Prédiction de la concentration du chlorophylle à la surface de l'océan à partir d'une série chronologique d'images satellitaires

Thèse de Recherche

par

Karim ASSAAD et Bilal DIAB

Tuteur de Stage: Carlos MEJIA

Laboratoire LOCEAN

Enseignant Référent: SYLVIE THIRIA

Université Paris Saclay

Master TRIED

Mars 2017

Remerciement

Nous tenons à remercier toutes les personnes qui ont contribué au succès de notre stage et qui nous ont aidé lors de la rédaction de ce rapport.

Nos remerciements les plus chaleureux s'adressent à notre encadrant , Mr Carlos MEJIA, Ingénieur de Recherche au sein de LOCEAN, pour son accueil, le temps passé ensemble et le partage de son expertise au quotidien. Grâce à sa confiance nous avons pu nous impliquer totalement dans l'étude. Il fut d'une aide précieuse dans les moments les plus délicats.

Nous remercions vivement Mme Silvie THIRIA pour l'aide précieuse et le support permanent qu'elle a bien voulu nous apporter ainsi que tous les conseils qu'elle nous a donné.

Nous remercions également toute l'équipe de LOCEAN pour leur accueil et en particulier Mr.Julien BRAJARD et Mr.Anastase CHARANTONIS pour leurs explications.

Enfin nos derniers remerciements, qui n'en sont pas pour autant les moindres, vont à tous nos professeurs pour tout ce qu'ils nous ont appris et ce que nous avons vraiment utilisé et appliqué pour réussir ce stage.

Listes des figures

Contents

What would an ocean be without a monster lurking in the dark? It would be like sleep without dreams.

Werner Herzog

1

Intorduction Générale

L'analyse des données s'applique sur de vastes sujets divers non pas justes mathématiques mais a dépassé pour toucher presque tous les domaines où des données existent. Dans cette étude, nous avons eu la chance de travailler sur un sujet où l'analyse des données a pris un état très évolué et où les méthodes utilisés sont parmi les plus performantes dans le traitement des données de ce type.

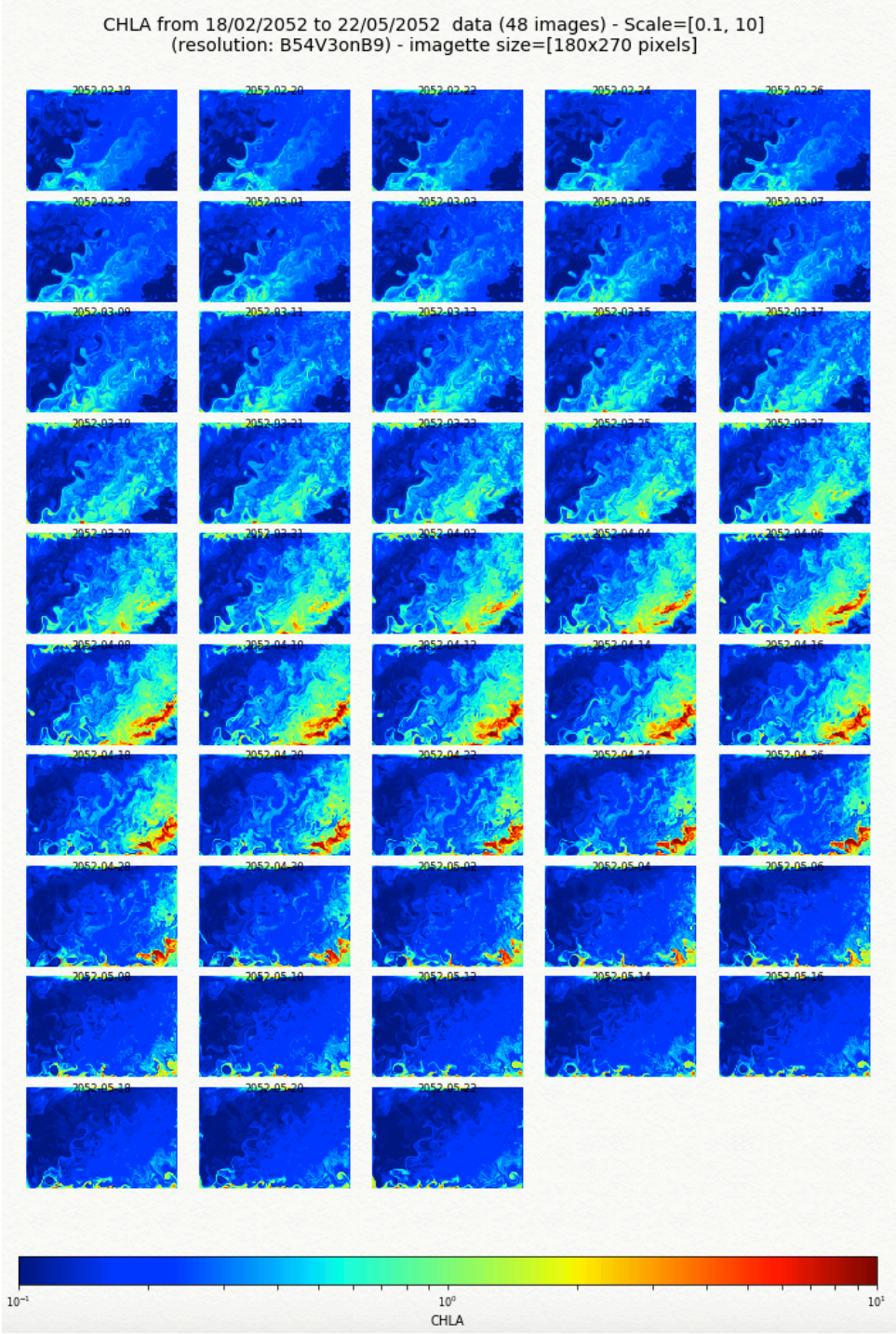
Dans notre cas, l'application de l'analyse des données est dans le domaine de la physique des océans: "Prédictions de la concentration de la chlorophylle dans l'eau et de la température à la surface de l'eau en appliquant des méthodes de Deep Learning sur des séries chronologiques d'images satellitaires". Pour faire, nous disposons d'images qui représentent les concentrations de la chlorophylle dans l'eau codés par des couleurs. Ces images ne sont des vraies, mais elles sont générés selon

un modèle dont nous n'avons pas entré dans les détails et c'est pour cette raison que les années sur lesquelles nous avons travaillé, comme on l'expliquera ci-après, sont des années dans le futur (dans les années 50 du 21ème siècle).

