Statistical analysis of the spatio-temporal variations of geophysical variables, application to the satellite-derived Ocean Color and the Sea Surface Temperature

Bertrand Saulquin

THÈSE / Télécom Bretagne
sous le sceau de l'Université européenne de Bretagne

pour obtenir le grade de Docteur de Télécom Bretagne
En accréditation conjointe avec l'Ecole doctorale Matisse
mention : Traitement du signal et télécommunications

présentée par
Bertrand Saulquin
préparée dans le département ITI
Laboratoire Labsticc

Statistical analysis of the spatio-temporal variations of geophysical variables, application to the satellite-derived Ocean Color and the Sea Surface Temperature

Thèse soutenue le 2 décembre 2014
Devant le jury composé de :

Valérie Monbet
Professeure, Université de Rennes 1 / Présidente

Jean-Yves Tourneret
Professeur, ENSEEIHT - Toulouse / rapporteur

Robert Frouin
Chercheur, Université de Californie - Scripps-/ rapporteur

Ronan Fablet
Professeur, Télécom Bretagne / examinateur

François-Régis Martin-Lauzer
Directeur - Docteur, ACRI-ST, Sophia Antipolis / examinateur

Cédric Jamet
Maître de conférences, Université du Littoral Côte d'Opale / examinateur

Grégoire Mercier
Professeur, Télécom Bretagne / directeur de thèse

Odile Fanton d'Andon
Docteur, ACRI-ST - Sophia Antipolis / invitée

N° d'ordre: 2014telb0332

Sous le sceau de l'Université européenne de Bretagne

Télécom Bretagne

En accréditation conjointe avec l'Ecole Doctorale Matisse

Ecole Doctorale – MATISSE

Statistical analysis of the spatio-temporal variations of geophysical variables, application to the satellite-derived Ocean Color and the Sea Surface Temperature

---

Thèse de Doctorat

Mention: Traitement du signal et télécommunications

Présentée par Bertrand Saulquin

Département: ITI

Laboratoire: Lab-STICC

Directeur de thèse: Grégoire Mercier

Soutenue le 02/12/2014

Jury:

Robert Frouin

Jean-Yves Tourneret

Ronan Fablet

François-Régis Martin-Lauzer

Valérie Monbet

Cédric Jamet

Grégoire Mercier

Odile Fanton d'Andon

Remerciements

Je tiens à remercier Mme Odile Fanton d'Andon, directrice de la société ACRI-ST pour laquelle je travaille, et Mr Antoine Mangin, directeur scientifique, de m'avoir donné l'opportunité de réaliser cette thèse. Je remercie également Grégoire Mercier d'avoir accepté de devenir mon directeur de thèse, sachant que mon profil était quelque peu atypique pour un thésard en France. Merci également Grégoire pour tes points de vue souvent décalés qui m'ont permis de prendre un peu de recul par rapport aux travaux engagés. Merci à Ronan Fablet qui m'a accompagné au long de cette thèse avec patience et pédagogie. Merci également aux membres du Jury pour avoir accepté d'évaluer mon manuscrit.

Contents

List of figures

List of Tables

Résumé étendu en Français

L'analyse de l'impact du changement climatique global, la caractérisation de phénomènes climatiques majeurs, la prévision de processus géophysiques d'intérêt, influencent les politiques des états. A titre d'exemple, depuis le protocole de Kyoto (signé en 1997 et appliqué en 2005) la Commission Européenne a posé en 2005 les bases d'une stratégie communautaire sur le changement climatique [1]. Cette stratégie repose entre autre sur l'élaboration de nouvelles mesures en coordination avec les autres politiques européennes, sur le renforcement de la recherche, de la coopération internationale, et sur la sensibilisation des citoyens. Les analyses scientifiques à l'origine de ces politiques sont réalisées soit directement à partir de séries spatio-temporelles d'observations in-situ ou satellitaires, ou, à partir de modèles numériques utilisant ces observations comme forçage.

Depuis une trentaine d'années, la majorité des séries temporelles d'observations de la surface des océans est fournie par des capteurs embarqués sur des plateformes satellites. Ces séries sont désormais assez longues pour caractériser de faibles variations temporelles et spatiales dans les variables géophysiques mesurées directement au sommet de l'atmosphère, ou estimées à partir de ces dernières.

La variation temporelle d'une variable géophysique est envisagée depuis trente ans suivant trois formes principales :

- la plus connue est **l'estimation d'une tendance à long terme**, qu'elle soit linéaire [2, 3, 4] ou non [5, 6]. Elle est largement utilisée pour étudier le réchauffement climatique [7].

- **l'analyse des corrélations spatio-temporelles dans un ou plusieurs jeux de données** [5, 8, 9]. Typiquement les analyses par 'Principal component analysis' (PCA) ou 'Empirical Orthogonal Functions' (EOF, [9]) décomposent une matrice de covariance spatio-temporelle en modes principaux. Chaque mode est ensuite associé à des conditions de forçage, comme par exemple des signaux à très large échelle, saisonniers ou locaux. Ce type d'approche, visant à rechercher des modes orthogonaux dans la covariance, souligne la nécessité d'être capable de découpler, au sein du signal étudié, les différents processus géophysiques pour pouvoir les étudier séparément.

- **la caractérisation de régimes**: Nous définissons un régime comme une relation liant la variable d'intérêt Y à ses prédicteurs X. La variable d'intérêt est-elle mieux estimée par plusieurs régimes plutôt qu'un seul régime linéaire ou non-linéaire? Cette question est particulièrement intéressante pour les variables géophysiques qui sont souvent marquées par des comportements saisonniers importants: la saisonnalité implique souvent des relations variant dans le temps entre la variable d'intérêt et ses conditions de forçage. L'estimation et la compréhension de ces liens sont par conséquent particulièrement importantes pour pouvoir estimer, inverser, et prédire la variable considérée.

Dans cette thèse, nous nous intéressons aux variations temporelles de la couleur de l'eau et la température de surface observées depuis l'espace. Nous abordons les quatre questions scientifiques suivantes:

- l'estimation de tendances, de biais, et de cycles saisonniers significatifs dans plusieurs séries temporelles géophysiques.
- l'analyse spatio-temporelle d'un signal climatique majeur.
- la modélisation et la prévision d'une variable géophysique soumise à des processus saisonniers.
- l'inversion d'une variable géophysique.

D'un point de vue méthodologique, la nature spécifique du signal géophysique, comme par exemple la discontinuité potentielle des observations et l'autocorrélation du bruit, nécessite d'aborder des problématiques spécifiques. Parmi celles-ci, et relativement aux questions scientifiques, nous distinguons notamment:

- la régression d'une variable aléatoire géophysique.
- l'estimation et la caractérisation de modes distincts dans des signaux multivariés. Un mode fait référence ici à une composante élémentaire d'un mélange.
- l'estimation et la caractérisation de régimes physiques distincts, soit de relations liant une (des) variable(s) à un ensemble de prédicteurs.

Ce manuscrit est organisé de la façon suivante. Ce résumé étendu en Français synthétise les travaux réalisés dans les chapitres I à V rédigés en Anglais. Au début de ce résumé, nous présentons également les mesures télédétectées et les principales méthodologies utilisées dans cette thèse. Le second chapitre fournit une introduction en Anglais. Les chapitres II à V sont structurés autour des questions scientifiques posées et d'un article publié ou soumis dans le cadre de cette thèse. Chaque chapitre est décomposé de la façon suivante: une présentation du contexte scientifique relativement à l'état de l'art, une présentation de la méthodologie mise en œuvre pour y répondre, et une illustration thématique correspondante.

- Le chapitre II présente notre contribution méthodologique pour caractériser les tendances, cycles saisonniers et biais significatifs dans plusieurs séries temporelles de couleur de l'eau [4].
- Le chapitre III présente notre contribution méthodologique, basée sur la détection d'évènements temps-fréquence dans la température de surface pour l'analyse spatio-temporelle du signal El Niño Soutern Oscillation (ENSO) [10].
- Le chapitre IV est consacré à la modélisation et la prévision de la turbidité de surface avec des modèles à changement de régimes Markoviens [11].
- Le dernier chapitre détaille nos recherches sur l'inversion des réflectances marines en milieux côtiers à l'aide de mélanges de lois gaussiennes multivariées [13].

Dans l'annexe A est présentée une publication réalisée pendant la première année de thèse sur l'analyse de la transparence de la colonne d'eau à partir des données MERIS haute résolution et de son impact sur les distributions observées des Posidonies en Méditerranée. Cet article n'est pas directement lié à la thèse mais contient de nombreuses informations thématiques et contextuelles d'intérêt pouvant aider à une meilleure compréhension des enjeux associés à cette thèse.

0.1 Couleur de l'eau et température de surface de la mer

En océanographie, le coût d'acquisition des mesures in-situ est extrêmement important. Par exemple, le navire océanographique 'Pourquoi Pas' de l'Ifremer, a des coûts intégrés sur certaines missions supérieurs à 50000 € par jour [14]. Par conséquent, même si les mesures in-situ représentent un jeu de données incontournable en océanographie, leur représentativité spatiale et temporelle est limitée pour caractériser des changements à l'échelle globale, et les observations satellitaires représentent aujourd'hui la principale source utilisée pour caractériser ces changements.

Nous présentons ici les données satellitaires de couleur de l'eau et de température de surface utilisées dans cette thèse, ainsi que le principe de leur estimation.

0.1.1 Les données satellitaires

Les données issues des capteurs satellitaires représentent une source d'observation unique en termes de couverture spatio-temporelle de la surface des océans. Le satellite TIROS-N, lancé en 1978, fut la première plateforme à orbite héliosynchrone embarquant un capteur permettant de mesurer la température de surface des océans (Sea Surface Temperature, SST) [15]. De 1981 à nos jours, les données de la température de surface acquises par 14 missions héliosynchrones de la 'National Oceanic and Atmospheric Administration' (NOAA), embarquant les capteurs 'Advanced Very High Resolution Radiometer' (AVHRRs) [16], constituent pour les océanographes la plus longue série d'observations continues, disponible à l'échelle globale.

Lancé également en 1978, CZCS[1], embarqué à bord de la plateforme Nimbus 7, fut le premier capteur mesurant la 'couleur de l'eau' depuis l'espace, c.-à-d. la luminance spectrale dans le domaine du visible de 400 à 700 nm. Il mesura jusqu'en 1986, selon une orbite héliosynchrone, la luminance aux longueurs d'ondes 443, 520, 550, et 670 nm. Après une attente de douze années, le satellite OrbView-2 fut lancé en 1998 (capteur 'Sea-Viewing Wide Field-of-View Sensor', SeaWiFS), suivi de TERRA et AQUA (respectivement en 1999 et 2002, embarquant le capteur 'Moderate Resolution Imaging Spectroradiometer', MODIS), et ENVISAT (2002, capteur embarquant le capteur 'MEdium Resolution Imaging Spectrometer', MERIS). De 1999 à nos jours, chaque jour, la couleur des océans est observée depuis l'espace à l'échelle du globe.

Dans la Table 1 sont listées les caractéristiques des capteurs de couleur de l'eau et de température de surface utilisés dans cette thèse.

[1] CZCS: Coastal Zone Color Scanner

Table 1: Résumé des caractéristiques des capteurs satellitaires utilisés dans cette thèse.

Capteur	MERIS	MODIS	SeaWiFS	AVHRR
Domaine	Couleur de l'eau	Couleur de l'eau	Couleur de l'eau	Température de surface
Satellite	Envisat	EOS-PM1&2	SeaStar	NOAA 9, 11, 14, 16, 18
Agence	ESA[2]	NASA	NASA	NASA[3]
Début	2002	2002	1997	1985
Fin	2010	Aujourd'hui	2010	2009
Heure du passage à l'équateur (h)	10:00	13:30 (TERRA) 10:30 (AQUA)	12:20	14:20, 13:30, 13:30, 20:37
Fauchée (km)	1150	2330	2806	2600
Résolution (m)	300	1000	1100	4000
Bandes spectrales utilisées	15 bandes de 412 à 900 nm	9 bandes de 412 à 900 nm	8 bandes de 412 à 900 nm	5 bandes de 0.6 à 12 µm

0.1.2 Principe de la mesure de "la couleur de l'eau": Introduction à la théorie du transfert radiatif.

Le capteur de 'couleur de l'eau' mesure la luminance directionnelle spectrale (une énergie dont l'unité communément utilisée est le $mW.m^{-2}.nm^{-1}.sr^{-1}$) émise par la surface de l'eau entre 400 et 900 nm, soit les parties visibles et le proche infra-rouge du spectre [17]. Dans le cas d'un capteur embarqué à bord d'un satellite, l'énergie mesurée au sommet de l'atmosphère (TOA, Top Of Atmosphere) est la somme des contributions de l'irradiance solaire reçue, qui est rétrodiffusée soit a/ directement par l'atmosphère, soit b/ par la réfraction spéculaire de la surface des Océans ('Glint', [18, 19]) ou c/ par les constituants dans la mer ayant des propriétés optiques non neutres (Figure 1). En moyenne, la fraction du signal rétrodiffusée par l'eau et ses constituants représente moins de 10 % du signal mesuré à la surface de l'atmosphère [20]. L'étape de déconvolution du signal mesuré TOA en a/ b/ et c/ est appelée 'corrections atmosphériques'. Celle-ci est particulièrement délicate dans le domaine spectral du visible du fait de la faible contribution de la partie marine du signal, et de signatures spectrales atmosphériques et marines potentiellement similaires [17].

[2] ESA : European Spatial Agencies
[3] NASA : National Aeronautics and Space Administration

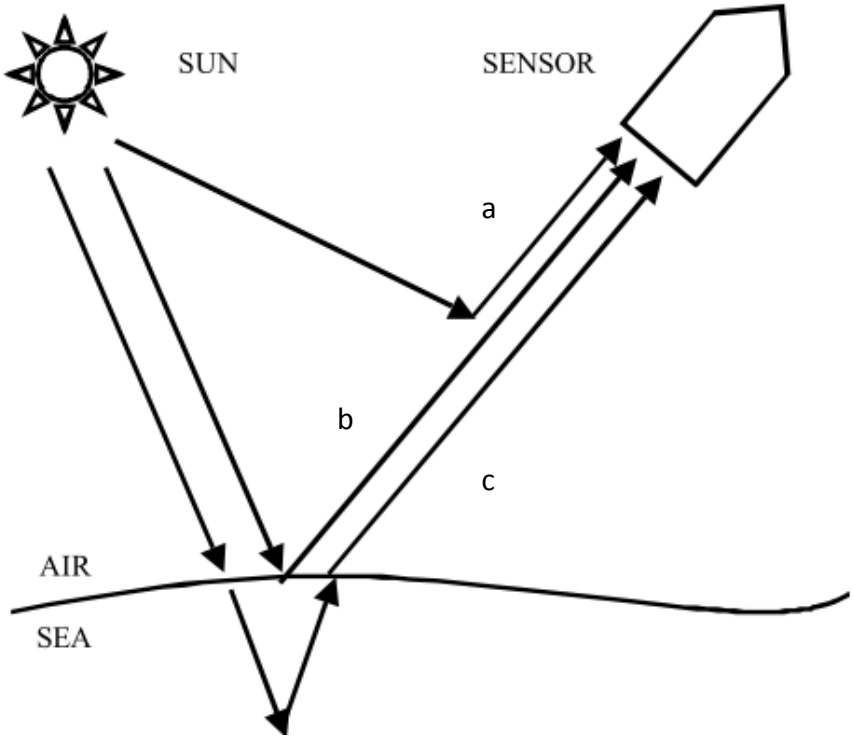


Figure 1: Trajectoire des photons dans l'atmosphère et l'eau avec une hypothèse de diffusion simple. Lors de chaque trajet, illustré ici avec une flèche, les photons sont potentiellement soumis à des phénomènes de diffusion ou d'absorption.

De la mesure de la réflectance de la surface de la mer aux constituants de l'eau.

La luminance spectrale (L, en $mW.m^{-2}.nm^{-1}.sr^{-1}$) dépend de la géométrie d'observation (position de l'instrument de mesure et de la cible par rapport à la source d'éclairement) et des propriétés anisotropes ou isotropes de diffusion des composants optiquement actifs de l'atmosphère et de l'eau. En intégrant la luminance spectrale sur l'angle solide [0; 2π] on obtient l'irradiance spectrale (E, en $mW.m^{-2}.nm^{-1}$). Cependant, la luminance spectrale montante Lu mesurée par le capteur dépend de l'irradiance solaire descendante Ed, qui varie en fonction de la saison et de la latitude.

Par conséquent, la quantité géophysique étudiée est la réflectance spectrale $\rho_{TOA}(\lambda)$, soit la luminance spectrale montante (observée) normalisée par l'irradiance spectrale descendante [20]:

$$\rho_{TOA}(\lambda) = \pi \frac{\mathrm{Lu}}{E_d . \cos(\Theta s)} \quad (1)$$

avec Θs l'angle solaire zénithal. Dans l'équation (1), $\rho_{TOA}(\lambda)$ dépend de Lu donc toujours de la géométrie d'observation (Figure 2).

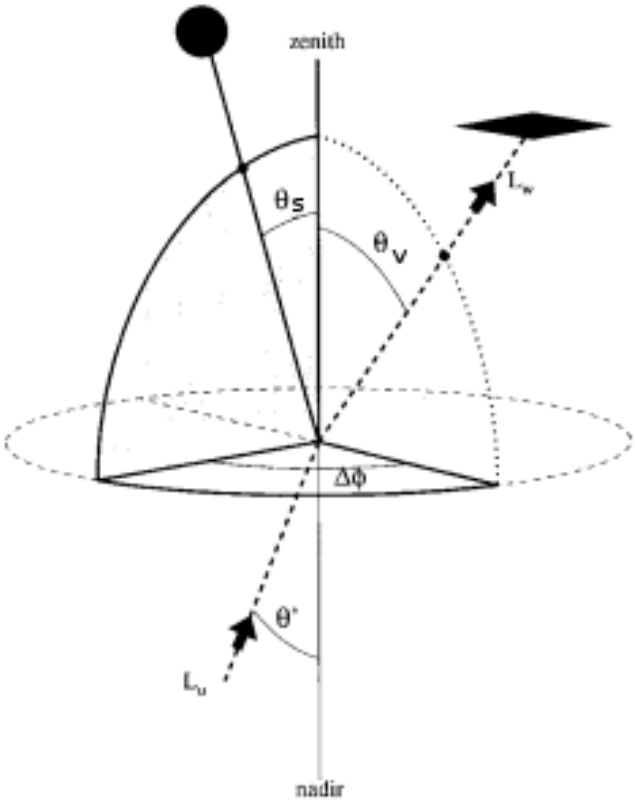


Figure 2: Géométrie de la mesure télédétectée.

Lors des corrections atmosphériques (cf. chapitre V), la réflectance de l'eau observée à la surface de la mer pour une géométrie donnée $\rho_w(\lambda)$ est estimée à partir de $\rho_{TOA}(\lambda)$. Pour obtenir une mesure complètement normalisée de $\rho_w(\lambda)$, qui soit comparable en toutes circonstances d'observations et d'ensoleillement, $\rho_w(\lambda)$ est normalisée pour $\Theta s = 0$ et $\Theta v = 0$ c.-à-d. corrigée de la bidirectionnalité de l'eau [21].

La réflectance normalisée mesurée à la surface de la mer $\rho_{wn}(\lambda)$ est liée aux propriétés optiques apparentes de l'eau, soit l'absorption et la rétrodiffusion [22]:

$$\rho_{\mathrm{wn}}(\lambda) = \pi \frac{f}{Q} R \frac{bb}{a + bb} \qquad (2)$$

Avec f et Q, deux termes de bidirectionnalité qui varient avec la géométrie et faiblement avec les propriétés optiques de l'eau (uniquement pour f) [21]. Le facteur R prend en compte les effets de réfraction et réflexion lors du passage air-eau [23, Figure 1]. Les paramètres bb et aa sont respectivement la rétrodiffusion et l'absorption des composants optiquement actifs de la colonne d'eau , soit l'eau pure, les matières organiques dissoutes (CDOM, Colored Dissolved Organic Matters), la concentration en chlorophylle-a (chl-a) et les matières minérales en suspension (SPM, Suspended Particulate Matters) [17] :

$$a = a_w + a_{chl} + a_{SPM} + a_{CDOM} \quad (3)$$

$$bb = bb_w + bb_{chl} + bb_{spm} \quad (4)$$

0.1.3 Principe de la mesure de “la température de surface”

Dans la gamme des longueurs d’onde de l’infrarouge thermique entre 4 et 13 µm, la luminance mesurée correspond au rayonnement électromagnétique directement émis par la surface de l’océan ou le sommet des nuages. La loi de Planck [24] permet de calculer, à partir de la luminance mesurée, une température que l’on appelle température de brillance ou température radiométrique de la surface observée. La température de brillance est la température estimée du corps noir qui émettrait la même énergie.

Les longueurs d’onde entre 4 et 13 µm sont très peu ou pas impactées par l’atmosphère et les constituants de l’eau de mer. Par conséquent l’étape de corrections atmosphériques est plus simple dans ces longueurs d’onde que dans le visible, et les estimations de température de surface sont également moins bruitées et plus proches des mesures in-situ que les estimations de couleur de l’eau. Cette propriété physique de la température de surface, l’influence directe de celle-ci sur les échanges atmosphère-océans [25], et le lien direct entre la circulation et la température de surface [26], ont contribué au développement de multiples missions satellitaires embarquant des capteurs mesurant dans l’infra-rouge.

0.2 Méthodologies

La nature spécifique du signal géophysique doit être prise en compte dans l’approche scientifique mise en œuvre. Le signal géophysique est typiquement caractérisé par:

- la présence d’un bruit autocorrélé. Dans le cas s’un signal temporel, cela signifie que le bruit observé au temps t est corrélé à celui observé au temps t-1 [27].
- de modes caractéristiques dans la variable d’intérêt. Par exemple, la réflectance observée de la surface de la mer est conditionnée par les constituants présents dans la colonne d’eau (chl-a, SPM et CDOM). La réflectance observée de la surface de la mer peut être considérée comme un mélange de composantes élémentaires, chaque composante élémentaire (mode) étant liée à type d’eau, i.e. une proportion spécifique des composants optiquement actifs dans l’eau (chl-a, SPM et CDOM).
- de régimes distincts, soit des relations variant dans le temps et l’espace entre une variable et ses prédicteurs. Par exemple, il existe un seuil d’énergie nécessaire pour remettre en suspension les sédiments sous l’effet de la houle. Cela implique deux régimes distincts entre la variable ‘turbidité’, et la variable ‘intensité de la houle’ en fonction du temps [11].

En fonction des spécificités du signal géophysique, nous distinguons dans cette thèse trois familles de méthodes à savoir : la régression d'une variable aléatoire géophysique; l'estimation et la caractérisation de modes distincts dans des signaux multivariés; l'estimation et la caractérisation de régimes physiques distincts.

0.2.1 Régression d'une variable aléatoire géophysique.

L'ensemble des travaux présentés dans cette thèse est directement ou indirectement lié à la régression d'une variable observée Y en fonction de prédicteurs X: $Y = f(X) + B$, avec f une fonction linéaire ou non. Parmi les estimateurs disponibles des paramètres de f nous nous intéresserons particulièrement aux:

- **Moindres carrés ordinaires et généralisés**. Pour le modèle linéaire $Y = AX + B$, l'estimateur des moindres carrés standard (Ordinary Least Square, OLS) cherche à minimiser les résidus: $\hat{A} = argmin_A(y - \hat{y})$. Dans cette thèse, nous considérons que les erreurs suivent une loi normale centrée, mais il est possible de considérer des modélisations plus complexes de la distribution du bruit. Dans le cas d'un bruit non gaussien, l'estimateur OLS est biaisé [28]. L'estimateur OLS suppose que les résidus sont non corrélés entre eux.

  La méthode des moindres carrés généralisés (Generalized Least Square, GLS) est une généralisation de l'estimateur OLS qui prend en compte la corrélation entre les résidus. Cette formulation est utilisée dans cette thèse pour la détection de tendances, biais et cycles saisonniers, parmi de multiples séries temporelles (cf. chapitre II). La matrice de covariance des résidus est alors exprimée en fonction des caractéristiques du bruit (Cochrane & Orcutt [29]).

- **Estimateur par maximum de vraisemblance.** Dans le cadre d'une régression linéaire, la variable aléatoire Y= {y_1,.., y_n} suit une loi normale f de moyenne $\hat{A}X$ et de variance σ² [30]. La vraisemblance L de A conditionnellement à Y est alors $L(A|Y) = P(Y|A) = \prod_{i=1}^{n} P(Y = y_i|\,\mathrm{A})$. Pour obtenir l'estimateur $\hat{A}$ du maximum de vraisemblance, on maximise la log-vraisemblance en la dérivant par rapport à A. L'estimateur de maximum de vraisemblance est utilisé dans les chapitres III à V.

- **Régresseurs non-linéaires.** Les régressions à vecteurs de support (Support Vector Machines, SVR [31]), et les réseaux de neurones (Neural Networks, NN [32]) sont parmi les techniques à apprentissage automatique typiquement utilisées pour l'estimation de relations non-linéaires. Par rapport aux modèles linéaires, les modèles non-linéaires, du fait de leur complexité et flexibilité, nécessitent davantage de données pour l'estimation des paramètres réalisée lors de la phase dite d'apprentissage. Nous les utiliserons principalement pour évaluer nos modèles statistiques de prévision (chapitre IV) et d'inversion (chapitre V).

La régression est souvent directement associée au choix des prédicteurs X dont la contribution est significative. L'**analyse linéaire discriminante** (LDA) vise à identifier les variables contribuant **linéairement et de façon significative** à la caractérisation du groupe étudié. Elle permet typiquement de réduire le nombre de variables à considérer parmi un ensemble pour caractériser la variable d'intérêt. Nous avons utilisé les LDA pour choisir les prédicteurs X de nos modèles de prévisions (chapitre IV).

0.2.2 Estimation et caractérisation de modes distincts dans des signaux multivariés.

Nous nous intéressons particulièrement dans cette thèse à l'identification de modes physiques distincts dans des signaux multivariées. Un mode correspond à une composante élémentaire d'un mélange. Le caractère 'physique' du mode considéré vient du fait qu'il correspond à un état physique reconnu thématiquement (par exemple un type d'eau ou un type d'aérosol). Dans cette thèse, nous essayons le plus souvent possible de caractériser thématiquement les modes identifiés. Différentes approches ont été envisagées:

- **Mélanges de distributions multivariées Gaussiennes [34]**. Les mélanges de gaussiennes servent à modéliser la densité d'une variable aléatoire X de dimension n avec une somme de K gaussiennes:

$$f(x) = \sum_{k=1}^{K} \lambda_k \frac{1}{\sqrt{2\Pi^n |\Sigma|}} e^{-0.5(x-\mu_k)' {\Sigma_k}^{-1} (x-\mu_k)} \quad (5)$$

  μ_k est la moyenne, Σ_k la matrice de covariance de la loi normale k, et λ_k la probabilité à priori du mode k dans le mélange de gaussienne: $\lambda_k = P(Z_n = k)$. Contrairement à la télédétection terrestre, nous ne disposons pas toujours en océanographie de références de terrain absolues, et par conséquent les segmentations considérées dans cette thèse sont donc non-supervisées et les modes Z sont inconnus (cachés). Les mélanges de gaussiennes sont utilisés dans le chapitres III pour estimer des échelles temporelles caractéristiques de la température de surface et dans le chapitre V pour estimer des spectres de référence des aérosols et de l'eau.

  Pour estimer μ_k et Σ_k nous utilisons dans cette thèse **l'algorithme de maximisation de la vraisemblance 'Expectation Maximisation (EM)'**. Comme son nom l'indique l'algorithme EM celui-ci est composé de deux étapes : l'étape 'Expectaction', qui consiste à estimer les probabilités d'appartenance aux modes en supposant les paramètres connus $P(Z_t = k|\theta)$, et l'étape de 'Maximisation' qui consiste à réestimer les paramètres θ de la loi de mélange. Nous utilisons ici la forme dite "par lot": toutes les observations sont utilisées pour mettre à jour les paramètres. On peut montrer [33], qu'à partir de valeurs initiales des paramètres du modèle, l'algorithme EM garantit la convergence vers le maximum local de la fonction de vraisemblance. L'algorithme EM est utilisé dans cette thèse pour l'inférence des paramètres des modèles utilisés dans les chapitres III à V.

- **Méthodes basées sur la diagonalisation de la matrice de covariances.** Les analyses en composantes principales (en anglais PCA) et leurs dérivées sont utilisées pour extraire les modes de covariance principaux d'une (PCA, EOF, [9, 35]) ou deux variables (Singular Value Decomposition, SVD, [9, 36]). Par conséquent, la segmentation de la covariance spatio-temporelle observée peut être directement associée à la caractérisation de modes distincts. Les analyses en composantes principales sont classiquement utilisées pour réduire la dimensionnalité des variables étudiées (principe de parcimonie). Nous les avons utilisées dans le chapitre IV pour caractériser l'influence spatio-temporelle d'une variable (la turbidité de surface) dans l'embouchure de la Gironde en résumant ce signal spatio-temporel à l'analyse de quatre composantes principales associées à leur coefficient d'expansion dans le temps.

- **Factorisation par matrices non négatives (Non Negative Matrix Factorisation NNMF).** Dans le même esprit que les ACP, elles visent à décomposer une matrice en un produit de matrices d'une base de projection par ses coordonnées. Contrairement aux ACP, la base de projection n'est pas forcément orthogonale. Pour les NNMF, la base et les coordonnées sont strictement positives. La positivé des coefficients de projections est particulièrement pertinente pour certaines variables géophysiques. Par exemple, nous utilisons les NNMF dans le chapitre V pour segmenter les spectres de réflectance in-situ en spectres de référence ou types d'eaux. La positivité des coefficients de projection nous permet de proposer en sortie de la procédure d'inversion des estimations des réflectances marines strictement positives, contrairement aux algorithmes de l'état de l'art.

- **Décomposition temps-fréquence d'une série temporelle.** Nous utilisons dans le chapitre III la représentation temps-fréquence du spectre d'énergie d'une série temporelle estimée par ondelettes [8]. La représentation temps-fréquence d'un signal peut être envisagée comme sa décomposition dans le temps en processus possédant des caractéristiques fréquentielles différentes. Dans le chapitre III, nous détectons des évènements dans le spectre d'énergie de la température de surface, séparant ainsi notre signal en une somme de contributions possédant des signatures fréquentielles ou temporelles distinctes.

0.2.3 Estimation et caractérisation de régimes physiques distincts

Nous distinguons ici la caractérisation de modes de celle de régimes. Dans la caractérisation de régimes, nous cherchons à distinguer des relations distinctes entre une variables observée Y et prédicteurs X. La caractérisation de régimes cachés, ou non, peut être alors considérée comme la caractérisation de modes dans la probabilité conditionnelle P(Y|X). Par exemple, nous déterminons dans le chapitre IV des régimes saisonniers pour la turbidité, soit des modes dans la distribution conditionnelle de la turbidité sachant la hauteur de la houle du vent et la marée. Les modèles de Markov cachés (HMM [38, 39]) sont une des familles des modèles espace-état:

- **Modèles espace-état**. Les modèles espace-état intègrent la distinction entre les variables observées (le signal) et les variables d'état Z potentiellement cachées. Ces modèles sont

constitués d'une ou plusieurs équation(s) d'observation et d'une ou plusieurs équation(s) d'état décrivant la dynamique entre les régimes. La distribution des erreurs peut suivre des lois normales ou non. Dans cette thèse, nous nous intéresserons au cas particulier des modèles espace-état linéaires à erreurs Gaussiennes. Dans la communauté statistique, ces modèles sont connus comme des modèles linéaires dynamiques [37].

Modèles Markoviens cachés. Le paradigme Markovien est que l'état Z observé au temps t dépend uniquement de l'état observé au temps t-1. Pour le modèle Markovien le plus classique nommé en Anglais 'Hidden Markov Model, HMM', cela signifie de manière sous-jacente que toute l'information nécessaire pour estimer la variable d'intérêt au temps t est entièrement contenue dans l'état estimé au temps t-1 (Z_{t-1}) et les X au temps t (X_t). Cette contrainte sur l'équation d'état permet également de pouvoir inférer plus aisément les paramètres des équations d'observation et d'état. Nous avons utilisé les HMM ainsi que les versions dérivées incluant des probabilités de transitions non-homogènes entre les états (NHMM) et un terme autorégressif (HMM-AR et NHMM-AR) pour la prévision de la turbidité dans l'estuaire de la Gironde (Chapitre IV).

0.3 Principaux résultats

Détection de tendances, biais et cycles saisonniers significatifs.

L'objectif scientifique est ici d'estimer l'impact des interruptions entre les différentes missions satellitaires, et inversement, la durée de recouvrement optimale entre celles-ci, pour minimiser les incertitudes sur une tendance, un biais, et un cycle saisonnier estimé. Cet aspect peut se révéler déterminant dans le futur pour les agences spatiales, à l'heure où les constellations de microsatellites embarquant un unique capteur sont désormais favorisées par rapport aux plateformes massives multi-capteurs de type ENVISAT[4]. En effet, la relative souplesse de ces micro-missions en termes de planification budgétaire permet désormais de prendre en compte de tels objectifs scientifiques dans leur planification. La méthodologie est également directement applicable aux réseaux d'observation in-situ, qu'ils soient côtiers (REPHY[5], MAREL[6], et SOMLIT[7]) ou hauturiers (stations Aeronet[8]) et notamment pour planifier les périodes de maintenance des instruments de mesures.

D'un point de vue méthodologique, nous généralisons le travail de Weatherhead [40] sur la détection d'une tendance dans une série avec un modèle linéaire Y=AX+b, à de multiples séries

[4] ENVISAT: ENVIronment SATellite
[5] REPHY: Réseau de surveillance du phytoplancton et des phycotoxines
[6] MAREL: Mesures Automatisées en Réseau pour l'Environnement
[7] SOMLIT: Service d'Observation en Milieu LITtoral
[8] Aeronet: Aerosol Robotic Network

[4]. Par rapport aux séries économiques et financières, pour lesquelles de nombreux travaux sur la détection de tendances ont été réalisés [41, 42], les spécificités des séries géophysiques sont la disponibilité relative des observations (discontinuités naturelles en cas de couverture nuageuse ou inhérentes à la mesure) et l'autocorrélation du bruit Φ (on parle souvent dans la littérature de bruits colorés [27]).

L'autocorrélation du bruit implique que les résidus de la relation linéaire ne sont plus indépendants entre eux, et la matrice de covariance γ des résidus n'est plus égale à la matrice identité multipliée par la variance du bruit. Dans ce cas, il n'est plus possible d'utiliser l'estimateur classique des moindres carrés (OLS) et l'estimateur à considérer est alors la méthode des moindres carrés généralisés (GLS).

L'autocorrélation naturelle est largement négligée, à tort, par la communauté scientifique car d'une part, elle augmente la complexité du système à résoudre, et d'autre part, cela ne change pas les valeurs des paramètres estimés, l'estimateur GLS étant non biaisé. Cependant, cela impacte directement les incertitudes associées aux paramètres estimés (tendances, biais, cycles). Dans le cadre d'études climatiques, Il est par conséquent essentiel de qualifier la significativité des paramètres A su modèle linéaire [43, 2] relativement à cet aspect.

Nous décrivons § 2.3.2.1 l'estimation de l'incertitude d'une tendance détectée à partir de deux séries temporelles en fonction, de leur situation temporelle relative (recouvrement ou interruption), du nombre d'observations disponibles dans chaque série, de σ_1^2, σ_2^2, ϕ_1, ϕ_2 les variances et autocorrélations des bruits supposés gaussiens de chaque série, et α le coefficient de corrélation entre les deux séries. Notre méthodologie est appliquée aux données de chlorophylle-a issues des capteurs satellitaires MERIS (2002-2010) et SeaWiFS (1998-2010). Les résultats sont aussi extrapolés à la future mission Sentinel-3/OLCI[9], programmée en 2015.

La Figure 6 illustre de façon synthétique l'impact d'une période de recouvrement ou d'interruption entre deux séries sur l'incertitude de la tendance estimée. Nous montrons qu'une période de recouvrement de 6 mois entre deux séries est optimale pour minimiser $\sigma_{\hat{\omega}}$, pour des conditions usuellement observées (niveau d'autocorrélation de 0.7 pour un bruit autorégressif du premier ordre, dit AR_1). Nous montrons également Figure 8 qu'il faudra en moyenne 53 mois d'observations du capteur de couleur de l'eau OLCI pour améliorer les estimations des tendances à l'échelle globale réalisées à partir des données de SeaWiFS et MERIS. Cette durée aurait été largement diminuée si une période de recouvrement entre OLCI et MERIS avait été observée, minimisant ainsi l'incertitude sur le biais inhérent à la mesure entre les observations de MERIS et d'OLCI (Figure 6, partie de droite).

[9] OLCI: Ocean Land Colour Instrument

Caractérisation d'échelles temporelles significatives dans la température de surface observée depuis l'espace de 1989 à 2005.

Nous estimons dans ce chapitre les échelles temporelles de référence d'un processus géophysique et caractérisons la distribution spatiale de ces échelles des interactions entre celles-ci. En rappelant les limitations inhérentes aux méthodes d'analyses classiques de type EOFs (Empirical Orthogonal Functions, similaires aux ACP), nous introduisons un nouveau concept, la représentation d'une série temporelle comme une somme d'événements significatifs par rapport aux conditions locales de bruit. D'un point de vue méthodologique, notre méthode est basée sur la détection automatique par contour de niveaux ('level-set', [44]) d'évènements significatifs dans des spectres d'ondelettes temps-fréquence, et la segmentation des descripteurs de la base de données évènements. Ce type d'approche est dérivée du 'datamining' plus connue de nos jours sous le nom de 'big-data' [45].

Pour illustrer notre approche, nous caractérisons l'impact d'ENSO sur la plus longue série d'observations disponibles à l'échelle de la planète: la température de surface de la mer de 1985 à 2009 [10]. L'estimation des échelles de référence d'ENSO est réalisée à partir d'une version modifiée de l'algorithme EM (cf §0.2.2) qui prend en compte la distribution fractale, naturellement observée dans la nature, des échelles temporelles des événements [10, 46]. Dans ce cas la distribution estimée est égale au produit d'un mélange de lois normales, identifiant les échelles de référence, par une loi exponentielle décroissante.

Nous avons estimé quatre échelles temporelles de référence à 1.54, 3.36, 5.03 et 7.11 années, répondant à des questions récurrentes de la littérature sur l'existence d'échelles distinctes dans la gamme de fréquences (1.5-7 années) généralement attribuées à ENSO [47, 48, 49, 50, 51]. L'analyse de la distribution spatiale des échelles de référence d'ENSO a permis de retrouver des signatures connues de la littérature, exhibées par les analyses de corrélations.

L'estimation des échelles de référence nous permet également de caractériser des changements de fréquences dans la signature d'ENSO dans la SST pendant l'épisode majeur de 1997 à 2000. Un changement maximal, de la haute vers la basse fréquence, est observé Figure 15 deux mois avant le pic MEI[10] en mai 1998 [10]. Ces changements de fréquence influencent directement les structures de SST observées, soit la dynamique de surface [52] et sont ignorés par les analyses classiques de corrélation.

Nous nous sommes également intéressés à la signature d'ENSO sur les évènements haute-fréquence de la SST, soulignant une distribution spatiale significativement différente selon les périodes dites 'normales', Niño et Niña. Cette signature physique d'ENSO souligne l'apport de notre approche car elle est ignorée, par construction, avec les analyses classiques de la covariance de la SST réalisées depuis 30 ans [9,35, 36].

[10] MEI : Multivariate ENSO Index

Caractérisation de régimes physiques dans des séries temporelles géophysiques et application à la prévision de la turbidité de surface.

Le troisième chapitre est relatif à la caractérisation de régimes physiques temporels entre une variable et ses prédicteurs. L'objectif est la modélisation d'une variable Y suivant un processus physique fortement non-stationnaire, car soumis à des forçages saisonniers, en utilisant des processus linéaires Y=AX+b à changements d'états Markoviens [11]. Quatre extensions du modèle Markovien à états cachés (HMM) sont développées, dans lesquelles la chaîne de Markov devient non-homogène et un terme autorégressif est ajouté. Dans le cas non-homogène, la matrice de transitions entre les états Z est exprimée en fonction d'une loi normale multivariée décrivant la densité de probabilité des covariables lors des transitions. Les covariables sont ici des paramètres géophysiques influençant les transitions entre deux régimes, et dans notre cas les prédicteurs X observés à t-2. L'estimation des paramètres des modèles d'observation, c.-à-d. les coefficients des régressions et la variance des résidus pour chaque régime (cf. § 4.2.1), et des paramètres de transition, est réalisée de façon simultanée par maximisation de la vraisemblance avec l'algorithme EM.

Nous abordons également un problème important de la modélisation statistique qui est celui de la sélection des prédicteurs, des covariables, et du nombre d'états cachés. Pour cela nous utilisons des analyses linéaires discriminantes pour la sélection des prédicteurs (variables explicatives) et un critère de type 'log-vraisemblance pénalisée', le BIC [12], pour sélectionner le modèle optimal (parmi les quatre extensions développées) et le nombre d'états cachés (le nombre de régimes).

L'application réalisée est la prévision de la turbidité estimée par satellite dans l'embouchure de la Gironde en fonction de ses variables de forçage : la houle, le vent, la marée et le débit de la Gironde. Nous montrons qu'un nombre optimal de trois régimes est nécessaire pour modéliser la relation complexe entre la turbidité et ses prédicteurs. Nous comparons ensuite les résultats obtenus avec les méthodes couramment utilisées en océanographie comme les régressions linéaires multivariées classiques, et un modèle non-linéaire de type SVR (Support Vector Regressions, [31]). Les résultats montrent un gain de performance de l'ordre de 150% pour prédire la variable d'intérêt par rapport aux régressions multivariées classiques et de 40% par rapport au modèle non-linéaire à apprentissage SVR, soulignant la capacité du mélange de relations linéaires cachées pour modéliser un processus physique fortement non-stationnaire.

La modélisation de la matrice transitions à partir de la distribution des covariables est particulièrement pertinente puisque le modèle NHMM-AR fournit globalement les meilleurs résultats. Cette modélisation des transitions permet également de caractériser physiquement les changements de régimes, comme par exemple pour notre application l'arrivée des houles d'automne sur la côte Landaise.

Nous montrons également qu'en l'absence d'observation (conditions nuageuses), et pour des périodes courtes inférieures à 15 jours, il demeure plus intéressant de conserver un modèle comportant un terme autorégressif Y_{t-1} dans les prédicteurs X. Dans ce cas, $\hat{Y}_{t-1}$ est estimé à partir de la dernière observation disponible, des covariables disponibles (ici des sorties de modèles), et de la matrice de transition.

Modèles Bayésiens cachés pour l'inversion de la réflectance de la surface de la mer en milieux côtiers complexes

Le quatrième et dernier chapitre détaille nos recherches sur l'amélioration des corrections atmosphériques en milieu côtier [13]. Ce travail se situe par conséquent en amont de l'étude des séries temporelles mais est aujourd'hui nécessaire pour pouvoir proposer des séries temporelles de réflectance de la surface de la mer, et des produits géophysiques dérivés, plus proches de la réalité dans les milieux côtiers. Historiquement, les données de couleur de l'océan ont été divisées, selon leurs propriétés optiques, en deux catégories par la communauté scientifique: les eaux claires dites de cas 1 et les eaux turbides de cas 2 (généralement côtières). L'estimation des réflectances des eaux de cas 2 à partir des observations satellitaires TOA est toujours, de par sa complexité, un axe de recherche d'actualité et prioritaire car les zones concernées sont celles directement liées aux activités anthropiques.

Notre méthodologie est basée sur l'estimation et la caractérisation de modes dans les distributions multivariées jointes de variables et covariables. Dans le cas de la correction atmosphérique, soit la séparation du signal marin et atmosphérique, l'identification des modes consistera en la caractérisation des formes spectrales de référence de la réflectance des aérosols, qui correspondent à une réponse de l'atmosphère aux aérosols présents, et des formes spectrales de référence de la réflectance marine, conditionnées par les constituants optiquement actifs de la colonne d'eau. Les covariables sont ici des paramètres géophysiques significativement corrélés aux variables aléatoires d'intérêt, et dans ce cas précis, les variables décrivant la géométrie d'observation et des estimations de la réflectance marine et des aérosols dans le proche infra-rouge.

Nous avons caractérisé 10 modes d'aérosols côtiers et 9 modes de spectres marins à partir d'un jeu de données colocalisées satellite/in-situ [53]. Les modes estimés sont ensuite utilisés pour optimiser l'inversion de la réflectance de la surface de la mer à partir d'observations TOA : lors de la phase d'inversion, les distributions à priori des variables sont corrigées en fonction des valeurs des covariables pour optimiser les initialisations de l'algorithme Bayésien (cf. [13]). Le modèle de réflectance de l'eau est une projection sur la base de spectres de référence issue de la factorisation en matrices non-négatives des spectres in-situ, utilisés pour entraîner nos modèles. La contrainte de positivité sur la réflectance marine assure une convergence vers des solutions ayant toujours un sens géophysique, ce qui n'est pas le cas avec les chaînes de traitement de l'ESA

et de la NASA pour lesquelles des réflectances négatives de la surface de la mer sont couramment observées en milieu côtier.

Nous comparons ensuite, en utilisant le jeu d'observations colocalisées, les estimations réalisées à partir de notre méthode, aux estimations fournies par la chaîne standard de l'ESA pour MERIS (MEGS v8, [20]) et à celles fournies par un réseau de neurones entrainé sur les mêmes eaux (C2R[11], [54]). Nous avons amélioré les estimations des réflectances marines entre 412 et 865 nm, en moyenne de 67 % par rapport à la chaîne ESA classique et 9% par rapport au réseau de neurones.

Nous détaillons finalement les étapes à accomplir pour arriver à un produit opérationnel pour la future mission OLCI/Sentinel 3.

0.4 Conclusions et perspectives

Dans cette thèse, plusieurs familles de modèles sont proposées pour analyser les séries temporelles géophysiques de couleur de la mer et de température de surface des océans. Parmi ceux-ci une attention particulière est portée aux approches de segmentation multivariées, de type mélange de modèles Gaussiens, et aux modèles espace-état à changement de régimes Markoviens (HMM, HMM-AR, NHMM, NHMM-AR). L'aspect multi-modes ou multi-régimes des analyses mises en œuvre est intégré dans l'analyse thématique des résultats obtenus: dans chaque cas nous essayons de caractériser physiquement les modes ou régimes. De notre point de vue, cette **caractérisation thématique des modes ou des régimes estimés renforce l'intérêt de nos approches par rapport aux modèles à apprentissage automatique de type réseaux de neurones**.

Le chapitre II souligne **la nécessité d'intégrer les spécificités de la variable géophysique** (nature du bruit, discontinuité des observations) **lors de l'inférence des paramètres à estimer**. Les perspectives directes de la méthodologie développée sont l'optimisation, dans un but de surveillance environnementale, de la planification des missions satellites et des réseaux in-situ. D'un point de vue académique, avant cette thèse, il n'y avait pas de cadre méthodologique publié sur la détection de tendances dans plusieurs séries géophysiques. La publication réalisée doit permettre par conséquent de sensibiliser les géophysiciens et thématiciens aux interactions entre les caractéristiques spécifiques du signal géophysique et les incertitudes des paramètres estimés.

Dans la continuité du chapitre II nous caractérisons dans le chapitre III un phénomène relativement à sa nature géophysique : un évènement significatif de SST par rapport aux

[11] C2R MERIS Regional Case 2 Water Algorithms (C2R)

conditions locales de variance et d'autocorrélation du bruit. D'un point méthodologique, le chapitre III souligne **l'apport de la double approche; discrétisation d'un signal spatio-temporel en une somme d'évènements significatifs temps-fréquence et segmentation des descripteurs, pour caractériser un phénomène géophysique**. La méthodologie permet une analyse plus fine, mais également plus complexe, du phénomène étudié par rapport aux approches classiques d'analyses de covariance. La caractérisation des échelles de référence du signal ENSO dans les basses fréquences, et de sa signature claire jusqu'alors insoupçonnée sur la haute fréquence de la température de surface, sont les exemples typiques de la valeur ajoutée de notre approche scientifique par rapport à l'état de l'art.

Les perspectives pour ce type d'approche sont multiples. Nous nous sommes principalement intéressés à la distribution des échelles temporelles des évènements. Les descripteurs de l'évènement contiennent de nombreuses informations pertinentes pour la description du processus physique considéré qui n'ont pas été traitées ici (par exemple la pente de l'ellipse décrivant l'évènement est un proxy de sa propagation). D'un point de vue global, **l'analyse des distributions jointes des descripteurs de différentes variables** (température, chlorophylle-a, vent…) **contribue à l'établissement d'un cadre méthodologique pour mieux appréhender les notions de dépendances et causalités entre les variables**.

La modélisation d'un processus géophysique fortement non stationnaire, la turbidité de surface, **à partir de modèles à régimes cachés suivant un processus Markoviens**, augmente considérablement la capacité de prévision par rapport aux régressions multivariées classiquement utilisées en océanographie. Elle ouvre clairement la voie **vers la prévision opérationnelle** de la turbidité de surface **avec ce type de modèle**. Cette modélisation sera particulièrement intéressante pour des zones où la bathymétrie et les conditions aux limites sont mal maitrisées : dans ce cas les modèles hydrodynamiques classiques ne pourront produire des simulations réalistes. L'extrapolation à la colonne d'eau pourra être réalisée à l'aide de nouveaux régimes, liant par exemple la morphologie du profil vertical de turbidité aux observations de surfaces. Les applications potentielles pour d'autres types de données sont multiples tant de nombreuses variables géophysiques sont intrinsèquement gouvernées par des changements de régimes. La chlorophylle-a, la croissance des coquillages et des poissons sont par exemple des variables typiquement marquées par des alternances de phases actives et passives gouvernées par des facteurs environnementaux.

La caractérisation à priori de spectres de référence de la réflectance des aérosols et de la surface de la mer, pour l'inversion de la réflectance de surface de la mer en milieux côtier, est novatrice par rapport aux approches de l'état de l'art, telles que l'inversion séquentielle (chaîne MEGS) et les réseaux de neurones (C2R). Nous avons obtenu des améliorations significatives sur les estimations entre 412 et 490 nm, soit un domaine spectral utilisé pour estimer de nombreux paramètres géophysiques: la chlorophylle-a, le CDOM et la transparence de l'eau. Au-delà de

l'aspect analytique, ce travail nous rappelle que les valeurs géophysiques télédétectées ne sont que des estimations, et qu'il convient par conséquent de conserver à tout moment un regard critique sur la qualité de ces estimations et des analyses subséquentes. Ceci est particulièrement vrai dans les zones côtières ou les milieux sont complexes et les interactions avec la côte importantes.

Les perspectives de l'approche Bayésienne avec a priori sont l'amélioration des produits satellitaires fournis par les agences et une meilleure surveillance des zones typiquement influencées par l'homme et soumises aux politiques environnementales des états. La généralisation du modèle Bayésien pour l'inversion devra aussi être envisagée. Actuellement nous discrétisons des spectres sur des critères de forme et d'amplitude. Il serait intéressant de généraliser l'approche en estimant indépendamment la forme de l'amplitude en utilisant par exemple les distributions multivariées pour les angles (Von Mises [55]). Dans ce cas, l'estimation de l'amplitude lors de l'inversion demeure complexe. Enfin, notre approche bayésienne peut être appliquée à d'autres variables montrant intrinsèquement des modes dans leur distribution. La caractérisation des types de phytoplanctons à partir de spectres de reflectances de la mer est un exemple typique qui pourra être traité avec cette approche. Dans ce cas, la segmentation sera réalisée soit directement sur les spectres, relativement à nos connaissances des espèces observées in-situ, ou sur des descripteurs de ces spectres associés à des covariables d'environnement.

L'aspect spatial a été peu traité dans cette thèse et l'analyse simultanée des dépendances spatiales et temporelles est de mon point de vue la suite logique de cette thèse. La covariance spatio-temporelle entre plusieurs variables géophysiques (comme par exemple la chl-a et la SST) montre typiquement des modes distincts en fonction de la localisation spatiale et du temps. Ces modes de corrélations entre variables peuvent être d'un intérêt certain pour l'estimation de données manquantes. Nous pourrons à l'avenir, mieux estimer un pixel non observé de chl-a (à cause de la présence d'un nuage) en utilisant les a priori sur les modes de covariances entre par exemple la chl-a et la SST. En comparaison avec les techniques classiques d'interpolation optimale, qui utilisent la modélisation de la covariance spatio-temporelle $\gamma(d)=f(d)$ d'une seule variable en fonction de la distance d, la covariance $\gamma'(d)$ sera alors estimée par l'espérance conditionnelle de γ', sachant les observations de SST et chl-a et de potentielles covariables.

D'un point de vue général, les perspectives méthodologiques à long terme de cette thèse sont de mon point de vue **la généralisation des approches Bayésiennes multi-modes, multi-régimes, pour l'inversion et la caractérisation des variables géophysiques**. Que ce soit pour l'inversion d'un paramètre à partir d'observations, les analyses des interactions entre variables, il apparait selon mon expérience, que beaucoup de ces problèmes ont finalement des aspects multi-régimes, multi-modes, de par la nature intrinsèque 'géophysique' des variables considérées. Il conviendra alors certainement de travailler sur des paramétrisations plus complexes: des régimes non-linéaires, des distributions non gaussiennes pour les mélanges et les résidus.

A la fin de cette thèse mon sentiment personnel est que le potentiel du traitement du signal et des statistiques pour traiter des questions scientifiques classiques, ou novatrices, de la Couleur de l'eau, est sous-exploité et par conséquent en devenir. Ceci vient de la nature interdisciplinaire requise pour mettre en place ce type d'approche. L'héritage historique des communautés scientifiques qui ont évoluées séparément, comme par exemple les thématiciens, les statisticiens et les modélisateurs, peut également compliquer la mise en place de ces collaborations. On observe également dans certains cas des appréhensions face au caractère novateur des méthodes proposées. La rareté et la structure même des financements des projets potentiellement concernés, qu'ils soient nationaux ou Européens, ne favorisent également pas toujours ces approches interdisciplinaires. Comme initié pendant cette thèse au travers des collaborations réalisées avec Telecom Bretagne, le CNRS et l'IRD, et au regard du potentiel de ces approches, mon objectif personnel sera à l'avenir le développement des approches pluridisciplinaires pour le traitement des données géophysiques.

1 Chapter 1: Introduction

Climate change analysis, characterization of major climatic events, and characterization and forecasting of geophysical processes, have an influence on state policies. For example, since the protocol of Kyoto (signed in 1997 and applied in 2005) the European commission has defined in 2005 the foundation of a strategy to address the climate change [1]. This strategy requires the development of new laws in coordination with the European's states, the reinforcement of international cooperation and research, and the organization of citizen awareness campaigns. Scientific analyses that drive these policies are performed using either spatio-temporal in-situ or satellite time series, or modelling simulations using these observations as forcing conditions.

For the past thirty years, sensors embedded on satellite platforms have provided most of the sea surface observation time series. These are now long enough to characterize weak spatio-temporal variations in the geophysical variables measured at the top of atmosphere, or estimated using these latest.

