Diagnostic de la MJO sur le bassin Sud Ouest de l'Océan Indien à partir des données de prévision intra-saisonnière de la base S2S avec des techniques d'intelligence artificielle

Rémy Köth [1]

Encadré par
Quoc-Phi Duong [2]
Sylvie Malardel [2]
Hélène Vérèmes [2]

[1] École Nationale de Météorologie (ENM),
Météo-France
[2] Laboratoire de l'Atmosphère et des Cyclones (LACy),
Université de La Réunion / Météo-France / CNRS

2021

Résumé

Cette étude s'inscrit dans un contexte de prévision aux échéances intra-saisonnières, de 2 semaines à 2 mois, qui a été assez peu explorée jusqu'à récemment. La prévision à ces échéances est particulièrement importante sur le bassin Sud-Ouest de l'Océan Indien (SOOI). En effet, la prévisibilité des cyclones tropicaux en été austral sur ce bassin est fortement liée à la prévisibilité de l'Oscillation de Madden-Julian (MJO) et des ondes tropicales (ondes de Kelvin et de Rossby-gravité). Nous nous concentrons dans le cadre de ce projet sur le diagnostic de la MJO par régression de l'indice RMM. L'objectif de ce stage est d'abord d'évaluer la capacité de différentes méthodes d'IA à détecter la phase et l'intensité de la MJO à l'échéance 0 (i.e. en mode analyse). Ensuite, l'objectif est d'appliquer ces méthodes à la prévision de MJO pour des échéances allant jusqu'à la semaine 4 voire au delà. On espère qu'elles permettront d'améliorer le diagnostic de la MJO, en particulier lorsque son amplitude devient faible (c'est un biais connu de la prévision à longue échéance de la MJO).

Notre diagnostic de la MJO repose sur l'indice RMM (Real-time Multivariate MJO), permettant de rendre compte à la fois de l'intensité et de la phase de la MJO. Notre approche est de faire une régression de cet indice. Le prédictand de la régression, c'est à dire l'information que l'on cherche, est donc l'indice RMM. Les prédicteurs, c'est à dire l'information utilisée pour cette régression, sont différents champs du modèle S2S global de ECMWF (European Centre for Medium-Range Weather Forecasts) : le vent zonal à 850 et à 200hPa, l'OLR (Outgoing Longwave Radiation) et l'eau totale sur la colonne d'eau (TCW : Total Column Water). Les méthodes d'intelligence artificielles mises en oeuvre relèvent à la fois du machine learning et du deep learning : régression linéaire, Support Vector Regressor (SVR), perceptron multicouche (MLP : MultiLayer Perceptron) et réseau de neurones convolutif 2D (CNN : Convolutive Neural Network).

Nous avons montré que les méthodes de régression par IA, de machine learning ou de deep learning, sont capables d'établir un bon diagnostic de l'indice RMM de la MJO. Le RMM considéré comme la vérité dans notre étude est celui de l'analyse du BoM. Il est bien retrouvé par régression à partir de l'échéance 0 du modèle S2S de ECMWF. Nous avons montré que cette tâche est double : elle consiste à la fois en un diagnostic de l'indice RMM et en un débiaisage du modèle par rapport à l'analyse du BoM. Le modèle SVR s'est avéré être le modèle le plus performant. Il permet une précision de 71% sur le diagnostic de la phase d'une MJO intense. En revanche, il faut noter sa tendance à sous-estimer l'intensité de la MJO, avec une erreur sur l'amplitude de -0.27. En mode prévision, cette méthode d'IA s'est montrée performante en permettant d'améliorer la prévision de MJO par rapport au calcul classique de RMM. Elle permet à la fois d'obtenir des prévisions plus fiables et de repousser le seuil de prévisibilité de 1 à 2 jours. Les méthodes de régression linéaires et de MLP se sont avérées moins intéressantes que le SVR mais tout de même performantes. Les modèles de CNN, de KNN et de forêt aléatoire n'ont pas permis d'obtenir de bons résultats.

Table des matières

Liste des figures

Liste des tableaux

Acronymes

ACP Analyse en Composantes Principales.

AI Artificial Intelligence.

BCORR Bivariate Correlation.

BoM Bureau of Meteorology.

BRMSE Bivariate Root Mean Squared Error.

CNN Convolutive Neural Network.

DIROI Direction Interrégionale de l'Océan Indien de Météo-France.

DJF Décembre, Janvier, Février.

ECMWF European Centre for Medium-Range Weather Forecasts.

ENM École Nationale de Météorologie.

ENSO El Niño Southern Oscillation.

EOF Empirical Orthogonal Functions.

EPS Ensemble Prediction System.

HPC High Performance Calculus.

IA Intelligence Artificielle.

k-NN k-Nearest Neighbors.

LACy Laboratoire de l'Atmosphère et des Cyclones.

LSTM Long Short-Term Memory.

MAE Mean Absolute Error.

MJO Madden Julian Oscillation.

MLP MultiLayer Perceptron.

MLR MultiLinear Regression.

MSE Mean Squarred Error.

NCAR National Center for Atmospheric Research.

NCEP National Centers for Environmental Prediction.

NOAA National Oceanic and Atmospheric Administration.

OLR Outgoing Longwave Radiation.

PC Principal Component.

PIROI Plate-forme d'Intervention Régionale de l'Océan Indien de la Croix-Rouge Française.

PISSARO Prévisions Intra-Saisonnières à Saisonnières avec AROme.
PNT Prévision Numérique du Temps.
PV200 Potentiel de Vitesse à 200 hPa.

RBF Radial Basis Function.
ReLU Rectified Linear Unit.
RMM Real-time Multivariate MJO.

S2S Subseasonal to Seasonal.
SGD Stochastic Gradient Descent.
SOOI Sud-Ouest de l'Océan Indien.
SST Sea Surface Temperature.
SVM Support Vector Machine.
SVR Support Vector Regressor.

TCW Total Column Water.
TTR Top net Thermal Radiation.

U200 Vent zonal à 200 hPa.
U850 Vent zonal à 850 hPa.

VPM Velocity Potential MJO index.

1 Introduction

La prévision aux échéances intra-saisonnières, c'est à dire de 2 semaines à 2 mois, a été assez peu explorée jusqu'à récemment. L'objectif du projet S2S (Subseasonal-to-Seasonal), débuté en 2013, est ainsi de réduire le fossé de prévisibilité entre les courtes et moyennes échéances et les échéances saisonnières et climatiques [Brunet et al., 2010] .

Cette prévision intra-saisonnière est particulièrement importante sur le bassin Sud-Ouest de l'Océan Indien (SOOI). Le projet PISSARO [1] (Prévisions Intra-Saisonnières à Saisonnières avec AROme) constitue une passerelle entre recherche, prévision opérationnelle et acteurs de terrain sur le bassin SOOI. Il rassemble ainsi plusieurs partenaires, dont la Direction Interrégionale de l'Océan Indien (DIROI) de Météo-France et la Plate-forme d'Intervention Régionale de l'Océan Indien (PIROI) de la Croix-Rouge Française. Le projet PISSARO s'intéresse aux phénomènes météorologiques à risques tels que l'activité cyclonique ou les fortes précipitations.

Sur le bassin SOOI, la prévisibilité des cyclones tropicaux est fortement liée à la prévisibilité de l'Oscillation de Madden-Julian (MJO) et des ondes tropicales (Ondes de Kelvin et de Rossby-gravité) qui se propagent en été austral. Dans le cadre de ce stage de fin d'étude, nous nous concentrerons sur la prévision de la MJO. Les modèles globaux actuels prévoient bien en général la MJO pour les échéances jusqu'à 2 semaines [Kim et al., 2018]. Au-delà, le signal s'affaiblit et devient plus complexe à détecter.

L'approche choisie dans ce projet est d'établir une prévision de MJO par régression de l'indice RMM, à partir de différents champs prévus par un modèle S2S de prévision numérique du temps (PNT). L"indice Real-time Multivariate MJO (RMM) permet de définir la phase et l'intensité de la MJO chaque jour et est couramment utilisé pour diagnostiquer la MJO [Wheeler and Hendon, 2004]. Il y a deux objectifs à ce stage. Tout d'abord, il s'agit d'explorer dans quelle mesure les méthodes d'intelligence artificielle permettent de diagnostiquer la MJO sans passer par le calcul classique, mais lourd, des indices RMM. Ensuite, il s'agit d'évaluer si le post-traitement de la prévision numérique par l'intelligence artificielle permet ou non une amélioration, par rapport au calcul de RMM classique à partir des champs prévus.

En premier lieu, nous exposerons le contexte scientifique de nos travaux (section 2) en présentant la MJO, l'intelligence artificielle et l'état de l'art. Ensuite, nous présenterons les données (section 3) et les différentes méthodes que nous avons déployées (section 4). Nous vous proposerons un récapitulatif des différents résultats ainsi que leur analyse (section 5). Enfin, nous conclurons et discuterons des perspectives (section 6).

2 Contexte scientifique

2.1 L'oscillation de Madden-Julian

2.1.1 Généralités

L'oscillation de Madden-Julian constitue le mode principal de variabilité de l'atmosphère à l'échelle intra-saisonnière. Elle impacte largement le globe, par exemple la mousson asiatique, l'océan superficiel, le temps aux latitudes moyennes ou ce qui nous intéresse ici : la formation des cyclones tropicaux. La MJO est un phénomène propagatif vers l'Est,

1. http://fr.pissaro.re/

sur une période de l'ordre de 30 à 80 jours. La MJO a été décrite pour la première fois dans [Madden and Julian, 1971].

La figure 1 est une représentation schématique de la propagation de la MJO à l'équateur, autour du globe (pour une séquence d'environ 50 jours). La MJO est caractérisée par deux phases, présentes en même temps sur le globe et opposées en longitudes. La première est dite active. Il s'agit d'une zone d'anomalie de convection profonde. Elle s'accompagne d'une anomalie de basse pression au niveau de la mer et d'une anomalie de vents zonaux convergents en basses couches et divergents à la tropopause. On remarque que l'anomalie de convection est bien marquée lorsque cette phase active se situe au niveau de la zone indo-pacifique, mais disparaît au delà de l'océan Pacifique central, sur l'Amérique et l'océan Atlantique. Le signal en vent persiste toutefois. La deuxième phase est dite inactive. Il s'agit d'une zone d'anomalie de subsidence de grande échelle, associée à une anomalie de haute pression au niveau de la mer. Ces deux phases font le tour du globe en se déplaçant lentement vers l'est, à une vitesse de l'ordre de 5 $m.s^{-1}$. À chaque instant, ce schéma est plus ou moins marqué. On parle alors de MJO plus ou moins intense.

2.1.2 Détection et suivi

Plusieurs paramètres permettent de détecter et suivre le schéma conceptuel présenté plus haut.

En zone tropicale et subtropicale, on utilise l'OLR (Outgoing Longwave Radiation) comme traceur de la convection profonde. Il s'agit d'une mesure par satellite de la quantité d'énergie émise dans l'infrarouge vers l'espace par la surface terrestre, les océans et l'atmosphère. L'unité est le $watt.m^{-2}$. Les faibles valeurs d'OLR correspondent à des températures froides, à des sommets nuageux froids, donc à des nuages dont le sommet est élevé.

Un autre moyen de détecter les zones de convection est de regarder les vents zonaux à 850 hPa et à 200 hPa car ils permettent de rendre compte, respectivement, de la convergence en basses couches et de la divergence d'altitude. La variable utilisée pour regarder le vent à ces deux altitudes est la vitesse de la composante zonale du vent à 850 hPa, notée U850, et à 200 hPa, notée U200.

Une autre variable, le potentiel de vitesse, est également utilisé pour le diagnostic de la divergence ou de la convergence. Sa construction se base sur la propriété selon laquelle pour tout champ de vecteur $\vec{U}$ en 2 dimensions, il existe une décomposition unique

$$\vec{U}(x,y) = \vec{k} \wedge \vec{\nabla}\psi(x,y) + \vec{\nabla}\chi(x,y)$$

avec ψ la fonction de courant et χ le potentiel de vitesse. Le potentiel de vitesse du vent caractérise ainsi la composante divergente du champ de vent. Le potentiel de vitesse à 200 hPa (noté PV200) renseigne sur les zones favorables à grande échelle au développement de la convection profonde. Les zones de divergence d'altitude, c'est à dire de PV200 négatif, correspondent à des zones favorables à la convection profonde. En revanche, les zones de convergence en altitude, c'est-à-dire PV200 positif, correspondent à des zones défavorables à la convection profonde.

Étant donné que la MJO correspond à la propagation d'un signal qui vient s'ajouter aux conditions normales atmosphériques, il est plus pertinent de travailler à partir des anomalies de champs. On regarde ainsi classiquement les anomalies par rapport à la climatologie.

La MJO a une période allant de 30 à 80 jours. À ce signal se superposent les phénomènes de période plus courte (par exemple les ondes de Kelvin) ainsi que ceux de période

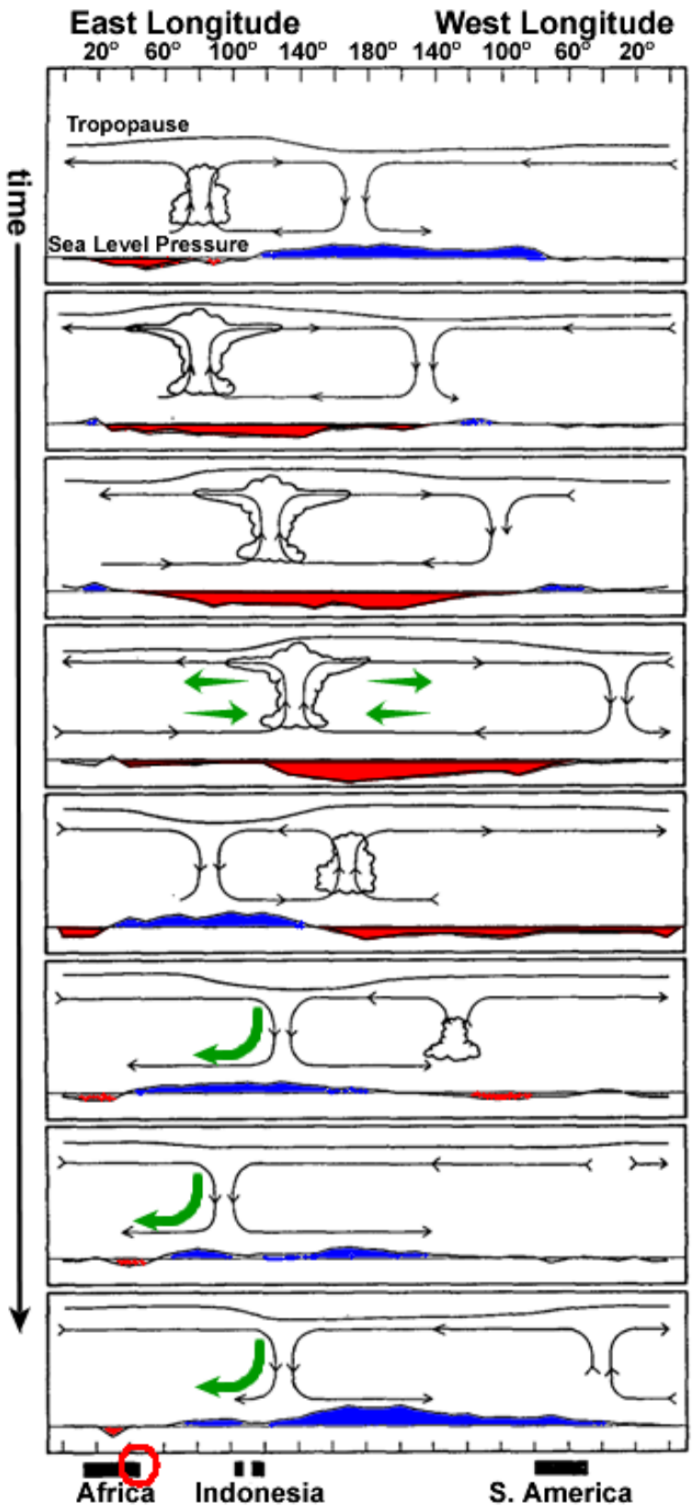


FIGURE 1 – Représentation schématique de la propagation de la MJO à l'équateur, autour du globe. Séquence d'environ 50 jours. *Source : cours de Frédéric Ferry, d'après [Madden and Julian, 1971].*

plus longue (par exemple le cycle saisonnier). Il est alors courant de filtrer en fréquence les champs que l'on regarde afin de se concentrer sur les échelles de temps auxquelles on s'intéresse, et libérer le signal des autres phénomènes. La visualisation de l'évolution de ces champs filtrés permet alors de repérer la propagation de la MJO.

