} de 6 et plus. Nous constatons que pour les am associés à des activités faibles à intenses ($am < 150$ nT), les TPR et FPR sont proches, pour tous les seuils considérés. Le comportement présenté par les courbes a) et b) pour ces valeurs sont similaires. En revanche, pour des activités plus intenses ($am > 150$ nT), si les FPR restent faibles, dans les deux cas considérés, les TPR sont meilleures en utilisant le sous-ensemble d'entraînement a), plutôt que le sous-ensemble d'entraînement b). Par exemple, si on considère un risque à 70% d'événement très intense, le TPR obtenu avec le sous-ensemble a) vaut 0.681 tandis que celui obtenu avec le sous-ensemble b) vaut 0.595. Il est donc important de définir un ensemble d'événements pour entraîner le réseau qui optimise au mieux les hyperparamètres.

³ La table de conversion de am vers K_{pm} est disponible sur isgi.unistra.fr/Documents/ConversionTables/am2Kpm.pdf

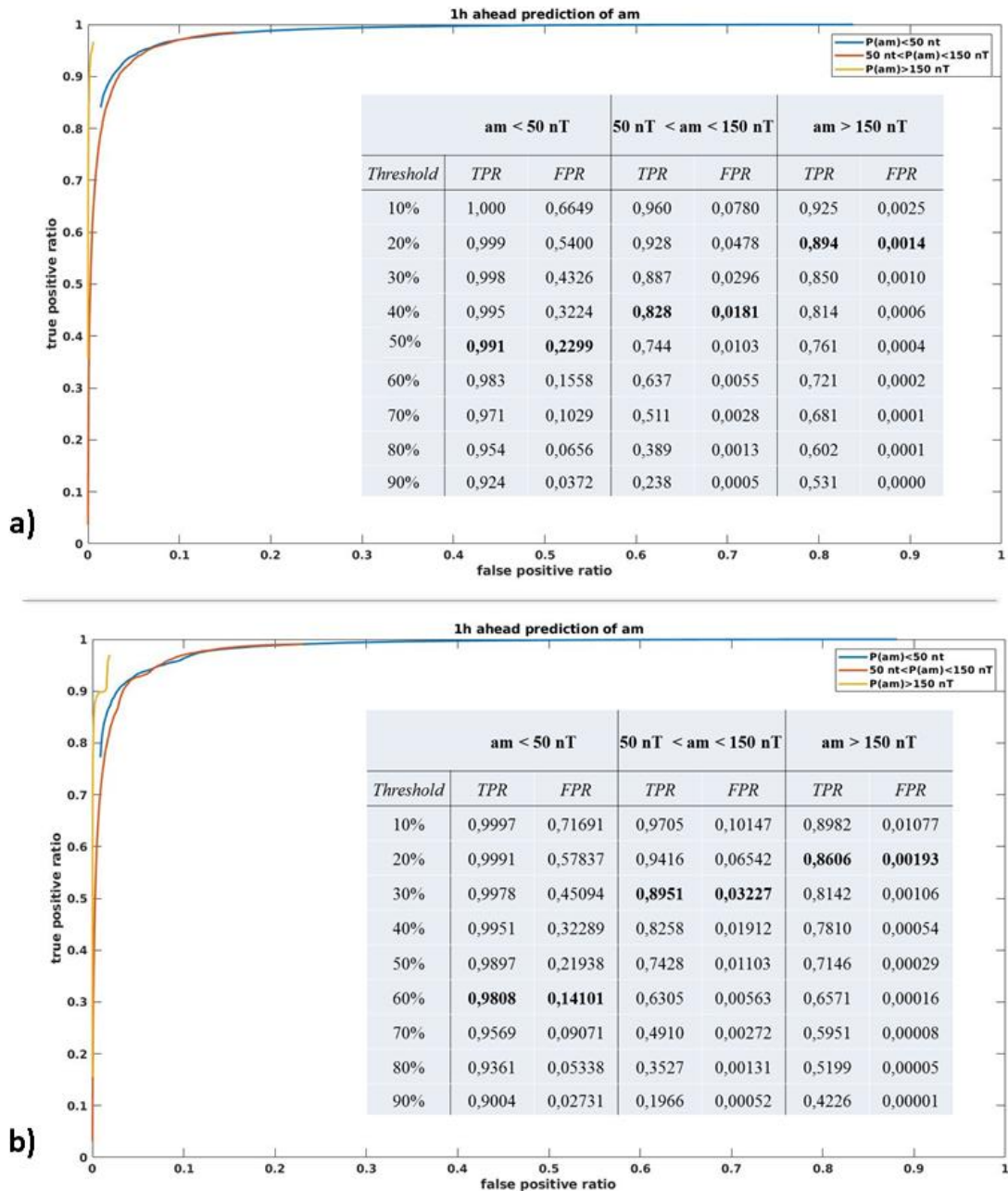


Figure 82-Courbes ROC associées aux sous-ensembles d'entraînement a) et b) de la Figure 81.

Cet effet est également observé sur la Figure 83 qui présente les diagrammes de fiabilité associés aux sous-ensembles d'entraînement a) et b). Dans les deux cas, pour une prédiction à une heure, le modèle a tendance à sous évaluer des événements qui ont de faibles probabilités d'occurrence, et surestimer des événements qui ont de fortes probabilités d'occurrence. Cela signifie qu'il est souhaitable de travailler sur la dispersion associée aux prédictions faites à une heure pour améliorer les performances à une heure. On constate cependant qu'en fonction du sous-ensemble d'entraînement, les prédictions faites de deux heures à six heures ne présentent pas les mêmes courbes d'évolution. En effet, le sous-ensemble d'entraînement a) aura moins tendance que le sous-ensemble d'entraînement b) à prédire plus d'événements qu'il n'y en a réellement. Si on se focalise sur le cas d'une prédiction à six heures, le diagramme de fiabilité présente une courbe proche de la diagonale lorsqu'on utilise le sous-ensemble d'entraînement a), tandis qu'avec le sous-ensemble d'entraînement b), on va avoir tendance à surévaluer la probabilité qu'un événement ait lieu.

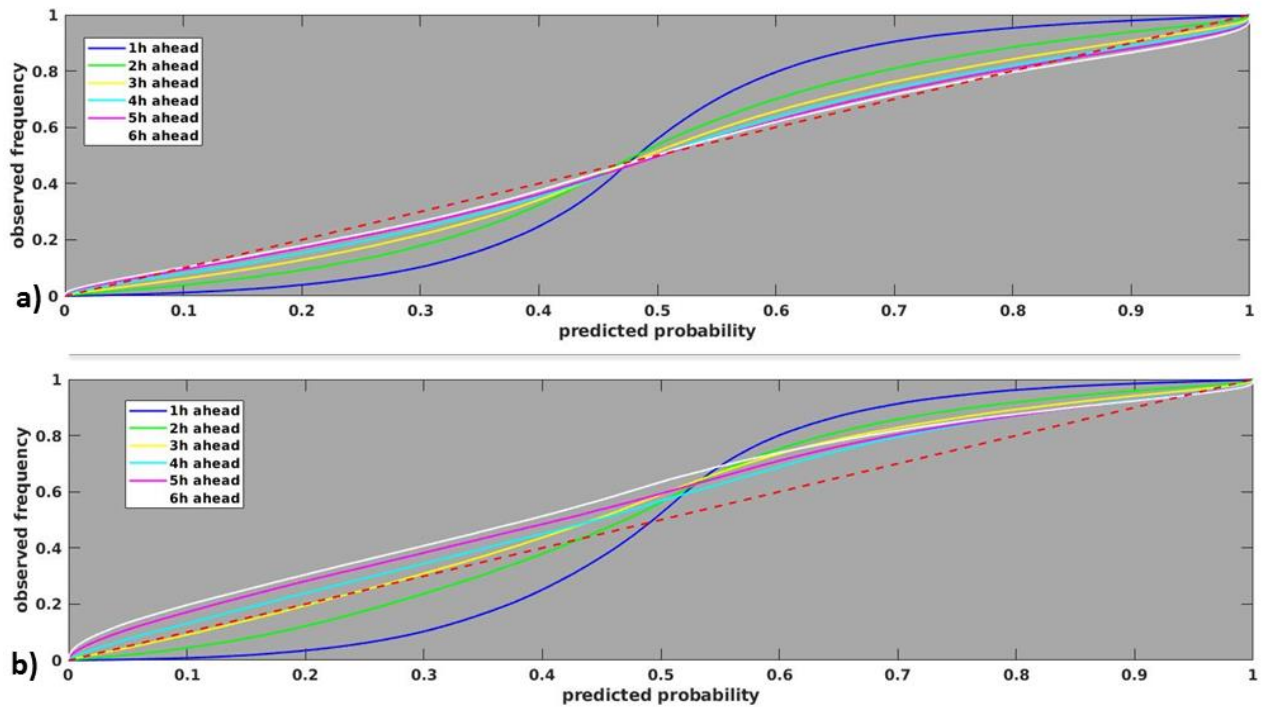


Figure 83- Diagrammes de fiabilité associés aux sous-ensembles d'entraînement a) et b) de la Figure 81.

Afin d'étudier l'importance du sous-ensemble d'entraînement sur un cas concret, nous avons considéré notre cas de validation pour l'étude de la prédiction de l'indice magnétique am , l'événement de Juillet 2004. La Figure 84 présente les résultats obtenus pour prédire cet indice de une heure à six heures, à partir des paramètres du vent solaire. A tout temps de prédiction, la valeur moyenne prédite obtenue avec le sous-ensemble d'entraînement a) est plus proche de la valeur réelle que celle prédite avec le sous-ensemble d'entraînement b). En fonction du sous-ensemble d'entraînement considéré, les prédictions obtenues peuvent fortement varier. Par exemple, pour une prédiction faite à cinq heures, le pic d'activité de 350 nT est prédit à 250 nT avec le sous-ensemble a) et à 175 nT avec le sous-ensemble b). Mais ces valeurs moyennes restent éloignées de la valeur réelle, et même si la barre d'erreur permet globalement d'anticiper une valeur extrême, il est nécessaire d'optimiser davantage l'entraînement du GPNN pour la prédiction de l'indice magnétique am .

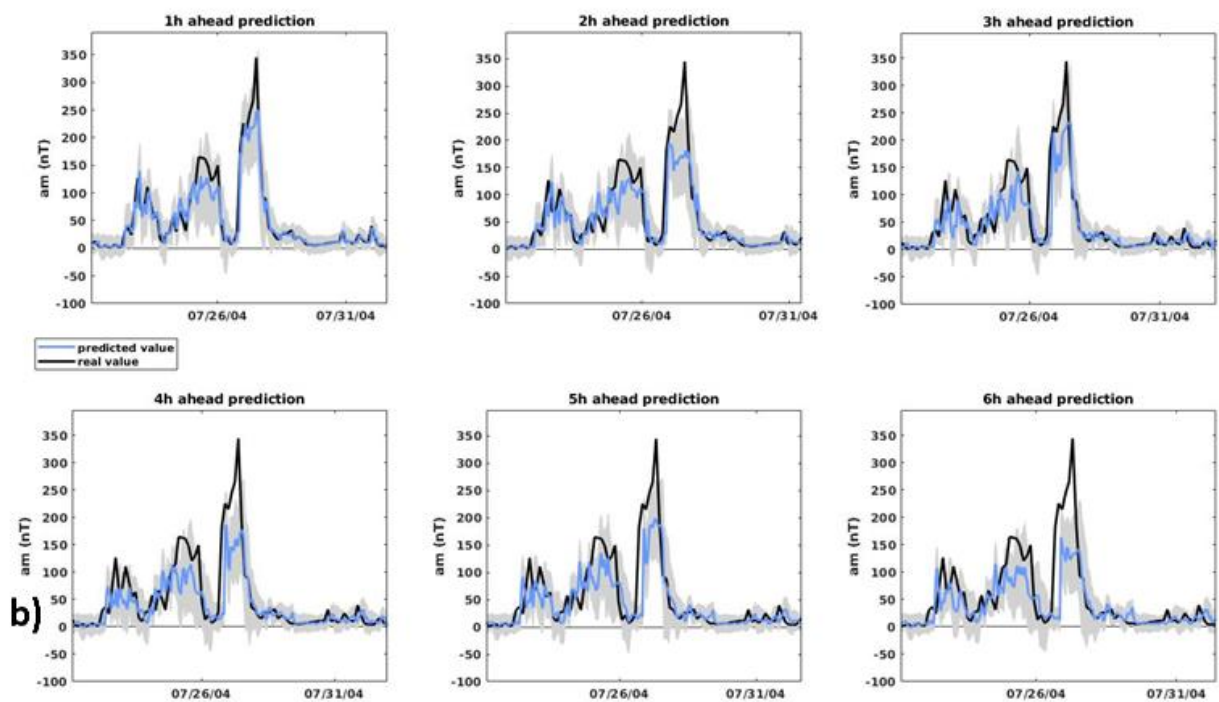
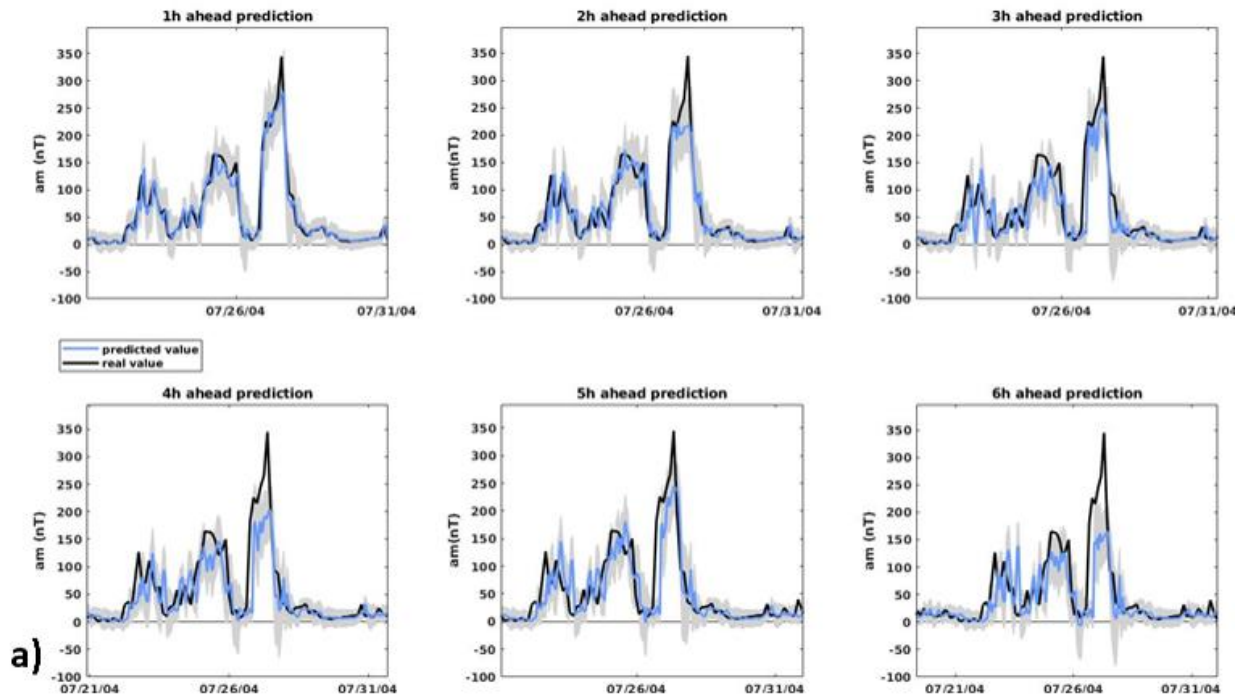


Figure 84 - Résultats de la prédiction de l'événement de Juillet 2004 obtenus à partir des sous-ensembles d'entraînement a) et b) de la Figure 13. La ligne noire représente la valeur réelle, la ligne bleue la valeur moyenne prédite et la partie grisée représente la dispersion à $\pm 2\sigma$ autour de la valeur moyenne.

4. BILAN SUR L'OPTIMISATION DES PREDICTIONS D'INDICES MAGNETIQUES A PLUS LONG TERMES AU MOYEN D'UNE NOUVELLE TECHNIQUE HYBRIDE

Dans ce chapitre sur l'optimisation des prédictions d'indices magnétiques jusqu'à six heures au moyen d'une nouvelle technique hybride, nous avons développé un processus associant réseaux de neurones et processus gaussiens appelé GPNN. Ceci a été fait afin de combiner les performances optimales de prédiction obtenues avec un réseau de neurones, et l'aspect probabiliste des processus gaussiens. Les processus gaussiens ont la possibilité de fournir une dispersion autour d'une moyenne, correspondant à un intervalle de confiance qu'un opérateur peut utiliser pour analyser le domaine d'activité dans lequel une perturbation magnétique se situera. Pour définir ce processus gaussien, les prévisions de une heure à six heures d'indices magnétiques fournies par le réseau de neurones sont utilisées en entrée du processus gaussien. Elles définissent la moyenne du processus gaussien.

Pour évaluer ces prévisions probabilistes, nous avons utilisé deux métriques basées sur l'analyse d'un événement au sens «one versus all the others», c'est-à-dire étudier l'appartenance de l'événement prédit à un domaine d'activité spécifique. Ceci a été fait afin de définir d'une part des seuils optimaux de déclenchement d'alerte grâce aux courbes ROC, et d'autre part la capacité de ces réseaux à fournir des prédictions en accord avec la fréquence réelle d'occurrence d'un événement. Ces techniques ont été utilisées pour fournir des prédictions d'indice magnétique de une heure à six heures.

En ce qui concerne l'application de cette technique à la prédiction de l'indice magnétique *Dst*, nous avons constaté que le GPNN permet de fournir une valeur moyenne prédite de *Dst* plus proche de la valeur réelle de l'indice que celle obtenue avec le réseau LSTM, notamment pour des prédictions faites à six heures. L'apport principal de cette technique réside dans l'estimation d'erreurs, qui permettent d'évaluer le risque maximal associé à un événement à six heures. Cependant, la dispersion augmente avec le temps de prédiction. Pour que cette prédiction soit utilisable par un opérateur, il faudrait donc travailler sur l'optimisation du GPNN pour des temps de prédiction éloignés, ou considérer d'autres types de données en entrée. En effet, avec l'étude du LSTM et du GPNN pour des prédictions de une heure à six heures, nous avons constaté qu'au-delà de six heures, les deux modèles ont des difficultés à établir une connexion entre les informations en provenance du vent solaire mesurées au niveau de l'onde de choc et au niveau des GPS, à la perturbation magnétique mesurée au sol. Il faudrait donc étudier l'apport que cela aurait de considérer des données d'observations solaires, et non des données prises au niveau de l'environnement magnétique terrestre, afin de gagner une réelle avance sur le temps de prédiction des effets associés à un événement solaire.

Enfin, nous avons appliqué la technique du GPNN à la prédiction de l'indice magnétique *am*. Le travail le plus important à fournir lorsqu'on développe un processus gaussien est de trouver une base d'entraînement des hyperparamètres qui soit optimale pour anticiper au mieux un événement à venir. En effet, le processus gaussien est basé sur l'inférence bayésienne, c'est-à-dire que la prédiction d'un événement futur est calculée à partir des connaissances a priori. Pour permettre d'obtenir une prédiction à six heures de l'indice magnétique *am* à partir du GPNN en se basant uniquement sur les paramètres du vent solaire, il sera donc nécessaire de continuer à travailler sur des sous-ensembles d'entraînement, en considérant des combinaisons d'événements, ou en étudiant une liste d'événements différente de celle de [Jian et al., 2006]. Si le nombre de variables considérées en entrée devait augmenter, le nombre d'hyperparamètres serait également plus important, la taille des hyperparamètres θ étant directement lié au nombre de variables. Il faudrait donc utiliser des processus gaussiens adaptés à un grand nombre de paramètres en entrée [Bouhleb et al. 2016].

CONCLUSION ET PERSPECTIVES

Cette thèse a eu pour but de développer des modèles de prédiction de l'indice magnétique am de type réseaux de neurones, en utilisant uniquement les paramètres du vent solaire enregistrés par un satellite d'observation solaire. Ceci est fait dans le but d'anticiper en temps réel l'impact de l'activité solaire sur l'environnement magnétique terrestre. Ces modèles de prédiction doivent répondre à des besoins opérationnels, c'est-à-dire fournir une prévision à plus ou moins long terme avec une mesure de fiabilité sur laquelle l'opérateur pourra se baser pour prendre une décision appropriée. Tout au long de ce travail, nous avons oscillé entre raisonnements physiques et modèles mathématiques. La physique de l'interaction Soleil-Terre est complexe, et si chaque jour de nouvelles réponses apparaissent, la physique à elle seule ne permet pas d'anticiper la réponse de la magnétosphère à un événement solaire extrême. Nous avons donc développé des modèles appartenant au domaine de l'intelligence artificielle afin de connecter d'une part les paramètres du vent solaire mesurés par des satellites en amont de la magnétosphère, et d'autre part les indices magnétiques mesurés au sol, plus spécifiquement l'indice magnétique global am . Nous avons démontré qu'il était possible de fournir des prédictions de l'indice magnétique am uniquement à partir des paramètres du vent solaire, mesurés par le satellite ACE au point de Lagrange L1. Cette prédiction est faite à court terme, c'est-à-dire à une heure, au moyen d'un réseau de neurones récurrent, le Long Short Term Memory ou LSTM. Pour obtenir des prédictions à plus long termes, nous avons travaillé dans le cadre d'une collaboration avec le CWI sur le développement d'un nouveau modèle combinant réseau de neurones et processus gaussiens pour obtenir des prédictions probabilistes jusqu'à six heures. Afin d'évaluer l'information fournie par les modèles de prédiction, nous avons fait appel à des métriques statistiques basées sur une matrice de confusion. Ces métriques sont la probabilité de détection (POD) et le taux de fausses alarmes (FAR), et ont été utilisées pour évaluer globalement les réseaux de neurones. Nous avons également fait appel aux courbes ROC et aux diagrammes de fiabilité pour évaluer le modèle probabiliste résultant de la combinaison d'un réseau de neurones et d'un processus gaussien. Grâce à ces analyses, un opérateur est capable d'analyser plus en détail la prédiction fournie par un modèle et d'adapter son jugement. Nous revenons alors sur les réponses clefs que chacune des études explicitées dans les différents chapitres a apportée :

Est-il possible de prédire l'indice magnétique am uniquement à partir des paramètres du vent solaire, au moyen du Time Delay Neural Network ou réseau à retard de temps (TDNN) ?

Cette première analyse de la capacité des réseaux de neurones à prédire l'impact de l'activité solaire sur l'environnement magnétique terrestre a été basée sur le développement, l'optimisation et la comparaison de trois réseaux de neurones. Ces trois réseaux éprouvés par le passé en météorologie de l'espace pour prédire d'autres indices magnétiques sont le réseau perceptron multicouche (MLP) également appelé « feedforward backpropagation », le réseau temporel TDNN et le réseau récurrent NARX. Pour définir les paramètres d'entrée des réseaux de neurones, nous avons effectué une analyse du coefficient de Kendall. Cette méthode basée sur une analyse des rangs a souligné la corrélation de trois paramètres du vent solaire avec l'indice magnétique am : la vitesse fs , l' $IMF |B|$ et la densité n . Dans un premier temps, nous avons comparé les performances des trois réseaux de neurones avec les paramètres fournis par la base OMNI. Cette étude a mis en évidence les performances du NARX, supérieures à celles des réseaux « feedforward » et TDNN. Elle a également souligné la capacité du TDNN à fournir des prédictions de l'indice magnétique am , ce qui est un véritable défi étant donné que ce réseau est basé uniquement sur les paramètres du vent solaire. Dans un second temps, nous

avons considéré les données directement fournies par le satellite ACE situé au point de Lagrange L1. Cette étude a eu pour but d'étudier les effets que pouvait avoir l'utilisation de données provenant d'un satellite et non d'une base prétraitée comme celle d'OMNI sur les performances des réseaux de neurones. Le réseau NARX restait le réseau le plus performant, et le TDNN a montré sa capacité à fournir des prédictions optimales notamment à seuil d'activité élevé, en comparaison avec le réseau « feedforward ». Nous avons dans un premier temps comparé les performances globales des réseaux au moyen des POD et FAR, puis nous avons analysé un cas d'événement extrême, celui de Juillet 2004. Nous avons alors démontré la capacité du TDNN à prédire des variations rapides et complexes d'activité. Cependant, les performances de ce réseau n'atteignent pas celles du réseau NARX, qui est un réseau contenant une partie autorégressive dont nous souhaitons nous affranchir. En effet, dans un cadre opérationnel il est risqué de considérer un modèle prenant en entrée la valeur prédite par celui-ci, étant donné que la valeur définitive d'un indice magnétique n'est pas fixée avant un certain délai. Suite à cette première étude, nous en avons conclu que le compromis serait donc de trouver une structure qui aurait les performances d'un réseau récurrent sans liaison autorégressive. Nous avons donc développé un réseau qui n'avait pas été utilisé auparavant en météorologie de l'espace, le réseau récurrent Long Short Term Memory (LSTM). Les travaux résultant de cette étude ont été soumis au Space Weather Space Climate Journal (*Prediction of the geomagnetic index am based on the development and the performance comparisons of static and dynamic Neural Networks*, M. A. Gruet, N. Bartoli, S. Rochel, R. Benacquista, A. Sicard and G. Rolland).

Pouvons-nous améliorer les performances de prédiction de l'indice magnétique am en faisant appel à un nouveau réseau de neurones récurrent ayant une mémoire à court et long termes, LSTM (Long Short Term Memory) ?

Dans cette seconde étude, nous avons analysé la capacité du réseau LSTM à améliorer les performances des prédictions de l'indice am à partir des paramètres du vent solaire. Dans un premier temps, nous avons souligné sa capacité d'optimisation plus rapide que les réseaux classiques de type « feedforward ». Nous avons ensuite comparé les performances de ce réseau récurrent au réseau TDNN, et mis en évidence le fait que le LSTM fournit une réelle amélioration dans le cadre du développement d'un réseau de neurones opérationnel. Non seulement il est plus performant que le TDNN, mais en plus il n'est pas nécessaire d'optimiser une fenêtre temporelle de taille fixe comme c'est le cas avec le TDNN. Les informations arrivent en temps réel en entrée du réseau LSTM, et la temporalité est prise en compte par la structure en chaîne de celui-ci. Dans une logique d'amélioration de performances du LSTM, nous avons fait appel aux fonctions de couplage en entrée du réseau de neurones. Une fonction de couplage permet d'évaluer l'entrée en énergie dans la magnétosphère associée à un événement solaire. Cette entrée en énergie est la nouvelle entrée du réseau de neurones. Nous avons utilisée celle développée par [Wang et al., 2014] et constaté que l'utilisation de la fonction de couplage permet d'améliorer les performances des réseaux de neurones pour l'indice am . Pour aller toujours plus loin dans le développement de modèles répondant aux besoins d'un opérateur, nous avons travaillé sur la prédiction de l'indice am sectoriel ou $a\sigma$. Cet indice développé par [Chambodut et al., 2013] permet de fournir à un opérateur une mesure de la perturbation magnétique en fonction du Temps Magnétique Local, et de voir l'évolution d'une perturbation en fonction de ceux-ci. Nous avons alors développé deux types de réseaux, un réseau multisortie fournissant quatre prédictions associés à $\{a\sigma_{dawn}, a\sigma_{noon}, a\sigma_{dusk}, a\sigma_{midnight}\}$, et quatre réseaux monosortie spécifiques à chaque indice sectoriel. Grâce à cette étude, nous avons pu souligner les limites actuelles des réseaux multisortie et conclure qu'il était mieux adapté d'utiliser quatre réseaux monosortie, notamment pour les cas d'activité les plus extrêmes.

Est-il possible de fournir à plus long termes des prédictions probabilistes d'indices magnétiques afin de répondre plus spécifiquement aux besoins d'un opérateur ?

Dans le cadre d'une collaboration avec le CWI (laboratoire d'informatique et de mathématique d'Amsterdam) sous la direction d'Enrico Camporeale, nous avons développé un nouveau modèle combinant les performances optimales de prédiction obtenues avec un réseau de neurones, et l'aspect probabiliste des processus gaussiens. Ce modèle a été appelé GPNN pour Gaussian Process-Neural Network. Ceci a été fait afin de répondre au besoin d'un opérateur qui est d'avoir une mesure d'erreur sur la prédiction fournie par le modèle. Les processus gaussiens ont la possibilité de fournir une dispersion autour d'une moyenne, correspondant à une barre d'erreur. Pour définir ce processus gaussien, les prévisions de une heure à six heures d'indices magnétiques fournies par le réseau de neurones sont utilisées comme moyenne du processus gaussien. Nous avons dans un premier temps appliqué cette technique à la prédiction de l'indice magnétique *Dst* et nous avons constaté que le GPNN permet de fournir une valeur moyenne prédite de *Dst* plus proche de la valeur réelle de l'indice que celle obtenue uniquement avec le réseau LSTM. Dans un second temps, nous avons appliqué cette technique pour la prédiction de l'indice magnétique *am* jusqu'à six heures. Nous avons alors démontré qu'il est possible d'utiliser cette technique à condition d'optimiser la base d'entraînement des hyperparamètres du processus gaussien. Dans les deux cas, nous avons souligné qu'à partir de six heures, les modèles de prédictions peinent à établir une connexion entre paramètres du vent solaire et indices magnétiques. Les travaux résultant de l'étude de la prédiction de l'indice magnétique *Dst* ont été acceptés dans le journal *Space Weather (Multiple-Hour-Ahead forecast of the Dst index using a combination of Long Short-Term Memory Neural Network and Gaussian Process, M.A. Gruet, M. Chandorkar, A. Sicard, E. Camporeale)*

Suite à ces trois études, des réponses clefs concernant la prédiction d'indices magnétiques ont été apportées, mais aussi de nombreuses perspectives. Le domaine de la météorologie de l'espace est riche et ne demande qu'à faire appel à de nouvelles techniques, nouvelles données, nouvelles mesures. Nous discutons des principales perspectives issues de ces études.

Lorsque nous avons développé le réseau multisortie pour la prédiction de l'indice magnétique sectoriel $\alpha\sigma$, nous avons constaté qu'à l'heure actuelle, les techniques existantes pour faire du multisortie ne sont pas suffisamment optimales en comparaison avec des modèles monosortie. Cependant, les réseaux multisortie sont bien plus en accord avec la physique de la magnétosphère que les réseaux monosortie, car ils prennent en compte la dépendance entre les sorties. Dans le cadre du développement du GPNN, nous avons dû développer six réseaux spécifiques à chaque temps de prédiction, de une heure à six heures, alors qu'il semble évident qu'un réseau multisortie prenant en compte les dépendances entre les indices magnétiques seraient bien plus adaptés. Mais il faudrait alors soit développer une technique de régression logistique multinomiale, méthode qui généralise les techniques de régression à des problèmes multiclassés, c'est-à-dire avec un nombre multiple de valeurs discrètes, soit transformer le problème actuel de prédiction fait à partir d'un modèle de régression, en problème de classification.

La question de la classification est régulièrement revenue durant cette thèse. Nous sommes partis du principe que nous étions face à un problème de régression, c'est-à-dire que pour déduire la valeur de l'instant suivant, nous nous basions sur les connaissances acquises aux instants précédents. La question de la classification permettrait sûrement de simplifier le modèle, en définissant des classes correspondant à des niveaux d'activité, et d'entraîner les modèles de prédiction à reconnaître des

classes d'activité et non des valeurs d'indices magnétiques. Cependant, il faut que cette méthode soit en accord avec le besoin d'un opérateur. Si dans nos méthodes d'optimisation et de comparaison des performances des modèles de prédictions, nous avons distingué des niveaux d'activité, en revanche nos modèles ont toujours eu pour but d'être au plus près de la valeur réelle à prédire.

Dans le cadre du développement du GPNN, nous avons mis en évidence l'apport principal de cette technique qui réside dans l'estimation d'erreurs. Ceci permet d'évaluer le risque maximal associé à un événement jusqu'à six heures dans le futur. Cependant, la dispersion augmente avec le temps de prédiction, soulignant non pas un défaut dans l'algorithme du modèle, mais la nécessité de travailler avec des données différentes. En effet, grâce à cette étude sur le GPNN, nous avons constaté qu'au-delà de six heures, il est complexe d'établir une connexion entre les informations en provenance du vent solaire, mesurées au niveau de l'onde de choc et au niveau des GPS, à la perturbation magnétique mesurée au sol. Il faudrait donc étudier l'apport que cela aurait de considérer des données d'observations du vent solaire, et non des données prises au niveau de l'environnement magnétique terrestre, afin de gagner une réelle avance sur le temps de prédiction des effets associés à un événement solaire. Des scientifiques du CWI en collaboration avec l'INRIA travaillent à l'heure actuelle sur l'analyse de la capacité des modèles de type réseaux de neurones convolutionnels pour détecter un événement solaire à partir d'images provenant de SDO (Solar Dynamic Observatory). Ce travail représente au-delà d'un algorithme complexe, un temps de calcul conséquent. En effet, le traitement d'images pixel par pixel nécessite des ressources bien plus importantes que le traitement de séries temporelles. Mais en terme de gains de temps de prédictions, une découverte dans ce domaine pourrait être un grand tournant dans le domaine de la météorologie de l'espace. On pourrait alors utiliser les données extraites directement des observations solaires, et anticiper leurs impacts des heures voir des jours avant l'arrivée des particules observées dans l'environnement magnétique terrestre.