During this time, the temporal variation of a geophysical variable has been envisaged using three different aspects:

- The best known is **the long term trend**, which may be either linear [2, 3, 4] or not [5, 6]. The long term trend has been largely used in the estimation of the climate change impact [7].

- **Analysis of the spatio-temporal correlations within a single or between datasets** [5, 8]. Typically, Principal Component Analysis (PCA) or Empirical Orthogonal Functions (EOF, [9]) approaches aim at decomposing the covariance in orthogonal modes. Each mode is then traditionally attributed to different forcing conditions such as seasonal, large time-scale events and local signals. Such approaches aim at retrieving spatio-temporal modes in the covariance matrix, underlying the need to unmix the geophysical processes within a global signal to be able to study them separately.

- **Characterization of time-varying physical processes**. We define a regime as the relationship between a variable of interest Y and its predictors. A reasonable question is whether the variable of interest better estimated using multiple regimes or a single linear or non-linear regime? This is a particular field of interest for geophysical processes that are often driven by seasonal signals: the seasonality often leads to varying relationships between the variable of interest and its forcing parameters. The estimation and the characterization of these relationships are thus particularly crucial to estimate, inverse, or forecast the considered variable.

In this thesis, we focus on the temporal variations of Ocean Color (OC) and Sea Surface Temperature (SST) observed from space. We address the four following scientific questions:

- The estimation of significant trends, bias, and seasonal cycles among multiple geophysical datasets.
- The spatio-temporal analysis of a major climatic signal.
- The modelling and forecasting of a geophysical variable driven by seasonal processes.
- The inversion of a geophysical variable.

From a methodological point of view, the 'geophysical nature' of the signal, which often implies some discontinuities of the observations and autocorrelation of the noise, requires specific topics to be addressed. Among them, we will distinguish:

- The regression of a random geophysical variable.
- The estimation and characterization of distinct modes in multivariate signals. A mode refers here to an elementary component of a mixture.
- The estimation and the characterization of distinct physical regimes, i.e. relationships between a variable Y and its predictors X.

The manuscript is organized as follows. Chapters II to V address the scientific questions raised and are structured around published (II-IV) or submitted (V) articles.

Chapter II describes the methodology we have developed to characterize significant trends, bias, and seasonal cycles in multiple geophysical time series. Our methodology accounts for missing data, as observed during cloudy conditions for the satellite-derived ocean color measurements, bias between time series and the local characteristics of the noise. The main objective of this research topic is to estimate the impact of gaps, and conversely the optimal overlap duration, between missions, to minimize the uncertainty on the estimation of a potential long term trend. Our methodology is applied to the MERIS[12] (2002-2010) and SeaWiFS[13] (1998-2010) chlorophyll-a datasets and we extrapolate our analysis to the incoming OLCI[14] sensor that will be embedded on the Sentinel 3 mission (scheduled for 2015). Our methodology is also directly applicable to in-situ monitoring networks, from the coastal networks (REPHY[15], MAREL[16], and SOMLIT[17]) to the seafaring networks (Aeronet[18] stations). In this latest case, it can be also used to schedule maintenance procedures.

In chapter III, we propose a methodological contribution to analyze a reference climatic signal. The objective is the estimation and the characterization of the reference time-scales of a geophysical process. Regarding drawbacks of the state of the art EOF, we introduce a new concept: the

[12] MERIS: MEdium Resolution Imaging Spectrometer
[13] SeaWiFS: Sea-Viewing Wide Field-of-View Sensor
[14] OLCI: Ocean Land Colour Instrument
[15] REPHY: Réseau de surveillance du phytoplancton et des phycotoxines
[16] MAREL: Mesures Automatisées en Réseau pour l'Environnement
[17] SOMLIT: Service d'Observation en Milieu LITtoral
[18] Aeronet: Aerosol Robotic Network

representation of a time series as a sum of time-frequency events showing a significant energy relative to the local conditions. Our method relies on the 'level-set' extraction of significant events in the wavelet estimated power spectrum, and the segmentation of the event descriptor database [44]. Our approach is derived from the 'datamining' process often referred as 'big data'. To illustrate our approach, we study the impact of the El Niño-Southern Oscillation (ENSO) on the longest dataset available at global scale: the sea surface temperature observed from space between 1985 and 2009.

The fourth chapter addresses the characterization of physical regimes between a variable and its predictors. The objective is to model a complex and non-stationary geophysical variable, driven by seasonal forcing conditions, using Markovian models and available observations and model outputs. Various extensions of the model are considered. In these extensions, transitions between states (i.e. the regimes) become non-homogeneous and are conditioned by external covariates. The inclusion, or not, of an autoregressive term is also discussed. In the non-homogeneous transition cases, the transition matrix is modelled using the MultiVariate Normal (MVN) distribution of the covariates. Estimation of the observation and state model parameters is then performed simultaneously by maximizing the likelihood with the 'Expectation Maximization' (EM) algorithm.

As an example, we analyze the time-varying relationships between the suspended matters concentration observed from space in the French Gironde's estuary, and its forcing variables (predictors), namely waves, winds, tides and Gironde's outflows. We compare our estimates with the state of the art approaches, namely standard multivariate regression, which is the most popular regressor used in the Oceanographic community, and Support Vector Regressions (SVR) non-linear machine learning model.

The fifth and last chapter details our research topic on the enhancement of ocean color products in coastal areas [13]. This topic is upstream to the time series analysis but is necessary to propose time series of marine reflectances, and derived geophysical products, closer to the in-situ observations in such areas. The methodological approach relies on the estimation and characterization of Gaussians modes in the joint multivariate distribution of the observed variables and covariables. Covariates are geophysical parameters significantly correlated with the variable of interest. The estimated modes are then used to optimize the statistical inference to be completed, i.e. the inversion of the marine reflectance. For that purpose, the prior distributions of variables, sometimes referred as priors, are corrected using the covariate observed values to optimize the 100 random initializations of our algorithm (MEETC2, [13]).

Using an observation collocated dataset between MERIS and in-situ observations [53], we compare the estimates provided by our method, with the actual MEGS[19] v8 ESA[20] processing chain for MERIS [20] and outputs provided by a neural network trained using the same in-situ and satellite dataset (C2R[21], [54]). Finally, we detail the steps necessary to provide an operational product for the incoming OLCI/Sentinel 3 mission.

Annex A contains a scientific article completed and published during the first year of this thesis. It describes water transparency estimated using high resolution MERIS observations (250m). Although not directly related to this thesis, this article may contain thematic and contextual information of interest to provide readers a better understanding of the thematic concerns that we address in this thesis.

1.1 Context

The thesis is has been carried out at partial times within a SME, the company ACRI-ST based in Sophia Antipolis and in collaboration with the French school Telecom Bretagne (ENSTB). ACRI-ST works for spatial agencies and has been involved in environmental monitoring projects for 20 years. For SME the valuation of the performed research is a fundamental aspect and the proposed thesis contains both methodological and applied dimensions. To achieve this work, different skills were gathered. Firstly, the research on the mathematical and statistical models is completed by me under the supervision and with inputs of Dr. Ronan Fablet & Dr. Grégoire Mercier, professors at ENSTB, and an external collaborator, Dr. Pierre Ailliot, professor of mathematics at the French University of Brest. The thematic aspects, the definition of the scientific questions and the result analysis, are completed by me. Some specific analyses were performed in collaboration with scientists, Dr. Ludovic Bourg, Dr. Odile Fanton d'Andon, Dr. Antoine Mangin from ACRI-ST, Dr.David Antoine & Dr.David Doxaran, oceanographers at the French CNRS research institute, Dr.Hervé Demarcq, oceanographer at the French IRD research institute. Lastly, large satellite derived dataset provision and programming were essentially performed principally by myself with some inputs from the ACRI-ST IT team, Julien Demaria, Dr. Philippe Garnesson, and The Globcolor team at ACRI-ST.

[19] MEGS: MERIS Ground Segment development platform
[20] ESA: European Space Agency
[21] C2R: Case2 Regional

2 Chapter II: Detection of linear trends in multi-sensor time series in presence of auto-correlated noise: application to the chlorophyll-a SeaWiFS and MERIS datasets and extrapolation to the incoming Sentinel 3 - OLCI mission.

This chapter addresses an important methodological issue in climate change studies: how to estimate long-term trends in multiple geophysical time series. From a methodological point of view it addresses the estimation of model parameters in geophysical time series, i.e. relative to their intrinsic specificities, and in particular the natural autocorrelation of the noise. Due to noise autocorrelation, the residuals of the linear regression Y=AX+B are no longer uncorrelated, and the estimation of A is not anymore completed using the standard Ordinary Least Square (OLS) estimator or the ordinary maximum likelihood estimator. In this case, the Generalized Least Square (GLS) must be considered as it involves, conversely to the OLS, an estimated covariance matrix for the residuals. In practice, a transformation is applied to the variables X and Y to take into account the noise autocorrelation [29], and the GLS estimator finally resorts to the standard OLS estimator.

Although the noise autocorrelation in geophysical time series does bias $\hat{A}$ (the GLS estimator is unbiased [57]), it affects $\sigma_{\hat{A}}$ and consequently the ability to detect, or not, a significant trend and other estimated parameters. Ignoring the noise autocorrelation level, as often seen in published climatic analysis, typically leads to an over detection of significant trends.

Due to satellite lifetime, usually between 5 and 10 years, satellite-derived time series do not cover the same period and are acquired by different sensors with different characteristics. These differences lead to unknown level shifts, often referred as biases in literature. The estimation of parameters A, in our case the level-shift between two time series, the trend, the seasonal cycles and the noises must be completed at the same time. A previous bias correction of one time series relative to the other, before the parameter estimation (as often seen in many published studies), resorts to a single time series analysis and to the underestimation of $\sigma_{\hat{A}}$.

This chapter was published in 2013 in the 'Journal of Geophysical Research Oceans (JGR)' [4].

2.1 Introduction

A variety of studies have addressed the detection of long-term trends in auto-correlated processes. Tiao *et al.* [2] showed that the trend estimation uncertainty is strongly affected by the variability and the autocorrelation of the underlying noise process. Environmental data typically involve strong autocorrelation level [Frankignoul *et al.,* 27]. For instance, a positive anomaly in the

observed wind or temperature on a given day is often associated with similar conditions the following days. This natural autocorrelation is the result of local conditions but also of large-scale signals such as for instance the well-known El-Niño-La-Niña oscillation [Philander, 58; Torrence *et al.*, 59]. It also implies that the day-to-day or month-to-month observations are no more independent one from each other, and that the 'real' number of independent observations available to detect a trend is significantly lower than in uncorrelated cases. [Clifford *et al.,* 60; Tiao 2; Dutilleul, 61].

Since the end of the 1970s, satellite ocean-color observations have been providing large-scale measurements of the water-leaving radiance [McClain, 62], i.e. the light intensity estimated at the surface of the ocean at different wavelengths in the visible from 400 to 700 nm and near infrared. These radiances are used as inputs of inversion algorithms to retrieve biogeochemical parameters. Among available ocean-color variables, the most popular is the chlorophyll-a (chl-a) concentration [Maritorena *et al*. 63; O'Reilly *et al.*, 64; Morel *et al.*, 65], which is used in this work. The limited lifetime of space-based sensors implies that the long-term variability of such geophysical parameter can only be evaluated using a combination of time series. Among the historical ocean color sensors, the most widely used are the NASA Sea-viewing Wide Field of view Sensor, SeaWiFS [Hooker *et al.*, 66] that operated from September 1997 to December 2010, the ESA MEdium Resolution Imaging Spectrometer Instrument, MERIS [Rast *et al.*, 67], in activity from April 2002 to April 2012, and the NASA Moderate Resolution Imaging Spectroradiometer, MODIS-Aqua [Salomonson *et al.*, 68] launched in July 2002 which is still operational. Ocean color data with limited wavelength range are also available from the Coastal Zone Color Scanner (CZCS), which operated from 1978 to 1986 [Evans and Gordon, 69].

Trend estimation using the single SeaWiFS dataset has been previously addressed using different methods. Gregg *et al.* [70] estimated trends in the chl-a over the period 1998-2003 using a classical linear trend estimation. Recently, Vantrepotte *et al.* [71] used the census X11 method (adapted from Pezzulli *et al*, [72]) and Henson *et al.* [36], a simple model based on a three-components decomposition according to a seasonal signal, a linear trend and an auto-correlated noise, to estimate trends over the period 1998-2007. Trend estimation from multi-sensor datasets have been impaired until now because of inter-calibration uncertainties among available data sets. For instance, Antoine *et al*. [73] reanalyzed the CZCS and SeaWiFS time series to study the chl-a changes between these two missions, but they could not attribute to the changes they observed to a long-term trend.

Here, we go beyond inter-calibration issues and deal with the detectability of a linear trend or its significance from multi-sensor datasets. From a methodological point of view, we extend the statistical analysis of linear trends in single-sensor time series in presence of auto-correlated noise [Tiao *et al.*, 2; Weatherhead *et al.*, 3; Henson et al., 36] to multi-sensor time series. In particular we address both time overlaps and time gaps between time series. We report and discuss an application to the MERIS and the SeaWiFS chl-a datasets, which clearly demonstrate the gain of a multi-sensor analysis. We propose here a simple oceanographic description of the observed trends

assuming that the full understanding of the long term trends in the chl-a should be studied in conjunction with the temperature and the sea surface level.

Besides, we investigate how the time overlap between successive satellite missions could be optimized to improve the detectability of long-term trends. The Global Monitoring for Environment and Security (GMES) Sentinel-3 (S3) mission should be launched in 2015. This mission will carry the Ocean and Land Color Instrument (OLCI), an imaging spectrometer that will deliver multichannel wide-swath optical measurements of ocean and land surfaces, providing a new time series of chl-a observed from space. We also exploit the proposed statistical methodology to evaluate the duration of the S3-OLCI observation series required to improve the joint SeaWiFS-MERIS trend detection based on the hypothesis that the OLCI-MERIS level shift uncertainty will be of the same magnitude as the SeaWiFS-MERIS one.

2.2 Trend estimation

In table Table 2 is listed the used symbols with their description.

Table 2: list of symbols. Units are relative to the studied parameter, here, the chl-a.

Symbol	Designation	Unit
y_t	2D time series	$mg.m^{-3}$
M	Intercept	$mg.m^{-3}$
ω	Linear trend	$mg.m^{-3}.year^{-1}$
σ_ω	Uncertainty of trend	$mg.m^{-3}.year^{-1}$
S_t	Seasonal component	n.a.
N_t	Auto-correlated (red) noise	n.a.
σ_N^2	Red noise variance	$mg^2.m^{-6}.year^{-2}$
ϵ_t	White noise	n.a
σ^2	White noise variance	$mg^2.m^{-6}.year^{-2}$
φ	Noise autocorrelation	No unit
δ	Level shift	$mg.m^{-3}$
$\lvert\omega\rvert / \sigma_\omega$	Trend detection variable	No unit

α	Correlation between Y_{1t} & Y_{2t}	No unit
n	Length of the time series	months
T_0	Start time of the second time series	months

2.3 Statistical modeling

2.3.1 Single-sensor dataset

The observed geophysical time series, y_t, are modeled as a sum of three components: a long-term linear trend, a seasonal pattern, and a noise process, as follows:

$$y_t = \mu + \omega.\mathrm{t} + S_t + N_t, \quad \mathrm{t} = 1..\mathrm{n} \tag{6}$$

where n is the length of the time series, μ is the intercept term, ω the linear trend, S_t the seasonal component which includes annual and semi-annual terms. We chose here a similar representation of S_t as in Weatherhead *et al.* [3]:

$$S_t = \sum_{i=1}^{4} \mathrm{a_i}.\cos\left(\frac{2\pi it}{12}\right) + \mathrm{b_i}.\sin\left(\frac{2\pi it}{12}\right) \tag{7}$$

Here S_t is identical from year to year with a null sum over a year (S_t does not contribute to a global trend). N_t is the correlated noise (red noise), assumed to be a first order autoregressive process, AR_1:

$$N_t = \phi.N_{t-1} + \epsilon_t \tag{8}$$

Where ϵ_t is a white noise, i.e. an independent random variable with zero mean and variance σ^2. The stationary condition for N_t imposes that -1< φ <1. In presence of autocorrelation, the residuals are no longer independent, and the calibration of model Eq.(6) involves a generalized least square estimator, GLS [Aitken, 74; Russel, 41]. The latter relies on the estimation of the covariance matrix γ of the residuals N_t, generally unknown. For φ=0, γ is diagonal with term value equal to the variance of the white noise. If $\phi \cong 0$, the diagonal terms are still equal to the variance of the red noise and the other terms are estimated as lagged covariance between noise realizations:

$$\mathrm{cov}(N_t, N_{t+n}) = \sigma^2 \frac{\phi^n}{1-\phi^2} \tag{9}$$

In practice, the estimation of the model parameters in Eq.(6) may be completed using several methods. Among them Prais-Winsten [Prais *et al.*, 75] and Cochrane and Orcutt [29] methods aim at transforming Eq.(6) in an expression involving an uncorrelated noise residual [Tiao *et al.*, 2; Weatherhead *et al.*, 3]:

$$y_t^* = \mu^* + \omega . t^* + S_{t^*}^* + \epsilon_t \quad (10)$$

Given Eq.(10), the standard OLS estimator may be used [Aitken, 74; Russel, 41]. The trend estimation is not affected by the noise autocorrelation but its uncertainty strongly depends on ϕ [Tiao, 2]:

$$\sigma_\omega = \frac{\sigma}{(1 - \phi).\sqrt{\sum_{i=1}^{n}(t - \bar{t})^2}} \quad (11)$$

σ_ω can be expressed as a function of the white noise variance, $\sigma^2 = \sigma_N^2.(1 - \phi^2)$, with σ_N^2 the red noise variance, G, the trend coefficient uncertainty defined as the uncertainty on the trend estimate normalized with respect to the white noise variance σ:

$$\sigma_\omega = \sigma \, . G(n, \phi) \quad (12)$$

with $G = \frac{1}{(1-\phi).\sqrt{\sum_{i=1}^{n}(t-\bar{t})^2}}$

2.3.2 Multi- sensor dataset

We investigate a generalization to datasets acquired by different sensors for possibly different time periods. While the increase of the number of observations may decrease the variance of the trend estimation compared to the single-sensor case, the presence of unknown level shifts between the time series may significantly affect the uncertainty of the trend estimation. For the sake of simplicity, we consider in the subsequent a two-sensor dataset, but the proposed framework generalizes to three or more sensors. Given a two-sensor dataset, we assume that the two time series share the same long-term trend and seasonal patterns but involve an unknown level shift and correlated noise processes:

$$\begin{aligned} y_t &= \mu + \omega . t + S_t + N_{1t}\,, \qquad t = 1..n_1 \\ y_t &= \mu + \omega . t + \delta . U_t + S_t + N_{2t}, \quad t = T_0..n_2 \end{aligned} \quad (13)$$

where the time t is in any case relative to the start of the first time series, which is considered as the reference. T_0 is the starting time of the second time series, and n_1, $n_2 - T_0 + 1$, are respectively the length of the first and second time series. When $T_0 < n_2$ we observe an overlap between the two series and when $T_0 > n_2$, a gap. μ and ω are respectively the intercept term and the linear trend

shared by the two time series. δ is the unknown level shift of the second time series compared to the first one, supposed here as constant in time. U=1 for t >= T_0 and U = 0 for t < T_0. N_{1t} and N_{2t} are the auto-correlated noises of the two time series.

The estimation of the level shift between the two time series using an inter-calibration procedure [Johnson et al.,76] prior to the estimation of the shared linear trend is statistically relevant if one accounts for the uncertainty of the level shift in the variance of the trend estimate. Neglecting this uncertainty, as sometimes observed, resorts to a null-shift case, i.e. the study of a single time series. This is equivalent to considering a single time series and would greatly underestimate the variance of the trend estimate. To fit model parameters in Eq.(13), we consider an iterative procedure adapted from the Cochrane & Orcutt transformation [29].

2.3.2.1 Resolution

Only the estimates obtained after convergence that satisfy the 95 % detection threshold are considered in our analysis. This procedure leads to the estimation of the model parameters u, ω, S, δ as well as the variance of these estimates, and the variance of the uncorrelated residuals σ ².

Equation (13) resorts to:

$$y_t = \mu + \omega.t + \delta.U_t + S_t + N_{1t} + N_{2t}$$

To handle with the autocorrelation term, the following transformation is applied to resort to uncorrelated variables. For periods where only one time series is present, the standard Cochrane & Orcutt transformation is applied:

$$y_t^* = y_t - \phi.y_{t-1}$$

- When only the first time series is present, Eq.(13) turns in:

$$y_{1t}^* = \mu(1-\phi_1) + \omega.\phi_1 + \omega.(1-\phi_1).t + \epsilon_{1t} \quad , t = 2..n_1 \;\&\; \epsilon_{1t} \sim N(0, {\sigma_1}^2)$$

- When the second time series is present, Eq.(13) turns in:

$$y_{2t}^* = \mu(1-\phi_2) + \omega.\phi_2 + \omega.(1-\phi_2).t + \delta.(1-\phi_2).t + \epsilon_{2t} \;, t = T_{0+1}..n_2 \;\&\; \epsilon_{2t} \sim N(0, {\sigma_2}^2)$$

- When both time series are present, we suppose that the colored noises are correlated together:

$$N_{1t} = \alpha N_{2t} + \epsilon_{3t}$$

with $\alpha = \mathrm{corr}\,(N_1, N_2)$ and ϵ_{3t} a white noise

The following transformation is applied:

$$y_{3t}^* = y_{1t} - \alpha.y_{2t}$$

$y^*_{3t} = \mu\,(1-\alpha) + \omega.(1-\alpha).t - \alpha.\delta + \epsilon_{3t}$, t = $T_0..n_1$, $\epsilon_{3t} \sim N(0, {\sigma_3}^2)$

with:

$$\sigma_3^2 = \frac{\sigma_{1.}^2}{(1-{\phi_1}^2)} + \alpha^2.\frac{\sigma_{2.}^2}{(1-{\phi_2}^2)} - 2\alpha^2.\frac{\sigma_1.\sigma_2}{\sqrt{(1-{\phi_1}^2)}.\sqrt{(1-{\phi_2}^2)}}$$

Model parameters estimation

The transformed equation can be expressed using the matrix form:

$$Y^* = X^* A + \epsilon$$

Where X* is either a Txdim(A) matrix (Tx3 if A={μ, δ, ω}) or a Tx11 matrix when considering a seasonal signal S(t) for A. Y* is a Tx1 matrix and γ a TxT diagonal covariance matrix of the residuals. In the simpliest case X^*, Y^* and γ resort to:

$$X^* = \begin{bmatrix} 1-\phi_1 & 0 & t_{2..T_0}(1-\phi_1) \\ 1-\alpha & -\alpha & t_{T_0..n_1}(1-\alpha) \\ 1-\phi_2 & 1-\phi_2 & t_{T_0+1..n_2}(1-\phi_2) \end{bmatrix}$$

$$Y^* = \begin{bmatrix} y^*_{1t} \\ y^*_{3t} \\ y^*_{2t} \end{bmatrix} \quad \gamma = \begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_3^2 & 0 \\ 0 & 0 & \sigma_2^2 \end{bmatrix}$$

The GLS estimator of A resorts to:

$$\hat{A} = (X'.\gamma^{-1}.X)^{-1}.X'.\gamma^{-1}.Y$$

X*' stands for the transpose of X*.

In practice the equation must be solved using an iterative process. First the values of $\hat{\phi}_1$, $\hat{\phi}_2$, $\hat{\sigma}_1$, $\hat{\sigma}_2$, $\hat{\sigma}_3$, and α must be evaluated from the data. Then $\hat{A}$ is estimated. The values of $\hat{\phi}_1$, $\hat{\phi}_2$, $\hat{\sigma}_1$, $\hat{\sigma}_2$, $\hat{\sigma}_3$, and $\hat{\alpha}$ are then revaluated. The iterative procedure is iterated until convergence. The estimated covariance matrix of $\hat{A}$ is obtained using:

$$cov(\hat{A}) = (X^{*'}.\gamma^{-1}.X^*)^{-1} = \begin{bmatrix} \sigma_\mu^2 & \sigma_{\mu\delta} & \sigma_{\mu\omega} \\ \sigma_{\mu\delta} & \sigma_\delta^2 & \sigma_{\delta\omega} \\ \sigma_{\mu\omega} & \sigma_{\delta\omega} & \sigma_\omega^2 \end{bmatrix} \quad (14)$$

for A={μ, δ, ω}

Estimation of the uncertainty with prior knowledge on the level shift level

This case is very particular and may appear using in-situ datasets. if the uncertainty σ_0 of the level shift between two time series may be estimated from external sources (independent cross calibration of sensors, theoretical model), the covariance matrix of the estimate $\hat{A}$ is given by:

$cov(\hat{A})=(X^{*'}.\gamma^{-1}.X^{*'}+\sigma^2.\gamma_2)^{-1}$

With $\gamma_2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \frac{1}{\sigma 0} & 0 \\ 0 & 0 & 0 \end{bmatrix}$

where σ^2 is thus the weighted average of the noise variance of the two time series :

$$\sigma^2 = \frac{\sum_{t=1}^{T1}(t - t_{median})^2 {\sigma_1}^2 + \sum_{t=T0}^{T}(t - t_{median})^2 {\sigma_2}^2}{\sum_{t=1}^{T1}(t - t_{median})^2 + \sum_{t=T0}^{T}(t - t_{median})^2}$$

With $t_{median} = median(t_1, t_2)$ and σ_1^2 and σ_2^2 the white noise variance of the two time series.

2.4 Detecting significant trends

The detectability of a trend or its significance may be treated from different but coincident points of views. It generally relies on the estimation of the standard deviation of the trend estimate (or the associated interval of confidence), and less usually on the number of observations required to detect a trend among a noise with a given variance. Formally, the statistical assessment of the significance of a trend in a time series of length n resorts to testing either the variable $|\hat{\omega}|/\sigma_{\hat{\omega}}$ or the variable $\frac{\hat{r}\sqrt{n-2}}{\sqrt{1-\hat{r}^2}}$, with r, the coefficient of correlation between the time series and the trend. Both tests are similar [Scherrer, 77] and both variables theoretically follow a student's T-distribution with n-2 degrees of freedom [Haan, 78; Legendre & Legendre, 79; Scherrer, 77]. Under the considered red noise model assumption, the 90% confidence level is reached for $|\hat{\omega}|/\sigma_{\hat{\omega}} > 1.64$ and the 95 % confidence level for $|\hat{\omega}|/\sigma_{\hat{\omega}} > 1.96$. In the subsequent, we consider a 95% confidence level, such that we test for $|\hat{\omega}|/\sigma_{\hat{\omega}} > 1.96$.

2.5 Application to the two-sensor SeaWiFS-MERIS dataset.

2.5.1 The dataset

Tiao *et al.* [2] showed that the existence of a moderate positive value of Φ in the daily measurements is enough to make the trend estimate insensitive to changes in the temporal

sampling. Compared to the daily data, the monthly averaged data will lower the length of the time series and the autocorrelation leading to similar trend detection. It implies that geophysical datasets, associated with high autocorrelation levels, may be analyzed using the monthly time series. Two datasets are used here, the global 1998-2010 SeaWiFS monthly chl-a products estimated using the OC4 algorithm [O'Reilly *et al.*, 64 & 80], and the global 2003-2011 MERIS chl-a monthly estimated using the MERIS OC4 algorithm [Morel *et al.*, 81]. Data were projected on a regular 1x1°grid and time series with more than 30% of missing data were withdrawn from the analysis, leaving 31829 for both datasets. For each location of the 1x1° grid, a climatology estimated from the available observations has been subtracted to the original time series to remove the seasonal signal, S_t. Neither spatial nor temporal interpolations were performed on the dataset. The value of T_0 used in Eq.(13), i.e. the starting time of the MERIS time series, is equal to 60 months.

2.5.2 Single-sensor linear trend detection using the SeaWiFS dataset

Figure 1 shows for the period 1998-2010 the estimated model parameters for the single-sensor model Eq.(6) using the SeaWiFS monthly chl-a dataset, namely the long term trend $\hat{\omega}$, the noise autocorrelation $\hat{\phi}$ and the white noise variance $\hat{\sigma}^2$. Overall we detect significant linear trends for 41 % of the 31829 time series (Figure 3). There are several coherent patches with significant trends. The typical magnitude of trends in the chl-a is ~ ± 0.003 $mg.m^{-3}$ $.year^{-1}$, with positive peak values of +0.009 $mg.m^{-3}.year^{-1}$ at the Eastern part of the Argentina, the South of Australia, the Behring Sea and specific coastal areas. Negative peak values of -0.009 $mg.m^{-3}.year^{-1}$ are reached in the North Atlantic & the Arabian Sea. We observe in the inter-tropical region a majority of negative trends in the chl-a concentration with a mean value of -0.002 $mg.m^{-3}.year^{-1}$.

Compared to previous trend estimations performed on the SeaWiFS dataset, Gregg [70] observed globally comparable trends on the period 1998-2003 with the exception of the eastern part of Africa that do not show anymore a positive trend. This difference may be explained either by the absence of a global linear trend for the entire 1998-2010 period, or by the ignorance of noise autocorrelation by Gregg *et al.* [70]. Henson *et al.* [36] showed a similar global distribution of the trend estimates for this period using the same single-sensor model, Eq.(6), and the 1998-2007 SeaWiFS dataset, with nevertheless less positive trends East of South Argentina. This suggests that the data from 2008 to 2010 contributed significantly to the detection of significant positive trends of chl-a in this area.

Concerning the noise autocorrelation $\hat{\phi}$, the mean observed value over the globe is 0.3 (Figure 3b). Minimum values of -0.2 are observed locally in the Southern part and specific coastal areas. Maximum $\hat{\phi}$ values of 0.75 are observed at 30°S in the Indian Ocean and the East Chile. Globally, the noise autocorrelation is greater in the tropical region with a mean value of 0.35 between 30°S and 30°N, compared to a mean value of 0.25 for latitudes South of 60°S or North of 60°N.

The estimated variance of the residuals (Figure 3c) shows latitudinal and coastward distribution with greater values observed at high latitudes and along the shores, and appears correlated to the mean values of the chl-a distribution [Blunden, 82]. In the inter-tropical zone, y_t variance is lower than the one observed at high latitudes and the variability is led by non-seasonal signals leading to a large correlation in the residuals (large values of $\hat{\phi}$).

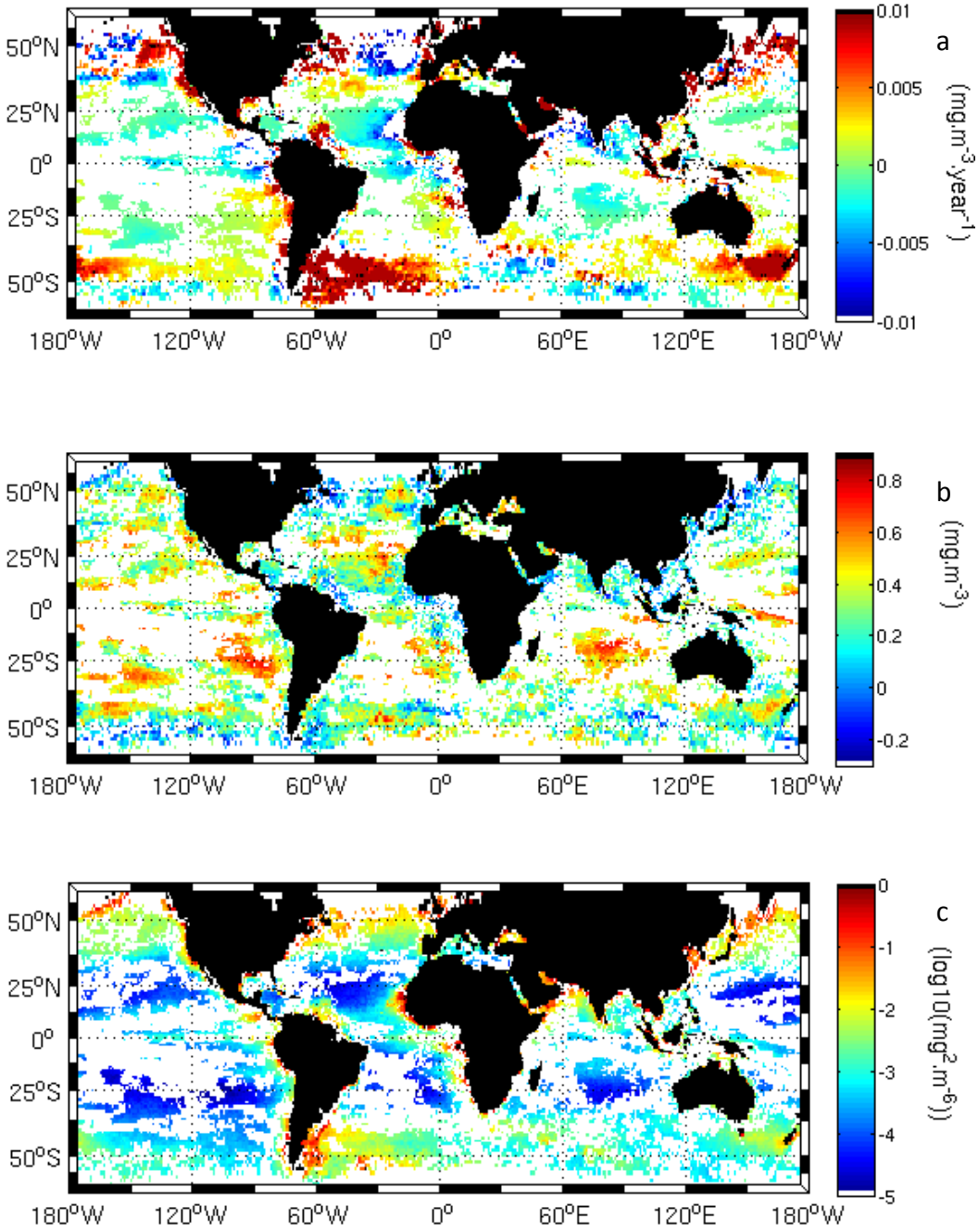


Figure 3: Estimated parameters for the single-sensor model, Eq.(6), using the SeaWiFS monthly data (1998-2010). (a) Significant linear trends, $\hat{\omega}$, with respect to a 95% confidence level. (b) noise auto-correlation $\hat{\phi}$. (c) noise variance $\hat{\sigma}^2$.

2.5.3 Single-sensor linear trend detection using the MERIS dataset

We also report the same analysis as above for the 2003-2011 MERIS dataset (Figure 5). Although the MERIS dataset is only a 10-year time series compared to the 13 years of data available for the SeaWiFS dataset, we detect significant linear trends for 50% of the 2003-2011 MERIS time series (resp. 41% for the SeaWiFS data). The Equatorial Pacific shows a linear decrease of -0.002 $mg.m^{-3}.year^{-1}$ surrounded by a large belt of positive trends, with a mean value of 0.006 $mg.m^{-3}.year^{-1}$, starting from the East Papua New Guinea and ending in the North in the Mexico and in the South in the North of Chile leading to a wishbone shape of positive trends in the Equatorial Pacific. This region is well known to be strongly influenced by the ENSO signal (http://www.srh.weather.gov/srh/jetstream/tropics/enso.htm). In this region, the difference in terms of detection between SeaWiFS and MERIS dataset is clearly visible. A major El-Niño event occurred during the 1997-1998 followed by a ENSO–Niña period during 1998-2000. Phytoplankton productivity relies on the availability of sunlight, macronutrients (e.g., nitrogen, phosphorous), and micronutrients (e.g., iron), and thus is sensitive to climate-driven changes in the delivery of these resources to the euphotic zone. Turk *et al.* [83] showed that the ENSO oscillation strongly impact the chl-a and the primary production in the Equatorial Pacific and one can clearly see Figure 3a that using the SeaWiFS 1998-2010 dataset a limited number of significant trends are detected in this area compared to the MERIS dataset over the period 2003-2011: non-stationary processes such as El-Niño-La-Niña tend to reduce the ability to detect a trend. The problem of estimating and removing ENSO-related variations from climate records has been addressed in many previous studies for the SST using a variety of methods. In this spirit, Compo *et al.* [84] used the ENSO pattern filter, EPF [Alexander *et al.,* 85] developed to remove the contribution of ENSO patterns in the Sea Surface Temperature (SST) from 1871–2006. Satellite-derived ocean color time series are nevertheless shorter and show a greater intra-seasonal variability. This alters the ability of filtering the long-term variability caused by such signals. This type of filtering has not yet been implemented for the chl-a time series and its evaluation is in any case beyond the scope of this analysis, as we do not discuss of the quality of the input data.

The estimated noise autocorrelation shows a similar geographical distribution as observed for the SeaWiFS dataset with nevertheless a large band of high auto-correlated noise in the South Pacific. The residual variance is distributed similarly to the one estimated using SeaWiFS with nevertheless a Northward extension of the detected trends.

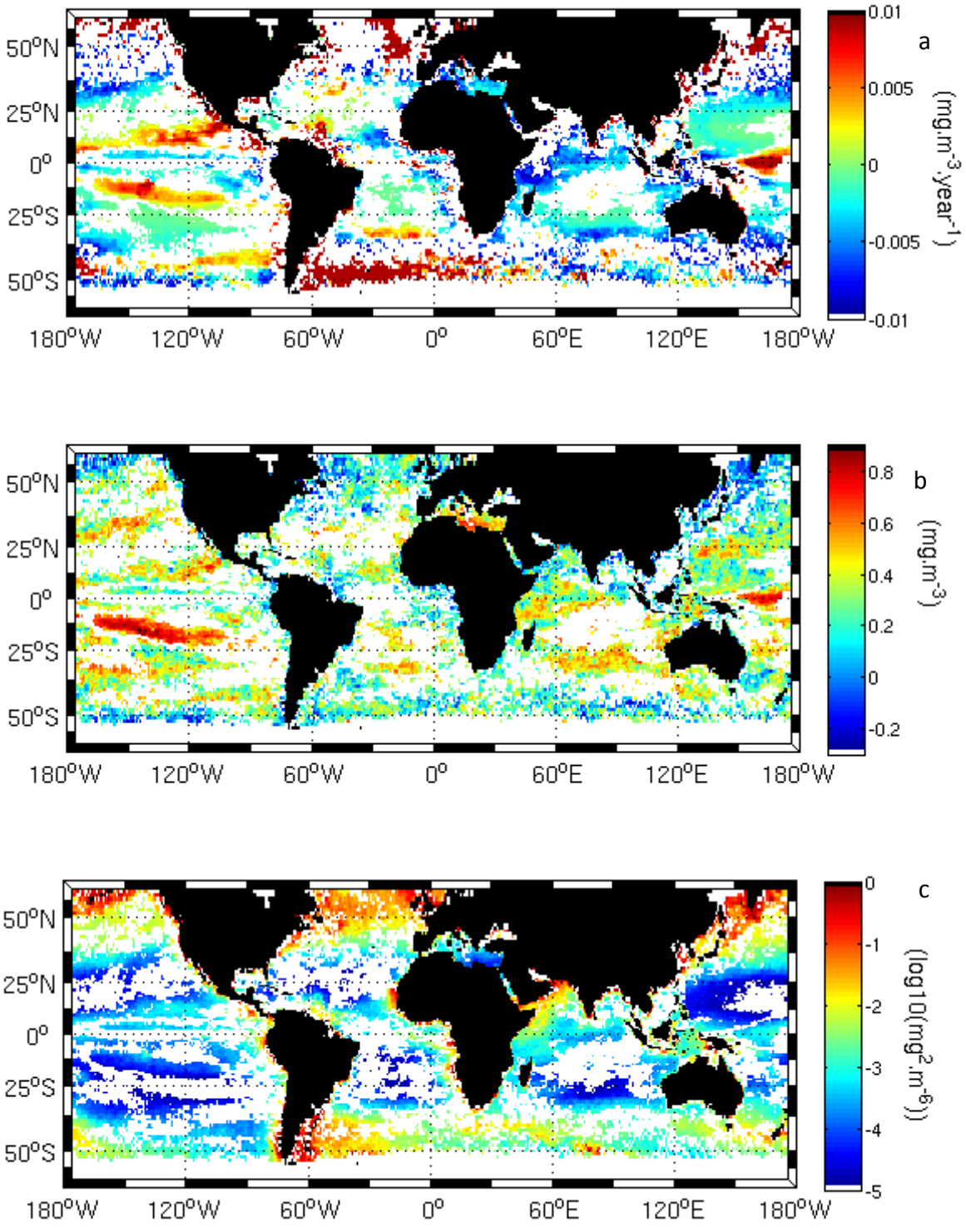


Figure 4: Estimated parameters for the single-sensor model, Eq.(6), using the MERIS dataset (2003-2011). (a) Significant linear trends, $\hat{\omega}$, with respect to a 95% confidence level. (b) noise auto-correlation $\hat{\phi}$. (c) noise variance $\hat{\sigma}^2$.

2.5.4 Two-sensor linear trend detection using both MERIS and SeaWiFS data

Using the two-sensor model, Eq. (13), the joint analysis (Figure 5) of the MERIS and SeaWiFS time series leads to 60% of significant detections of linear trends for the period 1998-2011 (resp. 50% and 41% for MERIS and SeaWiFS data alone). It resorts to much clearer patterns at a global scale for the period 1998-2011 compared to the period 1998-2010 & 2003-2011 considered individually. In Table 2 we summarize at ocean-scale trend estimated statistics. At global scale, the observed median value in the significant trends is 2.83 x 10-4 $mg.m^{-3}.year^{-1}$. This value is low and opposite to the estimated trend by Boyce *et. al.* [86] for the 20^{th} century using this time in-situ data. Indian Ocean shows the largest median value with a decreasing value of -1.40 x 10-3 $mg.m^{-3}.year^{-1}$ while the Pacific and the Atlantic show similar median positive trends of respectively of 7.27 x 10-4 and 8.27 x 10-4 $mg.m^{-3}.year^{-1}$. Regarding coastal areas, we detect positive trends especially in the Bering Sea, the Pacific shores of the United States, the Patagonian Shelf (Figure 5a). Regarding the open ocean, Southern regions show as observed in Figure 3a & Figure 4a, a majority of positive trends with local maximum at +0.009 $mg.m^{-3}.year^{-1}$ in the Eastern part of South Argentina and the South East part of Australia. Although we do not discuss here of the quality of the dataset, we underline nevertheless that both algorithms, SeaWiFS OC4 and MERIS OC4, are calibrated for open ocean waters where the observed radiance is constrained by the water and the chl-a absorption properties. In coastal areas and specific areas, the effect of the suspended matters and the colored dissolved organic matters may alter the observed radiances leading to positive biases in the estimated chl-a retrieval using the OC4 algorithms and possibly affecting the trend estimation in such areas. A full discussion on the estimation of optical properties in coastal areas is available in [IOCCG, 17].

The 'wishbone' pattern in the Equatorial Pacific clearly appears Figure 5b with this time some extensions of the positive trends from the Florida to the Mediterranean Sea and from Brazil to South Africa. This Atlantic extension of this structure was not visible using the SeaWifs (Figure 5b) dataset and only partially visible using the MERIS dataset.

Some coastal regions depict negative trends with a minimum of -0.009 $mg.m^{-3}.year^{-1}$ in the equatorial area, the North Atlantic and the North Pacific. From the joint analysis, the Indian gyre and more generally the Indian Ocean shows a global negative trend except for its Southern area. The decline in the global gyres in the productivity, directly linked to the chl-a, was also observed by Polovina *et al.* [87].

In the Atlantic, the inter-tropical zone shows a low decrease of -0.002 $mg.m^{-3}.year^{-1}$. In the South Atlantic and below 40°S, the trend increases positively. Regarding the North Atlantic, we detect an increase of the chl-a in the North Atlantic Current & the Gulf Stream, and Northwards the Atlantic Western part shows a positive trend conversely to the Eastern part.

The estimated shift between the two chl-a datasets is reported Figure 5b. Its magnitude, conversely to its uncertainty, does not affect the trend estimation. Maximum positive shift values

are observed in the North Atlantic with local values of 0.08 $mg.m^{-3}.year^{-1}$ of positive shift for the MERIS OC4 compared to the SeaWiFS OC4 chl-a. Conversely, negative maximum values are observed in the Tasman Sea. The large shift values observed at high latitudes might be related to local differences in the atmospheric corrections used for each sensor [IOCCG, 88].

The estimated variance of the residuals (Figure 5c) shows greater values for the high latitudes and on the shores directly correlated to the mean values of the chl-a distribution [Blunden *et. al.*, 82] as observed Figure 3c & Figure 4c.

Table 3: Statistics at large scale on the estimated significant trends in the chl-a ($mg.m^{-3}.year^{-1}$) over the period 1998-2011.

	Median($\widehat{\omega}$)	**Min**($\widehat{\omega}$)	**Max**($\widehat{\omega}$)	$\boldsymbol{\sigma}_{\widehat{\omega}}$
Global	2.83×10^{-4}	1.59×10^{-1}	1.0×10^{-2}	3.20×10^{-3}
Atlantic	8.27×10^{-4}	-1.59×10^{-1}	1.0×10^{-2}	4.60×10^{-3}
Pacific	7.27×10^{-4}	-7.49×10^{-2}	1.0×10^{-2}	2.80×10^{-3}
Indian Ocean	-1.40×10^{-3}	-1.14×10^{-1}	9.60×10^{-3}	2.00×10^{-3}

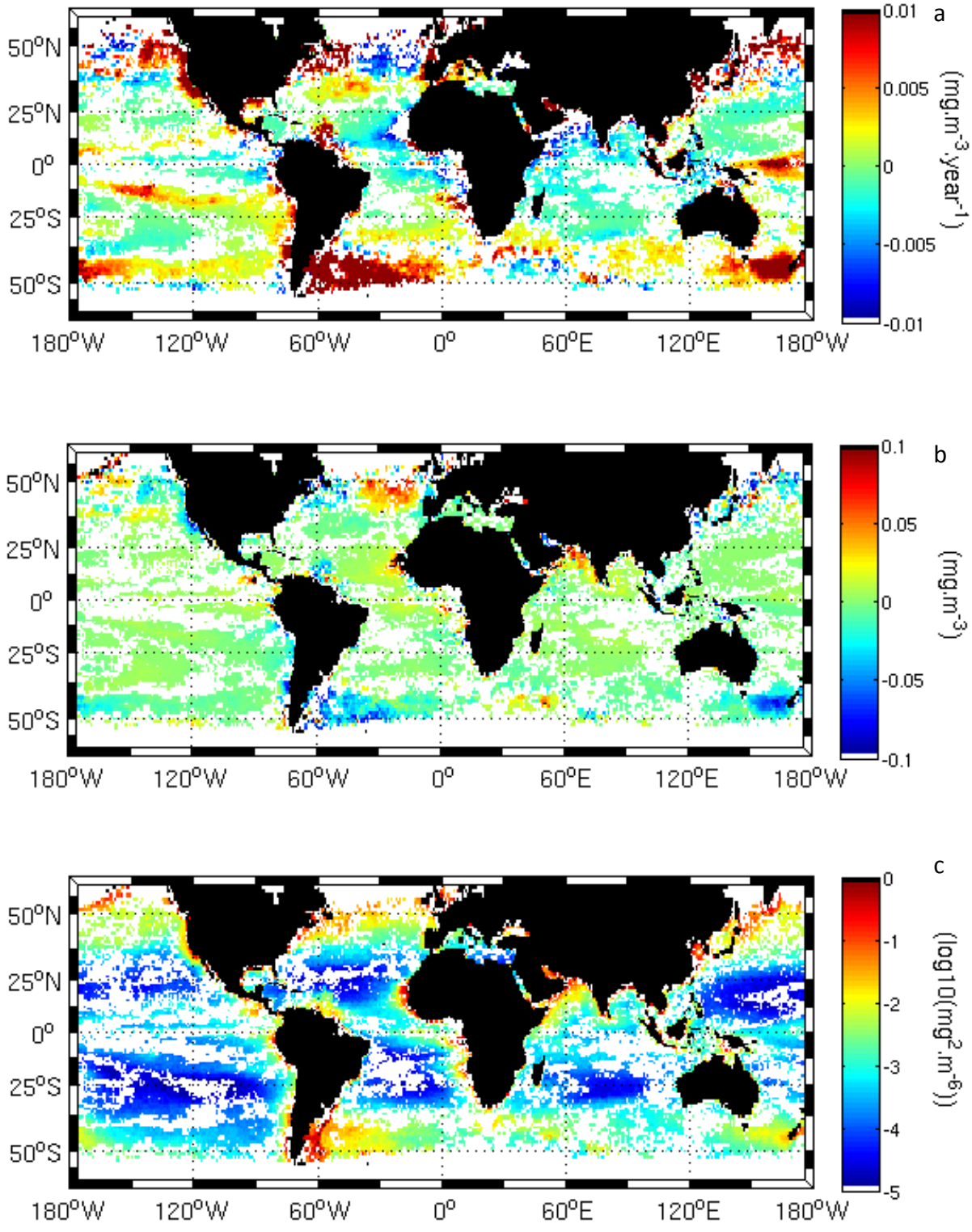


Figure 5: Estimated parameters for the multi-sensor model, Eq.(13), using the SeaWiFS and the MERIS dataset (1998-2011). (a) Significant linear trends, $\hat{\omega}$, with respect to a 95% confidence level. (b) level shift $\hat{\delta}$. (c) noise variance $\hat{\sigma}^2$.

2.6 Optimization of a time-overlap between successive missions for long term monitoring & impact of the incoming ESA Sentinel 3 – OLCI mission.

The proposed multi-sensor model, Eq.(13), provides the basis for investigating the extent to which the time overlap between successive missions may be optimized to reduce the uncertainty on the long-term detection of linear trends in geophysical time series.

From Eq.(12), the uncertainty of the trend estimation σ_ω of Eq.(14), could also be expressed as an function of the model parameters $\{n, \phi, DT, \alpha\}$. Nevertheless due to the complexity of its derivation we will consider the matrix form of $\sigma_{\hat{A}}$. σ_ω Eq.(14) may be expressed as:

$$\sigma_\omega = \sigma \,.\, G(n, \phi, DT, \alpha) \qquad (15)$$

where G is the trend uncertainty which, in case of the use of 2 time series is a function of n, φ, DT, n the total number of non-redundant months between two time series, φ the observed autocorrelation (we supposed here $\varphi_1 = \varphi_2 = \varphi$). DT is the starting time of the second time series, given the first time series is assumed to start at time t=0. Depending on parameter DT, we cover both time overlap between the two series (DT< n_1, the length of the first time series) as well as time gaps (DT> n_1). The parameter α is the correlation coefficient between the two white noise processes and σ² the weighted variance expressed as a function of the two white noise variances σ_1^2 and σ_2^2.

Given two time series of 60 months, we report the coefficient G values for the trend estimate as a function of parameters φ, n and DT (Figure 6 & Figure 7). Parameter α was set to 0.7, the mean correlation value observed between MERIS and SeaWiFS. The uncertainty coefficient G (and consequently σ_ω) increases with φ (Figure 6). When an overlap is present (DT < 60 months), G decreases with the time overlap until it reaches a minimum value which depends on the autocorrelation value. This minimum corresponds to the optimal value of the time overlap between two time series to optimize the balance between the uncertainty on the shift parameter δ and the length of the two time series. For φ = 0.3, i.e. the mean value observed for SeaWiFS (Figure 3b), the minimum is reached for 12 months of time overlap. When no overlap is present and the time gap increases, the uncertainty on the trend remains constant as the estimation of the trend resorts to analyzing independent time series only sharing a common trend, such that the overall uncertainty only depends on the sum of the lengths of the two series, here 120 months (Figure 6).

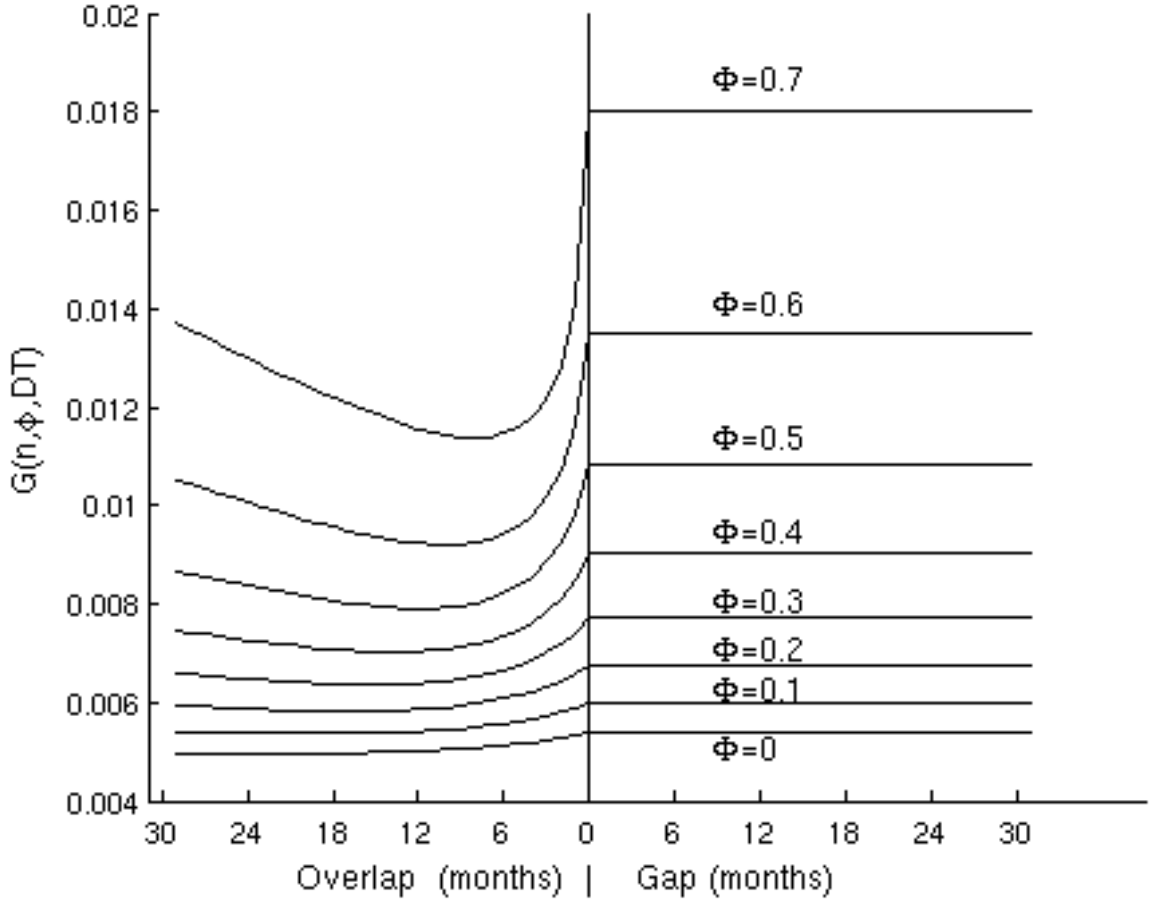


Figure 6 : Effect of the time overlap or the gap-time (in months) between two time series of 60 months on the trend uncertainty coefficient G, Eq.(15).