2.1.3 Détection et suivi par l'indice RMM

Une autre méthode est très largement utilisée afin de détecter la MJO : le calcul de l'indice RMM. Elle a été élaborée par [Wheeler and Hendon, 2004] et permet de caractériser la MJO en temps réel ou en mode prévision à partir de données quotidiennes non filtrées. Cette méthode repose sur l'analyse de trois variables combinées : l'OLR, U850 et U200.

Dans la publication originale, les données d'OLR sont des moyennes quotidiennes d'observations satellites de la National Oceanic and Atmospheric Administration (NOAA), pour le mode analyse, ou du National Centers for Environmental Prediction (NCEP), pour le mode temps réel. Les données en vent zonaux sont les réanalyses du NCEP – National Center for Atmospheric Research (NCEP–NCAR), pour le mode analyse, ou les analyses du modèle opérationnel du Bureau of Meteorology (BoM) australien, pour le mode temps réel. L'ensemble de ces données a une résolution de 2.5° en latitude et en longitude. La période couverte s'étend de juin 1974 à fin 2003, excepté mars à décembre 1978. La période dite de référence s'étend de 1979 à 2001.

Ces données sont tout d'abord pré-traitées pour supprimer l'influence du cycle saisonnier. On soustrait pour chaque jour la climatologie du jour de l'année, calculée sur la période de référence. On supprime également les 3 premières harmoniques du cycle annuel. Les données sont ensuite traitées pour supprimer la variabilité interanuelle, en particulier l'influence de El Niño Southern Oscillation (ENSO). Le schéma d'ENSO présentant des similarités avec certains schémas de MJO, cela impliquerait de fausses détections si l'on laissait ce signal. Cela est effectué en 2 étapes. La première consiste à supprimer la partie du champ considéré (OLR, U850 ou U200) liée linéairement à l'indice SST1 de ENSO. Pour cela, une régression linéaire du champ est calculée puis soustraite pour chaque jour, à partir de la valeur de SST1. Cet indice SST1 est la série temporelle correspondant à la première composante principale (PC : Principal Component) d'une rotation d'ACP de la température de surface de la mer (SST : Sea Surface Temperature) sur l'indo-pacifique. Voir [Wheeler and Hendon, 2004] et [Drosdowsky and Chambers, 2001] pour plus de détails sur ce pré-traitement. La deuxième étape pour écarter la variabilité interannuelle consiste à supprimer chaque jour une moyenne glissante des 120 jours précédents. Finalement, la donnée est allégée en moyennant chacun des champs sur une bande de latitude 15°N-15°S, sur l'ensemble des longitudes.

Les trois champs obtenus avec ce traitement préalable sont ensuite décomposés de manière combinée par Analyse en Composantes Principales (ACP). Cette technique est aussi qualifiée de décomposition orthogonale aux valeurs propres, ou de décomposition en EOF (Empirical Orthogonal Functions). Une analyse en composante principale d'un signal permet de le décomposer selon les modes dominant sa variabilité. À chaque mode est alors associé une série temporelle, correspondant à chaque instant à la proportion du signal décrite par le mode correspondant. Pour le calcul de l'indice RMM, chacun des trois champs est tout d'abord normalisé par sa variance globale. Cela permet de s'assurer que chaque champ contribue de manière égale. Ces trois champs sont ensuite décomposés de manière combinée par ACP, en retenant les deux premiers modes de variabilité. Cette décomposition s'effectue à partir des données de la période de référence (1979-2001). Une

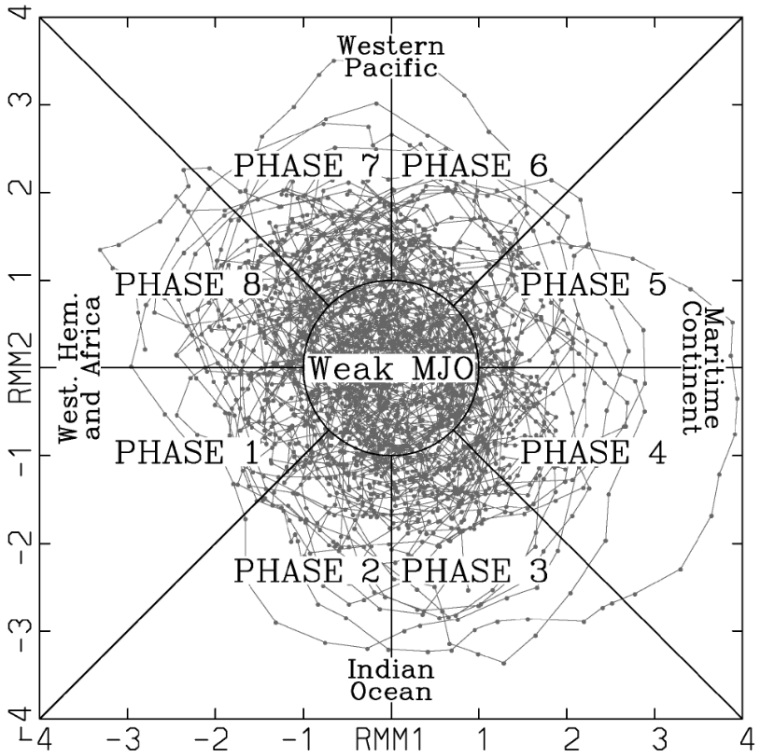


FIGURE 2 – Indice RMM journalier sur décembre, janvier et février (DJF) entre 1974 et 2003. *Source : [Wheeler and Hendon, 2004].*

FIGURE 3 – Position de la MJO selon sa phase. *Source : cours de Nicole Girardot.*

fois les deux modes identifiés, le signal combiné des trois champs est projeté sur chacun des deux modes, pour toute la période. On obtient finalement deux séries temporelles décrivant la variabilité des trois champs (OLR, U850 et U200) selon leurs deux modes principaux de variabilité. Ces deux séries temporelles sont notées RMM1 et RMM2, et constituent ce que l'on appelle l'indice RMM.

Une façon conventionnelle de visualiser l'indice RMM est de représenter son évolution temporelle sous forme de nuage de points sur un graphique de RMM2 en fonction de RMM1. On parle d'espace de phases. La figure 2, extraite de [Wheeler and Hendon, 2004], représente l'indice RMM journalier sur décembre, janvier et février (DJF) entre 1974 et 2003. D'une part, la distance au centre du graphique correspond à l'intensité de la MJO. Ainsi, un point dans le cercle de rayon 1 centré sur l'origine indique une MJO faible, ou peu intense. À l'inverse, un point à l'extérieur de ce cercle indique l'occurrence d'un phénomène marqué de MJO. On parle aussi de MJO intense. D'autre part, la position du point dans cet espace nous renseigne sur la localisation de la MJO sur le globe. L'espace

est en effet décomposé en 8 cadrants, correspondants aux 8 phases de la MJO. Chacune des phases correspond à une position de la phase active de la MJO sur le globe (figure 3). On note que les cadrants ne sont pas régulièrement répartis sur le globe. La propagation vers l'est de la MJO se traduit alors par un parcours dans le sens trigonométrique de l'espace des phases.

Il est utile de noter que cette représentation de la MJO n'est pas parfaite. En effet, l'indice RMM ne permet pas toujours de rendre compte correctement de l'état de la MJO, en particulier si elle est non canonique ou si elle est faible. Une autre approche est d'utiliser l'indice VPM (Velocity Potential MJO index), construit de manière similaire en remplaçant l'OLR par le potentiel de vitesse. En opérationnel, les indices RMM et VPM sont utilisés de manière complémentaire. Nous avons fait le choix pour ce projet de considérer l'indice RMM comme indicateur de la MJO.

2.2 L'intelligence artificielle

L'Intelligence Artificielle (IA) est définie par le Larousse comme un "ensemble de théories et de techniques mises en œuvre en vue de réaliser des machines capables de simuler l'intelligence humaine". Cette discipline n'admet pourtant pas unanimement de définition, de par ses domaines très vastes d'application. Bien qu'existant depuis les années 1950, l'IA a connu plusieurs vagues d'engouement. Depuis les années 2010, ce domaine est en plein essor, porté par l'augmentation rapide de la puissance de calcul et du volume de données disponible.

2.2.1 Intelligence artificielle pour la prévision météo-climatique

Le domaine des géosciences s'est naturellement penché sur les possibilités qu'offre l'IA. Dans le domaine de la prévision météo-climatique, l'IA ouvre certaines portes détaillées dans [Chantry et al., 2021].

La première approche est qualifiée de hard AI (Artificial Intelligence). Il s'agit de prédire l'état futur du système à partir des observations passées et présentes, en utilisant uniquement un modèle d'IA.

La deuxième approche est qualifiée de medium AI. Dans ce cas, le modèle numérique de prévision du temps est conservé, mais amélioré par des techniques d'IA. La première possibilité en medium AI est d'utiliser l'intelligence artificielle comme un outil de post-traitement des sorties du modèle. La deuxième possibilité est de remplacer une composante du modèle (une paramétrisation de la physique par exemple) par un modèle d'IA. Ce modèle aura été entraîné préalablement sur des observations ou sur un modèle plus performant (mais plus coûteux). L'idée est d'améliorer la performance du modèle sans en augmenter le coût.

La troisième approche d'utilisation de l'IA pour la prévision du temps est qualifiée de soft AI. Dans ce cas, une composante du modèle est remplacée par un modèle d'IA, mais cette fois-ci entraîné sur le modèle numérique lui-même. La performance de la prévision n'est donc pas améliorée directement, mais le coût du modèle est réduit. Ce coût gagné peut alors être réinvesti dans une autre composante du modèle pour en améliorer finalement sa performance, à coût égal.

Dans ce projet, nous avons opté pour une approche de medium AI, par post-traitement des sorties d'un modèle de prévision numérique du temps.

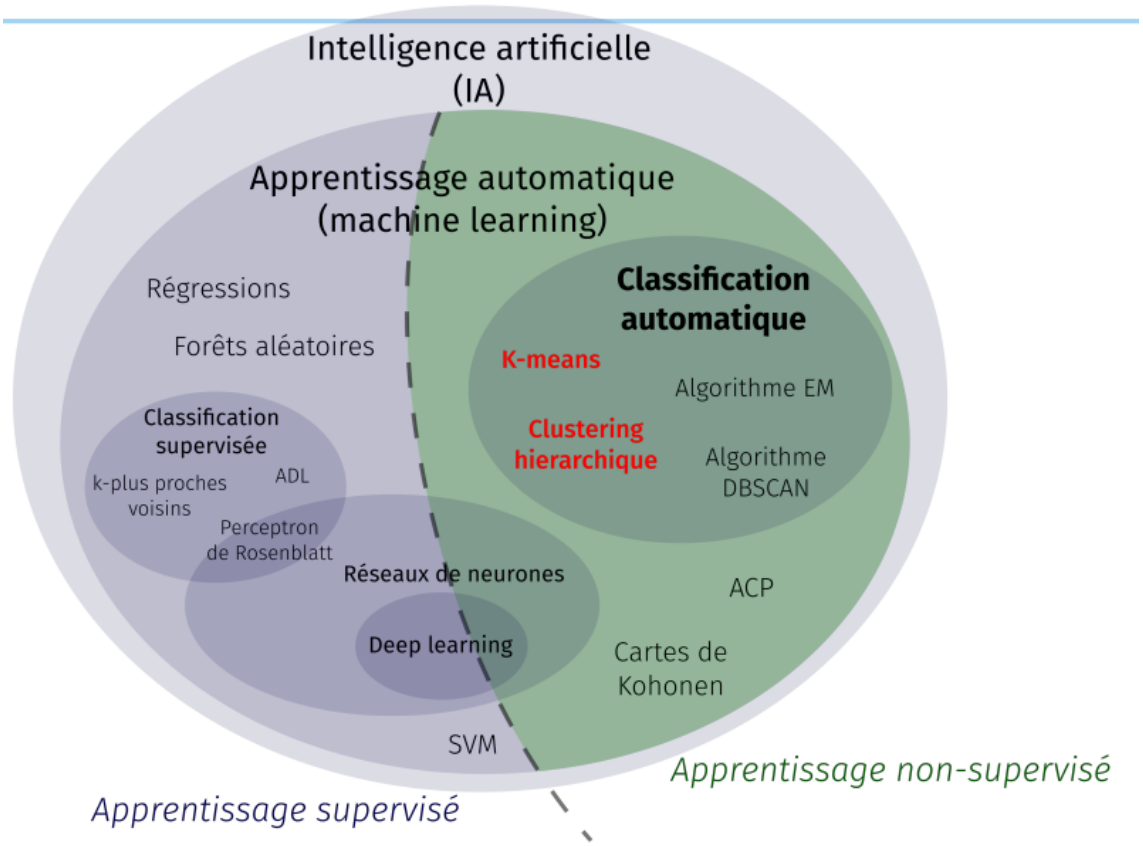


FIGURE 4 – Paysage de l'intelligence artificielle. *Source : cours de Thomas Rieutord.*

2.2.2 Paysage de l'intelligence artificielle

Le terme d'intelligence artificielle regroupe un ensemble de techniques très variées. La figure 4 en situe certaines dans ce paysage complexe de l'IA. Tout d'abord, l'apprentissage automatique (ou machine learning) désigne les techniques pour lesquelles l'algorithme apprend, s'affine au fur et à mesure de son utilisation et des cas qui lui sont présentés.