Les images de chlorophylle (CHL) utilisées sont des produits grillés ou mappés de niveau L3 journaliers de résolution de 9km fournis par le capteur Modis embarqué à bord des satellites. Ce choix a été fait grâce à la thèse de Manel JOUINI qui a obtenu des résultats des analyses spectaculaires effectuées sur les données modèles restituées par des méthodes statistiques et qui ont montré des limites éventuelles des méthodes pour les échelles inférieures à 10km.

Pour prédire la concentration de chlorophylle au temps t à partir de 3 images qui représentent la chlorophylle et qui correspondent au temps t-1 t-2 t-3. Ces images ne sont pas utilisés tel qu'elles sont mais elles sont découpées et redimensionnées afin de correspondre à nos modèles. Pour cela, on parcourt chaque image par une matrice de dimension 7x7 pour obtenir ce qu'on appelle des imagettes. Chaque imagette représente une zone de l'image et ce sont ces zones qui constitueront nos observations. Donc pour prédire une valeur de chlorophylle dans une zone, on se base sur 3 imagettes provenant des trois images précédentes et qui représentent la même zone. A noter que les valeurs prédite sont des centres d'imagettes et non pas des imagettes.

Dans cette étude, nous allons dans un premier temps faire une bref analyse descriptive qui nous permettra de mieux choisir les données dans la modélisation. Puis dans un deuxième temps nous allons choisir la meilleure architecture qui correspond le plus à nos données. Ensuite nous allons introduire une nouvelle variable explicative qui est la température de surface de l'océan.



.

Figure 1.0.1: Représentation des images satellitaires utiliser dans l'apprentissage

You can have data without information, but you cannot have information without data.

Daniel Keys Moran

2 Analyse des données

2.1 Introduction

La qualité des données est un facteur très important qui affecte la qualité des résultats obtenus par la modélisation. Pour cela, il est indispensable avant de commencer à les modéliser, d'analyser les données en regardant les relations entre elles et vérifier si elles sont cohérentes.

2.2 Division des donnees

La première étape du travail était la division des données en trois ensembles: apprentissage, validation et test. Pour l'ensemble d'apprentissage nous avons choisi de prendre une période de l'année 2052, une période où on voit bien une évolution au niveau de la concentration de la chlorophylle. Pour cela nous avons choisi

une période d'à peu près trois mois comprise entre 18/02/2052 et 22/05/2052 où nous disposons d'une image tous les deux jours. Concernant l'ensemble de validation, nous avons choisi de prendre une durée identique à celle de l'ensemble d'apprentissage mais pour l'année 2053 et ceci afin d'augmenter la probabilité d'avoir la même évolution de la concentration de la chlorophylle et pour éviter un biais dans les calcul d'erreurs (dans le cas où les deux ensemble sont pris de la même année). Pour cet ensemble, nous avons pris le quart de images de cette durée. Et dernièrement, pour l'ensemble test, nous avons choisi une longue période de l'année 2054 et ceci afin de tester le modèle construit, sur différents phases et surtout sur des phases où il y a des concentrations faibles et constantes de chlorophylle. Cette période est entre le 02/01/2054 et le 29/06/2054 inclus en ayant une image tous les deux jours.

Comme nous l'avons déjà mentionné dans le chapitre précédent, nous n'utilisons pas les images complètes mais plutôt des imagettes de dimensions 7x7 obtenues en balayant les images initiales. Et donc, pour chaque observation (qui correspond à une zone de l'image), l'entrée est formée de trois imagettes alors que la sortie n'est qu'une seule valeur. Dans un premier temps, nous avons tracé des nuages de points pour chacun des trois ensembles: l'ensemble d'apprentissage, l'ensemble de validation et l'ensemble test (Fig 2.2.1). En abscisse, la moyenne des 3 imagettes et en ordonnée la valeur référence qui est donnée par zone (donc par individu). Le moyennage de l'entrée a pour but de la transformer en une forme comparable à la sortie. Ces nuages de points ont pour but de pouvoir comparer les ensembles entre eux surtout celui d'apprentissage avec celui de validation et ceci afin de voir si des données (correspondant à un cas précis) sont présents dans un ensemble mais pas dans un autre. Dans le chapitre précédent nous avons parlé du choix de période pour l'ensemble de validation. En effet, le nuage de points de cet ensemble ne contenait pas tous les cas présents dans l'ensemble d'apprentissage c'est à dire toutes les évolutions de la concentration de chlorophylle possibles et ceci allait affecter les résultats de validation du modèle. Pour cela, nous avons modifié les données d'apprentissage pour englober plus d'informations.

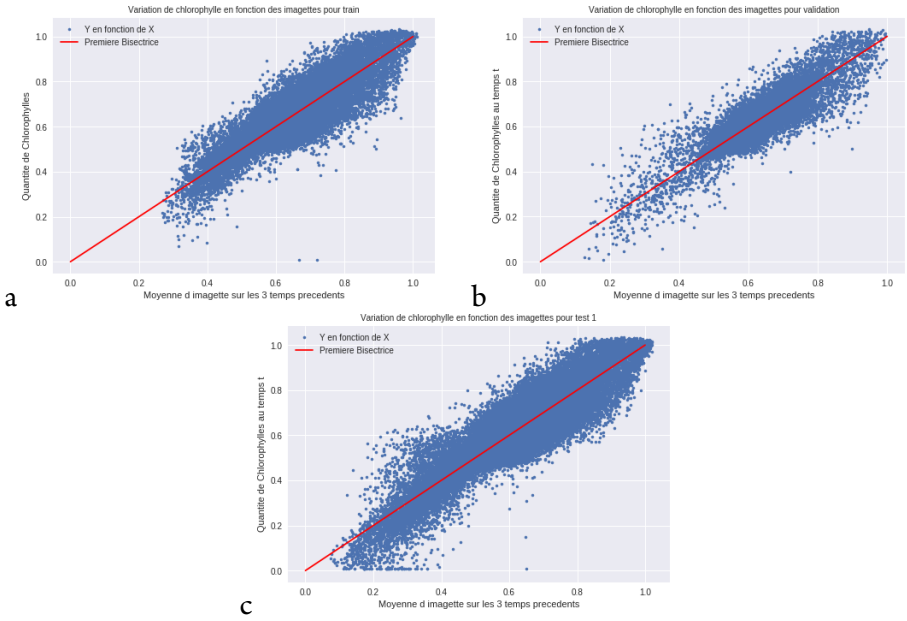


Figure 2.2.1: Représentation des nuage des données de l'ensemble train (a), validation (b) et test (c).