To illustrate how to use Figure 6, we simulate the detection of a ω value of 0.01/12 mg.m^{-3}.month^{-1} within the two time series of 60 months with a φ value equal to 0.6 and a σ value equal to 0.03. Considering an overlap of one year, the detection value, $|\omega| / \sigma_{\omega} = |\omega| / (G.\sigma) = (0.01/12)/(0.03*0.0095) = 2.19$, i.e. the 95 % level of confidence is reached. Conversely, for the one year gap situation, $|\omega| / \sigma_{\omega} = (0.01/12)/(0.05*0.013)=1.28$, i.e. this trend would not be detected if analyzed with the same number of monthly observations but with disjoint time series.

We also depict the evolution of uncertainty G as a function of the length of the second time series, with a given length of the first time series set to 60 months. We test for two different situations: a one year time overlap (Figure 7a) with α value set to 0.7, and a one year gap (Figure 7b). In both cases, uncertainty G increases with φ and decreases with the length of the second time series. For a one-year overlap and for a moderately–high value of φ of 0.6, a typical value observed in Figure 3b & Figure 4b, G values are respectively of 0.015 and 0.0125 after a duration of 12 and 36 months

for the second time series, i.e., σ_ω has decreased of 16%. For a one year gap, for the same value of φ, G values are respectively of 0.019 and 0.018, i.e., σ_w has decreased of 5% in two years and is 26% greater at 12 months and 44% at 36 months compared to the overlap situation.

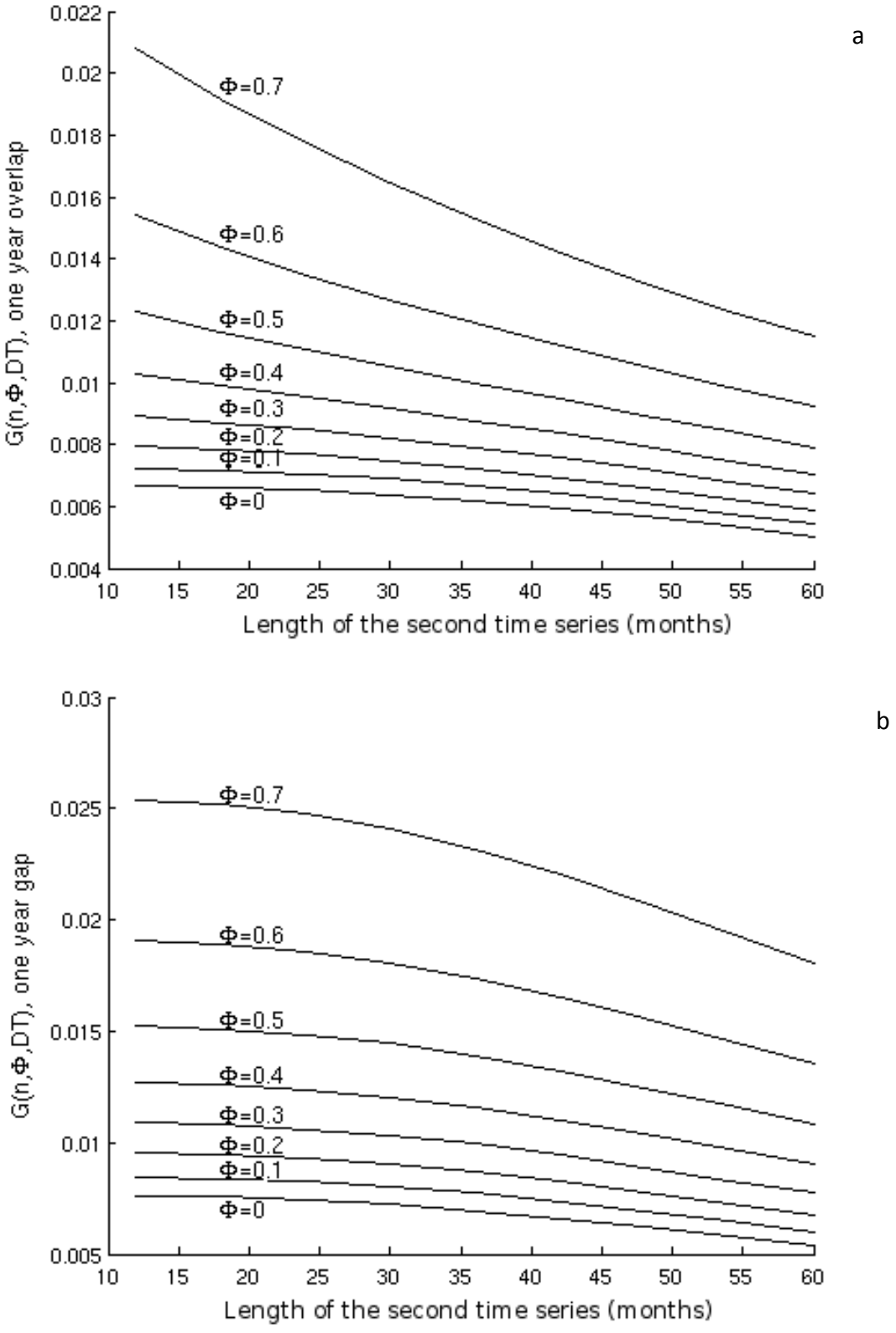


Figure 7: Effect of the length of the second time series on the uncertainty trend coefficient G, Eq.(15) with (a) a one year overlap and (b) a one year gap.

The impact of new available space-based observations such as provided by the incoming S3 satellite with onboard the OLCI sensor may also be evaluated using the proposed model, Eq.(15). This satellite should be launched at the end of 2014 and consequently no overlap will be observed with the MERIS, SeaWiFS and MODIS time series (MODIS-AQUA, launched in 2002, mission's initial lifetime was planned to be about 6 years). To evaluate the added value of the S3 mission regarding the long-term trend detection, we consider here that the uncertainty on the level shift between OLCI & MERIS OC4 derived chl-a will be of the same magnitude than the one estimated between MERIS OC4 and SeaWiFS OC4 (not shown). Given this assumption, we proceed as previously to determine the variance of the trend estimate.

We proceed as follows to derive a global map (Figure 5). For locations such that $|\widehat{\omega}|/\sigma_{\widehat{\omega}} > 0.5$, i.e. a 70 % significance level, we assume that the trend estimate $\widehat{\omega}$ might be relevant but was not detected as significant due to a too low number of MERIS-SeaWiFS observations compared to observed local noise level. From simulations, we determine the required duration of the OLCI time series to reach a 95% significance level, i.e. $|\widehat{\omega}|/\sigma_{\widehat{\omega}} > 1.96$. For locations with significant SeaWiFS-MERIS linear trend estimate with a 95% significance level, we determine from simulations the required duration of the OLCI time series to reduce the uncertainty $\sigma_{\widehat{\omega}}$. For these numerical derivations, we also assume in Eq.(15) that $\sigma_1^2 = \sigma_2^2$, i.e. the white noise variance measured from OLCI will be equal to the white noise variance estimated from the SeaWiFS - MERIS dataset (Figure 5c). The starting time of the OLCI time series is the first January 2015 leading to a T_0 value of 36 months (MERIS time series ends here in December 2011). Overall, the reported results show that a mean duration of 53 months of S3-OLCI observations will be necessary to actually enhance the detection of significant linear trends issued from the joint SeaWiFS-MERIS analysis (Figure 5). Interestingly, results are spatially homogeneous with local variability related to region-specific noise characteristics. In the South Pacific, the West of Senegal, and the Arabian Sea, a minimum of 40 months of S3-OLCI is needed to enhance the detection. In these areas a trend is nearly detected. In the South part of, America, South Africa and the South of Australia, high estimated values of $\widehat{\sigma}$ lead to an increase of $\sigma_{\widehat{\omega}}$ and a longer duration of S3-OLCI observations (typically about 68 months) will be necessary. In the Equatorial Pacific, the variance of the noise is low, but the significant estimated trends are very weak increasing the time of S3-OLCI observations necessary to actually improve the detection of significant long-term trends.

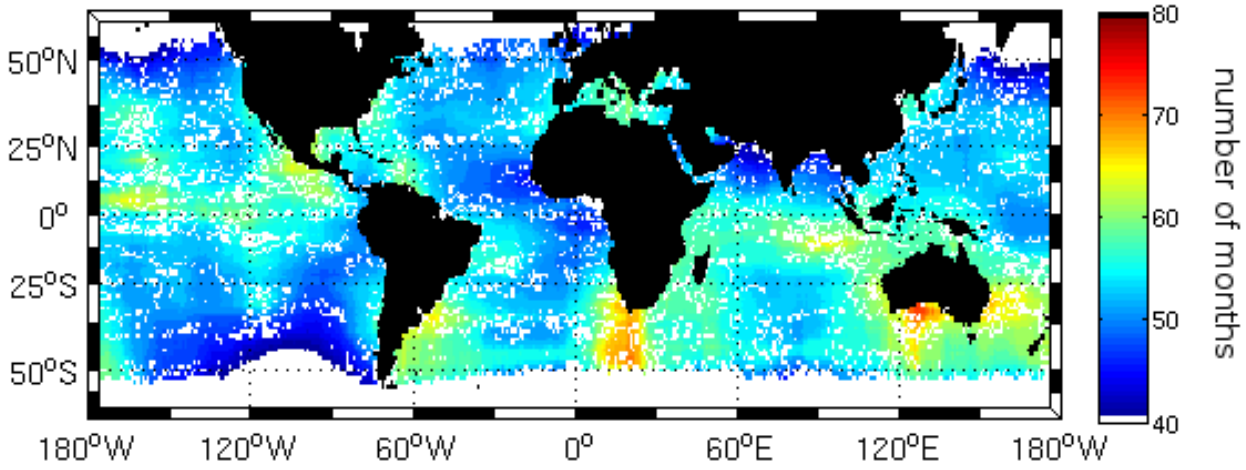


Figure 8: Estimated duration of needed Sentinel 3 - OLCI month measurements to enhance the joint SeaWiFS - MERIS detection of long-term linear trend: from simulations of model (Eq.(15), see text for details).

2.7 Conclusions

The two major statistical factors governing a trend estimation and detection in a single-sensor time series are the autocorrelation and the variance of the noise. The estimated noise autocorrelation showed latitudinal distribution with a mean value of 0.35 in equatorial zones compared to 0.25 at higher latitudes. This difference leads to an increase of 16% of the uncertainty on the estimation of the same trend in these two different areas. When two time series are available, the trend detection depends on the uncertainty on the level shift between the datasets. In case of an overlap, the shift uncertainty is diminished. The use of the joint chl-a SeaWiFS-MERIS dataset over the period 1998-2011 led to the detection of 60% of significant trends, compared to 41 % for the SeaWiFS dataset only and 50% for the MERIS dataset only, contributing to a better characterization of region-specific patterns in the detected trends.

It might be noted that in situ data generally involve greater variance levels which may not. Optimizing an observation network for the long term monitoring implies to minimize the effect of the unknown level shift by organizing time overlaps between successive missions. From our analysis and for a noise autocorrelation level greater than 0.3 as observed in average for our dataset, an overlap of 12 months has been found to be optimal to lower the uncertainty on the level shift and to minimize the uncertainty on the trend estimate within two time series of 60 months.

In case the time series present no time overlap, the estimation of a potential level shift and its uncertainty is needed. This can be derived from inter-calibration analyses based on the physical characteristics of the sensor measurements, as well as from inter-calibration based on comparison with consistent long-term field observations [Antoine *et al.*, 89; Clark *et al.*, 90]. This aspect that should be addressed in future woks, is crucial for a meaningful merging of the incoming Sentinel 3 – OLCI time series with previous ocean color missions. Savings, in terms of necessary duration of Sentinel 3 – OLCI observations and resulting costs is grandly constrained by this issue. In this respect, we estimated the minimal region-dependent duration of the Sentinel 3 - OLCI mission necessary to improve the detection of long-term linear trends issued from the SeaWiFS-MERIS dataset. We estimated a mean value of 53 months for the needed Sentinel 3 – OLCI observations, with some region-dependent fluctuations between 40 to 68 months. This simulation was carried out using an uncertainty level on the shift between OLCI and MERIS of the same magnitude than the one estimated between SeaWiFS and MERIS. These results are coherent with the expected lifetime of the Sentinel 3-OLCI mission, and suggest that the analysis of the global long-term patterns should actually benefit from the joint analysis of SeaWiFS, MERIS and Sentinel 3-OLCI datasets.

In the future, the methodology will be applied to other ocean-color variables such as the vertical attenuation of the light [Morel *et al.,* 81; Saulquin *et al*., 91]. Its application to in situ data might also be considered, for instance for validation purposes. Given the noise parameters of the considered remotely-sensed data, we were able to detect relatively weak trends, typically between +/- 0.01 $mg.m^{-3}$ per decade and +/- 0.1 $mg.m^{-3}$ per decade. In-situ measurements may involve greater variance levels caused by support effects, i.e. the local variability in the chl-a, caused by fine-scale structures such as filaments which are averaged using the 1°x1° satellite data [Saulquin *et al*., 92]. Such local variability especially at the shore might occlude such trends. The design of specific in-situ setting (e.g., sensor networks) might require to reduce these variances to detect linear trend levels similar to those issued from remotely sensed data. Regarding methodological aspects, refined models of level shift between time series (magnitude dependent models) and the effect of outliers in the estimation of the autoregressive parameters [Sarnaglia, 93] should also be evaluated. The influence of low frequency climatic signal such as ENSO on the ocean-color dataset should also be considered with care. In this respect, the development of specific filtering procedures to remove such contributions could be investigated.

3 Chapter III: Multi-scale event-based mining in geophysical time series: characterization and distribution of significant time-scales in the Sea Surface Temperature anomalies relative to ENSO periods from 1985 to 2009.

This chapter is motivated by state of the art drawbacks of some of the most popular climatic and oceanographic analysis methods. Since the 70's, EOF[22] [9] or SVD[23] [5, 8] have been largely used to extract main spatio-temporal covariance patterns within a single spatio-temporal dataset (EOF) or a multivariate dataset (SVD). These methods are convenient, but, by construction, highlight the low frequency modes of the covariance. Consequently, the impact of major climatic phenomenon, such as the El Niño Southern Oscillation (ENSO [8, 59]), is described to occur mainly in this domain of frequencies. It is for example common [59, 47] to study the ENSO signal using a low passband filter in the 1.5-7 years range, which finally limits the impact of the ENSO to such range. Nevertheless, the ENSO has some high frequency components or indirect contributions on the SST at these frequencies [58]. Within the 1.5-7 years range, specific signatures may also be discretized.

From a methodological point of view this chapter addresses firstly, in keeping with the first chapter, estimation of significant parameters in geophysical time series. In this case, an ellipsoid-designed event is detected in the time-frequency representation of the wavelet spectrum of single a SST time series, relative to the local conditions of noise. The second methodological aspect addresses estimation and characterization of Gaussian modes, referring to the reference time-scales, in the distribution of the detected events.

This chapter was published in 2014 in the IEEE 'Journal of Selected Topics in Applied Earth Observations and Remote Sensing (JSTARS)' [10].

3.1 Introduction

Many information sources, including instrumental in-situ data records and satellite observations, highlight the great variability and the non-stationarity of the earth's climate over a wide range of time-scales from months to decades. The most widely used technique to investigate the spatio-temporal variability of climate-relevant time series, such as temperature [94], wind [82], relies on the empirical orthogonal functions (EOF) [9], also referred to as principle component analysis (PCA) in the literature. This method combines the extraction of the main deformation modes of the covariance (correlation) of a univariate or bivariate (Singular Value Decomposition, SVD)

[22] EOF : Empirical Orthogonal Functions
[23] SVD : Singular Value Decompostion

dataset, and the analysis of the time correlation of these principal modes with potential causing factors. An introduction to univariate and multivariate EOF analysis may be found in [95] and a thorough review of advanced EOF-based methods along with inter-comparisons is presented by Benestad *et al.* [96]. These EOF-based analyses however suffer from intrinsic limitations, the main one being the assumption that the considered processes are stationary. Nevertheless, geophysical dynamics widely involve non-stationary processes (e.g., emergence of extreme events including for instance large Niño events, time shifts of seasonal cycles, propagation phenomena, trends), which may hardly be characterized as stationary. Environmental data also involve strong auto-correlation level [27]. For instance, a positive anomaly in the observed wind or temperature at a given day (week) is often associated with similar conditions the following days (weeks). This natural auto-correlation is the result of short time and local events but also of large-scale signals such as for instance the well-known El-Niño/La-Niña oscillation [8,58]. Such auto-correlated level clearly affects the determination of correlation significance level used as input of the EOF [97]. Both non-stationarity and auto-correlation may affect the interpretation of the extracted principal modes. Nevertheless these aspects are often overlooked by EOF-based approaches. Besides, EOFs are also known to be prone to outliers [97] and can hardly reveal fine time-scales signatures, which typically involve greater non-stationary variabilities.

Wavelet analysis is particularly appealing to address these issues. In contrast to EOF-based approaches, wavelet analysis actually addresses the decomposition of the fluctuations exhibited by non-stationary signals. An introduction to wavelet analysis related to climate research is given by Torrence & Compo [8]. Wavelet analysis has been used to investigate global climate changes in Sea Surface Temperature (SST) [98, 99] and interactions between physical parameters such as SST and Sea Surface Height (SSH) [100]. Whereas EOF-based schemes aim at extracting the main patterns of the covariance (or correlation) structure, wavelet analysis identifies and characterizes local time-scale patterns. The correlation between two processes can also be decomposed in the time-scale domain based on the wavelet coherency spectrum [8, 101].

Here we further investigate wavelet analysis for geophysical time series to develop an event-based representation and analysis of a geophysical dataset. Formally, we regard a time series as a collection of significant elementary time-scale events, and use an unsupervised clustering method, namely a Gaussian Mixture Mode (GMM) [102], to characterize the significant time-scales of the dataset.

To illustrate our approach, we study the SST anomalies (SSTA) observed over the globe from 1985 to 2009. We firstly characterize 4 characteristic low-frequency patterns in the SSTA time-scale distribution and study their space and time distribution regarding the ENSO modes. We also focus on the high-frequency SSTA and show a strong spatial signature of ENSO. Usually, only the low frequency in the SSTA (from 1.5 to 8 years) is attributed to ENSO [58, 103] and previous works (for example Enfield [103]) use an EOF decomposition of the filtered SSTA in the 1.5-8 year range. We underlie here that ENSO events also depict high-frequency signatures in the SSTA. We do not support here a full description of the ENSO phenomenon but use knowledge on its interactions with the SSTA to illustrate the added value, compared to standard methods, of both the

decomposition of the time series using an events-based framework and the proposed data mining approach.

3.2 Event-based analysis of geophysical times series

3.2.1 Wavelet-based extraction of elementary time-scale events

Wavelet analysis aims at characterizing non-stationary signals, i.e. signals whose statistical characteristics (e.g., mean and variance) may change over time. From the decomposition of a 1d signal in the time-scale domain, significant frequencies can be detected in any given time interval. Such decompositions typically achieve a better detection and description of the characteristic time-scale variabilities of the observed phenomenon [8] and represent a real added value to unmix non-stationary scale-dependent processes compared to classical covariance-based analysis (e.g., EOF-based schemes [9], and autoregressive models [104]). Formally, the wavelet transform of a 1D signal consists in computing the complex wavelet coefficients W(s, T) as the projection of the signal on scaled and translated versions of the selected mother wavelet Ψ [105].

$$W(s, T)=(1/\sqrt{s}) \int z(t)\, \Psi^{*} \left(\frac{t-T}{s}\right) dt \quad (16)$$

where s is the time-scale, t and T time instants and Ψ* stands for the conjugate complex of the mother wavelet Ψ. Since the wavelet transform computes the similarity between the wavelets and the signal, the choice of the mother wavelet is important. SST is often modeled using harmonics [107] and Gu and Philander [106] suggest that the ENSO signal may be represented by sinusoids. This supports the choice of the Morlet wavelet, stated as a time-windowed pure harmonic component. The wavelet power spectrum of the time-scale decomposition, Eq.(16), is defined as:

$$\mathcal{P}(s, T) = |W(s, T)|^2 \quad (17)$$

Here we propose to model the studied geophysical process as a collection of significant elementary time-scale events. Formally, this amounts to viewing the spectrum as a sum of K individual events and a red noise:

$$\mathcal{P}(s, T) = \sum_{j=1}^{K} \mathcal{P}_j(s, T) + \mathcal{P}_r(s, T) \quad (18)$$

where $\mathcal{P}_j(s, T)$ is the detected event j in the wavelet power spectrum, $\mathcal{P}_r(s, T)$, a theoretical red noise Fourier power spectrum [8], whose relevance, compared to the white noise model, is acknowledged for geophysical processes (a positive anomaly in the observed temperature on a

given day is often associated with similar conditions the following days). The first-order red noise is characterized as:

$$r(t) = \alpha\, r(t-1) + \epsilon \quad (19)$$

where α is the lag-1 autocorrelation, i.e., the mean correlation between samples at the current and preceding time steps. ϵ is a white noise process with zero mean and variance σ². If α = 0, Eq.(19), it resorts to the white noise model. For a given time series, we use a robust estimation of the noise model parameters, i.e. autocorrelation coefficient α [108] and variance σ² [109]. Regarding the significance level, we follow Torrence and Compo [8] who showed that if the original signal's Fourier components are normally distributed, then the wavelet power spectrum $\mathcal{P}$ is χ^2 distributed. The associated 95% confidence level is then obtained using:

$$\mathcal{P}(s;\ \alpha) >= 0.5\ \sigma^2\ \mathcal{P}_r(s\ ;\alpha\ )\ \chi_2^2(95\%) \quad (20)$$

where σ² is the variance of the noise model (in practice we used a robust estimation of the variance of the time series) and χ_2^2 is the chi square distribution with two degrees of freedom. $\mathcal{P}_r(s;\ \alpha)$ is the theoretical Fourier power spectrum of the red noise with variance's value set to one:

$$\mathcal{P}_r(s;\alpha) = (1 - \propto^2) \,/\, (1 + \propto^2 - 2\alpha\, cos\left(\frac{2\pi}{s}\right)) \quad (21)$$

Figure 9 shows the distribution of the theoretical Fourier power spectrum, $\mathcal{P}_{r,}$ Eq.(21), for both a red (α = 0.5, red curve) and a white noise (α = 0, blue curve). Their corresponding dashed lines represent the 95 % confidence level, Eq. (20), for σ value set to 1.

As illustrated (Figure 9), if the autocorrelation of the noise is ignored in the analysis, it typically leads to an over-detection of the low-frequency component of the signal and an under-detection of its high-frequency component.

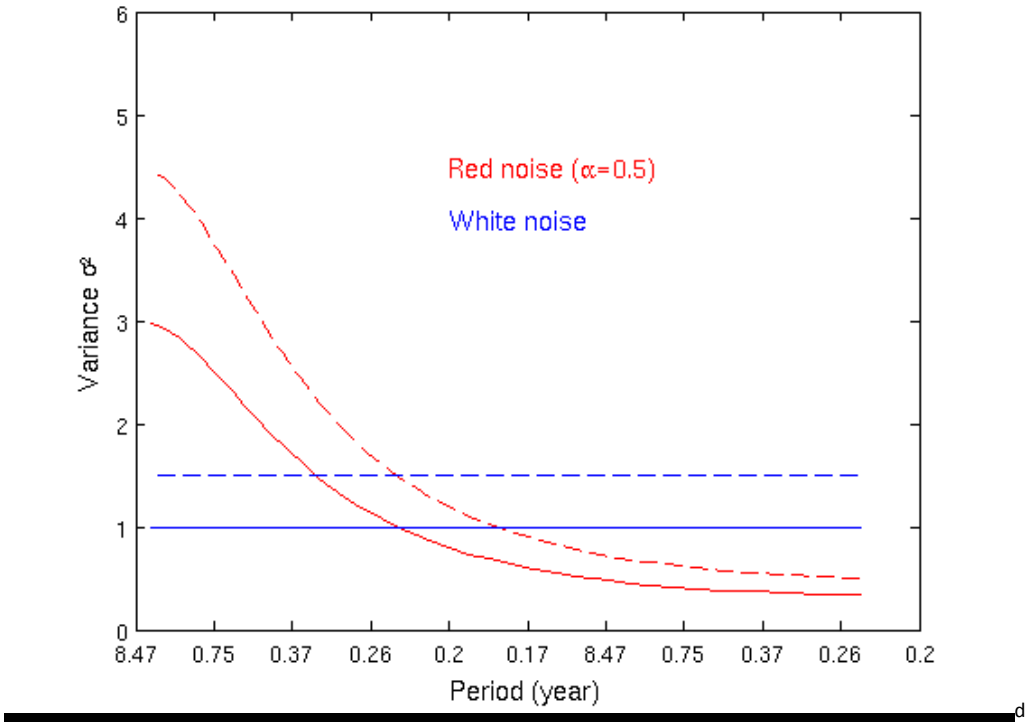


Figure 9: Theoretical Fourier spectrum, Eq. (21), as a function of the period for a white noise (blue curve) and a red noise (red curve), this latest being representative of a geophysical time series. In dashed lines the corresponding 95% confidence levels, Eq.(20)

To match the time-scale dimension of the wavelet power spectrum $\mathcal{P}(s,T)$, Eq.(17), the theoretical one dimensional Fourier power spectrum, $\mathcal{P}_r$, Eq.(21), is expanded for each time step t:

$$\mathcal{P}_r(s,t;\,\alpha) = \mathcal{P}_r(s;\,\alpha) \qquad \forall t \tag{22}$$

Our implementation of the wavelet analysis involves 69 time-scales s_j ranging from 0.2 year to 8.5 years:

$$s_j = s_0\, 2^{j+1}\, d_j\,, \quad j = 0 \ldots J \tag{23}$$

with s_0 = 1/6 year (2 months) and the time step d_j = 1/12 year (1 month) determine the smallest time-scale (2 months) and J the largest scale (8,5 years).

3.2.2 The determination of the elementary time-scale events

The elementary time-scale characterization from the wavelet spectrum $\mathcal{P}(s,T)$ [110] (see §3.3 for an example of wavelet power spectrum) involves the extraction of local regions of interest as

maximal level-sets [111, 112], i.e. areas of the wavelet spectrum which depict an energy level above the significance level [8]. In our implementation, we first detect all the significant local maxima [110] in the valid part of the wavelet spectrum, i.e. out of the cone of influence [8], and then determine their associated maximal level-set. A maximal level set is the largest time-scale area in the spectrum which contains only one maximum of energy.

Finally, an ellipse is fitted on the selected local regions of interest using the Eigen decomposition of the covariance matrix of the coordinates of elementary $\mathcal{P}(s_i,T_i)$ included in the detected event. The first eigenvector (e1) points in the direction of the greatest variance and defines the major axis for the prediction ellipse. The second eigenvector (e2) points in the direction of the minor axis. The ellipse can be represented using the parametric equation: $\text{sqrt}(\Gamma_1)\cos(t)*e1 + \text{sqrt}(\Gamma_2)\sin(t)*e2$ with $t=[0:2\pi]$ and Γ the Eigen values.

Finally, each ellipse is described by its time and time-scale extensions and the position of its center (local maximum of energy). Ellipse axes refer to the main axes of the time-scale covariance which are, even if not addressed here, directly linked to the frequency modulation observed during the propagation of the event [101].

3.2.3 Event-based mining of the event database

Our event-based mining strategy involves an unsupervised analysis of the time-scale distribution of the elementary events. This distribution is modeled using a mixture of Gaussian modes. By nature, when considering time-scale analysis, occurrences of high-frequency events are greater than those at low-frequencies. This natural distribution of the scale of the events is referred in the literature as a fractal distribution [46]. To account for this scale-dependent sampling, the mixture model involves a scale-dependent normalization factor. Formally, the considered normalized mixture model f(s) resorts to:

$$f(s) = \sum_{i=1}^{I} (\lambda_i N(s; \mu_i, \sigma_i)) / E(s) \quad (24)$$

where E(s) is the scale-related normalization accounting for the global scale-dependent sampling of the elementary events, I is the number of modes (Gaussians), λ_i the prior probability of the mode i of the mixture and N the normal probability density function (PDF) of the time-scale events for mode i:

$$N(s; \mu_i, \sigma_i) = \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma_i^2}(s-\mu_i)^2\right] \quad (25)$$

An exponential distribution for the normalization factor E(s) is proven meaningful (see § 4.4):

$$E(s) = \gamma\, e^{-\gamma s} \quad (26)$$

To infer the parameters of the mixture model we first fit the normalization factor γ and in a second step, mixture model parameters μ_i and σ_i are estimated using an EM (Expectation-Maximization) procedure [113], which aims at maximizing f(s), Eq.(23)(21), or minimizing the log likelihood:

L= -log(f(s)) (27)

For a given initialization for model parameters the EM procedure iterates expectation steps (E-step), which compute the posterior likelihoods given current model parameters, and maximization steps (M-step) to update the model parameters given the posteriors. The algorithm iterates until numerical convergence $|L(n+1) - L(n)| < 10^{-4}$. The estimation of the number of modes of the mixture model proceeds as follows: given 15 initial modes in the mixture model, only the modes with $\sigma_i > 0.05$ year are kept in the model after each EM step.

3.3 Application to the satellite-derived SSTA observed from 1985 to 2009

3.3.1 The pathfinder dataset

Satellite-derived SST data are extracted from the global AVHRR Pathfinder SST v5.2 [16] daily gridded product (http://www.nodc.noaa.gov/SatelliteData/pathfinder4km/). To avoid diurnal effect, we used the data acquired at night time. A quality control was performed by selecting pixels with a quality flag level greater than 3. This quality flag is provided in the Pathfinder v5.2 product, and its level was determined using Kilpatrick studies [114]. The estimation of the SSTA involves firstly the estimation of monthly mean SST fields at 36 km resolution. A minimum of 30 observations per grid cell is used to estimate the average. The seasonal signal (climatology) must be then removed from the SST to obtain the SSTA. The harmonic-based estimation of a climatology of is more accurate than the simple average estimation [115]. Hence, for each time series, a local climatology S_t composed of 4 harmonics and a linear trend [4] is estimated and subtracted from the SST to remove 12, 6, 4 and 3 months periodicities:

$$S_t = \sum_{i=1}^{4} a_i . \cos\left(\frac{2\pi i t}{12}\right) + b_i . \sin\left(\frac{2\pi i t}{12}\right) \quad (28)$$

Finally, SSTA monthly fields were spatially averaged on a 1°x1° grid. The resulting studied SSTA dataset is a 180x360x300 matrix. We removed land cells and obtained 32047 continuous time series of 300 months (no missing data) with a view to characterizing the spatio-temporal variability of the SSTA at global scale from 1985 to 2009. Figure 10 shows the standard deviation of the monthly SSTA for the 1985-2009 periods. It highlights three types of regions of low frequency

variability: the equatorial pacific up to 160°E (and in a less pronounced way the equatorial Atlantic), the temperate regions, mostly of the northern hemisphere, with a maximum amplitude in the north Pacific gyre and the north-western Atlantic gyre, and the equatorial borders of the major upwelling areas.

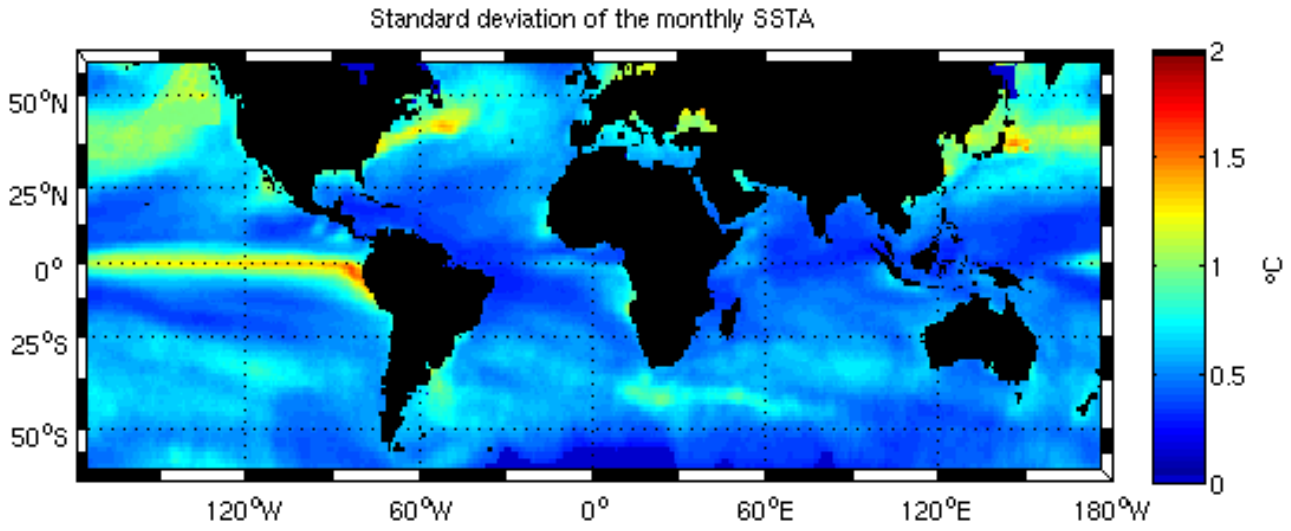


Figure 10: The standard deviation of the monthly SSTA for the period 1985-2009 (source Pathfinder v5.2).

3.3.2 The Multivariate Enso Index

The Niño/La Niña events have been thoroughly addressed in the literature and the reader may refer to http://elNiño.noaa.gov/ as an interesting entry point to understand the ENSO and its regional impacts. The low-frequency variability in the SSTA can be associated with low-frequency atmospheric climatic variations. As a peculiar example, El-Niño-La-Niña events relate to oceanic-atmospheric oscillations of the equatorial Pacific [58]. Among the numerous ENSO-related indexes, we consider the Multivariate ENSO Index (MEI, http://www.esrl.noaa.gov/psd/enso/mei/) which is based on six observed variables over the tropical Pacific: sea-level pressure, zonal and meridional components of the surface wind, sea surface temperature, surface air temperature, and total cloudiness fraction of the sky. As suggested by the NOAA (http://www.esrl.noaa.gov/psd/enso/mei/rank.html), Niño regimes were defined as the periods for which the MEI index value exceeded the percentile 30 of the positive values and conversely, La Niña regimes as significantly negative periods (below percentile 30 of the negative values).

3.3.3 Event detection examples in SSTA time series

Figure 11 shows two SSTA time series (top) and the corresponding wavelet power spectrum (bottom) in the East Pacific. We superimposed the Niño periods (pink) and Niña periods (light blue) on the SSTA time series (top of Figure 11). The detected events in the power spectrum are delimited using ellipses. Events refer to wavelet spectrum areas where the energy levels are significantly greater, at 95% of confidence, than the local red noise theoretical power spectrum, Eq.(20) & Figure 9.

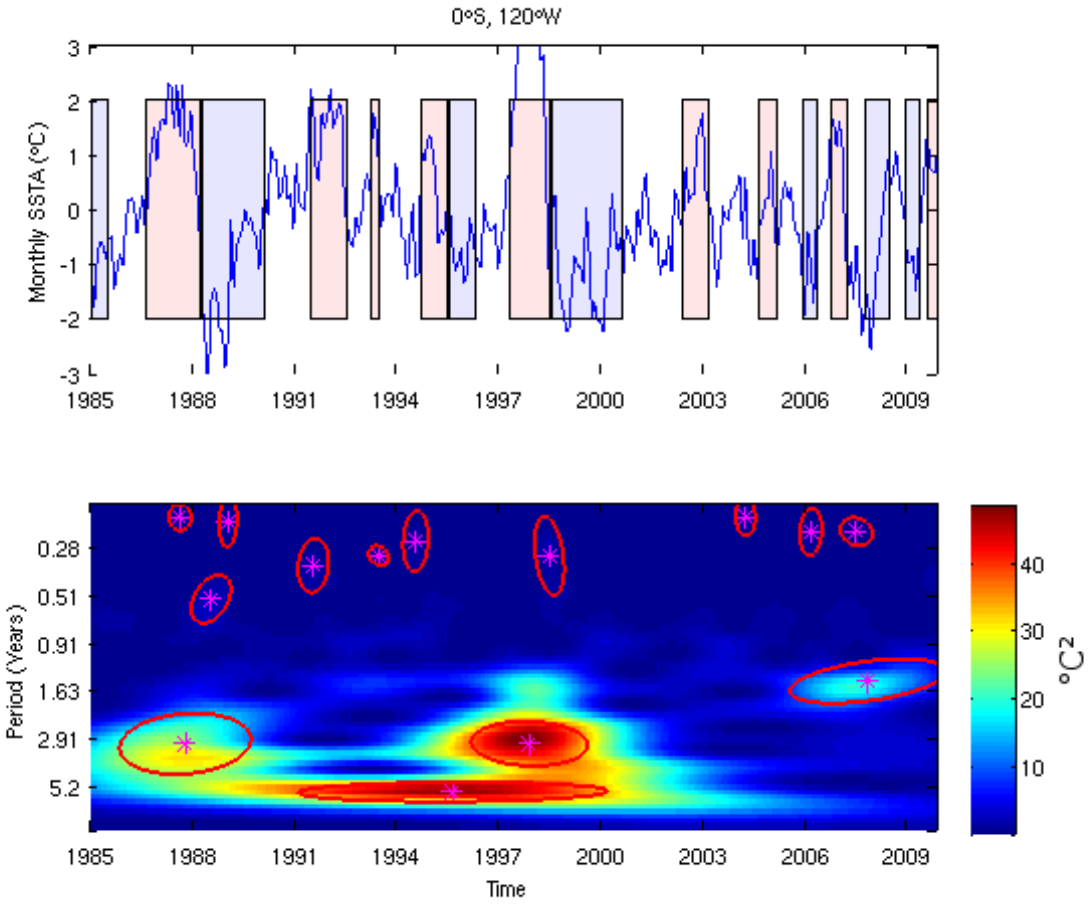


Figure 11 : illustration of the event-based analysis of SSTA time series. Top, SSTA time series observed at 0°N and 120°W, i.e. in the eastern equatorial Pacific known to be strongly affected by ENSO processes. Bottom, the corresponding wavelet power spectrum and the detected significant elementary events delimited by ellipses with the corresponding maximum of energy indicated by a cross. See §2.1 and [111, 112] for details on the detection of the elementary events as local significant spectrum areas with respect to the theoretical energy depicted by a red noise with the same correlation and variance statistics than the considered series.

In the eastern Pacific (Figure 11), at 0°S, 120°W (center of Niño 3 region, http://upload.wikimedia.org/wikipedia/commons/9/9d/Enso-index-map.png), 14 significant events were detected. Two major events occur at the 3.36-year scale from 1986 to 1990 and 1998 to 2000. These two periods correspond to two well-known major ENSO events, each characterized by a succession of a strong Niño and Niña periods, the second period corresponding to the

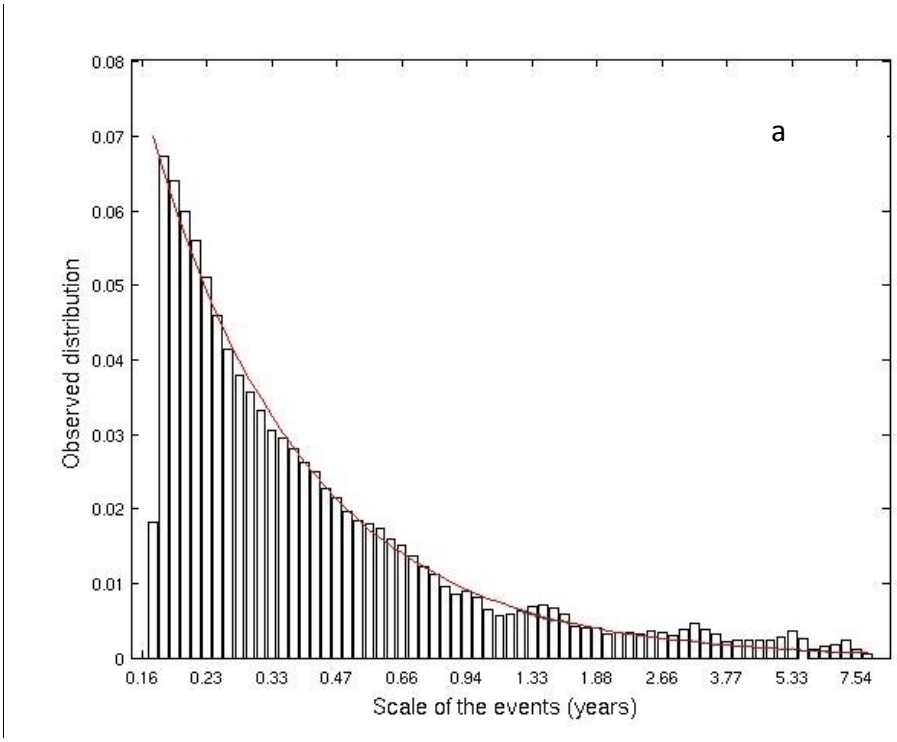
strongest Niño-Niña event recorded. As expected, the variability of the SSTA at low frequency is thus related to the ENSO signal.

3.3.4 Characteristic time-scales of SSTA elementary events.

From the 32047 SSTA time series we extracted 486144 significant elementary events with estimated mean scales from 0.2 to 8 years. Figure 12a shows the time-scale distribution of these events. The exponential distribution (red curve Figure 12a) is the normalization factor used to account for the scale-dependent sampling, Eq.(26). Figure 12b shows the normalized time-scale distribution. We divided our dataset of events in two main categories:

- Events with mean time-scale lower than 0.4 year showed a uniform normalized-scale distribution and were gathered in a single category, referred to the HF (High-Frequency) category.

- Events with scale greater than 0.4 year showed significant Gaussians modes in the normalized-scale-distribution. We fitted a Gaussian mixture model, Eq.(24) to this dataset. The parameter estimation needed 400 EM iterations, using as convergence criterion a log likelihood threshold value of 10^{-4}. The estimated model involved 7 Gaussian modes (Figure 12b).



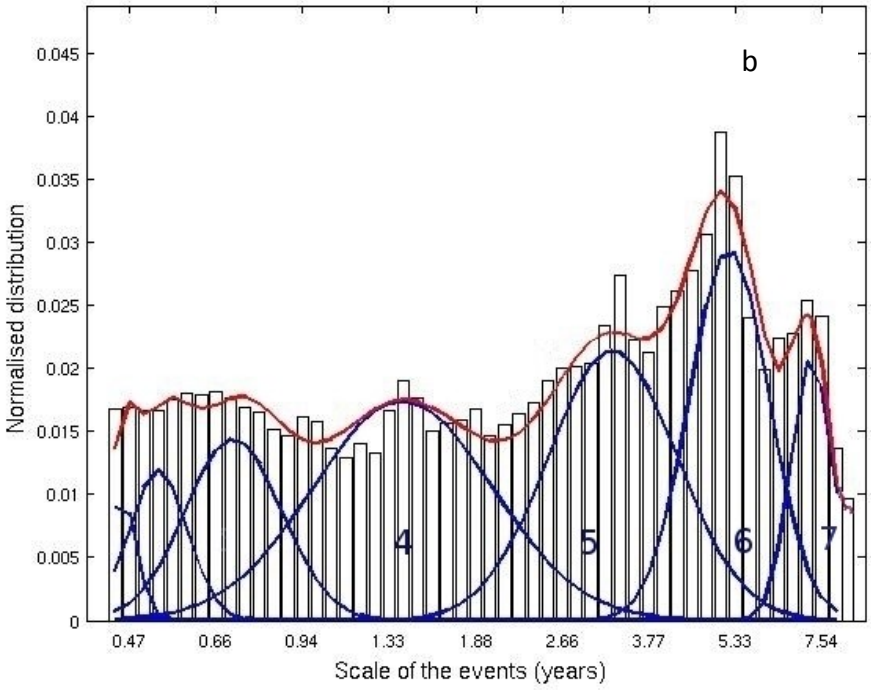


Figure 12: Time-scale distribution and characteristic time-scales of the elementary events extracted from the SSTA dataset. (a) the initial distribution across scales of all of the extracted elementary events and the fitted exponential decay, Eq. (26), corresponding to the natural fractal distribution of the event time-scales [110]. (b) the observed normalized distribution, Eq. (24), i.e. the initial distribution Figure 12a normalized by the red curve of Figure 12a. Figure 12b, the 7 Gaussian modes, Eq.(25) (blue), fitted onto the normalized SSTA time-scale distribution.

Modes 3 and 4 of the GMM showed respectively mean scale of 0.70 and 1.54-years. They both refer to the inter-annual variability, the seasonal component (mainly the energy at a one-year scale), being removed in the SSTA. Modes 5 to 7 at scales 3.36, 5.03 and 7.11 years, contain the very low frequency changes in the SSTA caused by large space-time climatic signals such as ENSO and are therefore considered as the three characteristic low frequency time-scales of ENSO influences on the SSTA.

We will see in the next paragraph that the spatial distribution of the 1.54-year scale events also relates to ENSO region of influence. For this reason we consider this scale as an additional characteristic ENSO time-scale. In Table 1 are summarized the characteristics (mean and standard deviation) of the Gaussian distributions for the four low frequency reference time-scales of the SSTA. Standard deviations for reference scales greater than 1.54 are relatively low ensuring a narrow distribution and a very good confidence in these three classes.

Table 4: Mean and standard deviation of the Gaussian distributions for the four low frequency reference time-scales of the SSTA from 1985 to 2009.

Mean time-scale (μ) in years	1.54	3.36	5.03	7.11
Sigma (σ) in years	1.34	0.56	0.22	0.17

3.3.5 The spatial distribution of the SSTA characteristic scales.

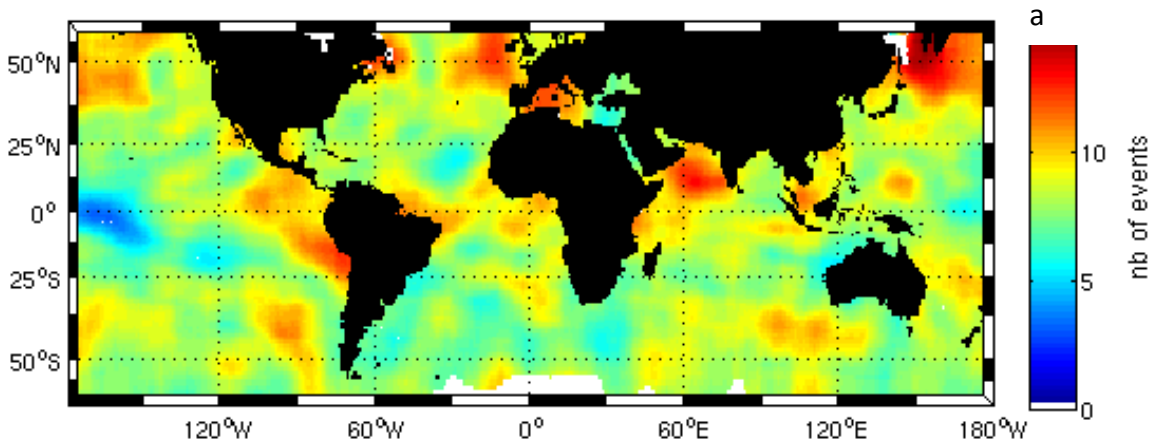
We investigate the spatial distribution of both the HF and the low-frequency characteristic time-scales. From the inferred Gaussian Mixture Model, and knowing all the parameters $\Theta=\{\lambda_{1\ldots I}.\mu_{1\ldots I},\sigma_{1\ldots I}\}$, we evaluate the posterior membership probability $\Pi_{ki} = \mathrm{P}(\mathrm{Y}_k = \mathrm{C}_i|\Theta)$ of the Y_k^{th} event with a mean time-scale s_k to be assigned to the category C_i:

$$\Pi_{ki} = \lambda_i.N(s_k \mid \mu_i,\sigma_i)/\sum_{j=1}^{I}\lambda_j.N\left(s_k \mid \mu_j,\sigma_j\right) \quad (29)$$

The spatial distribution of the given event category C_i is estimated for each pixel (time series) using the expectation of C_i:

$$\mu(\mathrm{C}_i) = \sum_{k=1}^{K}\Pi_{ki} \quad (30)$$

where K is the number of detected events in the time series. Figure 13 shows the estimated $\mu(C_i)$ for both the HF category (s < 0.4 year, Figure 13a) and the characteristic time-scales at 1.54, 3.36 and 5.03 years (Figure 13b-d).



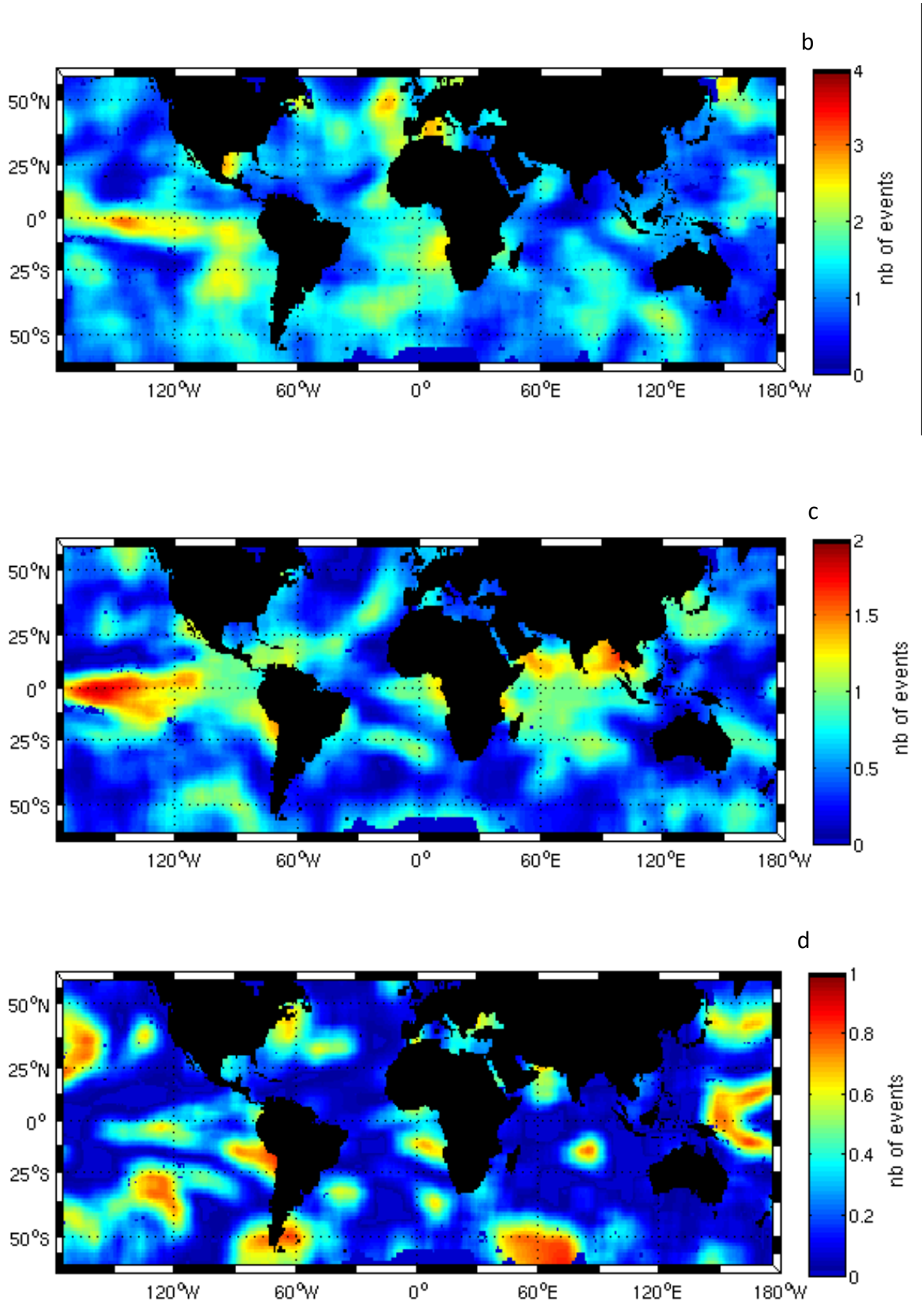


Figure 13: Spatial distribution of the estimated SSTA characteristic time-scales. Mean number of events by time-scale categories from 1985 to 2009. a) for the HF event category (mean time-scale < 0.4 year); b) to d) for characteristic time-scales of respectively 1.54, 3.36 and 5.03 years.

The HF events (Figure 13a) contributed for 61 % of the total number of the detected events. Overall the mean number of detected HF events is of 10.1 over the globe for the considered 25-year period. Local maxima of 18 detections are observed on the Peruvian shores, in the South-Eastern part of the Niño3 reference area, the western Mediterranean Sea, the Arabian Sea and the Okhotsk Sea. The Arabian Sea is strongly affected by monsoon winds reversal, whose effect on SST variability from climatology is already known [116] whereas the high latitudinal areas are commonly affected by winter storms that increase the SST variability, probably enhanced by the presence of the continents. No similar pattern is observed in the high latitudes of the southern hemisphere, probably because of the low interference of continental masses and the regularity of the circum-polar winds. A minimum of 3 events is observed in the equatorial part of the Eastern Pacific from 160 to 180°W, where the variance in the SSTA is mostly driven by the low frequency at 1.54 and 3.36-year scale (Figure 13b&c).

Events at 1.54-year scale (Figure 13b) represent 11% of the total number of events and show a mean number of 2 detected events over the globe for the 1985-2009 period. Local maxima were observed in the eastern part of the Niño3 area underlying the signature of ENSO at this scale on the SSTA. Local patches clearly appear at this scale in the Gulf of Mexico, the North Atlantic, the Namibian shores, the western Mediterranean Sea and the Okhotsk Sea. Eastern boundary systems (Humboldt and Benguela coastal upwelling regions) specially show a higher number of events, probably caused by a high interannual variability.

The 3.36-year event category (Figure 13c) accounts for 4% of all elementary events with a mean number of 0.8 event over the globe from 1985 to 2009. Its spatial distribution highlights regions known to be strongly affected by ENSO: the central equatorial Pacific, the central Humboldt system, and the northern Indian Ocean [9, 58, 116, 47, 117]. In the Indian Ocean, we observe that the low frequency signature of ENSO on the SSTA is also observed at this time-scale. This influence of ENSO on the monsoon in this region has been largely documented [116, 47, 117] but often without time-scale analysis [117] or at multi-decadal time-scale [47].

The 5.03-year event category (Figure 13d) represents 1.5% of the elementary events with a mean number of 0.3 event over the globe. The highlighted areas are the Western part of the Pacific Ocean, the central Humboldt and southern Argentinian shores, as well as in general the boundaries of the regions detected at 3.36-year scales.

ENSO signal is known to propagate [103] and for example ENSO signal generally occurs 4 months after it starts in the West of Peru [103] and until 9 months in the Philippines [103]. Figure 13b-d suggests that the propagation of the ENSO signal includes time-scale shifts as already envisaged by Torrence and Webster [47]. As a peculiar example, in the Eastern inter-tropical Pacific (Figure 13c), the detected events at 5.03-year scale (Figure 13d) geographically surround the detected 3.36-year scale events suggesting the shift between frequencies during the ENSO propagation. We note that the proposed methodology (cf §2.2) suits well to address such hypothesis compared to EOF method that does not involve such an explicit scale-related analysis [103].