Plusieurs types d'apprentissage sont à distinguer concernant le machine learning. Le premier est dit supervisé. Dans ce cas, les données à partir desquelles le modèle apprend doivent être "étiquetées", cela signifie que l'information recherchée doit être connue, au moins sur une partie des données. Une fois le modèle entraîné sur un ensemble d'apprentissage (train), sa performance est testée sur un ensemble test pour évaluer sa capacité à retrouver l'étiquette associée à la donnée. Si le modèle est suffisamment performant, il peut être utilisé de manière opérationnelle, sur des données non étiquetées. L'apprentissage supervisé permet par exemple de faire de la classification ou de la régression (estimer un prédictand [2] à partir de un ou plusieurs prédicteurs [3]). Le deuxième type de machine learning est l'apprentissage non-supervisé. Dans ce cas, les données ne sont pas étiquetées, et l'information pertinente doit être extraite par le modèle sans sa connaissance préalable. Ces techniques permettent par exemple de faire de la réduction de dimension (par ACP par exemple) ou de la classification automatique (on parle de clustering). Un troisième type de machine learning est l'apprentissage par renforcement (reinforcement learning). Dans ce cas, l'algorithme doit réaliser une tâche donnée et évolue de manière autonome

2. Prédictand : variable que l'on cherche
3. Prédicteur : variable à partir de laquelle on cherche à prévoir la valeur du prédictand

pour y parvenir de manière efficace. C'est cette forme d'IA qui est développée pour le déplacement autonome de drônes ou de voitures par exemple.

Un sous-domaine à distinguer au sein de l'apprentissage automatique est celui des réseaux de neurones. Il s'agit de techniques mimant la structure d'un cerveau biologique. Les réseaux de neuronnes sont construits par succession de couches de neurones formels. Si le réseau comporte plus de deux couches, on dit que c'est un réseau de neurones profond et on parle d'apprentissage profond (deep learning). Les réseaux de neurones permettent de faire aussi bien de la classification que de la régression. Leur fonctionnement sera décrit plus en détail dans la partie 4.3.

Nous avons choisi de faire une régression de l'indice RMM (voir 2.1.3) à partir de différents prédicteurs. Nous avons utilisé pour cela différentes techniques d'apprentissage automatique supervisé, dont du deep learning. L'apprentissage s'effectue sur une partie seulement des données, l'échantillon d'apprentissage (ou échantillon train). L'évaluation du modèle s'effectue sur le reste des données, l'échantillon test.

2.3 État de l'art des méthodes de diagnostic et de prévision de la MJO

Différentes études ont été menées concernant le diagnostic de MJO par IA. [Dasgupta et al., 2020] propose une reconstruction de RMM du passé par régression à partir de champs de pression au niveau de la mer. [Dasgupta et al., 2020] compare plusieurs modèles : MultiLinear Regression (MLR), Support Vector Regressor (SVR) et Convolutive Neural Network (CNN). Le dernier donne les meilleurs résultats, avec une amélioration de 4% par rapport à la méthode de régression multilinéaire. Le RMM reconstruit a permis d'étudier la variabilité multi décennale de la MJO.

Une approche différente est développée dans [Toms et al., 2020], qui met en oeuvre une classification des épisodes intenses de MJO selon ses 8 phases. Le modèle utilisé est un perceptron multicouche (MLP : MultiLayer Perceptron), appliqué à un grand nombre de champs de réanalyses NASA MERRA-2. Deux méthodes d'interprétation des réseaux de neurones sont déployées et mettent en évidence la capacité du modèle à apprendre les situations météorologiques pertinentes pour cette classification. Cela permet de comprendre ce qui est important pour la prise de décision du modèle, et lève l'aspect "boîte noire" souvent reproché à l'IA. Ces résultats sont utilisés pour étudier la saisonnalité de la MJO.

D'autres études se sont intéressées à la qualité de prévision de la MJO. L'article [Kim et al., 2018] en constitue une revue. Il conclue qu'actuellement, les prévisions restent correctes jusqu'à la semaine 2 à 4, selon le modèle et l'état de l'atmosphère à l'initialisation. Or, toujours d'après [Kim et al., 2018], la prévisibilité théorique de la MJO est de 6 à 7 semaines, ce qui laisse une marge importante de progression. Notre étude s'inscrit dans ce contexte en gagnant en prévisibilité grâce à l'utilisation de méthodes d'IA.

3 Données

3.1 L'indice RMM du BoM

Nous avons fait le choix dans cette étude de considérer l'indice RMM de l'analyse du BoM comme la "vérité". On parle de RMM "observé". C'est la référence de laquelle nous souhaitons nous approcher. Nous travaillons directement avec les analyses

temps réel du BoM[4]. Elles sont calculées quotidiennement selon la méthodologie de [Wheeler and Hendon, 2004] (section 2.1.3).

L'indice RMM est renseigné de deux manières. Tout d'abord par les indices RMM1 et RMM2, coordonnées cartésiennes dans l'espace des phases. Ces coordonnées peuvent être passées en polaire, et ainsi renseigner à la fois sur la phase de la MJO (par la coordonnée angulaire en polaire) et sur son intensité (la coordonnée radiale en polaire). La phase et l'intensité de la MJO sont directement fournies par le BoM.

Nous disposons d'un indice RMM par jour, correspondant à la situation de la MJO sur l'ensemble du globe. La période couverte s'étend de juin 1974 au présent, excepté mars à décembre 1978.

3.2 Données de prévision intra-saisonnière (modèle global ensembliste ECMWF)

La base de données du projet S2S[5] met à disposition les sorties de plusieurs modèles numériques de prévision intra-saisonnière. Nous nous concentrons pour ce projet sur le modèle ensembliste S2S de ECMWF (European Centre for Medium-Range Weather Forecasts).

3.2.1 Temps réel

Nous décrivons ici la version temps réel[6] du modèle S2S global ECMWF Ensemble Prediction System (EPS).

Il s'agit d'un ensemble de 1 membre de contrôle et de 50 membres perturbés. Il tourne 2 fois par semaine depuis janvier 2015, les lundis et jeudis. Les échéances vont de 0 à 1104 heures (46 jours), par pas de 6 heures. Selon le paramètre, le niveau vertical et la version du modèle, les échéances disponibles se restreignent parfois à 768 heures (32 jours), par pas de 24 heures. La résolution spatiale native de son modèle atmosphérique sur l'horizontale est d'environ 16 km (Tco639 en spectral) jusqu'au jour 15, puis d'environ 31 km (Tco319 en spectral). Sa résolution spatiale archivée est de 1.5° sur l'horizontale. La résolution verticale native du modèle atmosphérique est de 91 niveaux sigma (qui suivent la surface), entre la surface et le niveau 0.01 hPa. La résolution temporelle native du modèle est de 12 minutes jusqu'au jour 15, puis de 20 minutes.

Nous travaillons avec une résolution spatiale horizontale de 1.5°, à différents niveaux verticaux, sur un domaine 30°N-30°S pour toutes les longitudes et pour les échéances de 0 à 768 heures (32 jours, i.e. prévision mensuelle) par pas de 24 heures.

3.2.2 Reforecast

La version de re-prévision (reforecast) du modèle S2S global ECMWF EPS est identique dans ses caractéristiques à sa version temps réel, excepté le nombre de ses membres. Elle comporte 1 membre de contrôle et 10 membres perturbés (au lieu de 50).

Les reforecasts de ce modèle sont calculés "au fil de l'eau" ("on the fly"). C'est à dire qu'à chaque fois que le modèle tourne en temps réel, 2 fois par semaine, la version actuelle du modèle est relancée pour cette date de l'année, sur les 20 années précédentes. Les

4. https://iridl.ldeo.columbia.edu/SOURCES/.BoM/.MJO/.RMM/
5. http://s2sprediction.net/
6. https://apps.ecmwf.int/datasets/data/s2s-realtime-instantaneous-accum-ecmf

reforecasts permettent alors d'éviter les inhomogénéités liées aux changements de version du modèle.

Notre ensemble de reforecast se base sur la version CY46R1 du modèle [7] (opérationnelle du 11/06/2019 au 29/06/2020). Nous disposons donc des reforecasts au fil de l'eau calculés entre juillet 2019 et juin 2020, qui couvrent la période de juillet 1999 à juin 2019. Pour compléter l'année 2019, nous utilisons également les données temps réel (le membre de contrôle et les 10 premiers membres perturbés) de juillet à décembre 2019.

Finalement, nous disposons d'un ensemble homogène en terme de version (CY46R1) et de nombre de membres (1 contrôle et 10 perturbés), avec deux simulations par semaine entre 2000 et 2019. Dans le temps imparti au stage, nous avons exploité uniquement le membre de contrôle. Nous entraînons nos modèles statistiques sur les données de 2000 à 2014 (échantillon train) et les évaluons sur les années 2015 à 2019 (échantillon test).

3.2.3 Variables

Les variables du modèle seront pour nous, dans l'approche IA de notre projet, les prédicteurs de la régression.

Outgoing Longwave Radiation (OLR), en $W.m^{-2}$ Comme décrit dans la partie 2.1.2, en tant que traceur de la convection profonde, l'OLR nous renseigne sur la MJO. Cette mesure, à l'origine satellitaire, est simulée par le modèle. La variable disponible est la Top net Thermal Radiation (TTR). Or, $OLR = -TTR$. Le modèle donne une valeur cumulée ($W.m^{-2}.s$). Pour correspondre à la valeur instantanée d'OLR, la différence de valeur entre deux échéances successives est divisée par le nombre de secondes les séparant (dans notre cas : 1 jour, donc 86400 secondes). On obtient alors une valeur moyenne entre les deux échéances. On choisit de l'associer à l'échéance suivante plutôt qu'à la précédente. Pour l'échéance 0, on décide d'utiliser la même valeur que pour l'échéance 1 (24 heures).

Vents zonaux à 850 hPa (U850) et à 200 hPa (U200), en $m.s^{-1}$ Les vents zonaux sont un moyen de rendre compte de la convergence ou de la divergence, et permettent de caractériser la convection. Pour le modèle de ECMWF, ces champs sont disponibles en valeurs instantanées aux échéances qui nous intéressent. Il s'agit de champs instantanés (0H UTC).

Eau totale sur la colonne (TCW : Total Column Water), en $kg.m^{-2}$ L'eau totale sur la colonne représente en kg le contenu total d'eau, peu importe sa phase, sur l'intégralité de la colonne d'air à la verticale d'une surface. Contrairement à l'OLR, à U850 et à U200, cette variable n'intervient pas dans le calcul de l'indice RMM par ACP. En revanche, une forte valeur de TCW traduit un fort contenu en eau, et donc potentiellement une forte convection profonde. Il s'agit de moyennes journalières.

3.2.4 Indice RMM du modèle

Le projet S2S met à disposition certains indices issus du post-traitement de leurs modèles [8]. L'indice RMM est ainsi disponible pour le modèle global ensembliste intra-saisonnier de ECMWF. Quatre informations sont fournies : RMM1, RMM2, intensité

7. Description des versions du modèle : https://confluence.ecmwf.int/display/S2S/ECMWF+Model
8. ftp://s2sidx@acquisition.ecmwf.int/

et phase de la MJO. Cet indice RMM est calculé selon une méthode proche de celle de [Wheeler and Hendon, 2004], décrite en partie 2.1.3, et détaillée dans [Gottschalck et al., 2010]. Notons que la climatologie est calculée dans le cadre du projet S2S sur les données des modèles eux-même, et non pas sur les réanalyses NCEP-NCAR comme c'est la cas dans [Gottschalck et al., 2010]. La référence pour cette climatologie est la période 1999-2010. Notons également que la suppression de la variabilité associée à ENSO par régression linéaire sur l'indice SST1 (voir partie 2.1.3) n'est pas effectuée. La fenêtre glissante de 120 jours a été jugée suffisante par [Lin et al., 2008].

Comme pour les sorties directes du modèle, nous disposons de 20 ans (2000-2019) de données RMM homogènes issues de la version CY46R1.

Cet indice RMM, calculé de manière classique à partir des données du modèle à différentes échéances, constituera pour nous un point de comparaison. La performance de cet indice simulé, comparé à l'indice observé décrit en 3.1, sera la performance que nous souhaitons approcher, voire dépasser, par les méthodes d'IA mises en oeuvre. Le calcul classique de RMM est lourd. Utiliser un modèle d'IA bien entraîné, s'il est performant, sera plus léger

4 Méthodes

4.1 Normalisation et scores

Avant d'appliquer un modèle statistique, chaque prédicteur est normalisé. Cette normalisation permet au modèle d'avoir des valeurs comparables d'un prédicteur à l'autre. C'est une technique nécessaire à la convergence correcte de la majorité des modèles. D'un point de vue statistique, nous appliquons une standardisation, qui consiste à centrer et réduire la distribution. Pour tout élément x_i de l'échantillon d'apprentissage X_{train} ou de l'échantillon test X_{test}, nous remplaçons sa valeur par

$$\frac{x_i - \overline{X_{train}}}{\sigma(X_{train})}$$

avec $\overline{X_{train}}$ la moyenne de l'échantillon d'apprentissage et $\sigma(X_{train})$ son écart type [9]. Notons que la standardisation est toujours effectuée avec la moyenne et l'écart type de l'échantillon train, pour que le modèle n'ait aucune information sur l'échantillon test pendant la phase d'apprentissage. Cette opération est répétée pour chacun des prédicteurs. Il est souvent utile de normaliser également les prédictands. Dans notre cas, ce sont les RMM, construits par ACP. Ils sont donc déjà normalisés, par définition de l'ACP.

Pour évaluer la qualité de notre régression, l'indice RMM prévu par le modèle statistique est confronté à l'indice RMM observé du BoM. L'indice RMM prévu directement par le modèle de ECMWF est également confronté à l'indice RMM observé. Cette évaluation s'effectue par un ensemble de scores, présentés en annexe A.

4.2 Modèles de machine learning

Nous avons confronté plusieurs modèles de machines learning à notre problème. Le principe général d'IA est présenté en section 2.2. Seuls les modèles performants dans le

9. L'écart type $\sigma(X)$ d'un échantillon X est une mesure de la dispersion de ses valeurs. $\sigma(X) = \sqrt{\overline{(x_i - \overline{X})^2}} = \sqrt{\overline{x_i^2} - \overline{X}^2}$

cadre de notre étude sont présentés ici. Nous avons mis à l'épreuve de notre problème d'autres modèles, qui ne se sont pas avérés performants : un modèle de régression par k plus proches voisins (k-NN : k-Nearest Neighbors) et un modèle de forêt aléatoire (random forest).

Une part non négligeable du travail de ce stage a consisté à la mise en place d'un environnement de travail de machine learning. Cette thématique est en effet nouvelle pour l'équipe Cyclones du LACy. Tout cet aspect technique a été documenté pour accompagner au mieux les futurs membres de l'équipe qui s'intéresseront à ce domaine.

4.2.1 Régression linéaire

Le modèle de machine learning le plus simple est celui de la régression linéaire. Il est classiquement confronté aux modèles plus complexes pour évaluer si ces méthodes plus coûteuses sont utiles dans le cadre d'un problème donné. L'objectif de ce modèle est de chercher la relation linéaire optimale entre les prédicteurs et le prédictand.

