Table des matières

Chapitre I. INTRODUCTION.

I.1 QUESTIONNEMENT SCIENTIFIQUE

Les échanges d'eau au sein du cycle global hydrologique sont déterminés par des contraintes mécaniques et thermodynamiques complexes qui forment les bases du système de dynamique du climat. La salinité océanique (S) est l'une des variables de ce cycle les plus délicates à observer (U.S. CLIVAR Office, 2007), surtout en Atlantique où elle présente une forte variabilité spatiale. S est aussi un indice d'intensité et de localisation des échanges d'eau au niveau de l'interface océan-atmosphère (Dessier, et al., 1994). La mise au point de modèles mathématiques permettant de relier les paramètres de surface, plus accessibles telles que la SSH: Sea Surface Height, l'ADT : Absolute Dynamic Topography, la SST: Sea Surface Temperature, la SSS: Sea Surface Salinity,...le profil de température (T) et le profil de S en fonction de l'immersion, serait un atout précieux pour l'obtention d'informations à exploiter en terme de variabilité océanique et climatique.

Le but de la thèse est de reconstruire à l'échelle régionale le profil de S océanique à partir des données de surface observées et celles issues des modèles et à l'aide d'outils statistiques avancés adaptés à ce type de systèmes complexes.

Le chantier se situe en Atlantique tropical, sur une période allant de 2002 à 2012. Mesures in situ (réseau de navires marchands, flotteurs lagrangiens Argo), climatologies, expériences numériques et satellites y sont conjointement disponibles.

La méthode statistique est mise au point à partir des résultats d'une expérience numérique. Cette dernière est comparée à des données observées afin de juger de ses capacités à compléter les données disponibles. Les résultats obtenus avec la méthode statistique sont confrontés aux mesures in situ prises comme "vérité terrain". Le challenge futur étant de disposer d'un outil statistique qui relierait les profils de S aux observations de SST, de SSS, et de hauteur d'eau données par satellite, et donc d'atteindre des échelles de variabilité de temps et d'espace inégalées par les mesures traditionnelles.

Pour la mise en œuvre du modèle statistique, nous avons adapté une approche « systèmes complexe » qui est définie dans le paragraphe suivant.

I.2 L'APPROCHE « SYSTEMES COMPLEXES »

L'idée de systèmes de complexes existe depuis longtemps, mais lors de cette dernière décennie (depuis 2000), elle a suscité un intérêt particulier et d'importants travaux associant des scientifiques de diverses disciplines ont été faits. Il existe plusieurs définitions des systèmes complexes. Ils sont souvent définis à travers leurs caractéristiques. Ainsi, *Advances in Complex Systems Journal,* qualifie un système complexe comme un système comprenant un nombre (généralement grand) d'interactions (généralement fortes) entre entités, processus ou agents, dont la compréhension nécessite le développement ou l'utilisation, de nouveaux outils scientifiques, des modèles non linéaires hors de description d'équilibre et/ou de simulations informatiques. Herbert Simon définit un système complexe comme un système constitué de plusieurs composants ayant un nombre de relations relativement important entre eux, de sorte que le comportement de chaque élément dépend du comportement des autres (Simon, 1962). Dans de tels systèmes, l'ensemble est supérieur à la somme des parties (Simon, 1962). Jerome Singer, quant à lui, parle d'un système qui fait intervenir de nombreux agents en interaction dont la somme des comportements doit être comprise. Cette activité globale est non linéaire, donc ne peut pas être simplement déduite de la somme des comportements individuels des composants. Vus ainsi, ces types d'interactions nécessitent une nouvelle manière de voir, une nouvelle approche.

L'approche « Systèmes Complexes » dans l'étude des systèmes tels que les phénomènes géophysiques, la propagation des épidémies, l'Internet, est très récente, elle date des années 2000. Elle consiste à utiliser un ensemble de théories et de techniques pour comprendre et modéliser des entités en interaction. En l'espace de quelques années, la théorie des réseaux est devenue l'un des outils les plus populaires que l'on peut appliquer à la description, à l'analyse et à la compréhension des systèmes complexes. Plusieurs articles - (Strogatz, 2001) et (Albert et al., 2002)- et des livres de vulgarisation scientifique font leur apparition notamment celui controversé de Stephen Wolfram « A New Kind of Sciences » (Wolfram, 2002) qui aborde également les automates cellulaires et la modélisation orientée agent. D'autres outils plus anciens sont également utilisés dans l'étude des systèmes complexes. La théorie du Chaos -la dynamique non linéaire ou systèmes dynamiques- semble être établie de fait. L'utilisation des méthodes dynamiques

non linéaires était déjà profondément imbriquée dans la recherche scientifique avant cette période. Ce qui n'était pas le cas pour les modèles orientés agent (Ottino, 2003) et les modèles basés sur l'utilisation des réseaux de neurones.

Pour l'étude des systèmes géophysiques (océan, atmosphère,...) qui présentent de nombreuses propriétés caractéristiques des systèmes complexes, différentes approches sont utilisées. La plus courante d'entre elles est la modélisation numérique mais elle ne prend pas toujours en compte toute la complexité qui caractérise ces systèmes à travers leurs paramètres (Wu et al., 2012).

De nos jours, les méthodes utilisant la connaissance liée aux observations sont de plus en plus utilisées. Notre approche est basée sur ces dernières.

La méthodologie proposée repose sur une classification automatique des paramètres à l'aide de cartes auto-organisatrices de Kohonen (SOM : Self Organizing Map). Dans un premier temps, les profils verticaux de S associés aux paramètres physiques de surface pertinents et aux profils de T sont classifiés de façon à décrire toute la variabilité des situations rencontrées. Cette étape permet notamment de faire apparaître différents régimes océaniques pour lesquels on peut trouver une relation spécifique entre les paramètres de surface, les profils de T et les profils de S. Nous rappelons que le terme « profil » s'applique ici à la discrétisation de la variable (T ou S) en fonction de la profondeur en un point géographique et un instant donnés.

Dans un second temps, les paramètres observés sont associés à un ou plusieurs régimes déterminés lors de la classification précédente.

La mise au point d'une procédure d'optimisation globale permet de déterminer le profil vertical de S le plus cohérent. Cette optimisation est réalisée par un algorithme de marche aléatoire sur la carte de Kohonen qui optimise à la fois la connaissance de surface a priori induite par les observations satellitaires et la vraisemblance des profils verticaux qui sont eux non observés. La cohérence sera quantifiée en utilisant les corrélations spatiales et temporelles existantes entre les différents paramètres pris en compte.

C'est l'ensemble de ces données (modélisées, in situ et satellites) associées à la méthode statistique développée qui fournit à cette thèse son originalité. Le thème donne un cas d'école exemplaire de ce que peuvent apporter les méthodes statistiques de modélisation des systèmes complexes aux questionnements géophysiques, notamment quand ces

derniers se heurtent à des problèmes de manque d'observations (in situ), de limite d'expérience ("vision" du satellite limitée à la surface océanique) et de précision des forçages et de résolution (modèles numériques).

La méthodologie proposée repose sur plusieurs étapes:

- ✓ une étude statistique préalable des données issues d'un modèle numérique afin d'explorer les dépendances qui existeraient entre les niveaux d'immersion,
- ✓ une inter-comparaison de ces données de modèle et des données in situ,
- ✓ une classification automatique des paramètres à l'aide de SOM,
- ✓ un développement de méthodes d'inversion à partir des paramètres de surface.

Ces étapes de mise en œuvre du modèle d'inversion statistique sont suivies par son évaluation et son application. Ces dernières sont présentées via:

- ✓ l'application et évaluation du modèle sur des données in situ,
- ✓ la reconstruction et analyse de profils de S en Atlantique tropical.

Une exploitation préliminaire des données a permis de rendre compte de l'importance des paramètres qui sont utilisés dans l'inversion. Cette exploitation est complétée par l'étude de l'interdépendance des paramètres pour déterminer leur degré de liaison. Une extraction de caractéristiques pertinentes (feature selection) est ensuite appliquée aux données. Le chapitre suivant décrit les données et les méthodes utilisées dans ce travail.

Chapitre II. DONNEES ET PRETRAITEMENT.

Dans ce chapitre, nous allons présenter les données et l'ensemble des prétraitements majeurs effectués sur celles-ci. Les données concernent des profils de S et de T, des données d'ADT (ou de SSH), de SSS et de SST. Elles proviennent de différentes sources parmi lesquelles les campagnes de mesure in situ ARAMIS, les données simulées par l'expérience numérique DRAKKAR, celles du programme Coriolis et des missions altimétriques. Ainsi, le terme de "donnée" s'appliquera ici aussi bien aux résultats d'expériences numériques qu'aux données expérimentales issues de mesures océanographiques de terrain ou satellitales. Ces sources présentent différentes résolutions spatio-temporelles. Des prétraitements ont été nécessaires pour non-seulement résoudre les problèmes liés aux différences de résolution mais aussi pour établir un niveau de confiance sur les données modèles. L'objet (la donnée) principal de notre étude est S qui désigne la quantité de sels dissous dans un liquide, notamment l'eau qui est un puissant solvant pour de nombreux minéraux. La salinité de l'océan est la quantité totale de sels dissous dans l'eau de mer. Cette S est de 35 g par kg d'eau en moyenne, et reste généralement comprise entre 30 g/kg (Atlantique Nord) et 40 g/kg (mer Rouge). Bien que le sujet fasse débat dans la communauté océanographique, nous adopterons comme unité le psu (practical salinity unit, 1 psu=1 g/kg). Notons que ce choix n'altère en rien nos résultats. Dans la section suivante, quelques instruments de mesure de la salinité sont présentés.

II.1 LES INSTRUMENTS DE MESURE DE LA SALINITE

Aujourd'hui la plupart des mesures in situ de la salinité repose sur la mesure de la conductivité de l'eau de mer (Martin, 2013). Il existe différents instruments de mesure du profil de S, les plus utilisés sont les flotteurs et les CTDs (Conductivity-Temperature-Depth). Ces instruments sont présentés ci-après.

II.1.1 Les flotteurs Argo

Le flotteur est, d'une façon générale, un instrument (sous-)marin autonome qui mesure un paramètre (T, S, oxygène,…) au coeur des océans. Il est programmé à l'avance et est déployé à partir de navires océanographiques ou d'opportunité. Dès lors, il réalise des

cycles de mesures jusqu'à perte de son autonomie.

Les flotteurs « Argo » sont gérés par ARGO, un programme international d'observations de l'océan à partir de flotteurs autonomes (http://www.argo.ucsd.edu). Ces flotteurs mesurent donc en temps réel des grandeurs physiques permettant de caractériser l'océan et sa variabilité, comme T et S sur plusieurs niveaux d'immersion. Ils sont déployés via des équipes d'organismes scientifiques qui les paramètrent en fonction des capacités de leur modèle. Le cycle des flotteurs Argo est de 10 jours. Chaque cycle comporte une descente de quelques heures vers une immersion de référence (généralement 1000m), où le flotteur dérive pendant environ 9 jours, mesurant régulièrement T et S. Puis, il plonge à une immersion (généralement 2000m) à laquelle il démarre un "profil remontée" en échantillonnant T et S comme illustré par la Figure II-1. Dès son arrivée en surface, il transmet ses données vers un satellite puis entame le cycle suivant.

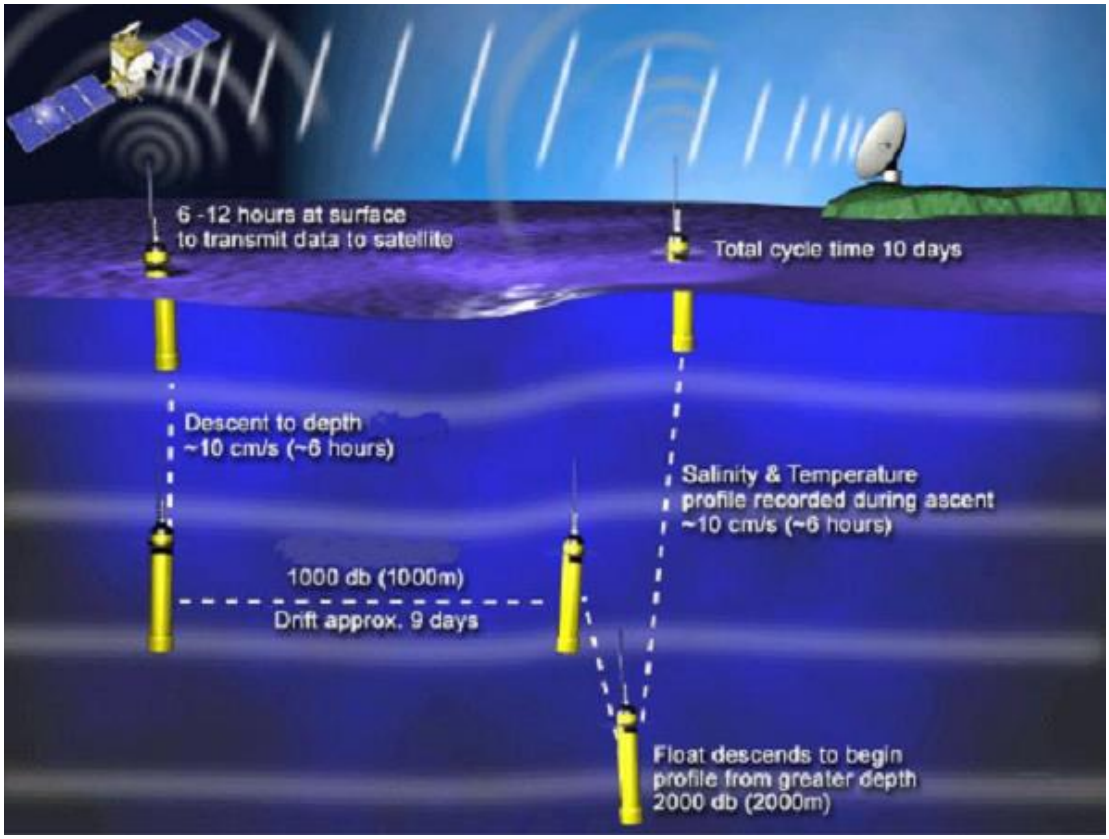


Figure II-1: Illustration du fonctionnement d'un flotteur Argo

Tous les flotteurs Argo permettent la mesure de T et un grand nombre sont aussi équipés de conductimètre afin de mesurer S. La résolution verticale des Argo peut fortement varier

entre les différents flotteurs.
Ces flotteurs constituent plus de 95% des fichiers que nous avons extraits de la base de données Coriolis sur la zone d'étude.

II.1.2 Les CTDs, XCTDs et XBTs

Les profils « CTDs » sont mesurés par des sondes CTDs. La CTD est un instrument électronique couramment utilisé par les océanographes lors des campagnes océanographiques. Elle enregistre en continu S (en mesurant la conductivité), T et la profondeur (par mesure de la pression). La CTD permet donc d'effectuer des profils verticaux à partir de la surface de la mer lors des stations hydrologiques. D'autres capteurs peuvent être installés sur une CTD: des capteurs d'oxygène dissous dans l'océan, des capteurs de fluorescence qui permettent de mesurer la chlorophylle, etc.

Les CTDs nécessitent toutefois un navire océanographique pour leur mise à l'eau, et du temps bateau pour réaliser les stations. C'est pourquoi s'est également développée dans les années 80 la possibilité de réaliser des profils T ou T et S à partir de bateaux en marche comme les navires marchands. Il s'agit des sondes XBTs (eXpendable BathyThermograph) pour T, et XCTDs (eXpendable Conductivity Temperature Depth) pour T et S. Le principe consiste à envoyer dans l'océan une sonde « jetable » reliée à un ordinateur par un fil conducteur qui va enregistrer et transmettre le long de la descente les données relevées par des mini-capteurs. A une certaine immersion (qui dépend du type de sonde employée et de la vitesse du navire et donc de l'étirement du fil, généralement entre 700 et 800m), le fil conducteur se casse par étirement, la sonde est perdue mais les données sont enregistrées. D'une précision moindre que les CTDs, l'énorme avantage de ce système réside en l'absence de mise en place de moyens « lourds ». Il est toutefois depuis 2010 fortement concurrencé par les flotteurs Argo et XCTDs et XBTs se font rares.
En plus de ces types de mesure in situ de profils de S, il existe d'autres moyens de mesurer S comme les thermosalinographes (cf http://www.legos.obs-mip.fr/observations/sss/) mais ces mesures se limitent à la « surface » océanique (même si l'immersion de prélèvement peut varier de plusieurs mètres).
La Figure II-2, présente 2 profils de S échantillonnés en Atlantique tropical. Les profils

sont donnés avec une mesure à chaque mètre d'immersion.

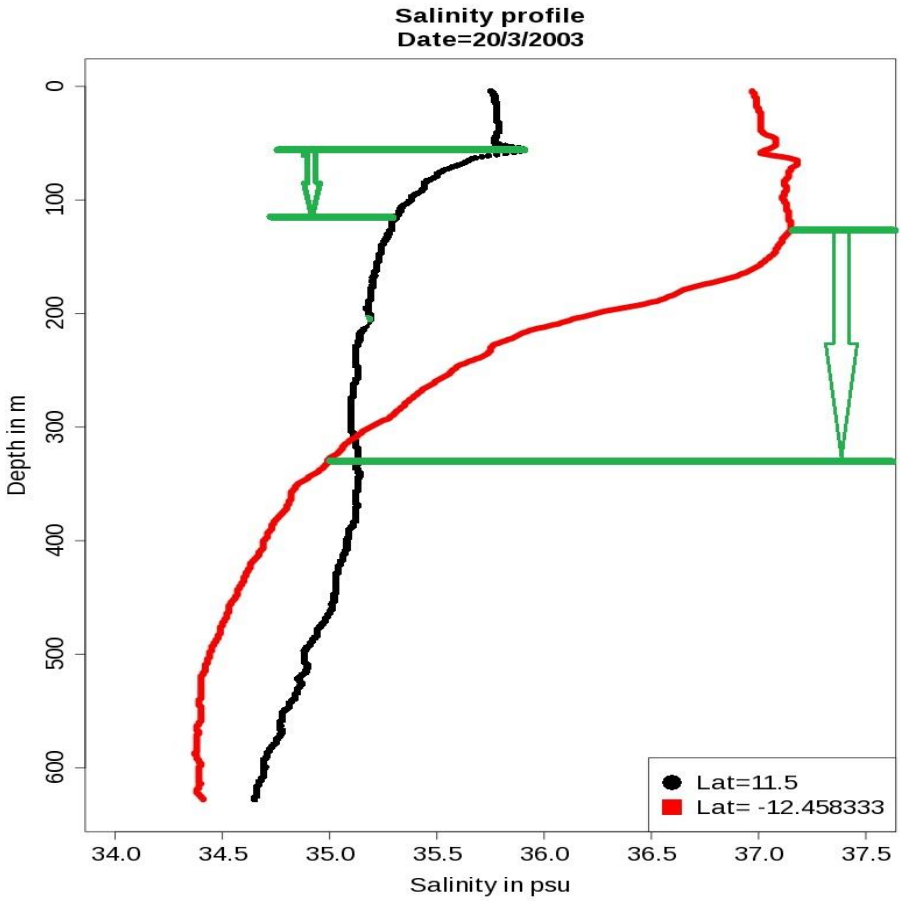


Figure II-2 : Exemples de profils de S. 2 profils in situ du projet ARAMIS (cf. II.2) du même jour à des latitudes différentes (couleur rouge mesuré à 12.46° S et noire mesuré à 11.5° N). Les traits verts indiquent les limites de l'halocline.

On remarque une grande différence entre ces profils pourtant réalisés à la même date. Cette différence traduit la localisation et les différentes masses d'eau échantillonnées par ces profils : eaux salées en surface, dessalées en sub-surface de l'hémisphère Sud, et inversement dans l'hémisphère Nord. Malgré ces différences, il existe une similitude dans leur forme. Ils peuvent être décomposés principalement en 3 parties délimitées par les traits verts sur la figure : entre la surface et le 1er trait vert, S varie légèrement c'est-à-dire qu'on a une couche relativement homogène en S ; en dessous, nous notons une décroissance très brusque des S. Ce fort gradient est connu sous le nom de halocline. Puis dans la dernière partie redevient presque constante. Ces profils sont donc constitués de 2

zones homogènes et entre les 2, on a une zone de gradient très fort. La délimitation de cette zone, notamment sa profondeur de début et de fin est très difficile à faire car dépendant des paramètres très variés parmi lesquels on peut citer la latitude, la longitude, les précipitations, les courants marins et surtout les sous-courants, la hauteur d'eau, etc.
Pour reconstruire ces profils de S, nous allons aussi utiliser des données de surface telles que la hauteur d'eau (SSH : Sea Surface Height ou ADT : Absolute Dynamic Topography), la SST (Sea Surface Temperature), la longitude et la latitude et aussi des profils de T. 4 ensembles de données ont été utilisés dans cette thèse. Ils concernent des données in situ et des données simulées issues de modèle numérique. Ces ensembles sont présentés dans les paragraphes suivants.

II.2 LE PROJET DE MESURES IN SITU ARAMIS

De 2002 à 2008, le projet ARAMIS (Altimétrie sur un Rail Atlantique et Mesures In Situ, http://aramis.locean-ipsl.upmc.fr/, responsable S. ARNAULT) a mis en place en Atlantique tropical un suivi basse fréquence des structures thermo-halines de la couche 0-1000m sur une ligne de bateau marchand. La ligne de bateaux retenue est la ligne World Ocean Circulation Experiment WOCE-AX11 entre l'Europe et l'Amérique du Sud comme indiqué par la Figure II-3. Cette ligne est particulièrement intéressante puisqu'elle traverse les grands courants zonaux de l'Atlantique tropical ainsi que la trace au sol de la position moyenne de la Zone de Convergence Intertropicale (ZCIT ou ITCZ). La ligne échantillonne également les 2 zones de formation des eaux de S maximum, véritable marqueur d'anomalies climatiques dans leur hémisphère respectif, et se superpose à une trace de la mission Jason.
L'expérience ARAMIS a été conduite de juillet 2002 à octobre 2008, soit 13 campagnes de mesures. 2 fois par an, en automne et au printemps boréal, des sondes permettant de recueillir profils de T (XBT pour eXpendable BathyThermograph) et de S (XCTD pour eXpendable Conductivity-Temperature-Depth, cf. II.1.2) entre la surface et 700m de profondeur sont lancées depuis le navire, entre 35°N et 20°S. La résolution finale est de 0.5° en T et 1° en S. Des mesures de thermosalinographe (SST, SSS) sont également recueillies le long de la route, ainsi que des indications météorologiques.

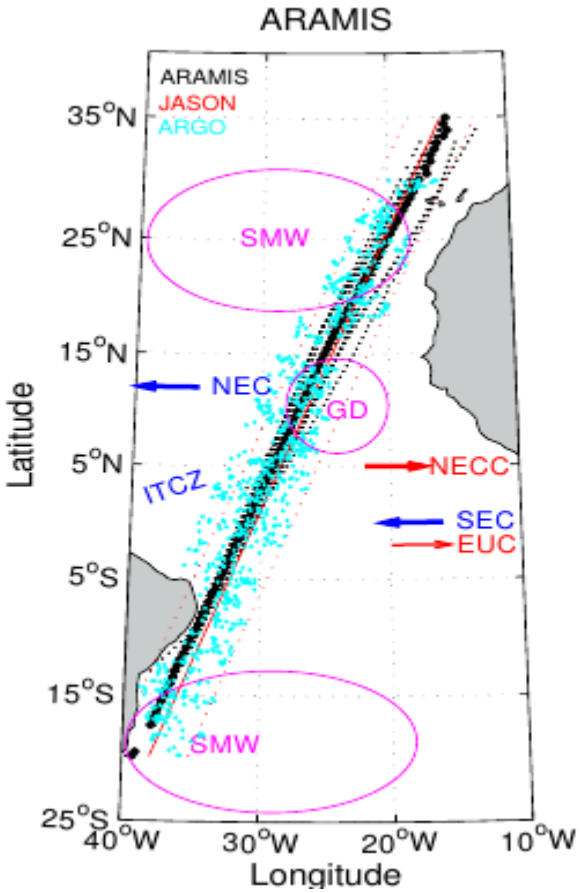


Figure II-3: Routes ARAMIS : les flèches indiquent les courants, NECC (North Equatorial ConterCurrent), NEC (North Equatorial Current), SEC (South Equatorial Current), EUC(Equatorial UnderCurrent). Les points noirs= différentes campagnes ARAMIS, points bleus ciels= flotteurs Argo, pointillés rouges= les traces des satellites JASON. D'après Arnault et al., 2011.

Concrètement, chaque « tirs » d'XBT ou XCTD donne lieu à la création d'un fichier numérique qui comporte les informations sur la longitude (x), la latitude (y), la date de lancement (t) de la sonde en entête de fichier et la valeur de S (XCTD) et de T (XBT, XCTD) ré-interpolée tous les mètres de profondeur.

Le prétraitement de ces fichiers a permis de recueillir dans une matrice les mesures sous forme de profils de S c'est-à-dire pour chaque point (caractérisé par une longitude x en degré, une latitude y en degré et une date t (jour/mois/année)), nous avons une mesure de S à chaque mètre, allant de la surface à 700m de profondeur (z).

Nous avons également utilisé des données issues du modèle numérique DRAKKAR.

II.3 L'EXPERIENCE NUMERIQUE DRAKKAR

Cette section comporte 2 parties, une présentation du modèle DRAKKAR et une étude comparative DRAKKAR/ARAMIS

II.3.1 Présentation du modèle

Le projet DRAKKAR (http://www.ifremer.fr/lpo/DRAKKAR/index.htm) construit une hiérarchie de modèles numériques sur la base du système NEMO «Nucleus for European Modelling of the Ocean» (Barnier, et al., 2006). Ce système inclut l'outil national français de modélisation océanique OPA9, le modèle de glace de mer LIM, et le code de traceurs et biogéochimie TOP.

Ces modèles simulent sur la période 1950-2007 l'évolution tridimensionnelle de la circulation et des masses d'eaux océaniques, des glaces de mer, et de composants chimiques dissous (CFCs, 14C) sur 46 niveaux d'immersion répartis irrégulièrement de la surface au fond (5875). Le *Tableau II-1* présente ces immersions.

Ces simulations sont forcées en surface par des reconstructions globales journalières des vents (modèle du Centre Européen de Prévision Météorologique à Moyen Terme de Readings), de T et de l'humidité atmosphériques depuis les années 50. Une résolution horizontale de 1/4° (soit ~30 km dans les tropiques), qui permet aux tourbillons de se développer, est utilisée dans une configuration de l'océan global (ORCA025), et dans une configuration resserrée autour de l'Atlantique Nord et des Mers Nordiques (de 30° S à 80° N).

Tableau II-1 : répartition des niveaux d'immersion de DRAKKAR

Niveau	Immersion (en m)	Niveau	Immersion (en m)
1	3.05	24	989.23
2	9.45	25	1136.92
3	16.36	26	1297.72
4	23.9	27	1470.89
5	32.21	28	1655.47
6	41.48	29	1850.37
7	51.95	30	2054.41
8	63.88	31	2266.45
9	77.62	32	2485.37
10	93.59	33	2710.13

11	112.28	34	2939.81
12	134.28	35	3173.59
13	160.28	36	3410.76
14	191.09	37	3650.71
15	227.62	38	3892.95
16	270.9	39	4137.05
17	322.02	40	4382.65
18	382.14	41	4629.48
19	452.44	42	4877.3
20	534.02	43	5125.92
21	627.85	44	5375.18
22	734.72	45	5624.95
23	855.11	46	5875.14

Les résultats utilisés dans la thèse concernent les simulations des profils de S, de la SST, les hauteurs d'eau (SSH) le long de la route des campagnes de mesures in situ ARAMIS (données ARAMIS, cf. II.2) de 2002 à 2007.

Suivant nos besoins, différentes extractions des données DRAKKAR ont été faites autour des routes ARAMIS ré-échantillonnées

Dans un premier temps, nous avons colocalisé DRAKKAR sur ARAMIS et les 2 ensembles ont été ré-échantillonnés.

II.3.2 Colocalisation spatiale DRAKKAR/ARAMIS

La co-localisation spatiale consiste à trouver pour chaque point (x,y) ARAMIS, le point DRAKKAR qui lui est le plus proche. La Figure V-2 **Erreur ! Source du renvoi introuvable.** présente en exemple la route ARAMIS1 réelle et son équivalent DRAKKAR. ARAMIS1 désigne la première campagne ARAMIS, dans la suite ARAMISN désignera la énième campagne d'ARAMIS et DRAKKARN désignera son équivalent DRAKKAR.

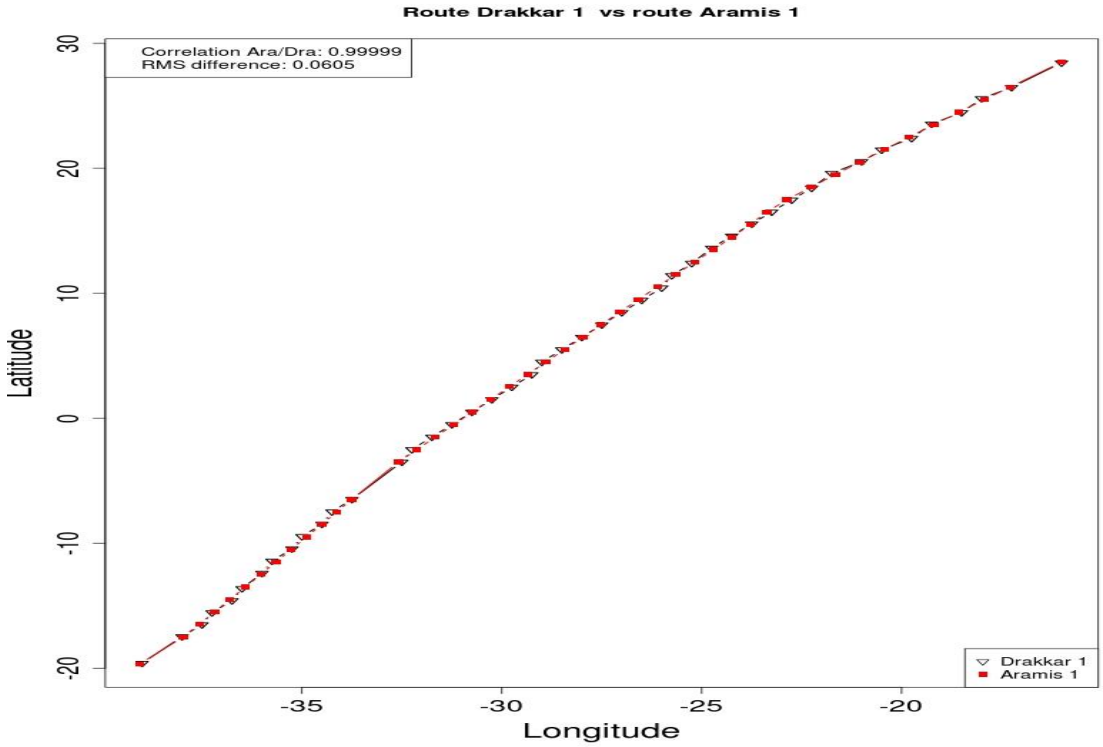


Figure II-4: Exemple de co-localisation d'ARAMIS1 par DRAKKAR. Les carrés rouges représentent la route ARAMIS 1 et les triangles vides son équivalent DRAKKAR (DRAKKAR1).

La corrélation qui est presque de 1 montre que la route DRAKKAR1 suit correctement la route ARAMIS1. Cette situation est celle rencontrée sur toutes les co-localisations spatiales de DRAKKAR sur les routes ARAMIS. Cette précision est due au fait que le modèle à une résolution de 0,25°. L'algorithme utilisé pour cette co-localisation spatiale considérant le point DRAKKAR qui est le plus proche en termes de distance euclidienne, un point ARAMIS sera au plus distant de 0,25/2 soit 0,125° de son équivalent DRAKKAR.

II.3.3 Co-localisation temporelle DRAKKAR/ARAMIS

Pour cette recherche nous avons utilisé pour chaque campagne ARAMIS les données DRAKKAR ayant la date la plus proche de la date moyenne de la campagne considérée. Le

Tableau II-2 suivant donne les dates DRAKKAR utilisées

Tableau II-2:Dates DRAKKAR co-localisées ARAMIS (JJ/MM/AAAA)

Campagne	Date début ARAMIS	Date FIN ARAMIS	Date DRAKKAR
1	20/07/2002	27/07/2002	24/07/2002
2	15/03/2003	23/03/2003	21/03/2003
3	12/10/2003	19/10/2003	17/10/2003
4	03/05/2004	10/05/2004	05/05/2004
5	10/05/2004	20/09/2004	17/09/2004
6	28/04/2005	05/05/2005	30/04/2005
7	05/10/2005	12/10/2005	07/10/2005
8	08/05/2006	14/05/2006	10/05/2006
9	23/10/2006	29/10/2006	27/10/2006
10	23/04/2007	29/04/2007	25/04/2007
11	24/09/2007	30/09/2007	27/09/2007

Cette utilisation d'une date unique DRAKKAR pour chaque campagne est justifiée par le fait que nous voulons éviter les erreurs de simulation dues au changement de dates rencontrées dans beaucoup de modèles. En effet le changement de dates est accompagné par une modification de certains paramètres ou de la valeur de ces derniers. Ces paramètres souvent forcés sont considérés comme étant les mêmes pour une même campagne ARAMIS.

Les écarts spatio-temporels sont acceptables et peuvent permettre une inter-comparaison pertinente.

En plus de cette co-localisation qui est utilisée dans l'inter-comparaison ARAMIS/DRAKKAR en II.4, une autre extraction de données DRAKKAR basée sur cette co-localisation a été faite. Cette nouvelle extraction est utilisée principalement pour la mise en place de la carte SOM et pour les modèles statistiques d'inversion développés dans cette thèse au Chapitre IV. Elle est basée sur la route moyenne ARAMIS.

En considérant la route moyenne ARAMIS, nous avons pris :

- ✓ en latitude, le point DRAKKAR équivalent au premier point « moyen » ARAMIS et ceux qui le suivent sur la même longitude dans la base de données DRAKKAR donc en se déplaçant vers le Nord, jusqu'à

atteindre la latitude du point DRAKKAR équivalent au point moyen ARAMIS suivant. Ce point ARAMIS moyen suivant correspond à celui qui est le plus au Sud des points ARAMIS à l'exception du point qui a été sélectionné. La même chose est répétée jusqu'à l'avant dernier point ARAMIS moyen, le dernier correspondant à la limite supérieure.

- ✓ En longitude, pour chaque point DRAKKAR sélectionné par l'algorithme ci-dessus, nous avons pris 20 points à l'Est et 20 à l'Ouest. Pour une résolution de ¼ °, la gamme de longitude varie de 5° pour chaque point et pour chaque côté.
- ✓ En temps, nous avons considéré tous les pas de temps DRAKKAR, soit un ensemble obtenu par la méthode ci-dessus tous les 5 jours sur 8 ans (2000 à 2007).

La Figure II-5 est une illustration de l'espace considéré.

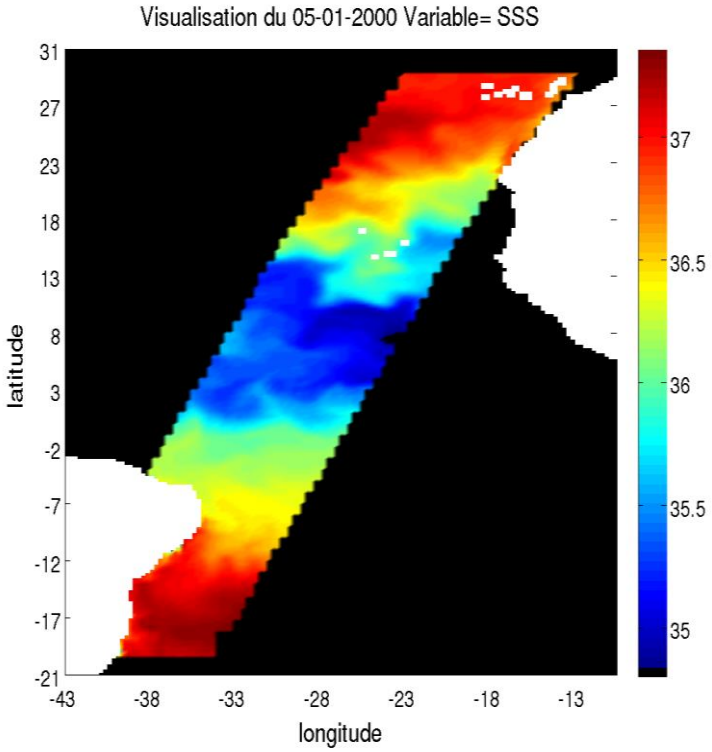


Figure II-5:Visualisation spatiale des données de SSS du 05/01/2000 pour l'étendue géographique sélectionnée. Cette bande entoure la route moyenne ARAMIS de 5° à l'Est et à l'Ouest. La couleur donne la valeur de S.

Cette image présente le découpage des données qui sont utilisées. Nous avons 7433 points pour chaque pas de temps (1 pas de temps=5 jours) sur 8 ans, ce qui fait un total de 7433 x 73 x 8=4340872 points.

Les variables considérées sont les 25 premières immersions de DRAKKAR qui vont de la surface à 1136.9m de profondeur, la SST, la SSH, la latitude et la longitude. La base de données est donc constituée de 4340872 données caractérisées par 29 variables. Ces données ont permis de mettre en place le réseau de neurones, comme explicité dans le Chapitre IV.

Dans une seconde phase de prétraitement, les 2 ensembles, ARAMIS et DRAKKAR colocalisé ARAMIS, ont été ré-échantillonnés. Suivant les opérations à effectuer, 3 méthodes de ré-échantillonnage ont été utilisées. Ces méthodes sont définies ci-après.

II.3.4 Les méthodes de ré-échantillonnage

Pour résoudre le problème lié aux différences de niveaux d'immersion (z), nous allons utiliser un ré-échantillonnage. Il consiste à modifier le nombre d'immersions des individus appartenant à un ensemble *E1* afin d'obtenir le même nombre d'immersion que les éléments d'un autre ensemble *E2*. Les éléments de ces ensembles peuvent être caractérisés par une latitude, une longitude et une date de prise de la mesure de S. Une fois le nombre d'immersions connu de même que les immersions, il faudra définir quelle sera la valeur de S des éléments de *E1* à chacune de ces nouvelles immersions. Suivant le cas à gérer, le nombre d'immersions dans *E1* peut être plus grand ou plus petit que le nombre d'immersions de *E2*. Pour prendre en compte ces différents cas, 3 méthodes ont été définies : une méthode directe, une méthode utilisant la moyenne des S et une utilisant l'interpolation linéaire.

- ***La méthode directe***

Pour utiliser cette méthode, il faut que le nombre d'immersions dans *E1* soit plus grand que le nombre d'immersions de *E2*. Il faut également quelques hypothèses afin de garantir la pertinence de cette méthode. Parmi ces hypothèses les plus importantes sont :

- Supposer que pour chaque immersion dans *E2*, il existe une immersion dans *E1* qui lui est très proche
- Soit h une immersion d'un élément $x \in E1$ qui est caractérisé par sa latitude,

sa longitude et sa date, k l'immersion dans *E2* qui lui est la plus proche et une fonctionnelle $f: x_h \rightarrow f(x_h)$ avec $f(x_h)$ la valeur de S de x pour l'immersion h, il faut supposer que $(x_k) \cong f(x_h)$.

k est appelé « l'immersion équivalente à h »

Partant de ces hypothèses, la méthode directe consiste à trouver un k (dans *E2*) pour chaque immersion h (dans *E1*) et remplacer $f(x_h)$ par $f(x_k)$. Ainsi, si on considère $y \in E2$ un équivalent (par co-localisation spatiale et temporelle) de $x \in E1$, k une immersion de y, h une immersion de x qui est aussi une équivalente à k, $f(x_h)$ sera désignée par S équivalente de $f(y_k)$. Les éléments de *E1* qui avaient *n* immersions ont maintenant *m* immersions, avec n le nombre d'immersions initial dans *E1* et m le nombre d'immersions des éléments de *E2*. Dans la suite cette méthode directe sera désignée par 1ère méthode.

- ***La méthode utilisant la moyenne des S***

La contrainte selon laquelle *« le nombre d'immersions dans E1 doit être plus grand que le nombre d'immersions de E2 »* reste valable. Cependant la manière de procéder est différente. Nous n'avons pas besoin de considérer les deux hypothèses comme cela a été le cas pour la première méthode mais il faut deux ou plusieurs niveaux d'immersion de *E1* entre deux niveaux d'immersion de *E2*. Le principe est le suivant :

Soient j les niveaux d'immersion des $y \in E2$, i les niveaux d'immersion des $x \in E1$. On sait $j \in \{1,2, \dots, m\}$, $i \in \{1,2, \dots, n\}$. Considérons maintenant les fonctions suivantes

$f: y_j \rightarrow f(y_j)$ avec $f(y_j)$ la valeur de S de y pour la $j$ième immersion ,

$h: j \rightarrow h(j)$ avec $h(j)$ l'immersion du niveau j exprimée en unité de longueur (le plus souvent en mètre). La méthode consiste à chercher la S de $x \in E1$ pour une immersion au niveau j en utilisant la moyenne des S de x qui sont dans un voisinage de j.

Ce voisinage est déterminé par les niveaux i qui sont entre $\frac{h(j-1)+h(j)}{2}$ et $\frac{h(j)+h(j+1)}{2}$. $h(j-1), h(j), h(j+1)$ sont des immersions successives d'élément de E2.

Pour trouver S, nous appliquons la formule suivante :

$$f(x_j) = \sum_{k=a}^{b} \frac{f(x_k)}{l}$$

où a désigne le niveau d'immersion de x correspondant à $\frac{h(j-1)+h(j)}{2}$, b le niveau d'immersion de x correspondant à $\frac{h(j)+h(j+1)}{2}$ et l le nombre d'immersions entre a et b.

Cette méthode sera appelée la 2ème méthode dans la suite.

- ***La méthode utilisant l'interpolation linéaire.***

Nous considérons toujours deux ensembles $E1\ et\ E2$. La méthode proposée ici consiste à rééchantillonner $E2$ cette fois-ci, en faisant une interpolation linéaire sur tous les niveaux d'immersion qui existent dans $E1$. On cherche à estimer la S de $y \in E2$ pour une profondeur au niveau i. Pour résoudre ce problème on considère les S de y pour les niveaux j et $j+1$ sont connues.

$$f(h_i) = f(y_j) + (h(i) - h(j))\frac{f(y_{j+1}) - f(y_j)}{h(j+1) - h(j)}$$

Cette méthode sera appelée la 3ème méthode tout au long de ce document.

Nous avons ré-échantillonné les profils de S issus des données DRAKKAR colocalisées de même qu'ARAMIS. Dans notre cas, nous avons des profils ARAMIS discrétisés tous les mètres de profondeur, et des profils DRAKKAR discrétisés sur 46 niveaux d'immersion répartis irrégulièrement (*Tableau II-1*) de la surface à 5900m de profondeur. Dans l'inter-comparaison présentée ici, les 22 premiers niveaux de DRAKKAR seront utilisés afin de couvrir la même profondeur d'étude qu'ARAMIS. Il convient donc de rééchantillonner ARAMIS sur les niveaux DRAKKAR concernés ou de faire l'inverse. Nous avons utilisé la 1ère et la 2ème méthode pour traiter les données ARAMIS et la 3ème méthode pour DRAKKAR.

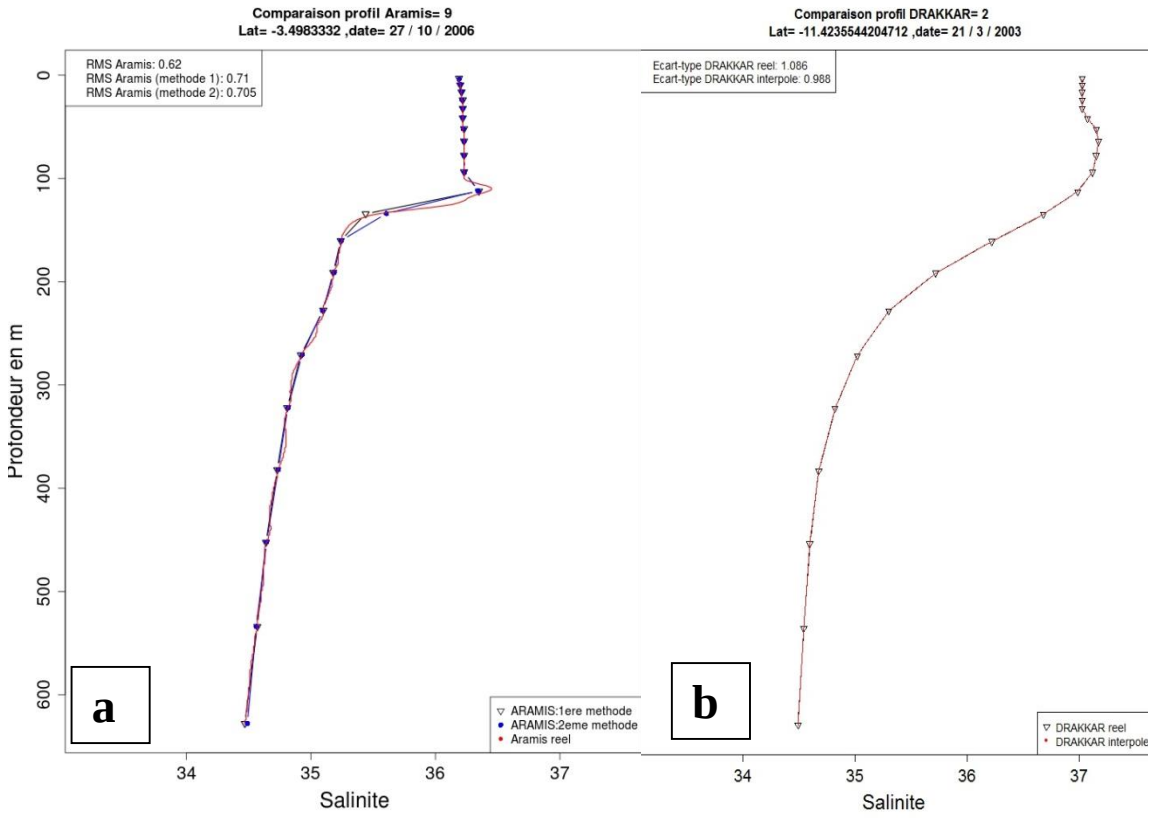


Figure II-6: Exemples de profils de S ré-échantillonnés. En (a) profil in situ ARAMIS réel en ligne rouge, ré-échantillonné par la 1ère méthode en triangle et par la 2ème méthode en cercle bleu. En (b) profil simulé par DRAKKAR en triangle noir et ré-échantillonné par la 3ème méthode rouge.

On voit, que le profil ARAMIS ré-échantillonné par la 2ème méthode n'a pas toujours une intersection avec le profil réel au niveau des profondeurs considérées, par exemple entre 100 et 150m. Ce profil représente donc a priori moins bien le profil réel que le profil ré-échantillonné par la 1ère méthode pour les immersions considérées. Mais les exemples de ce type restent marginaux et la majorité des profils sont bien représentés par les 2 méthodes.

Pour ré-échantillonner DRAKKAR, nous avons utilisé la 3ème méthode (la méthode d'interpolation, Figure II-6b) car il s'agit d'augmenter le nombre d'immersions. Tous les profils DRAKKAR sont bien ré-échantillonnés.

Dans les études comparatives ARAMIS/DRAKKAR, nous avons utilisé les données obtenues avec les 3 méthodes de ré-échantillonnage, 1 et 2 pour ré-échantillonner ARAMIS et 3 pour DRAKKAR.

II.4 ÉTUDE COMPARATIVE DRAKKAR VS ARAMIS A L'AIDE DE STATISTIQUES CLASSIQUES.

Les modèles sont souvent utilisés pour combler le manque de données in situ, d'observations. Il est donc important de déterminer le niveau de confiance d'un modèle. Au cours de ces dernières décennies, le développement de modèles numériques simulant des processus géophysiques est devenu une préoccupation majeure de la communauté des sciences physiques (Willmott, et al., 1985). Un aspect important du processus d'élaboration du modèle, l'évaluation de la performance du modèle, a reçu relativement peu d'attention dans la littérature géophysique. Le problème est aggravé par l'absence d'accord entre les scientifiques au sujet des mesures et des procédures pour déterminer la précision des modèles les plus appropriés. Mais depuis quelques années, les efforts sont en train d'être faits, notamment avec le projet CMIP (Coupled Model Intercomparison Project) initié en 1995 par le World Climate Research Programme (WCRP). La validation des résultats du modèle par comparaison avec les données observées est essentielle. Il s'agit de la mesure par laquelle nous pouvons évaluer la qualité d'un modèle et elle informe des utilisations appropriées des données.

Le travail décrit ici n'est toutefois pas une validation du modèle DRAKKAR, il a pour but d'établir un niveau de confiance des données sorties du modèle DRAKKAR le long de la route ARAMIS et sur les dates pour lesquelles nous disposons de données observées. Il permet aussi de définir les précautions à considérer dans l'utilisation du modèle.

L'inter-comparaison se décompose en 2 parties, la 1ère présente les résultats obtenus en utilisant des statistiques classiques et l'autre partie est basée sur l'utilisation des réseaux de neurones définis en III.2.

Pour comparer ces 2 ensembles DRAKKAR et ARAMIS, il faut que les données soient sur une même base spatiale (x,y,z) et temporelle (t). La co-localisation et le ré-échantillonnage DRAKKAR/ARAMIS ont permis de régler ce problème lié à la mise en œuvre d'une même base.

Les résultats issus de ces méthodes ont montré que les données ré-échantillonnées sont en accord avec les données originales. En se basant sur les nouvelles données, ce paragraphe présente des comparaisons directes entre DRAKKAR et ARAMIS sur la base

de statistiques classiques.

Nous allons nous intéresser à la variabilité de S autour de sa moyenne verticale à l'aide de l'écart-type. L'écart-type mesure la dispersion d'une série de valeurs autour de leur moyenne. Il est donné dans le cas présent par :

$$\sigma_{c,y} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{c,y,i} - \overline{x_{c,y}})^2}$$

Avec *c*=campagne concernée, *y*=latitude, *n*=nombre de couches et $x_{c,y,i}$ la S en une immersion i.

Il permet de voir si le modèle reproduit la même dispersion verticale, en termes de S, que les observations (ARAMIS) en un point donné de latitude (un individu) pour une campagne donnée.

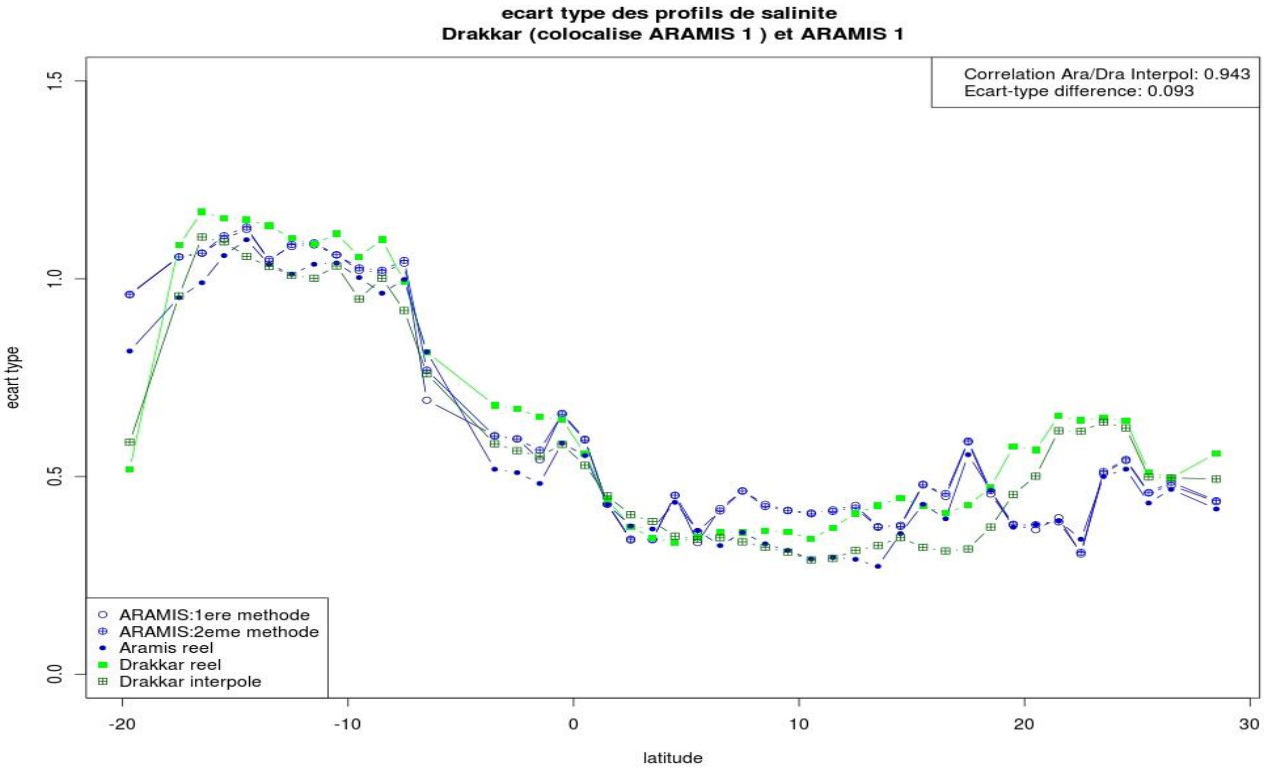


Figure II-7: Evolution latitudinale de l'écart-type autour de la moyenne verticale des données ARAMIS1 (en bleu) et DRAKKAR 1 (en vert)

La Figure II-7 présente la dispersion de S autour de sa moyenne verticale pour chaque point ARAMIS1 (juillet 2002, en cercle bleu) et son équivalent DRAKKAR (réel en vert

clair et interpolé en vert foncé) en fonction de la latitude. Les courbes ont la même allure générale avec 2 zones de variabilité extrême au nord et surtout au sud de la ligne. Ces régions correspondent aux zones de SMW. Entre les 2, la variabilité est plus faible sauf à l'équateur où un maximum relatif est sans doute à attribuer à la présence du Sous Courant Equatorial (EUC pour Equatorial UnderCurrent) vers 100m et aux eaux salées de l'Atlantique Sud qu'il transporte. Un autre extremum vers 5°N, sous la ZCIT est lié aux eaux dessalées de surface associées. L'accord entre modèles et données est meilleur au sud de 5°N approximativement qu'au nord. Nous observons en effet quelques différences non négligeables autour de 20-28°N: la variabilité présentée par DRAKKAR y est toujours supérieure à celle d'ARAMIS. Cette zone se situe au bord de la région des SMW du nord, donc une zone de haute variabilité spatiale et temporelle (Arnault, 2004) ce que reflète sans doute la comparaison puisque la tendance est inverse (écart type des observations supérieur à celui du modèle) entre 15 et 20°N.