3.4 The density of HF and 1.54-year events with respect to ENSO modes.

To address possible high frequency signatures of ENSO from the analysis of the space-time distribution of HF and 1.54-years event categories, we analyze the distribution of both HF and 1.54-year event categories conditionally to the three ENSO conditions. We use the estimation of the starting and ending times of the events of category C to analyze their density D relative to the ENSO regime E_r for normal, Niño and Niña periods (cf § 2.2 for the ENSO mode definition):

$$D(C_i|E_r) = L(C_i|E_r)/L(E_r) \quad (31)$$

where $L(C_i \mid E_r)$ is the number of months spent in events of class C_i during period E_r and $L(E_r)$ the number of months of period E_r. The density $D(C_i|E_r)$ is a time and energy normalized representation of the energy observed for each C_i and ENSO mode. It aims at estimating the number of months where the energy is significant at this time-scale compared to the local conditions.

During normal periods, i.e. out of ENSO periods, the HF (Figure 13a) and the 1.54-year frequency events (Figure 13b) show respectively a global mean density values of 0.19 and 0.23 event.month^{-1}. Local maximum values of HF density are observed in the West of Peru, in an extended area of the Peruvian-Chile upwelling cell, i.e. for most of the Humboldt upwelling system [125], in the Arabian Sea and the Okhotsk Sea. The lowest density value (0.1 event month^{-1}) is reached in the eastern part of the Pacific Ocean. At the 1.54-year scale (Figure 13b), the West African Benguela region (0-20°W; 0-30°S) [119], characterized by a strong upwelling, displays a high density of 0.4 event.month^{-1}, whereas the East equatorial Pacific shows similarly high values (>0.35), but on a relatively moderate extent. The Okhotsk and Mediterranean Sea show a high density of event at the 1.54-year scale, whose origin is probably linked to their specific regional climate and proximity to anticyclonic areas of St Helen and Libya [120].

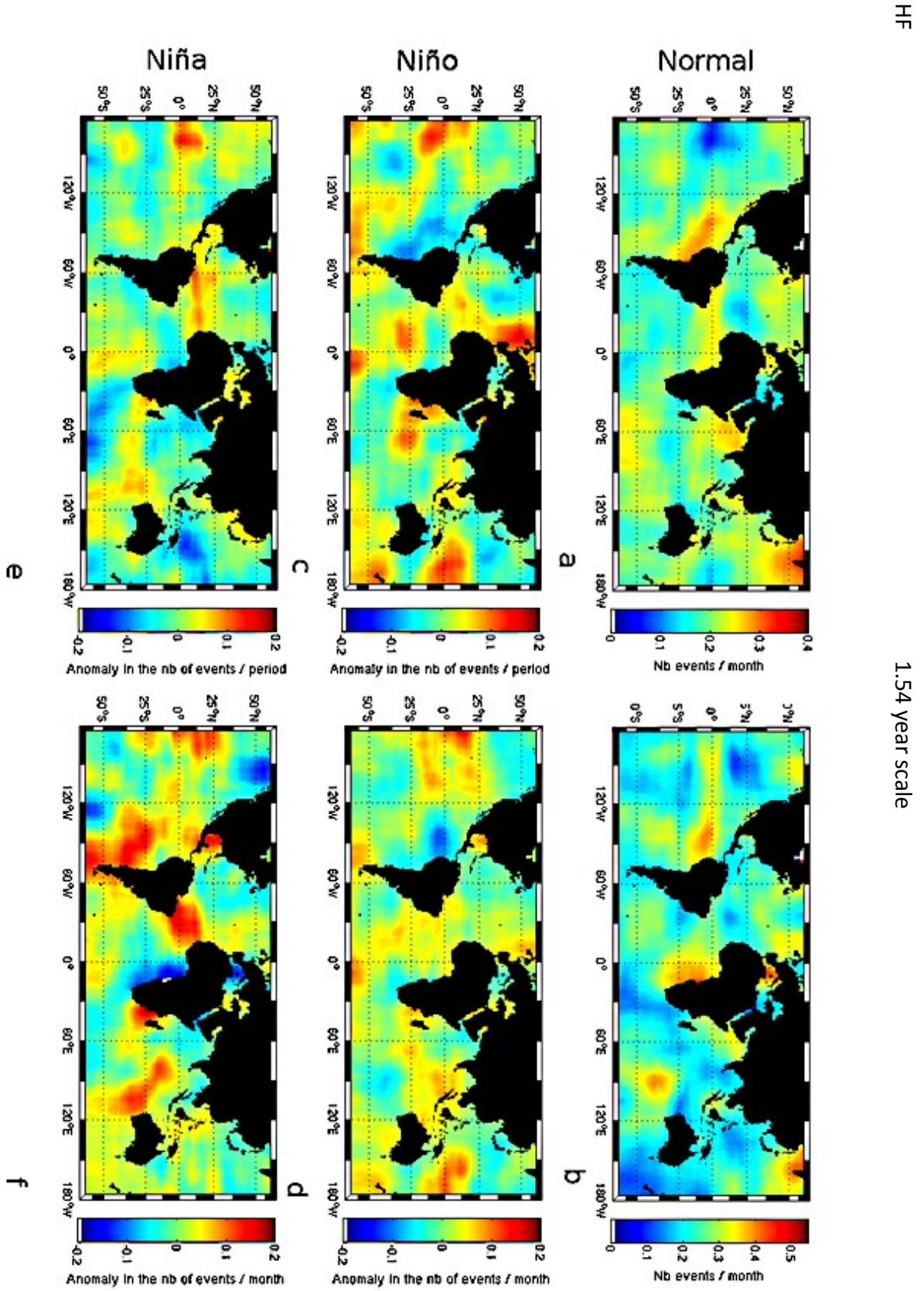


Figure 14: Observed spatial distributions of HF and 1.54 year scale event density for normal conditions (a and b), Niño (c and d) and Niña conditions (e and f).

To highlight the ENSO impact on the time-scale distribution ot the SSTA, event density anomaly maps are computed for both scales during Niño (Figure 14 c-d) and Niña periods (Figure 14 e-f), Figure 14 a-b being taken as the reference state.

During Niño periods (Figure 14 c and d), the intensity of the easterlies decreases, and the warm pool of SST, usually observed in the middle of the inter-tropical Pacific, moves Eastward. We observe globally an increase by 11% and 6% of the HF and 1.54 year frequency events with respect to the density outside ENSO events. A large positive anomaly in the HF density of 0.15 event $month^{-1}$ is observed over the East equatorial Pacific from 160°E to 160°W (Figure 14 c), in the South Atlantic, the Agulhas current [119] and in New Zealand. The last three patterns correspond to the border of the southern ocean anticyclonic regions, except for the eastern pacific, where the ENSO effects dominate. A large positive anomaly of 0.20 $event.month^{-1}$, i.e. an increase by 80% compared to the normal HF conditions, is observed in the North East Atlantic. This is in agreement with a known influence of El Niño in the North Atlantic [118, 121]. Negative anomalies in the HF density (Figure 14 c) are observed in an extended area of the Chile-Peruvian upwelling system, suggesting that the decrease of the easterlies intensity, and the resulting decrease of the upwelling intensity, tends to reduce the number of observed HF events in this area.

Niña periods (Figure 14 e and f) are characterized, in average, by a moderate positive anomaly of 8% for HF events and an increase by 6% for the 1.54-year scale, with nevertheless specific spatial patterns. In contrast to Niño phases, the easterlies strength increases during Niña periods (Figure 14 e and f) and the inter-tropical Pacific surface warm waters move westward. The signature of the southern Humboldt upwelling is clearly visible at the 1.54-year scale, with a positive anomaly of 0.18 $event.month^{-1}$ but off the stronger Peruvian upwelling. In the Guinea gulf and Benguela upwelling (5°N-30°S and 0°-20°W), Niña periods are characterized by a large negative anomaly in the SSTA events at 1.54-year scale. In this region the SSTA is mostly dependent on the upwelling intensity, suggesting a specific stabilization of its variability during Niña periods compared to normal periods (Figure 14b). Off South Africa, we observe a clear opposite influence on the SSTA between the West and the East shores for both Niño and Niña periods and both scales, a difference already highlighted by Rouault [119].

It is obvious that we cannot interpret all the local differences observed in the time-scale distributions of the SSTA anomalies. ENSO signal is particularly complex and involve both atmospheric and oceanic processes, the SSTA being the resulting interaction between these two factors. Other large scale processes such as the Pacific Decadal Oscillation (PDO) [122, 123], and the Atlantic Multidecadal Oscillation (AMO) [124], also affect the SSTA. It appears nevertheless that significant differences are found in the observed distributions of the HF and 1.54-year events time-scales SSTA between one hand the ENSO/normal periods, and the other hand the Niño and Niña phases.

This observation underlies clearly the signature of ENSO on the HF SSTA, and emphasizes the interest of dedicated time-scale decomposition methods, to improve our understanding of processes at various spatio-temporal scales, the reference scales exhibited from the Gaussian Mixture Mode (Figure 12b) being used to choose the time-scales to be studied and discretize the dataset in an optimal way (compared to the standard wavelet analysis).

3.5 Investigating frequency shifts in the SSTA and the inter-tropical Pacific during the ENSO 1997-2000 event.

Time-scale changes during Niño-Niña periods are suspected to occur [47, 48, 49, 50, 51] and Campo [49] underlined that ENSO time-scale variability at decadal scales may differ substantially from one ENSO event to another. An and Wang [52] also showed a significant relationship between the observed ENSO frequency and observed SSTA structures, underlying the influence of these frequency shifts on the circulation. To investigate the time-scale variability of ENSO, we focus on the 1997-2000 ENSO major event in the inter-tropical Pacific, the strongest ever recorded. We clearly see in Figure 13c the strong signature of this ENSO event in the inter-tropical Pacific at 3.36-year scale. Figure 15a shows the distribution of the observed mean time (center of the events) at 3.36-year scale. The maximum in the distribution is observed in August 1999, i.e. approximately in the middle of the Niño period. To investigate frequency shifts during we estimate the distribution of the variables δT_1 & δT_2:

$$\delta T_1 = T_{1.54} - T_{3.36} \qquad (32)$$
$$\delta T_2 = T_{HF} - T_{3.36}$$

where T3.36, T1.54 and THF are respectively the position (time) of the event centers (maximum of energy) at 3.36-years, 1.54-years and HF, collocated in the same spectrum. Figure 15b shows the probability density functions of both δT_1 and δT_2, estimated using respectively 1302 and 1474 pairs of events.

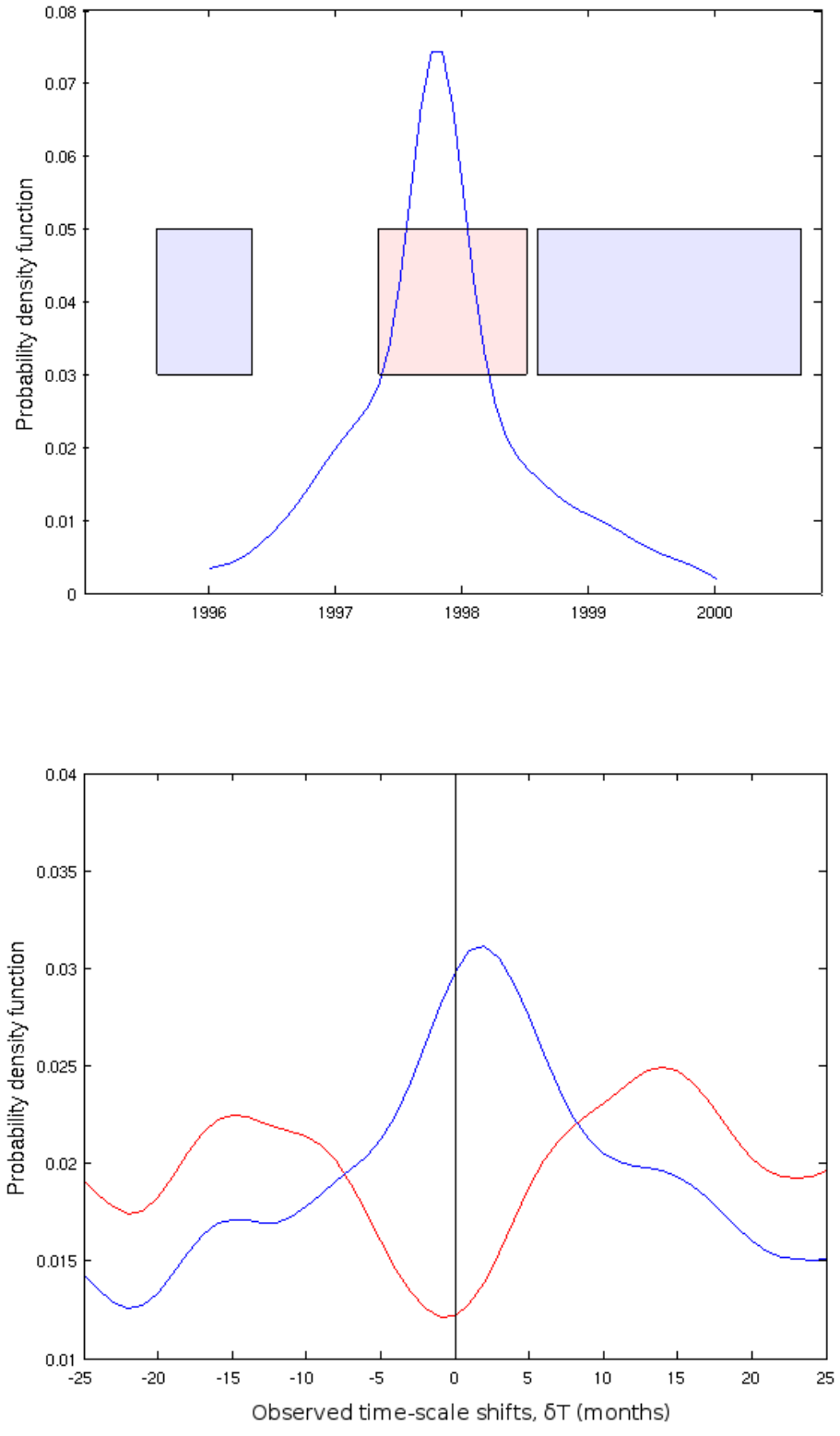


Figure 15: (top) temporal distribution of the maximum of energy (event centers) observed in the inter-tropical Pacific at 3.36-year scale (known as being a reference time-scale for ENSO cf Figure 13b and [8]). In pink are highlighted the Niño period and light blue the Niña periods. (bottom)

Distribution of the observed time shifts between the maximum of energy of the events at 3.36 and 1.54-years (blue) and 3.36 and HF (red).

In Figure 13b, we observe a clear increase of the number of 1.54-3.36 year pair of events, with a maximum observed 2 months later than the maximum at 3.36-years. Conversely, we observe a decrease of the number of HF-3.36 year pair of events in the inter-tropical Pacific during the maximum of the ENSO 1997-2000 event with a minimum observed 2 months before the maximum at 3.36 years. This underlies the time-scale shifts from high to low frequencies during this ENSO. This observation motivates the determination of both the reference scales and their relative distribution to characterize ENSO compared to a classical sum of the spectrum energy between the 1.5 and 7-year scales [8]: even if the 1.5-7 year sum of energy is constant in time, the shift between scales provides a significant signature of ENSO 1997-2000.

3.6 Conclusions and future work

Our proposed methodology involves the event-based mining of geophysical time series. We regard a time series as a collection of significant elementary time-scale events complemented by a red noise process. Our approach resorts to a normalized representation in variance of a time series through the detection of significant time-scale events. This is of key interest for SST anomalies that show a high spatial and temporal variability of the variance. The estimation of the threshold in energy to detect a significant event accounts for the auto-correlation and noise level of the local time series. This is also a key issue for geophysical time series, which depict naturally large autocorrelation noise levels (typically from 0.3 to 0.7 in the monthly SSTA dataset studied here).

The method is applied to the global SSTA observed from 1985 to 2009. We use a mixture of Gaussian to identify four reference time-scales at 1.54, 3.36, 5.03 and 7.11 years. The spatial distribution of these low-frequency reference scales highlights the inter-tropical Pacific, the West of Peru, the Indian Ocean and the South of the Atlantic, regions known as being strongly influenced by ENSO. In addition, we reveal that ENSO modes are also characterized by significant space-time differences in the distribution of high-frequency events (typically, with characteristic time-scale below 4 months). We show that the high frequency event density of the SSTA increases by in mean 11% over the globe during Niño events, with a maximum of 80% in the North East of Europe, and 6 % during Niña periods. Even if all this high frequency variability may not be attributed to ENSO, this large increase is a significant signature of the Niño periods and is minimized by the EOF approach, which tends to exhibit by construction the low frequency correlation modes. Our method also allows identifying times-scale shifts in the energy spectrum in the inter-tropical Pacific during the major 1997-2000 ENSO event with a maximum shift from the high frequency towards the reference (3.36-year scale) observed 2 months before the maximum of energy at the 3.36-year scale.

Compared to EOF-based time series analysis, the key contribution of the proposed event-based approach is to account for signal non-stationarity and noise autocorrelation in the time-scale decomposition of geophysical processes variability. Whereas EOF-based schemes mainly reveal low-frequency patterns, our wavelet-based approach can identify both low-frequency and high-frequency signatures and investigate their respective space-time distribution.

Compared to classical wavelet approach, our main methodological contribution lies in the characterization of significant times-scales in the SSTA taking in account the spatially-varying variance and the autocorrelation of each time series. The classical wavelet approach [8, 47] usually considers a time-scale sum of the energy between 1.5 and 7 years to depict the ENSO signatures in the SSTA [47]. Nevertheless, even if the 1.5-7 year sum of energy may be constant in time, the shift between scales is also a significant signature of ENSO 1997-2000 (Figure 15b). Compo [47] already pointed out that ENSO-induced changes of extratropical 500mb height variability are time-scale dependent and An [52] showed also a relationship between ENSO frequency changes and observed structure in the SSTA, raising the crucial question of the choice of discrete frequency bands. The present methodology addresses this question providing a quantitative mean to unmix the processes and study their time-scale relationships (Figure 15b). These time-scale dependencies may evolve in time, which make them particularly difficult to deal with from an extension of the cross-wavelet spectrum [8].

Our event-based methodology opens new perspectives for the analysis of multivariate time series such as wind and SST, light and chlorophyll-a. While we consider here the interaction between elementary events at different characteristic scales of the same geophysical variable, this methodology could be applied to two or more variables. Besides, the event-based detection could also be considered to address long term trend estimation [4] and correlation analysis while being robust to the presence of low-frequency non-stationary signals such as ENSO.

4 Chapter IV: Characterization of time-varying regimes in remote sensing time series: application to the forecasting of satellite-derived suspended matter concentrations.

This chapter addresses an important scientific issue for time series-analysis: optimization of observation and statistical-based forecasting models. It is a recurrent issue for marine scientists who often wish to forecast a high resolution geophysical variable such as SST, winds or surface velocity, using both observations (mainly satellite) and available model outputs [126, 127, 128, 129].

Inherently, physical processes such as turbidity (our variable of interest here) or SST, and biological processes such as chlorophyll-a, fish and shell growth, are often by definition non-stationary and time-varying processes as they are driven by seasonal signals. From a methodological point of view, this chapter addresses characterization of multiple regimes (expressed here using linear multivariate regressions, Y=AX) between Y and its predictors X. For each regime, Y thus has a specific response to X. This aspect is fundamentally different from a single linear (multivariate regression) or non-linear (such estimated with SVR) response of Y to X.

The regimes are identified using and hidden variable Z and the temporal dynamic of Z is considered as a Markovian process. Another important issue in the statistical modelling is also addressed in this section: the selection of predictors X that contribute significantly to the estimation of Y, and the estimation of the appropriate number of regimes.

This chapter was published in 2014 in the IEEE 'Journal of Selected Topics in Applied Earth Observations and Remote Sensing (JSTARS)' [11]

4.1 Introduction

The forecasting of a geophysical variable using statistical models is an alternative to model-based approaches which typically involve complex simulation and/or assimilation [131, 132]. For instance, coupled hydrodynamic and sediment transport models can be used to estimate the concentration of suspended particulate matters within the water column [133] while statistical approaches may use available satellite and model data to predict the same variable [134]. Many statistical approaches have been proposed and evaluated to forecast or infer a studied variable from predictors. Among them, linear multivariate regression [75] and non-linear (polynomial) multivariate regression [135] are the most known. Numerous specific models dedicated to time series analysis such as AutoRegressive Moving Average (ARMA) and AutoRegressive Integrated Moving Average (ARIMA) models [136] have also been developed initially to address financial time series. These latest, which aim at studying the behavior of a time series without considering

forcing factors, have also been applied to geophysical time series [137]. Non-linear regressions, based on supervised learning strategies, such as Neural Networks [138] and Support Vector Regressions (SVR) [139] may provide relevant alternatives to estimate a variable from predictors. In the context of geophysical studies, they may nevertheless suffer from two major drawbacks. First, though relevant regression performances may be reported, these models may not be physically interpretable and may be very sensitive to the training dataset. Second, multi-regime dynamics, often exhibited by geophysical processes driven by the seasonality [140], cannot be addressed by such non-linear models, contrary to latent-regime models as demonstrated in our study.

We propose here to characterize time-varying relationships between a variable and its forcing parameters using latent-regime models, and hence optimize forecasting results. As an illustration, we address the concentration of inorganic suspended particle matters (SPIM), estimated from satellite data using a regional algorithm [141, 142], and observed in the mouth of the Gironde estuary. In this area, sediments are mainly exported from the Gironde estuary [142, 143] and SPIM concentration clearly depends on the local physical forcing: swell, tide, wind and river outflow. A minimum of energy has to be brought by waves and tides to re-suspend cohesive sediments accumulated at the bottom. Conversely, when sediments have been re-suspended in the water column by wave influence, their settling velocity depends on their size and density [144] and physico-chemical properties [145]. This example stresses that the relationships between the studied variable (SPIM) and the causing factors evolve in space and time and potentially requires advanced statistical methods to identify the underlying geophysical regimes.

From a methodological point of view, “latent regime regressions” also referred as “clusterwise regressions” [146, 147] are particularly appealing to identify such non-linear and multi-regime patterns within a dataset. Each regime is associated with a linear regression and the overall non-linear patterns are thus estimated as a combination of the different linear contributions. Regarding the temporal dynamics of these regimes, we here consider Markovian processes [39], which state the transitions in time between two regimes. The standard Hidden Markov Model (HMM) and Non-Homogeneous Hidden Markov Model (NHHMM) are evaluated [39]. The inclusion of an autoregressive term (HMM-AR) and (NHHMM-AR) is also discussed. This aspect is motivated by the strong autocorrelation level depicted by geophysical time series [148]. When the observation of the previous day (referred as t-1) is available, it is obvious, considering the strong natural autocorrelation of geophysical data that the forecast at time t should take in account the observation at time t-1. Conversely, for specific applications, or if the observations are not available during long periods (such as winter storms, or after a sensor failure), one may need to estimate the variable without using the observations of the previous days. We discuss here the choice between autoregressive or non-autoregressive models for long lacks of observation period using forecasting results from t+1 to t+15.

Model parameter estimation is carried out from a dataset composed of 5862 time series of 1096 points in the mouth of the Gironde estuary in the [3°W-1°E ; 45-46.5°N] area during the period 2007-2009. Validation is performed on the same area for using the data for the year 2010. We

used EOFs to reduce the dimension of the space-time observations. This is a usual approach in spatio-temporal statistics [9, 149] although alternatives may be considered such as linear discriminant analysis [150], and, we could also introduce a latent variable to describe the regime at each location and interact with the regimes at other locations. Nevertheless, such models are known to be very difficult to fit on the data and remain a research challenge for statisticians. We infer our mixture model using the expansion coefficients of the first four modes of the EOF which explain 99% of the total variance. The variables used as predictors for the SPIM expansion coefficients (EC) are the wave height issued from a numerical model [151], the wind fields optimally interpolated from satellite observations [152], the tide coefficient [153] and the Gironde fresh water outflow (sum of the Garonne and Dordogne rivers contributions).

4.2 Methods

4.2.1 Markov switching models

We address here the study of a two dimensional scalar geophysical time series Y. In a hidden Markov model framework (HMM; [39]), one states two different processes, the observed process Y and a hidden process Z. The observed process (here the turbidity) is assumed to be temporally dependent of the hidden process. At a given time t, the hidden variable Z_t is a discrete value which states the regime in play at time t characterized by a latent [146] linear regression model with coefficient B_k between the variable Y_t and the predictor X_t. The conditional likelihood of the observation Y_t given predictor X_t and regime Z_t is thus expressed as [146]:

$$\mathrm{P}(\mathrm{Y_t}|\mathrm{X_t}, \mathrm{Z_t} = \mathrm{k}) \sim \mathrm{N}(\mathrm{X_t B_k}, \sigma_\mathrm{k}) \quad (33)$$

where N represents the Gaussian probability density function with mean $\mathrm{X_t B_k}$ and variance σ_k^2. The $\mathrm{B_k}$ linear coefficients are estimated using a weighted linear regression and the training dataset $\{Y_t, X_t\}$ for the 2007-2009 period. The hidden process Z_t is modeled as a first order Markov chain [39] characterized by its transition probability matrix between Z_{t-1} and Z_t

In the simplest case (HMM), one assumes homogeneous transitions, i.e. time and context-independent transition matrix. The NHHMM allows the transition matrix between the hidden regimes to depend on a set of observed covariates S_t. Hughes and Guttorp [155, 156] highlighted the added value of the NHHMM to characterize the links between the large-scale atmospheric measures and the small-scale spatially discontinuous precipitation field. In the NHHMM settings, the probability transition matrix is now time-dependent and conditioned by the covariates S_t:

$$P(Z_t = k \,|Z_{t-1} = l, \mathrm{S_t}) = \left[P\left(\mathrm{S_t} \,|\, Z_{t,} = k,\ Z_{t-1} = l\right) .\ P(Z_t = k|Z_{t-1} = l)\right] \Big/ \left[\sum_{k,l} P(\mathrm{St} \,|\, Z_t = k\ ,\ Z_{t-1} = l)\ .\ P(Z_t = k|Z_{t-1} = l)\right] \quad (34)$$

The non-homogeneous matrix transition is derived from the likelihood of the covariate S_t given transition from Z_{t-1} to Z_t. We suppose that the probability density function of the covariates during this change of regime follows a normal distribution:

$$P(S_t \mid Z_t = k, Z_{t-1} = l) = N(\mu_{l,k}, \Sigma_{l,k}) \quad (35)$$

Where N is a multivariate normal distribution of dimension n, the number of covariates used to estimate the transitions with mean $\mu_{l,k}$, and covariance matrix $\Sigma_{l,k}$. In the present application, and to reduce the number of parameters to be estimated, we consider that the predictors are uncorrelated (null covariance) and their relative influence is identical (same variance), i.e. $\Sigma_{l,k}$ is a multiple of the identity matrix.

Figure 16 shows a graphical representation of the conditional dependencies involved in the model, in the form of the general Directed Acyclic Diagram (DAG). It illustrates the interactions between the variable Y_t, the predictors X_t, the hidden regime Z_t and the covariate S_t which acts on regime switching. X_t and S_t are known, as they are either observations or numerical model outputs. X_t contains forcing variables such as wind, wave height, tide coefficient and river outflow, and eventual lagged values of Y_t (referred as autoregressive terms). Figure 16 defines a general family of model which encompasses the most usual ones with regime switching. When no covariate is considered i.e., Z_t only depends on Z_{t-1}, and, Y_t, only depends on (Y_{t-s} ..Y_{t-1}) and Z_t, we retrieve the usual Markov switching autoregressive (MS-AR) model. If we further assume that s=0 (without autoregressive component, Y_{t-1}) then we obtain a Hidden Markov Model (HMM). When Z_t does not dependent on Z_{t-1} but only on S_t it comprises the threshold autoregressive (TAR) model which is another important family of regime-switching models in the literature. HMMs, MS-AR and TAR have been used in many fields of applications including geosciences [154].

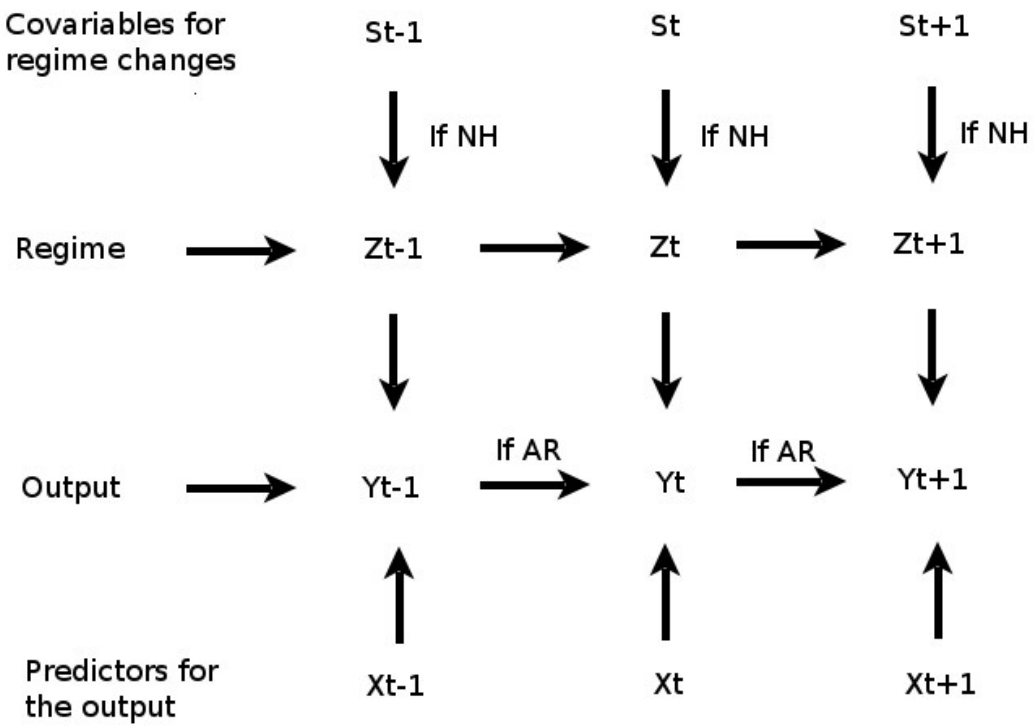


Figure 16: Graphical representation of the various Markov-Switching Models considered in this work: the arrows state the conditional dependencies between the random processes in play, namely hidden regime process Z, observed process Y, prediction process X and regime change covariate process S.

An homogeneous Markov chain (HMM, [39]) is characterized by its transition matrix $P(Z_t \,|Z_{t-1}\,)$, its initial law $P(Z_0 = k\,|X_0, Y_0\,)$ and the conditional probability $P(Y_t\,|X_t, Z_t\,)$ referred in the literature to the emission probability. A key property of the considered Markov switching models is the factorized expression of the joint likelihood of the observed and hidden processes.

The probability of the hidden Markov process is given by $P(\,\mathrm{Y}_0^s, \mathrm{Z}_0^s|\mathrm{X}_0^s)$, where Y_0^s (resp. X_0^s and Z_0^s) are Y values from t=0 to s. $P(\,\mathrm{Y}_0^s, \mathrm{Z}_0^s|\mathrm{X}_0^s)$ can be expressed using the Bayes rules as:

$$P(\,\mathrm{Y}_0^s, \mathrm{Z}_0^s|\mathrm{X}_0^s) = \; P\,(\mathrm{Y}_0^s|\,\mathrm{X}_{0,}^s\mathrm{Z}_0^s\,).\,\mathrm{P}(\mathrm{Z}_0^s|\mathrm{X}_0^s) \tag{36}$$

Given the Markovian memoryless property and dependencies $P\,(\mathrm{Y}_0^s|\,\mathrm{X}_{0,}^s\mathrm{Z}_0^s\,)$ can be factorized [Figure 16, 159]:

$$P\,(\mathrm{Y}_0^s|\,\mathrm{X}_{0,}^s\mathrm{Z}_0^s\,) = \prod_{t=0}^{s} P\,(\mathrm{Y}_t\;|\,\mathrm{X}_{t,}\,\mathrm{Z}_t\;\,) \tag{37}$$

and

$$\mathrm{P}(\mathrm{Z}_0^s|\mathrm{X}_0^s) = \mathrm{P}(\mathrm{Z}_0^s) = \left[\prod_{t=1}^{s} P(Z_t\,|Z_{t-1})\right].P(Z_0)$$

Eq. 36 finaly factorizes:

$$P(\,\mathrm{Y}_0^s, \mathrm{Z}_0^s|\mathrm{X}_0^s) = \prod\nolimits_{t=0}^{s} P(Y_t\,|X_t, Z_t\,).\prod\nolimits_{t=1}^{s} P(Z_t\,|Z_{t-1}\,).P(Z_0) \tag{38}$$

4.2.2 Estimation of the model parameters

The considered models involve two categories of parameters to be estimated: those of the observation model, θ_k, namely regression coefficient B_k and standard deviation σ_k for each regime and θ_s the parameters of the hidden Markov-swtiching process. For homogeneous models θ_s is the transition matrix $P(Z_t\,|Z_{t-1})$ while for non homogeneous models θ_s is the $(\mu_{l,k}, \Sigma_{l,k})$ parameters. Given observed Y and X series, we proceed to the estimation of model parameters according to the maximization of the log likelihood, using the Expectation Maximisation (EM) procedure [159, 157]:

$$\log(L(\theta)) = \log(\mathrm{P}(\mathrm{Y}_0^T|\mathrm{X}_0^T, \mathrm{S}_0^T, \theta)) \tag{39}$$

where T is the time-index of the last observation (i.e. all the series are observed) and $\theta = \{\theta_s, \theta_k\}$ the set of parameters to be estimated. The EM procedure proceeds iteratively as follows: for a given initialization for the parameters the procedure iterates estimation steps (E-step) of the posterior regime likelihood $P(Z_t = k|\,\mathrm{Y}_0^T, \mathrm{X}_0^T, \theta)$ [159], and the maximisation step (M-step), to update the parameters given these posteriors. The algorithm iterates until convergence between steps n and n+1, i.e. $|\,L\left(\theta^{(n)}\right) - L\left(\theta^{(n+1)}\right)| < 10^{-3}$.

More precisely, the posterior regime likelihood $P\left(Z_t = k\middle|\,\mathrm{Y}_0^T, \mathrm{X}_0^T, \theta^{(n)}\right)$ are estimated at step n+1 using the classical forward-backward recursions [39, 159] given series X and Y and current parameter estimate $\theta^{(n)}$. The M-step re-estimates $\theta^{(n+1)}$ using the the posterior regime likelihood at step n+1 and $\theta^{(n)}$. The EM algorithm maximizes the function Q, namely the expectation of the log of the incomplete likelihood, Eq. (40), conditionally to Y and X and $\theta^{(n)}$ [159]:

$$Q(\theta, \theta^{(n)}) = \sum_{z_0^T} \log(P(\,\mathrm{Y}_0^T, \mathrm{Z}_0^T = z_0^T|\mathrm{X}_0^T, \theta)).P(Z_0^t = z_0^T|\mathrm{Y}_0^T, \mathrm{X}_0^T, \theta^{(n)}) \tag{41}$$

Using Eq (38), Eq. (41) resorts to (see §4.2 of [159] for details)

$$Q\left(\theta, \theta^{(n)}\right) = \sum_{t=1}^{T}\sum_{k} \log\left(\mathrm{P}\left(\mathrm{Y}_t\ |\mathrm{X}_t\ , Z_t\,, \theta\right)\right).P\left(Z_t = k\middle|\mathrm{Y}_0^T, \mathrm{X}_0^T, \theta^{(n)}\right) + \tag{42}$$

$$\sum_{t=1}^{T}\sum_{k,l} \log\left(P(Z_t\ \middle|Z_{t-1}\,, \theta_s^{(n)}\right)).\ P(Z_t = k, Z_{t-1} = l|\mathrm{Y}_0^T, \mathrm{X}_0^T, \theta^{(n)})) +$$

$$\sum_k \log\big(P(Z_0 = k|\theta)\big) . P\left(Z_0 = k \middle| \mathrm{Y}_0^T, \mathrm{X}_{0,}^T \theta^{(n)}\right)$$

From Eq(42) we see that it is possible to break the optimization problem in three parts, the estimation of the observation model parameters $\hat{\theta}_k$ (first term of Eq(42)), the estimation of transition parameters $\hat{\theta}_s$ (second term) and the estimation of the initial state (last term).

Model parameters $\hat{\theta}_s$ of the Marvov switching process are estimated with maximizing the second term of Eq. (42).

$$\hat{\theta}_s^{(n+1)} = argmax_{\theta s}\left(\sum_t \log(P(Z_t = k\,|Z_{t-1} = l, \theta_s^n)\ .\ P(Z_t = k, Z_{t-1} = l\,|\mathrm{Y}_0^T, \theta^{(n)})\right) \tag{43}$$

For $\hat{\theta}_{k,}^{n+1}$ estimation of the regression parameters $\widehat{\mathrm{B}}_k^{n+1}$ involves the maximization of the first term, i.e. the estimation of the k weighted linear regressions parameters with a least square criterion, where the weights are given by the posterior likelihoods $P(Z_t = k|\mathrm{Y}_0^T, \mathrm{X}_0^T, \theta^{(n)})$ observed at step n. $\widehat{\sigma}_k^{n+1}$ is the weighted residual standard deviation given $\widehat{\mathrm{B}}_k^{n+1}$ and weights at step n+1:

$$\hat{\theta}_k^{(n+1)} \begin{cases} \widehat{\mathrm{B}}_k^{(\mathrm{n}+1)} = argmin_{B_k} \sum_t \left(P(Z_t = k|\mathrm{Y}_0^T, \mathrm{X}_0^T, \theta^{(n)})\ (Y_t - B_k X_t\right)^2) & (44) \\ \widehat{\sigma}_k^{(\mathrm{n}+1)} = \sum_t \left(P(Z_t = k|\mathrm{Y}_0^T, \mathrm{X}_0^T, \theta^{(n)})\ \ (Y_t - B_k X_t\right)^2) & (45) \end{cases}$$

4.2.3 Forecasting application

The considered multi-regime regression models are applied to the short-term forecasting of series Y. More precisely, at a given time t, we aim at predicting variable Y at time t+dt. We assume that prediction variables X and covariates S, typically numerical simulations, are available up to time t+dt whereas the variable Y is only known up to time t. Thus, the forecast at time t+dt, denoted by $\widehat{\mathrm{Y}}_{t+\mathrm{dt}}$ is given by the conditional expectation of variable $\mathrm{Y}_{\mathrm{t+dt}}$ given observations series up to time t and predictor series up to time t+ dt (see Krolzig [160] for details):

$$\widehat{\mathrm{Y}}_{t+\mathrm{dt}} = \mathrm{E}(Y_{t+\mathrm{dt}}|\mathrm{Y}_0^t, \mathrm{X}_0^{t+dt}) = \sum_k P\big(Z_{t+\mathrm{dt}} = k\big|\ \mathrm{Y}_0^t, \mathrm{X}_0^{t+dt}\big)\, E\,(\mathrm{Y}_{\mathrm{t+dt}}|\mathrm{X}_{\mathrm{t+dt}}, \mathrm{Z}_{\mathrm{t+dt}} = \mathrm{k}) \tag{46}$$

For HMM it resorts to:

$$\widehat{\mathrm{Y}}_{t+\mathrm{dt}} = \sum_{k} P\left(Z_{t+\mathrm{dt}} = k \middle| \mathrm{Y}_0^t, \mathrm{X}_0^{t+dt}\right) . \mathrm{X}_{t+dt}\, \mathrm{B_k} \quad (47)$$

For NHHMM it resorts to:

$$\widehat{\mathrm{Y}}_{t+\mathrm{dt}} = \sum_{k} P\left(Z_{t+\mathrm{dt}} = k \middle| \mathrm{Y}_0^t, \mathrm{X}_0^{t+dt}, \mathrm{S}_0^{t+dt}\right) . \mathrm{X}_{t+dt}\, \mathrm{B_k} \quad (48)$$

For autoregressive models HMM-AR and NHHMM-AR, i.e. X_{t+dt} contains Y_{t+dt-1} which is not available, $\widehat{\mathrm{Y}}_{\mathrm{t+dt}-1}$ is thus estimated using $\widehat{\mathrm{Y}}_{\mathrm{t+dt}-2,}$ X_{t+dt-1} and θ. The HMM-AR estimation resorts to:

$$\widehat{\mathrm{Y}}_{t+\mathrm{dt}} = \sum_{k} P\left(Z_{t+\mathrm{dt}} = k \middle| \mathrm{Y}_0^t, \mathrm{X}_0^{t+dt}\right) [\, \alpha \widehat{\mathrm{Y}}_{t+dt-1} + \mathrm{X}_{t+dt}\, \mathrm{B_k}] \quad (49)$$

and for the NHHMM-AR:

$$\widehat{\mathrm{Y}}_{t+\mathrm{dt}} = \sum_{k} P\left(Z_{t+\mathrm{dt}} = k \middle| \mathrm{Y}_0^t, \mathrm{X}_0^{t+dt}, \mathrm{S}_0^{t+dt}\right) [\, \alpha \widehat{\mathrm{Y}}_{t+dt-1} + \mathrm{X}_{t+dt}\, \mathrm{B_k}] \quad (50)$$

It might be noted that these predictions actually account for the uncertainties in the determination of the underlying regimes. Contrary to deterministic methods, confidence interval and uncertainties on $\widehat{\mathrm{Y}}_{t+\mathrm{dt}}$ can be derived [161] which is a key issue for modeling considerations.

4.2.4 Model performance estimation

A key issue in practice, which has received lots of attention in the last few years, is the problem of model selection which aims at finding the ”optimal” number of predictors and covariates [158]. Hereafter, we have chosen to use both the Bayes Information Criterion (BIC) and the explained variance (EVAR) as a first guides. BIC index generally permits to select parsimonious models which fit the data well [162]. It is defined as:

$$\mathrm{BIC} = -2 \log^*(\mathrm{L}) + p^* \log(\mathrm{S}) \quad (51)$$

Where L is the likelihood of the data, p is the number of parameters and S is the number of observations. We also use the classical explained variance, EVAR, to characterize the model relevance:

$$\mathrm{EVAR} = 1 - \mathrm{var}\left(\widehat{\mathrm{Y}}_{t+1} - \mathrm{Y}_{t+1}\right) / \mathrm{var}(\mathrm{Y}_{t+1}) \quad (52)$$

BIC and EVAR are partially linked [162]. BIC tends to penalize complex models whereas explained variance criterion only qualifies the result and may lead to the over-parameterization of a model

that typically leads to errors when other dataset are tested using the same parameterization. Therefore we consider both BIC and EVAR to assess the model performance.

4.3 The data

4.3.1 The studied variable

Non-algal SPM concentrations (SPIM) are estimated using an analytical algorithm [141] defined as the difference between total SPM and phytoplankton biomass, the latter derived from the chl-a concentration. It incorporates mainly mineral SPM and smaller amounts of organic SPM not related to living phytoplankton. This method to derive non-algal SPM from remote-sensing reflectance is based on the inversion of a simplified equation of radiative transfer, assuming that chlorophyll concentration is known. This merged dataset consists of fields of non-algal surface SPM concentrations, derived from the Sea-viewing Wide Field-of-view Sensor (SeaWiFS), the Moderate Resolution Imaging Spectroradiometer (MODIS) and the Medium Resolution Imaging Spectrometer (MERIS) sensors, provided by the Ocean Color TAC (Thematic Application Facility) of MyOcean, and interpolated with a kriging method [92] for the period 2007–2009 over the Gironde mouth river from 3°W-1°E ; 45-46.5°N. Finally 5682 continuous time series of 1096 days compose our initial dataset of mineral suspended matters concentration. We first account for the space-time variability of the dataset, previously detrended and centered for each time series [4] using a EOF decomposition [9], expressed here using the matrix form:

$$Cov(SPIM)=UVU^{t} \quad (53)$$

where U is a here 5682*5682 matrix containing the spatial modes (Eigenvectors) of the covariance decomposition (ordered by percentage of explained variance). Associated with each spatial mode k, its expansion coefficient (also referred in the literature as principal component) is the time evolution of the k^{th} mode:

$$EC\_SPIM_{k,t} = SPIM_{t} * U_{k} \quad (54)$$

Figure 17 shows the four first spatial modes of the EOF decomposition. Figure 18 depicts the four associated time series $EC\_SPIM_{i=1,4}$. The first mode (Figure 17a) comprises 85% of the total variance. It clearly addresses the seasonal cycle as shown in Figure 18a where the switch between winter (high values of EC_SPIM_1 correspond here to high values of SPIM observed in winter) and summer periods is clearly visible. The variability around the seasonal mean is captured by the other modes (Figure 17 c-e & Figure 18 c-e). Mode 2 refers to the inter-annual and the intra-seasonal variability in the shoreward gradient and represents 7% of the total variance. Mode 3 addresses some North-South gradients and represents 4% of the total variance and mode 4 is clearly influenced by the Gironde river (Figure 17d), which brings sediments during water outflow,

and represent 3% of the variance. By construction, EOF decomposition imposes the orthogonality [9] of the spatial modes (Figure 17).

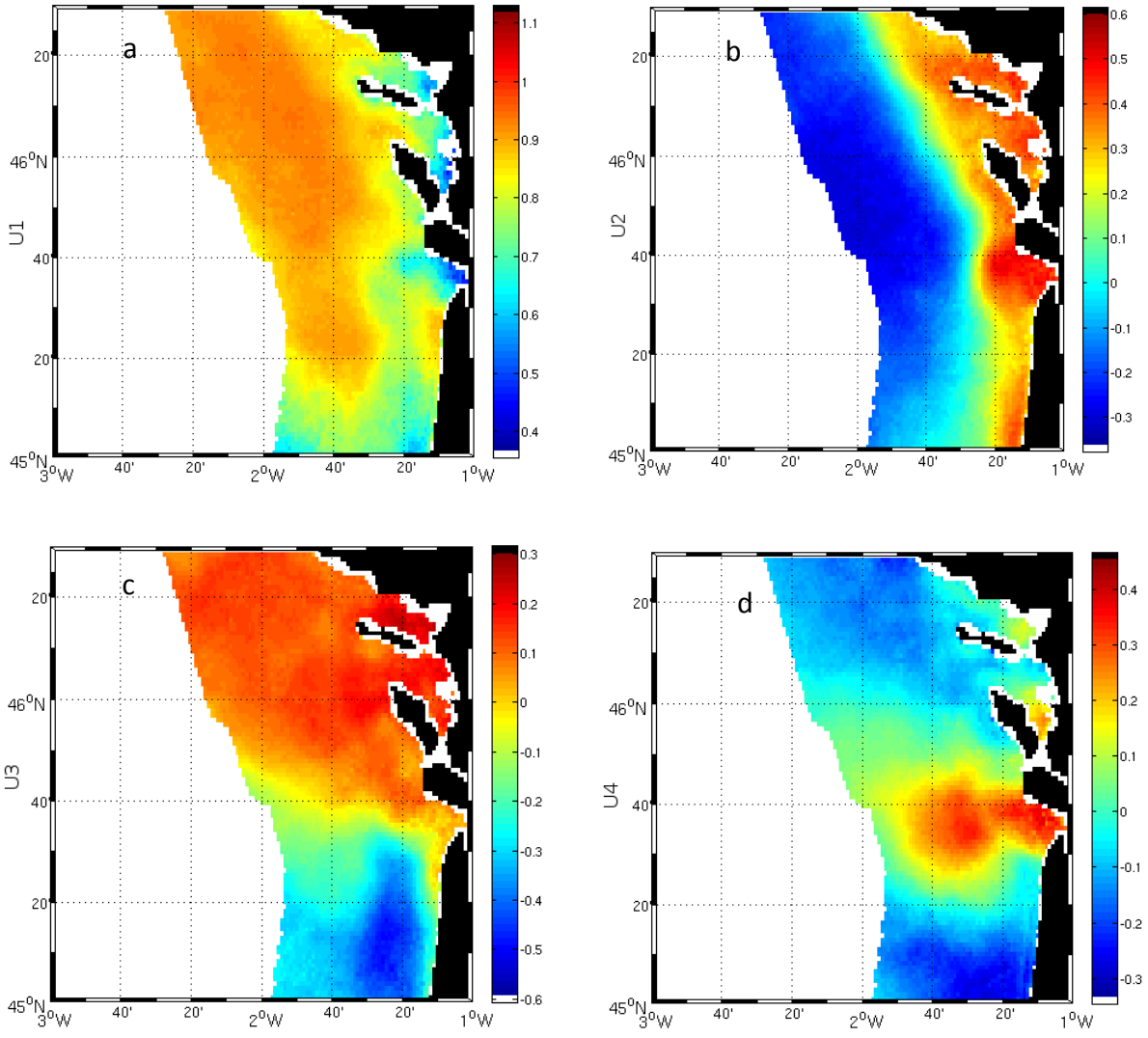


Figure 17: spatial modes of the EOF decomposition of the SPIM observed from satellite from 2007-2009 in the Gironde mouth river. From left to right and top to bottom the first four EOF modes account respectively for 85, 7, 4 and 3% of the total variance.

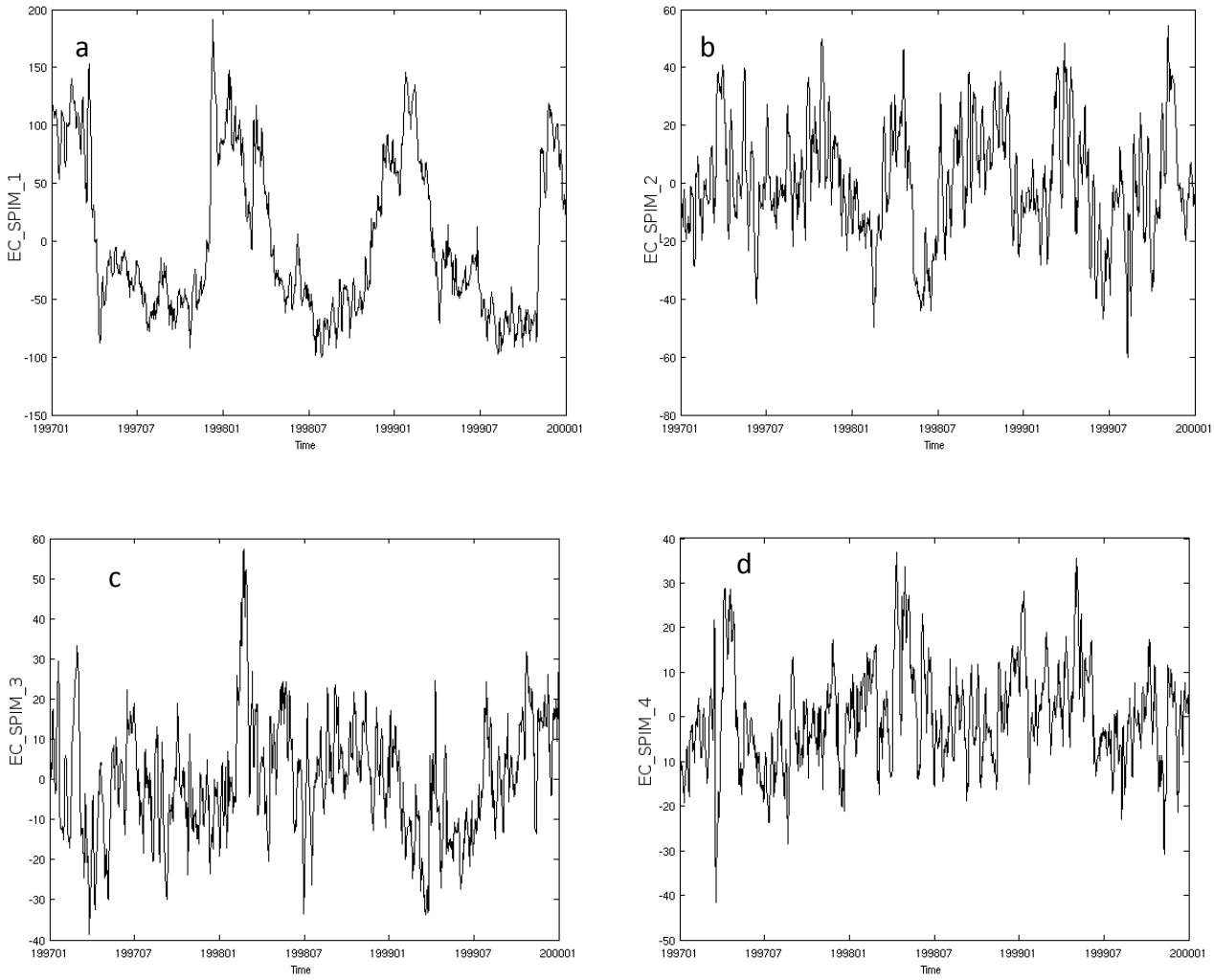


Figure 18: EOF decomposition of the SPIM observed from satellite from 2007-2009 in the Gironde mouth river: from left to right and top to bottom, the expansion coefficients ($EC\_SPIM_{1\text{-}4}$) associated with the first four EOF modes depicted in Figure 17, i.e. the time evolution of the spatial modes.

The reconstruction of the SPIM variable from the estimated ECs is performed as:

$$\widehat{\mathrm{SPIM}}_t = \sum_k \widehat{\mathrm{EC\_SPIM}}_{\mathrm{k,t}} \,.\, U_k \qquad (55)$$

The total explained variance using the 4 first modes is shown Figure 19b. On average, the explained variance represents 99 % of the total variability on the areas with some local minima of 60% observed at the very near-shore and the Southwestern part of the area.

