

Projet Long ENSIIE

Prédition des profils verticaux de la CHL-A à partir des données surfaciques

Réalisé par :

DORRA BENNOUR

Supervisé par:

M. ANASTASE CHARANTONIS

Table des matières

Introduction Générale	1
1 Etude exploratoire	1
1.1 Données	1
1.2 Analyse unidimensionnelle	2
1.2.1 Profils verticaux	2
1.2.2 Données de surface	3
1.2.3 Nouvelles variables	4
2 Méthodologie	4
2.1 Principe des TCN	4
2.1.1 Motivation	4
2.1.2 Modèle	5
2.1.3 Prédiction	7
2.2 Répartition de la base d'entraînement et de test	8
2.3 Optimisation du modèle	8
2.3.1 Validation Croisée	9
3 Résultats	9
3.1 Paramètres et performances	9
3.2 Reconstruction des années 2007-2008	10
3.3 Expérimentations	12
3.4 Etude comparative	14
Conclusion et perspectives	15
Annexe	16
Références	19

Table des figures

1	Localisation des points de mesure	2
2	Profils verticaux de la chlorophylle-a entre 1992 et 2008	2
3	Profils verticaux de la Température à la surface entre 1992 et 2008	2
4	Profil de l'année moyenne en CHL-A	3
5	Profil de l'année moyenne en SST	3
6	Exemple d'une couche convolutionnelle 1D	5
7	Exemple de couche 2-dilatée avec une longueur d'entrée de 4 et une taille du filtre de 3	6
8	Exemple d'un réseau convolutionnel temporel simple	7
9	Exemple d'une architecture de TCN complète [4]	8
10	RMSE en fonction des différentes profondeurs	10
11	Résultats de la reconstruction des profils verticaux de la CHL-A des années 2007 et 2008- La première figure correspond aux profils verticaux des données réels-La deuxième figure correspond aux profils verticaux des données prédites	11
12	Profils des résidus	11
13	Diagramme de dispersion des profils prédicts en fonction des profils prédicts	12
14	Résultats de la reconstruction des profils verticaux de la CHL-A des années 2007 et 2008- La première figure correspond aux profils verticaux des données réels-La deuxième figure correspond aux profils verticaux des données prédites	13
15	Les résidus par le modèle TCN en fonction de la variable cible	13
16	RMSE selon le modèle de l'expérience 1 en fonction des différentes profondeurs	14
17	Moyenne des variables de surface sur les 9 points en fonction du temps	16
18	Matrice de corrélation entre les différentes variables	16
19	Diagramme de dispersion entre les différentes variables	16
20	Matrice de corrélation des différentes variables en tenant compte des profondeurs	17
21	Résultats de la reconstruction des profils verticaux de la CHL-A des années 2007 et 2008 pour une fenêtre de temps égale à 20- La première figure correspond aux profils verticaux des données réels-La deuxième figure correspond aux profils verticaux des données prédites	18
22	Résultats de la reconstruction des profils verticaux de la CHL-A des années 2007 et 2008 pour une fenêtre de temps égale à 30- La première figure correspond aux profils verticaux des données réels-La deuxième figure correspond aux profils verticaux des données prédites	18

Liste des tableaux

1	Statistiques élémentaires des données de surfaces	3
2	Les paramètres du modèle TCN	9
3	Performances du modèle TCN sur en validation et en test	10
4	Les paramètres du modèle TCN pour la première expérience	12
5	Performances du modèle TCN sur en validation et en test	14
6	Tableau comparatif des modèles TCN et PROFHMM	15

Liste des abréviations

- CC** Couverture Nuageuse (Cloud Cover)
- CHL-A** Chlorophylle-A
- EWCV** Expanding Window Cross Validation
- PROFHMM** Reconstruction de profils par HMM (PROFile reconstruction through HMM)
- SR** Radiance Solaire à courte ondes (Shortwave Radiataion)
- SSH** Elevation du niveau de la mer (Sea-Surface Elevation)
- SST** Temperature de surface (Sea-surface Temperature)
- TCN** Temporal Convolutional Network
- WS** Vitesse du vent (Windspeed)

Introduction Générale

La connaissance de la structure verticale des propriétés biogéochimiques de l'océan est cruciale pour l'estimation de la production primaire, la distribution du phytoplancton et la modélisation biologique. Les données satellitaires, fournissent des mesures de la surface de l'océan qui sont caractérisées par une haute résolution spatiale et temporelle. Cependant, les variables physiques en profondeur sont mesurées via des mesures dites *in situ*. Ces mesures sont généralement rares et non uniformément distribuées dans le temps et plus particulièrement dans l'espace en comparant avec les données satellitaires. Par conséquent, l'obtention des estimations des profils verticaux du champ des variables biogéochimiques notamment la Chlorophylle-a (CHL-A) et de la température à partir d'observations de surface relève un défi scientifique critique pour mieux contraindre l'étude des océans. Les premières études établant ce sujet ont commencé en 1989, cependant, la plupart des travaux précédents ne se sont pas basés sur des techniques d'apprentissage automatique, dont l'utilisation s'est accentuée ces dernières années. Aujourd'hui, en effet, la plupart de ces techniques sont employées pour étendre les informations biologiques de surface à des couches plus profondes. Un exemple est celui de Charantonis et al. [1], qui ont présenté le modèle "PROFHMM" permettant l'utilisation combinée d'une carte auto-organisatrice et de modèles de Markov cachés pour inférer des champs de CHL-A tridimensionnels à partir de des données de surface de plusieurs variables.

Dans cette étude, on cherche à reconstruire les profils verticaux de la CHL-A au cours du temps à partir des données de surfaces par des modèles neuronales. Nous avons alors utilisé une récente architecture de réseaux neuronaux dite "Réseaux convolutionnels temporels", une technique permettant de modéliser la nature non linéaire de la relation entre les différentes variables d'entrées ainsi que leur évolution temporelle. Bien que les architectures de réseaux de neurones artificiels soient déployées auparavant par Sammartino et al. [3], Les réseaux TCN se différencient principalement par le fait qu'ils tiennent compte de la causalité entre l'entrée à une date t et une entrée à une date $t-1$. L'emploi ce lien causal semble alors adapté pour prédire l'évolution de la distribution verticale de la CHL-A à partir des différentes variables de surface.

1 Etude exploratoire

1.1 Données

L'étude de notre système porte principalement sur le système phytoplanctonique dans l'océan, particulièrement, dans un carré $2^\circ \times 2^\circ$ dans l'océan atlantique (Bermudes) BATS "Bermuda Atlantic Time Series". Comme les organismes phytoplanctoniques sont caractérisés essentiellement par la Chlorophylle-A (CHL-A), on s'intéresse donc à cette variable comme notre variable cible tout au long de l'étude. On dispose de 17 années (1992-2008) de données prélevées de la zone BATS (voir figure 1.1). Généralement la croissance du phytoplancton dépend principalement de cinq paramètres : le rayonnement disponible en ondes courtes, les nutriments disponibles, les herbivores et la biologie, la température de l'eau, la turbidité de l'eau. Ces paramètres ne peuvent cependant pas être facilement contrôlés par une approche directe. L'imagerie par satellite quant à elle peut nous fournir des informations de substitution, qui peuvent être utilisées dans une approche empirique pour déterminer les profils verticaux de chlorophylle-a. Plus précisément, dans cette étude, nous allons utiliser ces variables :

- Concentration en CHL-A à la surface de l'océan (SCHL)
- Température à la surface de l'océan (SST)
- Elévation du niveau de la mer (SSH)

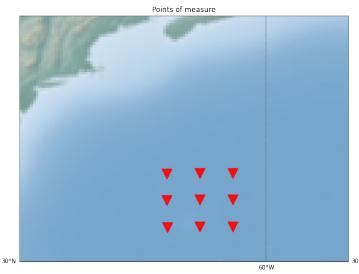


FIGURE 1 – Localisation des points de mesure

- Radiance solaire (SR)
- vitesse du vent (WS)
- Couverture nuageuse (CC)

Notre base de données, extraite à partir d'un modèle produit par le modèle de circulation océanique NEMO couplé au modèle biogéochimique PISCES, est décrite par dix-huit niveaux de profondeurs allant de 5 mètres jusqu'à 217 mètres, moyenné sur cinq jours sur chaque année pour la chlorophylle-a et la température.