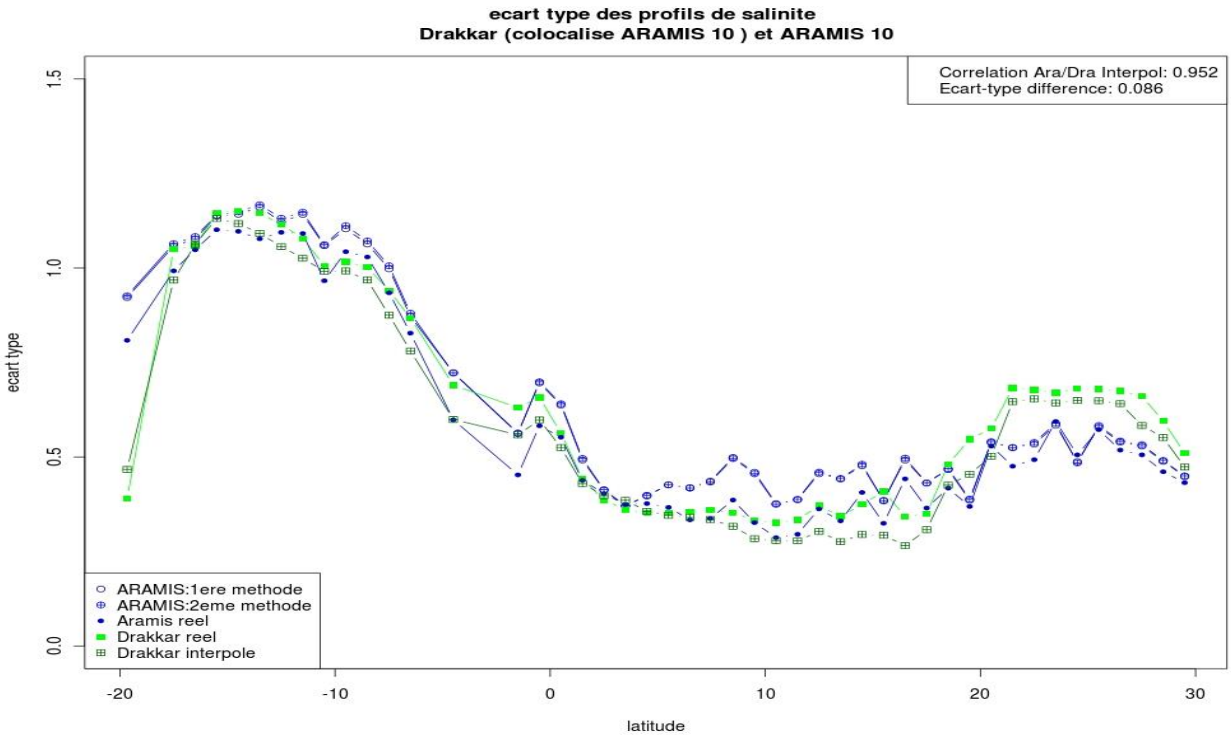
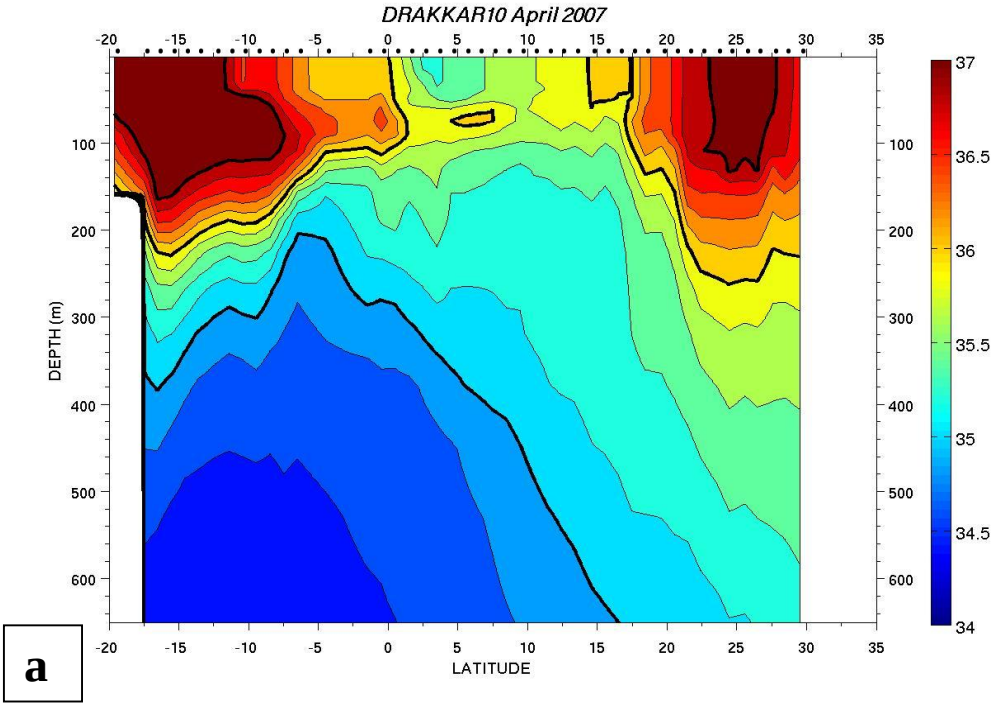


Figure II-8: Evolution latitudinale de l'ecart-type autour de la moyenne verticale des données ARAMIS10 et DRAKKAR10

La Figure II-8 confirme pour ARAMIS10 (avril 2007) la situation présentée par la Figure

II-7 précédemment décrite.

Cette comparaison montre que le modèle peut combler notre manque de données mais cependant il faut l'utiliser avec attention surtout sur les hautes latitudes comme le montre la Figure II-9.



a

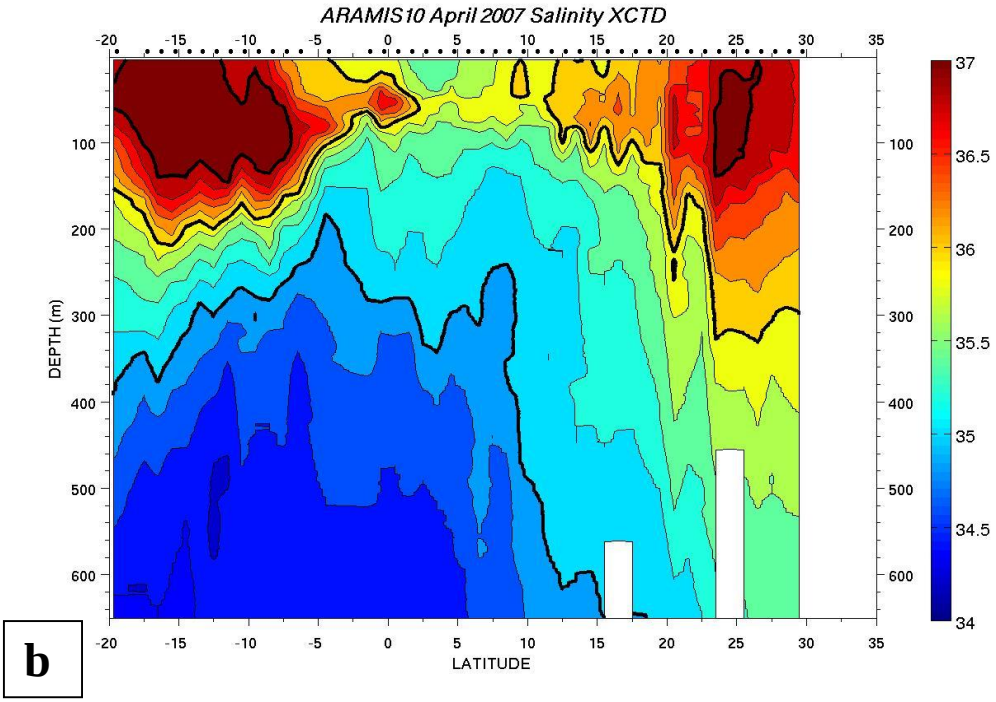


Figure II-9: exemple de section verticale (b) ARAMIS 10 en avril 2007 et son équivalent DRAKKAR (a)

La Figure II-9 présente la section des données (profils) obtenues en une campagne en fonction de la latitude.

On note quelques petites différences entre le modèle et les observations que la Figure II-10, qui présente la section des différences de S entre ARAMIS10 et DRAKKAR10, permet de mieux voir.

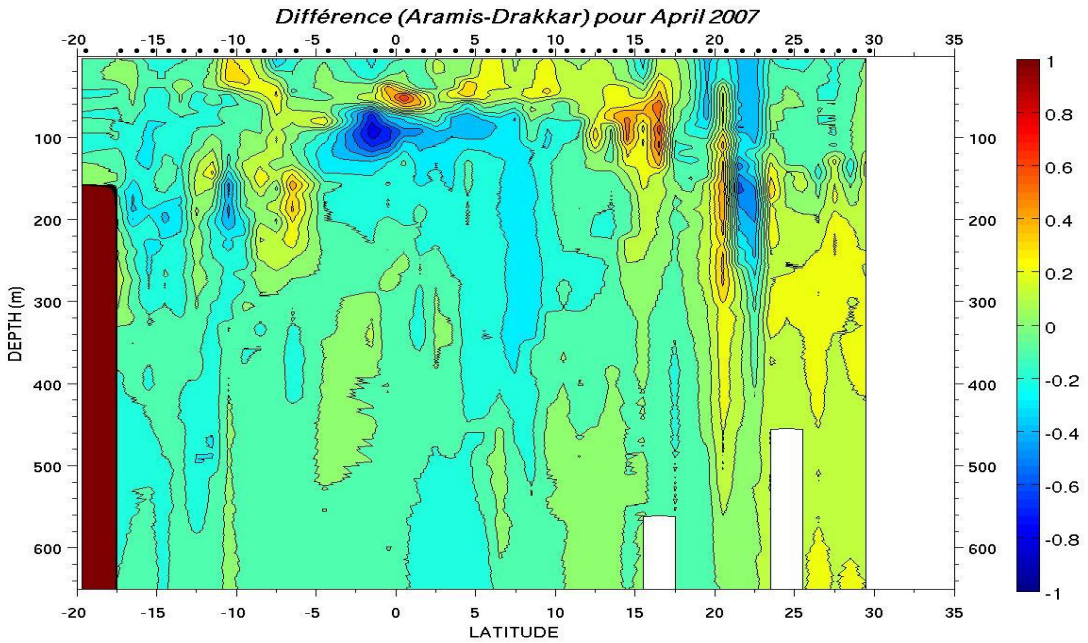


Figure II-10: Section Différence (ARAMIS -DRAKKAR)

Cette différence tourne globalement autour de 0,15 psu pratiquement partout. Les plus grandes valeurs (~0.4 psu) sont localisées vers 22° N, de la surface à 200m de profondeur, et au niveau de l'équateur autour de 100m de profondeur. Nous pouvons remarquer que ces différences extrêmes apparaissent sous la forme de doubles noyaux de signes opposés. Cela traduit donc un mauvais positionnement de l'halocline (zone équatoriale) ou du front halin (zone 24° N). Ces 2 zones correspondent respectivement aux régions de l'EUC et des SMW, zones hautement variables et donc où la moindre différence spatio-temporelle dans la localisation des données se traduira par une importante différence de S.

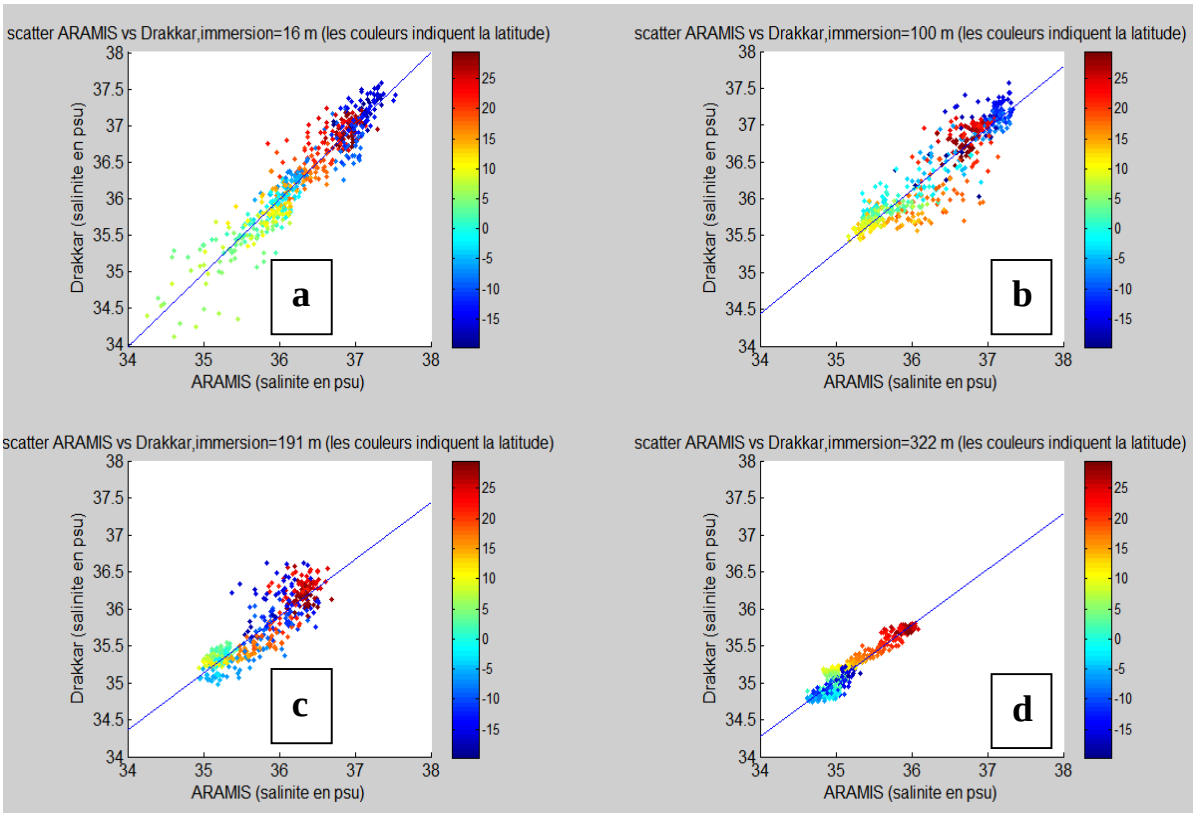


Figure II-11: Dispersion DRAKKAR en fonction d'ARAMIS pour les immersions indiquées dans le titre de la figure, toutes campagnes confondues. Le code couleur indique la latitude. La droite de régression est en bleu

La Figure II-11 donne la dispersion de DRAKKAR en fonction d'ARAMIS (toutes les campagnes étant confondues). La figure montre 4 situations : en surface (Figure II-11a, 16m) nous avons les maxima de S au sud vers 15° S et au nord de la route (autour de 20°N) et les minima autour de l'équateur, minima certainement dus aux fortes pluies qui dessalent la surface océanique. Le modèle comme ARAMIS fait ressortir cette situation comme le confirme la droite de régression qui correspond quasiment à la 1ère bissectrice. Cette situation est inversée dans les eaux plus profondes (Figure II-11d, 322m) où on note des eaux plus salées au nord. Considérant cette même immersion, 322m, S croit avec la latitude dans l'hémisphère nord. Cette immersion correspond aussi à la plus faible variabilité. Autour de 100m (Figure II-11b) nous avons une structure moins organisée. Cette partie correspond souvent à l'halocline, zone de gradient vertical de S intense. On

peut voir également que les plus faibles valeurs de S sont autour de 10°N et les plus fortes valeurs au Nord et au Sud, traces des SMW que l'on retrouve à 195m de profondeur (Figure II-11c). Les minima sont localisés autour de 5°S et 5°N. Entre ces 2 immersions, ARAMIS et DRAKKAR montrent de moins bons accords.

Cette étude comparative ARAMIS/DRAKKAR est complétée par une autre approche basée cette fois sur les réseaux de neurones qui sont définis en III.2. Cette approche nécessitant la définition et le rappel de quelques notions statistiques, elle est décrite dans le chapitre III.

Le dernier ensemble de données avec lequel nous avons travaillé est celui constitué des données collectées dans la base Coriolis.

II.5 DONNEES ISSUES DU PROJET CORIOLIS

Ces données sont recueillies et mises à disposition gratuitement par le projet Coriolis et les programmes qui y contribuent (http://www.coriolis.eu.org). Coriolis a démarré en tant que projet pilote en 2001 avec pour mission d'acquérir, à partir de profileurs flotteurs Argo et de navires de recherche et marchands, des données in-situ nécessaires à l'océanographie opérationnelle physique et la recherche sur le climat, de les traiter, de les qualifier et de les distribuer à Mercator et à la communauté de recherche française en temps réel (<48heures de l'acquisition).

Coriolis contribue aux réseaux Argo, SO-SSS, SO-Pirata, bouées dérivantes et accompagne l'acquisition de mesures à partir des navires océanographiques français opérés par l'Ifremer, le SHOM, l'INSU, l'IPEV et l'IRD. Elle gère la cohérence du service vers l'océanographie opérationnelle et la recherche (base de données commune, procédures communes de contrôle qualité, suivi du service, analyse des nouveaux besoins), la mise en commun de moyens et la gouvernance d'ensemble. Depuis 2009, dans le cadre du programme national français Previmer, et du programme européen MyOcean, Coriolis a étendu son périmètre à des données plus régionales (web, http://archimer.ifremer.fr/doc/00063/17406/14912.pdf, consulté en août 2012).

Le centre de données Coriolis distribue plusieurs types de données d'origines différentes. Dans le cadre de notre thèse, nous avons utilisé des données ayant 4 origines selon la nomenclature Coriolis: les « flotteurs Argo », les « CTDs », les « XBTs » et « Autres ». Ces données sont assez différentes par rapport à leur structure. Les immersions sur

lesquelles nous disposons de mesures varient d'un profil à l'autre.
Les données extraites sont localisées en Atlantique Tropical. Elles s'étendent de 20°E à 40°W et de 40°N à 40°S et sur la période allant de 01/2000 à 04/2012.
Comme rappelé auparavant, les données Coriolis ont diverses origines. Dans notre étude, nous avons utilisé les flotteurs Argos, les CTDs, les XBTs et les autres profils disponibles..
Les données Coriolis sont fournies dans des fichiers NetCDF[1] avec plusieurs paramètres (Web Coriolis, 2012). Parmi ces derniers ceux qui nous intéressent sont :

- REFERENCE_DATE_TIME : contient la date de référence ;
- JULD : contient la date julienne de la mesure (observation) en nombre de jours c'est-à-dire le nombre de jours compris entre la date de mesure et la date de référence ;
- LATITUDE ;
- LONGITUDE ;
- N_PROF : cette variable donne le nombre de profils contenu dans le fichier. Chaque profil est considérée une observation ;
- N_PARAM : donne le nombre de paramètres pour le profil considéré. S'il est égal à 3 alors la mesure contient un profil de S ;
- N_LEVELS : contient de le nombre d'immersion ;
- PRES : contient le profil de pression ;
- TEMP : contient le profil de T ;
- PSAL : contient le profil de S ;

En plus de ces paramètres, les données sont flaguées QC (Quality Control) en fonction de la qualité des variables mesurées. Ce contrôle concerne aussi bien les données de position que les données mesurées, ci-après la liste des flags qui sont utilisés dans l'extraction.

- JULD_QC : contient le flag de la date julienne ;
- POSITION_QC : contient le flag de la position du profil ;

[1] NetCDF was developped at the Unidata Program Center in Boulder, Colorado. For more informations on NetCDF : http ://www.unidata.ucar.edu/software/netcdf

- PROFILE_PARAM_QC : contient le flag du profil. <PARAM> peut prendre les valeurs PRES ou DEPTH, TEMP ou PSAL.
- PARAM_QC : un vecteur de la même taille que le profil et contenant le flag pour chaque niveau ;

Nous avons également des variables <PARAM>_ADJUSTED qui contiennent des profils corrigés. Ces variables sont remplies si le profil a été corrigé sinon elles ne contiennent aucune valeur.

L'extraction concerne principalement les paramètres cités <PARAM> en tenant compte des QC. Les QC sont de deux types. Il y a des QC qui qualifient la mesure de la donnée et peuvent avoir des valeurs variant de 0 à 9 et des QC qui donnent la qualité du profil du paramètre considéré. Ces derniers qui varient de "A" à "F" donnent le pourcentage de bonnes mesures que compte le profil. Le tableau suivant donne les valeurs possibles des QC et leur signification :

Tableau II-3:QC et leur signification. N représente le pourcentage de niveaux ayant de bonnes données.

Flag QC	Signification
0	Aucun contrôle
1	Bonne
2	Probablement bonne
3	Probablement mauvaise
4	mauvaise
5	Corrigée
8	interpolée
9	absente

(a) Flag sur les données

Flag QC	Signification
" "	Aucun contrôle
A	$\mathbf{N} = 100\%$
B	$75\% <= \mathbf{N} < 100\%$
C	$50\% <= \mathbf{N} < 75\%$
D	$25\% <= \mathbf{N} < 50\%$
E	$0\% < \mathbf{N} < 25\%$
F	$\mathbf{N} = 0\%$

(b) Flag sur profil

Seuls les fichiers contenant des profils de S donc ayant la valeur de N_PARAM à 3 ont été considérés. L'algorithme d'extraction est composé principalement de 2 parties.

1) **extraction des données** : Cette 1ère partie sert à extraire le maximum de « bonnes » données. Nous testons d'abord si <PARAM>_ADJUSTED est renseigné pour le prendre à la place de <PARAM>. Puis il faut voir si le profil de S existe. Si ce n'est pas le cas cet ensemble de profils est ignoré.

   Les flags QC sont pris en compte de la manière suivante :

- ✓ *si la position (POSITION_QC) ou la date (JULD_QC) est mauvaise,*
  le profil est exclus
- ✓ *si le profil contient moins de 50% de bonnes données*
  alors il est exclu aussi
- ✓ *sinon s'il contient moins de 100% de bonnes données*
  alors les mauvaises données sont remplacées par des NaN
- ✓ *sinon il est pris entièrement.*

   Le profil considéré peut être un profil de S, de T ou de pression ou immersion (ce profil de pression ou immersion donne les profondeurs auxquelles la S et la T sont mesurées). Les latitudes et les longitudes des profils (mesures) sont obtenues directement. La SST est extraite du paramètre TEMP ou TEMP_ADJUSTED, elle correspond à la première valeur de ce paramètre.
   A la fin de cette étape, les matrices préliminaires des profils de pression, de T et de S sont construites, ces matrices serviront à retrouver les S aux immersions considérées c'est-à-dire celles qui sont les plus proches des 25 premières immersions de DRAKKAR, définies en II.3.1.

2) **Recherche des bonnes immersions** : Dans la 2ème partie, l'algorithme recherche la bonne immersion dans la matrice des pressions/profondeurs construites précédemment. Cette immersion est la plus proche de celle de DRAKKAR considérée. La S à l'indice trouvé est considérée comme la S du profil pour cette immersion. La recherche s'arrête si une pression non renseignée est rencontrée dans la matrice des pressions.

Rappelons que les données Coriolis (S(z)) n'ont pas la même résolution. Sur certains profils de S, la résolution est de 5m, sur d'autres de 10m ou tout simplement irrégulière.

Le ré-échantillonnage des données Coriolis sur les immersions DRAKKAR peut être plus ou moins précis suivant la résolution du profil. Ce n'est pas le cas pour les données ARAMIS dont la résolution au mètre est assez haute pour une très bonne précision du ré-échantillonnage.

Après cette phase, les profils ne contenant que des NaNs ont été supprimés. 75936 profils de S ont été extraits, de même que pour les T et les pressions. La Figure II-12 illustre la répartition de ces profils dans la zone d'étude et dans le temps.

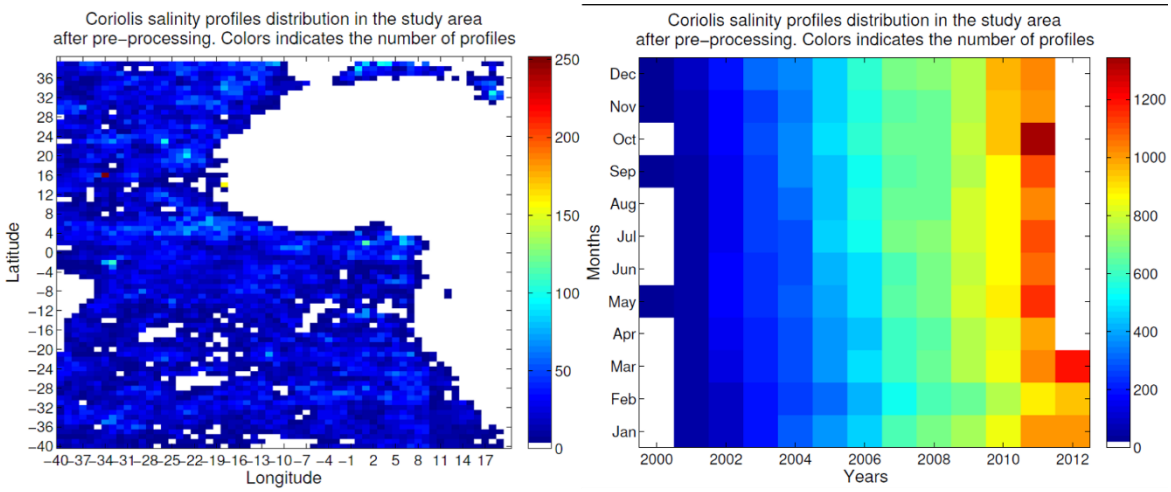


Figure II-12:Distribution du nombre de profils. A gauche(a): suivant l'espace, à droite(b) : suivant le temps. La couleur donne le nombre de profil par point (1x1°) et par mois/année respectivement pour la distribution spatiale (a) et temporelle (b).

Elle montre une répartition des profils assez homogène dans l'espace hormis quelques points. Le nombre de profils est inférieur à 50 dans presque toute la zone d'étude.

La répartition temporelle est dépendante de l'année considérée. On note une croissance du nombre de profils dans le temps :

- ✓ de 2000 à 2005 : le nombre de profils de S par mois est compris entre 0 et 500 ;
- ✓ de 2006 à mai 2008 : le nombre de profils de S par mois est compris entre 600 et 800 ;
- ✓ de juin 2006 à 2012 : le nombre de profils dépasse 800 par mois ;

Ceci illustre les avancées technologiques qui rendent la mesure de S des couches de surface et de proche surface plus facile à acquérir grâce notamment aux flotteurs Argo.

Malgré toutes les précautions prises lors de l'extraction et de la validation, il arrive que celles-ci se révèlent inexploitables. Les 74778 profils obtenus après ce premier prétraitement, ont subi une dernière étape d'épuration basée sur une classification par carte de Kohonen. Cette étape a permis de repérer les profils de S ayant une allure invraisemblable. Ces profils 'invraisemblables' se regroupent dans des classes caractéristiques qui permettent de les repérer facilement à l'œil. De ce fait, l'ensemble des profils appartenant à ces classes caractéristiques est supprimé de la base de données. Cette étape d'épuration a permis d'éliminer 179 profils, ainsi la base de données finale est constituée de 74599 individus.

II.6 LA HAUTEUR D'EAU ALTIMETRIQUE

La hauteur de l'eau, ou niveau de mer, est un produit typique utilisé en océanographie depuis les années 90 qui se mesure à partir de satellites. Grâce à la connaissance de longues séries temporelles de hauteur de l'eau, il a été possible de suivre l'évolution globale du niveau moyen des océans durant les deux dernières décennies, et non seulement au niveau des marégraphes comme auparavant, et ainsi de mettre en avant la tendance actuelle à la montée des eaux. Les produits altimétriques que nous avons utilisés sont issus du traitement des données altimétriques par la Division d'Océanographie Spatiale de CLS (Collecte Localisation Satellites) dans le cadre du projet DUACS (Developing of Altimetry for Climate Change), en tant que partie intégrante du projet européen pour l'environnement et le climat (EU ENACT). Ils sont distribués en ligne par le centre opérationnel AVISO (Archiving, Validation, and Interpretation of Satellite Oceanography Data, http://www.aviso.oceanobs.com/, consulté en août 2012).

Dans la perspective de l'inversion sur une large fenêtre temporelle, il est impératif d'utiliser un produit homogène sur la durée. Nous avons donc utilisé les produits en temps différé de type "Ref" pour référence et grillés pour lesquels les corrections sont basées sur l'orbite de 2 satellites (traces Jason-2/Envisat ou Jason-1 / Envisat ou TOPEX/Poséïdon / ERS). La présentation des données sous forme de grilles permet la co-localisation précise des données. Ces produits sont disponibles tous les 7 jours depuis octobre 1992 et les données sont réparties sur une grille Mercator d'1/3°.

Il existe 2 types de produits pour la hauteur de l'eau : la Sea Level Anomaly (SLA) qui est une anomalie et l'Absolute Dynamic Topography (ADT) qui se réfère au géoïde. Il est

généralement admis que l'ADT est surtout utilisée pour l'étude de la circulation grande échelle et la circulation océanique générale (gyre). A l'inverse la SLA est plus recommandée pour l'étude de la variation océanique, comme la circulation mésoéchelle, les variations saisonnières, El Niño ... Le produit le plus naturel serait donc d'utiliser la SLA. Cependant, en utilisant la SLA, l'information de la "pente" typique autour de l'équateur est perdue alors qu'elle permet de différencier par exemple la zone au nord de la zone au sud de l'équateur. En effet, le principe est de classer les profils en catégorie, chaque classe ayant une étiquette contenant les informations de surface comme l'ADT. La SLA ne permet donc pas de différencier ces classes car elle ne contient pas l'information à plus grande échelle qu'est la MDT (Mean Dynamic Topography). Nous avons donc préféré l'ADT en nous référant sur les travaux de thèse de Yves Tanguy (Tanguy, 2011) qui a fait des tests sur la SLA et l'ADT et a conclu que l'ADT donnait de meilleurs résultats pour des besoins similaires aux nôtres.

Ces données d'ADT d'origine satellitaire n'ont pas nécessité un traitement particulier. Une co-localisation spatiale et temporelle a permis de sélectionner l'ADT la plus proche en temps et en espace du point géographique à considérer. Ce point peut correspondre à une donnée issue de la base ARAMIS, DRAKKAR ou Coriolis selon le cas. La résolution spatiale de cette variable de 1/3° et 7 jours en temps donne des écarts inférieurs à 1/6° en espace et 3.5 jours en temps.

Chapitre III. METHODES STATISTIQUES DE BASE.

Dans cette partie nous décrivons les méthodes utilisées tout au long de cette thèse. Notre travail se basera principalement sur les cartes de Kohonen explicitées en III.2. Ces cartes permettent de classer d'une manière itérative l'ensemble des situations rencontrées dans une base de données. Nous avons également utilisé des méthodes de sélection et d'extraction de paramètres. Ces dernières permettent d'étudier la pertinence des variables sur l'expression de la variabilité du phénomène étudié et l'interrelation entre ces variables descriptives. Cependant, l'étude de ces méthodes n'est pas l'objet de notre travail. Elles sont utilisées comme outils d'analyse et de traitement de données pour mettre en place des modèles statistiques pertinents permettant la reconstruction de profils de S à partir de paramètres de surface sur la zone tropicale de l'Atlantique. Ces méthodes sont donc exposées brièvement avec parfois des illustrations tirées de notre base de données. Ce qui permet de mieux expliquer la méthode considérée et comprendre les données de notre étude. En revanche, les cartes topologiques dont l'utilisation est importante dans cette thèse ont été détaillées. Ce chapitre est structuré comme suit: une introduction générale sur les algorithmes de classification est d'abord faite, ensuite les cartes topologiques de Kohonen sont expliquées et enfin les méthodes d'extraction et de sélection de paramètres pertinents sont présentées.

III.1 METHODES D'ANALYSE DE DONNEES ET ETUDE DES DONNEES DRAKKAR

Ayant démontré dans les paragraphes précédents que les données DRAKKAR et ARAMIS donnent des images voisines de la variabilité de la zone Atlantique tropical qu'elles échantillonnent, nous présentons une description des données DRAKKAR. La méthode d'analyse utilisée pour l'exploration des données DRAKKAR est principalement basée sur les corrélations et l'Analyse en Composantes Principales (ACP). Rappelons que les données utilisées sont celles de DRAKKAR spatialement colocalisé sur la route moyenne ARAMIS à tous les pas de temps DRAKKAR et avec les immersions originales de DRAKKAR.

III.1.1 L'Analyse en Composantes Principales.

Dans la plupart des applications, un individu est caractérisé par un nombre de variables souvent élevé. Pour analyser des données ayant cet aspect multidimensionnel, ACP est une méthode factorielle particulièrement puissante utilisée pour un but:

- exploratoire. Elle propose des techniques qui aident à visualiser du mieux possible sur un plan des observations décrites par plus deux variables
- réducteur du nombre de variables. Elle permet de définir des variables artificielles appelées composantes principales, qui expliquent dans l'ordre de leur définition les plus grands pourcentages de la variance du nuage des observations

L'ACP est utilisée pour réduire p variables corrélées en un nombre q de variables non corrélées de telle manière que les q variables soient des combinaisons linéaires des p variables initiales, que leur variance soit maximale et que les nouvelles variables soient orthogonales entre elles suivant une distance particulière (Duby, et al., 2006).
L'ACP a pour but de trouver un espace de dimension inférieure à la dimension de l'espace d'origine. On souhaite également que les observations puissent être visualisables dans cet espace de nouvelle dimension. Ceci implique souvent une dimension idéale de 3 ou 2 car au-delà la visualisation sur une seule figure reste impossible. Cependant cette nouvelle dimension doit respecter une contrainte qui est de minimiser la perte d'information sur les observations initiales. Afin de déterminer le nombre de CP nécessaires, un certain nombre d'approches sont utilisées; Jolliffe (2002) a fait un résumé complet et toujours actuel sur l'état de ces approches. Ces pratiques peuvent être globalement classées en 4 catégories, bien qu'à l'intérieur de chaque catégorie, il puisse exister encore diverses variations mineures (Zhu et al.., 2006). Nous avons retenu la suivante :
Pourcentage de variance: il consiste à trouver le plus petit nombre de composantes qui capturent un certain pourcentage de la variance totale. Ceci équivaut à conserver les q CP où q désigne le plus petit nombre entier compris entre 1 et le nombre total de composantes (p) de telle sorte que

$$\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_q}{\lambda_1 + \lambda_2 + .. + \lambda_q + \cdots + \lambda_p} \geq \gamma$$

où λ_i est la part de la variance totale expliquée par la composante i et γ est une valeur

déterminée. Cette valeur est très souvent de l'ordre de 70% à 90%. Une valeur supérieure à 90% est appropriée quand 1 ou 2 CP représentent les sources dominantes et évidentes de variation (Jolliffe, 2002). Cette démarche montre que si le cumul de pourcentage de variance expliquée par des CP est > 70% alors on peut considérer ces CP comme pouvant résumer les variables initiales dans un nouveau repère ayant comme axes ces CP.

Après une ACP, il faut traduire le résultat sur les données d'origine (avant transformation), c'est-à-dire revenir à la base d'origine dans laquelle les axes correspondent aux variables initiale. La méthode la plus naturelle pour donner une signification à une composante principale est de la relier aux variables initiales et de calculer les coefficients de corrélation linéaire et en s'intéressant aux plus forts coefficients en valeur absolue (Saporta, 2011). Mais cette méthode doit être renforcée par les coefficients de contribution.

Nous avons considéré une quantité appelée coefficient de contribution donnée par l'équation suivante:

$$CTR(j,k) = \frac{cor^2\,(\,c_{k,}x_j)}{\sum_{i=1}^{n} cor^2\,(\,c_{k,}x_i)}$$

où *CTR(j,k)* désigne la contribution de la variable x_j sur la construction de la composante c_k et n le nombre de variables.

Dans les paragraphes suivants, nous allons maintenant décrire les méthodes de classification. Les opérations de classification sont utilisées pour identifier et classifier des données. Il existe une variété d'approches prises pour faire une classification numérique. Les 2 plus connues, souvent utilisées, sont la classification supervisée et la classification non supervisée.

La classification supervisée demande à un utilisateur d'instruire le système en désignant des données pré-existantes comme étant des échantillons représentatifs des classes à extraire. La classification non supervisée, quant à elle, ne demande aucune connaissance a priori de l'utilisateur. Soit il existe une collection donnée de classes définies indépendamment, soit le nombre de classes et leurs caractéristiques sont définis automatiquement lors de la classification.

Nous présentons ci-après les algorithmes de classification regroupés en 2 catégories à savoir les méthodes classiques et les méthodes avancées plus récentes. Ce regroupement

est justifié par le fait que les méthodes avancées, notamment les cartes auto-organisatrices que nous avons utilisées, sont basées sur les algorithmes classiques, d'une part et d'autre part parce qu'ils facilitent la compréhension de la notion de classification.

III.1.2 Les méthodes de classification classiques

III.1.2.1 Les K-means ou K-moyennes

L'algorithme des *K*-moyennes est une méthode simple de classification automatique qui sépare les données en *K* classes. On définit *K* centres de gravité appelés centroïdes. Puis, on assigne chaque élément à la classe dont le centroïde est le plus proche (par rapport à une distance). Après chaque affectation, on calcule le centre de gravité des éléments de chaque classe. Les deux dernières étapes sont répétées de façon itérative jusqu'à la convergence des centroïdes.

L'algorithme classique des *K*-moyennes propose de classifier les données en *K* ensembles en minimisant l'énergie suivante :

$$U = \sum_{i=1}^{k} \sum_{\overrightarrow{x_j} \in C_i} \left\| \overrightarrow{x_j} - \overrightarrow{\mu_i} \right\|^2$$

où $\overrightarrow{\mu_i}$ désigne la moyenne des éléments de la classe C_i. Dans ce cas, le nombre de classes est fixé et représente le seul paramètre de l'algorithme, les moyennes $\overrightarrow{\mu_i}$ étant estimées récursivement.

Cet algorithme présente néanmoins quelques inconvénients : il faut fixer au préalable le nombre de classes *K*, il converge vers un minimum local et n'est pas adapté au cas de classes de structures non convexes et de tailles différentes.

Palubinskas (Palubinskas, 1998) modifie l'algorithme des K-moyennes en lui ajoutant un terme entropique afin d'estimer automatiquement le nombre de classes. On définit l'entropie par la quantité $-\sum_{i=1}^{k} pi.\log(pi)$

où $pi = \frac{Card(C_i)}{N}$ représente la probabilité qu'une donnée de l'ensemble de taille N appartienne à la classe C_i.

En ajoutant ce terme d'entropie à l'énergie précédente on obtient :

$$U = \sum_{i=1}^{k} \sum_{\overrightarrow{x_j} \in C_i} \left\| \overrightarrow{x_j} - \overrightarrow{\mu_i} \right\|^2 - \alpha_E \sum_{i=1}^{k} pi.\log(pi)$$

Le paramètre α_E joue le rôle de compromis entre les deux termes. Après reformulation, on obtient :

$$U = \sum_{i=1}^{k} (\sum_{\overrightarrow{x_j} \in C_i} \left\| \overrightarrow{x_j} - \overrightarrow{\mu_i} \right\|^2 - \alpha.\log(pi))$$

Avec $\alpha_E = \alpha.N$

L'algorithme est initialisé avec un nombre élevé de classes. A chaque itération, on assigne chaque donnée x à la classe i la plus proche au sens de la distance définie ici $d(\vec{x}, \overrightarrow{\mu_i}) = (\|\vec{x} - \overrightarrow{\mu_i}\|^2 - \alpha.\log(pi))$

Il est à noter que le paramètre α décroît de manière exponentielle à chaque itération. Ainsi le terme entropique détermine dès les premières itérations le nombre optimal de classes, mais n'intervient presque plus lorsque l'algorithme a convergé.

L'algorithme des K-moyennes entropique se comporte alors comme un simple algorithme des K-moyennes.

L'algorithme des K-moyennes est donc le suivant :

– Initialisation : choix d'un grand nombre de classe (K = 50 par exemple) et positionnement aléatoire des K centroïdes sur des points de l'ensemble,

– Itérations : tant que les centroïdes changent de position :

• on assigne chaque point à la classe du centroïde le plus proche au sens de la distance définie par : $d(\vec{x}, \overrightarrow{\mu_i}) = (\|\vec{x} - \overrightarrow{\mu_i}\|^2 - \alpha.\log(pi))$

• si une classe ne comporte plus d'élément, elle est éliminée,

• on recalcule le centre de gravité $\overrightarrow{\mu_i}$ des éléments de chaque classe

La partition obtenue par l'algorithme des k-moyennes dépend des représentants initialement choisis. De façon à s'affranchir en partie de cette dépendance, on exécute l'algorithme des k-moyennes (k et d étant fixés) avec des initialisations différentes, et on retient la meilleure partition. La qualité d'une partition est mesurée par la quantité :

$$D = \sum_{i=1}^{k} \sum_{\overrightarrow{x_j} \in C_i} d(\vec{x}, \overrightarrow{\mu_i})$$

qui mesure la cohésion des classes obtenues.

III.1.2.2 K-Plus Proches Voisins ou Nearest- Neightbour

L'algorithme des K Plus Proches Voisins est une méthode de classification supervisée. Il affecte un élément à la classe la plus représentée parmi les K plus proches éléments de la base d'apprentissage.

Le paramètre K permet de réduire les effets du bruit sur la classification. Cependant, si K est trop grand, l'algorithme aura tendance à assigner un nouvel objet à la classe de la base d'apprentissage la plus représentée. La vitesse de cet algorithme est proportionnelle au nombre d'éléments de la base d'apprentissage.

L'algorithme des K-Plus Proches Voisins est donc le suivant :

Pour chaque point $\vec{x}$ de l'ensemble :

– On calcule les distances entre le point et tous les vecteurs $\overrightarrow{x_i}$ de la base d'apprentissage : $d(\vec{x}, \overrightarrow{x_i})$

– On cherche les K vecteurs de la base d'apprentissage les plus proches du point, c'est-à-dire, ceux qui ont les K plus petites distances

– On assigne le point à la classe la plus représentée parmi ces K vecteurs.

III.1.2.3 Maximum de Vraisemblance ou Maximum likelihood

L'estimation du Maximum de Vraisemblance (MV) est une méthode statistique supervisée. Au lieu d'affecter un objet à la classe dont le centre de gravité est le plus proche (K-moyennes) ou dont les éléments sont les plus proches (K-PPV), il se base sur l'analyse statistique de la distribution des éléments de la base d'apprentissage pour définir des probabilités d'appartenance à chaque classe. Le nouvel objet est assigné à la classe pour laquelle la probabilité d'appartenance est la plus élevée.

Un avantage de cette méthode est qu'elle fournit, en plus de la classe, un degré de confiance lié à ce choix. Il faut cependant faire une hypothèse sur le type de distribution des éléments de la base d'apprentissage. Dans le cas d'une distribution gaussienne, on cherche à maximiser pour chaque nouvel objet $\vec{x} \in \mathrm{R}^m$ la probabilité d'appartenance à la classe y_i :

$$arg \max_{y_i} P(\vec{x}/y_i) = arg \max_{y_i} \frac{1}{\sqrt{2\pi^m |Q_i|}} \exp\left(-\frac{1}{2}(\vec{x} - \overrightarrow{\mu_i})^T . Q_i^{-1} . (\vec{x} - \overrightarrow{\mu_i})\right)$$

où $\overrightarrow{\mu_i}$ et Qi désignent respectivement la moyenne et la matrice de covariance associées à

la classe y_i.

L'algorithme du Maximum de Vraisemblance est donc le suivant :

– On calcule des statistiques pour chaque classe de la base d'apprentissage : moyennes $\overrightarrow{\mu_i}$ et matrices de covariance Q_i,

– Pour chaque point :

- ✓ on calcule les probabilités d'appartenance à chaque classe : $P(\vec{x}/y_i)$
- ✓ on assigne le point à la classe ayant la plus grande probabilité.

Dans le paragraphe ci-dessus nous avons présenté les algorithmes classification classiques et nous avons remarqué que leur utilisation présente quelques inconvénients. Ci-après, nous avons présenté quelques algorithmes plus récents.

III.1.3 Les techniques de classification avancées.

Ces techniques de classification, certes plus complexes à implémenter, sont des améliorations des algorithmes classiques et par conséquent donnent de meilleurs résultats. Elles sont toutes basées sur des apprentissages. Nous présentons brièvement, dans le paragraphe suivant quelques-unes de ces techniques.

III.1.3.1 Support Vector Machine (SVM)

Les machines à vecteurs de support ou séparateurs à vaste marge (en anglais Support Vector Machine, SVM) sont un ensemble de techniques d'apprentissage supervisé destinées à résoudre des problèmes de discrimination et de régression. Les SVM sont les plus puissants algorithmes d'apprentissage automatique à ce jour. Elles sont basées sur des algorithmes optimisés pour repérer les limites entre les classes.

Les séparateurs à vastes marges reposent sur deux idées clés: la notion de marge maximale et la notion de fonction noyau. La marge est la distance entre la frontière de séparation et les échantillons les plus proches. Ces derniers sont appelés vecteurs supports. Dans les SVM, la frontière de séparation est choisie comme celle qui maximise la marge afin de pouvoir traiter des cas où les données ne sont pas linéairement séparables. La deuxième idée clé des SVM est de transformer l'espace de représentation des données d'entrée en un espace de plus grande dimension, dans lequel il est probable qu'il existe un séparateur linéaire. Ceci est réalisé grâce à une fonction noyau.

Pour rappel, le cas simple est le cas d'une fonction discriminante linéaire, obtenue par combinaison linéaire du vecteur d'entrée $x = (\mathbf{x}_1, ..., \mathbf{x}_N)^T$: $h(x) = w^T x + w_0$
La frontière de décision h(x) = 0 est un hyperplan, appelé hyperplan séparateur, ou séparatrice. Rappelons que le but d'un algorithme d'apprentissage supervisé est d'apprendre la fonction h(x) par le biais d'un ensemble d'apprentissage.

III.1.3.2 Les Algorithmes Génétiques

Les algorithmes génétiques sont des algorithmes d'optimisation s'appuyant sur des techniques dérivées de la génétique et de l'évolution naturelle: croisements, mutations, sélection (Alliot, 1996), etc. Les algorithmes génétiques ont déjà une histoire relativement ancienne puisque les premiers travaux de John Holland sur les systèmes adaptatifs remontent à 1962 (Holland, 1962).
Un algorithme génétique recherche le ou les extrema d'une fonction définie sur un espace de données. Pour l'utiliser, on doit disposer des cinq éléments suivants :

1. Un principe de codage de l'élément de population. Cette étape associe à chacun des points de l'espace d'état une structure de données. Elle se place généralement après une phase de modélisation mathématique du problème traité. La qualité du codage des données conditionne le succès des algorithmes génétiques. Les codages binaires ont été très utilisés à l'origine. Les codages réels sont désormais largement utilisés, notamment dans les domaines applicatifs pour l'optimisation de problèmes à variables réelles.
2. Un mécanisme de génération de la population initiale. Ce mécanisme doit être capable de produire une population d'individus non homogène qui servira de base pour les générations futures. Le choix de la population initiale est important car il peut rendre plus ou moins rapide la convergence vers l'optimum global. Dans le cas où l'on ne connaît rien du problème à résoudre, il est essentiel que la population initiale soit répartie sur tout le domaine de recherche.
3. Une fonction à optimiser. Celle-ci retourne une valeur de R^+ appelée *fitness* ou fonction d'évaluation de l'individu.
4. Des opérateurs permettant de diversifier la population au cours des générations et d'explorer l'espace d'état. L'opérateur de croisement recompose les gènes

d'individus existant dans la population, l'opérateur de mutation a pour but de garantir l'exploration de l'espace d'états.

5. Des paramètres de dimensionnement: taille de la population, nombre total de générations ou critère d'arrêt, probabilités d'application des opérateurs de croisement et de mutation.

III.1.3.3 Les cartes auto-organisatrices de Kohonen

Nous nous sommes intéressés aux cartes auto-organisatrices (Self-organizing maps, SOM) qui font partie des méthodes de classification non supervisées. Il s'agit d'analyser les structures d'un seul ensemble de données cohérentes entre elles, en ne donnant aucune indication a priori sur les structures recherchées. Cela signifie que, dans une première approche, ces modèles seront utilisés dans un but descriptif. Les données à analyser sont constituées d'observations dont on cherche à comprendre la structure éventuellement pour trouver des similitudes. Ainsi, dans le cadre de notre travail, elles sont utilisées pour décrire l'ensemble des situations rencontrées dans la base de données. Nous reviendrons plus en détail en III. 2 sur les SOMs. Les méthodes que nous venons de rappeler présentent des limites bien que beaucoup de progrès aient été faits pour comprendre et améliorer les algorithmes d'apprentissage.

Les SVM donnent un apprentissage assez lent et les paramètres optimaux sont difficiles à déterminer si les données d'entrainement ne sont pas linéairement séparables. De plus, l'ajustement d'un algorithme génétique est délicat. L'un des problèmes les plus caractéristiques des algorithmes est celui de la dérive génétique, qui fait qu'un bon individu se met, en l'espace de quelques générations, à envahir toute la population. Un autre problème surgit lorsque les différents individus se mettent à avoir des performances similaires : les bons éléments ne sont alors plus sélectionnés, et l'algorithme ne progresse plus.

Nous pouvons citer également le fait que certains de ces algorithmes ne sont applicables qu'à un certain nombre de domaines assez limités et que leur généralisation devient rapidement très compliquée. Les connaissances expertes utilisées dans le cas des apprentissages supervisés peuvent ne pas être totalement fiables et induire des erreurs de classification de données considérées.

Après cette brève description de ces algorithmes de classification ainsi que leurs limites, nous présentons ci-après en III.2, les cartes auto-organisatrices qui constituent la base des modèles d'inversion développés dans cette thèse.

III.2 LES CARTES AUTO-ORGANISATRICES

L'utilisation de ces cartes a permis de s'affranchir de certaines limites des techniques citées précédemment. Par exemple, dans le cas de données non linéairement séparables, comme celles que nous étudions, les SVM classiques ne peuvent pas être utilisés et le problème d'optimisation n'a pas de solution. Il faut faire appel à d'autres techniques (Pinquier, 2004). Les problèmes liés à une connaissance experte a priori ne se posent pas ici car l'apprentissage est non supervisé.

III.2.1 Généralités sur les Cartes Auto-organisatrice

Les SOM, envisagées par Kohonen en 1981 (Kohonen, 1981), cherchent, par apprentissage à partir des données, à partitionner l'ensemble des observations disponibles en groupements similaires. Les groupements proposés possèdent la particularité caractéristique d'avoir une structure de voisinage qui peut être matérialisée à l'aide d'un espace discret que l'on appelle « carte topologique ». Il s'agit le plus souvent d'un treillis de faible dimension (grille 1D, 2D ou 3D) sur lequel les structures de voisinages sont prises en considération par le modèle.

Les neurones sont disposés sur une grille régulière et sont connectés entre eux par une relation de voisinage ce qui crée la topologie de la carte. L'architecture de la carte est décrite ci-après.

Architecture

Une carte de Kohonen est un réseau de neurones à deux couches (cf. Figure III-1):

- La couche d'entrée sert à la présentation des observations à classer ; cette couche contient un vecteur de taille n (*n* étant la dimension de l'espace des observations). On note *D* l'espace des observations possibles et *A* un ensemble d'apprentissage contenu dans *D*.

$A = \{x_i \in R^n ; i = 1 \text{ à } m\}$ avec *m* le nombre de vecteurs réels de dimension *n*, représentatifs du problème à traiter.

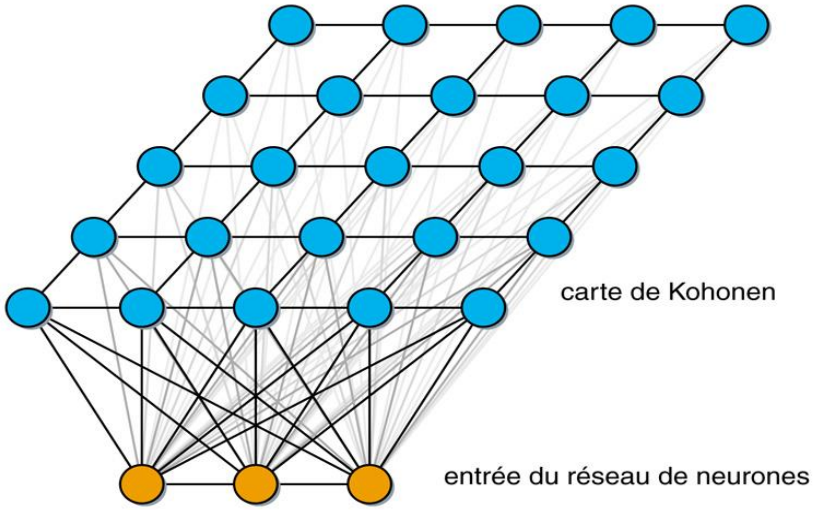


Figure III-1: Architecture d'une carte topologique en 2-D. Le réseau est constitué de deux couches : une couche d'entrée qui sert à la présentation des observations et une couche d'adaptation (pour laquelle il faut définir un voisinage, cf. figure ci-aprés) formée d'un treillis régulier à 2 dimensions dont chaque noeud est occupé par un neurone, qui est lui-même connecté à tous les éléments de la couche d'entrée. Chaque neurone c est affecté d'un référent y_c.

- La couche de sortie est formée du treillis des neurones qui forment la carte. Chaque neurone est connecté à tous les éléments de la couche d'entrée. Le vecteur référent y_c est le vecteur de poids associé au neurone *c*.

Ce treillis représente un espace discret, de faible dimension, muni d'une topologie engendrée par sa structure. Nous définissons une distance discrète δ pour toute paire de neurones *(c, r)* de la "carte" *C*, *δ(c, r)* représente la longueur du plus court chemin entre *r* et *c*. Pour chaque neurone *c*, cette distance discrète permet de définir la notion de voisinage d'ordre *v* de *c* : $V_c(v) = \{r \in C, \delta(c, r) \leq v\}$. Cette notion de voisinage définit, avec la forme de la carte, ce qu'on appelle la « topologie » de la carte. Le terme «topologie» est lié à la particularité de ce type de carte. Ces cartes prennent en compte la continuité des données dans *D*: c'est à dire que deux neurones voisins dans *C* (distance discrète) doivent représenter deux ensembles d'observations voisines dans *D* (distance euclidienne). Quant au terme «auto-organisatrice», il fait référence principalement à 4 constituants essentiels d'une carte auto-organisatrice qui sont les suivants (Kohonen, 1982):

1) Un tableau d'unités de traitement qui reçoit des entrées cohérentes d'un espace d'événements et forme de simples fonctions discriminantes des signaux d'entrée.
2) Un mécanisme qui compare les fonctions discriminantes et sélectionne l'unité avec la plus grande valeur de la fonction.
3) Une interaction locale qui active simultanément l'unité sélectionnée et ses voisins les plus proches.
4) Un processus adaptif qui fait que les paramètres des unités activées augmentent leurs valeurs de fonction discriminante en rapport avec la présente entrée.

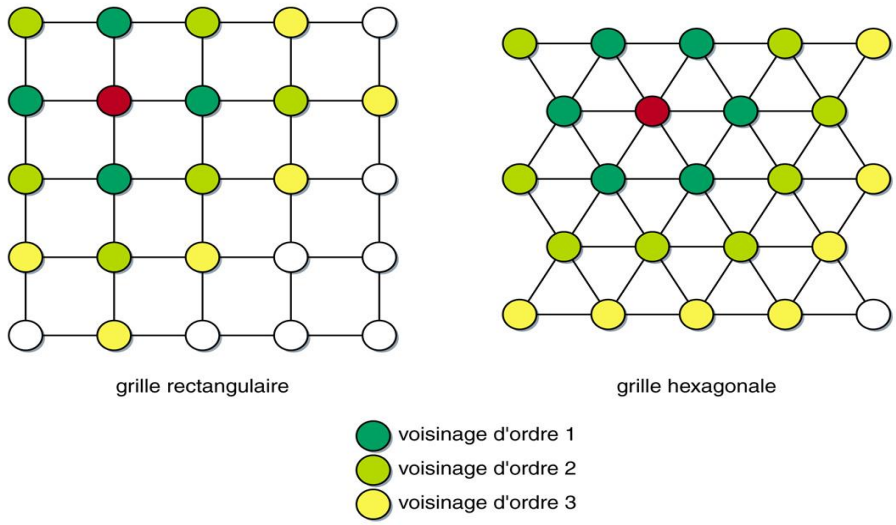


Figure III-2: Représentation de la topologie discrète d'une carte à deux dimensions; chaque point de la figure représente un neurone c. La distance δ entre deux neurones permet de définir le voisinage d'ordre v (ici v = 1, 2 ou 3), qui représente l'ensemble des neurones dont la distance au neurone c est inférieure ou égale à v.