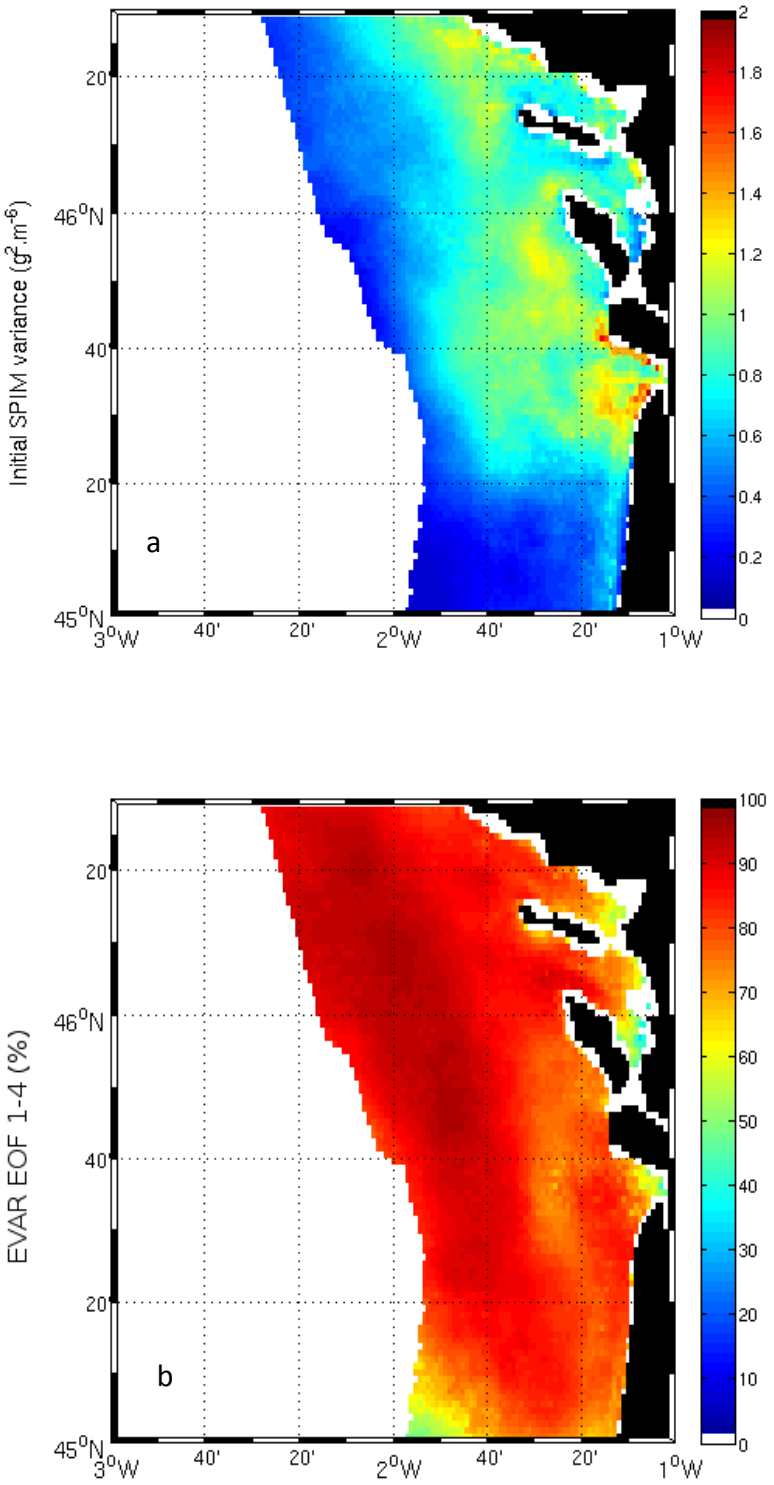


Figure 19: a) initial SPIM variance. b) Percentage of variance explained by the four first modes of the EOF decomposition of the suspended particulate matters.

4.3.2 Predictors and covariates

The predictors X are the variables used in the estimation of Y and Z any time, Eq. (47)&(48). We used here wave height (WH) daily means of the Wave Watch 3 model (WW3; [151, 164]) provided by the IOWAGA and PREVIMER programs, eastward and northward winds interpolated from QuickSCAT and ASCAT observations in conjunction with ECMWF forecasting [152], provided by Ifremer, tide index (SHOM, 2000) at Bordeaux and the flow measurement of river la Gironde. Similarly to the SPIM data, all the data were log transformed. For the wind data which is signed, the transformed log variable was signed negatively a posteriori to the log transformation. The WH first mode of the EOF decomposition explained 98 % of the total variance, 93% for the Northern wind (WND1), and 96% for the eastward wind (WND2).

The choice of the predictors is performed here as follows. We first select as predictors the variable showing a significant correlation with the studied EC. Given these predictor datasets, we tested all the possible configurations and chose the predictors which provide the lower BIC and the greatest EVAR on the training dataset. Covariates are the normalized predictors used in the estimation of the EC but considered at t-2. In the same way, this time-lag has been estimated as the optimal lag using BIC and EVAR results on the training dataset.

4.4 Results

We summarize in Table 1 the prediction performance for the first four ECs of the SPIM issued from four models: HMM, NHHMM, HMM-AR, and NHHMM-AR. The number of considered modes for the mixture varies from 1 to 3. The one-mode models refer to a simple multivariate regression analysis. For each configuration we provide the BIC and EVAR_train on the training dataset (2007-2009) and EVAR_valid on the validation dataset (2010). Note that the selection of the predictors and resulting covariates is completed as a prior step as described in §4.2.4.

The first mode of the EOF decomposition explains 85% of the total variance. EC_WH_1 and EC_WND2_1 (respectively the expansion coefficient of the first EOF of the eastward winds) are identified as being the relevant predictors (cf. §4.2). This mode captures the mean seasonal variability of the SPIM, which is mainly driven by WH of the North Atlantic storms and at a second order by the eastward winds. For EC_SPIM_1, when no autocorrelation term is used, the best fit is obtained for a 3-regime NHHMM model (BIC= 9873, EVAR_train=90% and EVAR_valid=85%). When a first order autocorrelation term is added, the 3-regime HMM-AR and NHHMM-AR models show the best results: BIC= 7997 (resp. 8000), EVAR_train = 98% and EVAR_valid = 97%. This stresses that when observations are available first-order autoregressive term (AR_1) should be included to enhance the performances. The lag-1 autocorrelation observed value is 0.85 for EC_SPIM_1, underlying the strong link between two successive observations. Compared to a single

autoregressive model, i.e. without predictors and covariates and regime discretization (not shown), the gain value provided by X and S on EVAR_train and EVAR_tvalid is of 15%.

The second mode of the EOF decomposition of the SPIM variability explains 7% of the total variance. The selected predictors are the first mode of the eastward wind, the tide, and the river flow. The variability captured by EC_SPIM$_2$ relates to the local eastward wind, which is not captured by the WH model, and the very coastal variability introduced by the tide and the river outflow. For the non-AR models the selected model was the three-regime NHHMM. It is interesting to note in this case that EVAR_valid increased from 50% to 73% between the HMM and the NHHMM, highlighting the contribution of the non-homogeneous transition model.

Table 5: Model performance for each EOF Expansion Coefficient (EC) of the SPIM variability. For each configuration we report the BIC (a) and the explained variance (EVAR_train, b) for the training dataset (2007-2009), and the explained variance (EVAR_valid, c) for the validation dataset (2010). In bold are highlighted for each EC the selected configurations (see § 5.2).

EC_SPIM	Number of modes, M					
	1	2	3	1	2	3
1	HMM (a) 1183 (b) 37 (c) 32	HMM 10037 84 70	HMM 9874 85 75	HMM-AR 8157 92 91	HMM-AR 7986 95 93	HMM-AR 7997 98 97
	NHHMM 11184 37 34	NHHMM 10037 84 71	NHHMM 9873 90 85	NHHMM-AR 8171 92 90	NHHMM-AR 7994 92 94	NHHMM-AR **8000** 98 97
2	HMM 9403 18 12	HMM 8579 67 33	HMM 8129 76 50	HMM-AR 7167 90 87	HMM-AR 7098 91 89	HMM-AR 7075 92 91
	NHHMM 9451 18 12	NHHMM 8614 67 44	NHHMM 8152 79 73	NHHMM-AR 7188 89 88	NHHMM-AR 7383 90 87	NHHMM-AR **7070** 92 91
3	HMM 8840 12 7	HMM 8222 57 44	HMM 7844 68 72	HMM-AR 6723 85 84	HMM-AR 6632 86 91	HMM-AR 6630 88 92
	NHHMM 8866 11 16	NHHMM 8246 59 45	NHHMM 7862 75 76	NHHMM-AR 6745 88 86	NHHMM-AR 6673 88 91	NHHMM-AR **6633** 88 92

4	HMM 8248 18 28	HMM 7596 60 63	HMM 7285 71 72	HMM –AR 6398 85 86	HMM -AR 6416 85 86	HMM -AR 6313 86 86
	NHHMM 8276 18 28	NHHMM 7628 62 59	NHHMM 7267 70 75	NHHMM-AR 6426 85 83	NHHMM-AR 6445 85 83	NHHMM-AR **6314** 86 85

The third mode of the EOF decomposition of the SPIM variability explains 4% of the total variance. It captures some inter-annual and intra-seasonal variability of the latitudinal gradient of the SPIM. The selected predictors are EC_WH_1, EC_WND1_1 (northward) and the tide. Once again, three-regime NHHMM and HMM-AR provide the best results.

Regarding the fourth mode of the EOF decomposition of the SPIM variability, which accounts 3% of the total variance, EC_WH_1, EC_WND2_1, the tide and the river flow are selected as contributive predictors. We reconstruct 75 % of EC_SPIM_3 variance of the validation dataset using a three-regime NHHMM and 92% using the three-regime HMM-AR and NHHMM-AR .

Globally, we observe from Table 3 that three regimes are needed for all models to forecast optimally the EOF ECs at t+1. NHHMM outperforms HMM for forecasting results at t+1. The inclusion of an AR term clearly improves the results. NHHMM-AR and HMM-AR shows similar results at t+1. We will see (Table 7) that the added value of non-homogeneous transitions for AR-model clearly appears for the long term forecasting.

4.4.1 Example with the estimation of EC_SPIM_1

We report in Figure 20 the temporal evolution of the three regimes of the NHHMM for $\widehat{EC\_SPIM1}$ estimated at t+1. In table 2 are shown the corresponding coefficients for each predictor and the intercept. The first regime (light grey, Z_t=1), characterized by high SPIM levels (intercept of 65), is referred as a 'winter regime'. The' winter regime' also strongly relates to the wave height (WH regression coefficient of 0.6). Dark grey periods (Z_t=3) are identified as a 'transition regime', and medium grey (Z_t=2) identified as the 'summer regime'. From the 'winter' to the 'summer' regime, coefficient for WH decreases from 0.6 to 0.12. In summer the energy brought by waves is not sufficient enough to re-suspend massively the sediments. It might be noticed that for all regimes the wind conditions show a small but significant effect on $\widehat{EC\_SPIM1}$. When an autocorrelation term is added (HMM-AR, table 2), the AR(1) coefficient value is 0.86 for 'winter' regime and 0.9 for 'summer' regime which underlies that during calm periods the SPIM concentration remains low.

Figure 20 compares the prediction of $\widehat{EC\_SPIM1}$ using a single multivariate regression (green) and the proposed multi-regime NHHMM. In this case the explained variance value (Table 1) is of 37% for the multivariate regression model compared to 90% for the three-regime NHHMM.

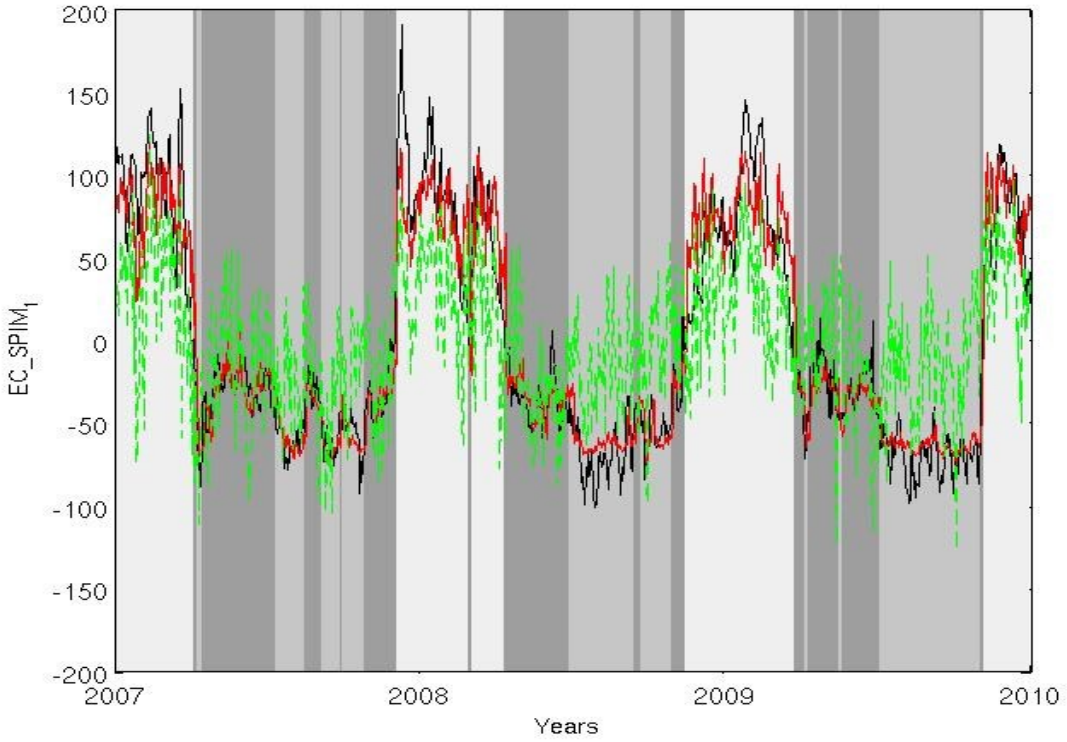


Figure 20: Estimation of the EC_SPIM$_1$ (in black) using EC_WH$_1$, EC_WND2$_1$ and a single regression (green) and a 3 regime NHHMM (red). The nuances of grey in the background highlight the temporal distribution of the regimes (1, light grey; 2, medium grey; 3 dark grey).

Table 6: Estimated regression parameters for each of the three regimes of the NHHMM and the HMM-AR for the first EOF EC of the SPIM: regression parameters involve an intercept and the regression coefficients of the significant forcing parameters i.e. the wave height and the eastward wind velocity.

	EC_WH$_1$	EC_WND$_2$	Intercept	
NHHMM	(1, light grey, winter) 0.6037 (2, dark grey, summer) 0.0910 (3, medium grey, transition) 0.1210	-0.0632 0.0006 0.0100	65.0672 -61.6442 -24.6578	
	EC_WH$_1$	EC_ WND$_2$	Intercept	AR(1)
HMM-AR	(1) 0.2383 (2) -0.0050 (3) 0.0354	-0.0033 0.0168 0.0035	4.4694 -3.9531 0.6079	0.86 0.90 0.90

Figure 21 illustrates the non-homogeneous transition probability used in the NHHMM between the 'transition' and 'winter' regimes as a function of the normalized values of EC_WH$_1$ (eastward swell) and EC_WND2$_1$ (eastward wind). The probability of switching from regime 3 to 1 increases with wave height and eastward winds normalized covariates. When the probability is greater than 0.5 the regime changes from 3 to 1.

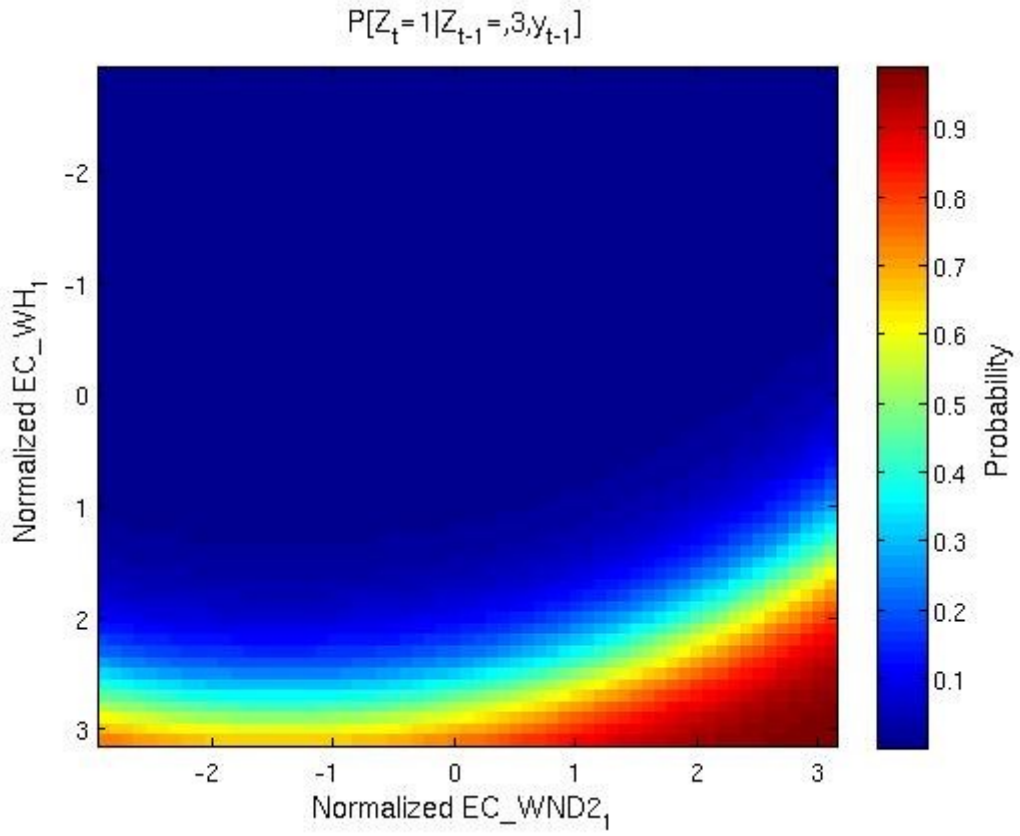


Figure 21 : Non-homogeneous transition between 'transition regime' (medium grey Figure 20) and 'winter regime' (light grey Figure 20) as a function of the normalized wave height WH_1 and eastward wind WND_2 covariates.

We forecast SPIM fields from the reconstructed $\widehat{EC}$s, Eq.(47), (48), (53). Figure 22a&b compare the explained variance of the initial field (SPIM) using the three-regime NHHMM and NHHMM-AR models. On average we were able to predict at t+1 80% of the variance using the NHHMM (Figure 22a) and 93% using the NHHMM-AR. The spatial distribution of the error is not homogeneous. Figure 22 shows that EVAR_valid value is of 90% in the Northern part with nevertheless poorer results in the South. Figure 22b shows that the AR_1 component of the model increases EVAR for the whole area.

We also consider the results of a standard multi-regression analysis. If only one regime is considered NHHMM and HMM resort to a standard multivariate regression and NHHMM-AR and HMM-AR to a standard multivariate regression including an AR_1 coefficient, the transition probability being equal to 1. Figure 22c shows the results obtained with the standard multivariate regression and Figure 22d the standard multivariate regression including an AR_1. From Figure 22c to Figure 22a, the gain in explained variance is in mean of 150% (from in mean 32% Figure 22c to 80% Figure 22a) while for the AR models, the gain is of 11% (from in mean 83% Figure 22d to 93% Figure 22b).

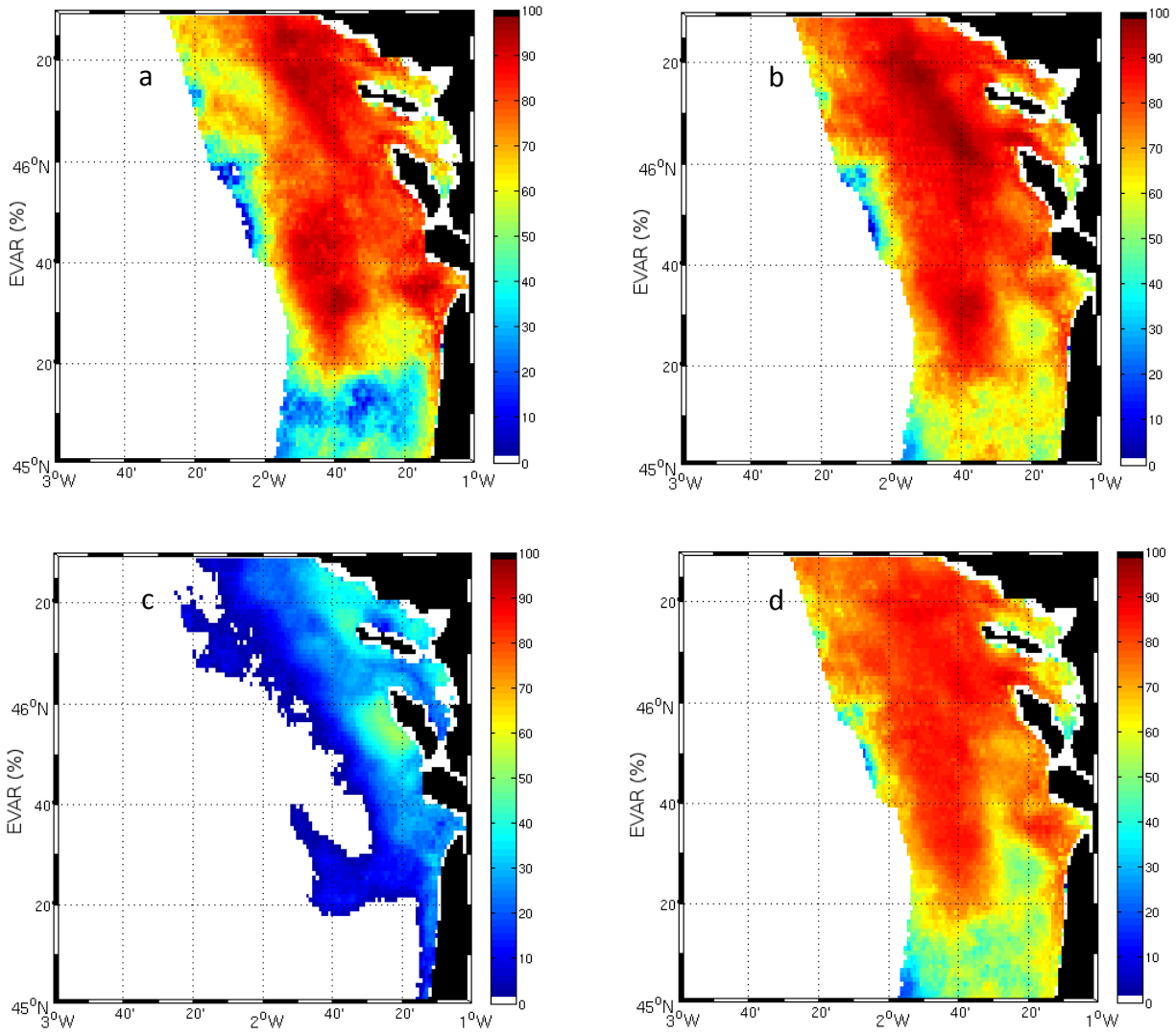


Figure 22: Explained variance for the 2010 validation dataset reconstructed using the 3-regime NHHMM (a) and NHHMM-AR (b), compared to a standard multivariate regression without AR_1 (c) and including an AR_1 (d).

Regarding the model forecasting performances, we report the short-term forecast results at different time steps using the 2010 validation dataset. Table 3 synthesizes the explained variance statistics using 3 regimes and the four tested models for the forecasting at t+1, t+5 and t+15.

The long term forecasting results are globally better with the NHHMM-AR. At t+15 using the NHHMM we are able to forecast 74% of the variance for 2010, compared to 40% for the HMM. In this case the time-varying regime transition probability $P(Z_t = k|Z_{t-1} = l, \mathrm{S_{t+dt}})$ helps in the estimation of $\widehat{\mathrm{Y}}_{\mathrm{t+dt}}$ (covariates S_t^{t+dt} are model outputs for which the short term predictions are assumed to be available). For autoregressive models, at dt=5, we were able to forecast 82% of the 2010 SPIM variance with the NHHMM-AR compared to 77% with the NHMM. In this case $\widehat{\mathrm{Y}}_{\mathrm{t+1}}^{\mathrm{t+dt}-1}$, estimated using $\mathrm{X}_{\mathrm{t+1}}^{\mathrm{t+dt}-1}$, $\mathrm{Y_t}$, and the inhomogeneous transition properties, help to estimate $\widehat{\mathrm{Y}}_{\mathrm{t+dt}}$. At t+15 NHHMM and NHHMM-AR show equivalent results underlying the maximal time-step for which the autoregressive term brings significant information.

Table 7: Validation results on year 2010. Explained variance, Eq. (50), for the forecast at t+1, t+5 and t+15 of the 2010 validation dataset. For each model, three latent-regimes are used.

dt (days)	EVAR for the 2010 validation dataset			
	HMM	HMM-AR	NHHMM	NHHMM-AR
1	73	93	80	93
5	63	80	77	82
15	40	70	74	75

A SVR model was also evaluated to evaluate the performances of a non-linear model on the studied dataset. To perform the comparison, we train the SVR model (http://www.csie.ntu.edu.tw) for each EC using the same training dataset (2007-2009) and performed forecasting using the same validation dataset (2010). We used the setting as following: model epsilon-SVR (s=3), linear or polynomial kernel (t=0 or 1) and the same inputs (predictors, covariates) for each EC. Parameters c and g [139] were optimized for each EC using the training dataset and the cross validation mode. On the 2010 validation dataset, the best forecasting results reached 40% at t+1 of the EVAR (without AR) and 85% with an AR coefficient. The results were significantly worse than those obtained using the time-varying models for increasing time steps. The SVR can address non-linear relationships. Nevertheless, it cannot deal with multi-regime processes. By contrast, the latent-regime model addresses by nature multi-regime processes and can approximate non-linear relationships as a series of linear models.

4.5 Discussion

We investigated the relevance of four regime-switching latent regression models, namely HMM, NHHMM, HMM-AR and NHHMM-AR to characterize time-varying linear relationships between the high-resolution SPIM data (inorganic suspended matter concentration) and forcing conditions i.e. the wave height, the northward and eastward winds, the tide and the river flow. SPIM data were issued from MODIS, SeaWiFS and MERIS satellite data. As a case study, we considered a coastal area in the mouth of the Gironde estuary in the [3°W-1°E; 45-46.5°N] area. Model calibration was carried out using 2007-to-2009 datasets, whereas 2010 dataset was used as an independent validation dataset of 1-to-15-day forecasting performances.

An optimal number of three regimes were identified to capture the different geophysical dynamics and optimize forecasting performances. Autoregressive and non-homogeneous models showed better performances. For the 2010 validation dataset, one-day NHHMM forecasts explained 80 %

of the variance, whereas a NHHMM-AR model explained 93 % of the variance. The natural high autocorrelation level observed in geophysical time series makes the observation of the previous day an important predictor to consider. The comparison to other models clearly stresses the relevance of the proposed latent-class models. Whereas the explained variance of one-day forecast for a standard multivariate linear regression was of 32% (resp. 83%) without (resp. with) an first-order auto-regressive term, the non-linear SVR model reached respectively, 40% and 80% of explained variance. The gain of 100% between the NHHMM and the SVR model (resp. 16% between the NHHMM-AR and the SVR including an AR_1 term) pointed out the relevance of the multi-regime approaches. The SVR model failed here in retrieving regime shifts.

As illustrated for the first SPIM EOF component (Figure 20), the proposed multi-regime setting identified three different relationships between the observed turbidity, the wave height and the wind. We did not drive the model to account for seasonal regimes but these regimes exhibited seasonally-discriminated patterns, with two leading factors: the mean SPIM level (intercept) and the wave height. The later was interpreted as a feature of the minimum of energy to be brought by the swell to re-suspend the sediments. This is regarded as a key characteristic of the latent-regime model compared to other non-linear regression models, such as Neural Networks [54] or SVR [139], which can hardly be interpreted in general.

Regarding long-term forecast performance, at t+15 best results obtained were of 74% of explained variance for the NHHMM and 75% for the NHHMM-AR. For short period, typically from 1 to 15 days, when the observed Y is not available, NHHMM-AR provided the best results. In this case the available predictors X_t^{t+dt}, covariates S_t^{t+dt} and the estimated $\widehat{\mathrm{Y}}_{t+1}^{t+dt-1}$ help in the estimation of $\widehat{\mathrm{Y}}^{t+dt}$. At t+15 NHHMM and NHHMM-AR showed similar results. Hence, a 15-day period could be regarded as the maximal time interval, beyond which one may only consider covariates. It may also be noted that, in case of sensor failure and/or long missing data periods (e.g., series of successive storms in the case-study region), though no satellite observations might be available, one could still reach relevant SPIM prediction accounting in average for about 75% of the variance.

In the future, we will address the forecasting of the chlorophyll-a using satellite-derived observations such as the photosynthetic available radiation, the temperature, the suspended matters (as index of available nutrients) and light attenuation [91]. In this more complicated case, second order relationships between the variable and its predictors have to be evaluated, the chlorophyll-a dynamic being not anymore a passive result of the forcing conditions, as expected with the SPIM, but having its proper characteristics depending on each phytoplankton specie. Extensions of the considered latent regime setting to other inverse problems in satellite sensing data analysis are also under investigation, such as latent regime inversion procedures for satellite-derived chlorophyll-a concentration to account for different water types (turbid or not turbid) and/or the presence of specific phytoplankton species.

5 Chapter 5: Ocean Color Atmospheric corrections in coastal complex waters using a Bayesian latent class model and potential for the incoming Sentinel 3 - OLCI mission.

This last chapter details our research on the enhancement of sea surface reflectance estimates in coastal areas. From the TOA observations, atmospheric correction aims to separately distinguish the atmosphere and the water contribution [20]. From a methodological point of view, our approach is based on characterization of prior modes in the joint distribution of the observed variable and covariates using Gaussian Mixture Models (GMM). Here, covariates are observed geophysical parameters, in this case the geometry acquisition conditions and pre-estimates of the reflectance in the near Infrared part of the spectrum, significantly correlated with the variable of interest.

The GMM prior modes characterize reference spectra of both aerosol and water reflectances. A reference spectrum for the aerosol characterizes the specific signature of the aerosols on the observed aerosol reflectance. A reference spectrum for the water characterizes the specific signature of chlorophyll-a, suspended particulate matters and colored dissolved organic matters on the observed sea surface reflectance.

The GMM prior modes are then used to optimize the inversion of sea surface reflectance from MERIS top-of-atmosphere observations. For that purpose, prior distributions of the marine and aerosol variables are corrected using the observed values of the covariates to optimize the 100 random initializations for our MEETC2 algorithm [13].

This chapter was submitted in October 2014 to the 'Remote Sensing of Environment Journal (RSE)' [13].

5.1 Introduction

The inversion of Ocean Color signal in coastal areas from top-of-atmosphere (TOA) measurements remains a scientific challenge. This is a crucial point for the ocean color community as many governmental policies such as the European Water Framework directive (WFD) rely on estimation of coastal water quality, itself possibly derived from space-based ocean-color measurements [165]. Hence, ocean color inversion is certainly among highest priority research topics for ocean–color community in the incoming years. Different aspects may explain the difficulties encountered in this inversion process. Firstly, the contribution of suspended matters to the reflectance in the near infrared (700-900 nm) is an issue as many algorithms expect these reflectances to be null.

This assumption is called the black pixel hypothesis [20] and relies on the strong natural absorption of the water in this domain [20]. Secondly, bio-optical modelling, i.e. the estimation of the water-leaving reflectance from the Inherent Optical Properties (IOPs, namely the absorption and backscattering of the sea water constituents) in complex coastal waters is also an issue. Despite accurate physical models exist for open clear waters [166] that cover 85% of the oceans, their derivation for coastal waters is more complex [17].

As a consequence, available operational standard level 2 reflectance products [167] may perform poorly in coastal areas, and consequently these products are often flagged as anomalous values [167] for such areas. Reflectances in the blue and green bands are often underestimated and may involve physically-meaningless negative values. Obviously [169], this strongly affects relevance of level-2 products for the end users, which typically use water-reflectance spectra as inputs to estimate the chlorophyll-a and the suspended particulate matter concentrations (SPM, [168]), or the vertical light attenuation (K_{dPAR} [91], K_{d490} [65]).

Over the last fifteen years, many regional algorithms have been developed to address user's needs for reliable water-reflectance data in coastal areas. Among them, the MERIS Case2-Regional (C2R, [54]) based on a non-linear learning machine model, namely a Neural Networks (NN) [138, 54], estimates water reflectance [54] over turbid areas. The learning paradigm relies on the calibration of a non-linear model to relate the available satellite-derived observations to the geophysical quantity of interest from a training dataset. This training dataset typically consists of a collection of in-situ measurements along with the satellite-derived measurements. This learning-based strategy may suffer from two major drawbacks: weak geophysical/biological interpretability of this 'black-box' model and the assumption on representativity of the training dataset. They may restrict the applicability of the model to a specific region and questions its validity with respect to the generally unknown variability of the atmospheric and water conditions.

Here, we develop a Bayesian latent class approach to address these limitations. The key feature of our model is the assumption that TOA-ocean-color relationships may not be well represented by a single model, linear or not, but are characterized by multiple and local relationships, the global inversion being addressed using a mixture of these identified elementary relationships. To our knowledge, Bayesian model mixtures have been seldom explored for ocean color inversion [171]. The proposed Baysian latent class models allow unmixing the diversity of TOA-ocean-color relationships, while keeping geophysical interpretability of the model. Such model involves training, in the same spirit as leaning machines. This training phase is necessary to estimate the model parameters, and may be completed using in-situ data [53] or simulated data using for example radiative transfer simulations [172]. Conversely to the machine learning approaches (NN, or Single Values Regressions, SVR [173]), each latent class model is a linear model which may be easily linked to existing and interpretable physical processes, namely here the coastal aerosol and the water types.

Our inversion scheme (MEETC2) estimates water reflectances in complex waters from the MEdium Resolution Imaging Spectrometer (MERIS) TOA observations. Model calibration and validation involve here the MEris MAtchup In-situ Database (MERMAID) radiometric in-situ dataset [53]. Quantitative comparisons with the standard MEGS v8 and the MERIS C2R Neural Network outputs [54] clearly demonstrate the relevance of our approach. We further discuss the potential of MEETC2 for the incoming OLCI / Sentinel 3 mission that should be launched in 2015.

5.2 Review of the standard Ocean Color inversion method

5.2.1 Atmospheric correction principles

Ocean-color sensor measures at TOA the upwelling radiance (L_u) in $mW.m^{-2}.sr^{-1}$ backscattered by the ocean-atmosphere system. This radiance originates from photons scattered by air molecules and/or aerosols, which may also have been reflected directly at the sea surface (glint effect, [18, 19]), and may potentially have penetrated in the ocean. The TOA measured reflectance (ρ_{TOA}) is the ratio between the upwelling irradiance L_u and the downwelling irradiance (E_d), i.e. L_u integrated over the solid angle [0;2π]. The water reflectance contribution measured at TOA, i.e. transmitted through the atmosphere, represents at maximum 10% of the signal. This low signal/noise ratio stresses the resulting difficulties to unmix the atmospheric contribution from the water one. The traditional signal decomposition [20] expresses measured TOA reflectance for each wavelength λ as a sum of elementary contributions:

$$\rho_{gc}(\lambda) = \rho_{Ray}(\lambda) + \rho_{aer}(\lambda) + t_d(\lambda).\rho_w(\lambda) + \varepsilon \quad (56)$$

where ρ_{GC} is the ρ_{TOA} (observations) corrected from the glint [20] and gaseous absorption [20], ρ_{Ray} (known) the reflectance of a purely molecular atmosphere (no aerosol) [175], ρ_{aer} (unknown) the reflectance of the aerosols including the coupling term between air and aerosol molecules [20], t_d (unknown) the diffuse transmittance of the atmosphere, ρ_w (unknown) the water reflectance which is the principle quantity to inverse. ε is a noise process i.e. with a normal distribution $N(0, \sigma^2)$. We consider here the Rayleigh corrected reflectance variable $\rho_{RC}(\lambda)$:

$$\rho_{RC}(\lambda) = \rho_{gC}(\lambda) - \rho_{Ray}(\lambda) = \rho_{aer}(\lambda) + t_d(\lambda).\rho_w(\lambda) + \varepsilon \quad (57)$$

The diffuse transmittance t_d is the product of both air molecules and aerosol particles scattering:

$$t_d(\lambda) = e^{-(0.5.\tau_{ray}(\lambda)+(1-w_a(\lambda).F_a(\lambda).\tau_a(\lambda)).M} \quad (58)$$

where $\tau_{ray}(\lambda)$ is the Rayleigh optical thickness (known), $\tau_a(\lambda)$ is the aerosol optical thickness (unknown), M the air mass factor (known), Wa the aerosol single scattering albedo (known), Fa

the forward probability scattering (known) [167]. τ_a is linked with the estimated aerosol reflectance for primary scattering [167]:

$$\rho_{aer}(\lambda) = \frac{Px.Wa(\lambda)}{4(\cos(\Theta s) + \cos(\Theta v))}(1 - e^{-\tau a(\lambda)M}) \quad (59)$$

where $Px.Wa$ is the aerosol phase function (known) times the single scattering albedo for the current scattering angle [167], Θs and Θv are respectively the sun and the view zenith angles (known). Eq. (59 is used to express $\tau a(\lambda)$ as a function of $\rho_{aer}(\lambda)$ to estimate the transmittance in Eq. (58).

Whereas, in open ocean waters, one can exploit null contribution of water reflectance in the near infrared (NIR) range to infer aerosol contributions, no such simple inversion scheme applies in coastal waters, which are characterized by a non-null contribution in this domain [176]. This is a major issue to be dealt with in the atmospheric corrections in coastal waters. For a fixed geometry, aerosols contributions are often assumed to follow an exponential decay [177]:

$$\rho_{aer}(\lambda) = \rho_{aer}(\lambda_0)e^{c(\lambda-\lambda_0)} \quad (60)$$

where $\lambda_0 = 865$ nm and c is the exponential decay of the aerosol spectrum, i.e. representative of the aerosol type. Thought relevant in the NIR domain, the assumption of an exponential decay appears to be too restrictive in the 400-700 nm range where multiple scattering between aerosol and air molecules may become significant [20]. Following [178], a polynomial model is considered to provide a more general model of aerosol contributions. Using our training dataset (cf § 5.5) a polynomial of order 3 was found as relevant to estimate the aerosol contributions:

$$\rho_{aer}(\lambda) = \rho_{aer}(\lambda_0) + a_1(\lambda-\lambda_0) + a_2(\lambda-\lambda_0)^2 + a_3(\lambda-\lambda_0)^3 \quad (61)$$

5.3 The standard processing atmospheric correction scheme

In the standard Level 2 processing of MERIS, MODIS and SeaWifs, the following four-step scheme is applied to estimate the water-leaving reflectances [20]:

- The signal is corrected from absorbing gaseous such as ozone, oxygen, water vapor and nitrogen dioxide.
- The estimated contribution of suspended matter particles in the NIR is removed from TOA observations after single scattering transmittance through the atmosphere. This step is known as the Bright Pixel Atmospheric Correction (BPAC) and detailed in the next section.
- A mixture of two aerosol models among 34 (for MERIS) is estimated from the values of ratio $\rho_{path} = \rho_{gc} / \rho_{ray}$ (Eq.57) at 779 and 865 nm, leading to the estimation of both the aerosol reflectance $\rho_{aer}(\lambda)$ and the multiple scattering transmittance $t_d(\lambda)$ (Eq.58).

- Water reflectance contribution is estimated by subtracting the estimated aerosol contribution from $\rho_{aer}(\lambda)$ using Eq. 57.

The Bright Pixel Atmospheric Correction (BPAC)

BPAC [195] is an iterative algorithm for pre-correction of TOA signal in the NIR. It aims at removing the water contribution, caused by suspended matters, of the TOA observed reflectance. This step is essential in the standard processing as the estimation of the aerosols is performed using the NIR bands under the assumption $\rho_w(NIR)$ =0. Moore [195] proposed for MERIS a two steps algorithm which iterates: the estimation of $\rho_{aer}(709,865)$, c and $\rho_{aer}(779)$ using $\rho_{path}(779,865)$, then, using the estimated residuals $\hat{\rho}_w$ in the NIR from Eq.57 and a parametric model, the estimation of the SPIM concentration and related $\hat{\rho}_w$ at TOA.

This converging algorithm suffers actually from drawbacks for very turbid waters where the used water model does not allow retrieving high concentrations of SPM. It typically leads in these areas to an over correction of the blue water-reflectance, i.e. an underestimation of ρ_w at 412 and 442 nm with the standard Level 2 processing, and may resort to geophysically-meaningless negative reflectance values.

5.4 Method

5.4.1 Spectral representations of the water contributions using Non-Negative Matrix Factorization

Given the spectral overlap of water and aerosol contributions in (Eq. 56 & 57), inversion of (Eq.1) requires some prior knowledge on water contributions. We propose here to determine from the training dataset [53] a parametric spectral representation of water contributions. We use here a Non-Negative Matrix Factorisation (NNMF) with projected gradients [180]. Similarly to PCA, it relies on additive decomposition of a water spectrum on a basis learnt from the data. In contrast to PCA, it does not involve orthogonality constraints but imposes non-negativity for both the basis function and the projection coefficients. NNMF is among most popular approach in multispectral and hyperspectral remote sensing [181], as a mean to unmix contributions issued from various sources in a sensed environment. Formally, NNMF leads to the following parametric representation of a given water spectrum $\rho_w(\lambda)$:

$$\rho_w(\lambda) = W(\lambda, n) * h(n) \quad (62)$$

where W>0 is a λxn matrix whose each column contains a reference water type spectra identified by NNMF using the training data, and h(n)>0 refer to the nx1 vector of coordinates of the

spectrum $\rho_w(\lambda)$ in the decomposition space. It may be noticed that NNMF decomposition could also be replaced here by a bio-optical model [166]. Nevertheless, to our knowledge none of this model is today performant enough to estimate, in coastal areas, the water leaving reflectance spectrum from the water's constituents. The NNMF decomposition, by imposing non-negativity of both the coordinates and the reference radiometric water shapes also appropriately constrains our inversion to converge toward physically realistic solutions (cf § 5.6.3), conversely to the standard Level 2 processing (ESA and NASA).

5.4.2 Bayesian Formalism

From Eq.56, the variables to be estimated are $x_w = \{h_i\}$, i.e. the coordinates of ρ_w in the basis W (Eq. 62), and $x_a = \{a_i\}$ i.e. the polynomial coefficients of the aerosol models (Eq.61). Conversely to standard least square fitting or minimization procedure such as Levenberg-Marquard [183], minimization relies not only on the likelihood of ρ_{RC} but also on the prior distributions of x_a and x_w. We consider the Maximum A Posteriori estimation (MAP) [184] which aims at maximizing the conditional probability $P(\mathrm{x_a}, \mathrm{x_w}|\rho_{RC}, \varphi)$:

$$P(\mathrm{x_a}, \mathrm{x_w}|\rho_{RC}, \varphi)\ \alpha\ P(\rho_{RC}|\mathrm{x_a}, \mathrm{x_w}, \varphi)\ .P(\mathrm{x_a}, \mathrm{x_w}|\ \varphi)$$

We suppose here that $\mathrm{x_a}$ *and* $\mathrm{x_w}$ are independent i.e.: (63)

$$P(\mathrm{x_a}, \mathrm{x_w}|\rho_{RC}, \varphi)\ \alpha\ P(\rho_{RC}|\mathrm{x_a}, \mathrm{x_w}, \varphi)\ .P(\mathrm{x_a}|\ \varphi).P(\ \mathrm{x_w}|\ \varphi)$$

In the proposed framework, $\mathrm{P}(\rho_{RC}\ |\ \mathrm{x_a}, \mathrm{x_w}, \varphi)$ is modeled using a multivariate normal distribution (MVN) with a null vector, $\mu 0_{\rho RC}(\lambda)$ and a full covariance matrix $\Sigma 0_{\rho RC}(\lambda)$. As detailed in the next sections, $\mathrm{P}(\mathrm{x_a}|\ \varphi)$ and $\mathrm{P}(\mathrm{x_w}|\ \varphi)$ are modeled using a mixture of MVN distributions, namely a Gaussian Mixture Models (GMM, [186]). $\varphi = \{\mu 0_{\rho RC}, \Sigma 0_{\rho RC}, \mu 0_{Xwi}, \Sigma 0_{Xw}, \mu 0_{Xa}, \Sigma 0_{Xa}\}$ is the vector of hyperparameters to be estimated.

The MAP criterion cost function is finally expressed as:

$$C = -\log(P(\mathrm{x_a}, \mathrm{x_w}|\rho_{RC}, \varphi)) \quad (64)$$

During the inversion and knowing all parameters of the Bayesian model, the numerical maximization of the MAP criterion (Eq. 64) is completed using the Sequential Quadratic Programming algorithm (SQP) gradient-based descent algorithm [185].

5.4.2.1 Covariates and prior distributions

5.4.2.2 Choice of the covariates

Covariates are here geophysical parameters significantly correlated with the variable of interest. From a physical point of view, the observed shape of aerosol reflectance spectrum $\rho_{aer}(\lambda)$, i.e. a_i coefficients of Eq.61, is correlated (cf § 5.6.1.1) with: the variables which describe geometry of acquisition conditions (Θs, the sun zenith angle, Θv, the view zenith angle, and δψ, the delta azimuth [20]), the variables $\rho_{aer}(865)$ and c (Eq. 5) estimated using the NIR part of the spectrum dutring the BPAC step. For these reasons these latest variables are referred here to covariates.

To characterize the correlation between variables and covariates, we use a linear discriminant analysis [174] and the training dataset. Table 8 reports the regression coefficients between the covariates $\{\rho_{aer}(865), c, \Theta_v, \Theta_s\}$ and the polynomial coefficients a_i of the aerosol model (Eq. 61). Table 8 outlines the significant regression coefficients (p-value << 0.05) between the coefficients a_i and the considered covariates underlying that the covariates provide significant information on the type (spectral shape) of the aerosols.

Table 8: Statistical analysis of the regression between the aerosol covariates $\{\rho_{aer}(865), c, \Theta_v, \Theta_s\}$ and the aerosol model coefficients a_i.

ai	$\rho_{aer}(865)$	C	Θ_v	Θ_s
a1	coeff=1.8625e-08 std=1.1048e-09 p-value =5.92e-57	coeff=-7.9458e-08 std=1.3105e-08 p-value =1.82e-09	coeff=1.89e-12 std=9.00e-13 p-value =0.03	coeff=5.4758e-12 std=9.5619e-13 p-value =1.3660e-08
a2	coeff=5.8402e-06 std=4.4532e-07 p-value =1.22e-36	coeff=6.4137e-06 std=7.7101e-06 p-value = 0.04	coeff=1.6669e-09 std=5.2647e-10 p-value =0.0016	coeff=2.8897e-09 std=5.5255e-10 p-value = 2.0804e-07
a3	coeff=-0.0021 std=7.1659e-05 p-value =4.75e-139	coeff=0.0181 std=8.5006e-04 p-value=7.20e-85	coeff=3.8955e-07 std=5.2184e-08 p-value =1.8553e-13	coeff=-2.5004e-08 std=5.5824e-08 p-value = 0.006543

Similarly, the spectral shape of $\rho_w(\lambda)$, i.e. NNMF coefficients h_i of Eq. 62, is strongly correlated with the geometry conditions and the values of $\rho_w(780)$. Table 9 reports the regression coefficients between the covariates $\rho_w(780), \Theta v, \Theta s$ and the water model coefficients h_i. It points out that the covariates provide significant information on the type (spectral shape) of water.

Table 9: Statistical analysis of the regression between the covariate $\rho_w(780), \Theta_v, \Theta_s$ and the water model coefficients h_i.

h_i coefficients	$\rho_w(780)$	Θ_v	Θ_s
h_1	coeff=2.3634 std=0.1143 p-value=9.7393e-83	coeff=1.7046e-04 std=2.0093e-05 p-value =4.6483e-17	coeff=1.1850e-04 std=2.0267e-05 p-value =5.9768e-09
h_2	coeff= 6.9345, std=0.0487 p-value=0	coeff=-9.9434e-05 std=7.2293e-06 p-value =6.5377e-41	coeff=-1.2726e-04 std=7.2918e-06 p-value =6.4574e-63
h_3	coeff=-0.4498 std=0.0326 p-value =1.2734e-40	coeff=2.2393e-05 std=7.0524e-06 p-value =0.0015	coeff=-4.2340e05 std=7.0809e-06 p-value =2.7169e-09
h_4	coeff=0.1616 sdt=0.0579 p-value=0.0053	coeff=8.5963e-0 std=1.5074e-85 p-value =2.7169e-09	coeff=8.6707e-06 std=3.1041e-140 p-value =2.7169e-09

5.4.2.3 Prior distributions

During operational inversion (cf § 5.6.2), geometry conditions are known and initial estimate of the covariates $\rho_{aer}(865), c \; and \; \rho_w(780)$ is performed during the Bright Pixel Estimation step (BPE, cf § 5.6.2). We thus consider the distribution of the extended variables $X_w = \{x_w, \rho_w(780), \Theta_v, \Theta_s\}$ and $X_a = \{x_a, \rho_{aer}(865), c, \Theta_v, \Theta_s\}$. X_a and X_w prior distribution are estimated using a GMM:

$$P(\mathrm{Xw} \mid \varphi) \; \alpha \sum_i \Lambda_i \; exp(-0.5.(X_w - \mu 0_{Xwi})^T . \Sigma 0_{Xwi}^{-1} . (X_w - \mu 0_{Xwi})) \quad (65)$$
$$P(\mathrm{Xa} \mid \varphi) \; \alpha \sum_j \Lambda_j \; exp(-0.5.\left(X_a - \mu 0_{Xaj}\right)^T . \Sigma 0_{Xaj}^{-1} . \left(X_a - \mu 0_{aj}\right))$$

where Λ_i is prior probability of mode i (or j) in the GMM:

$$\Lambda_j = \; P(Z_n = j) \quad (66)$$

Z is the hidden mode (aerosol type), n the number of spectra, $\mu 0_{Xwi}$, $\Sigma 0_{Xwi}$, $\mu 0_{Xaj}$ and $\Sigma 0_{Xaj}$ are respectively the vector of means and covariance matrices for X_w and X_a for mode i (or j). Xw and Xa distributions are modeled using a GMM, then, the conditional distribution of x_a and x_w given the covariates is estimated with the updated GMM [187]. For example, the updated GMM parameters to estimate the distribution of x_a given covariate $x_b = \{\rho_a(865), c, \theta v, \theta s\}$ resort for each mode to:

$$\mathrm{E}(x_a | x_b = z, \varphi) = \mu 0_{xa} + \Sigma 0_{x_a x_b} \Sigma 0_{x_b}^{-1} (z - \mu 0_{xb})$$
$$\Sigma 0(x_a | x_b = z, \varphi) = \Sigma 0_{x_a} - \Sigma 0_{x_a x_b} \Sigma 0_{x_b}^{-1} \Sigma 0_{x_b x_a} \quad (67)$$

$$\lambda_j = P(Z_n = j|\mathrm{x_b} = \mathrm{z}, \varphi) = \Lambda_j * \mathrm{P}(\mathrm{x_a}|\mathrm{x_b}, j, \varphi) \,/ \sum_z \mathrm{P}(\mathrm{x_a}|\mathrm{x_b}, z, \varphi)$$

The choice of the covariates for both x_a and x_w is of particular interest as it allows maximizing the probability to retrieve respectively the good aerosol and water models used in the inversion. We note here that δψ, the delta azimuth, is not considered here as a covariate as it was not found significantly correlated with x_a or x_w (cf § 5.6.1.1 & 5.6.1.2).

5.4.3 Performance valuation.

To validate the proposed methodology, radiometric in-situ profile dataset have been divided randomly in two sets of equal size: a training dataset and a validation dataset. Model parameters are estimated using the training dataset. The optimal number of clusters, k, used in the GMM to estimate X_a and X_w PDF, i.e. the number of hidden physical relationships to characterize, is estimated using the Bayes Information Criterion (BIC) [12] and the explained variable criterion [11]. Validation is performed with the validation dataset, using scatter plots between estimated and in-situ $\rho_w(\lambda)$, histograms, and related statistics estimators. For statistics, regressions of type II [188] are used, i.e. a regression model that considers uncertainties for both y and x as the in-situ measurements also involves uncertainties.

5.5 The in-situ MERMAID dataset.

The MERMAID (http://hermes.acri.fr/mermaid/home/home.php) in-situ matchup database is a comprehensive dataset that gathers in-situ measurements of water leaving radiances, IOPs, and MERIS TOA reflectances [53] measured at the same location. Many sites are available and among them, the most known are the NASA bio-Optical Marine Algorithm Dataset (NOMAD, [189]), the "BOUée pour l'acquiSition d'une Série Optique à Long termE" (BOUSSOLE) mooring program [190], the Aerosol Robotic Network (AERONET, [191]) stations, the Helgoland transect [192] that provides a full dataset of radiometric in-situ measurements in the Baltic Sea complex waters, and the MUMM Trios dataset [193]. Our initial dataset gathers 1976 matchups (without glint [18, 19]) in case 2 waters measured at the MERIS wavelengths: 412.5, 442.5, 490, 510, 560, 630, 665, 681, 708, 753.75, 778.75 and 865 nm. For each in-situ measurement, we use the corresponding 1km² MERIS pixel (no spatial averaging) as our dataset involves large shoreward gradients.

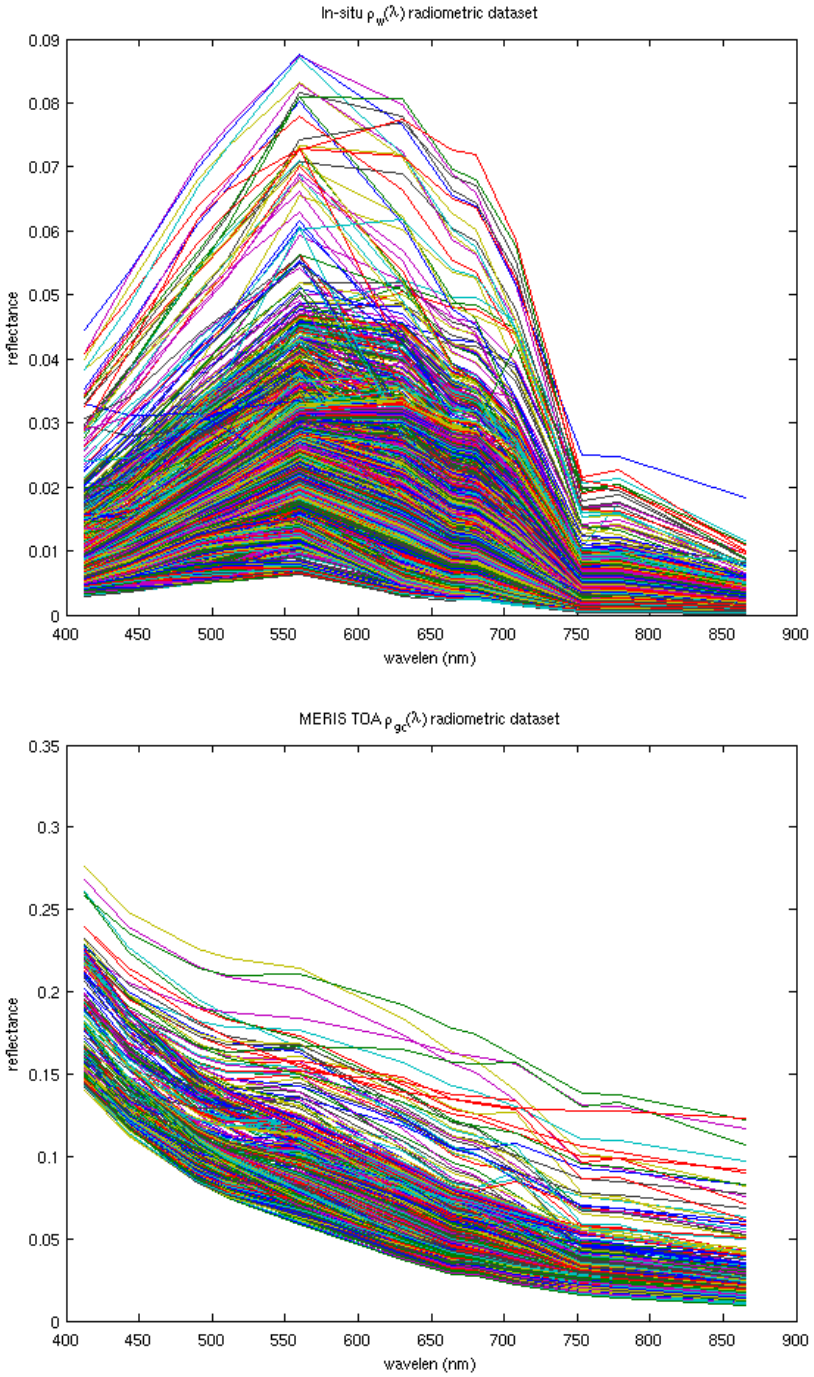


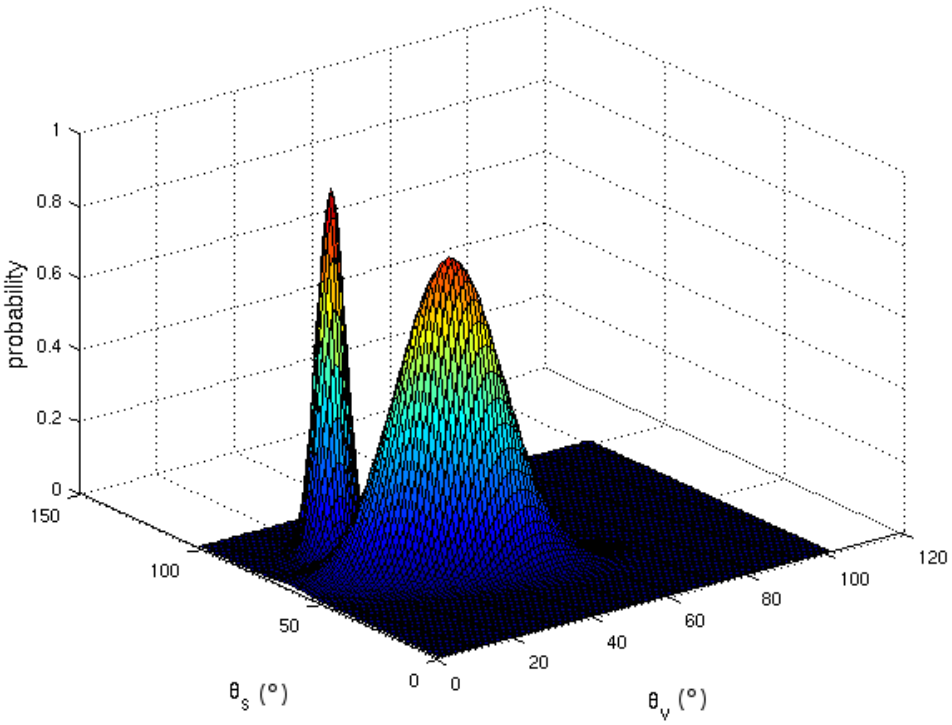
Figure 23: Top, the 1976 in-situ water reflectance profiles in complex waters. Bottom, the corresponding (matchups) ρ_{GC} (TOA) observed from the MERIS sensor.

