THESE DE DOCTORAT DE

L'UNIVERSITE PIERRE ET MARIE CURIE

Spécialité

Télédétection et méthodes statisitques

École Doctorale des Sciences de l'Environnement d'Ile de France

Présentée par

M. Charantonis Anastase Alexandre

Pour obtenir le grade de

DOCTEUR de l'UNIVERSITÉ PIERRE ET MARIE CURIE

Sujet de la thèse :

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.

soutenue le 24/01/2013

devant le jury composé de :

Mme. Sylvie Thiria	Directeur de thèse
Mme. Herlin Isabelle	Rapporteur
M. Blanke Bruno	Rapporteur
M. Gallinari Patrick	Examinateur
M. Jourdin Frédéric	Examinateur
M. Govaert Gérard	Examinateur
M. Moulin Cyril	Examinateur
M. Ghil Michael	Examinateur

RÉSUMÉ

Les observations satellitaires permettent d'estimer les valeurs de différents paramètres biogéochimiques à la surface des océans. D'une manière générale, les paramètres observés sont reliés à des grandeurs géophysiques de l'océan comme : les profils verticaux de concentrations en Chlorophylle-A, les profils de Salinité et de Température… La dimensionnalité de ces données environnementales est très grande, autant dans le cas des données de surface que des profils verticaux. A cause de leur grande dimensionnalité, et de la dynamique complexe qui relie ces données, il est difficile de modéliser leurs relations de façon linéaire. Il s'agit dans cette thèse d'élaborer une méthodologie d'inversion statistique des observations de surface afin de retrouver ces profils verticaux.

La méthodologie développée et que nous avons nommée PROFHMM, fait appel aux Cartes Topologiques Auto-organisatrices pour pouvoir modéliser le problème sous forme d'une chaîne de Markov cachée. PROFHMM utilise les capacités topologiques des Carte Auto-organisatrices non seulement pour déterminer les états et la topologie de la chaîne de Markov cachée générée, mais aussi pour améliorer l'estimation des probabilités qui sont essentielles pour son fonctionnement. Sur les applications géophysiques que nous avons traité dans cette thèse, l'introduction des cartes topologiques auto-organisatrices se révèle un élément essentiel pour assurer les performances obtenues.

Le manuscrit est structuré en quatre parties.

La première partie présente les méthodes statistiques qui forment la base des méthodologies proposées dans cette thèse. Il s'agit des chaînes de Markov Cachées , de l'algorithme de Viterbi et des cartes topologiques auto-organisatrices.

Chaque partie suivante représente un article :

Le premier présente la méthodologie générale de PROFHMM, et traite de l'application qui porte sur la reconstitution temporelle des profils verticaux de Chlorophylle-A. Cette application permet de voir qu'il est possible de synchroniser des données issues de modèles numériques avec des données d'observation satellitaires.

Le second article présente les résultats obtenus par l'application de PROFHMM pour reconstruire les données de la campagne ARAMIS à partir des données altimétriques AVISO et la température de surface fournie par la NOAA. Les performances obtenues prouvent qu'il est

possible de synchroniser une dynamique océanique apprise par des données in-situ et des donnes de surface.

Finalement, dans le troisième article nous présentons une modification à l'algorithme de Viterbi pour prendre en compte, durant la phase de reconstruction de trajectoires, une connaissance à priori sur la qualité des observations. La validité de l'approche est démontrée à partir d'expériences jumelles de reconstruction de séries temporelles de données surface.

Mots-clés : Inversion de données satellite, profils verticaux de Chlorophylle-A, profils verticaux de température, Chaînes de Markov Cachées, Cartes Topologiques Auto-organisatrices, Complétion de séries de données, Algorithme de Viterbi, introduction d'expertise

ABSTRACT

Satellite observations provide us with the values of different biogeochemical parameters at the surface layer of the ocean. These observations are highly correlated with the underlying vertical profiles of different oceanic parameters, such as the Chlorophyll-A concentration, the salinity and temperature of the water column… The sea-surface data and the vertical profiles of the oceanic parameters constitute multi-dimensional vectors. Due to their multi-dimensionality and the high complexity of the dynamics connecting these data sets, their links cannot be modeled linearly. In this thesis we present a methodology to statistically invert sea-surface observations in order to retrieve these vertical profiles.

The developed methodology, named PROFHMM, makes use of Self Organizing Maps in order to render the inversion problem compatible with the Hidden Markov Model formalism. PROFHMM makes full use of the topological aspect of the Self Organizing Maps in order not only to generate the topology and states of the Hidden Markov Model, but also improve the estimation of the probabilities essential to the accuracy of the model. The use of the Self Organizing maps was essential in obtaining the results for the geophysical applications of PROFHMM presented in this manuscript.

The manuscript was structured in three chapters, each consisting of an article. In the first one, the general methodology of PROFHMM is developed, then tested for the retrieval of vertical profiles of Chlorophyll-A by inverting sea-surface observations. This application demonstrated the ability to synchronize sea-surface data with the output data of numerical models.

The second article presents the application of PROFHMM on the inversion of sea-surface data obtained from the AVISO and NOAA projects, in order to retrieve the vertical profiles of temperature over the rail of the ARAMIS mission. The performances obtained demonstrate the ability of PROFHMM to synchronize sea-surface data with in-situ measurements.

Finally, in the third article, we present a modification to the Viterbi Algorithm in order to take into account an à priori knowledge of the quality of the observations when performing reconstructions. The proposed methodology, named PROFHMM_UNC, was applied for the reconstruction of the temporal evolution of sea-surface data, by taking into account the quality of the satellite observations used. The validity of the method was proven by performing a twin experiment on the outputs of a numerical model.

Keywords: Inversion of satellite data, vertical profiles of Chlorophyll-A, vertical profiles of temperature, Hidden Markov Models, Self Organizing Maps, Completion of time-series, Viterbi Algorithm

REMERCIEMENTS

[Tapez votre texte ici]

TABLE DES MATIÈRES

LISTE DES ABRÉVIATIONS ET DES SIGLES

SST	Temperature de surface (Sea-surface Temperature)
CHL-A	Chlorophylle-A
SSH	Elevation du niveau de la mer (Sea-Surface Elevation)
SR	Radiance Solaire à courte ondes (Shortwave Radiataion)
WS	Vitesse du vent (Windspeed)
CC	Couverture Nuageuse (Cloud Cover)
ADT	Topographie absolue dynamique (Absolute Dynamic Topography)
HMM	Chaine de Markov Cachée (Hidden Markov Model)
SOM	Carte Auto-Organisatrice (Self Organizing Map)
PCA	Analyse en composantes principales (Principal Components Analysis)
PROFHMM	Reconstruction de profils par HMM (PROFile reconstruction through HMM)
PROFHMM_UNC	Reconstruction de profils par HMM, en tenant compte des incertitudes (PROFile reconstruction through HMM, taking UNCertainties into account)

INTRODUCTION

Les observations satellitaires permettent d'obtenir des paramètres de surface de l'océan comme la température (SST) ou les spectres de réflectance marine qui permettent d'estimer la chlorophylle-a (CHL-A), qui est une signature de la biomasse marine.

La couverture spatiale et temporelle de ces paramètres est bonne mais reste souvent incomplète par l'effet des nuages. Un autre inconvénient de ces mesures satellitaires est qu'elles ne donnent pas d'information sur la répartition verticale des paramètres mesurés en surface comme la température et la chlorophylle. Or ces profils contiennent des informations essentielles sur la dynamique, la physique et la biologie de l'océan.

Plusieurs méthodes ont été proposées pour pallier cette difficulté : on peut utiliser des bases de données à entrées multiples pour essayer de trouver la situation de la base qui est la plus proche de la situation observée en surface ou assimiler les données de surface dans des modèles tri-dimensionnels de l'océan. La première méthode nécessite l'utilisation de grandes bases de données, opération coûteuse et délicate. La deuxième méthode qui nécessite l'emploi de modèles océaniques de haute résolution et de techniques d'assimilation lourdes, ne peut être réalisée que par une équipe dédiée. Une telle opération est réalisée en France avec la mise en place de programmes d'océanographie opérationnelle comme MERCATOR ou MERCATOR-VERT, ce qui implique une équipe de plusieurs personnes et l'utilisation de calculateurs puissants.

Une alternative moins coûteuse, puisqu'elle ne nécessite pas de faire de l'assimilation de donnée (et l'écriture de l'adjoint d'un modèle océanique si l'on fait de l'assimilation de données), consiste à utiliser des Chaine de Markov Cachées (Hidden Markov Models – HMM) afin de mettre en concordance des séquences de profils verticaux de paramètres biogéochimiques, avec les observations satellitaires. Les chaînes de Markov Cachées sont déterminées par un nombre d'états discrets dits inobservables (ou états cachés), leur topologie (connectivité entre les états cachés, possibilité de passer d'un état à un autre), des observations corrélées aux états cachés, ainsi que les probabilités de passer d'un état caché à un autre et aussi la probabilité d'avoir un état caché étant donnée une observation concordante.

La mise en œuvre des HMM sur des données continues et multi dimensionnelles requiert de déterminer une méthodologie spécifique qui puisse prendre en compte le caractère continu et multidimensionnel aussi bien des profils verticaux que nous voulons estimer, que des observations de surface.

La méthodologie développé durant cette thèse, que nous avons nommée PROFHMM, pour PROFile inversion through HMM, fait appel aux Cartes Topologiques Auto-organisatrices (Self Organizing Maps – SOM) pour pouvoir modéliser le problème sous forme d'une chaîne de Markov cachée.

Les Cartes Topologiques Auto-organisatrices sont des méthodes neuronales automatiques de classification des données. Elles sont bien adaptées pour résoudre des problèmes de discrétisation de données multidimensionnelles. Leur intérêt principal est de pouvoir regrouper des données multidimensionnelles de façon à représenter la variabilité sous-jacente dans l'espace des données, et de les projeter sur une carte topologique de dimension faible (2D) permettant ainsi leur visualisation. Les classes discrètes obtenues par les SOMs sont organisées par selon leur "ressemblance" (définie selon des critères statistiques) sur la carte 2D ; la proximité des classes sur la carte topologique indique une proximité des données appartenant à ces classes dans l'espace des données.

PROFHMM utilise les capacités topologiques des Carte Auto-organisatrices non seulement pour déterminer les états et la topologie de la chaîne de Markov cachée qui va être générée, mais aussi pour améliorer l'estimation des probabilités essentielles pour la construction de l'HMM. Une foi les composantes de l' HMM déterminées, il est facile de reconstruire la séquence de profils verticaux la plus probable, étant donnés une séquence d'observations de surface, en appliquant l'algorithme de Viterbi.

Les séquences obtenues ne sont, bien entendu, pas aussi précises que celles qui seraient déterminées par assimilation des données puisque l'ensemble des processus physiques n'est pas pris en compte. Elles constituent cependant des séquences cohérentes avec la dynamique et l'observation et constituent donc d'excellents candidats pour contraindre l'assimilation des données (détermination de la condition initiale ou « background »)

Le manuscrit est structuré en quatre parties.

La première partie présente les méthodes statistiques qui forment la base des méthodologies proposées dans cette thèse. Il s'agit des chaînes de Markov Cachées, de l'algorithme de Viterbi et des cartes topologiques auto-organisatrices.

Les résultats de ce travail sont ensuite présentés dans trois articles :

Le premier présente la méthodologie générale de PROFHMM, et traite d'une application portant sur la reconstitution temporelle de profils verticaux de Chlorophylle-A. Cette application permet de voir qu'il est possible de synchroniser des données issues de modèles numériques avec des données d'observation satellitaires.

Le second article présente les résultats obtenus en utilsant PROFHMM pour reconstruire les profils de la campagne ARAMIS à partir des données altimétriques AVISO et la température de surface fournie par des radiomètres satellitaires et traitées par la NOAA. Les performances obtenues prouvent qu'il est possible de synchroniser une dynamique océanique apprise sur des données in-situ et des données de surface.

Finalement, dans le troisième article nous présentons une modification à l'algorithme de Viterbi pour prendre en compte, durant la phase de reconstruction de trajectoires, une connaissance à priori sur la qualité des observations. La validité de l'approche est démontrée à partir d'expériences jumelles de reconstruction de séries temporelles de données surface.

CHAPITRE 1 : RAPPEL DES METHODES DES CHAINES DE MARKOV CACHEES ET CARTES TOPOLOGIQUES AUTO-ORGANISATRICES

1.1 Introduction :

Dans ce chapitre nous introduisons les méthodologies et les algorithmes qui forment la base des méthodologies proposées dans cette thèse. Il s'agit des chaînes de Markov Cachées, de l'algorithme de Viterbi et des cartes topologiques auto-organisatrices. Les algorithmes fondamentaux qui forment les socles sur lesquels reposent les méthodologies proposées dans ce manuscrit, sont présentés avec leur formulation mathématique complète. Elle sont reprises d'une manière plus condensée dans les articles des chapitres suivants, qui sont dédiés aux validations. En particulier on présente en détail les performances obtenues sur des problèmes réels dans le domaine de l'environnement.

1.2 Chaînes de Markov Cachées :

1.2.2 Chaîne de Markov

On considère, une suite de variables aléatoires discrète ayant un même ensemble fini d'états $\{X_1, X_2, \ldots, X_N\}$. Le temps étant supposé régulièrement discrétisé sous la forme d'une séquence temporelle $t_0, t_1, t_2 ; \ldots t_n , \ldots .$, on note par X^n la variable aléatoire qui se réalise au nième temps t_n. Par exemple si l'on considère une suite de lancers de dés pour laquelle on a observé les valeurs 1, 6, 2, 5 on notera X^0, X^1, X^2, X^3.

On suppose par la suite que la suite vérifie la propriété de Markov du premier ordre :

$$P(X^t/X^0, X^1, \ldots ; X^{t-1}) = P(X^t/X^{t-1}) \qquad (1)$$

Ainsi, lors de son évolution, l'état futur de la série ne dépend pas de son passé, mais uniquement de son état présent. Selon cette propriété, la probabilité du second membre de la relation (1) est dite probabilité de transition au temps t. Si en plus on fait l'hypothèse que la

probabilité de transition est constante et indépendante du temps t, on dit alors qu'elle vérifie en plus la propriété d'homogénéité. Sous cette hypothèse on pose alors :

$$tr_{i,j} = P(X^t = X_j / X^{t-1} = X_i),$$

qui correspond à la probabilité de transition de l'état X_i vers l'état X_j. Les probabilités de transitions vérifient

$$\sum_{J=1}^{N} tr_{i,j} = 1$$

Une suite de variable aléatoire ayant un ensemble d'état fini sera dite chaîne de Markov, si elle vérifie les deux propriétés précédentes : propriété de Markov d'ordre 1 et propriété de l'homogénéité des probabilités de transition.

Une Chaîne de Markov est déterminé alors par la connaissance des éléments suivants :

- L'ensemble de ses états $\{X_1, \ldots, X_N\}$, où N est le nombre d'états dans lesquels la chaîne peut se trouver.
- La matrice **Tr** carrée d'ordre N de l'ensemble de ses probabilités de transitions **Tr**=[$tr_{i,j}$]. Cette matrice décrit la dynamique de l'évolution temporelle de la chaine.
- Le vecteur de probabilité initiale $\Pi = [\pi_i]$ où $\pi_i = P(X^0 = X_i)$. Ce vecteur décrit l'initialisation de la chaîne de Markov au temps t_0.

Une chaîne de Markov peut être représentée par un graphe. L'ensemble de ses sommets correspondent aux N états de la chaîne et un arc de X_i vers X_j correspond à une probabilité de transition $tr_{i,j}$ non nulle. La figure 1.2.1 (a) montre une représentation d'une Chaîne de Markov à 3 états interconnectés.

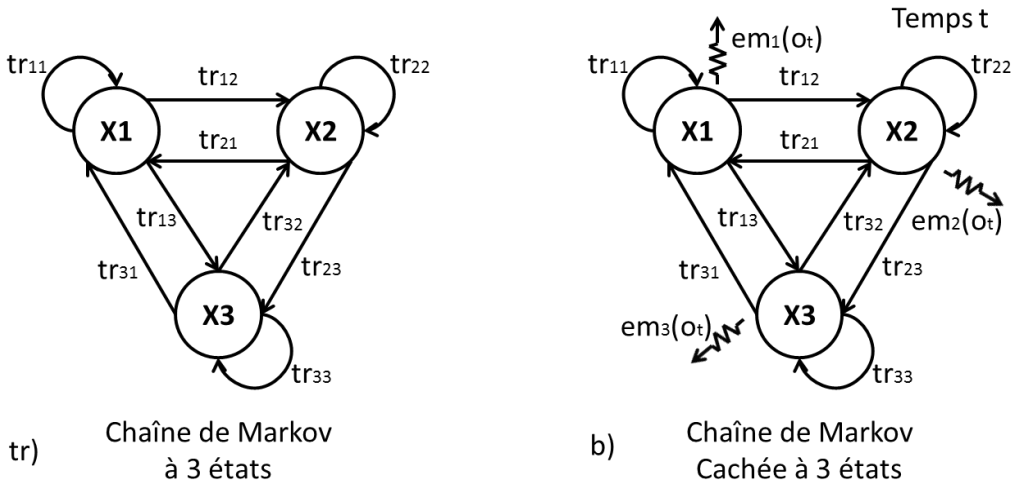


Figure 1.2.1 a) *Schématisation d'une Chaîne de Markov b) Schématisation d'une Chaîne de Markov Cachée*

Le plus souvent s'il est possible de trouver les états et la dynamique qui régit une chaîne de Markov, il est difficile de la faire fonctionner dans la réalité pour effectuer des prévisions du système dynamique qu'elle modélise. En effet, les valeurs qui pourraient être observées ne sont pas suffisantes pour connaître l'état exact dans lequel se trouve le système. Dans ce cas nous considérons que l'état de la chaîne est caché et que l'observation est une aide, qui contient de l'aléa, que nous pouvons utiliser pour estimer l'état dans lequel se trouve le système.

1.2.2 Chaîne de Markov Cachée

Une Chaîne de Markov cachée est constituée d'une chaîne de Markov dont les états ne sont pas observables, mais émettent cependant des signaux observables. On suppose qu'à chaque état de la chaîne de Markov est associée une loi de probabilité qui décrit la manière dont les observations sont émises par cet état. Nous ne pouvons donc pas savoir dans quel état se trouve la chaîne à chaque instant, mais nous avons une information de type « proxy ».

L'ensemble des observations peut être continu ou discret : dans le premier cas la loi d'émission d'un état donné est défini par sa fonction densité, par contre dans le second elle sera définie par le vecteur correspondant à la répartition des probabilités des observations.

Dans cette thèse nous limiterons notre étude aux chaînes de Markov avec un nombre fini d'émissions possibles. Les probabilités d'émission des états de la chaîne de Markov cachée sont contenues dans une matrice d'émission, notée **Em**. Chaque ligne de cette matrice correspond à un état donné et le contenu de la ligne correspond à la répartition des probabilités d'émission, des observations, par cet état. Ainsi, l'élément em_{ij} de cette matrice correspond à la probabilité d'émission de la j-ème observation par le i-ème état caché .

D'une manière synthétique, une Chaîne de Markov cachée est déterminée par :

- La chaîne de Markov cachée : ensemble des états, le vecteur des probabilités de l'état initiale Π^0 , la matrice de transition des états **Tr**. Les états de cette chaîne ne sont pas directement observables, c'est pourquoi ils seront dits cachés.
- L'ensemble des observations $\{obs_1, obs_2, \ldots, obs_M\}$ où M représente le nombre des observations possibles, qui dans le cadre de cette thèse sont supposées finies, ainsi que la matrice **Em** de leur probabilité d'émission par les états cachés.

Les probabilités de transition correspondent à une dynamique cachée du système modélisé, et les probabilités d'émission permettent de contraindre l'évolution des états du système par les observations. Les probabilités initiales correspondent à une connaissance à priori de la fréquence de chaque état caché, qui est sensée pénaliser l'initialisation de l'algorithme par des états cachés peu probables.

Une représentation d'une Chaîne de Markov Cachée est montrée dans la figure 1.2.1 (b). On remarque l'ajout des éléments correspondant aux probabilités d'émission de chaque état.

Un élément essentiel pour la bonne détermination d'une chaîne de Markov cachée (Hidden Markov Model - HMM) est l'estimation de ses trois matrices de probabilités. Ceci peut être effectué par des compteurs, puis raffiné par l'algorithme de Baum-Welch ou par d'autres techniques. Dans l'article du chapitre 2, nous présentons l'utilisation des cartes topologiques comme instrument d'amélioration des probabilités d'une HMM.

Le principe d'une HMM est de retrouver à partir d'une séquence temporelle d'observations, la séquence temporelle d'états cachés qui lui correspond. Ceci entraîne que les résultats fournis par la chaîne ne se focalisent pas d'une manière précise sur un état, mais produit la séquence la plus probable dans son ensemble.

Ainsi, si nous supposons observée une séquence de mesures définies par la séquence de leur indice $O_T=o_1o_2\ldots o_T$ où o_k représente l'indice de la kième observation, la détermination de la

séquence des indices des états cachés $I_T=i_1i_2\ldots i_T$ qui lui est sous-jacente, est déterminée en maximisant, relativement à I_T, la probabilité d'avoir I_T sachant qu'on a observé la séquence O_T. Ceci revient à maximiser :

$$P(I_T/O_T) = P(I_T \cup O_T) \Big/ P(O_T)$$

Le dénominateur étant indépendant de I_T, ceci revient à chercher une séquence I_T qui maximise le numérateur :

$$P\left(I_T \cup 0_T\right) = \pi_{i_0} tr_{i_0,i_1} em_{i_1,o_1} tr_{i_1,i_2} em_{i_2,o_1} \ldots tr_{i_{T-1},i_T} em_{i_T,o_T} = \pi_{i_0} \prod_{k=1}^{T} tr_{i_{k-1},i_k} em_{i_k,o_k}$$

(2)

Cette relation provient de la propriété des probabilités de transition de la chaîne de Markov et du fait que la réalisation de l'observation o_k ne dépend que de son état caché i_k. La résolution de ce problème se fait par l'algorithme dit de Viterbi.

1.3 Algorithme de Viterbi :

L'algorithme de Viterbi consiste à trouver un chemin de longueur maximal dans un graphe.

Pour une séquence d'observations, définie par la suite d'indice $O_T=o_1o_2\ldots o_T$, nous considérons le graphe valué suivant :

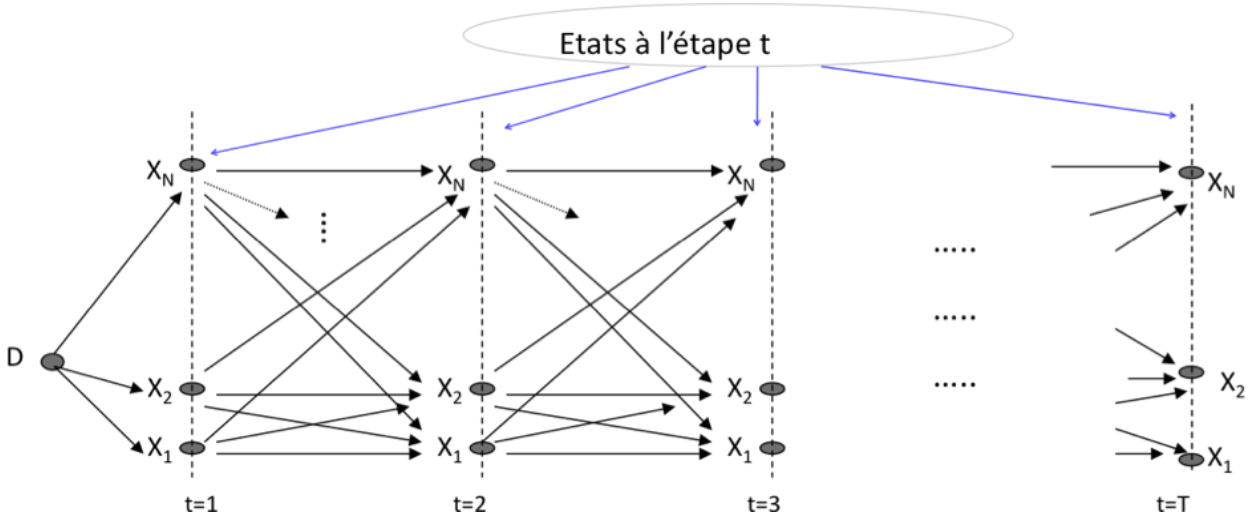


Figure 1.3.1 Graphe valué d'une chaîne de Markov.

Les sommets de ce graphe sont organisés en T niveaux, où chaque niveau t est formé de N sommets correspondant chacun à un état caché de la chaîne de Markov.

Chaque arc de type D→X_l est valué par $\pi_l em_{l,o_t}$ et chaque arc du type X_l → X_k entre les deux niveaux t-1 et t (où t>1) est value par $tr_{lk} em_{k,o_t}$. La maximisation de l'expression (2), relativement à I_t, revient à chercher le plus long chemin du sommet D à l'un des sommets du niveau t, sachant que la longueur d'un chemin est égale au produit des valuations de ses arcs. La séquence I_T des indices des états du chemin optimal correspond à la solution optimale de (2).

La résolution directe par énumération suppose le parcourt de N^T chemins possibles, ce qui correspond à $2TN^T$calcules. Ainsi pour une chaîne de Markov cachée à N=5, retrouver, par énumération directe, la séquence la plus probable correspondante à une séquence de 100 états observables, implique ≈ $200*5^{100}$ calculs. La théorie des HMM utilise pour déterminer la meilleure reconstitution un algorithme de type programmation dynamique pour retrouver la séquence d'états cachés la plus probable.

L'algorithme de Viterbi est un algorithme récursif qui permet de trouver à partir d'une suite d'observations, une solution optimale au problème d'estimation de la suite d'états cachés.

L'algorithme de Viterbi (figure 1.3.2), utilise les deux notations $\delta_t(i)$ et $\Psi_t(j)$.

L'Algorithme de Viterbi

Séquence d'observations :

$O_{k_t}^t$, avec $t = \{1, \dots, T\}$

Initialization:

For $1 \leq i \leq N$

$$\delta_1(i) = em_{k_1,i} * \Pi_i$$

$$\Psi_1(i) = 0$$

Calcul Itératif:

For $2 \leq t \leq T$

For $1 \leq j \leq N$

$$\delta_t(j) = \left(\max_{1 \leq i \leq N} \left[\delta_{t-1}(i) * tr_{i,j}\right]\right) * em_{k_t,j}$$

$$\Psi_t(j) = arg \max_{1 \leq i \leq N} \left[\delta_{t-1}(i) * tr_{i,j}\right]$$

Terminaison:

$$P = \max_{1 \leq i \leq N} [\delta_T(i)]$$

$$q_T = arg \max_{1 \leq i \leq N} [\delta_T(i)]$$

Backtracking:

For t=T to 2

$$q_{t-1} = \Psi_t(q_t)$$

Figure 1.3.2 Algorithme de Viterbi

- La notation $\delta_t(i)$ correspond à la probabilité maximale d'avoir une séquence d'états cachés qui culmine au temps t à l'état X_i. Ainsi, si l'on désigne par $I_t(i)=i_1 i_2 \dots i_t$ avec $i_t=i$ et par O_t, la séquence des t premières observations de O_T ; alors on a $\delta_t(i) = \underset{I_t(i)}{MAX}\ P(I_t(i), O_t)$. On a, d'une manière évidente, $\delta_1(i) = \pi_i em(i, o_1)$ et le principe de la programmation dynamique donne la formule de récurrence suivante :

$$\delta_t(j) = \left(\max_{2 \leq i \leq N} \left[\delta_{t-1}(i) * tr_{i,j}\right]\right) em_{j,o_t}$$

- La seconde notation est définie par $\Psi_t(j) = \underset{1<i\leq T}{\operatorname{argmax}}\left[\delta_{t-1}(i) * tr_{i,j}\right]$, elle enregistre, au pas de temps t, l'index i de l'état caché du temps t-1 l'ayant amené à calculée, par la formule précédente, la probabilité maximale $\delta_t(j)$.

La figure1.3.2 présente aussi la méthode de rétro propagation (backtracking) qui permet de reconstituer la séquence I_T des états cachés optimaux.

Pour le même nombre d'états N=5 et une séquence de 100 états observables, l'algorithme de Viterbi requiert approximativement $3T*N^2$ calculs, donc 7500 calculs.

Il existe des variantes de l'algorithme de Viterbi tel que le Lazy Vitterbi Algorithme et le Soft Output Viterbi Algorithme mais, étant donné notre choix de modélisation, l'Algorithme de Viterbi de base est suffisant pour remplir la tache exigée.

1.4 Cartes Topologiques Auto-Organisatrices :

L'algorithme des cartes topologiques proposé par Kohonen est un procédé d'auto-organisation qui cherche à réduire le nombre de données d'une très grande base en un nombre restreint de données significatives, et fournit conjointement une partition de la base de données initiale en sous ensemble cohérents. Les données peuvent être des vecteurs de grande dimension.

Les cartes de Kohonen sont une classe particulière de réseaux de neurones non supervisés. Elles sont formées de neurones répartis en deux couches. La première couche reçoit les données (vecteur multi-dimensionnels). La seconde couche est constituée d'un treillis (bi-dimensionnel dans le cas présent) de neurones interconnectés selon une loi de proximité. Deux neurones proches sur le treillis ont des caractéristiques proches. Dans la phase d'apprentissage chaque neurone du treillis capture via la première couche des données qui sont proches de ses caractéristiques. Les caractéristiques des neurones du treillis s'organisent de façon à représenter des sous-ensembles bien cohérents de l'ensemble d'apprentissage. A la fin de la phase d'apprentissage chaque sous-ensemble (chaque neurone) est caractérisé par un vecteur référent qui résume selon une certaine loi (la moyenne par exemple) les caractéristiques des vecteurs données qu'il a capturées. On a ainsi réalisé une partition cohérente de l'ensemble d'apprentissage caractérisée par les vecteurs référents associés à chaque neurone.

D'une manière générale, en phase de fonctionnement, les cartes topologiques vont projeter les données initiales sur les vecteurs référents. Grâce au procédé d'auto-organisation la topologie

qui lie les données initiales est conservée au niveau des réponses proposées par le réseau. La localisation des neurones actifs reproduit les liens existants au niveau des données initiales.

Décrivons d'une manière plus détaillée carte auto-organisatrice utilisée dans la présente étude; elle est constituée de deux couches de neurones (figure 1.4.1):

- La couche d'entrée sert uniquement à la présentation des formes à classer. Les états de tous ses neurones sont forcés aux valeurs des signaux d'entrées.

- La couche d'adaptation est formée du treillis de neurones évoqué précédemment. Le choix de la géométrie du réseau employé est fait à priori. Les neurones utilisés à ce niveau sont de simples neurones linéaires, chacun d'entre eux étant connecté à tous les éléments de la couche d'entrée. De manière à permettre le processus d'auto-organisation, les poids qui lient les deux couches du réseau sont adaptatifs.

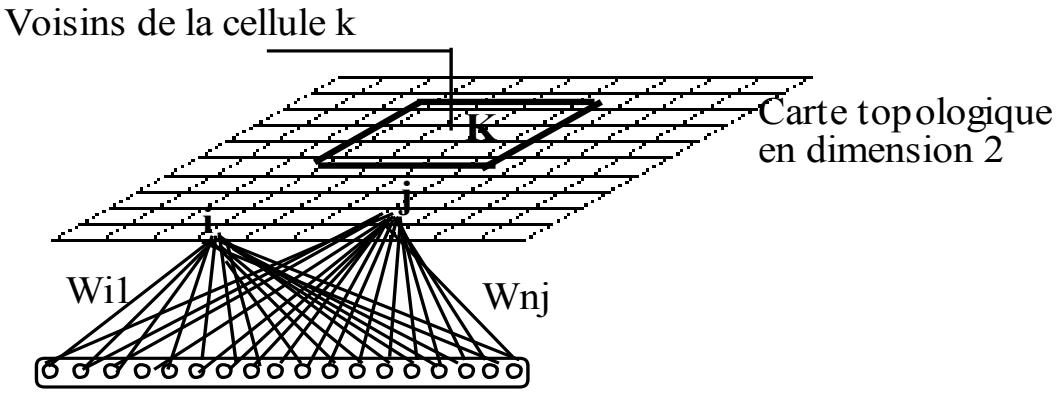


Figure 1.4.1 : Carte topologique :treillis de neurones muni d'un système de voisinage. Les neurones sont entièrement connectée à la couche d'entrée.

Les neurones de la carte calculent leur état, en parallèle, à partir des mêmes informations fournies par la forme présentée en entrée.

La principale caractéristique du processus d'auto-organisation est de permettre une adaptation des poids uniquement sur la région de la carte la plus "active". L'algorithme le plus simple détermine ce centre d'activité, comme étant le voisinage de la carte associé au neurone dont l'état est le plus grand. C'est l'utilisation de ce voisinage qui introduit les contraintes topologiques dans la représentation finale: ceci modélise de façon simplifiée un couplage latéral entre le neurone sélectionné et ses voisins dans la structure du réseau.

De cette façon, en fin d'apprentissage, les poids de chaque neurone vont converger vers des valeurs telles qu'un neurone ne sera plus actif que pour un sous ensemble précis d'éléments de la base d'apprentissage. Un neurone représente alors un exemple ainsi que la classe des données proches de cet exemple.

Rechercher le neurone de plus grande réponse revient à déterminer le vecteur de poids dont le produit scalaire (ou la corrélation non normalisée) avec l'exemple présenté est le plus grand.

Les vecteurs poids ayant la même dimension que celle des formes d'entrée, ils représentent des points à l'intérieur du même espace. Un second critère de sélection du neurone le plus actif peut être de chercher celui dont le vecteur de poids est le plus proche au sens de la distance euclidienne de la forme présentée. C'est ce critère qui est à l'heure actuelle utilisé dans l'algorithme des cartes topologiques par Kohonen. Son avantage est de permettre une formulation mathématique plus simple du problème.

Certaines variantes de l'algorithme cherchent à normaliser les vecteurs de poids à une longueur constante, les deux critères évoqués plus hauts sont alors identiques.

Différents procédés sont proposés pour l'adaptation des poids, nous ne présentons que le plus simple d'entre eux et renvoyons à [Kohonen 94] pour les variantes.

La description précise de l'algorithme nécessite donc de définir la notion de voisinage sur la carte. Le voisinage V_c de rayon d du neurone c est composé de tous les neurones du réseau qui se situent à l'intérieur d'un disque de rayon δ. La valeur choisie pour δ permet de considérer des voisinages de forme différents et de prendre en compte un nombre variable de neurones. Cette valeur va varier pendant le déroulement de l'algorithme et permettre d'améliorer la convergence.

La modification appliquée dans le voisinage choisi revient à rapprocher les vecteurs poids sélectionnés de l'exemple présenté. En reconnaissance des formes, les méthodes de classification automatique utilisent souvent des techniques similaires.

Si l'on note c le neurone sélectionné, V_c son voisinage et W_r le vecteur de poids d'un neurone r, les modifications qui vont avoir lieu après présentation de la forme z au temps t vont être:

$$\begin{cases} W_r^t = W_r^{t-1} + a(t)(W_r^{t-1} - z), \ si \ r \in V_c \\ W_r^t = W_r^{t-1}, \ sinon \end{cases} \quad (3)$$

Le procédé d'adaptation nécessite donc l'introduction d'un nouveau paramètre, le pas d'apprentissage, qui varie en fonction du temps: $a(t)$.

Nous présentons la version de l'algorithme (figure 1.4.2) pour laquelle le neurone qui réagit le plus fort est celui dont le vecteur de poids est le plus proche au sens de la distance euclidienne de la forme présentée. Il utilise deux paramètres qui sont la taille des voisinages et le pas d'apprentissage. L'initialisation des poids et la présentation des éléments de l'ensemble d'apprentissage sont aléatoires. Le critère d'arrêt peut être un nombre d'itérations fixé à priori.

Algorithme de Kohonen:

1. A l'itération t, présenter $z = (z_1, z_2, \cdots, z_n)$

2. Déterminer le vecteur de poids le plus proche W_c^t de z tel que:

$$\|z - W_c^t\| = \min_r \|z - W_r^t\|$$

où la minimisation est prise sur l'ensemble des vecteurs de poids associés à chaque neurone du réseau.

3. Déterminer la taille du voisinage de modification V_c et le pas d'apprentissage a(t).

4. Modifier les poids des neurones r selon le procédé décrit par la relation (3).

5. Itérer pour t= t+1

Figure 1.4.2 Algorithme de Kohonen.

La fonction qui contrôle le pas d'apprentissage $a(t)$ peut prendre des formes différentes, mais elle est décroissante de façon monotone. Si elle arrive à de très faibles valeurs, l'algorithme approxime le fonctionnement des algorithmes de « K-moyennes ».

Le rôle de la fonction (3) est de modifier, à chaque phase de modification des poids des neurones, la valeur des vecteurs voisins au vecteur sélectionné. Cette modification devient progressivement mineure, mais initialement force les neurones à s'adapter à la forme des données, et de les ordonner pour que, à la fin de la phase d'apprentissage, les référents proches dans l'espace des données soient représentés par des neurones proches sur la carte topologique.

Dans le cas de grandes bases de données multi-factorielles, on doit utiliser des cartes SOM avec un nombre relativement grand de neurones (de l'ordre de qq centaines de neurones) de façon à bien cerner la complexité des phénomènes représentés. Il est souvent difficile pour le scientifique d'analyser la signification scientifique d'un tel nombre de classes. Une technique

souvent employée est de réunir les classes ayant des propriétés voisine en un nombre restreint de groupes. Ces groupes pourront alors être analysés plus facilement en terme de processus physique, biogéochimiques. Ceci peut être réalisé à l'aide d'algorithmes dédiés comme la HAC (Hierarchical Ascendant Classification).

De nombreuses applications utilisant les cartes SOM ont été réalisées au LOCEAN au sein de l'équipe MMSA ; on peut citer la décomposition des situations météorologiques en type de temps où les situations météorologiques ont des caractéristiques voisines, ce qui permet des analyses fines de phénomènes complexes comme les occurrence de pluie [Travaux d'A. Gueye et al, (2009)] ou les relations reliant l'épaisseur optique atmosphérique mesurée par satellite et les aérosols [PM10 , Travaux de H. Yahi et al, (2011)], la détermination des différents types d'aerosols [Travaux de A. Niang et al (2003, 2006), D. Diouf (2012), J. Brajard, (2012)]. On doit aussi mentionner les travaux de M. Jouini et al, (2012) qui ont consisté à reconstituer sous les nuages les structures de Chlorophylle observés par satellite. M. Jouini a pour cela utilisée une méthode s'apparentant aux méthodes des "analogues" utilisées en météorologie, qui consiste à identifier à partir d'une très grande base de donnée, la structure la plus probable étant donné un certain nombre de contraintes statistiques. Cette structure est extraite d'une carte SOM qui rassemble les structures les significatives de la base de données. Tous ces travaux ont donné lieu à des publications dans des journaux internationaux (voir bibliographie).

CHAPITRE 2 : ETUDE DE L'EVOLUTION TEMPORELLE DES PROFILS VERTICAUX DE CHLOROPHYLLE-A PAR INVERSION DE DONNEES SATELLITE

2.1 Introduction :

L'activité bio-géochimique de l'océan joue un rôle important dans la régulation des flux de carbone en absorbant le CO2 atmosphérique. Pourtant, l'impact du changement climatique sur la distribution des organismes phytoplanctoniques, et, réciproquement, le rôle du phytoplancton dans le climat, à travers le cycle du carbone et la modification du forçage radiatif, restent à déterminer de façon précise.

La concentration océanique en Chlorophylle-a, pigment caractéristique des organismes phtyoplanctoniques, est généralement considéré comme bon indicateur de la production primaire. Un des avantages majeurs de ce pigment est que sa concentration en surface peut être estimée à partir d'images satellite (Seawiffs, Modis). Dans ce chapitre nous présentons une méthodologie, nommée PROFHMM pour « PROFile reconstruction through HMM », capable de déterminer des liens empiriques entre la concentration de Chlorophylle-A en surface et ses profils verticaux.

Le travail est présenté d'une manière synthétique sous forme d'article, soumis à JAOT. Les analyses préalables qui ont permis l'élaboration de la méthode sont présentés en annexe de l'article. Dans l'annexe 1 de la thèse on trouvera un article plus ancien, présenté au congres NCTA 2011, qui donne plus d'information sur les classifications par cartes topologiques. En annexe 2 on trouvera, en tant qu'exemple de la généralisation possible de PROFHMM, un des de deux rapport de projets longs, réalisés par des étudient de master. Il s'agit d'une application similaire de PROFHMM sur une zone géographiques différente (DYFAMED en mer méditerranée) qui valide la généralité de l'approche pour la restitution des profils de Chlorophylle-A.

2.2 ARTICLE 1: Retrieving the vertical profiles of Chlorophyll-A from satellite observations, by using hidden Markov models and self-organizing maps.

Résumé : Nous présentons une méthodologie statistique, nommée PROFHMM, pour estimer l'évolution des profils verticaux de paramètres biogéochimiques océanique à partir de données de surface. PROFHMM utilise des Cartes Topologiques Auto-Organisatrices pour déterminer les états et la topologie d'une chaîne de Markov Cachée. En plus de cela, l'aspect topologique des Cartes Topologiques Auto-Organisatrices est utilisé pour améliorer l'estimation des probabilités de la Chaîne de Markov cachée. Une foi les principes de PROFHMM expliqués, nous présentons les résultats obtenus sur une étude pour la détermination de l'évolution des profils verticaux de Chlorophyll-A à partir de l'inversion de données de surface.

Retrieving the vertical profiles of chlorophyll-a from satellite observations, by using hidden Markov models and self-organizing maps.

A A CHARANTONIS[1], F BADRAN[2], S THIRIA[1]
[1] Laboratoire d'Océanographie et du Climat – Expérimentation et Approches Numériques, Université Pierre et Marie Curie, Tour 45, 5ème étage 4, place Jussieu, 75005 Paris, France
[2] Laboratoire CEDRIC, Conservatoire National des Arts et Métiers, 292, rue Saint Martin, 75003 Paris, France

E-mail: anastase-alexandre.charantonis@locean-ipsl.upmc.fr

Abstract. We present a statistical method, denoted PROFHMM, to infer the evolution of the vertical profiles of oceanic biogeophysical parameters from sea-surface data. This method makes use of discrete hidden Markov models whose states are defined through self-organizing maps. The self-organizing maps are used to provide the states of the hidden Markov model, as well as improve its parameters. After introducing the general principles of PROFHMM, we present the results obtained in a case study in which the evolution of the vertical profiles of chlorophyll-*a* was inverted from sea-surface data.

1. Introduction

The current density of satellite observations has allowed a quasi-continuous observation of the global ocean surface. The two-dimensional images provided by this coverage contain information on physical or biogeophysical parameters but not on their vertical profiles [1, 2, 3]. Inverting the sea-surface data remotely sensed by satellite to obtain the vertical profiles of biogeophysical parameters, however, requires a numerical modeling of their relations. Such models are, however, often faced with problems of non-linearity, complexity and incomplete knowledge of the mechanisms that govern these profiles.
In the present paper we attempt to infer the vertical profiles of chlorophyll-*a* from observed sea-surface satellite images. The biogeochemical activity of the oceans and the carbon cycle are two parts of a complex feedback system. A change in climate and an increase in the amount of available carbon can affect the oceanic primary production. In return, a change in the biogeochemical activity affects the climate, by modifying the albedo and carbon fixation rates, as well as the atmospheric and oceanic carbon concentrations. It is therefore important to be able to determine the oceanic primary production, of which chlorophyll-*a* is a proxy.

In recent years, many algorithms have been developed to infer the chlorophyll-*a* concentration in ocean surface layers through satellite imaging [4]. It has also been proved that the vertical chlorophyll-*a* distribution is related to various types of sea-surface data [5]. The cost of determining the vertical

distribution of chlorophyll-*a* by in situ measurements is prohibitively high, and this can explain the lack of complete databases of such measurements. There are, however, large databases of extrapolated vertical profiles of chlorophyll-*a* calculated by biogeochemical models such as the MERCATOR-VERT and NEMO-PISCES models [6, 7]. It is generally accepted that these models are able to reproduce the dynamic processes that govern the evolution of the vertical profiles of chlorophyll-*a*. There is a large number of such databases, presenting time-series spanning over decades. Sea-surface measurements, on the other hand, are almost continuously accessible from satellite imagery.

These facts have lead us to investigate the possibility of inferring the evolution of the vertical profiles of chlorophyll-*a* solely from sea-surface data. Here below we consider that the chlorophyll-*a* vertical profiles correspond to a set of unobservable, so-called 'hidden states', and that the multidimensional sea-surface data are "emitted" from these. Formulated this way, this problem is similar conceptually to the statistical modeling method known as hidden Markov models (HMM) [8]. A HMM is fully defined by its states, its topology and its related probabilities.

We present a method that tackles the problem of inverting the surface data to infer the vertical profiles of chlorophyll-*a* and allows us to formalize it as a classical HMM. Due to the high dimensions of the vectors involved in the reconstruction of the chlorophyll-*a* profiles, the size of the database used to estimate the probabilities required to define the HMM is never sufficient. In order to get a robust determination of the parameters of the HMM we applied a self-organizing map (SOM) [9]. SOMs constitute a method that produces a topologically ordered classification of data sets. They have been used in other cases to generate the topology of an HMM [10]. In PROFHMM, that topology is further used to improve the quality of the estimated HMM probabilities, given the otherwise insufficient available data.

The method we have developed, which is a combination of HMM and SOM, was named PROFHMM for PROFile reconstruction through HMM. It inverts the evolution of sea-surface parameters to retrieve the evolution of the hidden vertical distribution of chlorophyll-*a* in the oceans.

PROFHMM is a very cost-efficient inversion method, since, once the training of the method is over, the computations needed are minimal and could be run on most personal computers.

After introducing the general principles of the method, we present the results obtained in a case study. The vertical profiles of chlorophyll-*a* were inverted from sea-surface data at the site of the Bermuda Atlantic Time Series (BATS) (32°N –64°W) of the JGOFS (Joint Global Ocean Flux Study) campaign [11], first using simulated data for both the vertical distribution profiles and the observation vectors, and then by using simulated data in conjunction with MODIS satellite data for the observation vectors [12].

2. Data

The study of the oceanic primary production is linked with phytoplankton mass and therefore with the chlorophyll-*a* distribution. One cannot determine the vertical distribution of chlorophyll-*a* without first understanding the parameters that influence the development of phytoplankton. It is generally accepted [13] that phytoplankton growth mainly depends on five parameters: available shortwave radiation, available nutrients, herbivores and biology, water temperature, water turbidity.

These parameters cannot be easily monitored through a direct approach. Satellite imaging, however, can give us proxy information, which can be used in an empirical approach to determining the vertical profiles

of chlorophyll-*a*. Specifically in this study, after conducting a preliminary principal-component analysis [14], we used:

- sea-surface chlorophyll-*a* concentration (SCHL)
- sea-surface temperature (SST)
- sea-surface height (SSH)
- downwelling shortwave radiation (SR)
- wind-speed (WS)

to infer the chlorophyll-*a* vertical profiles.

There is currently a lack of in situ measurements of chlorophyll-*a* with consistent revisit rates. The JGOFS Bermuda Atlantic Time Series (BATS), for example, makes in situ measurements every month for three consecutive days, which does not provide a good temporal sampling for inferring the dynamic processes that govern the development of phytoplankton in the area. This led us to use simulated chlorophyll-*a* data, produced by the NEMO oceanic circulation model coupled to the PISCES biogeochemical model [7] for testing the validity of our approach.

Figure 1. The location of BATS.

The data we wanted to classify as hidden states were the NEMO-PISCES output data vectors at BATS, which contained the average vertical chlorophyll-*a* distribution at 17 depth levels (from 5 to 217 meters) and the temperature distribution at nine of these depth levels. The inclusion of the temperature in the hidden chlorophyll-*a* data vectors permits a better representation of the physical conditions that constrain the chlorophyll-*a* development. These vertical distribution profiles (of dimension 26=17+9) were five-day averages of the model, spanning the period from 1992 to 2008 and located in a 2°×2° square centered on BATS. In order to have additional vectors from which to infer the possible states of vertical distribution of chlorophyll-*a*, the vectors at the neighboring grid points of the model were also taken into account. This gave us 9 points, with 73 five-day averaged profiles per year, during17 years for a total of 11,169 profiles for the determination of the hidden states. This is done under the assumption that all the chlorophyll-*a* profiles in a 6°×6° area are similar and can be used for the determination of the SOM maps. Doing so, we increase by a factor of nine the size of the training data set. We will refer to these vectors as hidden vertical distribution (HVD) vectors. A single such vector, taken at time *t*, is refered to as $\boldsymbol{x}_{hid}^{t}$.

In addition to the vertical profiles, the NEMO-PISCES model provides the associated observable surface parameter values, such as SSH, SST, SCHL, WS and SR. These values constitute the components of the model surface (MS) vectors.

We also obtained MODIS observations from 24 June 2002 up to 14 June 2011. These observations consist in SST and SCHL values, which are averaged over the valid pixels of the studied zone, and over the 5-day periods corresponding to those of the model outputs. Empty segments in the temporal series were estimated through linear interpolation. So we run a more realistic experiment by using five-dimensional remote-sensing (RS) vectors containing the SST and SCHL provided by MODIS, keeping only three of the NEMO-PISCES model inputs (SSH, WS and SR).

Consecutive time sequences of the vectors presented are used to generate time-series, denoted S_{hid}, for the sequences of HVD vectors, and S_{obs}, for the sequences of MS or RS vectors. A single vector, taken at time *t*, is denoted $\boldsymbol{x}_{hid}^{t}$ $(\in R^{26})$ for the HVD vectors, as $\boldsymbol{x}_{obs}^{t}$ $(\in R^{5})$ for an observable vector, or more specifically $\boldsymbol{x}_{MS}^{t}$ or $\boldsymbol{x}_{RS}^{t}$, for a MS or RS vector.

3. Method

In this section we discuss the statistical models known as Hidden Markov Models and Self Organizing Maps, and their combination in our method. The flowchart linking the different components of the PROFHMM method are shown in Figure 2.