Apprentissage

Il s'agit de réduire, dans l'espace discret C constitué d'un ensemble de P groupes Y_p représentatifs, les observations x_i ($x_i \in R^n$) d'une base d'apprentissage. A chacun des P groupes, aussi appelés neurones, on associera un vecteur $y_p \in R^n$ qui sera le vecteur référent du groupe Y_p. On a donc une fonction d'affectation $\chi(x_i) = y_p$ qui associe à chaque élément de la base d'apprentissage x_i son référent. La quantification vectorielle réalisée par les cartes de Kohonen encore appelée quantification par carte fait ainsi une réduction

de l'information différente de celle de la méthode des K-Means (leur ancêtre, cf. III.1.2.1) par l'introduction de l'ordre topologique et de la notion de voisinage entre les différents référents.

En réponse à la présentation au réseau d'un vecteur d'observation x_i, chaque neurone de C calcule sa distance par rapport à celui-ci. Celui qui est le plus proche au sens de la distance euclidienne, est le neurone gagnant (celui qui va capter l'observation) et cette distance est donnée par la formule suivante :

$$d(x_i, y_\chi(x_i)) = \sum_{p=1}^{P} K^T(\delta(y_p, \chi(x_i))) \left\| x_i - y_p \right\|^2$$

K^T est une fonction de voisinage qui pondère le voisinage.

A chaque fois que l'on cherche y et χ , on aboutit à une fonction du coût qu'il faut minimiser. Cette fonction de coût notée J est encore définie comme la somme des carrées des distances de chaque neurone par rapport à son réfèrent et est donnée par la formule suivante :

$$J(\chi, Y) = \sum_{i=1}^{N} \sum_{p=1}^{P} K(\delta(y_p, \chi(x_i))) \left\| x_i - y_p \right\|^2$$

Pour l'ensemble Y de référents fixé, la minimisation de la fonction de coût s'obtient en affectant chaque observation x_i au référent y_p qui est le plus proche selon une nouvelle fonction d'affectation χ^t : c'est la phase d'affectation.

$$\chi^t(x_i) = \underset{\chi}{\operatorname{argmin}}[d(x_i, y_\chi(x_i))]$$

La partition χ^t étant fixée, la fonction $J(\chi, Y)$ est minimisée par rapport à l'ensemble des référents Y. La fonction est convexe par rapport aux paramètres. Le minimum global est atteint pour :

$$y_p^t = \frac{\sum_{q \in C} K(\delta(p,q)) X_q}{\sum_{q \in C} K(\delta(p,q)) n_q}$$

Xq représente la somme de toutes les observations affectées au neurone q, nq le nombre de ces observations.

Le neurone de la carte, après avoir capté une donnée, peut se déplacer. Il modifie dans ce

cas sa position et son voisinage. Le vecteur poids qui déterminait sa position et son voisinage change, ce qui entraine le changement de la position de ses voisins. En effet, plus un neurone est proche du neurone sélectionné et plus il sera influencé. La proportion avec laquelle la modification du vecteur poids d'un neurone affecte celui-ci est déterminée par la fonction de voisinage. Celle-ci joue un rôle prépondérant dans l'adaptation du réseau à ses entrées et varie en fonction du nombre de cycles réalisés. Elle est régie par la donnée d'un paramètre communément appelé température qui décroit au cours de l'évolution de l'apprentissage. Lorsque celui-ci progresse, la fonction de voisinage change et plus particulièrement, le voisinage diminue.

La notion de voisinage peut être introduite à l'aide de fonction noyaux positives symétriques paramétrées par la température T comme fonctions de type gaussien ou de type voisinage à seuil notées $K^T(\delta)$ qui gèrent donc la taille du voisinage et définies de la sorte :

- pour la fonction de voisinage à seuil par :

$$K^T(\delta) = \begin{cases} 1 \; si \; \delta < T \\ 0 \quad sinon \end{cases}$$

- et pour la fonction de voisinage de type gaussien par :

$$K(\delta) = \exp(-|\delta| / T)$$

Dans la fonction de voisinage à seuil, les neurones du voisinage ont la même influence, en dehors ils n'en ont aucune.

Dans la fonction de voisinage de type gaussien, l'influence entre deux neurones dépend de la distance entre ces neurones. Ainsi donc, la distance $\delta(y_p, y_q)$ permet de faire varier l'influence relative des différents neurones, cette importance est quantifiée par $K^T(\delta(y_p, y_q))$.

Une fois l'apprentissage terminé, l'ensemble des données est partitionné en autant de groupes qu'il y a de neurones.

L'algorithme de Kohonen peut se résumer comme suit :

Début

1. **Phase d'initialisation** :

- *Choisir la structure et la taille de la carte et les p vecteurs référents initiaux (en général d'une manière aléatoire) ;*

- *Fixer les valeurs de T_{min} et T_{max}, le nombre d'itérations N_{iter}, et prendre t=0.*

2. Etape itérative t :

L'ensemble des vecteurs référents Y^{t-1} de l'étape précédente étant connu :

- *Choisir une observation x_i (en général d'une manière aléatoire) ;*
- *Calculer la nouvelle valeur de T en appliquant la formule :*

$$T = T_{max} * (\frac{T_{min}}{T_{max}})^{\frac{t}{N_{iter}-1}}$$

Pour cette valeur du paramètre T, effectuer les deux phases suivantes :

- ***Phase d'affectation** : on affecte l'observation x_i au neurone $\chi^t(x_i)$ défini à partir de la fonction d'affectation suivante :*

$$\chi^t(x_i) = \underset{c}{\operatorname{argmin}} \|x_i - y_c\|^2$$

- ***Phase de minimisation** : calcul de l'ensemble des nouveaux référents Y^t; les vecteurs référents sont modifiés selon la formule ci-dessous en fonction de leur distance au neurone sélectionné à l'étape d'affectation.*

$$y_c^t = y_c^{t-1} - \mu^t K^T(\delta(c, \chi^t(x_i))(y_c^{t-1} - x_i)$$

3. Phase finale :

Répéter l'étape itérative jusqu'à ce qu'on l'on atteigne $t = N_{iter}$.

Fin

Bien que la conservation de la topologie dans l'espace des observations soit la particularité principale de l'algorithme de Kohonen, on trouve d'autres différences et des similitudes avec les autres familles d'algorithmes de classification. Gueye (2010) a fait un exposé plus détaillé de certaines de ces différences.

Les SOM ont été appliquées sur les données DRAKKAR colocalisées ARAMIS, dans le but d'approfondir l'étude comparative de ces 2 ensembles dont les premiers résultats ont été discutés en II.4. Le paragraphe suivant décrit cette approche comparative sur la base de méthode neuronale.

III.2.2 Intercomparaison DRAKKAR/ARAMIS par approche neuronale.

Dans cette partie nous avons utilisé les réseaux de neurones et plus particulièrement les cartes de Kohonen présentées III.2.1 pour compléter notre analyse comparative DRAKKAR/ARAMIS.

L'idée est de construire une carte avec les données DRAKKAR et ensuite d'étudier le comportement des données in situ (ARAMIS) sur celle-ci. Le problème de structure de données réapparait ici. Pour y remédier en termes de profondeur, nous avons appliqué aux données DRAKKAR la 3ème méthode de ré-échantillonnage décrite dans le paragraphe II.3.4. C'est-à-dire que les données DRAKKAR ont été interpolées sur le nombre d'immersions d'ARAMIS. Rappelons que les profondeurs d'ARAMIS utilisées dans cette comparaison varient de 1 à 650m. Cette profondeur finale est encadrée par la 21ème et la 22ème immersion de DRAKKAR. Le nombre d'immersions de DRAKKAR passe de 22 à 650 après interpolation. Spatialement, les données DRAKKAR sont sélectionnées sur la route moyenne ARAMIS (route moyenne de toutes les campagnes) et à tous les pas de temps DRAKKAR c'est-à-dire tous les 5 jours de 2000 à 2007, soit 27480 données après prétraitement. Ces données ont donc chacune 650 niveaux d'immersion et 1/10 de ces niveaux est utilisé pour mettre en place la carte. Ainsi, les variables utilisées pour la carte sont les valeurs de S à 65 immersions (1/10 de 650).

Le réseau a les caractéristiques suivantes :

Type=SOM, carte de Kohonen

Apprentissage= 1/3 des données

Architecture :

- Nombre de neurones : 18x11 neurones : cette topologie permet une bonne couverture de la variabilité avec un temps de calcul optimal.
- Voisinage hexagonal => chaque neurone à 6 voisins

Nombre de variables=65

A l'issue de l'apprentissage, la carte obtenue a été analysée. Rappelons qu'il n'existe pas une méthode standard pour présenter les résultats d'un apprentissage avec une carte de Kohonen. La démarche proposée dans ce paragraphe permet d'introduire des concepts

qui sont utilisés dans cette thèse pour analyser d'autres cartes du même type. Cette analyse du résultat de l'apprentissage est faite ci-après.

Pour étudier globalement la manière donc DRAKKAR reproduit les données ARAMIS, ces dernières ont été projetées sur la carte : chaque donnée ARAMIS est captée par le neurone qui lui ressemble le plus dans l'ensemble des référents. Cette ressemblance est quantifiée par la distance euclidienne calculée sur les 65 immersions qui ont servi à construire la SOM. Chaque neurone de la carte a un profil de S qui correspond à son vecteur référent. La Figure III-3 et la Figure III-4 illustrent quelques neurones (en rouge) et leurs profils ARAMIS captés (en bleu).

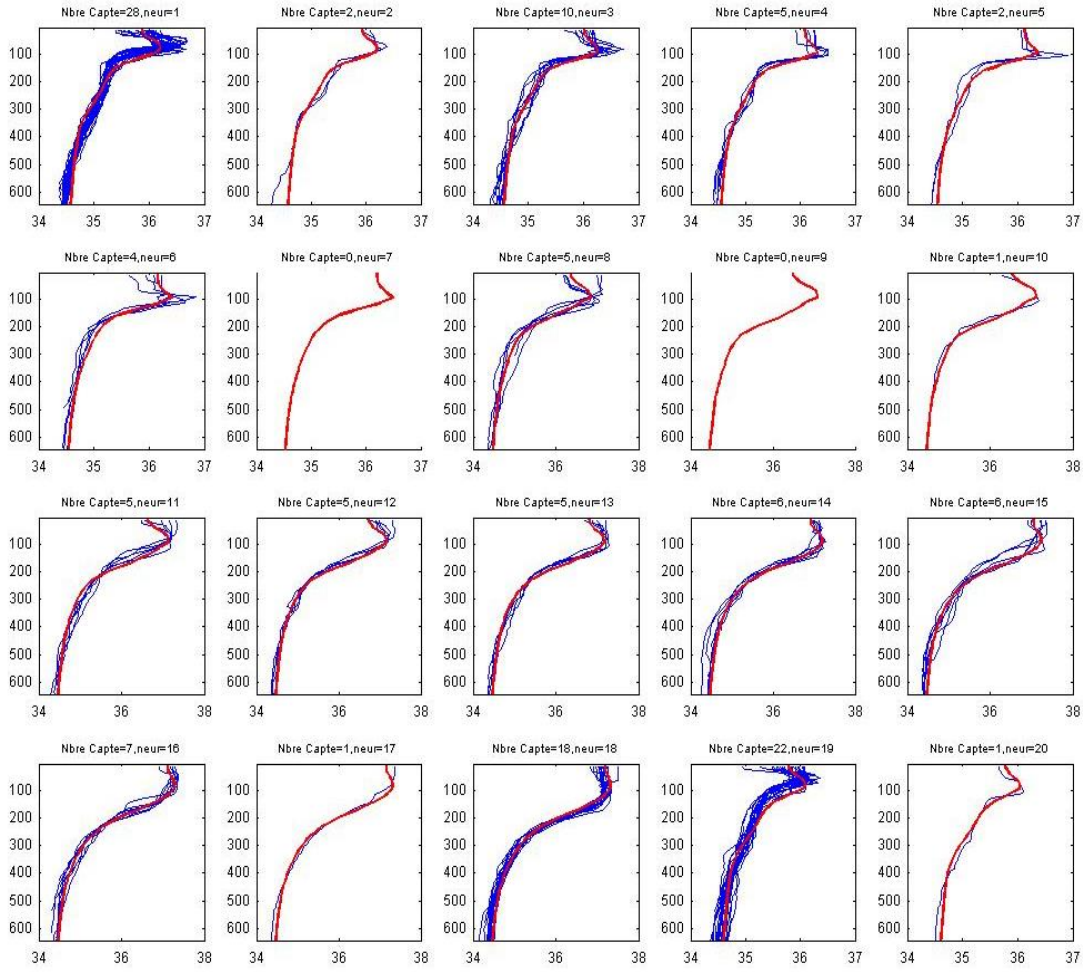


Figure III-3:exemple1 de neurones, en abscisse les S, en ordonnée les immersions. Neur donne le numéro du neurone sur la carte de Kohonen. Nbre Capte le nombre de profils ARAMIS captés par le neurone. Ces profils captés sont indiqués en bleu, tandis que le profil référent du neurone est en rouge.

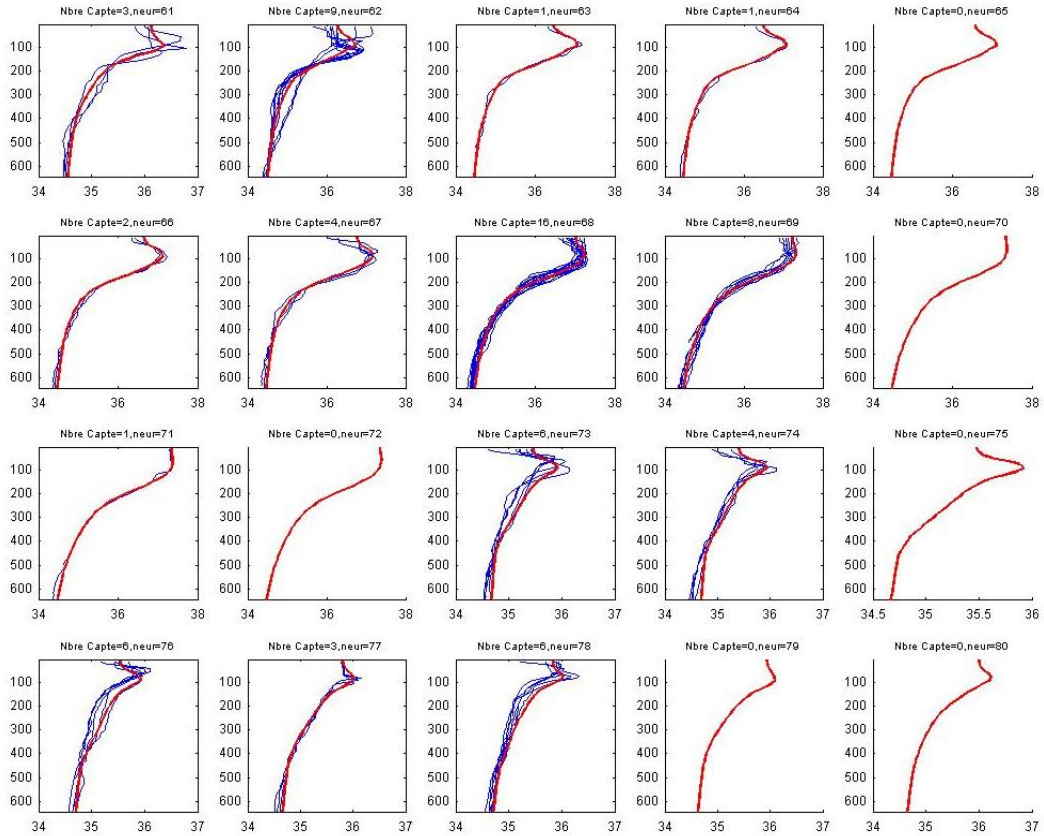


Figure III-4: Autres exemples de profils, idem Figure III-4.

Dans l'ensemble, les profils ARAMIS (en bleu) ont la même allure que le profil du neurone gagnant. Cependant, il y a des différences importantes, dans quelques cas, de la surface à 200m, là où la variabilité de S est importante. Cela confirme les résultats précédents obtenus en II.4 qui montrent que le modèle DRAKKAR reproduit mieux les couches de sub-surface que les couches de surface plus variables. Cette variabilité est très difficile à modéliser car elle résulte souvent de phénomènes complexes comme la pluviométrie, l'évaporation, donc les actions du vent et du soleil, les décharges fluviales, les entrainements liés aux courants marins y compris les mouvements verticaux. Les Figure III-3 et Figure III-4 donnent également des cas où les profils ARAMIS ressemblent un peu moins au profil du neurone gagnant, par exemple le neurone 62 de la Figure III-4. Ceci s'explique par le fait que les neurones ne sont pas créés avec les données ARAMIS mais avec les données DRAKKAR. Manifestement, le type de variabilité de ces profils ARAMIS n'existe pas dans DRAKKAR. Or, comme tout profil ARAMIS projeté doit être capté par un neurone, ces profils "spéciaux", qui ne sont pas représentés dans la carte

SOM, sont donc captés par des référents ayant une similitude approximative. On remarque que des neurones issus des données DRAKKAR n'ont pas capté de données ARAMIS, ceci est dû au fait que le réseau a été construit avec des données DRAKKAR qui couvrent un espace temporel beaucoup plus vaste que celui d'ARAMIS.

Dans cette partie nous avons caractérisé les neurones avec une variable qui n'a pas participé à l'apprentissage: la latitude. Le choix de cette variable est lié au fait qu'il existe une relation entre S et la latitude. Cette relation est explicitée plus loin dans la suite du document. Nous cherchons à voir si cette relation est présente dans la carte, cela veut dire aussi que ce paramètre est caractéristique des référents. Pour ressortir cette relation, la représentation de la latitude moyenne des données captées par les neurones est une solution, comme illustré par la Figure III-5.

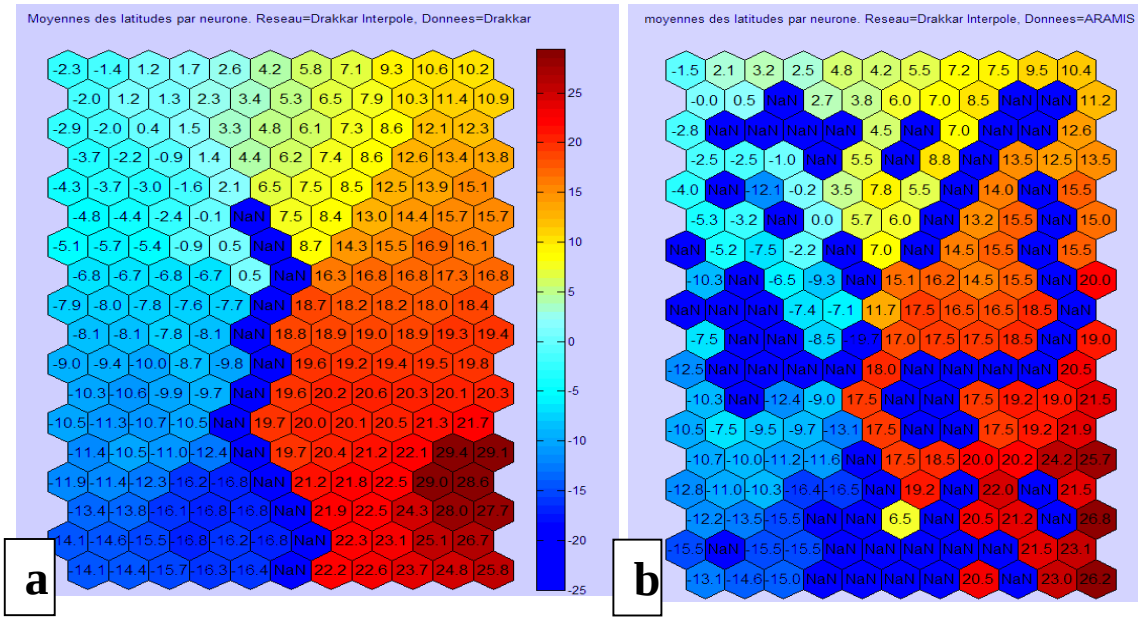


Figure III-5: Latitudes moyennes des profils des données captés par chaque neurone. (a) désignent la projection des profils DRAKKAR et (b) ceux d'ARAMIS sur la SOM. Les NaN indiquent les neurones qui n'ont pas capté de données.

Comme présenté en III.2.1, la SOM utilisée ici est en dimension 2 avec un voisinage hexagonal, donc la carte est formée d'un rectangle entrecoupé en plusieurs hexagones. Dans ces hexagones, nous pouvons mettre les valeurs correspondantes de ce qui est étudié. Ce type de carte peut être utilisé pour visualiser toutes les variables des référents des neurones (leur poids) ou des statistiques sur les données même concernant des variables qui n'ont pas participé à l'apprentissage. Dans les chapitres suivants, ce type de

visualisation sera très utilisé.

La Figure III-5 présente la topologie de la carte. A chaque neurone est associée la latitude moyenne des données qu'il a captées, soit pour DRAKKAR (Figure III-5a) soit pour ARAMIS (Figure III-5b). Les référents proches « topologiquement » présentent des latitudes moyennes proches numériquement, c'est-à-dire que les valeurs calculées sur les données captées par un neurone et celles calculées sur les données de ses voisins sont très proches. On peut dire que la topologie de la carte est totalement respectée sauf pour quelques rares cas visibles sur les données ARAMIS. Cependant cela ne suffit pas pour conclure que les neurones ont capté des profils géographiquement poches. La Figure III-6 donne la dispersion des latitudes autour de leur moyenne pour chaque neurone.

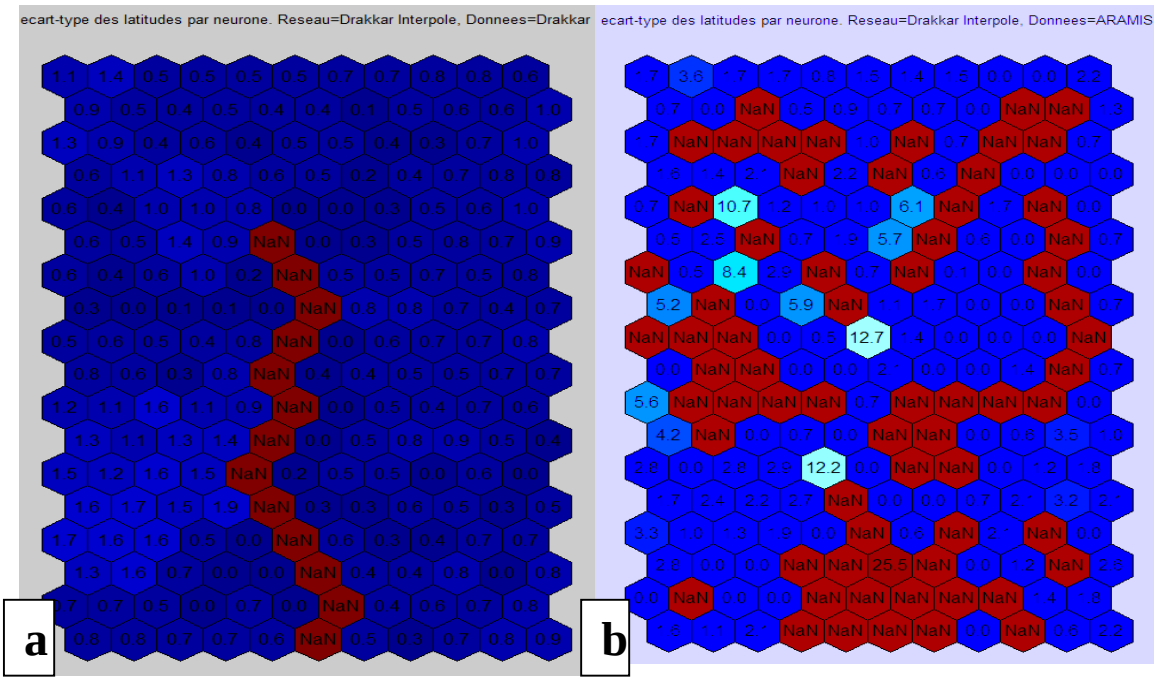


Figure III-6:Ecarts-types des latitudes des données captées par chaque neurone. (a) désigne la projection des données DRAKKAR et (b) celle des données d'ARAMIS sur la carte. Les NaN indiquent les neurones qui n'ont pas capté de données sur cette projection.

Les écarts-types sont très faibles, sauf pour quelques profils ARAMIS, ces référents sont ceux qui ont capté les profils particuliers. Cette dispersion des latitudes des profils étant faible, on en déduit que les profils captés par un même référent sont proches géographiquement.

Cette analyse est un complément de celle faite en II.4. Avec elle, nous montrons

également une nouvelle manière de comparer 2 ensembles de données en utilisant les cartes de Kohonen.

Les méthodes présentées ici ont permis de mieux cerner les limites du modèle DRAKKAR sur certaines zones et sur certaines profondeurs. Nous observons en effet quelques différences non-négligeables autour de 20-28°N: la variabilité présentée par DRAKKAR y est toujours supérieure à celle d'ARAMIS. Sur certaines immersions notamment entre 100m et 200m, immersions correspondant à l'halocline, on note un petit désaccord entre DRAKKAR et ARAMIS. Nous avons également défini les SOM qui ont été utilisées pour approfondir cette comparaison ARAMIS/DRAKKAR. Cela a permis de voir que toute la variabilité exprimée par les données ARAMIS n'existe pas dans DRAKKAR malgré une co-localisation et un ré-échantillonnage assez précis. Cependant, d'après les résultats de l'étude comparative, nous pouvons dire que, dans l'ensemble, le modèle DRAKKAR simule assez bien la S correspondant aux données ARAMIS pour l'espace et le temps considérés. Nous allons donc par la suite travailler avec les données DRAKKAR pour mettre en place les méthodes d'inversion qui constituent le sujet principal de cette thèse.

Chapitre IV. MODELES D'INVERSION DE PROFIL DE SALINITE A PARTIR DES DONNEES DE SURFACE.

Nous rappelons que ce chapitre est le cœur de la thèse car son objectif est de présenter un modèle optimisé de reconstruction du profil de S à partir de paramètres de surface notamment de la SST, de l'ADT ou de la SSH et de la SSS. Donc le questionnement scientifique est de savoir comment retrouver le bon profil de S connaissant les données de surface. Pour répondre à cette question, le problème qui consiste à trouver un lien entre surface et profil de S est considéré comme mal posé au sens de Hadamard (référence manquante) qui considère qu'un problème est bien posé si :

- ***la solution analytique existe***
- ***la solution est unique***
- ***la solution est stable vis-à-vis des erreurs de mesure.***

Dans notre cas le premier et le second point ne sont pas garantis donc le problème est mal posé. Pour trouver la relation entre surface et profil de S, une approche basée sur l'apprentissage des données est explorée, comme indiqué dans le chapitre introductif.

Dans ce présent chapitre nous proposons deux modèles d'inversion basés sur les SOM de Kohonen en utilisant une partie des données DRAKKAR pour la construction de ces modèles. Nous avons d'abord délimité les individus qui seront utilisés pour constituer une nouvelle base de données avant de décrire les méthodes d'inversion proposées et pour chacune d'elles les résultats obtenus. Le premier modèle est présenté puis commenté. L'étude approfondie des principales erreurs de ce modèle a donné certains résultats. L'exploitation de ces résultats et les solutions proposées aboutissent à un deuxième modèle. Ce nouveau modèle est appliqué aux données DRAKKAR, à celles de Coriolis et enfin à celles d'ARAMIS.

IV.1 MISE EN PLACE DU RESEAU POUR L'INVERSION

Pour l'inversion nous avons défini un réseau de neurones de type SOM de Kohonen avec les caractéristiques suivantes :

Architecture :

- *Nombre de neurones sur la carte* : 40x25 neurones soit 1000 sont choisis. Ce choix résulte d'un 1er apprentissage automatique en utilisant un ensemble

d'outils fourni par le SOM ToolBox sous Matlab (http://www.cis.hut.fi/somtoolbox/download/, consulté le 03/11/2013). Il permet une bonne couverture de la variabilité de la base de données avec un temps de calcul optimal.

- *Voisinage* : hexagonal => chaque neurone à 6 voisins.

Apprentissage :

- *Nombre d'individus*: 10% des données, présentées en II.3.3 soit 434087, centrées réduites sont utilisées pour l'apprentissage. Cette quantité d'individus est choisie en prenant 1 ligne sur 10 sans condition spécifique sur un critère.
- *Nombre de variables* : 29 (25 premières immersions de DRAKKAR, SST, SSH, Long, Lat).
- *Initialisation*: la base d'apprentissage est initialisée avec un algorithme qui permet de générer les meilleurs paramètres après une itération et un affinement de la carte. Cette carte issue de l'initialisation (1er apprentissage) a été optimisée par un nouvel apprentissage avec un autre algorithme où la température finale a été définie à 0.01, le nombre d'itérations à 100.

Les paramètres d'apprentissage ont été sélectionnés après plusieurs étapes de test avec différents paramètres.

IV.1.1 Résultat d'apprentissage.

A l'issue de l'apprentissage, différentes analyses peuvent être faites sur la carte obtenue afin de se prononcer sur sa qualité. La Figure IV-1 représente la cardinalité des neurones.

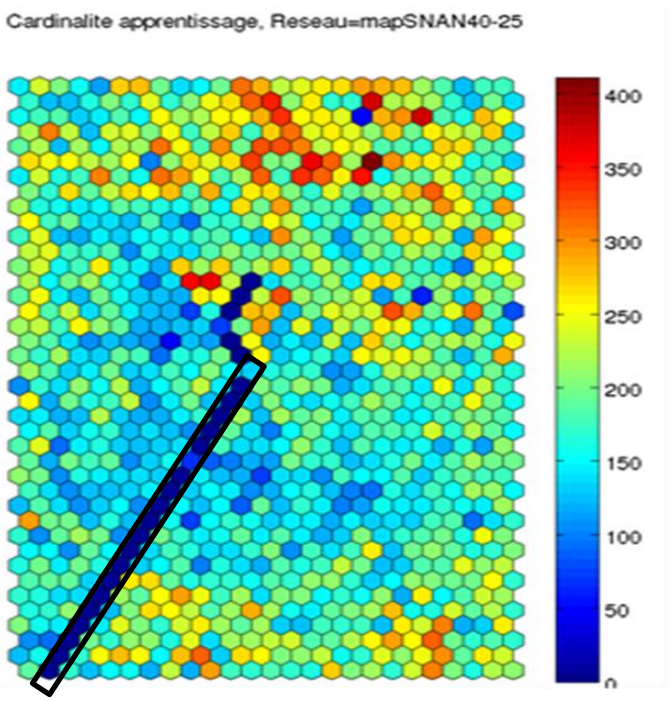


Figure IV-1:Carte des cardinalités (le nombre de données captées par chaque neurone). Une carte de 40x25 neurones, chaque petit hexagone de la figure représente un neurone et les hexagones autour sont ses voisins. Le nombre de données captées par un neurone est indiqué par sa couleur. La cardinalité varie des valeurs faibles (bleu foncé <50) aux valeurs fortes (rouge foncé > 400). Le rectangle noir encadre des neurones à cardinalité nulle.

Elle montre une zone frontière de neurones à cardinalité nulle. Cette zone sépare souvent des groupes de neurones différents et très discriminés. Les neurones qui sont de part et d'autre de cette zone ont certaines caractéristiques différentes c'est-à-dire que pour certaines variables, les référents de ces neurones présentent des valeurs très différentes. Cela permet de garder la topologie de la SOM. Nous pouvons remarquer que les cardinalités sont proches ce qui signifie qu'il n'existe pas de neurones ayant capté une partie importante des données. Nous pouvons également visualiser la carte des variables (Fig. IV. 2).

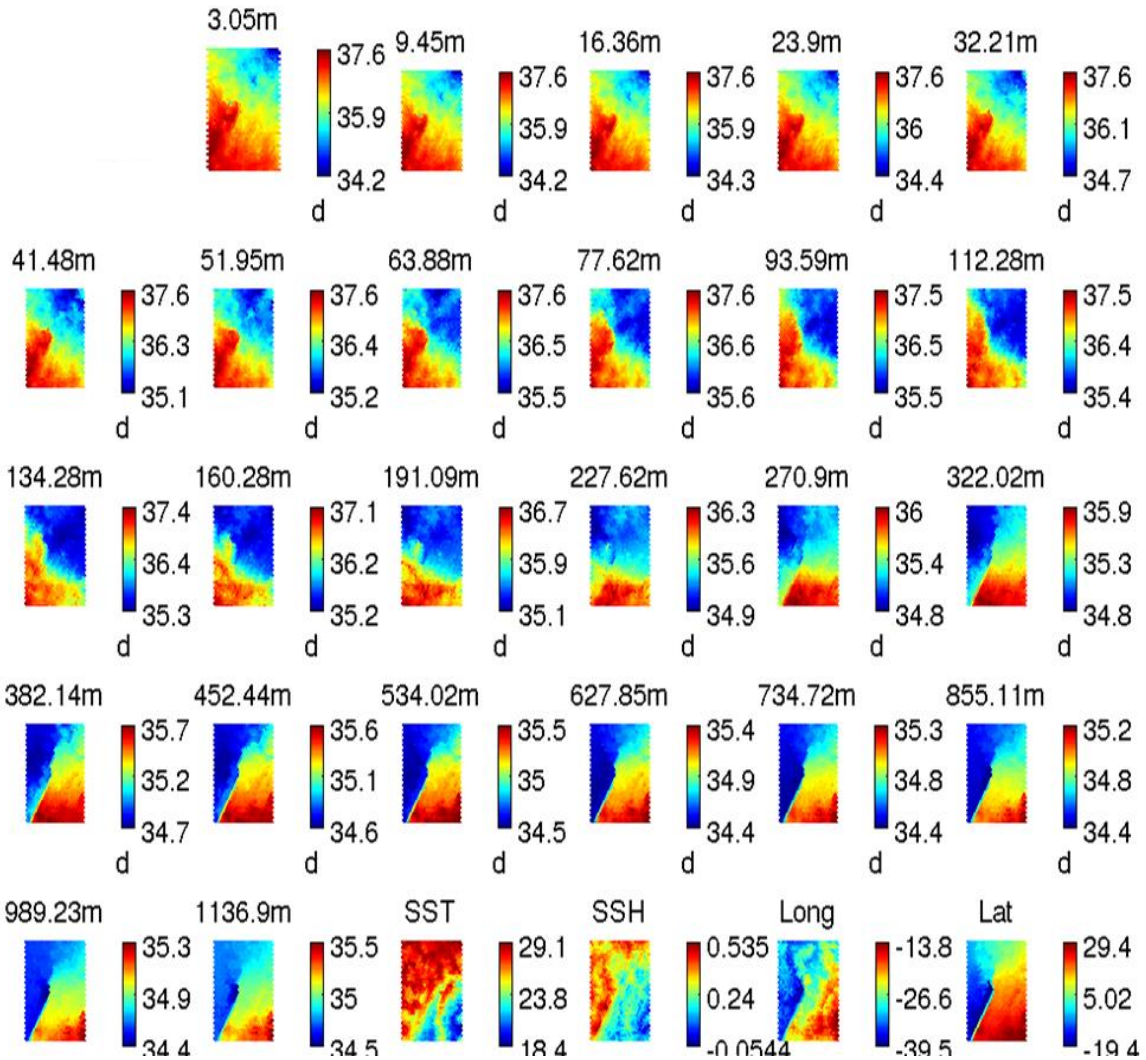


Figure IV-2:La carte des variables (les variables sont en titre de chaque rectangle). Chaque image correspond à une variable (S à différentes immersions et paramètres de surface). Chaque image est constituée de 40 x 25 hexagones qui correspondent aux neurones. Pour chaque variable les différentes couleurs donnent le poids (valeur) du neurone (l'échelle est indiquée par la barre de couleurs)

L'ensemble présente un ordre topologique bien organisé. En effet les neurones de même couleur (c'est-à-dire de poids similaires) apparaissent bien regroupés. Cette figure montre également que les immersions proches présentent des images très semblables : par exemple l'image de la variable 9,45m et celle de 16,36m sont très similaires. On peut aussi trouver différents sous-groupes selon l'organisation des poids des neurones en fonction des immersions. Les variables de 3,05m à 63,88m présentent des topologies similaires, de même que les variables de 63,88 à 160,28m et de 322,02 jusqu'à 1136,9m.

Nous pouvons noter cette ressemblance entre les immersions de sub-surface (322,14m à 1136,9m) et la latitude ; ces variables ont des valeurs faibles pour les neurones situés en bas à gauche de la carte, des valeurs fortes en bas à droite de la carte avec une frontière bien marquée et les valeurs intermédiaires sont situées dans les neurones du haut de la carte.

Pour mieux évaluer le résultat de l'apprentissage, la Figure IV-3 illustre les profils de S de quelques référents (neurones) et les profils des données qu'ils ont captées.

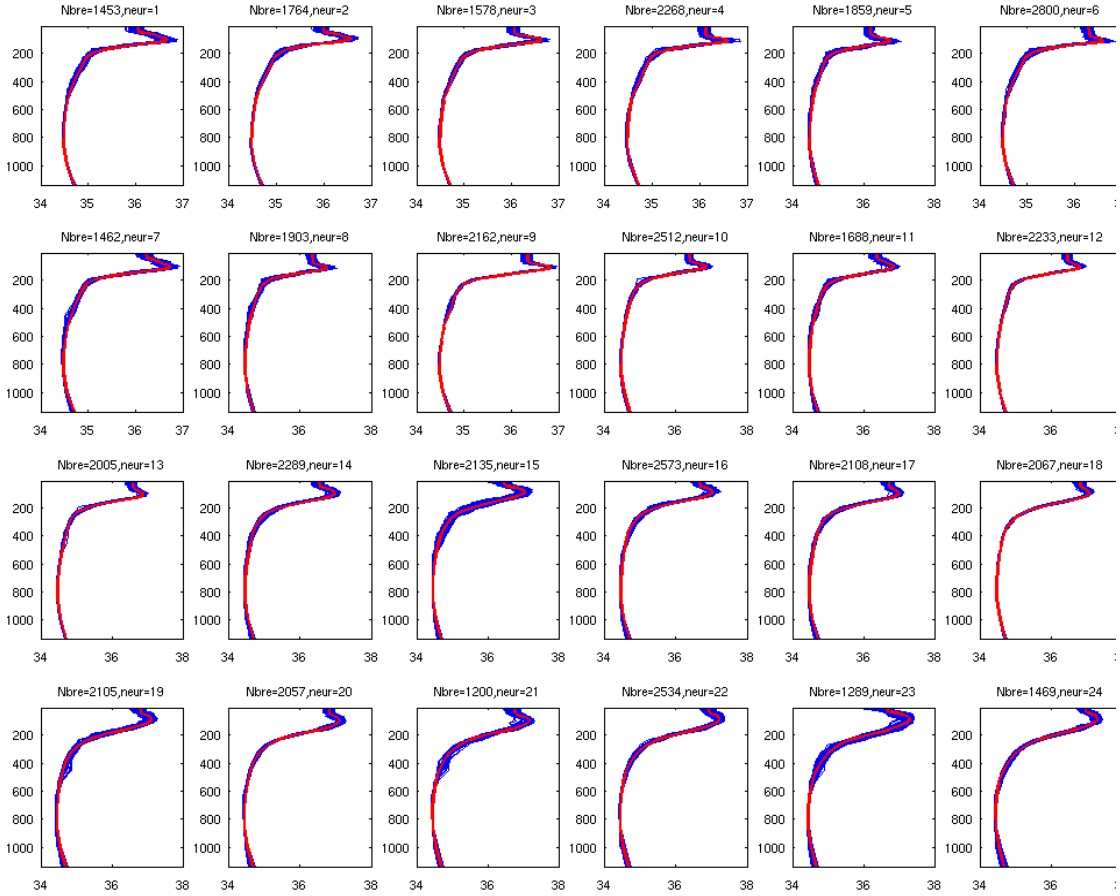
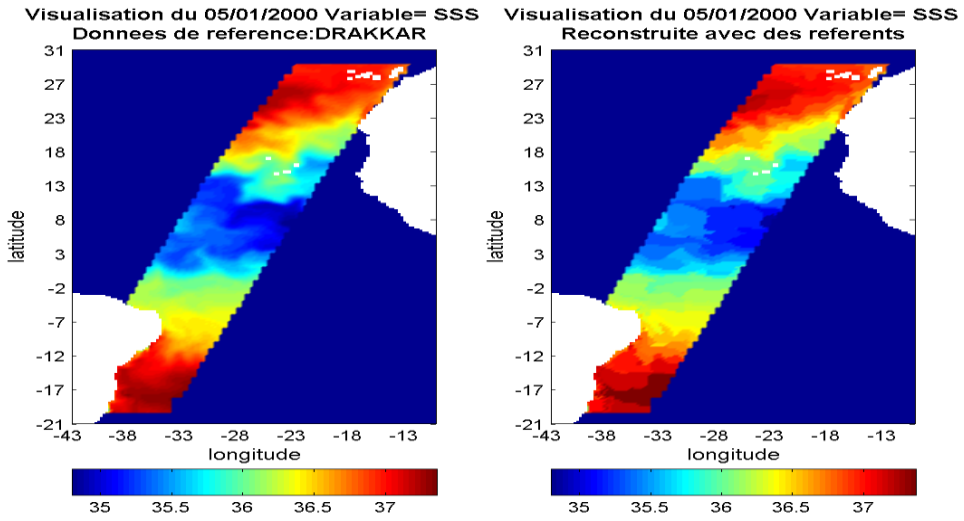


Figure IV-3:profils de quelques neurones. En rouge le neurone et en bleu les données captées par ce neurone (en abscisse la S en psu et en ordonnée les immersions en m).

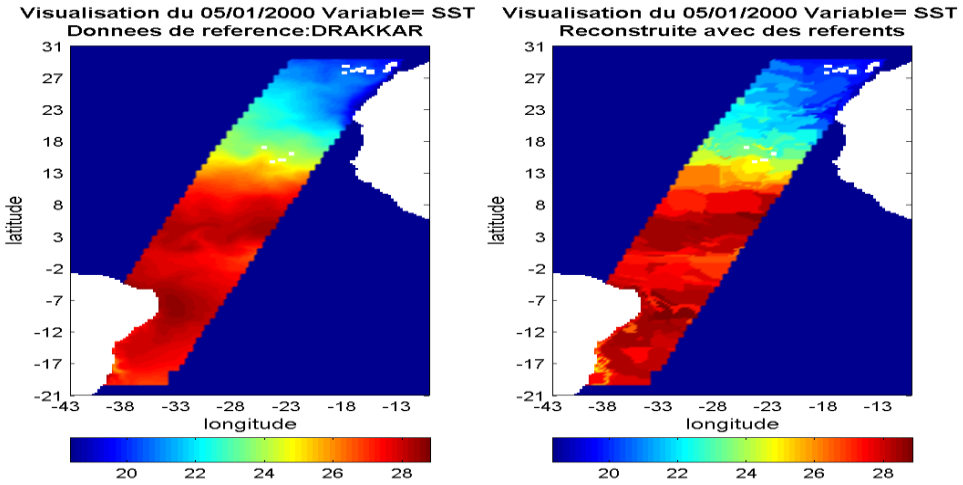
Le but principal de cette visualisation est de montrer le comportement des profils des données et de leur neurone gagnant ; elle permet de se prononcer sur la possibilité de travailler directement avec les neurones en lieu et place des données car il est plus facile de manipuler 1000 valeurs que d'en manipuler des millions. A travers cette figure, on peut voir que les données suivent généralement la même allure que celle de leur neurone

gagnant[2]. Elle montre aussi qu'effectivement les 1000 référents résument bien l'ensemble des données, et ceci est valable pour pratiquement tous les 1000 neurones. Chaque profil de neurone caractérise une classe de profils de S. La Figure IV-3 permet d'illustrer la capacité des neurones à résumer les données en termes de profils de S. Une autre manière de confirmer cette capacité est la comparaison, à partir de visualisation globale de certaines variables, des données de références et des référents des neurones. Ainsi, la Figure IV-4 donne des visualisations d'images de référence (construites avec les données DRAKKAR) et celles reconstruites avec les référents pour le 5 janvier 2000, de la SSS et la SST.



(a) SSS en psu

[2] Neurone gagnant : le neurone qui a capté la donnée donc le plus proche de la donnée en termes de distance euclidienne



(b) SST en °C

Figure IV-4:Visualisation d'images de référence DRAKKAR (à gauche) et celles reconstruites avec les 1000 vecteurs référents (à droite) pour SST (a) et SSS(b) à la date du 05/01/2000.

Pour obtenir les images reconstruites avec les 1000 vecteurs référents (à gauche de la Figure IV-4), nous avons procédé à une substitution de chaque valeur de la donnée et la variable considérées par le poids de son neurone gagnant pour cette variable. Rappelons que ces neurones gagnants sont obtenus après un apprentissage sur seulement 10% des individus de départ. Nous voyons que les images, prises deux à deux, sont similaires et les structures principales sont bien retrouvées.

IV.1.2 Variabilité des référents en fonction des paramètres de surface.

Cette étude permet de caractériser les neurones en fonction de la topologie, c'est-à-dire de leur position sur la carte ainsi que l'organisation de celle-ci. Sachant que les études précédentes ont montré que S, surtout aux plus profondes immersions, dépendait fortement de la latitude, nous avons commencé par étudier la topologie des neurones et la latitude comme le présentent les figures suivantes.

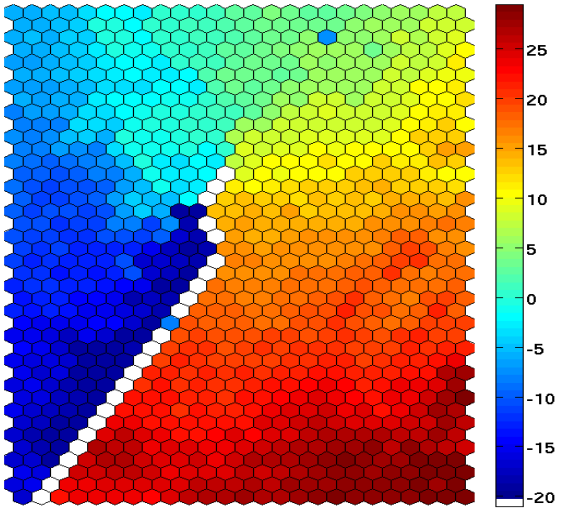


Figure IV-5: latitude moyenne des données captées par chaque neurone en °. Chaque hexagone représente un neurone et la couleur est la moyenne des latitudes des données que le neurone a captées quand on projette toutes les données de la base. La couleur blanche indique les neurones qui n'ont pas capté de données dans cette projection.

La Figure IV-5 montre que les neurones voisins ont des latitudes moyennes voisines sauf un seul neurone (en bleu, latitude autour de 10°S) localisé en haut de la carte autour de neurones caractérisés par des latitudes de 5° N. La carte peut être divisée, principalement en 5 parties. En haut de la carte, nous avons des moyennes qui varient de -7° (7°S) à gauche à 15° à droite avec des variations homogènes de la latitude. Cette zone (haut) de la carte est elle-même divisée en 3 parties. En bas, les deux parties (rouge et bleue) sont distinctes avec une zone frontière nettement visible. Cette frontière sépare donc les 2 parties, avec des moyennes variant de 20° S à 12° S d'une part et de 17° N à environ 30° N d'autre part. Nous ne pouvons toutefois pas encore conclure que l'ensemble des données projetées ont elles-mêmes cette structure, car la distribution des latitudes des données au niveau de chaque neurone peut être très étendue. L'étude de l'écart-type permet de mieux voir cette distribution.

La Figure IV-6 présente les écart-types autour de la moyenne latitudinale de chaque neurone. Cette figure présente des écarts-types faibles pour la plupart des neurones, considérant la gamme de latitudes des données étudiées (-20° à 30°). Ainsi, on peut affirmer que les neurones voisins topologiquement ont capté des données proches géographiquement, ce qui fait ressortir le pouvoir classifieur des SOM. Remarquons que le neurone qui présente un fort écart-type est celui qui se trouve dans une zone non caractéristique de sa latitude. Ce neurone a capté des données qui sont de part et d'autre de l'équateur météorologique. Cette figure présente également une zone (centre gauche de la carte) caractérisée par des neurones qui ont les plus grands écarts-types (jusqu'à 3) délimitée à droite par des neurones qui n'ont rien capté (couleur blanche).

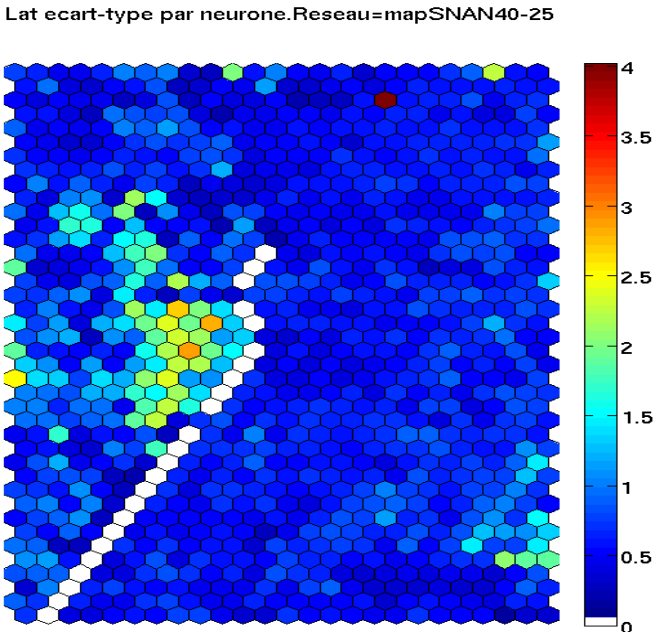
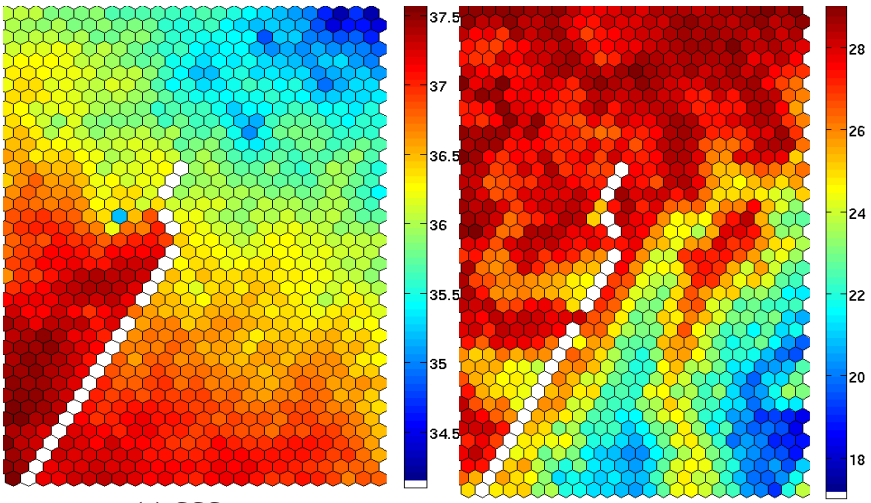


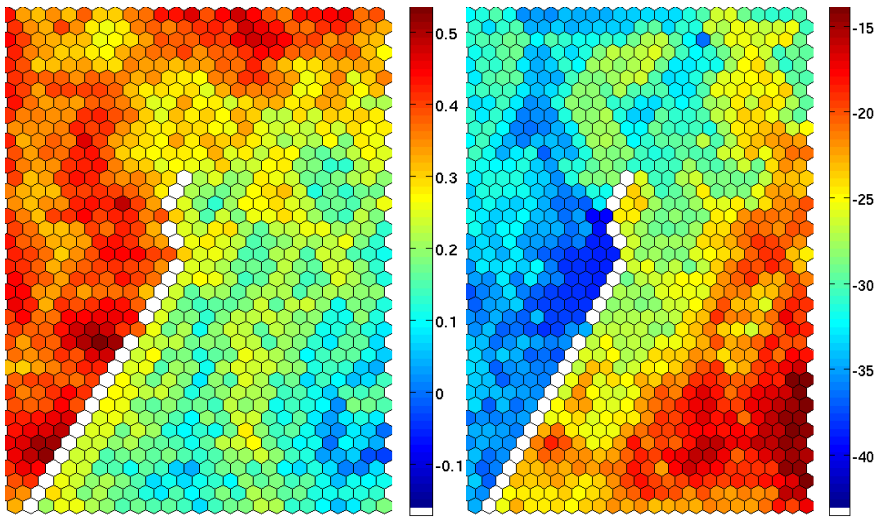
Figure IV-6 : écart-type en ° autour de la latitude moyenne des données captées par neurone (couleur blanche indique les neurones qui n'ont rien capté)

Cette étude peut être étendue aux autres paramètres de surface c'est-à-dire la SSS, la SST, la SSH et la longitude comme illustré à la Figure IV-7.



(a) SSS en psu

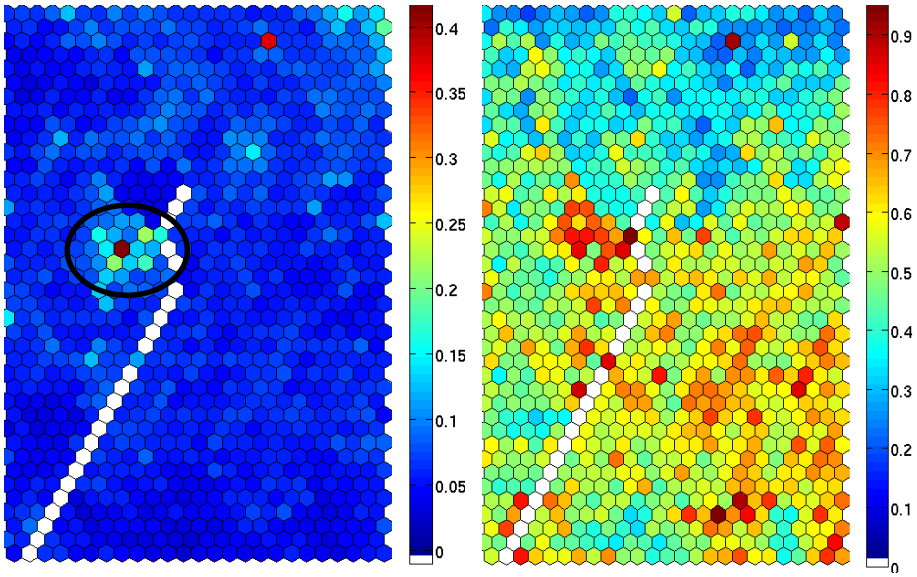
(b) SST en °C



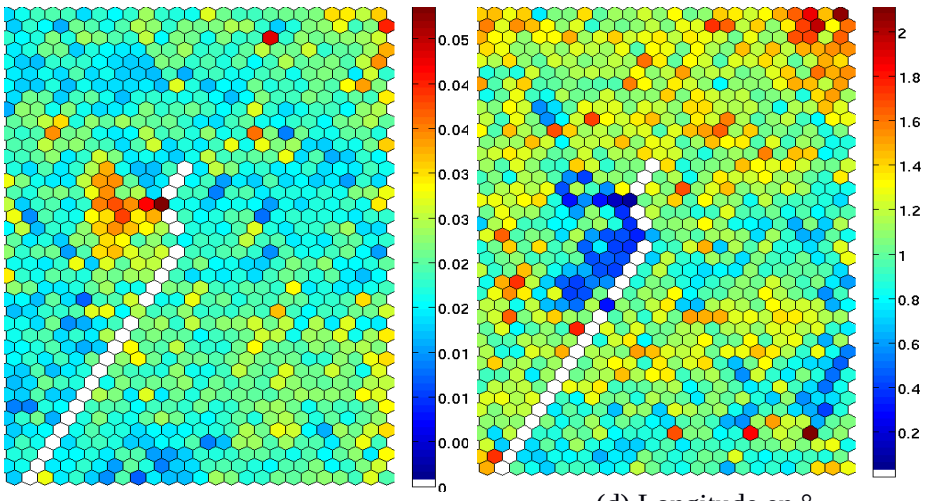
(c) SSH en m (d) Longitude en °

Figure IV-7: Moyenne par neurone des autres paramètres de surface : SSS, SST, SSH et longitude.

