MODÉLISATION SPATIO-TEMPORELLE D'UNE VARIABLE QUANTITATIVE À PARTIR DE DONNÉES MULTI-SOURCES APPLICATION À LA TEMPÉRATURE DE SURFACE DES OCÉANS

Pierre Tandeo

N° ordre : 2010-23
N° Série : G-8

THESE / AGROCAMPUS OUEST

Sous le sceau de l'Université Européenne de Bretagne

pour obtenir le diplôme de :

DOCTEUR DE L'INSTITUT SUPERIEUR DES SCIENCES AGRONOMIQUES, AGRO-ALIMENTAIRES, HORTICOLES ET DU PAYSAGE

Spécialité : Mathématique - Physique - Informatique

Ecole Doctorale : Vie-Agro-Santé

présentée par :

Pierre Tandeo

MODÉLISATION SPATIO-TEMPORELLE D'UNE VARIABLE QUANTITATIVE À PARTIR DE DONNÉES MULTI-SOURCES APPLICATION À LA TEMPÉRATURE DE SURFACE DES OCÉANS

Soutenance prévue le 15 octobre 2010 devant la commission d'examen

Composition du jury :	Denis Allard	Rapporteur
	Laurent Bertino	Rapporteur
	Bertrand Chapron	Membre
	Pierre Ailliot	Membre
	David Causeur	Directeur de thèse

Remerciements

Tout d'abord, je remercie David Causeur pour avoir accepté de devenir mon directeur de thèse. Merci également aux rapporteurs Denis Allard et Laurent Bertino pour l'intérêt qu'ils ont porté à mon travail et leurs commentaires judicieux. Je remercie conjointement Bertrand Chapron et Emmanuelle Autret pour m'avoir donné l'opportunité d'exercer cette thèse au sein du laboratoire d'océanographie spatiale de l'IFREMER de Brest. Ils m'ont laissé une grande liberté de travail tout en apportant leur expertise sur le sujet. Enfin, mes plus sincères remerciements vont au dernier membre du jury : Pierre Ailliot. En effet, il a été présent tout au long de cette thèse et a pris le temps de répondre à mes questions.

Pour ce qui est du soutien familial, il a été largement présent. Ainsi, j'en profite pour remercier mes parents Gilles et Anne et mon frère Maxime pour leurs relectures. Merci également à mon oncle Yves pour son aide lors de la rédaction. Enfin, merci à Morgane pour son soutien quotidien.

Résumé

Ce travail de thèse porte sur une variable océanographique importante dans le suivi du climat : la température de surface des océans. Au niveau global, les observations de cette température sont fournies principalement par des radiomètres embarqués sur des satellites. Afin de traiter ce flux important de données, un traitement statistique s'impose dans le but de synthétiser l'information en des cartes globales et quotidiennes de notre variable d'intérêt. Pour ce faire, nous proposons un modèle linéaire de type espace-d'état avec des erreurs Gaussiennes.

Nous commençons par présenter ce modèle sur des données issues de séries temporelles ayant un échantillonnage irrégulier. Suit un travail d'inférence avec la mise en place d'un schéma d'estimation des paramètres, basé sur la combinaison d'une méthode des moments et du maximum de vraisemblance au travers de l'algorithme EM et des probabilités de filtrage et lissage de Kalman. Nous appliquons enfin cette méthodologie pour estimer les variances d'erreurs et le paramètre de corrélation temporelle à tout l'océan Atlantique.

Nous ajoutons ensuite la composante spatiale et proposons une structure d'ordre deux, séparable, basée sur le produit d'une covariance temporelle et d'une covariance spatiale anisotrope. Les paramètres de cette dernière sont estimés sur l'océan Atlantique à partir de techniques géostatistiques usuelles et forment un atlas pertinent pour les océanographes. Finalement, nous montrons que l'apport de l'information spatiale augmente le pouvoir prédictif du modèle.

Abstract

In this thesis, an important oceanographic variable for the monitoring of the climate is studied: the sea surface temperature. At the global level, this variable is observed along the ocean by several remote sensed sources. In order to treat all this information, statistical methods are used to summarize our variable of interest in global daily map. For that purpose, a state-space linear model with Gaussian error is suggested.

We begin to introduce this model on data resulting from having an irregular sampling. Then, we work on the estimation of the parameters. This is based on the combination of the method of moments and the maximum likelihood estimates, with the study of the EM algorithm and the Kalman recursions. Finally, this methodology is applied to estimate the variance of errors and the temporal correlation parameter to the Atlantic ocean.

We add the spatial component and propose a separable second order structure, based on the product of a temporal covariance and a spatial anisotropic covariance. According to usual geostatistical methods, the parameters of this covariance are estimated on the Atlantic ocean and form a relevant atlas for the oceanographers. Finally, we show that the contribution of the spatial information increases the predictive behaviour of the model.

TABLE DES MATIÈRES

Introduction

Contexte

Le travail réalisé dans cette thèse est initié par le laboratoire d'océanographie spatiale de l'institut de recherche IFREMER[1] de Brest. Ce laboratoire, en collaboration avec l'ESA[2] a pour mission d'enregistrer, calibrer, homogénéiser puis distribuer de nombreuses variables physiques ayant trait à l'océan : la vitesse et la direction du vent, la salinité, la couleur de l'eau, la température de surface des océans, la hauteur des vagues ou encore les glaces. Ces variables sont mesurées directement dans les océans par différents flotteurs (nous parlerons de données *in situ*) ou indirectement par des instruments embarqués sur des satellites. Les observations satellitaires, du fait de l'amélioration constante des résolutions spatiales des capteurs, sont de plus en plus volumineuses. A l'heure actuelle, pour la seule température de surface des océans, environ cent gigaoctets de données brutes par jour (enregistrement direct des fauchées satellitaires) sont à gérer par le laboratoire. De plus, l'objectif principal du laboratoire est la distribution des données en un point fixe (séries temporelles) ou sur un champ spatial. Pour ce faire, il est important que les données brutes bénéficient, en amont, d'un traitement statistique de synthèse pour ne garder que l'information pertinente.

Dans cette thèse, la variable d'intérêt sera la température de surface des océans. Celle-ci intervient dans de nombreuses applications telles que la compréhension de l'évolution climatique ou la modélisation atmosphérique. Elle sert également de variable explicative dans les modèles de circulation océanique (cf. [12]). De nombreux laboratoires cherchent à estimer au mieux cette température de surface. Ceux-ci fournissent des cartes quotidiennes de notre variable sur tous les océans. L'analyse la plus connue est OIV2[3] (cf. [48]), établie à partir de données issues de différents capteurs satellitaires, avec une résolution spatiale faible, de l'ordre de 0.25° soit environ 25 km. Actuellement, le laboratoire d'océanographie spatiale

[1]Institut Français de Recherche pour l'Exploitation de la MER

[2]European Space Agency

[3]Optimal Interpolation Version 2

de l'IFREMER fournit également sa propre analyse quotidienne et globale, ODYSSEA[4] (cf. [1]), qui a pour but de réduire la résolution spatiale à 0.1°. Les résultats de ces différentes interpolations spatiales sont basés sur des modèles de type interpolation optimale (krigeage dans le langage géostatistique) et font l'objet de recherches au sein du groupe de travail GHRSST[5]. Cependant, les validations faites sur les sorties des modèles montrent certains défauts : un manque de continuité temporelle et spatiale entre les analyses des différents jours et le non-respect de la dynamique physique de la température de surface. Certains auteurs plus anciens (cf. [19] ou [38]) se sont intéressés à la modélisation du transport de la température de surface. L'objectif de cette thèse est de modéliser la température de surface des océans en prenant en compte les différentes remarques faisant défaut aux modèles actuels ainsi que les travaux des différents auteurs de référence cités ci-dessus. Nous chercherons à rendre le modèle explicatif en estimant les paramètres de la variabilité spatiale et temporelle par des méthodes statistiques.

L'association de plusieurs compétences est indispensable pour mener à bien ce travail. Premièrement, cette thèse s'inscrit dans un cadre statistique. Le traitement des données ainsi que la mise en place d'un modèle mathématique sont les tâches primordiales à accomplir ; plusieurs enseignants chercheurs de différents laboratoires de mathématiques appliquées ont suivi ces travaux de recherche. Deuxièmement, l'expertise physique, indispensable à la validation des hypothèses et des résultats, est faite par les chercheurs du laboratoire d'océanographie spatiale. Troisièmement, la collaboration avec les informaticiens et les ingénieurs chargés de la partie opérationnelle est importante, notamment pour la maintenance et la distribution des résultats de l'analyse ODYSSEA.

Plan

La température de surface des océans sera notre variable d'intérêt. Dans le chapitre 1, nous présenterons les différentes mesures possibles de cette température *in situ* et les observations satellitaires. Nous en dégagerons des problématiques sous-jacentes. Nous verrons que nous ne travaillerons pas directement sur cette température de surface mais sur son anomalie, différence entre un signal haute et basse résolution.

Le chapitre 2 présentera la forme générale du modèle espace d'état, largement utilisé en océanographie et météorologie. Il est particulièrement adapté au cas des observations satellitaires. Nous étudierons en détail la version linéaire à erreurs Gaussiennes de ce modèle avec une résolution par approche stochastique : le filtre de Kalman. Puis, nous détaillerons le schéma d'estimation des paramètres du modèle par maximum de vraisemblance.

Dans le chapitre 3, nous travaillerons sur une version simple du modèle espace d'état et sur des données de séries temporelles univariées sans composante spatiale. Nous appliquerons la méthode d'estimation des paramètres par maximum de vraisemblance présentée dans le chapitre 2, couplée à une approche basée sur les moments afin de déterminer les paramètres de variations temporelles de notre variable d'intérêt. Après une discussion autour de ces

[4]Ocean Data Analysis System for merSEA

[5]Group for High-Resolution SST

estimateurs et des résidus du modèle en un point fixe, nous généraliserons l'ajustement du modèle à différents points d'une grille régulière couvrant l'océan Atlantique. En effet, le modèle a pour finalité d'être appliqué à tous les océans. Il est donc important d'analyser longuement une grande zone telle que l'Atlantique où de nombreuses situations sont observées, avec des variabilités temporelles distinctes.

Le chapitre 4 s'intéressera à la variabilité spatiale de la température de surface de l'eau. Nous étudierons des champs de températures de surface fournis par les capteurs satellites. Nous en modéliserons la variabilité spatiale à partir de techniques classiques de géostatistique. Ici aussi, dans un but explicatif, nous généraliserons l'estimation de la variabilité spatiale en différentes régions couvrant tout l'océan Atlantique. Parallèlement, dans un but prédictif, nous chercherons à valider les interpolations de notre modèle spatio/temporel. Les sorties des deux modèles présentés dans les chapitres 3 et 4 seront comparées afin d'apprécier l'apport de l'information spatiale. Nous conclurons cette thèse par une discussion générale sur les résultats des différents modèles présentés ainsi que les perspectives à venir.

CHAPITRE 1

Température de surface des océans

La température de surface des océans sera notée, tout au long de cette thèse, SST[1]. Elle désigne globalement la température de l'eau dans les premiers mètres. C'est une variable fondamentale pour l'étude de l'océan, de l'atmosphère et de leurs interactions. Une bonne connaissance de la SST est également essentielle pour des applications dans plusieurs domaines : analyses et prévisions océaniques (assimilation dans les modèles, validation), modélisation atmosphérique (conditions aux limites), suivi du climat, suivi de qualité des eaux, halieutique et gestion des ressources marines.

Dans ce chapitre introductif, nous commencerons en 1.1 par présenter les différents moyens de mesure de la SST, que ce soit les données *in situ* ou celles provenant de capteurs satellitaires et nous nous efforcerons de mettre en évidence, au fur et à mesure, les différentes problèmatiques sous-jacentes à ces données de SST. Nous introduirons ensuite le contexte et nous synthétiserons les problématiques de l'étude en 1.2. Enfin, nous finirons par une discussion autour de ce premier chapitre introductif.

1.1 Différents types de mesures

La SST peut être déterminée par différents moyens de mesures : indirectement par observations de radiomètres, embarqués sur des capteurs satellites, selon différentes longueurs d'ondes (micro-onde ou infra-rouge) ou directement par mesures de bouées (fixes ou dérivantes). Comme l'indique la figure 1.1, cette température varie dans la première couche de l'eau de mer (inférieure à 1 mm) et sa définition n'est pas la même d'un moyen de mesure à un autre. Par exemple, les bouées mesurent plus profondemment notre variable d'intérêt alors que les données de capteurs satellitaires observent celle-ci juste à l'interface de la surface des océans. Il existe alors parfois un léger biais de température entre ces différentes sources de mesures.