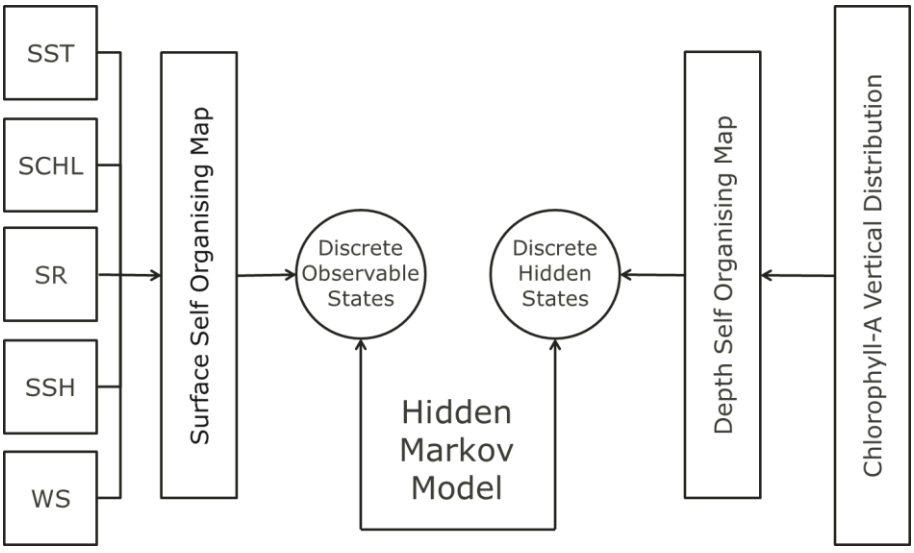


Figure 2. Flowchart showing the links between the different modules of PROFHMM and the observable and hidden vectors.[in the text you have used the spelling "organizing", so, if possible, you should do the same in this Figure; however, both spellings are correct in English; also the "A" in "Chlorophyll-A" should be "*a*"]

In section 3.1 we first introduce the HMM methodology, along with a brief description of the SOM that is necessary in order to understand the HMM architecture. The SOM and their role in PROFHMM are further detailed in section 3.2.

3.1. Hidden Markov models

A Markov model is a stochastic model that assumes the first-order Markovian property, meaning that each consecutive state of the model depends solely on the previous state of the model, that is, $P(X_t| X_1 X_2 ... X_{t-1}) = P(X_t| X_{t-1})$, where X_t is the state of the model at time *t*.

Expanding this principle, a hidden Markov model (HMM) is a stochastic model with two sequences: one sequence of hidden states that follows the first-order Markovian property, and one sequence of observable states which have a statistical link with the hidden states. The Viterbi algorithm [15] then finds the most likely sequence of hidden states, given a sequence of concurrent observations. There exist alternatives for reconstructing sequences [16, 17], but we focus on the Viterbi algorithm in this paper. PROFHMM could however be applied with a different reconstructing algorithm.

The HMM are algorithms which allow us to infer the most likely sequence of some discrete, hidden states, given a series of concurrent observations. To do so, we have to discretize all the available HVD vectors into a set of finite states, in such a way that each state corresponds to a referent vector of the vertical distribution of chlorophyll-*a*. The discretization needs to be very fine in order to obtain the most accurate reconstruction possible. It should therefore permit a partition of the set of HVD vectors into subsets, each one having a very small standard deviation.

The discrete, hidden states need to be connected among themselves through a probability matrix. The probabilities in this matrix correspond to a statistical learning of the dynamic processes governing the temporal transitions between the hidden states. These are referred to as transition probabilities.

The observations need to be consistent in nature and need to be linked, through a probability density function or matrix, with the hidden states. This density function, or the probability matrix elements, corresponds to the existing links between the observations and the hidden states and these elements are referred to as emission probabilities.

Therefore, when inverting sea-surface data to retrieve the vertical profiles of chlorophyll-*a* by using HMM, we are faced with three major problems: the determination of the hidden states, their transition probabilities and the emission probabilities. The continuous, multidimensional nature of parameters included in the sea-surface observations, in conjunction with the fine discretization of the hidden states, imposes a number of constraints on the determination of the conditional probability density functions. However, the calculation of the emission probability becomes easier when the observations are clustered into observable states.

To solve these problems we propose the use of the self organizing maps (SOM) in order to discretize both the sea-surface observations and the hidden vertical chlorophyll-*a* distribution data sets.

SOMs are unsupervised classification algorithms that cluster data into discrete classes. These classes are arranged on a map in such a way that classes that are close on the map represent situations that are close in the original data space. The general concepts of SOMs are further detailed in section 3.2. In the present study, the SOM classification is applied twice, once to the satellite sea-surface data and once to the vertical profiles of chlorophyll-*a* connected to these images. The resulting classes of these topological

maps applied to the vertical profiles correspond to the hidden states of the HMM, whereas the second application of the SOM is for the sea-surface observations and generates the observable states.

Yet, even when we discretize the observation space into states, only a limited region of the observation space may correspond to a particular hidden vertical distribution state. This causes a problem in the calculation of the different probability density functions: that of having a given observation among the seldom-visited specific hidden states.

To overcome such situations in our model, we make the assumption that each hidden state generates observation vectors according to a Gaussian probability density mixture. Two such distinct mixtures of Gaussians, issued from two hidden states close in terms of profiles of HVD vectors, need to determine proximate regions in the observation space. Thus, both the hidden states and the density mixtures need to present a "topology-preserving" aspect.

3.2. Self-organizing maps

Self-organizing topological maps (SOM) are clustering methods based on neural networks. They provide a clustering of a learning data set into a reduced number of subsets, called classes, which share some common statistical characteristics.

Each class is represented by its referent vector $r(\mathrm{i})$ which approaches the mean value of the elements belonging to it. The topological aspect of the maps can be justified by considering the map as an undirected graph on a two-dimensional lattice whose vertices are the N classes. This graph structure permits the definition of a discrete distance $d(\mathrm{C(i)}, \mathrm{C(j)})$ between two classes $\mathrm{C(i)}$ and $\mathrm{C(j)}$, defined as the length of the shortest path between $\mathrm{C(i)}$ and $\mathrm{C(j)}$ on the map. The nature of the SOM training algorithm forces a topological ordering upon the map and, therefore, any neighboring classes $\mathrm{C(i)}$ and $\mathrm{C(j)}$ on the map ($d(\mathrm{C(i)}, \mathrm{C(j)}) = 1$) have referent vectors $r(\mathrm{i})$ and $r(\mathrm{j})$ that are close in the Euclidean sense in the data space.

Let us consider a vector x that is of the same dimensions and nature as the data used to generate the topological map; we can find the index of the class to which it is classified by choosing: $\mathrm{index} = \mathrm{argmin_i}(||\mathrm{x} - \mathrm{r(i)}||)$, therefore assigning it to the class whose referent is closest to it in the Euclidean sense (Figure 3). A classified vector *x* will be represented by its class index, $\mathrm{C(index)}$.

In PROFHMM, we train two SOMs, the first one containing the observations, denoted $\mathrm{sMap_{obs}}$, and the second one containing the distributions of the hidden states, denoted $\mathrm{sMap_{hid}}$. The number of classes in $\mathrm{sMap_{obs}}$ and $\mathrm{sMap_{hid}}$ correspond to the number of states of the HMM, $\mathrm{N_{obs}}$ and $\mathrm{N_{hid}}$. Their respective classes and referent vectors will be denoted $\mathrm{C_{hid}}$ and $\mathrm{C_{obs}}$, $\mathrm{r_{hid}}$ and $\mathrm{r_{obs}}$. We used the algorithms provided by the matlab somtoolbox [9], specifically the functions som_make, som_batchtrain, som_bmus, in order to train our maps and classify our data. The determination of many of the SOM parameters is automatically calculated by this toolbox, by applying the default parameters.

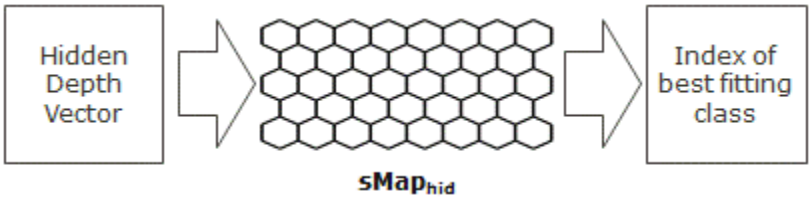


Figure 3. The classification of a HDV through the self-organizing map.

3.3. The PROFHMM method

In PROFHMM, we use the correspondence between the hidden vertical chlorophyll-*a* distribution profiles, $\boldsymbol{x}_{hid}^{t}$, and their class indexes, $C_{hid}(i^t)$, provided by the SOM map, sMap$_{hid}$, for translating the time-series of vectors, S$_{hid}$ ={[$\boldsymbol{x}_{hid}^{1}$, . . . , $\boldsymbol{x}_{hid}^{T}$}, into the corresponding times-series of class indexes, SI$_{hid}$ = [{(i^1), . . . , $C_{hid}(i^T)$ }. Similarly, we translate the time-series of observations, S$_{obs}$ = {$\boldsymbol{x}_{obs}^{1}$, . . . , $\boldsymbol{x}_{obs}^{T}$}, into the time-series of class indexes, SI$_{obs}$ = {$C_{obs}(i^1)$, . . . , $C_{obs}(i^T)$}.

We consider two phases: training and retrieval.

During the training phase, a number of concurrent couples of sequences of indexes, SI$_{hid}$ and SI$_{obs}$, whose length in consecutive time-steps is denoted as L$_{seq}$, are used in order to estimate the elements of the Transitions matrix, Tr$_{B\text{-}W}$, and the Emissions matrix, Em$_{B\text{-}W}$. We have then at our disposal a set of sequences, $A = \{seq_i, i \in 1 \cdots N_{seq}\}$, where N_{seq} is the number of sequences. These probabilities are estimated by using the Baum-Weltch algorithm [18], which is a particular case of a generalized expectation–maximization algorithm that takes as input all the concurrent sequences of SI$_{obs}$ and SI$_{hid}$ and outputs the most likely matrices to have generated them through a hidden Markov process.

Tr$_{B\text{-}W}$ contains the transition probabilities of the hidden states

$$tr_{i,j} = P(C_{hid}(i^t) = i \mid C_{hid}(i^{t-1}) = j) \quad (1),$$

$$\text{where } \sum_{i=1}^{N_{hid}} tr_{i,j} = 1 \quad (2).$$

Tr$_{B\text{-}W}$ corresponds, in a physical sense, to the underlying dynamics that govern the hidden states, containing the probabilities $tr_{i,j}$ of going from a hidden state *j* to the hidden state *i* at time *t*.

Em$_{B\text{-}W}$ contains the *a posteriori* probabilities of each observed state to have been emitted by a hidden state,

$$e_{i,j} = P(C_{obs}(i^t) \mid C_{hid}(j^t)) \quad (3),$$

$$\text{where } \sum_{i=1}^{N_{hid}} e_{i,j} = 1 \quad (4).$$

In a physical sense, $Em_{B\text{-}W}$ corresponds to the link existing between the observed quantities and the dynamics of the unobserved quantities, $e_{i,j}$, presenting the probability of having a given observed state *i*, given the concurrent hidden state *j* at time *t*.

Another probability matrix that needs to be calculated is the initial probability matrix Π, whose components, π_i, represent the average revisit rate of each hidden state given an infinite sequence. This matrix is used during the retrieval phase in order to optimize the starting point of the reconstruction of the most likely sequence.

For the retrieval phase, HMMs use the Viterbi algorithm, which is a well known dynamic programming algorithm, for inferring the most likely sequence of indexes, SI_{hid}, given the previously estimated parameters $Tr_{B\text{-}W,}$ $Em_{B\text{-}W}$ and Π of the HMM and a sequence of observation indexes, SI_{obs}.

3.4. Optimizing the estimation of probabilities

The method presented up to now, however, if applied without taking into account the specificities of the problem, namely the restricted number of observations in the data set used to estimate the probabilities, can present some inconsistencies. Therefore, once we have created the topological maps and have acquired the states of the HMM, we need to focus on some problems inherent in the hidden state reconstructions.

It is known that the Viterbi algorithm may present problems when performing a reconstruction based on probabilities computed with a training data set that misses transitions that exist in reality [5]. A balance needs to be found between the sizes of the SOM maps that determine the amount of discretization provided by the method and the correctness of the reconstruction. The size of the maps, which define the values of N_{obs} and N_{hid}, are optimized in order to get the best results for the HMM. This optimization is an iterative process of training both maps with different sizes, then running PROFHMM and performing a cross-validation using a separate validation data set [19].

Yet, even with an optimization, there will be some situations and transitions that are seldom encountered in the training data and result in null probabilities in the probability matrices, $Em_{B\text{-}W}$ and $Tr_{B\text{-}W}$ that we estimated in the first pass of the Baum-Weltch algorithm.

Due to this usual lack of sufficient data in the concerned domains, $Em_{B\text{-}W}$ and $Tr_{B\text{-}W}$ need to be adjusted. This is done by taking into account the properties of the SOM. A major characteristic of the present method is to use the topological order in order to improve the accuracy of the estimated probability matrices. SOMs allow us to modify the probabilities by allowing each state to communicate via a diminutive probability with each of its neighboring states.

This is done by considering the neighborhood matrices, NM_{obs} and NM_{hid}, of dimensions (N_{obs},N_{obs}) and (N_{hid},N_{hid}), where

$$NM_{SOM}(i,j)=\begin{cases}1, \text{if } d\left(C_{SOM}(i), C_{SOM}(j)\right) < 2 \\ 0, else\end{cases} \qquad (5)$$

with d(i,j) being the discrete distance on the map, with SOM representing alternatively either $sMap_{obs}$ or $sMap_{hid}$.

Taking into account the neighborhood states for calculating the final transition probabilities consists in increasing the probability of reaching a class j from a class i by an amount proportional to the sum of the previously calculated probabilities of reaching the neighbor classes of class j on either $sMap_{obs}$ or $sMap_{hid}$.

We therefore modified our algorithms in order to take into account multiple concurrent time sequences of indexes during the training. For each of those training sequences, we apply the Baum-Weltch algorithm and get two estimated initial probability matrice,s $Tr_{B\text{-}W(seq),}$ $Em_{B\text{-}W(seq)}$.

In order to favor the data observed during training, we add a weighting term, w_c, to the initial probabilities and we further multiply it by the square root of the total length of each training sequence used in the initial Baum-Weltch algorithm pass, noted L_{seq}, since this length is a measure of confidence in the correctness of the estimated parameters. The matrices obtained are increased by adding a constant to avoid null probabilities.

The final Em and Tr matrices, Em_{final} and Tr_{final}, are computed by applying for $1 \leq i \leq N_{obs}$ and for $1 \leq j \leq N_{hid}$, using all the sequences of the training data set A:

$$Em_{final}(i,j) = \sum_{N_{seq}} \left(w_c * \sqrt{L_{seq}} * Em_{B-W(\text{seq})}(i,j) + \sum_{k=1}^{N_{hid}} \left(NM_{hid}(j,k) * Em_{B-W(\text{seq})}(i,k) \right) \right) + 1, \quad (6)$$

Which is normalized to fit the constraint where $\sum_{i=1}^{N_{hid}} e_{i,j} = 1$ (7),

and for $1 \leq i,j \leq N_{hid}$, using all the sequences of the training data set A:

$$Tr_{final}(i,j) = \sum_{N_{seq}} \left(w_c * \sqrt{L_{seq}} * Tr_{B-W(\text{seq})}(i,j) + \sum_{k=1}^{N_{obs}} \left(NM_{hid}(i,k) * Tr_{B-W(\text{seq})}(i,k) \right) \right) + 1 \quad (8)$$

which is normalized to fit the constraint $\sum_{i=1}^{N_{hid}} tr_{i,j} = 1$ (9).

Each observation increases the emission probability of the most likely hidden state and the probability of its neighbors. In PROFHMM, we assume that when we classify a sea-surface observation on $sMap_{obs}$, it is classified correctly. However, if that is not the case, it would have been classified in one of the neighboring classes (Figure 4). This is taken into account by using the Em_{final} and Tr_{final} matrices.

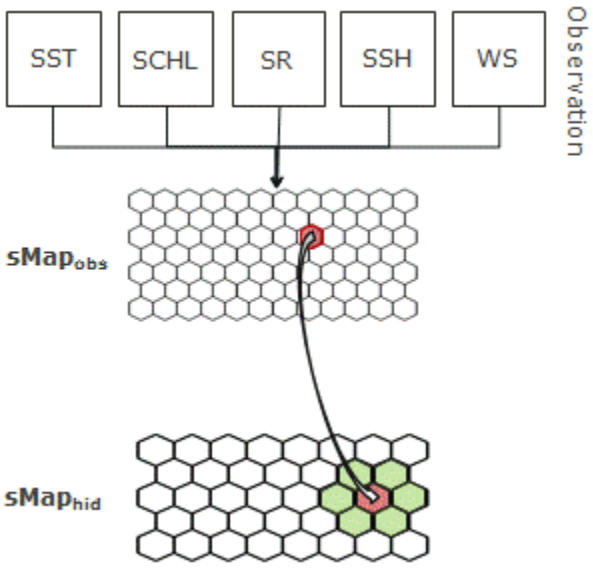


Figure 4. The emission probability of each class $C_{obs}(i)$ of the $sMap_{obs}$ from a class $C_{hid}(j)$ of $sMap_{hid}$ takes into account the probability of being emitted by a class $C_{hid}(k)$ neighboring $C_{hid}(j)$.

The above modifications permit the Viterbi algorithm to circumvent the problems of impossible transitions or emissions due to insufficient data in the training sequences that resulted in null probabilities in the estimated parameters.

4. RESULTS

As noted previously, the method was tested with two configurations, once by using model forcing and outputs in order to train the HMM and reconstruct the HVD vectors, and once when using a combination of satellite observations and model forcings and outputs in order to achieve the same reconstruction.

In both cases, the $sMap_{hid}$ was trained by taking into account all available profiles surrounding BATS, corresponding to 11,169 profiles (section 2).

An important point that we have to consider is the notion of "optimum reconstruction", given the discretization provided by the SOM. This term corresponds to the series of indexes, SI_{opt}, we would have obtained by projecting the complete HVD vectors on the $sMap_{hid}$. Indeed, the performance of PROFHMM is bound by this discretization, since even when we correctly reconstruct the SI_{hid} based on the SI_{obs}, a quantification error, due to the SOM discretization, still exists. This have led us to include the percentage of corresponding indexes between SI_{opt} and the reconstructed series of indexes, SI_{rec}, as a measure of the optimal performance. It must be noted that even when the algorithm does not recover exactly the same index, the effect of this misclassification on the reconstruction is low if the index retrieved is close to the optimum one. This again is due to the similarity of the referent vectors of neighboring classes.

The 10 first-time steps of each reconstruction by PROFHMM tend to have poor performances, since the reconstruction is greatly based on the observations, and the dynamic processes of the hidden states have

not been long enough to drive the system to the most likely state. The performances shown below, which include these 10 first-time steps, would be improved if we ignore these time steps.

4.1 Reconstruction from model surface vectors

We chose 2-D hexagonal SOM maps. In the NEMO-PISCES application, all the available HVD and MS vectors have been used during the learning phase of the SOM. The optimal architecture, after the cross-validation process, was found to have 294 (21x14) states both for the $sMap_{hid}$ and the $sMap_{obs}$.

For the estimation of the HHM parameters, we only took 13 years (1992–2005) for the training, each including 73 five-day-mean steps. Therefore we only have a unique training sequence, L_{seq}=1022 time steps. We kept three years (2006–2008), or 219 five-day-mean steps, to validate our approach.

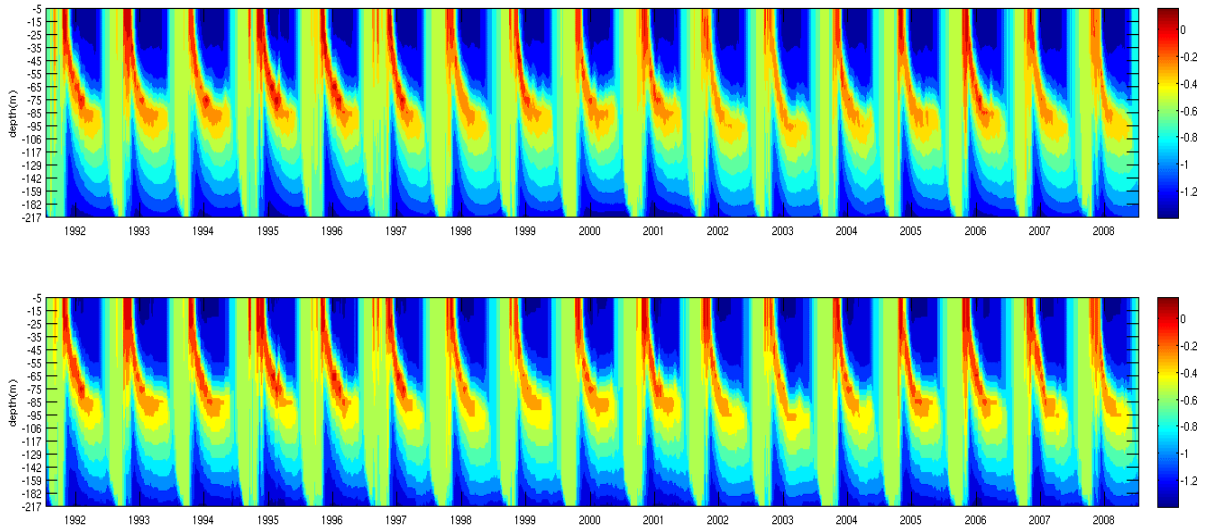


Figure 5. The chlorophyll-*a* values of the NEMO-PISCES model (top) and those given by the PROFHMM inversion (bottom), at BATS, for the period 1992 to 2008. The *X* axis represents time, the *Y* axis represents depth, and the colorbar is in $\log_{10}$ [ng/l] The last three years are validation years.

We can see, from Figure 5, that the PROFHMM reconstruction respects the form and general intensity of the chlorophyll-*a* profile evolution, throughout the period 1992–2008. At first glance there is no apparent difference between the reconstructions of the training (1992–2005) and the validation years (2005–2008).

4.1.1 Reconstruction of the year 2005

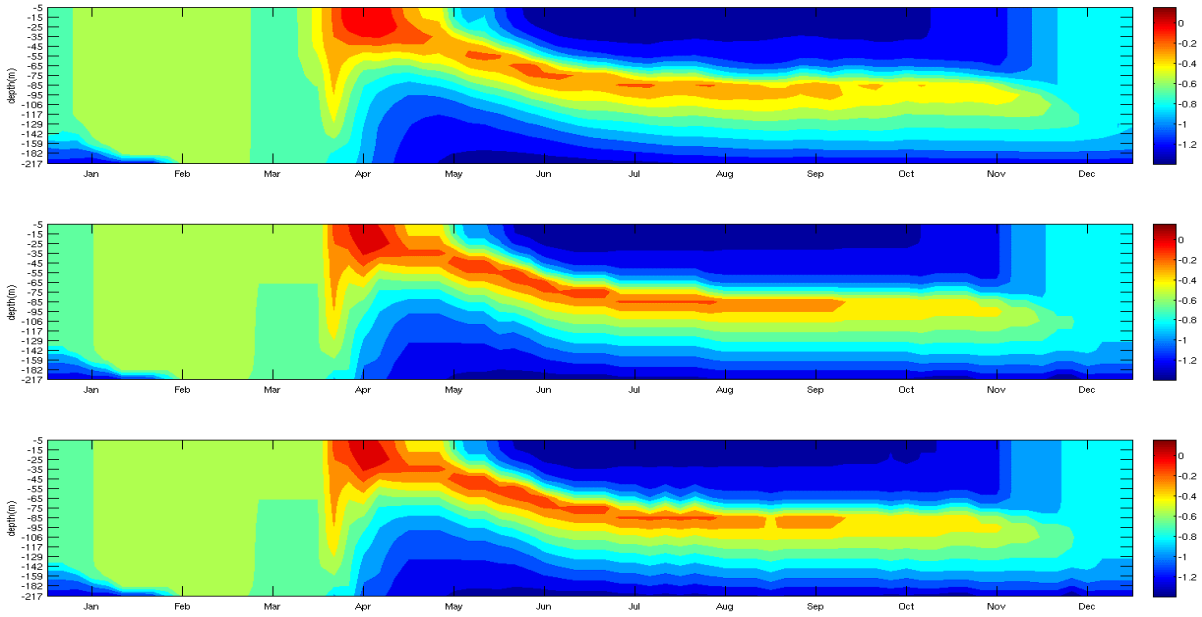


Figure 6. The chlorophyll-*a* values of the NEMO-PISCES model (top), the result of the PROFHMM inversion (middle), and the optimum reconstruction of the NEMO-PISCES model (bottom), at BATS for the year 2005. The *X* axis represents time, the *Y* axis represents depth, and the colorbar is in $\log_{10}$ [ng/l].

The year 2005 was the last year in the training data set. In Figure 6, which is a zoom on the year 2005, the reconstruction follows the form and intensity of the NEMO-PISCES chlorophyll-*a* values. The transition between states with different concentrations of chlorophyll-*a* in March is less accurate than in the rest of the reconstruction. This seems to be due to the fact that the situation present at the time in the model (top panel) is poorly represented in sMap_{hid} (bottom panel). We may assume that, given a finer discretization and longer training sequences, this drawback would disappear.

4.1.2. Reconstruction of the year 2008

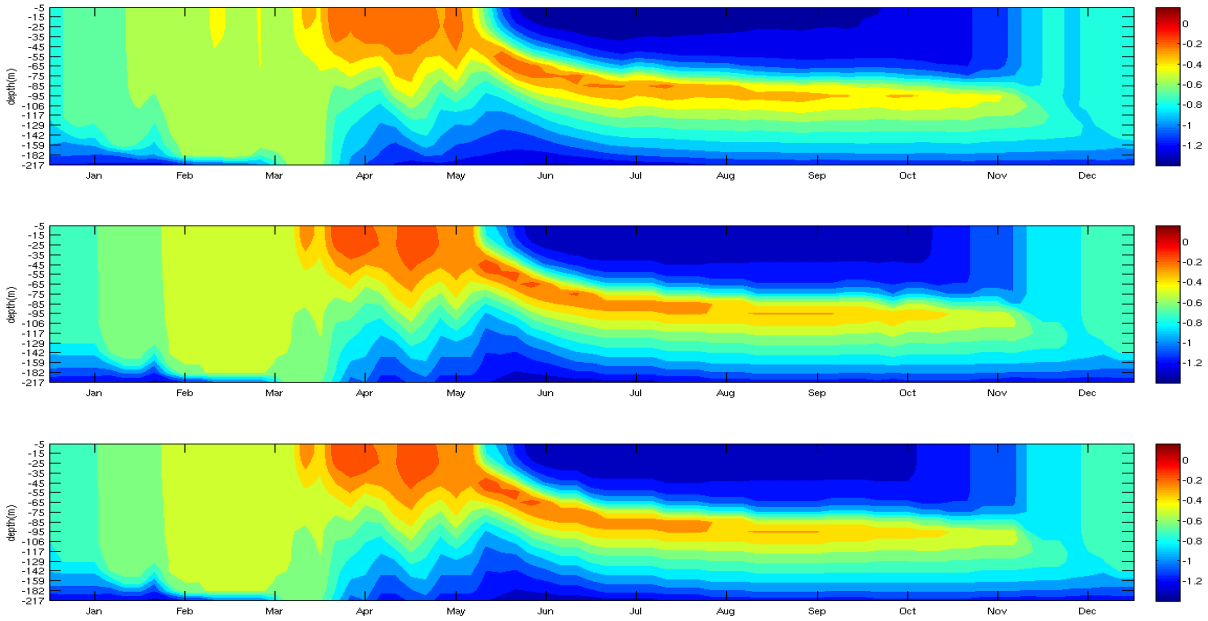


Figure 7. The chlorophyll-*a* values of the NEMO-PISCES model (top), the result of the PROFHMM inversion (middle), and the optimum reconstruction of the NEMO-PISCES model (bottom), at BATS for the year 2008. The *X* axis represents time, the *Y* axis represents depth, and the colorbar is in $\log_{10}$ [ng/l].

The year 2008 is a validation year for the PROFHMM reconstructions. The reconstruction still follows the form and intensity of the NEMO-PISCES values. As with the reconstruction of the year 2005, there are two quantification errors: one during late February and one during late May. This is seen in detail in Figure 8, where most of the errors are less than 0.02 $\log_{10}$ [ng/l] and none of the errors exceeds 0.2 $\log_{10}$ [ng/l].

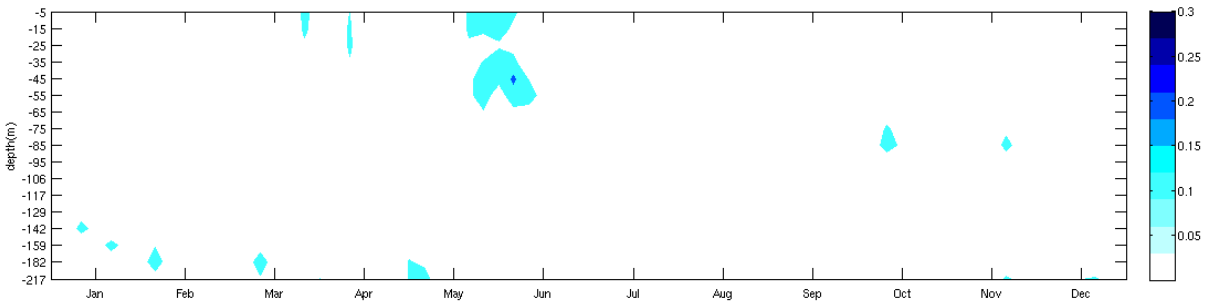


Figure 8. The absolute error in chlorophyll-*a* values between the NEMO-PISCES model and the result of the PROFHMM inversion The *X* axis represents time, the *Y* axis represents depth, and the colorbar is in $\log_{10}$ [ng/l].

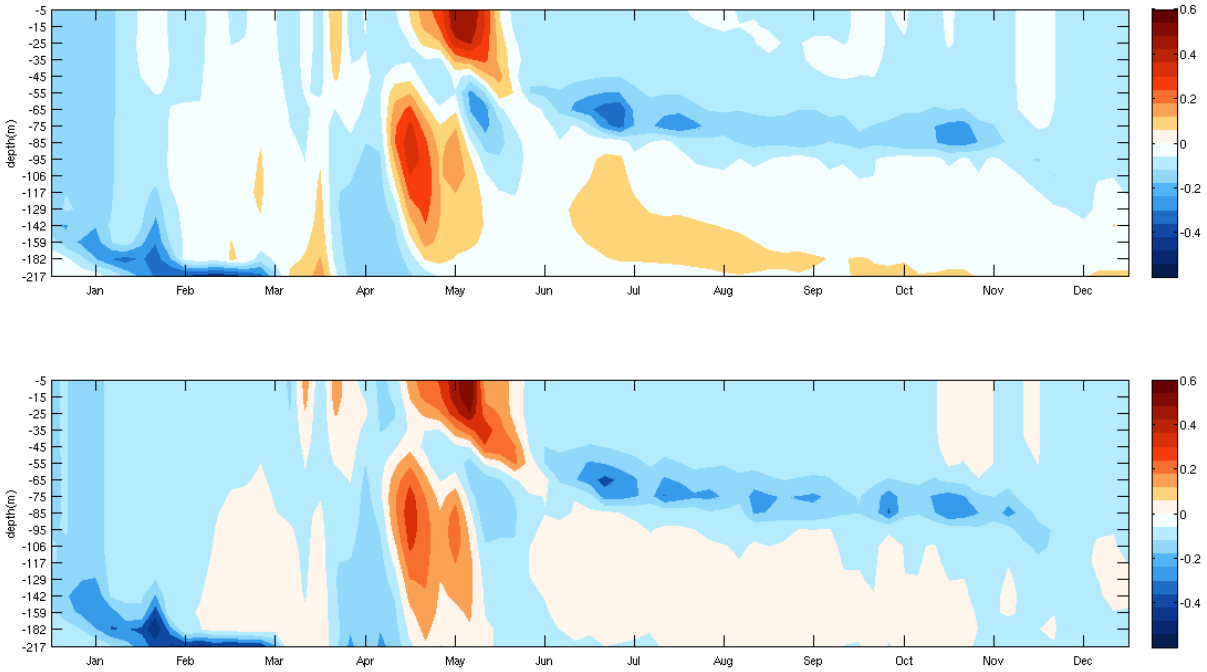


Figure 9. The difference in chlorophyll-*a* values between the 1992–2008 average year and the NEMO-PISCES model (top), and the difference in chlorophyll-*a* values between the 1992–2008 average year and the result of the PROFHMM inversion (bottom), for the test year 2008. The *X* axis represents time, the *Y* axis represents depth, and the colorbar is in $\log_{10}$ [ng/l].

In order to ensure that we are not repeatedly reconstructing the mean year, we calculated the climatology for the period 1992–2008 and subtracted it from both the NEMO-PISCES chlorophyll-*a* data and from the PROFHMM MS reconstruction. The climatology was computed by averaging over the 17 available years, the vertical distributions of chlorophyll-*a* of each of the 73 five-day steps that constitute each year.

4.2. Reconstruction from remote-sensing vectors

When applying PROFHMM to MODIS satellite data, we again used the NEMO-PISCES simulated data to represent the vertical profiles of chlorophyll-*a*. The available data provided by the NEMO-PISCES model and the MODIS sea-surface temperature and chlorophyll-*a* data were concurrent only through the last months of the year 2002 and the entire 2003–2008 period, which represents roughly six years of data. We kept the year 2008 as a validation set, and used the rest of the data for the training of both the sMap_{obs} and the HMM, corresponding to 402 five-day steps. This led us to decrease N_{obs} to 80 classes (arranged in an 8 × 10 matrix), due to the limited amount of vectors. This limited amount of vectors could provide a less discretized space and the omission of certain important states in our model. As an additional measure to overcome this, we initialized the training of sMap_{obs} by first training it on the MS data used before and continuing the learning with the RS data. As we used again the NEMO-PISCES data to perform our reconstructions, we kept the same sMap_{hid} as the one used in our first experiment. The degradation of the results due to the use of satellite images is minimal, as shown in Figure 10.

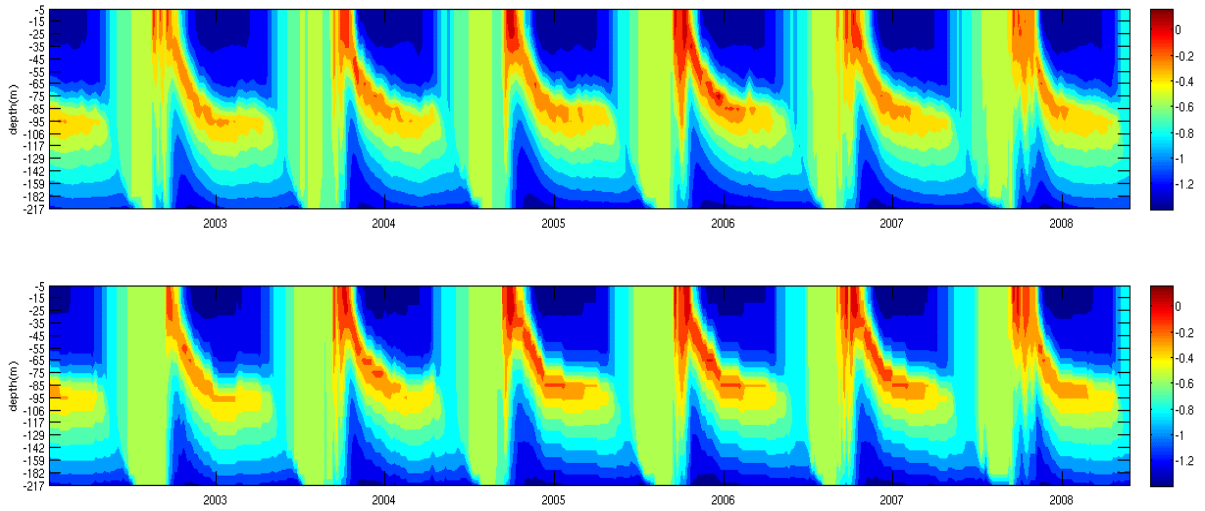


Figure 10. The chlorophyll-*a* values of the NEMO-PISCES model (top) and those given by the PROFHMM inversion based on MODIS satellite data (bottom), at BATS, for the period 2002 to 2008. The *X* axis represents time, the *Y* axis represents depth, and the colorbar is in $\log_{10}$ [ng/l]. The last year (2008) is a validation year.

4.2.1. Reconstruction of the year 2005

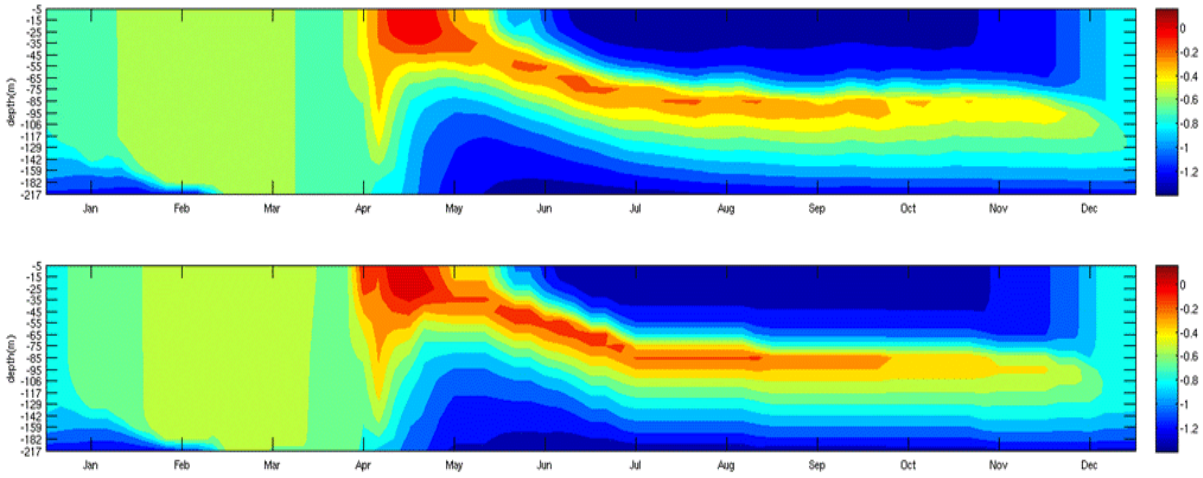


Figure 11. The chlorophyll-*a* values of the NEMO-PISCES model (top), the result of the PROFHMM inversion based on MODIS data (bottom), at BATS for the year 2005. The *X* axis represents time, the *Y* axis represents depth, and the colorbar is in $\log_{10}$ [ng/l].

The year 2005 was in the training data for both experiments (MS and RS). In Figure11, which is a zoom of the MODIS-based reconstruction of the year 2005, the reconstruction again follows the form and intensity of the NEMO-PISCES chlorophyll-*a* values. The reconstruction has slightly more pronounced high values.

4.2.2. Reconstruction of the year 2008

The year 2008 is, as stated above, a validation year for both PROFHMM reconstructions. Although we only used six years of data for the training of the HMM, and 402 RS vectors for the training of the $sMap_{obs}$, the reconstruction still follows the form and intensity of the NEMO-PISCES values. Again, the reconstruction of the test year presents slightly higher values, as shown in Figure 12.

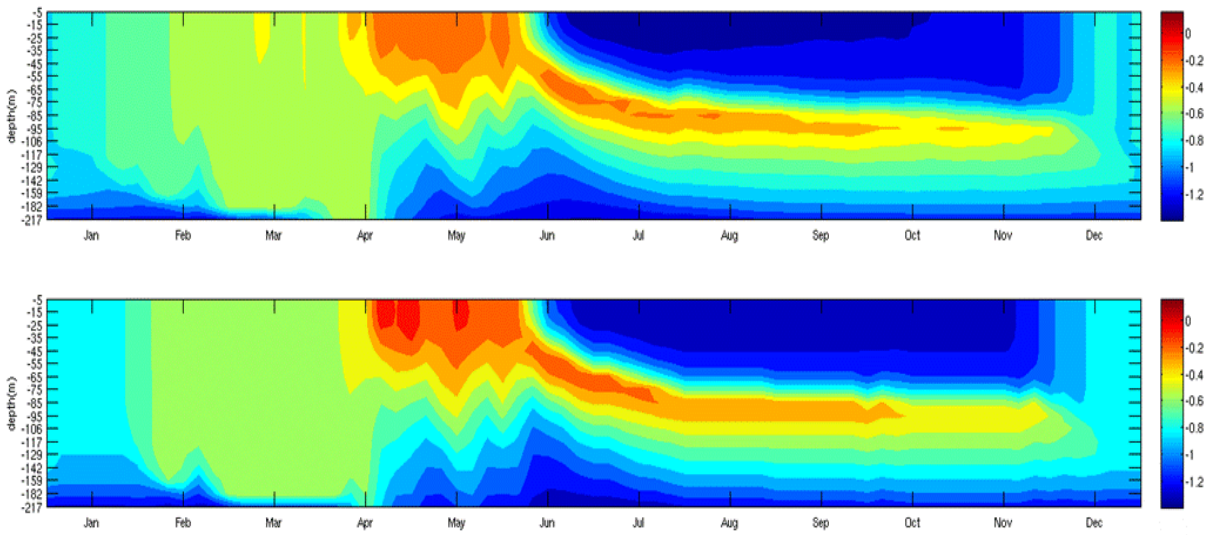


Figure 12. The chlorophyll-*a* values of the NEMO-PISCES model (top), the result of the PROFHMM inversion based on MODIS satellite data (middle), and the optimum reconstruction of the NEMO-PISCES model (bottom), at BATS for the year 2008. The colorbar is in $\log_{10}$ [ng/l].

4.3. Comparisons

In order to have some quantifiable results we computed the RMS between the PROFHMM reconstruction of chlorophyll-*a* and the NEMO-PISCES model values, as well as the RMS between the optimum reconstruction and the NEMO-PISCES model values.

We also located, in each case, the 10% of the minimum and 10% maximum values of chlorophyll-*a* in the NEMO-PISCES model, and calculated the RMS between them and their co-localized points in the PROFHMM and optimum reconstructions. This allows us to affirm that the extreme values are also well reconstructed by PROFHMM.

These RMSs are shown in Table 1, along with the percentage of retrieved indexes (SI_{ret}) corresponding to the optimum indexes (SI_{opt}).

The term inverted corresponds to the PROFHMM reconstruction, while the term optimum corresponds to the "optimum reconstruction". The years above these terms correspond to the period being reconstructed. MS DATA performances correspond to the reconstructions done with sea-surface observations taken from the model outputs, while RS DATA performances correspond to the reconstruction done with the help of MODIS satellite images.

Table 1 PROFHMM RMS performances in [ng/l].

	MS DATA PERFORMANCES				RS DATA PERFORMANCES			
	RMS	MIN	MAX	%	RMS	MIN	MAX	%
1992-2008 INVERTED	0.0455	0.0034	0.1192	99.97%	-	-	-	

1992-2008 OPTIMUM	0.0423	0.0032	0.1158		-	-	-	
2005 INVERTED	0.0411	0.0072	0.0422	93.15%	0.0499	0.0067	0.0514	84.93%
2005 OPTIMUM	0.0400	0.0069	0.0411		0.0400	0.0069	0.0411	
2008 INVERTED	0.0303	0.0076	0.0310	91.78%	0.0399	0.0096	0.0408	75.34%
2008 OPTIMUM	0.0302	0.0076	0.0309		0.0302	0.0076	0.0309	
2002-2008 INVERTED	0.0453	0.0097	0.0466	84.63%	0.0590	0.0217	0.0856	75.74%
2002-2008 OPTIMUM	0.0406	0.0077	0.0418		0.0406	0.0077	0.0418	

In Table 1, the RS performances for the year 2008 as seen in the difference OPTIMUM–INVERTED is less than the same difference computed over the complete period (2002–2008). This indicates that the performances on the training were poorer for this experiment, but that they would greatly improve given a longer training data set.

5. Conclusion

In the present paper we have introduced PROFHMM, which is an inversion method based on SOM and HMM. PROFHMM is able to reconstruct hidden profiles of biogeochemical parameters from observable data at the top layer of the profile. We applied this method for the reconstruction of the chlorophyll-*a* vertical profiles at BATS, using model outputs and satellite data as sea-surface observations. The method was also applied at the HOT (Hawaii Ocean Time-series) location of the JGOFS campaign and, for which we obtained similar performances (not shown in the article).

Upon developing PROFHMM, we realized that it could be modified to a more general, statistical, non-linear method that could be applied to the reconstruction of other oceanic parameters.

We intend to expand the method in order to reconstruct the spatial evolution of the temperature field on the transect of the ARAMIS campaign [20]. However PROFHMM is general enough to be applicable to a multitude of other problems in geophysics, for which it is possible to learn, in a statistical way, the dynamics of a geophysical model, while observing a sub-dimension of the parameters.

PROFHMM is very efficient in terms of calculations. Its cost efficiency could allow its integration into reanalysis or forecasting models for improving assimilation by producing accurate first guesses of the model evolution.

Long-term perspectives of PROFHMM include taking into account the amount of incomplete satellite observations in the inversions, and expanding the principle to Bayesian fields in order to reconstruct space–time evolutions of geophysical or biogeochemical parameters.

Acknowledgments

The research presented in this paper was financed by the Centre National de l'Etude Spatial (CNES, French National Center for Space Studies), and the Délégation Gouvernementale pour l'Armement (DGA, French Military Research Delegation), both of which we thank for their support. We also thank Cyril Moulin and Laurent Bopp, of the Laboratoire des Sciences du Climat et l'Environnement (LSCE; Climate and Environmental Sciences Laboratory), for their help with NEMO-PISCES, and Michel Crépon of the Laboratoire Océanographique et du Climat–Expérimentation et Approches Numériques (LOCEAN; Oceanographic and Climate Laboratory–Experimentation and Numerical Approaches) for his input on the method.

References

[1] Feldman G.C., N.A. Kuring, C. Ng, W.E. Esaias, C.R. McClain, J.A. Elrod, N.Maynard, D. Endres, R. Evans, J. Brown, S. Walsh, M. Carle, G. Podesta (1989). Ocean color: availability of the global data set. *EOS*, 70:634-641

[2] Dinnat E., J. Boutin, G. Caudal, J. Etcheto, P. Waldteufel. (2002) Influence of sea surface emissivity model parameters in L-band for the estimation of salinity. Int. J. Rem. Sensing, 23:5117-5122

[3] Krishna Rao P., W.L. Smith, R. Koffler (1972). Global sea-surface temperature distribution determined from an environmental satellite. *Monthly Weather Review* 100(1):10–14

[4] Brajard J., J. Cédric, M. Cyril, S. Thiria (2006). Use of a neuro-variational inversion for retrieving oceanic and atmospheric constituents from satellite ocean color sensor: Application to absorbing aerosols neural networks. Earth Sciences and Environmental Applications of Computational Intelligence, 19(2). Elsevier, Amsterdam

[5] Uitz J., H. Claustre, A. Morel, S.B. Hooker (2006). Vertical distribution of phytoplankton communities in open ocean: An assessment based on surface chlorophyll, *J. Geophys. Res.*, 111, C08005, doi:10.1029/2005JC003207

[6] Gehlen M., A. Moussaoui, C. Perruche, E. Dombrowsky, O. Aumont, P. Brasseur, P. Le Sommer, P. Lehodey (2010). Integration of biogeochemistry and ecology to Mercator ocean systems: Recent advances and future developments of the Green Mercator initiative. MyOcean Science Days 1–2.

[7] Gurvan M. and the NEMO team. (2012). NEMO ocean engine – version 3.4 – Note du Pôle de Modélisation de l'Institut Pierre-Simon Laplace, 27 (ISSN 1288-1619). Institut Pierre-Simon Laplace, Paris.

[8] Juang B.-H. (2003). Hidden Markov models. Encyclopedia of Telecommunications. Wiley Online Library (onlinelibrary.wiley.com)

[9] Kohonen T. (1990). The self-organizing map. *Proc. IEEE*, 78(9). (http://www.cis.hut.fi/projects/somtoolbox/package/docs2/somtoolbox.html)

[10] Jaziri R., M. Lebbah, Y. Bennani, J.-H. Chenot (2011) SOS-HMM: Self-organizing structure of hidden Markov model, artificial neural networks and machine learning – ICANN 2011, Lecture Notes in: *Computer Science*, 6792. pp87-94

[11] Doneya S.C., J.A. Kleypasa, J.L. Sarmiento, Falkowski P.G. (2002). The US JGOFS Synthesis and Modeling Project – An introduction. *Deep-Sea Res.* II, 49:1-20

[12] NASA OceanColor website (http://oceancolor.gsfc.nasa.gov/)

[13] Miller C.B. (2003). Biological Oceanography, ISBN 0632055367

[14] Jolliffe I.T. (2002). *Principal component analysis* (2nd ed.). Springer, Heidelberg, New York, Berlin

[15] Viterbi A.J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260-269 doi:10.1109/TIT.1967

[16] Viterbi A.J. (1998). An intuitive justification and a simplified implementation of a MAP decoder for convolutional codes. IEEE Journal on Selected Areas in Communications, 16(2):260-264

[17] Hagenauer J. and P. Hoeher (1989). A Viterbi algorithm with soft-decision outputs and its applications. Proc. IEEE GLOBECOM Conference, Dallas, Texas, USA, November 1989. pp47.11-47.17

[18] Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique

occurring in the statistical analysis of probabilistic functions of Markov chains. Annals

of Mathematical Statistics, 41:164-171

[19] Kohavi R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, 2(12):1137–1143. Morgan Kaufmann, San Mateo, California

[20] LOCEAN, 2002-2009 : Altimétrie sur un rail atlantique et mesures in situ. (ARAMIS) (http://aramis.locean-ipsl.upmc.fr)

2.3 Annexe de l'article : Analyse préalable des données utilisées pour la construction des cartes topologiques

Le phytoplancton est l'ensemble des micro-organismes végétaux qui vivent dans l'eau. Ceux-ci, grâce à la photosynthèse transforment le CO2 dissous en oxygène en carbone organique. Le développent du phytoplancton est lié à certains paramètres :

- La turbulence de l'eau : elle affecte la quantité de nutriments accessible au phytoplancton, ainsi que la quantité de temps passée au soleil pour chaque cellule de phytoplancton. La turbulence est liée aux vents de surface et à la circulation océanique de surface. Celle-ci est détectable quand on prend en considération la topographie de la surface de la mer.
- La quantité de rayonnement solaire accessible: elle affecte le taux de photosynthèse possible et dépend de la radiation solaire et de la quantité de nuages.
- La température de l'eau : différentes espèces se développent à différentes températures.

Une partie de ces phénomènes sont observables; ce sont ces observations ont été utilisés dans l'article pour mettre au point la méthode inverse permettant de passer de la surface à la profondeur et de reconstruire les profils de température et de chlorophylle.

La détermination des états observés et cachés de la Chaine de Markov Cachées est réalisée à l'aide des cartes topologiques. Les états recherchés doivent être représentatifs de l'ensemble des situations qui vont être observées à la station océanique BATS. Nous avons à notre disposition pour cela 1241 situations à BATS, ce qui ne pourrait pas nous donner une quantification par carte topologique très fine. Nous avons donc décidé d'y adjoindre l'ensemble des données relatives aux points voisins. Nous avons pour cela extrait les paramètres fournit par les sortie du modèle NEMO-PISCES au point de grille contenant BATS, et nous y avons adjoint les données relatives aux 8 points voisins. La grille est au 2° de degré et donc en ajoutant ces points nous supposons que les processus étudiés sur une région de ≈450.000 km² ne sont pas trop différents.