5.6 Results

5.6.1 Prior distributions of aerosol and water variables

5.6.1.1 Prior distribution of X_a

A 10-mode mixture model (cf § 3.3) was selected to model the joint distribution of X_a ={a_i, $\rho_{aer}(865)$, c, Θ_v, Θ_s}. As an illustration, Figure 23a depicts the marginal PDF of Xa for dimensions {Θ_v, Θ_s} for modes 8 and 9. Figure 23b shows the 10 aerosol modes reconstructed from the GMM centers, and their associated uncertainties for each wavelength. Priors are indicated in the legend. We remind that the PDF of Xa involves here a full covariance matrix $\Sigma 0_{Xai}$ for each mode. These covariance matrices, which differ among modes, influence the cost function in the maximization procedure (Eq. 63 & 65).



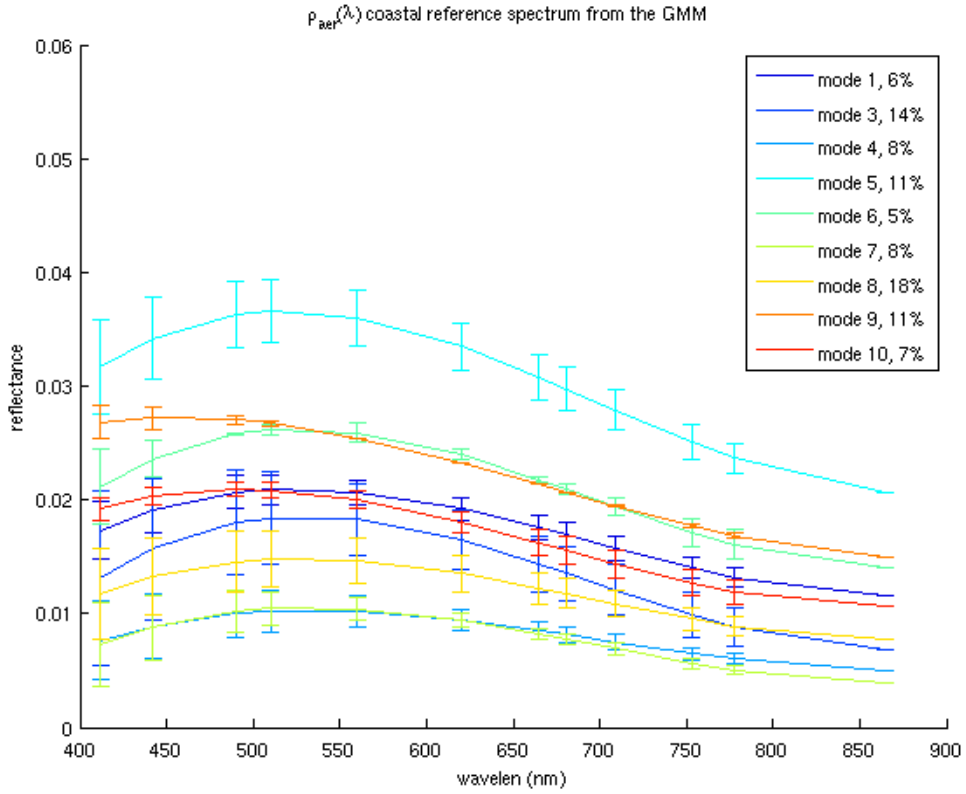


Figure 24: Top, marginal probability of Xa={$a_i, \rho_{aer}(865)$, c, Θv, Θs} for dimensions Θv & Θs and modes 8&9. Bottom, the 10 aerosol modes reconstructed from the GMM and Eq 61.

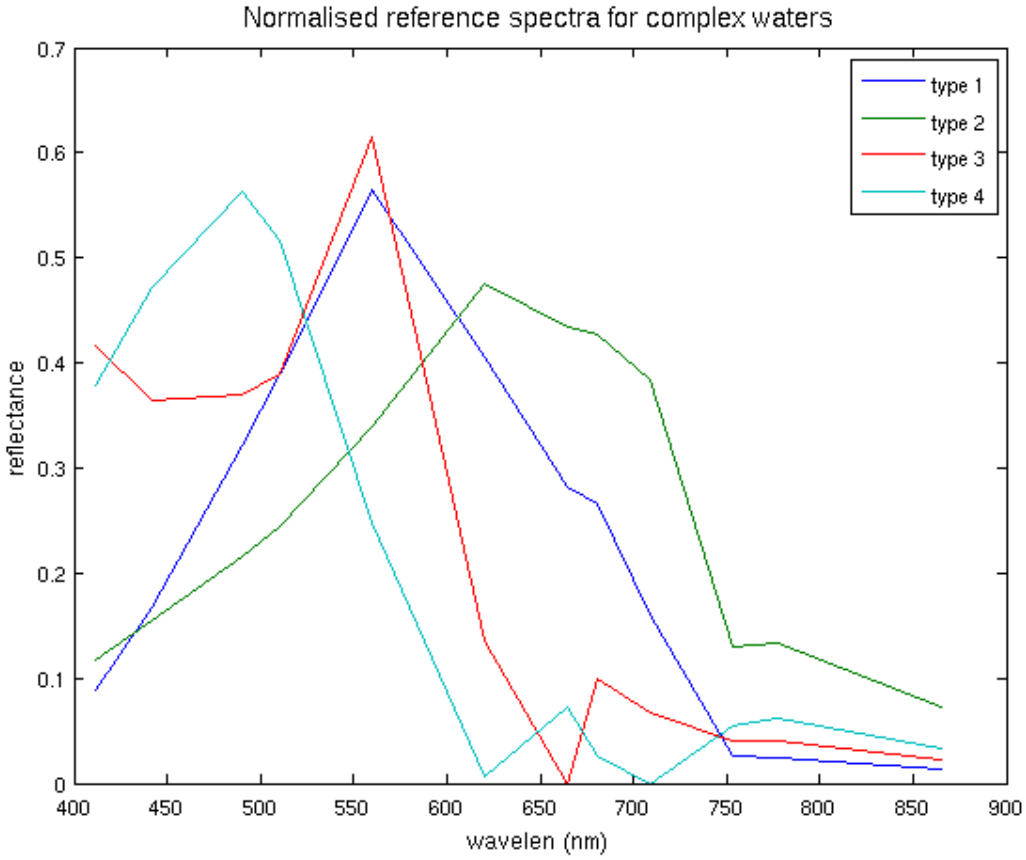
5.6.1.2 Prior distribution of distribution of Xw

From the NNMF applied to the in-situ water spectra (Eq. 62), 4 reference water types (W) were needed to reconstruct 99% of the variance of the in-situ spectra training dataset (Eq.62, Figure 25a):

- Spectrum n°1 (dark blue) highlights the strong signature of chlorophyll-a (chl-a) at 560nm on the reflectance.
- Spectrum n°2 (green) is a typical spectrum observed in presence of both high SPM concentration and Colored Dissolved Organic Matters (CDOM) absorption [168, 198].
- Spectrum n°3 (red) is characterized by a mixture of both signatures, of the pure water reflectance spectrum [194], and chl-a at 560 nm.
- Spectrum n°4 (light blue) is characterized by a mixture of both signatures of the pure water reflectance spectrum with CDOM absorption from 412 to 490 nm.

This characterization of the NNMF decomposition modes in water types is supported in section §5.6.5 by the spatial coherence of the distribution (h_i coefficients) of the water types inferred from the inversed water reflectances. It suggests that, although the NNMF decomposition is not a bio-optical inversion, it is related to true meaningful observed situations. A 9-mode GMM (Eq. 65) provides the best BIC indice to fit the prior distribution of X_w= {h_i, $\rho_w(780)$, θv, θs}.

Figure 25b shows the 4 most relevant modes for X_w and their associated uncertainty for each wavelength. Similarly to the prior distribution of aerosol contributions, a full covariance matrix $\Sigma 0_{Xwi}$ is estimated for each mode i.



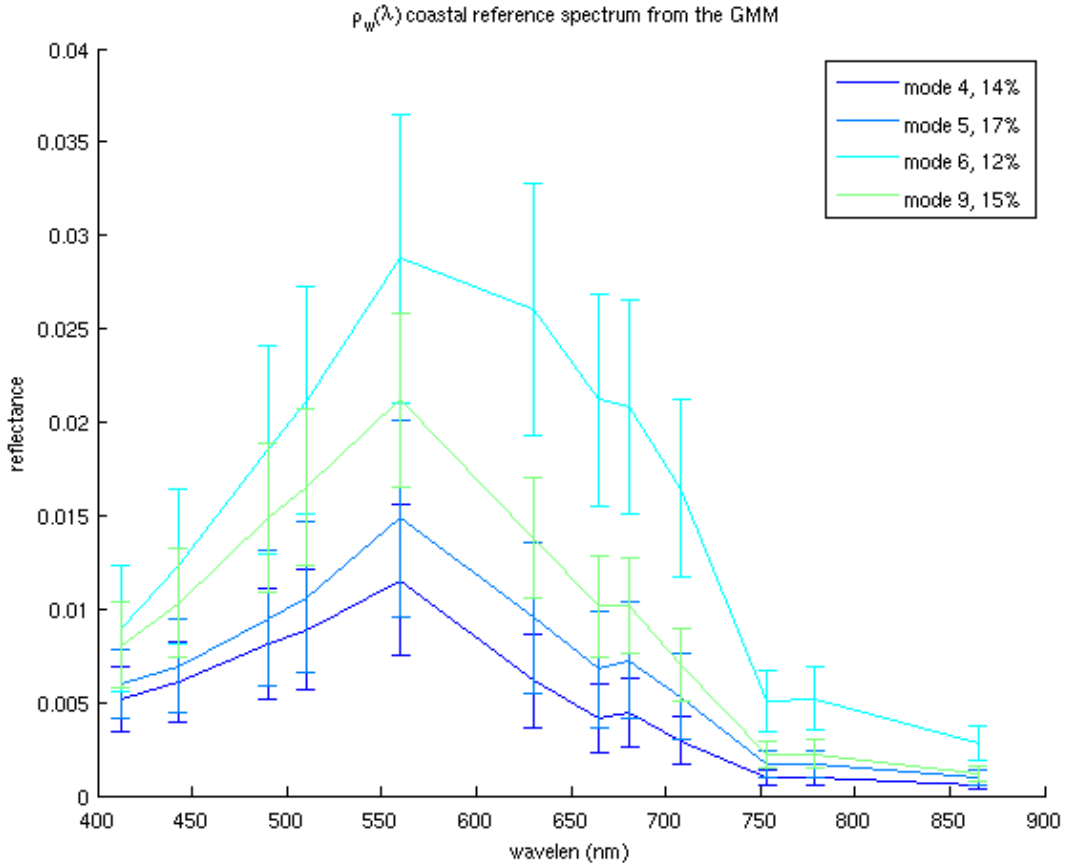


Figure 25: Top, the water type spectral shape, W(λ), estimated using NNMF with projected gradients. Bottom, 4 of the 9 reference water models reconstructed using the GMM centers and Eq 62.

5.6.1.3 Distribution of the observation model residuals $\rho_{gc}(\lambda)$.

The observation model residuals $\rho_{gc}(\lambda)$ (Eq.56), in the considered Bayesian setting (cf § 5.4.2), involves a null vector and a full covariance matrix. The covariance matrix is estimated using the distribution of the residuals obtained in the training phase for φ known (Eq. 63).

5.6.2 Bayesian ocean-color inversion

Given the calibrated priors and observation model in (Eq. 63), we achieve the Bayesian inversion of measured MERIS TOA reflectance according the MAP criterion (Eq. 64). It first involves the estimation of the covariates $\{\rho_{aer}(865), c, \rho_{aer}(780)\}$. We proceed similarly to the BPAC procedure (§5.3), but conversely, we do not correct the TOA signal. We refer this step as Bright Pixel Estimation (BPE). More precisely, we compared 3 BPAC implementations, the standard BPAC

(Moore [195]), a BPAC with varying backscattering slope (SAABIO, [196]), and a BPAC based on the water similarity spectrum [176]. We performed a quantitative evaluation according to the relative mean square error (%) for each variable over the full matchup dataset. Using our dataset, the best results were obtained with the following convergent algorithm based on the water similarity spectrum:

$$\begin{cases} initialisation\ of\ \rho_{aer}(753.75, 778.75, 865) = \rho_{Rc}(753.75, 778.75, 865) \\ \\ until\ convergence\ |\hat{\rho}_{RC}^{n+1} - \hat{\rho}_{RC}^{n}| < \varepsilon \\ \\ \hat{c}, \hat{\rho}_{aer}(865) = argmin(\rho_{aer}(\lambda) - \hat{\rho}_{aer}(\lambda)) \\ \rho_w = (\rho_{RC} - \hat{\rho}_{aer})/t_d\ for\ \lambda = 681, 708, 753.75, 778.75 \\ \hat{\rho}_w(780) = argmin(\rho_w(\lambda) - \hat{\rho}_w(\lambda)) \\ \rho_{aer} = \rho_{RC} - t_d\hat{\rho}_w\ \ for\ \lambda = 753.75, 778.75, 865 \\ \hat{\rho}_{RC}^{n} = \hat{\rho}_{aer}(\lambda) + t_d\hat{\rho}_w(\lambda)\ for\ for\ \lambda = 681, 708, 753.75, 778.75, 865 \\ \\ end \end{cases}$$

The Figure 26 shows the errors in the estimations of the covariates $\{\rho_{aer}(865), \mathrm{c}\ \text{and}\ \rho_w(780)\}$. On average, errors lower than 10% were reported for $\rho_{aer}(865)$ and c, while this error was more important (15% in mean) for $\rho_w(780)$. Though this latest error may affect the Bayesian inversion, the numerical experiments performed for the matchup database clearly pointed out better inversion performance using this initial covariate estimate than without.

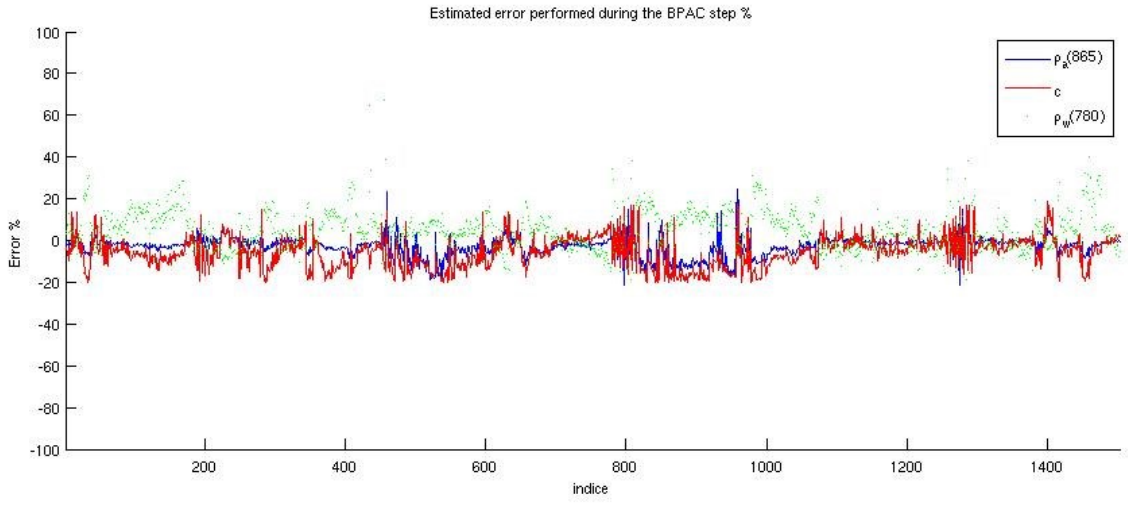


Figure 26: Errors performed (%) on the covariates $\rho_{aer}(865), \mathrm{c}\ \text{and}\ \rho_w(780)$ estimated during the BPE step.

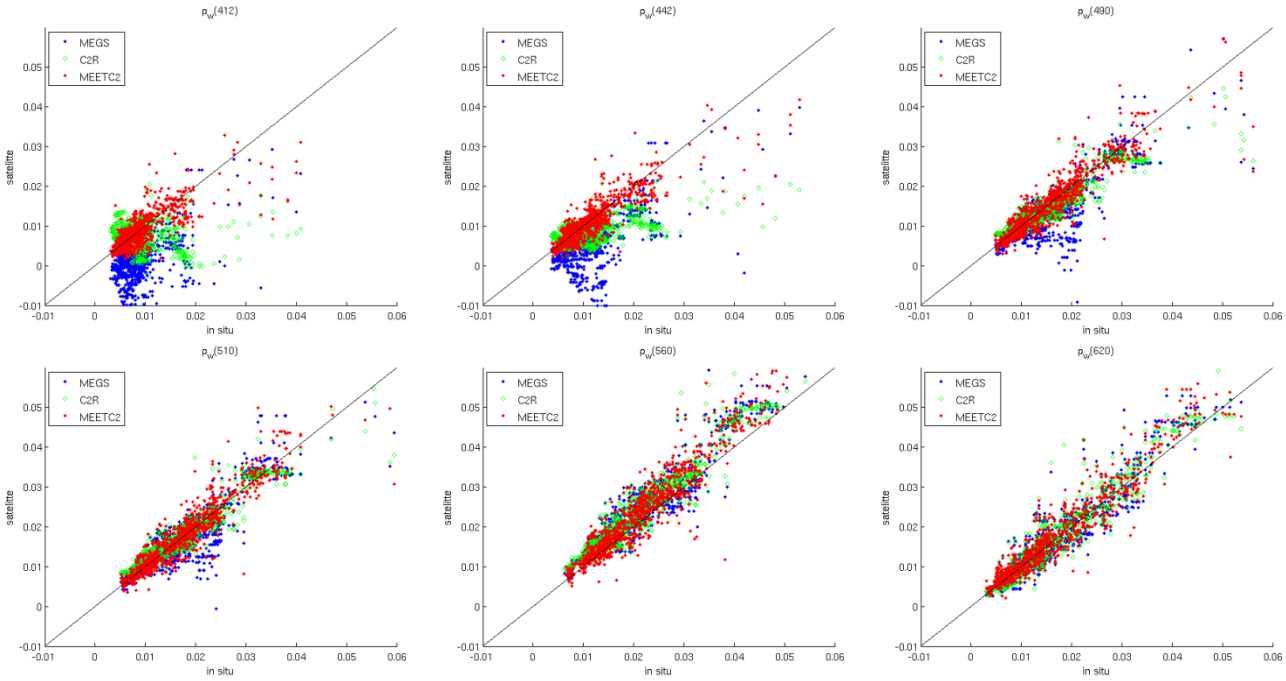
Given the estimated covariates, we update GMM for x_a and x_w conditionally to the covariates (Eq. 67). The initialization of the gradient descent is obviously a key issue as gradient-based maximization may converge toward local minima. We proceed as follows: 100 aerosol parameters are randomly generated using the updated distributions. x_w initialization is performed using

estimated $\hat{\rho}_w(780)$ and Eq.57. Overall, we select the solution of the gradient-based maximizations corresponding to the highest value of the MAP criterion (Eq.64).

5.6.3 Inversion performance for the Mermaid dataset

We perform a quantitative evaluation of the performance of the proposed Bayesian inversion model, MEETC2, for the Mermaid dataset and coastal waters. For the validation dataset, we analyze for each wavelength the estimated water reflectances $\hat{\rho}_w$ against in-situ measurements (Figure 27, red). In addition to the proposed Bayesian inversion, we also report on Figure 27 the inversion performed with MEGS v8 (blue [20]), and C2R (green [54]). Table 10 summarizes the corresponding statistical results.

On this validation dataset, MEETC2 clearly outperforms MEGS and C2R at bands 412, 442, 490 and in term of mean-bias, mean absolute error, slope, R^2 coefficient and σ. At 560 nm the three algorithms slightly overestimate the in-situ data with a minimum bias value of 0.0021. From 620 to 865 nm MEETC2 slightly outperforms the two other models. Overall, the gain value on the relative absolute error over the 12 bands is of 67% compared with MEGS and 9% compared with C2R.



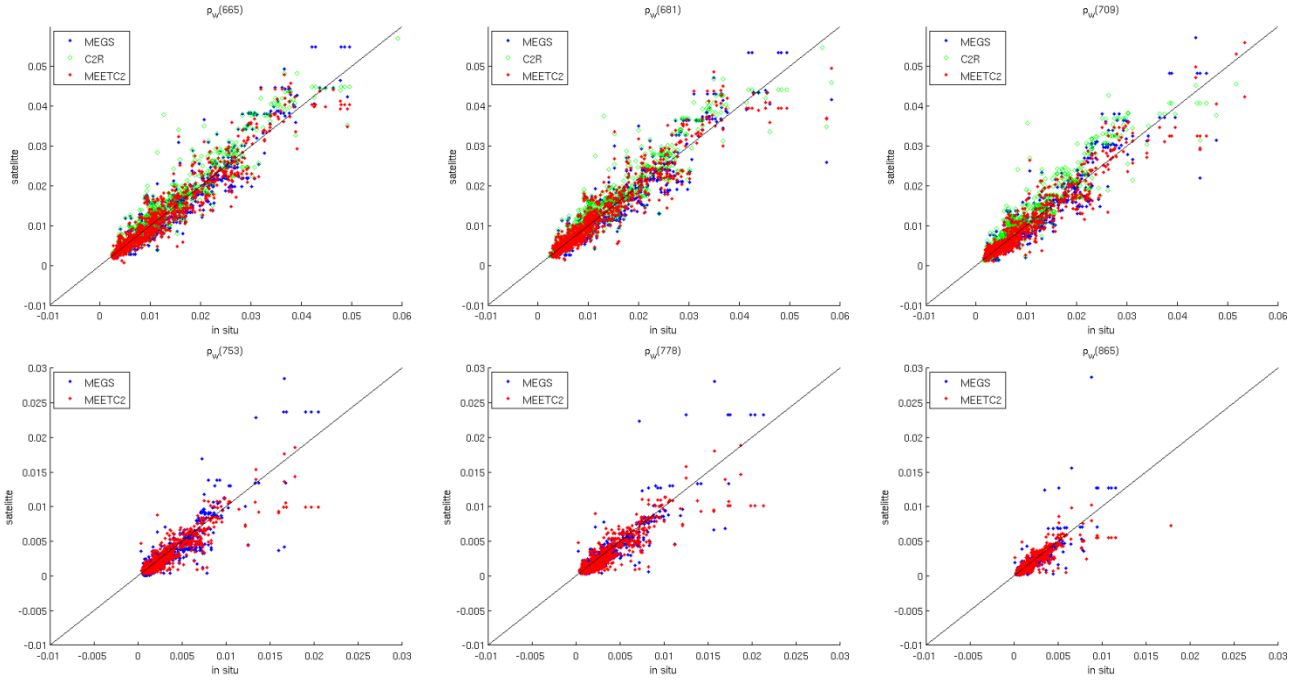


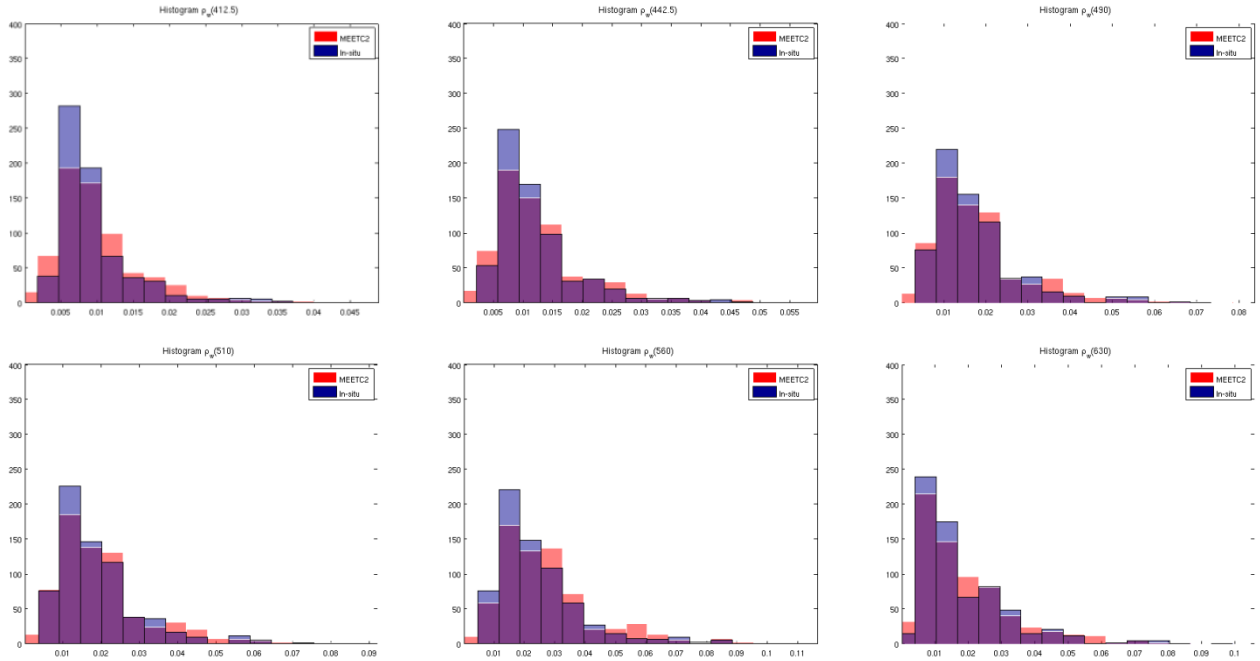
Figure 27: comparisons between $\hat{\rho}_w$ estimated using MEETC2 vs in-situ (red), MEGS 8 vs in-situ (blue) and C2R (NN) vs in-situ (green).

Table 10: Statistical analyses of the estimated water reflectances vs. in-situ data for the proposed Bayesian model (MEETC2), the standard MEGS processor and the neural-net-based algorithm C2R [54]. For each wavelength, we report the mean error (bias), the relative absolute mean error (%), the slope of the regression with the in situ data, the associated R^2 score and standard deviation (σ). We report in bold the algorithm which provided the best performance.

λ(nm)		Mean error	Relative absolute mean error (%)	Slope	R^2 (Pearson)	σ
412.5	MEETC2	**-0.0002**	**22.51**	**0.99**	**0.79**	**0.0039**
	MEGS	-0.0096	102.13	0.25	0.16	0.0101
	C2R	-0.0029	56.27	-0.04	0.15	0.0065
442.5	MEETC2	**-0.0004**	**18.31**	**0.98**	**0.87**	**0.0033**
	MEGS	-0.0073	59.48	1.39	0.38	0.0084
	C2R	-0.0040	41.04	0.31	0.60	0.0056
490	MEETC2	**-0.0003**	**16.46**	**0.97**	**0.92**	**0.0044**
	MEGS	-0.0033	28.25	0.92	0.76	0.0060
	C2R	-0.0019	27.10	0.77	0.87	0.0049
510	MEETC2	**0.0000**	**13.21**	**0.96**	**0.91**	**0.0039**
	MEGS	-0.0007	16.64	0.91	0.64	0.0053
	C2R	0.0006	16.44	0.72	0.88	0.0055
560	MEETC2	**0.0021**	**15.67**	**1.04**	0.88	**0.0046**
	MEGS	0.0028	18.01	0.95	0.81	0.0054
	C2R	0.0028	18.65	0.88	**0.90**	0.0061

620	MEETC2	**3.3684e-04**	**14.87**	**1.00**	**0.93**	**0.0042**
	MEGS	5.1299e-04	15.38	1.05	0.85	0.0053
	C2R	4.169e-04	14.97	0.97	0.90	0.0058
665	MEETC2	**-1.5457e-04**	**15.03**	**0.97**	**0.89**	**0.0045**
	MEGS	-1.9302e-04	17.24	1.07	0.85	0.0064
	C2R	2.3715e-04	17.02	1.02	**0.89**	0.0054
681	MEETC2	**7.3502e-05**	**15.90**	**0.99**	**0.92**	**0.0045**
	MEGS	-7.8276e-05	16.81	1.06	0.85	0.0062
	C2R	9.3321e-04	17.53	1.02	0.89	0.0053
708	MEETC2	**9.0173e-05**	**18.45**	**0.94**	**0.87**	**0.0038**
	MEGS	-2.2832e-04	19.93	1.13	0.83	0.0058
	C2R	0.0085	22.46	1.10	**0.87**	0.0051
753	MEETC2	**-5.6657e-05**	**18.51**	**0.90**	**0.90**	**0.0013**
	MEGS	-2.2260e-04	20.59	1.35	0.78	0.0028
778	MEETC2	**-1.9678e-05**	**19.11**	**0.90**	**0.89**	**0.0012**
	MEGS	-2.9360e-05	17.09	1.21	0.74	0.0034
865	MEETC2	**3.8509e-05**	**18.25**	**0.94**	**0.88**	**0.0007**
	MEGS	-5.6208e-05	20.85	1.31	0.75	0.0018

We further analyze the extent to which we recover realistic water reflectances from the proposed Bayesian inversion MEETC2. To this end, we compare for each wavelength, the distribution of in-situ measurements to the MEETC2 estimates (Figure 28). We see a global agreement between the distributions of $\hat{\rho}_w(\lambda)$, compared to the reference in-situ distributions, for each wavelength λ.



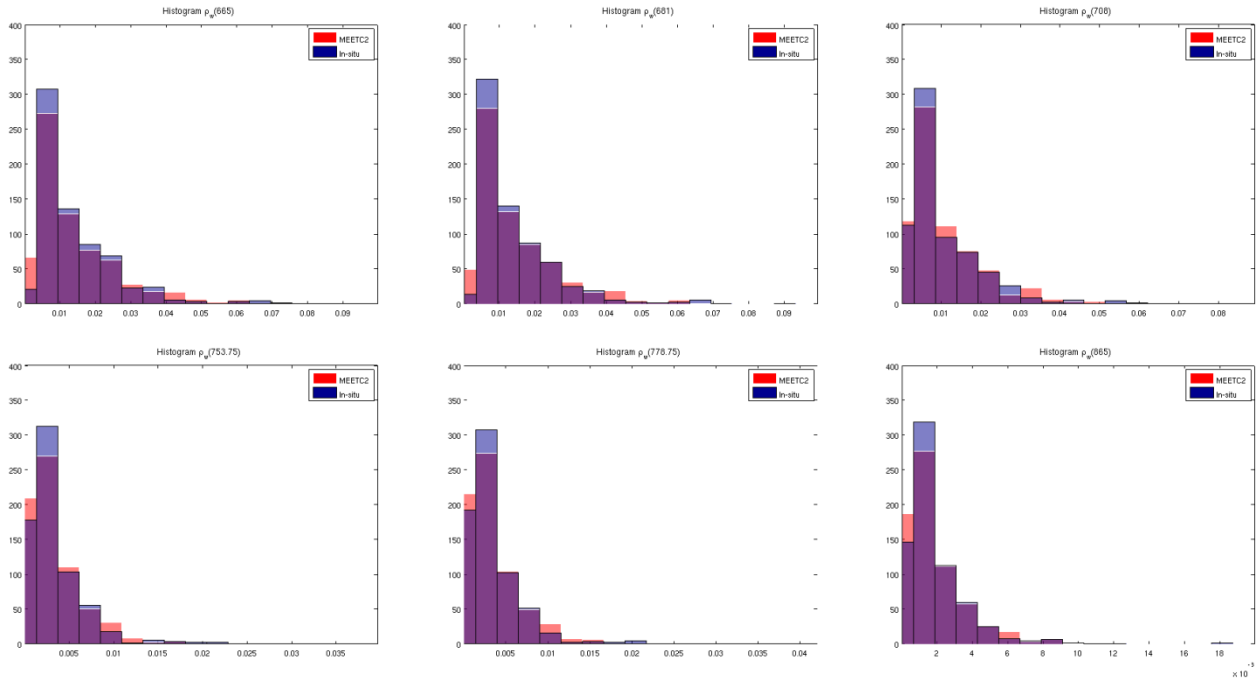
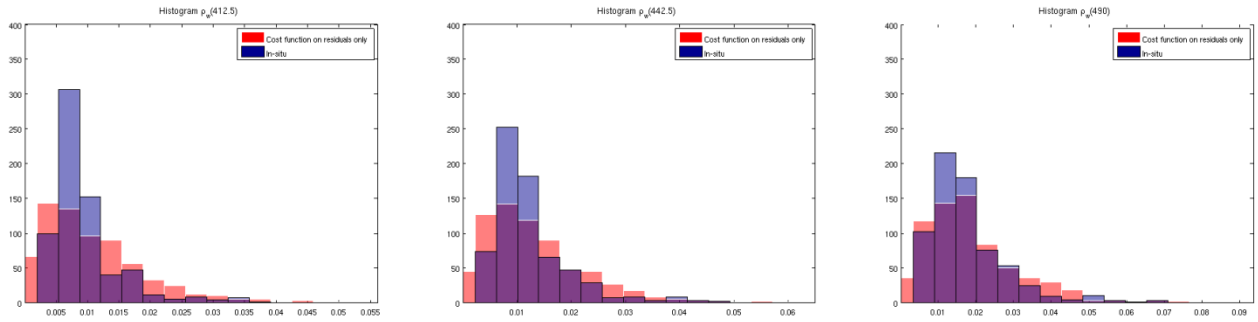


Figure 28: Comparison of the distributions of water reflectantes $\hat{\rho}_w$ for in-situ measurements (blue) and the proposed inversion (MEETC2 model, red).

To illustrate the added value of the introduction of priors on both water and aerosol spectra, we implement model (Eq. 63) without priors on Xa and Xw, i.e. the cost function of Eq. 64 is in this case equal to the cost on residual distribution: $C = -\log(P(\delta\rho_{RC}|x_a, x_w, \varphi))$. In that case, the MAP criterion reduces to the Maximum Likelihood criterion. Figure 29 shows the corresponding results obtained, using the same validation dataset. We clearly see in Figure 29a smoothing effects for bands 412, 443, 490, 560 and 680 nm on the estimated distributions of $\hat{\rho}_w$. The resulting bias with the in-situ is lower at 865 nm where the 9 water and the 10 aerosol models tend to converge at this wavelength.



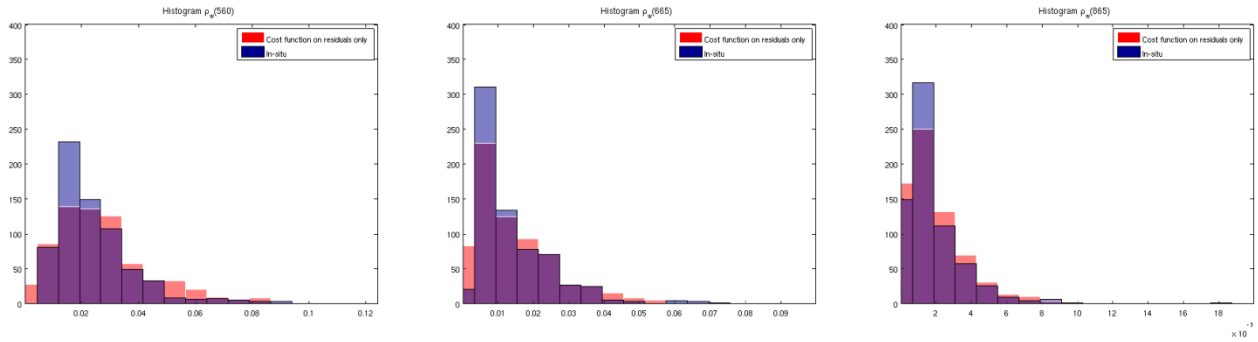


Figure 29: Distributions of $\hat{\rho}_w$ retrievals for wavelengths 412, 443, 490, 560, 680 and 865 nm using a cost function $C = -\log(P(\delta\rho_{RC}|x_a, x_w, \varphi))$ for the inversion (Eq. 64) vs in-situ. In that case, the MAP criterion reduces to the Maximum Likelihood criterion.

5.6.4 Example of estimated water reflectance on a very turbid area

Figure 30b shows the estimated $\hat{\rho}_w$, using the 20090318 MERIS Full Resolution (FR) level 1 observations over the French La Gironde's estuary, using the three algorithms. At springtime in this area, a bloom occurs leading to high chl-a concentrations (typically of magnitude from 5 to 15 $mg.m^{-3}$). In the same time, the seasonal river outflow involves high, SPM concentrations and CDOM absorption.

Chapter V: Ocean Color atmospheric corrections in coastal complex waters using a Bayesian latent class model

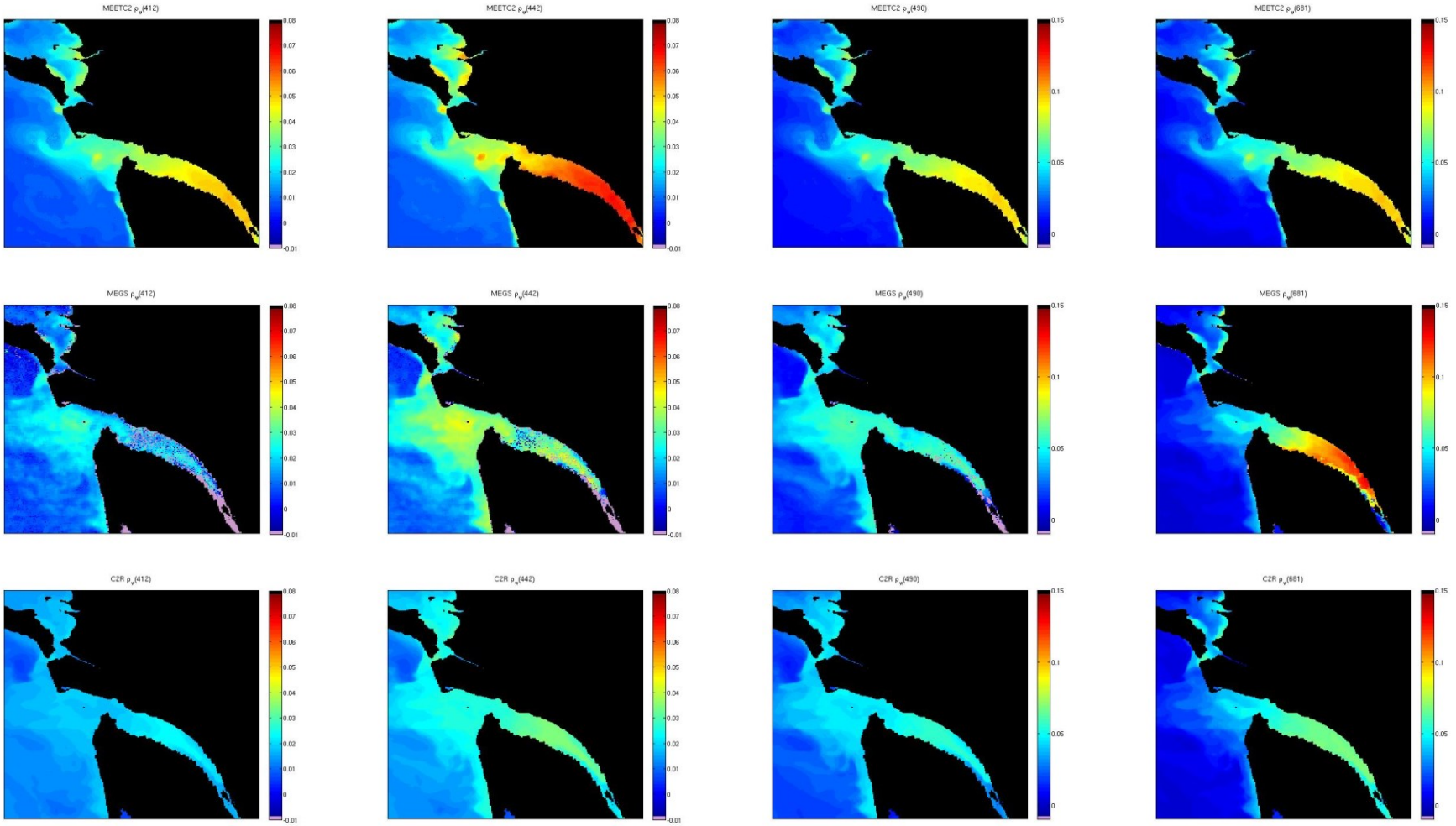


Figure 30: a/ true color representation of the 20090318 MERIS FR Level 1 image over the French river La Gironde's estuary. b/ Estimated $\rho_w(412, 442, 490, 680)$ (left to right) from the Level 1 data. Top, MEETC2 retrievals, middle, MEGS v8 and bottom C2R retrievals.

5.6.5 Estimated water types associated with the MEETC2 inversion

Figure 31 depicts the associated water type classification from the MEETC2 $\hat{\rho}_w$. Figure 31a depicts the presence (waters of type 1, Figure 25a) of chl-a over the all area as expected for this spring period and region. We observe clear contrasted situations between Figure 31b & c and Figure 31c & d. Waters of type 2 (Figure 25a), i.e. whose spectral shape is mainly constrained by SPM reflectance and CDOM absorption, are mainly located in the Gironde river, while clearer waters with chl-a (type 3) are located in the oceanic part of the estuary. The clearer waters (type 4) are observed in the more oceanic part of the area.

This spatial constistency of the distribution of the water types from the estimated $\hat{\rho}_w$, relative to our knowledge of the seasonal behavior in this area, contributes to validate the shapes of our estimated $\hat{\rho}_w$.

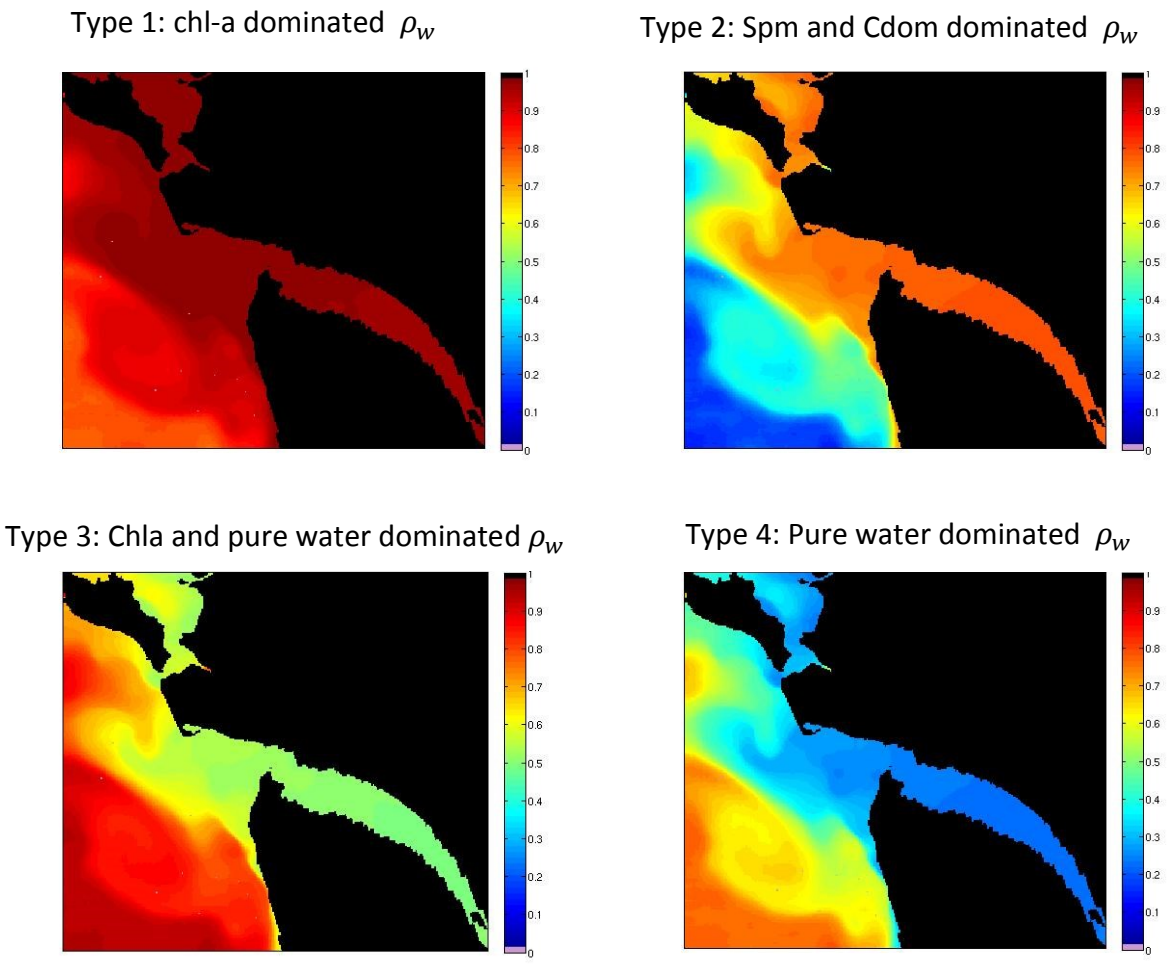


Figure 31: coordinated of the MEETC2 estimated $\hat{\rho}_w$ in the water reference spectrum basis.

5.7 Discussion

A significant improvement of ocean color inversion in coastal waters.

Retrieving reliable Ocean Color reflectances from space in coastal areas is of major challenge for a number of operational and scientific issues, including for instance delivery of reliable satellite-derived products in coastal areas for the spatial agencies, bio-optical and biological modeling, and environmental monitoring policies such as the WFD. Using the MERMAID satellite/in-situ collocated observation database, a Bayesian latent class model was shown to significantly enhance inversion of water reflectances for complex waters compared to the standard MEGS inversion scheme and the C2R, a Neural Network trained using similar in-situ data [54].

The improvements were especially large for the 412, 442 nm and 490 nm bands, which are used in Ocean Color for the estimation of the chl-a concentration, the CDOM absorption and the SPM concentration, underlying the potential of such approach to improve the standard level 2 products in coastal areas.

The complexity of the inversion is particularly stressed by the number of hidden models, respectively 10 for coastal aerosol reflectances and 9 for water reflectances, to address the spectral variability of both water and atmospheric contributions in such areas.

A physically-interpretable modelling framework

Conversely to Neural Network, the modes retrieved by the Gaussian Mixture Models correspond to identifiable aerosol and water signatures. Compared to neural networks, the fact that we explicitly distinguish parametric representations of aerosol and water spectra make also easier the independent calibration of the models. Machine learning typically requires all combinations of aerosol and water signatures in the training dataset with a view to learning a generic model. This appears to be very complex for non-specific areas. By contrast, our Bayesian model may benefit in a much simpler manner for newly collected and/or simulated dataset to improve each prior distribution independently. This is regarded as a key property for operational applications with respect to ongoing advances in bio-optical modelling, in-situ monitoring and future satellite missions.

Operational potential in the framework of the ocean sensor of upcoming Sentinel 3 platform

The incoming OLCI Ocean Color sensor, embedded on the Sentinel 3 platform, should succeed the MERIS sensor in 2015. The available spectral bands will be close to the MERIS ones. Beyond genericity of our Bayesian framework, we thus expect the considered parameterization, especially the NNMF-based representation, the GMM-based priors and

the covariance models, to be directly transferable to the future OLCI observations. Our ongoing work addresses the development of an operational product based on the proposed Bayesian model. First, the dependency of both aerosol and the water variable distribution to the observation geometry conditions is currently addressed using Hydrolight (© Curtis D. Mobley, 2008) and Modtrans [79] simulations to cover the full possible range of conditions (which is not the case using in-situ data). We also investigate additional covariates, e.g. humidity and wind conditions to further constrain the priors used by the model. Parallelized implementation is also under investigation, as, conversely to existing MEGS and C2R processors, our optimization is computationally more demanding than these as it relies on quasi-randomized initializations for the atmospheric initial model, i.e. multiple initializations given the observed covariates values. Optimal and noiseless results will be obtained with increased number of random initializations to converge towards the 'true' solution. This random initialization issue and the associated computing cost, is classic for genetic algorithms [199] and the new generation of satellite products such as the Soil Moisture Ocean Salinity (SMOS) product [200].

6 General discussion & perspectives

During this doctorate, several classes of models have been developed to analyze Ocean Color and Sea Surface temperature geophysical time series. Among them, particular attention has been given to multivariate clustering approaches, such as Gaussian Mixture Models, and hidden Markov Models (HMM, HMM-AR, NHMM, NHMM-AR). The multi-mode or the multi-regime nature of the proposed methods is integrated in the thematic analysis of the results: in each case, **we try to characterize the underlying physics behind the identified modes or regimes**. From our point of view, **this aspect reinforces the value of the described approaches in comparison to learning machines such as neural networks**.

Chapter II emphasizes **the need to integrate inherent characteristics of geophysical variables,** such as noise autocorrelation, discontinuity in the observations, **with inference of parameters to be estimated**. Perspectives of our proposed methodology are optimization of observation networks for environmental concerns, particularly, the planning of successive satellite or in-situ missions. From an academic point of view, the published paper is the first methodology proposed for estimation of trends in multiple geophysical time series. This paper should contribute to the raise awareness among the geophysical and thematic communities about the interactions between specific characteristics of geophysical signals as well as the uncertainties of estimated parameters.

In keeping with chapter II, we characterize in chapter III a signal relative to its geophysical nature: an event of SST relative to the local conditions of variance and autocorrelation of noise. From a methodological point of view, chapter III underlies **the potential of the double approach; discretization of a space-time signal and clustering of descriptors to characterize a geophysical process**. The proposed methodology allows a finer, but more complex, analysis of the studied phenomena compared to standard correlation analysis. The characterization of the low frequency reference time-scales of the ENSO, and its clear signature on the high frequency of SST, are typical examples of the added value of our approach compared with standard analyses.

Multiple perspectives emerge from such approaches. We have focused particularly on the time-scale distribution of detected events. Descriptors of the events also gather abundant information, relevant to characterize the considered physical process, which was not addressed here. An example is that the slope of the main axis of the event is a proxy of its propagation. From a general point of view, **analysis of joint distribution of descriptors of different variables** (such as temperature, chlorophyll-a and wind**) contributes to**

establishing a methodological framework for better understanding of dependency and causality relationships between variables.

Modelling a geophysical non-linear process, the surface turbidity, **using hidden Markov models** significantly improves forecasting accuracy compared with more classical multivariate regressions. It paves the way **towards operational forecasting** of the surface turbidity **using available observations and associated models**. This modelling will be of particular interest in areas where bathymetry and limit boundary conditions are not well known: in this case, classical hydrodynamic models won't provide accurate simulations. The turbidity in the water column may also be addressed using new regimes to model relationships between the morphology of the vertical turbidity profile and surface observations. Potential applications for other geophysical variables are also numerous since many variables are intrinsically driven by regime switching behaviors. The chlorophyll-a, shell and fish growth, typically depict active and passive phases driven by environmental factors.

Cluster analysis of aerosol and water reflectance spectra is an innovative approach to inverse sea surface reflectance compared with the state of the art approaches actually used in Ocean Color. At the term of this doctorate, the estimations of marine reflectance in turbid waters in the 412-490 nm range have been significantly improved, compared to the actual ESA processing chain and a dedicated neural network. This is of particular interest as these wavelengths have been used for the estimation of numerous biological and geophysical parameters such as chlorophyll-a, CDOM and the light attenuation. Beyond these analytical aspects, the improvements made remind us that remotely sensed estimations of a geophysical variable remain estimations. Such data and subsequent analyses should therefore be considered with a critical eye, especially regarding coastal areas.

The perspectives of Bayesian approaches with prior knowledge on distributions are the enhancement of satellite derived products provided by spatial agencies, and allow for a better monitoring of areas typically influenced by anthropogenic activities and subjected to environmental policies of European states. Generalization of such a model should be envisaged. Actually, spectra are clustered using both shapes and range criteria. It would be noteworthy to generalize this approach with independent analysis of shape and range of spectra using for example the multivariate distribution for angles (Von Mises, [55]). In this case the estimation of the range in the inversion remains difficult. Finally, our Bayesian approach may be applied to other variables showing multi-modes intrinsically in their distribution. The characterization of phytoplankton species is a typical example that may be addressed with the same methodology. In this case we will either cluster spectra directly, using in-situ observations, or cluster spectrum descriptors associated with covariates.