[1]Sea Surface Temperature

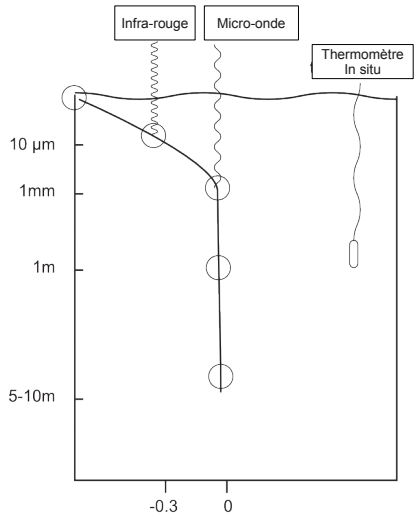


FIG. 1.1: Définition de la température de surface des océans (SST). L'axe des ordonnées renseigne sur la profondeur de mesure de la SST. L'axe des abscisses, en °C, désigne la différence entre la SST mesurée à une profondeur donnée et la SST à plus d'un mètre de profondeur.

Dans cette section, nous allons présenter les deux sources de données : *in situ* en 1.1.1 et satellitaires en 1.1.2. Ensuite, nous dirons quelques mots sur des données dites analysées et nous verrons leur utilité en 1.1.3. Pour finir, nous aborderons en 1.1.4 le sujet de la colocalisation, permettant de confronter différents moyens de mesure.

1.1.1 Données *in situ*

In situ vient du latin qui signifie "position d'origine". Depuis plusieurs dizaines d'années, lors de campagnes, de nombreuses bouées fixes et autonomes, dérivantes, sont lachées en mer. Celles-ci, immergées à quelques centimètres de la surface de l'océan, mesurent plusieurs variables dont la SST mais leur couverture spatiale est très réduite (cf. figure 1.2). Cependant, cette source de données qui mesure directement la SST est considérée comme valeur de référence.

Parmi les bouées fixes, on trouve un réseau célèbre TAO / PIRATA[2] (cf. [51]), qui mesure la SST et d'autres variables océaniques au niveau de l'équateur. Le pas d'échantillonnage des flotteurs est faible (10 minutes) et il y a relativement peu de données manquantes. Nous disposons donc de séries temporelles permettant de suivre l'évolution de la SST au cours du temps dans des régions précises. Prenons l'exemple d'une bouée du réseau PIRATA, se trouvant au point 0°N 23°W, au milieu de l'Atlantique. Celle-ci rend une mesure de SST toutes les 10 minutes et nous traçons la série temporelle des observations sur l'année 2008 en

[2]Tropical Athmosphere Ocean / Pilot Research morred Array in the Tropical Atlantic

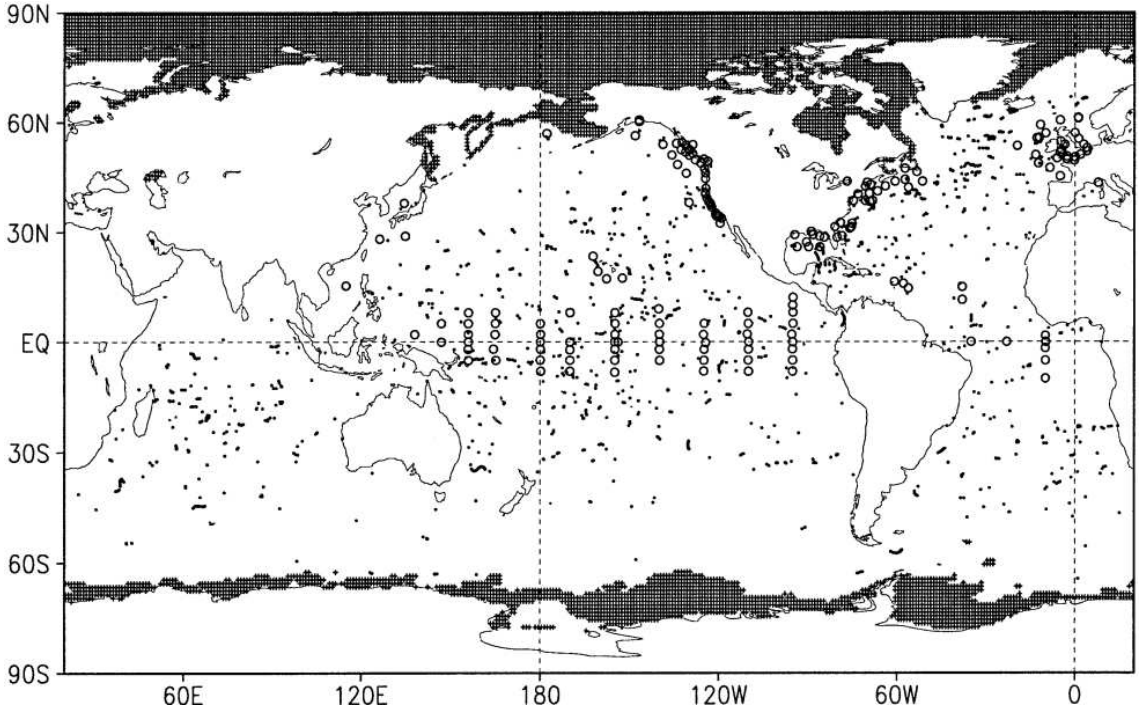
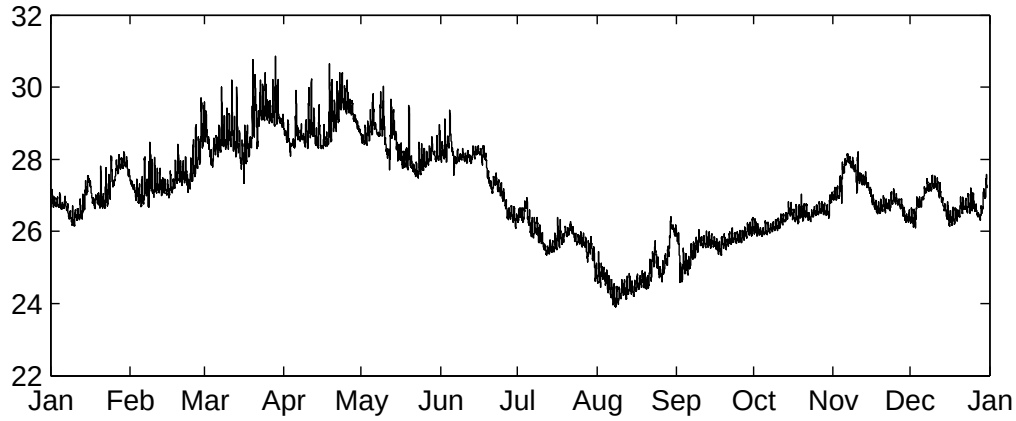


FIG. 1.2: Bouées fixes (cercles) et dérivantes (points) mesurant la SST sur une période d'une semaine, du 2 au 8 janvier 2000 (cf. [46]).

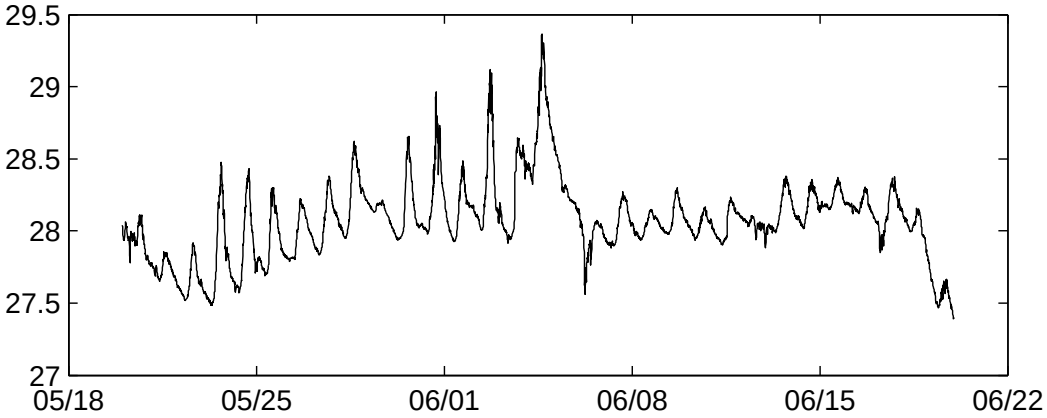
figure 1.3a. Nous remarquons un cycle saisonnier très marqué et une variabilité de la SST qui évolue légèrement au cours des saisons (par exemple, la variabilité semble plus forte de mars à juin). Si maintenant on s'intéresse à un mois de données, du 20 mai au 20 juin 2008 (cf. figure 1.3b), on identifie clairement un cycle journalier (cf. [53]) : la température de l'eau de surface, à cette latitude, s'élève rapidement dans la journée puis décroît lentement pendant la nuit, on parle de cycle diurne. Au cours de cette thèse, nous nous efforcerons de retirer ces cycles saisonniers et journaliers qui expliquent une grande partie de la variabilité de la SST.

1.1.2 Données satellitaires

Les bouées sont faiblement abondantes dans les océans. Ainsi, il est difficile d'avoir une information sur la distribution spatiale de la SST. Or, depuis une vingtaine d'années, les mesures provenant de radiomètres embarqués sur des satellites sont disponibles et sont majoritaires par rapport aux données *in situ*. Ces instruments échantillonnent le globe en quelques jours et, pour certains, plusieurs fois par jour à des résolutions spatiales de quelques kilomètres. Les capteurs utilisés sont de nature différente. Certains, infra-rouges, ont une résolution spatiale fine mais leurs mesures de SST sont altérées par les aérosols (poussières contenues dans l'atmosphère) et le contenu en vapeur d'eau. D'autres, micro-ondes, sont moins soumis aux conditions atmosphériques mais ont une résolution plus grossière. Ainsi, chaque instrument possède des caractéristiques propres en terme de couverture, de résolu-



(a)



(b)

FIG. 1.3: Série temporelle de SST en °C mesurée par une bouée fixe du réseau PIRATA au point 0°N 23°W sur l'année 2008 (a) et du 20 mai au 20 juin 2008 (b).

Capteur	Type	Résolution spatiale	Résolution temporelle
METOP	Infra-rouge	0.05°	1 / nuit
SEVIRI	Infra-rouge	0.10°	1 / nuit
AMSRE	Micro-onde	0.25°	1 / nuit

TAB. 1.1: Caractéristiques des données SST (produits L3) issues des capteurs satellitaires METOP, SEVIRI et AMSRE utilisés dans cette thèse.

tion, de précision, d'étalonnage et de niveau de profondeur mesurée. On compte environ une dizaine de capteurs mesurant la SST mais seuls quelques-uns sont utilisés dans le manuscrit. Le premier, METOP[3], est un capteur infra-rouge défilant de très bonne résolution spatiale. Le second, SEVIRI[4], est géostationnaire (restant au dessus d'un même point et ne mesurant que l'Atlantique) avec, à l'origine, une résolution temporelle élevée. Enfin, la dernière source de données provient du capteur micro-onde AMSRE[5] qui a la particularité de rendre des observations même lorsqu'il y a une couverture nuageuse. Dans cette thèse, nous travaillons sur des données échantillonnées sur des grilles régulières (appelées produits L3) dont la résolution temporelle moyenne sera de l'ordre d'environ deux observations par jour par capteur (soit une observation par nuit). Au final, les caractéristiques des trois capteurs cités sont référencées dans le tableau 1.1.

Comme nous l'avons signalé plus haut, nous travaillons sur des données de type L3, issues de traces satellites brutes et régulièrement grillées (on parle de champs de SST), pour lesquels un traitement a été apporté. Par exemple, chaque pixel sera référencé par une valeur de SST et un indice de qualité (compris entre 1 et 5, 5 étant l'indice de confiance la plus forte), indiquant l'incertitude de l'observation ; dans cette thèse, seuls les pixels de qualité 4 et 5 seront utilisés. Un biais et une variance d'erreur seront attribués à chaque pixel en fonction de l'indice de qualité : on parlera alors de SSES[6] et pour plus de renseignements, consultez [14]. Des exemples de champs de SST des trois capteurs présentés en tableau 1.1 sont donnés en figure 1.4 : il s'agit de données enregistrées dans la nuit du 10 au 11 avril 2008 dans l'océan Atlantique. On remarque clairement les traces satellitaires des capteurs défilants METOP et AMSRE (cf. figures 1.4a et 1.4e) avec, entre chacune d'entre elles, des données manquantes. Concernant METOP, la variance d'erreur des mesures donnée en 1.4b est plus forte aux bords des fauchées ainsi qu'autour des nuages (dont la résultante sont des données manquantes). Pour AMSRE, cette variance d'erreur est fortement liée à la présence de pluie. Ainsi, comme nous le montre la figure 1.4f, l'incertitude des mesures est forte dans les systèmes dépressionnaires dont celui présent ce jour là au dessus de l'Atlantique Nord. Au final, les variances varient selon le capteur, les conditions atmosphériques et l'état de mer et SEVIRI et METOP ont une incertitude de mesure (de l'ordre de 0.25) beaucoup moins forte que celle de AMSRE (qui varie de 0.5 à 1). Nous accordons donc plus d'importance à ces deux premières sources de données infra-rouges. Néanmoins, les données micro-ondes

[3]METeorological OPerational
[4]Spinning Enhanced Visible and Infrared Imager
[5]Advanced Microwave Scanning Radiometer - Earth observing system
[6]Sensor-Specific Error Statistics

AMSRE apportent, dans le cas d'une couverture nuageuse, une information non négligeable.