Ce qui est expliqué dans le paragraphe précédant se traduit dans ces trois nuages de données. Le nombre de points n'est pas le même dans les trois nuages, mais ce qui est semblable est que les différents valeurs sont présents dans les trois ensembles surtout si on compare l'ensemble d'apprentissage à celui de validation. De plus, les trois nuages sont bien centrés sur la première bissectrice.

2.3 Comparaison des Distributions

Ensuite, nous nous sommes intéressés à comparer la distribution des données de références et ceux obtenues en moyennant les imagettes. Pour cela, nous avons tracé leurs histogrammes pour l'ensemble d'apprentissage puisque les trois ensemble ont presque la même allure.

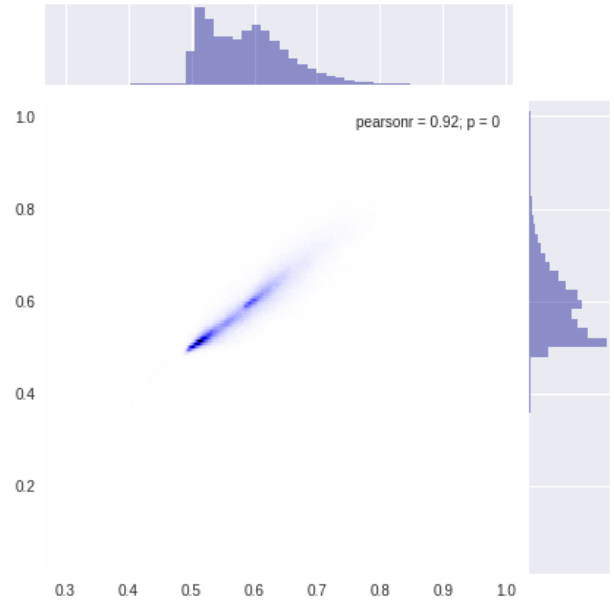


Figure 2.3.1: Représentation des histogramme de la variable a expliquer et de la moyenne de la variable explicative en plus de l'allure de leur dépendance.

On remarque d'après ces histogrammes que les valeurs de références dans l'ensemble d'apprentissage et les valeurs moyennes des imagettes dans le même ensemble suivent presque la même loi. Cela se conclut aussi par la grande corrélation entre ces deux valeurs, la corrélation est de 0,92. Ces résultats nous permettent de conclure que la fonction moyenne peut approximer la sortie attendue mais les résultats obtenus par les modèles construits (détaillés dans la suite de cette étude) nous ont donnés des résultats meilleurs que ceux obtenus juste avec la moyenne.

2.4 Conclusion

Nous avons donc bien choisi les données et nous avons testé leurs compatibilité entre elles. Ces données sont donc prêtes pour les utiliser dans la modélisation de notre problème.

All models are wrong, some are useful.

George BOX

3 Méthodes de validations

3.1 Introduction

Après avoir divisé les données en trois ensembles et testé leurs compatibilités, les données sont prêtes pour la modélisation. Mais vu que nous allons construire plusieurs architectures de modèles, il faut fixer des métriques de calculs d'erreurs afin de pouvoir juger les différents modèles.

Dans cette partie nous allons expliquer les différentes méthodes de calcul d'erreurs des résultats des modèles. Il est important de préciser que nous comparons toujours les valeurs issues de la prédiction se basant sur le modèle avec les valeurs références et donc le cas parfait sera quand le nuage de points représentant ces deux variables soit confondu avec la première bissectrice. Il est important aussi de mentionner que ces méthodes sont appliquées sur les résultats et les valeurs de référence de l'ensemble de validation.

3.2 Mesures choisies

3.2.1 RMS

La première métrique que nous calculons est la RMS (Root Mean Square) qui représente la racine carré de la moyenne des écarts entre la valeur prédite et la valeur référence. A priori cette méthode semble être la plus compatible avec le nuage de données de sortie. L'inconvénient de cette méthode est l'absence d'une valeur référence ou une valeur maximale à laquelle on peut comparer l'erreur obtenu. Mais cette méthode sert toutefois à comparer les modèles entre eux, ce qui est le but.

$$RMS = \sqrt{\frac{\sum (Di - Yi)^2}{n}}$$

D représente les valeurs référents et Y représente les valeurs prédites.

3.2.2 RMS relative

La RMS relative est calculée en divisant chaque écart par la valeur de référence. Cette méthode permet d'étaler, d'exprimer la valeur prédite sur l'échelle de la valeur référence afin de savoir l'ordre de grandeur de l'erreur.

$$RMSRelative = \sqrt{\frac{\sum \left(\frac{Di-Yi}{Di}\right)^2}{n}}$$

3.2.3 Erreur de Kullback

La méthode de Kullback se base sur le logarithme du rapport entre la valeur référence et la valeur prédite. La spécificité de cette méthode est qu'elle est très sensible au cas où la valeur prédite est plus petite que la valeur référence (rapport inférieur à 1) puisque le logarithme décroît fortement pour les valeurs inférieurs à 1, ce qui n'est pas le cas quand la valeur référence est la plus grande. De plus, les valeurs de cette méthode peuvent être négatifs. La valeur optimale de cette méthode est zéro,

lorsque la valeur prédite vaut la valeur référence.

$$DistanceDeKullback = \sum (Di * \log|\frac{Di}{Yi}|)$$

3.2.4 Erreurs de Corrélation

Il s'agit d'une simple méthode de calcul de corrélation entre deux vecteurs: le premier est le vecteur des valeurs références de l'ensemble de validation et le second est le vecteur de valeurs prédites du même ensemble par le modèle.

$$Corr(D, Y) = \frac{Cov(D, Y)}{\sqrt{Var(D)} * \sqrt{Var(Y)}}$$

3.2.5 Erreurs de l'angle de régression par rapport à la bissectrice

Dans cette méthode, nous commençons par faire une régression linéaire des données (Valeurs prédites en fonction de valeurs de références) ensuite nous nous intéressons à l'angle entre cette droite de régression et la première bissectrice: plus l'angle est grand, plus l'écart entre les deux droites est grand et donc plus le nuage de points est écarté de l'idéal. L'angle maximal atteint est de "$\pi/4$", pour cela nous divisons l'angle obtenu par "$\pi/4$" pour avoir un nombre compris entre 0 et 1. Le défaut de cette méthode est inverse à celui du RMS: si les points sont trop écartés de la première bissectrice mais de façon équilibré de part et d'autre, la droite de régression sera presque confondu avec la première bissectrice. On aura donc une erreur faible alors que les points sont très écartés de la bissectrice. Nous pouvons alors dire que cette méthode sera efficace avec un indice supplémentaire qui est la RMS.