The spatial aspect has been partially addressed in this PhD thesis manuscript, so the joint analysis of spatial and temporal behaviors would be the logical continuation for this work. Spatio-temporal covariance between different variables (e.g. SST and chl-a) also shows distinct modes depending on geographical location and period. These covariance modes between variables may be of particular interest to estimate missing data. In the future, we should be able to better estimate a missing chl-a pixel of an image (due to cloudy conditions) using pre-estimated spatio-temporal covariance modes between variables such as SST and chl-a. Compared with the standard optimal interpolation technique (OI) that uses the modelled spatial covariance γ(d)=f(d) for a single parameter as a function of the distance d, the covariance $\hat{\gamma}'(\mathrm{d})$ will be estimated using the conditional expectation of γ' given the observations of chl-a and SST and potential covariates.

In my opinion, perspectives of this PhD also lie in **the generalization of the multi-mode, multi-regime Bayesian approaches to inverse and characterize a geophysical variable**. For inversion of a geophysical parameter, or analysis of interactions between variables, it appears that these concerns are multi-mode or multi-regime aspects inherent to the intrinsic geophysical nature of the considered variables. For that purpose, more complex model parameterizations will be certainly required than those considered in this thesis: non-linear regressions, non-Gaussian distributions for mixture models and residuals.

As the end of this doctorate draws closer, my personal feeling is that the potential of both statistics and signal processing to address classical or new scientific questions in ocean color is underestimated and hence in progress. It originates from the required interdisciplinary aspect for these approaches. The historical legacies of the scientific communities that have evolved alone, such as thematicians, statisticians and modelers, may also complicate such collaborations. The scarcity and the structure of related national or European funding, do not always favor the development of these interdisciplinary connections. My personal objective will be the ongoing development of multidisciplinary approaches for analyzing geophysical datasets, as initiated during this thesis's collaboration with the ENSTB, the CNRS and the IRD, to explore the full potential of such cooperation.

7 Bibliography

1 Communication de la Commission, du 9 février 2005, 'Vaincre le changement climatique planétaire' COM(2005) 35 - Journal officiel C 125 du 21 mai 2005].

2 Tiao, G. C., G. C. Reinsel, D. Xu, J. H. Pedrick, X. Zhu, A. J. Miller, J. J. DeLuisi, C. L. Mateer, and D. J. Wuebbles (1990). Effects of auto-cor-relation and temporal sampling schemes estimates trend and on of spatial correlation, Journal of Geophysical Research, 95, 20,507-20, 517.1

3 Weatherhead, E.C., Reinsel, G.C., Tiao, G.C., Meng, X.L., Choi, D.S., Cheang, W.K., Keller, T., Deluisi, J., Wuebbles, D.J., Kerr, J.B., Miller, A.J., Oltmans, S.J., Frederick, J.E. (1998). Factors affecting the detection of trends: Statistical considerations and applications to environmental data. Journal of Geophysical Research-Atmospheres, 103(D14).

4 Saulquin, B., Fablet, R., Mangin, A., Mercier, G., Antoine, D., & Fanton d'Andon, O. (2013). Detection of linear trends in multisensor time series in the presence of autocorrelated noise: Application to the chlorophyll-a SeaWiFS and MERIS data sets and extrapolation to the incoming Sentinel 3-OLCI mission. Journal of Geophysical Research: Oceans, 118(8), 3752-3763.

5 Ghil, M., M. R. Allen, M. D. Dettinger, K. Ide, D. Kondrashov, et al. (2002). Advanced spectral methods for climatic time series. Rev. Geophys., 40(1): 1003.

6 Ghil, M., and R. Vautard (1991): Interdecadal oscillations and the warming trend in global temperature time series, Nature, 350, 324–327.

7 V. Vantrepotte and F. Mélin (2011). Inter-annual variations in the Sea-viewing Wide Field-of-view Sensor (SeaWiFS) global chlorophyll a (Chla) concentration - 1997-2007 Deep-Sea Research I(58) 429-441 doi: 10.1016/j.dsr.2011.02.003.

8 Torrence C. and Compo G. P. (1998). A practical guide to wavelet analysis. Bulletin of the American Meteorological Society, 79:61–78.

9 Preisendorfer, R.W. (1988). Principal Component Analysis in Meteorology and Oceanography, Elsevier, New York. pp 425.

10 Saulquin, B.; Fablet, R.; Mercier, G.; Demarcq, H.; Mangin, A; Fanton d Andon, O., (2014) Multiscale Event-Based Mining in Geophysical Time Series: Characterization and Distribution of Significant Time-Scales in the Sea Surface Temperature Anomalies Relative to ENSO Periods from 1985 to 2009, Selected Topics in Applied Earth Observations and Remote Sensing,IEEE,vol.PP,no.99,pp.1,10

11 Saulquin B, Fablet R, Ailliot P, Mercier G, Doxaran D, Mangin A, Fanton d'Andon O (2014). Characterization of time-varying regimes in remote sensing time series: application to the forecasting of satellite-derived suspended matter concentrations. In publication at Selected Topics in Applied Earth Observations and Remote Sensing, IEEE.

12 Bhat, H. S., & Kumar, N. (2010). On the derivation of the Bayesian Information Criterion. School of Natural Sciences, University of California.

13 Saulquin, B., Fablet, R., Bourg L., Mercier, G., Fanton d'Andon, O. (submitted at RSE, October 2014). MEETC2: Ocean Color Atmospheric corrections in coastal complex waters using a Bayesian latent class model and potential for the incoming Sentinel 3 - OLCI mission.

14 Le Traon, P. Y., Rienecker, M., Smith, N., Bahurel, P., & Bell, M. (1999). Operational oceanography and prediction-a GODAE perspective (No. NRL/PP/7323--99-0050). NAVAL RESEARCH LAB STENNIS SPACE CENTER MS.

15 Sidi, M. J. (1997). Spacecraft dynamics and control: a practical engineering approach (Vol. 7). Cambridge university press.

16 Casey, K.S., T.B. Brandon, P. Cornillon, and R. Evans (2010). The Past, Present and Future of the AVHRR Pathfinder SST Program. Oceanography from Space: Revisited, eds. Springer.

17 IOCCG (2000): Remote Sensing of Ocean Colour in Coastal, and Other Optically-Complex, Waters. Edited by Shubha Sathyendranath, pp. 140.

18 Cox, C. and W. H. Munk (1954a). Statistics of the sea surface derived from sun glitter. J. Mar. Res., 13, 198–227

19 Cox, C. and W. H. Munk (1954b): Measurement of the roughness of the sea surface from photographs of the Sun's glitter. JOSA, Vol. 44, Issue 11, pp 838-850

20 Antoine, D., & Morel, A. (1999). A multiple scattering algorithm for atmospheric correction of remotely sensed ocean colour (MERIS instrument): principle and implementation for atmospheres carrying various aerosols including absorbing ones. International Journal of Remote Sensing, 20(9), 1875-1916.

21 Morel, A., and Gentili, B. (1996). Diffuse reflectance of oceanic waters. III. implication of the bidirectionality for the remote sensing problem, Appl. Opt. 35, 4850-4862.

22 Mobley, C. D. (1994). Light and water: radiative transfer in natural waters. London, Academic Press.

23 MERIS ATBD 2.7 — Atmospheric Correction of the MERIS observations Over Ocean Case 1 waters.

24 Planck, M. (1914). The Theory of Heat Radiation. Masius, M. (transl.) (2nd ed.). P. Blakiston's Son & Co. OL 7154661M.

25 Kushnir, Y., Robinson, W. A., Bladé, I., Hall, N. M. J., Peng, S., & Sutton, R. (2002). Atmospheric GCM response to extratropical SST anomalies: synthesis and evaluation*. Journal of Climate, 15(16), 2233-2256.

26 Lau, N. C. (1997). Interactions between global SST anomalies and the midlatitude atmospheric circulation. Bulletin of the American Meteorological Society, 78(1), 21-33.

27 Frankignoul, C. and Hasselmann, K. (1977). Stochastic climate models. Part II: Application to SST anomalies and thermocline variability. Tellus, 29:289-305.

28 Plackett, R.L. (1950). Some Theorems in Least Squares. Biometrika 37 (1–2): 149–157. doi:10.1093/biomet/37.1-2.149. JSTOR 2332158. MR 36980.

29 Cochrane, D., & Orcutt, G. H. (1949). Application of least squares regression to relationships containing auto-correlated error terms. Journal of the American Statistical Association, 44(245), 32-61.

30 McCullagh, P., & Nelder, J. A. (1989). Generalized linear models.

31 Burges, C. J. C., (1998). A tutorial on support vector machines for pattern recognition.Data Mining and Knowledge Discovery 2:121–167. xliii, 81

32 Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. Proceedings of the national academy of sciences, 79(8), 2554-2558.

33 Wu, C. J. (1983). On the convergence properties of the EM algorithm. The Annals of statistics, 95-103.

34 Lindsay B.G. (1995). Mixture Models: Theory, Geometry, and Applications. NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. 5, Institute of Mathematical Statistics, Hayward.

35 Kaplan, A., Cane, M. A., Kushnir, Y., Clement, A. C., Blumenthal, M. B., & Rajagopalan, B. (1998). Analyses of global sea surface temperature 1856–1991. Journal of Geophysical Research: Oceans (1978–2012), 103(C9), 18567-18589.

36 Henson, S. A., Sarmiento, J. L., Dunne, J. P., Bopp, L., Lima, I., Doney, S. C., John, J., and Beaulieu, C. (2010). Detection of anthropogenic climate change in satellite records of ocean chlorophyll and productivity. Biogeosciences, 7: 621-640. doi:10.5194/bg-7-621-201.

37 S. Roweis and Z. Ghahramani (1999). A Unifying Review of Linear Gaussian Models. Neural Computation, 11 :305–345.

38 Rabiner, L. R., (1989). A tutorial on hidden Markov models and selected applicationsin speech recognition. Proceedings of the IEEE 77:257–286. xlii, 75, 79, 227, 229

39 B.H. Juang, and L.R. Rabiner (1991). “Hidden Markov models for speech recognition”, Technometrics, 33, 251-272.

40 Weatherhead, E.C., Reinsel, G.C., Tiao, G.C., Meng, X.L., Choi, D.S., Cheang, W.K., Keller, T., Deluisi, J., Wuebbles, D.J., Kerr, J.B., Miller, A.J., Oltmans, S.J., Frederick, J.E. (1998). Factors affecting the detection of trends: Statistical considerations and applications to environmental data. Journal of Geophysical Research-Atmospheres, 103(D14).

41 Russell D, Mackinnon, J. G. (1993). Estimation and inference in econometrics. Oxford University Press.

42 Prais, S. J.; Winsten, C. B. (1954). Trend Estimators and Serial Correlation. Discussion paper 383, cowls commission

43 Fisher, R. (1925). Statistical Methods for Research Workers. Edinburgh: Oliver & Boyd.

44 Samson, C., Blanc-Féraud, L., Aubert, G., & Zerubia, J. (2000). A level set model for image classification. International Journal of Computer Vision, 40(3), 187-197.

45 Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity.

46 Mandelbrot, B. B. (1982). The Fractal Geometry of Nature, W. H. Freeman, 2nd ed., 460 pp., New York.

47 Torrence, C, Webster P. J. (1998).Interdecadal Changes in the ENSO-Monsoon System. Journal of Climate, 12(8).

48 Kestin, Tahl S., David J. Karoly, Jun-Ichi Yano, Nicola A. Rayner, (1998) Time–Frequency Variability of ENSO and Stochastic Simulations’. Journal of Climate, 11, 2258–2272.

49 Compo, G. P., Sardeshmukh, P. D., & Penland, C. (2001). Changes of subseasonal variability associated with El Niño. Journal of climate, 14(16), 3356-3374.

50 Latif, M., & Barnett, T. P. (1996). Decadal climate variability over the North Pacific and North America: Dynamics and predictability. Journal of Climate, 9(10), 2407-2423.

51 Kirtman, B. P., & Schopf, P. S., (1998). Decadal variability in ENSO predictability and prediction. Journal of Climate, 11(11), 2804-2822.

52 An, S. I., & Wang, B. (2000). Interdecadal Change of the Structure of the ENSO Mode and Its Impact on the ENSO Frequency. Journal of Climate, 13(12), 2044-2055.

53 Barker, K., Mazeran, C., Lerebourg, C., Bouvet, M., Antoine, D., Ondrusek, M., Zibordi, Lavender, S. (2008), MERMAID : The MEris MAtchup In-situ Database,proceedings of the 2nd MERIS/(A)ATSR User Workshop, ESA/ESRIN, Italy, September 2008.

54 H. Schiller, R. Doerffer (1999). Neural network for emulation of an inverse model operational derivation of Case II water properties from MERIS data. International Journal of Remote Sensing, 20(9), 1735-1746.

55 Von Mises, R. (1964). Mathematical theory of probability and statistics (Vol. 75). H. Geiringer (Ed.). New York: Academic Press.

56 Nordhaus, William D., and Joseph G. Boyer (1998). Requiem for Kyoto: an economic analysis of the Kyoto Protocol. Cowles Foundation Discussion Paper.

57 Meyer, K., & Thompson, R. (1984). Bias in variance and covariance component estimators due to selection on a correlated trait. Zeitschrift für Tierzüchtung und Züchtungsbiologie, 101(1-5), 33-50.

58 Philander S.G. (1990). El Niño, La Niña, and the Southern Oscillation. International Geophysics (46).

59 Torrence, C., and P. J. Webster (1998). The annual cycle of persistence in the El Niño–Southern Oscillation Quarterly Journal of the Royal Meteorological Society, 124, 1985–2004.

60 Clifford, P., Richardson, S., and Hemon, D. (1989). Assessing the significance of the correlation between two spatial processes. Biometrics, 45(1): 123 – 134.

61 Dutilleul, P. (1993). Modifying the t-test for assessing the correlation between two spatial processes. Biometrics, 49(1): 305 – 314.

62 McClain, C.R. (2009). A decade of satellite ocean color observations, Annual Review of Marine Science. 1 1: 19-42. Carlson, C.A.; Giovannoni, S.J. (Ed.)

63 Maritorena S., Hembise Fanton d'Andon O., Mangin A., Siegel D.A. (2010). Merged Satellite Ocean Color Data Products Using a Bio-Optical Model: Characteristics, Benefits and Issues. Remote Sensing of Environment, 114, 8: 1791-1804 (doi: 10.1016/j.rse.2010.04.002)

64 O'Reilly, J. E., Maritorena, S., Mitchell, B. G., Siegel, D. A., Carder, K. L., Garver, S. A., et al. (1998). Ocean color algorithms for SeaWiFS. Journal of Geophysical Research, 103, 24937−24953.

65 Morel, A., Gentili, B., Chami, M., & Ras, J. (2006). Bio-optical properties of high chlorophyll Case 1 waters, and of yellow substance-dominated Case 2 waters. Deep-Sea Research, 53, 1439 − 1459.

66 Hooker, S. B., Esaias, W. E., Feldman, G. C., Gregg, W. W. and C.R. Mc Clain (1992). An overview of SeaWiFS and ocean color. In S. B. Hooker, and E. R. Firestone, SeaWiFS Technical report series, NASA Tech. Memo. 104566, Greenbelt, Maryland, NASA Goddard Space Flight Centre.

67 Rast, M., Bezy, J. L., & Bruzzi, S. (1999). The ESA Medium Resolution Imaging Spectrometer MERIS a review of the instrument and its mission. International Journal of Remote Sensing, 20(9), 1681-1702.

68 Salomonson, V. V., Toll, D. L. and W.T. Lawrence (1992). Moderate resolution imaging spectroradiometer (MODIS) and observations of the land surface. Proc. "1992 International Geoscience and Remote Sensing Symposium" (IGARSS'92), Houston, Texas.

69 Evans, R.H. and Gordon, H.R. (1994). Coastal zone color scanner 'system calibration': A retrospective examination. Journal of Geophysical Research, 99: 7293-7307.

70 Gregg, W. W., Casey, N. W, and McClain, C. R. (2005). Recent trends in global ocean chlorophyll. Geophysical Research Letter, 32, L03606, doi:10.1029/2004GL021808.

71 Vantrepotte, V., and Mélin, F. (2009). Temporal variability of 10-year global SeaWiFS time series of phytoplankton chlorophyll a concentration. ICES Journal of Marine Science, 66: 1547–1556.

72 Pezzulli, S., Stephenson, D. B., and Hannachi, A. (2005). The variability of seasonality. Journal of Climate, 18: 71–88

73 Antoine, D., Morel, A., Gordon, H. R., Banzon, V. F., and Evans, R. H. (2005). Bridging ocean color observations of the 1980s and 2000s in search of long-term trends. Journal of Geophysics Research, 110.

74 C. Aitken (1935). On Least Squares and Linear Combinations of Observations, Proceedings of the Royal Society of Edinburgh, vol. 55, pp. 42–48.

75 Prais, S. J.; Winsten, C. B. (1954). Trend Estimators and Serial Correlation. Discussion paper 383, cowls commission

76 Johnson, B.C., Bruce, S.S., Early., E.A., Houston, J.M., O'Brian, T.R., Thompson, A. , Hooker, S.B., and Mueller, J.L. (1996). The Fourth SeaWViFS Intercalibration Round-Robin Experiment (SIRREX-4), May 1995. NASA Tech. Memo. 104566, Vol. 37.

77 Scherrer, B. (1984). Biostatistique. Gaëtan Morin Ed., Boucherville. xix + 850 p.

78 Haan, C.T. (1977). Statistical methods in hydrology, The Iowa State University Press, Ames, 378 p.

79 Berk, A., Anderson, G. P., Bernstein, L. S., Acharya, P. K., Dothe, H., Matthew, M. W., ... & Hoke, M. L. (1999, October). MODTRAN4 radiative transfer modeling for atmospheric correction. In SPIE's International Symposium on Optical Science, Engineering, and Instrumentation (pp. 348-353). International Society for Optics and Photonics.Bhat, H. S., & Kumar, N. (2010). On the derivation of the Bayesian Information Criterion. School of Natural Sciences, University of California.

80 O'Reilly, J. E., Maritorena, S., Siegel, D. A., O'Brien, M. C., Toole, D., Mitchell, B. G., et al. (2000). Ocean color chlorophyll-a algorithms for SeaWiFS, OC2, and OC4: version 4. SeaWiFS postlaunch calibration and validation analyses, Part 3, NASA/TM 206892, Vol. 11 (pp. 9−23).

81 Morel, A., Huot, Y., Gentili, B., Werdell, P.J., Hooker, S.B. and B.A. Franz (2007). Examining the consistency of products derived from various ocean color sensors in open ocean (Case 1) waters in the perspective of a multi-sensor approach. Remote Sensing of Environment, 111, 69-88.

82 Blunden, J., D. S. Arndt, and M. O. Baringer (2011), State of the climate in 2010, Bull. Am. Meteorol. Soc., 92, S1–S236.

83 Turk, D., C. S. Meinen, D. Antoine, M. J. McPhaden, and M. R. Lewis (2011). Implications of changing El Niño patterns for biological dynamics in the equatorial Pacific Ocean, Geophysical Research Letter, 38, L23603.

84 Compo GP, Sardeshmukh PD (2010). Removing ENSO-related variations from the climate record. Journal of Climate, 23(8):1957–1978. doi: 10.1175/2009JCLI2735.1

85 Alexander, L. Matrosova, C. Penland, J. D. Scott, and P. Chang (2008). Forecasting Pacific SSTs: Linear inverse model predictions of the PDO. Journal of Climate, 21, 385–402.

86 Boyce, D. G., Lewis, M. R. & Worm, B. (2010). Global phytoplankton decline over the past century. Nature, 466, 591-596, doi:10.1038/nature09268.

87 Polovina J. J, Howell E. A., Abecassis, M., (2008), Ocean's least productive waters are expanding, Geophysical Research Letters. 35(3).

88 IOCCG (2010). Atmospheric Correction for Remotely-Sensed Ocean-Color Products. Wang, M. (ed.), Reports of the International Ocean-Color Coordinating Group, No. 10, IOCCG, Dartmouth, Canada.

89 Antoine, D., F. D'Ortenzio, S. B. Hooker, G Bécu, B. Gentili, D. Tailliez, and A. J. Scott (2008). Assessment of uncertainty in the ocean reflectance determined by three satellite ocean color sensors (MERIS, SeaWiFS and MODIS-A) at an offshore site in the Mediterranean Sea (BOUSSOLE project), Journal of Geophysical Research, 113, C07013.

90 Clark, D.K., M.A. Yarbrough, M. Feinholz, S. Flora, W. Broenkow, Y.S. Kim, B.C. Johnson, S.W. Brown, M. Yuen, and J.L. Mueller (2003). MOBY, a radiometric buoy for performance monitoring and vicarious calibration of satellite ocean color sensors: measurement and data analysis protocols, in Ocean Optics Protocols for Satellite Ocean Color Sensor validation, NASA Tech. Memo. 2003 – 211621/Rev4, VolVI, J.L. Mueller, G.S. Fargion and C.R. McClain Eds., NASA GSFC, Greenbelt, MD, 139 pp.

91 Saulquin B., Hamdi A, Gohin F., Populus J., Mangin A., Fanton d'Andon O (2012). Estimation of the diffuse attenuation coefficient KdPAR using MERIS and application to seabed habitat mapping. Remote Sensing of Environment, pp. 224-233.

92 Saulquin Bertrand, Gohin Francis, Garrello Rene (2011). Regional Objective Analysis for Merging High-Resolution MERIS, MODIS/Aqua, and SeaWiFS Chlorophyll-a Data From 1998 to 2008 on the European Atlantic Shelf. IEEE Transactions On Geoscience And Remote Sensing, 49(1), 143-154.

93 Sarnaglia A,J. (2010). Estimation of periodic autoregressive processes in the presence of additive outliers. Journal of Multivariate Analysis archive, 101(9): 2168-2183.

94 Chen Ge, Shao Baomin, Han Yong, Ma Jun, Chapron Bertrand (2010). Modality of semiannual to multidecadal oscillations in global sea surface temperature variability, Journal of Geophysical Research-oceans, 115.

95 Björnsson, H. and Venegas S.A. (1997). A Manual of EOF and SVD Analysis of Climatic Data", McGill University.

96 Bretherton C.S., Smith C., Wallace J.M. (2002). An Intercomparison of Methods for finding Coupled Patterns in Climate Data, Journal of Climate, 5:541-560.

97 Jackson D. A., Chen, Y. (2004). Robust principal component analysis and outlier detection with ecological data. Environmetrics environmetrics, 15: 129-13.

98 Lau K. and Weng H. (1995). Climate signal detection using wavelet transform: How to make a time series sing, Bulletin of American Meteorology Society, 76:2391–2402.

99 Sonechkin, D. M. and Datsenko, N. M. (2000). Wavelet Analysis of Nonstationary and Chaotic Time-series with an Application to the Climate Change Problem. Pure and Applied Geophysics, 157:653–67.

100 Leeuwenburgh, O., Stammer D. (2001). The Effect of Ocean Currents on Sea Surface Temperature Anomalies, Journal of Physical Oceanography, 31:2340–2358.

101 Rouyer, T., Fromentin, J.-M., Stenseth, N. and Cazelles, B. (2008). Analysing multiple time series and extending significance testing in wavelet analysis, Marine Ecology Progress Series, 359:11-2.

102 Flannery, B., (2007). Gaussian Mixture Models and k-Means Clustering. Numerical Recipes: The Art of Scientific Computing (3rd ed), Section 16.1.

103 Enfield, D. B. and Mestas-Nunez, A. M. (1999). Multiscale Variabilities in Global Sea Surface Temperature and their Relationships with Tropospheric Climate Patterns. Journal of Climate, 12, 2719-2733.

104 Brockwell, P.J., and Davis, R.A. (2009). Time-series: Theory and Methods. 2nd ed. Springer.

105 Gu, D., and S. G. H. Philander (1995). Secular changes of annual and interannual variability in the Tropics during the past century, Journal of Climate, 8, 865–87.

106 Mallat S. G., Multifrequency channel decomposition of images and wavelet models (1989). IEEE Transactions on Acoustics, Speech, and Signal Processing, 37(12):2091-110.

107 Levitus S. (1987). A comparison of the annual cycle of two sea surface temperature climatologies of the World Ocean. Journal of Physics and Oceanography, 17 (2): 197–214,.

108 Sarnaglia A,J. (2010). Estimation of periodic autoregressive processes in the presence of additive outliers. Journal of Multivariate Analysis archive, 101(9): 2168-2183.

109 Hoaglin, D. C., Mosteller F. and Tukey J. (2000). Understanding Robust and Exploratory Data Analysis. John Wiley & Sons. 404–414.

110 Fablet R. (2012). Multiscale geometric deformations along planar curves: application to satellite tracking and ocean observation data, Submitted to IEEE Transactions on Geoscience and Remote Sensing.

111 David G. L. (2004). Distinctive Image Features from Scale-Invariant Keypoints". International Journal of Computer Vision.

112 Monasse, P., Guichard, F. (2000). Fast Computation of a Contrast Invariant Image Representation", IEEE transactions on image processing, 9: 860-872.

113 Dempster, A.P. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm", Journal of the Royal Statistical Society, 1(39): 1–38.

114 Kilpatrick, K.A., Podesta, G.P. and Evans, R., (2001). Overview of the NOAA/NASA advanced very high resolution radiometer Pathfinder algorithm for sea surface temperature and associated match-up database. Journal of Geophysical Research, 106, pp. 9179–9197.

115 Narapusetty, Balachandrudu, Timothy DelSole, Michael K. Tippett (2009). Optimal Estimation of the Climatological Mean. Journal of Climate, 22, 4845–4859.

116 Krishnamurthy, V., Ben P. Kirtman (2009). Relation between Indian Monsoon Variability and SST, Journal of Climate, 22, 4437–4458.

117 Wang B., Yang J., Zhou T., Wang B. (2008). Interdecadal Changes in the Major Modes of Asian–Australian Monsoon Variability: Strengthening Relationship with ENSO since the Late 1970s. Journal of Climate, 21:8, 1771-1789.

118 Hastenrath, S., L, De Castro, C. and Aceituno P. (1987). The Southern Oscillation in the Atlantic sector. Contributions to Atmospheric Physics, 60:41987 47–463.

119 Rouault, M., B. Pohl and P. Penven (2010). Coastal oceanic climate change and variability from 1982 to 2009 around South Africa. African Journal of Marine Science, 32(2): 237–246.

120 Pedgley, D.E., Reynolds, D.R. & Tatchell, G.M. (1995). Long-range insect migration in relation to climate and weather: Africa and Europe. Insect Migration: Tracking Resources Through Space and Time. Cambridge University Press, pp. 3–29.

121 Wang X., Wang C., Zhou W., Wang D., Song J. (2010). Teleconnected influence of North Atlantic sea surface temperature on the Niño onset. Climate Dynamics Online publication.

122 Alexander, L. Matrosova, C. Penland, J. D. Scott, and P. Chang (2008). Forecasting Pacific SSTs: Linear inverse model predictions of the PDO, Journal of Climate, 21, 385–402.

123 Mantua, Nathan J.. (1997). A Pacific interdecadal climate oscillation with impacts on salmon production. Bulletin of the American Meteorological Society 78 (6): 1069–1079.

124 Knight, J. R., C. K. Folland, and A. A. Scaife (2006). Climate impacts of the Atlantic Multidecadal Oscillation. Geophysical Research Letter, 33, L17706.

125 Lett C., Penven P., Ayón P., Fréon P. (2007). Enrichment, concentration and retention processes in relation to anchovy (Engraulis ringens) eggs and larvae distributions in the northern Humboldt upwelling ecosystem. Journal of Marine Systems, Volume 64, Issues 1–4, January 2007, Pages 189-200.

126 Raillard N., Ailliot P., Yao J.F. (2014). Modelling extreme values of processes observed at irregular time step. Application to significant wave height. To appear in the Annals of Applied Statistics.

127 Ailliot P., Maisondieu C., Monbet V. (2013), Dynamical partitioning of directional ocean wave spectra. Probabilistic Engineering Mechanics, 33, pp. 95-102

128 Wright, C. J., Scott, R. B., Furnival, D., Ailliot, P., Vermet, F. (2013), Global Observations of Ocean-Bottom Subinertial Current Dissipation. Journal of Physical Oceanography, 43, pp. 402-417.

129 Ailliot P., Monbet V., (2012), Markov-switching autoregressive models for wind time series. Environmental Modelling & Software, 30, pp 92-101.

130 Carrère, L., & Lyard, F. (2003). Modeling the barotropic response of the global ocean to atmospheric wind and pressure forcing-comparisons with observations. Geophysical Research Letters, 30(6).

131 P. Lazure, V. Garnier, F. Dumas, C. Herry, M. Chifflet (2009). Development of a hydrodynamic model of the Bay of Biscay. Validation of hydrology. Continental Shelf Research, 29(8), 985-997.

132 L. Debreu, P. Marchesiello, P. Penven and G. Cambon (2012). Two-way nesting in split-explicit ocean models: algorithms, implementation and validation. Ocean Model., 49-50, 1-21.

133 A. Sottolichio, P. Le Hir, P. Castaing (2000). Modeling mechanisms for the turbidity maximum stability in the Gironde estuary, France. Coastal and Estuarine Fine Sediment Processes, pp 373-386.

134 Rivier, A., Gohin, F., Bryère, P., Petus, C., Guillou, N., & Chapalain, G. (2012). Observed vs. predicted variability in non-algal suspended particulate matter concentration in the English Channel in relation to tides and waves. Geo-Marine Letters, 32(2), 139-151.

135 William E. Wecker, Craig F. Ansley (1983). The Signal Extraction Approach to Nonlinear Regression and Spline Smoothing. Journal of The American Statistical Association 78(381):81-89, 1983.

136 Box, George; Jenkins, Gwilym (1976). Time Series Analysis: forecasting and control, rev. ed., Oakland, California: Holden-Day.

137 Tesfaye, Y. G., Meerschaert, M. M., and Anderson, P. L. (2006). Identification of periodic autoregressive moving average models and their application to the modeling of river flows. Water Resources Research, 42(1).

138 Krasnopolsky, V. M., & Schiller, H. (2003). Some neural network applications in environmental sciences. Part I: forward and inverse problems in geophysical remote measurements. Neural Networks, 16(3), 321-334.

139 Chih-Chung C., Chih-Jen L (2011). LIBSVM: A library for support vector machines Transactions on Intelligent Systems and Technology (TIST).

140 P. Ailliot, V., Monbet (2012). Markov-switching autoregressive models for wind time series. Environmental Modelling & Software, 30, 92-101.

141 Gohin, F., Loyer, S., Lunven, M., Labry, C., Froidefond, J. M., Delmas, D., ... & Herbland, A. (2005). Satellite-derived parameters for biological modelling in coastal waters: Illustration over the eastern continental shelf of the Bay of Biscay. Remote Sensing of Environment, 95(1), 29-46.

142 D. Doxaran (2002). Télédétection et modélisation numérique des flux sédimentaires dans l'estuaire de la Gironde, PhD thesis, University Bordeaux 1, France, 326 pp.

143 Doxaran, D., Froidefond, J. M., Castaing, P., & Babin, M. (2009). Dynamics of the turbidity maximum zone in a macrotidal estuary (the Gironde, France): Observations from field and MODIS satellite data. Estuarine, Coastal and Shelf Science, 81(3), 321-332.

144 Bowers, D. G., & Binding, C. E. (2006). The optical properties of mineral suspended particles: A review and synthesis. Estuarine, Coastal and Shelf Science, 67(1), 219-230.

145 Eisma, D., Bernard, P., Cadée, G. C., Ittekkot, V., Kalf, J., Laane, R. W. P. M., ... & Schuhmacher, T. (1991). Suspended-matter particle size in some west-European estuaries; Part II: A review on floc formation and break-up. Netherlands Journal of Sea Research, 28(3), 215-220.

146 DeSarbo, W. S., & Cron, W. L. (1988). A maximum likelihood methodology for clusterwise linear regression. Journal of classification, 5(2), 249-282.

147 Tandeo, P., Chapron, B., Ba, S., Autret, E., Fablet R. (2013). Segmentation of Mesoscale Ocean Surface Dynamics Using Satellite SST and SSH Observations Geoscience and Remote Sensing, IEEE Transactions on Volume, pp 1 – 9.

148 Frankignoul, C., & Hasselmann, K. (1977). Stochastic climate models, part II application to sea-surface temperature anomalies and thermocline variability. Tellus, 29(4), 289-305.

149 Cressie NAC, Wikle CK (2011). Statistics for Spatio-Temporal Data. Wiley; New York.

150 Abdi, H. (2007). Discriminant correspondence analysis. Encyclopedia of measurement and statistics, 270-275.

151 Ardhuin, F., Rogers, E., Babanin, A., Filipot, J. F., Magne, R., Roland, A., ... & Collard, F. (2009). Semi-empirical dissipation source functions for ocean waves: Part I, definition, calibration and validation. arXiv preprint arXiv:0907.4240.

152 A. Bentamy, A., & Fillon, D. C. (2012). Gridded surface wind fields from Metop/ASCAT measurements. International Journal of Remote Sensing, 33(6), 1729-1754.

153 Courants de marée et hauteurs d'eau. La Manche de Dunkerque à Brest. Service Hydrographique et Océanographique de la Marine, Brest, Rapport 564-UJA, 2000

154 Tong, H. (1990). Non-linear time series, a dynamical systems approach. Oxford University Press.

155 Hughes, J. P., & Guttorp, P. (1994). A class of stochastic models for relating synoptic atmospheric patterns to regional hydrologic phenomena. Water Resources Research, 30(5), 1535-1546.

156 Hughes, J. P., & Guttorp, P. (1994). Incorporating spatial dependence and atmospheric data in a model of precipitation. Journal of Applied Meteorology, 33(12), 1503-1515.

157 Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 1-38..

158 Castino, F., Festa, R., & Ratto, C. F. (1998). Stochastic modelling of wind velocities time series. Journal of Wind Engineering and industrial aerodynamics, 74, 141-151.

159 Bilmes, J. A. (1998). A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. International Computer Science Institute, 4(510), 126.

160 Krolzig H.M. (1997). Markov-switching vector Autoregressions. Modelling, statistical inference and applications to business cycle analysis. Lecture notes in economics and mathematical systems 454. Springer-Verlag, Berlin.

161 Cappé, O., Moulines, E., & Rydén, T. (2005). Inference in hidden Markov models (Vol. 6). New York: Springer.

162 Bhat, H. S., & Kumar, N. (2010). On the derivation of the Bayesian Information Criterion. School of Natural Sciences, University of California.

163 Binding, C. E., Bowers, D. G., & Mitchelson-Jacob, E. G. (2005). Estimating suspended sediment concentrations from ocean colour measurements in moderately turbid waters; the impact of variable particle scattering properties. Remote sensing of Environment, 94(3), 373-383.

164 Tolman, H. L. (2008). A mosaic approach to wind wave modeling. Ocean Modelling, 25(1), 35-47.

165 Gohin, F., Saulquin, B., Oger-Jeanneret, H., Lozac'h, L., Lampert, L., Lefebvre, A., & Bruchon, F. (2008). Towards a better assessment of the ecological status of coastal waters using satellite-derived chlorophyll-a concentrations. Remote Sensing of Environment, 112(8), 3329-3340.

166 Maritorena S., D.A. Siegel and A. Peterson. (2002). Optimization of a Semi-Analytical Ocean Color Model for Global Scale Applications. Applied Optics, 41(15): 2705-2714.

167 MERIS Level 2 Detailed Processing Model, Doc. no PO-TN-MEL-GS-0006, issue 7, revision 2, June 2005.

168 Doxaran, D., Froidefond, J. M., Lavender, S., & Castaing, P. (2002). Spectral signature of highly turbid waters: Application with SPOT data to quantify suspended particulate matter concentrations. Remote sensing of Environment, 81(1), 149-161.

169 Park, Y., De Cauwer, V., Nechad, B., & Ruddick, K. (2004). Validation of MERIS water products for Belgian coastal waters: 2002-2003.

170 Krasnopolsky, V. M., & Schiller, H. (2003). Some neural network applications in environmental sciences. Part I: forward and inverse problems in geophysical remote measurements. Neural Networks, 16(3), 321-334.

171 Frouin, R., & Pelletier, B. (2013). Bayesian Methodology for Ocean Color Remote Sensing.

172 Lenoble, J. (1985). Radiative transfer in scattering and absorbing atmospheres: standard computational procedures. Hampton, VA, A. Deepak Publishing, 1985, 314 p. No individual items are abstracted in this volume., 1.

173 Chang, C. C., & Lin, C. J. (2011). LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST), 2(3), 27.

174 McLachlan, G. J. (2004). Discriminant Analysis and Statistical Pattern Recognition. Wiley Interscience.

175 Santer, R., Carrere, V., Dubuisson, P., & Roger, J. C. (1999). Atmospheric correction over land for MERIS. International Journal of Remote Sensing, 20(9), 1819-1840.

176 Ruddick, K. G., De Cauwer, V., Van Mol, B. (2005). Use of the near infrared similarity reflectance spectrum for the quality control of remote sensing data, in Remote Sensing of the Coastal Oceanic Environment, edited by Frouin, Robert J., Babin, Marcel, Sathyendranath, Shubha. Proceedings of the SPIE, Volume 5885, pp. 1–12.

177 Gordon, H. R., & Wang, M. (1994). Influence of oceanic whitecaps on atmospheric correction of ocean-color sensors. Applied Optics, 33(33), 7754-7763.

178 Steinmetz, F., Deschamps, P. Y., & Ramon, D. (2011). Atmospheric correction in presence of sun glint: application to MERIS. Optics express, 19(10), 9783-9800.

179 Aiken, J., & Moore, G. (2000). Case 2 (S) Bright pixel atmospheric correction. MERIS ATBD, 2, 6-6.

180 Lin, C. J. (2007). Projected gradient methods for nonnegative matrix factorization. Neural computation, 19(10), 2756-2779.

181 Jia, S., & Qian, Y. (2009). Constrained nonnegative matrix factorization for hyperspectral unmixing. Geoscience and Remote Sensing, IEEE Transactions on, 47(1), 161-173

182 Frouin, R., Deschamps, P. Y., Gross-Colzy, L., Murakami, H., & Nakajima, T. Y. (2006). Retrieval of chlorophyll-a concentration via linear combination of ADEOS-II global imager data. Journal of oceanography, 62(3), 331-337.

183 Levenberg, K. (1944). A Method for the Solution of Certain Non-Linear Problems in Least Squares. Quarterly of Applied Mathematics 2: 164–168.

184 Harold W. Sorenson, (1980). Parameter Estimation: Principles and Problems, Marcel Dekker.

185 Bonnans, J. Frédéric; Gilbert, J. Charles; Lemaréchal, Claude; Sagastizábal, Claudia A. (2006). *Numerical optimization: Theoretical and practical aspects*. Universitext (Second revised ed. of translation of 1997 French ed.). Berlin: Springer-Verlag. pp. xiv+490.

186 Reynolds, D. A. (1995). Speaker identification and verification using Gaussian mixture speaker models. Speech communication, 17(1), 91-108.

187 Petersen, K. B., & Pedersen, M. S. (2008). The matrix cookbook. Technical University of Denmark, 7-15.

188 Laws, E. A. (1997). Mathematical methods for oceanographers: An introduction. John Wiley & Sons.

189 Werdell, P. J., & Bailey, S. W. (2005). An improved in-situ bio-optical data set for ocean color algorithm development and satellite data product validation. Remote Sensing of Environment, 98(1), 122-140.

190 Antoine, D., Chami, M., Claustre, H., d'Ortenzio, F., Morel, A., Bécu, G. & Adams, D. (2006). BOUSSOLE: a joint CNRS-INSU, ESA, CNES, and NASA ocean color calibration and validation activity.

191 Holben, B. N., Eck, T. F., Slutsker, I., Tanre, D., Buis, J. P., Setzer, A. & Smirnov, A. (1998). AERONET—A federated instrument network and data archive for aerosol characterization. Remote sensing of environment, 66(1), 1-16.

192 Petersen, W., Wehde, H., Krasemann, H., Colijn, F., & Schroeder, F. (2008). FerryBox and MERIS–Assessment of coastal and shelf sea ecosystems by combining in situ and remotely sensed data. Estuarine, Coastal and Shelf Science, 77(2), 296-307.

193 Ruddick K., De Cauwer V., Park Y. & Moore G. (2006). Seaborne measurements of near infrared water-leaving reflectance: The similarity spectrum for turbid waters. Limnology and Oceanography, Vol. 51(2), pp. 1167–1179

194 Pope, R. M., & Fry, E. S. (1997). Absorption spectrum (380–700 nm) of pure water. II. Integrating cavity measurements. Applied optics, 36(33), 810-8723.

195 Moore, G. F., Aiken, J., Lavender, S. (1999), The atmospheric correction scheme of water colour and the quantitative retrieval of suspended particulate matter in Case II waters : application to MERIS, International Journal of Remote Sensing, 20, 1713–1733.

196 SAABIO, http://kalicotier.gis-cooc.org/kalicotier_static/A1037-NT-010-ACR_v1.0.pdf

197 Morel, A., Claustre, H., Antoine, D., & Gentili, B. (2007). Natural variability of bio-optical properties in Case 1 waters: attenuation and reflectance within the visible and near-UV spectral domains, as observed in South Pacific and Mediterranean waters. Biogeosciences Discussions, 4(4), 2147-2178.

198 Bricaud, A., Morel, A., Babin, M., Allali, K., & Claustre, H. (1998). Variations of light absorption by suspended particles with chlorophyll a concentration in oceanic (case 1) waters: Analysis and implications for bio-optical models. Journal of Geophysical Research: Oceans (1978–2012), 103(C13), 31033-31044.

199 Davis, L. (Ed.). (1991). Handbook of genetic algorithms (Vol. 115). New York: Van Nostrand Reinhold.

200 Font, J., Camps, A., Borges, A., Martín-Neira, M., Boutin, J., Reul, N., Mecklenburg, S. (2010). SMOS: The challenging sea surface salinity measurement from space. Proceedings of the IEEE, 98(5), 649-665.

8 List of publications and communications during the thesis

Publications

- Saulquin, B., Fablet, R., Mangin, A., Mercier, G., Antoine, D., & Fanton d'Andon, O. (2013). Detection of linear trends in multisensor time series in the presence of autocorrelated noise: Application to the chlorophyll-a SeaWiFS and MERIS data sets and extrapolation to the incoming Sentinel 3-OLCI mission. Journal of Geophysical Research: Oceans, 118(8), 3752-3763.
- Saulquin, B.; Fablet, R.; Mercier, G.; Demarcq, H.; Mangin, A, Fanton d Andon, O., Multiscale Event-Based Mining in Geophysical Time Series: Characterization and Distribution of Significant Time-Scales in the Sea Surface Temperature Anomalies Relative to ENSO Periods from 1985 to 2009, *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE*, vol.PP, no.99, pp.1,10 doi: 10.1109/JSTARS.2014.2329921
- Saulquin B., Fablet R., Ailliot P., Mercier G., Doxaran D., Mangin A., Fanton d'Andon O Characterization of time-varying regimes in remote sensing time series: application to the forecasting of satellite-derived suspended matter concentrations, In publication at JSTARS.
- Saulquin, B., Fablet, R., Mangin, A., Mercier, G., Fanton d'Andon, O. (submitted at RSE, October 2014).MEETC2: Ocean Color Atmospheric corrections in coastal complex waters using a Bayesian latent class model and potential for the incoming Sentinel 3 - OLCI mission.
- Saulquin B., Hamdi A, Gohin F., Populus J., Mangin, A; Fanton d Andon, Estimation of the diffuse attenuation coefficient KdPAR using MERIS and application to seabed habitat mapping. *Remote Sensing of Environment* (2012), pp. 224-233, DOI information: 10.1016/j.rse.2012.10.002.

Presentations in workshops

- French national workshop on ocean color (GIS COOC) 30 Janvier 2014. http://www.gis-cooc.org/index.php?option=com_content&view=article&id=113%3Aatelier-2014&catid=51&Itemid=96&lang=en
- Présentations au colloque "Analyse de données spatio-temporelles en océano-météo", 28-29 novembre 2013, Landéda

- Présentation au colloque "Analyse de données spatio-temporelles en océano-météo", 03-05 juillet 2013, Ile de Berder.

- Workshop ESA, Leaving planet, 02/2013

- International Ocean Colour Science Meeting, 6-8 May 2013, Darmstadt

- French national workshop on ocean color (GIS COOC) 23 et 24 janvier 2012, http://www.gis-cooc.org/index.php?option=com_content&view=article&id=99&Itemid=91&lang=en

- International Workshop on Temporal Analysis of Satellite Images, Greece May 23-25, 2012 http://www.earsel.org/SIG/timeseries/pdf/Temp2012-Program-AbstractBook-Final.pdf

9 Annex

Remote Sensing of Environment 128 (2013) 224–233

Contents lists available at SciVerse ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

Estimation of the diffuse attenuation coefficient K_{dPAR} using MERIS and application to seabed habitat mapping

Bertrand Saulquin [a,*], Anouar Hamdi [b], Francis Gohin [c], Jacques Populus [c], Antoine Mangin [a], Odile Fanton d'Andon [a]

[a] ACRI-ST, Sophia-Antipolis, France
[b] Institut des Milieux Aquatiques, Bayonne, France
[c] Ifremer, Brest, France

ARTICLE INFO

Article history:
Received 25 August 2011
Received in revised form 27 September 2012
Accepted 6 October 2012
Available online xxxx

Keywords:
Ocean color
Light attenuation
Photosynthetic available radiation
Euphotic depth
Seabed mapping

ABSTRACT

The availability of light in the water column and at the seabed determines the euphotic zone and constrains the type and the vertical distribution of algae species. Light attenuation is traditionally quantified as the diffuse attenuation coefficient of the downwelling spectral irradiance at wavelength 490 nm (K_{d490}) or the photosynthetically available radiation (K_{dPAR}). Satellite observations provide global coverage of these parameters at high spatial and temporal resolution and several empirical and semi-analytical models are commonly used to derive K_{d490} and K_{dPAR} maps from ocean colour satellite sensors. Most of these existing empirical or semi-analytical models have been calibrated in open ocean waters and perform well in these regions, but tend to underestimate the attenuation of light in coastal waters, where the backscattering caused by the suspended matters and the absorption by the dissolved organic matters increase light attenuation in the water column.

We investigate two relationships between K_{dPAR} and K_{d490} for clear and turbid waters using MERIS reflectances and the spectral diffuse attenuation coefficient $K_d(\lambda)$ developed by Lee (2005). Satellite-derived fields of K_{d490} and modelled K_{dPAR} are evaluated using coincident in-situ data collected over the world in both clear and turbid waters, and by using Ecolight simulations. Temporal means at 250 m resolution of K_{dPAR} and euphotic depth were computed over the period 2005–2009 for European coastal waters. These mean data were cross-tabulated with in-situ data of kelp (*Laminaria hyperborea*) and seagrass (*Posidonia oceanica*), respectively observed at locations on Atlantic and Mediterranean shores where the light is taken as the limiting factor to the depth distribution for these species. The minima observed for *P. oceanica*, in percent of energy, are very close to 1% of surface irradiance, the historical threshold known as euphotic depth as defined by Ryther (1956). Real estimates of the surface irradiance (Frouin, 1989) are used in conjunction with the estimated K_{dPAR} to calculate the residual energy at the lower limit of *P. oceanica* and *L. hyperborea* in mol·photons·m^{-2}·day^{-1} as a complement to the usual fraction of the surface energy. We show that the observed values, in terms of energy, for both species were equivalent to the values reported in the literature.

1. Introduction

The light available in the water column at wavelengths between 400 and 700 nm in the visible part of the spectrum, termed photosynthetically active radiation (PAR), is utilised by phytoplankton for photosynthesis (Falkowski & Raven, 1997; Kirk, 1994) and constrains the type and distribution of algae species and benthic algae, which contribute significantly to total primary production (Cahoon et al., 1993; McMinn et al., 2005; Carter et al., 2005). The estimation of the light attenuation in the water column is also critical to understand physical processes such as the heat transfer in the upper layer of the ocean (Lewis et al., 1990; Morel & Smith, 1974; Sathyendranath et al., 1991; Rocheford et al., 2001; Wu et al., 2007). From an optical perspective, in addition to pure water, light attenuation is constrained by the concentration of three components (IOCCG Report 3, 2000): pigments, expressed here as the concentration of chlorophyll-a ([Chl-a]), dissolved yellow substances (gelbstoff or CDOM) absorption a_{cdm} and suspended particulate matter concentration ([SPM]). The in-situ spectral diffuse attenuation coefficient $K_d(\lambda)$ was traditionally measured by the ocean-colour scientific community at 490 nm (K_{d490}), following the primary studies in the 1970s (Jerlov, 1976). Concurrently, biologists have focused on the PAR measurement and attenuation (K_{dPAR}). Both K_{dPAR} and K_{d490} increase with increasing solar zenith angle and K_{dPAR} is significantly depth dependent (the longer wavelength, red in this example, is rapidly attenuated in the water column relatively to the shorter wavelength blue) even for well-mixed waters.

Since the launch of the Coastal Zone Color Scanner (CZCS) in 1978, the ocean-colour community has provided maps of K_{d490} or K_{dPAR} at large spatial scales offering a great improvement in spatial and temporal

* Corresponding author. Tel.: +33 492967128.
E-mail address: bertrand.saulquin@acri-st.fr (B. Saulquin).

0034-4257/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.rse.2012.10.002

resolution compared to in-situ data. Space based sensors measure top-of-atmosphere radiances at different wavelengths and the Medium Resolution Imaging Spectrometer (MERIS) sensor has 15 bands between 412 and 865 nm. The contribution from the atmosphere is firstly removed from the top-of-atmosphere radiance, through a process known as atmospheric correction (Gordon & Wang, 1994), to obtain the water-leaving radiance (Lw). The Lw values are normalised, i.e. expressed in standard solar conditions (sun at zenith) in the absence of the atmosphere, and corrected for bidirectional effects (viewing angle dependence and effects of seawater anisotropy, Morel et al., 2002) to obtain the normalised water-leaving radiance (nLw). Today, several empirical and semi-analytical models of K_{d490} and K_{dPAR} are commonly used to derive K_{d490} maps from satellite-derived nLw.

Mueller (2000) defines an empirical relationship between K_{d490} and the ratio between blue and green water-leaving radiances from the Sea-viewing Wide Field-of-view Sensor (SeaWiFS) (McClain et al., 2004), and the Moderate Resolution Imaging Spectroradiometer (MODIS) (Esaias et al., 1998). Morel et al. (2007) proposes an empirical relationship between the K_{d490} and the Chl-a concentration. Lee et al. (2002, 2005a, 2005b, 2007) provided a semi-analytical model for K_{d490} with dedicated versions for SeaWiFS, MERIS and MODIS nLw.

K_{dPAR} has historically been expressed as a function of [Chl-a] (Morel, 1988) for clear open ocean waters. This latest approach is routinely used in the open ocean where phytoplankton is the main contributor to attenuation (Claustre & Maritorena, 2003). In coastal waters however, the determination of K_{dPAR} is complicated by increased light attenuation by CDOM and SPM(Case 2 waters). In coastal areas regional approaches express K_{dPAR} as a function of the [Chl-a], and [SPM] (Devlin et al., 2009; Gohin et al., 2005). More recently, K_{dPAR} is more often related to K_{d490} using empirical approaches and the relationship between K_{d490} and K_{dPAR} has quite large regional variations (Barnard et al., 1999; Pierson, 2008; Morel et al., 2007; Pierson et al., 2008; Wang et al., 2009; Zaneveld et al., 1993).

In this paper, we show the performance of three models of K_{d490} (Lee et al., 2005a, 2005b; Morel et al., 2007; Mueller, 2000), routinely used as standard MERIS, SeaWiFS and MODIS Level 3 products, compared to an in-situ dataset collected near shore and in clear open ocean waters. We then derive two relationships between K_{dPAR} and K_{d490}, estimated by integrating the spectral irradiances over the euphotic depth and the visible spectrum using $K_d(\lambda)$ as estimated using Lee et al. (2005a, 2005b).

Our aim is to provide an estimation of K_{dPAR} for values greater than 0.06 m^{-1} and lower than 1 m^{-1}. For more turbid waters, dedicated algorithms may be used, and for oligotrophic waters (K_{dPAR}<0.06 m^{-1}), standard K_{dPAR} estimations (Morel et al., 2007; Mueller, 2000) are freely available at 4 km resolution on the Globcolour website (www.globcolour.info), and the oceancolor webpage (http://oceancolor.gsfc.nasa.gov/).

Secondly, temporal means of satellite derived K_{dPAR} and Z_{eu} were calculated for the European waters, from 2005 to 2009, to characterise a reference state for light and marine coastal fauna and flora in the intertidal zone at 250 m resolution.

Finally, six sites where selected by Ifremer in Corsica (Mediterranean Sea) and in Brittany (English Channel and Atlantic Ocean) to compare the satellite derived minimum light threshold values for *P. oceanica* and *L. hyperborea* to the literature. The threshold of 1% used to define Z_{eu} as the minimum light requirement for benthic primary production, was historically determined from in-situ observations of *P. oceanica* in the Mediterranean Sea. We therefore compare the satellite-derived 1% to the deepest depth at which *P. oceanica* is observed at the Corsican site. Nevertheless, some species can survive at lower light levels and the evaluation of the light available in fraction of the surface irradiance is biologically meaningless (Gattuso et al., 2006) as the fraction of moonlight is the same than the fraction of sunlight. Therefore, we propose the use of daily integrated PAR (Frouin et al., 1989) attenuated into the water column using K_{dPAR}, to arrive at an estimation of the PAR in the water column in mol·photons·m^{-2}·d^{-1}. This provides a more meaningful estimation of energy in the water column than fraction of the surface energy, generally used by the community.