La base de données est donc constituée de 11169 situations représentées par : l'élévation du niveau de la mer (SSH), la couverture Nuageuse (CC), l'Intensité du Vent (WS), la radiation

Solaire (SR), la température de l'eau aux profondeurs de 5,45, …, 145 mètres (THERM 1 à 7), la concentration de phytoplancton aux mêmes profondeurs (CHL 1 à 7), la latitude du point d'observation (latitude), la longitude du point d'observation (longitude). Ce qui donne un tableau de donné initial de 11196 lignes et 20 colonnes.

L'analyse Analyse en Composantes Principales qui suit cherche d'une part à vérifier l'hypothèse selon laquelle les points voisins représentent un même phénomène et peuvent donc venir compléter la base de données. Elle veut d'autre part montrer qu'une connaissance suffisante existe au niveau des données de surface pour retrouver les profils de CHL et de température qui nous intéressent. Etant donné le grand nombre de variables prises en compte (20) et les différences d'unité, nous avons réalisé une Analyse en Composantes Principales normé,

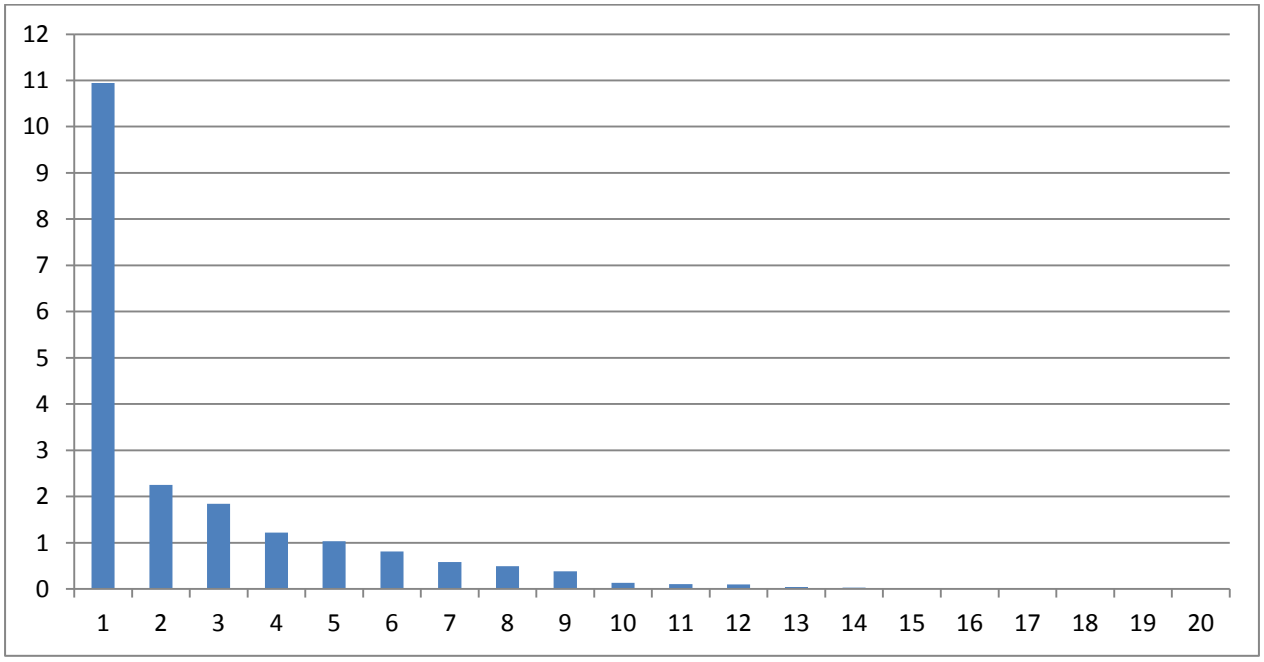


Figure 2.3.1 : Distribution de la variance expliquée par projection sur chaque axe principal de la PCA

La figure 2.3.1 présente la distribution de la variance en fonction des différents axes principaux issus de l'analyse. En utilisant le critère de Kaiser de sélection des axes en PCA normalisée, nous avons gardé les axes dont la valeur propre est supérieure à 1. Ainsi on peut considérer que le phénomène observé est explicable avec 5 axes, qui contiennent 86,46% de la variabilité

de l'ensemble des données (Tableau 2.3.1). La figure 2.3.2 présente la projection des variables sur le plan des deux premiers axes qui expliquent 65,99% de l'information.

Tableau 2.3.1 : Variances expliquées et valeurs propres des 10 premiers axes de la PCA.

Axe Principal	% Variance Expliquée	Valeur Propre	Axe Principal	% Variance Expliquée	Valeur Propre
1	54,72	10.9447	6	4,07	0.8145
2	11,27	2.2538	7	2,93	0.5863
3	9,20	1.8406	8	2,47	0.4943
4	6,11	1.2229	9	1,92	0.3834
5	5,16	1.0317	10	0,68	0.1356

Si l'on regarde le tableau 2.3.2 et la figure 2.3.2, on peut voir que la longitude et la latitude sont les variables les plus mal représentées sur le plan des deux premiers axes principaux. La latitude représente la principale variable de l'axe 4 et la longitude de l'axe 5. Dans cette région, la variation en latitude est due à la variabilité du Gulf Stream, elle est plus importante que celle due à la longitude, ce qui correspond aux résultats produits par l'analyse. Le fait d'avoir ajouté des points voisins à l'ensemble des données peut d'une certaine manière avoir ajouté une variabilité supplémentaire aux données (<10%). Cependant l'ACP nous montre que le bruit introduit n'est pas trop important.

Si l'on regarde maintenant sur la figure 2.3.2 la représentation des autres variables, il est clair que l'on est en présence d'un effet taille, les profils de température étant anti corrélés avec ceux de chlorophylle. La concentration en chlorophylle est, en générale, d'autant plus forte que la température de l'eau est froide et donc que l'apport en nutriment des eaux profondes est fort. Le premier axe oppose donc les situations chaudes avec peu de chlorophylle aux situations froides. Le second axe oppose les situations ensoleillées favorables à l'éclosion du phytoplancton à celles où il y du vent et des nuages.

Table 2.3.2 : Corrélation de variables avec les 5 premiers axes de la PCA.

	Axe 1	Axe 2	Axe 3	Axe 4	Axe 5
SSH	-0.7946	0.0724	-0.2688	-0.3467	0.2057
CC	0.1841	0.5050	0.1070	0.3432	0.1030
WS	0.2882	0.6419	0.3213	-0.0147	0.1133
SR	0.0900	-0.8929	-0.1403	0.1134	-0.0770
THERM 1	-0.9095	-0.2827	-0.1231	0.1610	-0.0659
THERM 2	-0.9237	-0.2328	-0.1428	0.1609	-0.0687
THERM 3	-0.9420	-0.1492	-0.1781	0.1563	-0.0665
THERM 4	-0.9522	-0.0412	-0.2129	0.1461	-0.0475
THERM 5	-0.9444	0.0739	-0.2352	0.1303	-0.0102
THERM 6	-0.9161	0.1791	-0.2429	0.1089	0.0399
THERM 7	-0.8717	0.2633	-0.2387	0.0839	0.0932
CHL 1	0.7934	0.1252	-0.5306	0.0279	-0.1050
CHL 2	0.7946	0.1140	-0.5429	0.0288	-0.1133
CHL 3	0.8165	0.0898	-0.5369	0.0286	-0.1084
CHL 4	0.8517	0.0260	-0.4513	0.0465	-0.0420
CHL 5	0.8474	-0.1212	-0.2283	0.1018	0.1051
CHL 6	0.7460	-0.3531	0.0919	0.1644	0.2541
CHL 7	0.5091	-0.5482	0.3234	0.1743	0.2437
latitude	0.1765	0.1983	0.1962	0.8327	-0.2786
longitude	0.0729	-0.0512	0.2796	-0.2694	-0.8270

La distribution en profondeur des profils de température et de chlorophylle qui varient en sens inverse reflète le même phénomène : les températures plus froides en profondeurs amènent des teneur en chlorophylle plus importante en surface. Cependant ce phénomène est moins fortement marqué que pour l'anti corrélation du premier axe. D'autre part, ce second axe principal correspond à l'opposition qui existe entre l'apport de nutriments par remonté d'eau froide (upwelling) dans la colonne d'eau et rayonnement solaire nécessaire à la photosynthèse qui provient de la surface. D'autre part, on voit le lien avec l'effet du vent de surface qui cause une homogénéisation de la colonne d'eau par un processus de mélange.

Il semble clair que les informations contenues dans la base de données doivent permettre de trouver un lien entre les variables de surface et celles de profondeur.

La variable de couverture nuageuse (CC) a été éliminée des données de surface utilisées. En effet elle est liée à la quantité de rayonnement à courtes ondes absorbée par la surface de l'océan, et ses effets sont inclus dans cette variable.

La Couverture Nuageuse (CC) apparaît au niveau de l'axe 2 et de l'axe 4. Cependant l'observation de la couverture nuageuse par satellite est peu fiable. Au moment des premières classifications avec les cartes topologiques, il est apparu que celle-ci était peu informative: les classes étant toujours très indifférenciées vis à vis de cette variable. Étant donné sa corrélation relativement faible et son faible pouvoir discriminant, nous l'avons écarté des variables utilisées pour la recherche des états observés.

Cette analyse nous a permis d'accepter l'inclusion dans la base de données d'apprentissage des états de la HMM des données relatives aux points voisins de BATS et de sélectionner pour l'apprentissage des cartes topologiques les 5 variables de surface (SST,SCHL,SSH,WS,SR).

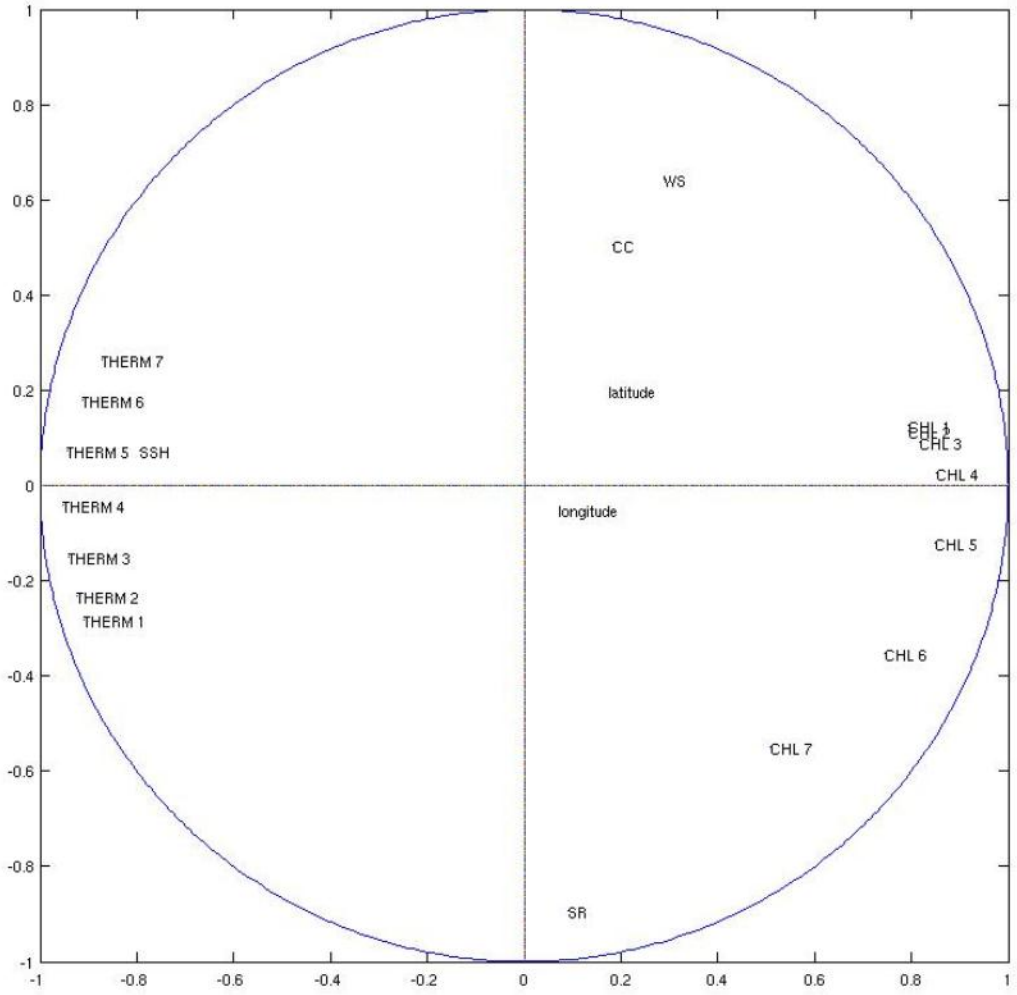


Figure 2.32 : Projection sur le premier plan factoriel des variables de l' Analyse en Composantes Principales

CHAPITRE 3 : INVERSION DE DONNEES SATELLITE POUR ESTIMER L'EVOLUTION SPATIALE DES PROFILS VERTICAUX DE TEMPERATURE SUR LE RAIL DE LA MISSION ARAMIS

3.1 Introduction :

Le premier article introduit la méthodologie générale utilisée par PROFHMM, il présente une application réelle qui porte sur la reconstitution temporelle des profils verticaux de Chlorophylle-A. La mise en œuvre de PROFHMM sur ce problème permet de voir qu'une chaîne de Markov cachée permet de synchroniser des données issues de modèles numériques avec des données d'observation satellitaires.

Le second article présente les résultats obtenus par l'application de PROFHMM pour reconstruire les données de la campagne ARAMIS à partir des données altimétriques AVISO et la température de surface fournie par la NOAA.

ARAMIS (Altimétrie sur un Rail Atlantique et Mesures In Situ - P.I.S. Arnault - http ://www.locean-ipsl.upmc.fr/aramis/) est un projet de recherche financé par le CNES (Centre National d'Etudes Spatiales), l'IRD (Institut de Recherche pour le Développement) et l'INSU (Institut National des Sciences de l'Univers) afin d'effectuer une surveillance des structures thermo halines des couches de surface océaniques en Atlantique tropical entre 2002 et 2008, via une ligne de bateaux marchands. Le principe du projet est de combiner des mesures in-situ facilement réalisables (sondes jetables eXpendable BathyThermograph XBT et eXpendable Conductivity Temprature Depth) et les mesures satellites afin de caractériser la variabilité à long terme de la circulation de surface en Atlantique tropical.

Cette seconde application permet de montrer l'aptitude de PROFHMM à retrouver la cohérence spatiale des profondeurs à partir de l'observation des variables de surface. Les performances obtenues montrent également que la méthode permet de synchroniser une dynamique océanique apprise à partir de donnée in-situ avec des données de surface.

L’application présentée s’est inspirée d’une première étude effectuée par Y. Tanguy durant ses travaux de thèse. En particulier le choix des variables de surface utilisées, qui avaient fait leurs preuves, sont celles utilisées dans son manuscrit.

3.2 ARTICLE 2: Retrieving the vertical profiles of temperature profiles along the ARAMIS rail from satellite observations, by using hidden Markov models and self-organizing maps.

Résumé : Nous appliquons la méthodologie d’inversion de donnée PROFHMM pour la reconstitution des profils verticaux de température issus de la mission ARAMIS. PROFHMM avait précédemment été utilisée pour reconstruire des séries temporelles de profils verticaux de concentration de Chlorophylle-A, à partir de données satellite. La méthodologie PROFHMM utilise des cartes topologiques auto-organisatrices pour modéliser des phénomènes géophysiques sous la forme de chaînes de Markov cachées. L’ application précédente de PROFHMM s’est basée sur des sorties du model numérique NEMO-PISCES pour apprendre ses profils verticaux et leur dynamique. En l’appliquant aux données ARAMIS nous démontrons la capacité de la méthodologie de reconstruire et d’apprendre la dynamique de profils verticaux de température à partir de données in-situ.

Retrieving vertical profiles of temperature profiles along the ARAMIS rail from satellite observations by using hidden Markov models and self-organizing maps.

A A CHARANTONIS[1], F BADRAN[2], S ARNAULT[2] S THIRIA[1]
[1] Laboratoire d'Océanographie et du Climat - Expérimentation et Approches Numériques, Université Pierre et Marie Curie - Tour 45, 5-ème étage 4, place Jussieu, 75005 Paris, France
[2] Laboratoire CEDRIC, Conservatoire National des Arts et Métiers - 292, rue Saint Martin, 75003 Paris, France

E-mail: anastase-alexandre.charantonis@locean-ipsl.upmc.fr

Abstract: We apply inverse method that uses Hidden Markov Models and Self-Organizing Maps in order to retrieve the vertical profiles of temperature of the ARAMIS mission from sea-surface data. This method, called PROFHMM, has been previously applied in order to reconstruct temporal series of vertical profiles of Chlorophyll-A concentration, based on sea-surface observations. PROFHMM makes use of Self-Organizing topological maps in order to model geophysical phenomena through Hidden Markov models. The vertical profiles on which the method was trained in the previous application were outputs of the NEMO-PISCES model. By applying it to the ARAMIS missions, we demonstrate the ability of the method to learn the ocean dynamics based on in-situ measurements and its applicability in the reconstruction of spatial transects.

1. Introduction

The prevalence of satellite missions monitoring the surface of the ocean has, during the last decades have been giving us access to an enormous database of past sea surface observations, as well as the possibility to access quasi real-time measurements [1,2,3]. However understanding, monitoring and study of the ocean dynamics require a good knowledge of its underlying structures that cannot be obtained directly from these satellite observations.

The acquisition of the vertical distribution of oceanic parameters is a complex task that usually requires in situ cruises or campaigns. Such campaigns provide us with high quality measurements of different oceanic parameters, but are limited in their spatial and temporal sampling.

Studies have proven the feasibility of connecting different sea-surface observations with the underlying distributions of oceanic parameters [4]. Such combination of sea-surface information and vertical distributions can help us generate more reliable databases containing the vertical distributions based on the existing data sets of observations.

The ARAMIS experiment [5] was a 6-year program, from 2002 to 2008 that recovered the vertical profiles of temperature and salinity from a merchant ship trade line. It gave us access to a large data set of vertical profiles of temperature and salinity. In the present paper we use this large ARAMIS in situ dataset, together with satellite information to retrieve the temperature profiles.

The feasibility of the inversion of the temperature distribution of the ARAMIS campaign from satellite imagery has been proven by Y.Tanguy [6]. In his method he linked the surface data with the underlying temperature profiles, acquired through both the ARGO floaters mission [7] and the ARAMIS campaign, by clustering them with the help of Self-Organizing Maps (SOM) [8]. This resulted in a larger learning data set (15760 available profiles, versus the 1222 profiles available solely through the ARAMIS campaign). The resulting clusters were used to retrieve the underlying distribution of temperature by considering that the average value of the temperature profile of each cluster corresponds to the average sea surface observation obtained for the same cluster of data. In his methodology, in cases for which the sea surface observations of many classes were similar, the determination of the correct underlying temperature profile was not evident, and a decision tree based on specific knowledge was introduced to facilitate the retrieval.

We propose to use the PROFHMM method [9] (PROfiles reconstruction through HMM) in order to exploit the ARAMIS database. PROFHMM is an inversion method that retrieves sequences of 3D distributions of different parameters based on observable 2D sequences of parameters, by making use of Hidden Markov Models and Self-Organizing Maps. It has previously been implemented in order to inverse satellite imaging to retrieve the temporal evolution of the vertical profiles of Chlorophyll-A. In that case the ocean dynamics were learned from the forcing and output of the NEMO numerical model under the PISCES configuration [10], while the sea-surface observations were taken from MODIS [11].

The advantage of applying PROFHMM to this particular problem is the unsupervised learning of the dynamics of the ocean and their links to the sea surface observations by using a Hidden Markov Model.

In this paper the PROFHMM methodology is applied to retrieve the vertical profiles of temperature over the ARAMIS transect by inverting sea-surface data. The ocean dynamics of the vertical profiles of temperature were estimated from the in-situ measurements of 12 of the ARAMIS missions, while the sea-surface observations were taken from the sea-surface data sets of Sea-Surface Temperature (SST) produced by NOAA and Absolute Dynamic Topography produced by Ssalto/Duacs.

2. PROFHMM METHODOLOGY

PROFHMM is an inverse method that retrieves sequences of 3D distributions of different parameters based on observable 2D sequences of parameters, by making use of Hidden Markov Models combined with Self-Organizing Maps.

A Markov model is a stochastic model that assumes the first order Markovian property, meaning that each consecutive state of the model depends solely on the previous state of the model such as $P(X_t| X_1 X_2 ... X_{t-1}) = P(X_t| X_{t-1})$, where X_t is the state of the model at the time t.

Expanding this principle, a Hidden Markov Model (HMM) is a stochastic model with two sequences: One sequence of hidden states that follows the first order Markovian property, and one sequence of observable states, that have a statistical link with the hidden states [12]. In order to model an HMM, the a priori knowledge of the transitions between the hidden states of the model and the probabilities of each observable state to have been emitted from a given hidden state must be found. The determination of these probabilities requires a data set containing concurrent sequences of hidden states and observable states.

If these a priori probabilities have been estimated, the Viterbi Algorithm [13], which is an algorithm often associated with HMMs, can find the most likely sequence of unobserved states, given a sequence of concurrent observations.

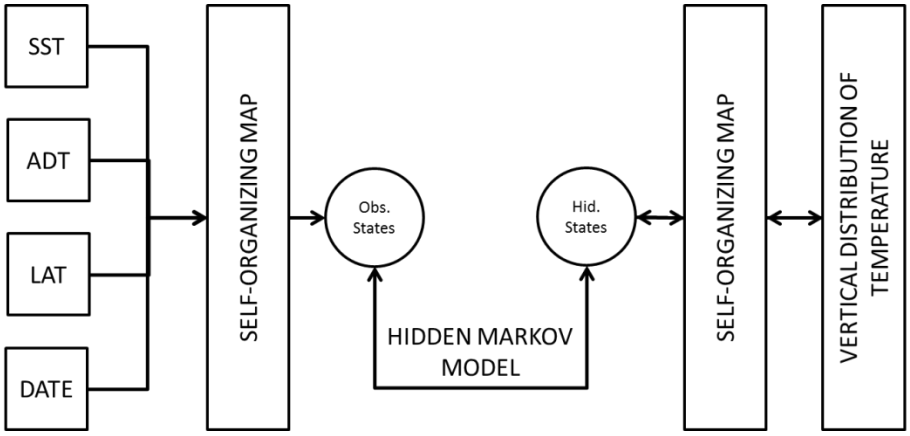


Figure 1: Flowchart relating the links between the different modules of PROFHMM and the observable and hidden vectors.

However the correct selection of states, and estimation of the a priori probabilities can be complicated. PROFHMM uses Self-Organizing Maps (SOM) in order to classify the available concurrent vertical distributions into the hidden states of the HMM and all observation vectors into the observable states. The SOM are unsupervised classification algorithms that cluster data into discrete classes. By applying SOM, we obtain a topological map containing states. Each state has a corresponding referent vector, whose value approximates the mean value of the data attributed to this class during the training. Thus, in PROFHMM, we have two maps which are used to generate sequences of indexes of the hidden and observable states used to train the HMM. These sequences are generated by attributing to each vector of a sequence the index of the state whose referent vector it “fits” the most.

When reconstructing a sequence of profiles of the vertical distribution of temperature, we first apply the Viterbi algorithm and obtain a sequence of indexes. We then consider that the value of the referent vector corresponding to the index obtain is the actual value of the vertical temperature distribution. We need to have finely discretized the available data, in order to get an accurate enough representation of the reality.

This would normally make the estimation of the a priori probabilities of the HMM harder, since we would observe less transitions and emissions per hidden state. However SOM dispatch the states on a 2 dimensional lattice according to their similarity, and PROFHMM makes use of this to improve the initial estimations of the HMM probabilities. A flowchart of the different interactions between the observations, vertical distributions and the methods employed in PROFHMM is shown in figure 1.

3. ARAMIS DATA

The vertical profiles of Temperature provided by the ARAMIS campaign, correspond to the hidden data of the HMM, while the satellite and auxiliary data correspond to the observable phenomena.

The ARAMIS project implemented a survey (2002-2008) of thermo-haline structures in the tropical Atlantic Ocean. Twice a year, once during the boreal spring and once in the fall, when the oceanic circulation is at its minimum and, maximum in the surface layers, respectively, expendable probes were launched along a merchant ship route, namely the line (AX11 : Santos, Brazil - Le Havre, France). This trade line spans from 20°S to 35°N on the tropical Atlantic Ocean. The different trajectories can be seen on figure 2, along with the point of intersection with ARGO profilers that were used to validate the measures or the ARAMIS mission [14,15] . The Temperature profiles were retrieved every 0,5° degrees of latitude.

Expendable BathyThermographs (XBTs) have become the most widely used method of retrieving in-situ temperatures for embarked missions since their use in the Tropical Ocean Global Ocean (TOGA) and World Ocean Circulation Experiment (WOCE) programs. Many studies dealing with thermal structures in the oceanic upper layers are based on these data sets and they are now commonly assimilated in operational global circulation models [14]. It is generally assumed that they provide temperature (T) profiles (0-800m) with an accuracy about +/- 0,1°C.

Table 1: Details on probes launched during the different ARAMIS missions

ARAMIS NUMBER	ARAMIS DATES (YYYY/MM/DD)	XBTs launched	XBTs lost	XBTs retained	XCTDs launched	XCTDs lost	XCTDs retained
ARAMIS1	2002/07/20–2002/07/28	60	3	57	48	2	45
ARAMIS2	2003/03/15–2003/03/23	60	5	54	51	2	48
ARAMIS3	2003/10/12–2003/10/20	59	0	57	56	8	48
ARAMIS4	2004/05/02–2004/05/11	55	4	51	49	0	49
ARAMIS5	2004/09/12–2004/09/20	59	0	58	52	3	48
ARAMIS6	2005/04/27–2005/05/05	55	0	54	49	0	49
ARAMIS7	2005/10/05–2005/10/13	54	0	53	44	1	42
ARAMIS8	2006/05/04–2006/05/17	54	2	52	49	0	49
ARAMIS9	2006/10/23–2006/10/29	60	0	59	48	0	48
ARAMIS10	2007/04/22–2007/04/29	58	4	54	53	5	47
ARAMIS11	2007/09/24–2007/09/30	52	0	51	51	2	49
ARAMIS12	2008/04/24–2008/05/02	53	0	53	48	0	47

Since the primary aim of the ARAMIS mission was to monitor thermo-haline patterns, they also made use of expendable salinity-temperature profilers (XCTD for eXpendable Conductivity-Temperature Depth). Their accuracy has been checked by different studies and determined to be of +/- 0,2°C (Miyake et al., 1981; Johnson, 1995; Alberola et al., 1996; Arnault et al., 2004a).

ARAMIS XBTs and XCTDs were carefully checked to eliminate spurious data. Less than 10% of the data were rejected due to quality constraints. In table 1 we can see the number of XBTs and XCTDs comprised in the ARAMIS data set.

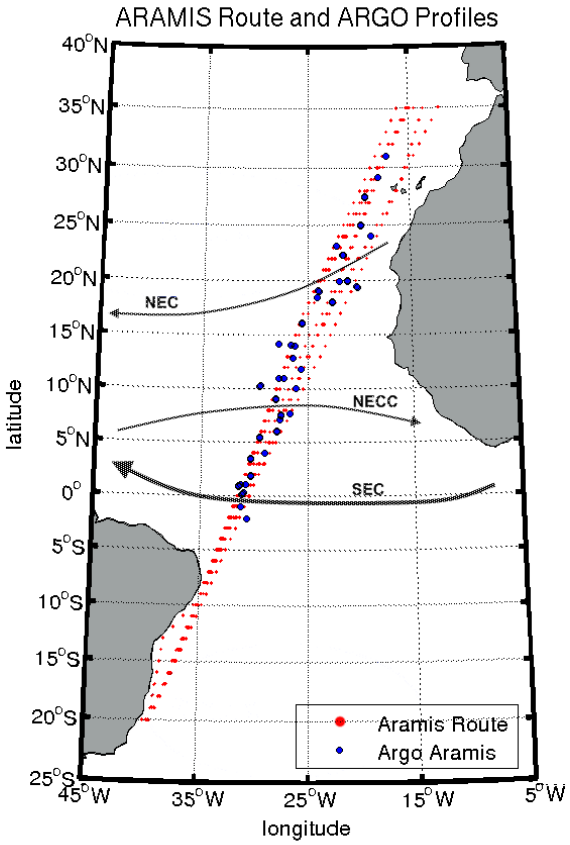


Figure 2: The trajectories of the ARAMIS missions.

The connection between the sea-surface phenomena and the underlying vertical distribution however is limited to the first layers of the ocean, since there exist many internal ocean processes that are unobservable from the sea-surface. In this application of PROFHMM we will work under the assumption that there is a link up to 300 meters between the surface and the vertical distribution of temperature, although this statement is not always true. The measurements of Temperature were taken every meter from 5 to 305 meters below the surface and were stocked in the $Data_{prof}$ matrix.

4. OBSERVABLE DATA

The observable data consisted of the sea surface Satellite observations and auxiliary data.

Each observation vector was taken at the same geographic and temporal coordinates as its corresponding hidden vector and contains the following parameters:

- SST - Weekly Sea Surface Temperature
- ADT - Absolute Dynamic Topography delayed time product
- Latitude
- Date: The sin(pi *month/12) and cos(pi*month/12)

The Sea-Surface Temperature was produced by NOAA's optimally interpolated satellite and in-situ dataset. The values were linearly interpolated from their 1° spatial resolution to 0.5°, over the ARAMIS rail.

The altimeter products were produced by Ssalto/Duacs and distributed by Aviso, with support from CNES [16]. They come from up-to-date datasets with up to four satellites at a given time (Jason-2 / Jason-1 / Envisat from 2009 or between October 2002 and September 2005, the association Jason-1 / Topex/Poseidon / Envisat / GFO). The ADT is important in order to inform on the potential upwelling/downwelling and water turbidity. The ADT product has a spatial resolution of 1/3° and therefore each point takes the value of the closest grid point of the ARAMIS rail.

The latitude is included here as a geographical indicator of the observable state. The inclusion of solely the latitude, instead of a combination of latitude and longitude, is due to the high regularity of the merchant ship route, which makes the longitude a function of the latitude for this given transect. The inclusion of the longitude therefore would not provide any additional information.

The Date is included as two variables: the cosine and sine of the month times pi/6, in order to have the same numerical distance between each consecutive month and prevent an abnormally large distance between December and January.

They observation vectors (SST, ADT, latitude, sin(month*pi/6), cos(month*pi/6)) were stored in the $Data_{obs}$ matrix and each value corresponds spatially and temporally with the respective ARAMIS in situ measurement stored in the $Data_{prof}$ matrix.

These surface parameters used were selected based on a prior research of Y.Tanguy [6].

5. PROFILE RECONSTRUCTIONS WITH PROFHMM.

In order to reconstruct the temperature profiles with PROFHMM, we trained two SOMs. The first one, named $sMap_{prof}$, was trained with all of the profile vectors of $Data_{prof}$, giving us 1222 vertical distributions of temperature retrieved during the ARAMIS missions 1 to 12.

$sMap_{prof}$ discretized the $Data_{prof}$ data set into 600 hidden states dispatched into the 30x20 states of the topological map. Each $sMap_{prof}$ state is associated with a referent vector, ref_{prof} containing the vertical profile of temperature.

The second one, named $sMap_{obs}$, was trained with $Data_{obs}$. $sMap_{obs}$ discretized the observations space into 651 observable states arranged into the 31x21 states of the topological map.

This number of states was selected after performing tests on the number of states of $sMap_{prof}$ and $sMap_{obs}$. The number of states was independently incremented for each map. For each combination of $sMap_{prof}$ and $sMap_{obs}$, the overall performance of PROFHMM was estimated by calculating on the HMM learning data set the root mean square errors (RMS). We kept the combination of the smallest number of states in each map that gave us the smallest RMS. The size selected (600 for $sMap_{prof}$ and 651 for $sMap_{obs}$) can be seen as a compromise between overlearning the data sets and maintaining a high level of discretization.

In order to verify that the referent vectors provided by the SOM maps were representative of the data, we performed a principal component analysis (PCA, [17]) of $Data_{prof}$ and $Data_{obs}$, and projected the data vectors and the referent vectors on the first plane of the analysis.

In figure 3 (a) we present, the projection on the first plane given by the PCA of $Data_{prof}$, of the referent vectors of the hidden states as red circles, and the data vectors of $Data_{prof}$ as blue crosses. Similarly, in figure 3 (b) we present the projection on the first plane of the PCA of $Data_{obs}$, of the referent vectors of the hidden states as red circles, and the data vectors from $Data_{obs}$ as blue crosses. The first plane of the PCA of the hidden data set corresponds to 98,3% of its variance, while the first plane of the PCA of the observable data set corresponds to 87,0% of its variance. Both hidden and observable states are well distributed over their respective data set. Therefore we can make the assumption that the selected states represent accurately the variance of the observed phenomenon.

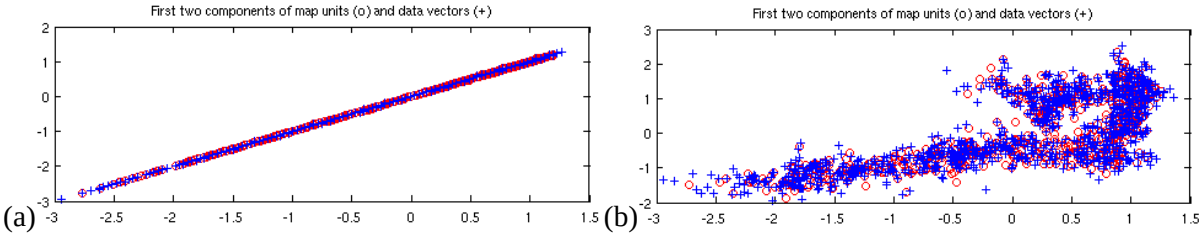


Figure 3: (a) Projection of the temperature profiles (in blue crosses) and the referent vectors of $sMap_{prof}$ (in red circles), onto the first plane of the PCA of $Data_{prof}$. (b) Respectively, projection of the observation vectors (in blue crosses) and the referent vectors of $sMap_{obs}$ (in red circles), onto the plane determined by the two first eigenvectors of the PCA of $Data_{obs}$.

Once the SOM maps were learned, we labeled each hidden and observable data vector with the corresponding index of the state whose referent vector is the closest to it.

We randomly excluded 122 data vectors (10% of the data) to use as a Validation set. The total number of profiles in each mission, the number of profiles in the HMM training and the number of validation profiles are shown in Table 2. The removal of these 122 vectors from the training database, along with the omission of the 35 incomplete temperature profiles in the training of the HMM provided us with 159 sequences of indexes. The Boreal spring ARAMIS missions, (highlighted in gray in Table 2) used were better represented in the training of the HMM while the Boreal fall ARAMIS missions were less represented and had more of their profiles included in the validation set.

We sort the remaining 1065 indexes obtained with $sMap_{prof}$ and the corresponding 1065 indexes obtained with $sMap_{obs}$ into 159 sequences of consecutive data, oriented from north to south. The initial probabilities of emission and transition are calculated from these sequences. We then use the topology aspect of the SOM states to improve them. This is discussed in more detail in [9].

Table 2: Profiles used during Training and for Validation

MISSION Number	Total Profiles	HMM training	Validation Profiles	MISSION Number	Total Profiles	HMM training	Validation Profiles
ARAMIS 1	102	96	6	ARAMIS 7	95	90	5
ARAMIS 2	102	93	9	ARAMIS 8	101	100	1
ARAMIS 3	105	94	11	ARAMIS9	107	72	35
ARAMIS 4	100	93	7	ARAMIS 10	101	99	2
ARAMIS 5	106	101	5	ARAMIS 11	100	71	29
ARAMIS 6	103	91	12	ARAMIS12	100	100	0
Boreal Spring	607	576	31	Boreal Fall	513	428	85

6. RESULTS

In the following section we present the results of the reconstruction of the vertical distribution of the temperature using sea-surface data by applying PROFHMM.

It is important to note that the errors presented in the reconstructions are often due to the rough discretization given by the SOMs. When reconstructing the sequence of profiles for an ARAMIS campaign, we select the referent vector corresponding to the index given by the Viterbi Algorithm. Since we finely tuned our SOMs, the referent vectors approximate the mean value of the data vectors attributed to that state during the training of the SOMs [8]. Since we reconstruct the temperature profile from this average value, the reconstructed value does not always completely fit the actual in-situ profile.

6.1 RMS tables

In order to estimate the precision of the reconstruction obtained through PROFHMM, we segmented the 300 meters of the profile into ten 30-meters segments. We considered two types of quantitative errors over these segments, and compute them for each segment of each mission. For each segment:

- The first error consists of calculating, the root mean square error (RMSE) between the average values of temperature and the average values obtained with PROFHMM.
- The second error consists of calculating the RMSE point by point between all temperature values of the segment and their reconstructions.

For each ARAMIS mission these errors are averaged over each latitude in the data set. The number of latitudes with profiles used by each mission is noted Nb_lat. We denote these errors mRMSE (1) and RMSE (2).

The mRMSE were calculated as:

$$mRMSE = \sqrt{\frac{1}{Nb\_lat}\sum_{lat}\left(\frac{1}{30}\left(\sum_{depth} Y_{ARAMIS}\right) - \frac{1}{30}\left(\sum_{depth} Y_{PROFHMM}\right)\right)^2} \text{ (1),}$$

While the RMSE were calculated as:

$$RMSE = \sqrt{\frac{1}{Nb\_lat}\sum_{lat}\frac{1}{30}\left(\sum_{depth}\left(Y_{ARAMIS} - Y_{PROFHMM}\right)^2\right)} \text{ (2).}$$

We calculated these errors twice:

- When using the total available data for each mission: noted mRMSE_all and RMSE_all
- When considering only the validation data set: noted mRMSE_val and RMSE_val

The mRMSE_all values are shown in table 3, while the mRMSE_val are shown in table 4. We chose to highlight in blue the values of table 3 that are over 0,1°C and the values of table 4 that are over 0,5°C, as the size of the validation set is small.

The RMSE_all values are shown in table 5, while the RMSE_val are shown in table 6. We chose to highlight in green the values of table 5 that are over 0,5°C and the values of table 6 that are over 1°C.

The RMSE thresholds are higher due to the point by point comparaison.

mRMSE Performances

When evaluating the results obtained, it is important to note that the ARAMIS missions were not equally represented in the training and validation sets, as seen in table 2.

The ARAMIS 1 mission was the only summer mission which means that the transition and emission probabilities of the states linked to this mission were very poorly learned.

The six Boreal spring missions (2, 4, 6, 8, 10, 12) were better represented in the training data set than the 5 Boreal fall ones (3, 5, 7, 9, 11). This was further accentuated by the random selection made for the determination of our validation set, which contained more values of the Boreal fall missions than of the Boreal spring ones. This was especially true in the case of the ARAMIS 9 and 11 missions with 35 and 29 vectors missing.

In the following tables, the boreal spring missions numbers have been highlighted in, while the Boreal fall missions numbers have been highlighted in purple.

We can observe that mRMSE_all values are generally below the margin of errors of +/- 0,1°C for the XBTs and +/-0,2°C for the XCTDs measurements. The ARAMIS 9 mission, part of the Boreal fall missions presents the highest values, but remains close to the in-situ precision values. We chose to highlight in blue the values of table 3 that are over 0,1°C and the values of table 4 that are over 0,5°C, as the size of the validation set is small.

Table 3: mRMSE_all by depth. In gray the Boreal spring missions, in purple the Boreal fall ones.

ARAMIS NUMBER	5 - 35m	35 - 65m	65 - 95m	95 - 125m	125 - 155m	155 - 185m	185 - 215m	215 - 245m	245 - 275m	275 - 305m	Average
1	0,004	0,011	0,011	0,016	0,015	0,014	0,088	0,009	0,001	0,035	0,020
2	0,072	0,060	0,055	0,050	0,040	0,059	0,068	0,118	0,034	0,048	0,060
3	0,073	0,071	0,038	0,053	0,097	0,135	0,055	0,046	0,000	0,085	0,065
4	0,023	0,031	0,044	0,052	0,036	0,087	0,073	0,019	0,011	0,100	0,048
5	0,003	0,022	0,068	0,102	0,074	0,032	0,029	0,010	0,010	0,020	0,037
6	0,041	0,011	0,012	0,024	0,064	0,101	0,100	0,065	0,088	0,114	0,062
7	0,061	0,058	0,021	0,017	0,048	0,006	0,068	0,007	0,021	0,003	0,031
8	0,020	0,009	0,014	0,021	0,009	0,021	0,050	0,021	0,026	0,040	0,023
9	0,096	0,120	0,146	0,130	0,101	0,192	0,273	0,198	0,009	0,107	0,137
10	0,006	0,000	0,015	0,013	0,027	0,055	0,011	0,038	0,009	0,009	0,018
11	0,063	0,040	0,041	0,007	0,054	0,035	0,027	0,087	0,166	0,107	0,063
12	0,009	0,010	0,006	0,019	0,004	0,025	0,011	0,019	0,004	0,057	0,016
Average	0,039	0,037	0,039	0,042	0,047	0,063	0,071	0,053	0,032	0,061	0,048

Table 4: mRMSE_val by depth. In gray the Boreal spring missions, in purple the Boreal fall ones.

ARAMIS NUMBER	5 - 35m	35 - 65m	65 - 95m	95 - 125m	125 - 155m	155 - 185m	185 - 215m	215 - 245m	245 - 275m	275 - 305m	AVERAGE	Validation Profiles
1	0,893	0,781	0,727	1,022	0,822	0,375	1,229	1,146	0,333	0,45	0,778	6
2	0,143	0,141	0,300	0,381	0,303	0,419	0,161	0,464	0,046	0,059	0,242	9
3	0,567	0,636	0,509	0,473	0,676	0,940	0,933	0,898	0,520	0,073	0,623	11
4	0,138	0,249	0,283	0,150	0,454	0,578	0,589	0,759	0,351	0,142	0,369	7
5	0,452	0,592	0,956	1,201	0,996	0,218	0,092	0,235	0,447	0,076	0,526	5
6	0,262	0,080	0,045	0,202	0,145	0,293	0,394	0,194	0,451	0,433	0,250	12
7	0,812	0,842	0,604	0,761	1,239	1,959	1,797	0,866	0,933	0,976	1,079	5
8	0,769	0,762	1,038	0,971	0,437	0,998	0,757	0,560	0,349	0,408	0,705	1
9	0,417	0,415	0,470	0,425	0,270	0,495	0,580	0,465	0,157	0,180	0,387	35
10	0,412	0,598	0,685	0,724	0,628	0,047	0,038	0,185	0,193	0,126	0,363	2
11	0,129	0,141	0,181	0,077	0,057	0,043	0,068	0,051	0,388	0,198	0,133	29
12	-	-	-	-	-	-	-	-	-	-	-	0
Average	0,354	0,356	0,386	0,390	0,361	0,446	0,496	0,432	0,331	0,231	0,378	

We can note that the average mRMSE_all values progressively increase after the first 95 meters in the ocean, when encountering the high-variability thermocline zone, then decrease when reaching relatively quiescent water layers.

The mRMSE_val values are higher than mRMSE_all ones. However these performances remain in acceptable ranges. Some of the values obtained, such as the errors in the Campaign 8, were values obtained by averaging a very small set of profiles.

We can notice that the mRMSE_val values of the ARAMIS 1 summer mission have higher error values. These are due to a great degree to the under-representation in the training data set of the transition and emission probabilities related to the typical summer hidden states.

Similarly we can notice that the ARAMIS missions held during the Boreal spring have generally slighter errors than those happening during the fall. This can be attributed to the better representation of the Boreal spring missions in the training data set.

RMSE Performances

The values obtained with RMSE are higher than those obtained with mRMSE. The results obtained constitute a finer, but stricter comparison between the in-situ and retrieved profiles. Every vertical shift of temperature by a meter or more generates cumulative errors that are filtered out when averaging with mRMSE. Therefore the RMSE is a very demanding performance test.

The RMSE_all values are shown in table 5, while the RMSE_val are shown in table 6. We chose to highlight in green the values of table 5 that are over 0,5°C and the values of table 6 that are over 1°C.

Table 5: RMSE_all by depth. In gray the Boreal spring missions, in purple the Boreal fall ones.

ARAMIS NUMBER	5 - 35m	35 - 65m	65 - 95m	95 - 125m	125 - 155m	155 - 185m	185 - 215m	215 - 245m	245 - 275m	275 - 305m	Average
1	0,431	0,378	0,399	0,498	0,453	0,417	0,508	0,708	0,414	0,432	0,471
2	0,318	0,347	0,392	0,430	0,447	0,623	0,772	0,931	0,582	0,245	0,509
3	0,416	0,387	0,408	0,439	0,503	0,592	0,658	0,699	0,739	0,65	0,549
4	0,385	0,368	0,435	0,557	0,623	0,573	0,650	0,835	0,635	0,283	0,534
5	0,252	0,272	0,349	0,411	0,451	0,412	0,587	0,395	0,525	0,346	0,400
6	0,377	0,433	0,542	0,727	0,969	0,989	0,765	0,77	0,577	0,369	0,652
7	0,402	0,437	0,422	0,474	0,623	0,938	1,042	0,846	0,923	0,846	0,695
8	0,270	0,228	0,251	0,293	0,346	0,397	0,448	0,491	0,408	0,269	0,340
9	0,515	0,519	0,609	0,674	0,643	0,759	0,877	0,886	0,797	0,404	0,668
10	0,271	0,288	0,317	0,411	0,459	0,493	0,439	0,459	0,347	0,246	0,373
11	0,459	0,430	0,535	0,588	0,725	0,796	0,913	0,732	0,818	0,659	0,665
12	0,186	0,194	0,209	0,204	0,221	0,305	0,319	0,409	0,399	0,271	0,272
Average	0,357	0,357	0,406	0,475	0,539	0,608	0,671	0,680	0,597	0,418	0,511

Table 6: RMSE_val . In gray the Boreal spring missions, in purple the Boreal fall ones.

ARAMIS NUMBER	5 - 35m	35 - 65m	65 - 95m	95 - 125m	125 - 155m	155 - 185m	185 - 215m	215 - 245m	245 - 275m	275 - 305m	Average	Validation Profiles
1	1,300	1,108	1,108	1,373	1,263	1,212	1,841	2,281	0,874	1,004	1,336	6
2	0,428	0,563	0,859	0,858	0,655	0,863	1,053	1,317	0,924	0,223	0,774	9
3	0,681	0,706	0,607	0,622	1,039	1,235	1,243	1,483	1,541	0,646	0,98	11
4	0,658	0,621	0,716	0,871	0,84	0,889	1,234	1,720	1,334	0,370	0,925	7
5	0,808	0,984	1,339	1,581	1,412	0,905	0,501	0,678	1,357	0,323	0,989	5
6	0,597	0,751	0,973	1,36	1,968	1,909	1,479	1,187	0,916	0,735	1,188	12
7	0,936	0,987	0,866	1,264	1,948	2,535	2,147	1,529	2,367	2,894	1,747	5
8	0,775	0,766	1,040	0,985	0,512	1,048	0,802	0,571	0,386	0,425	0,731	1
9	0,838	0,863	1,020	1,114	1,003	1,094	1,259	1,367	1,209	0,526	1,029	35
10	0,545	0,662	0,732	0,836	0,898	0,993	0,206	0,310	0,328	0,248	0,576	2
11	0,763	0,743	0,953	1,029	1,260	1,244	1,328	1,233	1,408	1,043	1,100	29
12	-	-	-	-	-	-	-	-	-	-	-	0
Average	0,762	0,791	0,939	1,073	1,190	1,249	1,292	1,344	1,259	0,756	1,065	

The errors presented are mainly due to two sources:

- The classification errors introduced with the referent vectors of $sMap_{prof}$
- The "dynamic reconstruction" errors introduced by the Viterbi Algorithm

Classification errors introduced when reconstructing with the referent vectors of $sMap_{prof}$

As mentioned in the beginning of section 6, the SOM clustering enables us to represent each state with a referent vector that is an approximation of the mean value of the data vectors attributed to that state during

the training phase. Therefore, when reconstructing a profile of temperature by attributing the value of this referent vector, we might have a slight differences leading to a shift in the temperature profiles compared to the in-situ data, which account for these higher error values.

"Dynamic reconstruction" errors introduced by the Viterbi Algorithm.

The Viterbi algorithm reconstructs the hidden states by maximizing the probability of the total transect, and it can sometimes reconstruct some not-best fitting states in order to allow a smoother transition between some situations further down in the transect. It can, therefore, at some latitudes, give a profile which does not fit well the in-situ measurements, in order to improve the global reconstruction.

For example, if a transition between two states has never been observed in the training data, by applying the Viterbi Algorithm the probability of retrieving the correct temperature profile on the validation data set would be null, even if that profile is well represented by a referent vector of $sMap_{prof}$. The modifications to the transition and emission probabilities of the HMM given by PROFHMM use the topological aspect of the self-organizing maps which increases the number of profiles considered when applying the Viterbi Algorithm. However this modification cannot be fully efficient in cases for which specific transitions or emissions have never been observed over large sections of the Self Organizing Map.

The Boreal fall ARAMIS missions present the higher RMSE both for RMSE_all and RMSe_val. We can also note that tendency of the RMSE to progressively increase while descending in the ocean layers persists.

Year to year variability

Since the ARAMIS missions were roughly separated in spring and fall missions, by averaging the profiles latitude by latitude for the 12 missions, we obtain a first order approximation of the mean seasonal state of the temperature profiles. We refer to this mean value as the "average year". By removing the average year value we can observe the year to year variations in the ARAMIS missions. Similarly if we average the reconstructions of the 12 campaigns obtained by inverting the sea surface data with PROFHMM, we obtain the "average reconstructed year". By removing it from the reconstructions we can observe whether our methodology correctly recovers the anomalies compares to the average year. This average year however was calculated only in the interval ranging from 13°N to 30°N in order to have a significant amount of data in order to perform a meaningful average.

In table 6 we present the mRMSE obtained for each mission by comparing the average distribution between all the available data of each ARAMIS campaign and their corresponding reconstruction, while having respectively subtracted the "average yearly" and "average reconstructed year" from the data. For a given segment the errors are computed as follows (3):

$$mRMSE\_av.year = \sqrt{\frac{1}{Nb\_lat}\sum_{lat}\left(\frac{1}{30}\left(\sum_{depth}(Y_{ARAMIS} - \overline{Y_{ARAMIS}})\right) - \frac{1}{30}\left(\sum_{depth}(Y_{PROFHMM} - \overline{Y_{PROFHMM}})\right)^2\right)} \quad (3).$$

In table 7, we present, respectively the RMSE obtained by removing the "average year" and "average reconstructed year", and considering the error segment by segment. For a given segment the errors are computed as follows (4):

$$RMSE\_av.year = \sqrt{\frac{1}{Nb\_lat}\sum_{lat}\frac{1}{30}\left(\sum_{depth}(Y_{ARAMIS} - \overline{Y_{ARAMIS}}) - (Y_{PROFHMM} - \overline{Y_{PROFHMM}})\right)^2} \quad (4).$$

Table 7: mRMSE_av.year by depth. In gray the Boreal spring missions, in purple the Boreal fall ones. In blue the values superior to 0.1°C.