$$a = \frac{\sum (Di - \bar{D}) * (Yi - \bar{Y})}{\sum (Di^2) - n * \bar{D}^2}$$

$$Anglelabissectrice = \frac{arctan(|(a-1)/(1+a)|)}{\frac{\pi}{4}}$$

3.2.6 Histogrammes

L'histogramme est en fait un générateur de sens. En pratique c'est une interface qui sise à l'intersection d'une problématique et d'un ensemble de données. Son rôle fondamental est bien de montrer la partie des données où le modèle se trompe au sens qu'il associe à ses données une autre information. Comme un dessin vaut mieux qu'un long discours, les représentations graphiques seront généralement préférées aux listes de données insipides dont la substantifique moelle tant espérée est bien difficile à extraire. En résumé, cette méthode consiste à tracer les histogrammes des valeurs référent et des valeurs prédites afin qu'on puisse les comparer et voir pour quelles valeurs nous avons eu les fausses prédictions.

3.3 Conclusion

Cette étape d'explication des méthodes d'évaluations des modèles est très importante, car une fois les modèles sont construits, il nous faut un critère qui nous permet de savoir quel modèle est le plus performant. Nous avons choisi d'en utiliser plusieurs afin d'affiner le plus possible la validation d'un résultat obtenu surtout que nos données d'entrée sont corrélées avec ceux de la sortie et de plus ce sont des données générées donc presque parfaites. Pour cela il faut augmenter la barre de validation des modèles.

In life ... Deep Learning comes from Deep Pain. But once the learning is complete, Deep Pain transforms into Deep Gain

Bishu Prusty

4

Construction des modèles

4.1 Introduction

Dans le chapitre précédent, nous avons fixé les critères d'évaluations des modèles. Dans ce chapitre, nous allons expliquer les différents types d'architecture, exposer leurs performances et choisir le meilleur avant de tester sa robustesse et son efficacité. Nous n'allons pas nous contenter d'un seul type de modèle mais nous allons en essayer plusieurs et même chaque type avec différents paramètres afin d'en choisir le meilleur modèle correspondant à notre problématique.

4.2 Dense

Le premier pas dans la construction des architectures de réseaux de neurones est d'essayer les architectures les plus simples qui sont formés uniquement de couches

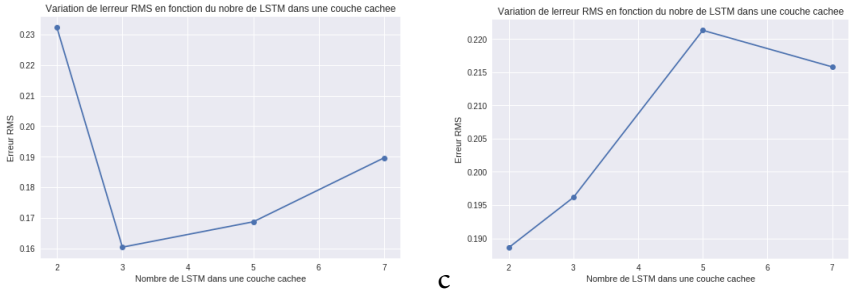
denses (couches entièrement connectées). Nous avons essayé ces architectures avec différents nombres de couches cachées et différents nombre de neurones dans chaque couche. Nous avons remarqué qu'il y a une diminution remarquable de l'erreur quand on passe d'une couche cachée à deux. Ceci montre que le modèle s'améliore en augmentant sa complexité. Pour cela, nous avons passé à des architectures plus profondes donc plus complexes. Le meilleur résultat de ce type de modèle (Dense) est obtenu pour une couche de 28 neurones reliée à une deuxième couche de 11 neurones. Le résultat est affiché dans le tableau récapitulatif dans la suite de ce chapitre.

4.3 Réseaux de neurones récurrents (LSTM)

Les réseaux de neurones récurrents existent sous différents formes. Ce type de réseaux de neurones est généralement utilisé quand il existe un contexte dans les données c'est à dire de l'information sous forme séquentielle comme les textes. Dans cette étude nous avons essayé deux types de réseaux de neurones récurrents: les réseaux récurrents simples (Simple RNN) et les Long Short-Term Memory (LSTM). L'avantage qu'a le LSTM sur le RNN est qu'il est capable de travailler sur des longues séquences et avec décalages temporels, ce qui est notre cas vu que nous avons des longues séquence d'imagettes obtenues en balayant les images par des matrice 7x7 et les images d'origine ne sont pas continués mais prises un jour sur deux. Cette remarque est confirmée par les résultats obtenus vu que les résultats données par la méthode des RNN, (meilleur résultat pour RNN est obtenu pour 3 RNN avec 3 récurrences) sont relativement moins bonnes comparant à ceux du LSTM (les résultats sont exposés dans la suite de ce chapitre). Pour cela, nous avons choisi de continuer en utilisant juste le LSTM comme type de réseaux récurrents.

4.4 Choix de l'architecture du LSTM

Deux paramètres importants étaient à choisir pour les réseaux LSTM, ces paramètres sont le nombre de récurrences de LSTM et le nombre d'unités LSTM. Il est important de signaler qu'une couche dense est ajoutée avant les LSTM et une après, juste avant la sortie. Les deux paramètres ont été choisis en essayant plusieurs possibilités et gardant celles qui donnent la plus petite erreur RMS. Les graphes ci-après l'erreur RMS en fonction du nombre d'unités de LSTM et du nombre de récurrences (respectivement à gauche et à droite).



c

Figure 4.4.1: Courbes montrant les choix optimaux des paramètres du LSTM.