La Figure IV-7 montre une topologie assez organisée pour tous les paramètres. La Figure IV-8 donne les écarts-types de ces cartes.



(a) SSS en psu (b) SST en °C



(c) SSH en m (d) Longitude en °

Figure IV-8: écarts-types par neurone des autres paramètres de surface. Le cercle indique une zone (maxSD) où les écarts-types sont plus forts.

Ils sont faibles en général mais présentent différentes répartitions suivant le paramètre. Pour la SSS (Figure IV-8a), ils sont autour de 0,05 sauf pour une zone où sont regroupés quelques neurones et pour un neurone localisé à la ligne 3-colonne 17. Ce neurone a été déjà repéré avec une latitude moyenne anormale par rapport aux autres. Pour la SST (Figure IV-8b), les écarts-types sont faibles en haut de la carte. La Figure IV-5 montre que cette partie de la carte est caractérisée par des neurones de la région 5°S-5°N. Les écarts-types deviennent plus importants en bas et à droite de la carte, une zone caractérisée par des neurones de l'Atlantique nord. On note quelques extrema dispersés dans la carte. Pour la SSH et la longitude, la dispersion est plus hétérogène.

Ces analyses ont permis de voir les capacités de la carte à représenter les données ainsi que son organisation. Nous avons également vu que toutes les situations observées sur les données DRAKKAR ont été représentées dans le réseau de neurones, en effet les plages des valeurs des observations sont les mêmes que celles des référents quelle que soit la variable. Cependant le challenge est de trouver un bon neurone pour chaque donnée à partir des paramètres de surface. Les paragraphes suivants présentent différents algorithmes permettant de trouver ces neurones, ainsi que leurs performances.

IV.2 METHODE DE PROJECTION DIRECTE DES PARAMETRES DE SURFACE.

IV.2.1 Description de la méthode.

La 1ère méthode de projection proposée consiste à projeter les données sur la carte en faisant un masque sur les variables de profondeur donc seuls les paramètres de surface c'est-à-dire la SSS (1er niveau d'immersion), la SST, la SSH, la latitude et la longitude sont utilisés et on ne considère que le meilleur neurone qui capte les données. Ce neurone est celui qui minimise la distance euclidienne avec la donnée projetée en ne prenant en compte que les paramètres de surface. L'équation suivante donne comment le neurone est choisi.

$$Neur^* = \underset{k}{\operatorname{argmin}} \sqrt{\sum_{i=1}^{q} (Neur_i^k - don_i)^2}$$

Avec $Neur^k$: le neurone *k*, *don* : la donnée projetée, $i=1, ..,q=5$: les variables de surface, $Neur_i^k$ le poids du neurone *k* pour la variable i. Les variables de surface sont la SSS (1er niveau d'immersion), la SST, la SSH, la latitude et la longitude

Donc $Neur^*$, le neurone qui modélise la donnée *don* est le neurone pour lequel $d(Neur^k, don) = \sqrt{\sum_{i=1}^{q} (Neur_i^k - don_i)^2}$ est minimal.

IV.2.2 Analyse des résultats de la méthode.

Les résultats de cette première méthode d'inversion sont présentés ci-après. Cette méthode donne des résultats différents de ce qui a été présenté ci-dessus. Considérant par exemple les cardinalités, tous les référents captent des données comme le montre la Figure IV-9. Il n'y a plus de cardinalités nulles.

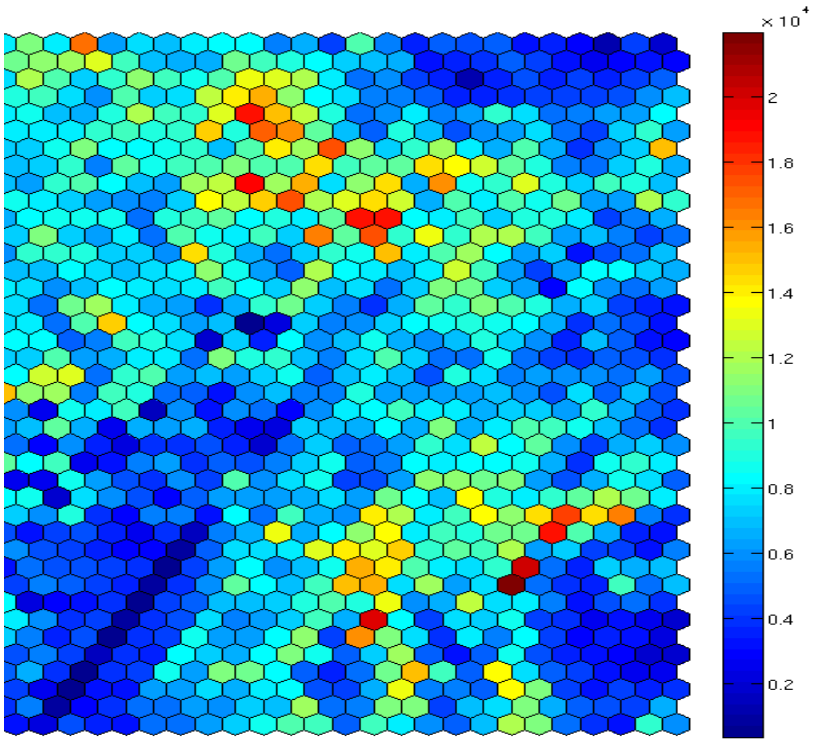
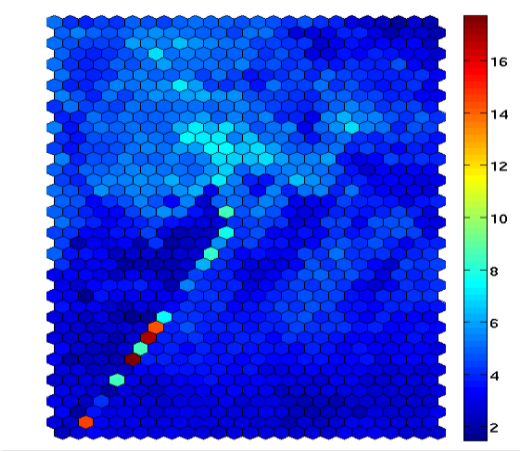
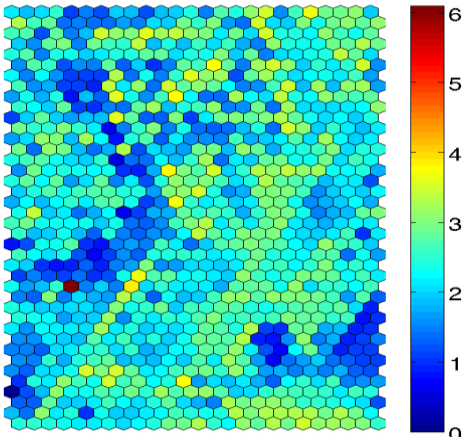


Figure IV-9:cardinalités des neurones en utilisant la première méthode d'inversion. Chaque petit hexagone représente un neurone et la couleur le nombre de données captées par le neurone lors de la phase d'apprentissage.

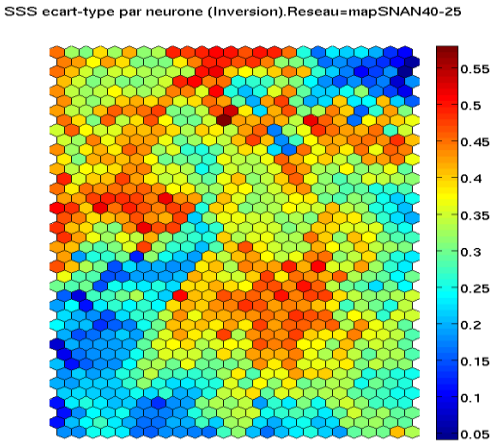
On peut voir que les cardinalités sont très différentes pour des données qui se ressemblent comme celles que nous étudions. En effet, si les données étudiées se ressemblent, la répartition des cardinalités doit être assez homogène c'est-à-dire que tous les neurones doivent capter un nombre de données proche. Ce n'est pas le cas sur cette figure. Les données captées par les neurones ne sont pas homogènes en termes de valeur avec des écarts-types importants et ce quel que soit le paramètre, comme illustré par la Figure IV-10 pour les variables de surface.



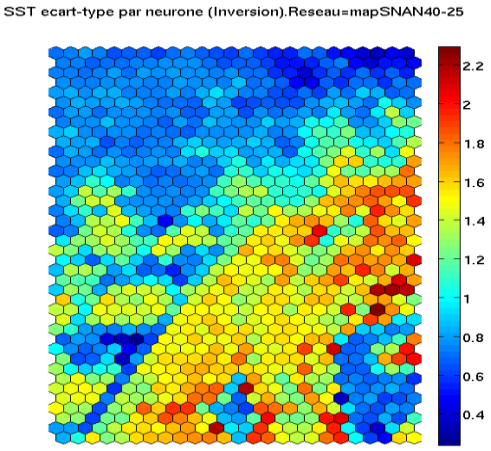
(a) Latitude en °



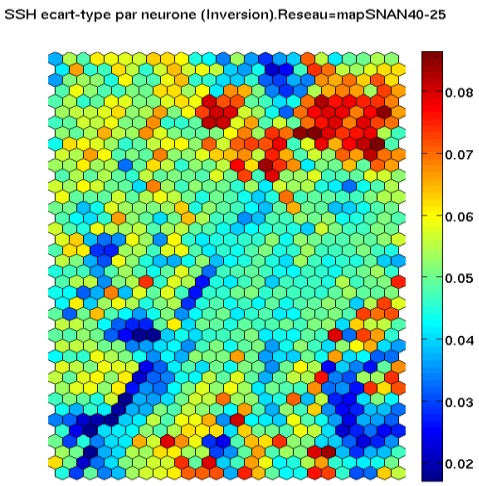
(b) Longitude en °



(c) SSS en psu



(d) SST en °C



(e) SSH en m

Figure IV-10: écarts-types des paramètres de surface.

Les différences de la variabilité moyenne sont traduites sur les écarts-types. Ces derniers sont beaucoup plus grands pour l'inversion que pour la projection normale c'est-à-dire quand toutes les variables sont projetées sur la carte comme l'ont montré les figures présentées en IV.1. Ce qui signifie que des neurones ont capté des données qui ont des valeurs différentes pour certains paramètres donc des données avec des profils qui ne sont pas toujours semblables à celui du référent du neurone lui-même

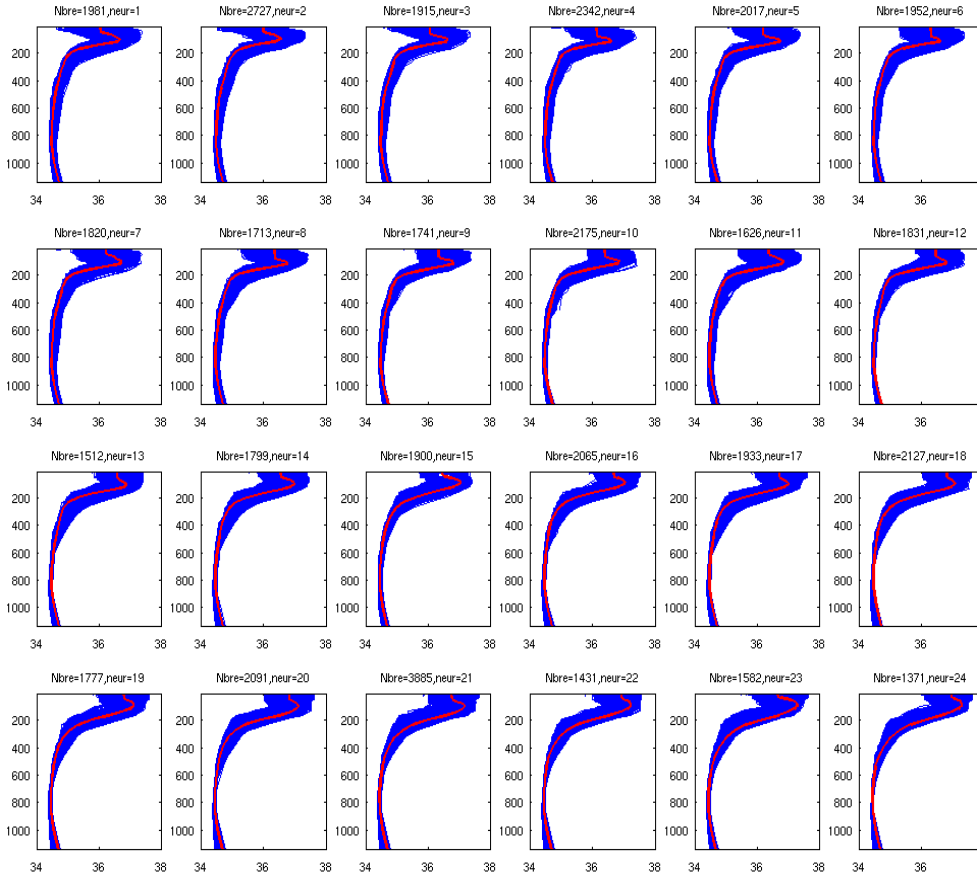


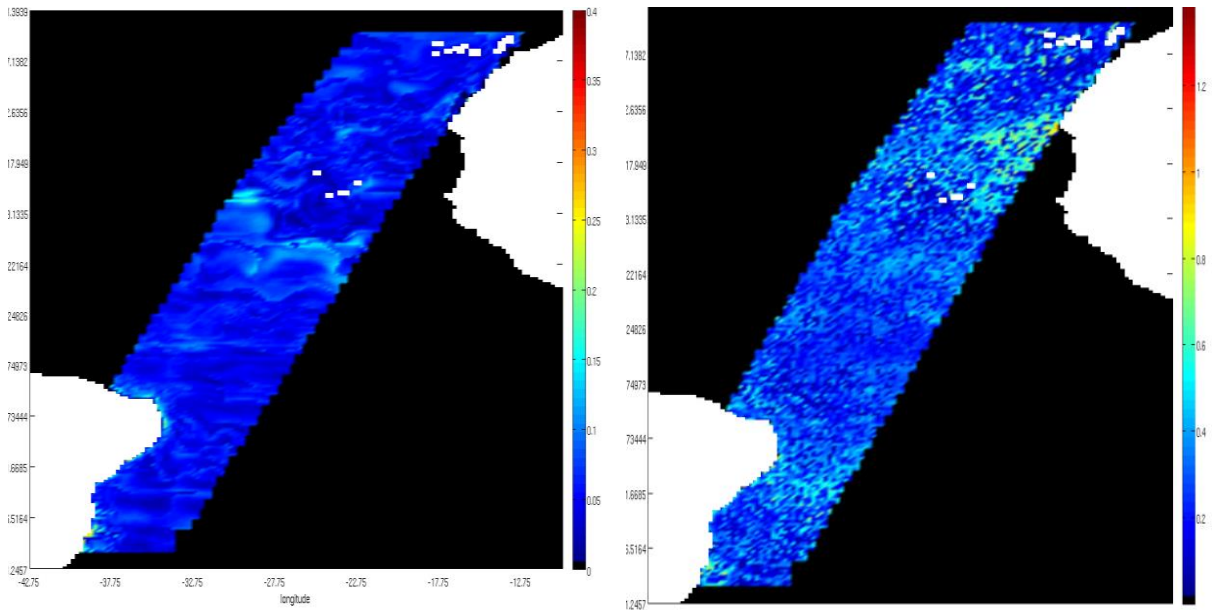
Figure IV-11: profils de quelques neurones (20 premiers) en rouge et en bleu les données captées par ces neurones (méthode d'inversion) (en abscisse la salinité en psu et en ordonnée les immersions en m).

comme l'illustre la *Figure IV-11*. Cette figure montre que les profils captés par un même neurone sont variés en comparaison de la Figure IV-3 (le faisceau autour du référent est plus large) même si les cardinalités sont très proches. Ceci est illustré également par la Figure IV-12 qui présente les racines carrées des erreurs quadratiques moyennes « root mean square errors (RMSE) » d'estimation des valeurs « réelles » de référence dans les deux cas. Dans notre cas, elle est donnée par la formule suivante :

$$\varepsilon = \sqrt{\frac{1}{p}\sum_{i=1}^{p}(x_i - y_i)^2}$$

Avec p=nombre de couches (immersions), x_i la S réelle (de référence) en une immersion i et y_i la S estimée par le modèle à la même immersion.

Cette erreur-type permet de connaitre la qualité de l'estimation. A travers son écriture on note que plus elle est petite meilleure est l'estimation. Cette quantité a été calculée aussi bien pour l'estimation donnée par la projection de tous les paramètres (a) que par la méthode d'inversion (b).



(a): rmse de la projection directe (échelle de couleur 0.001 à 0.4)

(b) : rmse de l'inversion (échelle de couleur 0.1 à 1.4)

Figure IV-12 : rmse entre profil réel et profil estimé dans les deux méthodes (a: sans masquage et b: inversion). Les échelles de couleurs ne sont pas les mêmes.

Notons que l'échelle de couleur diffère entre les 2 images de la figure. Si on prenait la même échelle pour les deux images (a et b), on ne verrait pas les structures diversifiées des rmse pour (a). Ces erreurs sont celles calculées entre la donnée (profil de S sur les 25 premières immersions) et le profil du neurone gagnant. La projection directe donne des erreurs comprises entre 0,001 et 0,4 psu alors que la projection sur les paramètres de surface donne des erreurs entre 0,1 et 1,8 psu. Ceci montre une différence entre les deux projections en termes de qualité. On en déduit aussi qu'il existe un bon neurone qui modélise à chaque fois la donnée considérée mais que le modèle d'inversion n'arrive pas à retrouver.

Ce modèle d'inversion présente beaucoup d'insuffisances qu'il convient d'étudier afin de proposer de meilleures solutions.

Pour rechercher l'origine de ces insuffisances, différentes pistes ont été explorées. Nous avons, par exemple, étudié les neurones qui présentaient les plus fortes erreurs d'estimation du profil de S ainsi que les données qu'ils modélisent. La Figure IV-13 montre quelques-uns des profils types de ces neurones et les données qu'ils ont captées.

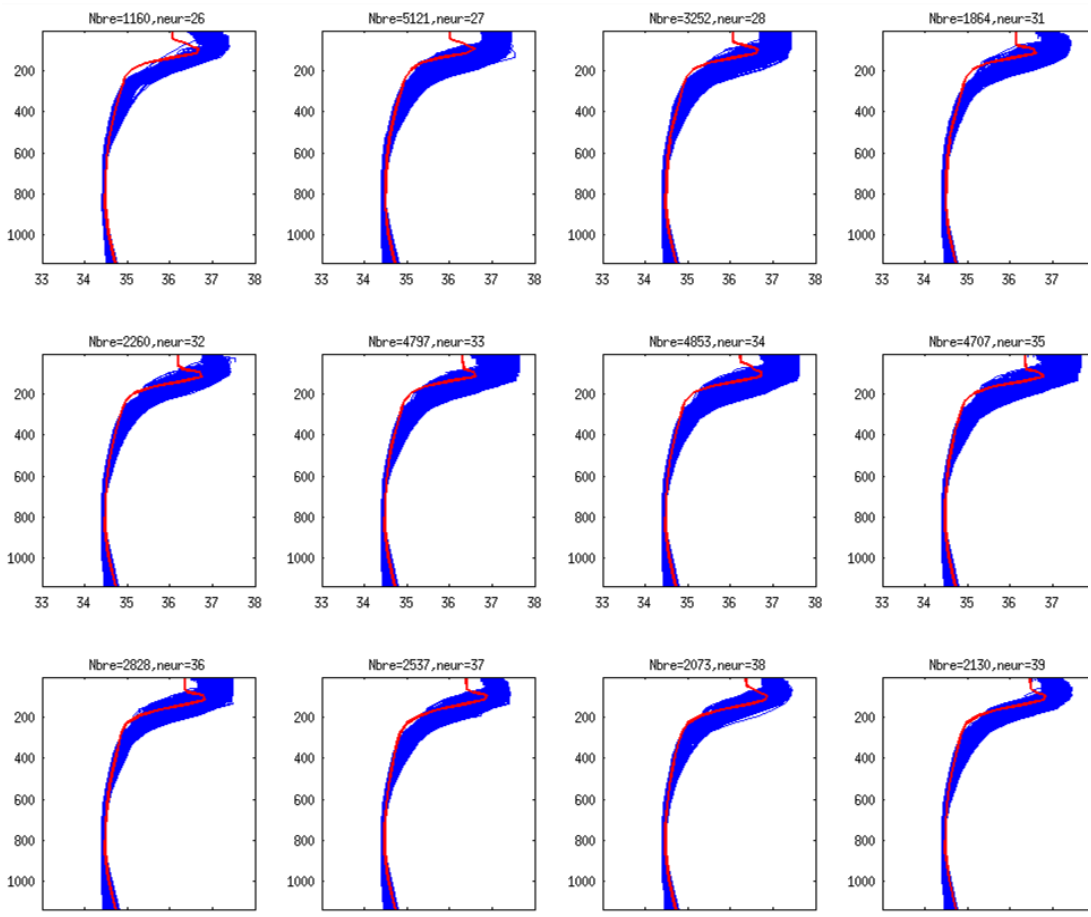


Figure IV-13 : quelques exemples de profils des référents (en rouge) et les données captées (en bleu) par les neurones de ces référents.

Dans ce groupe nous remarquons que les référents des neurones n'ont pas la même allure que les données captées sur les premières profondeurs (de 0m à environ 150m). Les neurones y présentent des valeurs en S moins importantes que les. L'analyse des vecteurs référents qui présentent des profils très différents des données captées nous a conduit à les classer en 5 groupes présentés à la Figure IV-14 suivante.

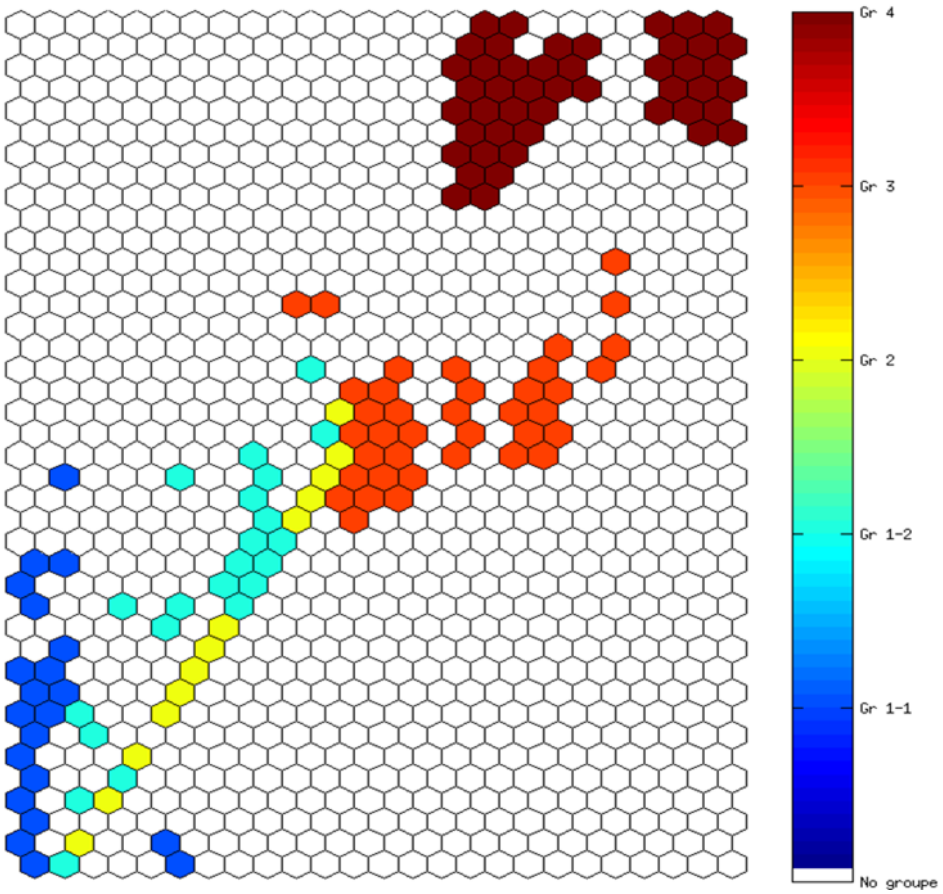


Figure IV-14: topologie des neurones qui captent des données très différentes suivant le groupe. Les neurones de même couleur appartiennent au même groupe.

La Figure IV-14 montre que les neurones d'un même groupe sont topologiquement proches (voisins). Chaque groupe a des caractéristiques bien particulières. Celles du groupe Gr1-1 ont été présentées dans la Figure IV-13. Les neurones du groupe Gr1-2 donnent une situation inverse. Cette fois, le profil du référent présente des valeurs de S plus grandes que celles des données captées par le neurone à partir de 300m jusqu'à la dernière immersion. Les neurones des groupes 3 et 4 ont capté des données ayant des S différentes entre la surface et 300m de profondeur ou de 300m à la plus basse immersion. En étudiant les données appartenant à ces neurones nous voyons une forte tendance latitudinale comme le montrent les neurones sur la carte de la Figure IV-14. Ce qui implique que la latitude est certainement très déterminante dans le choix du « bon » neurone. Cette analyse montre que les paramètres ne doivent pas être traités de la même manière car ils ont une influence différente sur le profil de S.

Pour solutionner ce problème nous avons pris en compte deux choses :

- d'abord le fait que les paramètres n'ont pas le même poids sur la détermination du profil de S. Puisque la latitude est très déterminante car elle caractérise les groupes de neurones qui modélisent des données ayant des profils différents de ceux de leur référent, on peut se demander ce qu'il en est pour les autres paramètres.
- ***Et*** le fait que certains paramètres dépendant d'autres reçoivent la même importance dans le traitement, ce qui revient à doubler le poids de l'un deux sur le profil.

Partant de ce constat, nous présentons ci-après un autre modèle d'inversion qui prend en compte ces remarques.

IV.3 METHODE DE PROJECTION SEQUENTIELLE.

Cette méthode, basée également sur le réseau construit dans la section IV.1, utilise la base de données décrite dans cette même section. Ce modèle est une amélioration du modèle d'inversion défini en IV.2. Il a pour point de départ les remarques faites au sujet de la 1ère méthode qui estime mal le profil. Pour prendre en compte les différentes remarques, nous avons réalisé une étude exploratoire des paramètres du profil de S, puis une analyse des relations qui existent entre les paramètres de surface.

IV.3.1 Étude exploratoire du profil de S.

L'analyse des données de S est présentée dans cette section. Elle concerne l'étude des dépendances entre les variables qui composent le profil de S ainsi que l'extraction de connaissances à partir de celles-ci. Les coefficients de corrélation et l'ACP sont principalement utilisés dans cette étude.

IV.3.1.1 L'interdépendance des immersions.

Cette étape permet de déterminer les variables liées (interdépendantes) et leur degré de liaison. Cette liaison est étudiée à partir du coefficient de corrélation donné par l'équation suivante pour deux variables aléatoires $X(x_1, \dots, x_N)$ et $Y(y_1, \dots, y_N)$

$$r_p = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

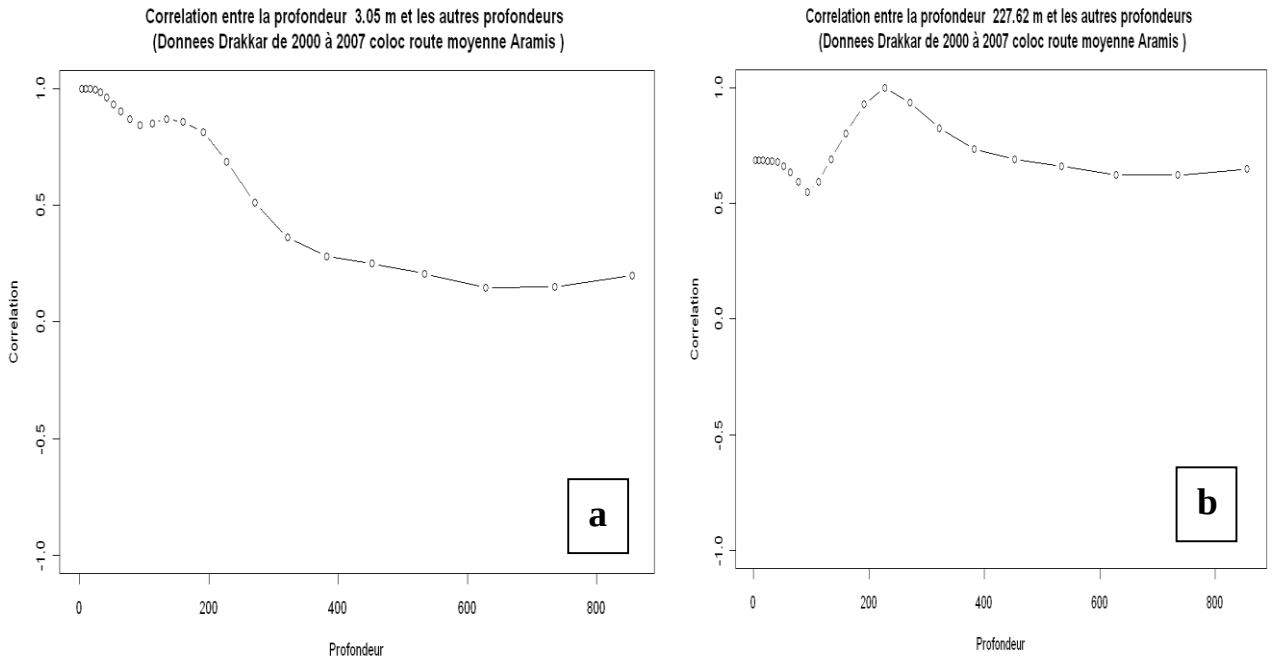
où r_p désigne le coefficient de corrélation, N : le nombre de données, $\overline{x}$ et $\overline{y}$ désignent respectivement les moyennes arithmétiques des variables X et Y.

Nous rappelons que le coefficient de corrélation entre deux variables est égal à

- 1 dans le cas où l'une des variables est fonction affine croissante de l'autre variable ;
- -1 dans le cas où la fonction affine est décroissante ;
- 0 signifie que les variables sont linéairement indépendantes.

Les valeurs intermédiaires renseignent sur le degré de dépendance linéaire entre les deux variables. Plus le coefficient est proche des valeurs extrêmes -1 ou 1, plus la corrélation entre les variables est forte ; on emploiera simplement l'expression « fortement corrélées » pour qualifier deux variables dont le coefficient de corrélation est proche de -1 ou 1.

Ce coefficient a permis d'étudier les interrelations entre les variables. Celles-ci correspondent à S à une immersion donnée. La Figure IV-15 donne la corrélation entre 3 niveaux d'immersion caractéristiques et les autres niveaux de S le long de la route moyenne ARAMIS.



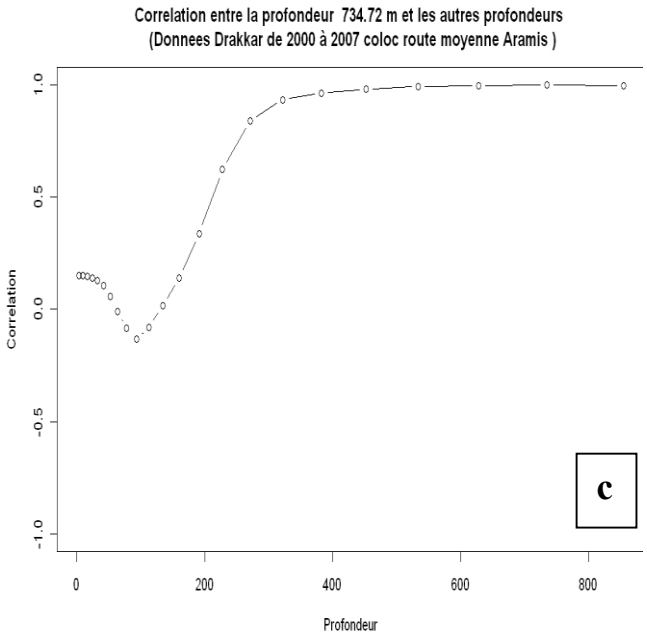


Figure IV-15: Coefficients de corrélation entre 3 niveaux d'immersion caractéristiques et les autres niveaux de S le long de la route moyenne ARAMIS. En abscisse la profondeur et en ordonnée les coefficients de corrélation. (a) corrélation entre le 1er niveau d'immersion (profondeur 3,05m) et les autres, (b) corrélation entre le 15ème niveau d'immersion (profondeur 227.6 m) et les autres, (c) corrélation entre le 22ème niveau d'immersion (profondeur 3,05m) et les autres.

Ces figures montrent une nette différence entre les niveaux allant de la surface à 192m de profondeur (groupe 1), de 192 à 322m (groupe 2) et les autres niveaux allant de 322 à 855m de profondeur (groupe 3). Elles montrent des coefficients de corrélation de fortes valeurs avoisinant 1 entre S à la surface et aux autres immersions du groupe 1. Puis, elles diminuent avec un fort gradient à partir de 192m jusqu'à environ 380m. A partir de cette profondeur les corrélations se stabilisent avec des coefficients proches de 0 jusqu'au dernier niveau considéré. On en déduit que les profondeurs du groupe 1 ont des S fortement liées avec celles de la surface. Puisque cette courbe est quasi-identique pour toutes les immersions du groupe 1, on peut dire que les S de toutes ces immersions sont interdépendantes. Ces niveaux d'immersion appartiennent tous à la première couche homogène du profil de S. En revanche, S de surface et S de profondeur offrent des variations différentes qui se traduisent par des corrélations presque nulles. La Figure IV-15c donne cette fois la corrélation entre S à 734m et les autres niveaux. On note une

forte dépendance entre les grandes profondeurs qui constituent également une couche homogène en dessous de l'halocline. En surface, en revanche la corrélation est presque nulle. Enfin la Figure IV-15b présente les corrélations entre la S à 227.62m et les autres niveaux. Elle montre une situation intermédiaire avec des corrélations qui avoisinent 0,5 de 3,05m à 160m puis de 380m jusqu'au fond. Nous pouvons noter que cette figure donne des coefficients de corrélation supérieurs à 0,8 pour les profondeurs comprises entre 160 et 322m. Seule cette immersion présente ces forts coefficients dans le groupe 3, contrairement au groupe 1 et 2 où nous avons des interdépendances entre les S de toutes les immersions, c'est-à-dire de fortes les corrélations à l'intérieur du groupe.

Cette analyse d'interrelation entre les variables a permis de voir les liens entre les immersions. Elle a également montré un aspect très important pour la suite de notre travail : le fort taux de corrélation entre la S de surface (SSS) et les immersions du groupe 1. Le but final de cette thèse étant de reconstruire les profils de S à partir de la surface, cet aspect est exploité. Cependant, pour les groupes 2 et 3, la relation avec la surface n'est pas directe d'où la nécessité d'approfondir cette exploration des données. Ainsi, nous avons utilisé l'ACP.

IV.3.1.2 L'Analyse en Composantes Principales et la pertinence des variables.

Dans le cas présent, les variables initiales correspondent à S aux 23 premières immersions définies sur le *Tableau II-1*. Les données ont été centrées et réduites afin d'éliminer les biais pouvant être introduits par l'ordre de grandeur des valeurs et la dispersion des valeurs pour une variable quelconque.

Les parts de variance expliquées par les 10 premières composantes principales (CP) obtenues après transformation sont représentées dans la Figure IV-16.

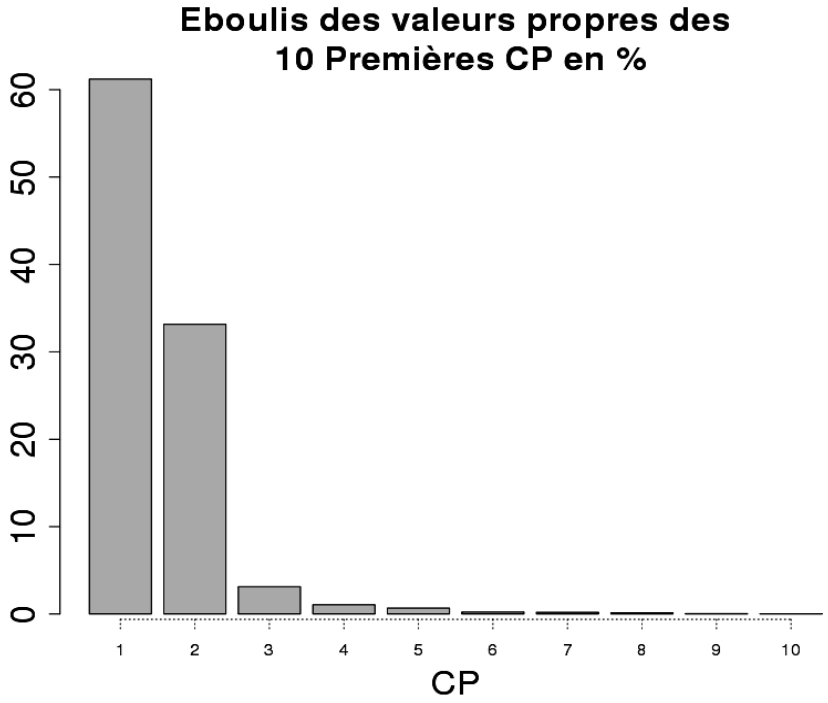


Figure IV-16: % de la variance totale expliquée par les 10 premières Composantes Principales (CP). Les hauteurs des barres grises représentent les pourcentages de variance

Le Tableau IV-1 donne le pourcentage de variance expliquée par chacune des 7 CP les plus importantes.

Tableau IV-1: pourcentage de variance expliquée par CP et leur cumul.

CP	1ère CP	2ème	3ème				7ème
Part	0,62600390	0,32022460	0,02962717	0,01037494	0,006452978	0,002280308	0,00183590
Cumul	0,62600390	0,94622850	0,97585567	0,98623061	0,992683588	0,994963896	0,996799796

Pour rappel l'analyse en composantes principales a pour but de trouver un espace de dimension inférieure à la dimension de l'espace d'origine. Ici la dimension initiale est de 23 autrement dit chaque individu à 23 coordonnées. On souhaite également que les observations puissent être visualisables dans cet espace de nouvelle dimension. Ceci implique, souvent une dimension idéale de 3 ou 2 car au-delàs la visualisation sur une seule figure reste impossible. Dans notre cas (cf. : Tableau IV-1) les 2 premières CP expliquent à elles seules plus de 94 % de la variance et sont supérieures à 1 (valeur seuil fixée pour valider l'importance d'une CP car l'ACP est normée dans notre cas). Les 2

premières CP représentent bien les 23 variables initiales avec une perte d'informations acceptable (5% de l'information totale).

Les 2 premières CP représentent bien les données, donc le repère ayant pour axes ces 2 CP permet une bonne visualisation.

La Figure IV-17 illustre les observations et les variables initiales sur la nouvelle base ayant comme axes les 2 premières CP.

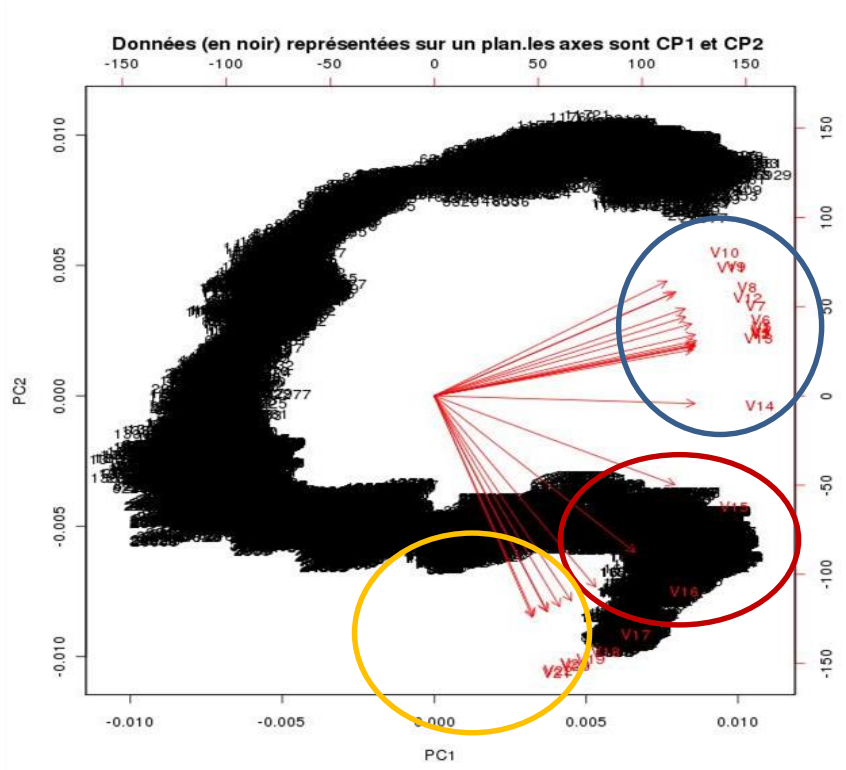


Figure IV-17: représentation des observations (en noir) et variables d'origine, S aux différents niveaux d'immersion, (en flèches rouges) sur le nouveau repère dont les axes sont les 2 premières CP. Les valeurs à l'abscisse inférieure et l'ordonnée de gauche concernent les variables, les valeurs supérieures et de droite concernent les observations. V1, V2, ..., V23 indiquent les niveaux d'immersion. Les cercles regroupent les V en 3 groupes : gr1 en bleu, gr2 en rouge et gr3 en jaune.

Elle montre les relations entre CP1, CP2 et les variables d'origine (S aux différents niveaux d'immersion, en flèche rouge) sur le nouveau repère. Elle illustre la qualité de la représentation des variables sur le plan factoriel. Une variable sera d'autant mieux représentée sur un axe que sa corrélation avec la composante principale correspondante est en valeur absolue proche de 1. En effet, le coefficient de corrélation entre une ancienne

variable et une nouvelle variable n'est autre que le cosinus de l'angle du vecteur joignant l'origine au point représentant la variable sur l'axe avec cet axe. Par rapport à notre cas, on peut observer 3 différents groupes de flèches rouges mis en évidence par des cercles sur la figure. Un premier groupe (gr1) constitué par les niveaux 1 à 14 correspondant aux profondeurs allant de la surface à 192m. Ces variables ont des cosinus proches de 1 avec le premier axe principal (l'axe des abscisses défini par la 1ère CP). Un deuxième groupe (gr2) constitué par les niveaux 15 à 17 correspondant aux profondeurs allant de 192m à 322m. Ces variables ont des cosinus plus faibles que les variables du gr1 par rapport à l'axe des abscisses mais plus forts par rapport à l'axe des ordonnées. Cette tendance est inversée entre les variables du gr2 et celles du gr3 qui présentent quant à elles de forts cosinus avec la deuxième CP, en valeur absolue, et de faibles cosinus avec la première CP. Ce découpage renforce l'analyse faite sur l'interdépendance des S présentée dans le paragraphe IV.3.1.1 qui présente le degré de liaison des variables dans chaque groupe.

Après cette analyse en CP, il faut traduire le résultat sur les données que nous étudions, c'est-à-dire revenir à la base d'origine dans laquelle les axes correspondent au S à différentes immersions.

La Figure IV-17 illustre un espace ayant comme axe les deux CPs retenues et on y a représenté les observations et les variables d'origine. Pour lire les nouvelles valeurs des observations dans ce nouveau repère, il faut considérer les chiffres de droite et du haut. Pour mieux analyser ces résultats, nous allons travailler sur les variables puis sur les individus. La méthode la plus naturelle pour donner une signification à une composante principale est de la relier aux variables initiales et de calculer les coefficients de corrélation linéaire et en s'intéressant aux plus forts coefficients en valeur absolue. Mais cette méthode doit être renforcée par les coefficients de contribution. La Figure IV-18 présente les coefficients de corrélation en valeurs absolue entre les CP 1 et 2 et les variables d'origine.

Nous voyons que les niveaux d'immersion de 0 à 192m sont anti-corrélés (coefficients de corrélation négative, figure non représentée) à la CP1 et non-corrélés à la CP2 mais les niveaux plus profonds allant de 380m à 800m c'est-à-dire sont corrélés à la 2ème CP et non-corrélés à la CP1.

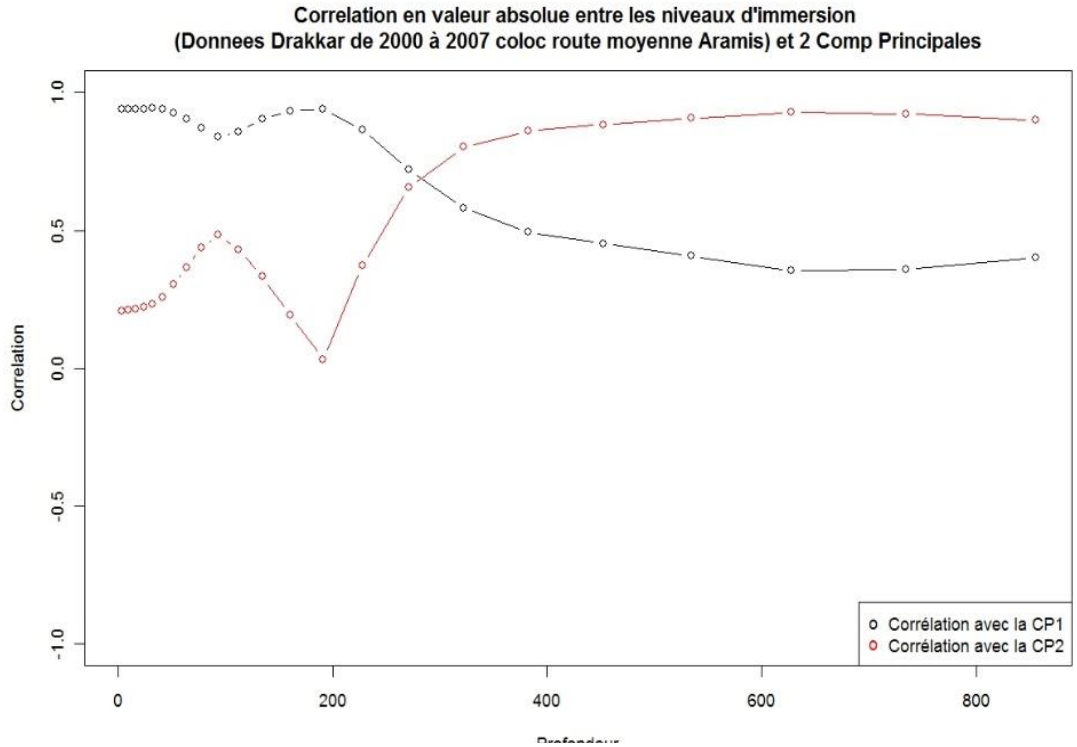


Figure IV-18: corrélation en valeur absolue entre les variables et les 2 premières Composantes. En noir avec la 1ère et en rouge avec la 2ème.

Cependant il faut trouver quelles sont les variables initiales qui ont le plus participé à la construction de ces deux CP dans la base d'origine (ancienne base) constituée de 23 axes car cette contribution donne plus de sens aux corrélations présentées à la Figure IV-18. Nous avons considéré le coefficient de contribution déjà défini en III.1.1. Le Tableau IV-2 et la Figure IV-19 résument les contributions de chaque variable sur les 2 premières CP.

Tableau IV-2: contributions des variables initiales sur les 2 premières CP.

Variables	Contribution à la		Supérieure moyenne contrib	
	CP1	CP2	CP1	CP2
V1 (3, 05m)	0,061362375	0,00589166	1	0
V2	0,061373892	0,00597204	1	0
V3	0,061392503	0,00618485	1	0

V4	0,061498603	0,0066026	1	0
V5	0,061895372	0,00729802	1	0
V6	0,061428576	0,00897001	1	0
V7	0,059697956	0,01242393	1	0
V8	0,056680266	0,01819513	1	0
V9	0,0526604	0,02592795	1	0
V10	0,048743978	0,03160865	1	0
V11	0,051223024	0,02510006	1	0
V12	0,056946807	0,01494933	1	0
V13	0,060465698	0,00512198	1	0
V14	0,061292196	0,00014045	1	0
V15	0,052143137	0,01891477	1	0
V16	0,036220563	0,05839842	0	1
V17	0,023476881	0,08736049	0	1
V18	0,016923	0,10039532	0	1
V19	0,014180404	0,10615193	0	1
V20	0,011563577	0,11160644	0	1
V21	0,008661597	0,11699809	0	1
V22	0,008910866	0,11575326	0	1
V23 (800m)	0,011258329	0,11003464	0	1
Moyenne	**0,043478261**	**0,04347826**		

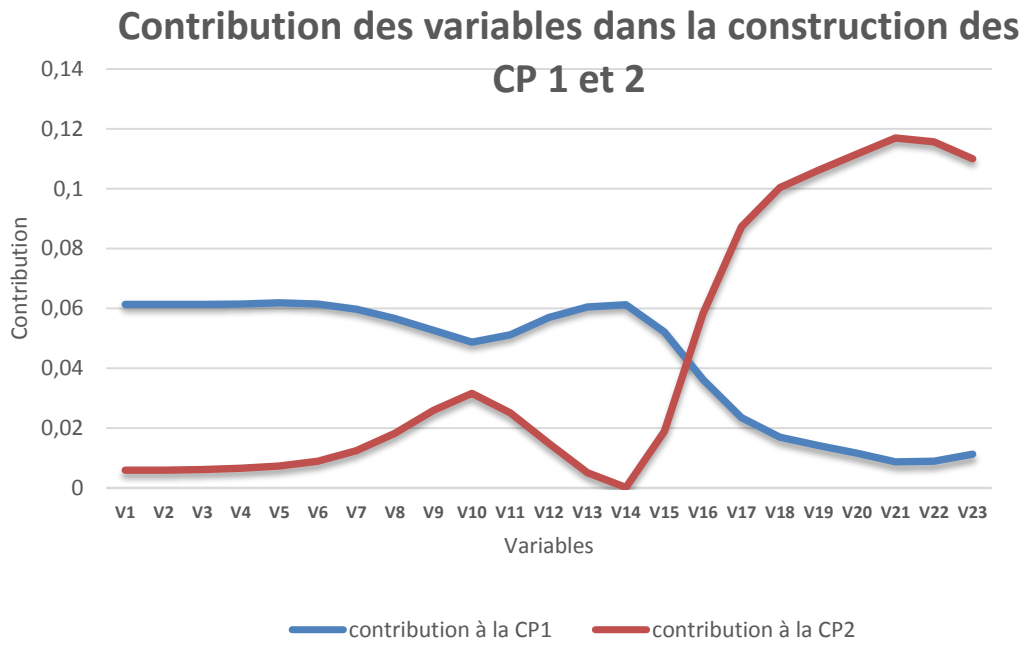


Figure IV-19: contribution des variables dans la construction des 2 premières CP

La Figure IV-19 montre la contribution de chaque variable sur la construction des 2 premières CP. Les variables 1 à 14 -les profondeurs de 3,05 à 191m- ont les plus grandes contributions. La plus grande contribution est celle de la variable 5 (32,21m) qui est approximativement de 0,06. Cette valeur est toutefois très proche de celle des variables représentant les niveaux d'immersion allant de la surface à 191 m de profondeur. Ces variables, très corrélées ont les plus faibles contributions à la CP2. Les contributions sur la CP2 montre un contraste avec d'une part des variables avec de grandes contributions (les niveaux les plus profonds) allant jusqu'à 0,116998085 et de très faibles contributions allant jusqu'à 0,005891663 soit 20 fois moins que la plus grande contribution.

Dans la zone qui concerne les variables 14 à 18 où les deux courbes de contribution se coupent, on note un fort gradient (descendant pour la CP1 et ascendant pour la CP2). Cette zone très particulière est également marquée dans le profil de S par une forte variation, elle rappelle ce que nous avions avec l'halocline.

Pour étudier la place et l'importance des individus par rapport à ces deux CP, les données ont été projetées sur le plan principal, comme illustré par la Figure IV-20 qui montre les valeurs des individus sur les axes principaux obtenus après l'ACP. Nous avons cherché à comprendre cette structure, en reliant les CP aux informations géographiques. Nous

avons trouvé que les variations de la CP2 sont fortement liées à la latitude comme le montre la figure ci-dessous.

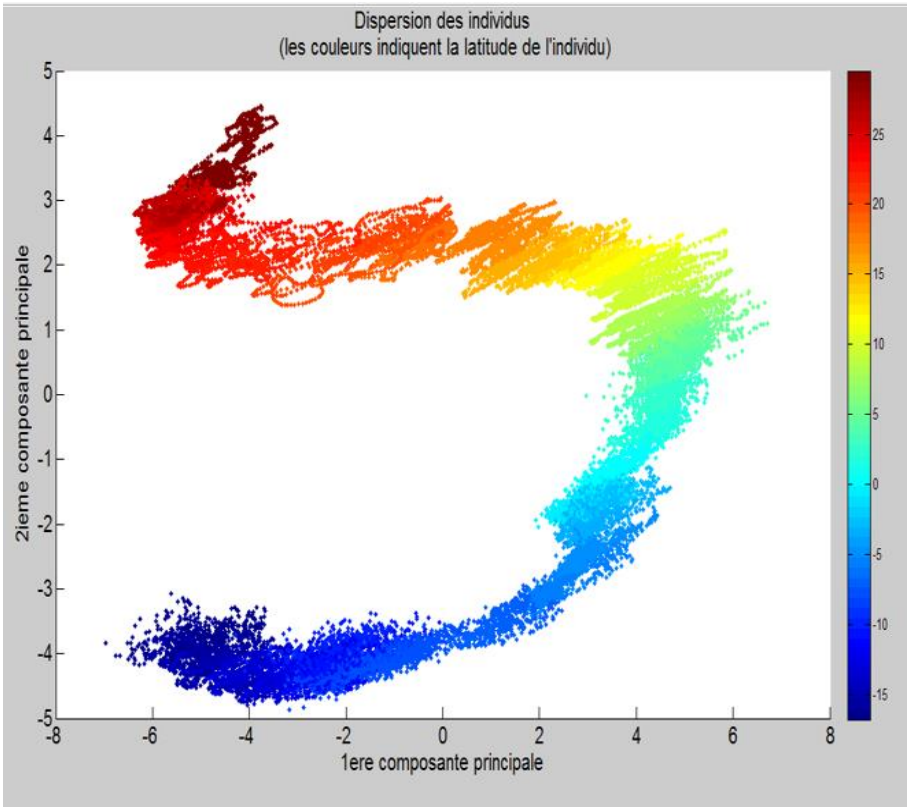


Figure IV-20: Nuage des points sur les axes principaux avec latitude (la couleur indique la latitude des individus).