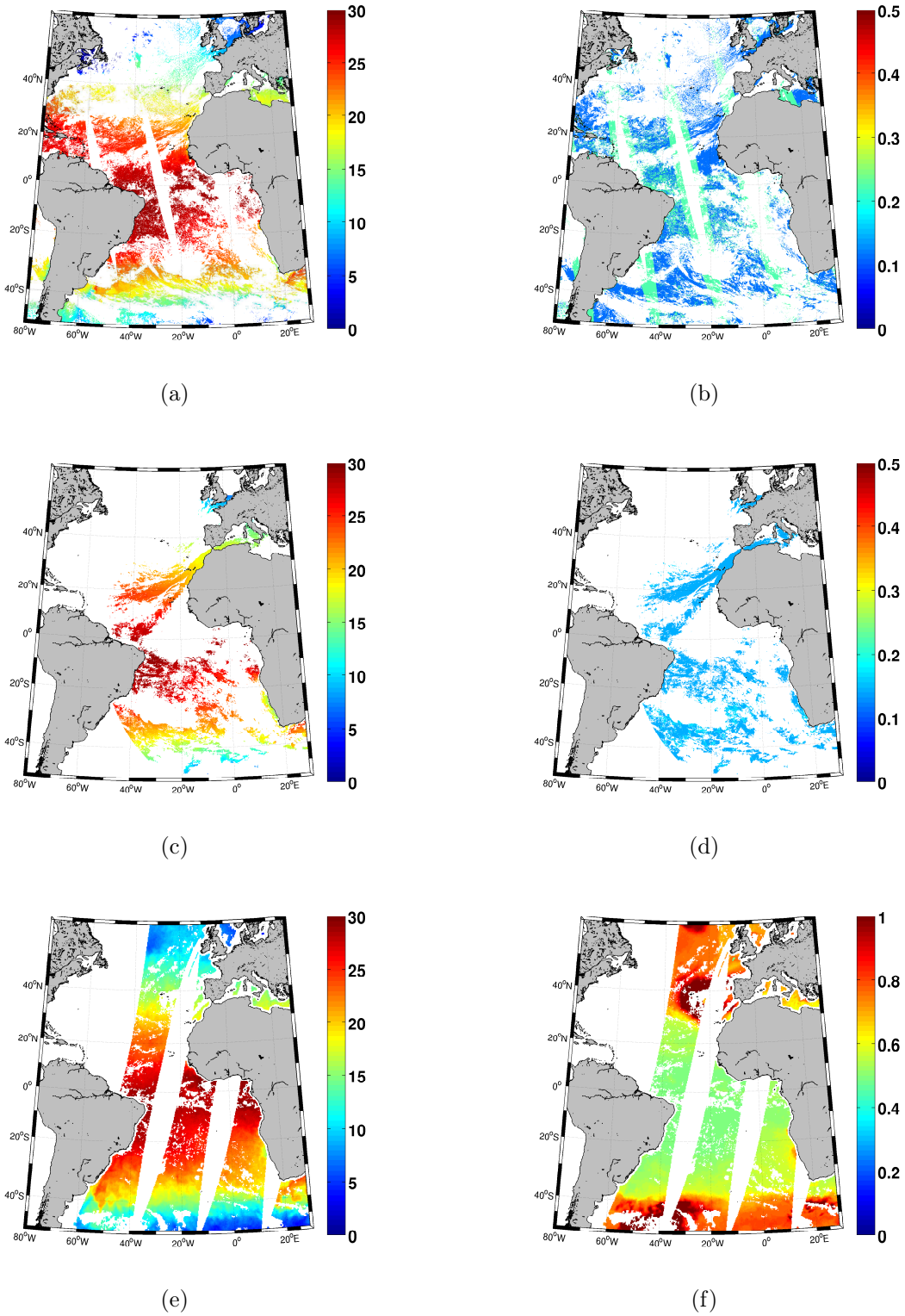
La résolution spatiale est également un critère de qualité des données et le maillage des grilles d'observations (cf. tableau 1.1) varie largement selon le capteur étudié. Pour s'en rendre compte, la figure 1.5 trace un zoom des observations METOP, SEVIRI et AMSRE (de la résolution spatiale la plus fine à la plus grossière) du 10 au 11 avril 2008 sur une zone $3° \times 3°$ autour du point situé en Atlantique Nord : 30°N 30°W. On remarque que les observations METOP (cf. figure 1.5a) sont beaucoup plus fines que celles des deux autres capteurs. Dans des conditions optimales, lorsque la couverture nuageuse est faible, les observations METOP permettent de distinguer des structures petites échelles telles que des fronts de SST, tourbillons, etc... Malheureusement, comme c'est le cas dans la nuit du 10 au 11 Avril 2008, les conditions atmosphériques devaient être mauvaises au-dessus de la zone 30°N 30°W. Ainsi, les capteurs infra-rouges METOP et SEVIRI sont entachés de nombreuses données manquantes alors que AMSRE, capteur micro-onde, a des valeurs renseignées sur presque tous ses pixels.

Dans cette thèse, nous allons projeter les observations sur une même grille fine de résolution 0.05°, correspondant à celle de METOP. Les champs des capteurs SEVIRI et AMSRE, de plus faible résolution, seront interpolés par plus proches voisins. Dans ce cas, on ne modifie en rien les observations initiales de ces deux capteurs. Les résultats de cette interpolation sont donnés en figures 1.5c et 1.5e.

Enfin, terminons par la résolution temporelle des trois sources de données en gardant l'exemple de la zone $3° \times 3°$ autour du point 30°N 30°W. Dans le figure 1.6, nous marquons les instants d'apparition des observations des différents capteurs quelques jours avant et après la nuit du 10 au 11 Avril 2008. Sachant que nous ne gardons que les données de nuit pour éviter l'effet diurne (cf. 1.3b), on observe au maximum une observation par nuit par capteur et il apparaît que, pour une même nuit, les trois sources d'observations ne soient pas enregistrées au même moment. Par exemple, dans la nuit du 10 au 11 avril 2008, les capteurs (dans l'ordre METOP, SEVIRI et AMSRE) nous retournent leurs observations à 22h21, 01h56 puis 03h45. Bien que la SST soit une variable évoluant lentement, un suivi chronologique tout au long d'une nuit peut apporter une information sur la dynamique de la SST. Notons également que parfois, pendant plusieurs jours, aucune donnée de SST n'est disponible pour un capteur donné : c'est par exemple le cas pour SEVIRI qui ne fournit rien deux nuits successives (7 au 9 avril). De plus, sachant que l'axe des z de la figure 1.6 représente le pourcentage de pixels renseignés dans chacune des images satellitaires, on remarque que le nombre de données manquantes reste élevé pour SEVIRI jusqu'au 11 avril. Pour AMSRE, ce pourcentage de pixels renseignés est élevé sur la période en question. Ce constat met en évidence la nécessité d'utiliser plusieurs sources d'informations de natures différentes (comme par exemple des capteurs infra-rouges et micro-ondes) pour pouvoir, en cas de défaillance d'un capteur ou de mauvaises conditions atmosphériques, disposer d'un minimum d'informations.

1.1.3 Données analysées

Les produits satellitaires L3 de SST grillés régulièrement, comme ceux présentés en 1.1.2, peuvent ensuite être interpolés avec une résolution plus ou moins fine. Ces champs de SST complets, sans données manquantes, sont appelés produits L4. Le plus utilisé est l'analyse

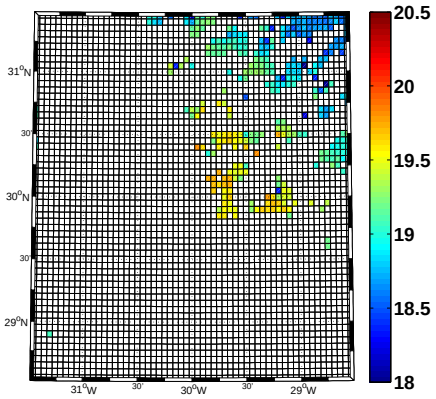


(a) (b)

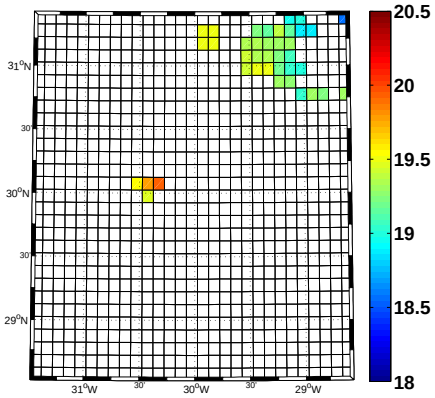
(c) (d)

(e) (f)

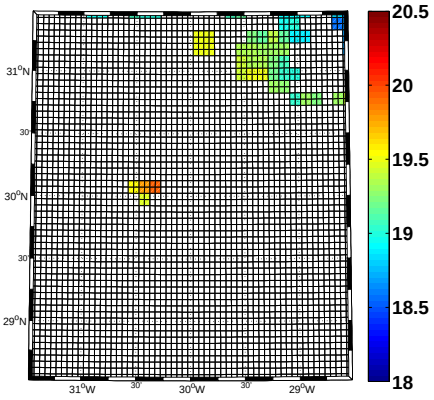
Figure 1.4: Données de SST en °C de trois capteurs satellitaires METOP (a), SEVIRI (c) et AMSRE (e) ainsi que de la variance d'erreurs des mesures correspondantes METOP (b), SEVIRI (d) et AMSRE (e) dans la nuit du 10 au 11 avril 2008 et dans l'océan Atlantique. Remarque : l'échelle des couleurs de la figure (f) n'est pas la même que celle des figures (b) et (d).



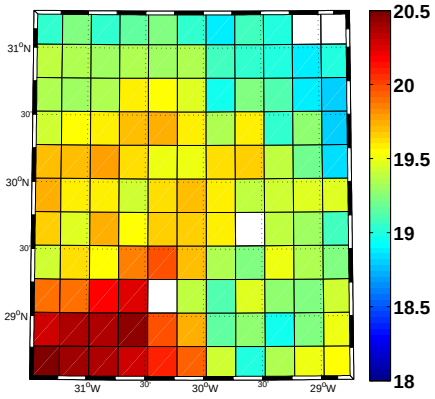
(a)



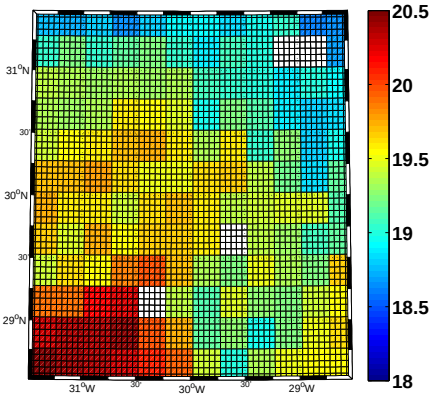
(b)



(c)



(d)



(e)

FIG. 1.5: Comparaison des résolutions spatiales initiales des trois capteurs satellitaires METOP (a), SEVIRI (b) et AMSRE (d) dans la nuit du 10 au 11 avril 2008 (respectivement 22h21, 01h56 puis 03h45) dans la zone $3° \times 3°$ autour du point 30°N 30°W. Les figures (c) et (e) représentent respectivement les champs SEVIRI et AMSRE interpolés par plus proches voisins sur la même grille que METOP soit 0.05°. Les trois figures de droite seront le type de données utilisées dans cette thèse.

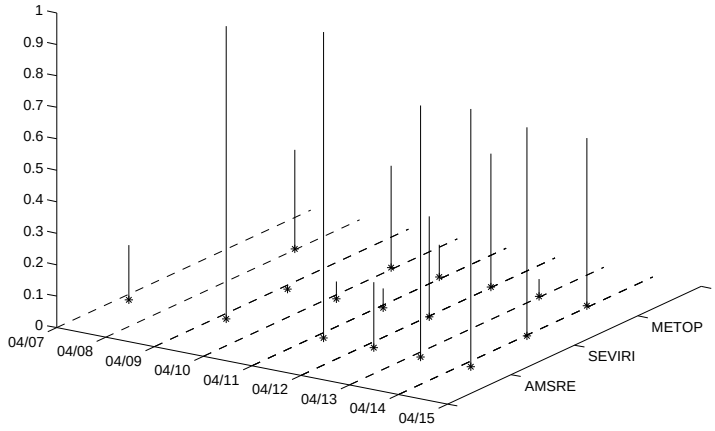
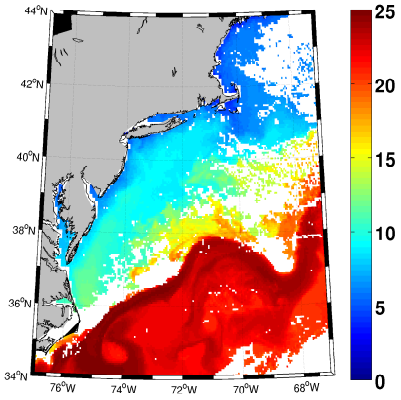


FIG. 1.6: Instants auxquels les observations METOP, SEVIRI et AMSRE sont disponibles, quelques jours avant et après la nuit du 10 au 11 avril 2008, dans la zone $3° \times 3°$ autour du point 30°N 30°W. Les traits pointillés représentent le temps à minuit et l'axe des z le pourcentage de pixels renseignés (dans la boîte $3° \times 3°$) pour chaque observation.