La mission Parker Solar Probe pourrait également fournir de nouveaux axes de réflexion dans le cadre de développement de modèles de prédiction. Cette mission décollera au courant de l'été 2018 et a pour but de s'approcher au plus près du Soleil afin de répondre à deux questions fondamentales en physique solaire. La première est liée à l'analyse du changement de température conséquent entre la couronne solaire et la surface de l'étoile. La seconde est quant à elle liée à l'analyse des vents solaires et porte sur l'analyse des mécanismes d'accélération des vents à des vitesses supersoniques. Comme le lecteur aura pu le constater tout au long de ce manuscrit, si les mathématiques sont au cœur du développement des modèles de prédictions, il est fondamental de revenir à la physique du problème pour améliorer les connexions entre vent solaire et indices magnétiques. Nous avons pu constater cet effet notamment en utilisant les fonctions de couplage pour prédire l'indice magnétique *am*. Parker Solar Probe apportera sûrement de nouvelles pistes de réflexions pour permettre de toujours mieux anticiper les effets du vent solaire sur l'environnement magnétique terrestre.

En attendant les résultats de la mission Parker Solar Probe, des scientifiques à travers le globe font des découvertes auxquelles il est important de prêter attention pour toujours optimiser les modèles comme le modèle EUHFORIA. Au début de ce manuscrit, nous avons présenté une illustration provenant de [Amari et al. 2018]. Les découvertes publiées au début de cette année 2018 par l'équipe de Tahar Amari ont montré qu'il existe un mécanisme de cage magnétique entourant une éruption, et que la probabilité finale de l'éruption dépend de la capacité de cette cage à maintenir une éruption. Des analyses futures permettraient alors sûrement d'évaluer le risque d'éruption basé sur ce modèle, afin d'en déduire le risque conséquent engendré au niveau de l'environnement magnétique terrestre.

ANNEXES

1. ANALYSE DE L'APPORT DES PARAMETRES V_x ET B_z POUR PREDIRE L'INDICE MAGNETIQUE AM AVEC LE TDNN

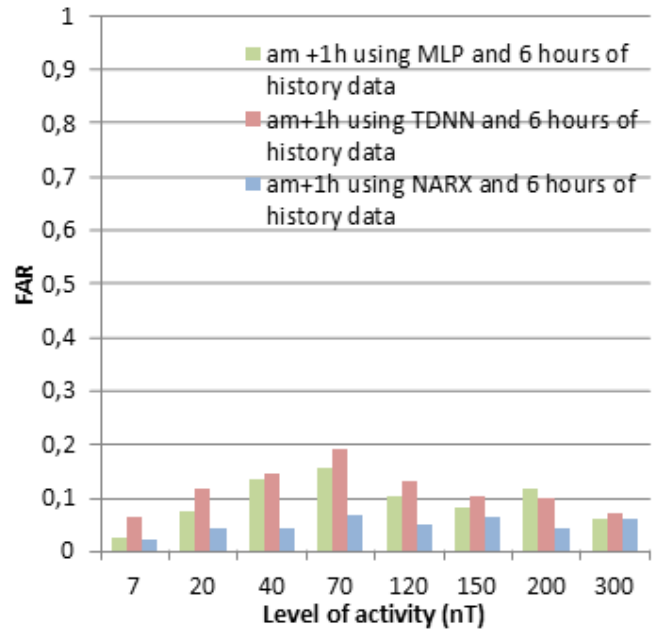
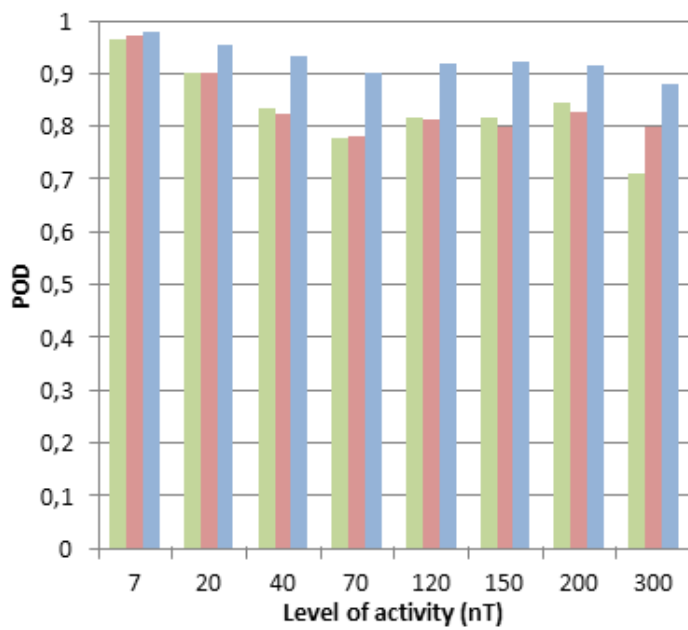
Dans les sections précédentes, nous avons étudié les performances des réseaux « feedforward », TDNN et NARX en considérant un trio de paramètres du vent solaire définis suite à l'analyse du coefficient de Kendall.

Cette analyse a mis en avant deux paramètres, la vitesse fs et l' $IMF |B|$, avec des coefficients de Kendall respectivement égaux à 0.346 et 0.371 dans le cas où les données manquantes sont interpolées. Ceci souligne l'existence d'une relation entre ces paramètres et l'indice magnétique am . Deux autres paramètres sont également ressortis, la composante V_x de la vitesse ainsi que la composante B_z . Nous avons alors fait le choix de ne pas considérer ces paramètres dans le but de minimiser le nombre d'entrées et avoir une meilleure analyse des résultats fournis par le réseau de neurones. En effet, plus le nombre de paramètres d'entrée est important, plus il est complexe d'étudier une relation entre les paramètres d'entrée et la sortie.

Nous avons alors souhaité analyser les performances du réseau de neurones en ajoutant V_x d'une part et B_z d'autre part aux trois paramètres initiaux $\{fs, IMF|B|, n\}$. Pour faire cette étude, nous avons considéré les données ACE qui représentent le meilleur intérêt pour un opérateur, avec un historique de temps de six heures.

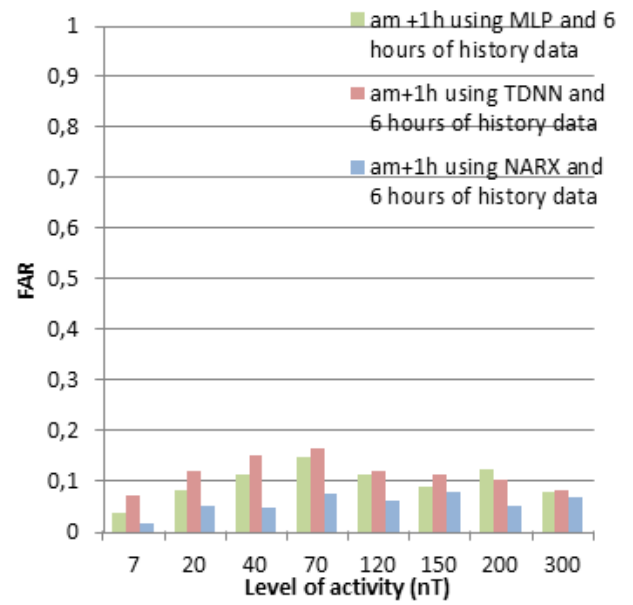
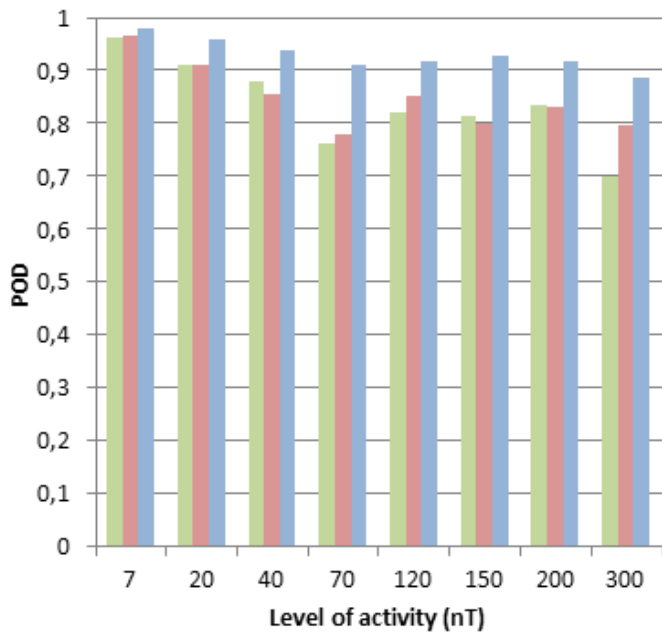
La Figure 85 présente les performances des réseaux en considérant en entrée $\{fs, IMF|B|, n, B_z\}$ et la Figure 86 celles des réseaux avec en entrée $\{fs, IMF|B|, n, V_x\}$. On constate dans tous les cas que le NARX reste le réseau le plus performant avec notamment pour le plus haut niveau d'activité, une POD de 0.881 et un FAR de 0.060 en considérant B_z , et une POD de 0.886 et un FAR de 0.0678 en considérant V_x . En comparant ces performances à celles du réseau NARX présenté Figure 43 basé uniquement sur les paramètres d'entrée $\{fs, IMF|B|, n\}$, on constate que ces nouvelles entrées ne permettent pas au réseau NARX de fournir de meilleures prédictions.

Les performances du TDNN sont modifiées en considérant le B_z et le V_x , et les résultats obtenus ne permettent pas de justifier l'utilisation de ces paramètres en entrée du réseau. Si la POD est légèrement améliorée pour certains niveaux d'activité, le FAR correspondant ne l'est pas nécessairement. Par exemple, en considérant le seuil d'activité compris entre 200 nT et 300 nT, avec en entrée $\{fs, IMF|B|, n\}$, on obtient une POD de 0.824 et un FAR de 0.071, en ajoutant la composante B_z en entrée on a une POD de 0.828 et un FAR de 0.0969. On a alors une meilleure POD mais un FAR plus élevé. Le constat est le même en ajoutant cette fois-ci V_x en entrée avec une POD de 0.832 et un FAR de 0.101.



	POD			FAR		
	Feed forward NN 6h	TDNN 6h	NARX NN 6h	Feed forward NN 6h	TDNN 6h	NARX NN 6h
7	0,967	0,971	0,978	0,0256	0,0619	0,0189
20	0,901	0,901	0,956	0,075	0,115	0,0401
40	0,835	0,825	0,935	0,135	0,145	0,0412
70	0,778	0,781	0,901	0,154	0,189	0,0654
120	0,817	0,8121	0,921	0,101	0,129	0,0501
150	0,818	0,8	0,924	0,079	0,1	0,0615
200	0,845	0,828	0,916	0,115	0,0969	0,041
300	0,712	0,798	0,881	0,061	0,0712	0,06

Figure 85- POD et FAR de chaque réseau avec les données ACE au point LI en considérant la densité, la vitesse, l'IMF|B| et B_z avec un historique de temps de six heures. Le réseau « feedforward » est en vert, le TDNN en rouge et le NARX en bleu.



	POD			FAR		
	Feed forward NN 6h	TDNN 6h	NARX NN 6h	Feed forward NN 6h	TDNN 6h	NARX NN 6h
7	0,965	0,967	0,98	0,035	0,072	0,0167
20	0,91	0,911	0,961	0,081	0,12	0,051
40	0,88	0,857	0,94	0,112	0,151	0,0452
70	0,761	0,781	0,91	0,145	0,165	0,0758
120	0,82	0,851	0,918	0,112	0,118	0,0615
150	0,813	0,801	0,928	0,089	0,112	0,0768
200	0,837	0,832	0,92	0,121	0,101	0,0485
300	0,7013	0,796	0,886	0,078	0,081	0,0678

Figure 86- POD et FAR de chaque réseau avec les données ACE au point L1 en considérant la densité, la vitesse, l'IMF|B| et V_x avec un historique de temps de six heures. Le réseau « feedforward » est en vert, le TDNN en rouge et le NARX en bleu.

2. DE MATLAB VERS PYTHON

Par définition, Python est un langage de programmation interprété, c'est-à-dire que les instructions qu'on lui envoie sont transcrites en langage machine au fur et à mesure de leur lecture. C'est un langage simple et portable sous différents environnements, ne demandant pas de licences spécifiques. Il y a une implémentation de base en C (appelée Cpython), et le programmeur peut, en fonction de ses besoins faire appel à différentes bibliothèques. Il peut également choisir son environnement de développement (ou Integrated development environment, IDE). Matlab quant à lui est un environnement numérique commercial de calcul qui utilise son propre langage de programmation et possède son IDE. Comme Python, Matlab propose des fonctions de calculs génériques, et pour faire des calculs spécifiques (traitement de données, statistiques, machine learning etc...) il faut acheter des toolbox. Sur la Figure 87, nous avons résumé les principales différences entre ces deux langages.

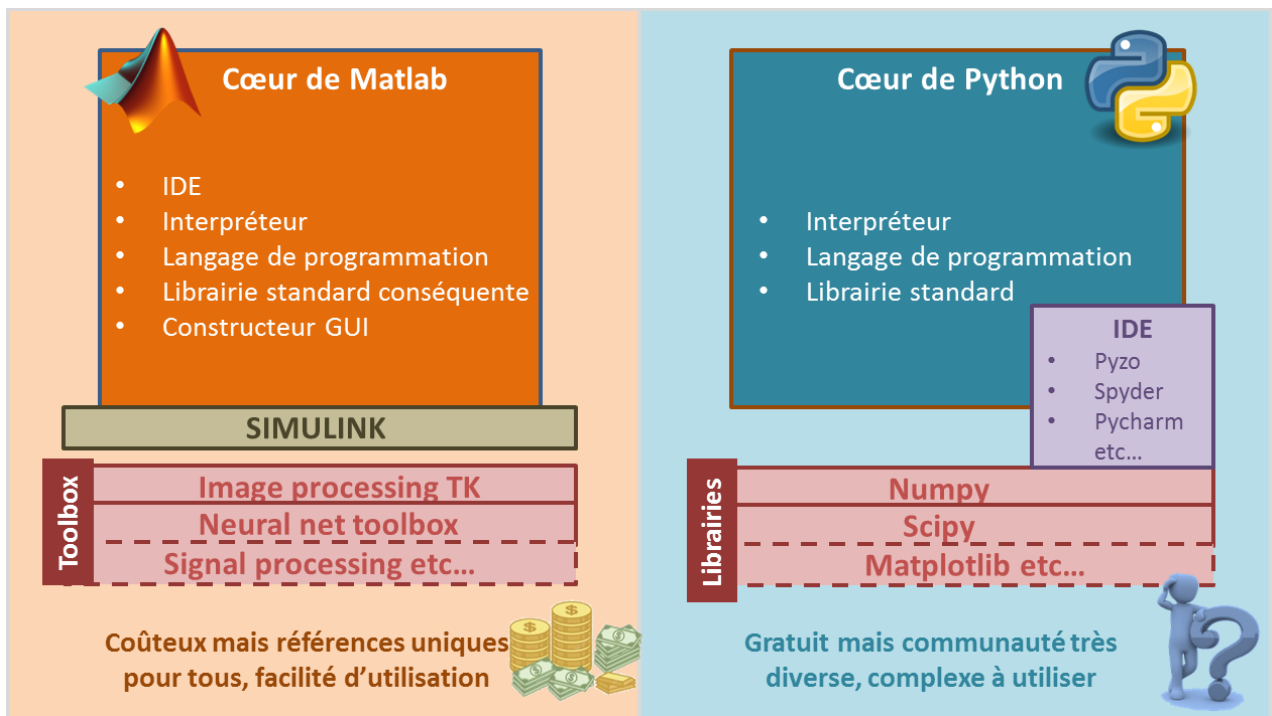


Figure 87- Matlab vs Python.

Si Matlab a des avantages comme être facile d'approche pour des débutants, avoir des packages combinant tout (alors qu'en Python il faut tout installer et choisir une IDE) et avoir une communauté qui utilise des toolbox uniques pour répondre à une problématique (ce qui permet d'avoir des repères communs et de faciliter la compréhension d'un problème). Ce langage a aussi de nombreux défauts, les principaux étant de notre expérience:

- Les algorithmes sont « proprietary » c'est-à-dire qu'on ne peut accéder au code de la plupart des algorithmes, notamment dans notre cas ceux de certaines fonctions utilisés pour les réseaux de neurones. On ne peut donc pas vérifier leur implémentation ni avoir une visibilité totale sur leur fonctionnement.
- Matlab est coûteux, tous les laboratoires ne font pas appel à ce langage et si l'on souhaite pouvoir échanger dans cette communauté, il est nécessaire d'avoir un langage facilement accessible.
- Matlab met des restrictions sur la portabilité des codes, et ce défaut rejoint le point précédent car si un autre utilisateur n'a pas exactement le même environnement que celui que nous utilisons, des erreurs peuvent apparaître et empêcher un autre utilisateur de s'en servir.

Dans une ère où la question du big data se pose et que des défi open source sont mis en place, notamment pour traiter les données du vent solaire, il est nécessaire d'utiliser un langage accessible par tous.

Si Matlab présente l'avantage de proposer une toolbox unique pour développer des réseaux de neurones, Python, étant open source, propose différentes bibliothèques et il est nécessaire de prendre le temps de les analyser pour choisir celle qui est la mieux adaptée. Dans la communauté du deep learning, les modèles de réseaux de neurones sont développés principalement pour faire de la classification ou de la reconnaissance d'image. Dans le cadre d'études et de prédictions de séries temporelles, il est plus complexe de trouver des bibliothèques adaptées à notre problème. C'est pourquoi nous avons lancé des bouteilles à la mer à différents laboratoires travaillant dans ce domaine afin de trouver une solution optimale. L'IRIT ou Institut de Recherche en Informatique de Toulouse, a répondu favorablement à notre requête. Nous avons alors travaillé avec Thomas Pellegrini, maître de conférences dans le domaine de l'utilisation du deep learning dans le cadre d'étude audiovisuelle. Si le domaine semble très éloigné de l'interaction Soleil-Terre, comme nous l'avons expliqué au Chapitre 2, il n'existe pas de structures définies spécifiquement pour notre problématique. Nous cherchons donc des structures ayant des caractéristiques assimilables à ce que nous cherchons à modéliser.

2.1. Définition des bibliothèques utilisées

Sous Python il faut choisir différentes bibliothèques pour construire l'environnement dans lequel on travaille. Après avoir travaillé sur la question avec l'IRIT, nous avons fait les choix suivants :

- Numpy qui est le package de base à considérer pour faire du calcul scientifique sous Python
- Theano qui est une surcouche de numpy, utilisée pour optimiser, définir et évaluer des expressions mathématiques
- Lasagne qui est une bibliothèque faite pour construire et entraîner des réseaux de neurones. Grâce à cette bibliothèque, il est possible de construire des réseaux de neurones comme de travailler avec des lego. Cette bibliothèque propose différents types de couche que le programmeur agence comme il le souhaite. Ceci nous permet de créer des architectures plus complexes comme le réseau LSTM, et nous résumons dans le Tableau 15 la « politique » associée à l'utilisation de Lasagne.

Tableau 15- Utilisation de Lasagne.

Simplicité	Facile à utiliser, étendre, applicable à la recherche
Transparence	Utilisation directe des fonctions mathématiques définies dans Theano et numpy
Modularité	Permet à tous les éléments d'être utilisés indépendamment (couches, régularisation, optimisation...)
Pragmatisme	Pas de surévaluation des cas inhabituels
Modération	N'entrave pas l'utilisateur avec des fonctions qu'il ne souhaite pas utiliser
Focus	« faire une chose et le faire bien »

2.2. Architecture globale du code

Nous avons représenté l'architecture globale du code sur la Figure 88, mais si au premier abord celle-ci semble similaire à celle proposée par Matlab, il y a quelques différences notables :

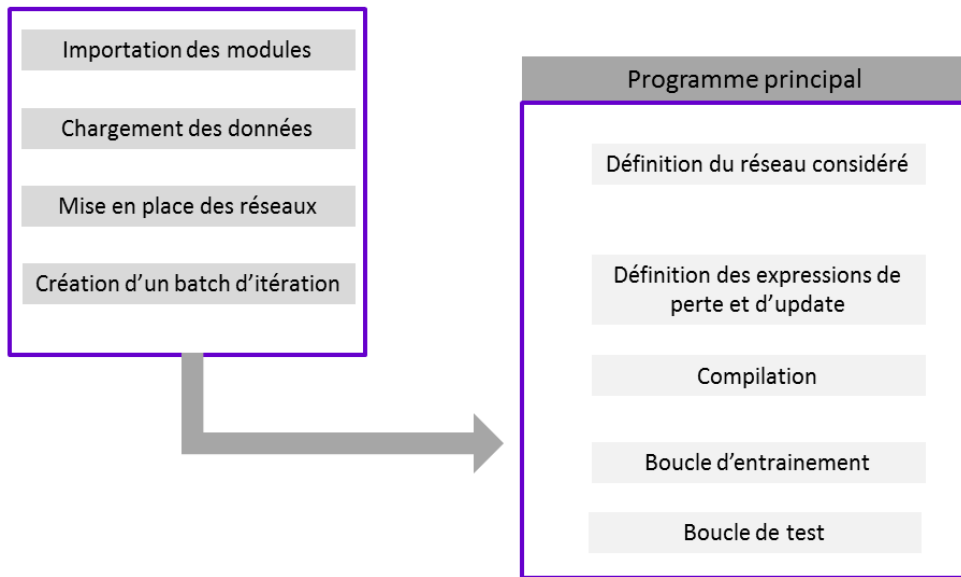


Figure 88- Architecture globale du code.

- Définition du réseau considéré : en Python, il faut définir des couches appropriées au réseau que nous souhaitons programmer. Sous Matlab, la structure de base est définie par une commande, et nous pouvons ensuite travailler sur le nombre de couches cachées, de nœuds etc.. pour minimiser l'erreur durant l'entraînement du réseau. Sous Python, le programmeur construit son réseau en utilisant des couches spécifiques (toutes détaillées sur le site <http://lasagne.readthedocs.io/en/latest/modules/layers.html>).
- Expressions de perte et update : Lasagne propose des fonctions d'update pour implémenter différentes méthodes pour contrôler le taux d'apprentissage, que l'on peut utiliser avec des algorithmes de gradient de descente. Matlab avait l'avantage de définir automatiquement des limiteurs et des régulateurs durant l'apprentissage notamment afin que les calculs ne divergent pas. Sur Python, il faut les définir et les contrôler car le calcul peut « exploser ». Ceci demande davantage de programmation mais permet d'avoir une meilleure visibilité sur l'état du réseau de neurones.
- Compilation : après avoir défini la fonction de perte, il est possible en compilant de tester le réseau sur un mini batch afin de voir comment il se comporte et s'il n'y a pas d'erreurs, avant de le lancer sur toute une série de données. Sur Matlab cette option n'est pas disponible, et aussi bien sous Matlab que sous Python, les temps de calcul pour la phase d'entraînement peuvent être longs donc il est intéressant de faire un test sur une série plus petite avant de lancer cette phase.
- Boucle d'entraînement et boucle de test : sur Python le programmeur n'est pas obligé de définir une phase de validation, comparable à la phase de test, ce qui permet de gagner du temps de calcul. Sur Matlab, les trois phases sont obligatoires.

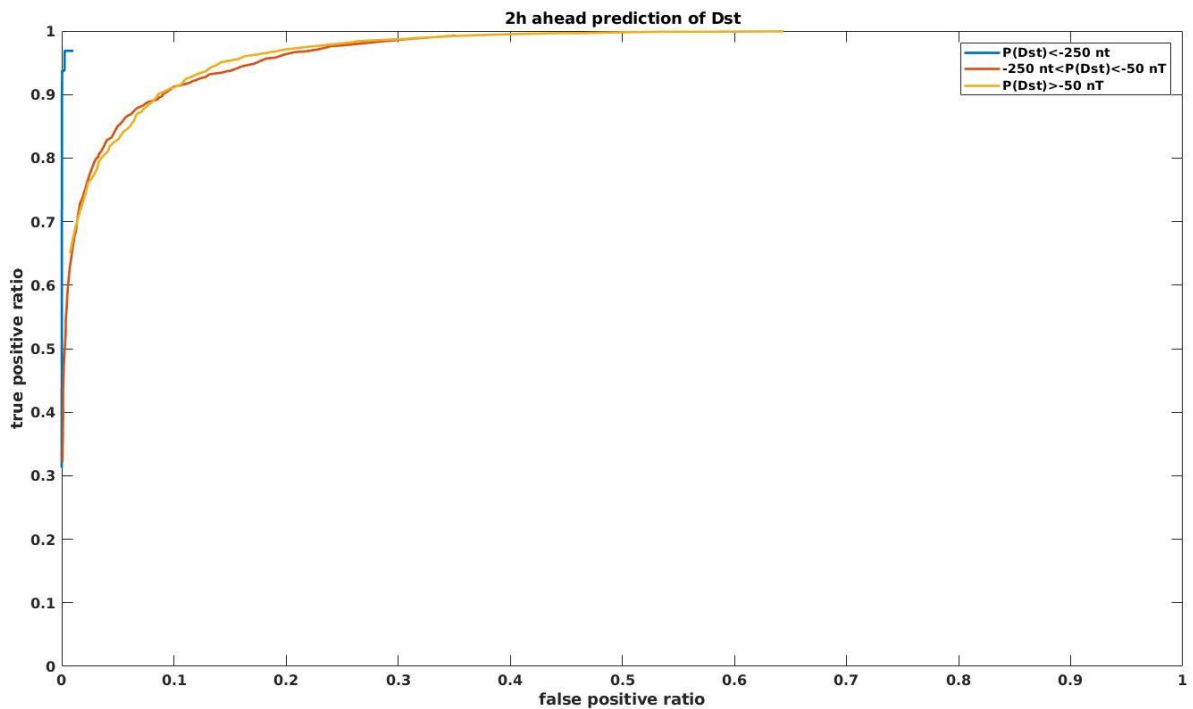
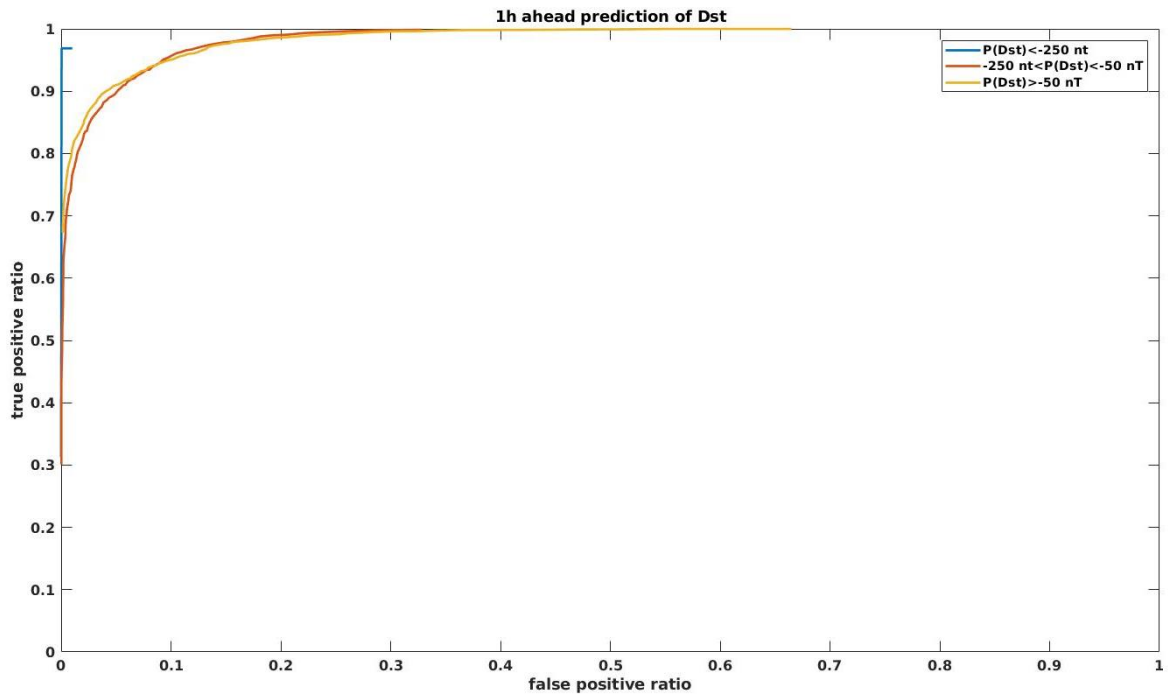
3. LISTE DES ORAGES UTILISES POUR LE TEST DU PROCESSUS GAUSSIEN COMBINE AU RESEAU LSTM

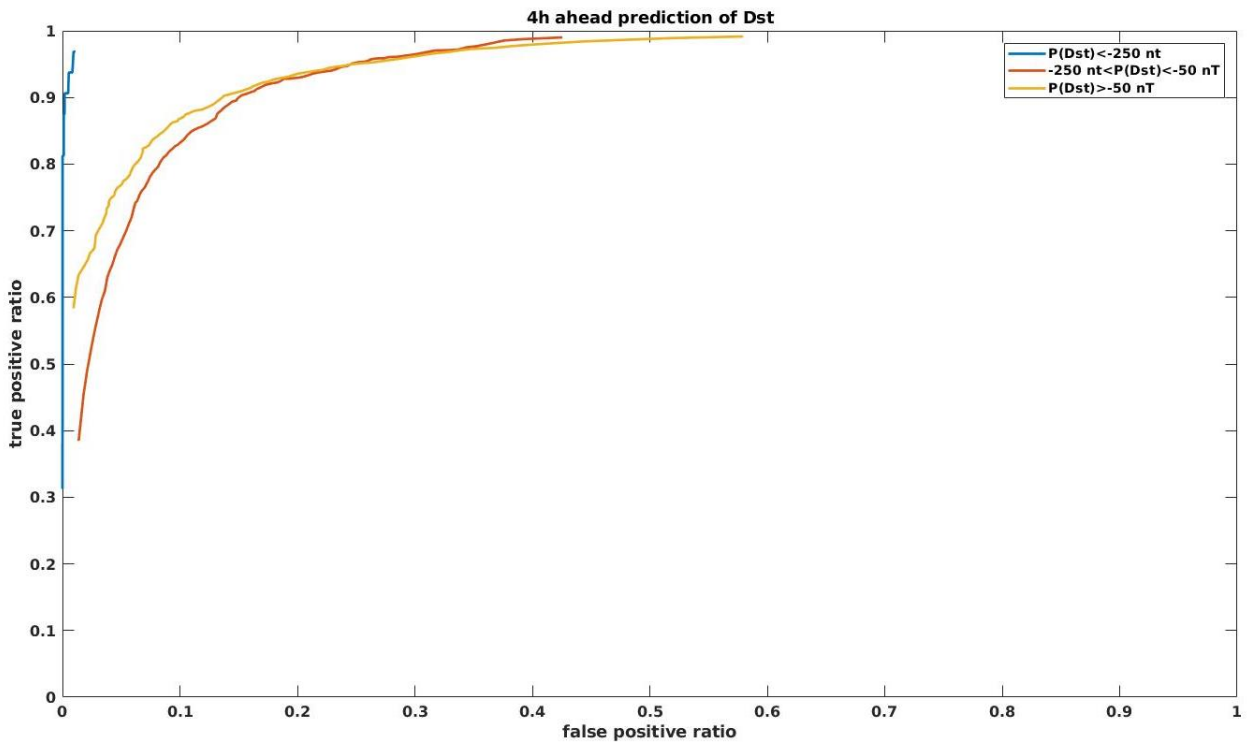
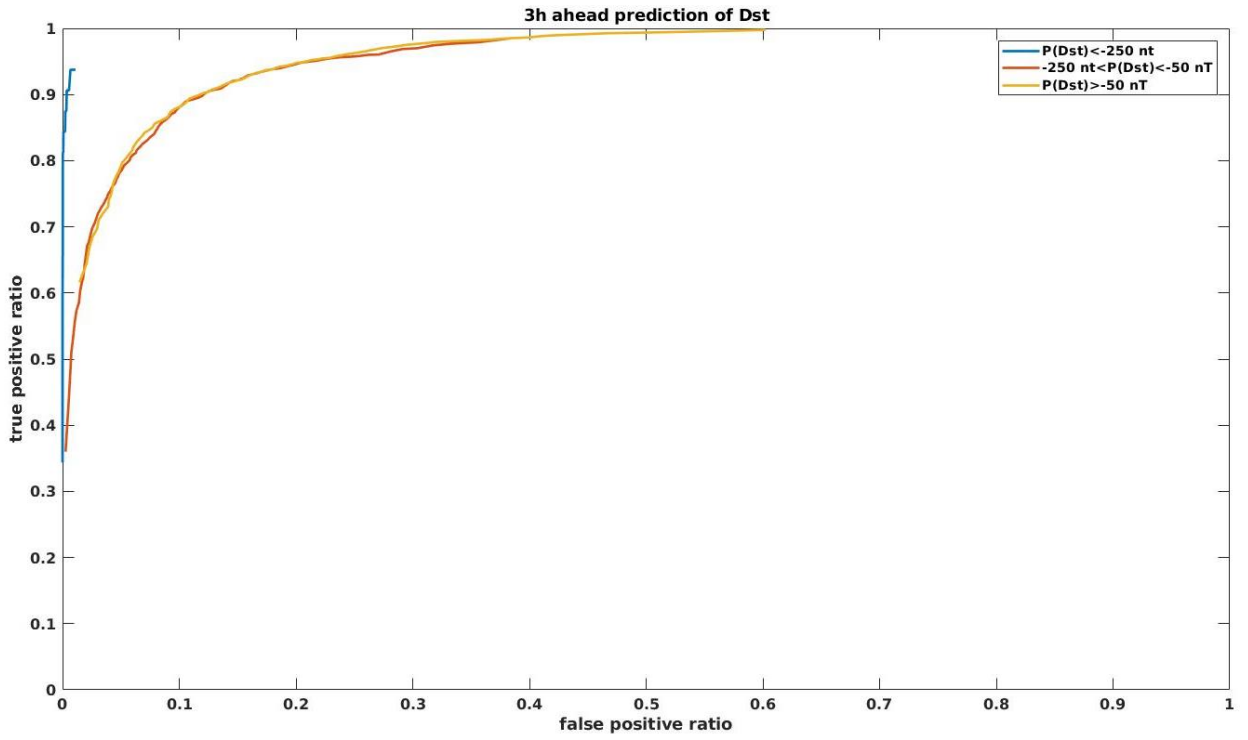
Start date	Start time	End date	End time	Min. <i>Dst</i> (nT)
2001 / 3 / 19	1500	2001 / 3 / 21	2300	-149
2001 / 3 / 31	400	2001 / 4 / 1	2100	-387
2001 / 4 / 18	100	2001 / 4 / 18	1300	-114
2001 / 4 / 22	200	2001 / 4 / 23	1500	-102
2001 / 8 / 17	1600	2001 / 8 / 18	1600	-105
2001 / 9 / 30	2300	2001 / 10 / 2	0	-148
2001 / 10 / 21	1700	2001 / 10 / 24	1100	-187
2001 / 10 / 28	300	2001 / 10 / 29	2200	-157
2002 / 3 / 23	1400	2002 / 3 / 25	500	-100
2002 / 4 / 17	1100	2002 / 4 / 19	200	-127
2002 / 4 / 19	900	2002 / 4 / 21	600	-149
2002 / 5 / 11	1000	2002 / 5 / 12	1600	-110
2002 / 5 / 23	1200	2002 / 5 / 24	2300	-109
2002 / 8 / 1	2300	2002 / 8 / 2	900	-102
2002 / 9 / 4	100	2002 / 9 / 5	0	-109
2002 / 9 / 7	1400	2002 / 9 / 8	2000	-181
2002 / 10 / 1	600	2002 / 10 / 3	800	-176
2002 / 11 / 20	1600	2002 / 11 / 22	600	-128
2003 / 5 / 29	2000	2003 / 5 / 30	1000	-144
2003 / 6 / 17	1900	2003 / 6 / 19	300	-141
2003 / 7 / 11	1500	2003 / 7 / 12	1600	-105
2003 / 8 / 17	1800	2003 / 8 / 19	1100	-148
2003 / 11 / 20	1200	2003 / 11 / 22	0	-422
2004 / 1 / 22	300	2004 / 1 / 24	0	-149
2004 / 2 / 11	1000	2004 / 2 / 12	0	-105
2004 / 4 / 3	1400	2004 / 4 / 4	800	-112
2004 / 7 / 22	2000	2004 / 7 / 23	2000	-101
2004 / 7 / 24	2100	2004 / 7 / 26	1700	-148
2004 / 7 / 26	2200	2004 / 7 / 30	500	-197