A l'issu de ces résultats, nous avons maintenant une idée générale de l'architecture du réseau à construire. Mais pour un bon fonctionnement du LSTM et une meilleure performance, nous avons cherché et trouvé que l'entrée du LSTM doit être distribuée, suivie et précédée d'une couche dense. La dernière couche dense ayant pour but de réunir les trois sorties du LSTM en une sortie qui est la sortie désirée. Donc nous avons obtenu l'architecture suivante:

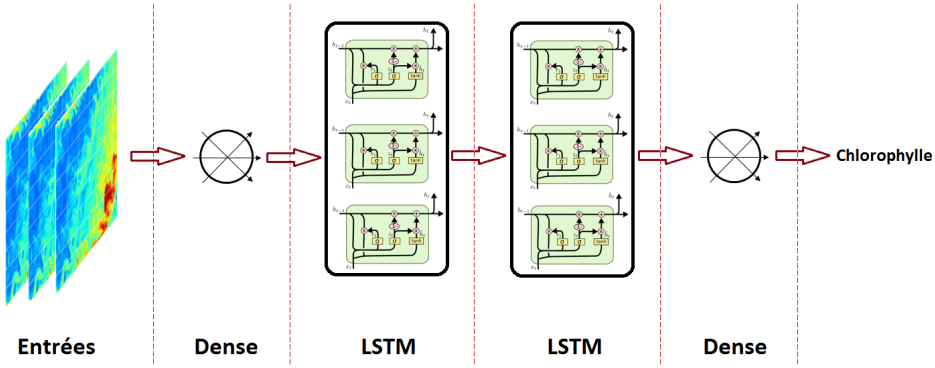


Figure 4.4.2: Architecture contenant les couches LSTM.

Les résultats de ces réseaux seront mentionnés dans le tableau résumé des résultats ci-après.

4.5 Réseaux de neurones à convolution (CNN)

Le premier défi était de choisir le nombre de filtres puis leurs grandeurs. Tout au long des expériences effectuées on a constaté que le maxpooling aide à la généralisation du modèle donc on a décidé de le présenter dans l'architecture. Le maxpooling ne fait que renvoyer le maximum d'une région (Il n'y a rien d'entraînable/ajustable sur le maxpooling) et donc il ne rectifie pas le nombre de poids. En revanche, la CNN utilisée nécessite de lourds poids. Ce qu'on entraîne dans une couche de convolution, ce sont les poids du masque de convolution de chaque filtre où les filtres sont attribués à leurs propres masques de convolution à entraîner.

On a fait plusieurs expériences (plusieurs architectures), la première était de mettre une CNN avec une dense, puis nous avons ajouté un maxpooling et nous avons varié le nombre de filtres (8, 16 et 32) et ceci pour différent nombre de dimensions (2, 5 et 7). Pour chaque combinaison filtre-dimension, nous avons reconstruit une nouvelle architecture mais en distribuant la CNN sur les entrées. A la fin nous avons ajouté une autre CNN de même nombre de filtres que la CNN précédente et qui a la même dimension que la sortie de maxpooling. Par la suite, en nous basant sur les mesures de validation, nous avons décidé de faire des CNN

distribuées sur les trois imagettes, en utilisant 16 filtres de dimensions 3 x 3 chacun. La distribution des CNN a beaucoup amélioré le résultat par rapport au cas où on fait un CNN sur les 3 imagettes simultanément. Dans le cas du maxpooling, nous avons choisi de réduire les fenêtres de dimension 2 x 2, dans le but d'extraire d'avantage d'informations du maxpooling. Après nous avons choisi d'effectuer une autre CNN de 16 filtres mais cette fois avec une dimension égale à 2x2. Comme dernière étape, nous avons relié la sortie du CNN à une dense pour obtenir la concentration de la chlorophylle. La figure ci-dessus représente l'architecture de ce cas.

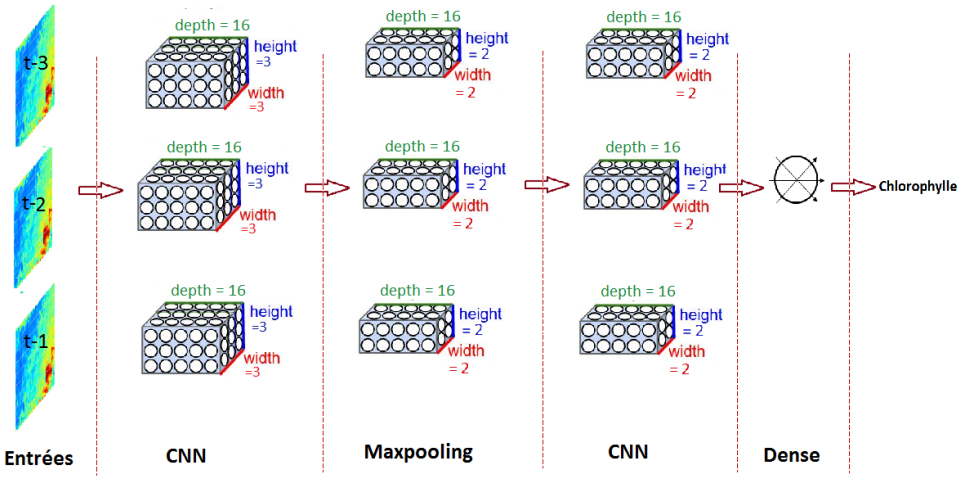


Figure 4.5.1: Représentation de l'architecture CNN.

4.5.1 Mixte: CNN + LSTM

Afin d'améliorer d'avantage les résultats on a décidé de combiner la meilleure architecture obtenue avec les LSTM avec la meilleure architecture obtenue avec les CNN. Donc nous avons utilisé la même architecture de CNN présentée dans la section précédente mais en ajoutant les 2 récurrences de LSTM entre la deuxième couche CNN et la couche Dense. De cette manière, nous avons augmenté la complexité du modèle mais nous allons voir dans le tableau des résultats dans la section

des résultats expérimentaux que cette architecture va avoir une très bonne performance.

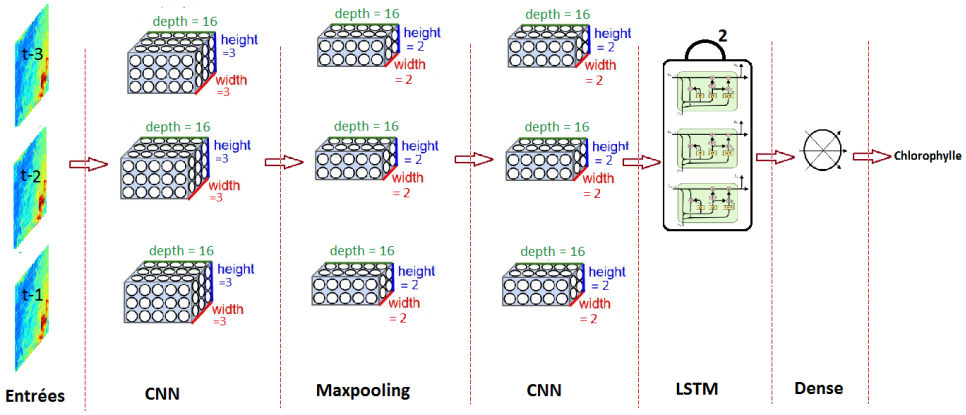


Figure 4.5.2: Représentation de l'architecture Mixte CNN+LSTM.