Dans notre problème, ce modèle se décrit par les relations :

$$\widehat{RMM}1 = w_0 + \sum_{j=1}^{n} w_j X_j$$

$$\widehat{RMM}2 = w'_0 + \sum_{j=1}^{n} w'_j X_j$$

avec X_j le prédicteur. n correspond au nombre total de prédicteurs. Les coefficients w_0,...,w_n et w'_0,...,w'_n sont les paramètres du modèles. Les paramètres optimaux sont déterminés lors de la phase d'entraînement (train). Lors de cette phase d'entraînement du modèle statistique, sa performance est évaluée par une fonction coût. L'objectif est de minimiser sa valeur. L'algorithme de régression linéaire simple que nous utilisons a pour fonction coût l'erreur quadratique moyenne (MSE), décrite dans l'annexe A. Les valeurs optimales des paramètres du modèle permettant de minimiser la fonction coût sont déterminés grâce à un algorithme d'optimisation. L'algorithme utilisé ici est celui des moindre carrés, qui consiste à minimiser la somme des carrés des résidus (la différence entre l'indice estimé et l'indice observé). Cet algorithme d'optimisation repose sur l'inversion de matrices et n'est pas itératif. Nous utilisons l'implémentation LinearRegression de la bibliothèque python scikit-learn [10] (aussi appelée sklearn).

Notons qu'une variante au modèle linéaire classique existe et est implémentée dans sklearn : la régression Ridge. Le principe est le même mais en appliquant une régularisation L^2. Cela consiste à ajouter à la fonction coût le carré de la norme L^2 des paramètres du modèle [11], pondérée par un coefficient : $\alpha||W||_2^2$, avec W un vecteur des paramètres du modèle. Cela contraint le modèle à garder des paramètres relativement faibles, et permet ainsi de réduire le sur-apprentissage. Le coefficient α est un hyperparamètre [12] du modèle Ridge : sa valeur n'est pas déterminée lors de l'apprentissage mais fixée au préalable par l'utilisateur.

10. https://scikit-learn.org/

11. La norme L^2 de W est définie comme ceci : $||W||_2 = \sqrt{|w_0|^2 + ... + |w_n|^2 + |w'_0|^2 + ... + |w'_n|^2}$. On parle également de norme euclidienne.

12. Certains paramètres sont à fixer par l'utilisateur. C'est ce qu'on appelle les hyperparamètres, en opposition avec les paramètres appris de manière automatique.

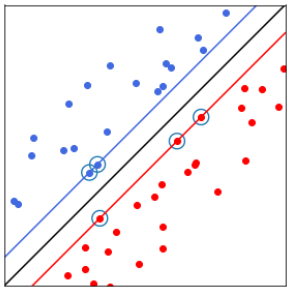


FIGURE 5 – Illustration de la SVM pour deux classes (bleu et rouge). Vecteurs supports cerclés en bleu clair. Droites supports, en bleu et rouge. Droite séparatrice, en noir. *Source : Data Analytics Post d'après Julien Audiffren.*

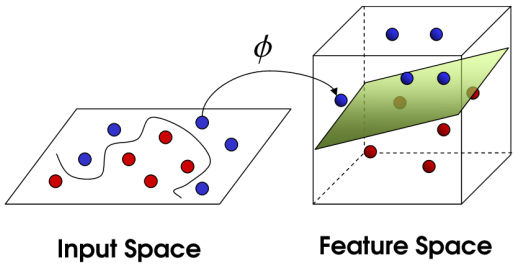


FIGURE 6 – Illustration de noyau de SVM. À gauche : espace d'entrée, et séparateur non linéaire. À droite : espace obtenu après transformation par le noyau ϕ et séparateur linéaire. *Source : article de Drew Wilimitis sur towardsdatascience.com.*

4.2.2 Support Vector Regressor (SVR)

Le Support Vector Regressor est un modèle de machine learning qui donne des résultats satisfaisants dans le cadre de notre étude. Cette méthode de régression est une variante d'une méthode de classification : la Support Vector Machine (SVM), également appelée en français Séparateur à Vaste Marge.

Une SVM permet la classification binaire grâce à un séparateur. La figure 5 présente son principe dans le cas de données en 2 dimensions. Dans cet exemple, deux classes sont à séparer : la classe bleue et la classe rouge. Chaque point représente un élément (un vecteur) de notre échantillon d'apprentissage, étiqueté par sa classe, bleu ou rouge. Pour chacune des deux classes, on identifie les vecteurs les plus proches de l'autre classe : ce sont les vecteurs supports, cerclés sur la figure en bleu clair. Les droites supports, bleu et rouge, sont parallèles entre elles et passent par les vecteurs supports. Le séparateur, la droite noire, est placé à égale distance de ces droites support. Il déterminera pour l'échantillon test la classe de chaque élément. La distance entre les deux droites supports est appelée la marge. L'objectif est de déterminer le séparateur qui permette la meilleure généralisation, c'est à dire les meilleurs résultats lors du passage de l'échantillon train à l'échantillon test. Pour ce faire, le problème d'optimisation est celui de la maximisation de la marge, qui garantit un séparateur optimal. Les vecteurs supports sont ainsi déterminés de manière à satisfaire à cette maximisation.

On notera que le séparateur est linéaire : c'est un hyperplan. Dans de nombreux problèmes, la séparation entre les classes n'est pourtant pas linéaire. Il est alors possible d'appliquer à l'espace d'entrée une transformation par un noyau, en augmentant le nombre de dimensions (figure 6). L'objectif de cette transformation est de passer à un espace dans lequel la séparation linéaire par SVM est possible. Un noyau couramment utilisé avec les SVM est le noyau RBF (Radial Basis Function). Notons qu'il prend un hyperparamètre γ. Dans l'implémentation sklearn, la valeur par défaut est

$$\gamma = \frac{1}{N * Var(X)}$$

avec N le nombre d'éléments dans l'échantillon train et $Var(X)$ la variance des prédictands dans cet échantillon. L'implémentation des SVM dans sklearn applique également

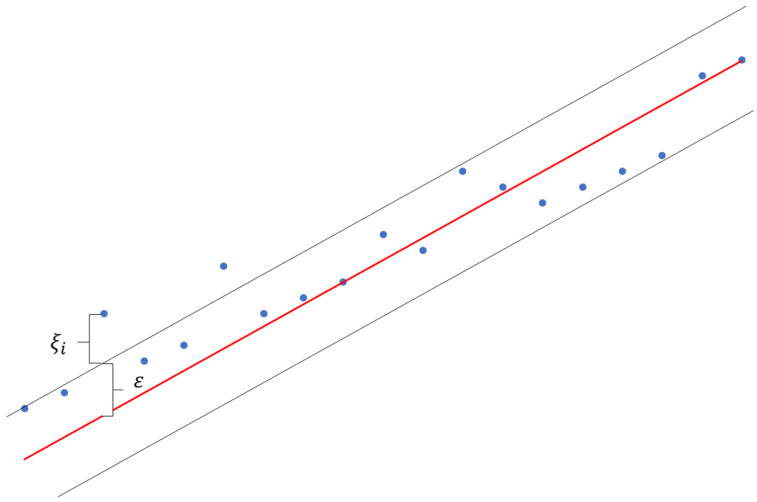


FIGURE 7 – Illustration de SVR. Les points bleus sont les données à modéliser. La droite rouge est la droite de régression. Les deux droites noires représentent le cône de régression, de diamètre ϵ. Les points en dehors de ce cône pénalisent le modèle par une erreur ξ_i. *Source : article de Tom Sharp sur towardsdatascience.com.*

une régularisation L^2 (voir section 4.2.1), avec pour hyperparamètre le coefficient noté C (égal à 1 par défaut).

L'objectif du SVR (figure 7 pour du 2D) est d'ajuster au mieux la droite de régression (en rouge) sur les données, en définissant une zone de tolérance (ligne fines grises). On parle également de cône de régression. Cette zone est définie par une distance ϵ de part et d'autre de la droite de régression (en rouge). ϵ est un hyperparamètre du modèle, par défaut à 0.1 dans l'implémentation SVR de sklearn. Les vecteurs étant dans cette zone de tolérance sont considérés comme correctement prévus par la régression et ne sont pas comptabilisés par la fonction coût. Les potentiels vecteurs supports sont dans le cas du SVR les points à l'extérieur de la zone de tolérance. Ils sont déterminés par un algorithme d'optimisation de manière à maximiser le nombre de points dans la zone de tolérance tout en minimisant les erreurs ξ_i. De la même manière que pour la classification par SVM, il est toujours possible de se ramener à un problème linéaire par application d'un noyau.

4.3 Modèles de deep learning (machine learning par réseau de neurones)

4.3.1 Neurone formel

Les réseaux de neurones profonds, ou modèles de deep learning, constituent un sous ensemble des modèles de machine learning. Leur construction repose sur l'utilisation de neurones formels.

La figure 8 schématise le fonctionnement d'un neurone formel. Les m entrées, représentées en bleu, passent d'abord par une fonction affine :

$$s = \sum_{i=1}^{m} w_i x_i + b$$

avec x_i l'entrée, w_i le poids qui lui est associé, b le biais et s la sortie de la fonction. On applique à cette sortie une fonction d'activation f pour obtenir en sortie de neurone

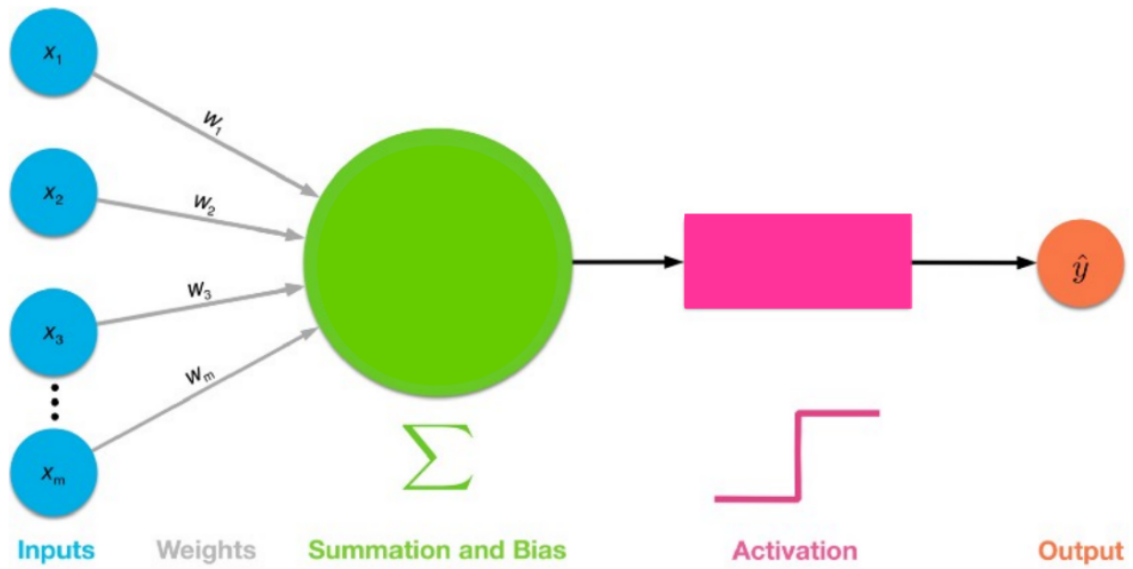
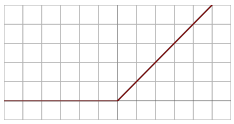


FIGURE 8 – Neurone formel. *Source : MOOC Deep Learning de Nicolas Thome.*

$\widehat{y} = f(s)$. La fonction d'activation a pour rôle de faire basculer le neurone dans une état soit activé soit désactivé, selon le dépassement ou non d'un seuil déterminé par le biais b. Plusieurs fonctions d'activation sont possibles. Nous utilisons une fonction populaire, la ReLU (Rectified Linear Unit) :

$$f(x) = \begin{cases} 0 & \text{si} \quad x < 0 \\ x & \text{si} \quad x \geq 0 \end{cases}$$



La fonction d'activation est essentielle pour permettre au réseau neuronal d'approximer tout type de fonction non linéaire. En utilisant la fonction identité, le neurone est équivalent à une modèle de régression linéaire. C'est toutefois utile lorsque l'on souhaite faire de la régression par exemple, pour les neurones de la couche de sortie.

Un neurone formel seul permet déjà de faire la régression ou de la classification, mais ne permet pas la non linéarité. En assemblant plusieurs neurones en couches successives, le modèle devient non linéaire.

4.3.2 Perceptron multicouche (MLP : MultiLayer Perceptron)

Un perceptron multicouche est un réseau de neurones complètement connecté. Plusieurs couches de plusieurs neurones se succèdent. Au sein d'une couche, chaque neurone est connecté à tous les neurones de la couche précédente et à tous les neurones de la couche suivante.

La figure 9 présente un exemple simple de MLP. Sur cet exemple, le réseau prend 3 valeurs en entrée. La première couche cachée contient 8 neurones. Chacun des neurones prend en entrée les 3 valeurs de la couche d'entrée, applique la fonction affine et la fonction d'activation puis renvoie en sortie la valeur résultante. La deuxième couche cachée contient 4 neurones. Ils prennent chacun pour entrée les sorties des 8 neurones de la couche précédente. Leurs sorties sont ensuite envoyées à la dernière couche du réseau : la couche de sortie. Elle comprend autant de neurones que de sortie désirée, dans notre cas 2 : un pour la valeur de RMM1 et un pour la valeur de RMM2. Les neurones des couches cachées

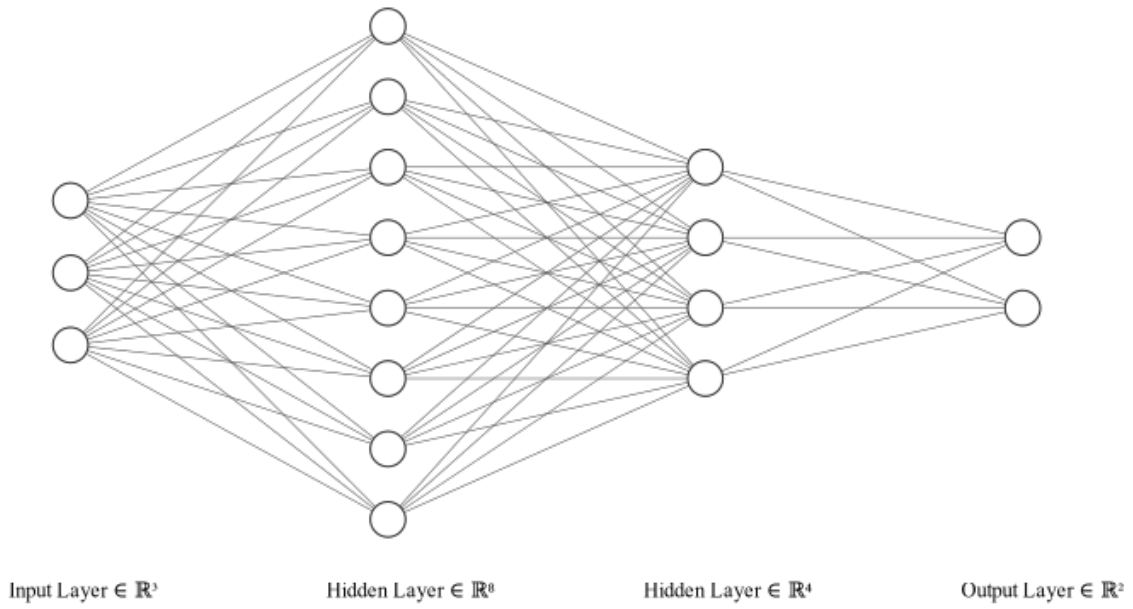


FIGURE 9 – Représentation d'un perceptron multicouche simple. Couche de 3 entrées, première couche cachée de 8 neurones, deuxième couche cachée de 4 neurones, couche de sortie de 2 neurones.

doivent appliquer une fonction d'activation, pour permettre la non linéarité. En revanche, les deux neurones de la couche de sortie ne doivent pas avoir de fonction d'activation, pour permettre la régression de l'ensemble des valeurs réelles ; on applique une fonction d'activation identité, ou linéaire.