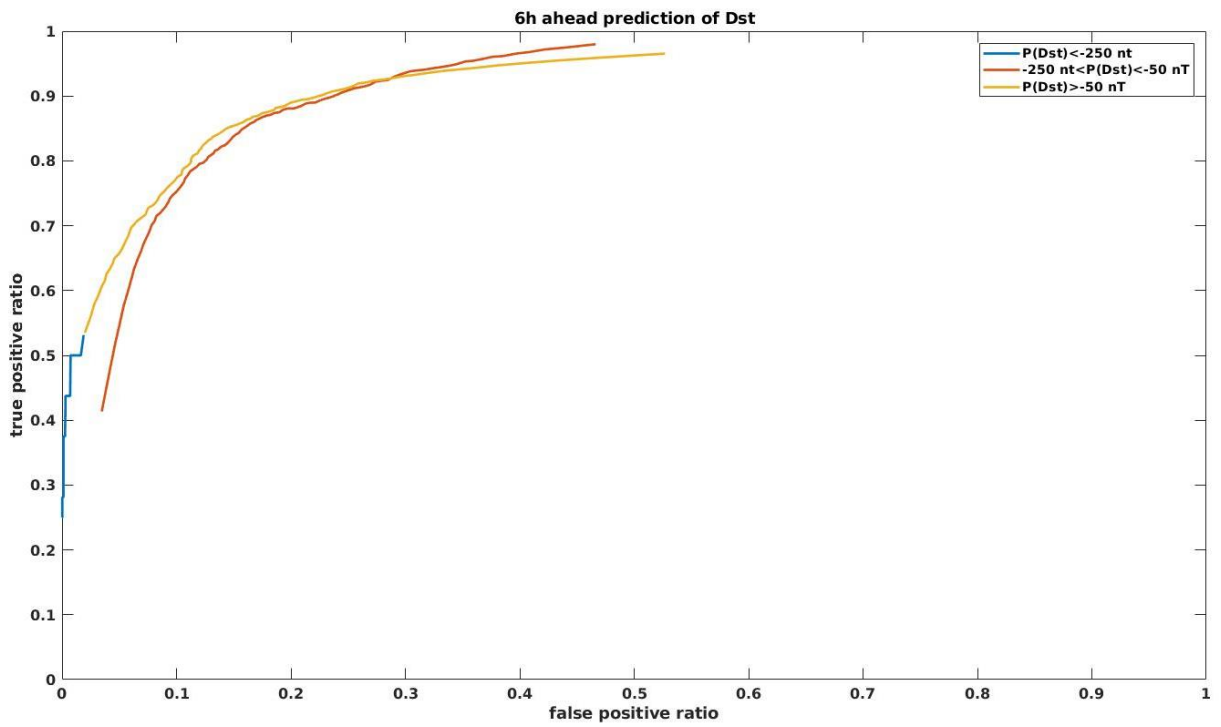
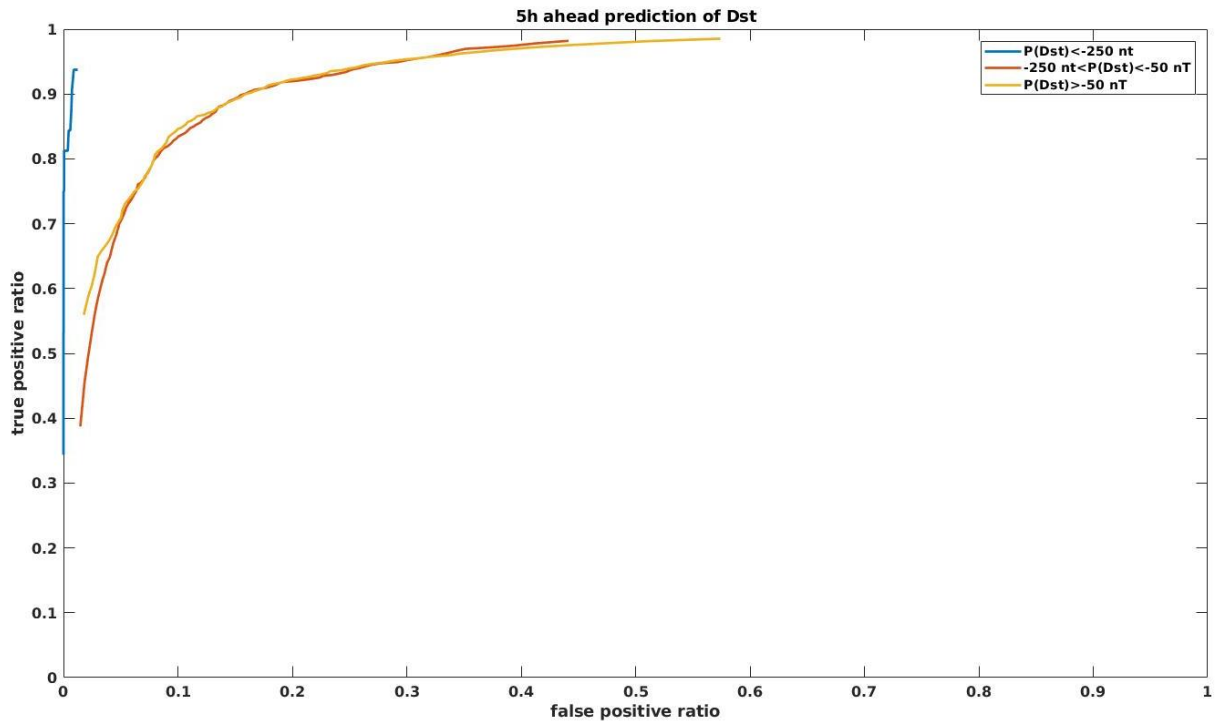
2004 / 8 / 30	500	2004 / 8 / 31	2100	-126
2004 / 11 / 11	2200	2004 / 11 / 13	1300	-109
2005 / 1 / 21	1800	2005 / 1 / 23	500	-105
2005 / 5 / 7	2000	2005 / 5 / 9	1000	-127
2005 / 5 / 29	2200	2005 / 5 / 31	800	-138
2005 / 6 / 12	1700	2005 / 6 / 13	1900	-106
2005 / 8 / 31	1200	2005 / 9 / 1	1200	-131
2006 / 4 / 13	2000	2006 / 4 / 14	2300	-111
2006 / 12 / 14	2100	2006 / 12 / 16	300	-147
2011 / 9 / 26	1400	2011 / 9 / 27	1200	-101
2011 / 10 / 24	2000	2011 / 10 / 25	1400	-132
2012 / 3 / 8	1200	2012 / 3 / 10	1600	-131
2012 / 4 / 23	1100	2012 / 4 / 24	1300	-108
2012 / 7 / 15	100	2012 / 7 / 16	2300	-127
2012 / 9 / 30	1300	2012 / 10 / 1	1800	-119
2012 / 10 / 8	200	2012 / 10 / 9	1700	-105
2012 / 11 / 13	1800	2012 / 11 / 14	1800	-108
2013 / 3 / 17	700	2013 / 3 / 18	1000	-132
2013 / 5 / 31	1800	2013 / 6 / 1	2000	-119
2014 / 2 / 18	1500	2014 / 2 / 19	1600	-112

-

4. COURBES ROC OBTENUES AVEC LA METHODE GPNN DANS LE CAS DE LA PREDICTION DE L'INDICE MAGNETIQUE DST







BIBLIOGRAPHIE

- [Amari et al., 2018] – Amari T., A. Canou, J.J. Aly, F. Delyon and F. Alauzet (2018), Magnetic cage and rope as the key for solar eruptions. *Nature*, 554(7691), 211.
- [Akasofu, 1981] - Akasofu, S. I. (1981), Energy coupling between the solar wind and the magnetosphere, *Space. Sci. ev.* 28, 121–190, doi:10.1007/BF00218810.
- [Andriambahoaka, Z., 2008] - Andriambahoaka, Z. (2008). Modélisation régionale du champ magnétique terrestre et établissement de cartes magnétiques détaillées appliqués à Madagascar (Doctoral dissertation, Strasbourg 1).
- [Armstrong, 1984] – Armstrong, J.S. (1984). Forecasting by extrapolation: Conclusions from 25 years of research. *Interfaces*, 14(6):52 – 66.
- [Ayala Solares et al., 2016] - Ayala Solares, J. R., H. L. Wei, R.J. Boynton, S. N. Walker and S. A. Billings (2016), Modeling and prediction of global magnetic disturbance in near-Earth space: A case study for Kp index using NARX models, *Space Weather*, 14(10), 899-916.
- [Bala and Reiff, 2012] - Bala, R., and P. Reiff (2012), Improvements in short term forecasting of geomagnetic activity, *Space Weather*, 10(6).
- [Bargatze et al., 1985] - Bargatze, L. F., D. N. Baker, R.L. McPherron and E.W. Hones (1985), Magnetospheric impulse response for many levels of geomagnetic activity, *Journal of Geophysical Research: Space Physics*, 90(A7), 6387-6394.
- [Bargatze et al, 1986] - Bargatze, L. F., R. L. McPherron, and D. N. Baker (1986), Solar wind-magnetosphere energy input functions, in *Solar Wind-Magnetosphere Coupling*, edited by Y. Kamide and J. A. Slavin, pp. 93–100, Terrapub/Reidel, Tokyo, Japan.
- [Bhaskar and Vichare, 2017] - Bhaskar, A. and G. Vichare (2017), Prediction of SYMH and ASYH indices for geomagnetic storms of solar cycle 24 including recent St Patrick’s day, 2015 storm using NARX neural network, arXiv preprint arXiv::1703.10583
- [Bjørnstad and Grenfell, 2001] - Bjørnstad O.N and B. T. Grenfell (2001), Noisy clockwork: Time series analysis of population fluctuations in animals. *Science*, Vol. 293(5530):638–643.
- [Boberg et al., 2000] - Boberg, F., P. Wintoft, and H. Lundstedt (2000), Real time Kp predictions from solar wind data using neural networks, *Physics and Chemistry of the Earth, Part C: Solar, Terrestrial & Planetary Science*, 25(4), 275-280.
- [Bouhlef et al. 2016] - Bouhlef, M.A., N. Bartoli, A. Otsmane and J.Morlier (2016), Improving kriging surrogats of high dimensional design models by Partial Least Squares dimension reduction, *Structural and multidisciplinary Optimization*, vol 53, 5, p 935-952.
- [Brown, 2004] - Brown R.G. (2004), Smoothing, forecasting and prediction of discrete time series. Dover Publications, Inc., Mineola, New York.
- [Cai et al., 2009] - Cai, L., S. Ma, H. Cai, Y. Zhou, and R. Liu (2009), Prediction of SYM-H index by NARX neural network from IMF and solar wind data, *Science in China Series E: Technological Sciences*, 52 (10), 2877–2885.

- [Camporeale et al., 2016] - Camporeale, E., Carè, A., and J.E. Borovsky. (2017), Classification of solar wind with machine learning. *Journal of Geophysical Research: Space Physics*, 122(11), doi:10.1002/2017JA024383.
- [Chambodut et al., 2013] - Chambodut, A., A. Marchaudon, M. Menvielle (2013), The K-derived MLT sector geomagnetic indices, *Geophysical Research Letters*,40(18), 4808-4812, DOI:10.1002/grl.50947.
- [Chapman and Ferraro, 1931] - Chapman, S. and Ferraro, V. (1931), A new theory of magnetic storms. *Terrestrial magnetism and atmospheric electricity*, 36(2) :77–97.
- [Chatfield, 2000] - Chatfield C. (2000), *Time-Series Forecasting*. CRC Press.
- [Chandorkar et al., 2017] - Chandorkar, M., Camporeale, E., and S. Wing. (2017), Probabilistic forecasting of the disturbance storm time index: An autoregressive Gaussian process approach. *Space Weather*, 15(8), 1004-1019, doi:10.1002/2017SW001627.
- [Chiplunkar, 2017] - Chiplunkar, A., Bosco, E., & Morlier, J. (2017, June), Gaussian Process for Aerodynamic Pressures Prediction in Fast Fluid Structure Interaction Simulations. In *World Congress of Structural and Multidisciplinary Optimisation* (pp. 221-233). Springer, Cham.
- [Coleman, 2005] – Coleman I.J. (2005), A multi-spacecraft survey of magnetic field lline draping in the dayside magnetosheath, *Ann. Geophys.*, 23, 885-900.
- [Costello, 1998] - Costello, K. A. (1998), *Moving the Rice MSFM into a real-time forecast mode using solar wind driven forecast modules* (Doctoral dissertation, Rice University).
- [Deutsch et al, 1994] – Deutsch T., E. Lehmann, E. Carson, A. Roudsari, K. Hopkins & P. Sönksen (1994) , Time series analysis and control of blood glucose levels in diabetic patients. *Comput Methods Programs Biomed.*, 41(3-4):167–182.
- [Finch and Lockwood, 2007] - Finch, I., & Lockwood, M. (2007, March), Solar wind-magnetosphere coupling functions on timescales of 1 day to 1 year. In *Annales Geophysicae* (Vol. 25, No. 2, pp. 495-506).
- [Franses, 1998] - Franses P.H. (1998), *Time Series Models for Business and Economic Forecasting*. Cambridge University Press.
- [Frigola-Alcalde, 2015] - Frigola-Alcalde, R. (2015), *Bayesian Time Series Learning with Gaussian Processes* . Doctoral thesis, University of Cambridge.
- [Gao and Er, 2005] - Gao, Y. and M. J. Er (2005), NARMAX time series model prediction: feedforward and recurrent fuzzy neural network approaches, *Fuzzy sets and systems*, 150(2), 331-350.
- [Gégout et al., 1995] - Gégout, C., B. Girau and F. Rossi (1995), Generic back-propagation in arbitrary feedforward neural networks. In *Artificial Neural Nets and Genetic Algorithms* (pp. 168-171). Springer, Vienna.
- [Gers et al., 2002] - Gers, F. A., Schraudolph, N. N., & Schmidhuber, J. (2002), Learning precise timing with LSTM recurrent networks. *Journal of machine learning research*, 3(Aug), 115-143.

- [Gleisner et al., 1996] - Gleisner, H., H. Lundstedt, and P. Wintoft (1996), Predicting geomagnetic storms from solar-wind data using time-delay neural networks, in *Annales Geophysicae*, 14 (7), 679.
- [Gleisner and Lunsdtedt, 1997] - Gleisner, H. and H. Lundstedt (1997), Response of the auroral electrojets to the solar wind modeled with neural networks, *Journal of Geophysical Research: Space Physics*, 102(A7), 14269-14278.
- [Godambe, 1966] - Godambe V.P. (1966), A new approach to sampling from finite populations. i sufficiency and linear estimation. *Journal of the Royal Statistical Society, Series B (Methodological)*, 28(2):310–319.
- [Gold, 1959] – Gold, T (1959), Motions in the magnetosphere of the earth « *journal of geophysical research* » 64 (9) 120 p 89.105.
- [Hawkins, 2004] – Hawkins D.M. (2004), The problem of overfitting. *Journal of chemical information and computer sciences*, 44:1–12.
- [Haykin, 1998] - Haykin, S. (1998), *Neural Networks: A Comprehensive Foundation* , Prentice Hall, Upper Saddle River, N. J.
- [Hill and Rassbach, 1975] - Hill, T. W., & Rassbach, M. E. (1975), Interplanetary magnetic field direction and the configuration of the day side magnetosphere. *Journal of Geophysical Research*, 80(1), 1-6.
- [Hochreiter and Schmidhuber, 1997] - Hochreiter, S. and J. Schmidhuber (1997), Long Short-term memory, *Neural Computation*, 9(8), 1735-1780.
- [Hochreiter, 1998] - Hochreiter, Sepp. (1998), The vanishing gradient problem during learning recurrent neural nets and problem solutions, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 6.02 (1998): 107-116.
- [James et al. 2013] – James G., D. Witten, T. Hastie and R. Tibshirani (2013), *An introduction to statistical learning*, volume 6. Springer.
- [Ji et al, 2012] - Ji, E. Y., Y. J. Moon, N. Gopalswamy, and D. H. Lee (2012), Comparison of Dst forecast models for intense geomagnetic storms, *Journal of Geophysical Research: Space Physics* , 117 (3), 1– 9, doi:10.1029 / 2011JA016872.
- [Jian et al., 2006] - Jian, L., Russell, C. T., Luhmann, J. G., & Skoug, R. (2006), Properties of stream interactions at one AU during 1995–2004. *Solar Physics*, 239(1-2), 337-392.
- [Kataoka and Miyoshi, 2008] - Miyoshi, Y. and R. Kataoka (2008), Flux enhancement of the outer radiation belt electrons after the arrival of stream interaction regions. *Journal of Geophysical Research: Space Physics*, 113(A3).
- [Kendall, 1938] – Kendall, M. (1938), A New Measure of Rank correlation », *Biometrika*, vol. 30, nos 1–2, 1938, p. 81–89, doi:10.1093/biomet/30.1-2.81, JSTOR 2332226.
- [Kilian and Siegelmann, 1993] - Kilian, J. and H. T. Siegelmann (1993), On the power of sigmoid neural networks, in *Proceedings of the sixth annual conference on Computational learning theory*, 137-143, ACM, Santa Cruz, California, USA, 26-28 July 1993.

- [Krige, 1951] - Krige, D. G. (1951), A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy*, 52(6), 119-139.
- [Kumar et al., 2012] - Kumar, S., Priyadarshi, S., Krishna, S. G., & Singh, A. K. (2012), GPS-TEC variations during low solar activity period (2007–2009) at Indian low latitude stations. *Astrophysics and Space Science*, 339(1), 165-178.
- [Lazzús et al., 2017] - Lazzús, J. A., Vega, P., Rojas, P., and I. Salfate. (2017), Forecasting the Dst index using a swarm-optimized neural network. *Space Weather*, 15(8), 1068-1089, doi:10.1002/2017SW001608.
- [Leontaritis and Billings, 1985] - Leontaritis, I. J. and S. A. Billings (1985), Input-output parametric models for non-linear systems part I: deterministic non-linear systems, *International journal of control*, 41(2), 303-328.
- [Levenberg, 1944] - Levenberg, K. (1944), A method for the solution of certain problems in least-squares, *Quarterly Applied Mathematics*, 2, 164-168.
- [Lilensten and Bornarel 2001] - Lilensten, J., and Bornarel, J. (2001), *Sous les feux du soleil: vers une météorologie de l'espace*. L'Editeur: EDP Sciences.
- [Lin et al, 1996] - Lin, T., B. G. Horne, P. Tino, and C. L. Giles (1996), Learning long-term dependencies in NARX recurrent neural networks, *IEEE Transactions on Neural Networks*, 7(6), 1329-1338.
- [Lundstedt and Wintoft, 1994] - Lundstedt, H., and P. Wintoft (1994), Prediction of geomagnetic storms from solar wind data with the use of a neural network, *Annales Geophysicae*, 12, 19-24.
- [Marquardt, 1963] - Marquardt, D. W. (1963), An algorithm for least-squares estimation of non-linear parameters, *Journal of the society for Industrial and Applied Mathematics*, 11(2), 431- 441.
- [Mayaud, 1968] - Mayaud, P. N. (1980), Derivation, meaning, and use of geomagnetic indices, *Washington DC American Geophysical Union Geophysical Monograph Series*, 22.
- [Mazouz et al., 2013] - Mazouz, F., C. Lathuillère, M. Menvielle, N. Sanchez-Ortiz (2013), Prototype Forecast of the DTM geomagnetic proxies, *Space call 3*.
- [McCulloch and Pitts, 1943] - McCulloch, W. S. and W. Pitts (1943), A logical calculus of the ideas immanent in nervous activity, *The bulletin of mathematical biophysics*, 5(4), 115-133.
- [McPherron et al., 1970] - McPherron, R. L. (1970), Growth phase of magnetospheric substorms. *Journal of Geophysical Research*, 75(28), 5592-5599.
- [Morley et al. 2017] - Morley, S. K., Sullivan, J. P., Carver, M. R., Kippen, R. M., Friedel, R. H. W., Reeves, G. D., and M.G. Henderson. (2017), Energetic particle data from the global positioning system constellation. *Space Weather*, 15(2), 283-289, doi: 10.1002/2017SW001604.
- [Murayama et al, 1982] - Murayama, T. (1982), Coupling function between solar wind parameters and geomagnetic indices. *Reviews of Geophysics*, 20(3), 623-629.

- [Neal, 1996] - Neal, R. M. (1996), Priors for infinite networks. In *Bayesian Learning for Neural Networks* (pp. 29-53). Springer, New York, NY.
- [Newell et al., 2007] - Newell, P. T., Sotirelis, T., Liou, K., Meng, C. I., & Rich, F. J. (2007), A nearly universal solar wind-magnetosphere coupling function inferred from 10 magnetospheric state variables. *Journal of Geophysical Research: Space Physics*, 112(A1).
- [Nielsen, 2015] - Nielsen, M. A. (2015). *Neural networks and deep learning*. Determination Press.
- [Papadopoulos et al., 1999] - Papadopoulos, K., Goodrich, C., Wiltberger, M., Lopez, R., and Lyon, J. G. (1999), The physics of substorms as revealed by the ISTP. *Physics and Chemistry of the Earth, Part C: Solar, Terrestrial & Planetary Science*, 24(1-3), 189-202.
- [Parker, 1958] - Parker E.N (1958), Dynamics of the interplanetary gas and magnetic fields , *Astrophysical journal*, vol 128 p 664.
- [Palmroth et al., 2003] - Palmroth, M., T. I. Pulkkinen, P. Janhunen, and C. C. Wu (2003), Stormtime energy transfer in global MHD simulation, *J. Geophys. Res.*, 108(A1), 1048, doi:10.1029/2002JA009446.
- [Palmroth et al., 2005] - Palmroth, M., P. Janhunen, T. I. Pulkkinen, A. Aksnes, G. Lu, N. Ostgaard, J. Watermann, G. D. Reeves, and G. A. Germany (2005), Assessment of ionospheric joule heating by GUMICS-4 MHD simulation, AMIE, and satellite-based statistics: Towards a synthesis, *Ann. Geophys.*, 23(6), 2051–2068.
- [Palmroth et al., 2006] - Palmroth, M., T. V. Laitinen, and T. I. Pulkkinen (2006), Magnetopause energy and mass transfer: Results from a global MHD simulation, *Ann. Geophys.*, 24(12), 3467–3480.
- [Palmroth et al., 2010] - Palmroth, M., H. E. J. Koskinen, T. I. Pulkkinen, P. K. Toivanen, P. Janhunen, S. E. Milan, and M. Lester (2010), Magnetospheric feedback in solar wind energy transfer, *J. Geophys. Res.*, 115, A00I10, doi:10.1029/2010JA015746.
- [Palmroth et al., 2012] - Palmroth, M., R. C. Fear, and I. Honkonen (2012), Magnetopause energy transfer dependence on the interplanetary magnetic field and the Earth's magnetic dipole axis orientation, *Ann. Geophys.*, 30(3), 515–526, doi:10.5194/angeo-30-515-2012.
- [Pearson, 1895] - Pearson K. (1895), Notes on regression and inheritance in the case of two parents, *Proceedings of the Royal Society of London*, 58 : 240–242
- [Perreault and Akasofu, 1978] - Perreault, P., and S.I. Akasofu, (1978) , A study of geomagnetic storms. *Geophysical Journal of the Royal Astronomical Society*, 54(3), 547-573.
- [Pulkkinen et al., 2002] - Pulkkinen, T. I., N. Y. Ganushkina, E. I. Kallio, G. Lu, D. N. Baker, N. E. Turner, T. A. Fritz, J. F. Fennell, and J. Roeder (2002), Energy dissipation during a geomagnetic storm: May 1998, *Adv. Space Res.*, 30(10), 2231–2240, doi:10.1016/S0273-1177(02)80232-0.
- [Pulkkinen et al., 2008] - Pulkkinen, T. I., M. Palmroth, and T. Laitinen (2008), Energy as a tracer of magnetospheric processes: GUMICS-4 global MHD results and observations compared, *J. Atmos. Sol. Terr. Phys.*, 70(5), 687–707, doi:10.1016/j.jastp.2007.10.01.

- [Pulkkinen et al., 2010] - Pulkkinen, T. I., M. Palmroth, P. Janhunen, H. E. J. Koskinen, D. J. McComas, and C. W. Smith (2010), Timing of changes in the solar wind energy input in relation to ionospheric response, *J. Geophys. Res.*, 115, A00I09, doi:10.1029/2010JA015764.
- [Rasmussen and Williams, 2006] - Rasmussen, C. E. and C.K. Williams. (2006). *Gaussian processes for machine learning*. 2006. The MIT Press, Cambridge, MA, USA, 38, 715-719.
- [Roberts et al. 2012] – Roberts S., M. Osborne, M. Edden, S. Reece, N. Gibson and S. Aigrain (2012). *Gaussian processes for time-series modelling*. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 371(1984).
- [Rochel et al. 2016] - Rochel, S., D. Boscher, R. Benacquista and J.F. Roussel (2016), A radiation belt disturbance study from the space weather point of view, *Acta Astronautica* 128, pp 650-656.
- [Rosenbauer et al. 1975] - Rosenbauer, H., Grünwaldt, H., Montgomery, M. D., Paschmann, G., and Scokopke, N. (1975), Heos 2 plasma observations in the distant polar magnetosphere: The plasma mantle. *Journal of Geophysical Research*, 80(19), 2723-2737.
- [Rosenblatt, 1958] - Rosenblatt, F. (1958), The perceptron: A probabilistic model for information storage and organization in the brain, *Psychological Review*, 65(6):386.
- [Šafránková et al., 2009] - Šafránková, J., Hayosh, M., Gutynska, O., Němeček, Z., and Přeč, L. (2009), Reliability of prediction of the magnetosheath BZ component from interplanetary magnetic field observations. *Journal of Geophysical Research: Space Physics*, 114(A12).
- [Siegelmaan and Sontag, 1991] - Siegelmann, H. T., and Sontag, E. D. (1991), Turing computability with neural nets. *Applied Mathematics Letters*, 4(6), 77-80.
- [Siegelmaan and Sontag, 1994] - Siegelmann, H. T., and Sontag, E. D. (1994), Analog computation via neural networks. *Theoretical Computer Science*, 131(2), 331-360.
- [Siegelmaan and Sontag, 1995] - Siegelmann, H. T. and Sontag, E. D. (1995), On the computational power of neural nets. *Journal of computer and system sciences*, 50(1), 132-150.
- [Shi et al. 2005] - Shi, Y., E. Zesta, L. R. Lyons, A. Boudouridis, K. Yumoto, and K. Kitamura (2005), Effect of solar wind pressure enhancements on the storm time ring current asymmetry, *J. Geophys. Res.*, 110, A10205, doi:10.1029/2005JA011019.
- [Shi et al. 2008] - Shi, Y., E. Zesta, and L. R. Lyons (2008), Modeling magnetospheric current response to solar wind dynamic pressure enhancements during magnetic storms: 2. Application to different storm phases, *J. Geophys. Res.*, 113, A10219, doi:10.1029/2008JA013420.
- [Stamper et al., 1999] - Stamper, R., Lockwood, M., Wild, M. N., and Clark, T. D. G. (1999), Solar causes of the long-term increase in geomagnetic activity. *Journal of Geophysical Research: Space Physics*, 104(A12), 28325-28342.
- [Sutton and Barto, 1998] - Sutton, R. S., and A. G. Barto. (1998). *Softmax action selection*. *Reinforcement Learning: An Introduction*.
- [Tao, 2011] Tao, T. (2011), *An Introduction to Measure Theory*, Graduate studies in mathematics, American 369, Mathematical Society.

- [Tipping, 2004] – Tipping M.E. (2004), Bayesian inference: An introduction to principles and practice in machine learning. In *Advanced lectures on machine Learning*, pp. 41–62. Springer.
- [Vasyliunas et al., 1982] - Vasyliunas, V.M, J.R. Kan, G.L. Siscoe and S.I. Akasofu (1982), Scaling relations governing magnetospheric energy-transfer, *Planet. Space Sci*, 30(4), 359-36.
- [Waibel et al., 1989] - Waibel, A., T. Hanazawa, G. Hinton, K. Shikano and K. J. Lang (1989), Phoneme recognition using time-delay neural networks, *IEEE transactions on acoustics, speech, and signal processing*, 37(3), 328-339.
- [Wang et al., 2014] - Wang, D. (2003), Temporal pattern processing. *The handbook of brain theory and neural networks*, 1163-1167.
- [Werbos, 1990] - Werbos P.J. (1990), Backpropagation through time: what it does and how to do it." *Proceedings of the IEEE* 78.10 : 1550-1560.
- [Wing et al. 2005] - Wing, S., Johnson, J. R., Jen, J., Meng, C. I., Sibeck, D. G., Bechtold, K., and K. Takahashi. (2005), Kp forecast models. *Journal of Geophysical Research: Space Physics*, 110(A4) doi:10.1029/2004JA010500.
- [Wohler and Anlauf, 1999] - Wohler, C., and Anlauf, J. K. (1999), An adaptable time-delay neural-network algorithm for image sequence analysis. *IEEE Transactions on Neural Networks*, 10(6), 1531-1536.
- [Wolpert and Macready, 1997] - D. H. Wolpert and W. G. Macready (1997), No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1(1):67–82.
- [Wu and Lundstedt, 1996] - Wu, J. G., & Lundstedt, H. (1996), Prediction of geomagnetic storms from solar wind data using Elman recurrent neural networks. *Geophysical research letters*, 23(4), 319-322.
- [Wu and Lundstedt, 1997] - Wu, J.-G., and H. Lundstedt (1997), Geomagnetic storm predictions from solar wind data with the use of dynamic neural networks, *J. Geophys. Res.*, 102, 14255–14268, doi:10.1029/97JA00975.
- [Xu and Shi, 1986] - Xu, W.-y., and E.-Q. Shi (1986), Numerical examination of Akasofu's energy coupling function, *Chin. J. Space Sci.*, 6(1), 24–3232.