4.6 Résultats Expérimentaux

Après la construction des modèles et le calcul de leurs taux d'erreurs suivant les différentes méthodes déjà expliquées précédemment, nous avons résumé les meilleurs résultats obtenus pour chaque type de modèle et ceci dans le tableau suivant.

Architecture	RMS	RMS R	Correlation	Kullback	Angle a la Bisectrice
Dense	0.25131	0.41893	0.88	4751.753	0.386
RNN	0.24049	0.32814	0.90	8556.689	0.374
LSTM	0.16299	0.22742	0.95	-303.985	0.056
CNN	0.18519	0.20667	0.95	1452.095	0.138
Mixte	0.15253	0.20207	0.96	-914.678	0.038

Table 4.6.1: Your caption here

L'architecture la plus simple qui est formée juste de couches denses est la moins performante. Les erreurs diminuent quand on passe aux architectures plus complexes. Le meilleur résultat est obtenu pour une architecture mixte entre CNN et

LSTM. Les architectures LSTM et mixte ont des résultats très proche mais d'après les résultats de prédiction à l'horizon (dans la suite de ce chapitre), on a choisi de prendre l'architecture mixte car au bout de plusieurs pas de temps, ses deux architectures ne sont plus comparables.

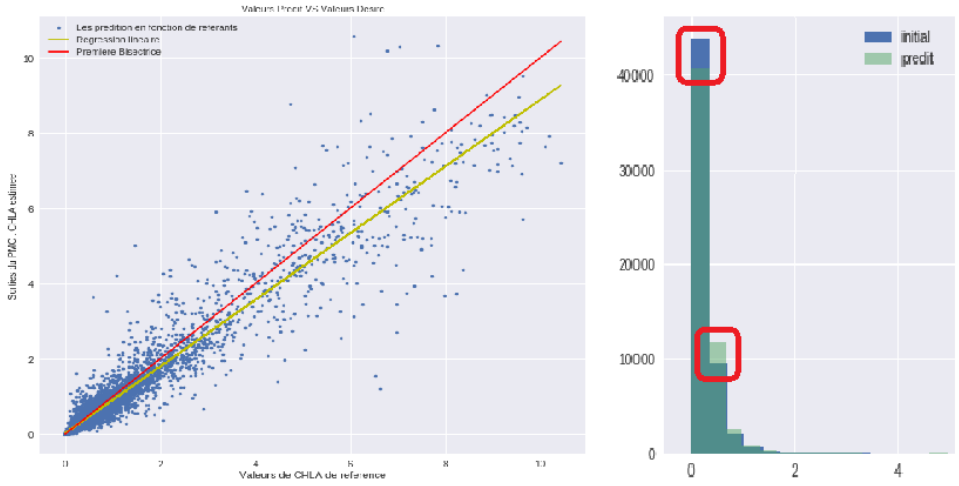


Figure 4.6.1: Représentation de l'angle à la bissectrice et de l'histogramme pour le modèle mixte.

On peut voir que le modèle échoue à prédire quelques valeurs qui sont à 0 et les prédit strictement entre zéro et 1 . Sinon pour le reste des valeurs, d'après la visualisation de l'histogramme on peut dire qu'il réussit assez bien à prédire les vraies valeurs.

4.7 Nombre d'itérations en apprentissage

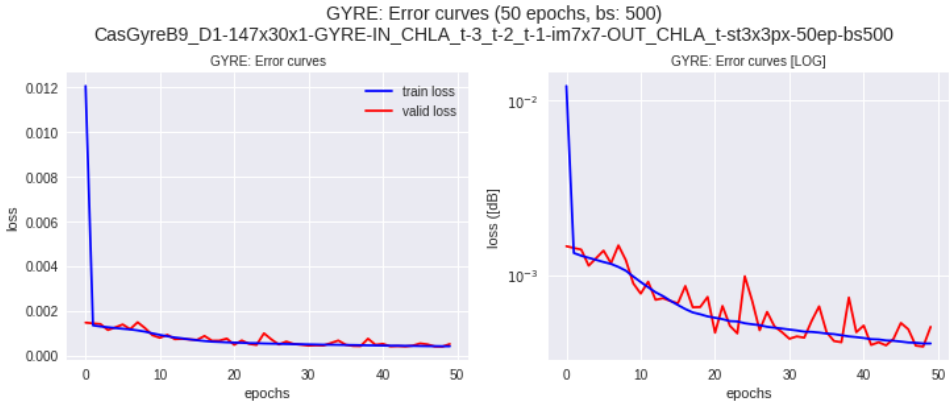


Figure 4.7.1: Décroissance des erreurs d'apprentissage et de validation en fonction des itérations.

Les résultats du tableau sont obtenus en faisant l'apprentissage sur 50 itérations. Mais on remarque que l'erreur de validation à chaque itération n'arrête pas de diminuer. Pour cela, après avoir choisi le meilleur modèle, nous avons décidé d'augmenter le nombre d'itérations jusqu'à 200 mais en utilisant une méthode de choix de la meilleure itération. Cette méthode consiste à sauvegarder les itérations où il y a une amélioration de l'erreur de validation et à la fin de l'apprentissage on ne garde que la dernière itération où une amélioration a eu lieu. Le but de cette méthode est de trouver localement le meilleur résultat. Mais, généralement parlant, le modèle peut supporter un nombre plus grand d'itérations car même avec 200 itérations on n'arrive pas à une erreur constante. L'apprentissage sur 200 itérations a amélioré les résultats: on a un augmentation de 0,004 pour la corrélation et l'angle est devenu 0,008 alors qu'il était à 0,038.

A mentionner que pour l'apprentissage avec 50 itérations a mis 15 minutes, pour 50 itérations avec 32 filtres il a mis 40 minutes, pour 70 itérations il a mis 35 minutes et pour 200 itération a mis 8 Heures.