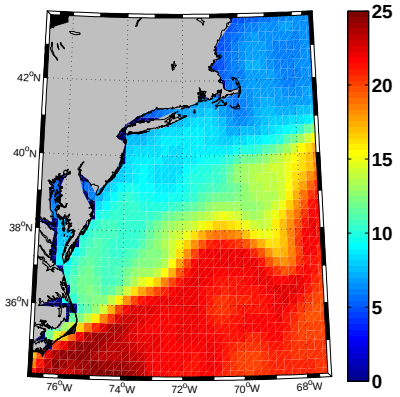
journalière globale de Reynolds OIV2[7] (cf. [46]). Cette interpolation se base sur des données satellitaires de nuit (notons que METOP et SEVIRI ne sont pas utilisés dans l'analyse OIV2) et sa grille spatiale est de résolution 0.25°. Un exemple de l'analyse sur l'Atlantique du 1^{er} janvier 2008 dans la région de Gulf Stream, au large du Canada est donné en figure 1.7b. OIV2 modélise, du fait de sa résolution, seulement les moyennes échelles de la SST et pas les fines structures comme par exemple les fronts ou tourbillons qu'on peut observer la même nuit par le capteur METOP en figure 1.7a. Nous verrons par la suite que ce produit L4 servira à stationnariser les données satellitaires afin de ne travailler que sur les variations petites échelles de la SST. Par exemple, si on prend les données du capteur METOP et qu'on y retranche l'analyse OIV2, l'échelle de SST étudiée sera comprise entre 0.05° et 0.25°, soit environ entre 5 km et 25 km. On parlera d'anomalie de SST et nous verrons que c'est cette variable qui nous intéressera par la suite. Ces résidus de SST de la nuit du 10 au 11 avril 2008 sont donnés en figure 1.7c.

D'autres analyses interpolées L4 de SST sont produites dans le monde et sont mises à disposition par le groupement GHRSST[8], accessible sur `http://www.ghrsst.org/`. Parmi elles, on trouve l'analyse de référence Française qui est pourvue par l'IFREMER de Brest, dans le laboratoire d'océanographie spatiale. Celle-ci est appelée ODYSSEA[9] (pour plus de

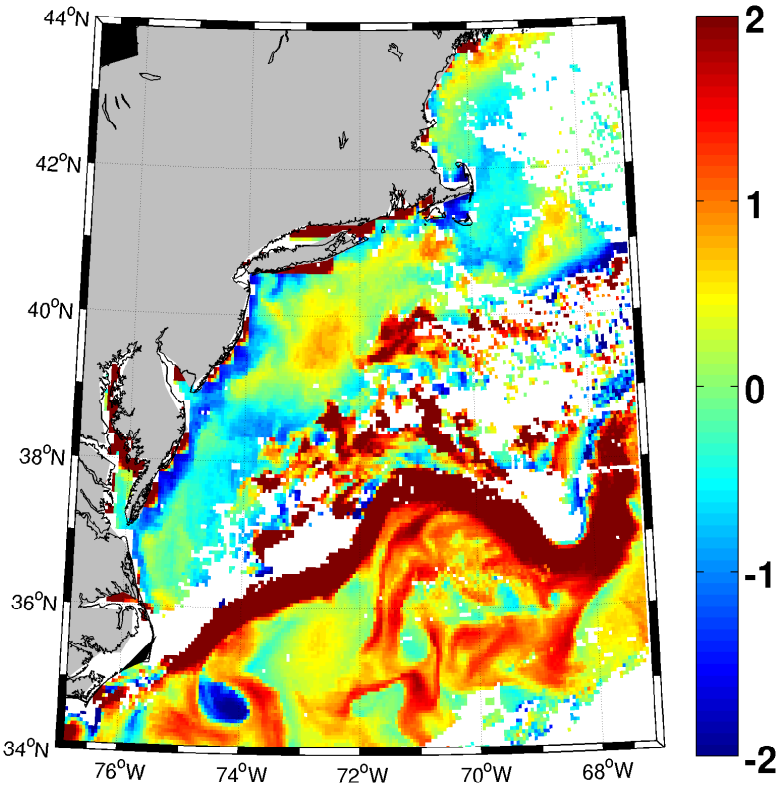
[7]Optimal Interpolation Version 2
[8]Group for High-Resolution SST
[9]Ocean Data analYsis System for merSEA



(a)



(b)



(c)

Figure 1.7: Observation METOP (a) et analyse OIV2 (b) en °C de résolution spatiale respectives 0.25° et 0.05° dans la nuit du 1^{er} janvier 2008 dans la région du Gulf Stream. (c) Anomalie de SST en °C correspondant à la différence entre METOP et OIV2.

renseignements, consultez [1]) et est produite de façon journalière à partir d'une dizaine de capteurs observant la SST. Pour résumer, les interpolations des différents produits L4 globaux sont basées sur un krigeage spatial classique (en océanographie, on parle d'interpolation optimale), sans inférence statistique propre (ni les paramètres de la variabilité spatiale ni ceux de la variabilité temporelle n'ont été estimés). De plus, les observations disponibles les jours avant et après l'analyse ne sont pas bien prises en compte alors que, comme le montre la figure 1.3a, la SST est une variable évoluant lentement dans le temps. Dans la section 1.2, nous insisterons sur les différents points à prendre en compte pour améliorer les résultats de cette interpolation.

1.1.4 Quelques mots sur la colocalisation

Aux vues de ce qui a été présenté jusqu'ici, nous disposons de plusieurs sources de données SST : *in situ*, satellitaires et analysées. Dans le but de confronter les mesures faites par des instruments embarqués sur des satellites aux mesures *in situ* qu'on considère comme référence, des bases de données dites colocalisées en temps et en espace ont été mises en place (cf. [43]). Ces bases de données enregistrent tous les couples de mesure SST capteur - *in situ* ainsi que d'autres variables physiques (vent, aérosols dans l'air, etc...) pouvant être des sources de perturbation de la donnée satellitaire SST. Lors d'un stage antérieur à cette thèse, j'ai étudié la colocalisation entre un capteur satellite précis et des mesures de bouées dérivantes. Il s'avère que ces critères de sélection des couples de mesures sont adéquats :

- distance de moins de 10 km
- période de moins de 2 heures afin d'éviter le problème du cycle diurne présenté en figure 1.3b
- une anomalie de SST entre les deux sources de mesure inférieure à 2°C afin d'écarter toutes les valeurs aberrantes (le plus souvent dues à des bouées défectueuses).

Etant donné que les données SST *in situ* sont considérées comme la référence, ces colocalisations permettent de comparer la qualité des observations des différents capteurs notamment en estimant un biais et une variance d'erreur (SSES, notion déjà abordée en 1.1.2). De plus, les covariables telles que les variables physiques peuvent servir à expliquer ces SSES. C'est un thème de recherche largement étudié dans la calibration des données satellitaires et pour la SST, on trouve beaucoup d'articles à ce sujet sur tous les différents capteurs. Par exemple, le capteur infra-rouge AATSR[10], longtemps considéré comme la référence, a fait l'objet de nombreuses publications à ce sujet (cf. [36] ou [6] par exemple). Notons qu'au cours de ces trois ans, un article a été publié sur ces mêmes données (cf. [54]) : l'objectif y est d'estimer, par régression linéaire multiple sur les covariables, le biais SSES du capteur en question. Celui-ci est ensuite comparé au biais SSES donné par le distributeur de données de AATSR et les résultats montrent une amélioration de la qualité des observations SST de ce capteur. Ce modèle simple, nécessitant peu de calculs, a ensuite trouvé son application dans la chaîne de traitement de données SST de l'AATSR au sein du laboratoire d'océanographie spatiale de l'IFREMER de Brest. Cet article est consultable en annexe A.

[10] Advanced Along Track Scanning Radiometer

Pour ce qui est de la variance d'erreur SSES, elle sera discutée tout au long de cette thèse. Nous la noterons R et nous l'estimerons dans le chapitre 3 sur l'ocean Atlantique pour un capteur donné (cf. figure 3.9d). Pour cela, nous chercherons à séparer la variabilité naturelle de la SST à celle des erreurs commises par les capteurs.

1.2 Objectifs de la thèse

Comme nous l'avons vu en 1.1.2, plusieurs sources d'observations mesurent directement ou indirectement la SST. Parmi elles, les données *in situ* et issues de capteurs satellitaires ont respectivement un échantillonnage temporel et spatial élevé avec des données manquantes dans les deux cas de figure. L'objectif de cette thèse est d'utiliser au mieux toutes ces informations afin de rendre, en temps réel, des cartes interpolées de SST qui serviront en grande partie d'entrées aux modèles océaniques physiques. Historiquement, l'IFREMER (le laboratoire d'océanographie spatiale de Brest) fournit une analyse L4 de ce type appelée ODYSSEA (cf. [1] et 1.1.3) combinant les données d'une dizaine de capteurs satellitaires. Cette thèse a pour but de répondre aux lacunes de cette analyse et de réfléchir à une méthode prenant en compte les remarques déjà abordées en 1.1 dont :

- la prise en compte de plusieurs sources d'informations provenant des trois différents capteurs. On aimerait également que le poids des observations soit inversement proportionnel aux variances d'erreurs données, pour chaque pixel, par les fournisseurs de données satellitaires
- la recherche d'une meilleure cohérence spatiale et temporelle. On souhaiterait notamment utiliser l'information de façon chronologique où chaque nouvelle observation satellite serait assimilée en temps réel. De plus, au niveau spatial, nous voudrions éviter, après interpolation, la création de structures physiquement impossibles, induites par de mauvais paramètres. Ces paramètres de dépendances spatiales et temporelles de la SST seront estimés à partir des observations pour qu'ils soient le plus réaliste possible et qu'ils apportent une information sur la physique de la SST
- la recherche d'une résolution temporelle et spatiale la plus fine possible. En effet, l'objectif est de rendre une analyse SST toutes les nuits avec une résolution spatiale comprise entre 0.05° et 0.25°, gardant au maximum en compte les structures petites échelles observées par exemple par le capteur haute résolution METOP
- la partie explicative sera aussi importante que la partie prédictive du modèle et une paramétrisation réaliste du modèle sera le principal axe de recherche.

De plus, comme nous l'avions soulevé en 1.1.3, les données analysées OIV2 ont une résolution spatiale moyenne de 0.25° et servent à stationnariser les observations des trois capteurs METOP, SEVIRI et AMSRE. Ainsi, nous ne travaillerons pas directement sur la SST mais sur l'anomalie ou résidus de SST. En effet, une grande partie du signal de la température de surface de l'eau est expliquée par un cycle annuel et un cycle journalier (cf. figure 1.3). Par la suite, nous nous efforcerons de retirer au mieux cette tendance (qui explique une grande partie de la variabilité) afin de ne travailler que sur les résidus de SST qui correspondent à la petite échelle.

1.3 Discussion

Dans ce chapitre, nous avons présenté les sources possibles de SST dont nous disposons à l'heure actuelle : données issues de bouées fixes et dérivantes ou issues de capteurs satellitaires micro-ondes et infra-rouges. Ainsi, nous avons pris conscience des composantes temporelles et spatiales du signal de SST et cette introduction a permis de mettre en évidence les principales problématiques à résoudre dans cette thèse. Le chapitre 2 présentera une famille de modèles permettant justement de prendre en compte ces différentes points.

A titre de remarque, notons qu'au cours de l'étude, nous n'aborderons pas les échanges et interactions possibles entre la surface de l'océan et l'atmosphère ainsi que la surface de l'océan et les couches d'eau plus profondes. Notre resterons dans un problème à deux dimensions, dans le sens où seule la variabilité spatiale, temporelle et les déplacements horizontaux de la SST seront étudiés. De plus, comme nous l'avons signalé au cours de ce chapitre, gardons bien en mémoire que la variable d'intérêt sera l'anomalie de SST, résultat de la différence entre les observations (satellitaires ou *in situ*) et les données analysées OIV2.

CHAPITRE 2

Modélisation espace d'état

Dans de nombreux domaines, comme en océanographie et en météorologie, on est amené à estimer une variable physique qui n'est pas forcément mesurable directement. On parlera alors de variable cachée, latente ou vraie que nous nommerons l'état. Dans cette thèse, cette variable d'état est la "vraie" SST qui, d'après 1.1, est partiellement observée et mesurée indirectement par différents capteurs satellites. Etant donné l'incertitude des observations satellitaires, nous aimerions qu'elles soient pondérées par un terme d'erreur. De plus, comme nous l'expliquons en 1.3, nous avons des connaissances sur le comportement dynamique de la SST. Il est alors possible d'écrire un modèle théorique de l'état. Or, nous supposons que celui-ci n'est pas parfait étant donné sa simplicité et nous l'entacherons d'un terme d'erreur. Au final, nous aimerions utiliser les sources d'information disponibles (observations et modèle) en prenant en compte leurs incertitudes respectives. C'est le principe de l'assimilation de données qui est utilisée en océanographie et météorologie (cf. [20]). Mathématiquement, les modèles espace d'état permettent de spécifier la relation entre l'état et les observations du système. Ces modèles font partie de la famille des processus de Markov cachés où la variable d'état est latente et continue.