PUBLICATIONS

Page 205 – 246 Prediction of the geomagnetic index am based on the development and the performance comparisons of static and dynamic Neural Networks, M. A. Gruet, N. Bartoli, S. Rochel, R. Benacquista, A. Sicard and G. Rolland, en révision au Space Weather Space Climate

Page 247 – 262 Multiple-Hour-Ahead forecast of the Dst index using a combination of Long Short-Term Memory Neural Network and Gaussian Process, M.A. Gruet, M. Chandorkar, A. Sicard, E. Camporeale, accepté au Space Weather.

1 Prediction of the geomagnetic index *am* based on the performance
2 comparisons of static and dynamic Neural Networks

3 M. A. Gruet¹, N. Bartoli¹, S. Rochel¹, R. Benacquista¹, A. Sicard¹ and G. Rolland²
4

5 ¹ONERA, The French Aerospace Lab, 2 avenue Edouard Belin,
6 31400 Toulouse, France

7 ²CNES, 18 avenue Edouard Belin, 31400 Toulouse, France
8

9 Corresponding author:

10 Marina Gruet (marina.gruet@onera.fr)
11
12
13

14 Key points 15

- 16 ● First Neural Network based predictions of the geomagnetic index *am*
- 17 ● Time Delay Neural Network relying on only solar wind parameter
18 inputs is a performant solution to predict the geomagnetic index *am*
- 19 ● Non linear AutoRegressive with eXogenous inputs Neural Network
20 present the best performance to predict the geomagnetic index *am*
21

22 Abstract 23

24 In the domain of Space Weather, Neural Networks are widely used to
25 predict the geomagnetic activity. Previous studies have shown that
26 depending on the predicted geomagnetic index (Kp, Dst, AE...), the solar
27 wind data considered (density n, flow speed V, IMF B) and the expected
28 forecast (1h, 3h, 4h...), specific structures for the Neural Network should be
29 favored in order to obtain the best performance of predictions. This article
30 focuses on the prediction of the global geomagnetic index *am* using Neural
31 Networks. Multiple structures of Neural Network exist. Three models were
32 considered in the present study: the feedforward backpropagation neural
33 network, the time delay neural network and the non-linear autoregressive
34 with exogenous inputs neural network. Two databases were used to train

35 those networks. First we used propagated solar wind data from the OMNI
36 database. Secondly we used measurements at Lagrangian point L1 based on
37 records from the Advanced Composition Explorer satellite. We then
38 compared their performance in order to find the one best suited for the
39 prediction of the global geomagnetic index am depending on operational
40 constraints. In all cases, the non-linear autoregressive with exogenous inputs
41 neural network offers the best performance in terms of global root mean
42 square error and correlation coefficient, but also in terms of Probability Of
43 Detection and False Alarm Rate at all levels of activity. If the user is
44 required to not use the geomagnetic index history, the time delay neural
45 network is an adapted solution, relying on only solar wind parameters and
46 providing him with good prediction performance.

47 1. Introduction

48
49
50 Recently, in the domain of Space Weather, efforts have been done to better
51 predict the effect of the solar activity on human technologies. As far as we
52 know, solar wind particles interact with the geomagnetic field of the Earth
53 and create electric currents in the magnetosphere. These currents produce
54 geomagnetic disturbances and are measured by magnetometers on the
55 ground. Thanks to the observation of disturbances along the horizontal
56 component, it is possible to estimate the magnetospheric activity. This
57 provides information for better understanding how our technologies can be
58 impacted. The need of more and more accurate predictions of geomagnetic
59 indices increases as Space Weather becomes critical in various domains
60 such as satellites instrumentation, communication and power grids.

61
62 Introduced by Bartels (1938), the K indices were the first indices describing
63 quantitatively the geomagnetic activity. They were then used to provide
64 integrated information, particularly about the magnetosphere behavior. The
65 most known of those indices is the global 3-hour magnetospheric index Kp
66 (Bartels, 1949). It is evaluated using K indices from 13 stations at the
67 subauroral zones. But a need for indices providing better spatial and
68 temporal description of the geomagnetic activity emerged. Thanks to
69 improvements in the computing of worldwide geomagnetic activity, and the

70 increasing number of observatories all over the world, a new derivation
71 process has been developed and led to new K indices (Mayaud, 1968). One
72 of them was the 3-hour geomagnetic index *am*. This geomagnetic index is
73 based on a network of 30 stations evenly spaced in longitude and latitude at
74 subauroral zones.

75
76 To connect geomagnetic indices and solar wind parameters, particularly the
77 solar wind density, the flow speed and the magnetic field, various methods
78 have been used such as linear filters (Bargatze, 1985). Those linear filters
79 demonstrated that in order to make accurate predictions, component of the
80 geomagnetic activity not directly driven by the solar wind had to be taken
81 into account. The magnetosphere has a nonlinear response to solar activity.
82 It is impacted by other mechanisms such as the loading-unloading behavior
83 of the magnetosphere (Russell and McPherron, 1973). **To model this**
84 **nonlinear behavior, Neural Networks (NNs) are used** (Gleisner, and
85 Lundstedt, 1997; Gleisner and Lundstedt, 2001). The NN approach to time
86 series is non parametric in the sense that it is not necessary to know any
87 information regarding the process that generates the signal. **As the processes**
88 **leading to magnetospheric activity are complex and not always fully**
89 **understood, NNs** find their places in the space weather community.

90 Different structures of NN (static, temporal or recurrent) have proved their
91 efficiency to predict **for geomagnetic indices and the auroral activity**.

92 In the Space Weather community, static NNs like the feedforward
93 backpropagation NNs were the first used. **Lundstedt and Wintoft (1994)**
94 **used** them to predict geomagnetic storms with the Dst index. Gleisner and
95 Lundstedt (1997) used them to predict the response of auroral electrojets to
96 solar activity. To predict the global geomagnetic activity, Boberg et al.
97 (2000) developed real time predictions of the Kp index using this NN.

98 Temporal networks are also powerful models. One of them, the Time Delay
99 Neural Network (TDNN), is based on time lagged solar wind data and has
100 also already proved its efficiency. One developed by Gleisner et al. (1996)
101 intends to predict the geomagnetic index Dst. With this model, they already
102 improved the performance of prediction in comparison to the performance

103 of feedforward NN. The TDNN was able to fit the structure of a
104 geomagnetic storm, like the recovery phase.

105 Recurrent networks, like Elman recurrent networks are structures in which
106 inputs nodes are fed back by the output of hidden nodes. They were used in
107 several models and provided accurate predictions. For example, the ring
108 current activity has been described using Elman recurrent network (Wu and
109 Lundstedt, 1996). Gleisner and Lundstedt (2001) developed this network to
110 predict the auroral activity. Wing et al. (2005) developed three dynamic
111 models for the prediction of Kp based on real time solar wind data recorded
112 by the Advanced Composition Explorer (ACE) satellite.

113
114 More recently, a specific recurrent network has been used in Space Weather,
115 the NARX NN. It offers great performance as it considers history of input
116 parameters and autoregressive output of the network for the prediction. It
117 was used to improve the understanding of the ring current activity with the
118 geomagnetic index SYM-H (Cai et al, 2009), to predict SYM-H and ASYH
119 (Bhaskar and Vichare, 2017) and to predict the global geomagnetic activity
120 with Kp (Ayala Solares et al, 2016).

121
122 Here, we will consider three models which already proved their efficiency
123 for the prediction of geomagnetic indices: the multilayer feedforward NN,
124 the Time Delay NN and the NARX NN. The aim of this study is to find the
125 NN structure which will offer the best predictions of the geomagnetic index
126 *am*, based on solar wind data taken either from the OMNI database or on
127 data recorded by ACE satellite at the Lagrangian point 1 (L1).

128
129 This paper is structured as follows: section 2 describes data used in this
130 study, section 3 presents NN developed for the prediction of *am*. Section 4
131 provides the results of NN performance and some comparisons between the
132 different considered NN structures. Section 5 discusses those performance
133 and opens to new techniques for future needs.

134 2. Data set description

135
136

137 Some descriptions of solar wind data and geomagnetic index will be given
138 in the following sections.

139
140
141

2.1. Solar wind data

142 Two types of solar wind data are considered, corresponding to two locations
143 in space between the Sun and Earth:

144

145 - At the Lagrangian Point L1. Data are provided by the Advanced
146 Composition Explorer (ACE) satellite (Stone et al. 1998) (available
147 on the website
148 <http://www.srl.caltech.edu/ACE/ASC/level2/index.html> maintained
149 by the Ace Science Center (ASC) hosted by the Space Radiation Lab
150 at California Institute of Technology (SRL Caltech)). Those are ACE
151 Level 2 data, which means that they are appropriate for serious
152 scientific study as explained on the website. Data provided by the
153 ACE satellite straight from the L1 point are taken from the 5th of
154 February 1998 to the 31st of December 2010, covering the solar
155 cycle 23. We use hourly averages data proposed by ASC. As Wing
156 et al. (2005) underlined, those estimations are rough estimates as the
157 solar wind at L1 and the location of the solar wind monitor may
158 vary. Data from the ACE satellite have already been used (Wing et
159 al. 2005; Bala and Reiff, 2012) and more recently by Wintoft et al.
160 (2017).

161 - At the bow shock. Data come from the OMNI database, maintained
162 by the National Space Science Data Center (NSSDC) of National
163 Aeronautics and Space Administration (NASA)
164 (<https://omniweb.gsfc.nasa.gov/ow.html>). Those data have already
165 been used in several studies (Watari, 2011). Data are taken from the
166 1st of January 1995 to the 31st of December 2014. Those data are
167 initially defined every minute, but we consider here hourly data.

168

2.2. Geomagnetic index

169 The geomagnetic index we intend to predict in this study is *am* (Mayaud,
170 1980). It comes from the database provided by the International Service of
171

172 Geomagnetic Indices (ISGI) hosted by the Ecole des Sciences de la Terre
173 (EOST) (<http://isgi.unistra.fr/>).

174
175 Global geomagnetic indices were the first to be defined. **The most widely**
176 **used is the global geomagnetic index Kp** (Bartels, 1949; Mayaud, 1980). Kp
177 is a 3-hour index, computed since 1932 thanks to 13 stations situated close
178 to the subauroral zone, as one can see on Figure 1. It has no unit and is
179 defined on a logarithmic scale from 0 to 9, value over which it is capped. In
180 addition, Figure 1 shows that most of the stations constituting the Kp
181 network are located in the northern hemisphere, so that the final measure of
182 Kp is heavily weighted toward the north. As a consequence of this
183 distribution, some geomagnetic events might be underestimated like
184 magnetic storms which happen over or below those locations, or at different
185 longitudes. As Menvielle and Berthelier (1991) underlined, the Kp index is
186 a « pioneer », so it is important in the history of geomagnetism and is widely
187 used nowadays. Mayaud (1968) introduced *am*, a new 3-hour geomagnetic
188 index, which provides a better definition and derivation of indices
189 monitoring the worldwide geomagnetic activity. To compute this index,
190 variations of the magnetic field are estimated for each of the 30 stations of
191 the network, to define the index *an* for the northern hemisphere and *as* for
192 the southern one. Then those values are averaged to obtain the final value of
193 *am* in nanoTesla, which contrary to Kp is not capped. As explained by
194 Menvielle and Berthelier(1991), in statistical studies, *am* is the best choice
195 as it is a more reliable witness of the geomagnetic activity than Kp. Thanks
196 to its extended and better distributed network, it ensures the same weight for
197 the different longitude sectors. Average amplitudes are more precise than
198 logarithmic classes. The geomagnetic index *am* offers a better detection of
199 localized phenomena and the best measurement for high levels of
200 disturbances.

201 In this study, we aim to predict *am* one hour ahead. Therefore, a spline
202 interpolation is done to obtain a 1-Hour index from the 3-Hour index.

203
204 **For all these reasons, the prediction of the global geomagnetic index *am* is**
205 **the key purpose of this paper. NNs developed to predict this index are**
206 **presented in following sections, as well as the optimization of its forecast.**

207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224

225
226
227
228
229
230
231
232

233
234
235
236
237

3. Method: Neural Networks

To predict geomagnetic indices, NNs are widely used. They are mathematical models which aim to fuse the learning capability of the human brain and computing performance of machines. In order to reproduce this biological functioning using artificial computer system, each biologic element of the biological neuron is programmed in terms of mathematical functions (McCulloch and Pitts, 1943). It is called the formal neuron, or node, represented by colored circles on Figures 2.a, 2.b, 2.c. The topology of a NN is defined by the organization of nodes in terms of layers, meaning its architecture, and the connection between these layers.

Depending on their structure, NN work like linear and nonlinear filters. Gleisner and Lundstedt (2001), defined that the geomagnetic activity measured by geomagnetic indices can be derived from time-lagged solar wind inputs. Here we take as an example the geomagnetic index am (see section 2 for a complete description of this index):

$$am = F(I_{t-m}) \tag{1}$$

with I_{t-m} a vector containing temporal information of solar wind inputs within a certain length m . Different values of m will be considered in the test cases. After a study based on different solar wind inputs to optimize NN performance, the three parameters considered in this article are the density n , the flow speed fs and the average amplitude of the Interplanetary Magnetic Field IMF $|B|$. The geomagnetic index am can also be computed from both time-lagged solar wind inputs and prior geomagnetic activity:

$$am = F(I_{t-m}, am_{t-n}) \tag{2}$$

with am_{t-n} a vector storing past geomagnetic activity observed during the last n hours and I_{t-m} a vector containing temporal information of solar wind inputs within a certain length m .

238 For a dynamic NN, the function F might be locally linear or non-linear
239 depending on the structure of the NN. This is a strong property of dynamic
240 NN because as Bargatze et al. (1985) showed when studying the link
241 between solar wind parameters and the auroral AL index, the response of the
242 magnetosphere is not completely linear. Vassiliadis et al. (1995) underlined
243 that a linear filter can be used if the filter is changed according to the
244 activity level, which is inadequate for practical prediction. So a non-linear
245 filter is more appropriate to describe the magnetospheric response. Sharma
246 (1995) explained that the required use of non-linear filters is due to the dual
247 behavior of the magnetosphere. There are time intervals during which the
248 magnetospheric behavior is dominated by solar wind drivers, and others
249 time intervals during which loading and unloading magnetospheric
250 mechanisms drive it. Russell and McPherron (1973) also demonstrated that
251 all the geomagnetic activity cannot be only driven by the solar wind.

252
253 According to this, NNs thanks to their non-linear specificities already found
254 their applications in the Space Weather domain. The aim here is to compare
255 their performance to predict the magnetic index am . We developed, trained
256 and compared three different NNs with different values for their respective
257 input parameter $sets$.

258 259 3.1. The static NN : the feed forward backpropagation NN

260
261 First we focused on a NN which already demonstrated great performance to
262 predict the magnetic activity, the standard multilayered feed-forward
263 backpropagation NN (Rumelhart and McClelland, 1988). It is also called
264 multilayer perceptron or MLP, according to its combination of perceptrons.
265 This is the simplest architecture, with one input layer represented by blue
266 circles on the Figure 2.a, connected to one hidden layer, on pink on the same
267 figure. This hidden layer produces the network's output. Each layer's node
268 is the weighted sum of the outputs coming from all the nodes in the previous
269 layer. The output of a node is given by different important elements: inputs
270 of the node, the nodal activation function which is a differentiable saturating
271 function, the bias and weights. The weights are the key to the network

272 performance; it determines how important the information for the node is.
273 We describe in Section 3.4 how they are computed.

274
275 As mentioned in Gleisner and Lundstedt (2001) and in the equation (2), the
276 geomagnetic activity can be described as a function of a vector of time
277 lagged solar wind inputs and by prior geomagnetic activity. Here, the input
278 vector x_i contains solar wind data like density (n), flow speed (fs) and the
279 Interplanetary Magnetic Field magnitude (IMF |B|). Depending on the NN
280 configuration this input vector can also take into account the nowcast
281 geomagnetic index *am*.

282
283 Outputs of such a network come from the equation (2) and can be written as:
284

$$\hat{y}_k = \sum_{j=1}^{N_{hidden}} (W_{jk} \tanh(\sum_{i=0}^{N_{input}} W_{ji} x_i)) \quad (3)$$

285
286 Index i refers to an input layer node, index j refers to a hidden layer node
287 and index k refers to an output layer node. The activation function of nodes
288 of the input layer and the output layer is a linear function. The activation
289 function of nodes of the hidden layer is a tangent hyperbolic function (tanh).
290 Weights W_{ji} connect input and hidden layer nodes, and W_{kj} is the connection
291 between hidden layer nodes and the output node. The way to compute the
292 number of hidden nodes N_{hidden} and the number of input nodes N_{input} is
293 described in Section 3.4. These notations will be kept for subsequent
294 equations.

295 In the present study, the output vector consists of a single value \hat{y} , which is
296 the predicted geomagnetic index.

297 So we can rewrite equation (3) as:

$$\hat{y} = \sum_{j=1}^{N_{hidden}} (W_j \tanh(\sum_{i=0}^{N_{input}} W_{ji} x_i)) \quad (4)$$

298 The information goes from the input to the output. This is a static structure.
299 The dynamic is given by the organization of the input data. This pushed us
300 to consider a temporal NN described in the following section.

301
302 3.2. The temporal NN : the Time Delay NN

303
304 The TDNN is a model widely used for speech recognition (Waibel et al.,
305 1989), image sequence analysis (Wöhler and Anlauf, 1999) and temporal

306 pattern processing (Wang, 2003). To keep a structure similar to the
 307 multilayer perceptron, one hidden layer is kept and is fed by one input layer,
 308 and one temporal layer call “the window of specialization”. This layer is
 309 also connected to the input layer with a time delay τ_d . The delay is also seen
 310 in the weights of the specialization window. The length of the time delay
 311 defines the length of the « window of specialization » represented in green
 312 on the Figure 2.b. and defining the vector x_l . As described in Peddinti et al.
 313 (2015), the TDNN uses shared weights. These weights correspond to
 314 weights which would have the same value for connections between neurons
 315 defined at the same time t_i . This way, the NN has the ability to deal with
 316 time translational invariance, and with fast transitions of the input signal,
 317 while slower variations are taken into account by the window of
 318 specialization.

319 We can consider the TDNN has a multilayer perceptron-type structure using
 320 shared weights. For the TDNN, input vectors x_i and x_l contain only solar
 321 wind data like density (n), flow speed (fs) and the IMF magnitude (IMF |B|).
 322 This is the same as described in equation (1). Our aim is to see how this NN
 323 can perform using exogenous information, without taking into account the
 324 geomagnetic index as an input.

325 The output of this network is linked to equation (1) and can be written as

$$326 \hat{y} = \sum_{j=1}^{N_{hidden}} (W_j \tanh(\sum_{i=0}^{N_{input}} W_{ji} x_i + \sum_{l=0}^{N_{window}} W_{lj} x_l)) \quad (5)$$

327
 328 Index l refers to the delay layer also called window of specialization which
 329 contains time series data x_l . W_{lj} is the connection between hidden layer and
 330 input layer with time delays. The way to define the number of nodes in the
 331 window of specialization N_{window} is described in Section 3.4.

332 3.3. The recurrent NN : the Non-linear autoregressive with exogenous input NN

333 The Nonlinear Autoregressive with eXogenous input (NARX) model
 334 (Leontaritis and Billings, 1985), therefore called NARX recurrent NN, is a
 335 powerful class of models which has been demonstrated to be well suited for
 336 modeling nonlinear systems and specially time series (Haykin, 1999; Lin et
 337 al., 1996; Gao and Er, 2005). The NARX NNs have shown in the past to be

340 as powerful as Turing machines and even demonstrated super-Turing
341 capabilities (Kilian et al., 1993; Siegelmann and Sontag, 1991, 1994, 1995).

342
343 It has been shown that gradient descending learning algorithms (see Section
344 2.4) is more effective in NARX networks than in other NNs and that they
345 converge faster and generalize better than other NNs (Lin et al., 1996 ; Gao
346 and Er, 2005).

347
348 The NARX model can be implemented in many ways, but the easiest seems
349 to be based on a multilayer perceptron–type structure, fed by different pieces
350 of information coming from two different input layers. On one hand, there is
351 an embedded memory represented by the blue input layer on the Figure 2.c.
352 which takes into account solar wind parameters represented by the vector x_i .
353 On the other hand, there is a delayed connection from the output of the
354 network to the orange input layer seen on the same figure which takes into
355 account the nowcast index represented by the vector x_m in equation (6).
356 Output of such a network is linked to equation (2) and can be written as

$$\hat{y} = \sum_{j=1}^{N_{hidden}} (W_j \tanh(\sum_{i=0}^{N_{input_i}} W_{ji} x_i + \sum_{m=0}^{N_{input_m}} W_{kmj} f_m x_m)) \quad (6)$$

357 Index m refers to the input layer connected to the output layer, W_{kmj}
358 represents the connection between this layer and the output layer, and f_m is
359 a nonlinear, differentiable, saturating function. The way to define the
360 number of input nodes containing solar wind parameters N_{input_i} and the
361 number of input nodes containing outputs of the network N_{input_m} is
362 described in Section 3.4.

363 364 3.4. Training and validation of NN

365
366 In Section 3.1, we saw that in order to define a NN, it is required to optimize
367 its weights. As NNs presented in this paper belong to the type of supervised
368 NNs, we train them using the NN Matlab toolbox with a Levenberg
369 Marquardt descent gradient method (Levenberg, 1944; Marquardt, 1963) to
370 minimize the root mean square error between the true value of the
371 geomagnetic index y and the estimated value \hat{y} . The root mean square error
372 can be computed as

$$RMSE = \sqrt{\frac{1}{N_y} \sum_{i=1}^{N_y} (\hat{y}_i - y_i)^2} \quad (7)$$

373 With N_y the size of the database, \hat{y}_i the predicted value and y_i the
 374 observed value. This notation will be kept in subsequent equations.

375
 376 To develop the network, the database is randomly divided into 3 sets, 70%
 377 for the training set, 20% for the test set and 10% for the validation set.
 378 Friedam et al. (2001) highlighted the difficulty to give a general rule for the
 379 partitioning ratio (70%,20%,10%), it depends mostly on the size of the
 380 database and on the data distribution. Here, a partition similar to the one
 381 used in studies quoted before has been considered. An exception to the
 382 random division is done for the July 2004 event that we decided to keep in
 383 the validation set and will be presented in section 4.4.

384 As various structures are developed and studied in this paper, the optimal
 385 number of hidden nodes was determined empirically to minimize the root
 386 mean square error. We varied the number of hidden nodes between 10 and
 387 60 for each NN. Overall, the optimal number of nodes was around 20.

388
 389 Table 1 shows the root mean square error (RMSE) and the correlation
 390 coefficient (CC) for the feedforward backpropagation NN, depending on the
 391 size of the solar wind dataset. The CC is estimated using equation (8) to
 392 estimate the similarities between time series of the observed *am* index, and
 393 the predicted one.

$$CC = \frac{\sum_{b=1}^{N_y} (y_b - \bar{y}_l) (\hat{y}_b - \bar{\hat{y}}_l)}{\sqrt{\sum_{b=1}^{N_y} (y_b - \bar{y}_l)^2 - \sum_{b=1}^{N_y} (\hat{y}_b - \bar{\hat{y}}_l)^2}} \quad (8)$$

394 With \bar{y}_l the average of the observed value and $\bar{\hat{y}}_l$ the average of the
 395 predicted value

396 One can observe that the lowest RMSE is reached for 6 hours of history data
 397 and is equal to 4.28 nT, while the highest CC is reached for 12 hours history
 398 data and is equal to 0.983. When Boberg et al. (2000) developed models to
 399 predict Kp, they also found that predictions were optimized when using 6

400 hours of history data with a RMSE of 1.01 and a CC of 0.66 during storm
401 times.

402
403 In Table 2, NN RMSE and CC using 6 hours of data are presented using
404 whether the OMNI or ACE databases. In all cases, the NARX NN provides
405 the lowest RMSE with a value of 3.32 nT using the OMNI database and
406 3.65 nT using the ACE database. It also provides the highest CC with a
407 value of 0.989 using OMNI and 0.988 using the ACE database.

408
409 **The contribution of the TDNN** can be seen in this Table 2. We said in
410 Section 3.2 that the aim of the TDNN is to use only solar wind parameters as
411 an input. We could have done so by developing the feedforward
412 backpropagation NN without using the nowcast index as an input, but as it is
413 shown in Table 2, the quality of the feedforward NN without the nowcast
414 index is worse than the one of the TDNN. For example, if we compare
415 RMSE and CC using the ACE database, the RMSE and the CC of the
416 feedforward NN without the nowcast index are respectively equal to 14.1
417 nT and 0.766, while for the TDNN they are equal to 6.85 nT and 0.958.

418
419 These global metrics are of interest to optimize the development of NN and
420 to have a global overview of their performance. To go deeper and extract
421 more information concerning the ability of NN to predict the geomagnetic
422 index *am*, we decided to compare performance using metrics described in
423 Section 3.5, defined for various thresholds of activities.

424

425 3.5. Comparison of Neural Networks performance

426

427 In Section 3.4, we measured the performance **of the trained NNs** by using
428 the RMSE and the CC between the outputs and targets **of the NNs** on the
429 training, test and validation sets. Here the term forecast means a prediction
430 of the future state of the magnetosphere, then forecast verification is the
431 process of assessing the quality of the forecast of it. The three most
432 important reasons to verify forecasts are to monitor the quality of the
433 forecast, to improve it, and compare the quality of different forecast
434 systems. To assess the quality of a forecast of a time series, various

435 verification methods exist such as multi-category, continuous, probabilistic
436 and dichotomous methods. Here we will consider dichotomous forecasts. It
437 says if an event will happen or not. As the impact of solar events on the
438 Earth's environment is not the same depending on the level of am index, it is
439 of interest to specify different thresholds. These thresholds are defined in
440 Table 3. It helps to estimate the NN performance dependency with the solar
441 and geomagnetic activity.

442
443 We use two parameters called the Probability Of Detection (POD) and the
444 False Alarm Rate (FAR). Those parameters are extracted from a
445 contingency table (Jolliffe and Stephenson, 2003). It is a useful way to see
446 which types of errors are being made. The POD is a number between 0 and
447 1, 1 being for the perfect score. It is described by

$$POD = \frac{H}{H + M} \quad (9)$$

448
449 Where H stands for “hits” (i.e. all correct positive-forecasts) and M for
450 “missed forecasts” (i.e. all incorrect negative forecasts). This notation
451 will be kept in subsequent equations.

452 It shows the fraction of observed geomagnetic events which were correctly
453 forecast. It should be used in conjunction with the FAR. The FAR is also a
454 number between 0 and 1, but here the perfect score is 0. It represents the
455 fraction of predicted geomagnetic events which did not occur. It is described
456 by

$$FAR = \frac{F}{H + F} \quad (10)$$

457 with F stands for “false alarms” (i.e. all incorrect positive-forecasts).

458 459 4. Results

460
461 To compare NNs performance, we observed variations of the POD and the
462 FAR versus the level of activity. As Gleisner and Lundstedt (1997)
463 underlined, a NN should be able to help improving our understanding of the
464 coupling between solar wind parameters and geomagnetic data. In Section

465 3.4, NN with various lengths of the solar wind history have been developed
466 and the CC and RMSE dependencies with this length have been discussed.
467 Thanks to this simple study, we can already improve our knowledge
468 concerning dissipation processes leading to magnetospheric disturbances
469 recorded by the geomagnetic index am . It raises the question of the stability
470 of the results due to the change of NN. But once the optimal length of solar
471 wind data history is found for a specific NN, varying its architecture gives
472 us important information about the influence of the architecture on the
473 modeling of the magnetospheric response to solar wind activity.
474 First, a study of the optimal length of solar wind data required to obtain the
475 highest POD and the lowest FAR with the most simple NN structure
476 presented in Section 3.1 is presented. It is for the multilayer feedforward
477 NN. Once the optimal time lag is found for the simplest NN, the same lag is
478 used for the other NN. Performance of the different NN using OMNI data or
479 data recorded at L1 by ACE are then discussed.

480 4.1. Looking for the optimal length of solar wind data history 481 using OMNI propagated data

482
483 As am is a global index, the multilayer feedforward NN is the first
484 considered, like did Boberg et al. (2000) who obtained interesting
485 correlations between expected Kp and the official one. One of the most
486 complicated parts of the development of a NN is to find the length of the
487 time sequences which has to be taken as an input. Figure 3 shows that
488 depending on the data history considered, the POD and FAR depend on the
489 level of activity. The main problem when training a NN is to be able to
490 reach great performance even at high level of activity. Most of the time
491 (99.703% of the data), the geomagnetic index am is below 150 nT, while
492 0.0487% of am values are over 300 nT. So NN does not have as much
493 training examples for high activity as for low activity, and the training is the
494 key of the NN performance.

495
496 The most interesting for an operator is to know when the solar activity will
497 have an important impact on the geomagnetic environment. We decided to
498 focus on POD and FAR obtained for the highest level of activity (over 300

499 nT) to compare the NNs. At this level, one can observe on Figure 3 that the
500 highest POD is equal to 0.794 (FAR = 0.122) corresponding to a data
501 history of 6 hours. As we explained in Section 3.4, Boberg et al. (2000) did
502 this observation during storm periods. We also observed in Section 3.4 that
503 the lowest RMSE is reached for this length of solar wind parameters. It is
504 also interesting to consider the data history for which the lowest FAR is
505 reached. This one corresponds to a data history of 12 hours, with a FAR of
506 0.047, around three times lower than the one obtained with 6 hours of data
507 history. But the POD decreases from 0.794 to 0.700. So the user would have
508 to make a compromise between choosing the multilayer feedforward NN
509 which provides the highest probability of predicting « real » events or the
510 lowest probability of predicting « fake » events.

511 One can observe that as the data history increases over 12 hours, the POD
512 decreases and stabilizes at 0.651 and the FAR increases at 0.0889. So, in
513 order to compare performance of various NN, we will work on POD and
514 FAR obtained with 6 hours and 12 hours of data history.

515 516 4.2. Comparison of NN performance using OMNI propagated data

517
518 The Figure 4 presents the results obtained with different NN structures when
519 various levels of activity are considered. As Wing et al. (2005) underlined
520 when describing the APL3 model, which takes as an input only solar wind
521 data, sometimes the Kp index is not available or reliable and it is the same
522 for *am*. Wing et al.(2005) then developed a recurrent model which took as
523 inputs only solar wind data (density n , flow speed $|V_x|$, IMF $|B|$ and B_z).
524 This model offered great predictions of Kp one hour ahead with a
525 correlation coefficient between predicted Kp and observed Kp of 0.84. This
526 is why we were interested in developing a dynamic model, to predict *am* by
527 using only solar wind data. We chose the TDNN which already showed
528 great performance for the prediction of the geomagnetic index Dst using
529 only solar wind data, as done by Gleisner et al. (1996), with the correlation
530 coefficient between predicted Dst and observed Dst reaching 0.92. We also
531 showed in the section 3.4 with the Table 2 that the performance of the
532 TDNN are higher than the one of the feedforward backpropagation NN
533 based on only solar wind parameters.

534
535 As observed on Figure 4, at low and medium activities ($am < 120$ nT), the
536 TDNN offers performance similar to the multilayer feedforward NN but as
537 the activity increases ($am > 120$ nT), the TDNN shows higher. If we consider
538 the highest level of activity (over 300 nT), for 6 hours of history data, the
539 TDNN has a POD of 0.810 and a FAR of 0.089 whereas for 12 hours of
540 history data, it has a POD of 0.857 and a FAR of 0.100. So the probability of
541 detecting « real » important geomagnetic event is higher using 12 hours of
542 history data, but there is also slightly more risks to predict « fake »
543 geomagnetic events.

544
545 In both cases, considering 6 or 12 hours of history data, the MLP is the
546 structure which shows the lowest performance at all levels of activity and
547 most of the time the highest FAR. On the contrary, the NARX recurrent NN
548 shows the highest POD (between 0.939 and 0.984) at all levels and the
549 lowest FAR (between 0.013 and 0.054) at all levels. We will discuss about
550 these differences in Section 5.

551 552 4.3. Comparison of NN performance using L1 data 553

554 After comparing performance of networks, we decided to train them using
555 data recorded at the L1 point. As described in Section 3.1, solar wind data
556 are provided by the ACE satellite. Costello (1998) developed a feedforward
557 backpropagation NN model based on ACE solar wind speed, IMF Bz and |B|
558 to predict Kp one hour ahead. Wing et al. (2005) also developed 3 models
559 based on ACE solar wind speed $|V_x|$, density n, IMF |B|, Bz and
560 depending on the model, nowcast Kp
561 (<http://www.swpc.noaa.gov/products/wing-kp>). Thus it is possible to make
562 short term forecasts based on ACE data.