La base de données est alors constitué de 11,196 lignes (17 années * 73 mesures/ année * 9 points de mesure)

et de sept variables (explicatives et cibles). Dans la suite, on considère la proximité des neufs points comme très faible et donc on ne prend pas les latitudes et les longitudes comme variables explicatives.

1.2 Analyse unidimensionnelle

1.2.1 Profils verticaux

Afin de mieux comprendre les données dont on dispose, on procède par une analyse unidimensionnelle de chaque variable à part. On commence alors par la représentation des profils verticaux au cours du temps des variables CHL-A (voir Figure 2) et SST (Voir Figure 3) :

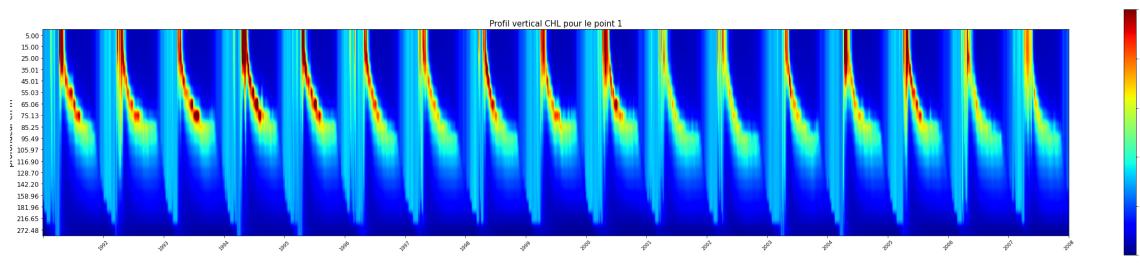


FIGURE 2 – Profils verticaux de la chlorophylle-a entre 1992 et 2008

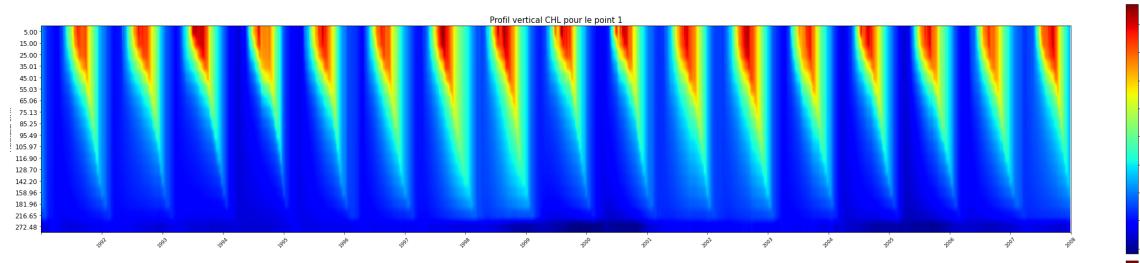


FIGURE 3 – Profils verticaux de la Température à la surface entre 1992 et 2008

Une saisonnalité considérable est mise en évidence par ces figures pour chacune des variables CHL-A et SST. En effet, concernant la chlorophylle-a, on remarque que sa concentration est d'autant plus forte durant le printemps jusqu'à l'automne où la concentration est plus forte plutôt en profondeurs (entre 5 et 85 mètres). Les fortes températures de surfaces, se manifestent généralement durant les saisons chaudes (Mai-Septembre) sur la surface et dans les profondeurs. Ainsi, afin d'éliminer l'effet de la saisonnalité et la non-stationnarité de nos variables on opte pour la désaisonnalisation de nos données, c'est à dire on soustrait la moyenne de toutes les années de chaque ligne.

Le profil de l'année moyenne en CHL-A et SST peut être vu par les Figures 4 et 5 suivantes :

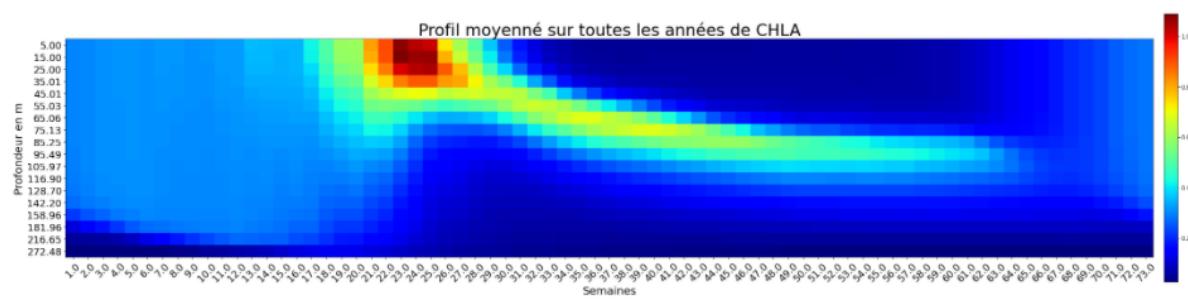


FIGURE 4 – Profil de l'année moyenne en CHL-A

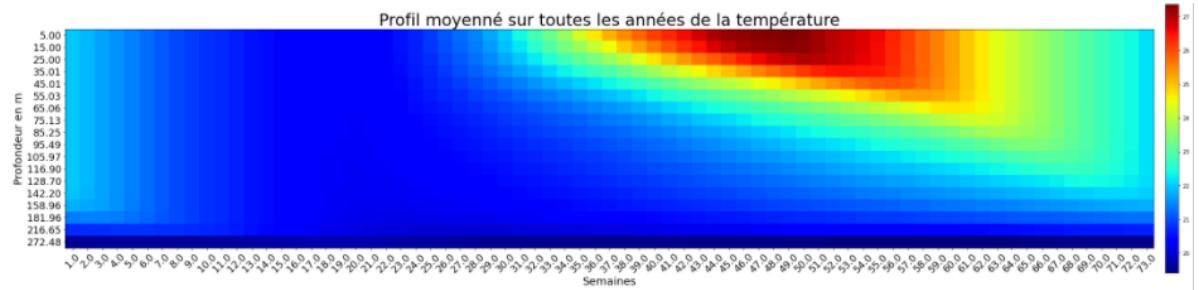


FIGURE 5 – Profil de l'année moyenne en SST

1.2.2 Données de surface

Afin de mieux saisir l'ordre de grandeur des différentes variables, on résume les statistiques élémentaire par la table ci-dessous :

	SCHL	SST	SSH	CC	WS	SR
Moyenne	0.23	22.05	-0.78	0.56	6.73	165.58
Ecarts type	0.07	1.28	0.07	0.11	1.87	60.26
Min	0.14	19.60	-0.99	0.16	2.05	60.54
Max	0.59	24.64	-0.57	0.86	12.66	290.52

TABLE 1 – Statistiques élémentaires des données de surfaces

On tire d'après les écarts de la table ci-dessus que les différentes variables n'ont pas le même ordre de grandeur, allant de 0.07 pour la concentration de la CHL-A à la surface de l'océan jusqu'à 60.26 pour la radiation solaire. Ce qui nous conduit par la suite au centrage et réduction des données pour que toutes les variables aient le même poids en apprentissage.