Les paramètres de ce modèle d'apprentissage sont les poids et les biais des fonctions affines de chaque neurone (voir section 4.3.1). Ces paramètres sont initialisés de manière aléatoire. L'objectif de l'apprentissage est de déterminer les paramètres optimaux afin que le modèle retrouve pour chaque situation (chaque date donnée) la configuration de MJO correspondante (RMM1 et RMM2 proches de la vérité). Cette correspondance entre régression et vérité est quantifiée par une fonction coût. Le jeu complet d'entraînement est divisé en un certain nombre de sous-ensembles appelées mini-batchs. À chaque itération, le modèle est entraîné sur un mini-batch. La fonction coût est calculée une fois que le modèle a parcouru un mini-batch complet et les poids et biais du modèle sont mis à jour. La taille du mini-batch a un impact sur la rapidité et la variance de convergence du modèle. L'utilisation de mini-batch permet également de réduire le sur-apprentissage, c'est à dire d'améliorer la capacité de généralisation du modèle, donc de moins perdre en performance lors du passage de la phase d'entraînement à la phase test.

Ce modèle présente plusieurs hyperparamètres. Pour l'optimisation de l'apprentissage, nous avons choisi l'algorithme adam. La taille du mini batch a été fixée à 32 et le nombre d'époques [13] à 200, suffisant à la convergence du réseau. Nous avons choisi une fonction coût que nous avons implémentée : la distance euclidienne avec seuil (voir annexe A). Elle permet au modèle de bien prévoir l'indice RMM lorsque la MJO est non intense, pour accorder plus d'importance à la bonne prévision de cas de MJO intenses. Le nombre et la taille des couches cachées est également un hyperparamètre. Nous avons déterminé la

13. Le nombre d'époques correspond au nombre de fois où l'algorithme voit l'échantillon d'apprentissage en entier. Ainsi, chaque fois que l'algorithme a vu tous les éléments de l'échantillon d'apprentissage, une époque s'est terminée.

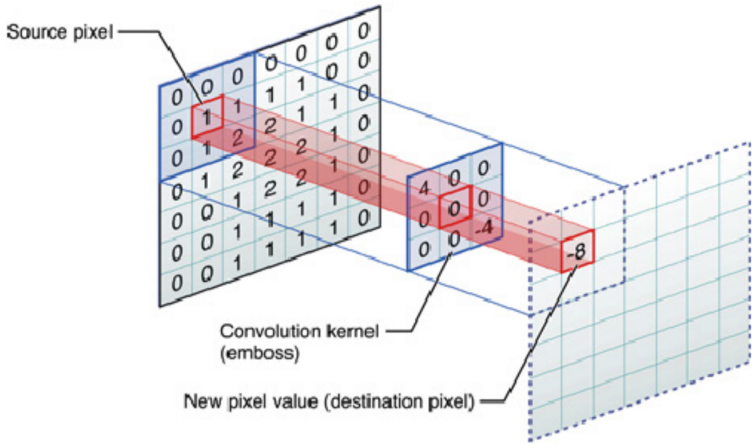


FIGURE 10 – Opération de convolution par un filtre 3x3. *Source : MOOC Deep Learning de Nicolas Thome.*

forme de réseau optimale pour notre étude grâce à la librairie AutoKeras [14]. Le réseau a quant à lui été implémenté avec la librairie Keras [15].

Finalement notre perceptron multicouches est construit comme ceci :

— couche d'entrée de taille 39360 (240 longitudes $*$ 41 latitudes $*$ 4 prédicteurs) ;
— 1ère couche cachée : 128 neurones, fonction d'activation ReLU ;
— 2ème couche cachée : 16 neurones, fonction d'activation ReLU ;
— couche de sortie : 2 neurones, fonction d'activation linéaire.

4.3.3 Réseau de neurones convolutif 2D (CNN 2D)

Les réseaux de neurones complètement connectés ne tirent pas parti a priori de la corrélation spatiale des données, comme par exemple de la cohérence spatiale de la MJO. Une façon d'y remédier est d'appliquer une ou plusieurs opérations de convolution avant de passer par un réseau complètement connecté. On parle de réseau convolutif, ici en 2 dimensions puisque l'on travaille sur des surfaces.

L'opération de convolution consiste a balayer l'ensemble des champs par un filtre (une matrice de dimension 2), de taille donnée. Il s'agit d'un produit de convolution entre deux matrices (figure 10). Les coefficients de cette matrice sont les coefficients multiplicateurs. Ils sont initialisés de manière aléatoire et c'est sur eux que porte l'apprentissage. Dans un réseau de neurones convolutif, ces opérations de convolution peuvent être effectuées en même temps, en faisant ainsi grandir la profondeur du tenseur de données. Elles peuvent également s'enchaîner de manière successives, en réduisant ainsi légèrement à chaque étape la taille des données en 2 dimensions (par perte sur les bords). Les hyperparamètres de ce modèle sont la fonction coût (distance euclidienne avec seuil pour notre modèle), la taille du mini batch (32), le nombre d'époques (200) et l'optimiseur (Stochastic Gradient Descent, SGD). La forme du réseau optimale est également un hyperparamètre et a été déterminée grâce à la librairie AutoKeras. Le réseau a ensuite été implémenté avec la librairie Keras. Il est construit comme ceci :

— couche d'entrée de taille 39360 (240 longitudes $*$ 41 latitudes $*$ 4 prédicteurs) ;
— 1ère couche de convolution : 16 filtres 3x3, activation ReLU ;
— 2ème couche de convolution : 64 filtres 3x3, activation ReLU ;

14. https://autokeras.com/
15. https://keras.io/

— dropout [16] 0.5.
— 3ème couche de convolution : 32 filtres 3x3, activation ReLU ;
— 4ème couche de convolution : 64 filtres 3x3, activation ReLU ;
— dropout de 0.5 ;
— 1ère couche cachée : 32 neurones, fonction d'activation ReLU ;
— 2ème couche cachée : 16 neurones, fonction d'activation ReLU ;
— couche de sortie : 2 neurones, fonction d'activation linéaire.

Dans la section suivante, nous présentons les résultats obtenus avec ces différentes méthodes d'IA.

5 Résultats

Nous présentons dans cette partie les résultats que nous avons obtenus. Dans le temps imparti au stage, nous nous sommes concentré sur l'étude du membre de contrôle du modèle ECMWF ensembliste. Dans un premier temps, nous avons travaillé en mode analyse, c'est à dire à l'échéance 0 du modèle. Dans un second temps, nous avons travaillé en mode prévision, jusqu'à l'échéance à 32 jours du modèle.

5.1 En mode analyse

Cette première étape de l'étude s'effectue en mode analyse, à l'échéance 0 du modèle. Nous commençons par une comparaison de deux jeux de RMM classiques, les deux utilisent des produits proches des observations mais sont néanmoins différents. La section 5.1.1 montre ces différences. Dans cette section, on n'utilise pas encore de méthode d'IA. Dans un second temps, nous mettons en oeuvre un certain nombre de méthodes d'IA permettant de diagnostiquer l'indice RMM à partir d'un ensemble de champs du modèle. Dans la section 5.1.2 c'est la première fois qu'on utilise une méthode d'IA, dans le contexte le plus simple. On n'utilise pas encore la "vérité" (RMM du BoM), mais seulement la vérité au sens ECMWF (RMM de ECMWF à l'échéance 0). Le modèle obtenu permettrait de remplacer le calcul lourd de RMM. Enfin, dans la section 5.1.3, on crée un modèle qui permet de retrouver des RMM indépendants du modèle ECMWF, ceux du BoM (qu'on considère comme notre vérité). L'enjeu est d'étudier le comportement des modèle d'IA dans ce cas. Dans l'ensemble de cette partie en mode analyse, nous effectuons la régression à partir des trois variables utilisées dans le calcul classique de l'indice RMM : OLR, U850 et U200 (sections 2.1.3 et 3.2.3 et article [Wheeler and Hendon, 2004]).

5.1.1 Référence : comparaison entre RMM du BoM et RMM de l'archive S2S

Afin d'évaluer la capacité du modèle de PNT à diagnostiquer correctement l'indice RMM de la MJO, nous comparons ses sorties à l'échéance 0 à l'analyse du BoM. Cela concerne l'ensemble de la période 2000 - 2019, avec 2100 éléments. Nous commençons par visualiser la dispersion de l'ensemble des évènements dans l'espace des phases. La figure 11 présente la dispersion de l'analyse du BoM. Les évènements sont bien répartis sur l'ensemble des phases. On note un grand nombre d'évènements non intenses, et peu d'évènements très intenses. La figure 12 présente la dispersion de l'analyse du modèle de

16. Le dropout consiste à désactiver de manière aléatoire un certain pourcentage de neurones lors de l'apprentissage (ici 50%). Cela permet de réduire le sur-apprentissage.

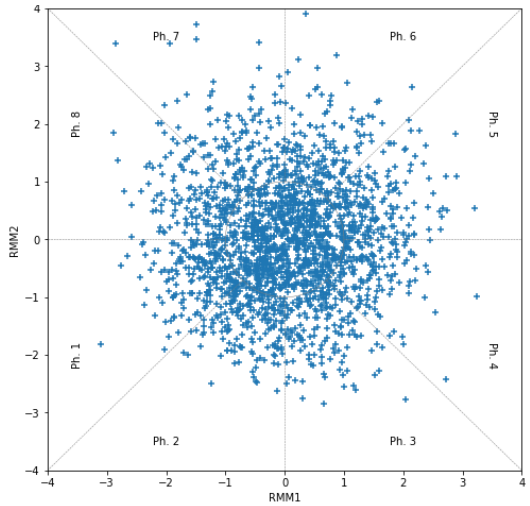


FIGURE 11 – RMM de l'analyse du BoM

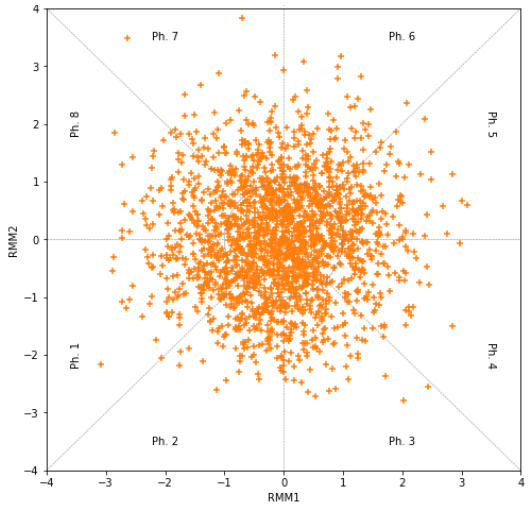


FIGURE 12 – RMM de l'analyse du modèle de ECMWF

MAE	0.33
MSE	0.18
Distance euclidienne	0.52
Distance euclidienne avec seuil	0.39
BRMSE	0.60
BCORR	0.91
Erreur sur l'amplitude	-0.05
Erreur absolue sur l'amplitude	0.30
Erreur sur l'angle	-2.8°
Erreur absolue sur l'angle	21°
Non intenses identifiés	78%
Intenses identifiés	82%
Précision	72%
Précision à une phase près	99.6%

Tableau 1 – Scores de l'analyse du modèle ECMWF comparée à l'analyse du BoM

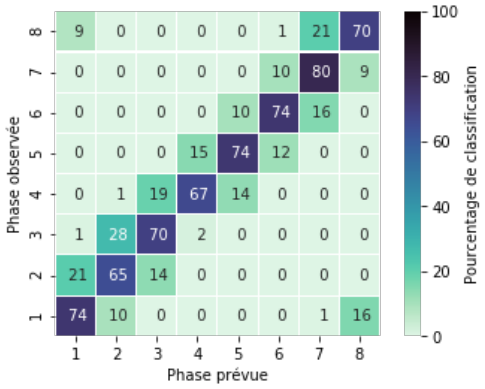


FIGURE 13 – Matrice de confusion de l'analyse du modèle ECMWF comparée à l'analyse du BoM

ECMWF. Dans l'ensemble, cette dispersion est similaire à celle de l'analyse du BoM. En revanche, si on regarde chaque évènement pris indépendemment, ils ne sont pas identiques dans les deux cas.

Pour quantifier cette analyse, nous avons calculé des scores dont les valeurs sont présentées dans le tableau 1. Le calcul de ces scores est détaillé en annexe A. Les scores MAE et MSE sont difficilement interprétables sans référence. Les deux scores calculés ici (respectivement 0.33 et 0.18) serviront de référence dans la suite de notre étude. La distance euclidienne peut être appréhendée grâce à la visualisation dans l'espace des phases (figure 2). Ici, un score de 0.5 correspond à une erreur déjà non négligeable. La distance euclidienne avec seuil, que nous utilisons plus tard pour l'apprentissage, est plus faible que la distance euclidienne. Cela découle du fait que les erreurs qui restent dans le cercle de MJO non intenses ne sont pas pénalisantes. La BRMSE, bien que non nulle, reste bien en dessous du seuil de $\sqrt{2} \approx 1.41$ au delà duquel l'information fournie par le modèle n'est plus pertinente. De même pour la BCORR, qui reste supérieure au seuil de 0.5, sans toutefois atteindre 1. L'erreur sur l'amplitude nous indique un léger biais de sous-estimation de l'intensité de la MJO. L'erreur sur l'angle nous indique un biais négatif du modèle, donc une tendance à un retard de MJO dans son parcours d'ouest en est (parcours trigonométrique de l'espace des phases). Le pourcentage d'évènements de MJO non intenses identifiés correspond au pourcentage d'évènements diagnostiqués non intenses par le modèle parmi ceux effectivement diagnostiqués non intenses par l'analyse du BoM. De la même manière, le pourcentage d'évènements de MJO intenses identifiés correspond au pourcentage d'évènements diagnostiqués intenses par le modèle parmi ceux effectivement diagnostiqués intenses par l'analyse du BoM. La performance d'un de ces scores ne doit pas se faire au détriment de l'autre. Ici, les deux pourcentages sont proches et relativement élevés, ce qui est satisfaisant. C'est en accord avec l'erreur sur l'amplitude quasi nulle. La précision correspond au pourcentage d'évènements pour lesquels la phase a été diagnostiquée correctement, pour les évènements à la fois observés (BoM) et diagnostiqués (ECMWF) intenses. Nous obtenons ici une précision de 72%. En opérationnel, une information avec un erreur de plus ou moins une phase est déjà intéressante et donne une idée de la localisation de la MJO. La précision à une phase près est de 100%. La matrice de confusion (figure 13 permet de détailler ces deux scores de précision. Chaque case indique le pourcentage d'évènements d'une phase selon l'analyse du modèle ECMWF sachant la phase de l'analyse du BoM (considérée comme la vérité). La somme sur chaque ligne vaut 100% (à l'erreur d'arrondi près). Les fortes valeurs sur la diagonale sont en accord avec la précision (72%). Le modèle est le plus performant dans le diagnostic de la phase 7 (80%), donc sur le pacifique, et le moins performant pour la phase 2 (65%), donc sur l'océan indien (voir figure 3 pour la position de la MJO selon la phase). Ce manque de performance sur la phase 2 peut s'expliquer par l'atténuation de la MJO en fin de cycle. On remarque également les fortes valeurs, de part et d'autres de cette diagonale. Dans l'immense majorité des cas, lorsque le modèle se trompe de phase c'est avec la phase directement précédente ou directement suivante. C'est en accord avec la précision à une phase près de 99.6% (à l'arrondi près).