563
564 Figure 5 presents the results of the different NN structures as a function of
565 the levels of activity when L1 data are used. By comparing Figure 4 and
566 Figure 5, we can observe that performance of NN vary depending on the
567 data location, at the bow shock of the magnetopause (OMNI) or L1 (ACE).
568 Using L1 data globally decreases the performance of each NN at all levels

569 of activity. This is mostly due to the fact that when the ACE database is
570 used, we consider parameters which are provided at the L1 point, so when
571 there is an important solar event, detectors onboard the satellite are
572 saturated, and there is a lot of missing data, even more at high levels of
573 activity. NASA scientists try to handle as much as possible those missing
574 data by using data provided by other satellites. This way, the OMNI
575 database contains more information to train the network at high levels of
576 activity, which optimizes the training phase and then positively impacts the
577 performance of NN.

578
579 When data are taken at the L1 point, we can also observe that the
580 feedforward NN offers better performance than a temporal structure like the
581 TDNN. If we consider performance obtained with 12 hours of solar wind
582 data at the highest level of activity ($am > 300$ nT), the POD of the
583 feedforward NN is 0.816 (FAR = 0.091) whereas the POD of the TDNN is
584 equal to 0.776 (FAR = 0.156). As a reminder, the TDNN takes as inputs
585 only solar wind data contrary to the feedforward backpropagation NN which
586 takes also as an input the nowcast geomagnetic index. It therefore seems that
587 if one wants to obtain performance as good as what the feedforward
588 backpropagation NN proposes, without using the geomagnetic index as an
589 input and with only L1 data, another structure is required if we do not want
590 to use the geomagnetic index as an input. There is more variability by
591 considering as inputs real time data because as Wing et al. (2005) noted, the
592 location of the monitor varies as the solar wind properties at L1 do, making
593 a time stable measurement at L1 difficult to obtain.

594
595 The NARX recurrent NN still offers great performance with 6 hours of solar
596 wind history, at all levels of activity. But we note that for the highest level
597 of activity, the POD decreases from 0.984 to 0.889 and the FAR increases
598 from 0.031 to 0.051.

599
600 Now that we have done a comparison of NN performance depending on the
601 level of activity, we will observe and compare them on a concrete case. We
602 present in the next section the ability of each NN to predict the event of July
603 2004, using OMNI and ACE database.

604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638

4.4. How do NN perform during geomagnetic storms? Application to the geomagnetic event of July 2004

To better observe and compare NN performance, we decided to confront them with an extreme geomagnetic event which occurred in July 2004. Figure 6 presents different physical characteristics of the solar wind parameters provided by the OMNI database, geomagnetic indices K_p and am provided by ISGI and the integral flux of electrons recorded by NOAA POES satellites from the 1st of July 2004 to the 1st of September 2004. On Figure 6, we can see that high geomagnetic activities were observed. **The solar wind parameters exhibit an important variability during this event. Growing peaks of activity were observed, until reaching a peak of activity corresponding to a severe geomagnetic storm on the 27th of July 2004.** The corresponding variations of am tell more about the real amplitude of the event and its variation than K_p . Indeed, during the event of July 2004, K_p varies between 7 and 9 while am varies between 90 nT and 350 nT. We know that the K_p is defined on a log scale but due to the larger variation of am we have a better indication of how disturbed the environment was. The event of July 2004 also shows that when K_p reaches values of 7, the corresponding values of am can be 120 nT, 160 nT or even 230 nT.

To better understand the impact on the geomagnetic terrestrial environment during this event, the integral flux of electrons is showed when the kinetic energy is higher than 0.300 MeV. These fluxes were recorded by the NOAA POES 15 satellite which orbits at LEO, at an altitude of around 800 km. They are represented as a function of L^* , which is the drift shell parameter (Roederer, 1970). This succession of geomagnetic events generates important electron fluxes between $L^*=6$, meaning when the satellite crosses the magnetic field line connected to geosynchronous orbit, and $L^*=2$ meaning in the inner magnetosphere.

So it would be interesting to see if **NNs** are able to predict such important events and to follow variations of geomagnetic disturbances. On the event of July 2004, am demonstrated important variations, with a first peak on the

639 22nd of July 2004 from 5 nT to 127 nT, a second peak on the 25th of July
640 2004 from 6 nT to 169 nT and a final important on the 27th of July from 3
641 nT to 350 nT. On figures 7 to 10, we focus on the ability of each NN to
642 predict those variations.

643
644 Figure 7 compares predicted values of am for different NN structures using
645 OMNI database. On Figure 7.a. predictions of the feedforward NN show
646 that variations are well predicted but sometimes this NN predicts higher
647 values of am than the real ones. For example instead of predicting a value of
648 223 nT, the network has predicted a value of 268 nT. Figure 7.b., displaying
649 performance of the TDNN, shows worse performance for this kind of event,
650 with for example predicted values of am of 162 nT instead of 226 nT. This
651 is probably due to the fact that this succession of events disturbed
652 progressively the magnetosphere which has a recovery time longer than the
653 event frequencies, as the particle fluxes increase seen on Figure 6 shows it.
654 So as the TDNN does not take into account the geomagnetic index as an
655 input, it does not have any information about the current disturbance of the
656 magnetosphere. The NARX NN, which contrary to the TDNN takes into
657 account as an input the geomagnetic index, predicts better values than the
658 TDNN does, as shown by Figure 7c.

659
660 Figure 8 shows the global performance of NN for this event using OMNI
661 database. This provides a visual understanding of the quality of the
662 prediction depending on the NN. It shows that for this event, the NARX NN
663 offers the best global performance with a RMSE of 5.49 and a CC of 0.981.
664 The TDNN is the less accurate with a RMSE of 10.53 and a CC of 0.925 for
665 this event.

666
667 Figure 9 compares predicted values of am for different NN structures using
668 ACE database. The first thing we can observe is that there are missing data
669 during this event. We made a polynomial interpolation to plot the variation
670 of the real value of am and underline the fact that when using data from
671 ACE database, missing data impact the ability of a NN to make predictions.

672

673 The feedforward backpropagation NN on Figure 9.a. is more accurate than
674 the TDNN on Figure 9.b. which has more difficulties to estimate the
675 variation of the magnetospheric disturbances. We make this observation in
676 Section 4.3 when we compare POD and FAR of NNs using ACE data, and
677 we observe that the TDNN shows some weaknesses relative to others NNs.
678 However, the TDNN predicts better than the feedforward backpropagation
679 NN the highest peak of activity. The real value is of 350 nT and the TDNN
680 predicts 330 nT while the feedforward backpropagation NN predicts 300 nT.

681
682 Figure 10 shows the global performance of NN for this event using ACE
683 database. In this case, the feedforward NN is the most accurate with a CC of
684 0.986 and a RMSE of 6.13. The NARX NN has a lower CC (0.937) and a
685 higher RMSE (9.81), but it is more accurate than other NNs at levels of
686 activity higher than 150 nT.

687
688 Even if there is a global decrease of performance for each NN when using
689 ACE data, the NARX NN predicts well the variation of the geomagnetic
690 activity. With this comparison of performance based on a concrete important
691 solar event, we compared the efficiency of the three networks, with the
692 NARX being the most efficient for both OMNI and ACE databases.

693 694 5. Discussions and Perspectives 695

696 In order to find the optimized structure for the prediction of the geomagnetic
697 index am , three models of NN have been developed. All those models
698 predict am one hour ahead. The first model is a feedforward
699 backpropagation NN which takes as inputs a history of solar wind data and
700 the nowcast geomagnetic index. The second model is a TDNN which takes
701 as inputs only solar wind data. The third model is a NARX NN which takes
702 as inputs vectors of solar wind data and nowcast geomagnetic index.

703
704 For each model, we used a specific length of solar wind data history. To
705 define the optimized length of history data of solar wind parameters we first
706 computed global metrics with the feedforward backpropagation NN which
707 are RMSE and CC. We found that using 6 hours of data history offers the

708 lowest RMSE. We also found that the highest CC is reached with 12 hours
709 of data history. Then this study was done against POD and FAR metrics
710 defined for various thresholds of activities. This study shows that 6 hours of
711 history data offers the best predictions with POD equal to 0.794 when $am >$
712 300 nT. We also decided to consider history data of 12 hours as it
713 corresponds to predictions with the lowest FAR (0.0465) when $am >$ 300
714 nT.

715
716 Then we compared performance of the three NNs using OMNI database and
717 solar wind parameters recorded at the L1 point with ACE. We studied the
718 TDNN with a window of specialization of 6 hours, and 12 hours. We also
719 studied the NARX NN with input vectors of solar wind parameters of 6
720 hours and 12 hours.

721
722 Considering OMNI data with 6 hours of history data, the prediction of am
723 based on the feedforward backpropagation NN is correct, but the NN
724 performance get worse when the geomagnetic activity increases. The POD
725 is the lowest at all levels of activity, and most of the time the FAR is the
726 highest. If we want to take into account the geomagnetic index, it is better to
727 consider a real dynamic network such as the NARX NN which offers the
728 best performance at all levels of activity, both in terms of POD (higher than
729 0.945), and FAR (lower than 0.054). The TDNN shows great performance
730 too even if it only takes into account solar wind data, particularly at high
731 level of activity where its POD becomes greater than the POD of the
732 multilayer feedforward network. But it also shows weaknesses for high
733 activity predictions, because the optimal length of the solar wind input
734 sequence depends on the geomagnetic activity, as detailed by Gleisner and
735 Lundstedt (2001). It has to be large enough to reveal the impact of the solar
736 wind, as it does not take as an input the geomagnetic index. When
737 considering OMNI data, a history data of 12 hours offers globally more
738 accurate predictions. We observed this concretely during the event of July
739 2004, as predicted values are sometimes lower than expected.

740
741 Considering ACE data, the NARX NN remains the structure that offers the
742 best predictions but this time the POD is globally lower than the one

743 obtained with OMNI data. This report is made as well as for 6h as for 12h
744 of history data. Data given afterward as comparison are for 6h of data
745 history. For example, the POD for the highest level activity goes from
746 0.9843 using OMNI data to 0.889 using ACE data. In addition, the FAR is
747 also globally higher. Again, for the highest level of activity, the FAR goes
748 from 0.013 using OMNI data to 0.05 using ACE data. At high level of
749 activity, the TDNN shows great performance with a POD of 0.800 and a
750 FAR of 0.071. Then the feedforward backpropagation NN is quite stable
751 with a POD of 0.727 and a FAR of 0.059 as it takes into account the
752 nowcast index.

753 We demonstrated in this study that no matter the location of data,
754 predictions of the geomagnetic index done by dynamic models still have a
755 POD over 0.7 and proved that these nonlinear filters find their places when
756 describing the link between solar wind inputs and the global geomagnetic
757 index *am*.

759 The NARX NN using OMNI data is the NN which offers the best
760 performance, but it would be of interest to use real time data provided by
761 satellites at L1. For future investigations, it could be interesting to develop
762 new models based only on solar wind parameters, like the TDNN does, but
763 which would offer higher POD and lower FAR than the TDNN. For
764 example, the Long Short Term Memory (Hochreiter and Schmidhuber,
765 1997) seems to be a great option for future developments. It is a recurrent
766 neural network which can rely on only solar wind parameters, and has the
767 ability to predict time series when there are important bounds between
768 events and time variability of unknown size. As the magnetosphere temporal
769 behavior is really variable, these features seem appropriate.

771 Recently, Chambodut et al. (2013) developed $a\sigma$ indices, which are four
772 new indices based on *am* network and calculation, defined with respect to
773 the Magnetic Local Time (MLT). In their article, it is shown that the
774 geomagnetic activity varies a lot depending on the MLT. Those indices
775 could offer precious information for space weather application. It would
776 therefore be interesting to work on NN which would provide multi
777

778 predictions with 4 outputs corresponding to each MLT sector. Then the
779 operator would have a better estimation of the evolution of the event in time
780 and space.

781
782
783
784

Acknowledgments

785 The solar wind plasma data of ACE were provided by the Caltech websites
786 cdaweb site (<http://www.srl.caltech.edu/ACE/ASC/level2/index.html/>). The
787 solar wind plasma data of OMNI were obtained from the National Space
788 Science Data Center (NSSDC) of National Aeronautics and Space
789 Administration (NASA) (<https://omniweb.gsfc.nasa.gov/ow.html>). The
790 results presented in this paper rely on geomagnetic indices calculated and
791 made available by ISGI Collaborating Institutes from data collected at
792 magnetic observatories. We thank the involved national institutes, the
793 INTERMAGNET network and ISGI. This research activity is supported by
794 the Centre National d'Etudes Spatiales (CNES) and the Office Nationale
795 d'Etudes et de Recherches Aérospatiales (ONERA).

796
797
798
799

800
801
802
803
804

805
806

References

807
808
809
810

Ayala Solares, J. R., H. L. Wei, R.J. Boynton, S. N. Walker and S. A. Billings (2016), Modeling and prediction of global magnetic disturbance in near-Earth space: A case study for Kp index using NARX models, *Space Weather*, 14(10), 899-916

811
812 Bala, R., and P. Reiff (2012), Improvements in short-term forecasting of geomagnetic activity,
813 *Space Weather*, 10(6).
814
815 Bargatze, L. F., D. N. Baker, R.L. McPherron and E.W. Hones (1985), Magnetospheric impulse
816 response for many levels of geomagnetic activity, *Journal of Geophysical Research: Space*
817 *Physics*, 90(A7), 6387-6394
818
819 Bartels, A. J. (1938), Potsdamer erdmagnetische Kennziffern: 5. Mitteilung, *Z. Geophys.*,14, 68-
820 78
821
822 Bartels, J. (1949), The standardized index, Ks, and the planetary index, Kp, *IATME Bull. 12B*,
823 97, 2010-2021.
824
825 Bhaskar, A. and G. Vichare (2017), Prediction of SYMH and ASYH indices for geomagnetic
826 storms of solar cycle 24 including recent St Patrick's day, 2015 storm using NARX neural
827 network, *arXiv preprint arXiv::1703.10583*
828
829 Boberg, F., P. Wintoft, and H. Lundstedt (2000), Real time Kp predictions from solar wind data
830 using neural networks, *Physics and Chemistry of the Earth, Part C: Solar, Terrestrial &*
831 *Planetary Science*, 25(4), 275-280.
832
833 Cai, L., S. Ma, H. Cai, Y. Zhou, and R. Liu (2009), Prediction of SYM-H index by NARX neural
834 network from IMF and solar wind data, *Science in China Series E: Technological Sciences*, 52
835 (10), 2877–2885.
836
837 Chambodut, A., A. Marchaudon, M. Menvielle (2013), The K-derived MLT sector geomagnetic indices,
838 *Geophysical Research Letters*,40(18), 4808-4812
839
840 Costello, K. A. (1998), Moving the Rice MSFM into a real-time forecast mode using solar wind driven
841 forecast modules (Doctoral dissertation, Rice University).
842
843 Friedam, J., T. Hastie and R. Tibshirani (2001), The elements of statistical learning, (1), 241-249, New
844 York: Springer series in statistics.
845

846 Gao, Y. and M. J. Er (2005), NARMAX time series model prediction: feedforward and recurrent fuzzy
847 neural network approaches, *Fuzzy sets and systems*, 150(2), 331-350.

848

849 Gleisner, H., H. Lundstedt, and P. Wintoft (1996), Predicting geomagnetic storms from solar-wind data
850 using time-delay neural networks, in *Annales Geophysicae*, 14 (7), 679

851

852 Gleisner, H. and H. Lundstedt (1997), Response of the auroral electrojets to the solar wind modeled
853 with neural networks, *Journal of Geophysical Research: Space Physics*, 102(A7), 14269-14278.

854

855 Gleisner, H., and H. Lundstedt (2001), Auroral electrojet predictions with dynamic neural networks,
856 *Journal of Geophysical Research: Space Physics*, 106(A11), 24541- 24549.

857

858 Haykin, S. (1999), *Neural networks: a comprehensive introduction*, Second Edition

859

860 Hochreiter, S. and J. Schmidhuber (1997), Long Short-term memory, *Neural Computation*, 9(8), 1735-
861 1780

862

863 Jolliffe, I.T., and D. B. Stephenson (2003), *Forecast Verification: A Practitioner's Guide in*
864 *Atmospheric Science*, John Wiley & Sons.

865

866 Kilian, J. and H. T. Siegelmann (1993), On the power of sigmoid neural networks, in *Proceedings of the*
867 *sixth annual conference on Computational learning theory*, 137-143, ACM, Santa Cruz, California,
868 USA, 26-28 July 1993.

869

870 Leontaritis, I. J. and S. A. Billings (1985), Input-output parametric models for non-linear systems part I:
871 deterministic non-linear systems, *International journal of control*, 41(2), 303-328.

872

873 Levenberg, K. (1944), A method for the solution of certain problems in least-squares, *Quarterly Applied*
874 *Mathematics*, 2, 164-168.

875

876 Lin, T., B. G. Horne, P. Tino, and C. L. Giles (1996), Learning long-term dependencies in
877 NARX recurrent neural networks, *IEEE Transactions on Neural Networks*, 7(6), 1329-1338.

878

879 Lundstedt, H., and P. Wintoft (1994), Prediction of geomagnetic storms from solar wind data
880 with the use of a neural network, *Annales Geophysicae*, 12, 19-24

881

882 Marquardt, D. W. (1963), An algorithm for least-squares estimation of nonlinear parameters,
883 *Journal of the society for Industrial and Applied Mathematics*, 11(2), 431- 441.

884
885 Mayaud, P.N, (1968), Indices Kn, Ks, Km, 1964-1967, CNRS, 156p
886
887 Mayaud, P. N. (1980), Derivation, meaning, and use of geomagnetic indices, Washington DC
888 American Geophysical Union Geophysical Monograph Series, 22.
889
890 McCulloch, W. S. and W. Pitts (1943), A logical calculus of the ideas immanent in nervous
891 activity, The bulletin of mathematical biophysics, 5(4), 115-133.
892
893 Menvielle, M. and A. Berthelier (1991), The K-derived planetary indices: Description and
894 availability, Reviews of Geophysics, 29(3), 415-432.
895
896 Peddinti, V., D. Povey and S. Khudanpur (2015), A time delay neural network architecture for
897 efficient modeling of long temporal contexts, Sixteenth Annual Conference of the International
898 Speech Communication Association, Dresden, Germany, 6-10 September 2015.
899
900 Roederer, J.G. (1970), Dynamics of Geomagnetically Trapped Radiation, Springer, New York.
901
902 Rumelhart, D. E., J. L. McClelland, and PDP Research Group (1988), Parallel distributed
903 processing, 1, 433-453.
904
905 Russell, C. T. and R. L. McPherron (1973), The magnetotail and substorms, Space Science
906 Reviews, 15(2), 205-266.
907
908 Surjalal Sharma, A. (1995), Assessing the magnetosphere's nonlinear behavior: its dimension is
909 low, its predictability, high, Reviews of Geophysics, 33(S1), 645-660.
910
911 Siegelman, H. and E. D. Sontag (1991), Turing compatibility with neural nets, Appl. Math.
912 Lett., 4(6), 77-80
913
914 Siegelman, H. and E. D. Sontag (1994), Analog computation via neural networks, Theoretical
915 Computer Science, 131(2), 331-360.
916
917 Siegelmann, H. and E. D. Sontag (1995), On the computational power of neural nets, Journal of
918 computer and system sciences, 50(1), 132-150.
919

920 Stone E., A. Frandsen, R. Mewaldt, E. Christian, D. Margolis, J. Ormes, F. Snow (1998), The
921 Advanced Composition Explorer, *Space Sci Rev*, 86,1-22.
922

923 Vassiliadis, D. A. J. Klimas, D. N. Baker and D.A. Roberts (1995), A description of the solar
924 wind-magnetosphere coupling based on nonlinear filters, *Journal of Geophysical Research:*
925 *Space Physics*, 100(A3), 3495-3512.
926

927 Waibel, A., T. Hanazawa, G. Hinton, K. Shikano and K. J. Lang (1989), Phoneme recognition
928 using time-delay neural networks, *IEEE transactions on acoustics, speech, and signal processing*,
929 37(3), 328-339.
930

931 Wang, D. (2003), Temporal pattern processing. *The handbook of brain theory and neural*
932 *networks*, 1163-1167.
933

934 Watari, S. (2011), Forecast of recurrent geomagnetic storms. *Advances in Space Research*,
935 47(12), 2162-2171.
936

937 Wing, S., J. R. Johnson, J. Jen, C. I. Meng, D. G. Sibeck, K. Bechtold, J. Freeman, K. Costello,
938 M. Balikhin and K. Takahashi (2005), Kp forecast models, *Journal of Geophysical Research:*
939 *Space Physics*,110.
940

941 Wintoft, P., M. Wik, J.Matzka and Y. Shprits (2017), Forecasting Kp from solar wind data: input
942 parameter study using 3-hour averages and 3-hour range values, *Journal of Space Weather and*
943 *Space Climate*, 7, A29.
944

945 Wöhler, C., and J. K. Anlauf (1999), An adaptable time-delay neural-network algorithm for
946 image sequence analysis, *IEEE Transactions on Neural Networks*, 10(6), 1531-1536.
947

948 Wu, J. G, and H. Lundstedt (1996), Prediction of geomagnetic storms from solar wind data using
949 Elman recurrent neural networks, *Geophysical Research Letters*, 23(4), 319- 322.
950
951

952 Table 1. Root mean square error and correlation coefficient depending on the length of solar
 953 wind history for the feedforward backpropagation NN using the OMNI database.

	RMSE (nT)	CC
3h	4.29	0.980
6h	4.28	0.982
9h	4.62	0.982
12h	4.39	0.983
15h	4.96	0.983
18h	4.77	0.979

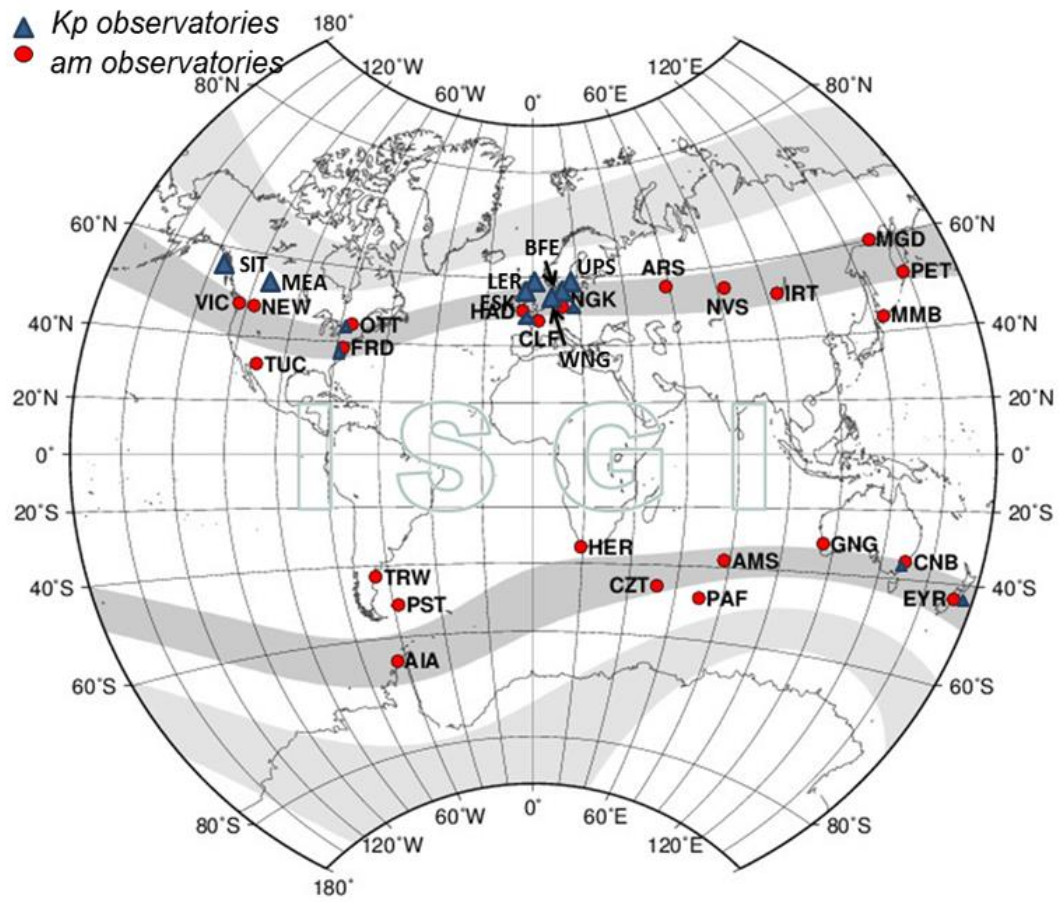
954
 955 Table 2. Root mean square error and correlation using OMNI and ACE data for each NN structure using 6
 956 hours of length of solar wind history

	RMSE (nT)		CC	
	OMNI	ACE	OMNI	ACE
Feedforward backpropagation	4.28	5.20	0.983	0.975
Feedforward backpropagation without nowcast index	10.6	14.1	0.912	0.766
TDNN	8.85	6.85	0.913	0.958
NARX	3.32	3.65	0.989	0.988

957
 958 Table 3. Definition of levels of geomagnetic activity (am) used for the study of NN performance

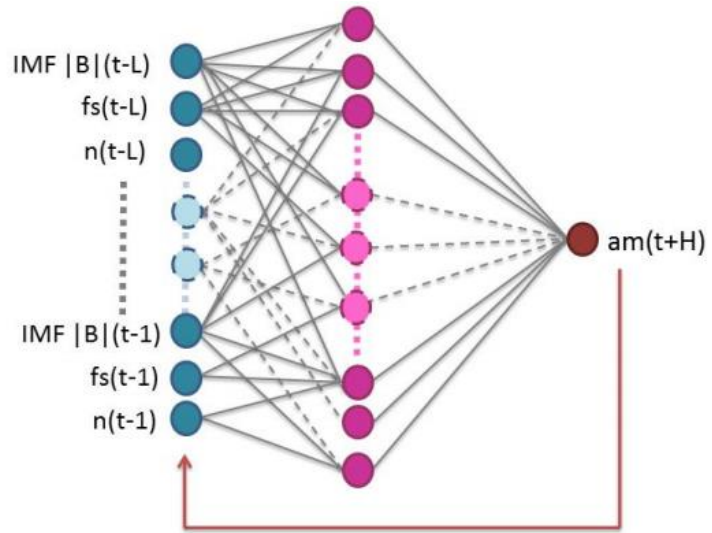
Value of the level of activity on graphs in nT	Domain in nT	Level of activity
7	<20	Very low
20	[20;40]	
40	[40;70]	
70	[70;120]	
120	[120;150]	
150	[150;200]	
200	[200;300]	
300	>300	Very high

959
 960
 961
 962



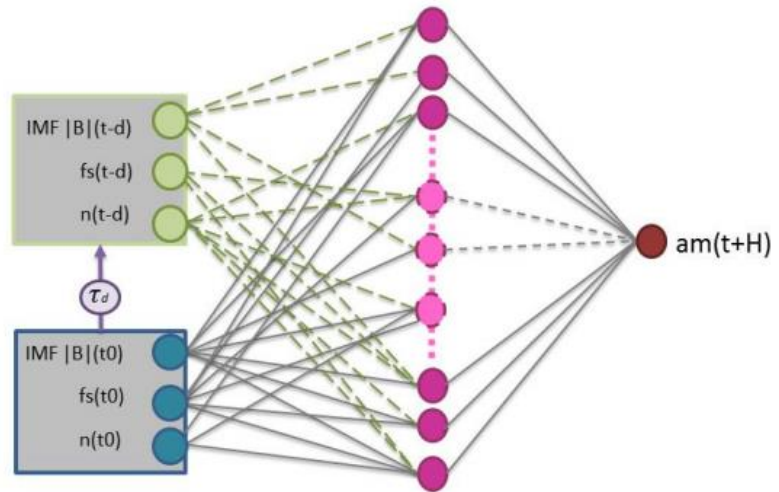
963
 964 Figure 1. Network of stations recording the disturbances leading to *Kp* and *am* indices.
 965
 966
 967
 968
 969

a)



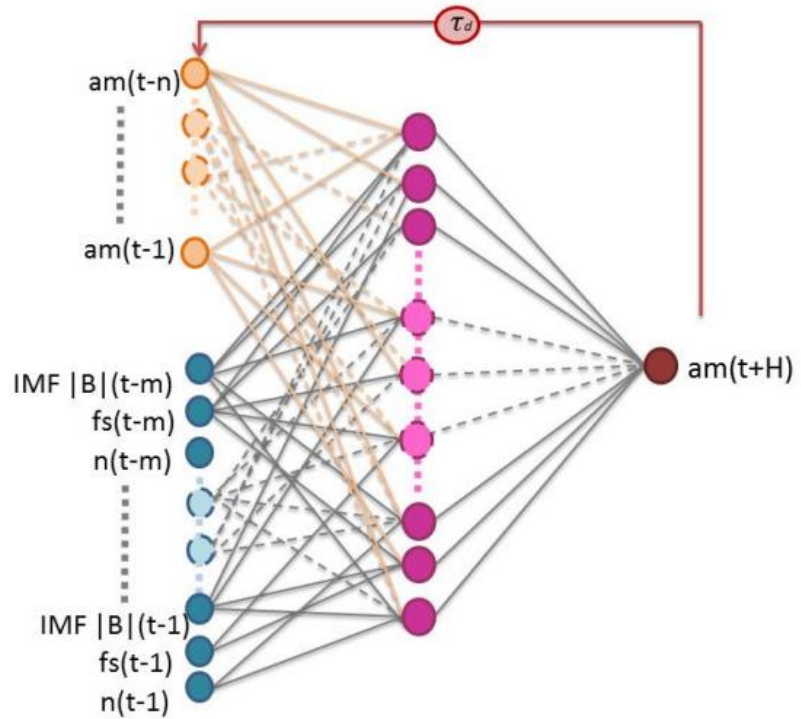
970

b)



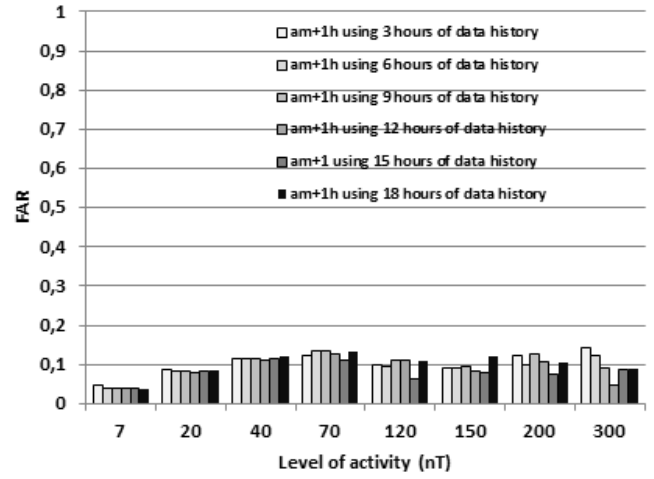
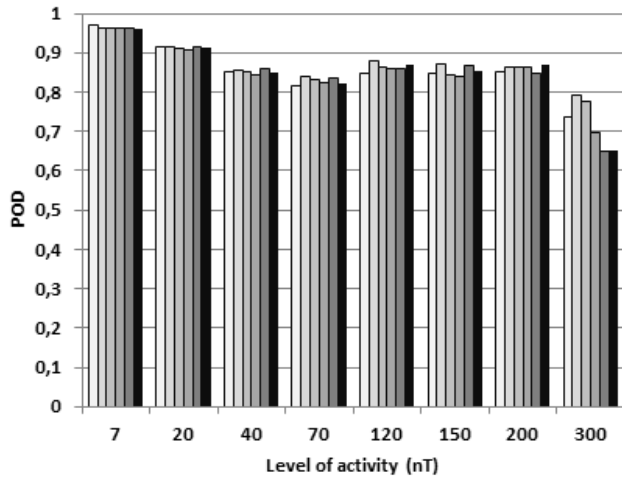
971

c)



972

973 Figure 2. For each NN, pink circles represent neurons of the hidden layer, and the red circle represents
974 the neuron of the output layer providing the prediction of am H hours ahead. Here H is equal to 1. a)
975 The feedforward multilayer NN, with blue circles representing the input layer containing solar wind
976 parameters with a time length of L b) the time delay NN, with blue circles representing the input layer
977 containing solar wind at time t. This input layer is connected with a time delay τ_d to the window of
978 specialization represented in green. The length of τ_d defines the length of d. And c) the NARX recurrent
979 NN, with blue circles representing the input layer containing solar wind parameters with a time length m
980 and orange circles representing another input layer connected to the output layer with a time delay τ_d
981 containing the geomagnetic index with a time length of n.
982