Dans un premier temps, en 2.1, nous présenterons un modèle espace d'état où les équations sont linéaires et où les termes d'erreurs suivent des lois normales. Ces hypothèses sur la structure des erreurs et la linéarité du modèle conduisent à une inférence statistique simple et plusieurs approches pour estimer l'état et les paramètres seront détaillées en 2.2. Enfin, nous terminerons ce chapitre par une discussion générale autour de cette modélisation espace d'état en 2.3.

2.1 Modèle linéaire à erreurs Gaussiennes

Les modèles espace d'état se déclinent en plusieurs sous-parties. La première segmentation se fait selon la linéarité ou non du modèle. En effet, dans le cas où les relations entre l'état et les observations ne peuvent pas s'écrire de façon exacte par produit matriciel, nous serons

dans le cas de non linéarité. La seconde différence se fait sur la distribution des erreurs du modèle. Celles-ci peuvent suivre des lois normales ou non.

Dans cette thèse, nous nous intéresserons au cas particulier des modèles espace d'état linéaires à erreurs Gaussiennes. Dans la communauté statistique, on parle également de modèles linéaires dynamiques (cf. [50]). Nous commencerons par introduire les notations spécifiques à ces modèles espace d'état en 2.1.1. Ensuite, nous présenterons en 2.1.2 une version simple du modèle, où nous avons une connaissance *a priori* sur l'état et où celui-ci est également observé. Enfin, nous introduirons en 2.1.3 une formulation séquentielle dans le temps de ce modèle espace d'état particulier.

2.1.1 Notations

Dans cette thèse, nous adoptons partiellement une notation spécifique aux domaines de l'océanographie et de la météorologie. Celle-ci a été proposée par [28]. Le tableau 2.1 donne les principaux vecteurs et matrices utilisés, leur dimension ainsi que leur description. La variable d'état est de dimension r alors que les observations sont de dimensions p (avec généralement $p \leq r$). Ainsi, nous insistons sur le fait que la variable d'état peut être complètement, partiellement ou même non observée.

2.1.2 Formulation ponctuelle

Nous commençons par étudier la forme la plus simple d'un modèle espace d'état linéaire à erreurs Gaussiennes. Notons $\{Y\}$ l'observation de la variable d'état $\{X\}$. On suppose que l'observation est reliée à l'état par l'équation de mesure suivante :

$$Y = HX + R^{1/2}\varepsilon \tag{2.1}$$

où $\{\varepsilon\}$ est un bruit blanc de loi normale de moyenne nulle et de variance unité. En pratique, R représente la variance de l'erreur d'observation et H est l'opérateur d'observation permettant une transformation linéaire entre l'état et l'observation. On définie $R^{1/2}$ comme étant une racine carrée possible de la matrice R.

Nous supposons que le processus latent $\{X\}$ est lié à une information connue *a priori* $x^{(b)}$ par l'équation d'état suivante :

$$X = x^{(b)} + B^{1/2}\eta^{(b)} \tag{2.2}$$

où $\left\{\eta^{(b)}\right\}$ est également un bruit blanc de loi normale de moyenne nulle et de variance unité. La variance de l'erreur commise par $x^{(b)}$ est notée B. Les erreurs Gaussiennes $\{\varepsilon\}$ et $\left\{\eta^{(b)}\right\}$ sont supposées indépendantes.

2.1.3 Formulation séquentielle

La formulation simple du modèle espace d'état est généralisable dans le temps. En effet, nous supposons maintenant que le processus latent est observé à plusieurs instants $(t_1, ..., t_n)$.

Notation	Dimension	Description
x	$\mathbb{R}^r$	vecteur d'état
M	$\mathbb{R}^{r\times r}$	matrice opérateur modèle
η	$\mathbb{R}^r$	vecteur d'erreur modèle
Q	$\mathbb{R}^{r\times r}$	matrice de covariance de η
y	$\mathbb{R}^p$	vecteur des observations
H	$\mathbb{R}^{p\times r}$	matrice opérateur d'observation
ε	$\mathbb{R}^p$	vecteur d'erreur d'observation
R	$\mathbb{R}^{p\times p}$	matrice de covariance de ε
$x^{(b)}$	$\mathbb{R}^r$	vecteur d'état *a priori* (b pour "background")
$\eta^{(b)}$	$\mathbb{R}^r$	vecteur d'erreur de l'état *a priori*
B	$\mathbb{R}^{r\times r}$	matrice de covariance de $\eta^{(b)}$
$x^{(f)}$	$\mathbb{R}^r$	vecteur d'état propagé (f pour "forecast")
$\eta^{(f)}$	$\mathbb{R}^r$	vecteur d'erreur de l'état propagé
$P^{(f)}$	$\mathbb{R}^{r\times r}$	matrice de covariance de $\eta^{(f)}$
$x^{(a)}$	$\mathbb{R}^r$	vecteur d'état filtré (a pour "analysis")
$\eta^{(a)}$	$\mathbb{R}^r$	vecteur d'erreur de l'état filtré
$P^{(a)}$	$\mathbb{R}^{r\times r}$	matrice de covariance de $\eta^{(a)}$
$x^{(s)}$	$\mathbb{R}^r$	vecteur d'état lissé (s pour "smooth")
$\eta^{(s)}$	$\mathbb{R}^r$	vecteur d'erreur de l'état lissé
$P^{(s)}$	$\mathbb{R}^{r\times r}$	matrice de covariance de $\eta^{(s)}$
d	$\mathbb{R}^p$	vecteur d'innovation
Σ	$\mathbb{R}^{p\times p}$	matrice de covariance d'erreur de d
K	$\mathbb{R}^{r\times p}$	matrice de gain

TAB. 2.1: Notation, dimension et description des vecteurs et matrices utilisés dans la modélisation espace d'état.

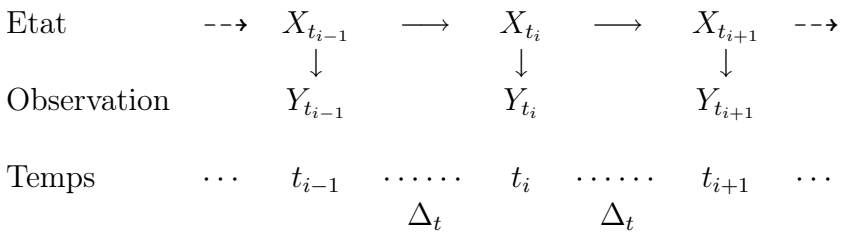


Figure 2.1: Graphique acyclique directionnel d'un modèle espace d'état avec un échantillonnage temporel régulier.

Par la suite, nous notons $\triangle_{t_i} = t_i - t_{i-1}$ la différence de temps entre deux observations successives. Dans ce chapitre, nous nous plaçons dans le cas d'un échantillonnage temporel régulier des données où $\Delta_{t_i} = ... = \Delta_{t_n} = \Delta_t$. Le cas irrégulier sera abordé dans le chapitre 3. Notons $y_{t_1}^{t_n} = (y_{t_1}, ..., y_{t_n})$ le vecteur des différentes observations.

Les processus $\{X_t\}$ et $\{Y_t\}$ sont supposés stationnaires dans le temps. Comme pour le modèle présenté en 2.1.2, l'observation est reliée à l'état par l'équation de mesure suivante :

$$Y_{t_i} = HX_{t_i} + R^{1/2}\varepsilon_{t_i} \tag{2.3}$$

Concernant l'équation d'état, une connaissance *a priori* du système ($x^{(b)}$ et B) est toujours requise puis l'état est propagé d'un temps t_{i-1} à t_i en suivant un modèle linéaire défini par :

$$X_{t_i} = MX_{t_{i-1}} + Q^{1/2}\eta_{t_i} \tag{2.4}$$

où $\{\eta_{t_i}\}_{i\in\{2,..,n\}}$ est un bruit blanc de loi normale de moyenne nulle et de variance unité, indépendant de $\{\varepsilon_{t_i}\}_{i\in\{1,..,n\}}$.

Au final, la structure Markovienne du modèle espace d'état (2.4-2.3) est résumée par un graphique acyclique directionnel (cf. figure 2.1). Remarquons qu'il existe des formes plus évoluées de ce type de modèles espace d'état. On peut notamment supposer que l'état ou les observations dépendent de covariables ou sont assujettis à un biais. Nous pouvons également penser que l'état est observé par plusieurs sources de données avec des paramètres H et R qui varient selon la nature de $\{Y_t\}$.

2.2 Inférence statistique

Une fois le modèle espace d'état présenté, nous allons parler de l'inférence statistique. Celle-ci se décompose en deux étapes que nous présenterons dans cette section. La première consiste à estimer les paramètres de notre modèle. La seconde réside dans l'estimation de l'état du système connaissant les paramètres du modèle ; par abus de langage, cette étape sera nommée résolution. Dans une formulation séquentielle comme celle présentée en 2.1.3, il y a plusieurs manières d'estimer l'état à un temps t_i donné. Nous parlerons de :

- filtrage si on utilise les informations passées $(t_1, ..., t_i)$
- lissage (plus connu sous le nom de réanalyse en océanographie) si on utilise toutes les informations passées et futures $(t_1, ..., t_n)$

– prédiction si on utilise les informations antérieures à t_i.

Nous commencerons, en 2.2.1, par présenter l'estimation de l'état conditionnellement aux observations et aux paramètres qui seront supposés connus et fixés. Enfin, nous estimerons les paramètres du modèle par maximum de vraisemblance en 2.2.2.

2.2.1 Résolution

Avant de discuter de l'estimation des paramètres, il est nécessaire d'aborder la résolution des modèles espace d'état présentés en 2.1. Nous commencerons par évoquer le cas du modèle ponctuel simple introduit en 2.1.2. Nous présenterons deux approches possibles permettant de résoudre ce modèle : la première sera basée sur la nature stochastique des processus étudiés en 2.2.1.1. On y présentera le meilleur estimateur linéaire non biaisé au sens des moindres carrés puis nous montrerons l'équivalence avec le point de vue des probabilités conditionnelles. La deuxième approche, proposée en 2.2.1.2, sera basée sur la minimisation d'une fonction de coût. Nous y montrerons l'équivalence entre les deux méthodes de résolution proposées. Enfin, nous insisterons sur l'approche stochastique qui sera utilisée tout au long de la thèse en 2.2.1.3. Nous travaillerons sur la formulation séquentielle du modèle défini en 2.1.3 et présenterons les probabilités de filtrage et de lissage qui peuvent être calculées grâce au filtre et lisseur de Kalman.

2.2.1.1 Approche stochastique

Nous travaillerons sur le modèle simple présenté en 2.1.2. L'objectif est de déterminer $x^{(a)}$, l'estimateur non biaisée de l'état $\{X\}$ et de variance $P^{(a)}$ minimale au sens des moindres carrés. $x^{(a)}$ est un compromis entre l'*a priori* $x^{(b)}$ et l'observation $\{Y\}$.

Nous aborderons deux points de vue pour répondre à cet objectif. Dans le premier, en 2.2.1.1.1, nous estimerons un coefficient de régression ou gain noté K permettant de prendre en compte au mieux l'*a priori* et l'observation. Le second point de vue, détaillé en 2.2.1.1.2, sera basé sur des probabilités conditionnelles. Finalement, on retombera sur un résultat classique qui, dans le cadre Gaussien, donne l'équivalence entre la prédiction par moindre carrés et celle par espérances conditionnelles.

2.2.1.1.1 Point de vue de l'algèbre linéaire Plaçons-nous dans le cas où le vecteur d'observation et d'état sont respectivement de dimension p et r, avec $p \leq r$. Ainsi, l'état peut être partiellement observé. On cherche un estimateur *a posteriori* $x^{(a)}$ prenant en compte, de façon linéaire, les deux sources d'information $x^{(b)}$ et $\{Y\}$. Introduisons alors le régresseur ou gain $K \in \mathbb{R}^{r\times p}$, de telle sorte que $x^{(a)}$ soit un estimateur non biaisé de $\{X\}$. On définit K de telle sorte qu'il vérifie la relation suivante :

$$x^{(a)} = (I - KH)\, x^{(b)} + Ky \tag{2.5}$$

où $H \in \mathbb{R}^{p\times r}$ est la matrice d'observation nécessaire pour que l'état et $x^{(b)}$ soient de même dimension.