ARAMIS NUMBER	5 - 35m	35 - 65m	65 - 95m	95 - 125m	125 - 155m	155 - 185m	185 - 215m	215 - 245m	245 - 275m	275 - 305m	Average
1	0,026	0,041	0,037	0,045	0,044	0,016	0,098	0,002	0,007	0,042	0,036
2	0,050	0,05	0,056	0,047	0,036	0,051	0,076	0,108	0,051	0,063	0,059
3	0,061	0,079	0,039	0,041	0,076	0,109	0,033	0,047	0,014	0,089	0,059
4	0,028	0,034	0,034	0,027	0,005	0,048	0,031	0,003	0,021	0,094	0,033
5	0,015	0,035	0,072	0,095	0,091	0,077	0,041	0,030	0,010	0,008	0,047
6	0,035	0,011	0,055	0,078	0,112	0,156	0,128	0,081	0,113	0,120	0,089
7	0,076	0,076	0,031	0,030	0,059	0,004	0,087	0,030	0,032	0,006	0,043
8	0,036	0,020	0,005	0,000	0,005	0,003	0,032	0,008	0,029	0,039	0,018
9	0,092	0,122	0,136	0,142	0,126	0,203	0,259	0,195	0,009	0,139	0,142
10	0,001	0,003	0,007	0,006	0,037	0,069	0,008	0,038	0,006	0,006	0,018
11	0,078	0,046	0,054	0,025	0,003	0,031	0,023	0,065	0,155	0,083	0,056
12	0,012	0,013	0,006	0,006	0,006	0,004	0,022	0,003	0,004	0,063	0,014
Average	0,043	0,044	0,044	0,045	0,050	0,064	0,070	0,051	0,038	0,063	0,051

Table 8: mRMSE_av.year by depth. In gray the Boreal spring missions, in purple the Boreal fall ones. In green the values superior to 0.5°C.

ARAMIS NUMBER	5 - 35m	35 - 65m	65 - 95m	95 - 125m	125 - 155m	155 - 185m	185 - 215m	215 - 245m	245 - 275m	275 - 305m	Average
1	0,426	0,363	0,372	0,451	0,430	0,430	0,608	0,754	0,433	0,444	0,471
2	0,287	0,327	0,368	0,406	0,454	0,646	0,831	0,998	0,613	0,218	0,515
3	0,419	0,398	0,426	0,452	0,513	0,605	0,698	0,75	0,788	0,697	0,575
4	0,339	0,270	0,314	0,412	0,466	0,485	0,609	0,848	0,644	0,284	0,467
5	0,264	0,286	0,343	0,380	0,424	0,410	0,629	0,404	0,456	0,364	0,396
6	0,338	0,383	0,482	0,665	0,971	1,047	0,823	0,815	0,615	0,378	0,652
7	0,408	0,446	0,435	0,484	0,639	0,968	1,074	0,849	0,939	0,874	0,712
8	0,256	0,215	0,235	0,273	0,333	0,406	0,466	0,487	0,402	0,279	0,335
9	0,529	0,518	0,582	0,633	0,58	0,806	0,942	0,956	0,864	0,423	0,683
10	0,278	0,300	0,328	0,421	0,464	0,512	0,458	0,477	0,341	0,227	0,381
11	0,287	0,258	0,282	0,289	0,363	0,656	0,913	0,738	0,857	0,681	0,533
12	0,181	0,187	0,209	0,181	0,206	0,296	0,312	0,414	0,405	0,280	0,267
Average	0,334	0,329	0,365	0,421	0,487	0,606	0,697	0,707	0,613	0,429	0,499

The results presented in tables 7 and 8 remain comparable to those obtained in table 3 and 5, with some slight ameliorations. The values in table 7 that are larger than 0,1°C were highlighted in blue, and the values in table 8 larger than to 0,5°C were highlighted in green. The ARAMIS 9, which was the least represented mission during the training continues to present the highest mRMSE and RMSE values.

6.2 Campaigns

We present now in more detail the reconstructions obtained over 3 ARAMIS missions. We chose to present the ARAMIS 1, 2 and 5 missions, in order to have one Boreal summer, spring and fall mission.

The first and last three 0,5° latitude steps of each reconstruction are omitted in the results presented as they are considered to be less reliable due to the way the Viterbi algorithm works.

ARAMIS 1

The ARAMIS 1 mission took place from the 20th to the 28th of July, 2002 and is the only mission done during the Boreal summer (Austral winter). The white columns in figure 4 (a) represent the missing profiles of temperature in the training data set, and separate the mission in 8 sequences that were used to learn the transition and emission probabilities of the HMM. The training data set contained 96 profiles while the validation set contained 6. The complete data is present in figure 4 (b) which represents the distributions of temperature obtained during the ARAMIS 1 mission while the reconstruction by PROFHMM is seen in figure 4 (c).

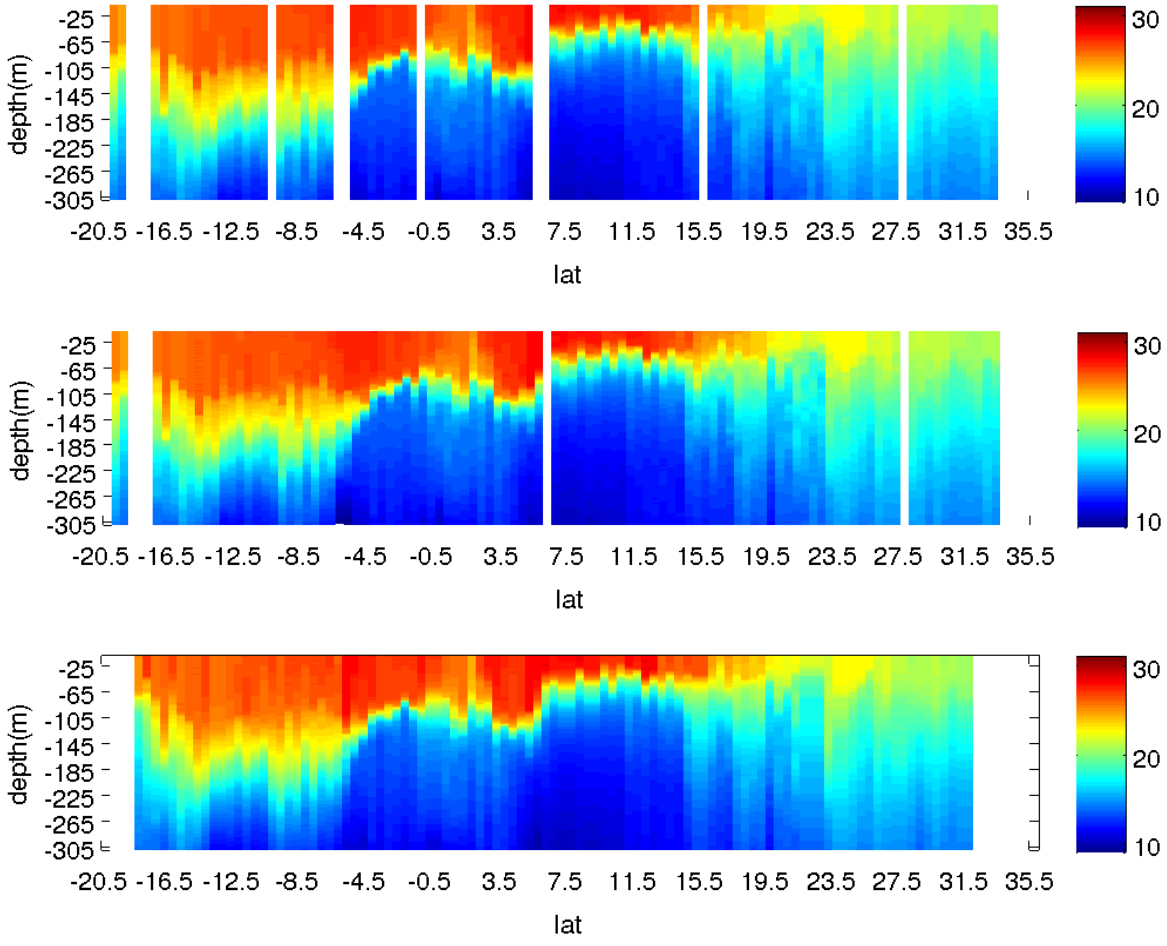
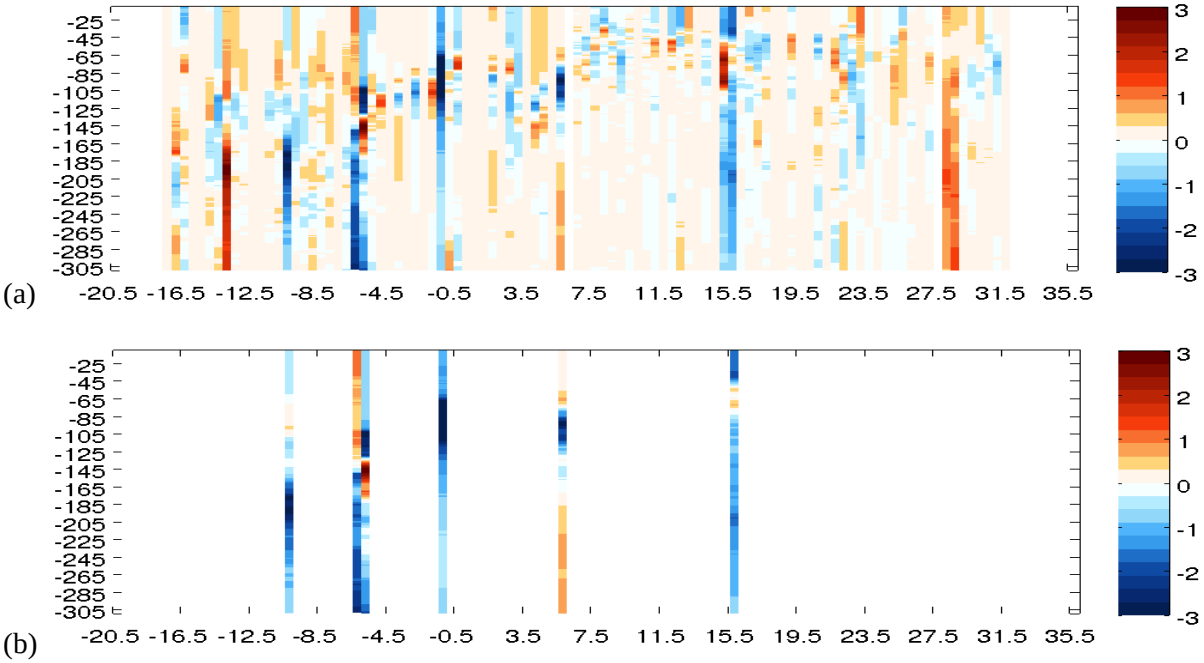


Figure 4: Vertical distribution of the validation data set of the ARAMIS 1 mission for every available latitude where such a profile was provided in the training data set (a), the complete ARAMIS 1 MISSION profiles (b) and the reconstruction of the ARAMIS rail based solely on surface data by the PROFHMM method (c). The colorbars are in °C.

During this period in the South Atlantic ocean there is a slight cooling of the surface layers, while in the Northern latitudes the surface layers are beginning to heat up. At the level of the equator we can observe an upwelling that slightly moves towards South. At 5°N there is the beginning of the hot equatorial counter-current structure. We can also note the almost constant vertical distribution of temperature in the higher northern latitudes, while in the latitudes from 5°S to 15°N we observe an equatorial 2-layer system. As seen in figure 4 (c) these phenomena are correctly reconstructed by PROFHMM.

The difference between the real values and the reconstructed ones is seen in figure 5 (a), while in figure 5 (b) we focus on the reconstruction of the values of ARAMIS 1 mission present in the validation data set. In figure 5 (a) we can observe that about 90% of the vectors reconstructed from the sea-surface images do not present significant differences with the in-situ measures, with the majority of the image being in white and corresponding to the error margins of the in-situ measuring instruments, as seen in table 4. The differences are well marked in the thermocline zone, where the variability of temperature is the strongest. This difference in values is mostly due to the quantification errors that stem from using the average values of a clustering method when reconstructing the temperature profiles.

In figure 5 (b) we can observe that the 6 profiles reconstructed show high values. However since the PROFHMM method was largely trained based with data from the Boreal spring and fall missions, some transitions were never encountered in the training dataset.



(a)

(b)

Figure 5: (a) Difference in values between the reconstruction and the complete ARAMIS 1 mission profiles. (b) A zoom on the errors over the values present in the validation set. The colorbar values are in °C. The vertical axis corresponds to the depth in meters while the horizontal one corresponds to the latitude over the ARAMIS transect.

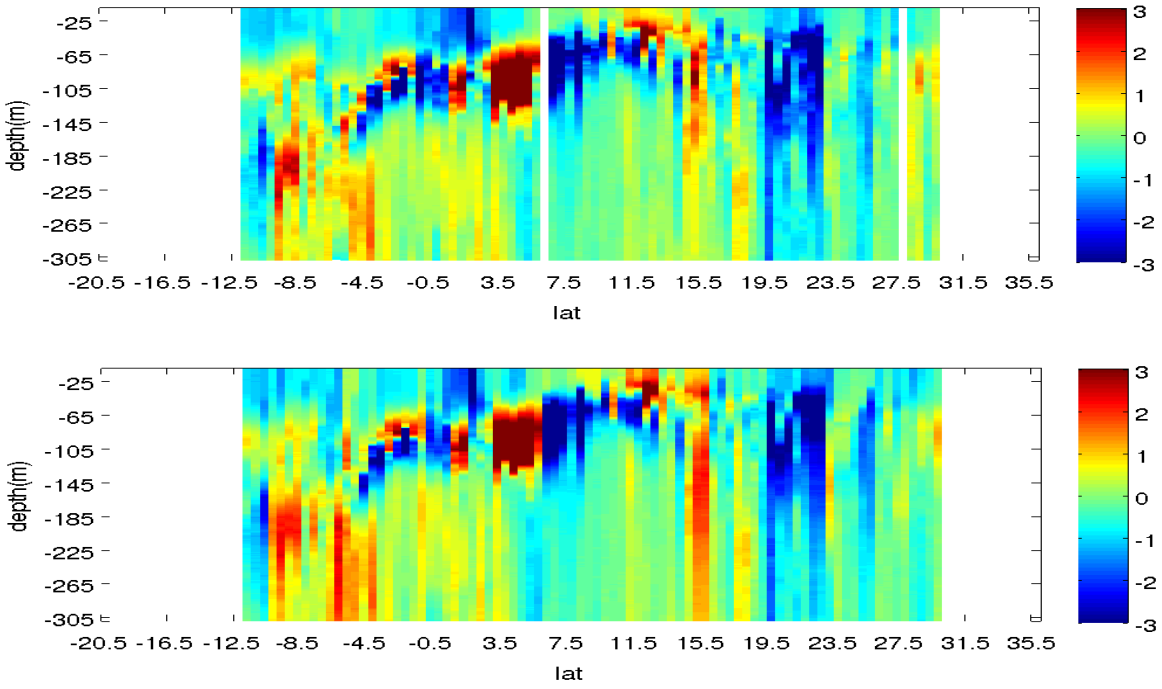


Figure 6: (a) Complete data set minus the "average year" (b) Reconstruction minus the "average reconstructed year". The colorbar values are in °C.

In figure 6 (a) and 6 (b) we present, respectively, the ARAMIS 1 mission minus the "average year" and the PROFHMM ARAMIS 1 mission reconstruction based on sea-surface observations minus the "average reconstructed year".

We can notice that the strong hot and cold anomaly present between 3,5°N and 7,5°N, and the barotropic anomaly in the 20°N neighborhood are present in both figures. The general form of the anomalies are retrieved both in amplitude and in position by the PROFHMM reconstruction.

ARAMIS 2

The ARAMIS 2 mission took place from the 15th to the 23rd of March, 2003. It is the first mission occurring during the Boreal spring (Austral fall). The white columns in figure 7 (a) represent the missing profiles of temperature in the training data set, and separate the mission in 8 sequences that were used to learn the transition and emission probabilities of the HMM. The training data set contained 93 profiles while the validation set contained 9. The complete data is present in figure 7 (b) which represents the distributions of temperature obtained during the ARAMIS 2 mission while the reconstruction by PROFHMM shown in figure 7 (c).

During this period the surface layers of the South Atlantic ocean continue to be heated by the atmosphere. The equator isotherms thin out, leading to a thermostad stable zone. The typical "W" equatorial structure is tightened, and we observe zones where the thermocline is almost flat, indicating a lack of currents.

These phenomena are present in figure 7 (b) which shows the distributions of temperature obtained during the ARAMIS 2 mission and are also well reconstructed by PROFHMM in figure 7 (c).

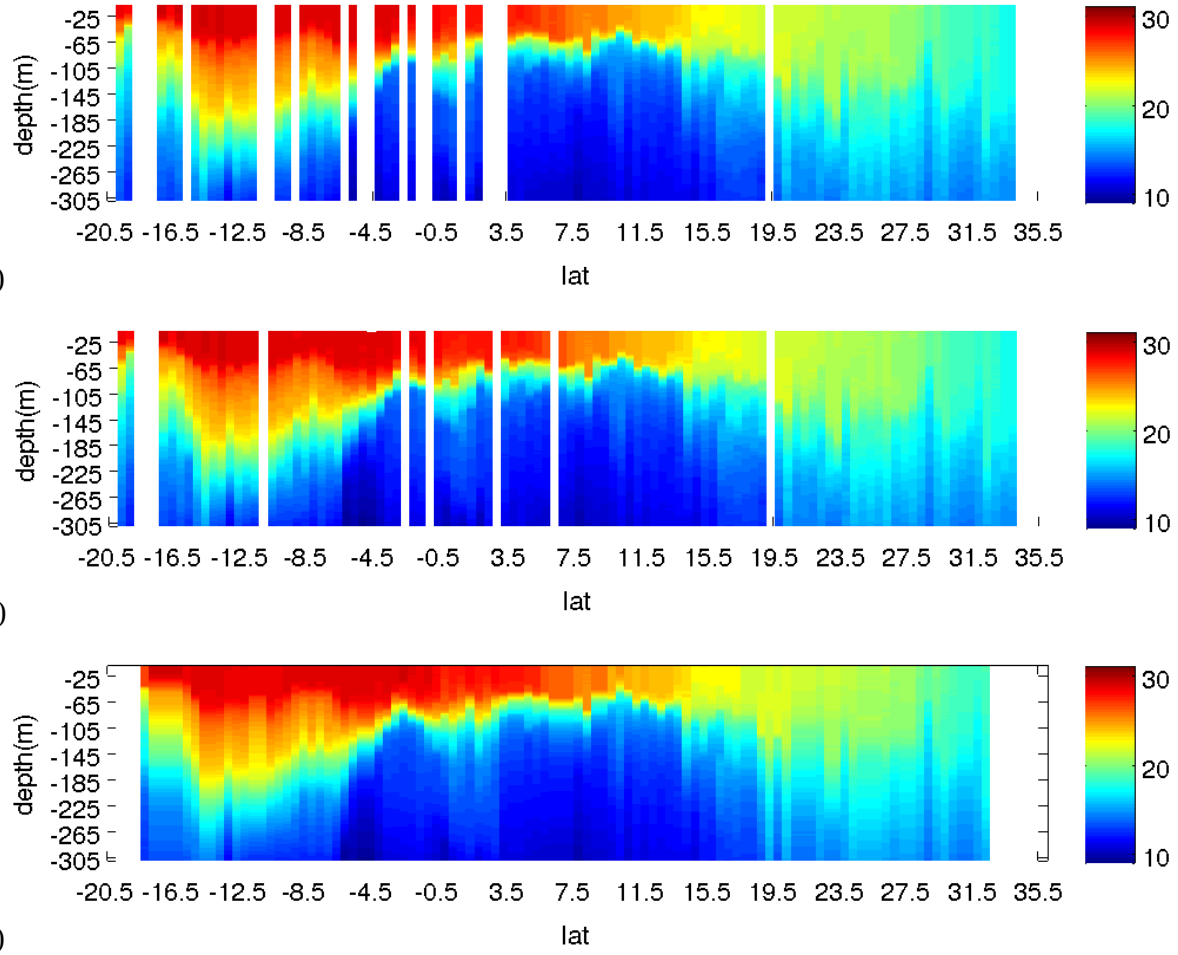


Figure 7: (a) Vertical distribution of the validation data set of the ARAMIS 2 mission for every available latitude where such a profile was provided in the training data set, (b) the complete ARAMIS 2 MISSION profiles and (c) the reconstruction of the ARAMIS rail based solely on surface data by the PROFHMM method. The colorbars are in °C.

The difference between the real values and the reconstructed ones is seen in figure 8 (a), while in figure 8 (b) we focus on the reconstruction of the values of ARAMIS 2 mission present in the validation data set. In figure 8 (a) we can observe that, again, about 90% of the vectors reconstructed from the sea-surface images do not present any significant differences with the in-situ measures, the majority of the image corresponding to the error margins of the in-situ measuring instruments. The differences are again well marked in the thermocline zone, where the variability of temperature is the strongest, due to the high variability of the zone and the quantification errors involved when using clustering methods. This is even more pronounced since there exist a principal and a secondary thermocline in the South hemisphere during this mission, as seen in figure 7 (b).

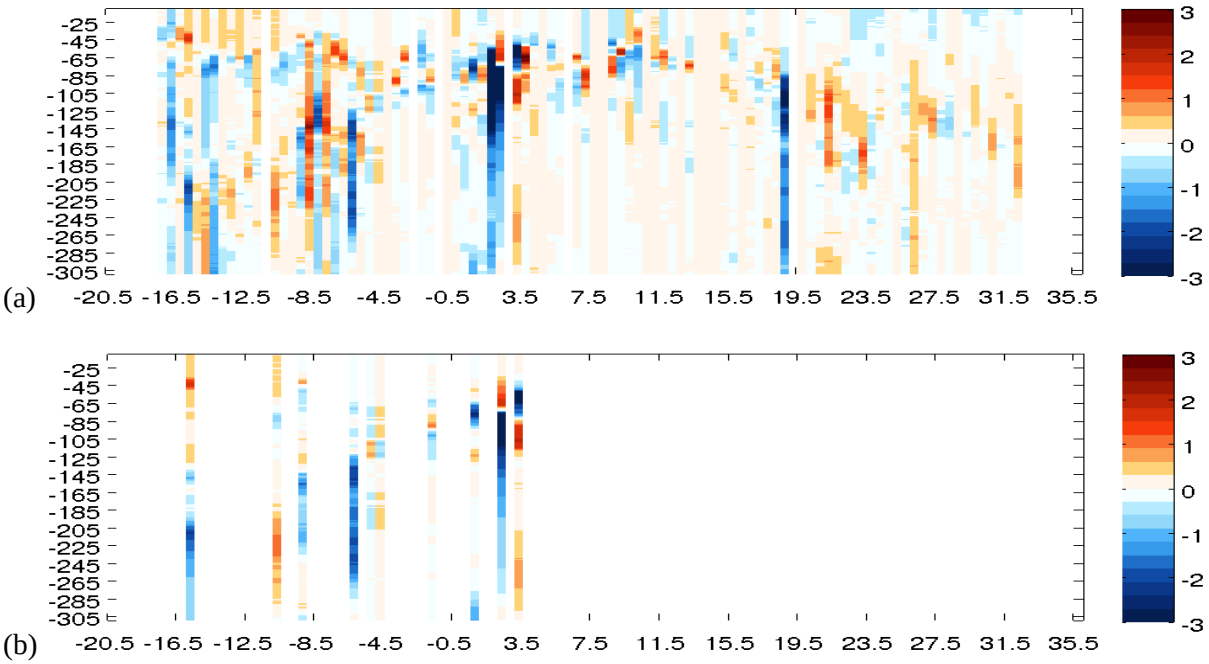


Figure 8: (a) The difference in values between the reconstruction and the complete ARAMIS 2 mission profiles. (b) A zoom on the errors over the values present in the validation set. The colorbar values are in °C. The vertical axis corresponds to the depth in meters while the horizontal one corresponds to the latitude over the ARAMIS transect.

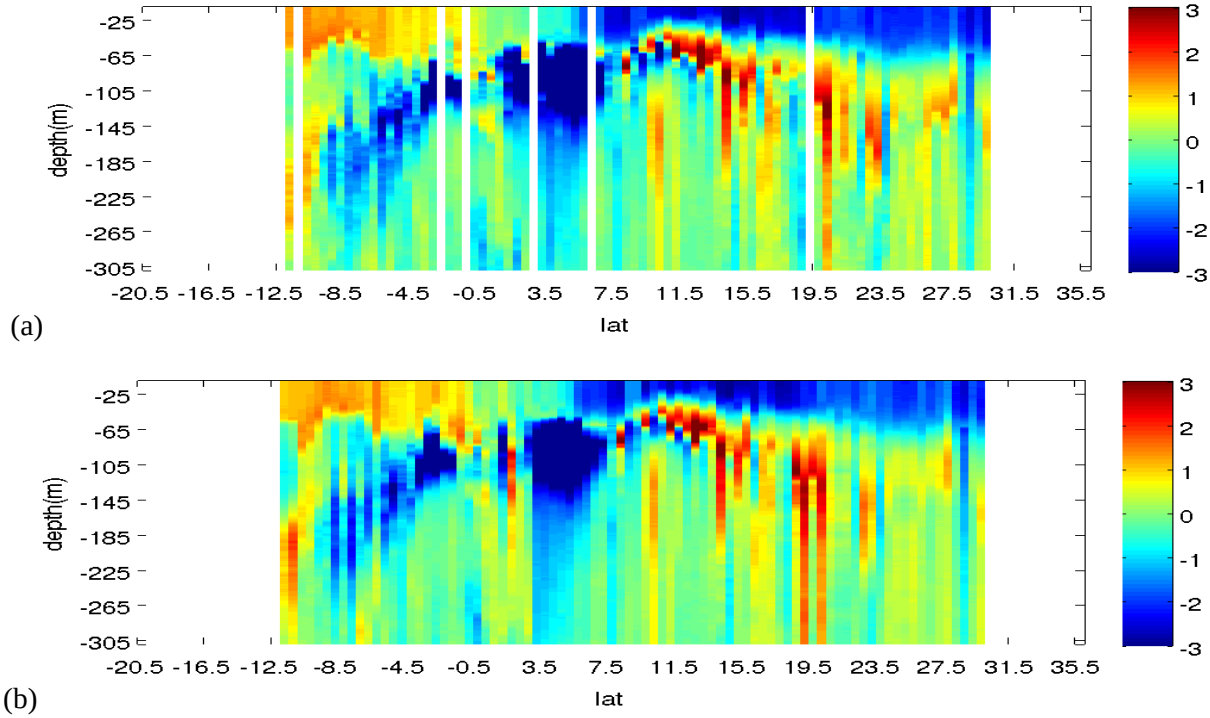


Figure 9: (a) The complete data set minus the “average year”. (b) Reconstruction minus the “average reconstructed year”. The colorbar values are in °C.

In figure 8 (b) we can observe that the 9 profiles reconstructed present lower difference in values than those observed in figure 5 (b). This is logical since there were more Boreal spring profiles of temperature in the training data-base, and therefore it is easier for the method to reconstruct them.

In figure 9 (a) and 9 (b) we present, respectively the ARAMIS 2 mission minus the" average year" and the PROFHMM ARAMIS 2 mission reconstruction based on sea-surface observations minus the "average reconstructed year". We notice that there is a warm anomaly in the South hemisphere and a cold one in the North when performing a year to year comparison, as well as a cold subsurface layer in 4°N. The general form of these anomalies is again retrieved both in amplitude and positioning by the application of PROFHMM.

ARAMIS 5

The ARAMIS 5 mission took place between the 12th and the 20th of September, 2004. It is the second mission occurring during the Boreal fall, and therefore during the Austral spring. The white columns in figure 10 (a) represent the missing profiles of temperature in the training data set, and separate the mission in 10 sequences that were used to learn the transition and emission probabilities of the HMM. The training data set contained 101 profiles while the validation set contained 5. The complete ARAMIS 5 mission profiles are presented in figure 10 (b) and the reconstruction by PROFHMM is seen in figure 10 (c).

During this period the South Atlantic ocean surface layers cool down , while the North Atlantic surface layers were heated. In the upwelling around Cape Verde front at 13°N, the isotherms are more concentrated creating an steep drop of temperature. The typical "W" equatorial structure is spread out. The counter current strengthens and transports warmer waters along its mean path at 6°N. These phenomena are present in figure 10 (b) which shows the distributions of temperature obtained during the ARAMIS 5 mission and are also reconstructed by PROFHMM in figure 10 (c).

The difference between the real values and the reconstructed ones is seen in figure 11 (a), while in figure 11 (b) we focus on the reconstruction of the values of ARAMIS 5 mission present in the validation data set. In figure 11 (a) we can observe that about 95% of the vectors reconstructed from the sea-surface images do not present significant differences with the in-situ measurements. The differences are again the strongest in the thermocline zone, where the variability of temperature is the strongest.

In figure 11 (b) we can observe that the 5 profiles reconstructed present higher difference in values than those observed in figure 5 (b) but higher than those presented in figure 8 (b). This is predictable, since more Boreal spring profiles of temperature are present in the training data-base, than Boreal fall ones and more than Boreal summer ones. This is consistent with the results in table 6.

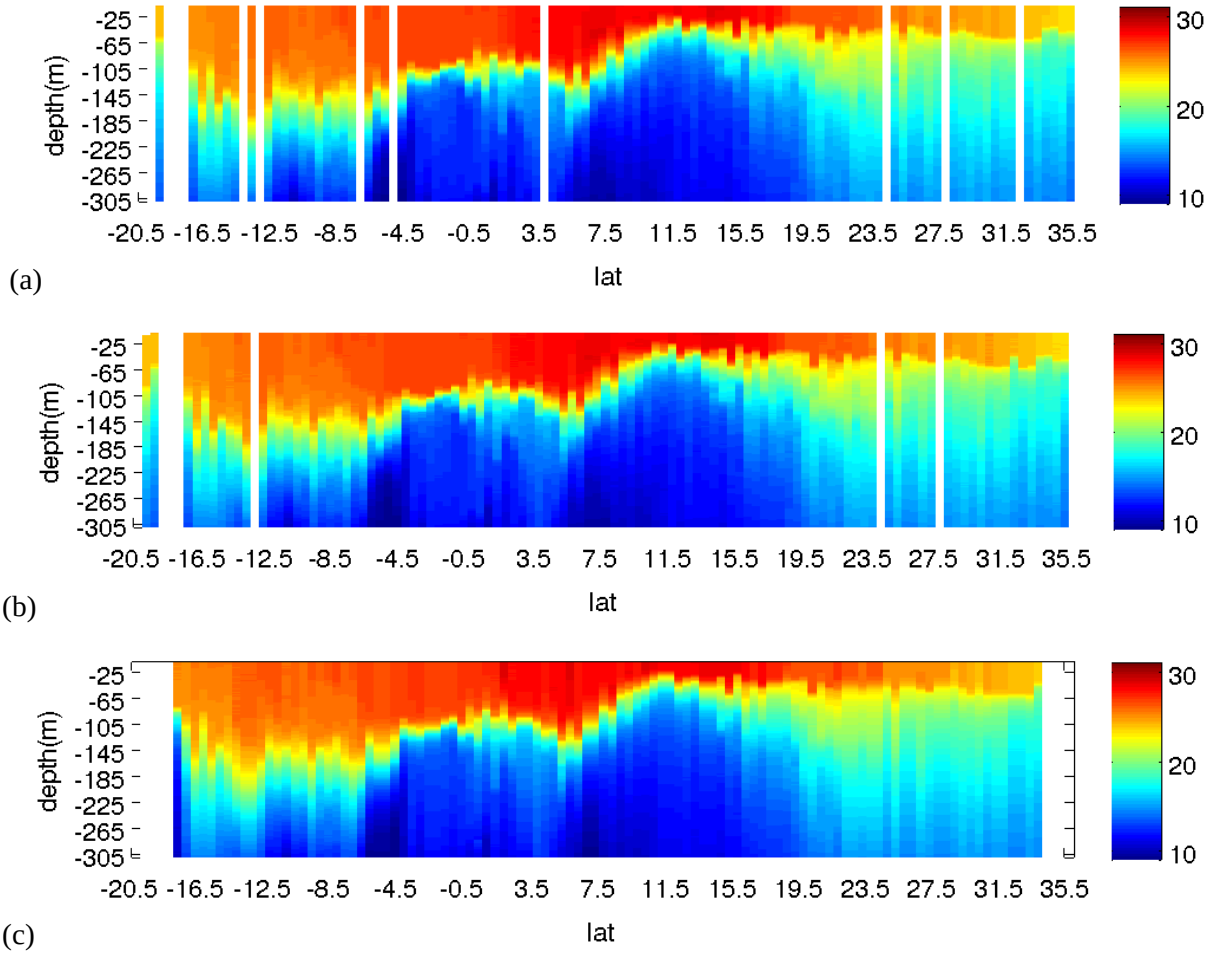


Figure 10: (a) Vertical distribution of the validation data set of the ARAMIS 5 mission for every available latitude where such a profile was provided in the training data set, (b) the complete ARAMIS 5 MISSION profiles and (c) the reconstruction of the ARAMIS rail based solely on surface data by the PROFHMM method. The colorbars are in °C.

In figure 12 (a) and 12 (b) we present, respectively the ARAMIS 5 mission minus the average year and the PROFHMM ARAMIS 5 mission reconstruction based on sea-surface observations minus the average reconstructed year. We notice that there is a hot anomaly in the surface of the North Atlantic ocean and a cold one in the surface of the South Atlantic. There is also a strong hot anomaly around 6°N latitude corresponding to the equatorial counter-current. The general form of these anomalies is again retrieved both in amplitude and position by the application of PROFHMM.

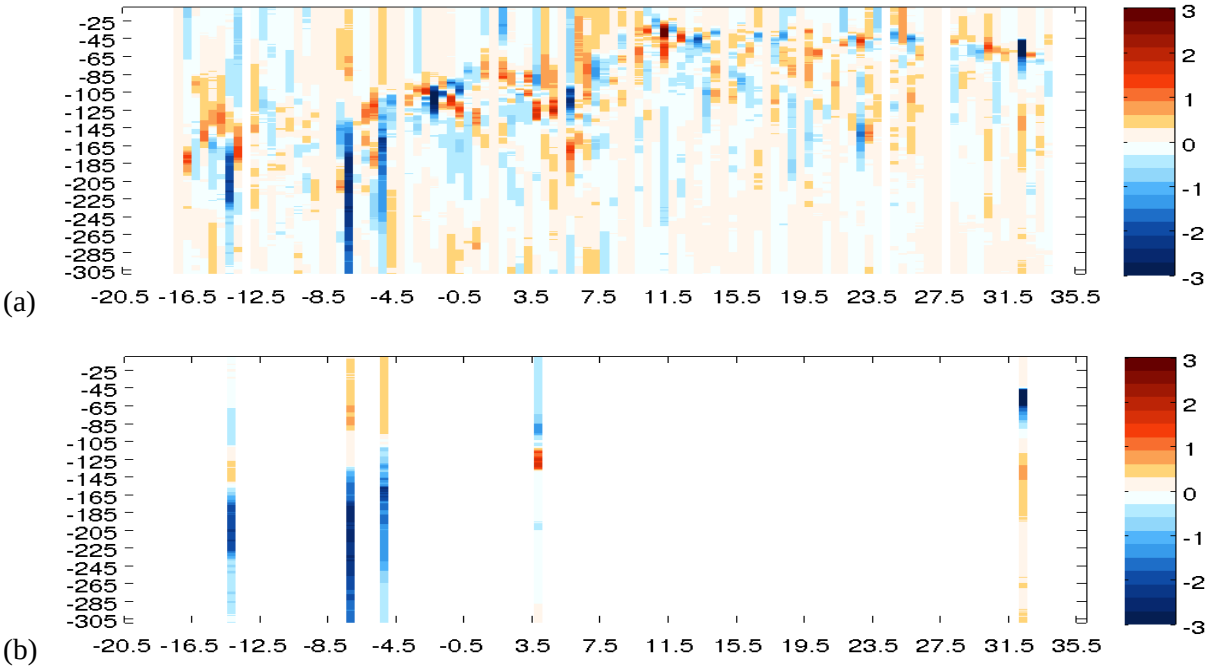


Figure 11: (a) The difference in values between the reconstruction and the complete ARAMIS 5 mission profiles. (b) A zoom on the errors over the values present in the validation set. The colorbar values are in °C. The vertical axis corresponds to the depth in meters while the horizontal one corresponds to the latitude over the ARAMIS transect.

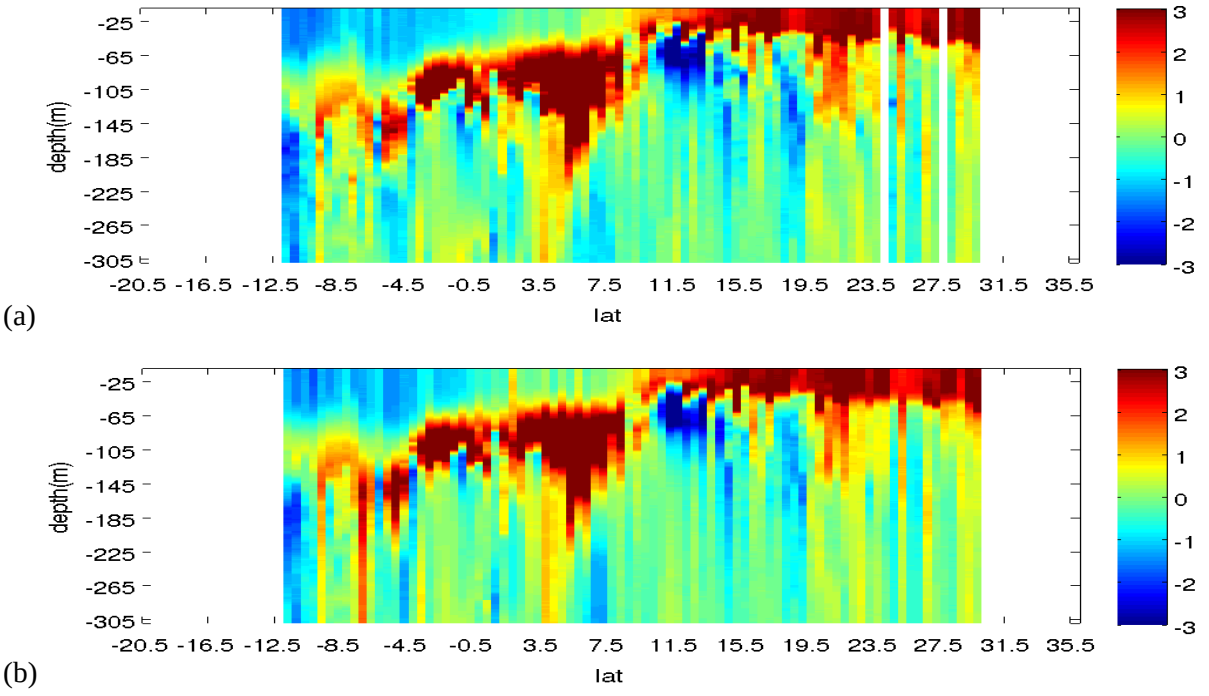


Figure 12: (a) Complete data set minus the “average year”. (b) Reconstruction minus the “average reconstructed year”. The colorbar values are in °C.

6. CONCLUSIONS:

The application of PROFHMM to the ARAMIS data set demonstrates that it is possible to statistically model the vertical distribution of temperature from sea-surface observation only. However, while applying the PROFHMM methodology, we are forced to work under a certain amount of limitations imposed by the modeling choice.

Given the performances obtained over the validation data set, however, we expect that we could use PROFHMM to retrieve the vertical profiles over the average ARAMIS rail with a similar precision to the one observed during the validation. Owing to the learning data base, the retrieval performances will be better during the Boreal spring.

The performances of PROFHMM and therefore its predictive power for the retrieval of vertical profiles of temperature by inverting sea-surface observations over the ARAMIS rail can be improved with a larger learning data base of concurrent vertical profiles and observations. Since PROFHMM is a statistical method, its performances depend on the sampling of the phenomena under study. The memorized dynamics reflect the observed cases, and are able to extrapolate in order to retrieve similar temperature profiles. We can therefore tolerate a certain amount of under-sampling. However, the retrieved profiles in extreme and infrequent situations will not be accurate, unless they are well introduced into the training data set.

As with most Hidden Markov Models, inverting with PROFHMM a trajectory with only few points does not give good results. The Viterbi Algorithm will give better results over a long series of data points, since that would maximize the information for the selection of each given state. Therefore, to acquire optimum results we require longer time sequences.

We also note that a larger training data base will allow a larger number of states with the SOM clustering, which in turn will lead to a finer discretization and, consequently, a better statistical modeling.

AKNOWLEDGMENTS

The research presented in this paper was financed by Centre National de l'Etude Spatial (CNES, French national center of spatial studies), and the Delegation Gouvernemantale pour l'Armement (DGA, French Military Research Delegation), which we wish to thank for their support.

REFERENCES

[1] Feldman, G.C., N.A. Kuring, C. Ng, W.E. Esaias, C.R. McClain and J.A. Elrod, N.Maynard, D.Endres, R. Evans, J.Brown, S.Walsh, M. Carle, G. Podesta (1989). Ocean Color: Availability of the Global Data Set. EOS, 70, 634-641

[2] Dinnat, E., J. Boutin, G. Caudal, J. Etcheto, and P. Waldteufel, Influence of sea surface emissivity model parameters in L-band for the estimation of salinity, International Journal of Remote Sensing, 23, 5117-5122, 2002.

[3] P. Krishna Rao, W. L. Smith, and R. Koffler (January 1972). "Global Sea-Surface Temperature Distribution Determined From an Environmental Satellite". Monthly Weather Review 100 (1): 10–14

[4] Uitz J, Claustre H, Morel A, Hooker S B 2006 Vertical distribution of phytoplankton communities in open ocean: An assessment based on surface chlorophyll, JOURNAL OF GEOPHYSICAL RESEARCH, VOL. 111, C08005, doi:10.1029/2005JC003207

[5] Altimétrie sur un Rail Atlantique et Mesures In Situ, PI S. Arnault, CNES,IRD and INSU

[6] Tanguy Y, 2011, « Variabilité de la dynamique et la thermodynamique dans l'Atlantique tropical : Projet ARAMIS » , phd Thèsis, UPMC

[7] CORIOLIS Data Centre (http://www.coriolis.eu.org/cdc/argo.htm)

[8] Kohonen T, 1990 The Self-organizing Map, PROCEEDINGS OF THE IEEE, VOL. 78, NO 9 and http://www.cis.hut.fi/projects/somtoolbox/package/docs2/somtoolbox.html

[9] Charantonis A, Badran F, Thiria S, 2012, Retrieving the vertical profiles of Chlorophyll-A from satellite observations, by using hidden Markov models and self-organizing maps. JAOT, submitted

[10] Gurvan M, and the NEMO team, 2012, NEMO ocean engine – version 3.4 – Note du Pôle de modélisation de l'Institut Pierre-Simon Laplace No 27 ISSN No 1288-1619.

[11] http://oceancolor.gsfc.nasa.gov/

[12] Juang B-H, 2003, Hidden Markov Models. Encyclopedia of Telecommunications

[13] A.J.Viterbi (April 1967). "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm". IEEE Transactions on Information Theory 13 (2): 260–269. doi:10.1109/TIT.1967

[14] ARGO Science Team 1998; see http://www.argo.ucsd.edu

[15] Tanguy, Y., et al., Isothermal, mixed, and barrier layers in the subtropical and tropical Atlantic Ocean during the ARAMIS experiment. Deep-Sea Research I (2010), doi:10.1016/j.dsr.2009.12.012

[16] http://www.aviso.oceanobs.com/duacs/

[17] Jolliffe, I.T. (2002). Principal Component Analysis, second edition (Springer)

3.3 ANNEXE DE L'ARTICLE/ PRESENTATION DES RECONSTRUCTION PROFHMM DES MISSIONS ARAMIS RESTANTES : 3,4 et 6 A 12.

Dans cette sections nous présentons, pour chaque campagne ARAMIS non présentée dans les résultats, les images correspondantes aux figures dans l'article. Les résultats restent cohérents avec ceux présentés dans l'article.

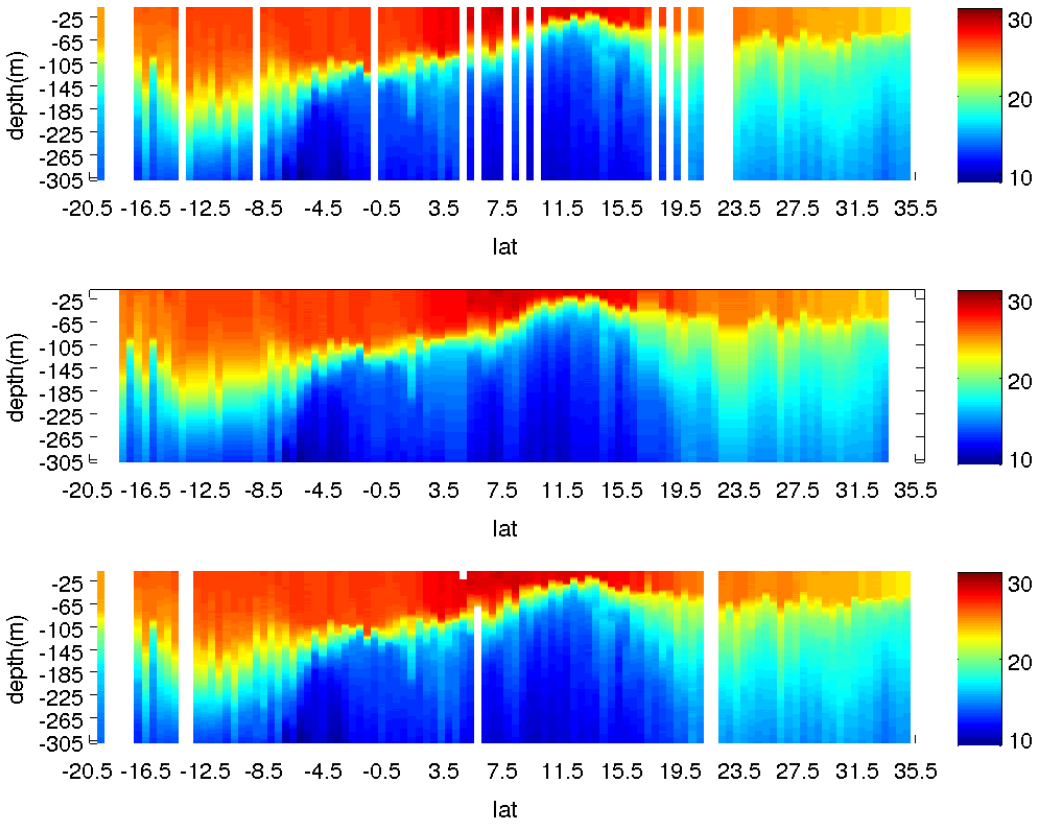


Figure 1: The vertical distribution of the validation data set of the ARAMIS 3 mission for every available latitude where such a profile was provided in the training data set is seen in figure 1 (a), the complete ARAMIS 3 MISSION profiles in figure 1 (c) and the reconstruction of the ARAMIS rail based solely on surface data by the PROFHMM method in figure 1 (b). The colorbars are in °C.

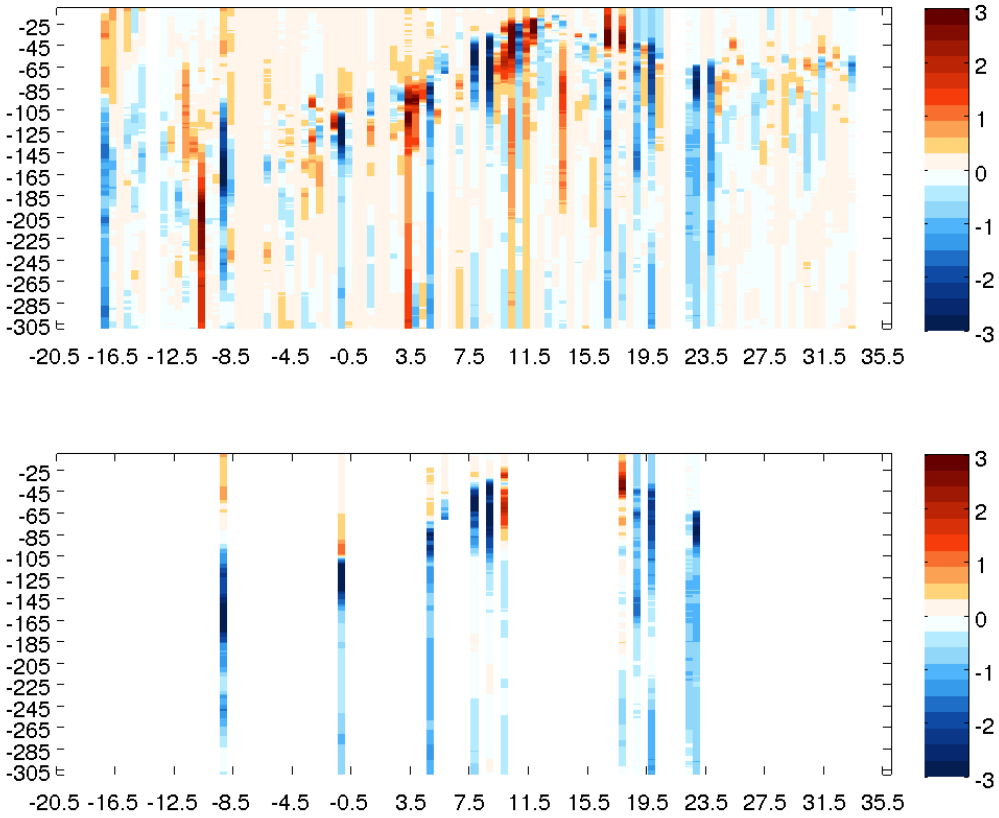


Figure 2: The difference in values between the reconstruction and the complete ARAMIS 3 mission profiles is shown in figure 2 (a). A zoom on the errors over the values present in the validation set is shown in figure 2 (b). The colorbar values are in °C. The vertical axis corresponds to the depth in meters while the horizontal one corresponds to the latitude over the ARAMIS transect.