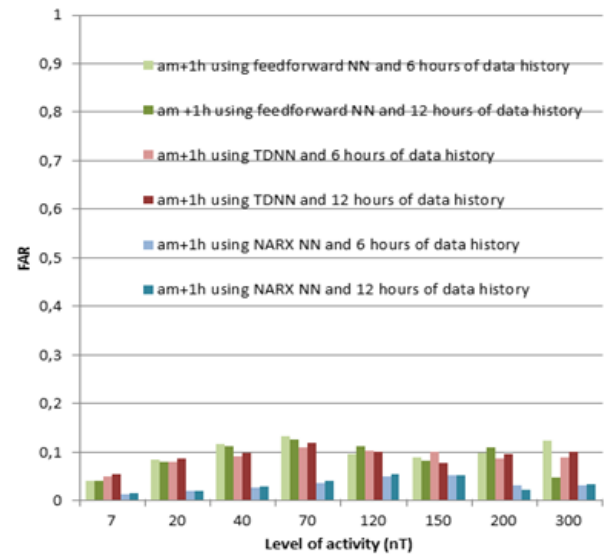
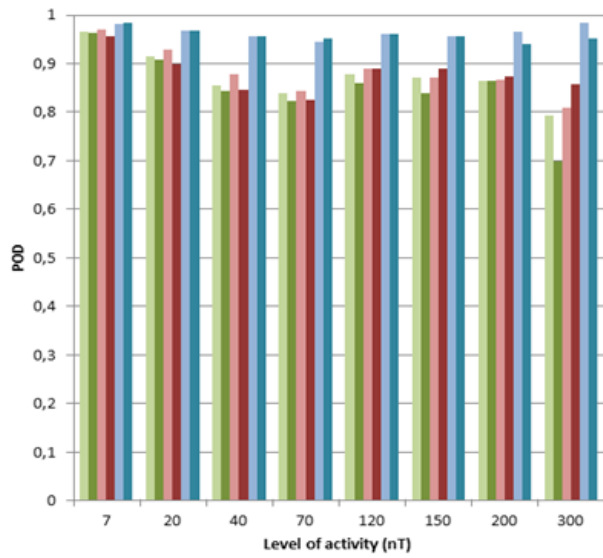


	POD						FAR					
	3h	6h	9h	12h	15h	18h	3h	6h	9h	12h	15h	18h
7	0,970	0,964	0,961	0,963	0,963	0,962	0,046	0,040	0,039	0,041	0,040	0,039
20	0,916	0,914	0,912	0,907	0,915	0,912	0,087	0,084	0,082	0,080	0,085	0,087
40	0,852	0,856	0,852	0,843	0,860	0,850	0,115	0,115	0,113	0,111	0,116	0,121
70	0,815	0,839	0,830	0,823	0,835	0,820	0,123	0,133	0,135	0,126	0,112	0,133
120	0,848	0,879	0,865	0,860	0,861	0,870	0,100	0,095	0,113	0,111	0,064	0,108
150	0,846	0,872	0,845	0,839	0,866	0,853	0,091	0,089	0,097	0,082	0,078	0,120
200	0,851	0,864	0,864	0,864	0,846	0,871	0,122	0,098	0,127	0,108	0,076	0,105
300	0,738	0,794	0,778	0,698	0,651	0,651	0,143	0,123	0,093	0,047	0,089	0,089

983

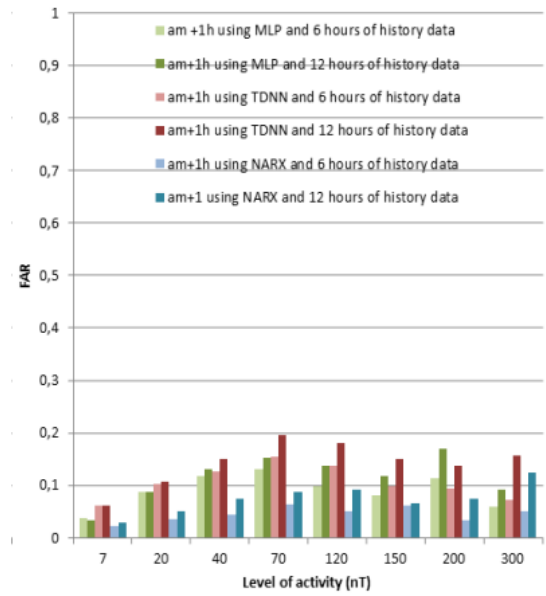
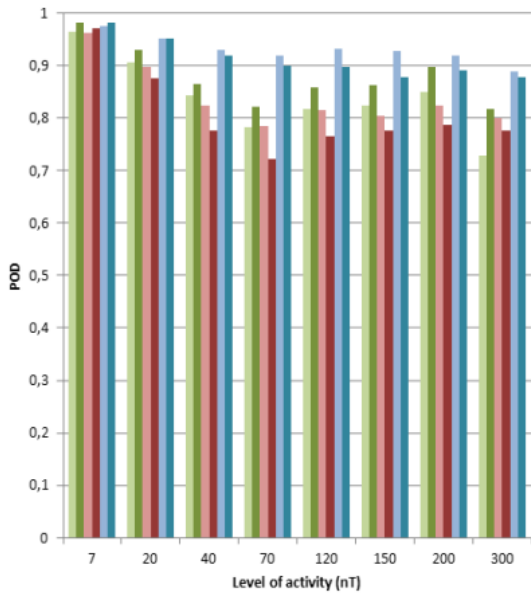
984 Figure 3. POD and FAR of the multilayer feedforward neural network depending on the length of solar
 985 wind data considered as inputs.

986



	POD						FAR					
	Feed forward NN 6h	Feed forward NN 12h	TDNN 6h	TDNN 12h	NARX NN 6h	NARX NN 12h	Feed forward NN 6h	Feed forward NN 12h	TDNN 6h	TDNN 12h	NARX NN 6h	NARX NN 12h
7	0,964	0,963	0,970	0,957	0,982	0,984	0,040	0,041	0,050	0,055	0,013	0,014
20	0,914	0,907	0,929	0,898	0,968	0,969	0,084	0,080	0,079	0,087	0,019	0,020
40	0,856	0,843	0,877	0,845	0,955	0,956	0,115	0,111	0,090	0,097	0,025	0,028
70	0,839	0,823	0,843	0,826	0,946	0,952	0,133	0,126	0,108	0,119	0,036	0,041
120	0,879	0,860	0,888	0,888	0,961	0,961	0,095	0,111	0,102	0,101	0,049	0,054
150	0,872	0,839	0,871	0,889	0,957	0,957	0,089	0,082	0,099	0,076	0,053	0,053
200	0,864	0,864	0,867	0,872	0,964	0,939	0,098	0,108	0,086	0,095	0,031	0,021
300	0,794	0,698	0,810	0,857	0,984	0,952	0,123	0,047	0,089	0,100	0,031	0,032

987
 988 Figure 4. POD and FAR of each network for 6 hours of history data and 12 hours of history data using
 989 OMNI database. Performance of the feedforward backpropagation NN are represented in green, those of
 990 the TDDN are in red and those of NARX NN are in blue.
 991



992

POD

FAR

POD

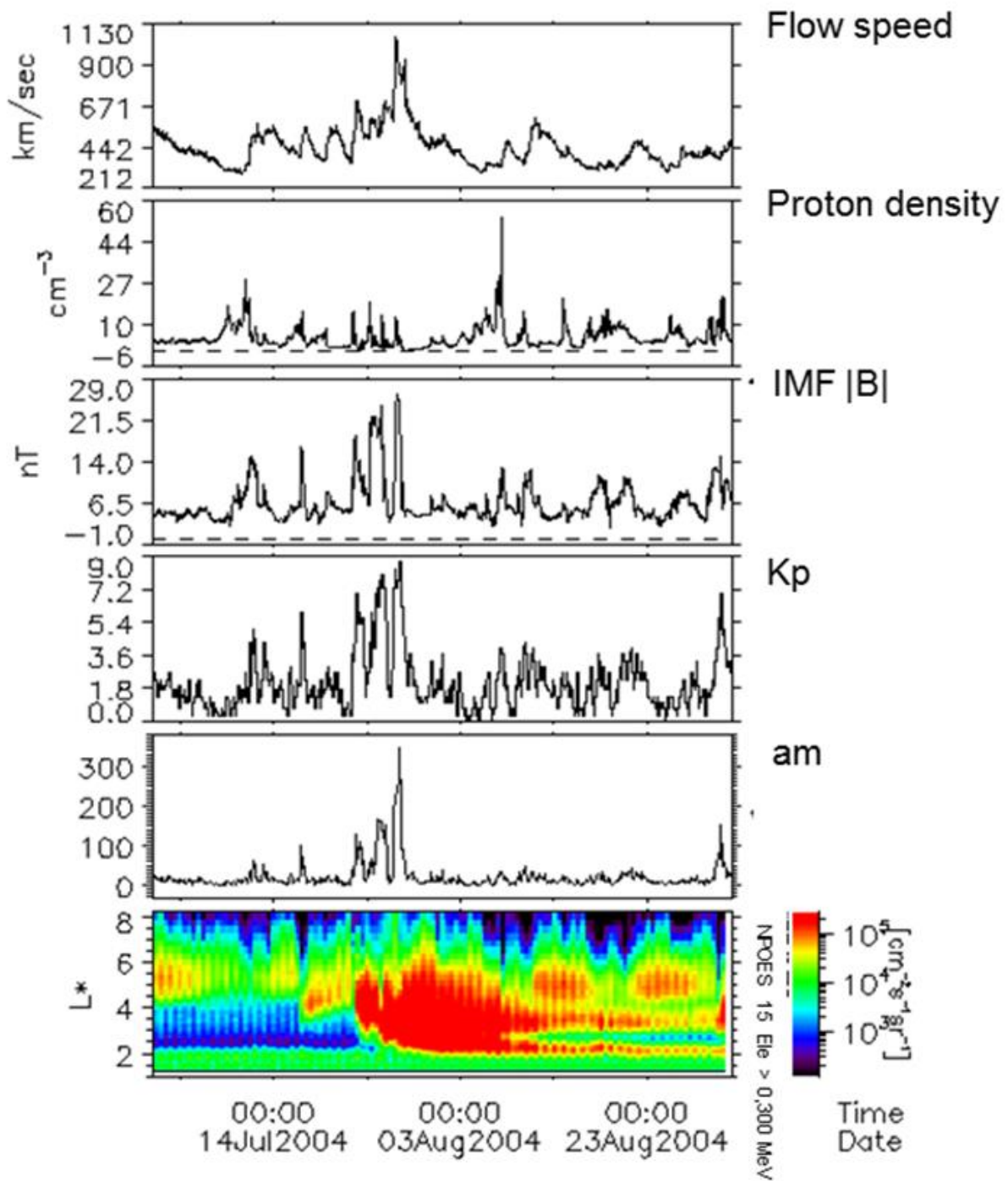
FAR

	Feed forward		TDNN		NARX NN		Feed forward		TDNN		NARX NN	
	NN 6h	NN 12h	6h	12h	NARX NN 6h	12h	NN 6h	NN 12h	6h	12h	NARX NN 6h	12h
7	0,965	0,982	0,962	0,971	0,976	0,982	0,036	0,033	0,062	0,060	0,022	0,029
20	0,906	0,929	0,898	0,875	0,952	0,951	0,088	0,087	0,103	0,106	0,035	0,050
40	0,843	0,864	0,824	0,777	0,929	0,920	0,117	0,130	0,127	0,149	0,044	0,074
70	0,782	0,822	0,784	0,721	0,919	0,900	0,131	0,152	0,155	0,195	0,062	0,086
120	0,816	0,858	0,814	0,765	0,932	0,896	0,098	0,136	0,136	0,179	0,051	0,091
150	0,824	0,863	0,805	0,776	0,927	0,877	0,081	0,117	0,098	0,149	0,061	0,064
200	0,850	0,896	0,824	0,786	0,918	0,890	0,112	0,169	0,094	0,136	0,032	0,074
300	0,727	0,816	0,800	0,776	0,889	0,878	0,059	0,091	0,071	0,156	0,051	0,122

993

994 Figure 5. POD and FAR of each network for 6 hours of history data and 12 hours of history data using
 995 ACE recorded data at L1 point. Performance of the feedforward backpropagation NN are represented in
 996 green, those of the TDDN are in red and those of NARX NN are in blue.

997



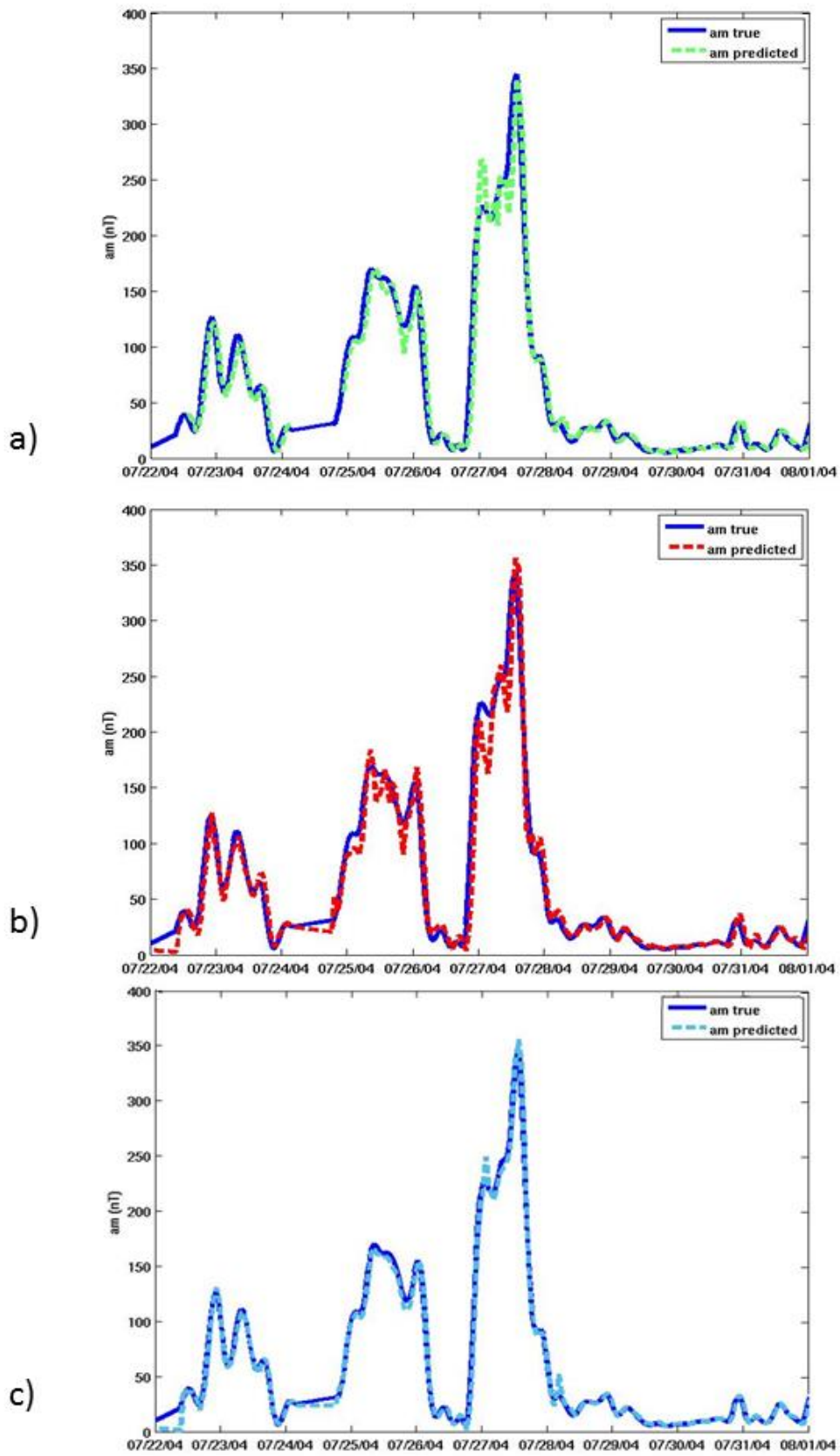
998

999 Figure 6. Description of the event of July 2004, from the top panel to the bottom panel: flow speed,

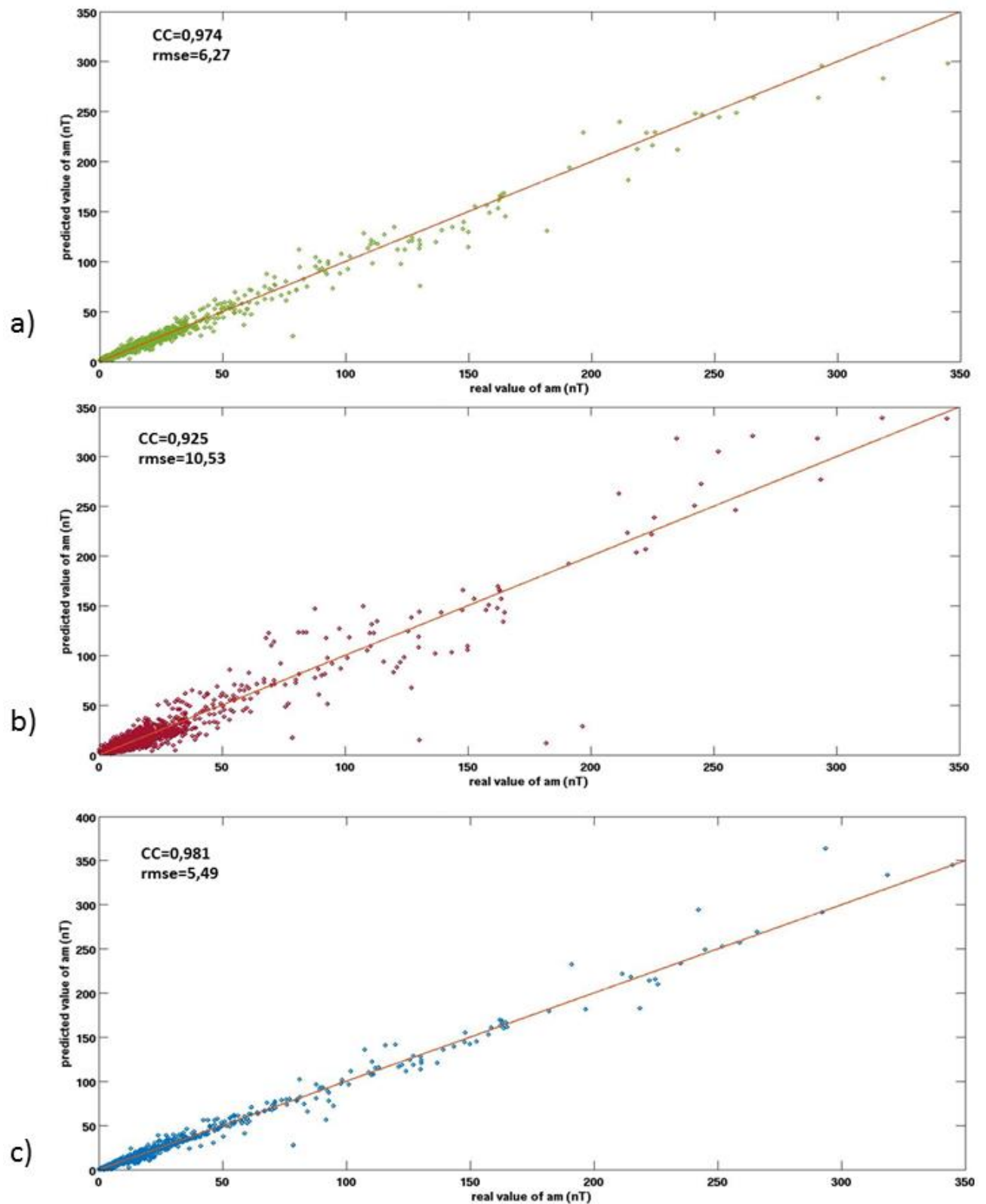
1000 proton density, IMF |B|, Kp, *am* and the integral electron flux recorded by NOAA POES 15.

1001

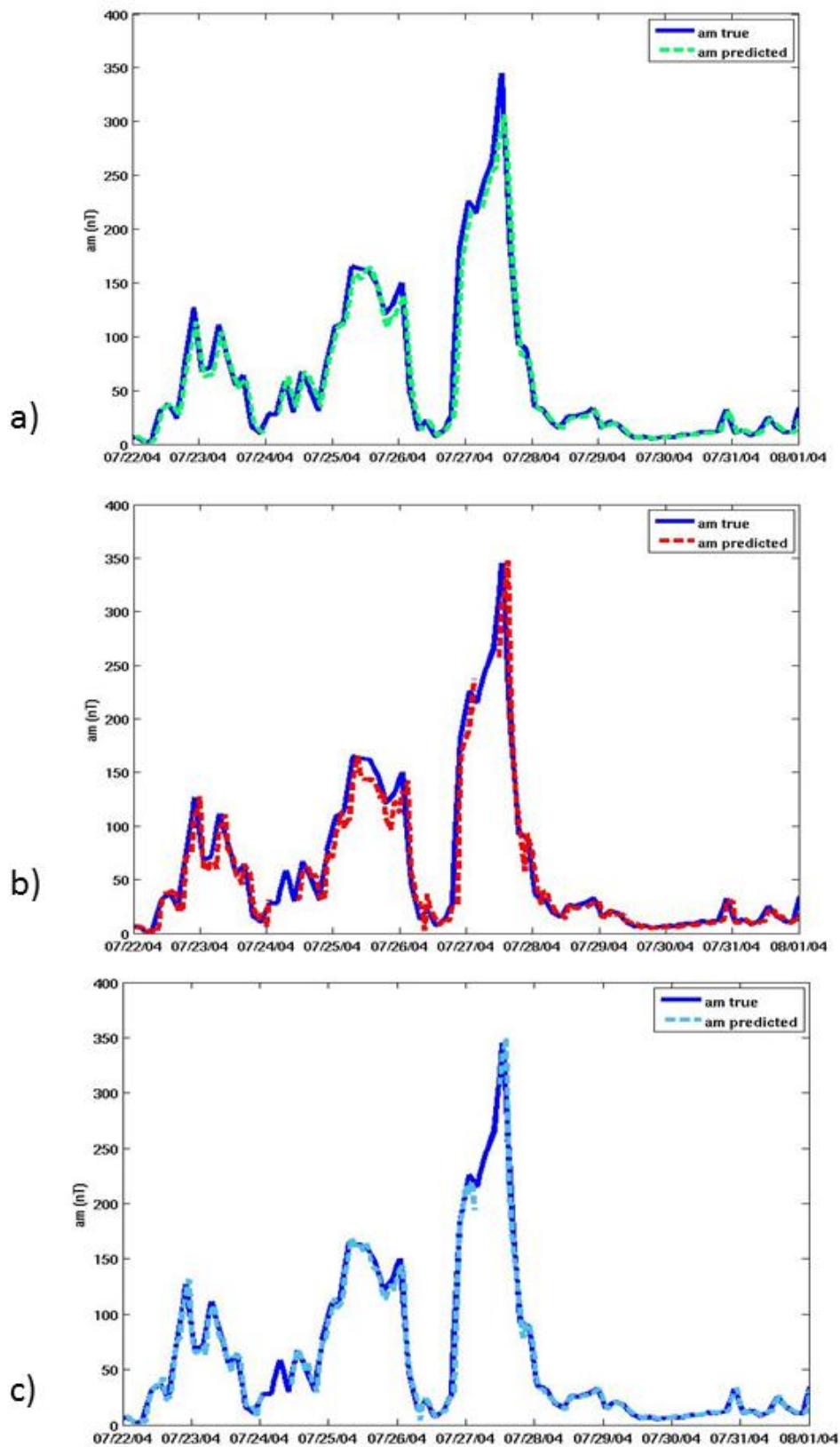
1002



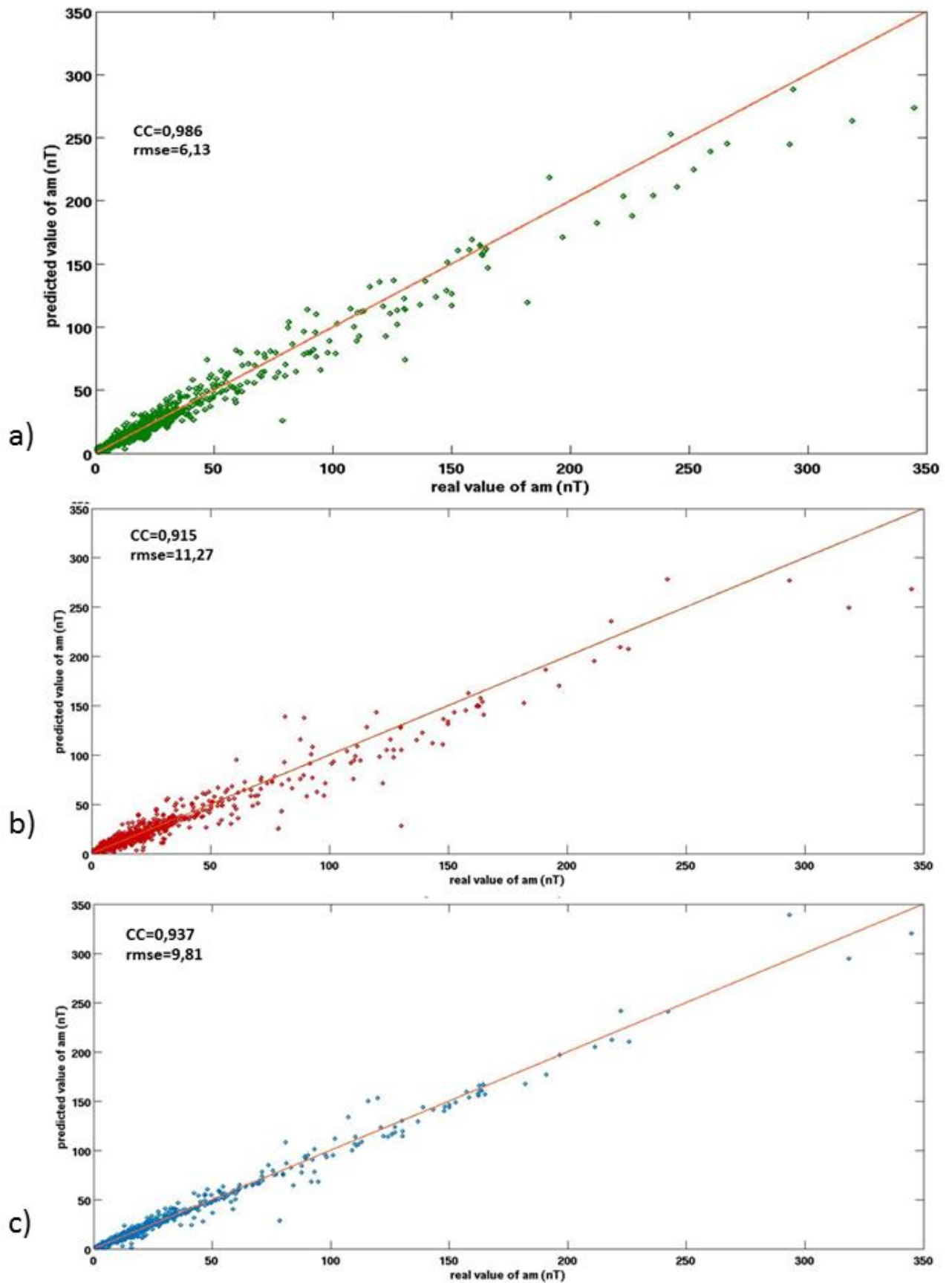
1003
 1004 Figure 7. Events of July 2004 using OMNI data, a) using feedforward multilayer network in green dot
 1005 line, b) using TDNN in red dot line, and c) using NARX recurrent network in light blue dot line. The
 1006 real values are in deep blue.



1007
 1008 Figure 8. NN performance for the event of July 2004 using OMNI data, a) using feedforward multilayer
 1009 network in green, b) using TDNN in red, and c) using NARX recurrent network in light blue.
 1010



1011
 1012 Figure 9. Events of July 2004 using ACE data, a) using feedforward multilayer network in green dot
 1013 line, b) using TDNN in red dot line, and c) using NARX recurrent network in light blue dot line. The
 1014 real values are in deep blue.
 1015



1016

1017 Figure 10. NN performance for the event of July 2004 using ACE data, a) using feedforward multilayer
 1018 network in green, b) using TDNN in red, and c) using NARX recurrent network in light blue.

1 Probabilistic Forecasting of the Disturbance Storm Time
2 Index: An Autoregressive Gaussian Process approach

M. Chandorkar,¹ E. Camporeale,¹ S. Wing²

Corresponding author: M. H. Chandorkar, Multiscale Dynamics, Centrum Wiskunde Informatica,
Science Park 123, 1098XG Amsterdam, Netherlands. (mandar.chandorkar@cwi.nl)

¹Multiscale Dynamics, Centrum Wiskunde
Informatica (CWI), Amsterdam, 1098XG
Amsterdam

²The Johns Hopkins University Applied
Physics Laboratory, Laurel, Maryland, 20723,
USA

3 **Abstract.** We present a methodology for generating probabilistic predictions
4 for the *Disturbance Storm Time (Dst)* geomagnetic activity index. We focus on
5 the *One Step Ahead (OSA)* prediction task and use the OMNI hourly resolu-
6 tion data to build our models.

7 Our proposed methodology is based on the technique of *Gaussian Process Re-*
8 *gression (GPR)*. Within this framework we develop two models; *Gaussian Pro-*
9 *cess Auto-Regressive (GP-AR)* and *Gaussian Process Auto-Regressive with eX-*
10 *ogenous inputs (GP-ARX)*.

11 We also propose a criterion to aid model selection with respect to the order
12 of auto-regressive inputs. Finally we test the performance of the GP-AR and GP-
13 ARX models on a set of 63 geomagnetic storms between 1998 and 2006 and
14 illustrate sample predictions with error bars for some of these events.

1. Introduction

15 The magnetosphere's dynamics and its associated solar wind driver form a complex dynam-
16 ical system. It is therefore instructive and greatly simplifying to use representative indices to
17 quantify the state of geomagnetic activity.

18 Geomagnetic indices come in various forms, they may take continuous or discrete values and
19 may be defined with varying time resolutions. Their values are often calculated by averaging
20 or combining a number of readings taken by instruments, usually magnetometers, around the
21 Earth. Each geomagnetic index is a proxy for a particular kind of phenomenon. Some popular
22 indices are the K_p , Dst and the AE index.

23 1. K_p : The Kp-index is a discrete valued global geomagnetic activity index and is based on
24 3 hour measurements of the K-indices [*Bartels and Veldkamp*, 1949]. The K-index itself is a
25 three hour long quasi-logarithmic local index of the geomagnetic activity, relative to a calm day
26 curve for the given location.

27 2. AE : The Auroral Electrojet Index, AE , is designed to provide a global, quantitative mea-
28 sure of auroral zone magnetic activity produced by enhanced Ionospheric currents flowing be-
29 low and within the auroral oval [*Davis and Sugiura*, 1966]. It is a continuous index which is
30 calculated every hour.

31 3. Dst : A continuous hourly index which gives a measure of the weakening or strengthen-
32 ing of the Earth's equatorial magnetic field due to the weakening or strengthening of the ring
33 currents and the geomagnetic storms [*Dessler and Parker*, 1959].

34 For the present study, we focus on prediction of the hourly Dst index which is a straight-
35 forward indicator of geomagnetic storms. More specifically, we focus on the *one step ahead*

36 (OSA), in this case one hour ahead prediction of *Dst* because it is the simplest model towards
37 building long term predictions of geomagnetic response of the Earth to changing space weather
38 conditions.

39 The *Dst* OSA prediction problem has been the subject of several modeling efforts in the
40 literature. One of the earliest models has been presented by *Burton et al.* [1975] who calculated
41 $Dst(t)$ as the solution of an *Ordinary Differential Equation* (ODE) which expressed the rate of
42 change of $Dst(t)$ as a combination of two terms: decay and injection $\frac{dDst(t)}{dt} = Q(t) - \frac{Dst(t)}{\tau}$, where
43 $Q(t)$ relates to the particle injection from the plasma sheet into the inner magnetosphere.

44 The *Burton et al.* [1975] model has proven to be very influential particularly due to its sim-
45 plicity. Many subsequent works have modified the proposed ODE by proposing alternative ex-
46 pressions for the injection term $Q(t)$ [see *Wang et al.* [2003], *O'Brien and McPherron* [2000]].
47 More recently *Ballatore and Gonzalez* [2014] have tried to generate empirical estimates for the
48 injection and decay terms in *Burton's* equation.

49 Another important empirical model used to predict *Dst* is the *Nonlinear Auto-Regressive*
50 *Moving Average with exogenous inputs* (NARMAX) methodology developed in *Billings et al.*
51 [1989], *Balikhin et al.* [2001], *Zhu et al.* [2006], *Zhu et al.* [2007], *Boynton et al.* [2011a],
52 *Boynton et al.* [2011b] and *Boynton et al.* [2013]. The NARMAX methodology builds mod-
53 els by constructing polynomial expansions of inputs and determines the best combinations of
54 monomials to include in the refined model by using a criterion called the *error reduction ratio*
55 (ERR). The parameters of the so called NARMAX OLS-ERR model are calculated by solv-
56 ing the *ordinary least squares* (OLS) problem arising from a quadratic objective function. The
57 reader may refer to *Billings* [2013] for a detailed exposition of the NARMAX methodology.