On a un coefficient de corrélation de 0,94 entre la latitude et la CP2, ce qui est un résultat assez intéressant dans la mesure où la latitude est une mesure de surface. Par rapport à l'objectif de la thèse qui est de reconstruire le profil de S à partir des mesures de surfaces, on pourra exploiter ces résultats. Nous avons vu que les 2 premières CP expliquent plus 94% de la variance totale obtenue après l'ACP. Le fait que la 1^{ière} CP soit fortement corrélée avec la S de surface, est un atout fort pour reconstruire les S dépendant de cette CP soit les S de la surface à 192 de profondeur. Ce qui n'est pas le cas de la CP2 qui est corrélée aux S des eaux plus profondes. Or nous venons de voir que cette CP est très liée à la latitude.

Cette analyse a permis de mieux comprendre, numériquement les relations entre les S aux différentes immersions.

IV.3.2 Étude préalable des paramètres de surface.

Cette étude fait appel à deux approches, la première basée sur les corrélations, c'est-à-dire sur l'interdépendance des paramètres de surface et la deuxième sur la capacité de ces derniers à exprimer la variabilité du profil.

IV.3.2.1 L'interdépendance des paramètres de surface.

Cette étape permet de déterminer les variables de surface liées. Cette liaison est étudiée à partir du coefficient de corrélation présenté en IV.3.1.1. La Figure IV-21 suivante présente les coefficients pour chaque couple de variable.

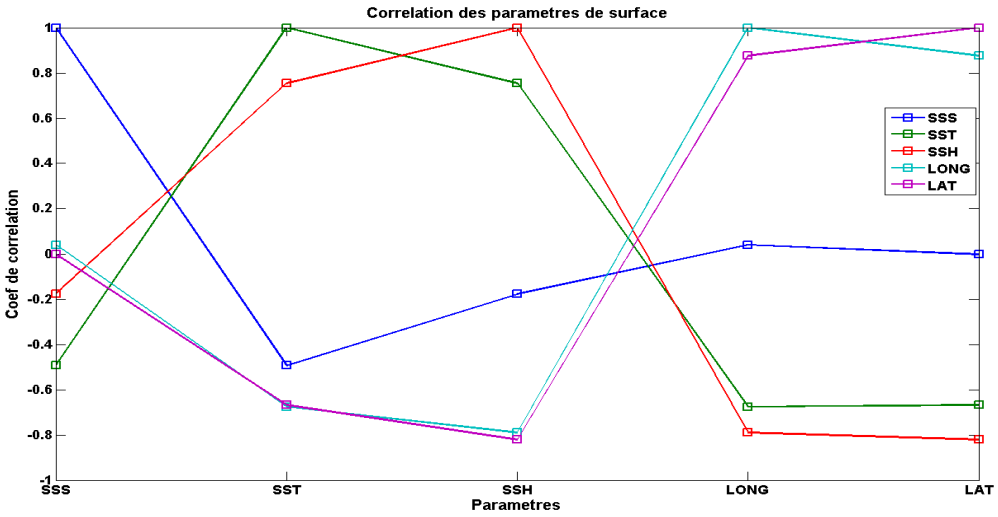


Figure IV-21:Coefficient de corrélation en fonction des paramètres de surface par couple de variables. En ordonnée nous avons le coefficient de corrélation entre la variable considérée (en abscisse) et les autres variables de la courbe. Par exemple pour la SST (abscisse), les petits carrés sur la même verticale indiquent la corrélation avec les autres variables en fonction de la couleur de la courbe. Ainsi, la SST a un coefficient de corrélation de 0.7 avec la latitude et la longitude respectivement en mauve et cyan.

Avec des coefficients variant de -0.5 à 0, la figure montre que la SSS est linéairement faiblement corrélée aux autres paramètres de surface. Alors que les 4 autres variables sont fortement corrélées (ou anti-corrélées) entre elles. On note également que la latitude est fortement corrélée à toutes les autres variables exceptée la SSS. Ce qui permet un premier regroupement basé sur la corrélation qui est un signe de redondance (Guyon et al., 2003). Ce regroupement concerne d'une part la SSS et d'autre part la latitude, la longitude, la

SST et la SSH. Donc une grande partie de l'information que comportent la longitude, la SST et la SSH est donnée aussi par la latitude. Cela signifie aussi que ces paramètres considérés un à un décrivent presque de la même manière la variabilité incluse dans les 4 pris ensemble. Ainsi, il est important de créer un nouvel espace où cette interdépendance sera supprimée et dans lequel nous aurons de nouvelles variables indépendantes et explicatives de ce phénomène. Cette étape qui correspond à une extraction de paramètres suivie d'une sélection de paramètres est décrite dans le paragraphe suivant.

IV.3.2.2 Extraction/sélection de paramètres pertinents avec l'Analyse en Composantes Principales (ACP).

Les techniques d'extraction de paramètres cherchent à transformer l'espace original des variables vers un nouvel espace habituellement de dimension plus petite et décrivant mieux les paramètres. Cette transformation peut être une projection (comme l'ACP) ou une compression basée sur la théorie de l'information. La meilleure méthode linéaire d'extraction de paramètres connue à ce jour est l'ACP (Mardia, et al., 1980).

L'ACP, décrite en III.1.1, est une méthode de la famille de l'analyse des données et plus généralement de la statistique multivariée, qui consiste à transformer des variables liées entre elles (dites « corrélées») en nouvelles variables décorrélées les unes des autres. L'ACP est faite sur 1/10 des données centrées et réduites (comme cela a été le cas dans la mise en place de la SOM). Cette opération de centrage-réduction est justifiée par le fait que les variables n'ont pas le même ordre de grandeur. Par exemple la SSS varie de 32 à 38 alors que la latitude varie de -20 à 30. Les résultats obtenus après l'ACP sont présentés ci-après :

Tableau IV-3:Pourcentage de variance expliqué par Composante Principale (CP)

Comp. Princ	% variance expl.	Cumul
CP1	66.9374	66.9374
CP2	23.1239	90.0613
CP3	4.1274	94.1887
CP4	3.4704	97.6591
CP5	2.3409	100.0000

Nous savons que l'ACP permet de réduire le nombre de variables donc la dimension, cependant cette réduction doit respecter une contrainte qui est de minimiser la perte d'information sur les observations. Dans notre cas les 2 premières CP expliquent plus de 90 % de la variance totale donc on pourra les retenir, les raisons de ce choix étant explicitées dans la section III.1.1. Ce qui implique que ces deux CP expliquent bien le phénomène et sont décorrelées, par définition de l'ACP.

Il faut maintenant qu'à partir de cette ACP nous puissions revenir à la base de départ qui comporte les variables utilisées dans notre étude. Pour ce faire, analysons les corrélations entre les CP1, CP2 et les variables d'origine comme l'indique le tableau suivant :

Tableau IV-4:Coefficients de corrélation entre CP et paramètres de surface.

	SSS	SST	SSH	Long	Lat
CP1	0.2685	-0.8746	-0.9254	0.9075	0.9109
CP2	-0.9430	0.3412	-0.0665	0.2477	0.2912
CP3	0.1279	0.1211	0.2968	0.2761	0.1051

Le Tableau IV-4 présente les 3 premières CP et leur coefficient de corrélation avec chacun des paramètres de surface. Pour la CP1, la SSH et la latitude présentent les plus grandes corrélations en valeur absolue, presque égales à 1. Ceci veut dire que ces 2 paramètres sont très déterminants dans la construction de cette CP et comme la latitude est très discriminante par rapport aux neurones et présente de meilleurs coefficients de corrélation avec les autres paramètres (c.f. IV.3.2.1) nous allons la choisir. En considérant le même raisonnement, la SSS est la variable qui explique le mieux la CP2. Ainsi la latitude et la SSS sont les variables retenues comme plus importantes dans la détermination et l'expression de la variabilité des profils de S. Ces 2 variables sont les moins dépendantes entre-elles car ont un coefficient de corrélation faible. Elles sont très corrélées aux CP1 et CP2 donc expliquent une part importante de la variabilité des données aussi bien de surface que de profondeur. Ce choix permet de lever les principales causes d'insuffisance du 1er modèle d'inversion.

IV.3.3 Algorithme utilisé pour la deuxième méthode d'inversion (INV2).

L'algorithme utilisé par ce deuxième modèle statistique de reconstruction de profils de S prend en compte les résultats de l'étude préalable. Ainsi, tous les paramètres de surface ne sont plus considérés avec la même importance dans leur projection sur la carte. L'algorithme proposé ici est constitué principalement de 3 étapes:

étape 1 : projection initiale

On fait une 1ère projection (Proj1) avec le masquage de tous les paramètres sauf la SSS et la latitude, ces variables étant celles retenues par l'étude préalable. A la suite de cette projection nous considérons les 25 meilleurs référents pour chaque donnée (individus), c'est-à-dire les 25 référents les plus proches en termes de distance euclidienne de la donnée si on ne considère que la SSS et la Lat. Donc les 25 neurones qui minimisent l'équation :

$$d(Neur^k, don) = \sqrt{\sum_{i=1}^{q} (Neur_i^k - don_i)^2}$$

Avec $Neur^k$: le neurone *k*, *don* : la donnée projetée, *i=1, ..,q=2* : (1er niveau d'immersion), et latitude , $Neur_i^k$ le poids du neurone *k* pour la variable *i*.

Étape 2 : 2ième projection

On fait une 2ième projection (Proj2) en masquant toutes les variables sauf la SSH, la longitude et la SST. Cette fois-ci on considère les 40 meilleurs référents pour chaque individu.

Le choix de 40 ou de 25 est basé sur la topologie de la carte. Une suite intéressante de notre travail consisterait à définir une méthode pertinente de détermination de ces nombres.

Étape 3 : recherche du neurone gagnant

Après les étapes 1 et 2, on cherche, pour chaque individu, si le meilleur référent (R1Proj1) de Proj1 appartient à l'ensemble (EProj2) des 40 référents sélectionnés en Proj2. Si oui, on le prend comme neurone gagnant, et c'est le neurone de ce référent qui va modéliser l'individu. Sinon on refait la même chose avec le deuxième meilleur référent (R2Proj1)

dans Proj1. Si tous les référents de Proj1 sont testés et qu'aucun d'entre eux n'appartient à EProj2, on conclut qu'il n'existe pas de bon référent qui peut modéliser cette donnée selon les critères cités.

Le schéma suivant explique les étapes de l'inversion :

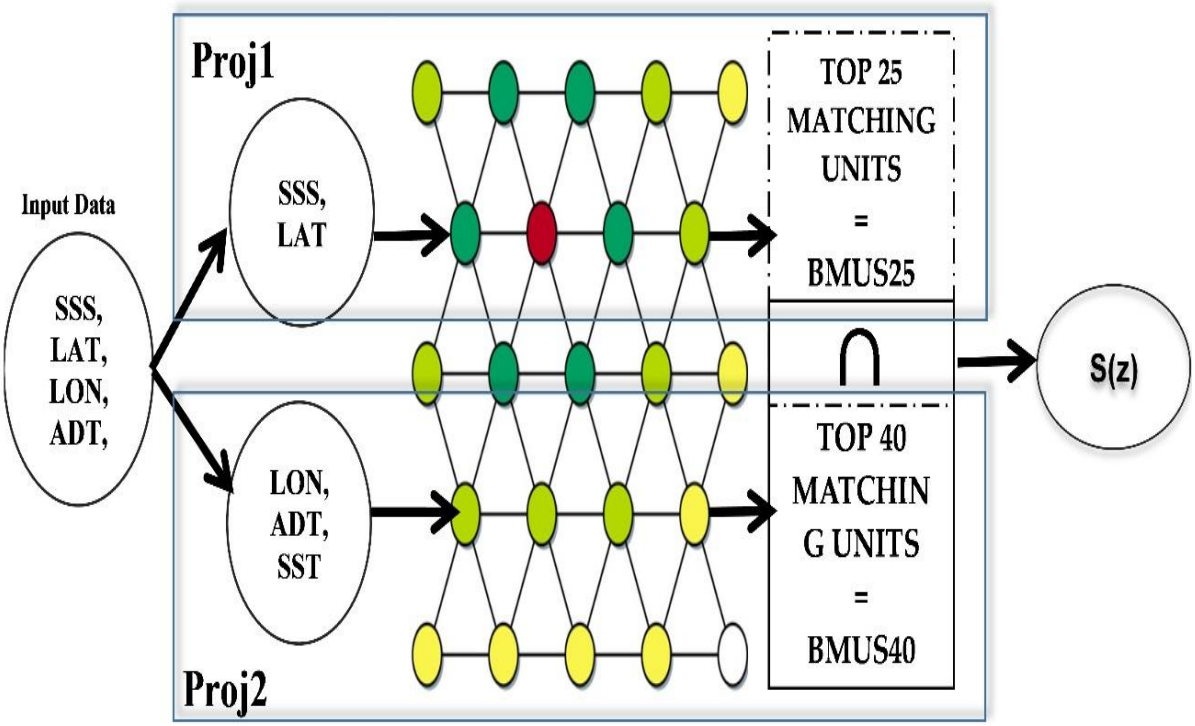


Figure IV-22: Schéma de la 2ème méthode d'inversion. Les ellipses correspondent aux données (Surface en entrée et profil en sortie), les rectangles correspondent aux ensembles de neurones sélectionnés après chaque projection. Les cercles en couleur correspondent à la SOM.

L'algorithme de projection est le suivant:

- Projection initiale :

  - *Masquage des paramètres sauf SSS et Lat sur la couche d'entrée « de présentation ».*
  - *Initialisation de la matrice des neurones gagnants à 0 (Bm1, taille =n (nombre de données à inverser) x 25 (nombre de neurones à conserver)).*
  - *Pour chaque donnée faire*
    Présentation de la SSS et de Lat à la couche d'entrée.
    Recherche des 25 neurones les plus proches en distance euclidienne.
    Affectation des numéros de neurones à la ligne de Bm1 correspondante

  Fin pour.

- 2nde Projection :

- *Masquage des paramètres sauf SSH, SST et Long sur la couche.*
- *Initialisation de la matrice des neurones gagnants à 0 (Bm2, taille= n x 40).*
- *Pour chaque donnée faire*
  Présentation de la SSH, SST et Long à la couche d'entrée.
  Recherche des 40 neurones les plus proches en distance euclidienne.
  Affectation des numéros de neurones à la ligne de Bm2 correspondante
  Fin pour.

- Recherche du neurone gagnant :

- *Initialisation du vecteur des neurones gagnants à 0 (BmFinal, taille= n).*
- *Pour chaque ligne i de Bm1 faire*
  a) *Rechercher le 1er élément Bm1[i,1] dans la ligne correspondante de Bm2*
  Si trouvé alors
  BmFinal[i]= Bm1[i,1]
  Sinon passer à l'élément suivant de la ligne considérée.
  b) *Répéter a) jusqu'à trouver l'élément recherché de la ligne dans Bm2*
  Si l'élément n'est pas trouvé alors BmFinal[i]=0
  Fin pour.

Le résultat est unique et stable mais l'existence de solution n'est pas toujours garantie. Le nombre de données non modélisées est de 40191 soit 0.93% de la base totale, une part que nous jugeons négligeable. Nous nous sommes intéressés à l'analyse de ces données non modélisées pour essayer de trouver des caractéristiques qui leur sont communes. Mais, il ne ressort aucune caractéristique particulièrement exploitable de ces données par rapport aux paramètres de surface comme l'illustre la figure suivante :

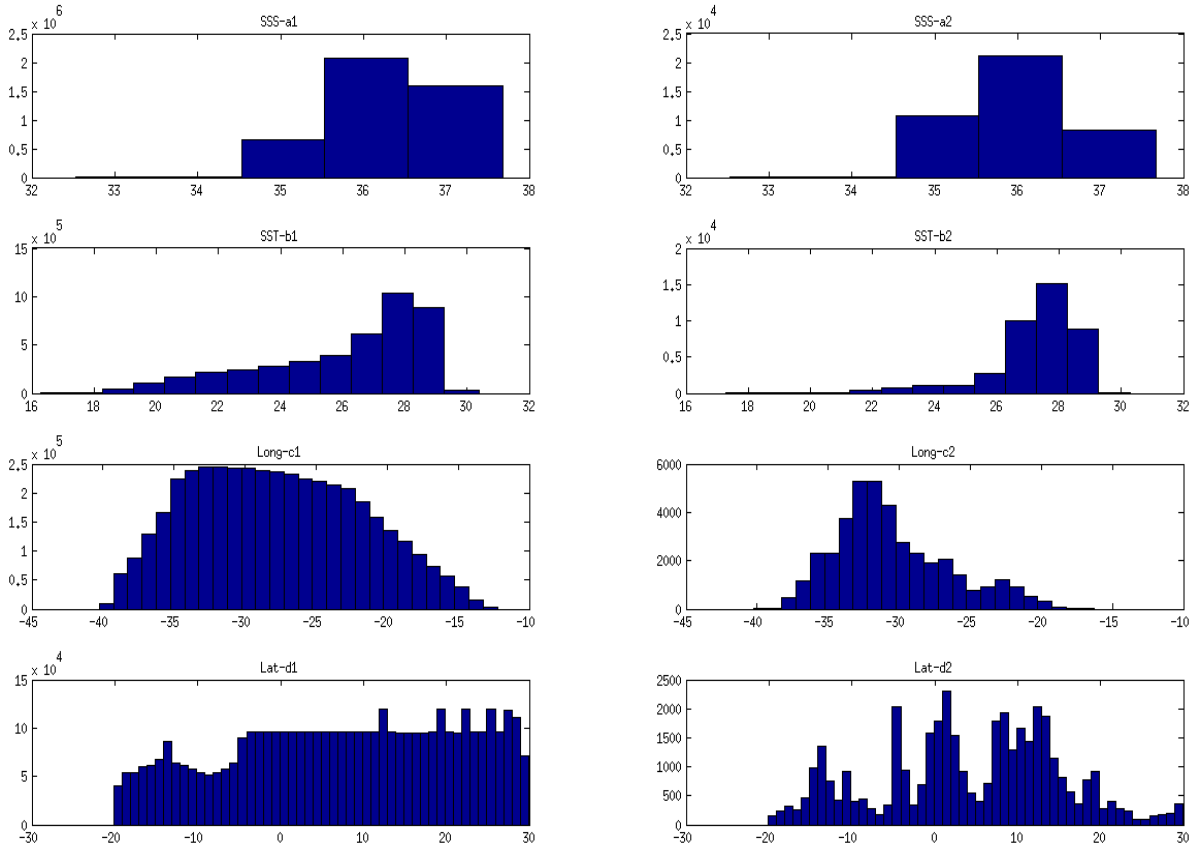


Figure IV-23 : comparaison de la distribution des données non modélisées (1, à droite) et celle des données de la base de données totale (2, à gauche). Chaque image représente un paramètre de surface, de haut en bas on a la (a) SSS, (b) la SST, (c) la longitude et (d) la latitude.

La Figure IV-23 montre les mêmes plages de valeurs aussi bien pour la base totale que pour les données non modélisées pour la SSS et la SST (images a et b), tandis pour la latitude et la longitude on peut remarquer que les données non modélisées sont localisées principalement autour de 15° (N et S), de 0°, et entre 30 et 35°W.

IV.3.4 Le 2ième modèle d'inversion : résultats.

Nous allons analyser ces résultats en termes de caractéristiques de données et de neurones. Commençons d'abord par les profils des neurones et ceux des données qu'ils ont captées. La Figure IV-24 présente les profils référents des neurones et ceux de leurs données captées, comme déjà vu. Les titres indiquent le numéro du neurone et sa cardinalité, c'est-à-dire le nombre de données qu'il modélise. Nous remarquons que les données suivent

bien le profil de leur neurone gagnant. Et malgré le nombre important de données captées le faisceau n'est pas très large, ce qui signifie que la variance n'est pas très grande par rapport à la première méthode d'inversion.

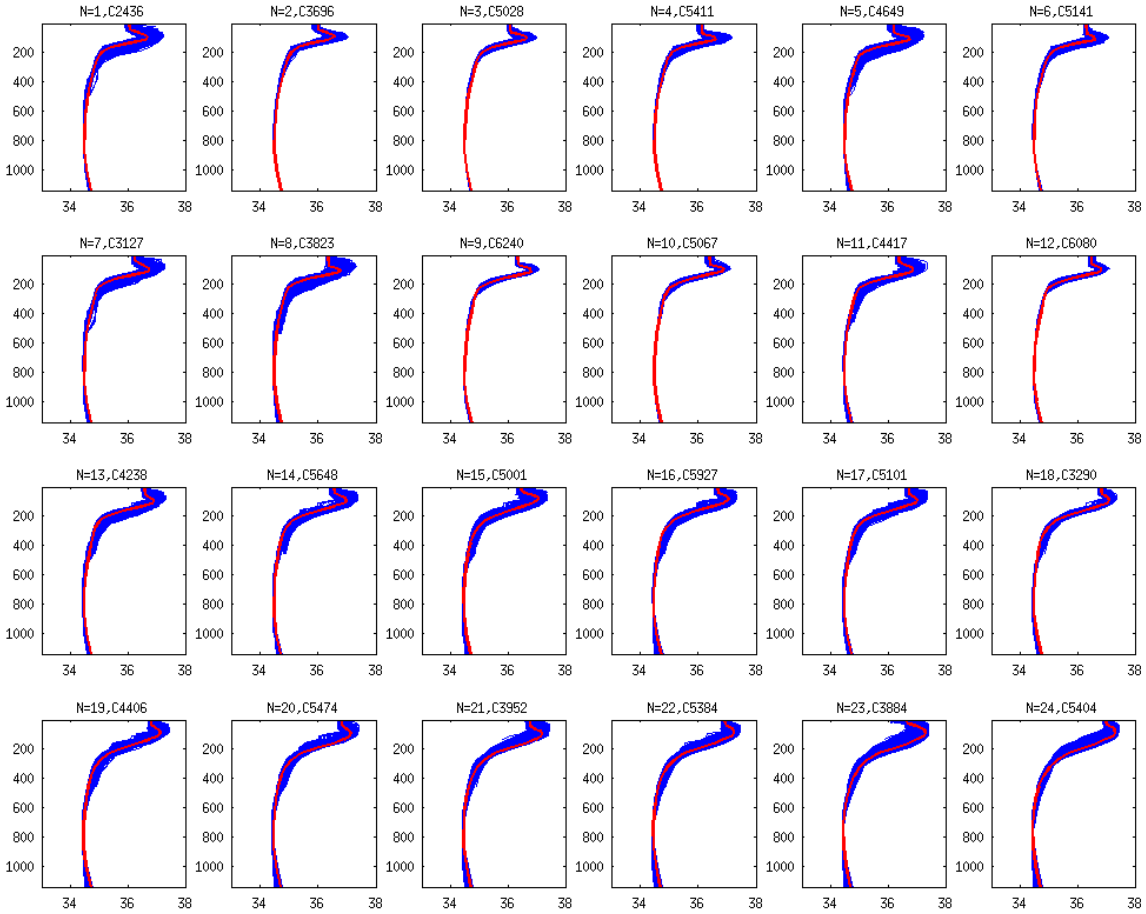


Figure IV-24:Quelques exemples de profils des neurones (en rouge) et des données captées (en bleu) par le neurone en utilisant la 2^{ième} méthode d'inversion.

Nous décrivons un autre critère d'analyse de la qualité de l'inversion basé sur les écarts-types. Pour cela, les neurones sont sélectionnés selon la base suivante,

- on calcule l'écart-type sur les données captées par chaque neurone (de référence) à chaque immersion en utilisant la projection directe (non inversée, projection de référence). Ce qui donne 25 écarts-types par neurone.
- Ensuite on sélectionne les données captées par chaque neurone et qui respectent la condition suivante :

$$| N_n(i) - donnee_m(i) | \leq a \times std_n(i)$$

où $N_n(i)$ est S du neurone du modélise $donnee_m$ *et* à l'immersion *i*, $donnee_m(i)$ est S de la donnée m à la même immersion et $std_n(i)$ est l'écart-type à l'immersion *i* des données captées par le neurone *n* quand tous les paramètres sont projetés et *a* un facteur multiplicatif entier.

Ce filtre a été appliqué sur les deux méthodes d'inversion avec les résultats ci-après.

Tableau IV-5:Nombre de données non éliminées par le filtre en fonction du nombre d'écart-type. Le pourcentage est donné entre parenthèse.

Nbre écart-type	meth1	meth2
2	24 391 (0,56%)	679 361 (15,65%)
3	68 265 (01,57%)	1 533 176 (35,32%)
4	161 719 (03,73%)	2 311 477 (53,25%)

Le tableau IV-3 présente le nombre de données qui passent le filtre en fonction du nombre d'écart-type utilisé entre les méthodes d'inversion meth1 et meth2.

Il illustre la qualité de la 2ème méthode d'inversion par rapport à la 1ère. Les pourcentages montrent que cette deuxième méthode est largement meilleure que la première, comme le confirme la figure suivante.

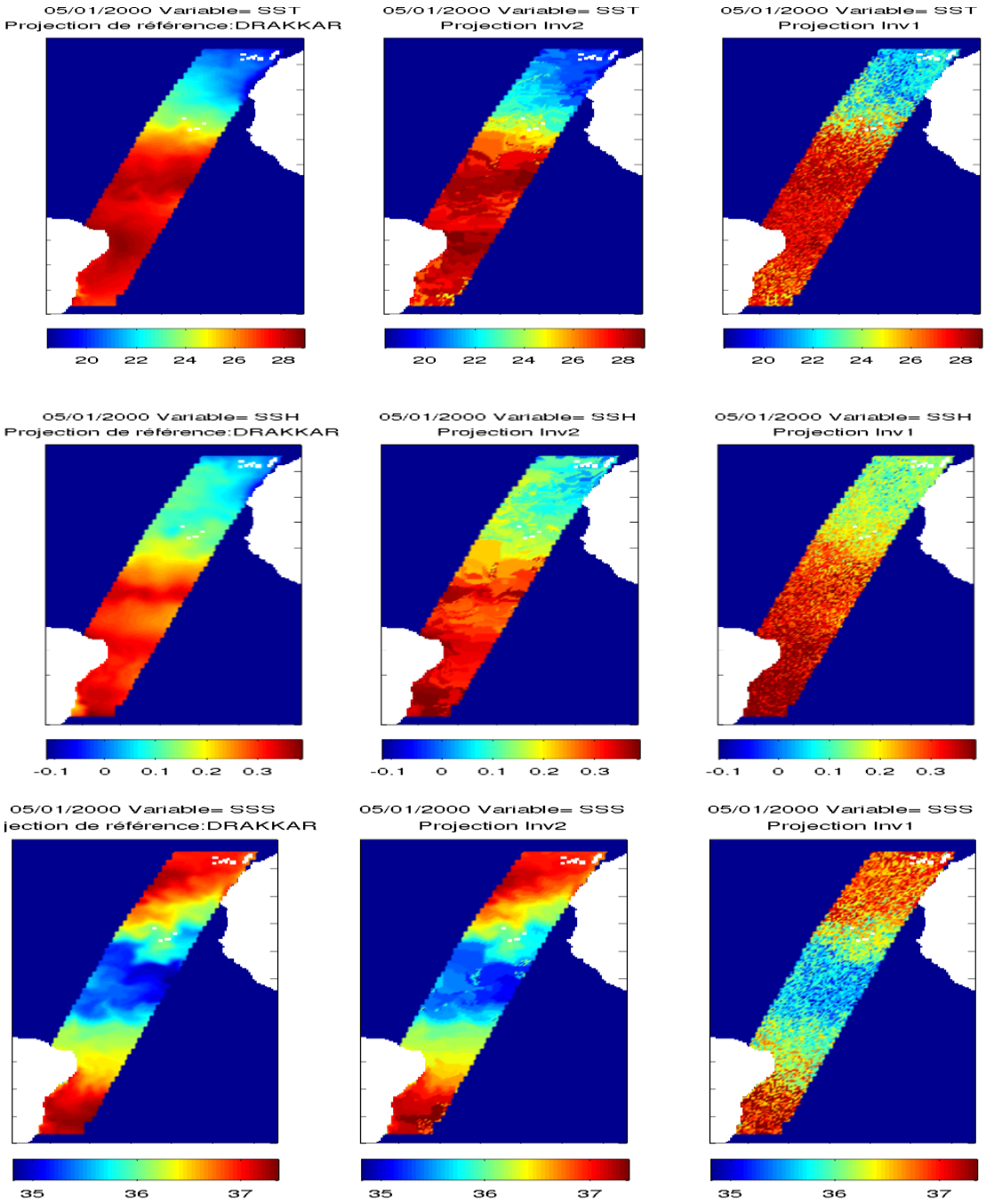


Figure IV-25 : Visualisation de 3 images (SST en haut, SSH au milieu et SSS en bas). (a), à gauche sont les références, celles reconstruites avec les neurones référents ((b), au centre) utilisant la 2^{ième} méthode d'inversion, ((c) à droite) utilisant la 1ère méthode d'inversion.

Nous avons reconstitué les images de référence en utilisant le 2ème modèle d'inversion, pour les variables SST, SSH et SSS du 05/01/2000 dans la *Figure IV-25*.

Elle montre de grandes différences entre les données réelles et les données inversées par

la 1ère méthode (*Figure IV-25*c) alors que ce n'est pas le cas pour la 2ème méthode d'inversion (*Figure IV-25*b) où les différences sont faibles même s'il y a de petites imperfections dans la reconstruction des structures fines.

IV.4 INTERCOMPARAISON DRAKKAR VS DRAKKAR INVERSEE.

Cette partie présente une comparaison entre les données DRAKKAR estimées par le modèle d'inversion INV2 et les données DRAKKAR de référence. Pour réaliser cette étude comparative, nous avons fait une projection des données DRAKKAR sur la carte en utilisant la méthode d'inversion INV2. Nous désignerons, dans la suite, les données DRAKKAR inversées par DRAKKAR-INV2 ou DRA-INV2.

Dans cette partie, les limites temporelles et spatiales sont basées principalement sur les campagnes ARAMIS. C'est-à-dire que nous considérerons une route moyenne ARAMIS comme référence spatiale et la date moyenne par campagne comme référence temporelle. Les comparaisons portent sur des critères globaux moyens. La zone d'étude est la bande oblique construite autour de la route moyenne ARAMIS présentée à Figure II-5.

Dans la comparaison nous avons calculé un profil moyen par gamme de latitudes. Chaque gamme est constituée par les données regroupées par latitude sur un intervalle de 0.5 °. Puisque les latitudes s'étendent de 20°S à 30°N, nous obtenons 100 profils moyens par jour pour les dates considérées. Les dates considérées sont celles moyennes des campagnes ARAMIS (cf. : Tableau II-2).

IV.4.1 Erreurs d'estimation

Nous aborderons cette phase d'étude des erreurs d'estimation par deux approches se basant sur les rmse et sur les erreurs absolues.

La rmse permet de connaitre la qualité de l'estimation dans sa globalité pour une latitude donnée. La *Figure IV-26* montre les rmse pour les dates choisies.

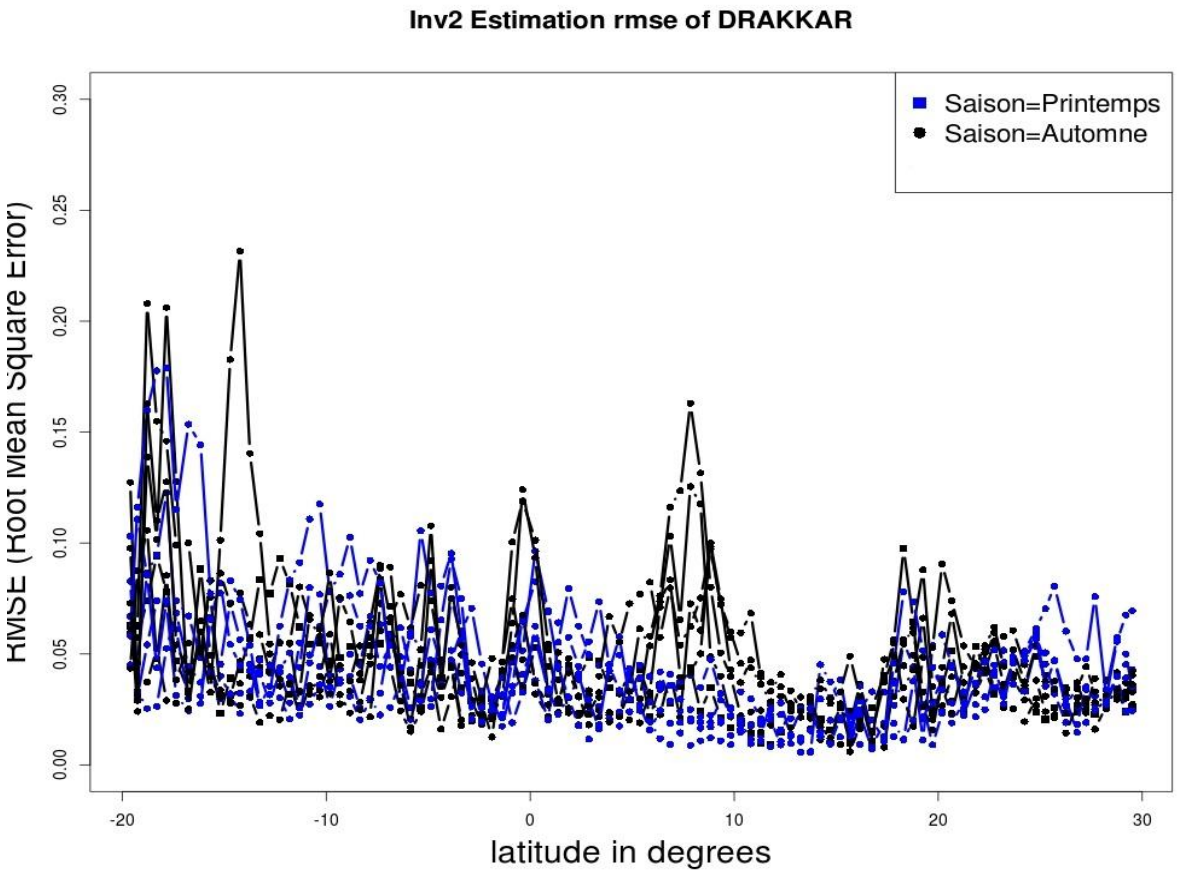


Figure IV-26 : Evolution latitudinale de l'erreur-type de l'estimation sur le profil. La courbe bleue marque les données de printemps et la courbe noire les données d'automne.

Nous notons des erreurs-types faibles, en général inférieures à 0.1, sur toutes les latitudes comprises entre 30°N et 12°S et des erreurs moins fortes pour les campagnes de printemps. Il n'y a pas de tendance latitudinale particulière des rmse même si nous pouvons constater que les rmse sont plus fortes au sud de la route surtout pour les campagnes d'automne. La figure suivante confirme cette situation.

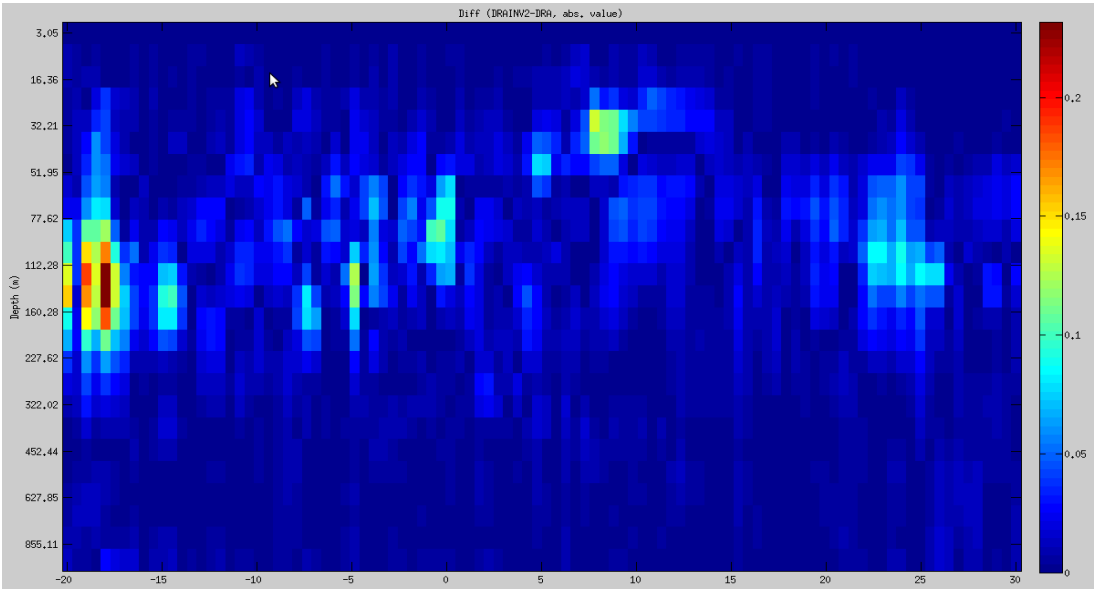


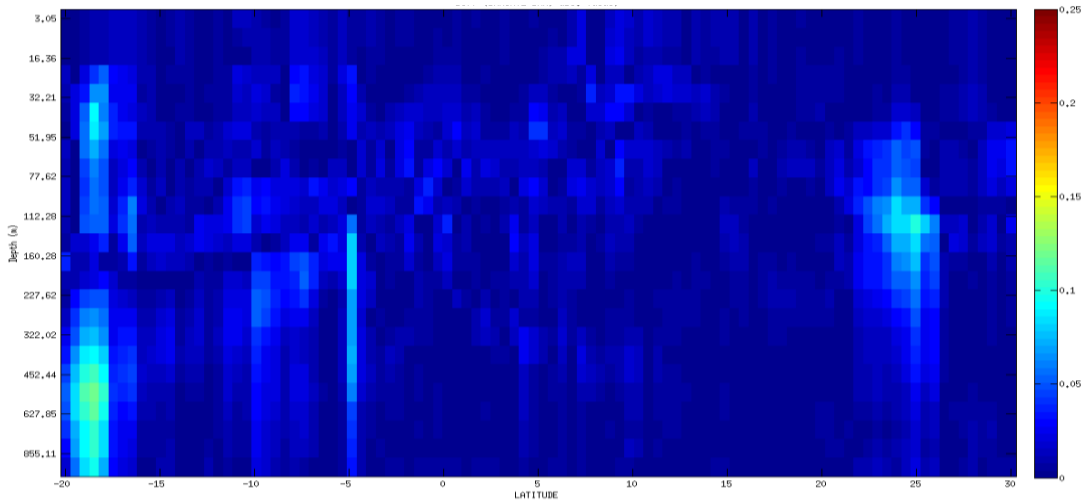
Figure IV-27 : Section des différences en valeur absolue entre DRAKKAR et DRAKKAR-INV2 moyennées.

Elle montre les différences absolues calculées sur les données moyennées par niveau d'immersion et par latitude entre DRAKKAR et DRAKKAR-INV2. Ces moyennes sont obtenues en calculant un profil moyen pour chaque gamme de latitude pour toutes les dates considérées.

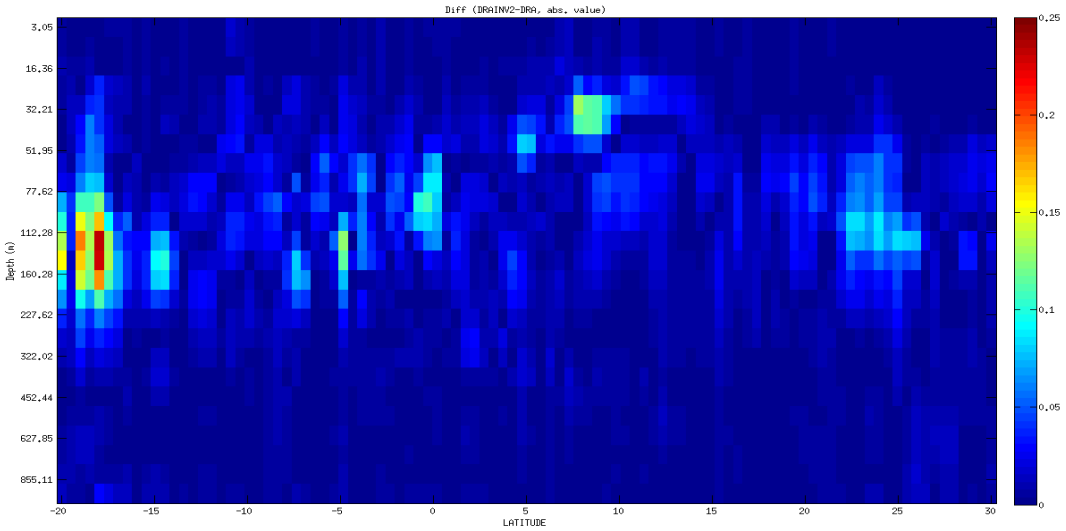
Les erreurs « moyennes » absolues d'estimation sont très faibles, inférieures à 0.075 psu sur toutes les latitudes et sur les immersions au-dessus de 25m et de 200m jusqu'à la fin du profil. Elles sont plus grandes autour de l'halocline.

Dans ces erreurs, il est important de définir la part de la quantification provoquée par la carte de Kohonen et celle provoquée par la méthode d'inversion elle-même. Les erreurs dues à la quantification vectorielle sont celles provoquées par la SOM, puisque ces dernières résument les données à un certain nombre de classes représentées par un référent. Et c'est cet espace de neurones qui correspond à l'ensemble des possibilités. Pour définir cette part, le neurone qui a le meilleur profil de S pour approcher celui d'une donnée (DRAKKAR), ayant participé ou pas à l'apprentissage a été considéré. Ce neurone est celui qui est le plus proche de la donnée en termes de distance euclidienne calculée seulement sur la S dans les immersions considérées ici soient de 0 à 1136m. L'ensemble construit avec ces neurones sera désigné par DRAKKAR-DEPTH (DRA-DEPTH) dans la suite.

La *Figure IV-28* présente la section des différences en valeur absolue entre DRAKKAR et DRAKKAR-DEPTH, et entre DRAKKAR et DRAKKAR-INV2 moyennées. Ces différences ont été calculées de la même manière que pour la *Figure IV-27*.



(a)



(b)

Figure IV-28 : la section des différences en valeur absolue entre DRAKKAR et DRAKKAR-DEPTH (a), DRAKKAR et DRAKKAR-INV2 (b) moyennées. La figure illustre les erreurs absolues DRAKKAR-INV2 et DRA-DEPTH, les latitudes sont en abscisses et les immersions en ordonnées, les couleurs donnent des erreurs allant de 0 à 0.25 psu.

La part qui vient des erreurs de quantification de la carte constitue une borne inférieure.

Et quelles que soient ses performances, le modèle d'inversion ne pourra pas faire mieux en termes de rmse sur tout le profil que cette borne inférieure. Cependant, l'inversion peut mieux estimer le profil sur certaines immersions.

La *Figure IV-28* montre quasiment sur tout le profil des écarts inférieurs à 0.05 psu pour DRA-DEPTH. Des écarts autour 0.1 psu sont localisés au sud de la route en dessous de 400m alors que pour DRA-INV2, les écarts localisés dans cette région aux mêmes immersions sont quasiment nuls. Ceci s'explique par le fait que le modèle d'inversion utilise la latitude comme principal argument et elle est très déterminante dans la formation de S à ces immersions. Mais, cette erreur est translatée vers le haut, autour de 100m, et ce niveau n'est relié par une forte relation à aucune des variables utilisées dans la première projection. L'analyse des profils de S présentée en III.1 a montré que les S de proche surface (de 0m à 100m) sont très liées à la SSS, les S de 300m à la dernière immersion très liées à la latitude mais aucune n'a montré aucune relation pertinente entre paramètres de surface et S autour de l'halocline.

Le paragraphe suivant présente d'autres méthodes de comparaison basées sur les profils.

IV.4.2 Expressions de la variabilité verticale des S DRAKKAR par le modèle.

L'étude de la variabilité verticale par les écarts-types verticaux telle que l'exprime les données DRAKKAR est présentée en *Figure IV-29*. Nous rappelons que l'écart-type vertical pour un profil de S est calculé de la manière suivante :

$$\sigma_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{y,i} - \overline{x_y})^2}$$

avec *y*=latitude, *n*=nombre de couches et $x_{y,i}$ S en une immersion i. Cette quantité représente la dispersion des valeurs de S d'un profil autour de S moyenne pour ce même profil en fonction des immersions.

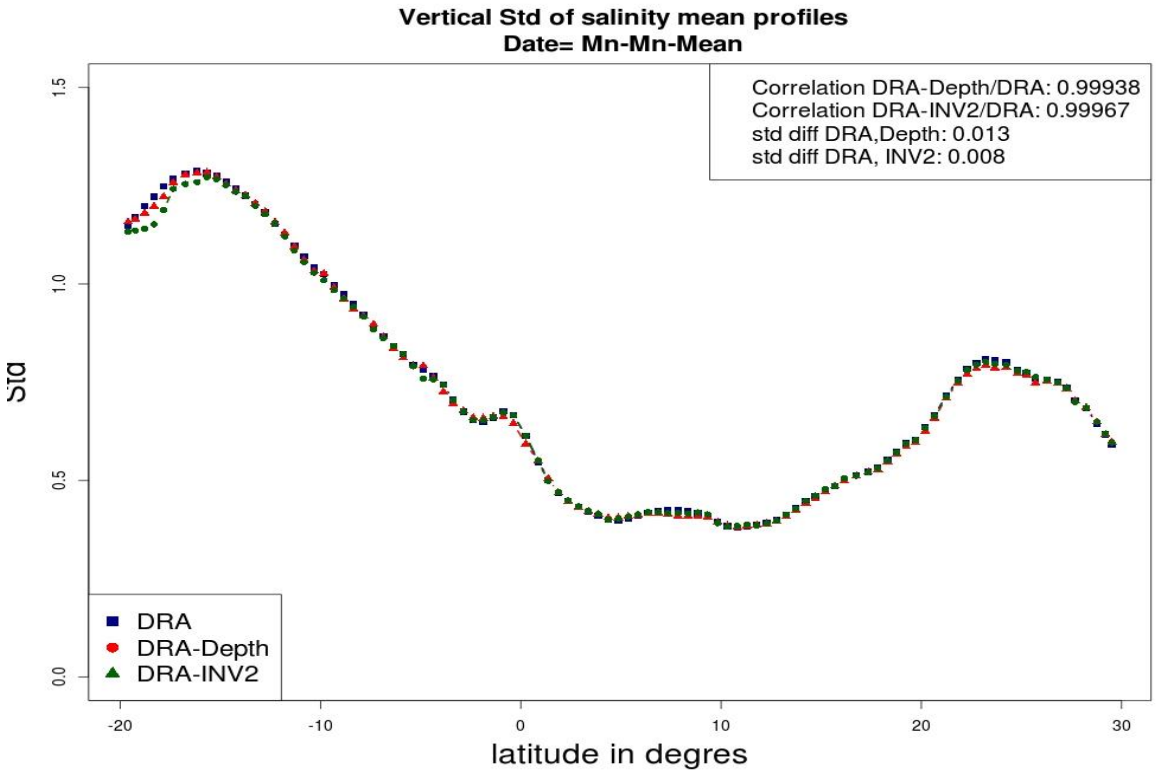


Figure IV-29: Evolution latitudinale de l'écart-type autour de la moyenne verticale des données (moyennées) DRA, DRA-DEPTH et DRA-INV2

Les corrélations entre les courbes d'évolution latitudinale des différents écarts-types sont fortes, ce qui signifie que les dispersions autour de la moyenne verticale réelle et estimée évoluent de la même façon dans l'espace (en latitude. A l'extrême sud de la route, on note une petite sous-estimation de la variabilité verticale par le modèle d'inversion. Les figures suivantes le confirment.

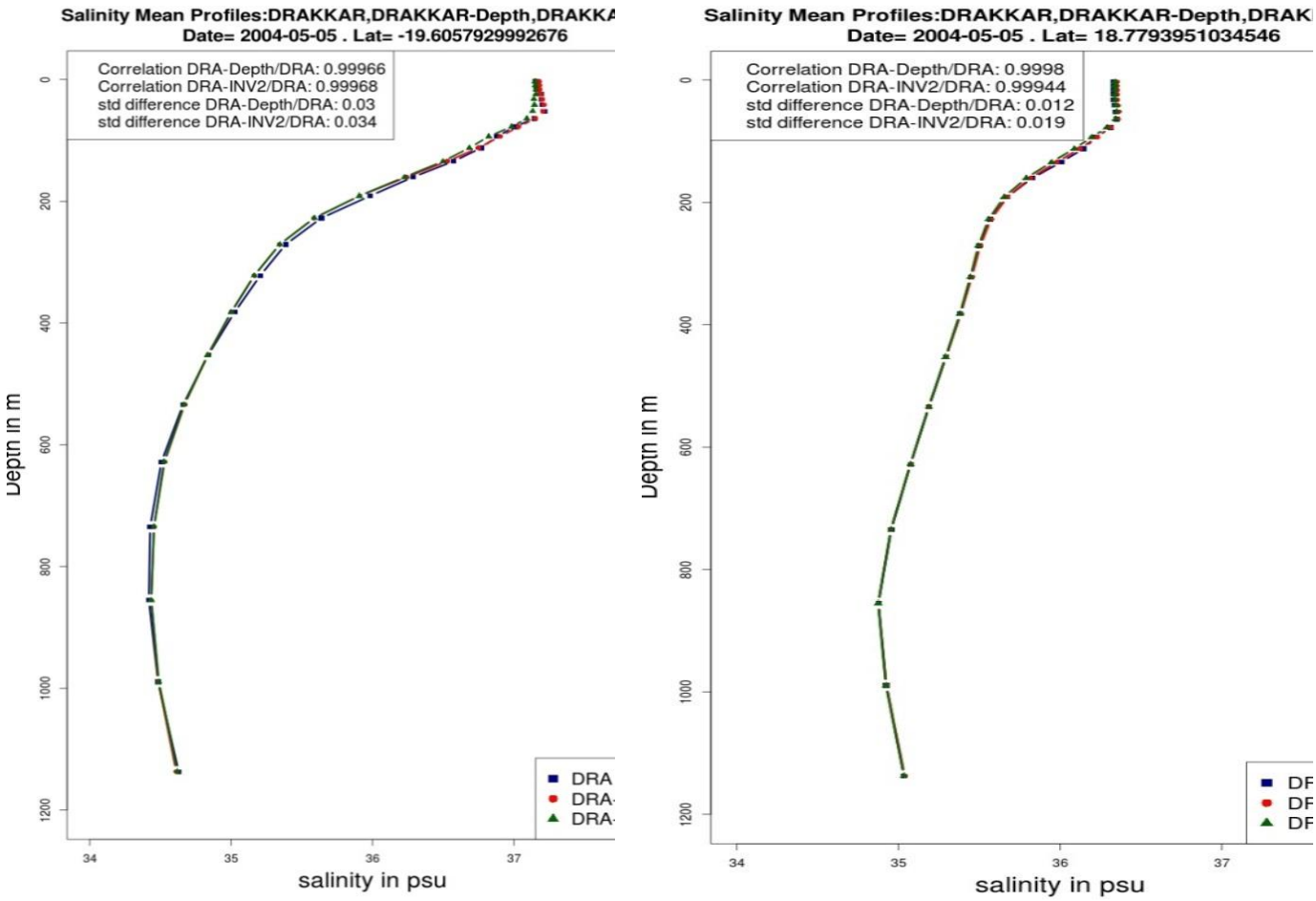


Figure IV-30: Exemples de quelques profils moyens de référence (blue), estimés inv2(Vert) et projDepth (rouge) sur différentes latitudes (a) 19S et (b) 18.7 N à la date du 05/05/2004.

La *Figure IV-30* montre deux profils de référence (DRAKKAR), leur inversion par le modèle INV2 et leur équivalent DEPTH i.e la projection des paramètres de profondeur. Nous voyons que pour la latitude 18.7°N (*Figure IV-30*b) le profil est mieux estimé.

Nous remarquons également que dans les deux cas le modèle reproduit avec plus d'exactitude les couches les plus profondes. S est très homogène dans ces immersions. Pour les couches de surface notamment vers 100m pour (a) et autour de 120m pour (b), les estimations sont moins exactes.

Cette inter-comparaison a permis de voir que le modèle d'inversion arrive bien à reproduire la référence qui est DRAKKAR, il est également très proche de la borne inférieure DRA-DEPTH.

Nous avons vu les bonnes performances de la méthode d'inversion sur les données DRAKKAR. Ces données sont issues de modèles numériques donc elles sont lissées et assez nombreuses. On peut ajouter le fait qu'elles aient une répartition spatiale et temporelle assez homogène. Ces 3 caractéristiques facilitent leur inversion. Nous avons également vu que les données DRAKKAR devraient être traitées avec prudence sur certaines latitudes et immersions où elles ne représentent pas très bien les données in situ ARAMIS.

La méthode d'inversion a été appliquée sur les données ARAMIS en utilisant la carte construite avec les données DRAKKAR. Il s'agissait de reconstruire ARAMIS à partir des données de surface avec cette carte. Les résultats obtenus sur cette expérience (non présentée dans le document) ne sont pas satisfaisants. D'une part parce que DRAKKAR utilise la SSH alors qu'avec ARAMIS, nous avons l'ADT, d'autre part nous avons vu que DRAKKAR présentait certaines différences avec ARAMIS ce qui augmente l'erreur de la borne inférieure.

Puisqu'il est important de tester la méthode d'inversion sur les données ARAMIS, véritable référence pour valider le modèle développé dans cette thèse, nous avons cherché des données in situ sur une zone qui englobe les trajets des campagnes ARAMIS. Ces données in situ sont celles extraites de la base Coriolis. Dans la section suivante, le modèle d'inversion est appliqué aux données Coriolis et aux données ARAMIS.

IV.5 INVERSION DES DONNEES CORIOLIS

Pour l'inversion des données Coriolis, seule la 2ème méthode d'inversion est utilisée. Dans le travail de thèse nous avons apporté une amélioration à ce modèle en intégrant les profils de T, mais l'inversion avec la 1ère version est d'abord présentée.

IV.5.1 Mise en œuvre du réseau pour l'inversion

Pour mettre en place la carte SOM, la démarche reste la même que celle utilisée en IV.1. Les seuls changements concernent les valeurs des paramètres. Ici 40% des données centrées réduites sont utilisées pour l'apprentissage, 40 x 25 neurones soit 1000 sont choisis et 29 variables (25 premières immersions des profils, correspondant à celles de DRAKKAR, SST, ADT, Long, Lat).

IV.5.2 Résultat d'apprentissage

Une fois l'apprentissage fini, nous pouvons faire différentes analyses de la carte obtenue afin de se prononcer sur la qualité de l'apprentissage. La Figure IV-31 représente la cardinalité des neurones.