Dans notre analyse on a discerné des indicateurs de saisonnalité des données à disposition (voir Figure 17 en annexe) pour chacune des variables SST et SSH, plus particulièrement ces deux variables semblent avoir le même comportement au cours du temps ce qui induit à une corrélation entre elles. Les radiations solaires quant à elles, définies par la variables SR, ont été très importantes entre les années 1997 et 2003. La chlorophylle-a à la surface de la mer cependant, n'est pas régit d'une saisonnalité apparente (La saisonnalité réside plutôt en son comportement interannuel, puisqu'il s'agit d'un pigment dépendant de beaucoup de facteurs).

1.2.3 Nouvelles variables

Variable Date

Comme les paramètres d'entrée sont fortement liés à l'écologie du problème étudié et comme les variables spatiales et temporelles tels que le jour de l'année ,latitude et longitude, sont importantes pour fournir des informations implicites sur la saison ou la géolocalisation du profil à prédire, ainsi que la température à la surface de la mer, qui peut être un facteur important dans la prédition du profil vertical de la température à prédire, qui, seule ou combinée aux autres entrées, influence profondément le profil vertical du CHL-A. Avant l'entraînement, les valeurs relatives à la date ont été mises à l'échelle dans un intervalle de [0, 1] afin de tenir compte de la cyclicité et la saisonnalité du problème étudié. La nouvelle variable date est alors défini comme suit :

$$\left[\cos\left(\frac{2\times5\times\pi\times\text{semaine de l'année}}{365}\right) \mid \sin\left(\frac{2\times5\times\pi\times\text{semaine de l'année}}{365}\right) \right]$$

Variable Cible

En se basant sur les travaux la thèse de Charantonis et al. [1], on introduit la normalisation logarithmique pour CHL-A. Notre nouvelle variable devient alors $\log(1+\text{CHL-A})$ pour chaque profondeur. Ce type de normalisation est en effet, couramment utilisé en analyse biogéochimique, puisqu'il permet la remise à échelle de la variable.

2 Méthodologie

2.1 Principe des TCN

Le but de ce projet est de trouver une méthode appropriée à la problématique étalée : "Inversion des données de surface pour la reconstruction des profils verticaux de la CHL-A". Tout au long de ce chapitre on expliquera la méthodologie suivie qui consiste principalement à appliquer des réseaux convolutionnels temporels dits TCN.

2.1.1 Motivation

Bien que communément associés à des tâches de classification d'images, les réseaux de neurones convolutifs (CNN) se sont révélés être des outils précieux pour la modélisation et la prévision de séquences. Un processus convolutionnel comprend principalement deux étapes :

premièrement, le calcul de caractéristiques de bas niveau en utilisant (généralement) des CNN qui codent des informations spatio-temporelles et deuxièmement, l'entrée de ces caractéristiques de bas niveau dans un classificateur qui capture des informations temporelles de haut niveau en utilisant (généralement) des RNN. Le principal inconvénient d'une telle approche est qu'elle nécessite deux modèles distincts. Le modèle TCN quant à lui, fournit une approche unifiée pour saisir les deux niveaux d'information de façon hiérarchique, tout en respectant la causalité qui régit les données.

2.1.2 Modèle

Un modèle simple d'un réseau convolutionnel temporel comprend trois couches principales :

1. Réseau convolutionnel 1D
2. Convolution causale
3. Dilatation

Réseau convolutionnel 1D

Un réseau convolutif 1D (Voir Figure 6) prend en entrée un tenseur tridimensionnel et sort également un tenseur tridimensionnel. Le tenseur d'entrée de notre implémentation TCN est de dimension (dimension du lot, longueur de l'entrée, dimension de l'entrée) et le tenseur de sortie est de dimension (dimension du lot, longueur de l'entrée, dimension de la sortie). Comme chaque couche d'un TCN a la même longueur d'entrée et de sortie, seule la troisième dimension des tenseurs d'entrée et de sortie diffère. Une seule couche convective 1D reçoit un tenseur d'entrée de forme (dimension du lot, longueur de l'entrée, profondeur de l'entrée) et produit un tenseur de forme (dimension du lot, longueur de l'entrée, profondeur de la sortie). Pour obtenir la sortie, nous prenons le produit scalaire de la sous-séquence de l'entrée et un filtre de poids appris de même longueur. Pour obtenir l'élément suivant de la sortie, la même procédure est appliquée, mais la fenêtre du filtre de la séquence d'entrée est décalée vers la droite d'un seul élément.

Remarque : Pour s'assurer que la séquence de sortie ait la même longueur que la séquence d'entrée, on applique du zero-padding au début de la séquence.

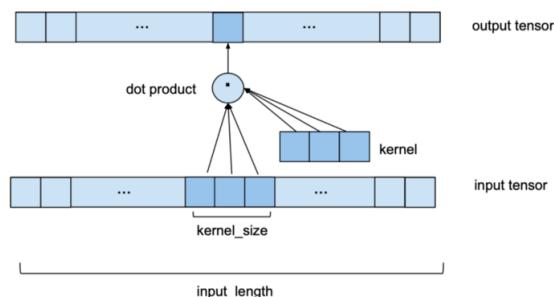


FIGURE 6 – Exemple d'une couche convolutionnelle 1D

Convolution causale

Pour qu'une couche convolutionnelle soit causale, il faut que pour chaque i dans $\{0, \dots, \text{input_length} - 1\}$ le i -ème élément de la séquence de sortie ne peut dépendre que des éléments

de la séquence d'entrée avec les indices 0, ..., i. En d'autres termes un élément de la sortie ne peut dépendre que des éléments qui lui précèdent.

Dilatation

Une qualité souhaitable d'un modèle de prévision est que la valeur d'une sortie spécifique dépend de toutes les entrées précédentes, c'est-à-dire de toutes les entrées qui ont un indice inférieur ou égal à lui-même. Ceci est réalisé lorsque le champ réceptif, c'est-à-dire l'ensemble des entrées de l'entrée originale qui affectent une entrée spécifique de la sortie, a la taille "longueur de l'entrée". Nous appelons également cela "couverture complète de l'historique". Plus généralement, un réseau convolutionnel 1D à n couches et un filtre de taille k a un champ réceptif de taille :

$$r = 1 + n * (k - 1)$$

Pour une couverture complète on a alors :

$$n = \frac{l - 1}{k - 1}$$

où l est la longueur de l'entrée.

k est la dimension du filtre.

Le nombre de couches étant proportionnel à la longueur de l'entrée, ceci peut alors causer une dégradation des performances puisque le nombre de paramètres augmentera considérablement. Dans le contexte d'une couche convulsive, la dilatation fait référence à la distance entre les éléments de la séquence d'entrée qui sont utilisés pour calculer une entrée de la séquence de sortie. Plus généralement, une couche d-dilatée avec un filtre de taille k a un champ de réception s'étendant sur une longueur de $1 + d * (k - 1)$.