58 Yet another family of forecasting methods is based on *Artificial Neural Networks* (ANN) that
59 have been a popular choice for building predictive models. Researchers have employed both the
60 standard *feed forward* and the more specialized *recurrent* architectures. *Lundstedt et al.* [2002]
61 proposed an *Elman* recurrent network architecture called *Lund Dst*, which used the solar wind
62 velocity, *interplanetary magnetic field* (IMF) and historical *Dst* data as inputs. *Wing et al.*
63 [2005] used recurrent neural networks to predict *K_p*. *Bala et al.* [2009] originally proposed
64 a *feed forward* network for predicting the *K_p* index which used the *Boyle coupling function*
65 *Boyle et al.* [1997]. The same architecture is adapted for prediction of *Dst* in *Bala et al.* [2009],
66 popularly known as the *Rice Dst* model. *Pallochia et al.* [2006] proposed a *neural network*
67 model called EDDA to predict *Dst* using only the IMF data.

68 Although much research has been done on prediction of the *Dst* index, much less has been
69 done on probabilistic forecasting of *Dst*. One such work described in *McPherron et al.* [2013]
70 involves identification of high speed solar wind streams using the WSA model (see *Wang and*
71 *Sheeley* [1990]), using predictions of high speed streams to construct ensembles of *Dst* trajec-
72 tories which yield the quartiles of *Dst* time series.

73 In this work we propose a technique for probabilistic forecasting of *Dst*, which yields a pre-
74 dictive distribution as a closed form expression. Our models take as input past values of *Dst*,
75 solar wind speed and the *z* component of the *Interplanetary Magnetic Field* (IMF) and output a
76 Gaussian distribution with a specific mean and variance as the OSA prediction of the *Dst*.

77 We use the *Gaussian Process Regression* methodology to construct auto-regressive models
78 for *Dst* and show how to perform exact inference in this framework. We further outline a
79 methodology to perform model selection with respect to its free parameters and time histories.

80 The remainder of this paper is organised as follows: Section 2 gives the reader an overview
81 of the history of *Gaussian Process* models as well as how they are formulated and how to
82 perform inference with them. Sections 3, 4 describe the GP-AR and GP-ARX models for OSA
83 prediction of *Dst* and how to choose their free parameters for better performance.

2. Methodology: Gaussian Process

84 *Gaussian Processes* first appeared in machine learning research in *Neal* [1996], as the limit-
85 ing case of Bayesian inference performed on neural networks with infinitely many neurons in
86 the hidden layers. Although their inception in the machine learning community is recent, their
87 origins can be traced back to the geo-statistics research community where they are known as
88 *Kriging* methods (*Krige* [1951]). In pure mathematics area *Gaussian Processes* have been stud-
89 ied extensively and their existence was first proven by Kolmogorov's extension theorem (*Tao*
90 [2011]). The reader is referred to *Rasmussen and Williams* [2005] for an in depth treatment of
91 Gaussian Processes in machine learning.

92 Let us assume that we want to model a process in which a scalar quantity y is specified
93 as $y = f(\mathbf{x}) + \epsilon$ where $f(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}$ is an unknown scalar function of a multidimensional
94 input vector $\mathbf{x} \in \mathbb{R}^d$, d is the dimensionality of the input space, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is Gaussian
95 distributed noise with variance σ^2 .

96 A set of labeled data points $(\mathbf{x}_i, y_i); i = 1 \cdots N$ can be conveniently expressed by a $N \times d$ data
97 matrix \mathbf{X} and a $N \times 1$ response vector \mathbf{y} , as shown in equations (1) and (2).

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix}_{n \times d} \quad (1)$$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}_{n \times 1} \quad (2)$$

98 Our task is to infer the values of the unknown function $f(\cdot)$ based on the inputs \mathbf{X} and the noisy
 99 observations \mathbf{y} . We now assume that the joint distribution of $f(\mathbf{x}_i), i = 1 \dots N$ is a multivariate
 100 Gaussian as shown in equations (3), (4) and (5).

$$\mathbf{f} = \begin{pmatrix} f(\mathbf{x}_1) \\ f(\mathbf{x}_2) \\ \vdots \\ f(\mathbf{x}_N) \end{pmatrix} \quad (3)$$

$$\mathbf{f} | \mathbf{x}_1, \dots, \mathbf{x}_N \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda}) \quad (4)$$

$$p(\mathbf{f} | \mathbf{x}_1, \dots, \mathbf{x}_N) = \frac{1}{(2\pi)^{n/2} \det(\boldsymbol{\Lambda})^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{f} - \boldsymbol{\mu})^T \boldsymbol{\Lambda}^{-1}(\mathbf{f} - \boldsymbol{\mu})\right) \quad (5)$$

101 Here \mathbf{f} is a $N \times 1$ vector consisting of the values $f(\mathbf{x}_i), i = 1 \dots N$. In equation (4), $\mathbf{f} | \mathbf{x}_1, \dots, \mathbf{x}_N$
 102 denotes the conditional distribution of \mathbf{f} with respect to the input data (i.e., \mathbf{X}) and $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda})$
 103 represents a multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Lambda}$. The
 104 probability density function of this distribution $p(\mathbf{f} | \mathbf{x}_1, \dots, \mathbf{x}_N)$ is therefore given by equation
 105 (5).

106 From equation (5), one can observe that in order to uniquely define the distribution of the
 107 process, it is required to specify $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$. For this probability density to be valid, there are
 108 further requirements imposed on $\boldsymbol{\Lambda}$:

- 109 1. Symmetry: $\Lambda_{ij} = \Lambda_{ji} \forall i, j \in 1, \dots, N$
- 110 2. Positive Semi-definiteness: $\mathbf{z}^T \boldsymbol{\Lambda} \mathbf{z} \geq 0 \forall \mathbf{z} \in \mathbb{R}^N$

111 Inspecting the individual elements of μ and Λ , we realise that they take the following form.

$$\mu_i = \mathbb{E}[f(\mathbf{x}_i)] := m(\mathbf{x}_i) \quad (6)$$

$$\Lambda_{ij} = \mathbb{E}[(f(\mathbf{x}_i) - \mu_i)(f(\mathbf{x}_j) - \mu_j)] := K(\mathbf{x}_i, \mathbf{x}_j) \quad (7)$$

112 Here \mathbb{E} denotes the expectation (average). The elements of μ and Λ are expressed as func-
 113 tions $m(\mathbf{x}_i)$ and $K(\mathbf{x}_i, \mathbf{x}_j)$ of the inputs $\mathbf{x}_i, \mathbf{x}_j$. Specifying the functions $m(\mathbf{x})$ and $K(\mathbf{x}, \mathbf{x}')$ com-
 114 pletely specifies each element of μ and Λ and subsequently the finite dimensional distribution
 115 of $\mathbf{f}|\mathbf{x}_1, \dots, \mathbf{x}_N$. In most practical applications of *Gaussian Processes* the mean function is often
 116 defined as $m(\mathbf{x}) = 0$, which is not unreasonable if the data is standardized to have zero mean.
 117 *Gaussian Processes* are represented in machine learning literature using the following notation:

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), K(\mathbf{x}, \mathbf{x}')) \quad (8)$$

2.1. Inference and Predictions

Our aim is to infer the function $f(\mathbf{x})$ from the noisy training data and generate predictions $f(\mathbf{x}_i^*)$ for a set of test points $\mathbf{x}_i^* : \forall i \in 1, \dots, M$. We define \mathbf{X}^* as the test data matrix whose rows are formed by \mathbf{x}_i^* as shown in equation (9).

$$\mathbf{X}_* = \begin{pmatrix} (\mathbf{x}_1^*)^T \\ (\mathbf{x}_2^*)^T \\ \vdots \\ (\mathbf{x}_M^*)^T \end{pmatrix}_{M \times d} \quad (9)$$

118 Using the multivariate Gaussian distribution in equation (5) we can construct the joint dis-
 119 tribution of $f(\mathbf{x})$ over the training and test points. The vector of training and test outputs $\begin{pmatrix} \mathbf{y} \\ \mathbf{f}_* \end{pmatrix}$
 120 is of dimension $(N + M) \times 1$ and is constructed by appending the test set predictions \mathbf{f}_* to the
 121 observed noisy measurements \mathbf{y} .

$$\mathbf{f}_* = \begin{pmatrix} f(\mathbf{x}_1^*) \\ f(\mathbf{x}_2^*) \\ \vdots \\ f(\mathbf{x}_M^*) \end{pmatrix}_{M \times 1} \quad (10)$$

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{f}_* \end{pmatrix} | \mathbf{X}, \mathbf{X}_* \sim \mathcal{N} \left(\mathbf{0}, \begin{bmatrix} \mathbf{K} + \sigma^2 \mathbf{I} & \mathbf{K}_* \\ \mathbf{K}_*^T & \mathbf{K}_{**} \end{bmatrix} \right) \quad (11)$$

122 Since we have noisy measurements of f over the training data, we add the noise variance σ^2
 123 to the variance of f as shown in (11). The block matrix components of the $(N + M) \times (N + M)$
 124 covariance matrix have the following structure.

- 125 1. \mathbf{I} : The $n \times n$ identity matrix.
- 126 2. $\mathbf{K} = [K(\mathbf{x}_i, \mathbf{x}_j)]$, $i, j \in 1, \dots, n$: Kernel matrix constructed from all couples obtained from
 127 the training data.
- 128 3. $\mathbf{K}_* = [K(\mathbf{x}_i, \mathbf{x}_j^*)]$, $i \in 1, \dots, n; j \in 1, \dots, m$: Cross kernel matrix constructed from all
 129 couples between training and test data points.
- 130 4. $\mathbf{K}_{**} = [K(\mathbf{x}_i^*, \mathbf{x}_j^*)]$, $i, j \in 1, \dots, m$: Kernel matrix constructed from all couples obtained
 131 from the test data.

132 With the multivariate normal distribution defined in equation (11), probabilistic predictions
 133 f_* can be generated by constructing the conditional distribution $\mathbf{f}_* | \mathbf{X}, \mathbf{y}, \mathbf{X}_*$. Since the original
 134 distribution of $\begin{pmatrix} \mathbf{y} \\ \mathbf{f}_* \end{pmatrix} | \mathbf{X}, \mathbf{X}_*$ is a multivariate Gaussian, conditioning on a subset of elements \mathbf{y}
 135 yields another Gaussian distribution whose mean and covariance can be calculated exactly, as
 136 in equation (12) (see *Rasmussen and Williams* [2005]).

$$\mathbf{f}_* | \mathbf{X}, \mathbf{y}, \mathbf{X}_* \sim \mathcal{N}(\bar{\mathbf{f}}_*, \Sigma_*), \quad (12)$$

where

$$\bar{\mathbf{f}}_* = \mathbf{K}_*^T [\mathbf{K} + \sigma^2 \mathbf{I}]^{-1} \mathbf{y} \quad (13)$$

$$\Sigma_* = \mathbf{K}_{**} - \mathbf{K}_*^T (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{K}_* \quad (14)$$

137 The practical implementation of *Gaussian Process* models requires the inversion of the train-
 138 ing data kernel matrix $[\mathbf{K} + \sigma^2 \mathbf{I}]^{-1}$ to calculate the parameters of the predictive distribution
 139 $\mathbf{f}_* | \mathbf{X}, \mathbf{y}, \mathbf{X}_*$. The computational complexity of this inference is dominated by the linear problem
 140 in Eq. (13), which can be solved via Cholesky decomposition, with a time complexity of $O(N^3)$,
 141 where N is the number of data points.

142 The distribution of $\mathbf{f}_* | \mathbf{X}, \mathbf{y}, \mathbf{X}_*$ is known in Bayesian analysis as the *Posterior Predictive Dis-*
 143 *tribution*. This illustrates a key difference between *Gaussian Processes* and other regression
 144 models such as *Neural Networks*, *Linear Models* and *Support Vector Machines*: a *Gaussian*
 145 *Process* model does not generate point predictions for new data but outputs a predictive distri-
 146 bution for the quantity sought, thus allowing to construct error bars on the predictions. This
 147 property of Bayesian models such as *Gaussian Processes* makes them very appealing for Space
 148 Weather forecasting applications.

149 The central design issue in applying *Gaussian Process* models is the choice of the function
 150 $K(\mathbf{x}, \mathbf{x}')$. The same constraints that apply to Λ also apply to the function K . In machine learn-
 151 ing, these symmetric positive definite functions of two variables are known as *kernels*. Kernel
 152 based methods are applied extensively in data analysis i.e. regression, clustering, classification,
 153 density estimation (see *Scholkopf and Smola* [2001], *Hofmann et al.* [2008]).

2.2. Kernel Functions

154 For the success of a *Gaussian Process* model an appropriate choice of kernel function is
155 paramount. The symmetry and positive semi-definiteness of *Gaussian Process* kernels implies
156 that they represent inner-products between some basis function representation of the data. The
157 interested reader is suggested to refer to *Berlinet and Thomas-Agnan* [2004], *Scholkopf and*
158 *Smola* [2001] and *Hofmann et al.* [2008] for a thorough treatment of kernel functions and the
159 rich theory behind them. Some common kernel functions used in machine learning are listed in
160 table 1.

161 The quantities l in the RBF, and b and d in the polynomial kernel are known as *hyper-*
162 *parameters*. Hyper-parameters give flexibility to a particular kernel structure, for example
163 $d = 1, 2, 3, \dots$ in the polynomial kernel represents linear, quadratic, cubic and higher order
164 polynomials respectively. The process of assigning values to the *hyper-parameters* is crucial in
165 the model building process and is known as *model selection*.

2.3. Model Selection

166 Given a GP model with a kernel function K_θ , the problem of model selection consists of
167 finding appropriate values for the kernel hyper-parameters $\theta = (\theta_1, \theta_2, \dots, \theta_i)$. In order to assign
168 a value to θ , we must define an objective function which represents our confidence that the GP
169 model built from a particular value of θ is the best performing model. Since GP models encode
170 assumptions about the probability distribution of the output data \mathbf{y} given inputs \mathbf{X} , it is natural
171 to use the negative log-likelihood of the training data as a model selection criterion.

$$\begin{aligned}
Q(\theta) &= -\log(p(\mathbf{y}|\mathbf{X}, K_\theta)) \\
&= -\frac{1}{2}\mathbf{y}^\top(\mathbf{K}_\theta + \sigma_n^2\mathbf{I})^{-1}\mathbf{y} - \frac{1}{2}|\mathbf{K}_\theta + \sigma_n^2\mathbf{I}| - \frac{n}{2}\log(2\pi) \\
\mathbf{K}_\theta &= [K_\theta(\mathbf{x}_i, \mathbf{x}_j)]_{n \times n}
\end{aligned}$$

172 The model selection problem can now be expressed as the minimization problem shown be-
173 low.

$$\theta^* = \arg \min_{\theta} Q(\theta)$$

174 The objective function $Q(\theta)$ in the general case can have multiple local minima, and evaluat-
175 ing the value of $Q(\cdot)$ at any given θ requires inversion of the matrix $\mathbf{K}_\theta + \sigma_n^2\mathbf{I}$ which has a time
176 complexity $O(n^3)$ as noted above. In the interest of saving computational cost, one cannot use
177 exhaustive search through the domain of the hyper-parameters to inform our choice for θ . Some
178 of the techniques used for model selection in the context of GPR include.

179 1. Grid Search: Construct a grid of values for θ as the cartesian product of one dimensional
180 grids for each θ_i , evaluate $Q(\cdot)$ at each such grid point and choose the configuration which yields
181 minimum value of $Q(\cdot)$.

182 2. Coupled Simulated Annealing: Introduced in *Xavier-De-Souza et al.* [2010], it follows
183 the same procedure as *grid search*, but after evaluation of the objective $Q(\cdot)$ on the grid, each
184 grid point is iteratively mutated in a random walk fashion. This mutation is accepted or rejected
185 according to the new value of $Q(\cdot)$ as well as its value on the other grid points. This procedure
186 is iterated until some stop criterion is reached.

187 3. Maximum Likelihood: This technique as outlined in *Rasmussen and Williams* [2005] is a
 188 form of *gradient descent*. It involves starting with an initial guess for θ and iteratively improving
 189 it by calculating the gradient of $Q(\cdot)$ with respect to θ . Although this method seems intuitive, it
 190 introduces an extra computational cost of calculating the gradient of $Q(\theta)$ with respect to each θ_i
 191 in every iteration and applying this method can sometimes lead to overfitting of the GPR model
 192 to the training data.

3. One Step Ahead Prediction

193 Below in equations (15) - (17) we outline a *Gaussian Process* formulation for *OSA* prediction
 194 of *Dst*. A vector of features \mathbf{x}_{t-1} is used as input to an unknown function $f(\mathbf{x}_{t-1})$.

195 The features \mathbf{x}_{t-1} can be any collection of quantities in the hourly resolution OMNI data set.
 196 Generally \mathbf{x}_{t-1} are time histories of *Dst* and other important variables such as plasma pressure
 197 $p(t)$, solar wind speed $V(t)$, z component of the interplanetary magnetic field $B_z(t)$.

$$Dst(t) = f(\mathbf{x}_{t-1}) + \epsilon \quad (15)$$

$$\epsilon \sim \mathcal{N}(0, \sigma^2) \quad (16)$$

$$f(x_t) \sim \mathcal{GP}(m(\mathbf{x}_t), K_{osa}(\mathbf{x}_t, \mathbf{x}_s)) \quad (17)$$

$$(18)$$

198 We consider two choices for the input features \mathbf{x}_{t-1} leading to two variants of *Gaussian Pro-*
 199 *cess* regression for *Dst* time series prediction.

3.1. Gaussian Process Auto-Regressive (GP-AR)

200 The simplest auto-regressive models for *OSA* prediction of *Dst* are those that use only the
 201 history of *Dst* to construct input features for model training. The input features \mathbf{x}_{t-1} at each time
 202 step are the history of *Dst*(*t*) until a time lag of *p* hours.

$$\mathbf{x}_{t-1} = (Dst(t-1), \dots, Dst(t-p+1))$$

3.2. Gaussian Process Auto-Regressive with eXogenous inputs (GP-ARX)

203 Auto-regressive models can be augmented by including exogenous quantities in the inputs
 204 \mathbf{x}_{t-1} at each time step, in order to improve predictive accuracy. *Dst* gives a measure of ring
 205 currents, which are modulated by plasma sheet particle injections into the inner magnetosphere
 206 during sub-storms. Studies have shown that the substorm occurrence rate increases with solar
 207 wind velocity (high speed streams) *Kissinger et al.* [2011]; *Newell et al.* [2016]. Prolonged
 208 southward interplanetary magnetic field (IMF) *z*-component (B_z) is needed for sub-storms to
 209 occur *McPherron et al.* [1986]. An increase in the solar wind electric field, VB_z , can increase the
 210 dawn-dusk electric field in the magnetotail, which in turn determines the amount of plasma sheet
 211 particle that move to the inner magnetosphere *Friedel et al.* [2001]. Therefore, our exogenous
 212 parameters consist of solar wind velocity and IMF B_z .

213 In this model we choose distinct time lags p , p_v and p_b for *Dst*, *V* and B_z respectively.

$$\mathbf{x}_{t-1} = (Dst(t-1), \dots, Dst(t-p+1),$$

$$V(t-1), \dots, V(t-p_v+1),$$

$$B_z(t-1), \dots, B_z(t-p_b+1))$$

3.3. Choice of Mean Function

214 Mean functions in GPR models encode trends in the data, they are the baseline predictions
 215 the model falls back to in case the training and test data have little correlation as predicted
 216 by the kernel function. If there is no prior knowledge about the function to be approximated,
 217 *Rasmussen and Williams* [2005] state that it is perfectly reasonable to choose $m(\mathbf{x} = 0)$ as the
 218 mean function, as long as the target values are normalized. In the case of the *Dst* time series,
 219 it is known that the so called *persistence model* $\hat{D}st(t) = Dst(t - 1)$ performs quite well in the
 220 context of OSA prediction. We therefore choose the *persistence model* as the mean function in
 221 our OSA Dst models.

3.4. Choice of Kernel

222 In this study, we construct Gaussian Process regression models with a combination of the
 223 *maximum likelihood perceptron* kernel and *student's T* kernel as shown in equations (19). The
 224 *maximum likelihood perceptron* kernel is the *Gaussian Process* equivalent of a single hidden
 225 layer feed-forward neural network model as demonstrated in *Neal* [1996].

$$K_{osa}(\mathbf{x}, \mathbf{y}) = K_{mlp}(\mathbf{x}, \mathbf{y}) + K_{st}(\mathbf{x}, \mathbf{y}) \quad (19)$$

$$K_{mlp}(\mathbf{x}, \mathbf{y}) = \sin^{-1}\left(\frac{w\mathbf{x}^\top\mathbf{y} + b}{\sqrt{w\mathbf{x}^\top\mathbf{x} + b + 1}\sqrt{w\mathbf{y}^\top\mathbf{y} + b + 1}}\right) \quad (20)$$

$$K_{st}(\mathbf{x}, \mathbf{y}) = \frac{1}{1 + \|\mathbf{x} - \mathbf{y}\|_2^d} \quad (21)$$

4. Experiments

Training

226 We selected OMNI data sections 00:00 January 3 2010 to 23:00 January 23 2010 and 20:00
 227 August 5 2011 to 22:00 August 6 2011 for training the GP-AR and GP-ARX models. The
 228 first training data section consists of ambient fluctuations of *Dst* while the second contains a
 229 geomagnetic storm.

230 The computational complexity of calculation of the predictive distribution is $O(N^3)$, as dis-
 231 cussed in section 2.1. This can limit the size of the covariance matrix constructed from the
 232 training data. Note that this computation overhead is paid for every unique assignment to the
 233 model hyper-parameters. However, our chosen training set has a size of 243 which is still very
 234 much below the computational limits of the method and in our case solving equation 13 on a
 235 laptop computer takes less than a second for the training set considered in our analysis.

Selection

236 In order to find appropriate values of the hyper-parameters of the chosen kernel K_{osa} , we
 237 apply *grid search*, *coupled simulated annealing* and *maximum likelihood* methods. We fix the
 238 parameters d and σ^2 of K_{st} and model noise to values 0.01 and 0.2 respectively, the remaining
 239 parameters w and b are kept free to be calculated by model selection. Table 2 summarizes the
 240 settings used to run each model selection procedure.

Validation

241 Apart from selecting the kernel parameters, one also needs to choose appropriate values for
 242 the auto-regressive orders p in the case of GP-AR and p, p_v, p_b in the case of GP-ARX. For this
 243 purpose we use a set of 24 storm events listed in table 4 and for every assignment of values

244 to the model order, we perform model selection with the routines in table 2 and record the
 245 performance on this validation set.

246 For measuring performance of model instances on the validation set storm events, the follow-
 247 ing metrics are calculated.

1. The mean absolute error.

$$MAE = \sum_{t=1}^n |(Dst(t) - \hat{D}st(t))| / n \quad (22)$$

2. The root mean square error.

$$RMSE = \sqrt{\sum_{t=1}^n (Dst(t) - \hat{D}st(t))^2 / n} \quad (23)$$

3. Correlation coefficient between the predicted and actual value of *Dst*.

$$CC = Cov(Dst, \hat{D}st) / \sqrt{Var(Dst)Var(\hat{D}st)} \quad (24)$$

248 In the case of GP-AR we let the model order p vary from 5 to 12 while for GP-ARX we vary
 249 the total model order $p_t = p + p_v + p_b$ vary from 3 to 12 and for each p_t evaluate every possible
 250 combination of p , p_v and p_b such that $p_t = p + p_v + p_b$ and $p, p_v, p_b > 0$.

Evaluation

251 After selecting the best performing GP-AR and GP-ARX models in the validation phase,
 252 we test and compare the performance of these models with the predictions generated from the
 253 *persistence model* $\hat{D}st(t) = Dst(t - 1)$, on a set of 63 storm events occurring between 1998 and
 254 2006 as given in table 5, which is the same list of storm events as used in *Ji et al.* [2012].

5. Results

255 Figures 1 and 2 show how the mean absolute error and coefficient of correlation as calculated
 256 on the validation set storm events of table 4, vary with increasing model order for GP-AR and

257 GP-ARX. The results are represented as boxplots, in which a rectangle is drawn to represent
258 the second and third quartiles, with a vertical line inside to indicate the median value while
259 outlying points are shown as dots. In both cases, the predictive performance first improves and
260 then stagnates or worsens with increasing model order.

261 Figures 3 and 4 break down the results for GP-ARX by the model selection routine used.
262 Apart from the general trend observed in 1 and 2, we also observe that *grid search* and *cou-*
263 *pled simulated annealing* give superior performance as compared to gradient based *maximum*
264 *likelihood*.

265 From the validation results, we can chose the model order which yields the best performance,
266 for GP-AR it is $p_t = 6$ while for GP-ARX it is $p_t = 11$. Further examination of the validation
267 results shows that in the scheme $p_t = 11$ choosing $p = 7, p_v = 1, p_b = 3$ gives superior results.

268 After choosing the best performing GP-AR and GP-ARX models, we calculate their perfor-
269 mance on the test set of table 5. The results of these model evaluations are summarized in table
270 3, the GP-AR and GP-ARX models improve upon the performance of the *persistence model*.

271 Figures 5, 6 and 7 show OSA predictions of the GP-ARX model with $\pm\sigma$ error bars for three
272 storm events in the time period between 1998 and 2003. The GP-ARX model gives accurate
273 predictions along with plausible error bars around its mean predictions.

6. Conclusions

274 In this paper, we describe a flexible and expressive methodology for generating probabilistic
275 forecasts of the *Dst* index. We proposed two *Gaussian Process* auto-regressive models, *GP-*
276 *ARX* and *GP-AR*, to generate hourly predictions and their associated error bars. We also describe
277 how to carry out model selection and validation of GP-AR and GP-ARX models.

278 Our results can be summarized as follows.

279 1. *Persistence* model plays an important role in the model building and evaluation process
280 in the context of *one step ahead* prediction of the *Dst* index. It is clear that the persistence
281 behavior in the *Dst* values is very strong i.e. the trivial predictive model $\hat{D}st(t) = Dst(t - 1)$
282 gives excellent performance according to the metrics chosen.

283 2. *Gaussian Process* AR and ARX models give encouraging benefits in OSA prediction.
284 Leveraging the strengths of the Bayesian approach, they are able to learn robust predictors from
285 data. If one compares the training sets used by all the models, one can appreciate that the models
286 presented here need relatively small training and validations sets: the training set contains 243
287 instances, while the validation set contains 782 instances.

288 3. Since the GP models generate predictive distributions for test data and not just point pre-
289 dictions they lend themselves to the requirements of space weather prediction very well because
290 of the need to generate error bars on predictions.

291 4. The *Gaussian Process* regression framework described in this study can also be extended
292 to multiple hour ahead prediction of *Dst*, which is currently a work in progress.

293 **Acknowledgments.** We acknowledge use of NASA/GSFC's Space Physics Data Facility's
294 OMNIWeb (or CDAWeb or ftp) service, and OMNI data. Simon Wing acknowledges supports
295 from CWI and NSF Grant AGS-1058456 and NASA Grants (NNX13AE12G, NNX15AJ01G,
296 NNX16AC39G).

References

297 Bala, R., P. H. Reiff, and J. E. Landivar (2009), Real-time prediction of magnetospheric activity
298 using the boyle index, *Space Weather*, 7(4), n/a–n/a, doi:10.1029/2008SW000407.

- 299 Balikhin, M. A., O. M. Boaghe, S. A. Billings, and H. S. C. K. Alleyne (2001), Terrestrial
300 magnetosphere as a nonlinear resonator, *Geophysical Research Letters*, 28(6), 1123–1126,
301 doi:10.1029/2000GL000112.
- 302 Ballatore, P., and W. D. Gonzalez (2014), On the estimates of the ring current injection and
303 decay, *Earth, Planets and Space*, 55(7), 427–435, doi:10.1186/BF03351776.
- 304 Bartels, J., and J. Veldkamp (1949), International data on magnetic disturbances, second quarter,
305 1949, *Journal of Geophysical Research*, 54(4), 399–400, doi:10.1029/JZ054i004p00399.
- 306 Berline, A., and C. Thomas-Agnan (2004), *Reproducing Kernel Hilbert Spaces in Probability*
307 *and Statistics*, 355 pp., Springer US, doi:10.1007/978-1-4419-9096-9.
- 308 Billings, S. A. (2013), *Nonlinear system identification: NARMAX methods in the time, fre-*
309 *quency, and spatio-temporal domains*, John Wiley & Sons.
- 310 Billings, S. A., S. Chen, and M. J. Korenberg (1989), Identification of mimo non-linear systems
311 using a forward-regression orthogonal estimator, *International Journal of Control*, 49(6),
312 2157–2189, doi:10.1080/00207178908559767.
- 313 Boyle, C., P. Reiff, and M. Hairston (1997), Empirical polar cap potentials, *Journal of Geophys-*
314 *ical Research*, 102(A1), 111–125.
- 315 Boynton, R. J., M. A. Balikhin, S. A. Billings, A. S. Sharma, and O. A. Amariutei
316 (2011a), Data derived narmax dst model, *Annales Geophysicae*, 29(6), 965–971, doi:
317 10.5194/angeo-29-965-2011.
- 318 Boynton, R. J., M. A. Balikhin, S. A. Billings, H. L. Wei, and N. Ganushkina (2011b), Using
319 the narmax ols-err algorithm to obtain the most influential coupling functions that affect the
320 evolution of the magnetosphere, *Journal of Geophysical Research: Space Physics*, 116(A5),
321 n/a–n/a, doi:10.1029/2010JA015505.

- 322 Boynton, R. J., M. A. Balikhin, S. A. Billings, G. D. Reeves, N. Ganushkina, M. Gedalin,
323 O. A. Amariutei, J. E. Borovsky, and S. N. Walker (2013), The analysis of electron fluxes
324 at geosynchronous orbit employing a narmax approach, *Journal of Geophysical Research:*
325 *Space Physics*, 118(4), 1500–1513, doi:10.1002/jgra.50192.
- 326 Burton, R. K., R. L. McPherron, and C. T. Russell (1975), An empirical relationship between
327 interplanetary conditions and dst, *Journal of Geophysical Research*, 80(31), 4204–4214, doi:
328 10.1029/JA080i031p04204.
- 329 Davis, T. N., and M. Sugiura (1966), Auroral electrojet activity index ae and its uni-
330 versal time variations, *Journal of Geophysical Research*, 71(3), 785–801, doi:10.1029/
331 JZ071i003p00785.
- 332 Dessler, A. J., and E. N. Parker (1959), Hydromagnetic theory of geomagnetic storms, *Journal*
333 *of Geophysical Research*, 64(12), 2239–2252, doi:10.1029/JZ064i012p02239.
- 334 Friedel, R. H. W., H. Korth, M. G. Henderson, M. F. Thomsen, and J. D. Scudder (2001), Plasma
335 sheet access to the inner magnetosphere, *Journal of Geophysical Research: Space Physics*,
336 106(A4), 5845–5858, doi:10.1029/2000JA003011.
- 337 Hofmann, T., B. Scholkopf, and A. J. Smola (2008), Kernel methods in machine learning, *Ann.*
338 *Statist.*, 36(3), 1171–1220, doi:10.1214/009053607000000677.
- 339 Ji, E. Y., Y. J. Moon, N. Gopalswamy, and D. H. Lee (2012), Comparison of Dst forecast models
340 for intense geomagnetic storms, *Journal of Geophysical Research: Space Physics*, 117(3), 1–
341 9, doi:10.1029/2011JA016872.
- 342 Kissinger, J., R. L. McPherron, T.-S. Hsu, and V. Angelopoulos (2011), Steady magnetospheric
343 convection and stream interfaces: Relationship over a solar cycle, *Journal of Geophysical*
344 *Research: Space Physics*, 116(A5), n/a–n/a, doi:10.1029/2010JA015763, a00I19.

- 345 Krige, d. g. (1951), *A Statistical Approach to Some Mine Valuation and Allied Problems on the*
346 *Witwatersrand*, publisher not identified.
- 347 Lundstedt, H., H. Gleisner, and P. Wintoft (2002), Operational forecasts of the geomagnetic
348 dst index, *Geophysical Research Letters*, 29(24), 34–1–34–4, doi:10.1029/2002GL016151,
349 2181.
- 350 McPherron, R. L., T. Terasawa, and A. Nishida (1986), Solar wind triggering of substorm
351 expansion onset, *Journal of geomagnetism and geoelectricity*, 38(11), 1089–1108, doi:
352 10.5636/jgg.38.1089.
- 353 McPherron, R. L., G. Siscoe, N. U. Crooker, and N. Arge (2013), *Probabilistic Forecasting of*
354 *the Dst Index*, pp. 203–210, American Geophysical Union, doi:10.1029/155GM22.
- 355 Neal, R. M. (1996), *Bayesian Learning for Neural Networks*, Springer-Verlag New York, Inc.,
356 Secaucus, NJ, USA.
- 357 Newell, P., K. Liou, J. Gjerloev, T. Sotirelis, S. Wing, and E. Mitchell (2016), Substorm
358 probabilities are best predicted from solar wind speed, *Journal of Atmospheric and Solar-*
359 *Terrestrial Physics*, 146, 28 – 37, doi:http://dx.doi.org/10.1016/j.jastp.2016.04.019.
- 360 O’Brien, T. P., and R. L. McPherron (2000), An empirical phase space analysis of ring current
361 dynamics: Solar wind control of injection and decay, *Journal of Geophysical Research: Space*
362 *Physics*, 105(A4), 7707–7719, doi:10.1029/1998JA000437.
- 363 Pallochia, G., E. Amata, G. Consolini, M. F. Marcucci, and I. Bertello (2006), Geomagnetic
364 Dst index forecast based on IMF data only, *Annales Geophysicae*, 24(3), 989–999.
- 365 Rasmussen, C. E., and C. K. I. Williams (2005), *Gaussian Processes for Machine Learning*
366 *(Adaptive Computation and Machine Learning)*, The MIT Press.