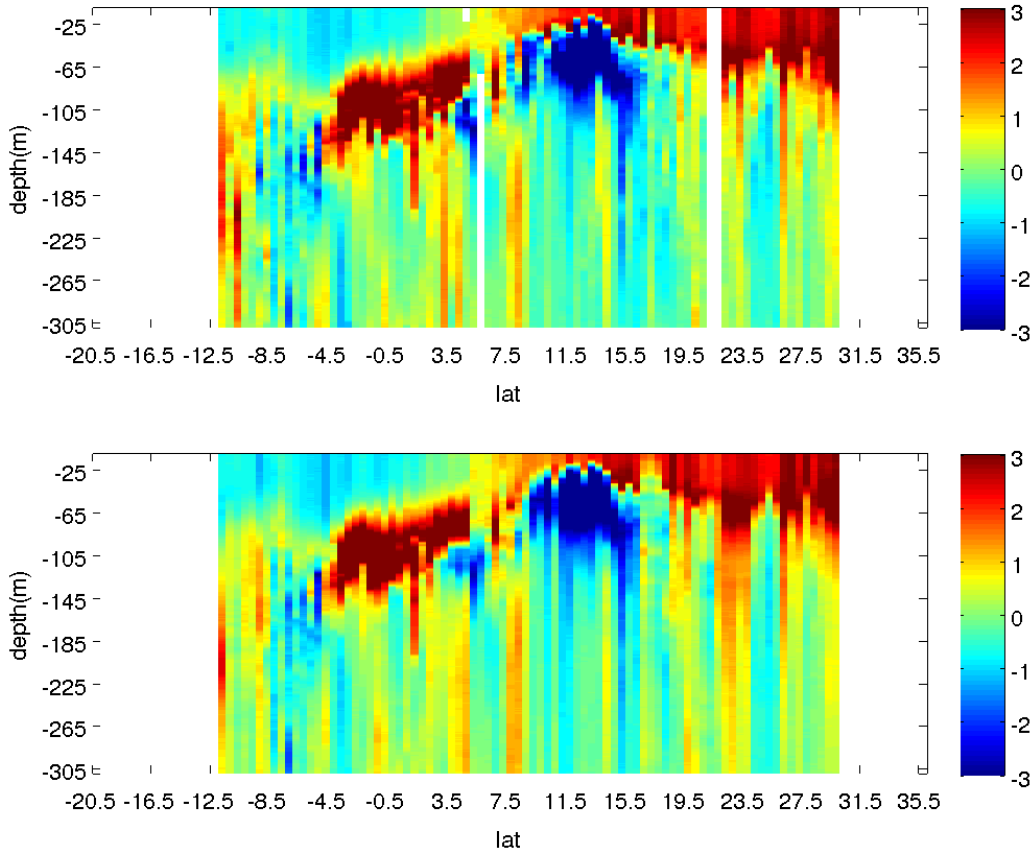


Figure 3: In figure 3 (a) we can see the complete data set minus the “average year”, and in figure 3 (b) the reconstruction minus the “average reconstructed year”. The colorbar values are in °C.

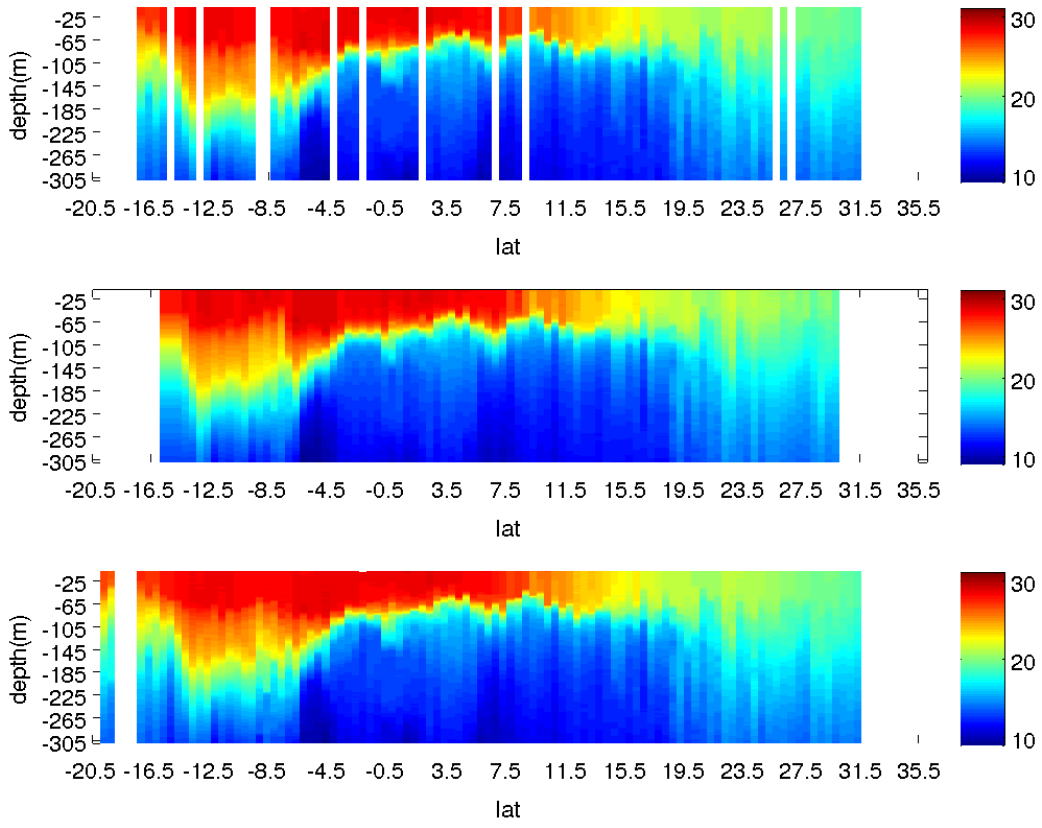


Figure 4: The vertical distribution of the validation data set of the ARAMIS 4 mission for every available latitude where such a profile was provided in the training data set is seen in figure 4 (a), the complete ARAMIS 4 MISSION profiles in figure4 (c) and the reconstruction of the ARAMIS rail based solely on surface data by the PROFHMM method in figure 4 (b). The colorbars are in °C.

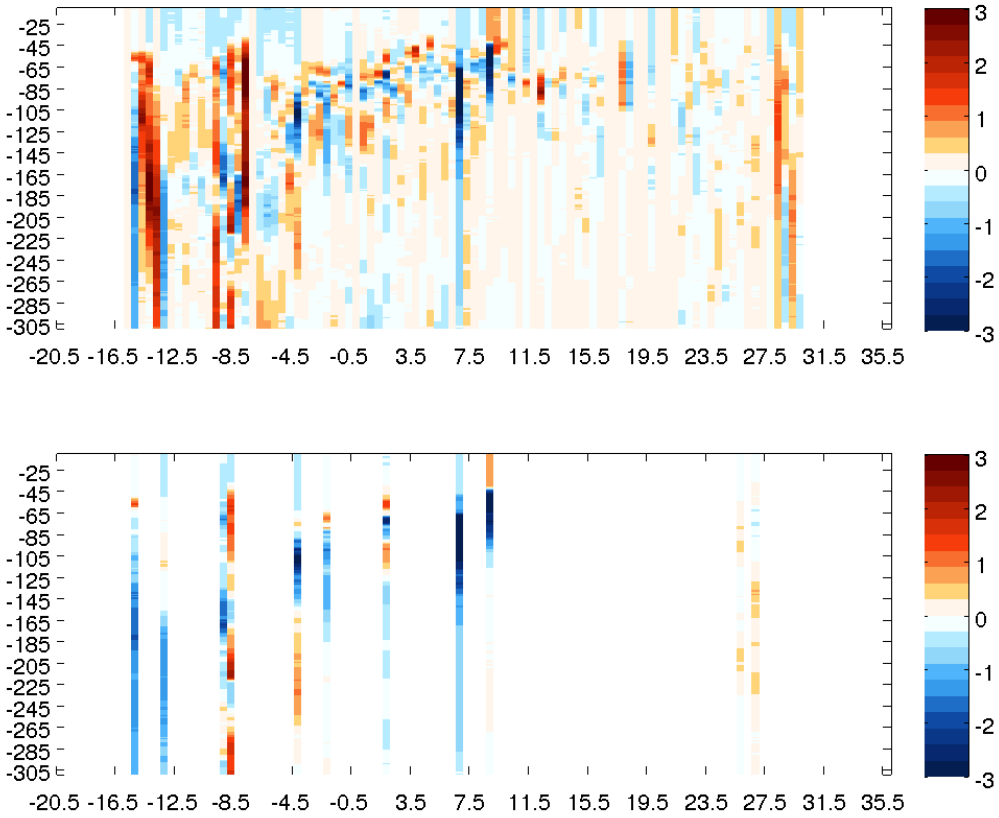


Figure 5: The difference in values between the reconstruction and the complete ARAMIS 4 mission profiles is shown in figure 5 (a). A zoom on the errors over the values present in the validation set is shown in figure 5 (b). The colorbar values are in °C. The vertical axis corresponds to the depth in meters while the horizontal one corresponds to the latitude over the ARAMIS transect.

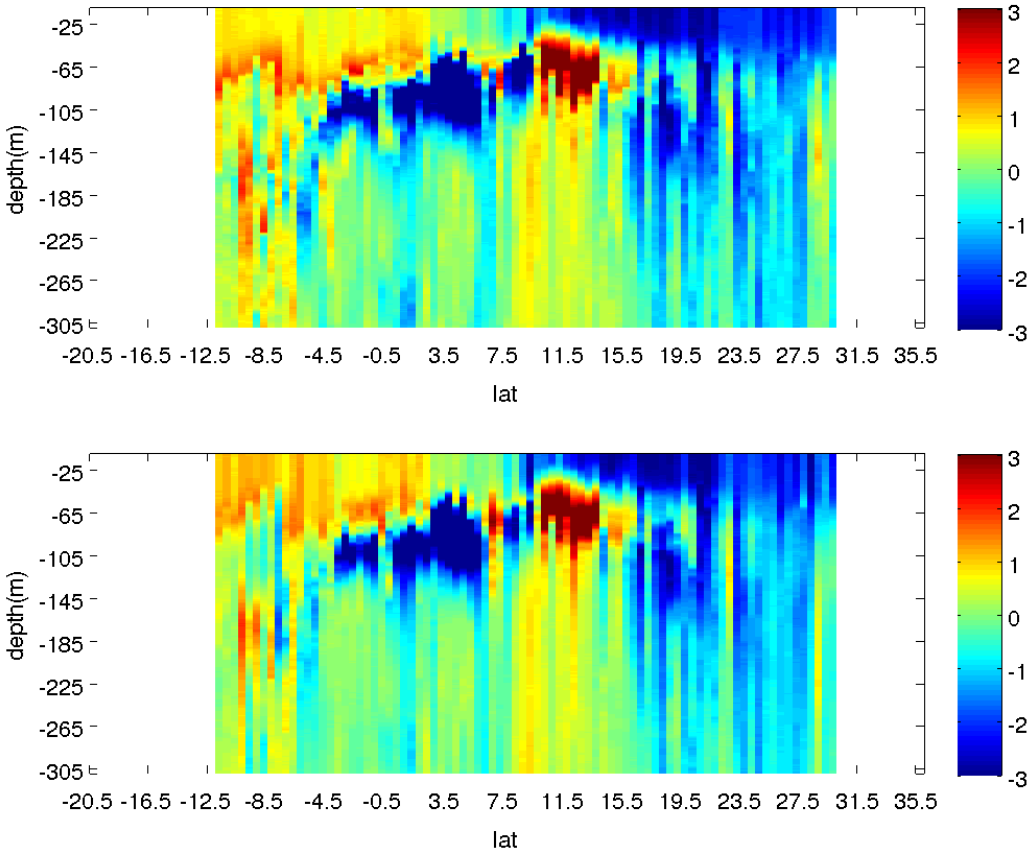


Figure 6: In figure 6 (a) we can see the complete data set minus the “average year”, and in figure 6 (b) the reconstruction minus the “average reconstructed year”. The colorbar values are in °C.

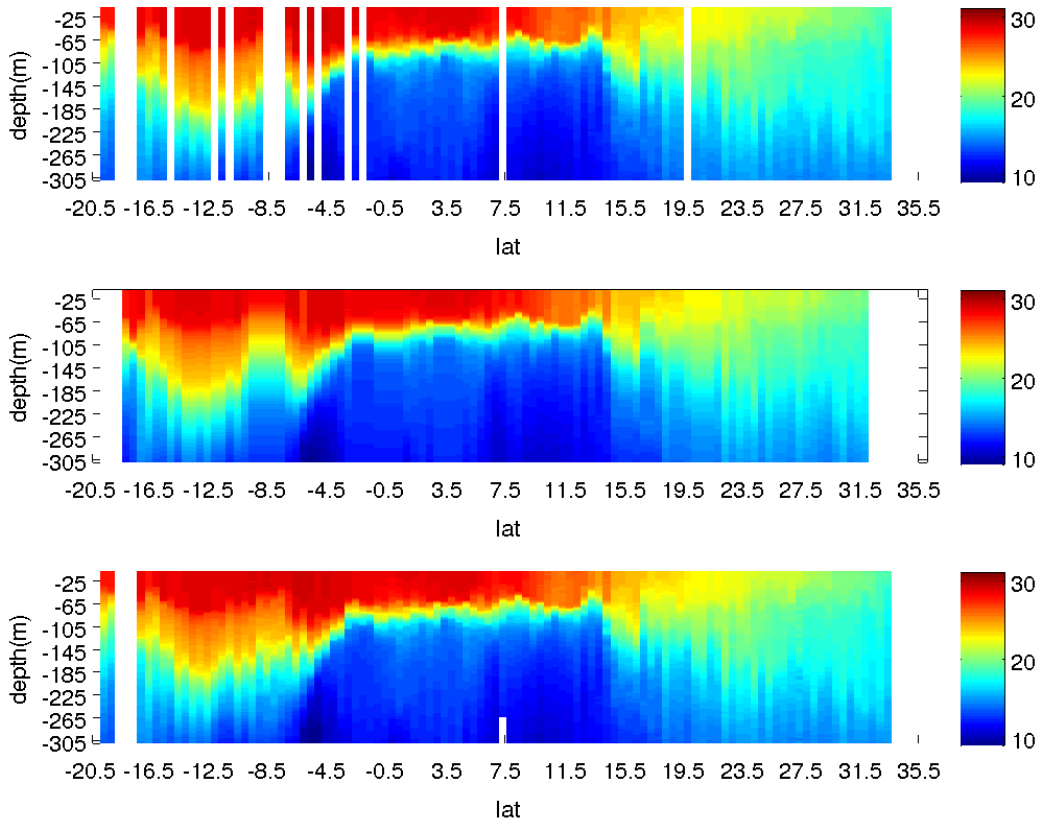


Figure 7: The vertical distribution of the validation data set of the ARAMIS 6 mission for every available latitude where such a profile was provided in the training data set is seen in figure 7 (a), the complete ARAMIS 6 MISSION profiles in figure 7 (c) and the reconstruction of the ARAMIS rail based solely on surface data by the PROFHMM method in figure 6 (b). The colorbars are in °C.

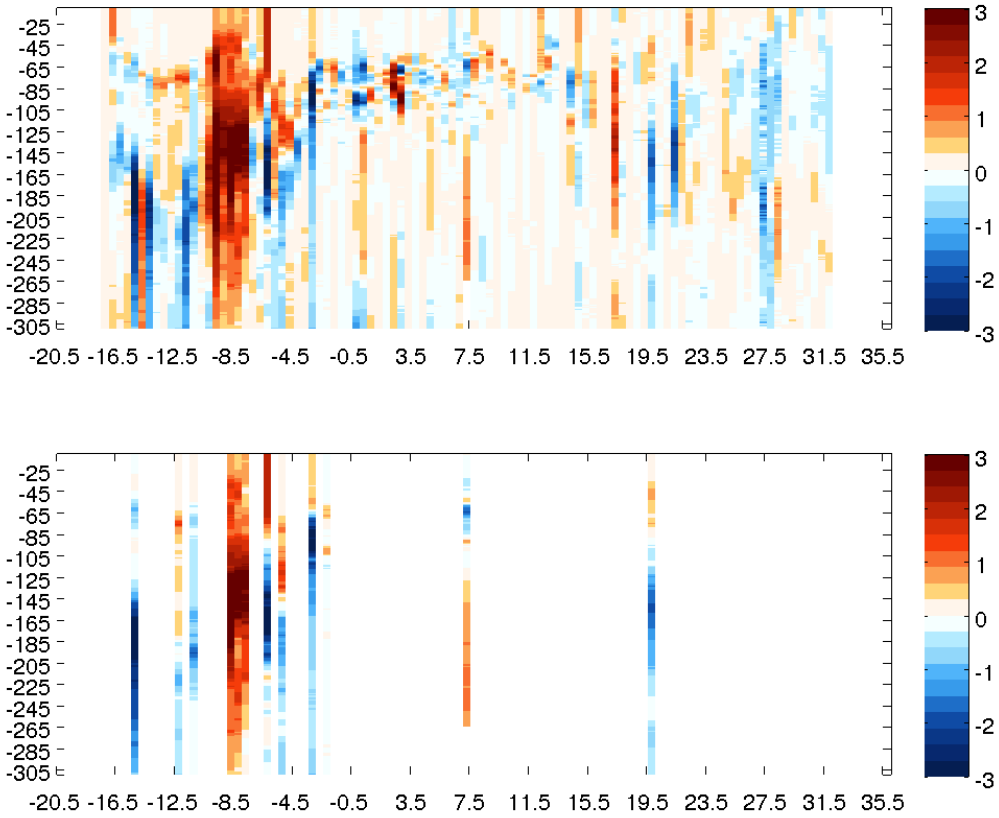


Figure 8: The difference in values between the reconstruction and the complete ARAMIS 6 mission profiles is shown in figure 8 (a). A zoom on the errors over the values present in the validation set is shown in figure 8 (b). The colorbar values are in °C. The vertical axis corresponds to the depth in meters while the horizontal one corresponds to the latitude over the ARAMIS transect.

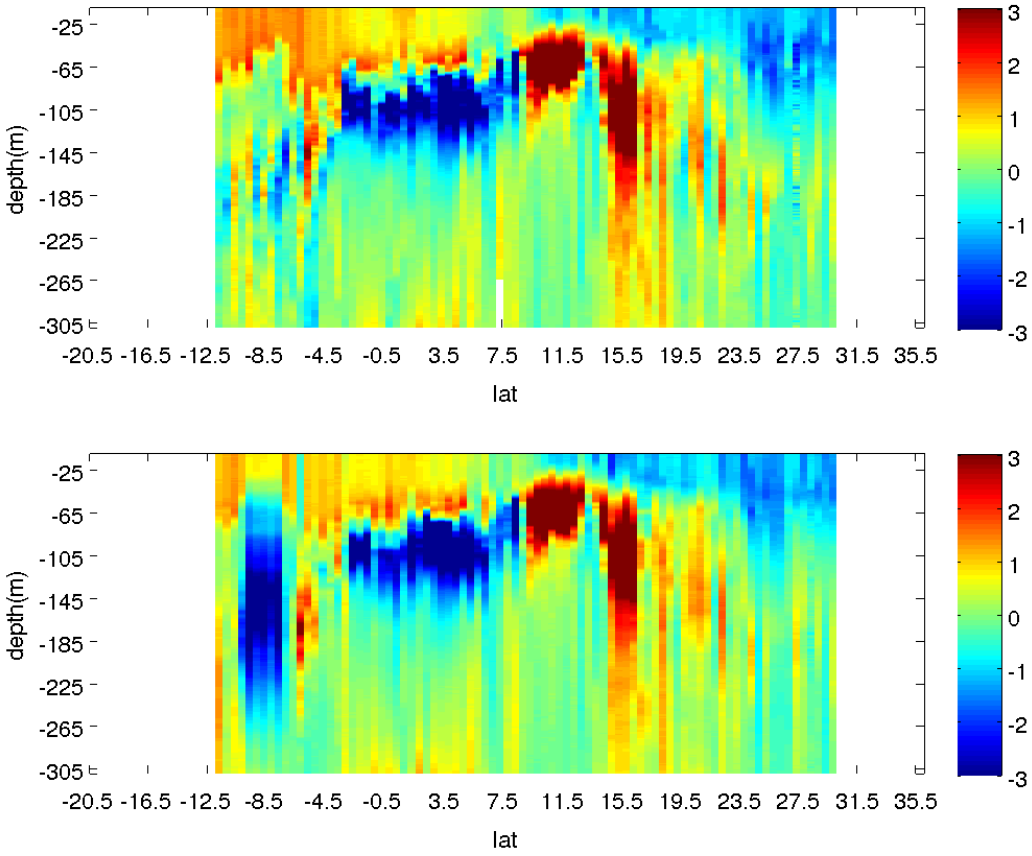


Figure 9: In figure 9 (a) we can see the complete data set minus the “average year”, and in figure 9 (b) the reconstruction minus the “average reconstructed year”. The colorbar values are in °C.

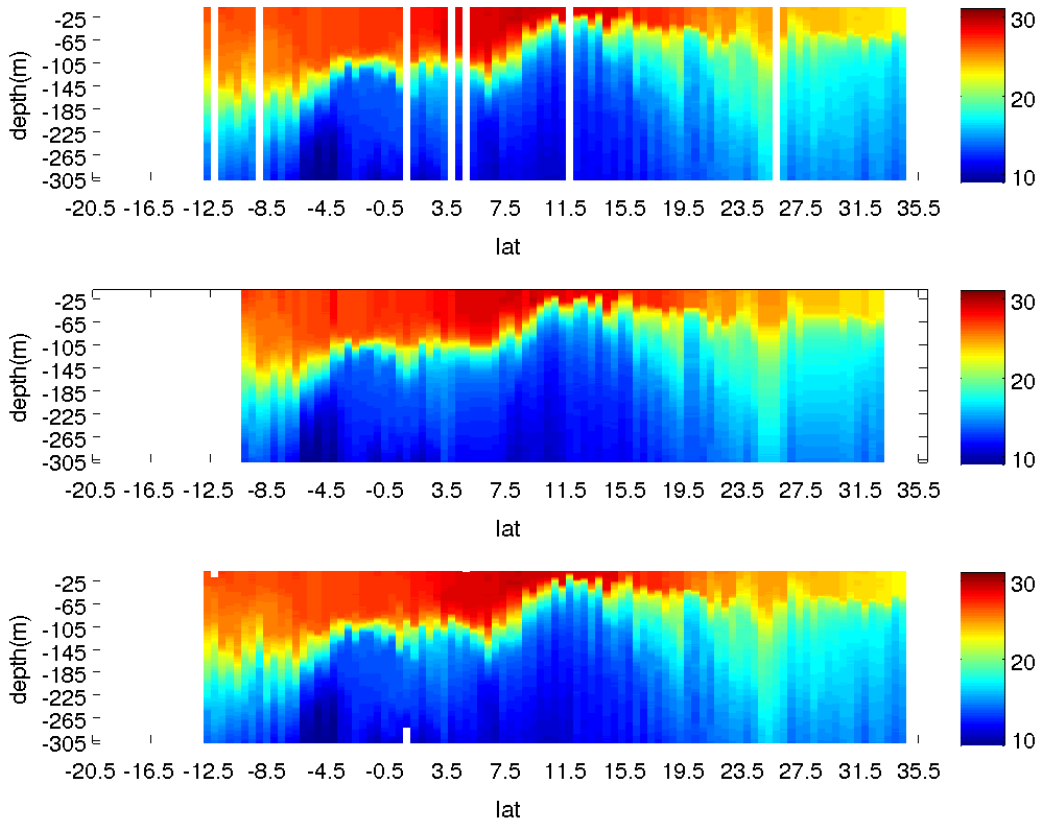


Figure 10: The vertical distribution of the validation data set of the ARAMIS 7 mission for every available latitude where such a profile was provided in the training data set is seen in figure 10 (a), the complete ARAMIS 7 MISSION profiles in figure 10 (c) and the reconstruction of the ARAMIS rail based solely on surface data by the PROFHMM method in figure 10 (b). The colorbars are in °C.

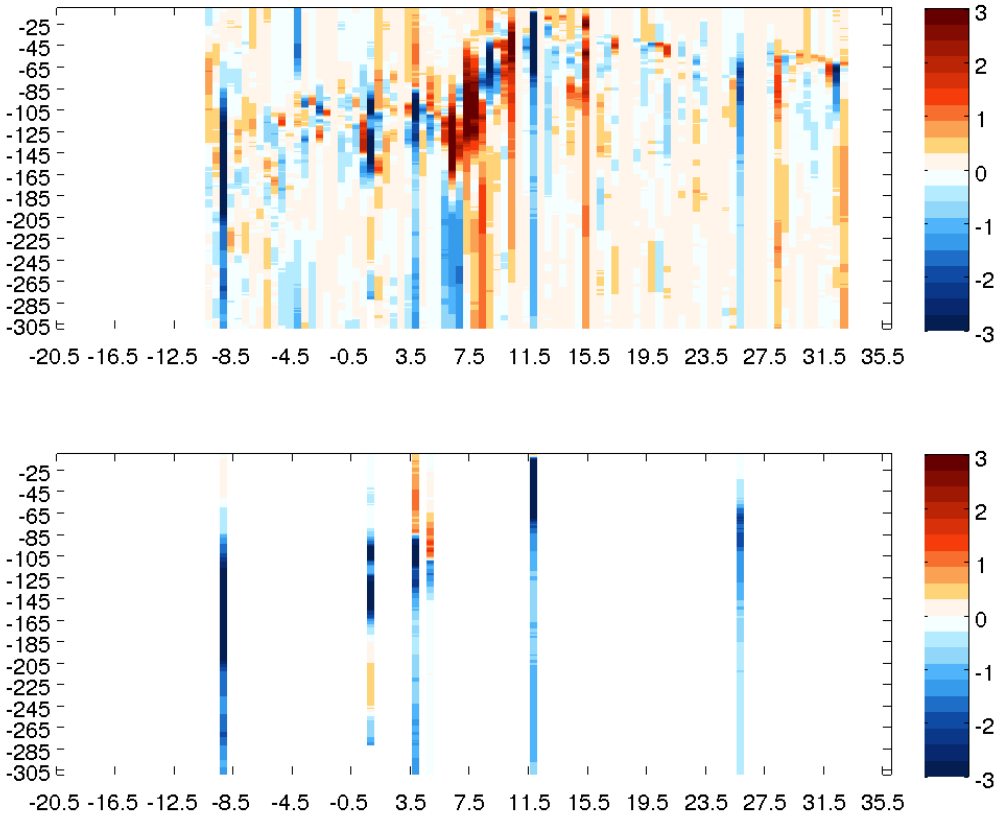


Figure 11: The difference in values between the reconstruction and the complete ARAMIS 7 mission profiles is shown in figure 11 (a). A zoom on the errors over the values present in the validation set is shown in figure 11 (b). The colorbar values are in °C. The vertical axis corresponds to the depth in meters while the horizontal one corresponds to the latitude over the ARAMIS transect.

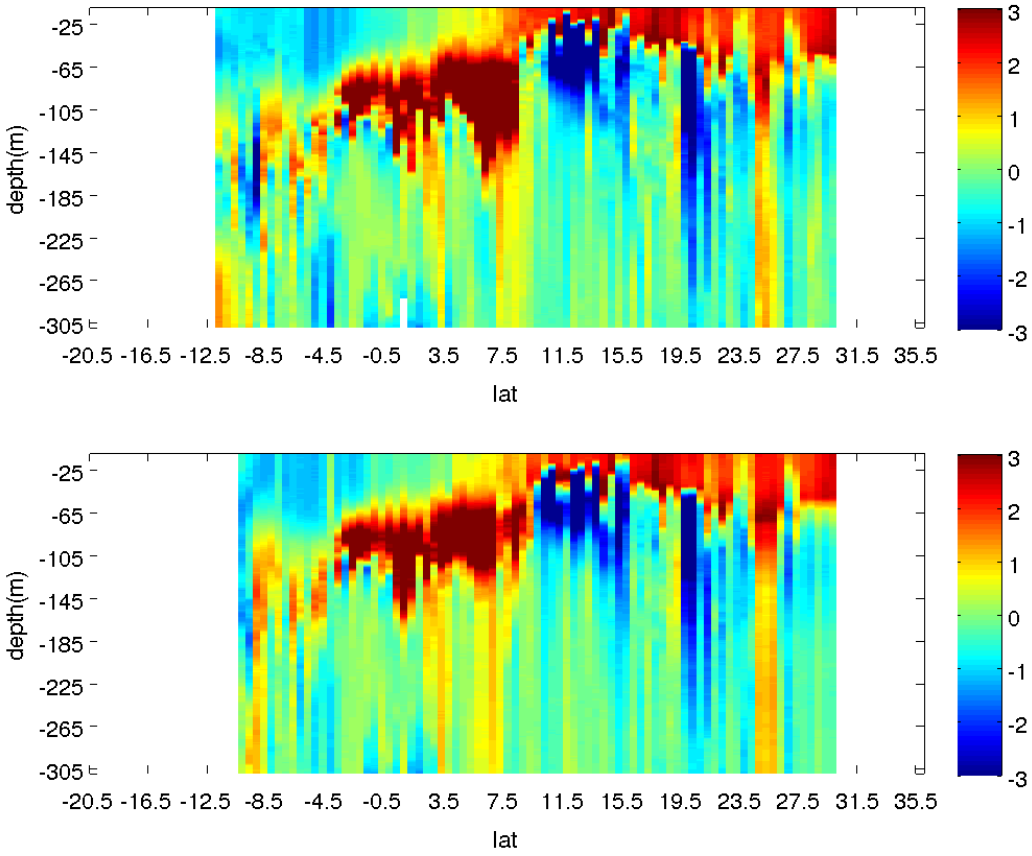


Figure 12: In figure 12 (a) we can see the complete data set minus the “average year”, and in figure 12 (b) the reconstruction minus the “average reconstructed year”. The colorbar values are in °C.

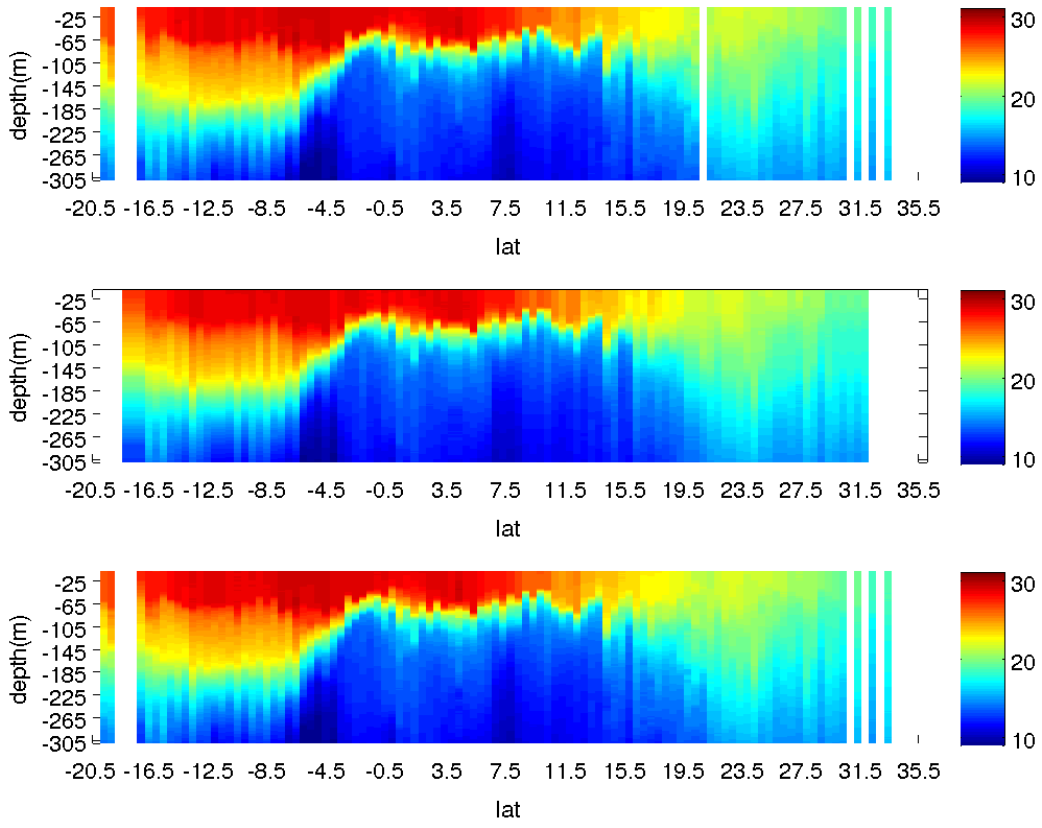


Figure 13: The vertical distribution of the validation data set of the ARAMIS 8 mission for every available latitude where such a profile was provided in the training data set is seen in figure 13 (a), the complete ARAMIS 8 MISSION profiles in figure 13 (c) and the reconstruction of the ARAMIS rail based solely on surface data by the PROFHMM method in figure 1 (b). The colorbars are in °C.

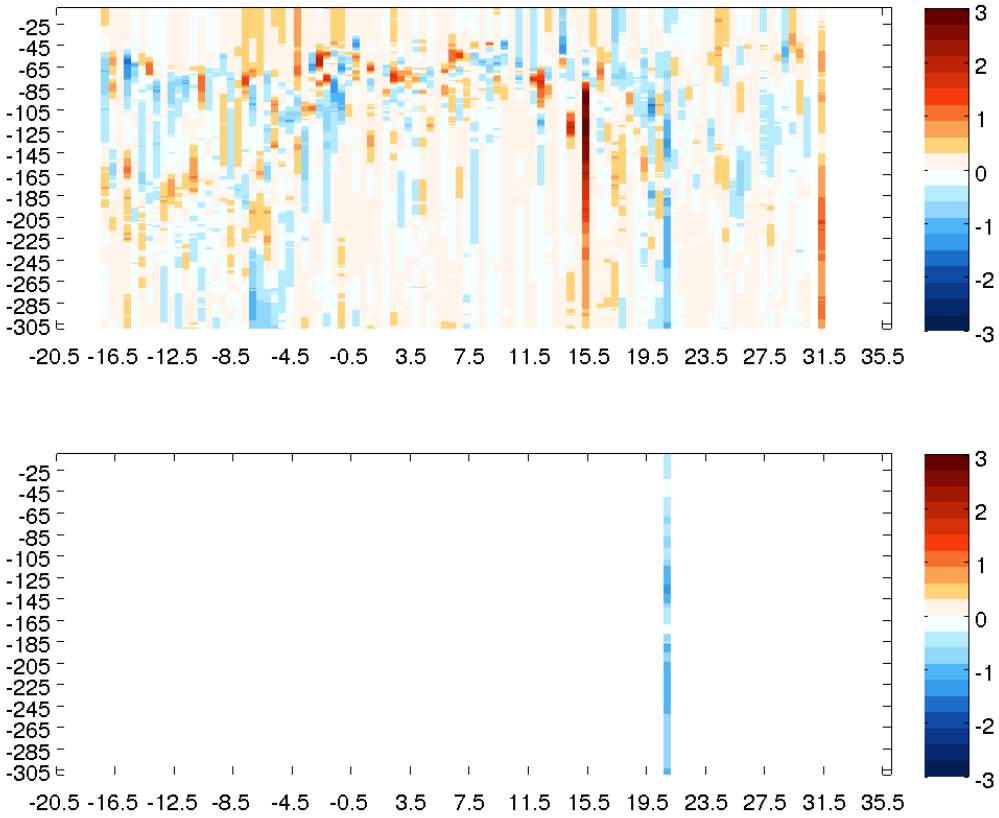


Figure 14: The difference in values between the reconstruction and the complete ARAMIS 8 mission profiles is shown in figure 13 (a). A zoom on the errors over the values present in the validation set is shown in figure 13 (b). The colorbar values are in °C. The vertical axis corresponds to the depth in meters while the horizontal one corresponds to the latitude over the ARAMIS transect.

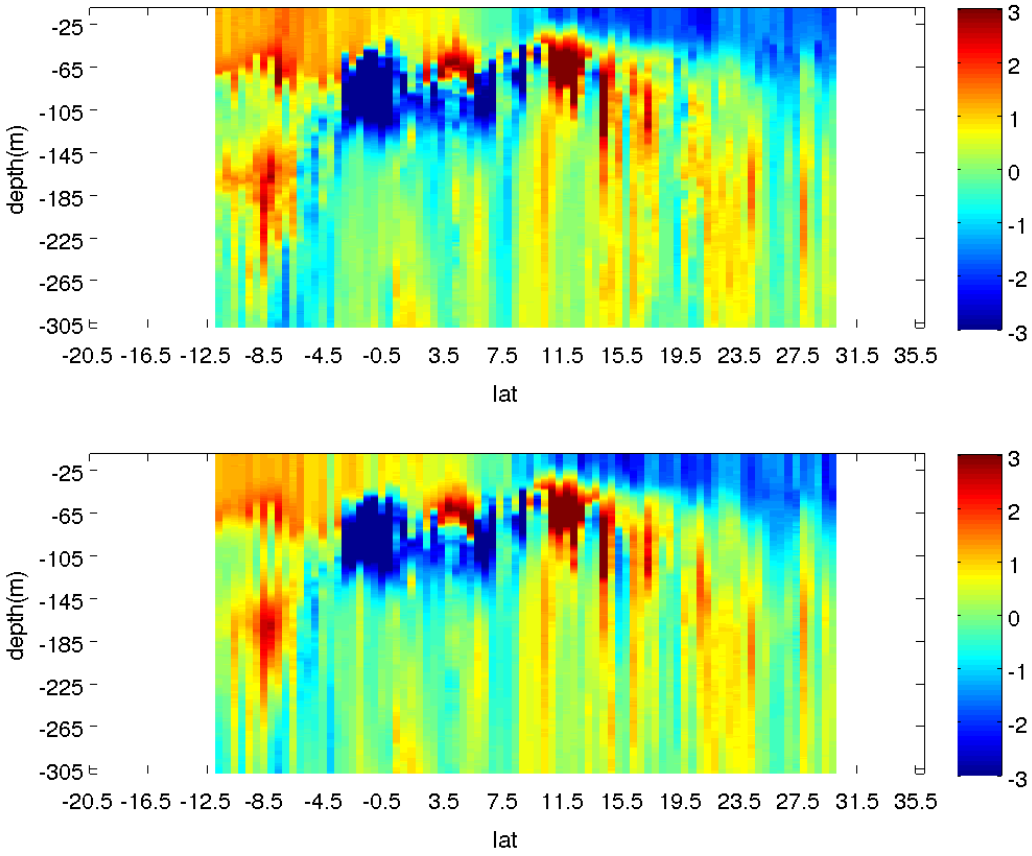


Figure 15: In figure 15 (a) we can see the complete data set minus the “average year”, and in figure 15 (b) the reconstruction minus the “average reconstructed year”. The colorbar values are in °C.

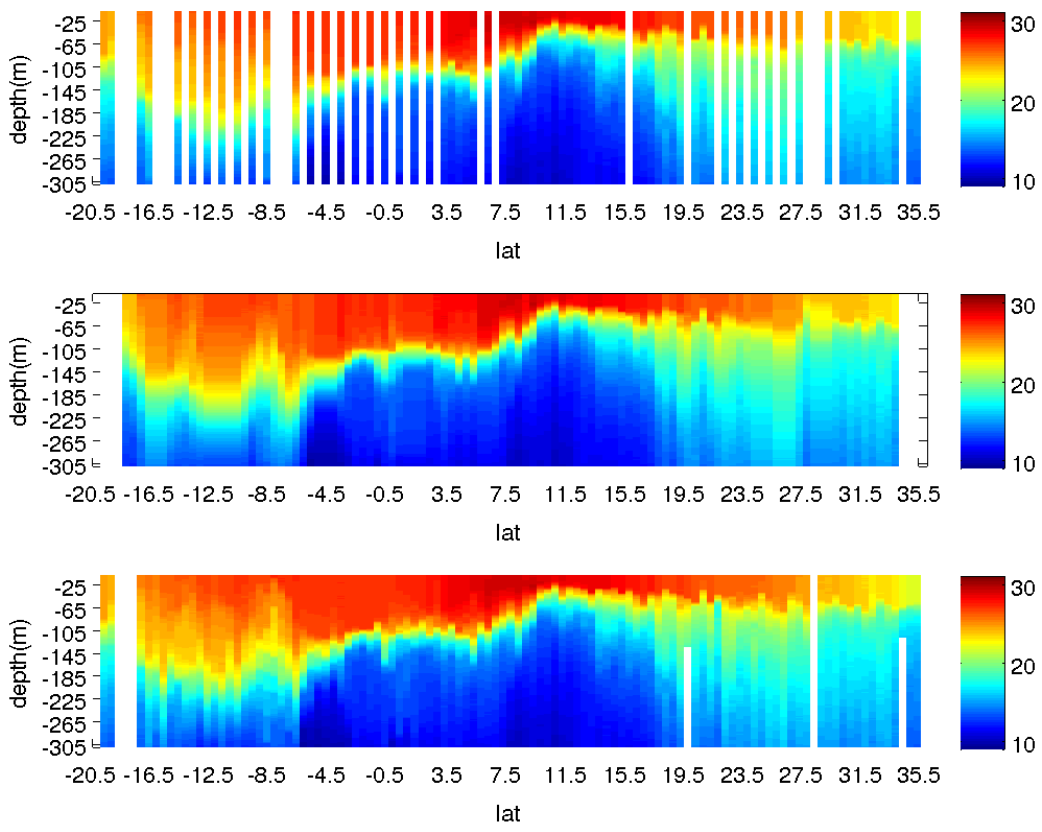


Figure 16: The vertical distribution of the validation data set of the ARAMIS 9 mission for every available latitude where such a profile was provided in the training data set is seen in figure 16 (a), the complete ARAMIS 9 MISSION profiles in figure 16 (c) and the reconstruction of the ARAMIS rail based solely on surface data by the PROFHMM method in figure 16 (b). The colorbars are in °C.

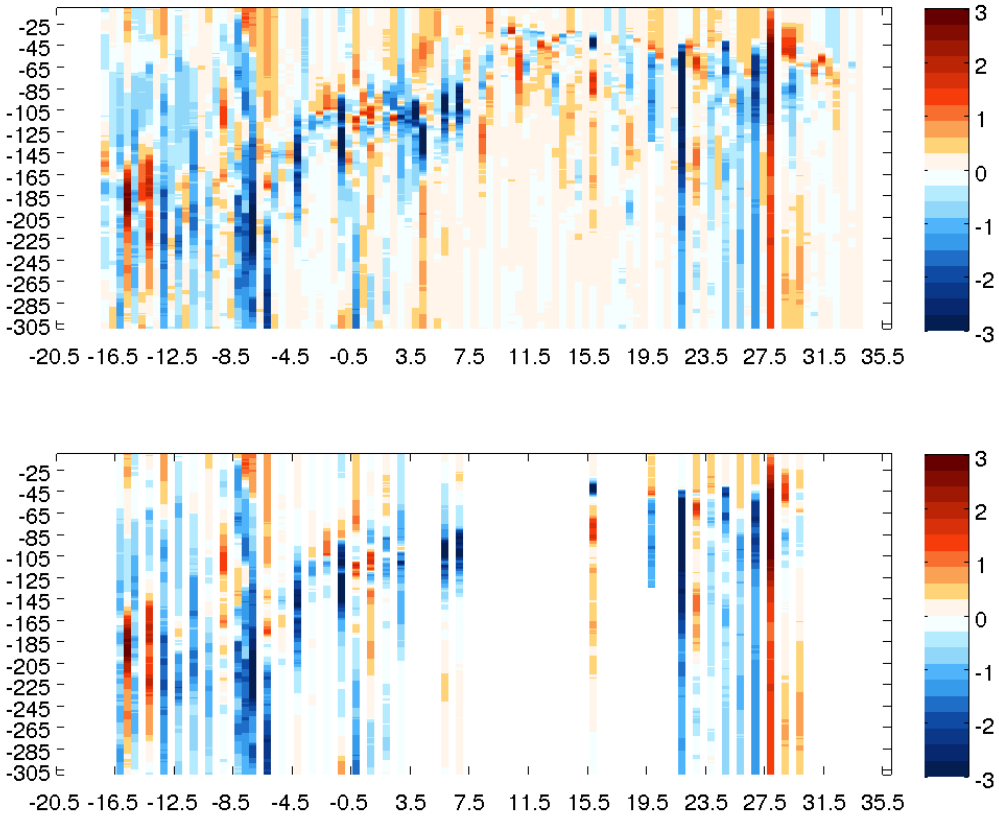


Figure 17: The difference in values between the reconstruction and the complete ARAMIS 9 mission profiles is shown in figure 17 (a). A zoom on the errors over the values present in the validation set is shown in figure 17 (b). The colorbar values are in °C. The vertical axis corresponds to the depth in meters while the horizontal one corresponds to the latitude over the ARAMIS transect.

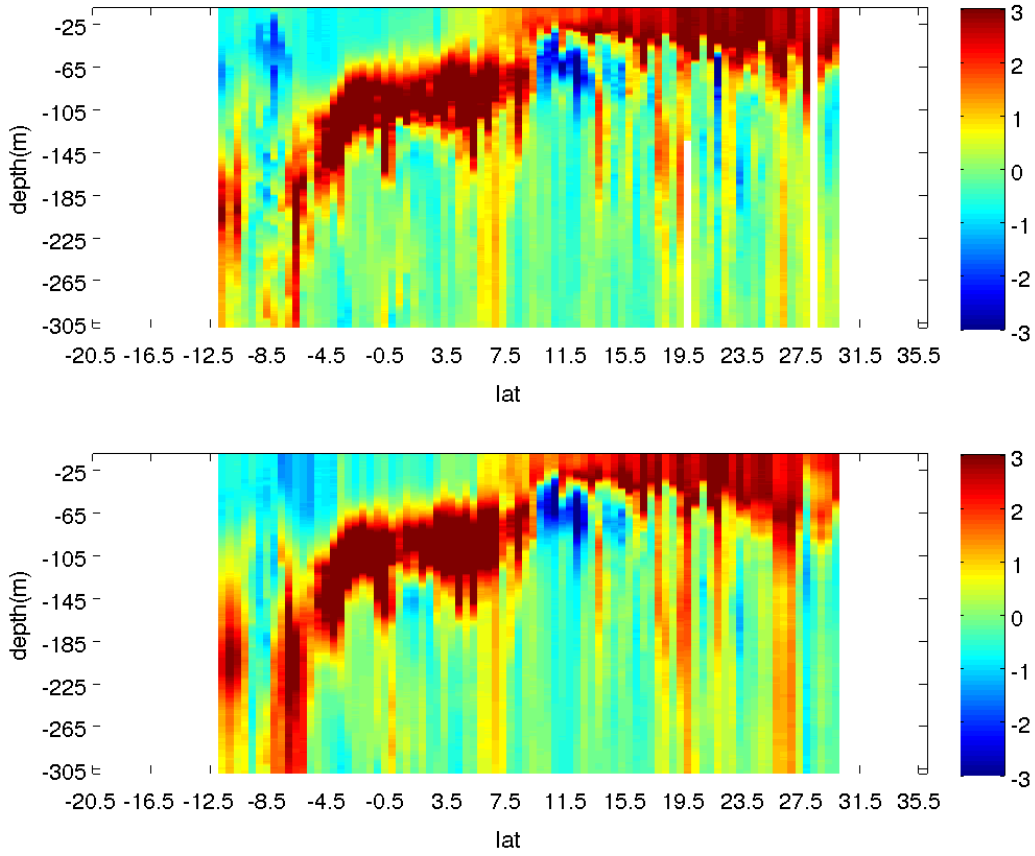


Figure 18: In figure 18 (a) we can see the complete data set minus the “average year”, and in figure 18 (b) the reconstruction minus the “average reconstructed year”. The colorbar values are in °C.

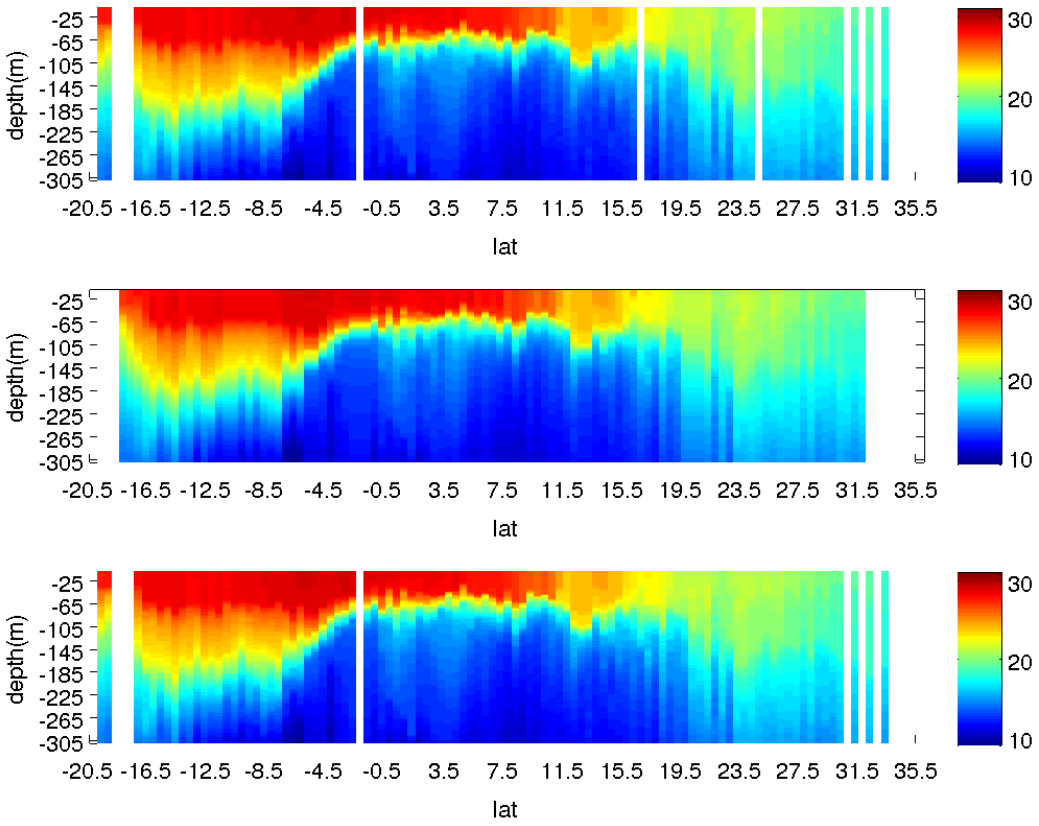


Figure 19: The vertical distribution of the validation data set of the ARAMIS 10 mission for every available latitude where such a profile was provided in the training data set is seen in figure 19 (a), the complete ARAMIS 10 MISSION profiles in figure 19 (c) and the reconstruction of the ARAMIS rail based solely on surface data by the PROFHMM method in figure 19 (b). The colorbars are in °C.