Tous ces scores vont nous servir de référence pour la suite de notre étude. Ils nous indiquent que le modèle ECMWF présente un biais, même à l'échéance 0. En effet, il reste relativement éloigné de l'information fournie par l'analyse du BoM.

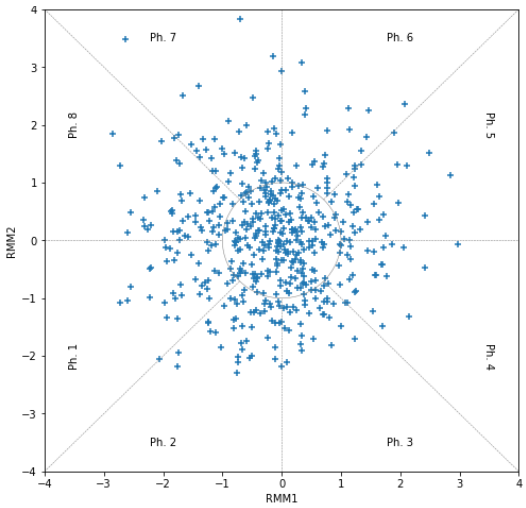


FIGURE 14 – RMM de l'analyse du modèle de ECMWF, sur l'échantillon test

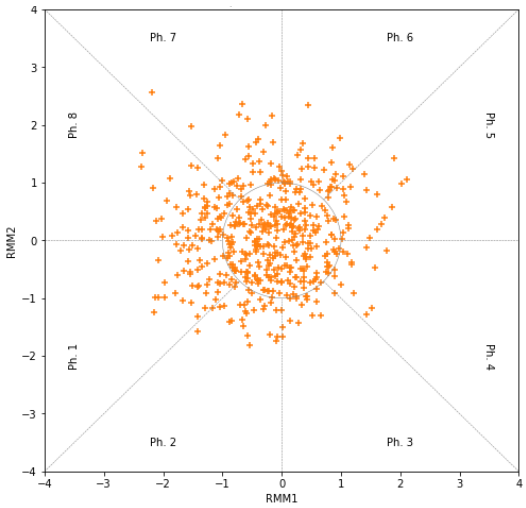


FIGURE 15 – RMM de l'analyse par régression de l'indice du modèle de ECMWF, sur l'échantillon test

MAE	0.34
MSE	0.18
Distance euclidienne	0.53
Distance euclidienne avec seuil	0.38
BRMSE	0.60
BCORR	0.91
Erreur sur l'amplitude	-0.20
Erreur absolue sur l'amplitude	0.36
Erreur sur l'angle	0.59°
Erreur absolue sur l'angle	19°
Non intenses identifiés	85%
Intenses identifiés	72%
Précision	72%
Précision à une phase près	100%

Tableau 2 – Scores de l'analyse par régression comparée à l'analyse du modèle de ECMWF

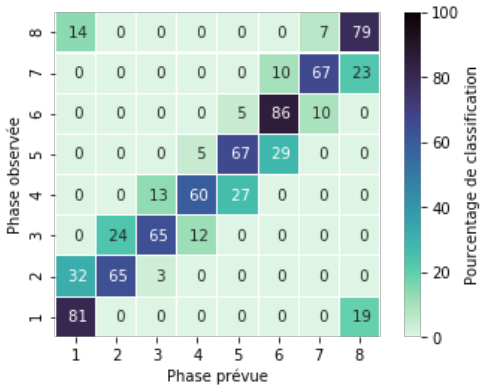


FIGURE 16 – Matrice de confusion de l'analyse par régression de l'indice du modèle de ECMWF

5.1.2 Régression de l'indice RMM du modèle de ECMWF

Dans cette section nous évaluons la capacité des méthodes d'IA à diagnostiquer l'indice RMM de l'analyse du modèle de ECMWF, à partir des champs d'analyse du modèle de ECMWF, par régression. Nous présentons ici les résultats les plus concluants, obtenus par la méthode SVR (voir section 4.2.2). L'objectif de la régression est ici d'approcher au mieux le calcul des RMM par la méthode classique, c'est à dire par filtrage et ACP.

Les figures de dispersion 14 et 15 ne présentent que les évènements sur l'échantillon test (2015-2019). La dispersion par régression (fig. 15) montre une répartition correcte sur l'ensemble des phases, mais une tendance à la sous-estimation de l'amplitude par rapport à la dispersion attendue (fig. 14). Les scores (tableau 2) confirment cette tendance, avec un biais négatif non négligeable sur l'amplitude (-0.20). Cela est encore confirmé par les pourcentages de non intenses identifiés et d'intenses identifiés, déséquilibrés en faveur des non intenses. L'erreur absolue sur l'amplitude est plus élevée que pour la référence (section 5.1.1). L'erreur absolue sur l'angle nous indique une légère avance de phase, et l'erreur absolue sur l'angle est légèrement plus faible que celle de la référence. Les autres scores sont quasiment identiques à ceux de la référence. La matrice de confusion (fig. 16) montre une bonne détermination globale des phases de la MJO, avec toutefois une plus grande disparité que pour la référence. La phase 6 (pacifique ouest) est la mieux déterminée (86%) et la phase 4 (continent maritime) est la moins bien déterminée (60%).

Finalement, ces résultats montrent que les méthodes d'IA, en particulier le SVR, sont capables de déterminer de façon satisfaisante l'amplitude et la phase de la MJO, sans filtrage des données.

5.1.3 Régression de l'indice RMM de l'analyse du BoM

Dans cette section nous évaluons la capacité des méthodes d'IA à diagnostiquer l'indice RMM de l'analyse du BoM, à partir des champs d'analyse du modèle de ECMWF, par régression. L'objectif de la régression est ici double : approcher au mieux le calcul des RMM par la méthode classique (objectif de la section 5.1.2) tout en tentant de réduire le biais du modèle de ECMWF par rapport à l'analyse du BoM (section 5.1.1).

Le tableau 3 rappelle les scores de référence (section 5.1.1) et synthétise les résultats obtenus avec quatre modèles d'IA : régression linéaire, SVR, perceptron multicouche et réseau convolutif en deux dimensions (décrits en sections 4.2 et 4.3).

Le modèle linéaire simple donne des résultats imparfaits mais déjà intéressants. Les scores MAE (0.42), MSE (0.27), distance euclidienne (0.66), distance euclidienne avec seuil (0.51), BRMSE (0.73) et BCORR (0.88) donnent des résultats globalement moyens en comparaison avec les autres méthodes, parfois proches de la référence (en particulier pour la BCORR). L'erreur sur l'amplitude montre un très léger biais négatif. En revanche, l'erreur absolue sur l'amplitude est la moins bonne de toute les méthodes. Les pourcentages de non intenses identifiés et d'intenses identifiés sont proches, ce qui confirme la performance du modèle concernant le biais sur l'amplitude. Ils sont en revanche tous les deux relativement bas, ce qui est en accord avec l'erreur absolue sur l'amplitude élevée. L'erreur sur l'angle et l'erreur absolue sur l'angle sont quant à elles toutes les deux les moins bonnes, en comparaison aux autres méthodes, avec un biais négatif conséquent. On note toutefois que le biais sur l'angle (-1.6°) est moins important que celui de la référence (-2.8°). Même le plus simple de nos modèle a donc réussi à débiaiser le modèle de PNT de ECMWF en terme de retard ou d'avance de phase de la MJO. La précision et la précision à une phase près ne sont pas très performantes mais restent correctes.

	Ref	Lin	SVR	MLP	CNN
MAE	0.33	0.42	0.38	0.39	0.43
MSE	0.18	0.27	0.22	0.24	0.28
Distance euclidienne	0.52	0.66	0.59	0.61	0.67
Distance euclidienne avec seuil	0.39	0.51	0.44	0.46	0.53
BRMSE	0.60	0.73	0.66	0.68	0.74
BCORR	0.91	0.88	0.91	0.88	0.86
Erreur sur l'amplitude	-0.05	-0.08	-0.27	-0.15	-0.01
Erreur absolue sur l'amplitude	0.30	0.41	0.40	0.38	0.21
Erreur sur l'angle	-2.8°	-1.6°	0.60°	0.91°	-0.90°
Erreur absolue sur l'angle	21°	26°	22°	23°	25°
Non intenses identifiés	78%	72%	89%	78%	65%
Intenses identifiés	82%	74%	71%	75%	86%
Précision	72%	67%	71%	66%	66%
Précision à une phase près	99.6%	98.8%	99.4%	99.7%	99.6%

Tableau 3 – Scores de l'analyse par régression comparée à l'analyse du BoM. La meilleure méthode concernant un score est mise en évidence en vert, et la moins bonne en rouge.

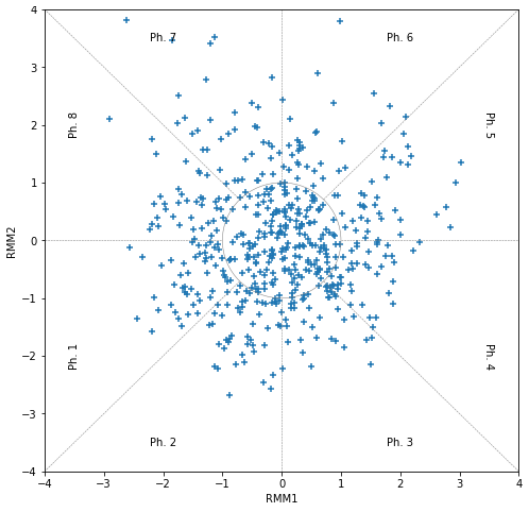


FIGURE 17 – RMM de l'analyse du BoM, sur l'échantillon test

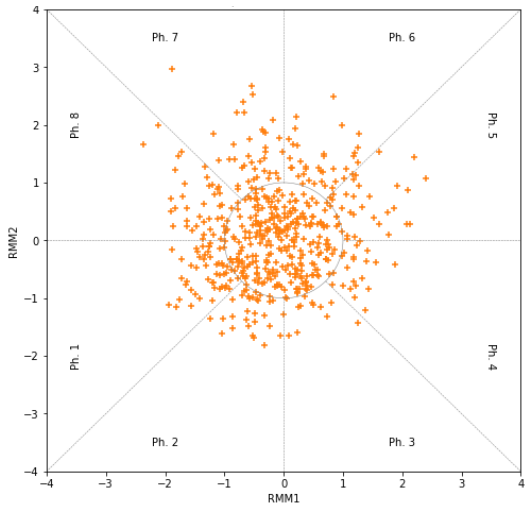


FIGURE 18 – RMM de l'analyse par SVR de l'indice du BoM, sur l'échantillon test

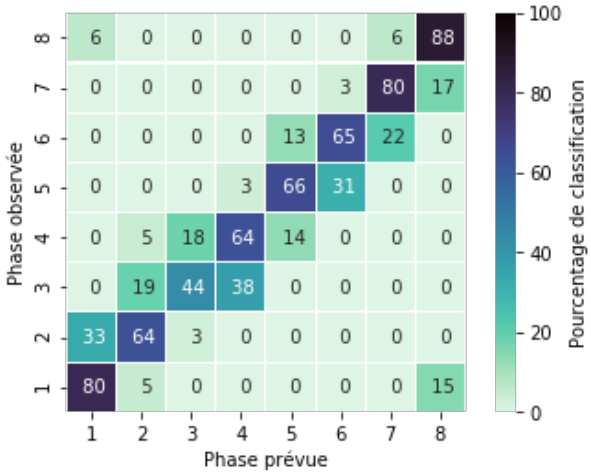


FIGURE 19 – Matrice de confusion de l'analyse par SVR de l'indice du BoM

Le SVR est globalement le modèle le plus performant, avec la majorité des scores qui sont les meilleurs. La BCORR égale même celle de la référence, l'erreur absolue sur l'angle l'égale presque et l'erreur sur l'angle la bat largement. On note en revanche une erreur sur l'amplitude qui est la plus mauvaise, avec une claire tendance à sous-estimer l'intensité de la MJO. Cela se confirme par le net déséquilibre des pourcentages d'intenses et de non intenses identifiés. L'erreur absolue sur l'amplitude n'est pas non plus performante.

Le MLP obtient des résultats moyens, moins bons que ceux du SVR. Notons qu'il a la précision la plus basse (66%).

Le CNN obtient globalement les résultats les moins bons. L'erreur sur l'amplitude ainsi que l'erreur absolue sur l'amplitude sont toutefois faibles. En revanche, les pourcentages d'intenses (86%) et de non intenses (65%) identifiés ne sont pas équilibrés, ce que indique une difficulté particulière du modèle à diagnostiquer l'intensité de la MJO spécifiquement lorsqu'elle est proche de 1, entre intense et non intense. Cette mauvaise performance du modèle CNN nous indique que la convolution ne permet pas de mettre en évidence les caractéristiques de l'atmosphère pertinentes au diagnostic de la MJO, contrairement à ce que l'on espérait.

Finalement, le modèle le plus performant que nous retiendrons est le SVR, malgré son biais négatif important en intensité de MJO. C'est celui qui permet de diagnostiquer au mieux la phase d'une MJO intense.

5.2 En mode prévision

Cette deuxième étape de l'étude s'effectue en mode prévision, de l'échéance J+1 à l'échéance J+32. Dans l'ensemble de cette étape, nous effectuons la régression à partir des trois variables utilisées dans le calcul classique de l'indice RMM (OLR, U850 et U200) ainsi que de TCW (section 3.2.3), qui permet d'améliorer légèrement les résultats. Nous avons utilisé dans cette partie les deux méthodes qui se sont avérées les plus performantes en mode analyse : SVR et MLP. Nous ne présentons ici que la méthode donnant les meilleurs résultats en mode prévision : le SVR. En effet, le MLP ne permet que rarement de battre la référence, qui est l'indice RMM fourni directement par le modèle de ECMWF, quels que soient le score et l'échéance.

Avec ce modèle SVR, deux méthodes d'apprentissage sont mises en oeuvre. La première consiste à entraîner le modèle à une échéance puis à l'évaluer sur cette même échéance, et ceci pour toutes les échéances. Les paramètres du modèle sont réinitialisés puis réentraînés pour chaque échéance. Nous désignons par la suite ce modèle par *SVR*. La deuxième méthode consiste à entraîner le modèle à une échéance donnée (1, 7, 14 ou 21 jours) puis à l'évaluer sur l'ensemble des échéances. Ces méthodes sont désignées par *ech1*, *ech7*, *ech14* ou *ech21*.