Soit $\eta^{(a)} = x^{(a)} - x$, le vecteur d'erreur de l'état filtré. D'après (2.5), on développe son expression et il devient :

$$\begin{aligned} \eta^{(a)} &= x^{(b)} - x + K\left(y - Hx^{(b)} + Hx - Hx\right) \\ &= \eta^{(b)} + K\left(\varepsilon - H\eta^{(b)}\right) \end{aligned} \tag{2.6}$$

La matrice de covariance de $\eta^{(a)}$ est notée $P^{(a)} = \eta^{(a)}\eta^{(a)'}$. D'après (2.6), $P^{(a)}$ s'écrit :

$$P^{(a)} = \eta^{(b)}\eta^{(b)'} + K\left(\varepsilon - H\eta^{(b)}\right)\left(\varepsilon' - \eta^{(b)'}H'\right)K' + \eta^{(b)}\left(\varepsilon' - \eta^{(b)'}H'\right)K' + K\left(\varepsilon - H\eta^{(b)}\right)\eta^{(b)'}$$

Or, on rappelle que ε et $\varepsilon^{(b)}$ sont indépendants. Ainsi, $P^{(a)}$ se simplifie et devient :

$$P^{(a)} = B + K\left(R + HBH'\right)K' - BH'K' - KHB \tag{2.7}$$

On rappelle que notre objectif est d'estimer $x^{(a)}$ de telle sorte qu'il soit de variance $\left(P^{(a)}\right)$ minimale. Matriciellement, cela revient à minimiser la trace de $P^{(a)}$ notée $Tr\left(P^{(a)}\right)$ et à en déduire la matrice de gain K. D'après (2.7), on écrit alors :

$$\begin{aligned} \frac{\partial Tr\left(P^{(a)}\right)}{\partial K} &= \frac{\partial Tr\left(B + K\left(R + HBH'\right)K' - BH'K' - KHB\right)}{\partial K} \\ &= \frac{\partial Tr\left(B\right)}{\partial K} + \frac{\partial Tr\left(K\left(R + HBH'\right)K'\right)}{\partial K} - \frac{\partial Tr\left(BH'K'\right)}{\partial K} - \frac{\partial Tr\left(KHB\right)}{\partial K} \end{aligned}$$

En supposant la symétrie de B et R, on peut alors écrire que

$$\frac{\partial Tr\left(P^{(a)}\right)}{\partial K} = 2K(R + HBH') - BH' - BH'$$

Ainsi, $P^{(a)}$ est minimale pour K s'écrivant

$$K = BH'\left(R + HBH'\right)^{-1} \tag{2.8}$$

Finalement, d'après (2.5) et (2.8), $x^{(a)}$ s'écrit :

$$\begin{aligned} x^{(a)} &= \left(I - KH\right)x^{(b)} + Ky \\ &= \left(I - BH'\left(R + HBH'\right)^{-1}H\right)x^{(b)} + BH'\left(R + HBH'\right)^{-1}y \end{aligned} \tag{2.9}$$

Et d'après (2.7) et (2.8), $P^{(a)}$ s'écrit :

$$\begin{aligned} P^{(a)} &= B + K\left(R + HBH'\right)K' - BH'K' - KHB \\ &= B + BH'K' - BH'K' - KHB \\ &= \left(I - KH\right)B \\ &= \left(I - BH'\left(R + HBH'\right)^{-1}H\right)B \end{aligned} \tag{2.10}$$

2.2.1.1.2 Point de vue des probabilités conditionnelles Essayons maintenant de répondre à la même problématique en utilisant une approche basée sur les probabilités conditionnelles. L'équation d'observation (2.1) et l'équation d'état (2.2) nous donnent respectivement que :

- X suit une loi normale d'espérance $x^{(b)}$ et de variance B. Il admet donc pour densité

$$p(x) = \frac{1}{(2\pi)^{r/2}|B|^{1/2}} \exp\left(-\frac{1}{2}\left(x - x^{(b)}\right)' B^{-1}\left(x - x^{(b)}\right)\right)$$

- la loi conditionnelle de Y sachant $X = x$ suit une loi normale d'espérance y et de variance R. Ainsi, la loi conditionnelle de Y sachant $X = x$ admet pour densité

$$p(y|x) = \frac{1}{(2\pi)^{r/2}|R|^{1/2}} \exp\left(-\frac{1}{2}(y - Hx)' R^{-1}(y - Hx)\right)$$

Or, nous cherchons à déterminer la densité de probabilité conditionnelle $p(x|y)$ et d'après Bayes,

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

Pour des raisons de lisibilité dans les calculs, nous choisissons de travailler dans le cas où l'état et l'observation sont des scalaires soit $r = p = 1$. De plus, nous supposons qu'aucun opérateur d'observation n'est nécessaire pour rendre l'observation et l'état homogènes, soit $H = 1$. D'après les équations ci-dessous, B et R désignaient des matrices ; elles seront maintenant des scalaires. On a :

$$p(y) = \int_{-\infty}^{\infty} p(y|x)p(x)\ \mathrm{d}x$$

De plus,

$$p(y|x)p(x) = \frac{1}{\sqrt{2\pi B}}\frac{1}{\sqrt{2\pi R}} \exp\left(-\frac{1}{2}\left(\frac{\left(x - x^{(b)}\right)^2}{B} + \frac{(y-x)^2}{R}\right)\right) \tag{2.11}$$

Posons σ et A tels que :

$$\begin{aligned}\frac{1}{\sigma^2} &= \frac{1}{B} + \frac{1}{R} \\ &= \frac{B+R}{BR}\end{aligned}$$

et

$$\begin{aligned}A &= \frac{\left(x - x^{(b)}\right)^2}{B} + \frac{(y-x)^2}{R} \\ &= \frac{x^2}{B} + \frac{x^2}{R} - 2\frac{x^{(b)}}{B}x - 2\frac{y}{R}x + \frac{x^{(b)^2}}{B} + \frac{y^2}{R} \\ &= \frac{1}{\sigma^2}\left(x^2 - 2\sigma^2\left(\frac{x^{(b)}}{B} + \frac{y}{R}\right)x + \frac{\sigma^2}{B}x^{(b)^2} + \frac{\sigma^2}{R}y^2\right)\end{aligned}$$

Posons également μ et c tels que :

$$\begin{aligned}\mu &= \sigma^2\left(\frac{x^{(b)}}{B}+\frac{y}{R}\right)\\ &= \frac{R}{R+B}x^{(b)}+\frac{B}{R+B}y\end{aligned}$$

et

$$c=\frac{\sigma^2}{B}x^{(b)^2}+\frac{\sigma^2}{R}y^2$$

A s'écrit alors :

$$\begin{aligned}A &= \frac{1}{\sigma^2}\left(x^2-2\mu x+c\right)\\ &= \frac{1}{\sigma^2}\left(\left(x-\mu\right)^2-\mu^2+c\right)\end{aligned}$$

Soit,

$$\begin{aligned}c-\mu^2 &= \frac{R}{B+R}x^{(b)^2}+\frac{B}{B+R}y^2-\frac{R^2}{\left(B+R\right)^2}x^{(b)^2}-\frac{B^2}{\left(B+R\right)^2}y^2-2\frac{BR}{\left(B+R\right)^2}x^{(b)}y\\ &= \frac{BR}{\left(B+R\right)^2}x^{(b)^2}+\frac{BR}{\left(B+R\right)^2}y^2-2\frac{BR}{\left(B+R\right)^2}x^{(b)}y\\ &= \frac{\sigma^2}{B+R}\left(x^{(b)}-y\right)^2\end{aligned}$$

On rappelle que :

$$\begin{aligned}p\left(y|x\right)p\left(x\right) &= \frac{1}{\sqrt{2\pi B}}\frac{1}{\sqrt{2\pi R}}\exp\left(-\frac{A}{2}\right)\\ &= \frac{1}{\sqrt{2\pi B}}\frac{1}{\sqrt{2\pi R}}\exp\left(-\frac{1}{2}\left(\frac{\left(x-\mu\right)^2}{\sigma^2}\right)\right)\exp\left(-\frac{1}{2}\left(\frac{\left(x^{(b)}-y\right)^2}{B+R}\right)\right)\end{aligned}$$

Or,

$$\begin{aligned}p\left(y\right) &= \int_{-\infty}^{\infty}p(y|x)p\left(x\right)\,\mathrm{d}x\\ &= \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi B}}\frac{1}{\sqrt{2\pi R}}\exp\left(-\frac{1}{2}\left(\frac{\left(x-\mu\right)^2}{\sigma^2}\right)\right)\exp\left(-\frac{1}{2}\left(\frac{\left(x^{(b)}-y\right)^2}{B+R}\right)\right)\,\mathrm{d}x\\ &= \frac{1}{\sqrt{2\pi B}}\frac{1}{\sqrt{2\pi R}}\exp\left(-\frac{1}{2}\left(\frac{\left(x^{(b)}-y\right)^2}{B+R}\right)\right)\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2}\left(\frac{\left(x-\mu\right)^2}{\sigma^2}\right)\right)\,\mathrm{d}x\end{aligned}$$

On sait que pour une variable aléatoire suivant une loi normale d'espérance μ et de variance σ^2,

$$\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\left(\frac{\left(x-\mu\right)^2}{\sigma^2}\right)\right)\,\mathrm{d}x=1$$

Soit,

$$\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\left(\frac{(x-\mu)^2}{\sigma^2}\right)\right) \mathrm{d}x = \sqrt{2\pi}\sigma$$

Donc, $p(y)$ s'écrit

$$p(y) = \frac{1}{\sqrt{2\pi B}} \frac{1}{\sqrt{2\pi R}} \exp\left(-\frac{1}{2}\left(\frac{\left(x^{(b)}-y\right)^2}{B+R}\right)\right) \sqrt{2\pi}\sigma \tag{2.12}$$

Finalement, d'après Bayes, (2.11) et (2.12),

$$\begin{aligned} p(x|y) &= \frac{p(y|x)p(x)}{p(y)} \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{(x-\mu)^2}{\sigma^2}\right)\right) \end{aligned}$$

L'espérance μ et la variance σ^2 de la densité de probabilité de la loi normale conditionnelle $p(x|y)$ sont respectivement égaux (pour $H = 1$) à l'estimateur optimal linéaire non biaisée $x^{(a)}$, de variance $P^{(a)}$ donnés respectivement en (2.9) et (2.10). Donc, la loi conditionnelle de X connaissant $Y = y$ suivant une loi normale $N\left(x^{(a)}, P^{(a)}\right)$ est optimale au sens des moindre carrés. Ce résultat est illustré par un exemple (cf. figure 2.2) où, à partir d'une loi sur l'état *a priori* $N\left(x^{(b)}, B\right)$ ainsi que la loi conditionnelle de Y sachant $X = x$ suivant $N(y, R)$, on obtient un compromis suivant la loi normale :

$$P(X|Y=y) = N\left(R(B+R)^{-1}x^{(b)} + B(B+R)^{-1}y, BR(B+R)^{-1}\right)$$

Notons que ces résultats sont obtenus uniquement dans le cas d'un modèle linéaire avec des erreurs suivant des lois normales. Pour les cas des erreurs non-Gaussiennes, il existe d'autres expressions analytiques. Il est également possible d'avoir une estimation de l'état *a posteriori* grâce à des simulations de différentes lois comme la méthode Monte-Carlo.

2.2.1.2 Approche variationnelle

L'approche variationnelle consiste à minimiser une fonction de coût J. Celle-ci est une mesure de la distance entre la variable d'état et les observations ainsi que la distance à une information *a priori*, toutes les deux normées par leurs matrices de covariance respectives. Dans l'approche stochastique définie en 2.2.1.1, nous cherchions une densité de probabilité de loi normale *a posteriori* et estimions la moyenne et la variance. Ici, on cherche à minimiser J, qui, dans le cas d'un modèle espace d'état non linéaire, nécessite l'utilisation de techniques d'optimisation numérique de type descente de gradient et dans le cas linéaire peut être résolue par algèbre linéaire.

Nous aborderons la construction de la fonction de coût J en 2.2.1.2.1 puis nous la minimiserons en 2.2.1.2.2 en montrant l'équivalence avec l'approche stochastique. Enfin, en 2.2.1.2.3, nous discuterons du choix de l'approche utilisée dans cette thèse.

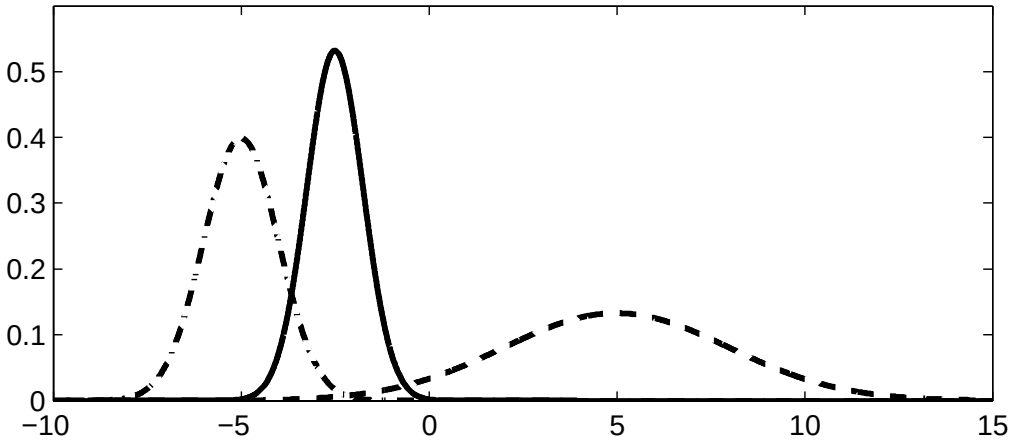


FIG. 2.2: Estimation scalaire de la densité de probabilité de l'état *a posteriori* (trait continu) à partir des densités de probabilité de lois normales de l'état *a priori* (trait tiret-pointillé) et de l'observation (trait discontinu).