4.8 Robustesse du modèle choisi

Pour étudier la physique du modèle nous avons décidé de le tester sur un autre ensemble de données de la même année que l'ensemble test précèdent mais qui englobe la 2ème partie de l'année: en effet le meilleur modèle choisi a été testé sur un ensemble constitué des images des six premier mois de l'année 2054, cette durée englobe celle utilisées dans l'ensemble d'apprentissage et de validation (mais sur différentes années). Pour cela, il était intéressant de voir la robustesse du modèle sur la deuxième partie de l'année 2054 surtout que dans cette période la concentration de la chlorophylle est basse et varie de façon négligeable. Les deux tableaux suivants montrent les résultats du modèle sur le premier ensemble test (les six premiers mois) et sur le deuxième ensemble test (dernier six mois).

Données	RMS	RMS R	Correlation	Kullback	Angle a la Bisectrice
Test1	0.12892	0.24866	0.96	-6229.355	0.038
Test2	0.043258	0.19698	0.93	-5648.190	0.013

Table 4.8.1: Mixte (TD CNN+LSTM)

Nous remarquons que les résultats obtenus pour le deuxième ensemble test sont meilleurs (sauf pour la corrélation). Ceci s'explique par le fait que dans le deuxième ensemble, il y a eu peu d'évolution au niveau de la concentration de la chlorophylle, alors que dans le premier il y a des évolutions remarquable au niveau de la concentration, donc plus de détails dans les images ce qui rend difficiles la prédiction. De plus, la majorité des valeurs dans l'ensemble test2 sont compris entre zéro et 1 et le modèle a une tendance à prédire plus des valeurs dans cet intervalle.

4.9 Étude de l'horizon

Jusqu'à présent, nous avons utilisé les modèles pour prédire une valeur de la chlorophylle à l'instant t en se basant sur sa valeur aux trois instants précédents. Mais le but est de pouvoir prédire à des instants postérieurs à l'instant t. Pour cela nous avons testé notre modèle sur un horizon de 10 pas de temps: on se base sur des

images de chlorophylle prises à trois temps consécutives t-3, t-2, t-1, on prédit t, et on utilise cette valeur prédite en plus des deux dernières valeurs (c'est à dire t-2 et t-1) pour prédire la valeur à l'instant t+1 et ainsi de suite jusqu'à arriver à prédire à l'instant t-10 en se basant sur les valeurs prédites aux instant t-7, t-8, t-9. De cette façon, nous avons pu prédire les valeurs de la chlorophylle sur 10 pas de temps en n'ayant que les valeurs à trois instants de temps.

D'autres part, nous nous sommes intéressés à évaluer la persistance et ceci en comparant la concentration de la chlorophylle à l'instant t à celle à l'instant t-1, donc voir l'évolution de la concentration pas à pas. Cette approche doit être normalement la moins bonne comparant au modèle.

Les figures suivantes montrent l'évolution de trois critères de validations qui sont la RMS, la corrélation et l'angle à la bissectrice en fonction des pas de temps et ceci pour la persistance et la prédiction à long terme (horizon).

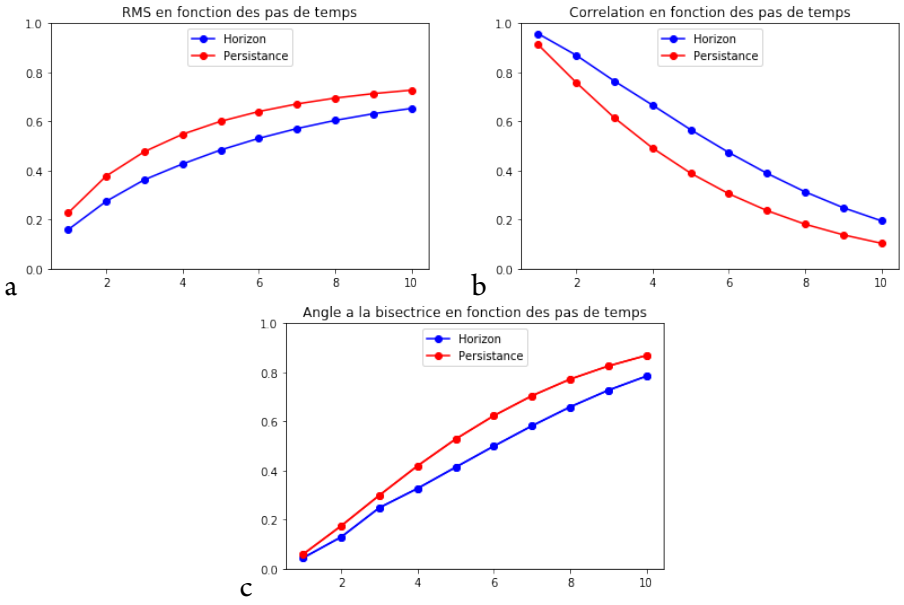


Figure 4.9.1: Représentation de la RMS (a), la corrélation (b) et l'angle à la bissectrice (c) en fonction des pas de temps.

Ces résultats nous montrent que le modèle et la persistance ont presque la même performance au premier pas de temps, celle du modèle commence à devenir

meilleure que la persistance au 4ème pas de temps et cet écart reste presque constant jusqu'au 10ème pas de temps où la RMS est de 0.72 pour la persistance et de 0.65 à l'horizon de même l'angle à la bissectrice est de 0.1 pour la persistance et de 0.2 à l'horizon alors que la corrélation est de 0.86 pour la persistance et de 0.78 à l'horizon . On peut donc conclure qu'à long terme le modèle a une efficacité remarquable.

4.10 Introduction de la SST

La variable SST est la température à la surface de l'océan. L'utilisation d'une autre variable n'était pas cohérente avec notre problème surtout dans le cas de prédiction de la chlorophylle à un temps t à partir des trois temps précédents car la variable à expliquer (la chlorophylle) est très corrélée avec les images utilisées donc l'utilisation d'une autre variable ne fait que réduire la performance du modèle. On peut dire que les modèles considèrent cette variable comme un bruit. Pourtant l'utilisation de cette variable supplémentaire peut être intéressante dans le cas de prédiction des valeurs manquantes.

Ceci se confirme par les résultats obtenus: RMS de 0.19678, une erreur Kullback de 2684.748, une corrélation de 0.95 et un angle à la bissectrice de 0.10

En comparant ces résultats aux résultats obtenus précédemment (ceux sans SST), on remarque que la performance du modèle a diminué.