Les résultats sont présentés en figures 20 et 21. Les colonnes de gauche présentent les scores. La ligne marron représente les scores de la référence (RMM du modèle ECMWF comparé à RMM du BoM). Les autres lignes présentent chacune des méthodes d'entraînement, en fonction de l'échéance. Les colonnes de droite présentent l'écart relatif entre le score et le score de référence, en fonction de l'échéance :

$$Delta\ score = \frac{score - score_{ref}}{score_{ref}} * 100$$

La visualisation de cet écart relatif permet de faciliter la comparaison à la référence, par le croisement avec l'axe 0. Pour les scores BCORR, précision et précision à une phase

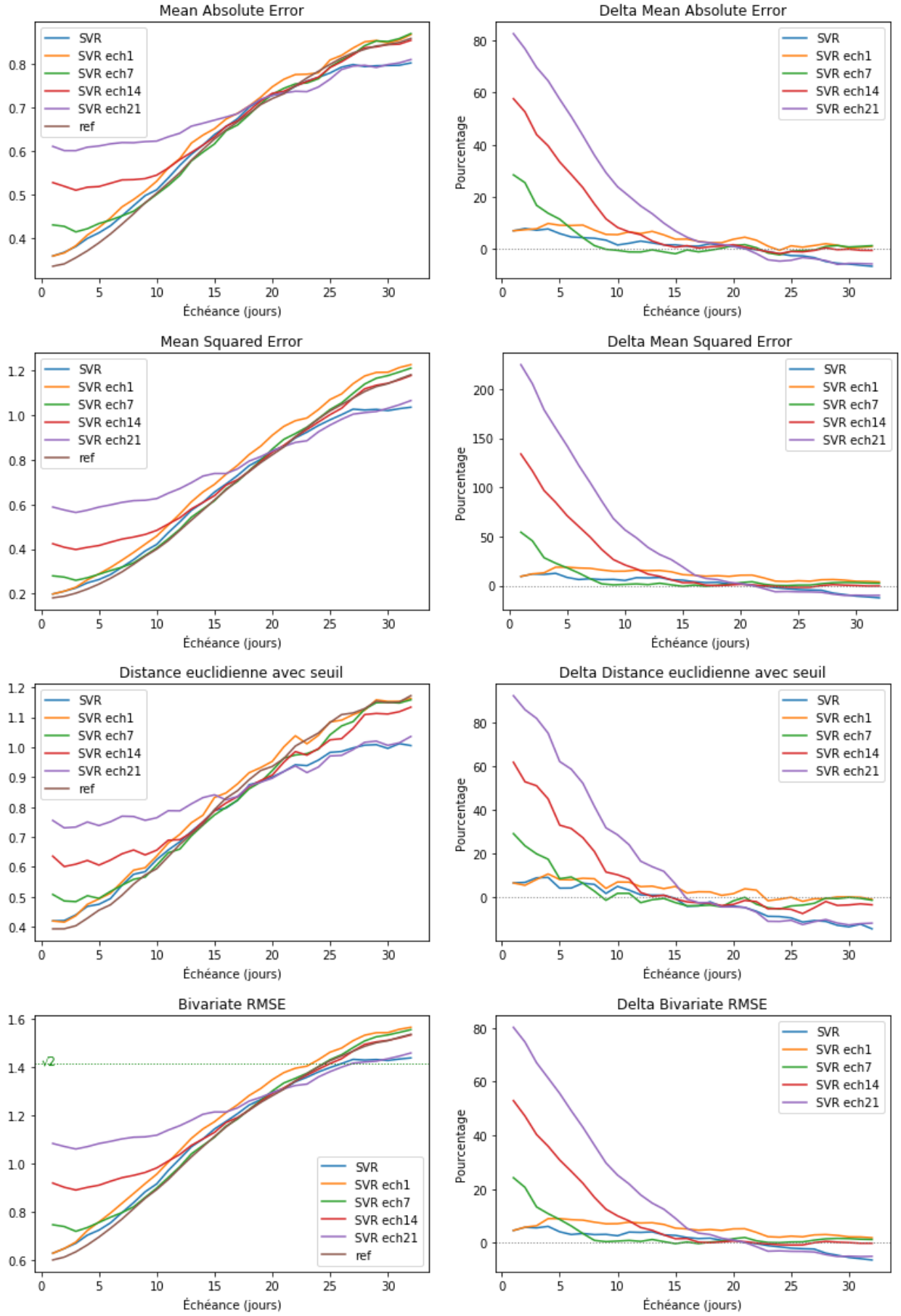


FIGURE 20 – Scores (à gauche) et scores comparés à la référence (à droite), en mode prévision

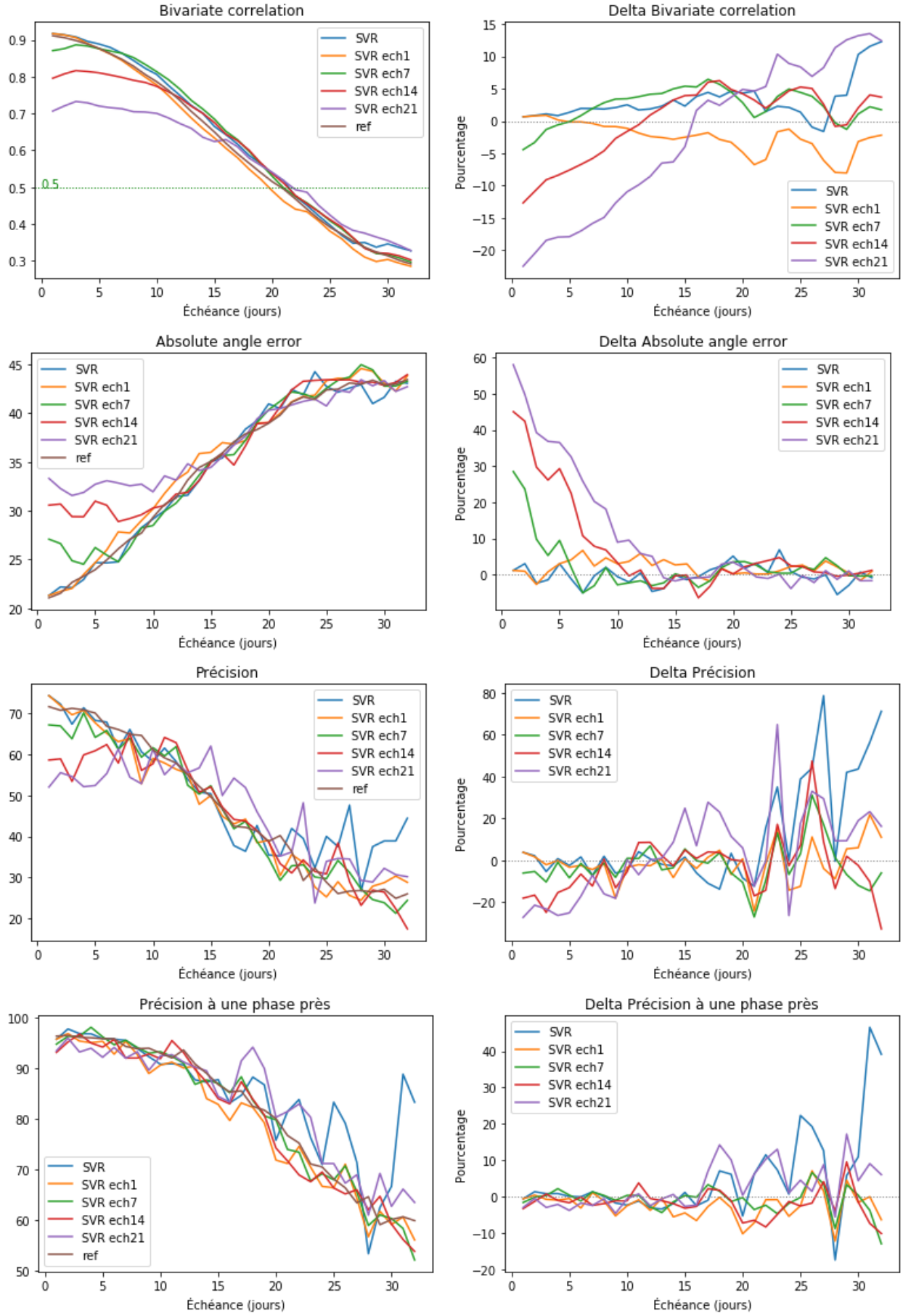


FIGURE 21 – Scores (à gauche) et scores comparés à la référence (à droite), en mode prévision — Suite

près, l'objectif est d'avoir le score le plus grand possible, donc un score plus élevé que la référence, donc un écart relatif positif. Pour tous les autres scores, l'objectif est d'avoir un écart relatif négatif pour battre le modèle de ECMWF.

De manière générale, tous les scores se dégradent naturellement avec l'échéance. Pour toutes les méthodes d'IA, leurs scores sont moins bons que la référence sur les premières échéances. Selon l'échéance d'entraînement choisie, elles battent cependant la référence pour certaines échéances. D'autre part, l'échéance d'entraînement choisie calibre et optimise le modèle pour la prévision autour de cette échéance. Cela apparaît nettement sur les courbes de score aux courtes échéances (5 premiers jours), avec l'entraînement à l'échéance 21 étant de loin le plus mauvais, suivi par l'entraînement à l'échéance 14, 7 puis 1. Cette tendance est inversée sur les échéances les plus lointaines (5 derniers jours).

Concernant la MAE, le modèle *ech7* permet de battre la référence entre les échéances 8 et 20. Au delà, les modèles *SVR* et *ech21* sont plus performants que la référence. Ces méthodes d'IA permettent donc d'améliorer les résultats entre les échéances 8 et 32, en terme de MAE. Pour la MSE, *ech7* égale presque la référence à partir du jour 8, sans toutefois la battre. En revanche, au delà du jour 20, *SVR* et *ech21* sont meilleurs. L'analyse est similaire pour la distance euclidienne (non montrée). Pour la distance euclidienne avec seuil en revanche, les modèles battent la référence dès le jour 12. D'abord *ech7* est le meilleur, jusqu'au jour 19, puis *ech21* et *SVR*. La BRMSE permet d'avoir comme information supplémentaire la prévisibilité. En effet, au delà de $\sqrt{2}$, l'information donnée par le modèle n'est pas jugée fiable. Le seuil de prévisibilité avec la référence est alors de 24 jours. Avec *ech1* il est plus faible, donc moins intéressant. Avec *ech7* et *ech14* il est similaire. Avec *SVR* il est de 25 jours et avec *ech21* il est de 26 jours. L'application de la méthode d'IA *ech21* permet alors de gagner 2 jours de prévisibilité de MJO. L'écart relatif nous indique que le modèle *ech21* bat la référence dès le jour 21. De la même manière, la BCORR informe sur la prévisibilité avec le seuil 0.5 sous lequel il ne faut pas passer. Avec *ech1*, la prévisibilité est de 20 jours, avec la référence ou *ech7*, elle est de 21 jours et avec ech 21 elle est de 22 jours. D'autre part, d'après la figure d'écart relatif, il y a toujours une méthode meilleure que la référence. La meilleure des méthodes est *SVR* jusqu'au jour 7 puis *ech7* jusqu'au jour 19 et enfin *ech21* jusqu'au jour 32. Concernant l'erreur absolue sur l'angle, les méthodes d'IA suivent d'assez près la référence, sans qu'une méthode ne se distingue en particulier. Elles la battent néanmoins fréquemment. Pour la précision et la précision à une phase près, les courbes sont moins stables que pour les scores précédents. Cette grande variabilité s'explique par la petite taille des échantillons, c'est à dire le petit nombre d'évènements de MJO intenses détectés. C'est lié au biais négatif en amplitude mis en évidence en section 5.1.3. Ce biais reste toujours moins bon que celui de la référence, avec toutes les méthodes et à toutes les échéances, tout comme l'erreur absolue sur l'amplitude (figures non montrées). Malgré cette forte variabilité sur les courbes de précision, les méthodes d'IA battent la référence dès le jour 11. Pour la précision à une phase près, elles battent la référence après le jour 16.

La figure 22 présente les prévisions du 18 novembre 2018 avec le modèle *ech1* et avec le modèle *ech14*. Il s'agit d'un exemple de ce que nous pourrions obtenir en opérationnel. Les analyses de J-4 à J0 sont en bleu, pour renseigner l'état passé de l'atmosphère. Les analyses de J+1 à J+32 sont en vert et constituent la vérité à atteindre. Les prévisions de J+1 à J+32 par IA sont en orange. Les prévisions par le modèle de PNT sont en marron, c'est la référence. Sur ces deux figures, seules les courbes de prévision par IA sont différentes. On est ici dans le cas d'une MJO qui reste intense sur toute la période, selon l'analyse. On retrouve une propagation vers l'est, avec un parcours de l'espace des

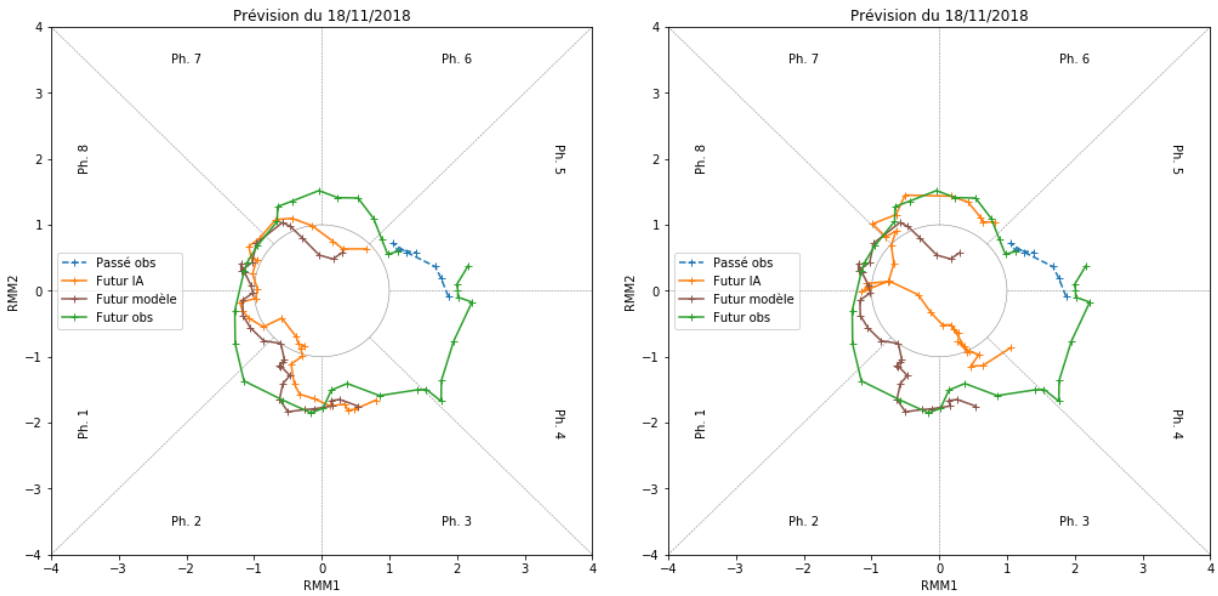


FIGURE 22 – Prévisions du 18 novembre 2018 avec le modèle *ech1* (à gauche) et avec le modèle *ech14* (à droite). Les analyses de J-4 à J0 sont en bleu et de J+1 à J+32 en vert. Les prévisions de J+1 à J+32 par IA sont en orange et par le modèle de PNT en marron.

phases dans le sens trigonométrique. Le modèle de ECMWF présente un fort décalage à la fois en angle et en amplitude à J+1. Le modèle ech1 corrige cette erreur et le modèle ech14 d'autant mieux. De J+1 à J+6, le modèle de ECMWF est en avance sur l'analyse, et indique une intensité trop faible. ech1 se montre légèrement meilleur et ech14 bien meilleur sur ces échéance. Ensuite, les modèles de ECMWF et ech1 suivent relativement bien l'analyse, excepté de J+13 à J+16 de l'analyse. À partir de J+10 du modèle ech14, La MJO prédite devient non intense jusqu'à J+28 et s'éloigne de manière importante de l'analyse. Les trois modèles simulent une propagation plus lente de la MJO que celle de l'analyse (phase 3 ou 4 à J+32 au lieu de phase 5). Ce retard est amélioré par les modèles d'IA par rapport au modèle de PNT. Selon l'échéance choisie pour l'entraînement (ech1 ou ech14), le modèle de SVR est plus ou moins performant, selon le cas auquel on l'applique et l'échéance à laquelle on s'intéresse.