2.2.1.2.1 Fonction de coût J Le problème est identique à la formulation stochastique. On cherche un estimateur *a posteriori* de l'état $\{X\}$ connaissant un *a priori* $x^{(b)}$ et une observation $\{Y\}$ avec leurs erreurs respectives $\eta^{(b)} = x^{(b)} - x$ et $\varepsilon = y - Hx$. Rappelons que :

$$\begin{aligned} E\left(\eta^{(b)}\right) &= 0 \\ E\left(\varepsilon\right) &= 0 \end{aligned}$$

et

$$\begin{aligned} Var\left(\eta^{(b)}\right) &= B \\ Var\left(\varepsilon\right) &= R \end{aligned}$$

On écrit alors une fonction de coût J prenant en compte les erreurs $\left\{\eta^{(b)}\right\}$ et $\{\varepsilon\}$, normalisées par leurs variances respectives :

$$J(x) = \frac{1}{2}\left(Hx - y\right)' R^{-1}\left(Hx - y\right) + \frac{1}{2}\left(x - x^{(b)}\right)' B^{-1}\left(x - x^{(b)}\right)$$

On reconnait la prise en compte de deux sources d'informations : l'observation et l'état *a priori*. Celles-ci sont pondérées par leur covariances d'erreurs respectives.

2.2.1.2.2 Equivalence avec l'approche stochastique La solution de l'approche variationnelle est obtenue pour $x^{(a)}$ défini par

$$x^{(a)} = \arg\min_x\left(J(x)\right) \tag{2.13}$$

Résolvons (2.13) par algèbre linéaire. On cherche la valeur $x^{(a)}$ qui annule la dérivée de J par rapport à x qui vaut :

$$\begin{aligned}
\frac{\partial J(x)}{\partial x} &= -H'R^{-1}\left(y - Hx\right) + B^{-1}\left(x - x^{(b)}\right) \\
&= -H'R^{-1}y + H'R^{-1}Hx + B^{-1}\left(x - x^{(b)}\right) \\
&= -H'R^{-1}y + H'R^{-1}Hx^{(b)} + H'R^{-1}H\left(x - x^{(b)}\right) + B^{-1}\left(x - x^{(b)}\right) \\
&= -H'R^{-1}\left(y - Hx^{(b)}\right) + \left(H'R^{-1}H + B^{-1}\right)\left(x - x^{(b)}\right) \\
&= 0
\end{aligned}$$

Ainsi, grâce à la formule de Sherman–Morrison–Woodbury (cf. [52]),

$$\begin{aligned}
x^{(a)} &= x^{(b)} + \left(B^{-1} + H'R^{-1}H\right)^{-1} H'R^{-1}\left(y - Hx^{(b)}\right) \\
&= x^{(b)} + BH'\left(R + HBH'\right)^{-1}\left(y - Hx^{(b)}\right) \\
&= \left(I - BH'\left(R + HBH'\right)^{-1}H\right)x^{(b)} + BH'\left(R + HBH'\right)^{-1}y
\end{aligned}$$

On retombe sur le même résultat que l'équation (2.9). Donc, uniquement dans le cas linéaire, l'estimation par l'approche variationnelle est équivalente à celle de l'approche stochastique. Cependant, remarquons que l'approche présentée ici ne rend pas d'estimation de la variance $P^{(a)}$.

2.2.1.2.3 Pourquoi l'approche stochastique ? Les deux approches présentées en 2.2.1.2 et 2.2.1.1 sont utilisées dans la communauté océanographique et météorologique. L'approche variationnelle est utilisée essentiellement par les numériciens alors que l'approche stochastique, basée sur une estimation non biaisée de variance minimale, est préférée par les statisticiens. Dans le cas d'un modèle espace d'état linéaire et Gaussien, l'état estimé $x^{(a)}$ par les deux méthodes est identique mais seule l'approche stochastique rend une estimation de la variance analysée $P^{(a)}$. Nous verrons plus loin, en 2.2.2, l'utilité de celle-ci dans l'estimation des paramètres d'un modèle espace d'état. Cette thèse s'inscrivant dans un cadre statistique et probabiliste, l'approche stochastique sera celle adoptée.

2.2.1.3 Résolution séquentielle *via* Kalman

La résolution séquentielle est une généralisation, aux différents temps $(t_1, ..., t_n)$, de la résolution ponctuelle stochastique présentée en 2.2.1.1. Nous travaillerons donc ici sur le modèle espace d'état linéaire à erreurs Gaussiennes proposé en 2.1.3. Nous utiliserons les propriétés Markoviennes du modèle afin d'estimer les densités des probabilités conditionnelles successives au cours du temps.

Nous estimerons la densité de probabilité *a posteriori* de l'état, conditionnellement à des observations et à une valeur initiale. Lorsque seules les observations passées seront utilisées, comme en 2.2.1.3.1, nous estimerons les probabilités du filtrage de Kalman. Dans le cas où toutes les observations (passées et futures) seront prises en compte, nous aborderons le lisseur de Kalman présenté en 2.2.1.3.2.

2.2.1.3.1 Filtre de Kalman Nous utilisons ici l'indiçage temporel des vecteurs d'état et d'observation ainsi que le modèle présenté en 2.1.3. Dans celui-ci, nous utiliserons les deux équations pemettant l'estimation du processus $\{X_t\}$. D'une part, l'équation séquentielle (2.4) propagera l'information de l'état précédent grâce à la matrice opérateur modèle M et la covariance d'erreur du modèle Q. D'autre part, l'équation (2.3) permettra de prendre en compte l'observation grâce à la matrice opérateur H et la covariance d'erreur des observations R. Ainsi, l'information apportée par les observations à différents instants $(t_1, ..., t_n)$ sera propagée tout au long de l'analyse. Ceci permet à l'analyse d'avoir une meilleure continuité et d'être plus réaliste.

Dans le cas du filtrage, l'objectif est d'estimer la densité de probabilité conditionnelle de X_{t_i} connaissant un *a priori* $x^{(b)}$ de l'état à l'instant initial, les observations passées $y_{t_1}^{t_i} = (y_{t_1}, ..., y_{t_i})$ et les paramètres du modèle que nous noterons $\theta = (M, H, Q, R)$. Cette densité de probabilité conditionnelle sera notée $P\left(X_{t_i}|y_{t_1}^{t_i};\theta\right)$. Comme nous l'avons démontré dans le cas de la formulation ponctuelle en 2.2.1.1, $P\left(X_{t_i}|y_{t_1}^{t_i};\theta\right)$ est une loi normale d'espérance $x_{t_i}^{(a)}$ et de variance $P_{t_i}^{(a)}$. Les estimations successives $x_{t_i}^{(a)}$ et $P_{t_i}^{(a)}$, $\forall i \in \{1, .., n\}$, seront données par l'algorithme du filtre de Kalman décrit ci-dessous et schématisé en figure 2.3a.

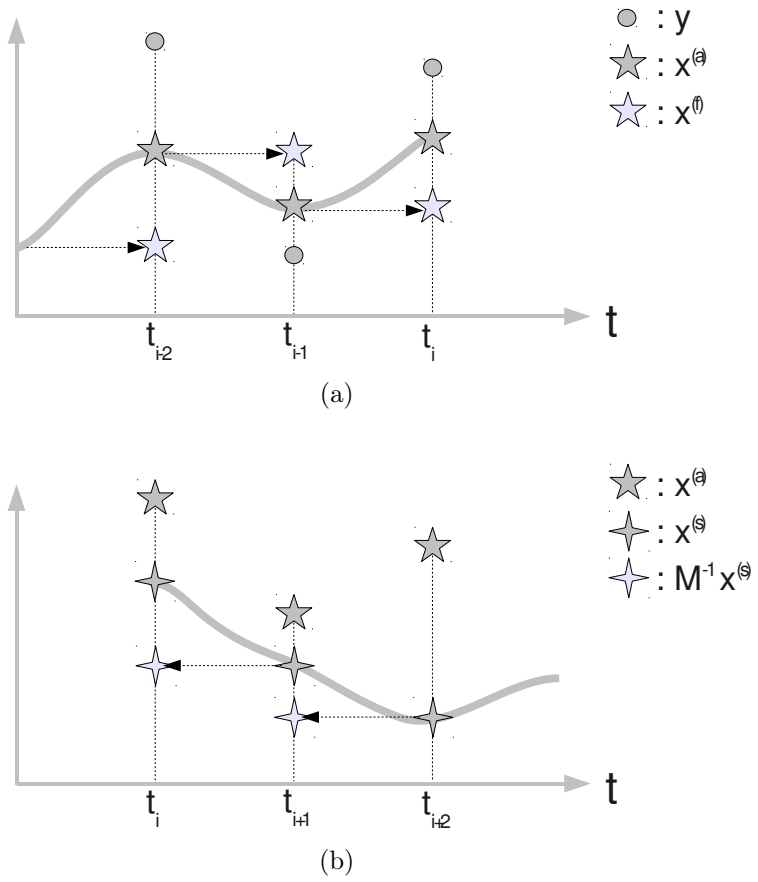


FIG. 2.3: Schématisation du filtre de Kalman (a) et du lisseur de Kalman (b).

Initialisation Au dernier temps t_1 de l'analyse, nous utilisons une information provenant d'une connaissance *a priori* de l'état, de moyenne $x^{(b)}$ et de covariance B. Ce choix est déterminant pour les premières itérations du filtre de Kalman. Il est important de partir de valeurs réalistes afin d'éviter une "période de chauffe" trop importante du filtre.

Etape de prévision Cette étape consiste à estimer la densité de probabilité conditionnelle $P\left(X_{t_i}|y_{t_1}^{t_{i-1}};\theta\right)$ dont l'espérance et la variance sont notées $x_{t_i}^{(f)}$ et $P_{t_i}^{(f)}$. Pour cela, nous utilisons l'équation d'état (2.4). Ainsi,

$$\begin{aligned} x_{t_i}^{(f)} &= E\left(X_{t_i}|y_{t_1}^{t_{i-1}};\theta\right) \\ &= ME\left(X_{t_{i-1}}|y_{t_1}^{t_{i-1}};\theta\right) + E\left(\eta_{t_{i-1}}|y_{t_1}^{t_{i-1}};\theta\right) \\ &= Mx_{t_{i-1}}^{(a)} \end{aligned} \tag{2.14}$$

et

$$\begin{aligned} P_{t_i}^{(f)} &= Var\left(X_{t_i}|y_{t_1}^{t_{i-1}};\theta\right) \\ &= MVar\left(X_{t_{i-1}}|y_{t_1}^{t_{i-1}};\theta\right)M' + Var\left(\eta_{t_{i-1}}|y_{t_1}^{t_{i-1}};\theta\right) \\ &= MP_{t_{i-1}}^{(a)}M' + Q \end{aligned} \tag{2.15}$$

Etape d'analyse Dans cette étape, nous utilisons l'équation d'observation (2.3) : y_{t_i} et sa matrice de covariance R vont venir finaliser l'estimation de la densité de probabilité $P\left(X_{t_i}|y_{t_1}^{t_i};\theta\right)$ d'espérance $x^a\left(t_i\right)$ et de matrice de variance $P^a\left(t_i\right)$. Ceci passe par le calcul de la matrice de gain de Kalman K_{t_i} dont la démonstration est donnée en 2.2.1.1.1 :

$$K_{t_i} = P_{t_i}^{(f)}H'\left(HP_{t_i}^{(f)}H' + R\right)^{-1} \tag{2.16}$$

De plus, nous définissons d_{t_i}, le vecteur d'innovation, défini par

$$d_{t_i} = y_{t_i} - Hx_{t_i}^{(f)} \tag{2.17}$$

Au final, $x_{t_i}^{(a)}$ et $P_{t_i}^{(a)}$ sont donnés par

$$x_{t_i}^{(a)} = x_{t_i}^{(f)} + K_{t_i}d_{t_i} \tag{2.18}$$

$$P_{t_i}^{(a)} = \left(I - K_{t_i}H\right)P_{t_i}^{(f)} \tag{2.19}$$

2.2.1.3.2 Lisseur de Kalman L'objectif est ici d'estimer la densité de probabilité conditionnelle de X_{t_i} connaissant un *a priori* $x^{(b)}$ de l'état, toutes les observations passées et futures $y_{t_1}^{t_n}$ et les paramètres θ du modèle. Ainsi, on va filtrer dans l'ordre chronologique ("forward") et dans le sens contraire ("backward"). On parle alors de lissage, plus connu sous le nom de "ré-analyse" ("hindcast" en Anglais) en océanographie et météorologie. Cette densité de probabilité sera notée $P\left(X_{t_i}|y_{t_1}^{t_n};\theta\right)$ et notons $x_{t_i}^{(s)}$ et $P_{t_i}^{(s)}$ son espérance et sa variance. $x_{t_i}^{(s)}$ sera un estimateur de X_{t_i}, compromis entre $x_{t_i}^{(a)}$ et $x_{t_{i+1}}^{(s)}$, schématisé en figure 2.3b.