Donc malgré l'apport que devait donner cette nouvelle variable, nous n'avons pas observé d'améliorations.

4.11 Application sur les données réelles

Jusqu'à présent, toutes les données que nous avons utilisées sont des données générées pour les années 2052, 2053 et 2054. Il était aussi intéressant de voir le comportement du modèle sur des vraies données. Pour cela, nous avons testé le modèle sur des données réelles: il s'agit de toutes les images de la chlorophylle prises en 2008. Avant de les utiliser, nous avons nettoyé les données vu qu'il y avait

des données manquantes: nous avons tous d'abord supprimer chaque 3 imagettes si l'une d'elles a un centre vide, ensuite nous avons supprimer les 3 imagettes si dans l'une d'elles contient plus que la moitié des pixels vides et sinon on remplace les pixels vides par la moyenne de l'imagette. Les résultats obtenus pour ces données sont très mauvais: on obtient une erreur de l'angle à la bissectrice qui vaut 0,98, une corrélation qui vaut 0,1 et une erreur RMS qui vaut 0,24. Ceci n'est pas inattendu vu le nombre non négligeables de données manquantes présents, la qualité moyenne des données réelles et la on compatibilité entre ces données et les données simulés. Ceci était le premier résultat obtenu, mais nous n'avons pas eu le temps de traiter ce problème .

4.12 Conclusion

Dans ce chapitre, nous avons choisi le meilleur modèle qui performe le plus pour nos données. Comme prévu, les modèles profonds, de récurrence et les modèles à convolution, ont été les meilleurs surtout le mixte entre ces deux types. Sa puissance aussi se traduit par le fait qu'il contient environ 1500 poids sans avoir de sur-apprentissage donc cela nous montre que la structure des donnes n'est pas simple. Le résultat attendu de toute cette modélisation se résume dans les figures ci-dessous où nous avons prédit une image de chlorophylle.

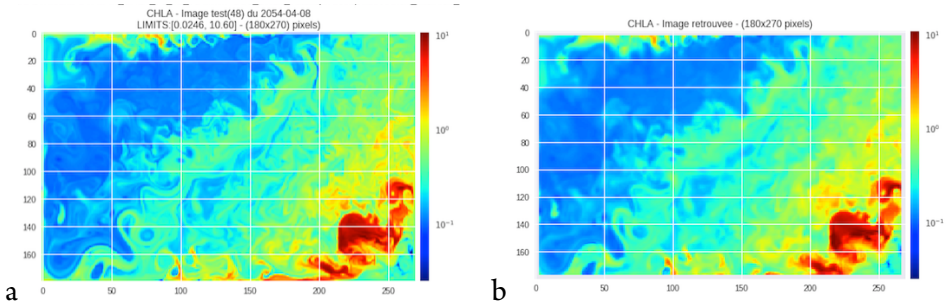


Figure 4.12.1: Image réelle (a) Vs image Prédite (b).

That's been one of my mantras - focus and simplicity. Simple can be harder than complex: You have to work hard to get your thinking clean to make it simple. But it's worth it in the end because once you get there, you can move mountains.

Steve Jobs

5 Conclusion générale

Dans cette analyse, les données sont des séries d'images satellitaires qui représentent la concentration de la chlorophylle. Pour cela, nous nous sommes intéressés à construire des architectures profondes de réseaux de neurones tel que les réseaux récurrents et les réseaux de convolution. Dans la validation du modèle, nous avons choisi plusieurs critères afin d'augmenter la barre de validation. Pourtant, nous avons remarqué que les erreurs calculées dépendent du nombre d'itérations choisi lors de l'apprentissage. Il est tout à fait normale que l'apprentissage sera meilleur quand on augmente le nombre d'itérations, mais pour des raisons de performance de machine nous étions limités de ce côté: pour l'architecture mixte avec 32 filtres, l'apprentissage a pris 1 heure et 40 minutes pour 50 itérations. Mais même avec cette limitation, nous avons eu un modèle qui dépasse en performance la persistance des données, ce qui signifie que la modélisation peut être améliorée d'avantage si les ressources le permettent.

Concernant la prédiction à des instants futurs (horizon), pour aller plus loin, nous pourrions prédire des images complètes c'est à dire pour chaque imagette prédire 7x7 valeurs au lieu d'une unique valeur par imagette et ceci se fait simplement en ajoutant une phase de déconvolution à la fin de l'architecture. L'importance de cette idée est d'éviter les pertes d'informations.

Le Deep Learning a pu construire des modèles de plus en plus complexes grâce à l'augmentation de la puissance de calcul des outils informatiques. Laisser la machine chercher seule les tendances cachées dans les données d'observation plutôt que de définir des modèles puis en circonscrire les limitations à force de simulations numériques pour finalement sélectionner celle s'approchant le plus de la réalité. Mais bien sûr on peut atteindre la réalité plus rapidement selon l'architecture. De nombreuses disciplines scientifiques ont bénéficié. D'après les résultats obtenus dans ce travail on peut aussi dire que la discipline de l'environnement peut aussi en bénéficier.

Références

[1] JOUINI. *Reconstruction des images satellitaires de chlorophylle dont la couverture est très irrégulière étant donné le nombre de données manquantes dû principalement à la couverture nuageuse.*

[2] Miguel Angel PEREZ CHAVARRIA. *Restitution de paramétres atmosphériques hydrologiques sur l'océan par radiometrie hyperfréquence spatiale. Méthodologie neuronale.* 2007.

[3] Julien BRAJARD, LOCEAN/IPSL, Paris, France; and A. A. CHARANTONIS and F. JOURDIN. *Predicting Ocean Dynamics through Machine Learning: Application on Sea-Surface Suspended Particulate Mater.* 2017.

[4] KARPATHY, Stanford. *CS231n Convolutional Neural Networks for Visual Recognition.*

[5] Carlos MEJIA. *DeepGyreNb1-base.* 2017

[6] Carlos MEJIA. *lib-gyre-imagettes.* 2017

[7] Carlos MEJIA. *pour-test-une-image.* 2017

[8] https://keras.io/layers/core/

[9] https://keras.io/layers/convolutional/

[10] https://keras.io/layers/recurrent/

[11] http://colah.github.io/posts/2015-08-Understanding-LSTMs/

[12] http://machinelearningmastery.com/time-series-prediction-lstm-recurrent-neural-networks-python-keras/