En général, pour un champ réceptif sans trous, la taille du filtre k doit être au moins aussi grande que la base de dilatation b. Ainsi, une couche convulsive classique peut être considérée comme une couche 1-dilatée, puisque les éléments d'entrée pour une valeur de sortie sont adjacents. La figure suivante montre un exemple d'une couche 2-dilatée avec une longueur d'entrée de 4 et une taille du filtre de 3.

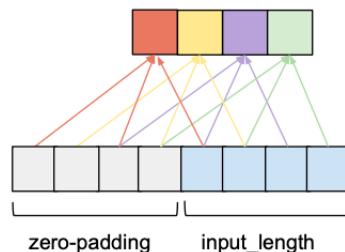


FIGURE 7 – Exemple de couche 2-dilatée avec une longueur d'entrée de 4 et une taille du filtre de 3

Si d est fixe, il faudra encore un nombre linéaire dans la longueur du tenseur d'entrée pour obtenir une couverture complète du champ réceptif. Ce problème peut être résolu en augmentant la valeur de d de manière exponentielle au fur et à mesure que l'on monte dans les couches. Pour cela, nous choisissons une constante entière b dite "base de dilatation" qui nous permettra de

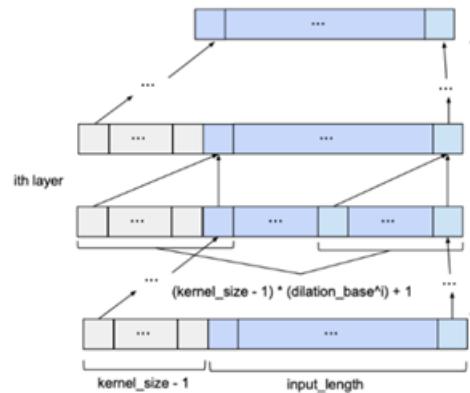


FIGURE 8 – Exemple d'un réseau convolutionnel temporel simple

calculer la dilatation d d'une couche spécifique en fonction du nombre de couches inférieures, i , puisque $d = b^i$.

Ainsi, compte tenu de la longueur d'entrée, de la taille du filtre, de la base de dilatation et du nombre minimum de couches requises pour une couverture historique complète, le réseau TCN de base ressemblerait à Figure 8

2.1.3 Prédiction

Pour former un réseau TCN à la prévision, l'ensemble d'entraînement consistera en (séquence d'entrée, séquence cible) : paires de sous-séquences de même taille de la série temporelle donnée. Une série cible sera alors une série qui est décalée vers l'avant par rapport à sa série d'entrée respective d'un certain nombre de longueur de sortie "ls". (une série cible de longueur "le" contient les derniers éléments (le - ls) de sa séquence d'entrée respective comme premiers éléments, et les éléments "ls" qui viennent après la dernière entrée de la série d'entrée comme ses éléments finaux. On choisit dans notre cas un horizon de prédition de 73 valeurs (nombre de 5 jours) et donc une année, puisqu'on s'intéresse au comportement interannuel en profondeurs de la chl-a et de la température.

Finalement un exemple d'un modèle TCN ressemblera à la structure de la Figure 9 :

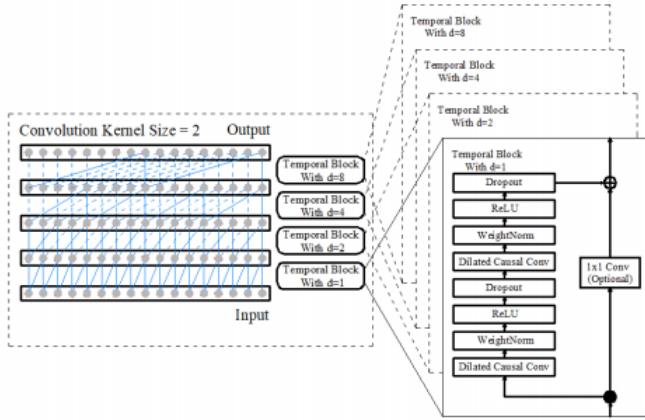


FIGURE 9 – Exemple d'une architecture de TCN complète [4]

Parmi les améliorations apportées au modèle, les blocs résiduelles permettant de diminuer davantage le facteur de dilatation. En effet, un bloc résiduel empile deux couches de convolution causale dilatée, et les résultats de la convolution finale sont rajoutés aux entrées pour obtenir les sorties du bloc. Si la largeur (nombre de canaux) des entrées et la largeur (nombre de filtres) des deuxièmes couches de convolution causale dilatées diffèrent, nous devrons appliquer une convolution 1D aux entrées avant d'ajouter les sorties de convolution pour faire correspondre les largeurs. D'une manière plus générale l'entraînement du réseau convolutionnel temporel consiste en l'entraînement des blocs résiduels où chaque bloc peut être défini par deux couches de ce type :

1. Une couche de convolution causale à dilatation d
2. Une normalisation (de lot, de couches ou de poids)
3. Une fonction d'activation
4. Autres hyperparamètres (Dropout, Régularisation,... etc)

2.2 Répartition de la base d'entraînement et de test

Pour entraîner le réseau temporel convolutionnel on choisit de répartir notre base de données comme suit :

- **Données tests** : Comme nous nous intéressons à l'aspect temporel de données, on choisit de prendre les données relatives aux deux dernières années (2007 et 2008) comme données test.
- **Données d'entraînement** : On considère le reste des données comme données d'entraînement.
- **Validation** : Pour optimiser les paramètres du modèle, on procède pour une validation croisée sur l'ensemble de la base d'entraînement.

2.3 Optimisation du modèle

On opte à l'optimisation du modèle par deux types de validation croisée. Les hyperparamètres (entre autres le type de la normalisation des données et son impact) sont recherchés aléatoirement.

2.3.1 Validation Croisée

Dans le cas des séries temporelles, il n'est pas aussi évident d'utiliser les validations croisées classiques puisque on évalue le modèle sur des ensembles indépendants. Ainsi on optera pour deux approches similaires pour la validation croisée :

- **Validation croisée à fenêtre croissante (Expanding window cross validation EWCV)** : La méthode consiste à évaluer en premier lieu un petit sous-ensemble de données, à chaque fois on rajoute un autre sous-ensemble à l'ensemble précédent et on ré-évalue le modèle jusqu'à la fin du tableau de données. Un inconvénient de cette méthode consiste en la possibilité d'introduction du biais lors de l'entraînement puisque certaines données ne sont pas vues pour la première fois. Dans ce cas on ajoute à chaque fois 73 valeurs, ce qui correspond à une année de mesures au sous-ensemble précédent.
- **Validation croisée par blocs de séries temporelles (Blocking Time Series)** : Comme le nom l'indique il s'agit une validation croisée où on évalue le modèle sur des sous-ensembles de séries temporelles successives. Dans notre cas on entraîne à chaque fois sur deux années de mesures et on évalue sur une seule année.