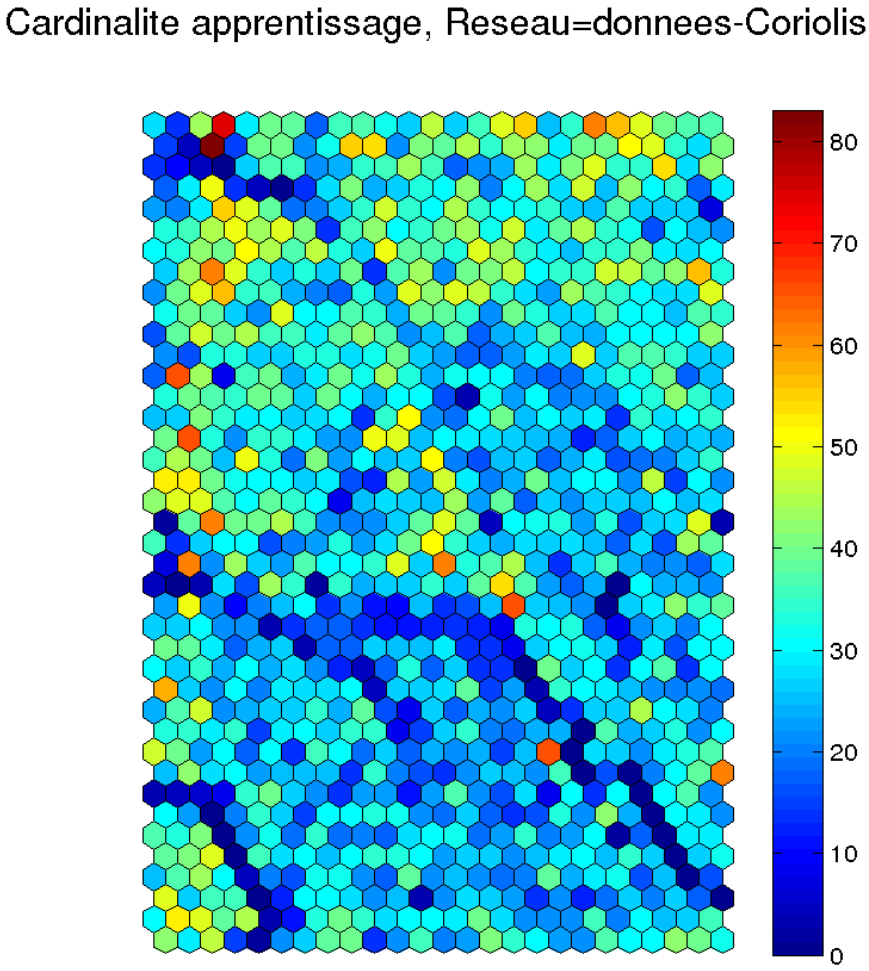


Figure IV-31:Carte des cardinalités (le nombre de données captées par chaque neurone). Figure déjà présentée à la page 60 en IV.1.1.

Nous pouvons remarquer que les cardinalités sont proches ce qui signifie qu'il n'existe pas par exemple de neurones qui ont capté une partie importante des données autrement dit la répartition est homogène. Nous pouvons également visualiser la carte des poids pour chaque variable par neurone.

La figure représente la carte des variables.

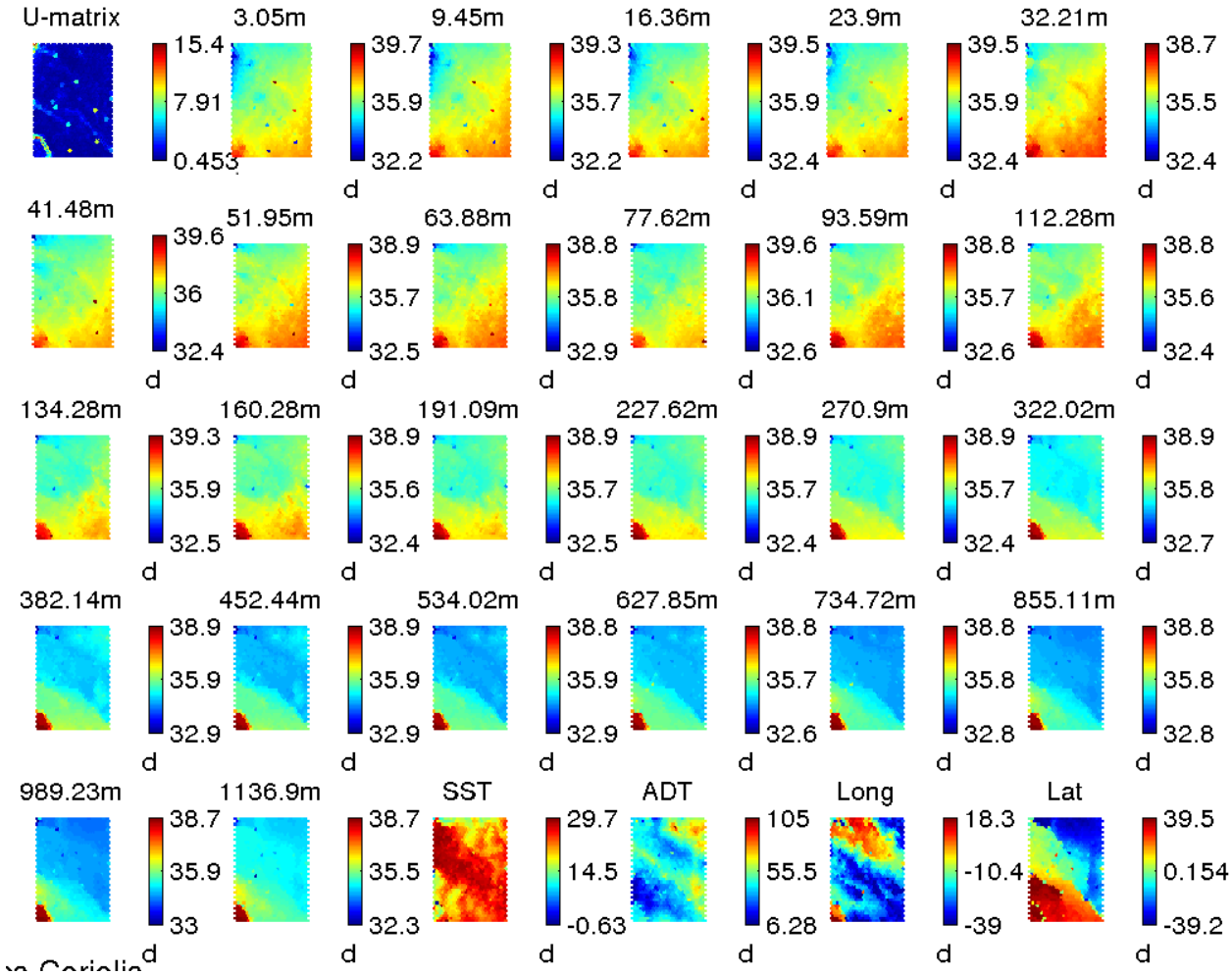


Figure IV-32:La carte des variables (les variables sont en titre de chaque rectangle). Chaque image correspond à une variable (S et paramètres de surface) à l'exception d'U-matrix qui indique plutôt la distance entre chaque neurone et ses voisins. Chaque petite image est constituée de 40 x 25 hexagones qui correspondent aux neurones. Pour chaque variable les différentes couleurs donnent le poids (valeur) du neurone (l'échelle est indiquée par la barre de couleurs).

L'ensemble présente un ordre topologique bien organisé dans la carte : en effet les neurones de même couleur (c'est-à-dire de poids similaires) apparaissent regroupés. Quasiment tous les tests présentés en IV.1.1 ont été appliqués à ce nouvel apprentissage pour en juger les résultats.

IV.5.3 Performance de l'inversion sur les données Coriolis

La validation d'un modèle et la quantification des erreurs sont des étapes importantes

dans toute étude de modélisation. Tout d'abord, elles donnent le degré de confiance que l'on peut avoir envers le modèle pour la prévision des phénomènes. De plus, elles peuvent aider à améliorer les modèles actuels en révélant les mécanismes les plus mal représentés, ainsi que les variables manquantes.

Malheureusement, ces étapes de validation et de quantification des erreurs sont négligées dans la plupart des études de modélisation. Arhonditsis et Brett (2004) ont analysé 153 modèles biogéochimiques publiés entre 1990 et 2002. Seulement 47% de ces modèles étaient validés avec des données expérimentales ou des observations en mer et seulement 30% présentaient une estimation des erreurs du modèle. Pour étudier les performances du modèle d'inversion sur ces données in situ, 2 approches seront considérées : une approche basée sur les erreurs sur les individus directement et une autre sur les erreurs globales moyennes.

IV.5.3.1 Erreurs directes de l'estimation

Cette étude est réalisée via les écarts et erreurs faits sur l'estimation d'un point *(x,y,t)* caractérisé par sa latitude, longitude et sa date.

L'étude des erreurs directes pose le problème de quantité de données. Pour y pallier, nous étudierons toutes les données ensemble ou une partie de celles-ci sans passer par une moyenne.

La quantification vectorielle, considérée comme une borne inférieure est également incluse dans la mesure des performances de l'inversion. La borne inférieure pour une donnée correspond au meilleur résultat qu'on puisse trouver pour la donnée considérée dans l'ensemble des résultats possibles après apprentissage. Comme expliqué en IV.4.1. ce résultat coïncide au meilleur neurone qui capte la donnée lorsque la projection concerne les paramètres de profondeur c'est-à-dire que seules les valeurs de profondeur sont présentées à la couche d'entrée. Cette borne inférieure est désignée par DEPTH. Elle est désignée par DEPTH (Coriolis) et l'inversion par INV2(Coriolis) dans la suite. La Figure IV-33 donne la répartition des rmse faites sur chaque profil pour les mois de janvier et de juillet.

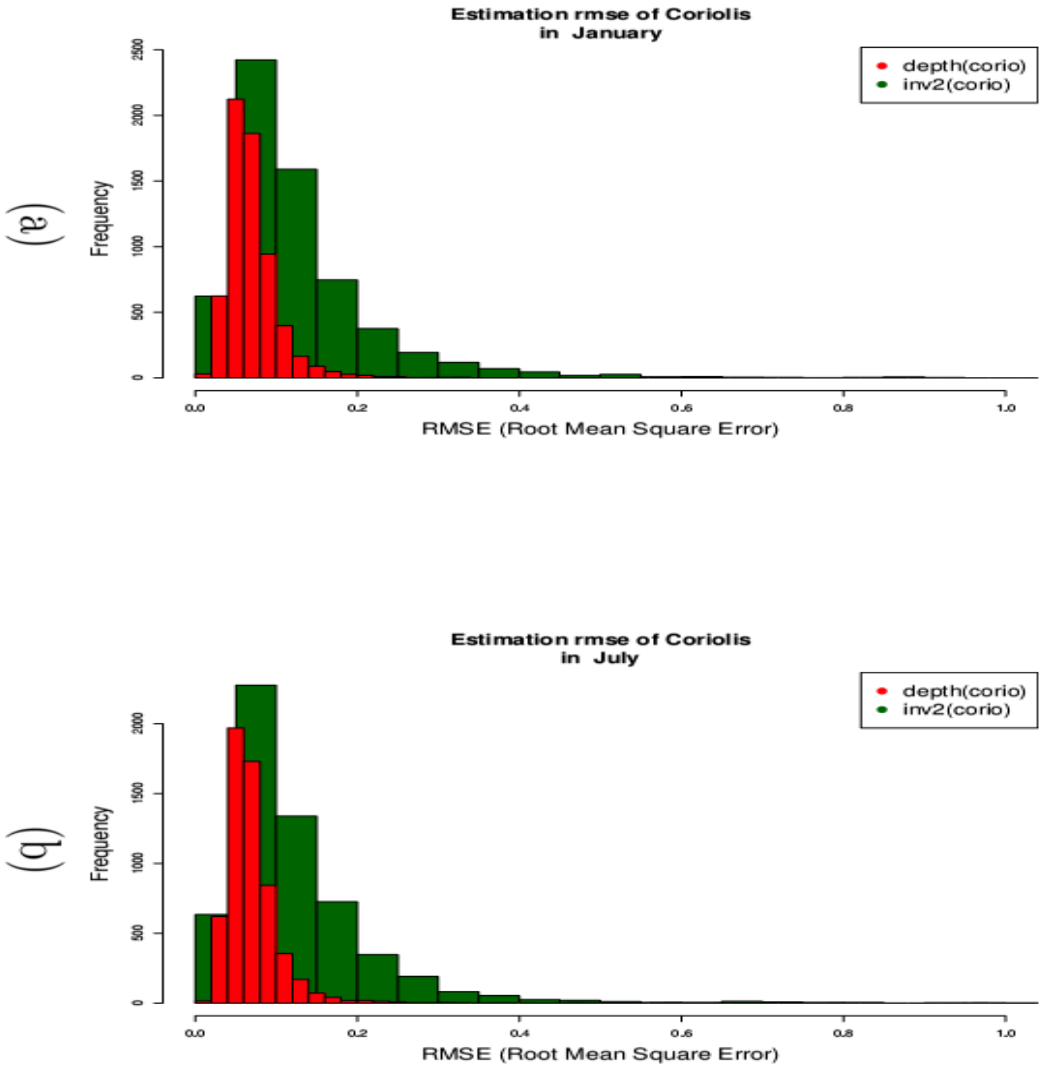


Figure IV-33:Fréquence des rmse pour les mois de janvier(a) et juillet(b) sur toute la base. Rmse sur le modèle d'inversion (vert foncé) et de la projection DEPTH (rouge).

Le nombre total de données par mois varie entre 5300 et 6400, il est de 6380 pour le mois de janvier et 5914 pour le mois de juillet. Les figures (a) et (b) permettent de voir plus clairement les vrais écarts, en rmse. Ainsi les écarts sont en majorité inférieurs à 0.1 psu. Nous ne notons pas de tendance particulière liée aux mois aussi bien pour INV2 que pour DEPTH, car quelque soit le mois, la distribution des rmse est la même. Cette situation est quasiment la même pour l'analyse spatiale des erreurs d'estimation.

IV.5.3.2 Erreurs globales de l'estimation

Dans cette comparaison nous avons calculé un profil moyen par gamme de latitudes.

Chaque gamme est constituée par les données regroupées par latitude sur un intervalle de 0.5 °. Puisque les latitudes s'étendent de 40°N à 40°S, nous obtenons 161 profils moyens.

Nous aborderons cette phase d'étude des erreurs globales d'estimation par le calcul des rmse et des erreurs absolues.

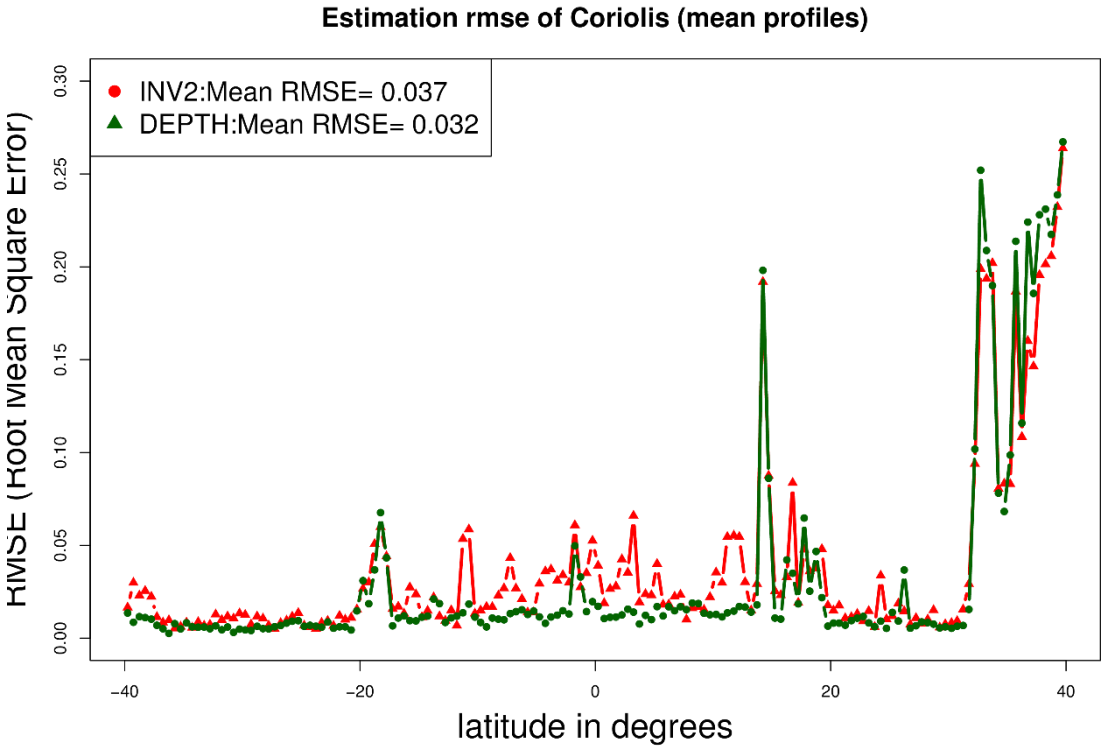


Figure IV-34:Évolution latitudinale de l'erreur-type de l'estimation sur le profil moyen. En cercle rouge les rmse faites par l'inversion et en triangle vert celles faites la projection DEPTH.

La Figure IV-34 représente les rmse des estimations par latitude et par profil moyen. Les profils ont été inversés puis pour chaque gamme de latitude, un profil moyen est calculé aussi bien à partir des données Coriolis réelles qu'à partir de ces mêmes données inversées, ce qui donne pour chaque gamme de latitude 2 profils moyens (des données réelles et des données inversées). INV2 représente la méthode d'inversion et DEPTH la projection du profil sur la carte, c'est-à-dire le meilleur neurone qu'on puisse trouver donc les erreurs minimales.

Dans l'ensemble, les erreurs sont faibles avec des valeurs moyennes de 0.037 psu et 0.032 psu, respectivement pour INV2 et DEPTH. Cependant, sur certaines latitudes, elles

atteignent des valeurs non négligeables allant jusqu'à 0.275 psu autour de 40°N et 0.20 psu autour de 17°N. Comme ces valeurs importantes se retrouvent aussi bien pour INV2 que pour DEPTH, on peut dire que cela est dû à la quantification vectorielle.
La Figure IV-35 présente le diagramme de dispersion des profils moyennés pour certaines immersions caractéristiques (cf. :Chapitre III.4).

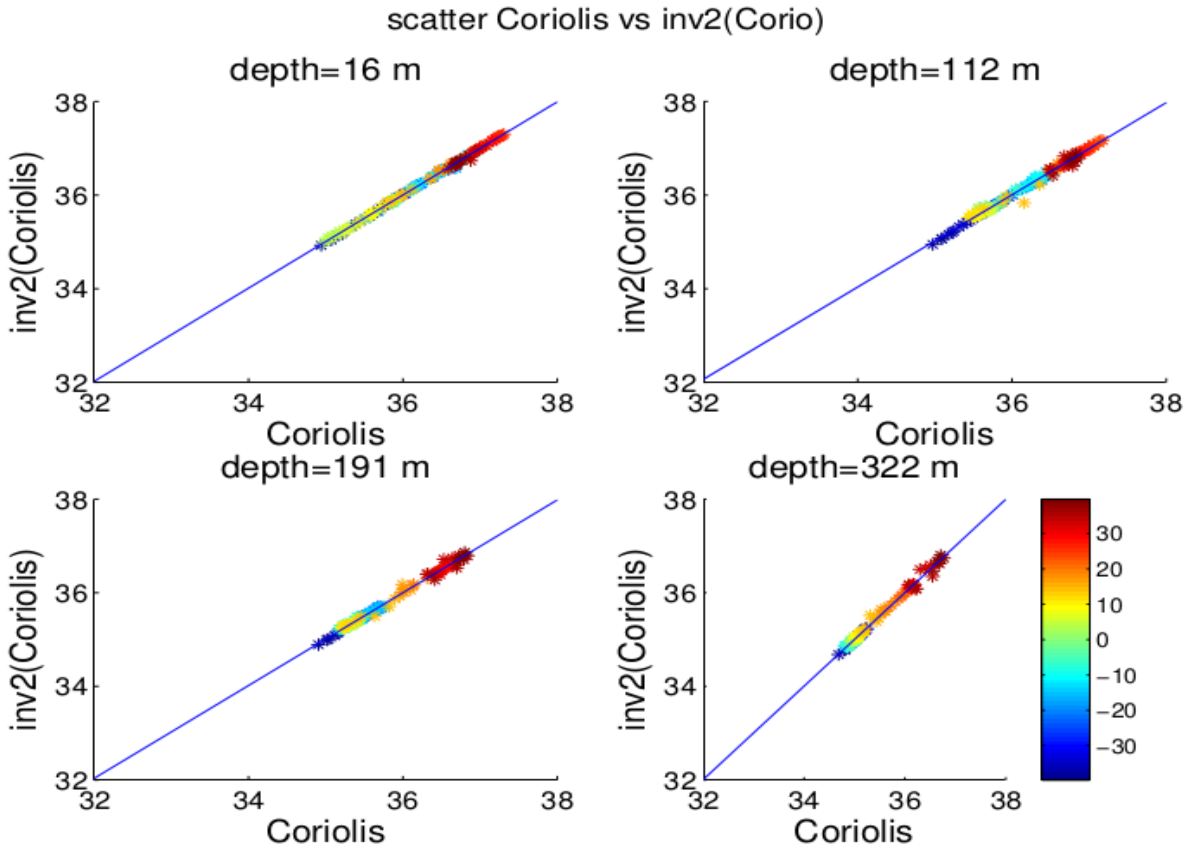


Figure IV-35: Dispersion Inv2(Coriolis) en fonction de Coriolis pour les immersions indiquées en titre. Le code couleur indique la latitude.

Les faisceaux de dispersion sont très fins sur toutes les immersions donc on en conclut que les profils sont bien reproduits.

En raffinant l'analyse, on voit que sur les couches de surfaces, l'inversion estime bien les S, en moyenne, sur toutes les latitudes. Alors que pour les immersions moyennes, l'estimation est moins bonne pour les latitudes supérieures à 35°N. Ceci est dû à une mauvaise projection, c'est à dire que des neurones ont capté des données se trouvant sur

des latitudes éloignées. Puisque l'inversion dépend de la latitude, qui est très caractéristique des plus basses immersions, ces données seront mal reproduites.

La Figure IV-34 présente la latitude "moyenne" de chaque neurone et les écarts-types par neurone en latitude. Cet écart-type est calculé sur les données captées par neurone.

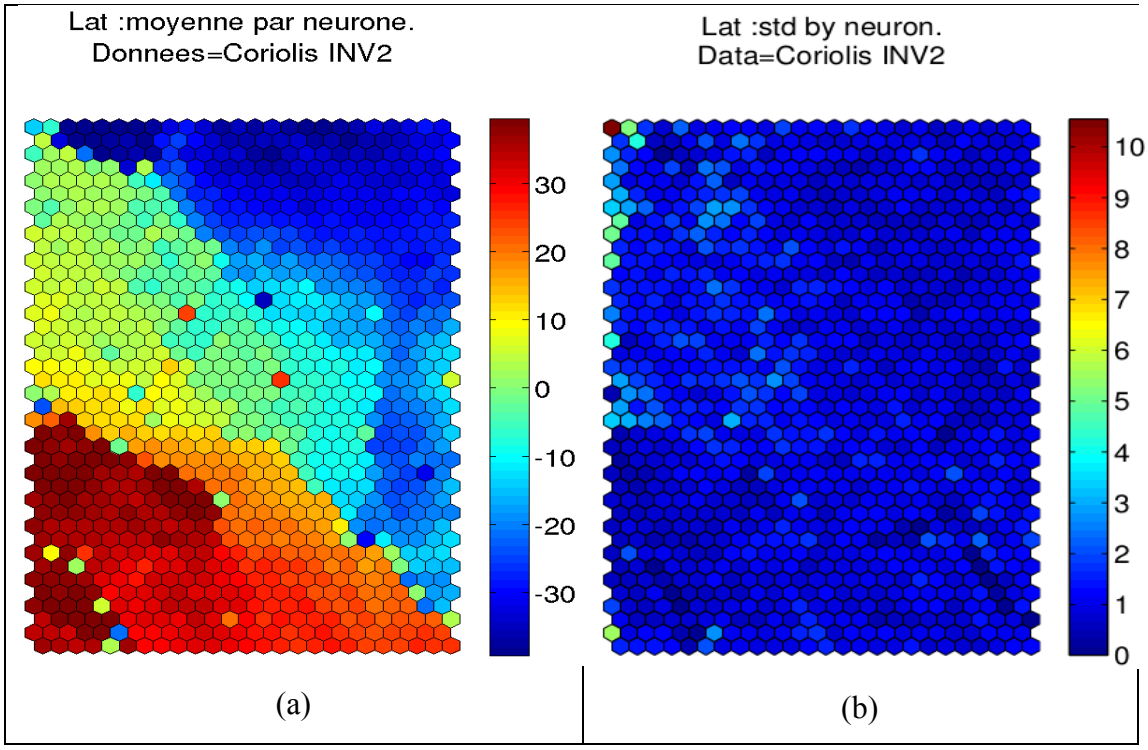


Figure IV-36:Moyenne(a) et Écarts-types(b) de la latitude pour chaque neurone.

La structure est bien organisée topologiquement. Les neurones voisins ont capté des données proches en termes de latitude. L'image b, donne les écarts-types qui sont généralement faibles (< 1°). Seuls 5 cas ont des écarts-types supérieurs à 5°.

Cette évaluation des performances de l'inversion montre que le modèle reproduit bien les profils de S à partir des paramètres de surface notamment la SSS, la Latitude, la Longitude, l'ADT et la SST. Considérant que le référent d'un neurone est formé par les valeurs moyennes de l'ensemble des données qu'il a captées, on pouvait s'attendre à ce que les neurones reproduisent bien les profils moyens des données. Cependant, la robustesse du modèle d'inversion réside dans sa capacité à bien estimer la projection du profil (DEPTH) c'est-à-dire à faire presque exactement ce que cette projection parvient à faire. Nous avons vu également qu'il est bien moins facile d'estimer le profil quand il est pris isolément.

IV.6 INVERSION DES DONNEES ARAMIS

Dans cette section nous présentons les résultats de l'inversion des données ARAMIS. La route ARAMIS est particulièrement intéressante puisqu'elle traverse plusieurs des grands courants océaniques zonaux de l'Océan Atlantique tropical comme présenté en II.2.

Dans ce paragraphe les résultats concernent l'inversion INV2 basée sur la carte réalisée avec les données Coriolis ainsi les référents des neurones qui modélisent les données seront sélectionnés dans cette carte. La carte sera désignée par SomCorio dans la suite, les données ARAMIS reconstruites par une méthode METH par le biais de cette carte sera désignée par $METH_{SomCorio}$(ARA). Les données ARAMIS ont été ré-échantillonnées ainsi seules les immersions correspondant à celles de DRAKKAR entre 0m et 850m ont été considérées.

Cette méthode de ré-échantillonnage a été testée et validée dans nos travaux précédents présentés dans la section II.3.4.

L'étude de l'inversion se fera en utilisant les techniques de comparaison comme celles définies en IV.5.3 et d'autres axées sur les niveaux d'immersion.

Une borne inférieure sera comparée en même temps à INV2.

Pour étudier les performances du modèle d'inversion sur ces observations, 2 approches seront considérées : une approche basée sur les individus directement et une autre basée sur des erreurs globales "moyennes".

IV.6.1 Comparaison directe entre données réelles et données inversées

Comme dans le paragraphe précédent, la racine carrée de l'erreur quadratique moyenne, est utilisée comme étant la base des comparaisons. La Figure IV-37 donne la répartition des rmse par centile.

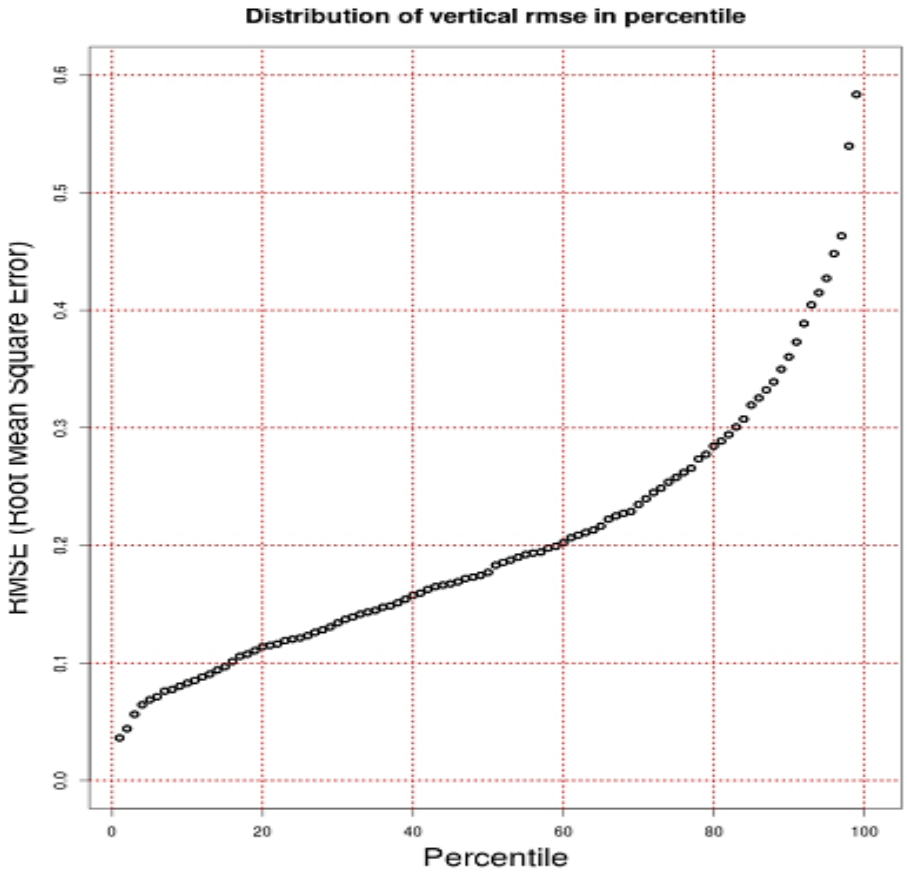


Figure IV-37:Dispersion des rmse en fonction des centiles.

L'axe des ordonnées de la figure représente la rmse en psu faite par la reconstruction sur chaque profil et l'axe des abscisses représente les centiles, c'est-à-dire indique le pourcentage de rmse dont la valeur est inférieure à ce qui est indiqué en ordonnée. Nous pouvons voir par exemple que 60% des profils ont été reconstruits avec une rmse verticale (faite sur le profil) < 0.2 psu et que 80% des profils ont été reconstruits avec une rmse < 0.28 psu. Ce résultat est encourageant, cependant il convient de déterminer la répartition de ces erreurs aussi bien en temps (t), en espace (x,y équivalents à la longitude et latitude) qu'en profondeur (z). Par rapport au temps et à l'espace, les rmse d'estimation des profils de S des campagnes ARAMIS le long de la route sont utilisées. La Figure IV-38 présente ces erreurs pour toutes les campagnes ARAMIS en fonction de la latitude (en abscisse) et de la saison (en couleur).

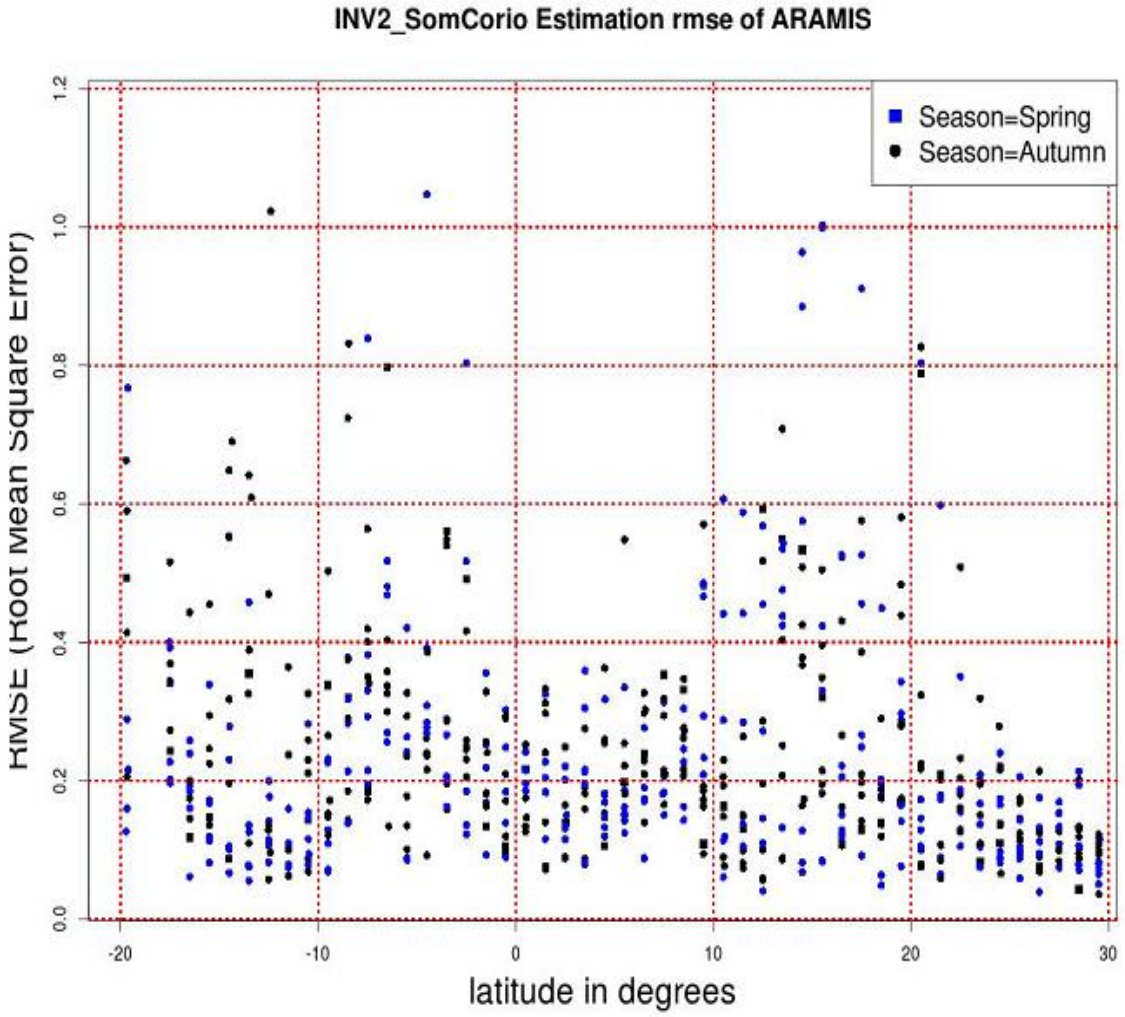


Figure IV-38:Racine de l'Erreur Moyenne Quadratique par Campagne sur les profils en fonction de la latitude. En bleu les campagnes de printemps et en noir les campagnes d'automne.

Le projet ARAMIS s'est déroulé sur 2 saisons (automne et printemps) de 2002 à 2008. Ceci justifie la classification des rmse en deux saisons. A priori, il n'y pas une tendance assez frappante d'une saison par rapport à l'autre en fonction de la latitude. Ce qui montre que, par rapport au temps, le modèle inverse de la même manière.

Étudions maintenant les rmse en fonction de l'immersion ce qui entraîne une redéfinition de l'équation de calcul de la rmse. L'équation suivante spécifie ceci

$$\sigma_{im} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{im,i} - y_{im,i})^2}$$

avec *im*=immersion donnée, *n*=nombre de points estimés sur la couche *im*, $x_{im,i}$ la S réelle

du point i à l'immersion im, $y_{im,i}$ la S estimée par le modèle $INV2_{SomCorio}$ du point *i* à l'immersion *im*

La Figure IV-39 donne les profils de rmse. Elle illustre les erreurs en fonction de la profondeur (en ordonnée) et de la saison (en couleur).

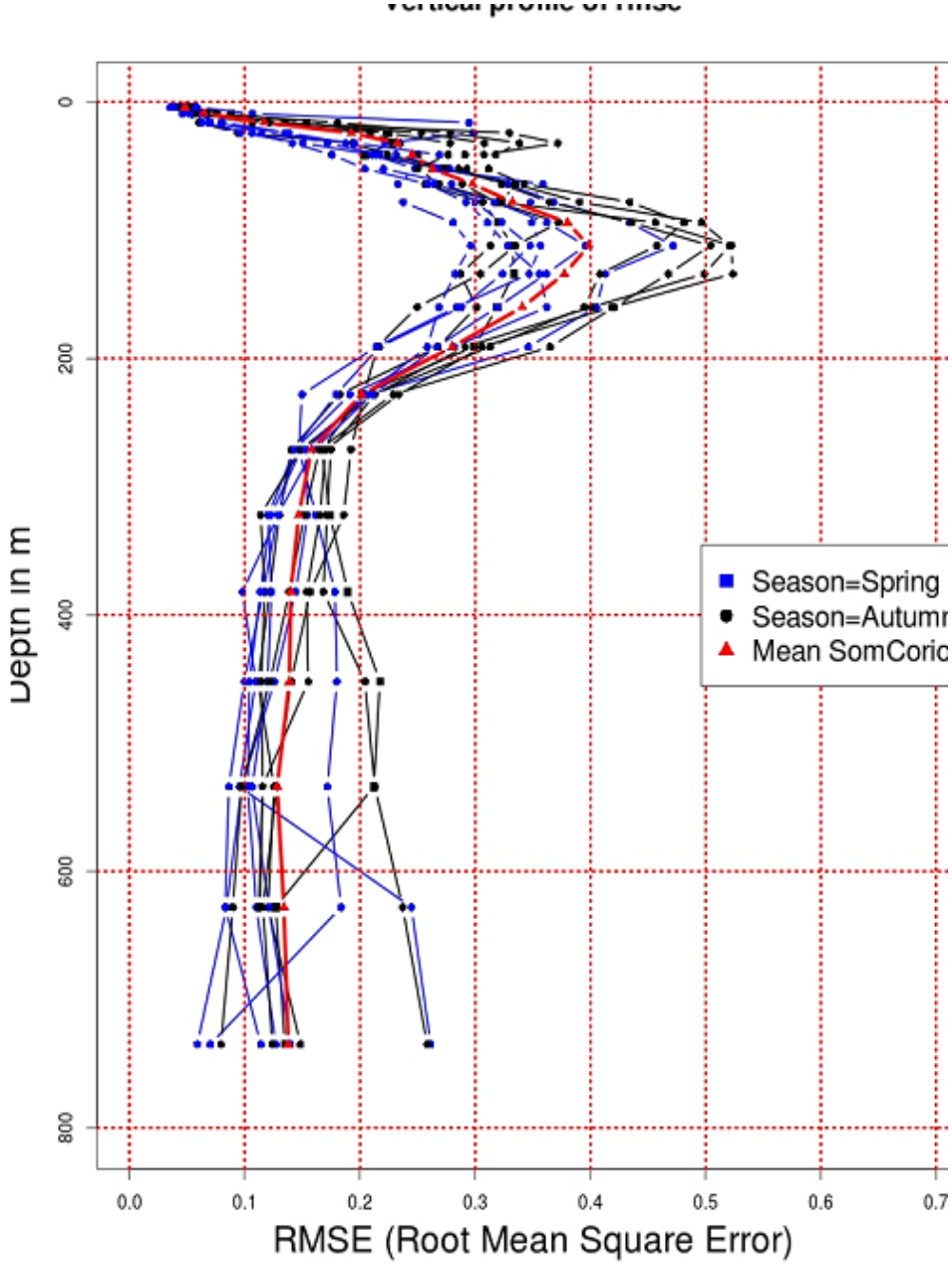


Figure IV-39:Profils de rmse du modèle $INV2_{SomCorio}$ sur les données ARAMIS. En bleu les rmse concernant les campagnes d'automne, en noir celles des campagnes de printemps et en rouge le profil moyen des rmse.

La figure montre globalement des erreurs plus fortes entre 50 et 200m de profondeur

quelle que soit la saison même si elles sont plus fortes dans les campagnes d'automne où elles atteignent 0.5 psu. En dessous des 200m et entre 0 et 50m, les rmse sont comprises entre 0 et 0.2 psu. Ceci s'explique par le fait que la variabilité est faible en dessous de 200m donc plus facile à modéliser. Ces immersions sont également très dépendantes de la latitude. Le modèle d'inversion étant principalement basé sur la latitude et SSS, les immersions en dessous de 200m sont mieux reproduites. L'utilisation de la SSS dans le modèle d'inversion a permis de mieux reproduire la S de 0 à 50m car la S à cette couche est très liée à la SSS. Cette figure, contrairement à la Figure IV-38 précédente, montre une tendance assez forte d'une saison à l'autre en fonction de la profondeur. Ce qui signifie que le modèle reconstruit différemment les S en fonction de l'immersion suivant la saison considérée.

Cette étude a permis de voir la distribution de la rmse en fonction des profondeurs et des saisons. La Figure IV-40 permet de mieux expliquer les différences au niveau des profondeurs, elle illustre les diagrammes de dispersion sur quelques immersions spécifiques.

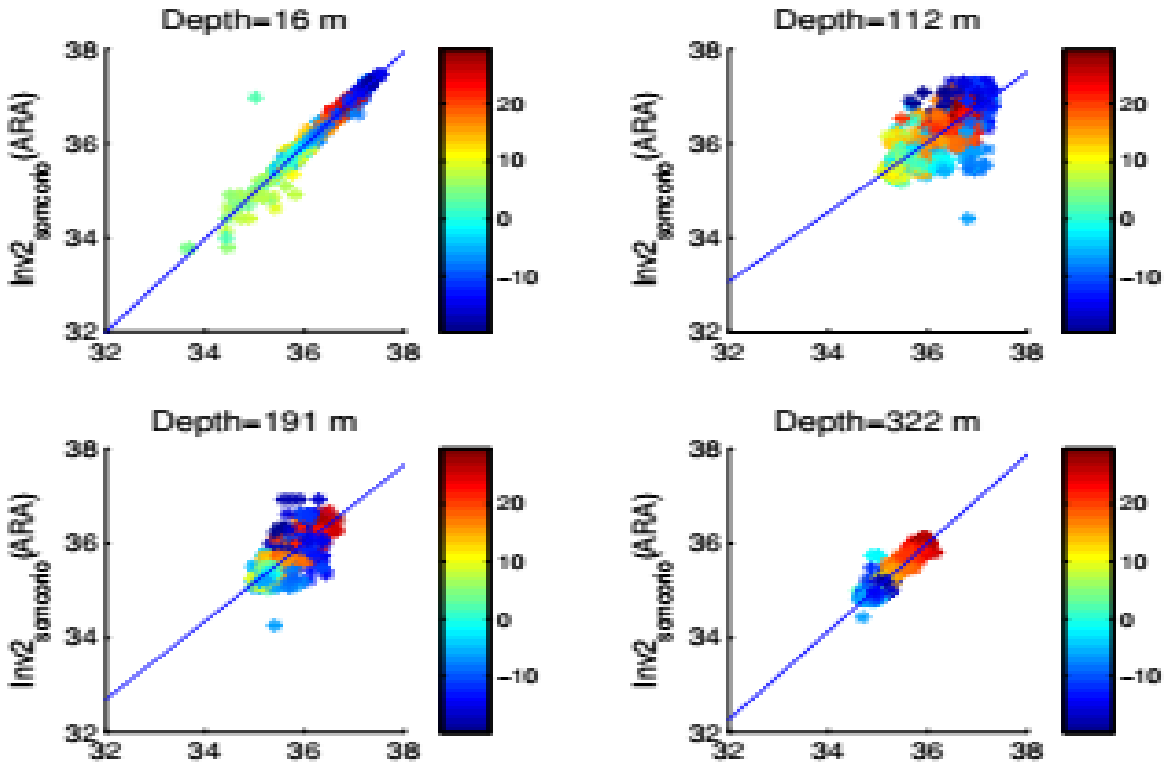
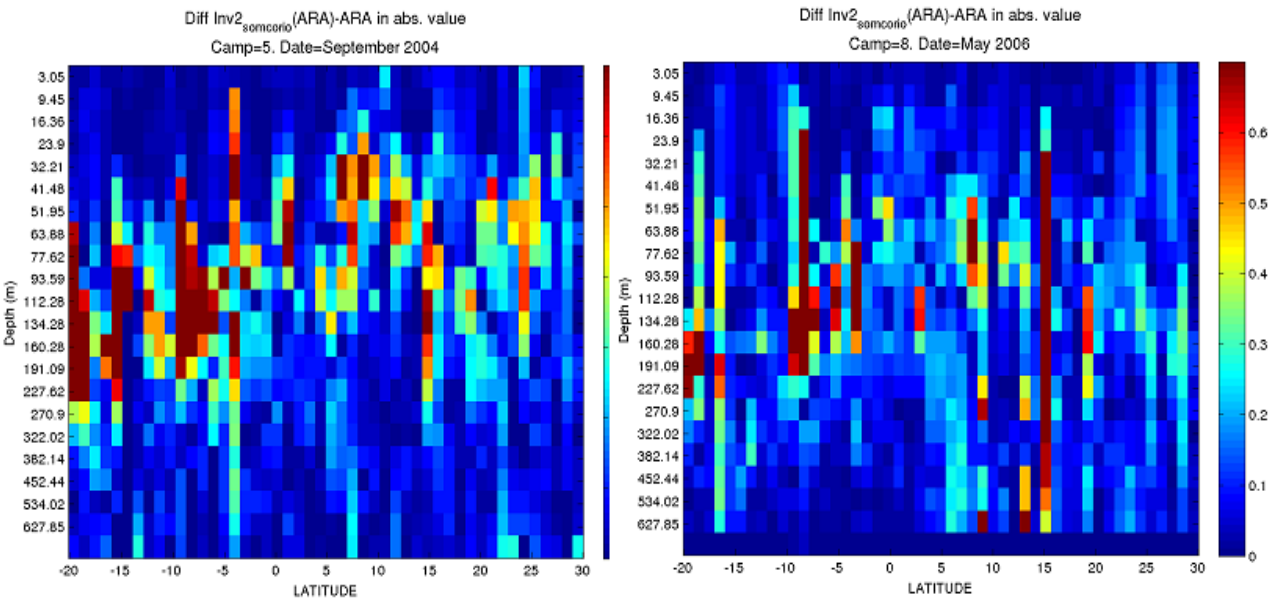


Figure IV-40: Dispersion de $INV2_{SomCorio}$ (ARAMIS) en fonction ARAMIS en abscisse pour quelques immersions spécifiques caractéristiques. Le code couleur indique la latitude.

On observe 4 situations : en surface (a) nous avons les maxima de S au nord vers 30° N et au sud vers 20°S, ces 2 zones correspondent aux SMW symétriques par rapport à l'équateur météorologique. Les minima sont autour de l'équateur, minima certainement dus aux fortes pluies qui dessalent les couches de surface. Aussi bien le modèle qu'ARAMIS font ressortir ces 2. En surface (Figure IV-40a), la droite de régression (trait bleu) correspond quasiment à la 1ère bissectrice (x=y) et les points décrivent un faisceau, ce qui veut dire que le modèle reproduit bien S à cette immersion comme l'a d'ailleurs montré la Figure IV-39 même si pour quelques points l'estimation est moins précise. Dans les eaux plus profondes (b, c), les points sont plus dispersés, donc le modèle reproduit moins bien S en ces immersions qui correspondent à l'halocline. Mais le modèle ressort bien les structures générales de S à ces immersions. Par exemple, aussi bien ARAMIS que le modèle montrent les maximas de S vers 15°S. Considérant l'immersion 322m, S croit avec la latitude dans l'hémisphère nord. Cette situation est bien reproduite par le modèle. Cette immersion correspond aussi à la plus petite variabilité comme le montre la figure, les données sont moins étalées sur la droite.

Les sections permettent de voir plus en détails les écarts par immersion en fonction de la latitude. Les 2 figures suivantes illustrent les sections des écarts absolus entre $INV2_{SomCorio}$ (ARAMIS) et ARAMIS réelles pour 2 campagnes, ARAMIS5 en septembre 2004 et ARAMIS8 en mai 2006. Elles correspondent respectivement aux campagnes dont les profils sont les moins bien reproduits et les mieux reproduits.



(a) :ARAMIS 5 en septembre 2004 (b) :ARAMIS 8 en mai 2006

Figure IV-41: Sections des écarts absolus entre INV2 et ARAMIS réelle pour les campagnes 5 en septembre 2004 et 8 en mai 2006.

Dans les couches plus profondes (de 225 à 800m) et de proche surface (0 à 30m), les différences sont plus faibles, les raisons sont expliquées dans le paragraphe précédent. Cependant dans d'autres couches nous avons de fortes valeurs en fonction de la latitude. Par exemple dans les campagnes impaires d'ARAMIS (de septembre-octobre, en automne), il y a de fortes valeurs (différences) autour 10-5° N et entre 50 et 150m. Ce qui montre que certaines structures fines telles que l'épaisseur de l'halocline ne sont pas toujours reproduites correctement par le modèle d'inversion.

Dans cette section, nous avons présenté les erreurs d'estimation sur la reconstruction directe des profils. Cette reconstruction est critique et constitue une bonne méthode pour évaluer un modèle dans la mesure où elle fait ressortir la moindre erreur. Dans la section suivante les erreurs globales c'est-à-dire sur les profils moyens sont étudiées.

IV.6.2 Performances globales de reconstruction des profils ARAMIS

Dans cette comparaison nous avons calculé un profil moyen par gamme de latitudes. Chaque gamme est constituée par les données regroupées par latitude sur un intervalle de

1° au lieu de 0.5° comme on l'avait défini avec Coriolis. Puisque les latitudes s'étendent de 30°N à 20°S, nous obtenons 50 profils moyens.

Nous aborderons cette phase d'étude des erreurs globales de la reconstruction par le calcul des mse et des erreurs absolues.

La rmse sur les profils moyens est représentée sur la figure IV.42 en fonction de la latitude.

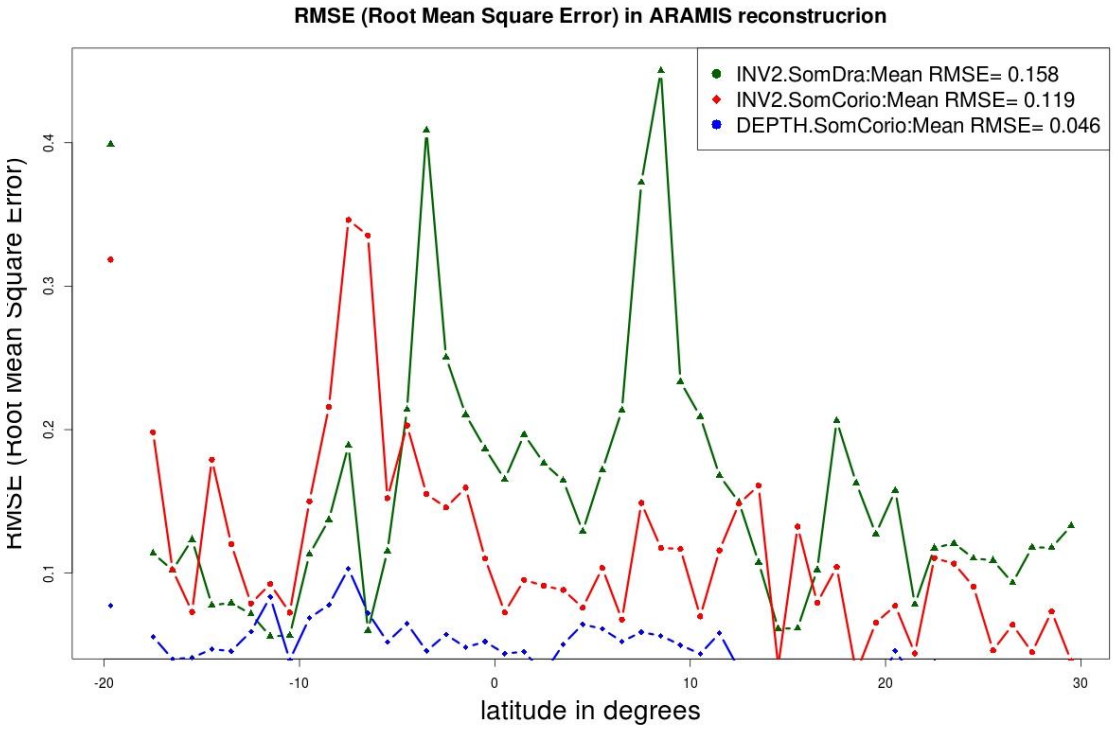


Figure IV-42:Évolution latitudinale de la rmse de l'estimation sur le profil moyen. En cercle rouge les rmse faites par l'inversion quand la carte Coriolis est utilisée comme ensemble d'apprentissage, en triangle vert les rmse faites par l'inversion quand la carte DRAKKAR est utilisée comme ensemble d'apprentissage et en cercle bleu celles faites par la projection DEPTH.

Cette représentation montre, dans ce type comparaison qui utilise les profils moyens, que SomCorio donne de meilleurs résultats que la carte qui utilise les données DRAKKAR dans sa base d'apprentissage. Dans l'ensemble, les rmse sont faibles avec des valeurs moyennes de 0.119 et 0.04, respectivement pour INV2 et DEPTH. Tous les commentaires faits sur l'inversion des données Coriolis sont valables.

Les erreurs-types sur les profils moyens par immersion présentée à la Figure IV-43 donnent des résultats aussi intéressants.

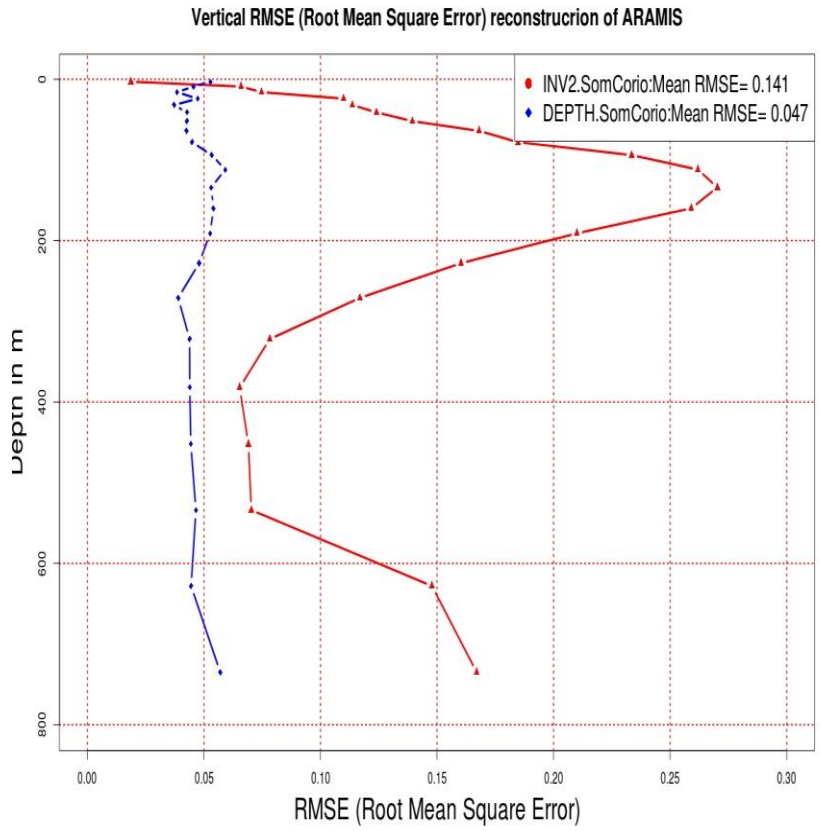


Figure IV-43:Profils de rmse du modèle INV2SomCorio sur les données ARAMIS (en rouge) et profil de rmse de la projection DEPTH en rouge.

En effet, sur les couches plus profondes (de 250 à 520m) et de proche surface (0 à 90m), les différences sont plus faibles. La SSS et la Latitude qui sont très importantes dans le modèle sont fortement liées à la S à ces immersions. Ceci a permis d'avoir des rmse inférieures à 0.07 psu à ces profondeurs. Cependant dans d'autres parties du profil, nous avons de plus fortes valeurs comprises entre 0.15 et 0.26 psu, le maximum est atteint autour de 100m approximativement dans la halocline.

Cette évaluation des performances de l'inversion sur les données Coriolis et ARAMIS montre que le modèle reproduit bien les profils de S à partir des paramètres de surface notamment la SSS, la Latitude, la Longitude, l'ADT et la SST. Cependant, certaines structures ne sont pas bien reproduites, ainsi il est important de créer plus de liaisons entre le profil de S et d'autres paramètres et d'introduire plus de variabilité dans la carte au

moment de l'apprentissage. Ainsi, le profil de T a été introduit dans l'étape de l'apprentissage, la taille de la carte a été augmentée. Les résultats avec cette nouvelle carte sont présentés ci-dessous.