- 367 Scholkopf, B., and A. J. Smola (2001), *Learning with Kernels: Support Vector Machines, Reg-*
368 *ularization, Optimization, and Beyond*, MIT Press, Cambridge, MA, USA.
- 369 Tao, T. (2011), *An Introduction to Measure Theory*, Graduate studies in mathematics, American
370 Mathematical Society.
- 371 Wang, C. B., J. K. Chao, and C.-H. Lin (2003), Influence of the solar wind dynamic pressure on
372 the decay and injection of the ring current, *Journal of Geophysical Research: Space Physics*,
373 *108(A9)*, n/a–n/a, doi:10.1029/2003JA009851, 1341.
- 374 Wang, Y.-M., and N. R. Sheeley, Jr. (1990), Solar wind speed and coronal flux-tube expansion,
375 *Astrophys. J.*, , 355, 726–732, doi:10.1086/168805.
- 376 Wing, S., J. R. Johnson, J. Jen, C.-I. Meng, D. G. Sibeck, K. Bechtold, J. Freeman, K. Costello,
377 M. Balikhin, and K. Takahashi (2005), Kp forecast models, *Journal of Geophysical Research:*
378 *Space Physics*, *110(A4)*, n/a–n/a, doi:10.1029/2004JA010500, a04203.
- 379 Xavier-De-Souza, S., J. A. K. Suykens, J. Vandewalle, and D. Bolle (2010), Coupled simulated
380 annealing, *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, *40(2)*,
381 320–335, doi:10.1109/TSMCB.2009.2020435.
- 382 Zhu, D., S. A. Billings, M. Balikhin, S. Wing, and D. Coca (2006), Data derived continuous
383 time model for the dst dynamics, *Geophysical Research Letters*, *33(4)*, n/a–n/a, doi:10.1029/
384 2005GL025022.
- 385 Zhu, D., S. A. Billings, M. A. Balikhin, S. Wing, and H. Alleyne (2007), Multi-input data
386 derived dst model, *Journal of Geophysical Research: Space Physics*, *112(A6)*, n/a–n/a, doi:
387 10.1029/2006JA012079.

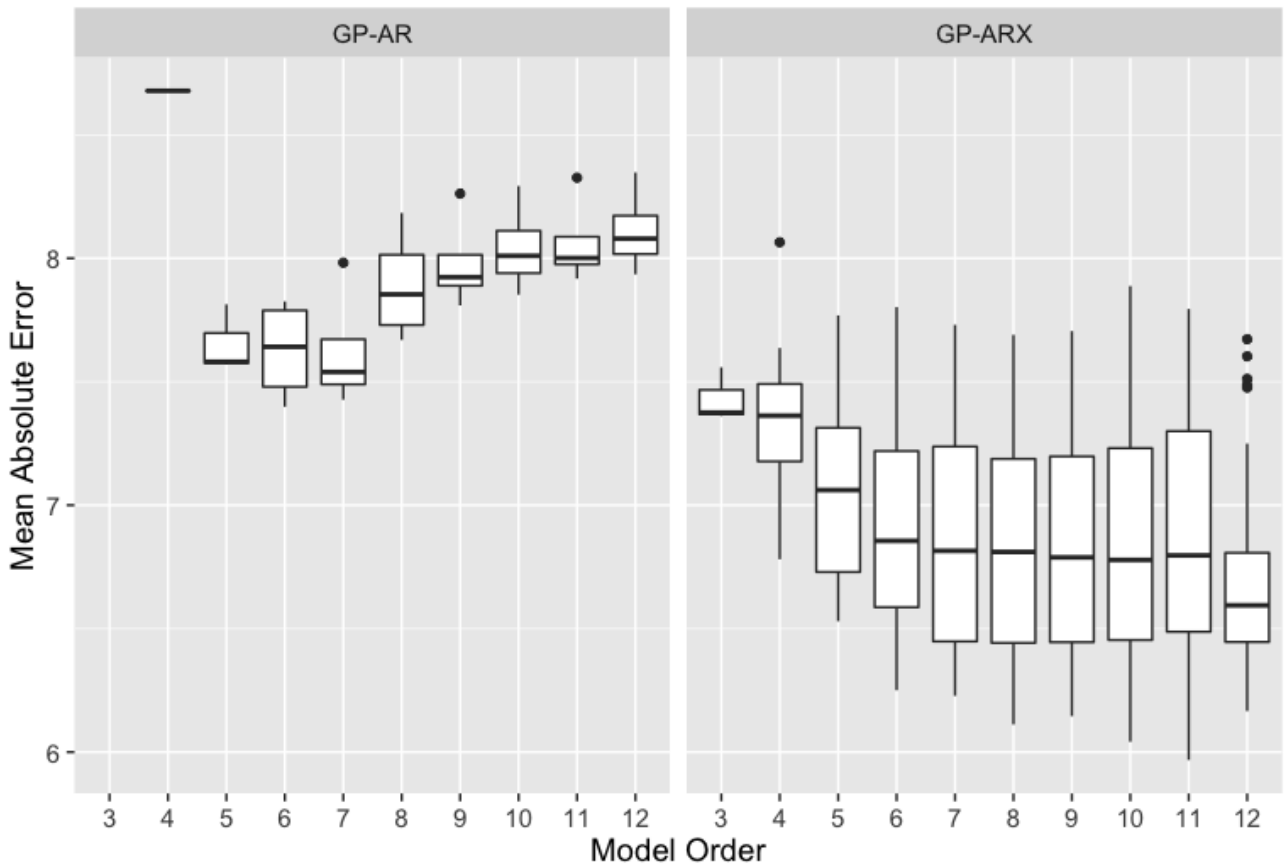


Figure 1. Mean Absolute Error on validation set storms vs model order for GP-AR and GP-ARX.

Key: Rectangle borders represent the second and third quartiles, with a horizontal line inside to indicate the median value while outlying points are shown as dots

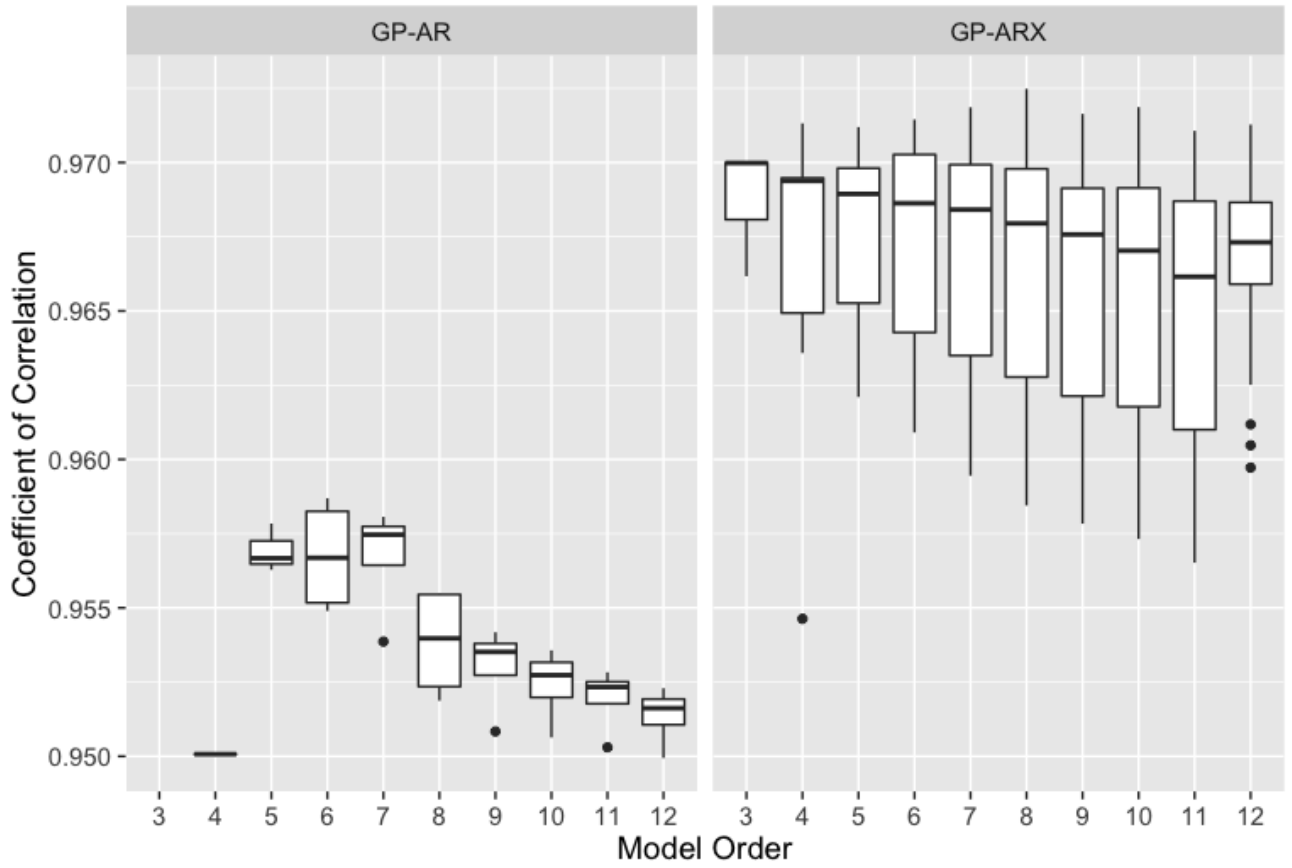


Figure 2. Coefficient of Correlation on validation set storms vs model order for GP-AR and GP-ARX

Key: Rectangle borders represent the second and third quartiles, with a horizontal line inside to indicate the median value while outlying points are shown as dots

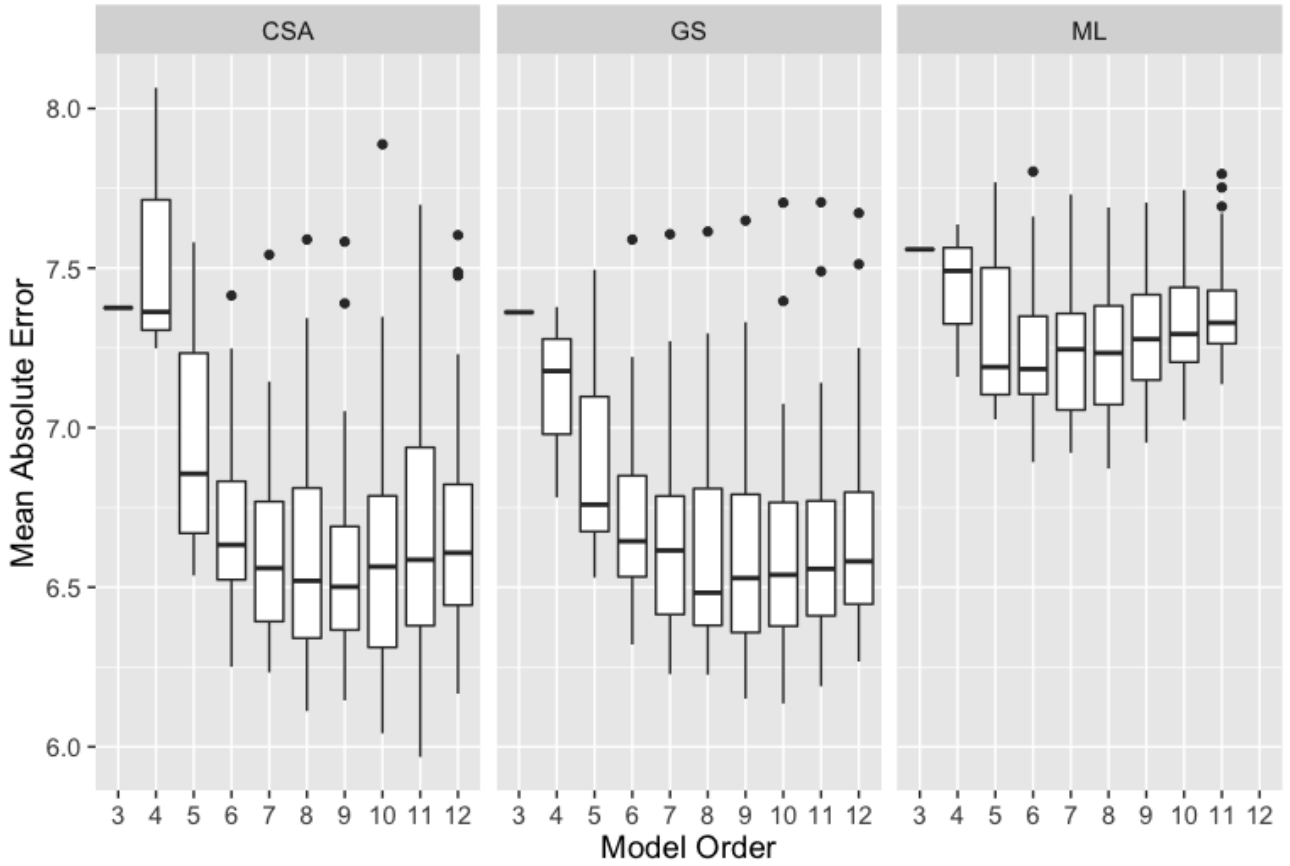


Figure 3. Mean Absolute Error on validation set storms vs model order for GP-AR and GP-ARX for *CSA*, *GS* and *ML* model selection routines

Rectangle borders represent the second and third quartiles, with a horizontal line inside to indicate the median value while outlying points are shown as dots

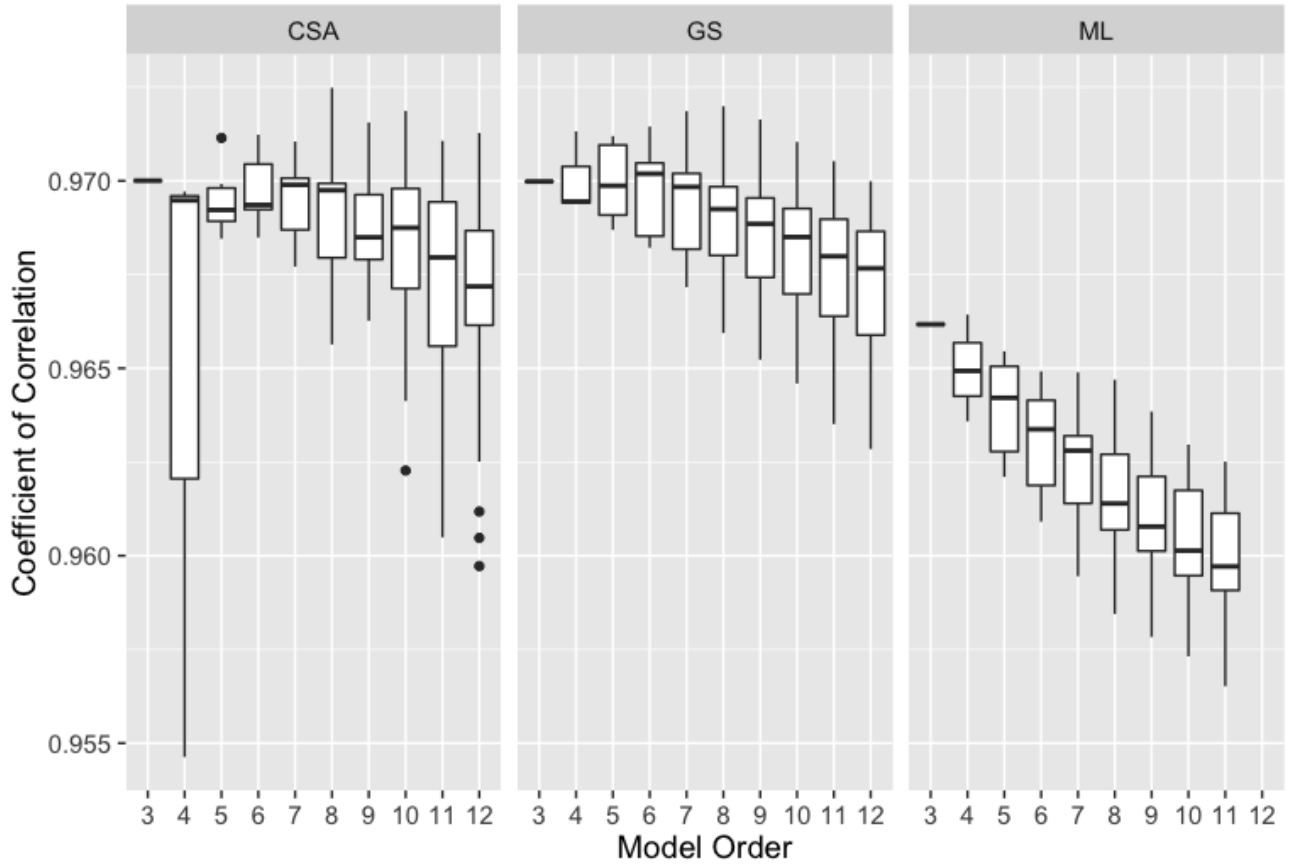


Figure 4. Coefficient of Correlation on validation set storms vs model order for GP-AR and GP-ARX for *CSA*, *GS* and *ML* model selection routines

Rectangle borders represent the second and third quartiles, with a horizontal line inside to indicate the median value while outlying points are shown as dots

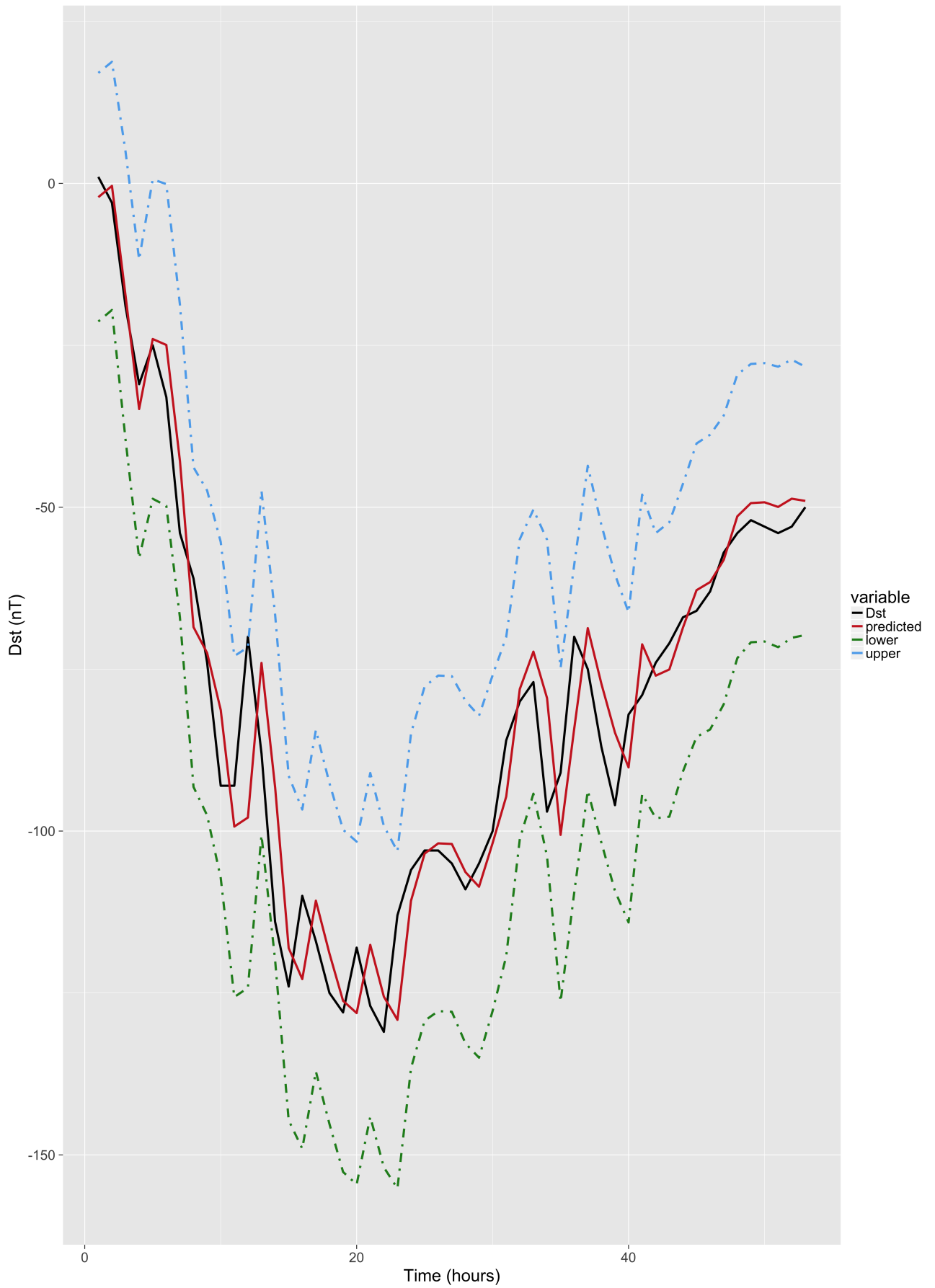


D R A F T

April 11, 2017, 2:38pm

D R A F T

Figure 5. OSA Predictions with $\pm\sigma$ error bars for event: 2003/06/17 19:00 to 2003/06/19 03:00

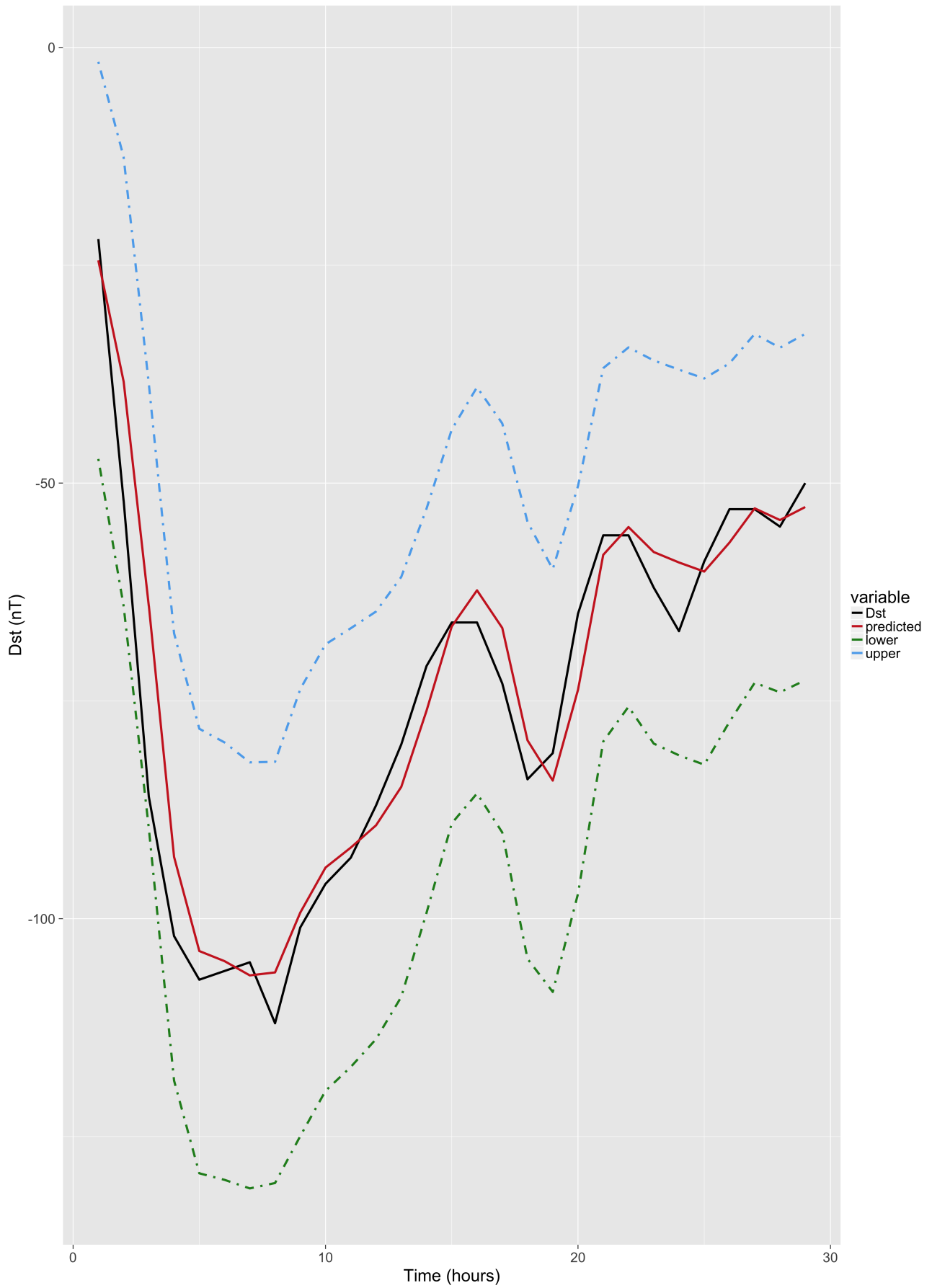


D R A F T

April 11, 2017, 2:38pm

D R A F T

Figure 6. OSA Predictions with $\pm\sigma$ error bars for event: 1998/11/13 00:00 to 1998/11/15 04:00



D R A F T

April 11, 2017, 2:38pm

D R A F T

Figure 7. OSA Predictions with $\pm\sigma$ error bars for event: 1999/01/13 16:00 to 1999/01/14 20:00

Table 1. Popular Kernel functions used in GPR models

Name	Expression	Hyperparameters
Radial Basis Function (RBF)	$\frac{1}{2} \exp(-\ \mathbf{x} - \mathbf{y}\ ^2 / l^2)$	$l \in \mathbb{R}$
Polynomial	$(\mathbf{x}^\top \mathbf{y} + b)^d$	$b \in \mathbb{R}, d \in \mathbb{N}$
Laplacian	$\exp(-\ \mathbf{x} - \mathbf{y}\ _1 / \theta)$	$\theta \in \mathbb{R}^+$
Student's T	$1 / (1 + \ \mathbf{x} - \mathbf{y}\ _2^d)$	$d \in \mathbb{R}^+$
Maximum Likelihood Perceptron	$\sin^{-1}\left(\frac{w\mathbf{x}^\top \mathbf{y} + b}{\sqrt{w\mathbf{x}^\top \mathbf{x} + b + 1} \sqrt{w\mathbf{y}^\top \mathbf{y} + b + 1}}\right)$	$w, b \in \mathbb{R}^+$

Table 2. Settings of model selection procedures

Procedure	Grid Size	Step	Max Iterations
Grid Search	10	0.2	NA
Coupled Simulated Annealing	4	0.2	30
Maximum likelihood	NA	0.2	150

Table 3. Evaluation results for models on storm events listed in table 5

Model	Mean Absolute Error	Root Mean Square Error	Coefficient of Correlation
GP-ARX	7.252	11.93	0.972
GP-AR	8.37	14.04	0.963
Persistence	9.182	14.94	0.957

Table 4. Storm events used for model selection of GP-AR and GP-ARX

Event Id	Start Date	Start Hour	End Date	End Hour	min. Dst
1	1995/03/26	0500	1995/03/26	2300	107
2	1995/04/07	1300	1995/04/09	0900	149
3	1995/09/27	0100	1995/09/28	0400	108
4	1995/10/18	1300	1995/10/19	1400	127
5	1996/10/22	2200	1996/10/23	1100	105
6	1997/04/21	1000	1997/04/22	0900	107
7	1997/05/15	0300	1997/05/16	0000	115
8	1997/10/10	1800	1997/10/11	1900	130
9	1997/11/07	0000	1997/11/07	1800	110
10	1997/11/22	2100	1997/11/24	0400	108
11	2005/06/12	1700	2005/06/13	1900	106
12	2005/08/31	1200	2005/09/01	1200	122
13	2006/12/14	2100	2006/12/16	0300	162
14	2011/09/26	1400	2011/09/27	1200	101
15	2011/10/24	2000	2011/10/25	1400	132
16	2012/03/08	1200	2012/03/10	1600	131
17	2012/04/23	1100	2012/04/24	1300	108
18	2012/07/15	0100	2012/07/16	2300	127
19	2012/09/30	1300	2012/10/01	1800	119
20	2012/10/08	0200	2012/10/09	1700	105
21	2012/11/13	1800	2012/11/14	1800	108
22	2013/03/17	0700	2013/03/18	1000	132
23	2013/05/31	1800	2013/06/01	2000	119
24	2014/02/18	1500	2014/02/19	1600	112

Table 5. Storm events used to evaluate GP-AR and GP-ARX models

Event Id	Start Date	Start Time	End Date	End Time	min. Dst
1	1998/02/17	1200	1998/02/18	1000	-100
2	1998/03/10	1100	1998/03/11	1800	-116
3	1998/05/04	0200	1998/05/05	0200	-205
4	1998/08/26	1000	1998/08/29	0700	-155
5	1998/09/25	0100	1998/09/26	0000	-207
6	1998/10/19	0500	1998/10/20	0800	-112
7	1998/11/09	0300	1998/11/10	1600	-142
8	1998/11/13	0000	1998/11/15	0400	-131
9	1999/01/13	1600	1999/01/14	2000	-112
10	1999/02/18	0300	1999/02/19	2100	-123
11	1999/09/22	2000	1999/09/23	2300	-173
12	1999/10/22	0000	1999/10/23	1400	-237
13	2000/02/12	0500	2000/02/13	1500	-133
14	2000/04/06	1700	2000/04/08	0900	-288
15	2000/05/24	0100	2000/05/25	2000	-147
16	2000/08/10	2000	2000/08/11	1800	-106
17	2000/08/12	0200	2000/08/13	1700	-235
18	2000/10/13	0200	2000/10/14	2300	-107
19	2000/10/28	2000	2000/10/29	2000	-127
20	2000/11/06	1300	2000/11/07	1800	-159
21	2000/11/28	1800	2000/11/29	2300	-119
22	2001/03/19	1500	2001/03/21	2300	-149
23	2001/03/31	0400	2001/04/01	2100	-387
24	2001/04/11	1600	2001/04/13	0700	-271
25	2001/04/18	0100	2001/04/18	1300	-114
26	2001/04/22	0200	2001/04/23	1500	-102
27	2001/08/17	1600	2001/08/18	1600	-105
28	2001/09/30	2300	2001/10/02	0000	-148
29	2001/10/21	1700	2001/10/24	1100	-187
30	2001/10/28	0300	2001/10/29	2200	-157
31	2002/03/23	1400	2002/03/25	0500	-100
32	2002/04/17	1100	2002/04/19	0200	-127
33	2002/04/19	0900	2002/04/21	0600	-149
34	2002/05/11	1000	2002/05/12	1600	-110
35	2002/05/23	1200	2002/05/24	2300	-109
36	2002/08/01	2300	2002/08/02	0900	-102
37	2002/09/04	0100	2002/09/05	0000	-109
38	2002/09/07	1400	2002/09/08	2000	-181
39	2002/10/01	0600	2002/10/03	0800	-176
40	2002/10/03	1000	2002/10/04	1800	-146
41	2002/11/20	1600	2002/11/22	0600	-128
42	2003/05/29	2000	2003/05/30	1000	-144
43	2003/06/17	1900	2003/06/19	0300	-141
44	2003/07/11	1500	2003/07/12	1600	-105
45	2003/08/17	1800	2003/08/19	1100	-148
46	2003/11/20	1200	2003/11/22	0000	-422
47	2004/01/22	0300	2004/01/24	0000	-149
48	2004/02/11	1000	2004/02/12	0000	-105
49	2004/04/03	1400	2004/04/04	0800	-112
50	2004/07/22	2000	2004/07/23	2000	-101
51	2004/07/24	2100	2004/07/26	1700	-148
52	2004/07/26	2200	2004/07/30	0500	-197
53	2004/08/30	0500	2004/08/31	2100	-126
54	2004/11/07	2100	2004/11/08	2100	-373
55	2004/11/09	1100	2004/11/11	0900	-289
56	2004/11/11	2200	2004/11/13	1300	-109
57	2005/01/21	1800	2005/01/23	0500	-105
58	2005/05/07	2000	2005/05/09	1000	-127
59	2005/05/29	2200	2005/05/31	0800	-138
60	2005/06/12	1700	2005/06/13	1900	-106
61	2005/08/31	1200	2005/09/01	1200	-131
62	2006/04/13	2000	2006/04/14	2300	-111
63	2006/12/14	2100	2006/12/16	0300	-147