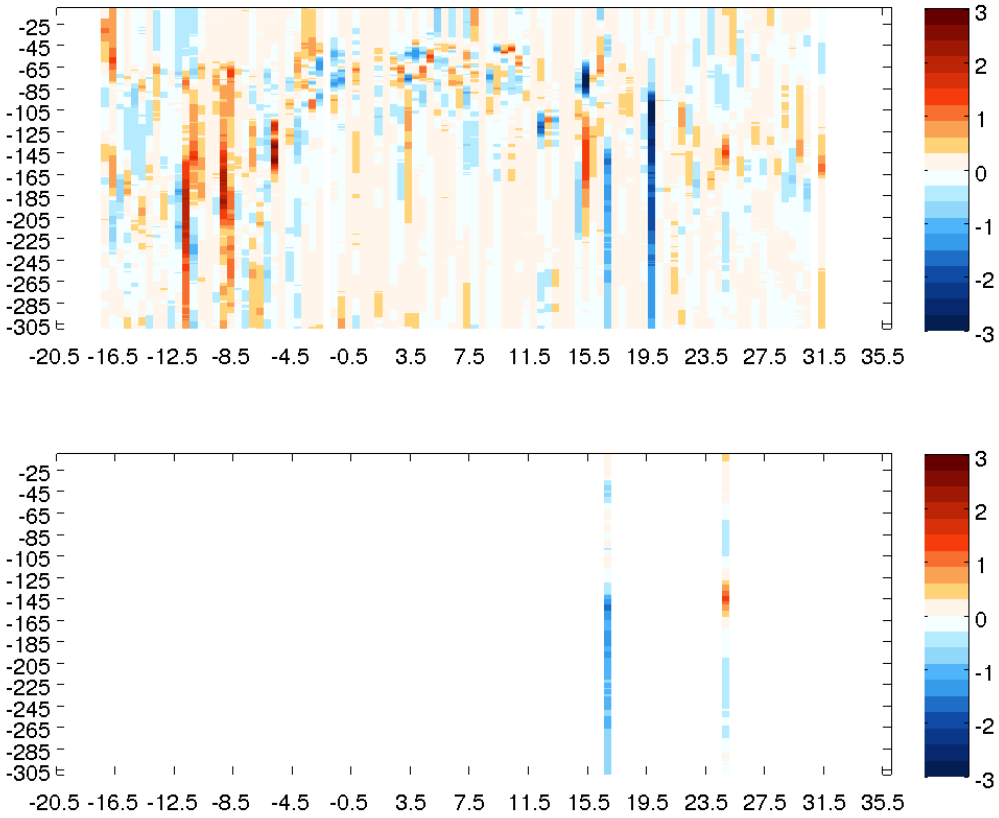


Figure 20: The difference in values between the reconstruction and the complete ARAMIS 10 mission profiles is shown in figure 20 (a). A zoom on the errors over the values present in the validation set is shown in figure 20 (b). The colorbar values are in °C. The vertical axis corresponds to the depth in meters while the horizontal one corresponds to the latitude over the ARAMIS transect.

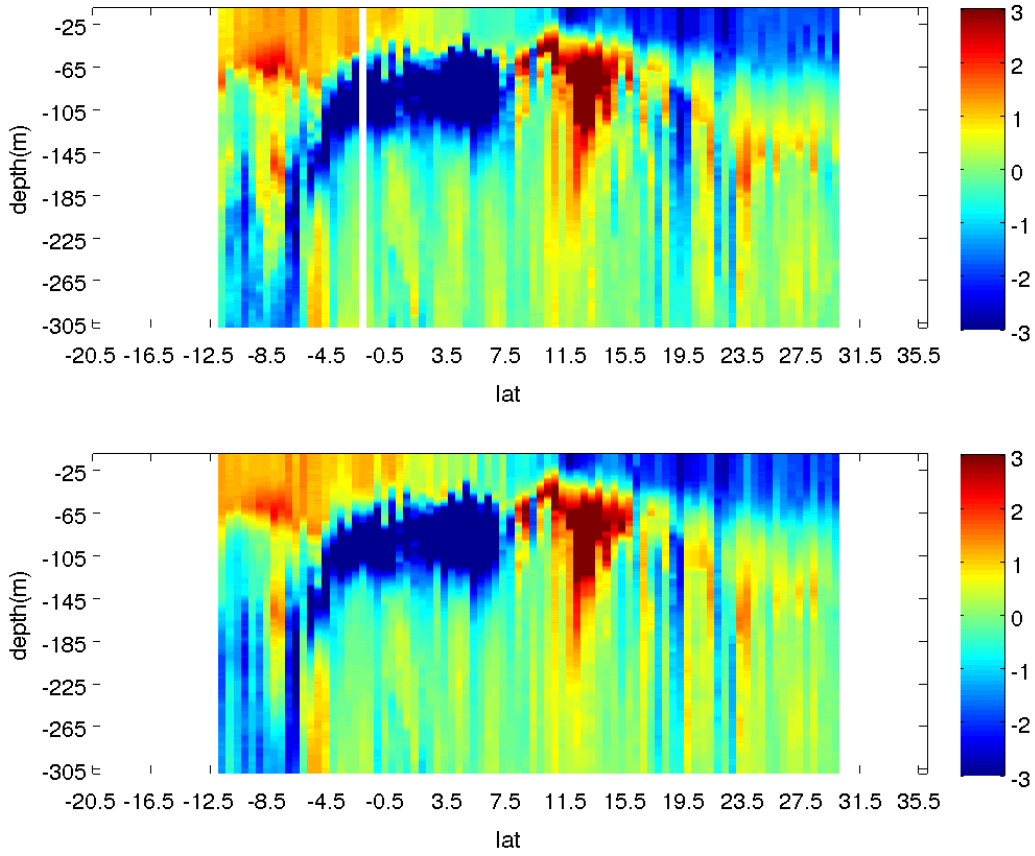


Figure 21: In figure 21 (a) we can see the complete data set minus the “average year”, and in figure 21 (b) the reconstruction minus the “average reconstructed year”. The colorbar values are in °C.

ARAMIS 11

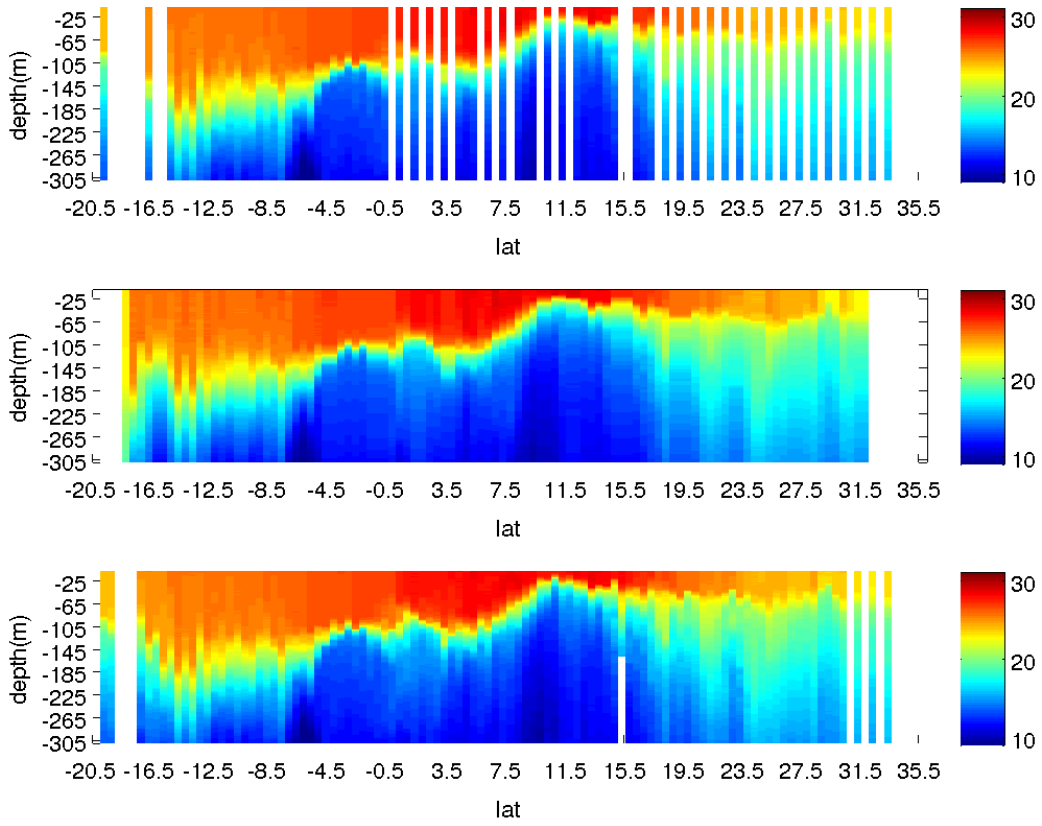


Figure 23: The vertical distribution of the validation data set of the ARAMIS 11 mission for every available latitude where such a profile was provided in the training data set is seen in figure 23 (a), the complete ARAMIS 11 MISSION profiles in figure 23 (c) and the reconstruction of the ARAMIS rail based solely on surface data by the PROFHMM method in figure 23 (b). The colorbars are in °C.

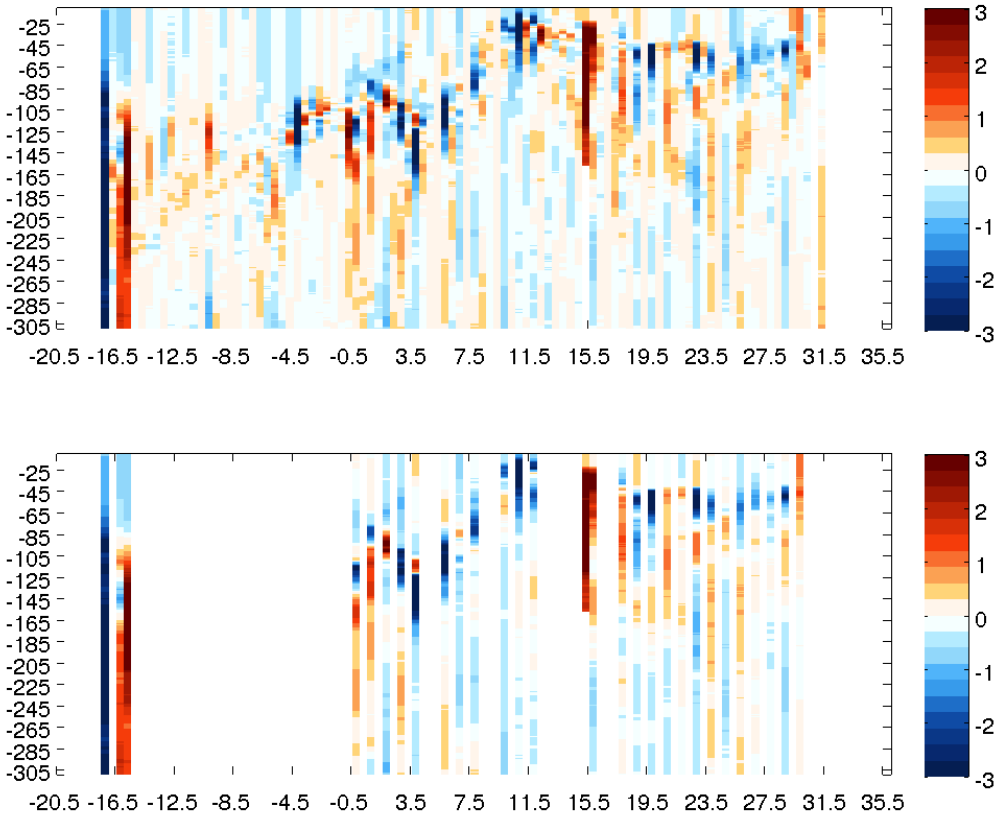


Figure 23: The difference in values between the reconstruction and the complete ARAMIS 11 mission profiles is shown in figure 23 (a). A zoom on the errors over the values present in the validation set is shown in figure 23 (b). The colorbar values are in °C. The vertical axis corresponds to the depth in meters while the horizontal one corresponds to the latitude over the ARAMIS transect.

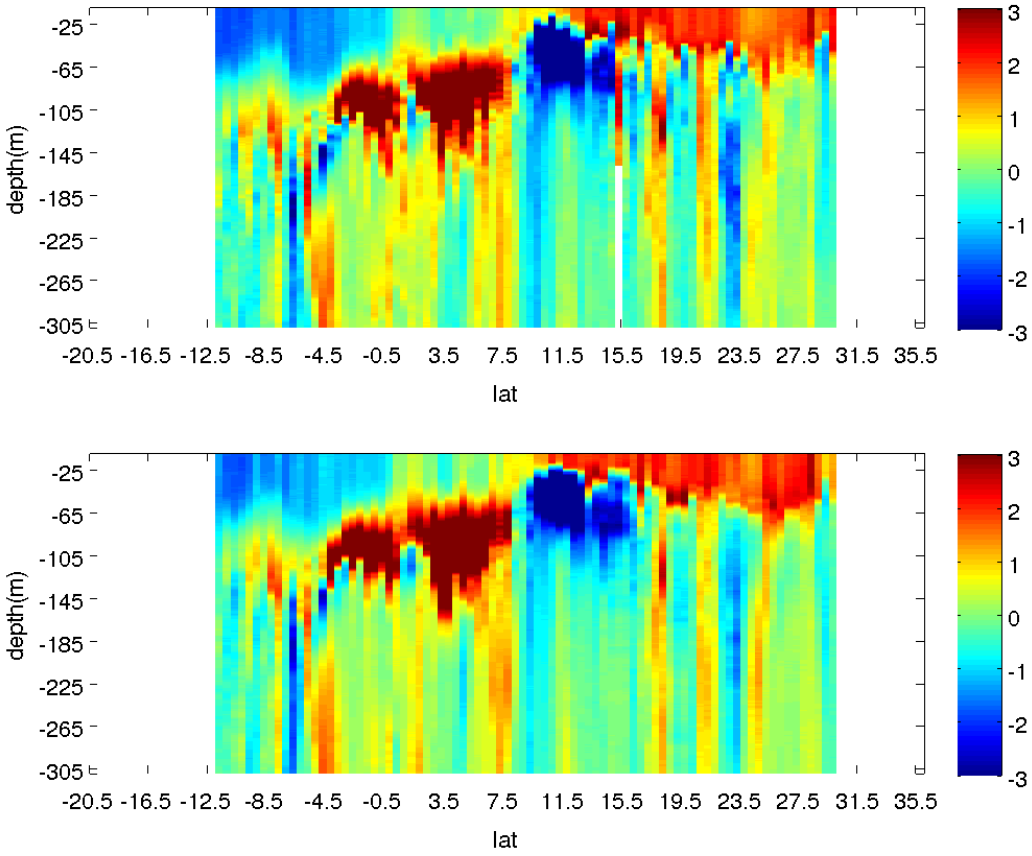


Figure 24: In figure 24 (a) we can see the complete data set minus the “average year”, and in figure 24 (b) the reconstruction minus the “average reconstructed year”. The colorbar values are in °C.

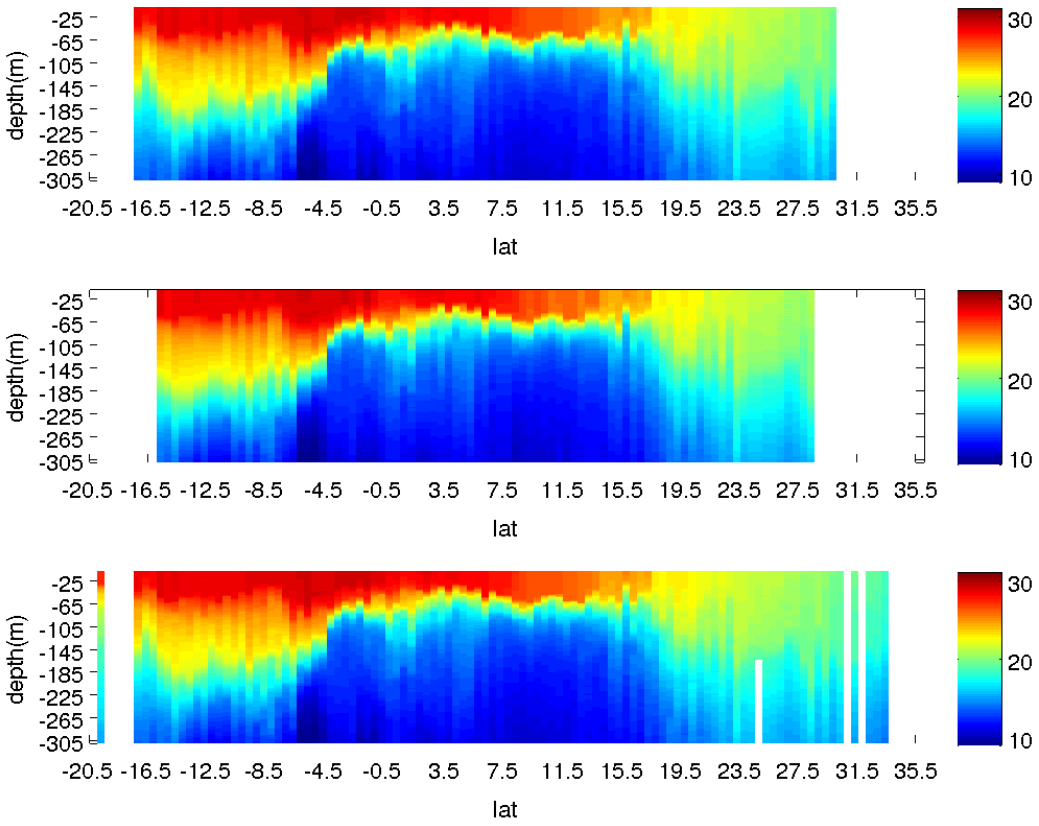


Figure 25: The vertical distribution of the validation data set of the ARAMIS 12 mission for every available latitude where such a profile was provided in the training data set is seen in figure 25 (a), the complete ARAMIS 12 MISSION profiles in figure 25 (c) and the reconstruction of the ARAMIS rail based solely on surface data by the PROFHMM method in figure 1 (b). The colorbars are in °C.

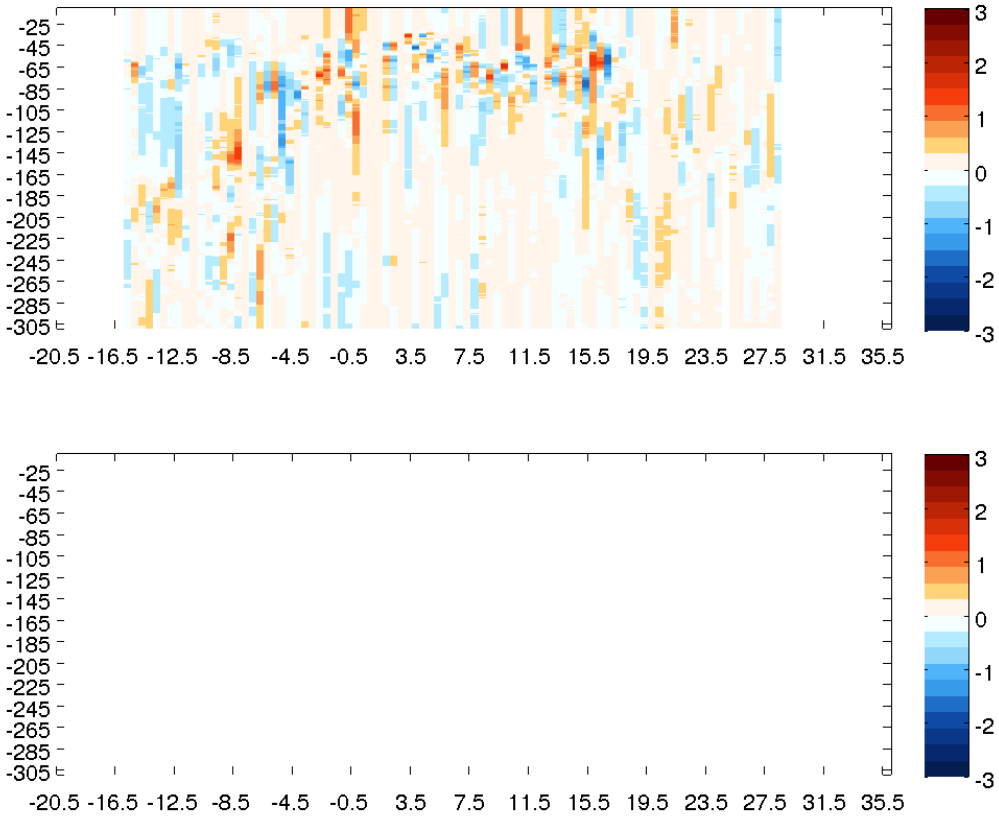


Figure 26: The difference in values between the reconstruction and the complete ARAMIS 12 mission profiles is shown in figure 26 (a). A zoom on the errors over the values present in the validation set is shown in figure 26 (b). The colorbar values are in °C. The vertical axis corresponds to the depth in meters while the horizontal one corresponds to the latitude over the ARAMIS transect.

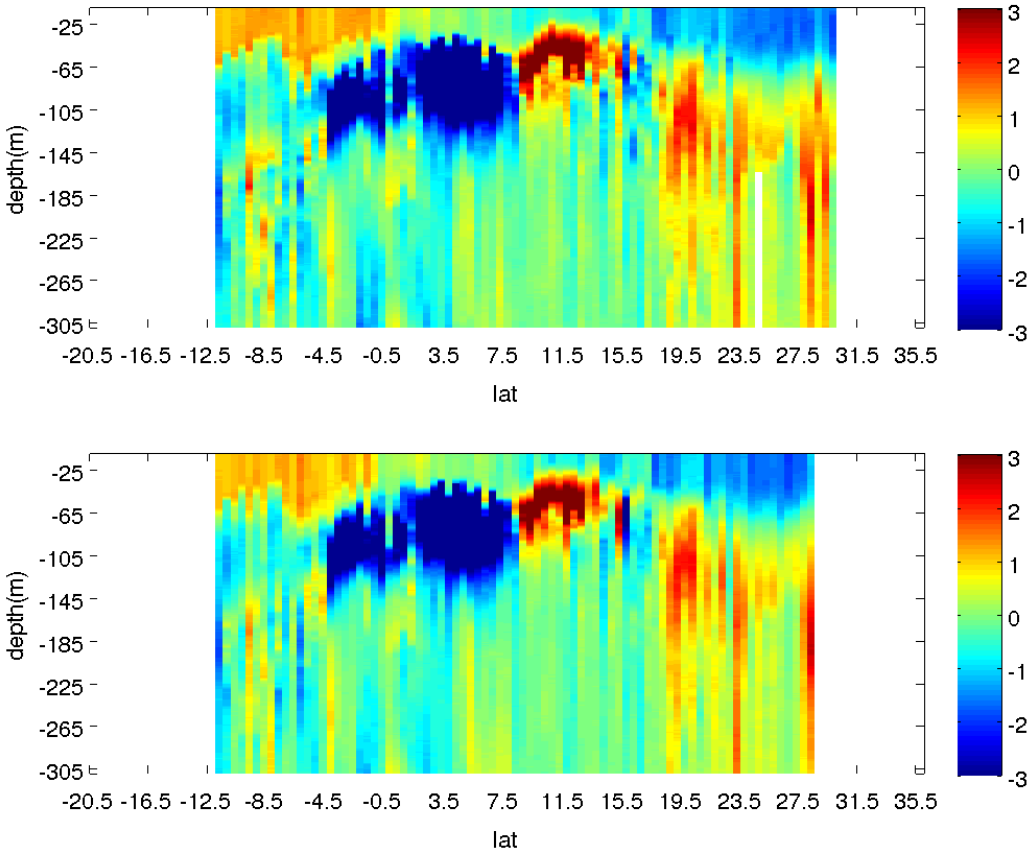


Figure 27: In figure 27 (a) we can see the complete data set minus the “average year”, and in figure 27 (b) the reconstruction minus the “average reconstructed year”. The colorbar values are in °C.

CHAPITRE 4 : MODIFICATION DE L'ALGORITHME DE VITERBI POUR PRENDRE EN COMPTE UNE CONNAISSANCE A PRIORI SUR LA CONFIANCE AUX OBSERVATIONS

4.1 Introduction :

Durant les applications précédentes de PROFHMM nous avons traité des séries d'observations continues :

- Dans le cas de la reconstitutions des profils verticaux de concentration de Chlorophylle-A nous avons complété les données manquantes en surface par des splines, qui sont des fonctions d' interpolation définies par morceaux à l'aide de polynômes. Seules les données manquantes ont été ainsi interpolées : nous avons conservé l'ensemble des données MODIS existantes même si celles-ci étaient de manière évidente erronées. D'autre part les splines introduisent un grand nombre d'erreurs sur les débuts et fins des séries interpolées.
- Dans le cas de la reconstitution des profils verticaux de température de ARAMIS, les observations de surface étaient issue de analyse optimales effectuées par l' équipe AVISO.

Étant donné la nature des données qui présentent des défauts dus à la précision des mesures et aux traitements que nous leur avons fait subir, il existe un degré d'incertitude sur chaque donnée qu'il peut être possible de prendre en compte.

En remarquant que PROFHMM était déjà capable de mémoriser la dynamique d'un système complexe, nous avons considéré qu'il était possible d'améliorer la complétion des données afin de prendre en compte une confiance au niveau de chaque observation. Pour cela nous présentons dans le quatrième chapitre une deuxième méthodologie que nous appellerons PROFHMM_UNC. Cette méthode utilise toute la modélisation introduite dans PROFHMM et modifie une partie de celle-ci pour introduire une expertise extérieure.

La modification porte sur l'algorithme de Viterbi et permet de compléter les séries de données initiales en estimant les données manquantes et en corrigent les données en fonction de la

confiance qui leur est attribuée. On modélise alors le problème sous la forme d'une chaîne de Markov cachée proche de celle de PROFHMM, mais maintenant :

- Les états cachés sont générés par une carte topologique entrainée avec les séquences complètes des données disponibles,
- Les états observés sont générées par une carte topologique entrainée avec toutes les données disponibles.

Par la suite, en introduisant une fonction de confiance aux observations et une fonction de pondération qui utilise cette confiance dans l'algorithme de Viterbi, nous pouvons compléter et corriger les séquences d'observations.

4.2 ARTICLE 3: PROFHMM_UNC: Introducing a priori knowledge for completing missing values of multidimensional time-series.

Résumé: Nous présentons une nouvelle méthode pour estimer les valeurs manquantes ou pour corriger les observations de mauvaise qualité de séquences de données manquantes. Cette méthode, basée sur les chaînes de Markov cachées et les cartes topologiques auto-organisatrices, est nommée PROFHMM_UNC. Les cartes topologiques sont utilisées pour discrétiser l'ensemble des sorties modèle et ainsi générer les états de la chaîne de Markov cachée. PROFHMM_UNC introduit une connaissance statistique du phénomène étudié dans le processus de reconstruction de la série tronquée d'observations. Pour ce faire, une modification à l'algorithme de Viterbi est introduite. Elle force l'algorithme à prendre en compte un information à priori de la qualité de observations durant la phase de reconstitution du chemin optimal. La validité de la méthodologie est démontrée par une expérience jumelle effectué sur les sorties du modèle numérique de la bio-géochimie marine NEMO-PISCES

PROFHMM_UNC: Introducing a priori knowledge for completing missing values of multidimensional time-series.

A A CHARANTONIS[1], F BADRAN[2], S THIRIA[1]

[1] Laboratoire d'Océanographie et du Climat - Expérimentation et Approches Numériques, Université Pierre et Marie Curie - Tour 45, 5-ème étage 4, place Jussieu, 75005 Paris, France

[2] Laboratoire CEDRIC, Conservatoire National des Arts et Métiers - 292, rue Saint Martin, 75003 Paris, France

E-mail: anastase-alexandre.charantonis@locean-ipsl.upmc.fr

Abstract: We present a new method for estimating missing values or correcting unreliable observed values of time dependent physical fields. This method, is based on Hidden Markov Models and Self-Organizing Maps, and is named PROFHMM_UNC. PROFHMM_UNC combines the knowledge of the physical process under study provided by an already known dynamic model and the truncated time series of observations of the phenomenon. In order to generate the states of the Hidden Markov Model, Self-Organizing Maps are used to discretize the available data. We make a modification to the Viterbi algorithm that forces the algorithm to take into account a priori information on the quality of the observed data when selecting the optimum reconstruction. The validity of PROFHMM_UNC was endorsed by performing a twin experiment with the outputs of the ocean biogeochemical NEMO-PISCES model.

1.Introduction

Initialization is one of the main factors for the computation of accurate predictions in most of the numerical prediction models. Some of these models require a complete time-sequence in order to generate their predictions. Time series encountered in many research fields, however, often contain missing or unreliable data due to reasons such as malfunctioning sensors and human factors. The issue of completing such multidimensional time series has been addressed by many different statistical or machine learning methods, such as the Maximum likelihood algorithm [1], expectation maximization algorithm [2], K-Nearest Neighbor [3], Varies Windows Similarity Measure [4] or Regional Gradient Guided Bootstrapping [5]. All these methods tend to reconstruct missing data that are subsequently used by the

corresponding prediction models. Thus the reconstruction of the initial time-series is disconnected from the dynamic model.

Most dynamic numerical models that have been developed over the years can reproduce the available observations of the phenomena under study, with varying degrees of success. In Geophysical sciences there exists a large amount of data sets and dynamic models [6] related to different physical phenomena. The accuracy of such numerical models is measured by comparing their output values to these observations. After the initial implementation of the model, there are often further studies that use the available data sets and the first implementation of the model in order to modify its internal parameters and improve its accuracy. The most prominent field of study attempting to combine model and data for improving our knowledge of the phenomena under study is data assimilation [7].

In this paper, we present a new method, which we will referred to as PROFHMM_UNC, for "PROFile reconstruction with HMM, taking into account UNCertainties" that combines the dynamic of the model and the available time series of observations in order to estimate the missing values or correct unreliable observed values. This is done by simplifying the dynamic model by transforming it into a multiple-state Hidden Markov Model (HMM). The reconstruction of the missing values and correction of the unreliable observations is done by applying a modified version of the Viterbi algorithm[8] which we introduce in this paper. This modification we introduce to the Viterbi algorithm uses a specific weighting function that modifies, during the optimum path selection process, the impact of the emission probability of an observation based on it's a priori confidence. PROFHMM_UNC makes use of Self Organizing Maps to generate the hidden and observable states of the model as used in PROFHMM [9], or SOS-HMM [10].

In the following, we present the general methodology we used to achieve that task and give an example of its implementation by performing a twin experiment for reconstructing the oceanic sea-surface Chlorophyll-A distributions and sea-surface Temperature based on the NEMO-PISCES model [11].

2.METHODOLOGY

The general theory behind the Hidden Markov Models is given in this section, followed by the introduction of our proposed modification to the Viterbi algorithm. This modification is used by the HMM for finding the most likely sequence of hidden states that results in a given sequence of observed events, when given an external indicator of the confidence in the data. We then briefly overview some of the advantages of discretizing multidimensional models into states through the use of self-organizing maps when trying to translate it into an HMM.

2.1 HIDDEN MARKOV MODEL

2.1.1 MODELISATION

A first order Markov model is a stochastic model made of a set of possible states $X_i\ i \in [1, \cdots, N_{hid}]$, and a transition probability matrix, noted Tr. First order Markov models assume the first order Markovian property, meaning that each consecutive state of the model depends solely on its previous state. Therefore the transition probabilities of a temporal sequence of states $X_{i_t}, t \in [1, \cdots, T]$, which are noted a_{ij}, are equal to $a_{ij} = Tr(i,j) = P\left(X_{i_t} = X_j \middle| X_{i_1} \cdots X_{i_{t-2}} X_i\right) = P\left(X_{i_t} = X_j \middle| X_{i_{t-1}} = X_i\right)$. The transition probabilities are considered invariant with time. Tr corresponds to a statistical learning of the dynamic processes governing the temporal transitions between these states.

Expanding this principle, a Hidden Markov Model (HMM) is a stochastic model with two sequences: one sequence of unobservable states, and one sequence of observations that have a statistical link with the unobservable states. We will henceforth refer to the unobservable states as hidden states, and symbolize them with X_i. The hidden states are assumed to follow the first order Markovian property.

The observations are linked with the unobservable states through a probability density function or matrix. This density function, or the probability matrix elements, correspond to the existing links between the observations and the unobserved states, and are referred to as emission probabilities. The probability of having observed an observation, *Obs*, given that we are in the state *i* is called its emission probability, and is denoted $b_i(Obs)$. In the following we chose to restrict our presentation to a HMM with discrete observable states, noted $Y_k, k \in [1, \cdots, N_{obs}]$. A hidden state X_i emits its observations according to an emission probability matrix, noted Em. The matrix elements connects the hidden states X_i to the observable ones such as $Em(i,k) = b_i(Y_k) = P(Y_k|X_i)$.

All the probabilities are determined during the training phase, by using an appropriate data set containing concurrent sequences of observed and known hidden states.

2.1.2 RECONSTRUCTION

After having determined the transitions and emissions probabilities, the Viterbi algorithm is then applied to find the most likely sequence of hidden states, given a sequence of concurrent observations. This is done by calculating, for each step of the observed sequence, the most likely sequence of states to end up at a given state, given the sequence of observations obtained up to that moment. The algorithm stocks these probabilities in a matrix, and the indexes of the states that generate these maximum probabilities for each state in another matrix. The algorithm then backpropagates to find the most likely sequence of indexes to have generated that sequence of observations.

The maximum probability to reach state *j* at time *t* is noted $\delta_t(j)$ and can be formulated as: $\delta_t(j) = \left(\max_{1 \le i \le N}\left[\delta_{t-1}(i) * a_{ij}\right]\right) * b_j(o_t)$, with o_t corresponding to the observation at time *t*. We use the matrix

$\Psi_t(j) = arg\max_{1 \le i \le N}[\delta_{t-1}(i) * a_{ij}]$, to store the index of most likely previous state of the Markov model to reach the state *j* at time *t*. This is primarily used when backpropagating through the algorithm in order to generate the most likely sequence of hidden states, which is noted q_t. The probability of the sequence $q_T = arg\max_{1 \le i \le N}[\delta_T(i)]$ is noted P, and T correspond to the index of the final time-step. The complete algorithm is shown in figure 1, and a graph representation of an HMM is shown in figure 2.

The Viterbi Algorithm

Initialization:

For $1 \le i \le N_{hid}$

$$\delta_1(i) = \pi_i * b_i(o_1),$$
$$\Psi_1(i) = 0$$

with π_i the initial probabilities the state i.

Iterative calculation

For $2 \le t \le T$

For $1 \le j \le N_{hid}$

$$\delta_t(j) = \left(\max_{1 \le i \le N_{hid}}\left[\delta_{t-1}(i) * a_{ij}\right]\right) * b_j(o_t)$$
$$\Psi_t(j) = arg\max_{1 \le i \le N_{hid}}\left[\delta_{t-1}(i) * a_{ij}\right]$$

Ending values

$$P = \max_{1 \le i \le N_{hid}}[\delta_T(i)]$$
$$q_T = arg\max_{1 \le i \le N_{hid}}[\delta_T(i)]$$

Backpropagating

For t=T to 2

$$q_{t-1} = \Psi_t(q_t)$$

Figure 1 The Viterbi Algorithm

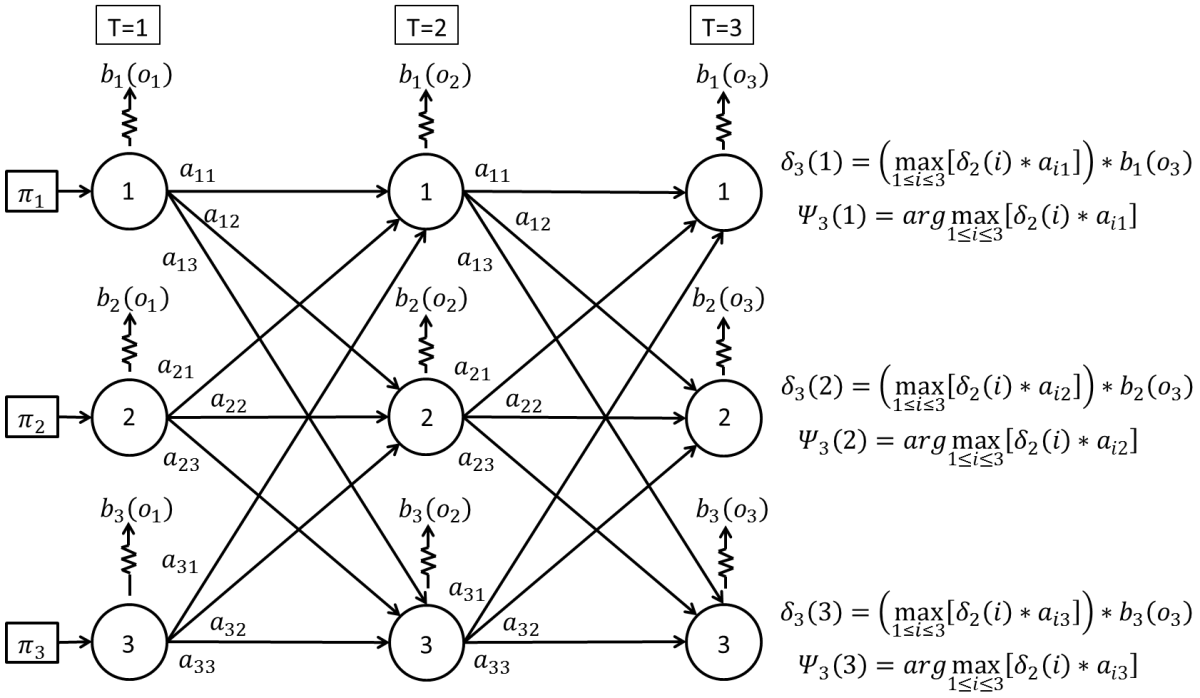


Figure 2: A graph of the evolution of an HMM with 3 hidden states over three time-steps. The Viterbi Algorithm calculated the $\delta_t(i)$ at each time-step t = 1,2,3 the maximum probability to reach a the state i, given the observations up to that point, and kept the indexes $\Psi_t(i)$ of the previous state that generated this maximum probability. When it reaches the final step it finds the index that generates the maximum $\delta_t(i)$, and backpropagates through the most likely states to have generated it.

2.2 TAKING INTO ACCOUNT UNCERTAINTIES

When applying HMMs it is generally assumed that the observation acquisition procedure and quality remain constant. However there are cases for which a combination of human errors and exterior parameters interfere and prevent the obtaining of sequences in which we have complete confidence in.

Given a Hidden Markov Model for which there exists a method for determining observation probabilities $b_i(Obs)$ for which we have full confidence on the observation, we present a modification of the Viterbi algorithm which takes into account a change of confidence in a given observation.

To do so, we first introduce a confidence function, named $conf(Obs)$. This function gives an external numerical evaluation of the quality of the observation. The $conf(Obs)$ function is scaled from 0 to 100, with 0 corresponding to a complete lack of confidence in the data (or a lack of data), and 100 corresponding to acquisition of a fully-trustworthy observation.

The confidence function is used, along with the $b_i(Obs)$ by a weighting function $F_w\big(b_i(Obs), conf(Obs)\big)$, in order to introduce in the HMM the confidence we have in the observed data.

The function $F_w\big(b_i(Obs), conf(Obs)\big)$ needs to be monotonically decreasing for both $b_i(Obs)$ and $conf(Obs)$ and takes values from 1 in the case of a non-trustable observation ($conf(Obs) = 0$) up to $b_i(Obs)$ for a fully trustable observation ($conf(Obs) = 100$). A typical form of F_w which is parameterized by different values of $b_i(Obs)$ can be seen in figure 3. Other functions could be chosen depending on the a priori information we want to introduce.

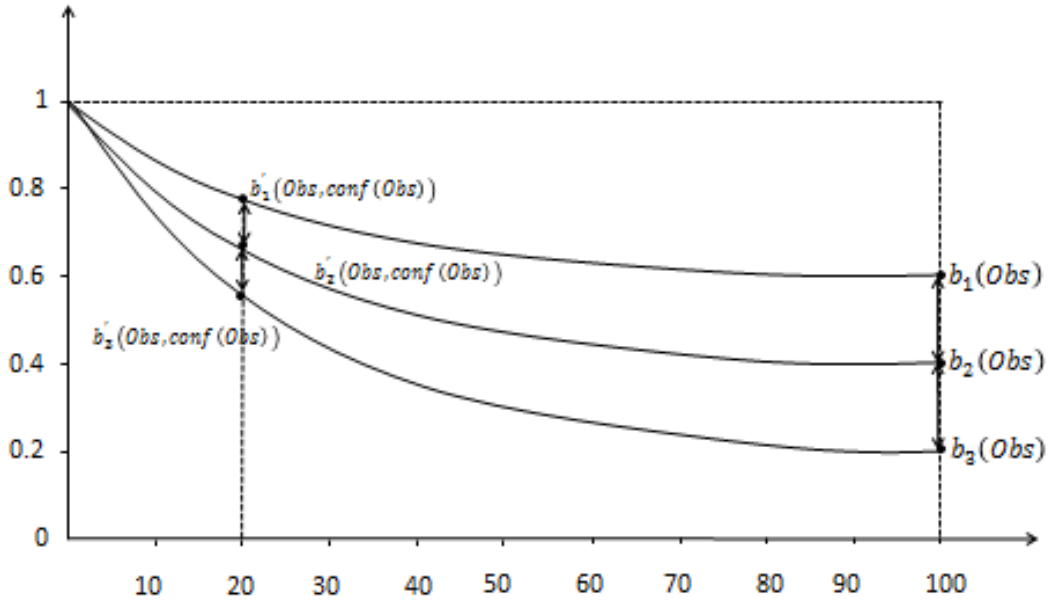


Figure 3 $F_w\big(b_i(Obs), conf(Obs)\big)$ for different values of $b_i(Obs)$, with respect to $conf(Obs)$. The x axis represents $conf(Obs)$.

The two functions are introduced in the Viterbi Algorithm when calculating the maximum probability to reach the state *j* at time *t*, by transforming it, from $\delta_t(j) = \big(\max_{1\le i\le N}\big[\delta_{t-1}(i) * a_{ij}\big]\big) * b_i(o_t)$, to $\delta_t(j) = \big(\max_{1\le i\le N}\big[\delta_t(i) * a_{ij}\big]\big) * F_w\big(b_i(o_t), conf(o_t)\big)$. This corresponds to the transformation of the probability $b_i(o_t)$ into a weighting term $b_i'\big(o_t, conf(o_t)\big) = F_w\big(b_i(o_t), conf(o_t)\big)$, which is no longer a probability.

Given a number of states with increasing a priori emission probabilities $b_i(o_t)$ for the observation o_t, their weighting terms $b_i'\big(o_t, conf(o_t)\big)$ will remain ordered in the same way for any non-null value of $conf(o_t)$. Since all a priori emission probabilities $b_i(o_t)$ are calculated from the same observation o_t vector, they will also have the same confidence $conf(o_t)$. We note that a decrease of the common value of $conf(o_t)$ increases all the weighting terms according to the curves representing the F_w function family (Figure 3). As $conf(o_t)$ decreases, the weighting terms $b_i'\big(o_t, conf(o_t)\big)$, converge towards 1, therefore

progressively decreasing the impact of the a priori probabilities $b_i(o_t)$ in the path selection of the Viterbi algorithm. A visual representation of one possible form of the F_w functions family is shown in figure 3

When the confidence is null, $(conf(o_t) = 0)$, the function $F_w(b_i(o_t), 0) = 1$, and therefore $\delta_t(j) = \max_{1 \le i \le N}[\delta_t(i) * a_{ij}]$, making the determination of the best path at the time *t* depend solely on the transition probabilities. Similarly, when the confidence is maximal, $(conf(o_t) = 100,)$ the function $F_w(b_i(o_t), 100) = b_i(o_t)$ and the determination of the best path at time *t* is done with the same process as with the unmodified Viterbi algorithm. The modification therefore can be considered a trade-off function between a regular Markov model and an HMM.

This method is general and can be applied in situations for which we have a high degree of confidence in the model and an exterior indicator of the quality of the observations.

2.3 USE OF SELF ORGANIZING MAPS WITH HIDDEN MARKOV MODELS

In order to build the HMM to model such a problem, it is necessary to discretize the dynamical model outputs into a discrete set of states. This can be a complicated task. A common method which is used in cases were the dynamical model can be described with few states, is to create independent states by clustering the available data [12]. A way to cluster the data is to reduce the dimension of the problem by using a Principal Component Analysis (PCA) [13], and then make use of Learning Vector Quantization to generate states of the HMM [14]. However reducing of the dimension of the data through a PCA, would hinder the reconstruction of highly multidimensional vectors, and would not permit a fine discretization of the data, which is important in time-series completion. In our method we use Self-Organizing Topological Maps (SOM) which are clustering methods based on neural networks [15]. They provide a discretization of a learning dataset into a reduced number of subsets, called classes, which share some common statistical characteristics. Each class corresponds to an index. For each class, the hidden data attributed to it is represented by a referent vector, which approximates the mean value of the elements belonging to it. These referent vectors are used to label any other data of the same dimension with the index of the "nearest" referent vector. The indexes represent the states of the HMM and their referent vectors are used to generate sequences of indexes of states that serve to learn the emissions and transition probabilities.

The SOM training algorithm forces a topological ordering upon the map, and therefore any neighboring classes have referent vectors that are close in the Euclidean sense in the data space. This particularity is used by PROFHMM, and by extension by PROFHMM_UNC, to improve the emissions and transitions probabilities of the HMM. It permits the inclusion of a high number of states in the HMM modeling of a phenomenon for which we have relatively few concurrent hidden and observable vectors. The process of improving these probabilities is detailed in the Annex.

3. APPLICATION

PROFHMM_UNC can be applied to real-world data for which we have a model that is consistent with the observed quantities. However, for the scope of this article we chose to perform a twin experiment with the outputs of the NEMO-PISCES model [10], which allow us to present the general behavior and some quantified performances of the PROFHMM_UNC. Doing so we can control the behavior of PROFHMM_UNC for different situations: low or high confidence.

3.1 THE MODEL

NEMO-PISCES is an ocean modeling framework which is composed of "engines" nested in an "environment". The "engines" provide numerical solutions of ocean, sea-ice, tracers and biochemistry equations and their related physics. The "environment" consists of the pre- and post-processing tools, the interface to the other components of the Earth System, the user interface, the computer dependent functions and the documentation of the system. We obtained the output of this model by running the ORCA2_LIM_PISCES version of NEMO, which is a coupled ocean / sea-ice configuration based on the ORCA tripolar grid at 2° horizontal resolution forced with climatological forcing (winds, thermodynamic forcing) in conjunction with the PISCES biogeochemichal model [10].

We extracted the five-day averaged outputs of this model at the grid points representing the BATS station (32 N -64 W), shown in figure 4. This station is one of the model calibration sites due to the existence of the Bermuda Atlantic Time Series (BATS) of the JGOFS campaign [12]. From the available data, we processed the values of the Sea Surface Temperature (SST), Sea Surface Chlorophyll-A (SCHL), Wind Speed (WS), the incident Shortwave Radiation (SR) and Sea Surface Elevation (SSH), averaged every five days. These averaged time steps are denoted t_{NEMO}. This gave us a complete data set of these five parameters for 1239 t_{NEMO} time-steps spanning from 1992 to 2008. We then generated a matrix containing the mean value of these parameters averaged for three consecutive t_{NEMO} time-steps. This average corresponds to the mean values of the parameters for a fifteen consecutive-days period, denoted t_{HMM}. The data set containing the values of the five “observable” parameters at the different t_{HMM} time-steps is noted $Data_{hid}$ and is a five-dimensional matrix with 413 t_{HMM} time-steps.

Figure 4 the location of BATS.

In order to generate the observable situations and to simulate satellite data, we added to each geophysical parameter at the t_{NEMO} temporal resolution, a white noise following a Gaussian $N(0,0.35*\sigma_{param})$, where σ_{param} is the standard deviation of each parameter. The data was then once more averaged every 3 consecutive t_{NEMO} time steps, in order to reach the t_{HMM} temporal resolution. This generated the $Data_{obs}$ matrix.

3.2 STATISTIC LEARNING AND WEIGHTING FUNCTION CONFIGURATION

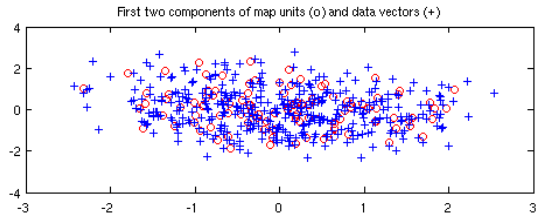
The SOM map (denoted $sMap_{hid}$ in the following) providing the hidden states of the HMM was trained with $Data_{hid}$. As described in section 2.3, by classifying the $Data_{hid}$ vectors, we generated a sequence of indexes, denoted SI_{hid}. These indexes correspond to the hidden states of the model at these consecutive t_{HMM} time-steps.

The SOM map (denoted $sMap_{obs}$ in the following) providing the observable states of the HMM was trained with $Data_{obs}$. Since the generation of $Data_{obs}$ included the calculation of the mean value at the t_{NEMO} temporal resolution, the white noises added to the signal is smoothed. $Data_{obs}$ is a five-dimensional data set (containing SCHL, SST, SSH, WS, SR) with 413 rows. The sequence of indexes of the observable data, SI_{obs} coincides temporally with SI_{hid}, and was generated by classifying the $Data_{obs}$ vectors.

The generation of the hidden and observable states was done by using the two SOMs. Both $sMap_{hid}$ and $sMap_{obs}$ contain 108 neurons that represent, respectively, the hidden and observable states of the HMM, distributed on an array formed of 12 by 9 lattices. They were generated using $Data_{hid}$ and $Data_{obs}$ from 1992 to 2005, each data set corresponding to 340 t_{HMM} time steps; the sequence of observations from the year 2006, corresponding to 24 t_{HMM} time steps, were used as a validation set, and the years 2007 and 2008, corresponding to a sequence of 49 t_{HMM} time steps, were used to test the performance of the method.

The size of the maps were set by iteratively increasing the number of states of each map, selecting the dimensions that had the smallest root mean square (RMS) errors between the actual data of the validation year 2006 and its reconstruction by PROFHMM.

In figure 5 (a) we present, projected on the first plane given by the PCA of $Data_{hid}$, the spatial distribution of the referent vectors of the hidden states as red circles, while the blue crosses correspond to the data vectors from $Data_{hid}$. Similarly, in figure 5 (b) we present, projected on the first plane of the PCA of $Data_{obs}$, the spatial distribution of the referent vectors of the hidden states as red circles, while the blue crosses correspond to the data vectors from $Data_{obs}$. The first plane of the PCA of the hidden data set corresponds to 69,3% of its variance, while the first plane of the PCA of the observable data set corresponds to 68,2% of its variance. Both hidden and observable states are well distributed over their respective data set. Therefore we can make the assumption that the selected states represent accurately the variance of the observed phenomenon.. It is important to note that this is just a projection of the data on the first plane and that we did not reduce the dimension of our vectors by applying this PCA.

(a) 

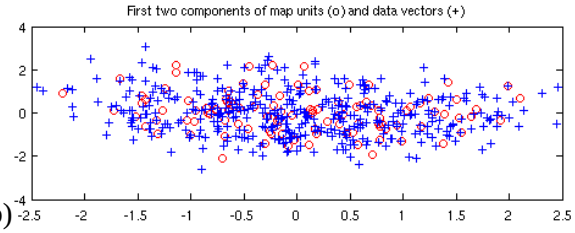
(b) 

Figure 5: (a) Projection of the temperature profiles (in blue crosses) and the referent vectors of $sMap_{hid}$ (in red circles), onto the first plane of the PCA of $Data_{hid}$. (b) Respectively, projection of the observation vectors (in blue crosses) and the referent vectors of $sMap_{obs}$ (in red circles), onto the plane determined by the two first eigenvectors of the PCA of $Data_{obs}$.