3 Résultats

3.1 Paramètres et performances

Nos données d'entraînement correspondent à une matrice à 11,169 lignes et 48 variables (6 variables de surfaces, 18 variables de profils verticaux de la température, 18 variables de profils verticaux de la CHL-A, année, le nombre des 5 jours, sinus de la date et cosinus de la date). Pour constituer notre entrée, on translate chaque année par un pas de temps (5 jours dans notre cas). La variable cible quant à elle, consiste en les profils verticaux de la CHL-A, c'est à dire une matrice à 18 variables de profondeurs.

Nous avons optimisé le modèle de TCN selon les deux types de validation croisée, cependant, dans la suite on montrera seulement les résultats trouvés par la validation croisée à fenêtre croissante. L'architecture du modèle retenu après optimisation aléatoire est alors défini comme suit :

Paramètres	Valeurs
Fonction coût	Erreur quadratique
Fonction d'activation	Leaky relu : $f(x) = \max(\epsilon x, x)$
padding	causal (Puisqu'on s'intéresse à des séries temporelles)
Diatations à base 2	[1,2,4,8]
Nombre de filtres	64
taille des filtres	3
Optimisateur	Adam
Nombre de couches résiduelles	2
Type de normalisation	Normalisation de couche
ratio dropout	0.05
Régularisation	0.02

TABLE 2 – Les paramètres du modèle TCN

Tous ces paramètres nous donnent un champs de réception égale à 241.

Comme on manipule des données numériques, on évalue le modèle selon six erreurs :

- Erreur quadratique : $MSE = \frac{1}{n} \sum (y_i - \hat{y}_i)^2$
- Erreur absolue moyenne : $MAE = \frac{1}{n} \sum |y_i - \hat{y}_i|$
- Racine carrée de l'erreur quadratique (fonction coût) : $RMSE = \sqrt{\frac{1}{n} \sum (y_i - \hat{y}_i)^2}$
- RMSE sur 9 tranches de profondeurs
- RMSE pour les 10% des valeurs les plus fortes qu'on note par la suite RMSE 10FO
- RMSE pour les 10% des valeurs les plus faibles qu'on note par la suite RMSE 10FA

L'application du modèle défini par l'architecture décrite précédemment nous donne les résultats suivants :

TCN BTSCV	MSE	MAE	RMSE	RMSE 10FA	RMSE 10FO
Données test	0.0032	0.0351	0.0518	0.03663	0.4734

TABLE 3 – Performances du modèle TCN sur en validation et en test

Pour explorer davantage ces erreurs on représente la courbe des erreurs RMS en fonction des différentes profondeurs :

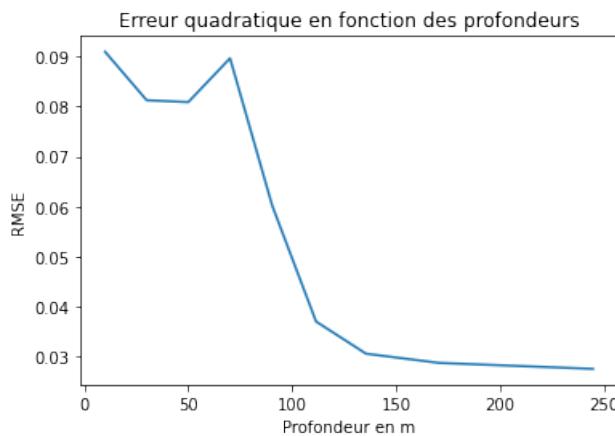


FIGURE 10 – RMSE en fonction des différentes profondeurs

On remarque d'après la Figure 10, que l'erreur en question est très élevée pour les 80 premiers mètres. En effet, ces profondeurs peuvent correspondre à des couches de mélanges, c'est à dire, pour un paramètre physique donné de l'état de l'océan par exemple : la température, la densité..., qui est supposé être mélangé et homogène à un certain niveau (profondeurs de l'océan dans notre cas). Ce qui peut entraîner des valeurs très proches des concentrations de la CHL-A et donc c'est plus difficile à les prédire. Ceci peut-être aussi constaté depuis la matrice de corrélation des profondeurs (Figure 20 en annexe), où les concentrations de la CHL-A pour les premières profondeurs sont très corrélées.

3.2 Reconstruction des années 2007-2008

Les données test qu'on a considéré, consiste en les données relatifs aux années 2007 et 2008. Pour évaluer la capacité reconstructive du modèle on représente la reconstruction des profils verticaux des prédictions, montrés par la Figure 11 suivante :

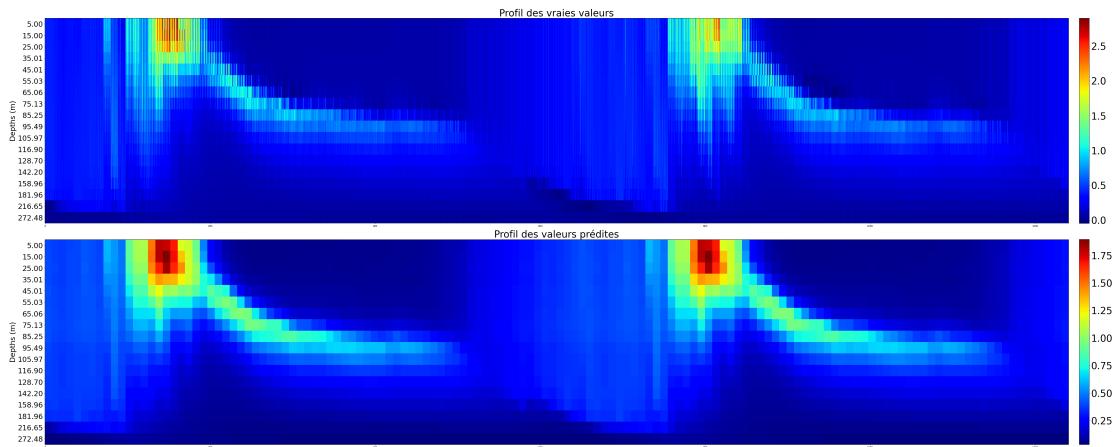


FIGURE 11 – Résultats de la reconstruction des profils verticaux de la CHL-A des années 2007 et 2008- La première figure correspond aux profils verticaux des données réels-La deuxième figure correspond aux profils verticaux des données prédictes

Comme on ne peut apercevoir une grande différence entre ces figures, on représente le profil de la différence :

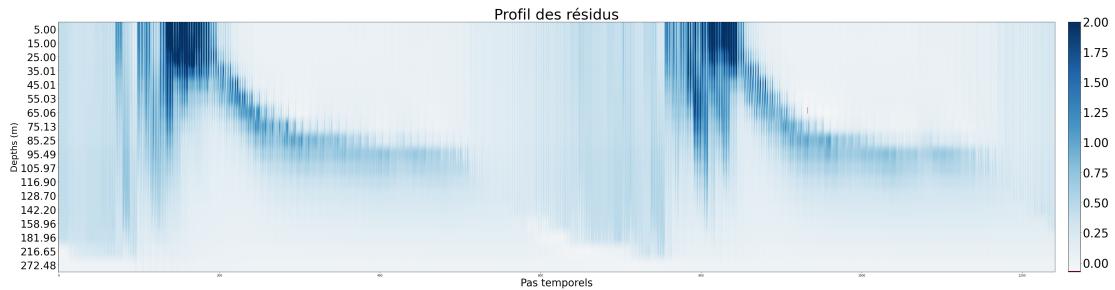


FIGURE 12 – Profils des résidus

La Figure 12 met en lumière les résultats trouvés par le biais du modèle optimale. En effet, les résidus sont en général très proches de zéro ce qui met en valeur la qualité de la prédiction en question. On peut consater celui-ci aussi à partir de la figure 13 des valeurs prédictes en fonction des valurs réelles :