Finalement, le post-traitement du modèle de PNT par des méthodes d'IA permet effectivement d'améliorer ses performances de prévision de la MJO par diagnostic de l'indice RMM mais de manière inégale selon les échéances. Elle permet également d'améliorer le seuil de prévisibilité de la MJO de 1 ou 2 jours.

6 Conclusions et perspectives

Cette étude s'inscrit dans un contexte de prévision aux échéances intra-saisonnières, de 2 semaines à 2 mois, qui a été assez peu explorée jusqu'à récemment. La prévision à ces échéances est particulièrement importante sur le bassin SOOI. En effet, la prévisibilité des cyclones tropicaux est fortement liée à la prévisibilité de l'Oscillation de Madden-Julian et des ondes tropicales (ondes de Kelvin et de Rossby-gravité) qui se propagent en été austral. Nous nous sommes concentrés dans le cadre de ce projet sur le diagnostic de la MJO par régression de l'indice RMM. L'objectif de ce stage a d'abord été d'évaluer la capacité de différentes méthodes d'IA à détecter la phase et l'intensité de la MJO à

l'échéance 0 (i.e. en mode analyse). Ensuite, l'objectif a été d'appliquer ces méthodes à la prévision de MJO pour des échéances allant jusqu'à la semaine 4 voire au delà (J+32).

Nous avons montré que les méthodes de régression par IA, de machine learning ou de deep learning, sont capables d'établir un bon diagnostic de l'indice RMM de la MJO. L'analyse du BoM est bien retrouvée par régression à partir de l'échéance 0 du modèle S2S de ECMWF. Nous avons montré que cette tâche est double : elle consiste à la fois en un diagnostic de l'indice RMM et en un débiaisage du modèle par rapport à l'analyse du BoM. Le modèle SVR s'est avéré être le modèle le plus performant. Il permet une précision de 71% sur le diagnostic de la phase d'une MJO intense. En revanche, il faut noter sa tendance à sous-estimer l'intensité de la MJO, avec une erreur sur l'amplitude de -0.27. En mode prévision, cette méthode d'IA s'est montrée performante en permettant d'améliorer la prévision de MJO. Elle permet à la fois d'obtenir des prévisions plus fiables et de repousser le seuil de prévisibilité de 1 à 2 jours. Les méthodes de régression linéaire et de MLP se sont avérées moins intéressantes que le SVR mais tout de même performantes. Les modèles de CNN, de k-NN et de forêt aléatoire n'ont pas permis d'obtenir de bons résultats.

Les perspectives à cette étude sont multiples. Il serait intéressant de mettre en oeuvre des méthodes d'IA exploitant la nature propagatrice de la MJO, grâce à l'information temporelle des données. Cette piste est en cours d'exploration dans l'équipe Cyclones du LACy, avec un réseau de neurone récurrent : le LSTM (Long Short-Term Memory). D'autre part, il pourrait être intéressant d'améliorer certains de nos modèles d'IA, notamment concernant le biais négatif sur l'intensité de la MJO. Des méthodes de correction statistiques à ajouter pourraient s'avérer utiles. Enfin, il s'agit maintenant d'exploiter le caractère ensembliste du modèle S2S global de ECMWF pour extraire d'autres informations sur l'évolution possible de l'atmosphère.

Remerciements

Merci tout d'abord à Quoc-Phi Duong, Sylvie Malardel et Hélène Vérèmes pour leur encadrement bienveillant. Ils ont su être à l'écoute de mes propositions tout en maintenant un cadre pertinent pour cette étude. Merci également à toute l'équipe Cyclones du LACy, pour leur accueil et leur bonne humeur. Merci à mes collègues stagiaires en compagnie de qui j'ai découvert une partie de l'île de la Réunion. Enfin, un grand merci à ma compagne, ma famille et mes amis pour leur soutien.

Annexes

Annexe A Scores

Notons $\widehat{RMM}1$ l'estimateur du prédictand $RMM1$ et $\widehat{RMM}2$ celui de $RMM2$. La moyenne sur l'échantillon (test, train ou l'intégralité disponible) est notée avec une barre supérieure. Chaque élément de l'échantillon correspond à une date, car nous travaillons avec des données journalières. Chacun de ces scores est appliqué à une échéance donnée. La performance du modèle de prévision numérique du temps dépend en effet de son échéance.

Pour chacun de ces scores, nous calculons également sa variance pour l'échantillon, au lieu de sa moyenne. Dans le cadre de l'apprentissage statistique, une variance trop élevée est signe de sur-apprentissage.

Coefficient de détermination ($\mathbf{R^2}$) Caractérise l'erreur sur chacun des deux prédictands indépendamment. Prend ses valeurs entre 0 et 1. Notons que ce score est d'autant meilleur qu'il est proche de 1.

$$R^2 = \frac{1}{2}\left[\left(1 - \frac{\sum_i (\widehat{RMM}1 - RMM1)^2}{\sum_i (RMM1 - \overline{RMM1})^2}\right) + \left(1 - \frac{\sum_i (\widehat{RMM}2 - RMM2)^2}{\sum_i (RMM2 - \overline{RMM2})^2}\right)\right]$$

avec l'indice i correspondant aux éléments de l'échantillon.

Erreur absolue moyenne ou Mean Absolute Error (MAE) Caractérise l'erreur sur chacun des deux prédictands indépendamment. Valeurs positives. Doit être le plus faible possible.

$$MAE = \frac{1}{2}\left(\overline{|\widehat{RMM}1 - RMM1|} + \overline{|\widehat{RMM}2 - RMM2|}\right)$$

Erreur quadratique moyenne ou Mean Squared Error (MSE) Caractérise l'erreur sur chacun des deux prédictands indépendamment, en pénalisant de manière quadratique et non pas linéaire. Valeurs positives. Doit être le plus faible possible.

$$MSE = \frac{1}{2}\left(\overline{(\widehat{RMM}1 - RMM1)^2} + \overline{(\widehat{RMM}2 - RMM2)^2}\right)$$

Distance euclidienne Distance euclidienne entre le point prévu et le point observé dans l'espace des phases RMM1 / RMM2. Caractérise l'erreur sur les deux prédictands de manière dépendante. Valeurs positives. Doit être le plus faible possible.

$$\overline{\sqrt{(\widehat{RMM}1 - RMM1)^2 + (\widehat{RMM}2 - RMM2)^2}}$$

Distance euclidienne avec seuil Pour chaque élément de l'échantillon :

- Si MJO prévue et MJO observée toutes deux non intenses (c'est à dire d'intensité inférieure à 1), alors le score est de 0.
- Sinon, le score est la distance euclidienne.

Le score global est la moyenne des scores de l'échantillon.

Cette distance euclidienne avec seuil peut-être utilisée comme fonction coût d'un modèle d'apprentissage. C'est une manière d'indiquer au modèle de se concentrer sur les cas intenses. La détection du caractère non intense est suffisant. Valeurs positives. Doit être le plus faible possible.

Bivariate Root Mean Squared Error (BRMSE) Il s'agît de la racine de l'erreur quadratique moyenne, en 2 dimensions. Caractérise l'erreur sur les deux prédictands de manière dépendante. Valeurs positives. Doit être le plus faible possible. Il est admis qu'un score BRMSE inférieur à $\sqrt{2}$ traduit une capacité prédictive raisonnable du modèle [Kim et al., 2018].

$$BRMSE = \sqrt{\overline{(\widehat{RMM}1 - RMM1)^2 + (\widehat{RMM}2 - RMM2)^2}}$$

Bivariate Correlation (BCORR) Corrélation entre série de MJO prévue et série de MJO observée, en 2 dimensions. Caractérise la corrélation sur les deux prédictands de manière dépendante. Prend ses valeurs entre 0 et 1. Notons que ce score est d'autant meilleur qu'il est proche de 1. Il est admis qu'un score BCORR supérieur à 0.5 traduit une capacité prédictive raisonnable du modèle ([Kim et al., 2018]).

$$BCORR = \frac{\sum_i (RMM1_i * \widehat{RMM}1_i + RMM2_i * \widehat{RMM}2_i)}{\sqrt{\sum_i (RMM1_i^2 + RMM2_i^2)}\sqrt{\sum_i (\widehat{RMM}1_i^2 + \widehat{RMM}2_i^2)}}$$

avec l'indice i correspondant aux éléments de l'échantillon.

Pour ce score, le calcul de la variance n'est pas possible, du fait de la construction de son calcul.

Erreur sur l'amplitude Correspond à l'erreur sur l'intensité de la MJO. Chaque erreur peut être négative ou positive. En moyennant, on obtient une information sur le biais. Un score négatif indique une tendance à la sous-estimation de l'intensité de la MJO. Un score positif indique une tendance à sa sur-estimation. Ce score doit être le plus proche possible de 0.

$$\overline{\widehat{AMP} - AMP}$$

Avec $\widehat{AMP}$ l'estimateur de l'amplitude AMP. L'amplitude de la MJO est définie par un passage des coordonnées cartésiennes aux coordonnées polaires de la manière suivante :

$$AMP = \sqrt{RMM1^2 + RMM2^2}$$

Erreur absolue sur l'amplitude Moyenne sur la valeur absolue de l'erreur sur l'amplitude. Caractérise alors non plus le biais positif ou négatif, mais la capacité du modèle à être juste en intensité de MJO. Valeurs positives. Doit être le plus faible possible.

$$\overline{|\widehat{AMP} - AMP|}$$

Erreur sur l'angle Correspond à l'erreur sur l'angle de la MJO dans l'espace des phases. Chaque erreur peut être négative ou positive. En moyennant, on obtient une information sur le biais. Un score négatif indique une tendance à un retard de la MJO. Un score positif indique une tendance à une avance. Ce score doit être le plus proche possible de 0.

$$\overline{tan^{-1}\left(\frac{RMM1 * \widehat{RMM}2 - RMM2 * \widehat{RMM}1}{RMM1 * \widehat{RMM}1 + RMM2 * \widehat{RMM}2}\right) * \frac{360}{2\pi}} \quad \text{(en degrés)}$$

Notons que l'angle de la MJO est défini par un passage des coordonnées cartésiennes aux coordonnées polaires de la manière suivante :

$$atan2(RMM2, RMM1) \quad \text{(en radians)}$$

Erreur absolue sur l'angle On moyenne sur la valeur absolue de l'erreur sur l'angle. Caractérise alors non plus le biais positif ou négatif, mais la capacité du modèle à être juste en angle de MJO. Valeurs positives. Doit être le plus faible possible.

$$\overline{\left|tan^{-1}\left(\frac{RMM1 * \widehat{RMM}2 - RMM2 * \widehat{RMM}1}{RMM1 * \widehat{RMM}1 + RMM2 * \widehat{RMM}2}\right)\right| * \frac{360}{2\pi}} \quad \text{(en degrés)}$$

Références

[Brunet et al., 2010] Brunet, G., Shapiro, M., Hoskins, B., Moncrieff, M., Dole, R., Kiladis, G., Kirtman, B., Lorenc, A., Mills, B., Morss, R., Polavarapu, S., Rogers, D., Schaake, J., and Shukla, J. (2010). Collaboration of the weather and climate communities to advance subseasonal-to-seasonal prediction. *Bulletin of the American Meteorological Society*, 91(10) :1397–1406.

[Chantry et al., 2021] Chantry, M., Christensen, H., Dueben, P., and Palmer, T. (2021). Opportunities and challenges for machine learning in weather and climate modelling : hard, medium and soft ai. *Philosophical Transactions of the Royal Society A : Mathematical, Physical and Engineering Sciences*, 379(2194) :20200083.

[Dasgupta et al., 2020] Dasgupta, P., Metya, A., Naidu, C. V., Singh, M., and Roxy, M. K. (2020). Exploring the long-term changes in the madden julian oscillation using machine learning. *Scientific Reports*, 10(1) :18567.

[Drosdowsky and Chambers, 2001] Drosdowsky, W. and Chambers, L. (2001). Near global sea surface temperature anomalies as predictors of australian seasonal rainfall. *Journal of Climate*, 14 :1677–1687.

[Gottschalck et al., 2010] Gottschalck, J., Wheeler, M., Weickmann, K., Vitart, F., Savage, N., Lin, H., Hendon, H., Waliser, D., Sperber, K., Nakagawa, M., Prestrelo, C., Flatau, M., and Higgins, W. (01 Sep. 2010). A framework for assessing operational madden–julian oscillation forecasts : A clivar mjo working group project. *Bulletin of the American Meteorological Society*, 91(9) :1247 – 1258.

[Kim et al., 2018] Kim, H., Vitart, F., and Waliser, D. E. (2018). Prediction of the madden–julian oscillation : A review. *Journal of Climate*, 31(23) :9425 – 9443.

[Lin et al., 2008] Lin, H., Brunet, G., and Derome, J. (01 Nov. 2008). Forecast skill of the madden–julian oscillation in two canadian atmospheric models. *Monthly Weather Review*, 136(11) :4130 – 4149.

[Madden and Julian, 1971] Madden, R. A. and Julian, P. R. (01 Jul. 1971). Detection of a 40–50 day oscillation in the zonal wind in the tropical pacific. *Journal of Atmospheric Sciences*, 28(5) :702 – 708.

[Toms et al., 2020] Toms, B. A., Kashinath, K., Prabhat, and Yang, D. (2020). Testing the reliability of interpretable neural networks in geoscience using the madden-julian oscillation. *Geoscientific Model Development Discussions*, 2020 :1–22.

[Wheeler and Hendon, 2004] Wheeler, M. C. and Hendon, H. H. (01 Aug. 2004). An all-season real-time multivariate mjo index : Development of an index for monitoring and prediction. *Monthly Weather Review*, 132(8) :1917 – 1932.