The SOM maps were trained wtih the algorithms provided by the matlab somtoolbox [15], specifically the functions som_make, som_batchtrain, som_bmus in order to train our maps and classify our data.

The SOM maps were used to classify the datasets and generate two sequences of state indexes, SI_{obs} and SI_{hid}. These sequences were subsequently used to train the Hidden Markov Model according to the procedure presented in section 2, and to estimate the HMM parameters.

The weighting function was set to $F_w(X,Y) = \frac{(1-X)*\exp(-0.035*Y)}{1-\exp(-3.5)} + \frac{X-exp(-3.5)}{1-\exp(-3.5)}$ whose form can be seen in figure 3. The determination of the form of this function and the specific values for the exponential, was a modeling choice made to force a slight degradation of $b_i(o_t)$ for high $conf(o_t)$, while greatly increasing them for very low ones.

We made the assumption that, due to exterior factors such as heavy cloud coverage or satellite instrument malfunction, only during some (or none) of these time-steps there were observations available. The confidence function was therefore defined as the percentage of available time-steps used to generate the observation. The flowchart of PROFHMM_UNC for the twin experiment is shown in figure 6.

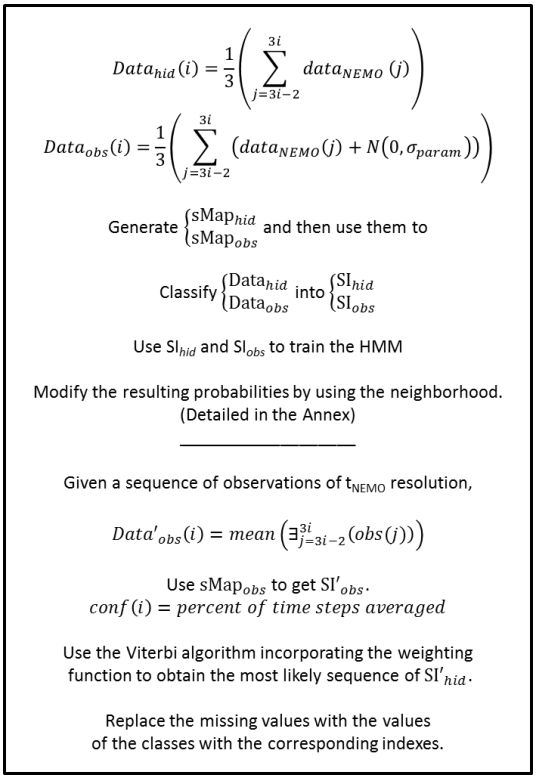


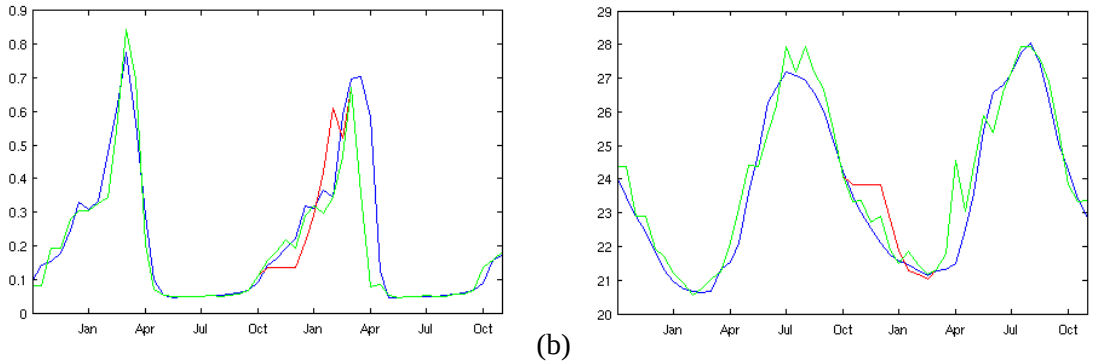
Figure 6: The flowchart of the twin experiment. The expression $mean\left(\exists_{j=3i-2}^{3i}(obs(j))\right)$ must be read as: "compute the mean for the existing values in the time sequence from $j = 3i - 2$ to $3i$".

3.3 PERFORMANCES

To test the performances of the model, we classified the hidden and observable data from the years 2007 and 2008 according to their respective sMaps. However, we simulated a perturbed sequence of data for which we introduced exterior indicators of confidence: for twelve consecutive t_{HMM} time steps we considered that the observable data was not given from the mean of three consecutive t_{NEMO} time steps, but by the value of only one of these. Doing so we increased the noise level of those specific data points. Therefore, an empiric way to set the confidence value $conf(Obs)$ at approximately a third of the maximum confidence, $conf(Obs)$=35, since we sampled the data at the rate of one out of the three consecutive time-steps.

The twelve consecutive time-steps data shown in figure 7, correspond to a period spanning from October 2007 to March 2008. We focused on the reconstruction of the Chlorophyll-A, in figure 7 (a) and the

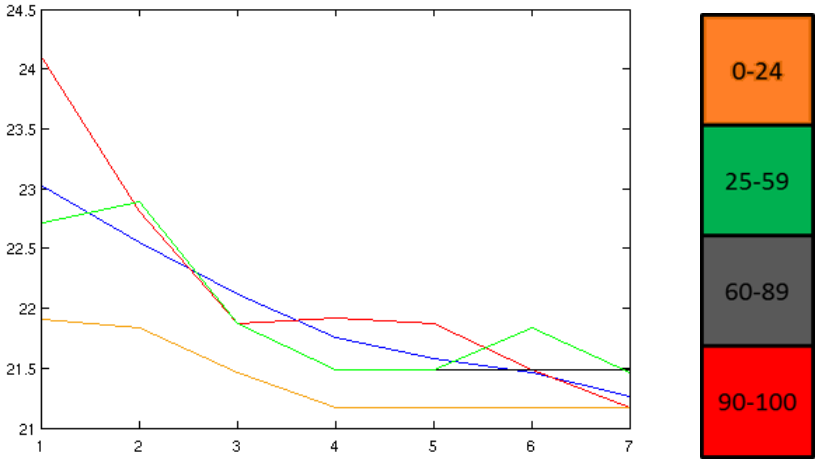
temperature, in figure 7 (b). The curves in this figure correspond to the reconstructions: in red for a complete confidence ($conf(Obs)$=100) in the observations and in green for the aforementioned $conf(Obs)$=35. The real model values are in blue. We can see that, by applying the PROFHMM_UNC, we increased our trust in the transitions probabilities of the HMM and better fitted the curve of the real data.



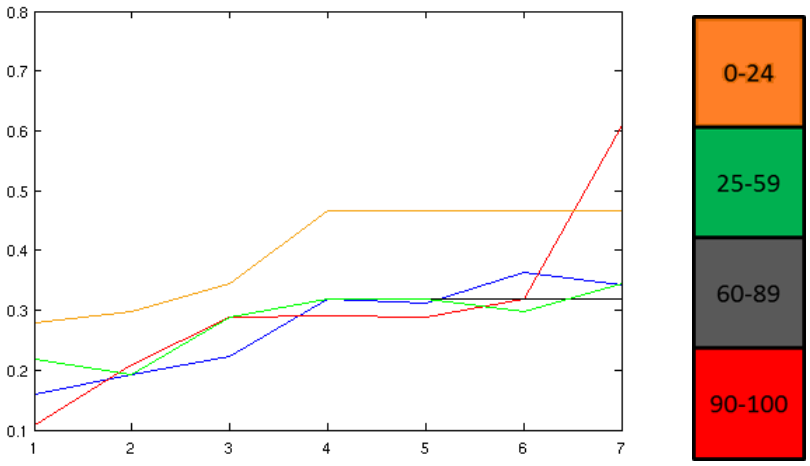
(a) (b)

Figure 7 (a) and (b) : Reconstruction, for the years 2007 and 2008, of Sea Surface Chlorophyll-A (a), in [10^{-6} mg/L] and Sea Surface Temperature (b), in °C. The blue line corresponds to the unmodified data of $Data_{hid}$. The red one corresponds to the values of the reconstruction while using the HMM without a modification of the emission probablity, considering that the modified observations have a confidence of 100. The green one corresponds to the result obtained by using PROFHMM_UNC, and using a confidence of 35 for each observation between, October and March. Out of that time period, the two curves coincide and we cannot differentiate the green and red curve.

After presenting these results, we progressively varied the value of $conf(Obs)$ by increments of 5, and plotted the resulting curves of CHL and SST for the 7 t_{NEMO} time step period which had their sampling modified. As seen in figure 8, we only obtain 5 different curves when performing this experience. In orange, we can see that if we give zero to small confidence in the data ($conf(Obs) \in [0 \cdots 24]$), we obtain a curve that almost does not take into account the observations and chooses the hidden states based on transitions only. The reconstruction therefore is far away from reality. In green we have the result obtained when we have a small, but not null confidence in the observations ($conf(Obs) \in [25 \cdots 59]$), this curve is closer to the NEMO values. In black, we have the values obtained with a higher confidence in the data ($conf(Obs) \in [60 \cdots 89]$). The black curve follows the NEMO-PISCES values quite well; is hidden by the green curve up to the fifth time step, then approximates the real data slightly better. Finally when we completely trust the data, ($conf(Obs) \in [90 \cdots 100]$) the model takes too much into account the modified observed data, increasing therefore the error of the reconstruction.



(a)



(b)

Figure 8 a) and b) : variation of the reconstructed values of Sea Surface Chlorophyll-A (a) and Sea Surface Temperature (b) with respect to the value of the $conf(Obs)$ function, for a 7 t_{HMM} period (from October 2007 to the first half of January 2008) assuming that each t_{HMM} observation is computed from a single t_{NEMO}. In blue we have the actual values of the model. The colorbar indicates the valued of $conf(Obs)$.

It is interesting to note that there are only 5 curves obtained when varying the values of $conf(Obs)$. This is due to the way the Viterbi Algorithm functions, whose principle is to select the optimum path: slight changes in the values of confidence will create slight modifications, which are often not enough to overcome a threshold value needed to change the index of the selected state, therefore generating the same path. The choice of the form of family of functions F_w controls the speed of the decrease of the impact of the a priori emissions probabilities on the selection of the optimum path by the Viterbi Algorithm. This, in

turn also changes the length and placement of the intervals of $conf(Obs)$ values that present the same reconstruction.

Figure 8 also raises another important point on the determination of the $conf(Obs)$ function. In the experiment above, we initially considered an empirical value of 35, trying to approximate the fact that we only sampled the data at the rate of one out of the 3 t_{NEMO} time steps. This was equivalent to assuming that they were independently selected using a simple probability distribution. However, as seen in figure 8, the results obtained in the interval $[60 \cdots 89]$ better fitted the data. This happens because the t_{NEMO} time step sampling is strongly correlated with the two ignored ones, and its associated data value contains a significant percentage of the total information we would have obtained by sampling all of the time steps. Therefore, it is important to perform a preliminary study to determine the confidence function that best fits the problem. This is especially true in the cases for which PROFHMM_UNC could be applied, since spatial and temporal sampling at different resolutions is often encountered in different fields of study, such as geophysical problems related to satellite information.

In order to present a quantifiable measure of the improvement obtained by applying PROFHMM_UNC, we performed a dedicated test by generating 10.000 different $Data_{obs}$. This was done while varying the placement of the consecutive 7 t_{HHM} time steps period of low confidence throughout the 2 testing year period. The $Data_{obs}$ matrices were generated by adding a, significantly stronger, white noise, that follows $N(0,1.5*\sigma_{param})$ distribution, to the $Data_{hid}$. We then calculated, for the years 2007-2008 the RMS errors between the reconstructed and NEMO outputs of the sea surface chlorophyll-a and sea surface temperature for each confidence interval. The results are shown in Table 1.

Table 1: RMS errors of the Chlorophyll-A and Temperature

	$conf(Obs) \in [90 \cdots 100]$		$conf(Obs) \in [60 \cdots 89]$		$conf(Obs) \in [25 \cdots 59]$		$conf(Obs) \in [0 \cdots 24]$	
$1.5*\sigma_{param}$	Chl-A	SST	Chl-A	SST	Chl-A	SST	Chl-A	SST
Max RMS	0.4828	5.6496	0.5496	4.7629	0.5771	5.2840	0.4972	4.3496
Min RMS	0.0013	0.1858	0.0011	0.0693	0.0011	0.1032	0.0013	0.1311
Mean RMS	0.1384	1.2895	0.1135	0.9303	0.1180	0.9919	0.1722	1.4349

The Chlorophyll-A values are in [10^-6 mg/L] while the temperature is in °C.

The values obtained indicate an improvement when taking into account the uncertainty of the observations. Once more we can see that, by applying $conf(Obs) = 70 \in [60 \cdots 89]$ the results are globally improved. This highlights the importance of performing a preliminary study to determine the appropriate confidence function for each phenomenon.

There exist variations of the Viterbi algorithm such as Lazy Viterbi [16] and the Soft Output Viterbi [17]. We limited ourselves to the Viterbi Algorithm, yet the modification could easily be applied to those approches.

4 CONCLUSIONS

In this paper we have presented PROFHMM_UNC, a new methodology, that combines a dynamic model and observations in order to constrain the outputs of a dynamic model to better fit the observations, while respecting the dynamic processes of the model. This was used to complete time-sequences with missing data, and to correct observations for which we have an external indicator of unreliability. The improvements obtained by using the method were illustrated through a twin experiment. The results of this experiment highlight the importance of performing a preliminary study to determine a case-appropriate confidence function. PROFHMM_UNC is very general and could be applied to model a system which is not described numerically, by learning the dynamic processes using a large amount of sequences of observations.

Going forward we might intend to apply this methodology for generating realistically complete time series of states, based on real satellite observations, to further be used by PROFHMM in order to retrieve 3D fields of parameters based on discrete observations generated by PROFHMM_UNC.

AKNOWLEDGMENTS

The research presented in this paper was financed by Centre National de l'Etude Spatial (CNES, French national center of spatial studies), and the Delegation Gouvernemantale pour l'Armement (DGA, French Military Research Delegation), which we wish to thank for their support. We wish to also thank Cyril Moulin and Laurent Bopp from LSCE for their help with NEMO-PISCES and Michel Crepon for his input on the method.

REFERENCES:

1 A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," Journal of the Royal Statistical Society. Series B, vol. 39, no. 1, pp. 1–38, 1977.

2 Z. Ghahramani and M. I. Jordan, “Supervised learning from incomplete data via an EM approach,” in Advances in Neural Information Processing Systems (NIPS 6), pp. 120–127, Morgan Kauffman, San Fransisco, Calif, USA, 1994.

3 I. Wasito and B. Mirkin, “Nearest neighbour approach in the least-squares data imputation algorithms,” Information Sciences, vol. 169, no. 1-2, pp. 1–25, 2005.

4 S. Chiewchanwattana, C. Lursinsap, and C.-H. H. Chu, “Imputing incomplete time-series data based on varied-window similarity measure of data sequences,” Pattern Recognition Letters, vol. 28, no. 9, pp. 1091–1103, 2007.

5 S. Prasomphan, C. Lursinsap, and S. Chiewchanwattana, “Imputing time series data by regional-gradient-guided bootstrapping algorithm,” in Proceedings of the 9th International Symposium on Communications and Information Technology (ISCIT '09), pp. 163–168, Incheon, South Korea, September 2009

6 E. Grayzeck, 2011, NATIONAL SPACE SCIENCE DATA CENTER, ARCHIVE PLAN FOR 2010 – 2013, NSSDC Archive Plan ’10-13.

7 A. R. Robinson and P.F.J. Lermusiaux “Overview of data assimilation” Harvard Reports in Physical/Interdisciplinary, Ocean Science, NUMBER 62

8 A.J.Viterbi (April 1967). "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm". IEEE Transactions on Information Theory 13 (2): 260–269. doi:10.1109/TIT.1967

9 A.A.Charantonis J. Brajard C. Moulin F. Bardan S. Thiria, Inverse Method for the Retrieval of Ocean Vertical Profiles using Self Organizing Maps and Hidden Markov Models - Application on Ocean Colour Satellite Image Inversion. IJCCI (NCTA) 2011: 316-321

10 R. Jaziri, M. Lebbah, Y. Bennani, J-H. Chenot,2011, SOS-HMM: Self-Organizing Structure of Hidden Markov Model, artificial Neural Networks and Machine Learning – ICANN 2011,Lecture Notes in Computer Science Volume 6792, 2011, pp 87-94

11 Madec G. 2008: "NEMO ocean engine". Note du Pole de modélisation, Institut Pierre-Simon Laplace (IPSL), France, No 27 ISSN No 1288-1619

12 A.S. Willsky, 2002, Multiresolution Markov Models for Signal and Image Processing, Proceedings of the IEEE, Vol. 90, No. 8.

13 Jolliffe, I.T. (2002). Principal Component Analysis, second edition (Springer)

14 C. Kwan, X. Zhang, R. Xu, and L. Haynes, 2003,"A Novel Approach to Fault Diagnostics and Prognostics" Proeerdlngs of the 2003 lEEE, lnlernafioosl Conference Robotics & Aulornstion,Taipei, Taiwan

15 T. Kohonen, 1990 The Self-organizing Map, PROCEEDINGS OF THE IEEE, VOL. 78, NO 9 and http://www.cis.hut.fi/projects/somtoolbox/package/docs2/somtoolbox.html

16 Doneya S C, Kleypasa J A, Sarmientob J L, Falkowski P G, 2002, The US JGOFS Synthesis and Modeling Project – An introduction Deep-Sea Research II 49 (2002) 1-20

17 A.J. Viterbi, 1998, "An intuitive Justification and a Simplified Implementation of a MAP Decoder for Convolutional Codes," IEEE Journal on Selected Areas in Communications, vol. 16, No. 2, Feb. 1998, pp. 260-264.

18 J. Hagenauer, P. Hoeher,A Viterbi algorithm with soft-decision outputs and its applications, Proc. IEEE GLOBECOM, pp. 47.11-47.17, Dallas, TX, Nov 1989.

ANNEX

Advantages of using Self-Organizing Maps for the determination of HMM states:

Self-Organizing Topological Maps (SOM) which are clustering methods based on neural networks. They provide a discretization of a learning dataset into a reduced number of subsets, called classes, which share some common statistical characteristics. Each class is represented by a referent vector, which approaches the mean value of the elements belonging to it, since the training algorithm can be forced to perform like the K-means algorithm at the final stages of its training.

The topological aspect of the maps can be justified by considering the Map as an undirected graph on a two-dimensional lattice whose vertices are the classes. This graph structure permits the definition of a discrete distance, noted *d*, between two classes, defined as the length of the shortest path between them on the map.

Any vector that is of the same dimensions and nature as the data used to generate the topological map, can be classified by assigning it to the class whose referent it resembles most. Therefore a sequence of data vectors can be classified in order to generate a sequence of indexes that correspond to the indexes of the classes to which they were assigned.

In our method, we trained two SOMs, the first containing the observations, called $sMap_{obs}$ and the second containing the hidden states, called $sMap_{hid}$. The hidden states correspond to the discretization of the numerical dynamic model.

The classes of $sMap_{obs}$ and $sMap_{hid}$ correspond respectively to the discretization of the observation vectors into a set amount of observable states, $Y_k, k \in [1, \cdots, N_{obs}]$, and to the hidden states of the HMM, $X_i\ i \in [1, \cdots, N_{hid}]$.

The topological aspect of the SOMs is useful in overcoming the usual lack of sufficient data in estimating the transition and emission probabilities of the HMM. After an initial estimation of the probabilities over each available training sequence, noted *seq*, these transitions, Tr_{seq}, and emissions Em_{seq}, can be combined and adjusted by taking into account the neighboring properties of the topological maps..

This is done by considering the neighborhood matrices NM_{obs} and NM_{hid}, of dimensions (N_{obs}, N_{obs}) and (N_{hid}, N_{hid}) respectively, where

$NM_{obs}(k,l)=\begin{cases}1, \text{if } d(Y_k, Y_l) < 2 \\ 0, \text{else}\end{cases}$ and $NM_{hid}(i,j)=\begin{cases}1, \text{if } d(X_i, X_j) < 2 \\ 0, \text{else}\end{cases}$

with d being the discrete distance on the respective maps.

The final Em and Tr matrices we used, noted Em_{final} and Tr_{final}, were computed by applying for $1 \leq i \leq N_{obs}$ and for $1 \leq j \leq N_{hid}$:

$$Em_{final}(i,j) = 1 + \sum_{seq}\left(w_c\ * \sqrt{L_{seq}} * Em_{\text{seq}}(i,j) + \sum_{k=1}^{N_{hid}}\left(NM_{hid}(j,k) * Em_{\text{seq}}(i,k)\right)\right),$$

Which was normalized to become $e_{i,j}$ where $\sum_{i=1}^{N_{hid}} e_{i,j} = 1$,

and for $1 \leq i, j \leq N_{hid}$

$$Tr_{final}(i,j) = 1 + \sum_{seq}\left(w_c\ * \sqrt{L_{seq}} * Tr_{\text{seq}}(i,j) + \sum_{k=1}^{N_{obs}}\left(NM_{hid}(i,k) * Tr_{\text{seq}}(i,k)\right)\right),$$

, which is normalized to become $tr_{i,j}$ where $\sum_{i=1}^{N_{hid}} tr_{i,j} = 1$.

The term w_c corresponds to a weighting constant that prevents the actual measured probabilities from being overshadowed by their neighborhood values. Its value is determined by an iterative optimization process on an independent test data set. In the case of the application (section 3.2) its value is taken to be equal to 9. The impact of neighboring states on the probabilities has been schematized in figure A1.

It is important to retain that in order to apply PROFHMM_UNC, we require a training data set of concurrent hidden and observable states sequences. This, unlike more complicated cases of HMMs (such as voice recognition software), permits for an estimation of the initial emissions and transition probabilities through the use of a counting algorithm such as hmmestimate of the matlab stat_toolbox.

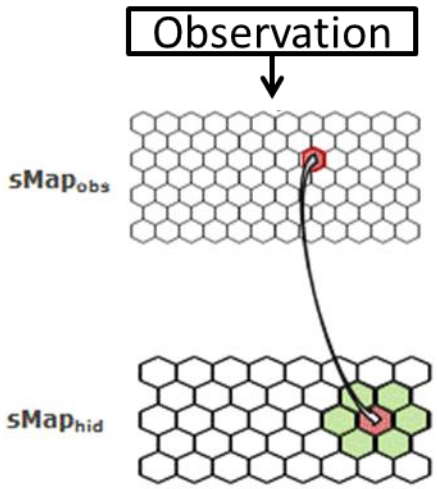


Figure A1. The emission probability of each class Y_k of the sMap$_{obs}$ is emitted from a class X_i of sMap$_{hid}$ that takes into account the probability of being emitted by a class X_j neighboring X_i.

CONCLUSIONS ET PERSPECTIVES

Cette thèse nous a permis de développer deux méthodologies performantes basées sur les cartes topologiques auto-organisatrices et les chaînes de Markov cachées:

- PROFHMM, est une méthodologie qui permet la reconstitution de champs à partir d'un sous ensemble de ces champs. PROFHMM fait initialement appel aux cartes topologiques pour générer les états et la topologie de la chaîne de Markov cachée, puis utilise leurs propriétés de conservation de l'ordre topologique dans le but d'améliorer l'estimation des probabilités. Cette amélioration se trouve être un élément essentiel au bon déroulement d'une HMM si les données concernées sont en dimensions multiples et donc trop souvent trop peu échantillonnées.
- PROFHMM_UNC, est une méthodologie de complétion et de correction de séries de données. Elle est basée sur PROFHMM, qui modifie l'algorithme de Viterbi pour prendre en compte une expertise sur la qualité des observations.

La validité de PROFHMM a été montré sur deux applications prises dans le domaine de l'océanographie : la reconstruction temporelle des profils de chlorophylle ou de température et la reconstruction spatiale des profils de température. Durant la première expérience PROFHMM a permis de bien synchroniser les données surface du modèle numérique NEMO-PISCES avec les profils verticaux de chlorophylle-a et de température de ce modèle. Les expériences ont porté sur la reconstruction temporelle en un point fixe de l'évolution des profils verticaux, et ont démontré que PROFHMM pouvait apprendre la dynamique d'un modèle numérique complexe et la reproduire correctement en utilisant uniquement les données de surface. Notons que le cout de calcul de la reconstruction d'une trajectoire est extrêmement rapide. .

La seconde expérience a montré qu'une même synchronisation pouvait être obtenue spatialement entre les observations satellitaires et les profils de température observés durant les campagnes ARAMIS. Les expériences ont porté sur la reconstitution des profils verticaux de température à partir des données altimétriques AVISO et de la température de surface fournie par la NOAA. Elles ont prouvé que PROFHMM pouvait :

- Etre appliquée pour la reconstruction de l'évolution spatiale de profils verticaux.
- Apprendre la dynamique océanique à partir de données in-situ.

Les données utilisées dans ces applications présentent des défauts dus à la précision des mesures et aux traitements que nous leur avons fait subir. Il existe donc un degré d'incertitude sur chaque donnée. La seconde méthodologie, nommé PROFHMM_UNC, permet de prendre en compte cette incertitude. C'est à dire qu'elle propose d'introduire cette expertise au niveau des observations avant que soit effectué la mise au point de PROFHMM. PROFHMM_UNC propose une modification à l'algorithme de Viterbi, en introduisant deux fonctions : une fonction de confiance aux observations et une fonction de pondération qui utilise cette confiance durant le calcul du chemin optimal pour pondérer l'impact des probabilités d'émissions de chaque observation. La validité de l'approche a été montrée à partir d'expériences jumelles de reconstruction de séries temporelles de données surface.

En conclusion, il est possible de dire que PROFHMM a montré sa généralité que l'on veuille s'appuyer sur la connaissance issue des modèles où que l'on considère des bases de données suffisamment représentatives de la dynamique du phénomène que l'on étudie. L'introduction d'expertise à l'aide de PROFHMM_UNC a permis d'envisager une amélioration des performances de reconstruction par l'amélioration de l'estimation des données manquantes. PROFHMM_UNC et PROFHMM ont été mises au point pour être enchaînées : la première HMM permettant de générer des séries complètes d'états observables plus fiables en entrée de la seconde HMM qui reconstruit les profils verticaux. Ce travail doit faire l'objet d'un stage de master.

Les perspectives de ce travail sont nombreuses: on peut citer :

Études de sensibilité

Tout d'abord une série d'étude peut être envisagée pour compléter l'étude des possibilités de PROFHMM. Des mémoires de projets longs effectués par des étudiants de master 2 (Annexe 2) ont montré que la méthodologie pouvait être appliquée sur deux zones géographiques différentes et donner de bonnes reconstructions. Il serait maintenant important de généraliser afin de déterminer si les variables de surface utilisés pour les reconstructions dans les trois cas (CHL-A, SST, WS, lat, ???), permettent d'appliquer PROFHMM dans n'importe quelle région de l'océan.

1. Une autre étude de sensibilité consisterait à regarder s'il est possible de faire fonctionner, avec de bonnes performances, la HMM apprise avec PROFHMM sur un lieu voisin de la zone apprise. Il serait important de délimiter d'une manière précise la zone

de validité de la reconstruction. Cela permettrait comme nous l'avons montré dans le premier article, où nous avons utilisé les 8 points voisins de BATS, d'augmenter les bases données utilisables pour l'apprentissage.

2. Les reconstructions des profils verticaux de température obtenues dans le chapitre trois, indique que nous pouvons reconstruire des séries spatiales de profils verticaux de température le long du rail ARAMIS, à partir de données de surface. La variabilité longitudinale des phénomènes thermiques de la zone étant plus faible que la variabilité latitudinale, il serait intéressant d'effectuer une analyse de sensibilité de la méthode pour déterminer l'étendue spatiale sur laquelle nous pouvons reconstruire des profils verticaux de température, en utilisant les cartes et les probabilités apprises sur les données ARAMIS.
3. Des problèmes similaires comme la reconstruction des profils de température et salinité obtenus par gliders dans le bassin méditerranéen (étudiés par les European Gliding Observatories -EGO), présentent des caractéristiques qui permettent d'envisager leurs traitements par les deux méthodologies.

Utilisations possibles de PROFHMM:

Un domaine conceptuellement proche des travaux présentés dans ce manuscrit est le domaine de l'assimilation des données. L'assimilation des données est le procédé qui consiste à corriger, à l'aide d'observations, les paramètres d'un modèle numérique, comme son état initial, ou les valeurs de certaines constantes internes à partir d'observations. Les méthodes d'assimilation variationnelle nécessitent de connaître une valeur approchée des solutions recherchées (« background ») . Les bonnes performances obtenues par les expériences que nous avons présentées indiquent que les profils reconstitués par PROFHMM pourraient être utilisés comme « background » par ces méthodes d'assimilation variationnelle des données.

Un objectif un peu plus lointain serait d'utiliser les connaissances mémorisées dans PROFHMM afin de simuler des données de surface ayant une cohérence dynamique et d'utiliser ces données simulées pour faire des études de sensibilité. L'intérêt dans ce cas serait d'obtenir un temps de calcul extrêmement faible comparé à l'utilisation des méthodes de Monté Carlo qui demandent de faire tourner le code numérique dans sa globalité. Le générateur probabiliste mis au point par PROFHMM est « comparable » à un modèle simplifié à quelques variables. L'Annexe 3 présente un peu plus en détail ce générateur.

A l'heure actuelle PROFHMM ne reconstruit que des séries 1D, qu'elles soient temporelles ou spatiales. Cette limitation peut être levé si l'on considère la théorie des champs de Markov. Un

but long terme serait de faire évoluer la méthode pour assurer de reconstructions multidimensionnelles.

ANNEXE 1 : NCTA 2011 ARTICLE

In this Annex we include the paper presented at the NCTA 2011 – International Conference on Neural Computation Theory and Applications, in Paris.

INVERSE METHOD FOR THE RETRIEVAL OF OCEAN VERTICAL PROFILES USING SELF ORGANIZING MAPS AND HIDDEN MARKOV MODELS.

Application on ocean colour satellite image inversion.

Charantonis Anastase Alexandre[1], Brajard Julien[1], Moulin Cyril[2], Bardan Fouad[3],and Thiria Sylvie[1]

[1] *Laboratoire d'Océanographie Climat et Analyses Numériques, Université Pierre et Marrie Curie, Tour 45-55, 4, Place Jussieu, 75252, Paris, France*

[2]*Laboratoire des Sciences du Climat et de l'Environnement, L'Orme des Merisiers, CEA Saclay, bat 712, 91191, Gif-sur-Yvette, France*

[3]*Laboratoire CEDRIC,Conservatoir National des Arts et Metiers (CNAM), 292, rue Saint Martin, 75003, Paris, France*

{Anastase-Alexandre.Charantonis, Julien.Brajard, Sylvie.Thiria }@locean-ipsl.upmc.fr,

cyril.moulin@lsce.ipsl.fr,

fouad.badran@cnam.fr

Keywords: Self Organising Maps; Hidden Markov Models; Inversion; Geophysical; Chlorophyll-A; Satellite imaging; Inversion.

Abstract: This paper presents a statistical inversion method used to infer 3D data from 2D imaging. The methodology is based on a combination of the Self Organising Maps and the Hidden Markov Models. The method has been validated by inferring the oceanic vertical profiles of Chlorophyll-A based on sea-surface data.

1 INTRODUCTION

The density of satellite observations allowed a semi-continuous observation of the global ocean surface. The two-dimensional images provided by this coverage often contain information on integrated quantities whose vertical distribution is unknown. Depending on the field of study there exist different dynamic approaches for inverting this type of data. However, these approaches are often faced with the problem of non-linearity, and can also be hampered by a lack of knowledge of the complete mechanisms that govern the distributions.

The present paper deals with the inversion of observed sea-surface satellite images, noted $\tilde{x}_{obs}^{t}$ $t \in [1 \cdots T]$, for retrieving of the vertical distribution of Chlorophyll-A, noted $\tilde{x}_{dis}^{t}$ $t \in [1 \cdots T]$, using a statistical, non-linear approach.

The methodology we have developed is a mixture of the neuronal algorithms known as Self Organizing Topological Maps (SOM) and the Hidden Markov Models (HMM). The SOM are unsupervised classification algorithms, that allow us to cluster our availiable data into classes. The classes are arranged on a topological map and connected to each other by a topological similarity distance. In the present study the SOM classification is applied twice, once on the sea-surface data, and once upon the vertical profiles connected to these images. The resulting topological maps allow us to discretize both data spaces into set amounts of classes.

The second statistical algorithm, the HMM allows us to infer the most likely sequence of some discrete, unobservable states, given a series of discrete, observable states. To do so a set of probability matrices are calculated, corresponding to the dynamic processes of the unobserved states and the links existing between the observed and unobserved states. We use the classes created through the SOMs to discretize the availiable data and therefore represent both the observable and the unobservable states.

In this paper, we present the results obtained with the methodology developed on a case study at the site of the Bermuda Atlantic Time Series (BATS) (32 N -64 W) of the JGOFS campaign.

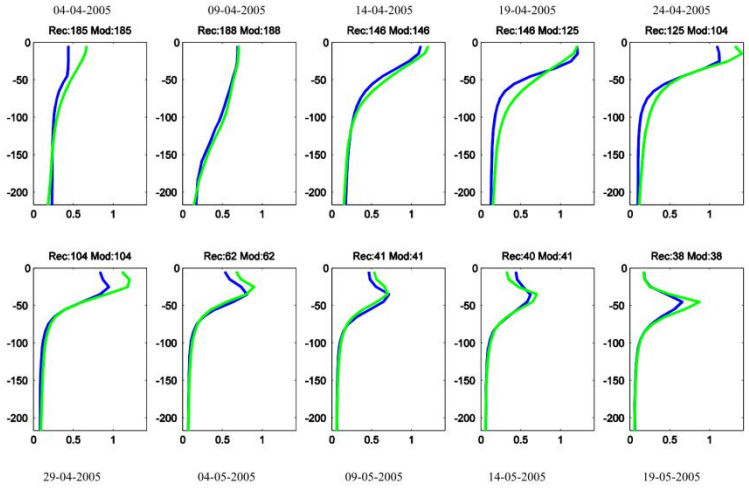


Figure 1: Inversion of ten 5-days steps for the period from 04-04-2005 to 05-19-2005, at BATS. In green, the states provided by the inverse method and in blue the vertical distribution of Chlorophyll-A according to the NEMO-PISCES model. The horizontal axes are in 10^6 µmol/L of Chlorophyll-A, while the vertical ones are in meters from the sea surface. The numbers on top correspond, after ***Rec***, to the indexes of the classes on M_{dis} attributed to that 5-day step by the inverse method, and, after ***Mo*** to the indexes attributed by projection of the total profile on M_{dis}. These are not show in the figure.

2 SELF-ORGANIZING TOPOLOGICAL MAPS

Self-Organising Topological Maps (SOM) are clustering methods based on neural networks (S, Haykin 1999). They provide a discretization of a learning dataset A= $\{ \tilde{x}_k \in R^p$, k =1...N $\}$ into a reduced number of subsets, called classes, P_i , { i = 1...M } that share some common statistical characteristics. Each subset is represented by its referent vector $\boldsymbol{r_i}$ which approaches the mean value of the elements in the class Pi. In our case, we trained two SOMs, one containing the observations, called M_{obs} and one containing the distributions of the unobservable states, called M_{dis}. The number of classes in M_{obs} and M_{dis} are respectively noted N_{obs} and N_{dis}.

The topological aspect of the maps can be justified if we consider the Map as an undirected graph on a two-dimensional lattice whose vertices are the m classes. This graph structure therefore allows the definition of an discrete distance d(i,j) between two classes i and j, defined as the length of the shortest path between i and j on the map. The nature of the SOM training algorithm forces a topological ordering upon the map, and therefore any neighbouring classes $\boldsymbol{c_i}$ and $\boldsymbol{c_j}$ on the map have referent vectors $\boldsymbol{r_i}$ and $\boldsymbol{r_j}$ that are close in the Euclidian sense in the data space R^P.The topological ordering constitutes a major element of our inverse method, since it allows us to make, latter on, the ergodic assumption for our Markov states.

We define a series of observable events by taking the data from observations related to a given period of time and we label each observation by the index of the class to which it is assigned by using M_{obs}. This classification is done by allocating to each observation $\tilde{x}_{obs}^t \;\; t \in [1 \cdots T]$, in the sequence the index of the class of M_{obs} whose referent is the closest to it in the Euclidian sense. Therefore we obtain the series

$$S_{obs} = \left\{ s_{obs}^t = \; argmin_i \left(\left| \tilde{x}_{obs}^t - \boldsymbol{r}_i^{obs} \right| \right), \; i = 1 \cdots N_{obs} \right\} \qquad (1)$$

In the same way, we obtain the time-series of distributions

$$S_{dis} = \left\{ s_{dis}^t = \; argmin_i \left(\left| \tilde{x}_{dis}^t - \boldsymbol{r}_i^{dis} \right| \right), \; i = 1 \cdots N_{dis} \right\} \qquad (2)$$

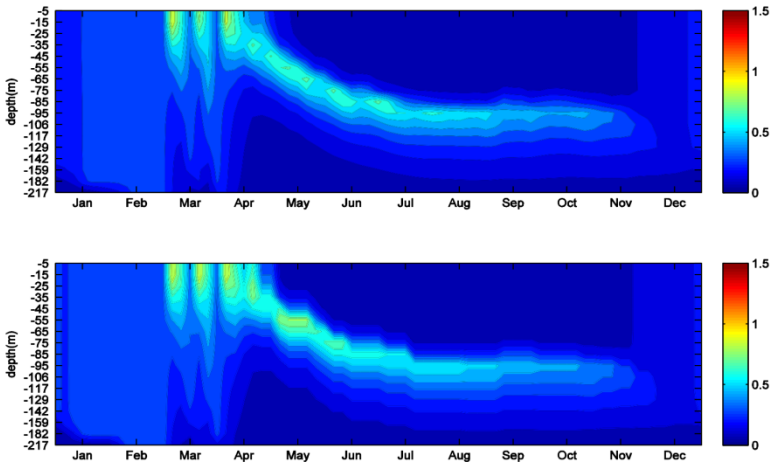


Figure 2: The reconstruction at BATS of the validation year 2005, according to, the NEMO-PISCES MODEL (top graph) and the inverse method result (bottom graph). The colorbar indicates the Chlorophyll-A concentration in 10^6 µmol/L.

These series S_{obs} and S_{dis} are used by the HMM in order to estimate the probabilistic links that exist between the observable states and the unobservable ones.

The referents $\boldsymbol{r}_{dis}$ of M_{dis} are also used as the vertical profiles in our sequence reconstructions.

3 HIDDEN MARKOV MODELS

A Markov model is a stochastic model that assumes the first order Markovian property, meaning that each consecutive state of the model depends solely on its previous stat of the model such as

$$P(X_t| X_1 X_2 \dots X_{t-1}) = P(X_t| X_{t-1}) \quad (3)$$

Expanding this principle, a Hidden Markov Model (HMM) is a stochastic model with two sequences. One sequence of unobservable states that follow the first order Markovian property, (represented in our method by S_{dis}), and one sequence of observable states, (represented by S_{obs}), that have a statistical link with the unobservable states (O. Cappé et al. 2005).

We consider two phases, a training one, and a retrieval one. During the training, the Transitions matrix Tr and the Emissions matrix Em are estimated. Tr contains the transition probabilities of the unobserved states

$$tr_{i,j} = P(C_{dis(i,t)} \mid C_{dis(j,t-1)}) \quad (4)$$

where

$$\sum_{i=1}^{N_{dis}} t_{i,j} = 1 \quad (5)$$

Tr corresponds, in a physical sense, to the underlying dynamics that govern the unobserved states.

Em contains the *à posteriori* probabilities of each observed state to have been emitted by an unobserved state,

$$e_{i,j} = P(C_{dis(i,t)} \mid C_{obs(j,t)}) \quad (6)$$

where

$$\sum_{j=1}^{N_{dis}} e_{i,j} = 1 \quad (7)$$

Em corresponds, in a physical sense, to the link existing between the observed quantities and the dynamics of the unobserved quantities. Another probability matrix that needs to be calculated is the initial probability matrix Π, with components π_i which represent the average revisit rate of each

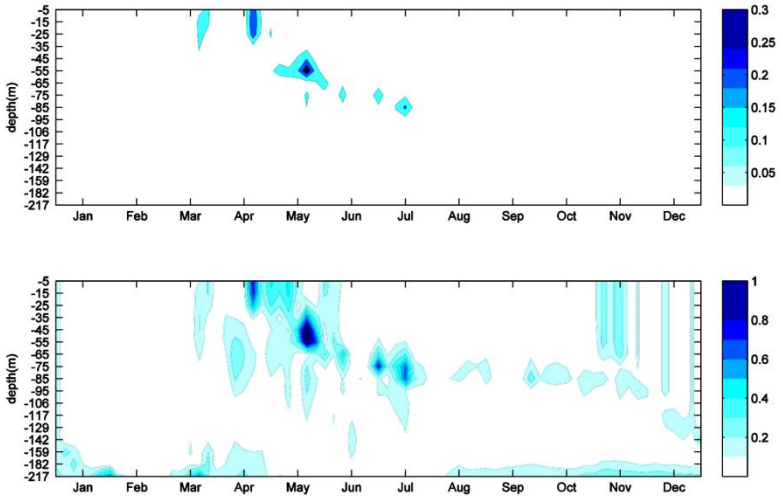


Figure 3: The top image contains the absolute error between the NEMO-PISCES model and the result of the inverse method, at BATS for the validation year 2005. The bottom image contains the absolute relative error between the NEMO-PISCES model and the result of the inverse method.

unobserved state given an infinite sequence. All mentioned probabilities are estimated by using the Baum-Welch algorithm (L. E. Baum et al. 1970), which is a maximum likelihood optimization algorithm, that takes as input the sequences S_{obs} and S_{dis} and outputs the most likely matrices to have generated them through a hidden Markov process.

During the recognition phase, we used the Viterbi algorithm, which is a well-know dynamic programming algorithm (Viterbi AJ, 1967), for inferring the most likely sequence of indexes $S_{dis\text{-}est}$ representing the unobserved states, given the previously estimated parameters Tr, Em and Π of the HMM and a sequence of observations $S_{obs\text{-}new}$. It is well documented (M.S. Ryan and G.R. Nudd. 1993) that the Viterbi algorithm can face problems due to transitions that were not observed in the training data set. A balance needs to be found between the sizes of the SOM maps that will determine the amount of discretization provided by the method, and the correctness of the allocation of indices. The dimensions of each map are therefore optimized using a validation set. Yet, even with an optimization there will be some situations and transitions that are seldomly encountered in the training data and result in null probabilities in the probability matrices $Em_{B\text{-}W}$ and $Tr_{B\text{-}W}$ that we estimated in the first pass of the Baum-Weltch algorithm.

Due to this usual lack of sufficient data in the concerned domains, $Em_{B\text{-}W}$ and $Tr_{B\text{-}W}$ need to be adjusted. This is done by taking into account the properties of the topological maps. A major characteristic of the present method is to use the topological order in order to improve the accuracy of the estimated probabilities matrices. The topological maps allow us to modify the probabilities by allowing each state to communicate via a diminutive probability with each of its neighbouring states.

This is done by considering the neighbourhood matrices NM_{obs} and NM_{dis}, of dimensions (N_{obs},N_{obs}) and (N_{dis},N_{dis}), where

$$NM_{SOM}(i,j)=\begin{cases}1, \text{if } d\left(c_i^{SOM}, c_j^{SOM}\right) < 2\\ 0, else\end{cases} \quad (8)$$

with d(i,j) being the discrete distance of the map. Taking into account the neighbourhood consists in increasing the probability of reaching a class j from a class i, by an ammount proportional to the sum of the previously calculated probabilities of reaching the neighbour classes of class j on SOM. In order to favour the data observed during training, we add a weighting term, noted w_c, to the initial probabilities, and we further multiply it by the total length of the training sequences used in the intitial Baum-Weltch algorithm's pass, noted $T_{training}$, since this length is a

measure of confidence in the correctness of the estimated parameters. The matrices obtained are then normalized. The final Em and Tr matrices we use, noted Em_{final} and Tr_{final}, are computed by applying:

$$Em_{final}(i,j) = w_c * T_{training} * Em_{B-W}(i,j) + \sum_{k=1}^{N_{obs}} \left(NM_{obs}(i,k) * Em(i,k)\right) + 1 \quad (9)$$

Which is normalized to fit the constraint (7), and

$$Tr_{final}(i,j) = w_c * T_{training} * Tr_{B-W}(i,j) + \sum_{k=1}^{N_{dis}} \left(NM_{dis}(i,k) * Em(i,k)\right) \quad (10)$$

Which is normalized to fit the constraint (5). For this application w_c is set to 9.

These modifications permit the Viterbi algorithm to circumvent the problems of impossible transitions, or emissions due to insufficient data in the training sequences that resulted in nul probabilities in the estimated parameters.

4 APPLICATION FOR THE RESTITUTION OF THE VERTICAL CHLOROPHYLL-A CONCENTRATION THROUGH SEA SURFACE DATA

The bio-geochemical activity of the oceans and the carbon cycle are two parts of a complex feedback system. A change in climate and an increase of the amount of available carbon can affect the primary oceanic production, and in return a change in the bio-geochemical activity affects, by modifying the albedo and carbon fixation rates, the climate and carbon concentration. It is therefore important to be able to determine the oceanic primary production. In recent years, many algorithms have been developed that infer the Chlorophyll-A concentration in ocean surface layers through satellite imaging (Brajard et al. 2008). It has also been proved that the vertical Chlorophyll-A distribution, is correlated with sea surface data (Uitz et al. 2006). Therefore, the determination of the vertical distribution of Chlorophyll-A from sea surface data is a problem that can be solved by the methodology we propose.

One cannot determine the vertical distribution of Chlorophyll-A without first understanding the parameters that influence the development of phytoplankton. It is generally accepted that phytoplankton growth depends on 5 parameters: available radiation, available nutriments, predators and biology, water temperature, water turbidity.

These parameters cannot easily be monitored through a direct approach. Satellite imaging, however, can give us proxy information, which can be used in an empirical approach for determining the vertical distribution of Chlorophyll-A. Specifically, in this study we used: Sea Surface Chlorophyll-A concentration (SCHL), Sea Surface Temperature (SST), Sea Surface Elevation (SSH), Shortwave Radiation (SR) and Wind-speed Intensity (WS).

Since our objective is to validate the theoretical methodology, we used simulated data in order to test the validity of our approach. We therefore approximated the satellite values of the previous parameters by using the input and output values provided by the NEMO oceanic circulation model coupled to the PISCES bio-geochemical model (C. Moulin, 2008). In order to better simulate the noise and errors inherent to satellite images we added a white noise $z \sim N(0,\hat{e})$, $\hat{e}=1/2 * (\sigma_{schl}, \sigma_{sst}, \sigma_{ssh}, \sigma_{ws}, \sigma_{sr})$ to the parameters that could be gathered from satellite imaging. σ represents the standard derivation of each corresponding surface parameter, as computed on the training data. The application was set at the site of BATS.

The unobserved states that were classified, were the output data vectors containing the average vertical Chlorophyll-A distribution at 17 depth levels (from 5 meters to 217) and temperature distribution at 9 depth levels. These vertical distribution profiles were 5-day averages spanning the period from 1991 to 2007 located in a 2°x2° square centred on BATS. Therefore M_{dis} belongs to R^{26} (17 levels of Chlorophyll-A + 9 levels of Temperature). M_{obs} belongs to R^5.

We trained M_{obs} and M_{dis} by taking into account all available profiles at BATS, as well as any adjacent points included in the model. This gave us 9*73*17=11169 profiles for the construction of the maps. The optimum map sizes, N_{obs} and N_{dis} were determined to be 21*14=294 classes.

For the estimation of the HHM parameters on the other hand, we take a total of 14 years (1991-2004) for the training, each including seventy-three 5-day steps. Therefore $T_{training}$=1022. We maintained 3 years (2005-2007), or 219 5-day steps, to validate our approach.

The results shown in Figure 1 present the temporary evolution of Chlorophyll-A profiles in ten 5-day steps sequence, from 04-04-2005 to 19-05-2005, at BATS. In green we see the Chlorophyll-A distribution profiles, taken from the referents $\boldsymbol{r_{dis}}$ of M_{dis}, corresponding to the indexes of the reconstructed time series $S_{dis\text{-}rec}$. In blue we can see

the vertical distribution of Chlorophyll-A according to the NEMO-PISCES model at the same 5-day steps.

In Figure 1 we also have, preceded by the acronym ***Rec***, the indexes that constitute time series $S_{dis\text{-}rec}$ and, preceeded by ***Mod***, the indexes we obtain by projecting the corresponding profiles of the NEMO-PISCES model on M_{dis}. In order to avoid confusion, the profiles corresponding to the indexes after ***Mod*** are not displayed. When the vertical distribution of Chlorophyll-A is know, these indexes would correspond to the optimum reconstruction we could get with M_{dis}. We note this optimum time series as $S_{dis\text{-}opt}$. It is interesting to notice that even when the indexes are not equal, the classes are neighbours on M_{dis}, and the estimated profiles are quite similar to the observed ones.

If we define $S_{dis\text{-}2005to2007}$ as the reconstructed time series of indexes of the validation years from 2005 to 2007 and as $S_{dis\text{-}opt\text{-}2005to2007}$ the corresponding optimum reconstruction we observe that they are in agreement 84,59% of the time. This perforance reaches 88,58% when applied on the reconstruction $S_{dis\text{-}2005}$ of the year 2005 alone, as compared to its optimum reconstruction. This was probably due to the validation year 2005 having a small variation from the mean year, and presenting often-observed transitions. In the training date we had on average an agreement of 86,46%

In Figures 2 and 3, we applied the inverse method to the full 73 5-day steps serie of the validation year 2005. We can observe that the reconstruction closely fits the results provided by the NEMO-PISCES model. The correlation index of the two images in Figure 2 is 97,30%.

We can notice that the discretization induced by the SOM is apparent, yet the general form and intensity are correctly represented, as it becomes clear in Figure 3, where the error graphs tend to have small values.

5 CONCLUSIONS

In the present paper we have introduced an inversion method based on SOM and HMM, that is able to reconstruct the vertical profiles of Chlorophyll-A based on satellite images. One of its main advantages in the inversion of Chlrophyll-A is that it does not use Gaussian approxiamtions used in other methods (Morel 1988, Uitz et Al 2006), allowing the reconstruction of situations where the distribution is not conforming to a single gaussian curve. An additional benefit of the inversion method presented, is its efficiency in terms of calculations. The method is open ended enough to be applicable for the inversion of the profiles of different bio-geophysical parameters based on satellite imaging.

We plan to further validate this method by testing its robustness with satellite imaging and in-situ data, as well as to apply it on different types of profiles, such as oceanic salinity or temperature profiles. A latter goal is to expand the method to take spatial constraints into consideration, and reconstruct 3D profiles.