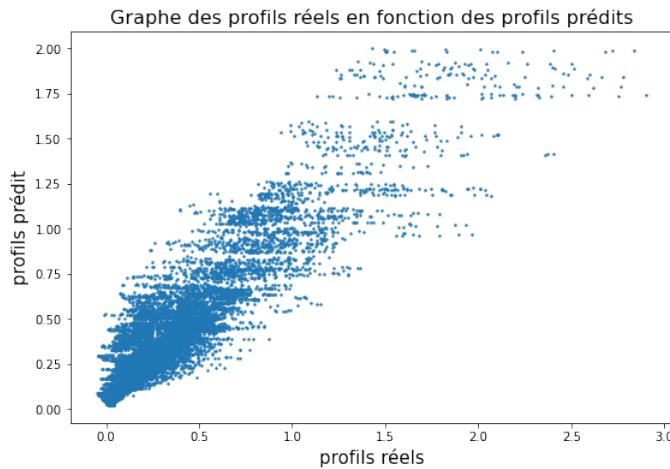


FIGURE 13 – Diagramme de dispersion des profils prédicts en fonction des profils réels

D'après la figure ci-dessous, on peut discerner une relation linéaire entre les deux profils, ce qui confirme les propos avancés précédemment.

3.3 Expérimentations

Afin de mieux évaluer le modèle, et sa capacité à reconstruire les profils verticaux de la chlorophylle-a à partir des données de surface au temps t et des profils verticaux au temps t-1. On évalue le modèle en optant deux approches principales qui sont expliquées comme suit :

1. **Expérience 1** : Une première expérience consiste à évaluer les performances du modèle en lui introduisant seulement les données de surfaces.

On réoptimise alors le modèle (Architecture donné par la table 4) et on analyse les résultats de la même manière qu'auparavant. Les erreurs trouvées sont résumées par la Table 5

Paramètres	Valeurs
Dilatations à base 2	[1,2,4,8]
Nombre de filtres	32
taille des filtres	3
Nombre de couches résiduelles	4
Type de normalisation	Normalisation de couche
Ratio dropout	0.15

TABLE 4 – Les paramètres du modèle TCN pour la première expérience

D'une manière visuelle, on reconstruit le profil vertical de la CHL-A des années 2007 et 2008 (Données de test) à partir du modèle optimale :

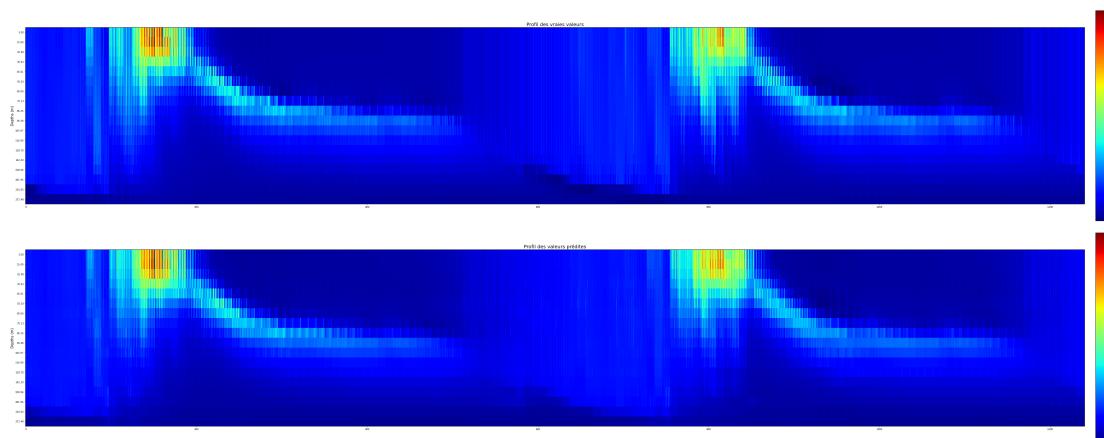


FIGURE 14 – Résultats de la reconstruction des profils verticaux de la CHL-A des années 2007 et 2008- La première figure correspond aux profils verticaux des données réels-La deuxième figure correspond aux profils verticaux des données prédictes

On peut voir, sur la figure 14, que la reconstruction des TCN respecte la forme et l'intensité générale de l'évolution du profil de la chlorophylle-a, tout au long des périodes 2007-2008. Pour pouvoir discerner encore plus la différence, on représente le graphe des résidus ($y_{réel} - y_{prédict}$) :

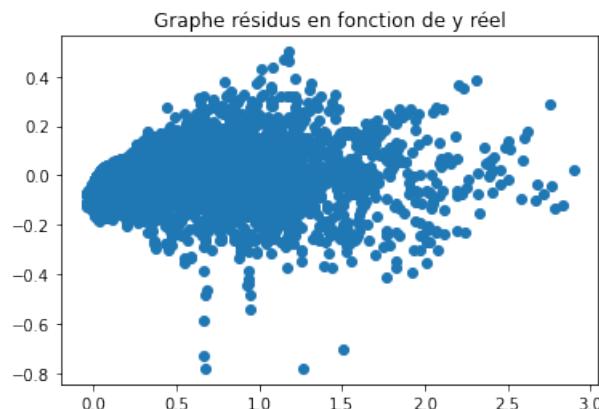


FIGURE 15 – Les résidus par le modèle TCN en fonction de la variable cible

On constate alors d'après la Figure 15 que les résidus n'ont pas une forme particulière, de plus, ils oscillent autour de 0, ce qui met en lumière que l'espérance de ces résidus est presque nulle. On peut dire alors que le modèle a de bonne performances.

Comme on dispose d'un vecteur cible multivariable, où les variables correspondent aux profils verticaux de la CHL-A, on évalue la racine carrée de l'erreur quadratique pour chaque paire de profondeur (Voir Figure 16)

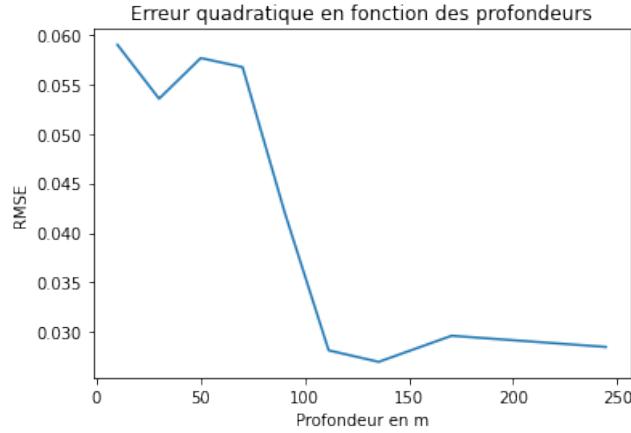


FIGURE 16 – RMSE selon le modèle de l'épérience 1 en fonction des différentes profondeurs

Tout comme la Figure 10, les racines carrées des erreurs quadratiques sont assez élevées pour les 60-80 premiers mètres de profondeur de l'océan. Bien qu'on dispose des données surfaciques en entrée, on s'attend plutôt à avoir des erreurs plus faibles quant à ces profondeurs. Ces résultats peuvent alors s'expliquer par le phénomène des couches mélangées (mixed layer depths) qui se présente à l'océan.