IV.7 L'APPORT DU PROFIL DE TEMPERATURE

Température et salinité océanique sont étroitement liées via l'équation d'état (Millero et Poisson, 1981),une équation très complexe dépendant donc de T, S et P. C'est ce lien fort entre S et T, ou Θ la température potentielle, qui permet aux océanographes de faire des hypothèses sur les transports des masses d'eau et donc les propriétés thermo-halines qu'ils véhiculent. Dans l'Atlantique tropical par exemple, les eaux en provenance de l'hémisphère Sud (SAW pour South Atlantic Water) auront une T (>25°C) et S (>36.5 psu) assez élevées en surface, une S très faible (< 34.5) et une T froide (< 5°C) au dessous de la halocline. Les NAW (North Atlantic Water) du Nord seront salées (> 37) comme les SAW mais plus froides (< 24° voire 20°C en hiver boréal) en surface et plus salées (> 35) et chaudes (> 7°C) en subsurface. Les eaux venant de l'Est seront pauvres en sel partout (35.6) quoique plus salées que les SAW en subsurface, mais très chaudes (près de 30°C) en surface (Bourlès et al., 1999 ; Urbano et al., 2008 ; Arnault et al., 2011).

La S peut être déduite des profils de T si les régimes d'eau de mer sont bien connus et qu'ils sont stables (Sverdrup, et al., 1942). Sous l'hypothèse de l'existence d'une relation entre θ-S, plusieurs auteurs proposent différentes techniques de calcul des profils de S et/ou de la hauteur dynamique de la mer quand seuls les profils de T sont disponibles (Reseghetti, 2007), même si de telles approches ont des bilans d'erreur importants. Le profil de T est également très utilisé dans les méthodes d'assimilation variationnelle.

Donc, bien que la relation T-S ne soit pas simple à retrouver, elle existe et dépend du temps, de l'espace et de la profondeur (pression. Donc à chaque classe (type) de profils de S correspondra une classe de profils de T dans nos classifications. En abordant cette relation comme un problème mal-posé, c'est-à-dire que ni l'existence, ni l'unicité d'une équation-relation n'est supposée encore moins une solution à cela, une approche basée sur l'apprentissage automatique est proposée. Elle consiste à intégrer le profil de T dans la base d'apprentissage afin de contraindre le profil de S à s'adapter au profil de T associé qui est celui mesuré au même moment et au même lieu.

La méthode d'inversion utilisant les profils de T a été appliquée sur les données Coriolis.

IV.7.1 Mise en œuvre de la base de données et du réseau pour l'inversion

En intégrant la T, une nouvelle extraction des données Coriolis a été faite. Les données varient maintenant de fin mars 2000 à fin juin 2012 avec la même couverture spatiale que précédemment c'est-à-dire de 40°S à 40°N et de 40W à 20E.

Un traitement particulier a été appliqué à ces données. Puisqu'elles n'ont pas les mêmes niveaux d'immersion, elles ont été ré-échantillonnées sur le profil vertical à l'aide d'une interpolation tous les 10m de 5 à 1100m. Puis la 2ème méthode de ré-échantillonnage défini en II.3.4 est appliquée aux nouveaux profils (S et T) afin d'obtenir les valeurs de S et de T aux mêmes 25 premières immersions que DRAKKAR (c.f. *Tableau II-1 : répartition des niveaux d'immersion de DRAKKAR*)

Le réseau de neurones plus précisément la carte de Kohonen est plus grande car non seulement les données ont augmenté mais le nombre de variables également dans l'apprentissage qui passe de 29 à 53 (24 T en plus). Ceci implique principalement 2 choses :

- ✓ La variabilité à exprimer est plus importante car le fait d'augmenter le nombre de dimensions d'un espace donne plus de possibilités de valeurs aux données donc la variabilité augmente. Pour prendre en compte cette nouvelle variabilité une augmentation du nombre de neurones dans la SOM est nécessaire;
- ✓ La carte convergera plus difficilement, le temps d'apprentissage doit être accru avec une augmentation du nombre d'itérations.

Pour mettre en place la carte SOM, la démarche reste la même que celle utilisée en IV.1. Les seuls changements concernent les valeurs des paramètres. Ici 40% des données centrées réduites sont utilisées pour l'apprentissage, 70 x 40 neurones soit 2800 sont choisis et 53 (25 premières immersions des profils de S et de T correspondant à celles de DRAKKAR, SST, ADT, Longitude, Latitude).

Type : SOM, carte de Kohonen.

Apprentissage : 40% des données centrées réduites sont utilisées pour l'apprentissage. Cette quantité d'individus est choisie en prenant une ligne sur 2,5 sans condition

spécifique sur un critère.

Nombre de neurones sur la carte : 70 x 40 neurones soit 2800 sont choisis. Ce choix fait suite à un 1er apprentissage automatique fait en utilisant un ensemble d'outils fourni par le SOM ToolBox sous Matlab. Il permet une bonne couverture de la variabilité de la base de données avec un temps de calcul optimal.

Voisinage : hexagonal => chaque neurone à 6 voisins.

Nombre de variables : 53 (25 premières immersions des profils de S et de T correspondant à celles de DRAKKAR, SST, ADT, Longitude, Latitude). Le nombre d'itérations a été augmenté à 50000.

L'apprentissage a été étudié de la même manière qu'en IV.5.2, et quasiment les mêmes résultats sont obtenus aussi bien sur l'organisation topologique de la carte SOM que sur les profils des neurones et ceux des données qu'ils ont captées malgré l'augmentation du nombre de variables.

Pour vérifier l'apport des profils de T à l'inversion, deux cartes SOM ont été utilisées: une carte sans le profil de T dans l'apprentissage (dénommée SomCorioS), c'est celle qui est utilisée jusque-là et une autre dans laquelle le profil de T est introduit dans l'apprentissage (dénommée SomCorioST). La Figure IV-44 présente les rmse de ces deux inversions sur les données Coriolis.

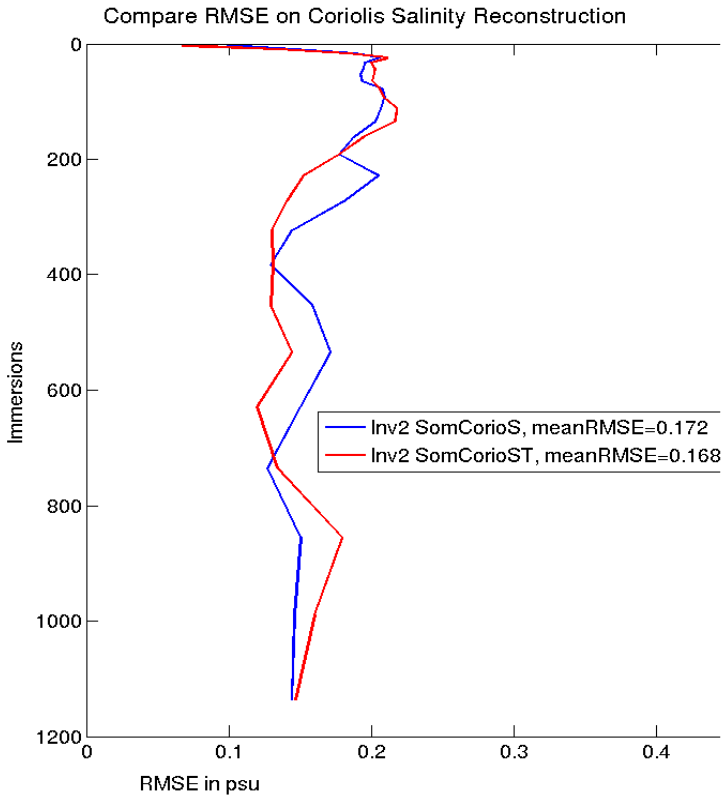


Figure IV-44: Comparaison des RMSE de la reconstruction des profils de S Coriolis en utilisant une carte SOM apprise avec les profils de T (Inv2_SomCorioST, en rouge) et une autre sans ces profils (Inv2_SomCorioS, en bleu).

Les deux profils de rmse présentent des valeurs inférieures à 0.2 psu sur toutes les immersions avec des valeurs moyennes de 0.172 pour la carte SomCorioS et 0.168 pour la carte SomCorioST. T apporte donc plus de précision dans la reconstruction. Même si cette différence est petite par rapport au nombre d'informations ajoutées à la carte qui est de 24 variables, elle reste significative. L'amélioration est due principalement aux niveaux d'immersions concernés par la diminution de l'erreur notamment de 200m à 700m.

IV.7.2 Comparaison des profils de S Coriolis vs Coriolis Inversé avec la carte SomCorioST

Les résultats présentés dans cette section concernent la reconstruction de profils de S Coriolis faite à l'aide de la carte SomCorioST, c'est-à-dire la carte SOM apprise en

ajoutant les profils de T.

La figure suivante présente les rmse moyennées sur une grille de 1x1°, toutes les dates confondues.

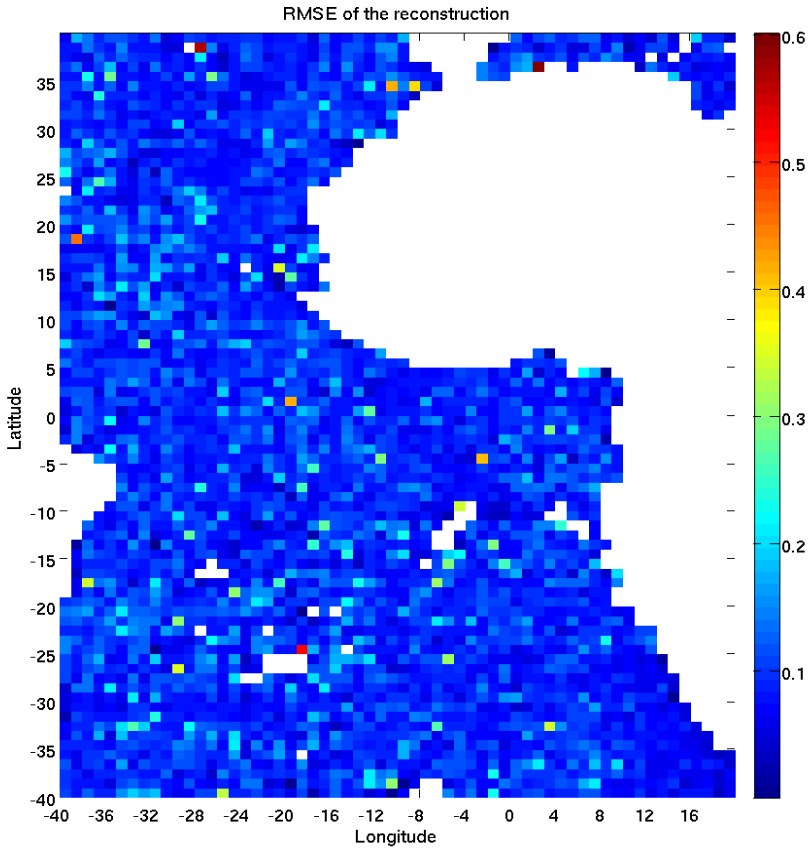


Figure IV-45: RMSE grillées de la reconstruction des profils de S Coriolis avec l'inversion utilisant la carte SomCorioST.

Chaque « point » de la figure correspond à un pixel 1° x 1° en longitude et latitude. Les rmse sont presque toutes inférieures à 0.1 psu. On ne note pas de tendance particulière des performances de la reconstruction en fonction de zones géographiques. Cependant il faut noter quelques rares points où les moyennes des rmse dépassent 0.2 psu. La Figure IV-46 donne la distribution des rmse.

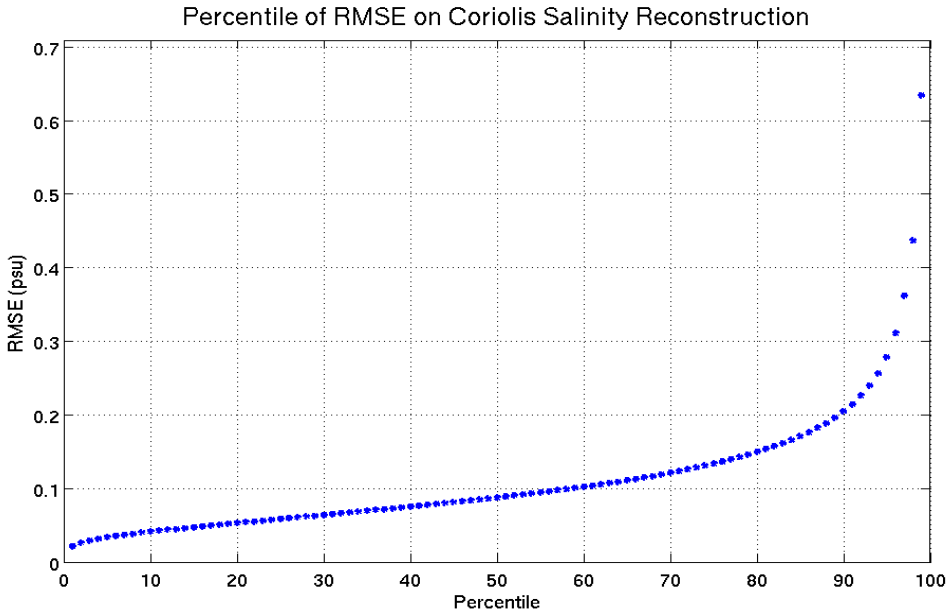


Figure IV-46: Distribution en Centiles des RMSE

Comme pour la figure lV.37 l'axe des ordonnées représente la rmse en psu et l'axe des abscisses représente les centiles qui correspondent au pourcentage de points dont la valeur de rmse est inférieure ou égale à ce qui est indiqué en ordonnée. Nous pouvons voir par exemple que 50% des profils ont été reconstruits avec une rmse verticale (faite sur le profil) < 0.1 psu et que 90% des profils de S ont été reconstruits avec une rmse < 0.2 psu. Ce qui confirme bien les résultats présentés par la Figure IV-45.

IV.7.3 Reconstruction des profils de S ARAMIS avec la carte SomCorioST.

Globalement la reconstruction avec la carte SomCorioST est meilleure aussi sur ARAMIS que la reconstruction faite avec la carte SomCorioS comme le montre la Figure IV-47 ci-après

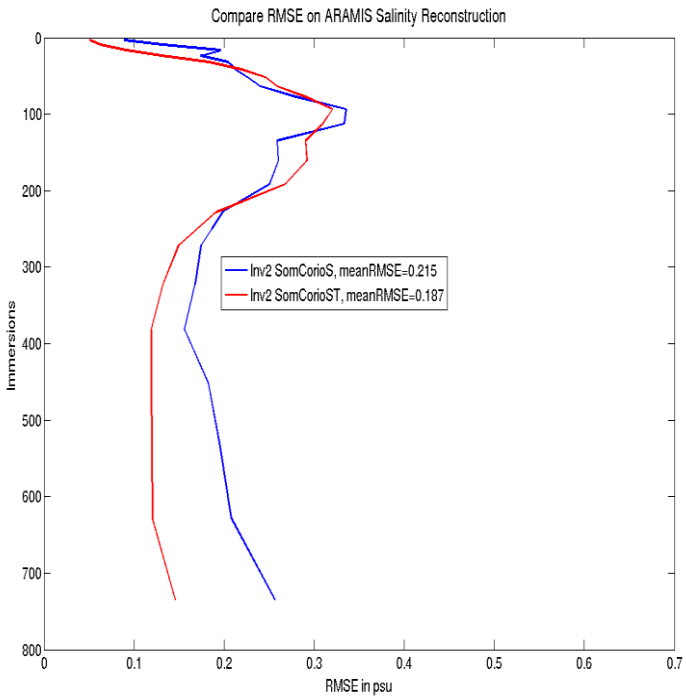


Figure IV-47:Comparaison des profils de RMSE: Reconstruction des profils de S ARAMIS avec SomCorioST en rouge et avec SomCorioS en bleu.

La reconstruction est meilleure sur toute la colonne d'eau de 250 à 800m de profondeur et sensiblement identique en surface. L'apport de T est plus visible que dans la figure IV.44 car la rmse passe de 0.215 à 0.187 psu soit de 0.02 psu, ce qui est significatif.

La répartition des valeurs de S sur certaines immersions caractéristiques a été étudiée sur cette reconstruction.

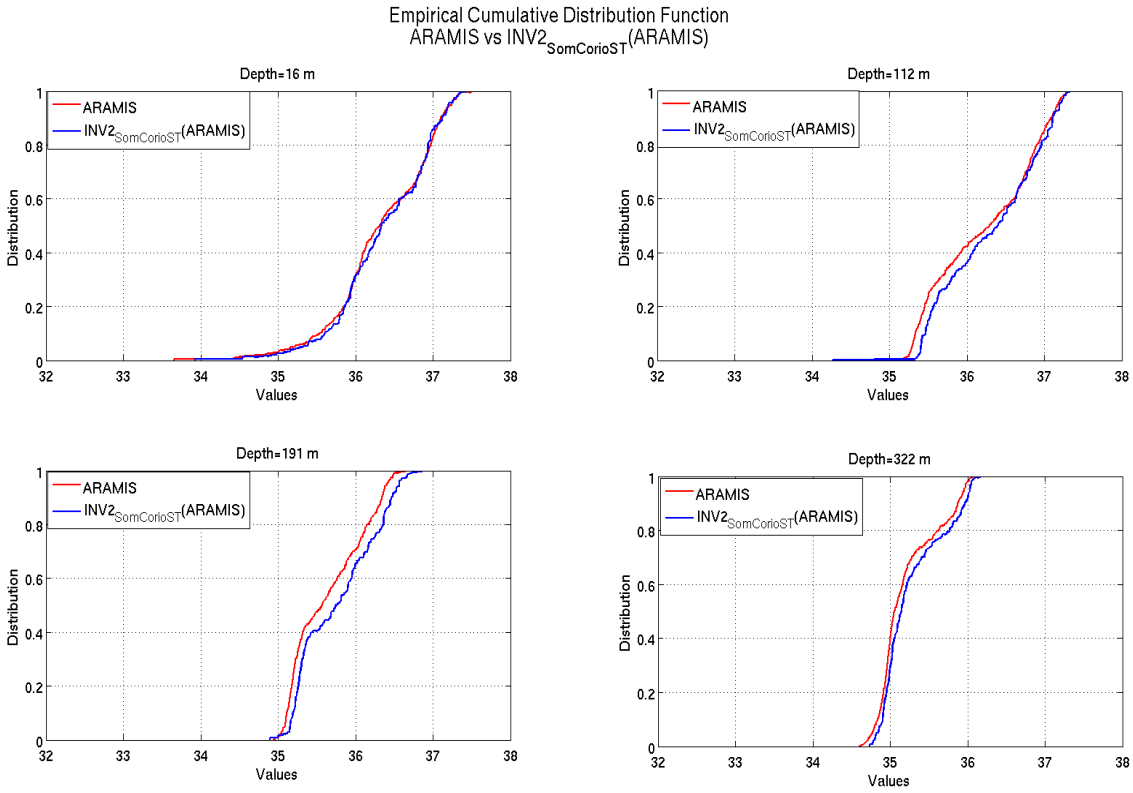


Figure IV-48:Empirical Cumulative Distribution Function ARAMIS vs INV2_SomCorioST(ARAMIS) sur les immersions caractéristiques (16, 112, 191 et 322m) de l'ensemble des situations rencontrées sur le profil, d'après l'analyse de données de S.

La Figure IV-48 présente les fonctions de distribution cumulée de ARAMIS et de son inversion. Le 1er niveau d'immersion (16m) montre une courbe aplatie de 33.75 à 35.75 psu avec des proportions inférieures à 20%, c'est-à-dire que moins de 20% des données ont des valeurs de S comprises entre 33.75 et 35.75. Cette situation vaut aussi bien pour ARAMIS « reconstitué » que pour ARAMIS « réel ». Puis, de 35.8 à 37.5 environ psu, le cumul monte assez linéairement à 100% pour une variation de 1.7 psu. Les immersions 112m et 191m offrent des situations quasi-identiques à l'exception du plateau asymptotique à 0 sur la partie basse de 112m qui, ne concernant qu'une donnée, reste négligeable. Les valeurs varient de 35 à 37.3 psu sur toutes les routes ARAMIS confondues, et quelle que soit la saison. Le modèle et ARAMIS sont en accord sur ces immersions même si on peut noter que le modèle a une tendance à dessaler légèrement de 0.1 psu à 191m. Enfin à 322m, zone homogène où S varie peut, les valeurs vont de 34.75 à 36 psu aussi bien sur le modèle INV2 que sur ARAMIS.

Cette étude sur les performances du modèle de reconstruction de profils de S a permis de mieux voir les capacités du modèle d'inversion proposé dans cette thèse. L'océan Atlantique, hautement variable « halinement » parlant, reste une zone difficile à modéliser. Rappelons aussi que S est une variable encore méconnue surtout en profondeur comparée aux autres variables telles que T.

Comme toutes données in situ, les données Coriolis présentent quelques problèmes liés à la qualité de la mesure et à la mesure elle-même. Dans le Chapitre V suivant, nous proposons une complétion de la base de données Coriolis sur l'Atlantique tropical et une étude géophysique de l'inversion par le biais d'analyses d'iso-halines et la reconstruction de quelques structures qui caractérisent l'Atlantique tropical.

Chapitre V. RECONSTRUCTION ET ANALYSE DE PROFILS DE SALINITE EN ATLANTIQUE TROPICAL.

Le chapitre précédent a analysé, statistiquement, les capacités des méthodes de reconstruction de profils de S que nous avons mises en place dans cette thèse. Ces résultats montrent que le modèle INV2 reconstitue les profils de S avec de faibles erreurs et de fortes corrélations. Cette analyse est valable aussi bien pour Coriolis que pour ARAMIS même si les résultats sont meilleurs avec Coriolis, ce qui est normal dans la mesure où le modèle d'inversion est construit avec les données Coriolis.

Dans ce présent chapitre, les résultats vont être analysés d'un point de vue plus géophysique que statistique dans un premier temps et ensuite une complétion de la base de données Coriolis pour la période 2003-2012 sera proposée. Le modèle d'inversion utilisé est INV2 avec la carte SomCorioST (profils de S + profils de T + variables de surface) qui sera désigné par $INV2_{SomCorioST}$ dans la suite.

Le chapitre est organisé en 2 parties. Dans la première partie, il sera question d'une étude plus géophysique des résultats obtenus avec la 2ème méthode d'inversion en se basant sur ARAMIS et dans la deuxième partie, nous proposerons une solution pour compléter la base de données Coriolis qui comptent beaucoup de profils de S de qualité moyenne ou mauvaise.

V.1 ETUDE GEOPHYSIQUE DE L'INVERSION : ANALYSE D'ISO-HALINE / RECONSTRUCTION DE STRUCTURES.

Dans cette section, nous présentons la manière dont le modèle d'inversion $INV2_{SomCorioST}$ reconstitue les structures halines dans l'Atlantique tropical, une zone à haute variabilité. Les données ARAMIS sont utilisées en guise de comparaison, elles sont intéressantes car elles constituent un bon échantillon dans la mesure où elles traversent toutes les zones dans cette région (cf. II.2) et elles restent indépendantes de l'apprentissage.

En Atlantique tropical, la variabilité climatique est assez particulière et ceci influe notamment sur S. Ainsi, il est important de comprendre la dynamique océanique en Atlantique tropical compte tenu de ses impacts sur d'autres régions avoisinantes. Par

exemple, il est maintenant établi des liens entre variabilité océanique Atlantique et climat du Nord-Est du Brésil, du Nord-Ouest de l'Afrique, de l'Amérique Centrale et des Caraïbes (Muñoz, et al., 2012). L'Atlantique tropical pourrait impacter aussi le Pacifique tropical ((Ding, et al., 2012 ; Losada, et al., 2010 ; Saravanan, et al., 2000) et l'Océan Indien (Kucharski, et al., 2008). Des réserves sont toutefois à prendre dans l'établissement de ces possibles téléconnexions. A cause de ses impacts sur toutes ces régions géographiques, l'étude de cette zone tropicale de l'Atlantique ainsi que sa compréhension demeurent primordiales et suscitent beaucoup d'intérêts dans la communauté scientifique aussi bien de la part des thématiciens que des modélisateurs.

Même si la zone tropicale de l'Atlantique a été reconnue comme une région importante dans le système climatique couplé, les phénomènes géophysiques s'y produisant sont difficiles à modéliser de manière adéquate même par les modèles climatiques couplés (Muñoz, et al., 2012). Cette difficulté est due principalement à des courants et des mouvements climatiques dépendant de plusieurs paramètres. Ces courants provoquent des échanges d'eaux entre les différentes sous-régions de l'Atlantique tropical. Ces échanges sont souvent marqués par des isoclines (isothermes pour la T et isohalines pour la S). Nous allons présenter une étude de la reconstruction de l'isohaline 35.2 psu et des structures halines sur certaines zones critiques.

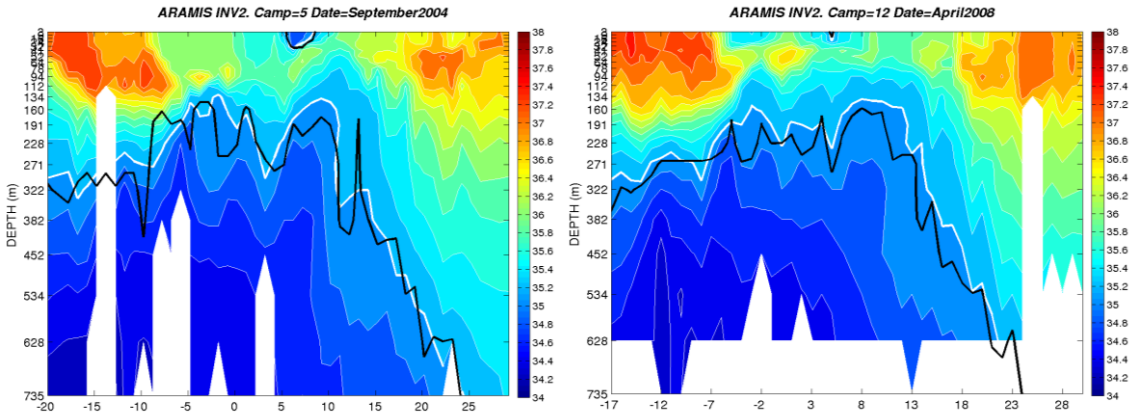
V.1.1 La reconstruction de l'isohaline 35.2 PSU

Une isohaline est définie comme une ligne reliant des points de même taux de S. L'isohaline 35.2 psu ainsi que sa reconstruction sont présentées dans la suite.

Les eaux de surface se déplacent vers l'intérieur de l'océan. La S venant de la surface de l'océan, transportée avec ces flux d'eaux permet d'identifier l'origine et la circulation de ces eaux. Dans la zone tropicale, l'isohaline 35.2 marque la fin de l'halocline, c'est-à-dire la profondeur de cette forte variation de S, elle trace ainsi l'halostad.

Le modèle d'inversion, $INV2_{SomCorioST}$ est appliqué aux données ARAMIS. Ce résultat est comparé à ARAMIS réelle sur 22 immersions. Nous rappelons que ces 22 immersions correspondent aux 22 premières immersions de DRAKKAR qui sont 3.05, 9.45, 16.36, 23.90, 32.21, 41.48, 51.95, 63.88, 77.62, 93.59, 112.28, 134.28, 160.28, 191.09, 227.62, 270.90, 322.02, 382.14, 452.44, 534.02, 627.85, 734.72m.

La Figure V-1 présente une section de S d'ARAMIS5 (a) et d'ARAMIS12 (b) le long de la route ARAMIS avec l'isohaline 35.2 issue des observations et reconstituée par l'inversion..



(a) ARAMIS 5 en septembre 2004 (b) ARAMIS 12 en mai 2008

Figure V-1: Section ARAMIS (5 et 12) avec mise en gras de l'isohaline 35.2 psu. Elle est mise en ligne blanche en gras et sa reconstruction par le modèle $INV2_{SomCorioST}$ est mise en trait gras noir.

Nous pouvons noter que les structures générales de variations latitudinales de l'isohaline sont bien reproduites. Sa profondeur moyenne l'est également. Pour ARAMIS5 (septembre 2004), l'isohaline 35.2 se situe vers 322m au sud de la ligne à 20°S, sa hauteur croit avec la latitude. Elle atteint 2 maxima à 130m environ à 2.5°S et 7.5°N puis descend brutalement en dessous de 750m à 25-30°N. Cette situation est quasi-identique à celle donnée par l'inversion même si sur quelques latitudes, on peut noter de petits décalages notamment vers 10°N à 160m où l'inversion sous-estime l'isohaline d'un niveau (191m). Des traces de cette isohaline sont visibles en surface autour de 7° N marquant ainsi la présence de l'ITCZ qui est localisée, en début d'automne (septembre) vers 6–8°N (Montégut et al., 2007). La conséquence de cette présence de l'ITCZ est le rafraîchissement des eaux de surface apportées par le NECC à l'intérieur du bassin. Ces traces de l'isohaline sont bien reproduites par le modèle pour la campagne ARAMIS5. En début de printemps, l'estimation de cette isohaline est encore meilleure pour ARAMIS12 (avril 2008). L'allure générale est mieux respectée. L'isohaline est aux alentours de 310m vers 20° S, dans la zone de SMW du sud, même si à ce niveau la profondeur de l'isohaline est un peu surestimée par le modèle. Elle monte ensuite en faisant des oscillations jusqu'à

environ 100m où elle atteint son maximum global vers 2°S. On note aussi un maximum local vers 200m autour de 8°N. L'isohaline descend brutalement en dessous de 800m vers 25°N. Cette situation est bien reprise par l'inversion sur toute la ligne ARAMIS 12. Comme, déjà signalé, la reconstruction de cette isohaline pour la campagne 12 d'ARAMIS est meilleure que celle de la campagne 5. Cependant on peut noter quelques imperfections notamment vers 3°S, l'isohaline est estimée à 271m par le modèle alors qu'elle est localisée vers 191m en réalité, soit 2 immersions plus en haut. Or, dans cette zone se trouve la branche centrale du SEC (Araujo et al., 2011) donc la S à ces niveaux (entre 190m et 300m) y dépend moins de la SSS qui est un des paramètres les plus importants du modèle d'inversion. Cette zone est très proche des côtes brésiliennes avec une faible SSS ; comme le modèle d'inversion est fortement lié à la SSS, il a tendance à dessaler légèrement à ces latitudes.

L'isohaline de 35.2 psu reste globalement bien reconstruite sur toutes les campagnes ARAMIS, même si, dans les détails, il existe quelques différences dans la reproduction parfaite de cette isohaline surtout dans les zones critiques. Ci-après sont présentées les performances de l'inversion dans quelques zones critiques.

V.1.2 Inversion des données ARAMIS sur quelques zones critiques.

Les courants océaniques sont des flux d'eau froide ou chaude, superficiels ou profonds (sous-courants). A la surface des océans, des bandes d'eau de 50 à 500 km de large et de quelques centaines de mètres d'épaisseur, entraînées par les vents dominants, tournent sans cesse. En profondeur, les mouvements d'eau se créent en raison de la différence de densité entre les diverses couches de l'océan. Cette différence de densité est fonction de T (l'eau plus froide est plus dense, et donc descend plus en profondeur) et de S (l'eau plus salée est plus dense que l'eau douce) des masses d'eau. À l'échelle du globe, les courants marins de profondeur constituent la circulation océanique profonde (ou circulation thermohaline). La circulation océanique dans les océans tropicaux se compose principalement de courants zonaux au large et d'intenses courants limitrophes près des côtes (Arnault, 1987) .

Une des difficultés, bien connue, de la modélisation des phénomènes physiques en Atlantique tropical est due principalement à la représentation de ces courants et à des

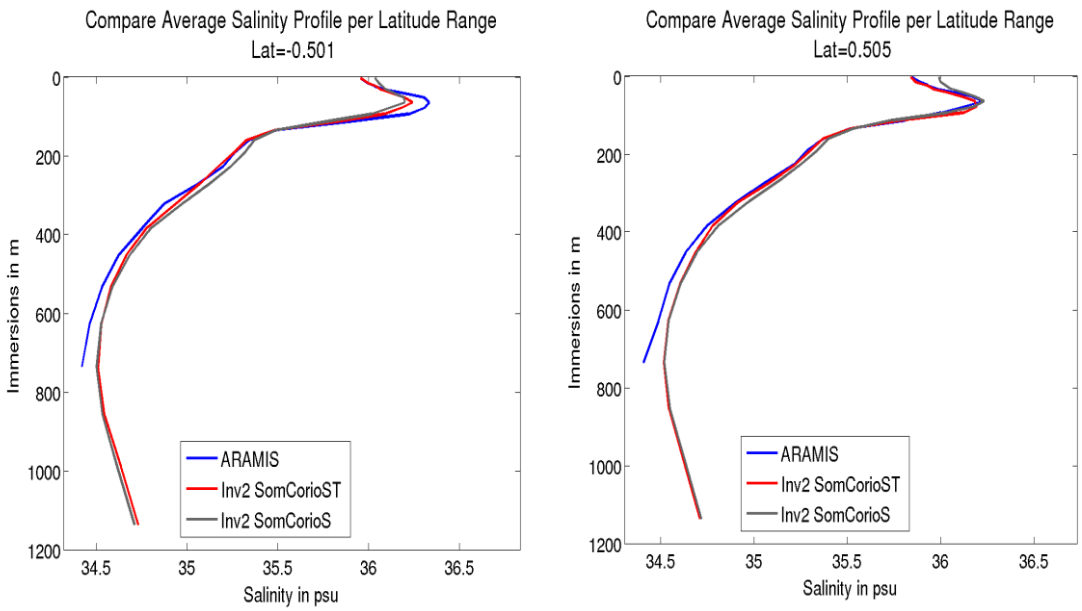
mouvements climatiques instables sur certaines zones.

Dans ce paragraphe, nous présentons les résultats du modèle $INV2_{SomCorioST}$ sur certaines de ces zones "critiques" à savoir :

- ✓ la région du EUC autour de 0.30°S ;
- ✓ et les 2 zones des SMW (Nord autour 25°N et Sud autour 15° S).

V.1.2.1 Région du Sous-Courant Equatorial ou EUC.

De nombreux éléments de preuve sont établis sur l'existence d'un sous-courant équatorial en Atlantique (EUC) (Neumann, 1960 ; Metcalf et al., 1962). Dans l'océan Atlantique, la découverte de ce sous-courant doit être attribuée à Buchanan (1886) qui a mesuré, vers 55m de profondeur, des courants en direction sud-est plus rapides que le déplacement des eaux de surface (Neumann, 1966). Mais c'est (Metcalf et al., 1962), qui a fourni, le premier, la preuve définitive de son existence et décrit ses principales caractéristiques. Ces auteurs ont présenté des distributions équatoriales transversales de T, de S et de l'oxygène. Le EUC est la manifestation la plus spectaculaire de la particularité équatoriale. C'est un courant très rapide (1 m/s) en forme de jet étroit (200m d'épaisseur, 300 km de largeur) qui est dirigé vers l'est, sous le Courant Equatorial Sud (SEC pour South Equatorial Current). On l'observe le long de l'équateur à une profondeur de 100 ou 150m dans les océans Atlantique et Pacifique (Frankignoul, 2012). Ce sous-courant comme son nom l'indique est localisé autour de l'équateur, la figure suivante présente les profils moyens ARAMIS et leur reconstruction au niveau de ce courant.



(a) (b)

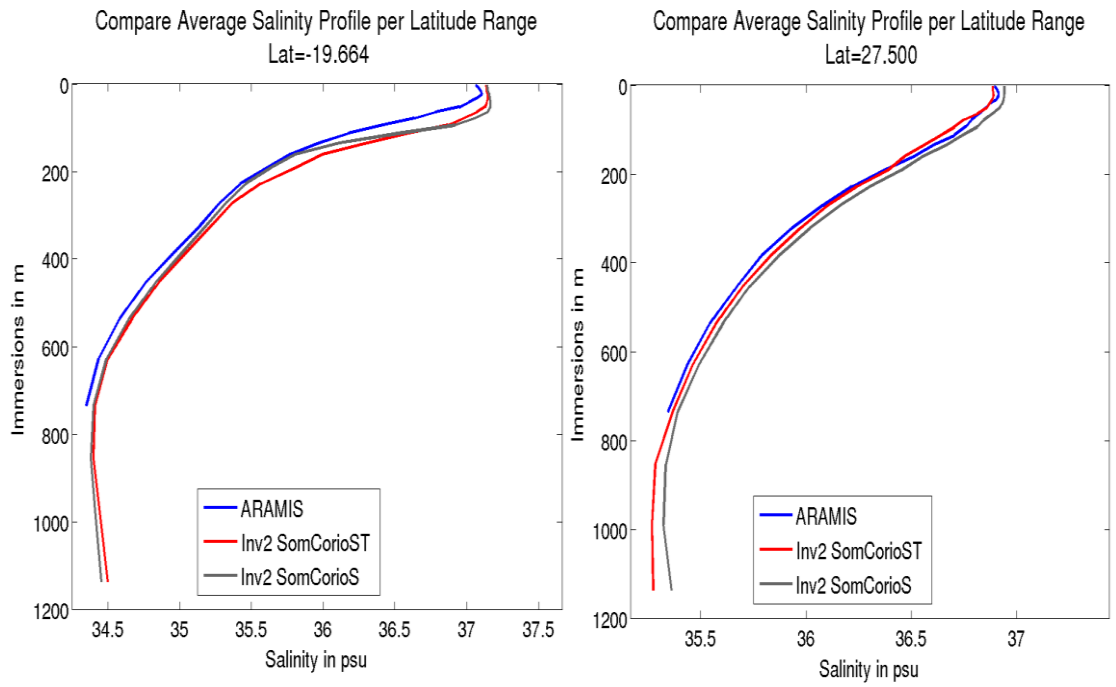
Figure V-2:Comparaison de profils moyens de S sur la région du EUC (a) à 0.5°S et (b) 0.5°N. La couleur bleue indique ARAMIS réelle, rouge : profil reconstitué en utilisant la carte $INV2_{SomCorioST}$ *, grise :reconstitué utilisant la carte* $INV2_{SomCorioS}$

Les auteurs qui ont décrit en premier ce sous-courant (EUC) notent des flux d'eaux salées et riches en oxygène qui coulent vers l'est le long de l'équateur, entre 1 ° S et 1 ° N et de quelques dizaines de mètres à environ 150m de profondeur, avec des vitesses maximales au-delà de 1 $m.s^{-1}$ vers 60m (Claret et al., 2012). Cette situation est reproduite aussi bien par ARAMIS que par l'inversion. Les maximas de S sont notés entre 75 et 175m de profondeur. On peut noter que le modèle dessale un peu à ce niveau d'immersion. Mais l'allure générale de la courbe reste bien reproduite. On remaruqe à nouveau que les profils moyens dans cette zone sont mieux modélisés par l'inversion utilisant la carte $INV2_{SomCorioST}$.que par celle liée aux S uniquement.

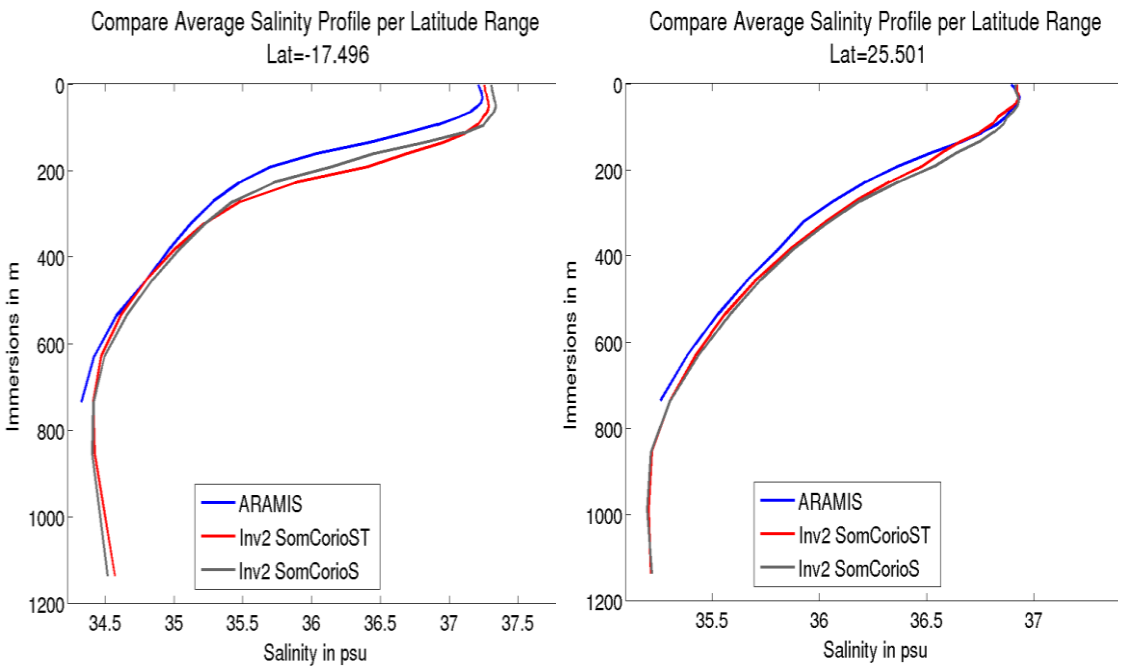
V.1.2.2 Zones des Salinity Maximum Waters.

Les SMW, comme le nom l'indique sont les zones de S maximale formées par évaporation intense. En Atlantique tropical, il existe 2 zones de SMW véritables marqueurs d'anomalies climatiques dans leur hémisphère respectif ; une au Nord entre 20 et 30° et l'autre au Sud entre 15 et 25° comme illustré par la Figure II-3. L'étude de ces zones n'est pas une chose facile dans la mesure où il est rare de trouver dans la littérature des

documents dédiés à ces zones. (Blanke et al., 2002) notent ce manque de bibliographie surtout en Atlantique tropical. Ils remarquent que la seule étude dédiée aux SMW est de O'Connor (O'Connor et al., 1998) dans l'Océan Pacifique. Pour l'Atlantique, on peut avoir des informations partielles à partir des observations et de récents travaux sur les Sub-Tropical Cells (STC) qui subduquent les anomalies T-S vers l'équateur et constituent des éléments importants de la circulation thermohaline (Hazeleger et Drijfhout, 2006 ; Goni, et al., 2003) mais nous n'avons pas vu d'études dédiées réellement à leur analyse spatio-temporelle. Les deux zones de S maximale de surface sont centrées sur des longitudes très proches (30°W et 35°W), et aux latitudes (25°N, 18°S) qui sont symétriques par rapport à la zone de convergence intertropicale (ZCIT). Dans ces zones, les plus fortes valeurs de S sont rencontrées dans les couches de surface, comme le montre la Figure V-3. Cette situation est assez bien reproduite par le modèle d'inversion.



(a) (b)



(c) (d)

Figure V-3: Profils moyens de S sur les eaux de S maximum au sud ((a) 19.66° S, (c) 17.5°S) et au nord ((b) 27.5°N, (d) 25.5°N). La couleur bleue indique ARAMIS réelle, rouge : profil reconstitué en utilisant le modèle $INV2_{SomCorioST}$ *, grise :reconstitué utilisant la carte* $INV2_{SomCorioS}$

Le modèle est capable de reproduire les fortes valeurs de S associées aux MSW (> 37 psu) en surface ainsi que l'allure générale du profil. On peut noter aussi que les S de sub-surface au Nord sont plus grandes que celles du Sud. ARAMIS et modèle sont en accord car ils donnent à ces immersions des S moyennes autour de 34.5 psu au Sud et 35.2 psu au Nord.

V.2 COMPLETION DES DONNEES CORIOLIS EN ATLANTIQUE TROPICAL SUR LA PERIODE DE 2000 A 2012.

V.2.1 Introduction

En statistique, la notion de « données manquantes » ou « valeurs manquantes » se produit lorsqu'aucune valeur n'est stockée dans une variable pour une observation (ou individu) donnée. Les données manquantes sont fréquentes surtout dans les mesures in situ et peuvent avoir un effet significatif sur les conclusions qui peuvent être tirées des données. Elles peuvent être de diverses origines. On peut citer l'impossibilité de la mesure de la valeur, la perte de la valeur mesurée, la valeur mesurée non utilisable car jugée aberrante. Ces problèmes liés à l'acquisition de la bonne valeur d'une donnée concernent plus particulièrement les sciences de l'environnement et notamment la mesure des données de profondeur des océans telles que la salinité de sub-surface. Ainsi, il s'avère très utile de compléter les données à chaque fois que c'est possible.

Compléter des ensembles de données collectées dans des environnements hétérogènes est un problème assez fréquent (Polasek et al., 2010). La difficulté réside surtout dans le respect de la consistance des données. Ce respect consiste à garder la plage de données telle quelle c'est-à-dire que les nouvelles valeurs qui servent de compléments ne doivent pas être elles-mêmes aberrantes.

Il existe plusieurs approches dans la complétion de données. Parmi elles, nous pouvons citer l'interpolation, l'extrapolation, la distribution qui sont des approches empiriques et d'autres approches plus théoriques qui consistent à considérer ces problèmes de complétion comme des problèmes mal-posés au sens de Hadamard. Chow et Lin (1971) sont les premiers à développer une plateforme unifiée d'approches empiriques basée sur l'interpolation, l'extrapolation et la distribution.

Cette partie est constituée de l'article suivant.

A neural approach for salinity profile completion using a recursive algorithm.

Mbaye Babacar Gueye[a,b] , Awa Niang[a], Sabine Arnault[b], Sylvie Thiria[b,c],

[a]*Laboratoire de Traitement de l'Information/ESP,Université Cheikh Anta Diop, BP 5085, Dakar Fann, Sénégal*

[b]*Sorbonne Universités, UPMC Univ Paris 06, LOCEAN/IPSL, F-75005, Paris, France*

[c]*UVSQ, 78035, Versailles, France*

Abstract

Complementing data sets collected in a heterogeneous environment is a frequent problem encountered in many scientific domains. We present a recursive method based on a neural network model for complementing ocean salinity profiles. The method is applied to the tropical Atlantic observations provided by drifting buoys sampling the ocean. The proposed method utilizes a Kohonen Self-Organizing Map based model to select a set of possible salinity profiles and then complements the missing values of the concerned profile by using a recursive algorithm. The performance of the method is checked with respect to the percentage of missing data. The completion misses some salinity features in areas with high time-space variability for which the limited existing dataset was unable to provide the complete variability ranges during the learning process. However good performances of the completion are observed both in near-surface immersions as well as in the deeper ones in cases for which the percentage of missing values is less than 50%. The methodology demonstrated reliable results with a fairly good overall agreement between the complemented data and actual measurements

Introduction

This paper presents an application of a robust non parametric method, the Self Organizing Map (SOM) (Kohonen, 1989; Kohonen, 2001), to complement a multidimensional database polluted by large patches of missing data. We demonstrate the efficiency of the method by complementing an actual database constituted by oceanic measurements provided by the french Coriolis data center (http://www.coriolis.eu.org). Such a problem frequently occurs with geophysical archive of experimental in-situ measurements (Kondrashov et al., 2006; Stefanakos et al, 2001). For these data, a lot of difficulties may arise due to bad weather conditions or malfunction of sophisticated sensors. However, these data are particularly critical for the scientific community that carries out research related to actual environmental problems such as those dedicated to management studies in the context of the societal impacts of global warming. These environmental data are collected and distributed by data centers in which provide friendly software to the scientific community to process the data and extract pertinent information. In this context, an important quality control must be carried out before making the data available. Given the difficulty in obtaining in situ data, one must do our best to keep the maximum amount of data available. The best way to provide the largest database is to apply advanced statistical methods to complement the missing items of incomplete records.

The completion of missing data is a general problem and many techniques for recovering them have been proposed in literature. For example in the remote sensing area, a standard one is the construction of spatial or temporal composite images obtained by combining images from multiple sensors or aggregating them over time conserving their major statistical properties (Reynolds and Smith, 1994, Smith and Reynolds, 2003). Some methods for handling missing values such as optimal analysis based techniques or advanced statistical tools such as wavelets (Pottier et al., 2008, Beckers and Rixen, 2003) were applied with success. Other methods belonging to machine learning field, such as Self Organizing maps, which are efficient unsupervised neuronal classifiers, have been used to exploit the homogeneity of underlying data classes for data completion (Cottrel and Letremy, 2007, Jouani et al., 2013).

The Self Organizing Map algorithm, the so-called *SOM*, proposed by Kohonen (2001) constitutes a powerful nonlinear classification tool. It aims at clustering samples of a multidimensional database *D* into classes represented by a fixed network the *SOM* map. *SOM* enables the partition of *D* in such a way that each subset is associated with a node of the map and is represented by a referent that is a synthetic multidimensional sample similar to the initial data. Each vector ***v*** of *D* is affected to the node whose referent is the closest, in the sense of the Euclidean Norm (*EN)*. This referent is denoted the Best Matching Unit (*BMU*) and represents the projection of the vector *v* on the map. Several modifications of the standard *SOM* algorithm have been proposed in the case for which *D* has incomplete data, to estimate the missing components. The *SOM* algorithm we propose can be decomposed in the following manner:

- *SOM*_Comp_1: The referent vectors of the *SOM* are determined through a learning algorithm using the complete records of D only. At the end of the learning process, the vectors that exhibit missing components are projected on *SOM* and complemented by using a truncated distance. This distance between an incomplete vector ***v*** and the referent vectors is the Euclidean Distance that considers only the existing components of ***v***; it is denoted Truncated Euclidian Distance hereinafter (*TED*). First the vector ***v*** determines the *BMU* with which it is associated by using the *TED* and then replaces its missing components by the corresponding component of the *BMU*.

- *SOM*_Comp_2: A modification of the learning phase permits to improve this methodology in the case for which the percentage of missing values is not too large. This is done by including in the learning dataset, the records that have less than a threshold number of missing components. At each iteration, the BMU is determined using the Euclidian distance (*ED*) or the *TED* depending on whether or not a vector has missing values. Then the algorithm proceeds as usual for the update of the referent vectors. At the end of the learning phase the vectors with missing components are complemented by using the values of their referents.

In the present paper we deal with an actual problem related to oceanography in which the number of missing data by record may reach 100%. More specifically, we focus on the

relationship that links the observed surface data to the deep ocean data. An inverse method (*SOM_SPI*) based on *SOM* (Gueye et al, 2014) has been recently proposed that allows inverting the sea-surface data remotely sensed by satellite to obtain the associated vertical profiles of salinity. *SOM_SPI* is calibrated during a learning phase using the in-situ observations done by drifting buoys (ARGO floats) equipped of salinity and temperature sensors sampling the ocean. They are distributed by the French Coriolis Data Center. Due to the complexity of the problem, the accuracy of the inversion highly depends on the quality of the database. Increasing the number of in-situ measurements can provide a substantial improvement for the reconstructed profiles. We first present the oceanographic problem that seeks to reconstruct salinity profiles using surface data provided by satellite remote sensing sensors and the Coriolis database that presents a large amount of missing data. The second section is dedicated to the presentation of the methodology, the third shows the performances and the last section is dedicated to conclusion and perspective.

The Oceanographic Problem

The emergence of operational oceanography in the last 10 years together with the increased number of multi sensors and multi format oceanographic data assembly centres implies to:

- integrate data coming from a wide variety of platforms and providers (including scientists, national data centres, satellite data centres and operational agencies);

- get enough information from the originators to be able to know exactly how the data have been acquired and processed;

- distribute them in an agreed standard format and with high quality control.

Coriolis (www.coriolis.eu.org) is one of these Data Assembly Centers that has originally been designed to provide real-time qualified and integrated products to the French ocean forceasting centers such as MERCATOR-Océan and French Hydrographic Service (SHOM). Coriolis is a joint contribution of seven institutes in France (CNES, CNRS,

Ifremer, IPEV, IRD, Météo-France, SHOM) in order to organize and maintain data acquisition in real-time and delayed modes of in-situ measurements needed for operational oceanography. The Coriolis centre distributes several types of data from different origins.

The nature of the measurements and the experimental difficulties associated with the observation processes make that these data may present different accuracies, sampling rates, observation depths. This explains why the Coriolis data base contains a large amount of missing data.

The constraints of scientific research imply a high quality control of these data. Given the importance of salinity in oceanographic studies associated with climate change and their societal impacts, it is of major importance to complement the Coriolis data files at best. Indeed the number of accessible measurements is a major factor for conducting high quality scientific studies in oceanography. In the present study we used the data provided by the ARGO floats.

The data we have extracted from the Coriolis database are localized in the Tropical Atlantic Ocean in an area ranging from 20°E to 40°W and from 40°N down to 40° S and during the time period ranging from January 2000 to April 2012. Each datum is characterized by 6 fields: the observation date, the latitude, the longitude, the number of observation levels, the temperature and the salinity profiles. As the number of observation levels is variable, the temperature and salinity profiles are associated with vectors whose dimension is variable. The sampling may also be variable; for some profiles the depth resolution is 5m while it is 10 m for others or simply irregular. Since most of statistical software deal with arrays whose dimensions are fixed, the profiles were linearly interpolated in order to obtain data on a regular grid whose depths correspond to the 25 upper depth levels used in the DRAKKAR experiments (see table 1 for the definition of the levels or the web site: (http://www.ifremer.fr/lpo/DRAKKAR/index.htm). DRAKKAR project is a scientific and technical coordination project between French research teams and other European and Canadian teams to design, carry out, assess, and distribute high-resolution global ocean/sea-ice numerical simulations based on the NEMO platform (www.nemo-ocean.eu), performed over long time periods (five decades

or more), and to improve and maintain a hierarchy of state-of-the-art ocean/sea-ice model configurations for operational and research applications.

Each temperature and salinity profile is therefore given by a 25 component vector after the depth interpolation to the DRAKKAR immersions.

Table 1: Depth levels of the DRAKKAR model used in the interpolation

Level	Immersion (in m)	Level	Immersion (in m)
1	3.05	14	191.09
2	9.45	15	227.62
3	16.36	16	270.9
4	23.9	17	322.02
5	32.21	18	382.14
6	41.48	19	452.44
7	51.95	20	534.02
8	63.88	21	627.85
9	77.62	22	734.72
10	93.59	23	855.11
11	112.28	24	989.23
12	134.28	25	1136.92
13	160.28		

Quality variables were associated with each profile provided by the Coriolis Data Assembly Center in order to rate the quality of the measurements (on a scale ranging from A to F). Table 2 gives the significance of each flag.

Table 2: Control quality table of Coriolis Data Assembly Center (N corresponds to the percentage of good data in a profile)

Flag QC	Signification
A	N=100%
B	75% <=N<=100%

C	50%<=N<=75%
D	25% <=N<=50%
E	0% <=N<=25%
F	N= 0%

In the following, we only used datasets whose missing data percentage is less than 50% (flags ranging from A to C). This dataset is denoted $Coriolis_{ABC}$, and contains 74599 data.

Figure 1 shows the time and space distribution of the $Coriolis_{ABC}$ dataset. The number of profiles per box of (1°longitude x 1°latitude, Fig.1a) is usually less than 50 except for a few boxes having up to 250 profiles. Besides we note an increasing number of profiles as the time is progressing (Fig.1b). The period between 2000 and 2012 is decomposed in three periods: a first one between 2000 and 2005 with less than 500 profiles per month, a second one between 2006 and 2008 with 500-600 profiles per month and a third period beyond 2008 with more than 800 profiles per month. The time repartition reflects the importance of the international ARGO project (http://www.argo.ucsd.edu/) in the oceanographic databases. Starting in 2002, ARGO is now a global array of more than 3,000 free-drifting profiling floats that measure temperature and salinity of the upper 2000 m of the ocean. Thus, as time is progressing it should become easier to reconstruct long time series, which must be also more robust.