ACKNOWLEDGEMENTS

We would like to thank the “Délégation Generale de l’Armement” for financing this work.

REFERENCES

S, Haykin 1999. "9. Self-organizing maps". *Neural networks - A comprehensive foundation* (2nd ed.). Prentice-Hall

L. E. Baum, T. Petrie, G. Soules, and N. Weiss, 1970. "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains", Ann. Math. Statist., vol. 41, no. 1, pp. 164–171.

F. Badran, M. Berrada, J. Brajarda M. Crépon,C. Sorror, S. Thiria, J.-P. Hermand M. Meyer, L. Perichob and M. Asch, 2008. "Inversion of satellite ocean colour imagery and geoacoustic characterization of seabed properties: Variational data inversion using a semi-automatic adjoint approach", Journal of Marine Systems,Volume 69, Issues 1-2, Pages 126-136

C. Moulin, A. Kremeur, A. El Moussaoui, C. Ethe, L. Bopp, E. Dombrowsky, E. Greiner, O. Aumont, P. Brasseur, 2008. "Understanding the interannual variability of the oceanic carbon cycle: Results from the coupled biogeochemical-physical global model PISCES-NEMO", American Geophysical Union, Fall Meeting 2008, abstract #OS31A-1226Journal of Marine Systems,Volume 69, Issues 1-2, Pages 126-136

J. Uitz, H. Claustre, A. Morel, S. B. Hooker, 2006 "Vertical distribution of phytoplankton communities in open ocean: An assessment based on surface chlorophyll", Journal of Geophysical Research Vol.111.

Olivier Cappé, Eric Moulines, and Tobias Rydén, 2005. "*Inference in Hidden Markov Models*." Springer.

Viterbi AJ, 1967. "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm". IEEE Transactions on Information Theory 13 (2): 260–269.

M.S. Ryan and G.R. Nudd. 1993. The Viterbi Algorithm. Technical Report University of Warwick RR-238

ANNEXE 2 : RAPPORT DE PROJET LONG SAYAD – HALIMI

Reconstitution de profils de chlorophylle à partir de données satellites sur le site d'observation DYFAMED

Présenté par :
- Mahmoud SAYAD
- Said HALIMI

Encadré par :
- Anastase-Alexandre CHARANTONIS

2010/2011

Table des matières

1. Introduction

La compréhension du rôle de l'océan dans la régulation des flux de carbone représente actuellement l'un des défis majeurs de l'océanographie. Dans ce contexte, l'étude de la biomasse phytoplanctonique et de la production primaire marine sont des thématiques essentielles.

Les études concernant la production primaire et les flux de matière organique ont pour référentiel le carbone organique, qui constitue *a priori* le meilleur estimateur de la biomasse phytoplanctonique. Cependant, il n'existe pas de méthode de mesure directe pour quantifier le carbone strictement phytoplanctonique. Ainsi la **chlorophylle *a*** (Chl*a*) est utilisée comme indicateur universel de la biomasse algale dans l'océan [1].

La télédétection de la "couleur de l'océan" permet d'estimer, depuis l'espace, le contenu en chlorophylle des eaux océaniques de surface. Cependant la couleur de l'océan ne donne accès qu'à la couche supérieure de l'océan. Et vu que le phytoplancton se développe dans les zones profondes de l'océan il est indispensable de trouver un moyen d'estimer la chlorophylle en profondeur en utilisant celle de surface.

Le but de notre projet est d'essayer de reconstituer les profils de la chlorophylle dans les différents niveaux de profondeur à l'aide des mesures effectuées en surface, et cela en utilisant les techniques d'apprentissage statistiques.

2. Généralités

2.1. Phytoplancton et chlorophylle

Le phytoplancton est l'ensemble des organismes aquatiques chlorophylliens du plancton. Il dispose d'antennes pigmentaires grâce auxquelles il peut capturer l'énergie des photons. Au premier rang des pigments du phytoplancton se trouve la chlorophylle a. En absorbant ainsi la lumière, tout spécialement dans la partie bleue du spectre visible, la chlorophylle modifie sélectivement le flux de photons qui transite dans la partie éclairée de l'océan, de telle sorte que la lumière solaire qui ressort de l'océan est moins bleue que lorsqu'elle y a pénétré. Cette propriété a été exploitée par les agences spatiales qui ont lancé des satellites munis de capteurs de couleur afin de pouvoir estimer le contenu en chlorophylle a de l'océan et sa variabilité dans l'espace et le temps.

La chlorophylle est un pigment, situé dans les chloroplastes des cellules végétales, qui intervient dans la photosynthèse pour intercepter l'énergie lumineuse, première étape dans la conversion de cette énergie en énergie chimique. Son spectre d'absorption du rayonnement lumineux est responsable de la couleur verte des végétaux ; la longueur d'onde la moins absorbée étant le vert, c'est donc cette couleur qui est perçue dans la lumière réfléchie vers l'œil par la feuille.

2.2. Présentation du site

Notre zone d'étude, pour réaliser ce projet, est centrée sur le site d'observation DYFAMED (**DY**namique des **F**lux **A**tmosphériques en **MED**iterranée). Il est situé dans la

zone centrale de la mer Ligure à environ 50 km sur la radiale Nice-Calvi (position : 43°25N 07°52E) par 2350m de profondeur (voir *fig.1*). Il a donc des caractéristiques hauturières. En mer Ligure la bande côtière est alimentée par le courant liguro-provençal qui isole complètement la zone du large. Dans cette zone centrale, la production primaire dépend des apports profonds de sel nutritifs mais aussi, pour une part encore mal déterminée, d'apports atmosphériques aussi bien pour l'azote, le phosphore (en période estivale) que pour certains métaux traces **(Marty et Chiaverini 2002 ; Marty et al. 2002).**

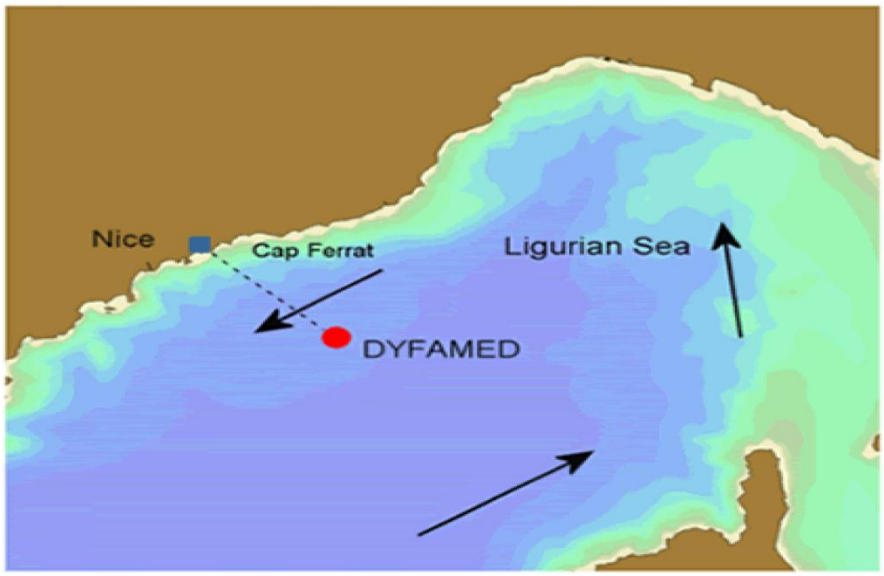


Fig.1 : Le site DYFAMED

3. Outils statistiques utilisés

3.1. Cartes topologiques

Les cartes topologiques, appelées aussi cartes de Kohonen, sont des réseaux de neurones artificiels orientés, constitués de 2 couches :

- Une couche d'entrée dont les neurones correspondent aux variables décrivant les observations ;
- Une couche de sortie qui est le plus souvent organisée sous forme de grille (de carte) de neurones à 2 dimensions. Chaque neurone représente un groupe d'observations similaires.

Les cartes topologiques sont des méthodes de classification et de visualisation de données qui ont les caractéristiques suivantes [2] :

- Chaque neurone c de la carte C est associé à un vecteur référent W_c de l'espace des données D

- L'apprentissage approxime la densité sous-jacente des données tout en cherchant à respecter une contrainte de conservation de la topologie de la carte *C*
- Deux neurones *c* et *r* «voisins» par rapport à la topologie discrète de la carte *C* sont associés à deux vecteurs référents W_c et W_r proches dans l'espace des données *D*.

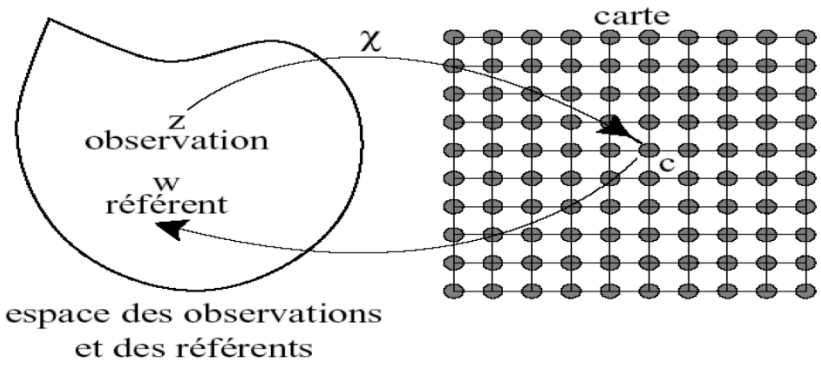


Fig.2 : Principe des cartes topologiques.

3.2. Modèles de Markov cachés

Les modèles de Markov cachés (*Hidden Markov Model : HMM*) sont des modèles statistiques permettant de modéliser des processus stochastiques. Pour mieux les comprendre, prenons un exemple. Considérons deux personnes séparées par un mur. La première personne possède trois dés biaisés. Tour par tour, elle choisit un dé, le lance et énonce à voix haute le numéro de la face résultante à la deuxième personne. La deuxième personne de l'autre côté du mur, ne connaît que la séquence des numéros de face des dés, pas la séquence des dés qui a été choisie par la première personne. La séquence des dés est appelée séquence des états cachés et la séquence des numéros de face est appelée la séquence d'observations [3].

Un modèle de Markov cachés λ permet de modéliser un tel processus à l'aide de certaines hypothèses simplificatrices. La première hypothèse permet de dire que la séquence des états cachés est régit par un processus de Markov de degré 1 en temps discret c'est-à-dire que la probabilité d'apparition d'un état caché ne dépend que de l'état caché précédent dans la séquence et que ces probabilités de dépendances ne changent pas au cours du temps. Si on note $Q = (q_1, q_2, \dots, q_T)$ la séquence des états cachés, alors on a :

$$P(Q|\lambda) = P(q_1/\lambda)\prod_{t=1}^{T-1} P(q_{t+1}|q_t, \lambda)$$

La deuxième hypothèse dit que la probabilité d'émission d'un symbole ne dépend que de l'état caché dans lequel est le processus. En notant $O = (o_1, o_2, \dots, o_T)$ la séquence d'observation, on a :

$$P(O|\lambda) = \sum_{Q} P(O|Q,\lambda)P(Q|\lambda)$$

Graphiquement, les dépendances des probabilités sont données par la figure () :

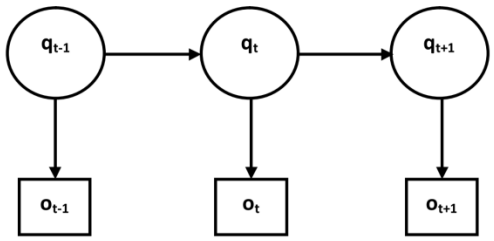


Fig.3 : Dépendances des probabilités d'un HMM

Classiquement, les HMM introduisent trois problèmes fondamentaux :

- Comment évaluer la vraisemblance d'une observation par rapport à un modèle c'est-à-dire comment calculer $P(O|\lambda)$? Ce problème est efficacement résolut par l'algorithme de Forward [4].
- Comment connaître la séquence des états cachés qui a le plus probablement généré une séquence d'observations ? Pour résoudre ce problème efficacement, on utilise l'algorithme de Viterbi [4].
- Comment ajuster les paramètres d'un HMM (probabilités initiales, probabilités des transitions et des émissions) pour qu'il reconnaisse mieux une séquence d'observations ? pour résoudre ce problème efficacement on utilise généralement l'algorithme de Baum Welch [4].

4. Méthodologie suivie

On rappel ici la problématique qui est d'essayer de reconstituer les profils de la chlorophylle dans les différents niveaux de profondeur à l'aide des mesures effectuées en surface. Pour cela on a suivie la méthodologie suivante :

4.1. Préparation des données

Vu le manque de données réelles à haute fréquence de revisite, on a utilisé les données issues du modèle NEMO/PISCES (modèle de circulation générale océanique et de biogéochimie marine), considérées ici comme données réelles.

On dispose des mesures de chlorophylle (surface et profondeurs) de 18 années (1992-2009) comme base d'apprentissage à l'aide de laquelle on va pouvoir définir les différents profils de chlorophylle.

La concentration de la chlorophylle est déterminée par sa biologie, la quantité de nutriments disponible et du rayonnement reçu. Ces paramètres ne sont pas observables. Pour les estimer on utilise les paramètres proxy suivants :

- En surface : température de l'eau (**TEMP**), quantité de lumière captée par le phytoplancton (Shortwave radiation : **SWR**), vitesse du vent (**WS**), et le niveau de la mer (**SSH**).
- En profondeur : température de l'eau (**TEMP**) aux différents niveaux de profondeur.

Les sorties (mesures) du modèles NEMO/PISCES sont des valeurs moyennées sur 5 jours pour les différents niveaux de profondeur (31 niveaux). On se contentera des 10 premiers niveaux car au-delà y a pratiquement pas de chlorophylle.

Donc les données concernant la surface sont des vecteurs de dimension 5 qui correspond aux 5 variables : *CHL, TEMP, WS, SWR ET SSH*. Et celles concernant la profondeur sont des vecteurs composés par le taux de chlorophylle (*CHL*) des 10 premiers niveaux de profondeur et les températures (*TEMP*) correspondantes.

Le taux de chlorophylle est fortement lié à la température. L'introduction de cette dernière va nous aider à bien créer les classes (profils) de chlorophylle mais ne doit pas trop peser sur l'apprentissage. De plus la variation de la température entre un niveau de profondeur et le suivant est faible. Pour remédier à ça on va prendre une température à chaque 2 niveau.

Pour pouvoir valider notre modèle, on va diviser chacun de nos 2 ensembles de données (surface et profondeur) en deux sous-ensembles :

- Un sous-ensemble d'apprentissage : mesures des années 1992 à 2005 ;
- Un sous-ensemble de validation : mesures des années 2006 à 2009.

Pour éviter de biaiser les cartes topologiques, notons bien que les données seront normalisées en les centrant et réduisant par leur variance car elles ne sont pas du même ordre.

4.2. Calcul des profils de chlorophylle

Pour déterminer les états de chlorophylle de surface et les profils de chlorophylle en profondeur, on va discrétiser les données initiales en utilisant les cartes topologiques qui vont regrouper les états de chlorophylle les plus similaires en formant des profils moyens.

On aura besoin de deux cartes, une pour la surface et l'autre pour les profondeurs. L'apprentissage des deux cartes topologiques est fait en utilisant la *som toolbox* de *Matlab*.

L'apprentissage est réalisé en utilisant les 2 fonctions suivantes :

1) *som_make* : elle reçoit en entrée les données d'apprentissage et le nombre de neurones constituant la carte. Elle nous permet l'initialisation, l'apprentissage et détermine les bonnes dimensions de la carte en réalisant une ACP sur les données d'entrées.

2) *som_batchtrain* : elle nous permet, en utilisant la carte construite par la fonction *som_make*, de préciser les paramètres d'apprentissage comme : la fenêtre de voisinage, le nombre d'itérations, l'algorithme d'apprentissage, …

4.3. Reconstitution des profils de chlorophylle

Pour la reconstitution des profils de chlorophylle on fait recours à un modèle de Markov cachés (*HMM*). Les états de chlorophylle formés par la carte de surface vont être utilisés comme observations et ceux de profondeur vont jouer le rôle des états de notre HMM.

Apprentissage :

En projetant les données d'apprentissage des mesures de surface (respect. de profondeur) sur la carte de surface (respect. de profondeur) on obtient pour chaque donnée le neurone auquel elle est affectée. La séquence de neurones obtenue par la carte de surface va être notre séquence d'observations et celle donnée par la carte de profondeur jouera le rôle des états cachés correspondant.

Une fois les deux séquences d'états cachés et d'observations connues, on applique la fonction Matlab *« hmmestimate »* pour estimer les paramètres de notre HMM : la matrice des transitions *T* et celle des émissions *E*.

« hmmestimate », basée sur l'algorithme Baum Welch, nous permet l'estimation des deux matrices *T* et *E* à partir d'une séquence d'observations et les états correspondant en utilisant la technique du maximum de vraisemblance. Cependant, elle n'estime que les transitions et émissions apparaissant dans la séquence d'apprentissage. Alors pour les transitions et/ou émissions qui sont possibles (par exemple 2 neurones voisins sur la carte) mais qui n'apparaissent pas dans la séquence d'apprentissage auront des probabilités nulles.

Pour éviter cela on a dû modifier les matrices *T* et *E* de telle sorte que les transitions et/ou émissions possibles non présentes dans la séquence d'apprentissage auront une probabilité non nulle mais tout en restant faible par rapport à celle présentes dans l'apprentissage. Cela est fait comme suit :

$T=T*(\text{nombre d'exemples d'apprentissage})^2+Voisinage\_Etats$

$E=E*(\text{nombre d'exemples d'apprentissage})^2+ Etats\_Observations$

Où :

- *Voisinage_Etats* est une matrice de dimension $(\text{nombre d'états})^2$ où chaque case *(i,j)* prend la valeur 1 si les 2 neurones *i* et *j* sont voisins sur la carte de profondeur et la valeur 0 sinon.
- *Etats_Observations* est une matrice de dimension *(nombre d'états*nombre d'observations)* où chaque case *(i,j)* prend la valeur 1 pour préciser à la fonction *« hmmestimate »* que chaque état de notre HMM peut émettre n'importe quel symbole de notre ensemble d'observations.

- *nombre d'exemples d'apprentissage* : utilisé pour donner plus d'importance pour les transitions qui apparaissent dans la séquence d'apprentissage et ça en fonction de la taille de cette séquence. Car plus la séquence est longue plus est meilleure l'estimation des paramètres du HMM.

Après cette modification, les valeurs de T et E seront supérieures à 1. Pour cela on applique à nouveau la fonction *« hmmestimate »* pour ajuster ces valeurs.

5. Résultats et discussion

On a fait plusieurs expériences avec des cartes de différentes tailles (de 100 à 400 neurones) et la meilleure reconstitution obtenue est celle avec une carte de profondeur à 300 neurones et une carte de surface à 150 neurones.

La figure (fig.4) illustre les deux cartes (profondeur et surface) retenues pour l'apprentissage de notre HMM :

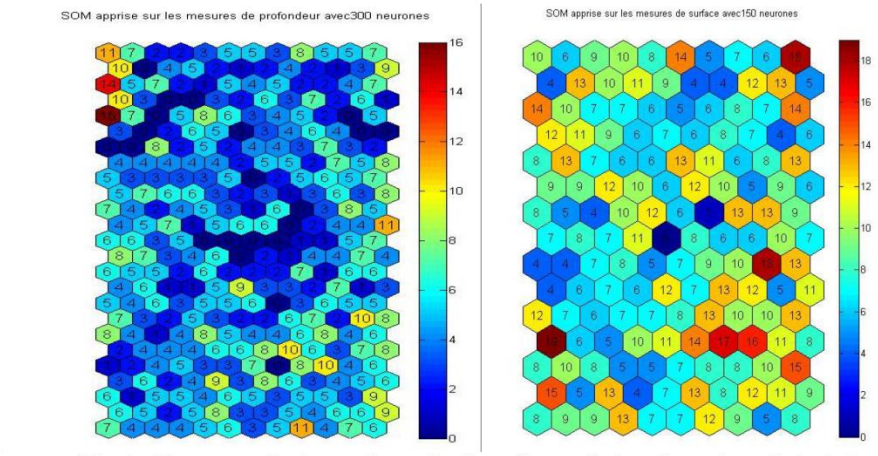


Fig.4 : Cartes topologiques de profondeur (à gauche) et de surface (à droite).

Les deux cartes (fig.4) se sont bien déployées et les données bien réparties. La figure suivante (fig.5) nous donne un aperçu sur les profils de chlorophylle captés par un neurone (courbes vertes) et le profil moyen (courbe noire). On voit bien que les profils captés par un neurone sont très proches (similaires).

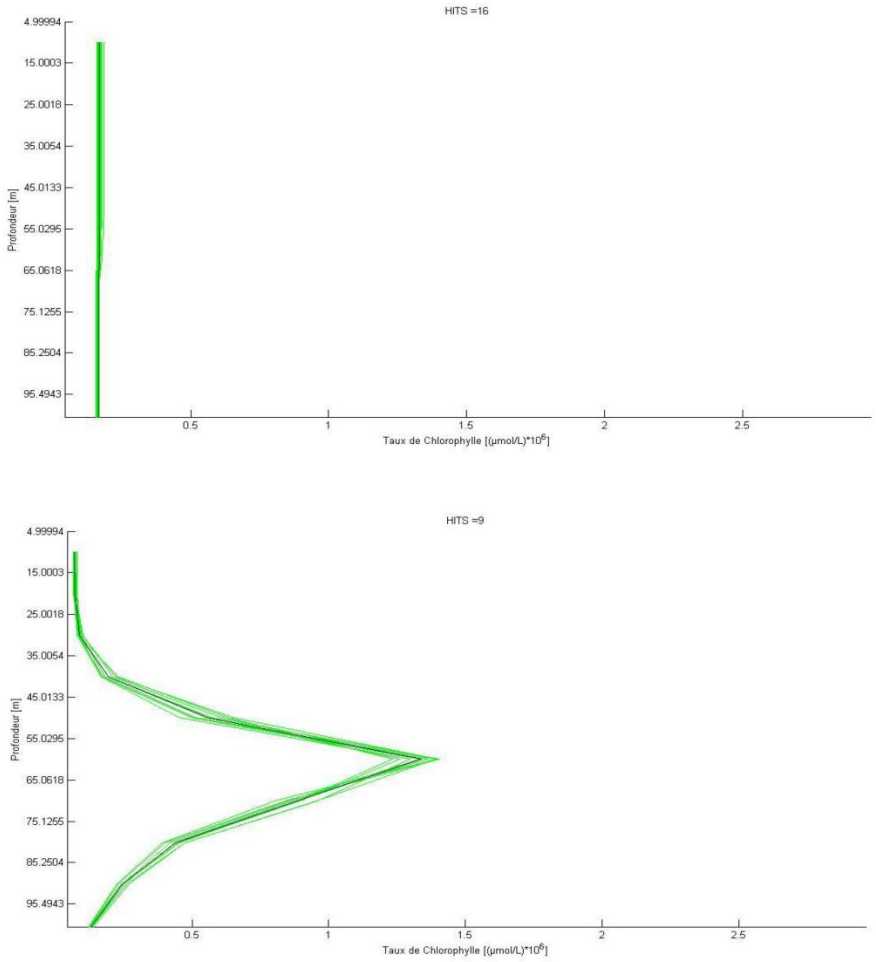


Fig.5 : Profils de chlorophylle captés par un neurone de la carte de profondeur et le profil moyen.

Dans la figure ci-dessus (fig.6), on a en haut la distribution de chlorophylle réelle pour toutes les années (1992 à 2009) en fonction de la profondeur et en bas celle reconstituée par notre modèle.

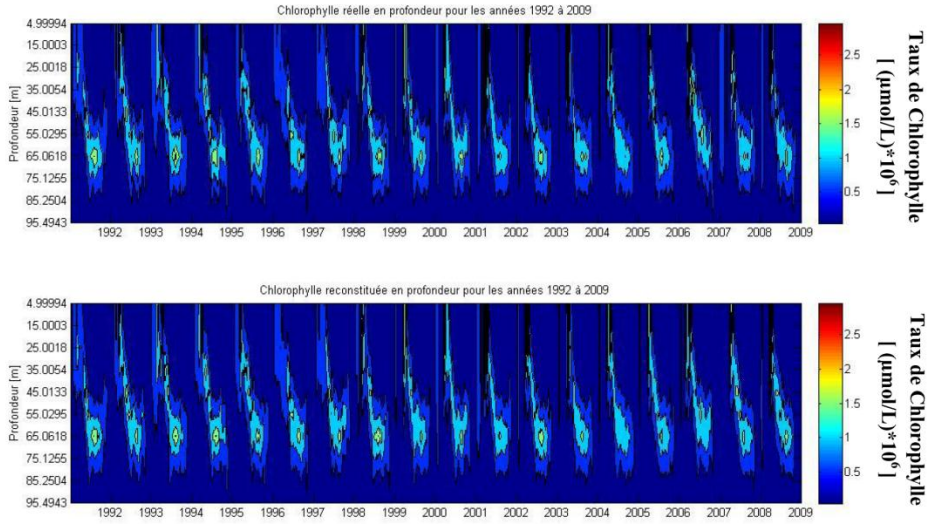


Fig.6 : Chlorophylle réelle et reconstituée pour toutes les années (1992-2009).

On remarque que la chlorophylle est bien reconstituée en forme et en intensité pour les années d'apprentissage (1992-2005). Pour les années de validation (2006-2009), les formes sont bien reconstituées mais nous observons des écarts au niveau des intensités.

La figure suivante (fig.7) nous donne un aperçu (agrandi) sur les années de validation.

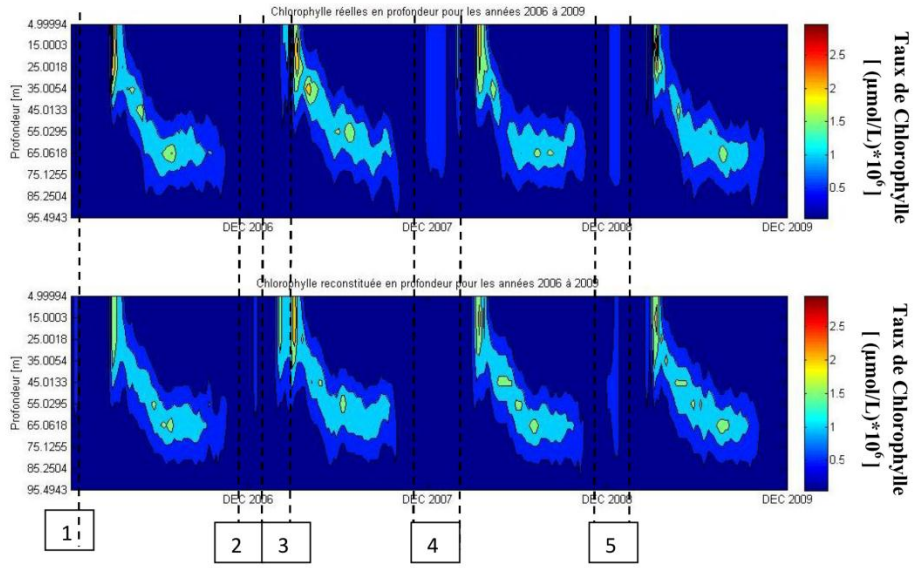


Fig.7 : Chlorophylle réelle et reconstituée pour les années de validation (2006-2009).

On voit bien que la forme est généralement bien reconstituée mais pas les intensités de chlorophylle. Cela est dû au fait qu'on a discrétisé un phénomène continu, en appliquant les cartes de topologiques, pour travailler avec des profils moyens.

La région 1 (carré 1) nous montre que le modèle a reconstitué de la chlorophylle au début de l'année 2006 alors que dans la réalité il n'y avait pas de chlorophylle. Ce problème est dû à l'algorithme de Viterbi qui a besoin d'un certain nombre de pas pour se stabiliser.

Dans la région 4 le modèle n'a pas pu reconstituer la chlorophylle à cause de son taux faible. C'est une petite portion d'un taux de chlorophylle faible au milieu d'une zone où il n'a pas de chlorophylle. Cela est dû aux probabilités de transitions du HMM : il y a plus de chance de rester dans l'état où il n'y a pas de chlorophylle que de passer à un état à taux de chlorophylle faible.

Par contre dans les régions 3 et 5, le modèle a reconstitué le profil mais pas avec la même intensité. Cela est normal vu qu'on a discrétiser l'espace des données initiales et qu'on travaille avec des profils moyens.

Et enfin, la région 2 nous montre une erreur de reconstitution qui est dû la aussi aux états de transitions du modèle.

Notons aussi que la carte de surface joue un rôle très important dans la reconstitution. En effet, une carte de surface plus grande nous permet une discrétisation plus importante mais cela peut engendrer un sur-apprentissage des années d'apprentissage.

La figure suivante (fig.8) nous montre les erreurs commises lors de la reconstitution :

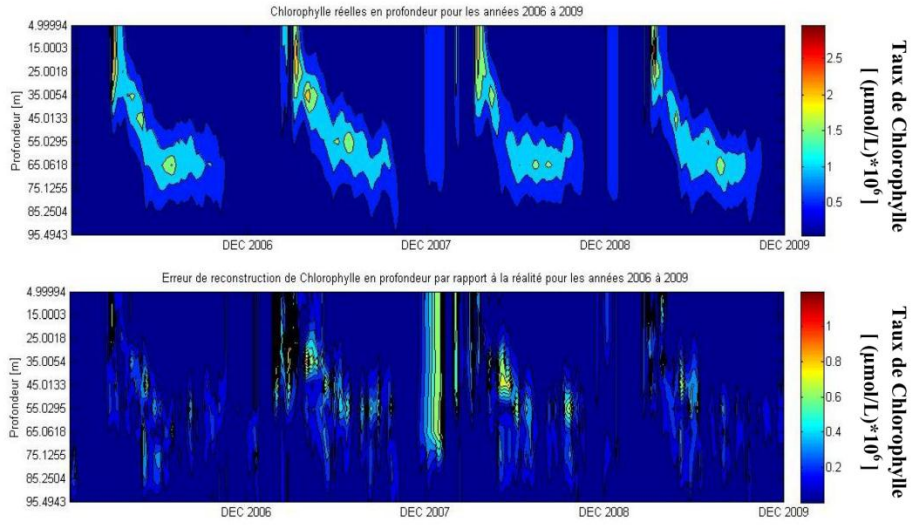


Fig.8 : Erreur de reconstitution de chlorophylle pour les années de validation (2006-2009).

On remarque qu'on commet plus d'erreurs dans les zones de transitions. En effet, on a du mal à reconstituer les transitions entre un milieu où il n'y a pas de chlorophylle et un autre où il y a une augmentation faible en chlorophylle. Cela est dû aux situations faibles en subsurface et les probabilités d'émission du HMM.

On a calculé l'erreur moyenne de reconstitution en utilisant la norme d'ordre 1 :

$$\frac{1}{|D|}\sum_{d\in D}\frac{1}{dim}\sum_{i=1}^{dim}|d_i - d_i^c|$$

Où :

- D : Ensemble de données d'apprentissage ou de validation ;
- dim : Dimension de l'espace des données ;
- $|D|$: cardinal de D ;
- d_i^c : i^{éme} composante de la donnée reconstituée.

L'erreur moyenne commise sur les données d'apprentissage est de 0.0297 et celle commise sur les données de validation est de 0.1039.

Pour montrer que notre modèle ne reconstitue pas seulement les profils de chlorophylle moyens mais il prend aussi la variabilité interannuelle, la figure ci-dessous (fig.9) nous illustre la chlorophylle réelle et reconstituée de toutes les années auxquelles on a enlevé la chlorophylle de l'année moyenne.

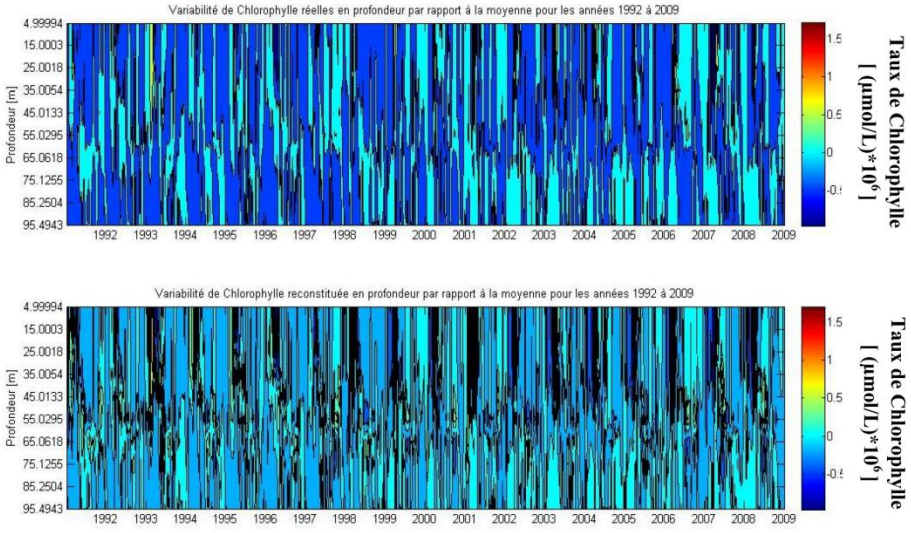


Fig.9 : Reconstitution de Chlorophylle interannuelle pour toutes les années (1992-2009).

On remarque ici que la variabilité interannuelle est faible (de l'ordre de 0.5µmol/L) ce qui nous donne des années très similaires. On voit aussi que notre modèle reconstitue bien la variabilité interannuelle. Comme pour les profils de chlorophylle, la variabilité est bien reconstituée en forme avec un écart en intensité.

Ce modèle, en plus de reconstituer les profils de chlorophylle, reconstitue bien les profils de température comme le montre la figure suivante (fig.10) :

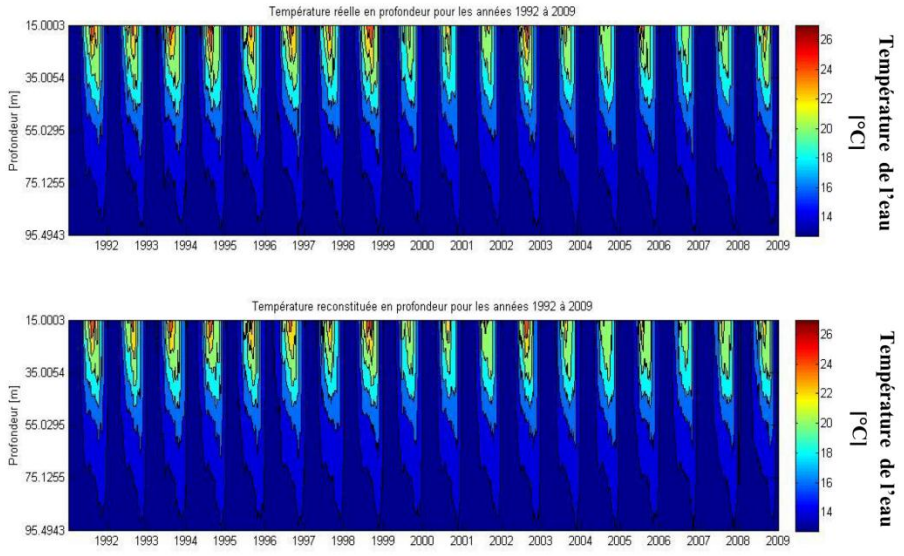


Fig.10 : Température réelle et reconstituée pour toutes les années (1992-2009).

7. Conclusion

Dans ce projet, on a testé un modèle utilisant les cartes topologiques et les modèles de Markov cachés pour la reconstitution de profils de chlorophylle à par des mesures de surface sur le site DYFAMED. Un autre binôme l'a testé aussi sur le site HOT (Hawaii) et les résultats sont assez satisfaisants.

Cependant, Il serait très intéressant de tester ce modèle avec des données in situ pour le valider vu qu'on ne dispose pas de données réelles et qu'on a dû travailler avec les données du modèle NEMO/PISCES.

Références bibliographiques

[1]	Julia UITZ. *Structure des communautés phytoplanctoniques et propriétés photophysiologiques dans les océans ouvert : paramétrisation en vue d'applications à la couleur de l'océan*. Thèse de doctorat, université de la Méditerranée, 2006.
[2]	Cours sur les cartes topologiques, deptinfo.cnam.fr/new/spip.php?pdoc2839.
[3]	Sébastien AUPETIT, Nicolas MONMARCHE et Mohamed Slimane. *Utilisation des Chaînes de Markov Cachées à Substitution de Symboles pour l'apprentissage et la reconnaissance robuste d'image.* http://www.hant.li.univ-tours.fr/webhant/pub/AupMonSli04a.majestic.pdf.
[4]	Rabiner, L. R.. *A tutorial on hidden Markov models and selected applications in speech recognition*, Proceedings of IEEE, vol. 77, num. 2, pp 257-286 (1989).

ANNEXE 3 : PROFHMM COMME GÉNÉRATEUR PROBABILISTE.

PROFHMM est un générateur probabiliste de données cohérentes entre-elles dont les paramètres ont été déterminés durant la phase d'apprentissage. Ainsi si on considère que :

- Chaque état observable émet des observations à partir d'une loi normale centrée sur son vecteur réfèrent (Neural Networks Methodology and Applications, chapter 7, « Self-Organizing Maps and Unsupervised Classification » G. Dreyfus F. Badran, M. Yacoub, and S. Thiria, SPRINGER),
- Chaque état caché choisit aléatoirement d'après ses probabilités d'émission quel état observable va émettre,
- L'évolution temporelle des états caches est déterminé par une marche aléatoire utilisant les probabilités de transition,(Berrada et al, 2009)

Nous pouvons concevoir la méthode comme un générateur probabiliste d'observations qui tient compte de la dynamique d'un phénomène physique (figure A3.1).

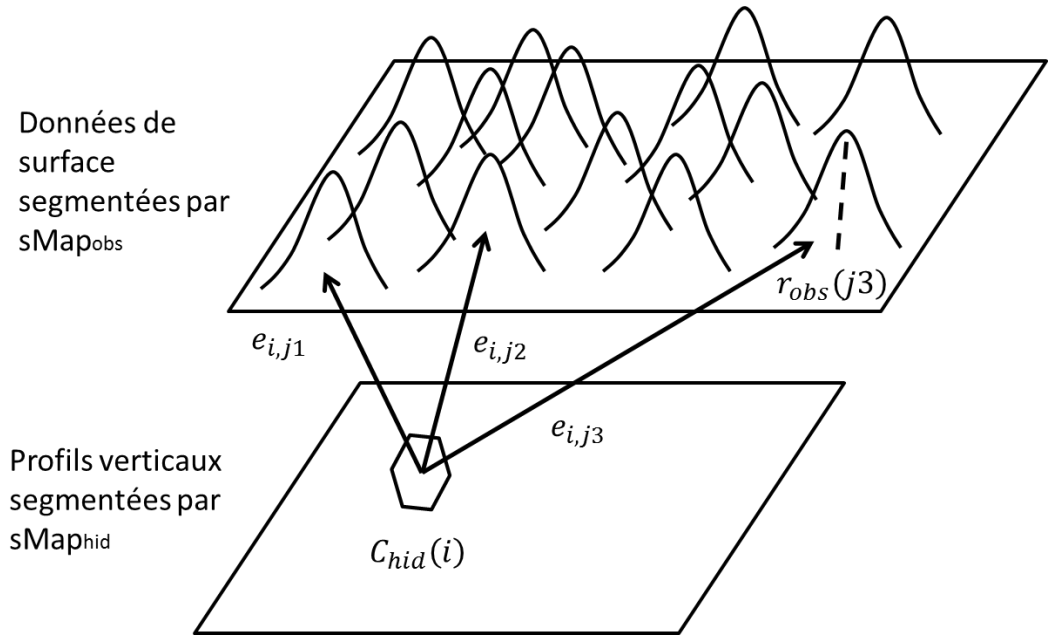


Figure A3.1 : représentation de PROFHMM comme un générateur probabiliste d'observations. Chis correspond à l'état caché émetteur, $e_{i,j}$ correspondent aux probabilités d'émissions à partir de l'état j, et r_{obs}(j3) correspond à la valeur moyenne de la gaussienne qui génère des observations si l'état observable j3 est sélectionné.

BIBLIOGRAPHIE

ARTICLE 1 :

[1] Feldman, G.C., N.A. Kuring, C. Ng, W.E. Esaias, C.R. McClain and J.A. Elrod, N.Maynard, D.Endres, R. Evans, J.Brown, S.Walsh, M. Carle, G. Podesta (1989). Ocean Color: Availability of the Global Data Set. EOS, 70, 634-641

[2] Dinnat, E., J. Boutin, G. Caudal, J. Etcheto, and P. Waldteufel, Influence of sea surface emissivity model parameters in L-band for the estimation of salinity, International Journal of Remote Sensing, 23, 5117-5122, 2002.

[3] P. Krishna Rao, W. L. Smith, and R. Koffler (January 1972). "Global Sea-Surface Temperature Distribution Determined From an Environmental Satellite". Monthly Weather Review 100 (1): 10–14

[4] Brajard J, Cédric J, Cyril M, Sylvie T, 2006 Use of a neuro-variational inversion for retrieving oceanic and atmospheric constituents from satellite ocean color sensor: Application to absorbing aerosols Neural Networks Volume 19, Issue 2, Earth Sciences and Environmental Applications of Computational Intelligence

[5] Uitz J, Claustre H, Morel A, Hooker S B 2006 Vertical distribution of phytoplankton communities in open ocean: An assessment based on surface chlorophyll, JOURNAL OF GEOPHYSICAL RESEARCH, VOL. 111, C08005, doi:10.1029/2005JC003207

[6] Gehlen M, Moussaoui A El, Perruche C, Dombrowsky E, Aumont O, Brasseur P, Le Sommer P, Lehodey P and Green Mercator consortium 2010 MERCATOR VERT - GREEN MERCATOR Integration of biogeochemistry and ecology to Mercator Ocean systems: Recent advances and future developments of the Green Mercator initiative. MyOcean Science Days 1 – 2 Mercator

[7] Gurvan M, and the NEMO team, 2012, NEMO ocean engine – version 3.4 – Note du Pôle de modélisation de l'Institut Pierre-Simon Laplace No 27 ISSN No 1288-1619.

[8] Juang B-H, 2003, Hidden Markov Models. Encyclopedia of Telecommunications

[9] Kohonen T, 1990 The Self-organizing Map, PROCEEDINGS OF THE IEEE, VOL. 78, NO 9 and http://www.cis.hut.fi/projects/somtoolbox/package/docs2/somtoolbox.html

[10] R. Jaziri, M. Lebbah, Y. Bennani, J-H. Chenot,2011, SOS-HMM: Self-Organizing Structure of Hidden Markov Model, artificial Neural Networks and Machine Learning – ICANN 2011,Lecture Notes in Computer Science Volume 6792, 2011, pp 87-94

[11] Doneya S C, Kleypasa J A, Sarmientob J L, Falkowski P G, 2002, The US JGOFS Synthesis and Modeling Project – An introduction Deep-Sea Research II 49 (2002) 1-20

[12] http://oceancolor.gsfc.nasa.gov/

[13] Miller C B, 2003, Biological Oceanography, ISBN 0632055367

[14] Jolliffe, I.T. (2002). Principal Component Analysis, second edition (Springer)

[15] A.J.Viterbi (April 1967). "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm". IEEE Transactions on Information Theory 13 (2): 260–269. doi:10.1109/TIT.1967

[16] A.J. Viterbi, 1998, "An intuitive Justification and a Simplified Implementation of a MAP Decoder for Convolutional Codes," IEEE Journal on Selected Areas in Communications, vol. 16, No. 2, Feb. 1998, pp. 260-264.

[17] J. Hagenauer, P. Hoeher,A Viterbi algorithm with soft-decision outputs and its applications, Proc. IEEE GLOBECOM, pp. 47.11-47.17, Dallas, TX, Nov 1989.

[18] R. Kohavi (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection". Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence 2 (12): 1137–1143.(Morgan Kaufmann, San Mateo, CA)

[19] ARAMIS : Altime´trie sur un Rail Atlantique et Mesures In Situ, http://aramis.locean-ipsl.upmc.fr)

ARTICLE 2 :

[1] Feldman, G.C., N.A. Kuring, C. Ng, W.E. Esaias, C.R. McClain and J.A. Elrod, N.Maynard, D.Endres, R. Evans, J.Brown, S.Walsh, M. Carle, G. Podesta (1989). Ocean Color: Availability of the Global Data Set. EOS, 70, 634-641

[2] Dinnat, E., J. Boutin, G. Caudal, J. Etcheto, and P. Waldteufel, Influence of sea surface emissivity model parameters in L-band for the estimation of salinity, International Journal of Remote Sensing, 23, 5117-5122, 2002.

[3] P. Krishna Rao, W. L. Smith, and R. Koffler (January 1972). "Global Sea-Surface Temperature Distribution Determined From an Environmental Satellite". Monthly Weather Review 100 (1): 10–14

[4] Uitz J, Claustre H, Morel A, Hooker S B 2006 Vertical distribution of phytoplankton communities in open ocean: An assessment based on surface chlorophyll, JOURNAL OF GEOPHYSICAL RESEARCH, VOL. 111, C08005, doi:10.1029/2005JC003207

[5] Altimétrie sur un Rail Atlantique et Mesures In Situ, PI S. Arnault, CNES,IRD and INSU

[6] Tanguy Y, 2011, « Variabilité de la dynamique et la thermodynamique dans l'Atlantique tropical : Projet ARAMIS » , phd Thèsis, UPMC

[7] CORIOLIS Data Centre (http://www.coriolis.eu.org/cdc/argo.htm)

[8] Kohonen T, 1990 The Self-organizing Map, PROCEEDINGS OF THE IEEE, VOL. 78, NO 9 and http://www.cis.hut.fi/projects/somtoolbox/package/docs2/somtoolbox.html

[9] Charantonis A, Badran F, Thiria S, 2012, Retrieving the vertical profiles of Chlorophyll-A from satellite observations, by using hidden Markov models and self-organizing maps. JAOT, submitted

[10] Gurvan M, and the NEMO team, 2012, NEMO ocean engine – version 3.4 – Note du Pôle de modélisation de l'Institut Pierre-Simon Laplace No 27 ISSN No 1288-1619.

[11] http://oceancolor.gsfc.nasa.gov/

[12] Juang B-H, 2003, Hidden Markov Models. Encyclopedia of Telecommunications

[13] A.J.Viterbi (April 1967). "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm". IEEE Transactions on Information Theory 13 (2): 260–269. doi:10.1109/TIT.1967

[14] ARGO Science Team 1998; see http://www.argo.ucsd.edu

[15] Tanguy, Y., et al., Isothermal, mixed, and barrier layers in the subtropical and tropical Atlantic Ocean during the ARAMIS experiment. Deep-Sea Research I (2010), doi:10.1016/j.dsr.2009.12.012

[16] http://www.aviso.oceanobs.com/duacs/

[17] Jolliffe, I.T. (2002). Principal Component Analysis, second edition (Springer)

ARTICLE 3 :

1 A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," Journal of the Royal Statistical Society. Series B, vol. 39, no. 1, pp. 1–38, 1977.

2 Z. Ghahramani and M. I. Jordan, "Supervised learning from incomplete data via an EM approach," in Advances in Neural Information Processing Systems (NIPS 6), pp. 120–127, Morgan Kauffman, San Fransisco, Calif, USA, 1994.

3 I. Wasito and B. Mirkin, "Nearest neighbour approach in the least-squares data imputation algorithms," Information Sciences, vol. 169, no. 1-2, pp. 1–25, 2005.

4 S. Chiewchanwattana, C. Lursinsap, and C.-H. H. Chu, "Imputing incomplete time-series data based on varied-window similarity measure of data sequences," Pattern Recognition Letters, vol. 28, no. 9, pp. 1091–1103, 2007.

5 S. Prasomphan, C. Lursinsap, and S. Chiewchanwattana, "Imputing time series data by regional-gradient-guided bootstrapping algorithm," in Proceedings of the 9th International Symposium on Communications and Information Technology (ISCIT '09), pp. 163–168, Incheon, South Korea, September 2009

6 E. Grayzeck, 2011, NATIONAL SPACE SCIENCE DATA CENTER, ARCHIVE PLAN FOR 2010 – 2013, NSSDC Archive Plan '10-13.

7 A. R. Robinson and P.F.J. Lermusiaux “Overview of data assimilation” Harvard Reports in Physical/Interdisciplinary, Ocean Science, NUMBER 62

8 A.J.Viterbi (April 1967). "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm". IEEE Transactions on Information Theory 13 (2): 260–269. doi:10.1109/TIT.1967

9 A.A.Charantonis J. Brajard C. Moulin F. Bardan S. Thiria, Inverse Method for the Retrieval of Ocean Vertical Profiles using Self Organizing Maps and Hidden Markov Models - Application on Ocean Colour Satellite Image Inversion. IJCCI (NCTA) 2011: 316-321

10 R. Jaziri, M. Lebbah, Y. Bennani, J-H. Chenot,2011, SOS-HMM: Self-Organizing Structure of Hidden Markov Model, artificial Neural Networks and Machine Learning – ICANN 2011,Lecture Notes in Computer Science Volume 6792, 2011, pp 87-94

11 Madec G. 2008: "NEMO ocean engine". Note du Pole de modélisation, Institut Pierre-Simon Laplace (IPSL), France, No 27 ISSN No 1288-1619

12 A.S. Willsky, 2002, Multiresolution Markov Models for Signal and Image Processing, Proceedings of the IEEE, Vol. 90, No. 8.

13 Jolliffe, I.T. (2002). Principal Component Analysis, second edition (Springer)

14 C. Kwan, X. Zhang, R. Xu, and L. Haynes, 2003,“A Novel Approach to Fault Diagnostics and Prognostics” Proeerdlngs of the 2003 lEEE, lnlernafioosl Conference Robotics & Aulornstion,Taipei, Taiwan

15 T. Kohonen, 1990 The Self-organizing Map, PROCEEDINGS OF THE IEEE, VOL. 78, NO 9 and http://www.cis.hut.fi/projects/somtoolbox/package/docs2/somtoolbox.html

16 Doneya S C, Kleypasa J A, Sarmientob J L, Falkowski P G, 2002, The US JGOFS Synthesis and Modeling Project – An introduction Deep-Sea Research II 49 (2002) 1-20

17 A.J. Viterbi, 1998, “An intuitive Justification and a Simplified Implementation of a MAP Decoder for Convolutional Codes,” IEEE Journal on Selected Areas in Communications, vol. 16, No. 2, Feb. 1998, pp. 260-264.

18 J. Hagenauer, P. Hoeher,A Viterbi algorithm with soft-decision outputs and its applications, Proc. IEEE GLOBECOM, pp. 47.11-47.17, Dallas, TX, Nov 1989.

Annexe 3:

1. Neural Networks Methodology and Applications, chapter 7, « Self-Organizing Maps and Unsupervised Classification » G. Dreyfus F. Badran, M. Yacoub, and S. Thiria, SPRINGER