2. **Expérience 2 :** Le modèle optimale a été obtenu en fixant la fenêtre du temps en une année (73 mesures moyennées sur 5 jours), cependant, il est d'autant plus intéressant d'étudier l'impact de la variation de cette fenêtre. Pour ce faire on réapplique le modèle choisi en variant la fenêtre temporelle. On choisit alors des fenêtres de pas temporels égale à 20 et à 30. D'après la Table 5 les erreurs sont très proches l'une de l'autre, la différence n'est donc aussi pas assez perceptible depuis les figures de reconstruction des profils verticaux (Voir Figures 21 et 22 de l'annexe). De plus, il est notable, que le changement des pas temporels a un impact sur la performance du modèle. En effet, on trouve des erreurs de plus en plus faibles en allant d'une fenêtre de 73 pas de temps à une fenêtre de 20 pas de temps. Par contre, une expérience qui suscite l'intérêt c'est d'optimiser cette fenêtre en visualisant les différents filtres des couches convolutionnelles et d'observer jusqu'à quel temps du passé il faut s'arrêter.

Expérience	MSE	MAE	RMSE	RMSE 10FA	RMSE 10FO
Expérience 1	0.0063	0.0583	0.0802	0.026	1.8548
Expérience 2					
Fenêtre = 20	0.0056	0.0504	0.0771	0.0245	0.0245
Fenêtre = 30	0.0059	0.0553	0.0769	0.0248	0.0248

TABLE 5 – Performances du modèle TCN sur en validation et en test

3.4 Etude comparative

En s'appuyant les travaux et les recherches réalisés précédemment, notamment de AA Charantonis et al. [1], les RMSE obtenues sont considérés comme légèrement plus faibles. En outre, comme le montre la Table 6, La sensibilité aux plus fortes valeurs n'est pas assez satisfaisante

dans certains cas. Ceci peut en effet être contourner en optimisant entre autres la fenêtre de pas temporels.

Modèle	TCN			PROFHMM		
	RMSE	RMSE 10FA	RMSE 10FO	RMSE	RMSE 10FA	RMSE 10FO
Erreurs	0.0531	0.040	0.047	0.0302	0.0076	0.0309
Données Test						

TABLE 6 – Tableau comparatif des modèles TCN et PROFHMM

Conclusion et perspectives

Plusieurs approches ont été avancées pour caractériser la structure phytoplanctonique. En tant qu'indicateur bien connu de la biomasse algale, la CHL-A joue un rôle fondamental dans la recherche sur la surveillance des écosystèmes marins. Aujourd'hui, l'emploi de capteurs satellites permet d'estimer la CHL-A à haute résolution spatiale et temporelle. Ceci est devenu un outil efficace pour suivre l'évolution de la surface de ce pigment, ainsi que l'évolution de l'écosystème marin. Dans le présent document, nous avons introduit les réseaux temporels convolutionnels, qui sont des réseaux de neurones convolutionnels. A partir du travail réalisé, les TCN semblent être capables de reconstruire des profils cachés de paramètres biogéochimiques à partir de données observables à la couche supérieure du profil. Nous avons alors appliqué cette méthode pour la reconstruction des profils verticaux de la chlorophylle-a au BATS, en utilisant les sorties du modèle et les données satellitaires comme observations de la surface de la mer. Notre méthode a été validée par rapport à des données in situ, montrant des résultats prometteurs qui peuvent être d'autant plus optimisés. En effet, nos résultats fournissent des statistiques comparables à celles obtenues dans d'autres travaux, citons la méthode de PROFHMM [1] comme exemple. Ces résultats peuvent être ainsi étendu à d'autres applications. En effet, nous avons l'intention d'étendre la méthode afin de reconstruire l'évolution spatiale du champ de température avec la chlorophylle-a comme variables cibles tout en jouant sur les données d'entrée et leur structure. Il sera aussi intéressant de tester le modèle en ne lui donnant que l'année moyenne des profils verticaux en entrée et regarder à partir de quel moment il converge.

Tous les codes des résultats démontrés dans ce rapport ont été développé sous python et sont accessible depuis <https://github.com/dorrabennour/Projet-Long-ENSIIE>.

Annexe

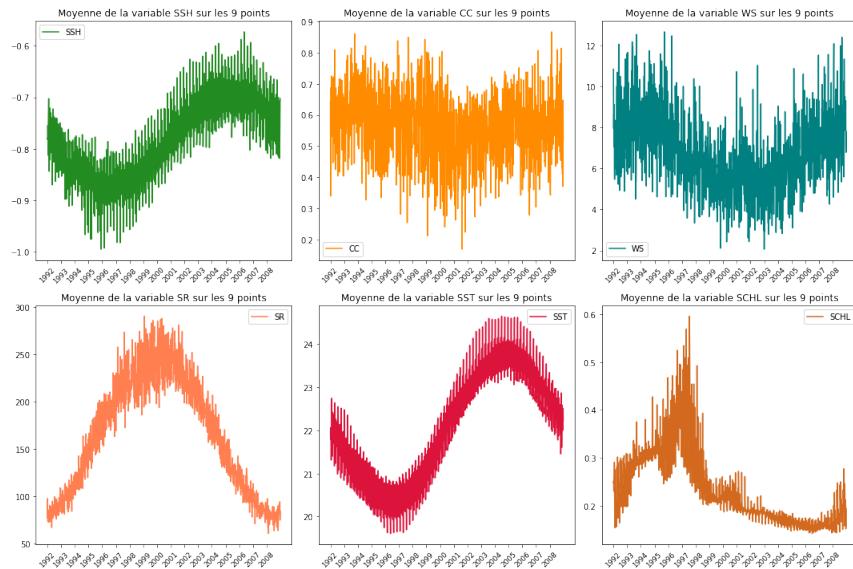


FIGURE 17 – Moyenne des variables de surface sur les 9 points en fonction du temps

Analyse bidimensionnelle

Pour étudier les relations entre les différentes variables, on procède par l'analyse des corrélations entre les variables en moyennant les profondeurs pour les variables chl-a et température en profondeur . La matrice de corrélation est alors montré par la Figure 18 suivante :

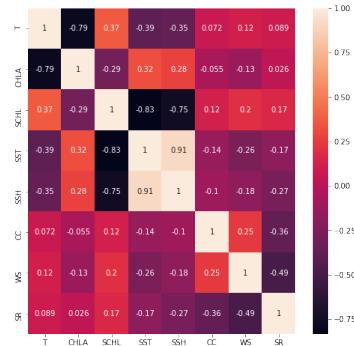


FIGURE 18 – Matrice de corrélation entre les différentes variables

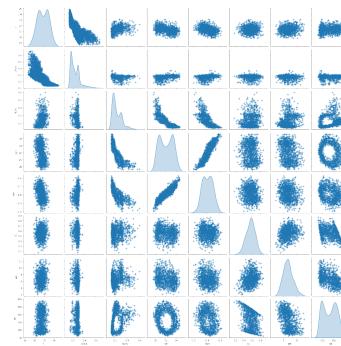


FIGURE 19 – Diagramme de dispersion entre les différentes variables

La matrice de corrélation met en évidence l'absence de corrélation linéaire entre quelques variables, notamment entre les profils verticaux de la chl-a et les autres variables de surface. Il en est de même pour les profils verticaux de la température et les autres variables de surface. Ceci est aussi mis en relief par le diagramme de dispersion (voir Figure 19) où la seule relation linéaire

apparente est celle des variables SST et SSH comme constaté auparavant d'après les courbes des séries temporelles. La matrice de corrélation en considérant chaque profondeur comme variable à part donne les mêmes résultats (Voir Figure 20).