Besides we also added the dynamics topography of the sea surface as a pertinent information for retrieving the temperature and salinity profiles. The dynamics topography is observed by satellite altimeters: it is provided by the AVISO data center (http://www.aviso.oceanobs.com). It consists in satellite Ssalto/Duacs Absolute Dynamic Topography (ADT) merged products.

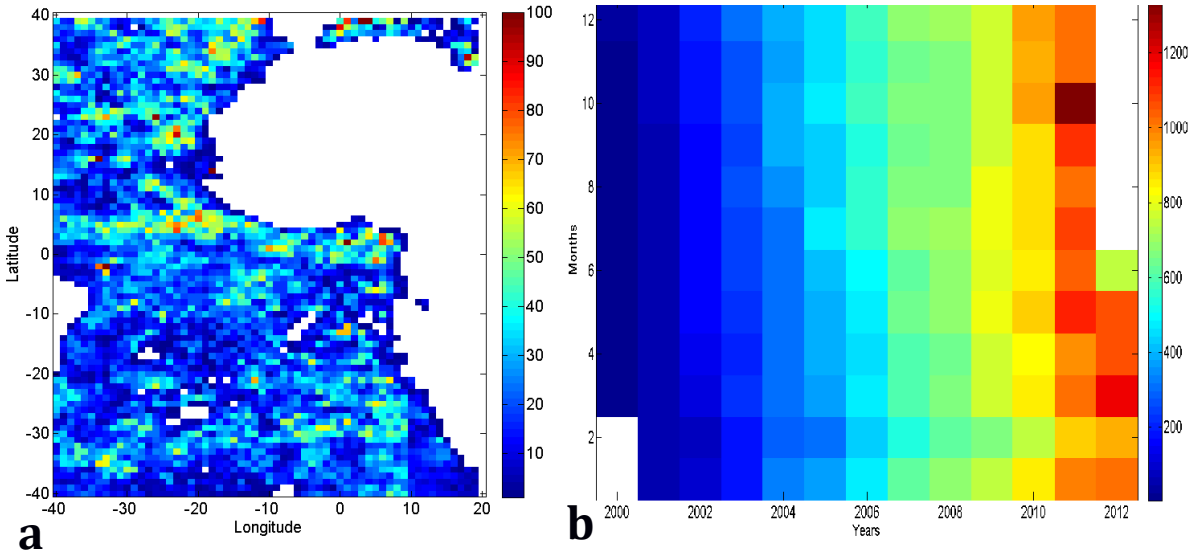


Figure 1: Spatial (a) and temporal (b) distribution of the Coriolis$_{ABC}$ dataset.

The data from all altimeter missions (Jason1 and 2, TOPEX/Poséïdon, ENVISAT, GEOSAT Follow On, ERS1 and 2, and GEOSAT) were processed by the Ssalto/Duacs system in order to provide a consistent and homogeneous catalogue of products. The series of weekly dynamics topography maps are distributed with a grid resolution of 1/3°x1/3°.

Complementation Implementation.

Completion permits to complement the salinity profiles in most cases even in difficult cases such as in flag C for which missing data percentage can reach 50%. An inversion method for retrieving salinity from sea surface data has been recently proposed by Gueye et al., 2014, which is denoted SOM-SPI. The data processed by this method are: salinity (S) profile, Sea Surface Temperature (SST), longitude, latitude and SSH provided by DRAKKAR. The basic algorithm used by the SOM-SPI method to retrieve the optimal referent has been adapted to the present situation to process the Coriolis observations. This new algorithm is denoted SOM-STI in the following.

The Salinity Data Completion Method proposed here (SDCOMP) is a two-step process: the first step consists in determining the referents associated with the neurons of the SOM.

This is done during a learning phase which processes the $Coriolis_{ABC}$ data set; the second step complements the missing data by selecting the values from a set of salinity profiles given by SOM-STI as described hereafter.

First step : SOM-STI:

1-The completion method first starts by a learning of the $SOM\_Coriolis_{ABC}$ map on the Coriolis data set, which have missing data. In this step, the learning Coriolis dataset (CoriolisApp) is constituted of 28668 elements (2/5 of $Coriolis_{ABC)}$). Some profiles can have up to 50% of missing data. Each sample of the training set represents a geophysical situation with 29 components: 25 *S* (S profile), and SST, Longitude and Latitude from the Coriolis data, ADT provided by AVISO. These last four data are surface parameters.

As mentioned above, at the end of the training, the $SOM\_Coriolis_{ABC}$ map is calibrated; its referent vectors have statistical characteristics similar to those of the learning set and constitute a compression of the Coriolis data set.

2- SOM-STI: After the training phase, we start the analyze phase. The vector to be analyzed is projected onto SOM. The adequate neuron, which is the closet to this vector, is selected by doing two distinct projections and then a selection phase.

- The pair (SSS, Latitude) is first projected using the *TED* and a subset of 25 best referents is determined.
- The second projection, that concerns the all remaining parameters - the surface parameters (SST, ADT, Longitude) as well as the available S values - selects a second subset of 40 best referents
- The referent that presents the best compromise between both subsets is selected.

Second step: Missing data Completion

SOM-STI is first applied to the Coriolis database to retrieve the appropriate referent (neuron N*) for each S profile. To complement a Sp profile with missing data, the following recursive algorithm is used.

1) Identify the winning neuron (N*) associated with Sp and all S profiles won by N* (En*) using SOM-STI.

2) Find the closest profile to Sp in terms of *TED* (denoted as Sp1),

3) Replace the Sp missing values with those of Sp1 at concerned depths

4) If there are missing values in Sp1 at these immersions, then steps 2) and 3) are repeated with Sp1.

5) Repeat steps 2, 3 and 4 until all missing values of Sp have been complemented or all elements in En* have completed steps 2, 3 and 4.

We note that if only surface data are available, no flag F (with 100% of missing data) profile can be complemented by this method.

Performances of SDCOMP

To validate the SDCOMP method, we compute some major statistics estimators over the completed database and the initial one. They must be equivalent on the two data bases. The initial data set, Coriolis_Val, is a set of profiles without missing data. The learning phase of SOM_STI was done on 2/5 of $Coriolis_{ABC}$. Thus Coriolis_Val was created using the remaining 3/5 of $Coriolis_{ABC}$ but keeping only flagged A data: the 25 S measurements are present in all profiles. Coriolis_VAL contains 38 467 data.

Knowing that the presence of missing data is mainly due to problems associated with measuring instruments, we assumed that the data are MCAR (Missing Completely At Random), in the sense that the probability that a value is missing does not depend on the missing value (Rubin, 1976, M. Ghannad-Rezaie et al., 2010). We ran several experiments with an increasing number of missing data in order to simulate the four flags (B, C, D, E) datasets that the Coriolis data center provides to users (see Tab. 1). We therefore constructed four datasets denoted Coriolis-GapB, Coriolis-GapC, Coriolis-GapD, Coriolis-GapE composed each of 38467 data simulating the four flags. For each S profile belonging to Coriolis_Val, we have randomly chosen a number n whose value is within the gap interval corresponding to the percentage associated with the missing data flag (example : for Coriolis –GapB, we have $0{,}75<n<1.0$), in such a way that the number of depths with good data is equal to 25xn. The four Coriolis_Gap datasets we have built, are incomplete datasets that we have complemented by using SDCOMP. The datasets

with the exact profiles corresponding to the Coriolis_Gap datasets are denoted Coriolis_ActualB (C, D, E repectively).

Table 3 gives the percentage of completed data for the different flags defined in tab. 2.

Table 3: Percentage of complemented values with respect to missing values for the different flags defined in tab. 2

Flag	B	C	D	E
Total number of missing data in Coriolis_Gap	115 501	327 237	557 762	769 267
Total number of missing data after completion	6 944	7 871	12 102	72 744
Percentage	93,9879	97,5947	97,8303	90,5437

The completion gives the highest rates for flags C and D. As the total number of missing data increases with the flags, this result shows that the completion algorithm have processed almost all the gaps.

Figure 2 shows for each Coriolis_Gap dataset, the RMSE corresponding to the reconstructed profiles. A continuous increase of the error with the missing data rate is observed but it remains below 0.14 in practical salinity unit (psu) in the case for which the percentage of missing data rate is up to 85% in average. We have clustered the depths into four depth intervals ([3m ; 52m], [63,9m ; 191m], [227,6m ; 322m], [382,1m ; 1136,9m]). Figure 3 shows the RMSE as a function of depth corresponding to the four flags (B, C, D, E) dataset.

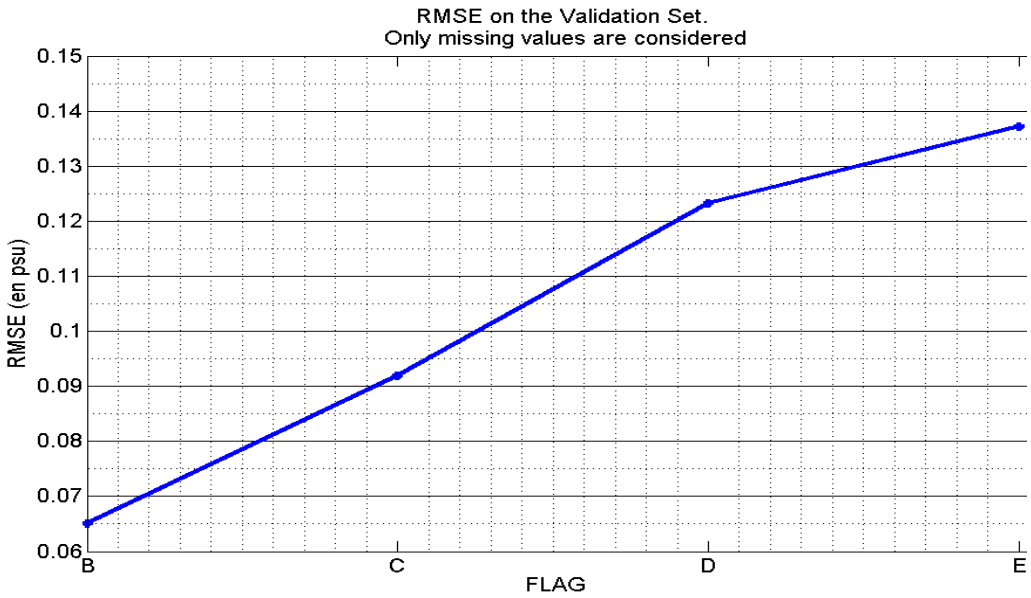


Figure 2: Performances of the SDCOMP method. The horizontal axis corresponds to the percentage of missing data (Flags B,C,D,E) for the different profiles, the vertical axis to the averaged RMSE computed on the reconstructed depth levels of 30 salinity profiles with the same missing data rate chosen at random.

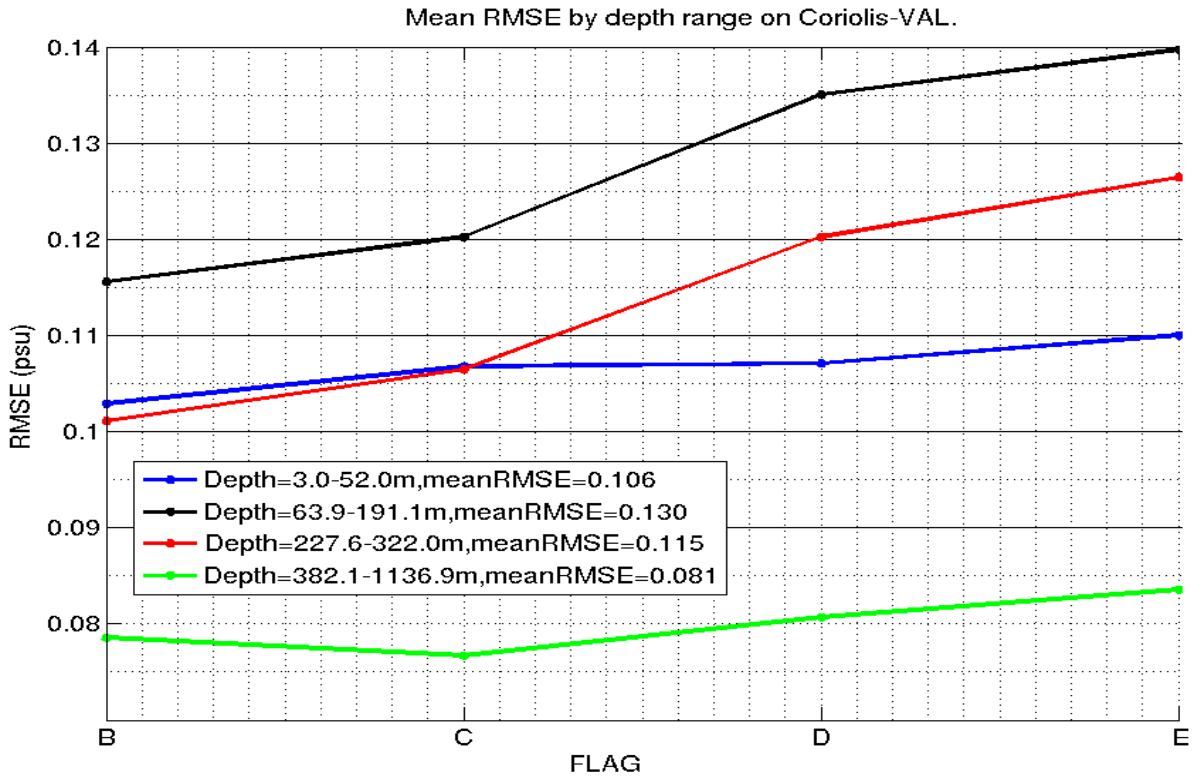


Figure 3: RMSE computed as a function of missing data percentage (Flags B, C, D, E) for SDCOMP for four different immersion intervals.

Different trends are observed. Black (depths between 63.9 and 191 m.) and red (depths between 227.6 and 322 m.) curves show the same behavior, i.e. that the RMSE continuously increases as a function of the missing data rate. The error is maximum at these depths (63.9 to 322 m), which correspond to the halocline; this strong salinity gradient where many complex geophysical interactions occur (Pailler et al., 1999) is highly variable and remains very difficult to model, especially in the tropical Atlantic region. The green curve (depths between 382.1 to 1136.9 m.) shows the mean RMSE of the subsurface layers which are always less than 0.085 psu while the blue curve (surface salinity) shows errors between 0.1 and 0.11 psu. The error is thus minimum in the subsurface layer and maximum in the halocline region. The RMSE is greater for the flags D and E. SOM-STI works with less information in these two cases and then complements S with less accuracy.

A third validation consists in analyzing the reconstructed datasets with SDCOMP. Let us denote Coriolis_CompB (C, D, E respectively) the four Coriolis_Gap datasets which have been complemented by the use of SDCOMP; their statistic characteristics must be very close to these of Coriolis_Val. We have represented these statistical parameters by using a Taylor diagram (Taylor, 2001) for each flag dataset and for each immersion cluster. This diagram permits to estimate the statistical differences existing between actual and reconstructed dataset. It computes the standard deviation, the root mean square difference (RMSD) and the correlations between these two sets. We have computed the mean salinity for the four depth intervals and the four complemented Coriolis_Comp datasets, and also for the Coriolis_Val. On the diagram, the standard deviation is represented in the vertical axis, the RMSD on the green circles and the correlation on the external circle. The red dot (A) is the reference which gives the statistics values corresponding to Coriolis Val.

Figure 4 shows the Taylor diagram obtained for the four flag datasets. The red dots corresponding to Coriolis_Val are located in regions close to these of flag B and flag C.

Correlation coefficients are higher than 0.95 for all the flags and increase from flag E to flag B. Performances (RMSD, correlation and standard deviation) are always better for the flag B. We can note that there are real differences between the flags in terms of correlations and RMSD (RMSE). However, these differences are less visible when we consider the standard deviation which means that the actual variability is respected by the completion for all the flags.

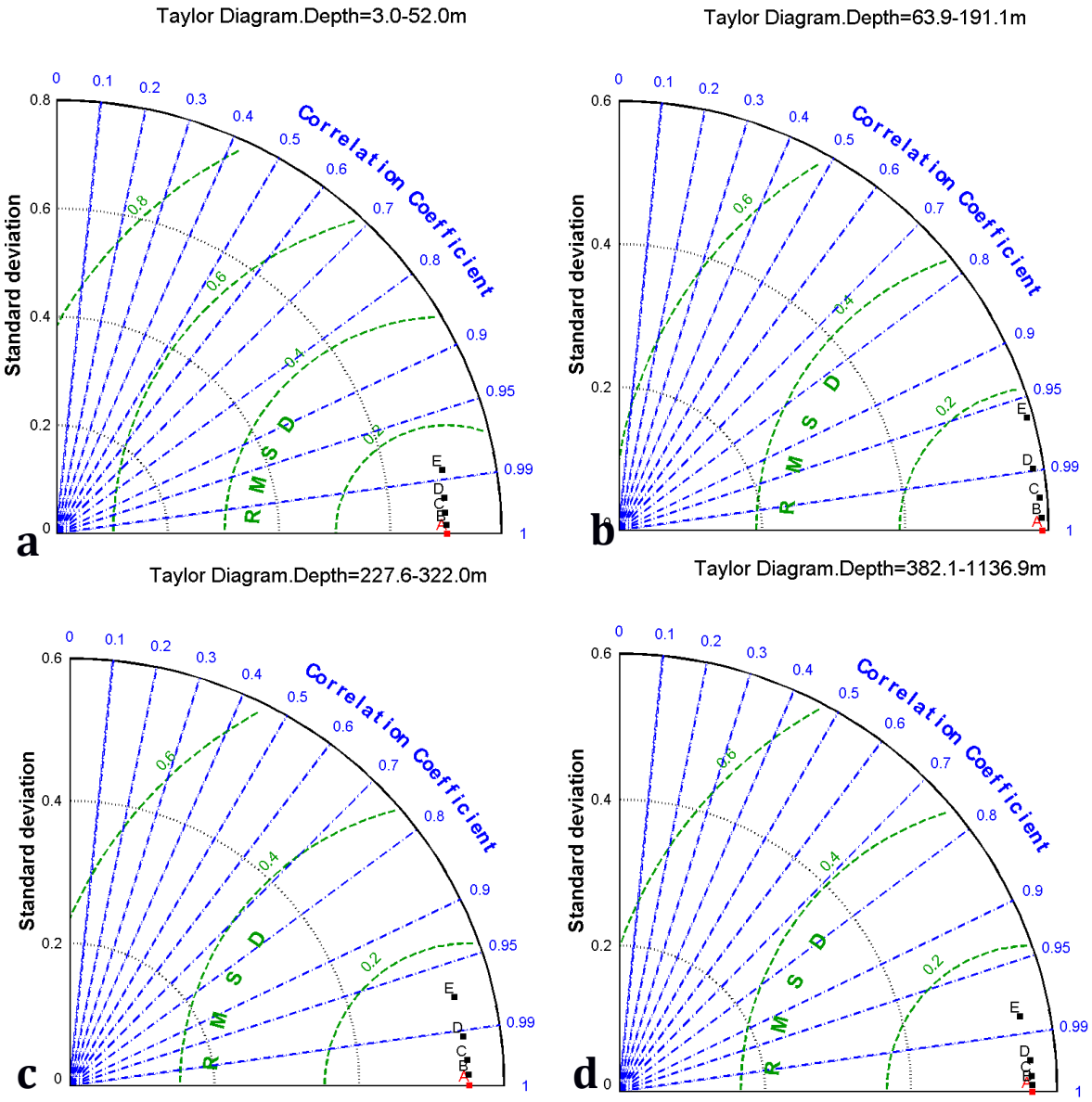
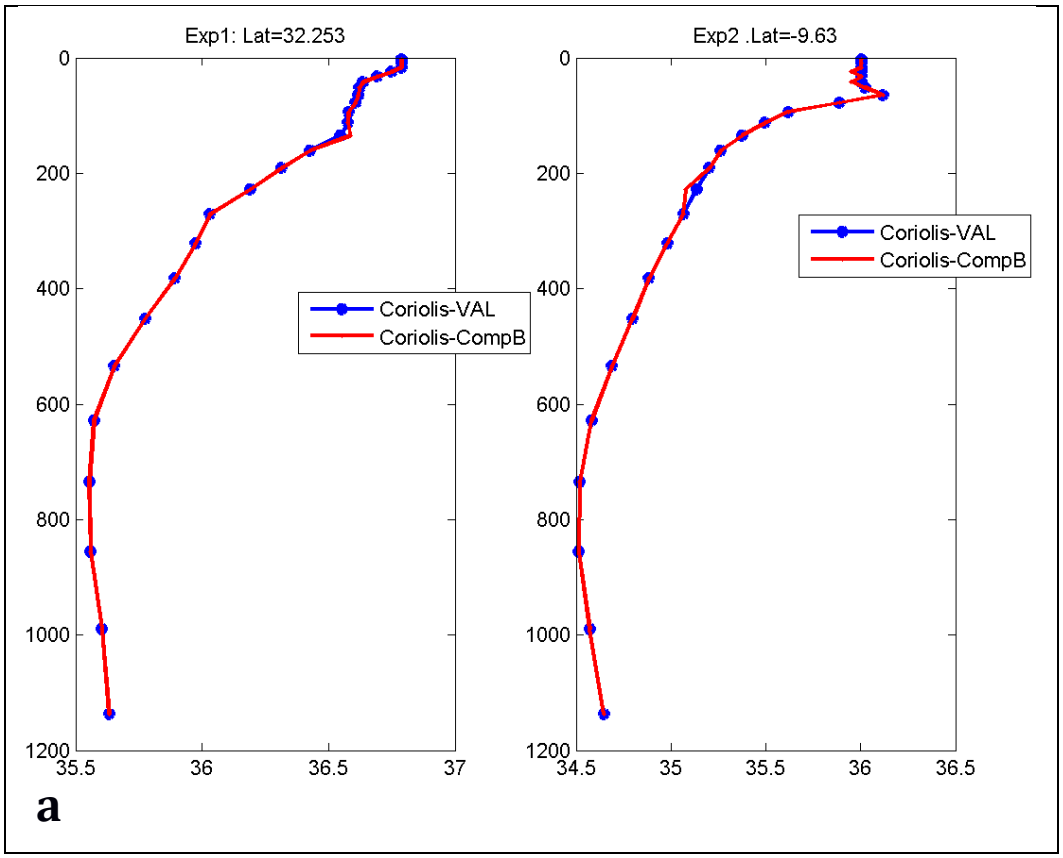
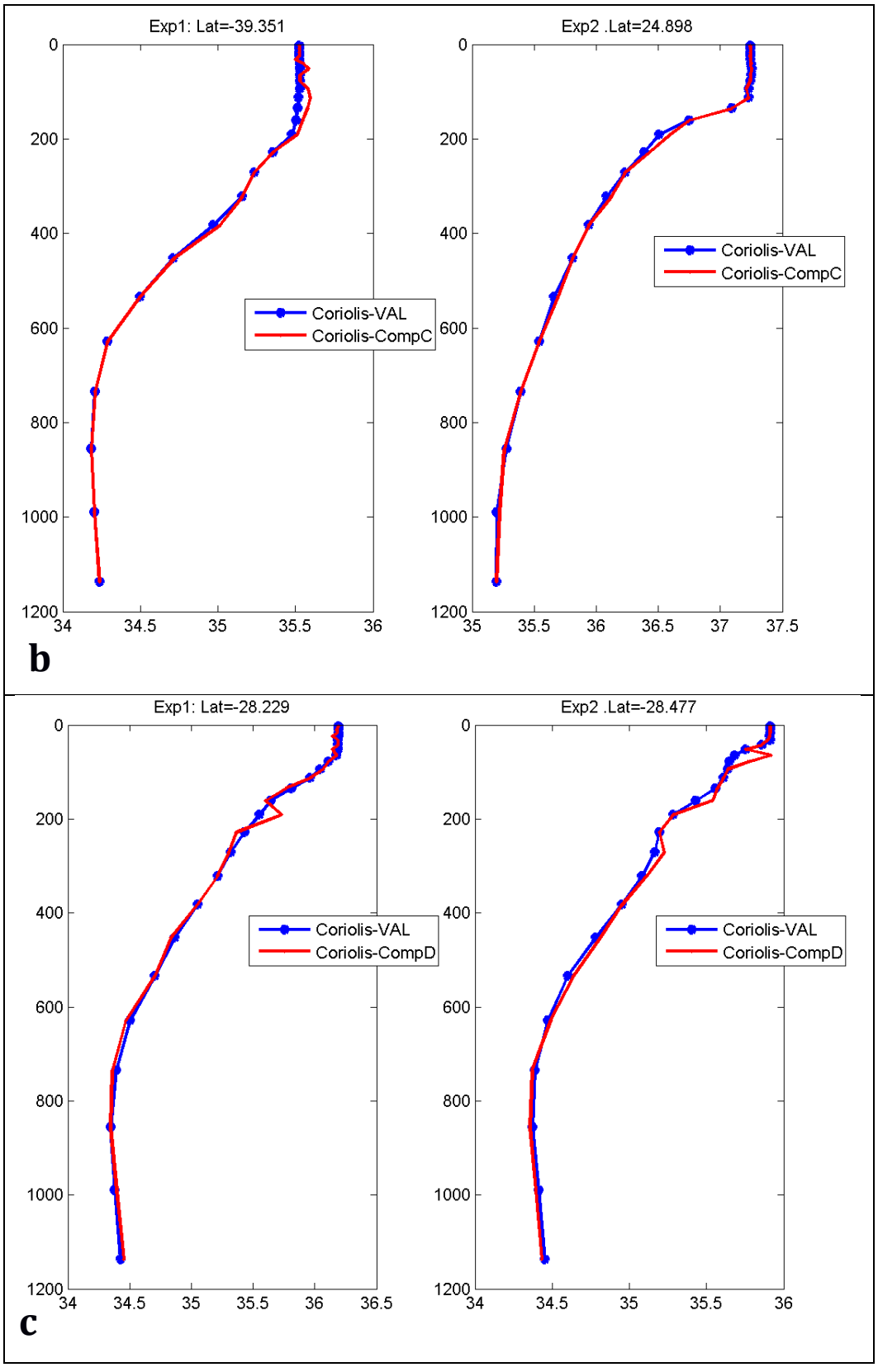


Figure 4: Taylor Diagrams for flags B, C, D and E datasets for four immersion

intervals [3m ; 52m](a), [63.9m ; 191m] (b) , [227.6m ; 322m] (c) , [382.1m ; 1136.9m] (d) in comparison of Coriolis Val, (A red dots).

To demonstrate the physical consistency of the reconstructed profiles, figure 5 shows for each flagged dataset, an "actual" salinity profile with no missing data and its associated complemented profile. These profiles have been randomly chosen from the Coriolis database profiles which have not been used for the learning.




Exp1: Lat=-39.351
Exp2 .Lat=24.898
Coriolis-VAL
Coriolis-CompC
b
Exp1: Lat=-28.229
Exp2 .Lat=-28.477
Coriolis-VAL
Coriolis-CompD
c

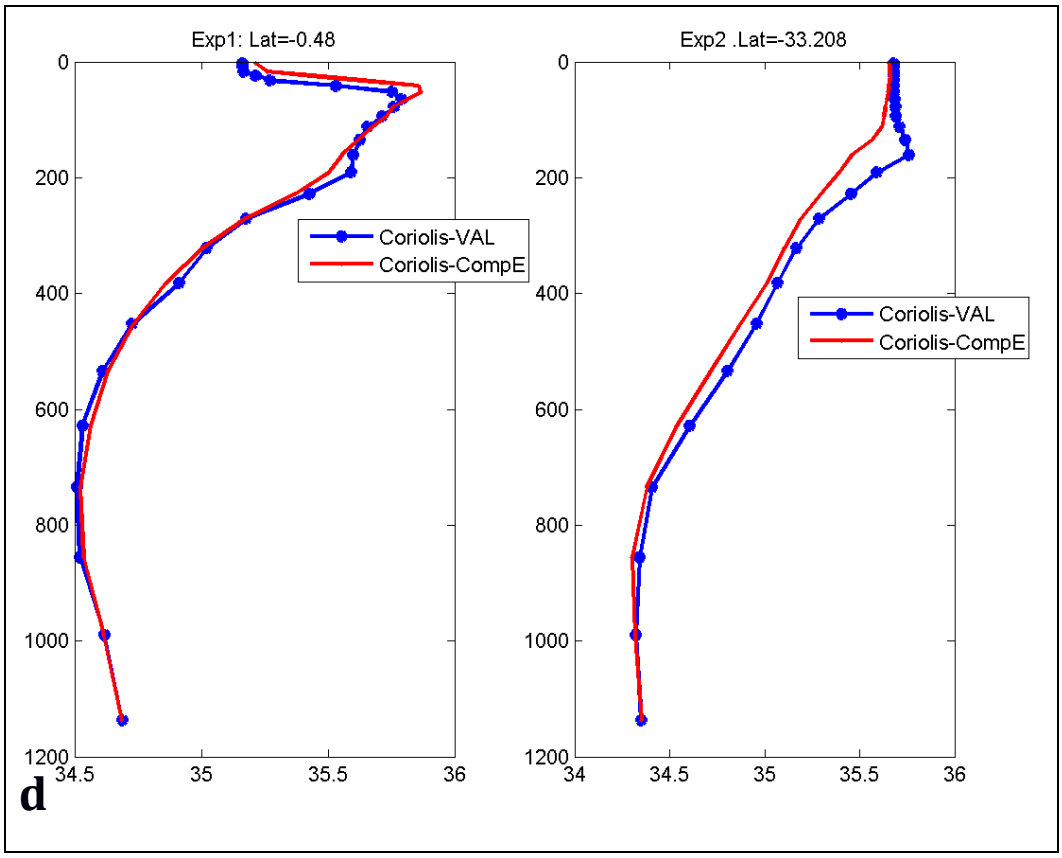


Figure 5: Examples of complemented profiles (red lines) and their actual reference (dotted blue lines) for flag B (a), flag C (b), flag D (c), flag E (d).

Figure 5 shows that the profiles are well complemented for flags B and C. But completion performance deteriorates for flags D and E which present more than 50% of missing values. Figure 5 shows also the difficulty of the completion algorithm to work efficiently around the halocline between 75 and 200 m. Completion of the halocline is less accurate for profiles in the southern hemisphere, particularly in the south of the Maximum Salinity Waters (MSW) region. But the subsurface is well complemented as it has been already noted in Figure 3.

Conclusion

We have developed a new method, the SDCOMP method, to increase the total number of available data of the ocean Coriolis database (http://www.coriolis.eu.org) which is a

highly correlated multidimensional database. SDCOMP, which uses the inverse model SOM-SPI proposed in Gueye et al., (2014), allows to reach a high degree of confidence on the completion of missing data. This method is based on the properties of the Self organizing maps which are robust statistical estimators, able to deal with data vectors presenting missing values (Jouini et al, 2014). SDCOMP is based on a recursive algorithm that complements missing values by processing the most similar data sets observed in the studied region. Some difficulties are encountered in specific regions, especially for the reproduction of the variability around the southern MSW and in the halocline, around 150 m depth. In these regions, the completion is less accurate mainly for flag D and E. If we can state that the method allows a satisfactory estimation of surface and deep-layer data, the challenge of the S profile completion around 150 m depth remains. The completion method is dependent on the dataset and the data quantity available. SDCOMP can be improved by considering not only the « best profile » provided by SOM-STI but a larger set of candidates that could be determined after the end of the SOM-STI phase. After completion, the final database presents a good data rate between 90.5% and 98% that will help oceanographers to better monitor the ocean and understand geophysical phenomena arising in areas where instruments failed to give good and reliable measurements.

Acknowledgments:

This study was funded by the French CNES (Centre National d'Etudes Spatiales) and the IRD (Institut de Recherche pour le Développement) organisations.We thank Dr M. Crepon for stimulating discussions. Ssalto/Duacs altimeter products were produced by Ssalto/Duacs and distributed by Aviso, with support from CNES (http://www.aviso.oceanobs.com/duacs/). The Argo data were collected and made freely available by the International Argo Program and the national programs that contribute to it.

(http://www.argo.ucsd.edu, http://argo.jcommops.org). The Argo Program is part of the Global Ocean Observing System.The numerical simulation is a contribution of the DRAKKAR project which is funded by the Centre National de la Recherche Scientifique (CNRS), the Institut National des Sciences de l'Univers (INSU), the Groupe Mission

Mercator Coriolis (GMMC) and Ifremer. Mbaye Babacar Gueye and Sabine Arnault are suported by the IRD.

Bibliography

Beckers, J.-M., Rixen, M. Eof calculations and data filling from incomplete oceanographic datasets. Journal of Atmospheric & Oceanic Technology 20 (12).

Cottrell, M., Patrick, L. "Missing values: Processing with the kohonen algorithm." arXiv preprint math/0701152 (2007).

Ghannad-Rezaie, M., Soltanian-Zadeh, H., Ying, H., Dong , M. Selection-fusion approach for classification of datasets with missing values, Pattern recognition 43 (6) (2010) 2340-2350.

Gueye, M. B., Niang, A., Arnault, S., Thiria, S., Crépon, M. Neural approach to inverting complex system: Application to ocean salinity profile estimation from surface parameters, Computers & Geosciences, (2014), in press. http://dx.doi.org/10.1016/j.cageo.2014.07.012.

Jouini, M., Lévy, M., Crépon; M., Thiria, S. Reconstruction of satellite chlorophyll images under heavy cloud coverage using a neural classification method, Remote Sensing of Environment 131 (2013) 232-246.

Kohonen, T. (2001). Self Organizing Maps (3rd ed.). Berlin Heidelberg: Springer Verlag; (501 pp).

Kondrashov, D., Ghil, M. Spatio-temporal filling of missing points in geophysical datasets. Nonlinear Processes in Geophysics, 13 (2) (2006) 151-159.

Pailler, K., Bourlés, B., Gouriou, Y. The barrier layer in the western tropical Atlantic Ocean, Geophysical research letters 26 (14) (1999) 2069-2072.

Pottier, C., Turiel, A., V. Garçon. Inferring missing data in satellite chlorophyll maps using turbulent cascading, Remote sensing of Environment 112 (12) (2008) 4242-4260.

Reynolds, R. W., Smith, T. M., Improved global sea surface temperature analyses using optimum interpolation, Journal of climate 7 (6) (1994) 929-948.

Smith T.M. and R. W. Reynolds, 2003: Extended Reconstruction of Global Sea Surface Temperatures Based on COADS Data (1854–1997). *J. Climate*, **16**, 1495–1510.

Stefanakos, Ch N., and G. A. Athanassoulis. A unified methodology for the analysis, completion and simulation of nonstationary time series with missing values, with application to wave data. Applied Ocean Research 23 (4) (2001): 207-220.

Chapitre VI. CONCLUSION ET PERSPECTIVES

Dans cette thèse, nous avons développé un modèle d'inversion du profil de salinité à partir de mesure de surface, notamment la SST, SSS, ADT (ou SSH), la latitude et la longitude. Ce modèle a été construit avec des données du modèle numérique DRAKKAR. Ensuite il a été appliqué et validé sur les données in situ du projet Coriolis et celles du projet ARAMIS.

L'utilisation des données DRAKKAR a été précédée par une étude comparative avec les données in situ ARAMIS. Cette étude a permis de déceler les limites du modèle DRAKKAR et d'établir un niveau de confiance de ce modèle numérique. Elle a montré une estimation moins bonne de S dans les zones à forte variabilité telles que les zones de courants ou/et de sous-courants marins et les SMW. Cependant dans l'ensemble, nous avons jugé que le modèle DRAKKAR estime bien le profil de S.

Après cette analyse comparative nous avons extrait une partie des données DRAKKAR qui ont servi à construire un réseau de neurones. Cette première étape consiste en une classification non supervisée d'un grand ensemble de profils de S avec 25 niveaux immersions dont la SSS, et les quatre autres paramètres de surface. Dans cette classification, nous avons utilisé une carte auto-organisatrice de Kohonen (SOM), qui agrège des profils similaires en un nombre réduit de groupes pertinents (neurones). La carte choisie dans cette première étape de la construction du modèle d'inversion est à deux dimensions avec 1000 (40 x 25) neurones, permettant une bonne représentation des données et l'apprentissage a été réalisé avec 1/10 des données centrées réduites. Chaque neurone de la carte est associé à un vecteur référent qui caractérise un groupe de données. Cette classification des profils verticaux de salinité associés aux paramètres physiques de surface pertinents (SSS, SST, SSH...) a permis de décrire la variabilité de toutes les situations rencontrées lors de l'apprentissage. Ces profils ont permis notamment de faire apparaître différents régimes océaniques pour lesquels on a trouvé une relation spécifique entre les paramètres de surface et les profils de salinité. Ces relations ont été marquées dans la carte par une topologie bien organisée.

La classification des profils associés aux paramètres de surface a été utilisée pour mettre en place un premier algorithme d'inversion. Cet algorithme basé sur une projection directe des paramètres de surface sur la carte topologique n'a pas donné une bonne estimation du

profil de S. L'étude des causes de cette mauvaise estimation a permis de dégager principalement deux hypothèses. La première est que les paramètres de surface n'ont pas la même influence sur le profil de S et la deuxième hypothèse est le fait que certains paramètres de surface sont liés entre eux. Ainsi l'utilisation simultanée de ces paramètres liés dans une même projection crée une redondance de l'information extraite et favorise les paramètres liés sur les autres alors que ces paramètres favorisés par cette simultanéité peuvent avoir une moindre explication de la variabilité des profils verticaux de S. L'extraction et la sélection de paramètres a montré que la SSS est très déterminante sur les S de la surface à 192m et que la latitude quant à elle est très déterminante pour les S allant de 380m à 1136m.

Partant de cette étude exploratoire, nous avons développé un modèle d'inversion basé sur un d'algorithme de projection séquentielle INV2 qui favorise la SSS et la latitude par rapport à la SST, la SSH (ADT) et la longitude. Cet algorithme considère, dans un premier temps, les 25 meilleurs neurones quand seules la SSS et la latitude (BMUS25) sont considérées ensuite les 40 meilleurs neurones quand la SST, la SSH et la longitude sont considérées (BMUS40). Finalement, le référent du meilleur neurone de BMUS25 appartement également à BMUS40 est sélectionné pour modéliser la donnée considérée. Ce modèle a été appliqué aux données DRAKKAR avec de bons résultats. Il a pu reconstruire les profils de S avec des erreurs absolues moyennes inférieures à 0.08 psu sur presque tout le profil. Ce deuxième modèle INV2 a donné des résultats largement meilleurs que la première version basée sur une projection simultanée de tous les paramètres de surface. Cependant, il faut noter que les données DRAKKAR sont issues d'un modèle numérique, elles sont lissées et donc l'apprentissage est peut-être moins complexe qu'avec des mesures in situ. De ce fait, la reproduction des profils est moins difficile. Il est également important de noter l'apport considérable de l'étude exploratoire des données à travers les techniques de feature selection (sélection de paramètres) et de feature extraction (extraction de paramètres). En effet celles-ci ont permis d'enlever l'information inutile (redondante) et de considérer chaque variable à sa juste valeur par rapport à son influence sur le profil de S.

Pour valider la réalité des données reproduites avec INV2-DRA (modèle construit avec les données DRAKKAR), les profils in situ issus du projet ARAMIS ont été inversés. Les

résultats ont montré que la reproduction de ces derniers présentait certains problèmes parmi lesquels le fait que nous disposions des données d'ADT pour ARAMIS alors qu'avec DRAKKAR, nous avions la SSH. Puisqu'il est important de tester la méthode d'inversion sur les données ARAMIS, véritable référence pour valider le modèle développé dans cette thèse, nous avons cherché des données in situ sur une zone qui englobe les trajets des campagnes ARAMIS. Ces données in situ sont celles distribuées sur le site du projet Coriolis.

Dans le cadre initial de la thèse, l'objectif était de reconstituer les profils de salinité sur l'Atlantique tropical et utiliser les données ARAMIS pour la validation. Suite à l'apprentissage, la carte avait donc pour but de permettre au mieux une représentation correcte et équivalente des profils de S dans toutes les régions de l'Atlantique tropical. Or les données DRAKKAR présentent quelques différences notamment autour de l'équateur et dans les SMW. Ces différentes raisons ont fait que les données Coriolis ont été utilisées pour une meilleure reconstruction d'ARAMIS.

Avec les données Coriolis, une première inversion a été faite. Dans l'ensemble, les erreurs moyennes tournent autour de 0.172 psu sur les profils individuels et 0.037 psu sur les profils moyens. Les plus fortes erreurs ont été notées au niveau de l'halocline autour de 150m. Pour améliorer les performances du modèle, une deuxième base d'apprentissage a été mise en place en intégrant les profils de température (SomCorioST) qui sont très liés aux profils de S même si cette liaison est non-linéaire. L'utilisation des profils de température a montré une évolution des performances du modèle d'inversion sur les données Coriolis surtout entre 200 et 700m de profondeur avec une rmse moyenne de 0.168 psu. SomCorioST a été utilisée pour inverser les données in situ d'ARAMIS et là aussi l'erreur moyenne passe de 0.215 à 0.187 psu. La comparaison de la fonction de distribution cumulée des données réelles ARAMIS et celle reconstruites sur quelques immersions caractéristiques (16m, 112m, 191m et 322m) a montré les bonnes performances du modèle sur les données ARAMIS même si elles n'ont pas été apprises par le réseau de neurones. Les courbes de ces deux fonctions présentent les mêmes allures quasiment sur toutes les immersions concernées. La reconstruction de quelques structures halines telles que l'isohaline 35.2 psu, les profils de la région du EUC et des SMW dans l'Atlantique tropical, une zone à haute variabilité, a été effectuée avec succès.

Nous avons également montré les capacités de la méthode d'inversion dans le cadre d'une complétion de données de S.

Les performances obtenues en utilisant le modèle prouve que cette approche est adéquate pour la reconstruction 3D de S océanique quand la connaissance dont nous disposons se limite à la surface. La méthode d'inversion neuronale présentée dans ce document alliée à l'extraction et à la sélection de paramètres est générale et pourrait être appliquée à la reconstruction d'autres variables océaniques telles que les profils de température et de chlorophylle.

Nous pouvons également noter les limites du modèle autour de l'halocline. Cette même zone n'est pas non plus bien reproduite par le modèle numérique DRAKKAR. En tant que méthode basée sur l'apprentissage de données, INV2 pourrait être moins performant quand la base d'apprentissage et les paramètres à partir desquels la reconstruction est faite sont dans des domaines géographiques différents. L'autre difficulté « temporaire » notée est liée à l'acquisition de la SSS qui est un paramètre de surface très important pour l'inversion. Ce problème pourrait être résolu dans un futur très proche avec les lancements des missions satellitaires telles que la mission européenne Soil Moisture and Ocean Salinity (SMOS) ou américaine AQUARIUS. L'autre possibilité prometteuse est d'estimer la SSS à partir des autres paramètres tels que l'ADT, la SST, la Latitude et la Longitude déjà disponibles pour la communauté scientifique. Une approche neuronale par apprentissage supervisé comme les perceptrons multicouches pourrait efficacement estimer la SSS. Ainsi, le modèle INV2 combiné à la SSS (estimée ou mesurée par télédétection) peut être utilisé pour inverser les longues séries temporaires de données surface disponibles depuis 1996 pour la reconstruction de profils de S dans un contexte d'étude du climat.

Ce travail de thèse pourrait déboucher sur la question de la fin des mesures in situ comme l'a discuté Tanguy (2011). Hors comme nous l'avons constaté, le bon échantillonnage des différentes régions est impératif pour la réussite de l'inversion. L'inversion ne peut donc venir qu'en complément des mesures satellites et in-situ, mais ne peut actuellement pas remplacer les mesures in-situ. Par contre l'un des principaux avantages de la méthode que nous avons développée est qu'elle peut reconstruire des profils à haute résolution à partir d'échantillons. Elle peut également fournir un nombre important de données et permettre

le suivi de certains phénomènes naturels tels qu'El Nino, les isohalines, etc.

BIBLIOGRAPHIE

Albert, R. & Barabási, A.-L., 2002. Statistical Mechanics of Complex Networks. *Reviews of modern physics,* 01.74(47).

Alliot, J.-M., 1996. *Techniques d'optimisation stochastique appliquées aux problèmes du controle aérien,* s.l.: s.n.

Araujo, M.; Limongi, C.; Servain, J.; Silva, M.; Leite, F. S.; Veleda, D.; Lentini, C. A. D., 2011. Salinity-induced mixed and barrier layers in the southwestern tropical Atlantic Ocean off the northeast of Brazil. *Ocean Science,* Volume 7, p. 63–73.

Arhonditsis, G. B. & Brett, M. T., 2004. Evaluation of the current state of mechanistic aquatic. *Marine Ecology Progress Series,* Volume 271, pp. 13-26.

Arhonditsis, G. B. & Brett, M. T., s.d. Evaluation of the current state of mechanistic aquatic. *Marine Ecology Progress Series,* Volume 271, pp. 13-26.

Arnault, S., 1987. Tropical Atlantic GeostrophicCurrents and Ship Drifts. *Journal of Geophysical Research,* 15 Mai, Volume 92, pp. 5076-5088.

Arnault, S., 2004. Variabilité océanique par altimétrie satellitaire. *Contemporary Puishing International:Télédétection,* 4(2), p. 103–124.

Arnault, S., Pujol, I. & Mélice, J.-L., 2011. In situ validation of Jason-1 and Jason-2 altimetry missions in the tropical Atlantic Ocean. *Marine Geodesy,* 34(3-4), pp. 319-339.

Barnier, B.; Madec, G.; Penduff, T.; Molines, J-M.; Treguier, A-M; Le Sommer, J.; Beckmann, A.; Biastoch, A.; Böning, C.; Dengg, J.; Derval, C.; Durand, E.; Gulev, S.; Remy, E.; Talandier, C.; Theetten, S.; Mathew, M.; McClean, J.; De Cuevas, B. 2006. Impact of partial steps and momentum advection schemes in a global ocean circulation model at eddy-permitting resolution. *Ocean Dynamics,* Issue 56, p. 543–567.

Blanke, B., Arhan, M., Lazar, A. & Prévost, G., 2002. A Lagrangian numerical investigation of the origins and fates of the salinity maximum water in the Atlantic. *Journal of Geophysical Research,* 107(C10).

Bourlès, B., Gouriou, Y., & Chuchla, R., 1999. On the circulation in the upper layer of the western equatorial Atlantic, Journal of Geophysical Research., 104(C9), p. 21151–21170.

Chow, G. C. & Lin, A.-L., 1971. Best Linear Unbiased Interpolation, Distribution, and Extrapolation. *The Review of Economics and Statistics,* 53(4), p. 372–375.

Claret, M., Rodriguez, R. & Pelegri, J. L., 2012. Salinity intrusion and convective mixing in the Atlantic Equatorial Undercurrent. *Advances in Spanish Physical Oceanography,* September.pp. 117-129.

Dessier, A. & Donguy, J. R., 1994. The sea surface salinity in the tropical Atlantic between 10°S and 30°N-seasonal and interannual variations (1977-1989). *Deep-Sea Research I,* Vol. 41(1), pp. 81-100.

Ding, H., Keenlyside, N. S. & Latif, M., 2012. Impact of the equatorial Atlantic on the El Nino Southern Oscillation. *Climate Dynamics,* Volume 38, p. 1965–1972.

Duby, C. & Robin, S., 2006. *Analyse en Composantes Principales.* s.l.:s.n.

Emery, W. J. & Dewar, J. S., 1982. Mean Temperature-salinity, salinity-depth and Temperaturedepth curves in the North Atlantic and North Pacific. *Progress in Oceanography,* Volume 11, pp. 219-305.

Frankignoul, C., 2012. Courants Equatoriaux. Dans: *Cours de circulation océanique.* Paris: s.n.

Goni, G. & Malanotte-Rizzoli, P., 2003. Subtropical cells in the Atlantic Ocean: An observational summary. Dans: *Interhemispheric Water Exchange in the Atlantic Ocean.* Elsevier Oceanography Series éd. s.l.:s.n., pp. 287 - 312.

Gueye, A. K., 2010. *Modélisation statistique des précipitations quotidiennes au sénégal,* Dakar: s.n.

Guyon, I. & Elisseeff, A., 2003. An Introduction to Variable and Feature Selection.. *Journal of Machine Learning Research,* Volume 3, pp. 1157-1182.

Hazeleger, W. & Drijfhout, S., 2006. Subtropical cells and meridional overturning circulation pathways in the tropical Atlantic. *Journal Of Geophysical Research,* Volume 111.

Holland, J., 1962. Outline for a logical theory of adaptive systems. *Journal of the Association of Computing Machinery,* Volume 3.

Jolliffe, I. T., 2002. *Principal Component Analysis. second ed.* 2e éd. Berlin(Berlin): Springer.

Kohonen, T., 1981. *Automatic formation of topological maps of patterns in a self-organizing system.* s.l., s.n., pp. 214-220.

Kohonen, T., 1982. Self-Organized Formation of Topologically Correct Feature Maps.

Biological Cybernetics, Volume 43, pp. 59-69.

Kohonen, T., 1990. *The Self-Organizing maps.* s.l., s.n.

Kucharski, F., Bracco, A., Yoo, J. H. & Molteni, F., 2008. Atlantic forced component of the Indian monsoon interannual variability. *Geophysical Research Letters,* Volume 35.

Losada, T. et al., 2010. Tropical response to the Atlantic equatorial mode: AGCM multimodel approach. *Climate Dynamics,* Volume 35, p. 45–52.

Mardia, K., Kent, J. T. & Bibby, J. M., 1980. *Multivariate Analysis*.. London: Academic Press.

Martin, A., 2013. *Analyse des mesures radiométriques en bande-L au-dessus de l'océan : campagnes CAROLS,* s.l.: s.n.

Metcalf, W., Voorhis, A. & Stalcup, M., 1962. The Atlantic Equatorial Undercurrent. *Journal of Geophysical Research,* Volume 67, pp. 2499-2508.

Millero, F. & Poisso, A., 1981. International one-atmosphere equation of state of seawater. *Deep-Sea Research,* Volume 28, p. 625–629.

MMSE, L. r., 2011. *Modélisation Mathématique et Statistique de l'Environnement (MMSE).* s.l., s.n.

Montégut, C. B., Mignot, J., Lazar, A. & Cravatte, S., 2007. Control of salinity on the mixed layer depth in the world ocean: 1. General description. *Journal Of Geophysical Research,* Volume 112.

Muñoz, E. et al., 2012. Mean and Variability of the Tropical Atlantic Ocean in the CCSM4. *Journal of Climate,* Volume 25, pp. 4860-4882.

Neumann, G., 1960. Evidence for an equatorial undercurrent in the Atlantic Ocean. *Deep-Sea Research,* Volume 6, pp. 328-334.

Neumann, G., 1966. The Equatorial Undecurrent in the Atlantic Ocean. *Symposium on Oceanography and Fisheries ressources of the tropical Atlantic,* October.

O'Connor, B., Fine, R. A., Maillet, K. A. & Olson, D. B., 1998. The rate of formation of the Subtropical Underwater (STUW) in the North and South Pacific from drifter and tracer data. *WOCE News,* Volume 31, pp. 18-20.

Ottino, J. M., 2003. Complex Systems. 49(2).

Palubinskas, G., 1998. An Unsupervised Clustering Method by Entropy Minimization. Volume 105, pp. 327-334.

Pinquier, J., 2004. Indexation sonore : recherche de composantes primaires pour une structuration audiovisuelle*,* Toulouse: s.n.

Polasek, W., Llano, C. & Sellner, R., 2010. Models, Bayesian Methods for Completing Data in Spatial. *Review of Economic Analysis,* Volume 2, p. 194–214.

Reseghetti, F., 2007. Empirical reconstruction of salinity from temperature profiles with phenomenological constraints.. *Ocean Science Discussions.,* Volume 4, p. 1–39.

Saporta, G., 2011. *Probabilités, analyses de données et statistique.* s.l.:Editions Technip.

Saravanan, R. & Chang, P., 2000. Interactions between the Pacific ENSO and tropical Atlantic climate variability. *Journal of Climate,* Volume 13, p. 2177–2194.

Simon, H. A., 1962. The Architecture of Complexity. *Proceedings of the American Philosophical Society,* 12 12.pp. 467-482.

Strogatz, S. H., 2001. Exploring Complex Networks," , 410, 268 (). *Nature,* Issue 410, pp. 268-276.

Sverdrup, H. V., Johnson, M. W. & and Fleming, R. H., 1942. The Ocean: Their Physics. *Chemistry and general Biology.*

Tanguy, Y., 2011. Variabilité de la dynamique et la thermodynamique en Atlantique tropical : Projet ARAMIS*,* s.l.: s.n.

U.S. CLIVAR Office, 2007. *Report of the U.S. CLIVAR Salinity Science Working Group. U.S,* s.l.: s.n.

WCRP, s.d. *CMIP Coupled Model Intercomparison Project.* [En ligne] Available at: http://cmip-pcmdi.llnl.gov/index.html [Accès le 29 07 2013].

Willmott, C. J., Ackleson, S. G. & Davis, R. . E., 1985. Statistics for the Evaluation and Comparison of Models. *Journal Of Geophysical Research,* 90(C25).

Wolfram, S., 2002. *A New Kind of Science.* s.l.:Wolfram Media.

Wu, X., Yan, X.-H., Jo, Y.-H. & Liu, W., 2012. Estimation of Subsurface Temperature Anomaly in the North Atlantic Using a Self-Organizing Map Neural Network. *Journal of Atmospheric and Oceanic Technology,* Volume 29.

Zhu, M. & Ghodsi, A., 2006. Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics & Data Analysis,* Volume 51, p. 918 – 930.

TABLE DES FIGURES

LISTE DES TABLEAUX

Titre : **Inversion neuronale pour la reconstruction de profils de salinité océanique en Atlantique tropical à partir de mesures de surface et de hauteurs d'eau.**

Auteur : Mbaye Babacar GUEYE

Résumé : Les échanges d'eau au sein du cycle global hydrologique sont déterminés par des contraintes mécaniques et thermodynamiques complexes qui forment les bases du système de dynamique du climat. La salinité océanique (S) est l'une des variables de ce cycle les plus délicates à observer. Les mesures in-situ ne permettent pas une bonne connaissance de la structure interne de l'océan car ne couvrent que des résolutions spatiales et temporelles souvent limitées. Alors que les mesures de surface offrent une bonne couverture spatio-temporelle.

L'objectif principal de cette thèse est de reconstruire le profil de S océanique à partir des données de surface (SSH: Sea Surface Height, l'ADT : Absolute Dynamic Topography, la SST: Sea Surface Temperature, la SSS: Sea Surface Salinity) en Atlantique tropical.

Le modèle d'inversion (INV2) que nous avons développé est décliné en 2 parties.

La partie exploratoire a permis de savoir, non seulement, que la latitude est liée aux S de subsurface et de profondeur alors que la SSS est liée au S de proche surface et de surface mais aussi que la SSS n'est linéairement liée à aucun des paramètres de surface étudiés alors que la latitude est fortement liée à la SSH (ADT) et à la SST. Partant de cette étude préalable, nous avons mis en place INV2 qui est basé sur un algorithme séquentiel à 3 étapes (2 projections sur une SOM et une recherche optimale).

INV2 a pu reconstruire les profils de S simulés par le modèle DRAKKAR avec des erreurs absolues moyennes inférieures à 0.08 psu sur presque tout le profil. Avec les mesures in situ Coriolis, les erreurs moyennes de INV2 tournent autour de 0.172 psu sur les profils individuels et 0.037 psu sur les profils moyens. Les plus fortes erreurs ont été notées au niveau de l'halocline et dans les zones de MSW. Après une évaluation, un algorithme de complétion de profils de S a été également proposé.

Les performances obtenues en utilisant le modèle prouve que cette approche est adéquate pour la reconstruction 3D de S océanique quand la connaissance se limite à la surface.

Mots-clefs : Océan Atlantique tropical, Inversion, Réseau de neurones, SOM, Paramètres de surface.