2. Methods

The spectral diffuse attenuation coefficient $K_d(\lambda)$ is the coefficient of the exponential attenuation of the spectral downwelling irradiance:

$$Ed(\lambda) = E0(\lambda).e^{-Kd(\lambda)z}. \quad (1)$$

Here $E_d(\lambda)$ is the spectral downwelling irradiance in W.m^{-2}.nm^{-1} at depth z and wavelength λ and $E_0(\lambda)$ is the energy just beneath the surface). All symbols and acronyms cited in the text are summarized in Table 1 for a better understanding. If the visible spectral domain is considered, the PAR at depth z can be related to $K_d(\lambda)$ and $E_d(\lambda)$ using energetic (Eq. 2a) or quantum units (Eq. 2b.) (Baker & Frouin, 1987; Morel & Smith, 1974):

$$PAR(z) = \int_{400nm}^{700nm} Ed(\lambda; z=0).\exp^{-Kd(\lambda).z} d\lambda \left[W.m^{-2}\right] \quad (2a)$$

$$PAR(z) = \frac{1}{h.c}\int_{400nm}^{700nm} \lambda.Ed(\lambda; z=0).\exp^{-Kd(\lambda).z} d\lambda \left[photons.m^{-2}.s^{-1}\right]. \quad (2b)$$

An expression of the instantaneous $K_{dPAR}(z)$ is:

$$KdPAR(z) = -\frac{\ln(PAR(z+dz)) - \ln(PAR(z))}{dz}. \quad (3)$$

K_{dPAR} changes with depth as the red photons are absorbed in the top layers. The spectral diffuse attenuation coefficient of downwelling irradiance $K_d(\lambda)$ also changes with depth, but its magnitude of variation is significantly smaller than that of K_{dPAR} (Lee, 2009; Zaneveld et al., 1993). The Hydrolight/Ecolight (© Curtis D. Mobley, 2008) is a radiative transfer model that computes radiance distributions and related quantities (irradiance, reflectances, diffuse attenuation functions, etc.) in any water body starting from the Chl-a and SPM concentration and CDOM absorption. Fig. 1 shows two Ecolight simulations of K_{dPAR} for clear (blue plot) and coastal turbid waters (orange plot). In this simulation the water is assumed to be well mixed and scattering of particulates is based on the model of Gordon and Morel (1983). The sky is assumed to be cloudless

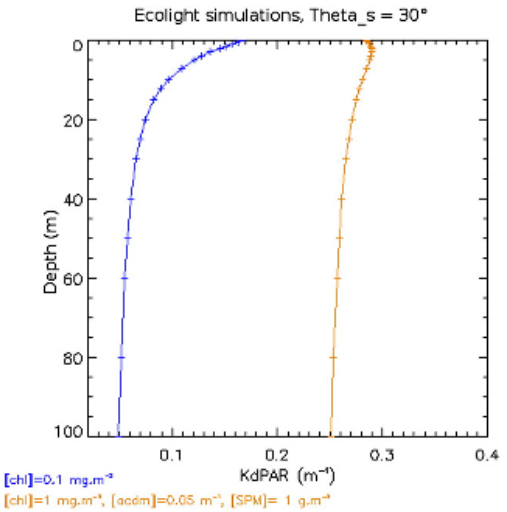


Fig. 1. Simulated $K_{dPAR}(z)$ in the water column using Ecolight for clear water with low [Chl-a] (case1, blue) and coastal water (case 2, orange). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

with the sun at 30° from the zenith. For coastal water simulation (orange), [Chl-a] is set to 1 $mg \cdot m^{-3}$, a_{cdom} to 0.05 m^{-1} and [SPM] to 1 $g \cdot m^{-3}$. The instantaneous K_{dPAR} (Fig. 1) is estimated using Eq. (3). Fig. 1 verifies that $K_{dPAR}(z)$ is more constant for coastal turbid waters (Wang et al., 2009).

We consider in this paper the vertical average value of K_{dPAR} between the surface and the euphotic depth, $\overline{KdPAR}$ (Eq. 4) because K_{dPAR} values reported in the literature or in-situ databases used for validation are estimated using Eq. (5) in this expression.

$$\overline{KdPAR} = \frac{\ln(PAR(0)) - \ln(PAR(z))}{z} \quad (4)$$

We use $z = Z_{eu}$ in this study. Using $\overline{KdPAR}$, instead of $K_{dPAR}(z)$ will lead to an accurate estimation of PAR near the surface and Z_{eu}. Between these two depths, PAR will be slightly over-estimated. Further in this paper, $\overline{KdPAR}$ is noted K_{dPAR}.

3. In-Situ data

3.1. K_{d490}, K_{dPAR} measurements

In-situ Ed(λ,z) or PAR(z) measurements must be collected following a community-vetted protocol, (Werdell & Bailey, 2005a, 2005b) to avoid ship shadow and reflectance. If required (not here), PAR irradiance data expressed in $W \cdot m^{-2}$ can be converted to molar units using the following approximation: 2.5×10^{18} quanta $\cdot s^{-1} \cdot W^{-1}$ or 4.2 $\mu E \cdot m^{-2} \cdot s^{-1} \cdot W^{-1}$ (Morel & Smith, 1974). In-situ data of K_{d490} and K_{dPAR} available through global datasets such as NOMAD (http://seabass.gsfc.nasa.gov/data/nomad_seabass_v2.a_2008200.txt), SeaBASS (http://seabass.gsfc.nasa.gov/) were extracted over the period 2005 to 2009.

Data from the instrumented buoy BOUSSOLE located near Villefranche (France) in the Mediterranean sea were also used (http://www.upmc.fr/en/research/living_earth_and_environment_section/laboratories/villefranche_sur_mer_oceanography_laboratory_umr_7093.html). Additional data in the Cheasapeake Bay, which is traditionally a turbid area (Wang & Shi, 2005), and data obtained from Ifremer and OPTIC-MED (2008) and OPTIC-PCAF (2004) cruises were also added as they provide some in-situ measurements on shores where SPM backscattering and CDOM absorption may be important.

In-situ K_{d490} and K_{dPAR} values reported in public databases are calculated using Eq. (4) and integrated over the first optical depth of E_{d490} (Z_{90}=1/K_{d490}) (Morel et al., 2007). To validate either satellite-derived K_{d490} or K_{dPAR}, we produced "matchups", i.e., data pairs of satellite-derived K_d and in-situ collocated in space (same pixel) and obtained during the same day. The satellite K_{d490} (Figs. 2 to 6) is directly comparable to the in-situ K_{d490}. We estimated, using Ecolight and the IOP available for the matchups from NOMAD and Seabass dataset, a correction for $K_{dPAR}(Z_{eu}) = 0.94 * K_{dPAR}(Z_{90})$ as we do not have the irradiance profiles to re-estimate $K_{dPAR}(Z_{eu})$. This correction is applied to Figs. 7a and 9. For OPTICs (12 matchups of Fig. 7) and Ifremer dataset (18 matchups of Fig. 7), the higher values of Fig. 7, we calculated $K_{dPAR}(Z_{eu})$ from the irradiance profiles and Eq. (4).

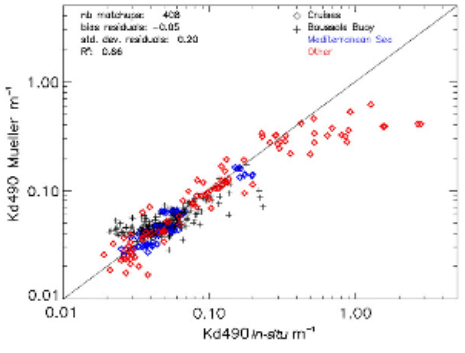


Fig. 2. Mueller's K_{d490} vs. in-situ.

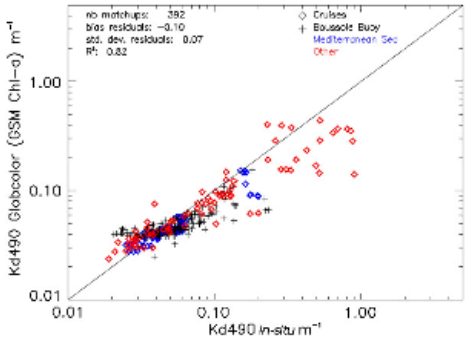


Fig. 3. Morel's K_{d490} vs. in-situ.

Matchups are used to produce statistical comparisons for the two fields. Bias and Pearson correlation coefficient (R) for Figs. 2 to 7 are calculated on log-transformed data.

3.2. Seagrass and kelp data

In-situ coverage of *P. oceanica* in Corsica and single beam sounder survey data acquired on rocky seabed covered by kelp (*L. hyperborea*) in Brittany (Méléder et al., 2010), are used to compare satellite-derived residual energy observed at the macrophytes lower limits to the known minimum thresholds reported in the literature. Six sites were selected by Ifremer according to accurate knowledge of species distribution and state of conservation, and the availability of an accurate bathymetry (resolution of 100 m horizontally and 1 to 5 m vertically).

Table 1
List of symbols and abbreviations.

Symbol	Definition	Unit
Lw	Water leaving radiance	$W \cdot m^{-2} \cdot sr^{-1} \cdot m^{-1}$
a(λ)	Absorption coefficient at wavelength λ	m^{-1}
bb (λ)	Backscattering coefficient at wavelength λ	m^{-1}
CDOM	Coloured dissolved organic matters	
Chl-a	Chlorophyll-a	
DTM	Digital terrain model	
GSM	Garver–Siegel–Maritorena	
Globcolour	Global ocean colour ESA funded project	
SPM	Suspended particulate matter	
$E_d(\lambda, z)$	Spectral downwelling irradiance at depth z	$W \cdot m^{-2} \cdot m^{-1}$
IOP	Inherent optical properties	
$K_d(\lambda, E\%)$	Spectral diffuse attenuation coefficient for downwelling irradiance between $E_d(\lambda,0)$ and % of E_d (λ,0)	m^{-1}
K_{dPAR}	Diffuse attenuation coefficient of PAR	m^{-1}
$\overline{KdPAR}$	Vertical average value of mean diffuse attenuation coefficient over the euphotic layer	m^{-1}
MERIS	Medium resolution imaging spectrometer	
MODIS	Moderate resolution imaging spectroradiometer	
PAR	Photosynthetically available radiation	photons $\cdot m^{-2} \cdot s^{-1}$ or $W \cdot m^{-2}$
Rrs	Remote sensing reflectance (ratio of water-leaving radiance to downwelling irradiance above the surface)	
SHOM	Service Hydrographique et Océanographique de la Marine	
SeaWiFS	Sea-viewing wide field-of-view sensor	
Z_{eu}	Euphotic depth	m
Z_{490}	Depth at which E(Z,490) = 1%E(0,490)	m
Z_{90}	First optical layer = $1/K_{d490}$	m
θs, Theta_s	Above surface solar zenith angle	Radians

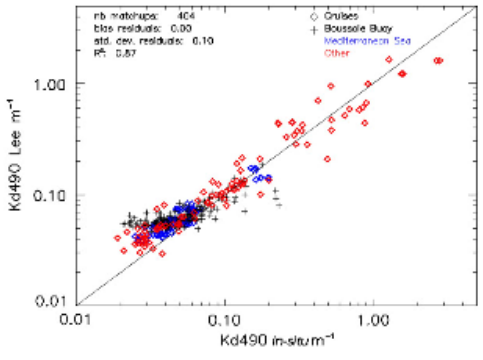


Fig. 4. Lee's K_{d490} vs. in-situ.

4. Satellite data

MERIS Level 2 Reduced Resolution (RR, 1 km resolution) data were used to match up with in-situ for the validation exercise. MERIS Full Resolution (FR) data were used to provide temporal means of Z_{eu} and K_{dPAR} over Europe. Coastal areas are characterised by strong gradients of Chl-a and SPM, which strongly affect the absorption and scattering of light. Therefore, the use of FR data when available is clearly relevant. The level 2 MERIS RR archive is available at ACRI-ST and MERIS FR data for Europe were downloaded from ESA facilities. Pixels flagged (MERIS Level 2 Detailed Processing Model) as CLOUD and HIGLINT were discarded. FR daily nLw were then projected on a regular grid of 250*250 m^2. Daily fields of K_{d490} and K_{dPAR} were subsequently calculated from nLw and temporally averaged over the period 2005 to 2009 as required by the EuseaMap project. Daily mean PAR (in $mol \cdot photons \cdot d^{-1} \cdot m^{-2}$) was evaluated using the algorithm developed by Frouin in 1989 and recently updated in 2011 for MERIS using Level 1 RR. The daily fields are averaged temporally over the period 2005 to 2009. Then the temporal averaged mean PAR is attenuated using the averaged K_{dPAR} at 250 m resolution and (Eq. 3) to provide an estimation of residual PAR in the water column in $mol \cdot photons \cdot m^{-2} d^{-1}$.

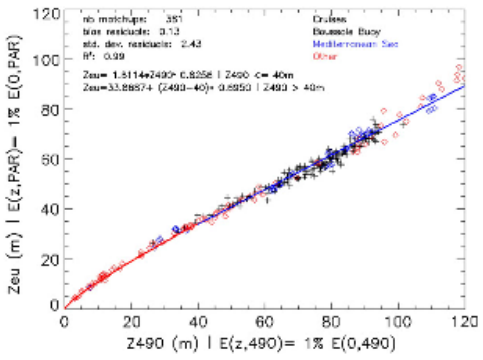


Fig. 5. Euphotic depth (Z_{eu}) related to Z_{490} (depth Z at which E(Z,490)) = 1%E(0,490) for the selected matchups.

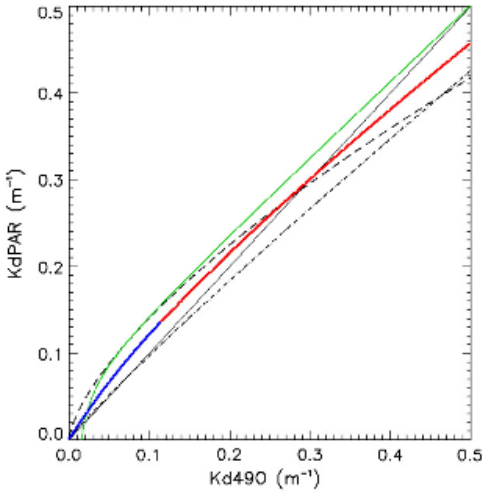


Fig. 6. Relationships between K_{dPAR} and K_{d490}. Blue curve for clear waters (Eq. 9b), red for coastal waters (Eq. 9a). The green curve shows the Morel (Eq. 10) for clear waters. The short black dotted curve for the Wang & Son's relationship for the Chesapeake bay (Eq. 11) and long black dashed line the relationship derived by Pierson and Kratzer for the Baltic Sea (Eq. 12). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

5. Results

5.1. Evaluation of existing K_{d490} model compared to our in-situ dataset

5.1.1. Mueller's algorithm

Mueller (2000) proposed an empirical model for non-turbid waters based on the ratio of the nLw at wavelengths 490 and 555 nm, i.e.:

$$Kd490 = Kw490 + A(nLw490/nLw555)^{B}. \quad (5)$$

$K_{w490} = 0.016\ m^{-1}$ is the diffuse attenuation coefficient for pure water. Parameter A was set initially to 0.15645 and B to −1.5401. Werdell (2005a) updated (Eq. 5) to improve the algorithm performance for the clearest ocean waters. K_{w490} was suppressed, A set to 0.1853 and B set to −1.349.

Fig. 2 shows that Mueller's K_{d490} estimation is accurate for clear water ($K_{d490} < 0.2\ m^{-1}$). Above $0.2\ m^{-1}$ the algorithm saturates and the K_{d490} is clearly under-estimated compared to the dataset used. 'Other' in Fig. 2 represents matchups not collected in the Mediterranean Sea. From this same dataset, the number of matchups may vary as we progress from Figs. 2 to 7 as the spectral bands used and the algorithms may be different.

5.1.2. Morel's approach

An empirical K_{d490} model based on chlorophyll-a concentration has been proposed by Morel in 2004. This model has been recently revised (Morel et al., 2007) using in-situ measurements from the NASA Bio-Optical Marine Algorithm Dataset (NOMAD) (Werdell & Bailey, 2005a, 2005b). The revised formula is given as:

$$Kd490 = 0.0166 + 0.0773.[Chl]^{0.6715}. \quad (6)$$

Fig. 3 shows that for $K_{d490} < 0.2\ m^{-1}$, the estimated K_{d490} fits the in-situ retrievals. For turbid $K_{d490} > 0.3\ m^{-1}$ the model underestimates the attenuation. We recall that the Mueller and Morel's algorithms have been calibrated and dedicated for open sea clear waters.

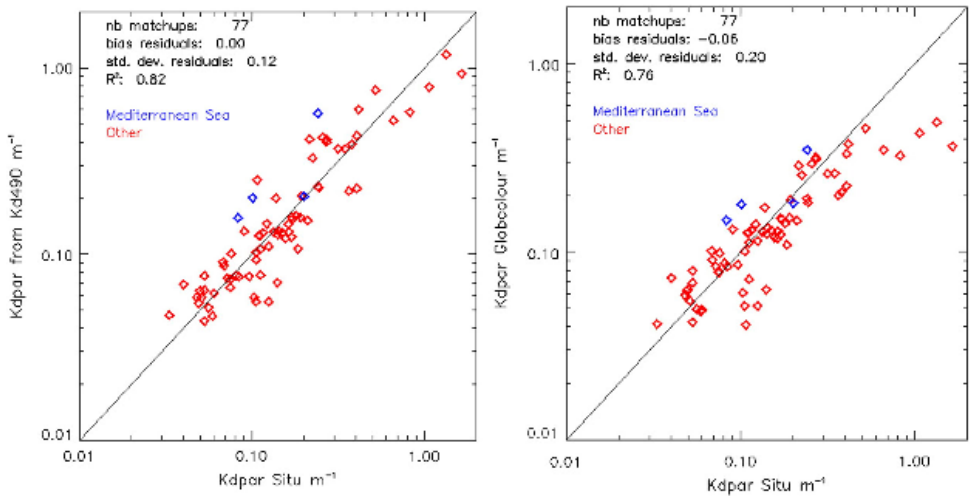


Fig. 7. a) satellite-derived K_{dPAR} from equations (Eqs. 9a and 9b) vs. in-situ K_{dPAR}. b) Globcolour standard K_{dPAR} (Eq. 10).

5.1.3. Lee's semi-analytical algorithm

Lee et al. (2005a, 2005b, 2009) proposed a semi-analytical approach to derive the mean Kd(λ) based on a radiative transfer model. The model has been revised recently (Lee et al., 2007), and Kd(λ,10%) i.e. integrated from the surface to the depth where (E(z, λ) = 10% E0(λ)) can be written as (Lee et al., 2005a):

$$Kd(\lambda,E10\%) = (1+0.005.\theta s).a(\lambda) + 4.18.\left(1-0.52.e^{-1.8\,a}\right).b_b(\lambda). \quad (7)$$

Where θs is the solar-zenith angle in the air, a(λ) the total absorption at λ and $b_b(\lambda)$ the total backscattering at λ. It is interesting to note that the semi-analytical approach developed by Lee allows the derivation of K_d at any wavelength. In our case, at λ = 490 nm, K_{d490} is derived from Eq. (7) and the absorption and backscattering coefficients at 490 nm, a(490) and $b_b(490)$ are themselves calculated using Lee's QAA v5 algorithm applied to the MERIS Rrs at wavelengths 443, 490, 555, and 670 nm.

We observe in Fig. 4 a good agreement between the satellite-derived K_{d490} and in-situ measurements between 0.06 and 1 m^{-1}. For very clear waters, K_{d490} tends to be overestimated when compared to our in-situ dataset. For in-situ K_{d490} greater than 0.08 m^{-1} the estimated K_{d490} compares well with the in-situ data.

5.2. From K_{d490} to K_{dPAR}

To derive the relationships between K_{d490} and K_{dPAR}, we have calculated using (Eq. 2b) the PAR values at the selected matchups at any depth, using a 0.1 m step for z, until PAR(z) = 1% PAR(0). $K_d(\lambda)$ at the wavelengths 412, 443, 489, 509, 559, 620, 664, and 709 were derived from Eq. (7) and applied to MERIS Rrs. $E_d(\lambda,z)$ were evaluated using Eq. (8) (Gordon & Wang, 1994), the theoretical extra-terrestrial solar irradiances $F_0(\lambda)$, the Rayleigh optical thicknesses T(λ) and the theoretical values of the ozone transmittance $T_{O3}(\lambda)$.

$$E0(z,\lambda) = F0(\lambda).\exp^{-T(\lambda)/2}.TO3(\lambda) \quad (8)$$

Two relationships were derived between Z_{eu} and the depth at which E(Z,490) = 1%E(0,490), Z_{490}, using the Lw observed at the in-situ matchups (Fig. 5). The two relationships between K_{dPAR} and K_{d490} are directly derived from the two relationships between Z_{eu} and Z_{490}. Relating K_{dPAR} to K_{d490} was not absolutely necessary as we could have integrated the spectral K_d provided by Lee using Eq. (2b). Nevertheless, we decided to propose a relationship between K_{dPAR} and K_{d490} as this link is meaningful and useful to derive K_{dPAR} from several K_{d490} in-situ datasets that are available. The threshold of 40 m for Z_{490} (K_{d490} = 0.115 m^{-1}) was set arbitrarily to separate clear from turbid waters.

An exponential model is fitted for turbid waters (Z_{490} < 40 m) and a linear model for clear waters ($Z_{490} \geq 40$ m). The proposed equations between K_{dPAR} and K_{d490} are shown in Fig. 6 (Eqs. 9a in blue and 9b in red).

$$K_{dPAR} = 4.6051.K_{d490}/(6.0700.K_{d490} + 3.200), \text{for} K_{d490} <= 0.115 m^{-1} \quad (9a)$$

$$K_{dPAR} = 0.8100.K_{d490}^{0.8256}, \text{for} K_{d490} > 0.115 m^{-1}. \quad (9b)$$

Morel et al. (2007) expressed K_{dPAR} as a function of Kd_{490} for clear waters:

$$K_{dPAR} = 0.0665 + 0.874.K_{d490} - 0.00121/K_{d490}. \quad (10)$$

Similar approaches to Eq. (9b) have been recently developed by Wang & Son (Eq. 11, 2009) and Pierson & Kratzer (Eq. 12, 2008) for respectively the Cheasapeake Bay turbid waters and Baltic Sea, where CDOM absorption is important.

$$K_{dPAR} = 0.8045.K_{d490}^{0.9170} \quad (11)$$

$$K_{dPAR} = 0.6677.K_{d490}^{0.6763} \quad (12)$$

Relationships between K_{dPAR} and K_{d490} directly depend on [Chl-a], a_{cdom} and [SPM]. In clear waters K_{d490} values are less than K_{dPAR} values as the attenuation is greatest in the red with a resulting stronger PAR attenuation (which includes the red). In coastal areas, Pierson (2007)

suggests that increasing a_{cdom} has the result of increasing more rapidly K_{d490} than K_{dPAR}.

Fig. 7 shows the scatterplot of the estimated K_{dPAR} (Fig. 7a) and the Globcolour (Fig. 7b) vs. in-situ data. The estimated K_{dPAR} is higher than the case 1 Globcolour standard algorithm for values greater than 0.3 (Fig. 6). Although the number of matchups available is small for $K_{dPAR} > 0.3\ m^{-1}$ we can observe the saturation effect on the standard K_{dPAR}. The number of available K_{dPAR} matchups is too small (Fig. 7, 77 matchups) and we propose therefore an alternative validation using Ecolight simulations (Fig. 8). The Ecolight configuration is provided in Appendix A. To obtain a realistic distribution of the IOPs we start from those gathered in the NOMAD dataset. The NOMAD dataset does not provide [SPM] and an estimation of this concentration was done using Babin et al. (2003):

$$[SPM] = 1.73/0.015.b_{bp443}. \quad (13)$$

B_{bp443} is the particular backscattering measured at 443 nm. The sun zenith angle, Θs, a required input parameter for Ecolight, was estimated for each in-situ data using the date, time, longitude and latitude. Finally the satellite-derived K_{dPAR} is compared to the K_{dPAR} estimated using Ecolight (K_{dPAR} is calculated from the PAR provided in the Ecolight output files and averaged using Eq. (4) and the depth at which $E(z) = 1\%\ E_0$).

Eqs. (9a, 9b) can also be used to derive an estimate of K_{dPAR} from K_{d490}. Fig. 9 shows a comparison for the NOMAD dataset between the K_{dPAR} estimated from the in-situ K_{d490} and the corresponding in-situ K_{dPAR}, i.e. a validation of Eqs. (9a, 9b) and Fig. 6.

Fig. 9 shows an overestimation for the very clear waters. As Lee's algorithm slightly overestimated K_{d490} for clear waters (Fig. 4) and K_{dPAR} is derived from satellite data and Lee's spectral K_d, this slight overestimation occurs for $K_{dPAR} < 0.1\ m^{-1}$, i.e. $Z_{eu} > 46$ m. For K_{dpar} greater than $0.1\ m^{-1}$ the estimated value fits to the in-situ data (Fig. 9).

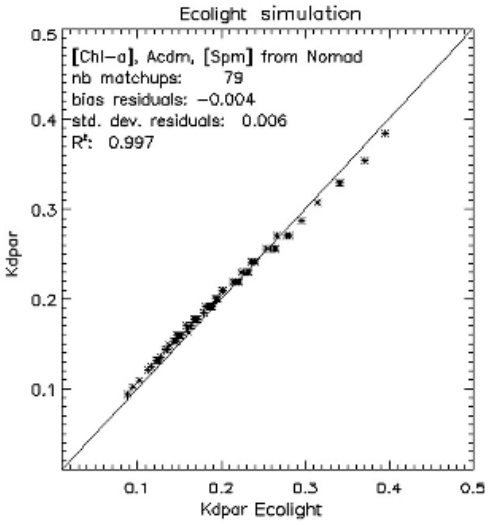


Fig. 8. Scatterplot between the Ecolight simulations and estimated K_{dPAR} based on K_{d490} by Lee et al. (2005a, 2005b).

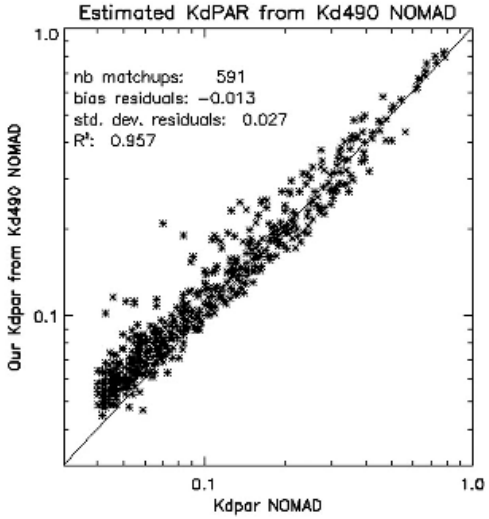


Fig. 9. Estimated K_{dPAR} from in-situ K_{d490} measurements compared to in-situ K_{dPAR}.

5.3. High resolution maps of Z_{eu} and K_{dPAR}

Fig. 10 shows the temporal mean of K_{dPAR} and Z_{eu} for Brittany and the Gulf of Lions at 250 m resolution. The covered area (Europe) by the EuSeaMap project was divided in 25 zones (not shown).

5.4. Application to seabed habitat mapping

5.4.1. P. oceanica in Corsica

In Corsica three sites where *P. oceanica* meadows are known to be in a natural state were selected for comparison. Fig. 11 shows the distribution of *P. oceanica* at two sites in north-west (Calvi) and north-east (south Bastia) Corsica. The orange line shows Z_{eu} ($1\%\ E_0$) estimated from MERIS 250 m over the period 2005–2009. It is interesting to note that the lower extension for *P. oceanica* follows the satellite-derived Z_{eu}.

Gattuso et al. (2006) proposed a light range of 0.1 to 2.8 mol·photons·m^{-2}·d^{-1} for the minimum requirements of *P. oceanica*. Table 2 shows for the 3 selected sites the observed value in percentage of the surface irradiance and energy at the lower limit of the *Posidonia* beds. Using GIS software, in-situ points (black dots, Fig. 11) were selected manually on fine scale *Posidonia* maps at locations representing the deep boundary of the meadows. Statistics (Table 2) were computed for depth, percentage of the surface irradiance and energy by retrieving at these locations the values of pixels from respectively a 100 m resolution depth DTM, the temporal mean at 250 m resolution of K_{dPAR} and the temporal mean at 1 km resolution of PAR. The observed mean values, weighted means for the 3 sites by the number of observations, are 0.94% and 0.26 mol·photons·m^{-2}·day^{-1} for *P. oceanica*. These values are very close to the 1% threshold and in the lower part of the energy range proposed by Gattuso et al. (2006).

5.4.2. Kelp in Brittany

In the same manner as the previous analysis, we have evaluated the minimum light requirements for kelp using single beam sounder acoustic data acquired in 2006 and 2007 at three sites in Brittany (les Abers in North Brittany, îles de Glénan and île de Groix in South Brittany). The bathymetry used here is calculated from the

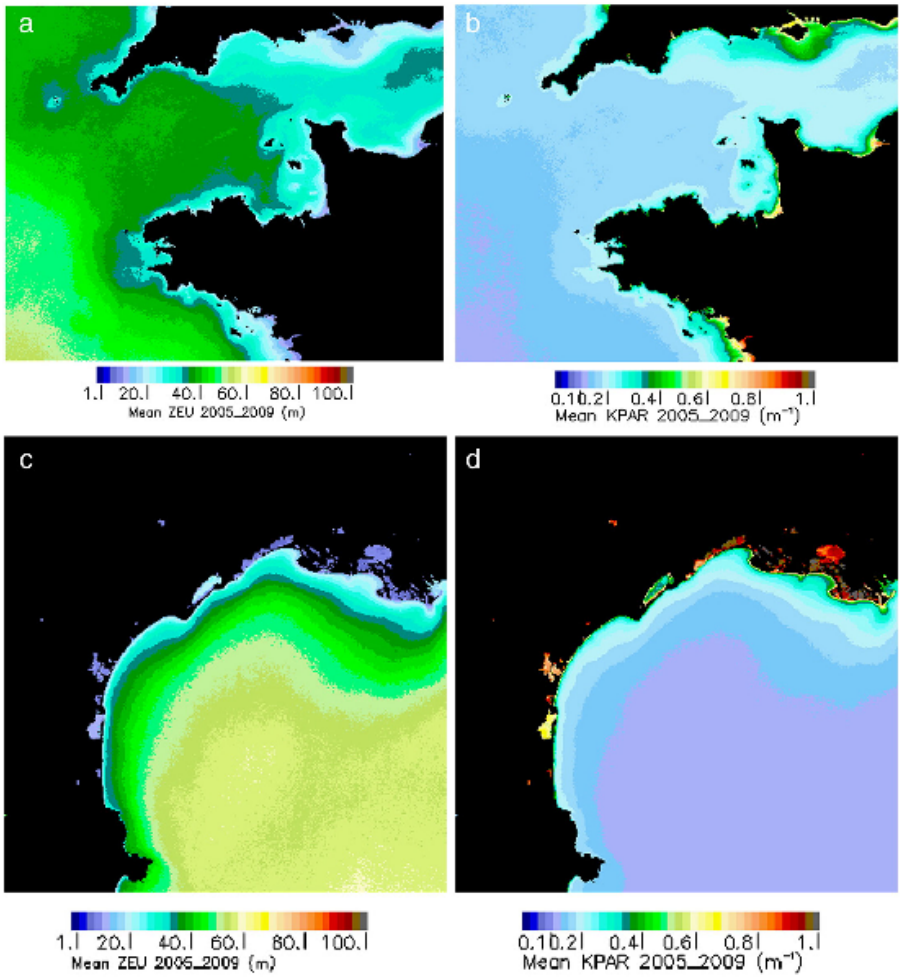


Fig. 10. Mean of Z_{eu} and K_{dPAR} at 250 m resolution over the period 2005–2009 for the Brittany (a and b) and the Gulf of Lions (c and d).

hydrographic zero, which in France corresponds to the lowest observed sea level. While in the Mediterranean Sea the tide range is very small, a tidal range of several metres in Brittany is normal. Therefore the half of the annual mean tide value at Brest (0.5 ∗ 6.1 m, Service Hydrographique et Océanographique de la Marine, SHOM) was added to the bathymetry (Table 3).

Fig. 12 shows the distribution of *L. hyperborea* in the French Abers. Kelp forest presence was obtained by echo-integrating the acoustic signal (Méléder et al., 2010), which enables distinguishing dense kelp forest from sparse kelp or bare rock. The sounder also provided in-situ depth measurements that account for the effect of tide, resulting in a relatively accurate estimate of the depth (with an uncertainty of 0.5 m). The values observed for the minimum in les Abers (mean of 2.3%) is significantly higher for the two other sites. This can be explained by the hydrodynamic energy regime at the seabed, which differs greatly between the North and South Brittany. Kain (1971, 1976) proposed a minimum percentage of incidental light ranges from 1% to 1.9% for *L. hyperborean* and Lüning (1979, 1990), 0.7% and 70 mol·m^{-2}.year^{-1} for this species, i.e. 0.19 mol·photons·m^{-2}·d^{-1}.

The calculated mean weighted values are 1.73% and 0.42 mol·photons·m^{-2}·d^{-1} in the range proposed by Kain (in fraction of surface energy) and slightly higher than the threshold approach proposed by Lüning.

6. Conclusions

We propose two relationships between the mean K_{dPAR}, integrated over the euphotic layer, and the K_{d490} estimated according to Lee et al. (2005a, 2005b), for very clear waters (K_{dPAR}<0.115 m^{-1}) and turbid waters (K_{dPAR}≥0.115 m^{-1}). The empirical relationship for coastal areas suggests a correction to the underestimation of K_{dPAR} by the standard Globcolour case 1 algorithm, and also provides an estimation of the $K_{dPAR}(Z_{eu})$ from the in-situ K_{d490}. Satellite derived K_{d490} and K_{dPAR} have been validated using available matchups between the MERIS data, in-situ measurements and Ecolight simulations. Evaluation results suggest that the Lee et al. (2005a, 2005b) algorithm derived for MERIS is valid for estimation of K_{d490} and the subsequent K_{dPAR} in coastal areas.

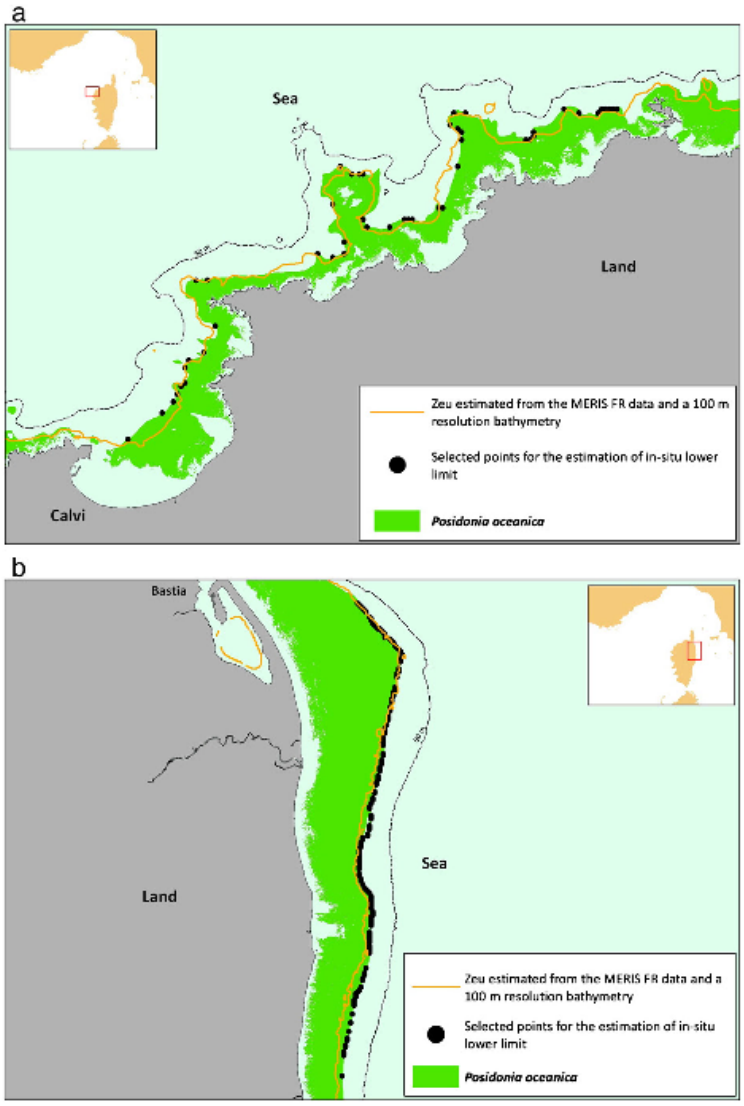


Fig. 11. Distribution of *P. oceanica* compared to Z_{eu} derived from MERIS FR daily data from 2005 to 2009 at Calvi a) and Bastia b).

The mean values of the observed threshold for the three selected sites in Corsica were 0.94% and 0.13 mol·photons·m^{-2}·d^{-1} for *P. oceanica*. These estimates are very close to the 1% definition of Z_{eu} and in the lower limit of the energy range proposed by Gattuso et al. (2006). For *L. hyperborea* surveys in Brittany, our estimated values from the satellite data were 1.73% and 0.42 mol·photons·m^{-2}·d^{-1}, in the range (1–1.9%) proposed by Kain (1971, 1976) and slightly higher than the energy threshold proposed by Lüning (0.7%, 0.19 mol·photons·m^{-2}·d^{-1} 1979, 1990). The bathymetry used in this work is calculated from the hydrographic zero, which in France corresponds to the lowest observed level of the sea. The influence of the tide has been considered in Brittany by adding the half of the mean tidal level. Bowers and Brubaker (2010) showed also that because of tide and non-linearity of the light attenuation, the light gained at low tide exceeds the loss at high tide leading to a deeper colonisation of the species in such areas. Therefore, future works will integrate accurate local estimations of annual mean tide values.

The estimation of minimum light requirements in mol·photons·m^{-2}·d^{-1}, a true physical quantity, is meaningful compared to an estimation expressed in fraction of surface energy. This residual energy reaching the bottom at high resolution is also a good candidate as input parameter in the predictive modelling of seabed habitats such as proposed by Méléder et al. (2010).

Acknowledgements

This work was funded by the EuseaMAP project http://www.jncc.gov.uk/page-5027 of the European Commission's Directorate-General

Table 2
Statistics of fraction of the surface light and corresponding energy in mol·photons·m^{-2}·d^{-1} observed at the lower limit of *P. oceanica* beds.

	Min	Mean	Max	St. dev.	Nb points
Aléria, depth DTM accuracy: 1 m					
% E_0	0.46	**1.15**	1.92	0.20	165
mol·photons·m^{-2}·day^{-1}	0.13	**0.33**	0.56	0.06	165
Depth (m)	26.0	**31.8**	37.0	2.68	165
South Bastia, depth DTM accuracy: 1 m					
% E_0	0.4	**0.73**	1.24	0.24	171
mol·photons·m^{-2}·day^{-1}	0.12	**0.22**	0.36	0.06	171
Depth (m)	33.0	**35.4**	38.0	1.1	171
Calvi depth DTM accuracy: 5 m					
% E_0	0.44	**0.96**	2.04	0.12	48
mol·photons·m^{-2}·day^{-1}	0.13	**0.28**	0.60	0.04	48
Depth (m)	28.0	**31.0**	33.0	1.6	48

Table 3
Statistics of the fraction of surface light and the corresponding energy in mol·photons·m^{-2}·d^{-1} observed at the lower limit of *L. hyperborea* in Brittany.

	Min	**Mean**	Max	St. dev.	Nb points
Abers					
% E_0	1.72	**2.3**	2.74	0.43	74
mol·photons·m^{-2}·day^{-1}	0.42	**0.57**	0.68	0.11	74
Depth (m)	21.1	**22.0**	24.7	0.65	74
Glénan					
% E_0	0.39	**0.85**	1.24	0.41	32
mol·photons·m^{-2}·day^{-1}	0.12	**0.20**	0.39	0.10	32
Depth (m)	26.0	**28.2**	32.1	1.96	32
Groix Sud					
% E_0	1.00	**1.25**	1.54	0.29	28
mol·photons·m^{-2}·day^{-1}	0.24	**0.31**	0.38	0.08	28
Depth (m)	19.1	**19.9**	21.0	0.60	28

for Maritime Affairs and Fisheries. The authors thank the SHOM and Frédéric Jourdin for providing OpticMed data and high resolution DTMs Michel Lunven from Ifremer for the provision of PAR profiles on the French shores, Robert Frouin for the provision of a revised version of the PAR estimation using MERIS data, Zhong-Ping Lee for useful advice and the Medbenth project for access to Mediterranean seagrass maps and Nick Stephens, from the Plymouth Marine Laboratory, for additional comments.

Appendix A. Hydrolight/Ecolight settings

Inherent optical properties

- Pure water absorption coefficient for 400–720 nm from (Pope and Fry, 1997)
- [Chl-a] constant with depth with values extracted from NOMAD
- Default Hydrolight Chl-a absorption coefficient
- Default Hydrolight Chl-a backscattering coefficient
- A_{cdm443} from NOMAD
- CDOM γ coefficient = − 0.0176 nm^{-1}
- [SPM] from NOMAD and Eq. (13)
- Mineral particles specific scattering coefficient at 555 nm = 0.51 m^{2}g^{-1}
- Mineral particles specific absorption coefficient at 443 nm = 0.041 m^{2}g^{-1}
- Wavelengths similar to MERIS
- Chlorophyll fluorescence effects not included

Geometry

- Solar zenith angle of 40°
- Nadir viewing

Atmospheric and air–sea interface

- Surface wind speed of 5 m·s^{-1}
- Real index of refraction of water = 1.34

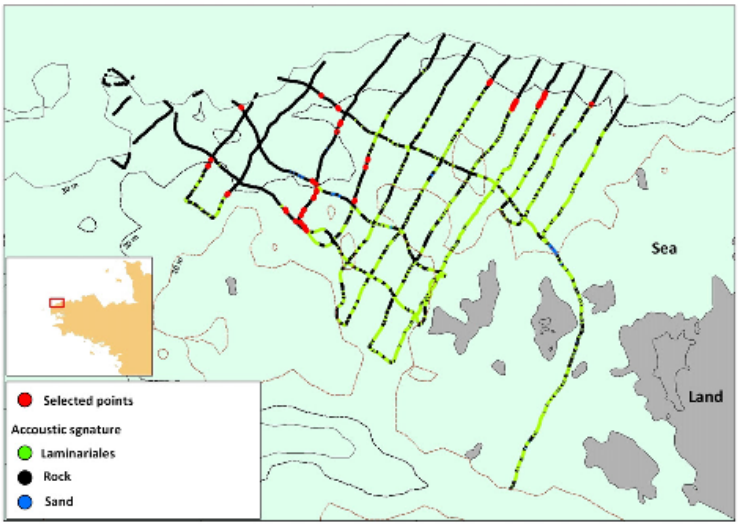


Fig. 12. Single-beam survey lines (thick lines) for the site Abers. Green dots denote the presence of kelp forest. Red dots are deepest occurrences of kelp forest. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Annex

References

Babin, M., Stramski, D., Ferrari, G. M., Claustre, H., Bricaud, A., Obolensky, G., et al. (2003). Variations in the light absorption coefficients of phytoplankton, non-algal particles, and dissolved organic matter in coastal waters around Europe. *Journal of Geophysical Research*. http://dx.doi.org/10.1029/2001JC000882.

Baker, K. S., & Frouin, R. (1987). Relation between photosynthetically available radiation and total insolation at the ocean surface under clear skies. *Limnology and Oceanography, 32*, 1370–1377.

Barnard, A. H., Zaneveld, J. R. V., Pegau, W. S., Mueller, J. L., Maske, H., Lara, R. L., Borrego, S. A., Duarte, R. C., & Holguin, E. V. (1999). The determination of PAR levels from absorption coefficient profiles at 490 nm. *Ciencias Marinas, 25*, 487–507.

Bowers, D. G., & Brubaker, J. M. (2010). Tidal amplification of seabed light. *Journal of Geophysical Research, 115*, C09008.

Cahoon, L. B., Beretich, G. R., Thomas, C. J., & Mcdonald, A. M. (1993). Benthic microalgal production at Stellwagen Bank, Massachusetts Bay, USA. *Marine Ecology Progress Series, 102*(1–2), 179–185.

Carter, C. M., Ross, A. H., Schiel, D. R., Howard-Williams, C., & Hayden, B. (2005). In situ microcosm experiments on the influence of nitrate and light on phytoplankton community composition. *Journal of Experimental Marine Biology and Ecology, 326*, 1–13.

Claustre, H., & Maritorena, S. (2003). The many shades of ocean blue. *Science, 302*, 1514–1515.

Devlin, M. J., Barry, J., Mills, D. K., Gowen, R. J., Foden, J., Sivyer, D., et al. (2009). Estimating the diffuse attenuation coefficient from optically active constituents in UK marine waters. *Estuarine, Coastal and Shelf Science, 82*(1), 73–83.

Esaias, W. E., Abbott, M. R., Barton, I., Brown, O. B., Campbell, J. W., Carder, K. L., et al. (1998). An overview of MODIS capabilities for ocean science observations. *IEEE Transactions on Geoscience and Remote Sensing, 36*(4), 1250–1265.

Falkowski, P. G., & Raven, J. A. (1997). *Aquatic photosynthesis*. Blackwell Science (349 pp.).

Frouin, R., Lingner, D. W., & Gautier, C. (1989). A simple analytical formula to compute clear sky total and photosynthetically available solar irradiance at the ocean surface. *Journal of Geophysical Research, 94*(7), 9731–9742.

Gattuso, J. -P., Gentili, B., Duarte, C. M., Kleypas, J. A., Middelburg, J. J., & Antoine, D. (2006). Light availability in the coastal ocean: impact on the distribution of benthic photosynthetic organisms and their contribution to primary production. *Biogeosciences, 3*, 489–513.

Gohin, F., Loyer, S., Lunven, M., Labry, C., Froidefond, J., Delmas, D., et al. (2005). Satellite-derived parameters for biological modelling in coastal waters: Illustration over the eastern continental shelf of the Bay of Biscay. *Remote Sensing of Environment, 95*(1), 29–46.

Gordon, H. R., & Morel, A. (1983). *Remote assessment of ocean color for interpretation of satellite visible imagery: A review, Vol. 114*, New York: Springer-Verlag.

Gordon, H. R., & Wang, M. (1994). Retrieval of water-leaving radiance and aerosol optical thickness over the oceans with SeaWiFS: A preliminary algorithm. *Applied Optic, 33*(3), 443–452.

IOCCG Report 3 (2000). *Remote sensing in coastal and other optically-complex waters*.

Jerlov, N. G. (1976). Optical oceanography. New York: Elsevier.

Kain, J. M. (1971). Continuous recording of underwater light in relation to Laminaria distribution. *Marine Biology Symposium* (pp. 335–346). London: Cambndge Univ. Press.

Kain, J. M. (1976). The biology of *Laminaria hyperborea*. *Oceanography and Marine Biology Annual Revue, 17*, 101–161.

Kirk, J. T. O. (1994). *Light and photosynthesis in aquatic ecosystems*. Cambridge University Press.

Lee, Z. P. (2009). KPAR: An ambiguous optical property. *Journal of Lake Sciences, 21*, 159–164.

Lee, Z. P., Carder, K. L., & Arnone, R. A. (2002). Deriving inherent optical properties from water color: A multiple quasi-analytical algorithm for optically deep waters. *Applied Optic, 41*, 5755–5772.

Lee, Z. P., Darecki, M., Carder, K., Davis, C., Stramski, D., & Rhea, W. (2005). Diffuse attenuation coefficient of downwelling irradiance: An evaluation of remote sensing methods. *Journal of Geophysical Research, 110*, C02017.

Lee, Z. P., Du, K. P., & Arnone, R. (2005). A model for the diffuse attenuation coefficient of downwelling irradiance. *Journal of Geophysical Research, 110*, C02016.

Lee, Z. P., Lubac, B., Werdell, J., & Arnone, R. (2009). *An update of the quasi-analytical algorithm (v5)*. IOCCG software report www.ioccg.org/groups/Software_OCA

Lee, Z. P., Weideman, A., Kindle, J., Arnone, R., Carder, K., & Davis, C. (2007). Euphotic zone depth: Its derivation and implication to ocean-color remote sensing. *Journal of Geophysical Research, 112*, C03009.

Lewis, M., Carr, M., Feldman, G., Esaias, W. E., & Mcclain, C. R. (1990). Influence of penetrating solar radiation on the heat budget of the equatorial Pacific ocean. *Nature, 347*, 543–545.

Lüning, K. (1979). Growth strategies of three *Laminaria* species (Phaeophyceae) inhabiting different depth zones in the sublittoral region of Helgoland (North Sea). *Marine Ecology Progress Series, 1*, 195–207.

Lüning, K. (1990). *Seaweeds: Their environment, biogeography, and ecophysiology*. New York, US: John Wiley & Sons.

McClain, C. R., Feldman, G. C., & Hooker, S. B. (2004). An overview of the SeaWiFS project and strategies for producing a climate research quality global ocean bio-optical time series. *Deep Sea Research Part II, 51*, 5–42.

Mcminn, A., Hirawake, T., Hamaoka, T., Hattori, H., & Fukuchi, M. (2005). Contribution of benthic microalgae to ice covered coastal ecosystems in northern Hokkaido, Japan. *Journal of the Marine Biological Association of the United Kingdom, 85*(2), 283–289.

Méléder, V., Populus, J., Guillaumont, B., Perrot, T., & Mouquet, P. (2010). Predictive modelling of seabed habitats: Case study of subtidal kelp forests on the coast of Brittany, France. *Marine Biology, 157*(7), 1525–1541.

MERIS level 2 detailed processing model. http://earth.esa.int/pub/ESA_DOC/ENVISAT/MERIS/MERIS_DPML2_i7r2A_re-issued.pdf

Morel, A. (1988). Optical modeling of the upper ocean in relation to its biogenous matter content. *Journal of Geophysical Research, 93*, 10749–10768.

Morel, A., Antoine, D., & Gentili, B. (2002). Bidirectional reflectance of oceanic waters: Accounting for Raman emission and varying particle phase function. *Applied Optics, 41*, 6289–6306.

Morel, A., Huot, Y., Gentili, B., Werdell, P. J., Hooker, S. B., & Franz, B. A. (2007). Examining the consistency of products derived from various ocean color sensors in open ocean (Case 1) waters in the perspective of a multi-sensor approach. *Remote Sensing of Environment, 111*, 69–88.

Morel, A., & Smith, R. C. (1974). Relation between total quanta and total energy for aquatic photosynthesis. *American Society of Limnology and Oceanography, 19*(4), 591–600.

Mueller, J. L. (2000). SeaWiFS algorithm for the diffuse attenuation coefficient, $K_{(490)}$, using water-leaving radiances at 490 and 555 nm. *SeaWiFS Postlaunch Technical Report Series*Greenbelt, Maryland: NASA Goddard Space Flight Center (24–27 pp.).

Pierson, D. C., Kratzer, S., Strombeck, N., & Hakansson, B. (2008). Relationship between the attenuation of downwelling irradiance at 490 nm with the attenuation of PAR (400 nm– 700 nm) in the Baltic Sea. *Remote Sensing of Environment, 112*, 668–680.

Pope, R. M., & Fry, E. S. (1997). Absorption spectrum (380–700 nm) of pure water. II. Integrating cavity measurements. *Applied Optics, 36*, 8710–8723.

Rochford, P. A., Kara, A. B., Wallcraft, A. J., & Arnone, R. A. (2001). Importance of solar subsurface heating in ocean general circulation models. *Journal of Geophysical Research, 106*, 30923–30938.

Ryther, J. H. (1956). Photosynthesis in the ocean as a function of light intensity. *Limnology and Oceanography, 1*, 61–70.

Sathyendranath, S., Gouveia, A. D., Shetye, S. R., Ravindran, P., & Platt, T. (1991). Biological control of surface temperature in the Arabian Sea. *Nature, 349*, 54–56.

Wang, M., & Shi, W. (2005). Estimation of ocean contribution at the MODIS near-infrared wavelengths along the east coast of the U.S.: Two case studies. *Geophysical Research Letters, 32*, L13606.

Wang, M., Son, S., & Harding, L. (2009). Retrieval of diffuse attenuation coefficient in the Chesapeake Bay and turbid ocean regions for satellite ocean color applications. *Journal of Geophysical Research, 114*, c10011.

Werdell, P. J., & Bailey, S. W. (2005a). An improved bio-optical data set for ocean color algorithm development and satellite data product validation. *Remote Sensing of Environment, 98*(1), 122–140.

Werdell, P. J., & Bailey, S. W. (2005b). *NASA technical memorandum 104566, vol. 25*.

Wu, Y., Tang, C., Sathyendranath, S., & Platt, T. (2007). The impact of bio-optical heating on the properties of the upper ocean: A sensitivity study using a 3-D circulation model for the Labrador Sea. *Deep Sea Research Part II, 54*, 2630–2642.

Zaneveld, J. R. V., Kitchen, J. C., & Mueller, J. L. (1993). Vertical structure of productivity and its vertical integration as derived from remotely sensed observations. *Limnology and Oceanography, 38*, 1384–1393.

Résumé

Dans cette thèse sur articles nous proposons des méthodes statistiques pour l'analyse de séries temporelles géophysiques. Nous nous intéressons à la couleur de l'eau et la température de surface de la mer observées depuis l'espace. La nature du signal géophysique, i.e. l'autocorrélation du bruit, la discontinuité des observations, et le mélange des processus physiques et biologiques pouvant comporter des modes distincts de variabilité, est partie intégrante des méthodes d'analyses proposées.

Le premier chapitre contient des informations sur la mesure physique des variables d'intérêt. Nous décrivons ensuite une méthode d'estimation de tendances linéaires et des incertitudes associées, à partir de plusieurs séries temporelles, dans le but d'optimiser les réseaux d'observations satellitaires et in-situ pour la surveillance à long terme de l'environnement. Puis nous caractérisons les échelles temporelles du signal climatique «El Niño Southern Oscillation» dans la température de surface à partir d'un mélange de Gaussiennes prenant en compte un facteur de normalisation pour corriger de la distribution naturelle des évènements. Le quatrième chapitre décrit la prévision temporelle d'un processus non-stationnaire soumis à des forçages saisonniers, la turbidité de surface, en utilisant quatre différents modèles de Markov cachés. Les variables cachées sont utilisées pour identifier des relations distinctes entre la variable à estimer et ses prédicteurs. Le dernier chapitre décrit un modèle Bayésien avec mise à jour dynamique des modèles a priori pour l'inversion de la réflectance marine en milieux côtiers complexes pour le capteur OLCI embarqué sur le satellite Sentinel 3.

Les perspectives sont l'amélioration des produits satellitaires fournis par les agences spatiales, la prévision opérationnelle avec des modèles statistiques basés sur des observations et de l'apprentissage, et l'optimisation des réseaux de surveillance.

Mots-clés : Statistiques, Signal géophysique, Observations satellitaires, Détection de tendances, Analyses en cluster, Chaînes de Markov, Prévision, Inversion Bayésienne, Variables cachées

Abstract

In this manuscript-based thesis we propose statistical methods to analyse geophysical time-series. We particularly focus on ocean colour and sea surface temperature observed from space. The specific characteristics of the geophysical signal, i.e. noise autocorrelation, discontinuities in the observations, and the mixing of physical and biological processes showing distinct modes of variability, is integrated in the proposed methodologies.

The first chapter includes details on the physical measurement principles of the variables of interest. Then we describe a methodology to estimate linear trends and associated uncertainties, using several time series, to optimise in-situ and satellite-based observation networks for the long term monitoring. We characterise after the significant time-scales of «El Niño Southern Oscillation» in the sea surface temperature using a normalised Gaussian mixture model to take into account the natural distribution of the scales of the events observed in the nature. The fourth chapter details the forecasting of a non-stationary process subject to seasonal forcing conditions, the sea surface turbidity, using four hidden Markov models. The hidden variables are used to estimate different relationships between the variable to estimate and its predictors. The last chapters details a Bayesian model, with dynamical updates of the a priori models, to inverse the sea surface reflectance in complex coastal areas for the OLCI sensor embedded onto the Sentinel 3 satellite.

The perspectives are enhancements of satellite products provided by spatial agencies, operational forecasting using statistical models based on observations and learning, and the optimisation of observation networks.

Keywords : Statistics, Geophysical signal, Satellite-derived observations, Trend detection, Cluster analysis, Markov chains, Forecasting, Bayesian inversion, Hidden variables

TELECOM Bretagne

n° d'ordre : 2014telb0332

Télécom Bretagne

Technopôle Brest-Iroise - CS 83818 - 29238 Brest Cedex 3

Tél : + 33(0) 29 00 11 11 - Fax : + 33(0) 29 00 10 00