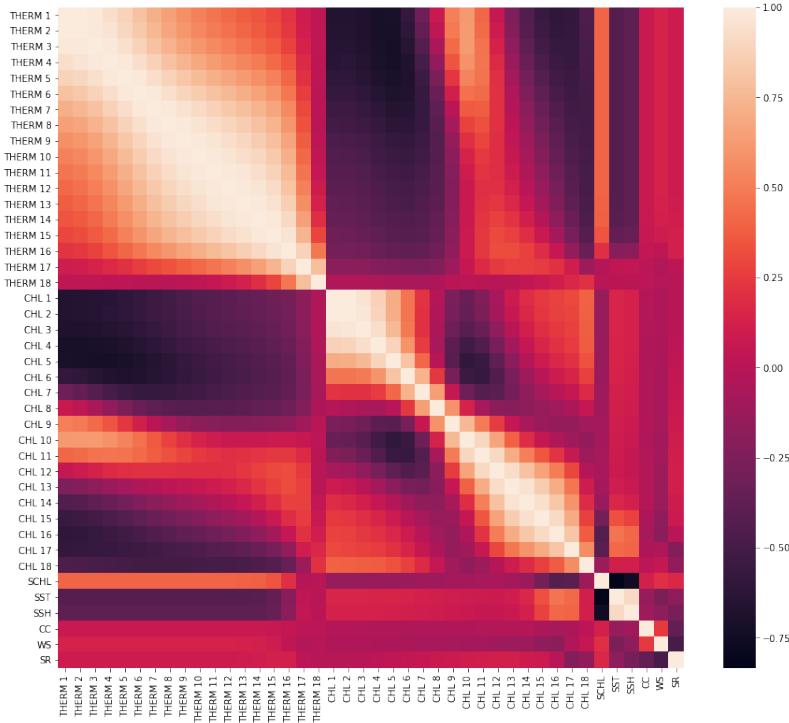


FIGURE 20 – Matrice de corrélation des différentes variables en tenant compte des profondeurs

En effet, seuls les profils verticaux de la température sont corrélés entre eux. Par contre, les profils verticaux des premières profondeurs (première profondeur jusqu'à la huitième profondeur) sont très corrélés entre eux, et très peu corrélés aux autres profondeurs.

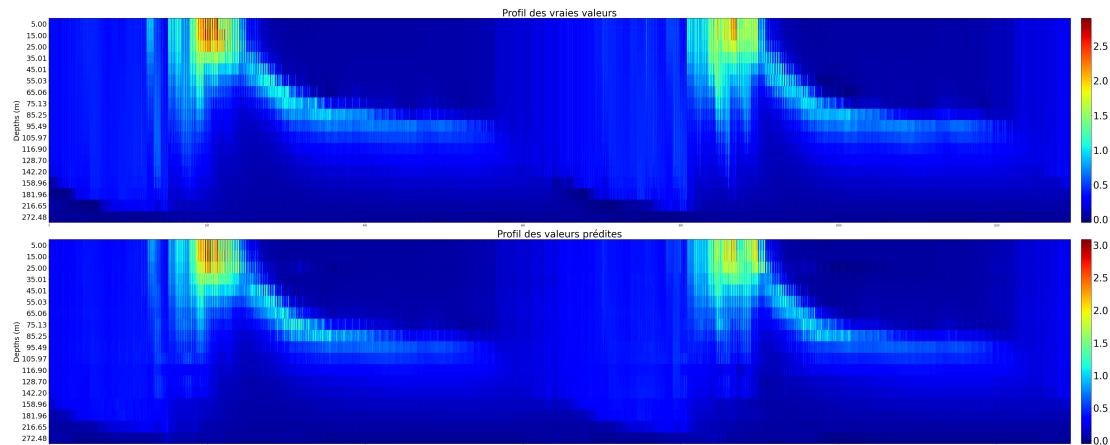


FIGURE 21 – Résultats de la reconstruction des profils verticaux de la CHL-A des années 2007 et 2008 pour une fenêtre de temps égale à 20- La première figure correspond aux profils verticaux des données réels-La deuxième figure correspond aux profils verticaux des données prédites

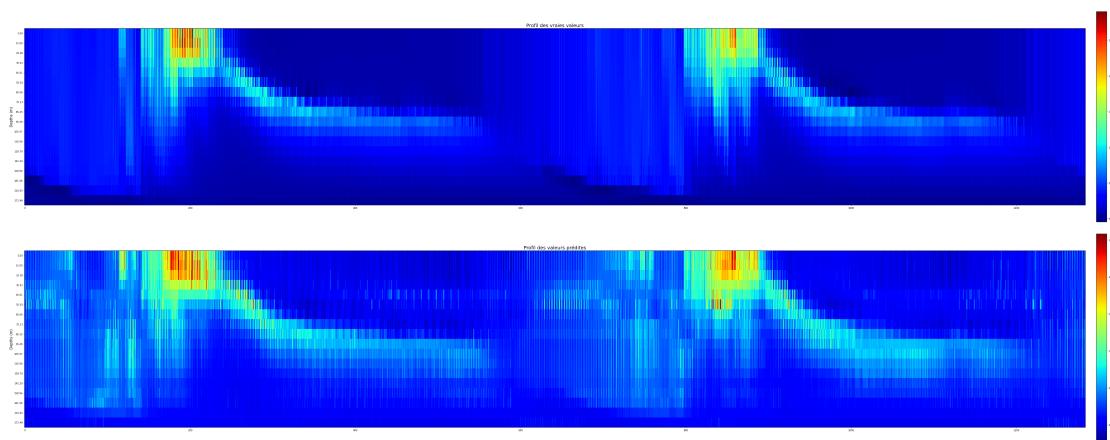


FIGURE 22 – Résultats de la reconstruction des profils verticaux de la CHL-A des années 2007 et 2008 pour une fenêtre de temps égale à 30- La première figure correspond aux profils verticaux des données réels-La deuxième figure correspond aux profils verticaux des données prédites

Références

- [1] Charantonis A.A (2013) *Méthodologie d'inversion de données océaniques de surface pour la reconstitution de profils verticaux en utilisant des chaînes de Markov cachées et des cartes auto-organisatrices* [Thèse de Doctorat,UNIVERSITE PIERRE ET MARIE CURIE]
- [2] JiningYan, Lin Mu, LizheWang, Rajiv Ranjan AlbertY. Zomaya *Temporal Convolutional Networks for the Advance Prediction of ENSO* Scientific Reports | (2020)
- [3] Michela Sammartino , Salvatore Marullo, Rosalia Santoleri Michele Scardi *Modelling the Vertical Distribution of Phytoplankton Biomass in the Mediterranean Sea from Satellite Data : A Neural Network Approach* Remote Sens. 2018, 10, 1666
- [4] Yunxiao Wang , Zheng Liu, Di Hu Mian Zhang *Multivariate Time Series Prediction Based on Optimized Temporal Convolutional Networks with Stacked Auto-encoders* Proceedings of Machine Learning Research 101 :157–172, 2019