Algorithm 2.1 Filtre de Kalman

Initialisation à t_1 :

$$x_{t_1}^{(f)} = x^{(b)} \text{ et } P_{t_1}^{(f)} = B$$

Pour i allant de 1 à n :

– **Etape de prévision**

$$x_{t_i}^{(f)} = M x_{t_{i-1}}^{(a)}$$

$$P_{t_i}^{(f)} = M P_{t_{i-1}}^{(a)} M' + Q$$

– **Etape d'analyse**

$$K_{t_i} = P_{t_i}^{(f)} H' \left(H P_{t_i}^{(f)} H' + R \right)^{-1}$$

$$d_{t_i} = y_{t_i} - H x_{t_i}^{(f)}$$

$$x_{t_i}^{(a)} = x_{t_i}^{(f)} + K_{t_i} d_{t_i}$$

$$P_{t_i}^{(a)} = \left(I - K_{t_i} H \right) P_{t_i}^{(f)}$$

Initialisation Au temps t_n de l'analyse, nous utilisons les résultats du filtre de Kalman $x_{t_n}^{(a)}$ et $P_{t_n}^{(a)}$.

Etape de lissage Soit $K_{t_i}^{(s)}$ la matrice de gain du lisseur de Kalman au temps t_i :

$$K_{t_i}^{(s)} = P_{t_i}^{(a)} M' \left(P_{t_{i+1}}^{(f)} \right)^{-1}$$

$$x_{t_i}^{(s)} = x_{t_i}^{(a)} + K_{t_i}^{(s)} \left(x_{t_{i+1}}^{(s)} - x_{t_{i+1}}^{(f)} \right) \tag{2.20}$$

$$P_{t_i}^{(s)} = P_{t_i}^{(a)} + K_{t_i}^{(s)} \left(P_{t_{i+1}}^{(s)} - P_{t_{i+1}}^{(f)} \right) K_{t_i}^{(s)\prime} \tag{2.21}$$

Notons $P_{t_{i-1},t_i}^{(s)} = Cov\left(X_{t_{i-1}}, X'_{t_i} | y_{t_1}^{t_n}; \hat{\theta}_k \right)$ l'auto-covariance entre deux temps successifs t_{i-1} et t_i. Celle-ci est donnée par :

$$P_{t_{i-1},t_i}^{(s)} = \left(I - K_{t_i} H \right) M P_{t_{i-1}}^{(a)} + \left(P_{t_i}^{(s)} - P_{t_i}^{(a)} \right) \left(P_{t_i}^{(a)} \right)^{-1} \left(I - K_{t_i} H \right) M P_{t_{i-1}}^{(a)} \tag{2.22}$$

avec K_{t_i} le gain du filtre de Kalman calculé en (2.16).

2.2.2 Estimation des paramètres

La résolution du modèle espace d'état séquentiel présenté en 2.1.3 est obtenue par le filtre de Kalman (cf. 2.2.1.3). Or, ce modèle est régi par plusieurs paramètres : les opérateurs modèle M et d'observation H ainsi que les erreurs modèle Q et d'observation R. Tout l'art de l'utilisation de modèles espace d'état réside dans la paramétrisation optimale de ces matrices

Algorithm 2.2 Lisseur de Kalman.

Initialisation à t_n :

$$x_{t_n}^{(s)} = x_{t_n}^{(a)} \text{ et } P_{t_n}^{(s)} = P_{t_n}^{(a)}$$

Pour i allant de $n-1$ à 1 :

– **Etape de lissage**

$$K_{t_i}^{(s)} = P_{t_i}^{(a)} M' \left(P_{t_{i+1}}^{(f)}\right)^{-1}$$

$$x_{t_i}^{(s)} = x_{t_i}^{(a)} + K_{t_i}^{(s)} \left(x_{t_{i+1}}^{(s)} - x_{t_{i+1}}^{(f)}\right)$$

$$P_{t_i}^{(s)} = P_{t_i}^{(a)} + K_{t_i}^{(s)} \left(P_{t_{i+1}}^{(s)} - P_{t_{i+1}}^{(f)}\right) K_{t_i}^{(s)'}$$

$$P_{t_{i-1},t_i}^{(s)} = \left(I - K_{t_i} H\right) M P_{t_{i-1}}^{(a)} + \left(P_{t_i}^{(s)} - P_{t_i}^{(a)}\right) \left(P_{t_i}^{(a)}\right)^{-1} \left(I - K_{t_i} H\right) M P_{t_{i-1}}^{(a)}$$

(dans ce chapitre, nous noterons $\theta = (M, H, Q, R)$ alors que dans les chapitres suivants, ces matrices dépenderons de paramètres). Nous verrons que la forme particulière du modèle linéaire à erreurs Gaussiennes permet l'utilisation de méthodes d'estimations classiques en statistique.

Nous commencerons par visualiser en 2.2.2.1, sur des séries temporelles simulées, l'effet des paramètres du modèle sur les probabilités de lissage obtenues par Kalman. Ensuite, nous étudierons l'estimateur du maximum de vraisemblance (que nous noterons EMV par la suite) de ce modèle particulier. Notamment, en 2.2.2.2, nous discuterons du calcul de l'EMV grâce à l'algorithme Expectation-Maximisation (noté EM).

2.2.2.1 Rôle des paramètres

Pour se rendre compte de l'effet des paramètres θ, on se place dans le cas où $r = p = 1$. Ainsi, M, H, Q et R sont des scalaires. On utilise le schéma suivant :

1. on fixe la valeur des paramètres θ de façon suivante : $M = 0.95$, $H = 1$, $Q = 1$, $R = 1$. La matrice auto-régressive M est choisie inférieure à 1 garantissant la stabilité de l'état (si $|M| < 1$, le processus auto-régessif d'ordre 1, noté AR(1), est stable et possède une unique solution). Notons également que le ratio entre les erreurs R et Q est égal à 1
2. on simule séquentiellement une série temporelle $(x_{t_1}, ..., x_{t_n})$ de taille $n = 200$ (avec un échantillonnage temporel régulier $\Delta_{t_1} = ... = \Delta_{t_n}$) correspondant à notre variable d'état initial, en accord avec (2.4). Nous prenons comme valeurs initiales de l'état $x^{(b)} = 0$ et $B = 1$
3. on simule une série temporelle d'observations $(y_{t_1}, ..., y_{t_n})$ de taille $n = 200$ en utilisant (2.3)
4. on fait varier un des paramètres θ et on estime la série temporelle résultante $\left(x_{t_1}^{(s)}, ..., x_{t_n}^{(s)}\right)$ ainsi que sa variance $\left(P_{t_1}^{(s)}, ..., P_{t_n}^{(s)}\right)$ par lisseur de Kalman. Nous supposons, comme

c'est le cas dans beaucoup d'études, que l'opérateur d'observation H est égal à 1 : aucune transformation entre l'état et l'observation n'est nécessaire. Nous ne faisons donc pas varier ce paramètre. Par contre, nous prenons successivement $M = 0.5$, $Q = 0.1$ et $R = 0.1$

5. on compare cette série temporelle lissée à celle de l'état.

On utilise le schéma présenté ci-dessus pour visualiser l'impact des variations M, Q et R sur l'estimation *a posteriori* de l'état. Les séries temporelles de l'état, des observations et celles obtenues par lissage de Kalman après avoir modifié l'un ou l'autre des paramètres sont représentées en figure 2.4.

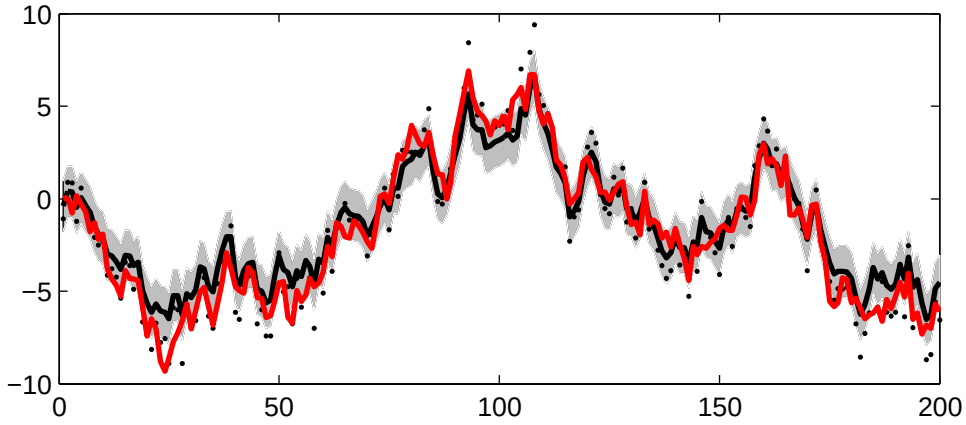
Nous commençons par prendre $M = 0.5$ (cf. figure 2.4a), correspondant à un coefficient autorégressif faible entre deux temps successifs. Les variations de $x^{(s)}$ sont, du fait de la faible valeur de M, autorisées à s'écarter de l'état x. Quand maintenant nous faisons varier les erreurs modèle et observation, on s'aperçoit que le rapport entre Q et R conditionne la vraisemblance entre x et y. En effet, le quotient Q/R rentre en compte dans le calcul du gain de Kalman (2.16) qui est une pondération entre les observations y et l'état propagé $x^{(f)}$ (2.14) par notre équation modèle. Ainsi, lorsque ce rapport est inférieur à 1 (cf. figure 2.4b), les variations de la série temporelle *a posteriori* suivent celles de l'état latent avec néanmoins un lissage important. Inversement, quand le rapport est supérieur à 1 (cf. figure 2.4c), les observations sont privilégiées par rapport au modèle.

La figure 2.4 présente également, grâce aux estimations de $P^{(s)}$, les intervalles de confiance à 95% autour des séries temporelles lissées par Kalman. Graphiquement, on remarque que plus le produit QR est élevé, plus l'incertitude autour de $x^{(s)}$ est forte ; ceci est vérifié par (2.16).

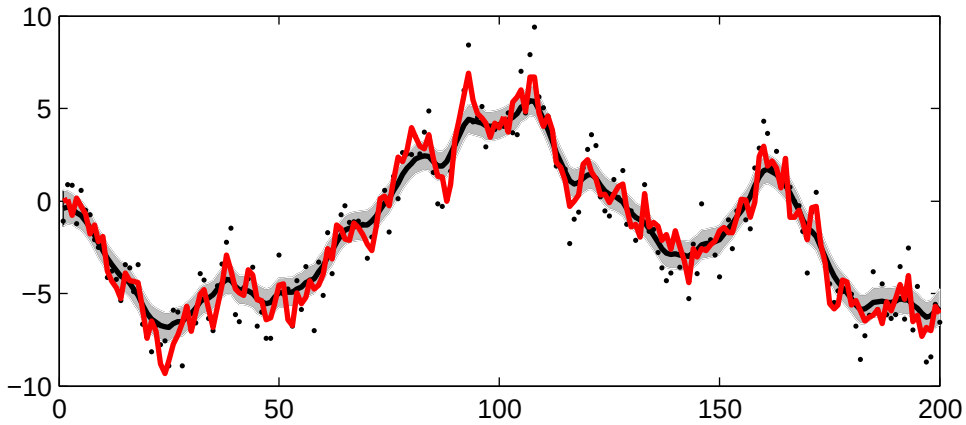
Au final, ces quelques exemples de densité de probabilité conditionnelles où nous faisons varier θ de sa vraie valeur montrent l'importance du choix des paramètres d'un modèle espace d'état. Il semble donc impératif, dans le cas où ceux-ci sont inconnus, d'inférer sur ces paramètres. Nous utiliserons dans la section suivante la vraisemblance du modèle.

2.2.2.2 Algorithme EM

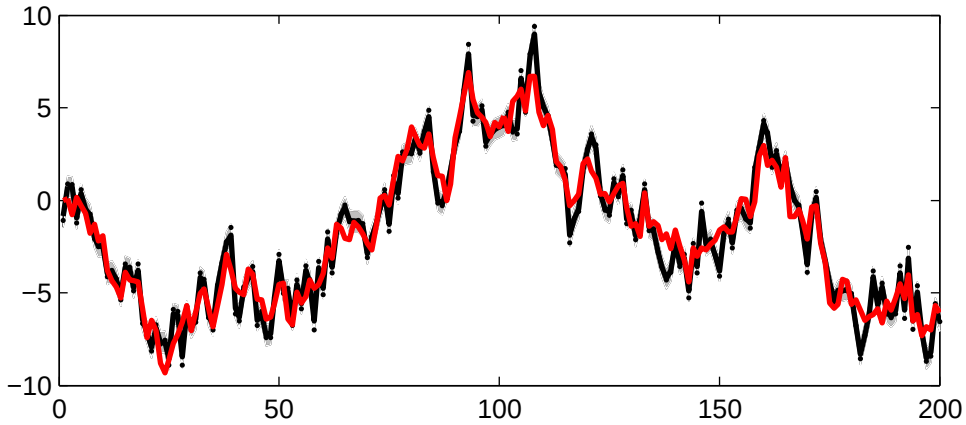
L'algorithme EM (Expectation maximisation) est classique en statistique. Il permet de trouver l'EMV d'un modèle stochastique à variable latente. La forme la plus courante de cet algorithme est dite "par lot" : toutes les observations sont utilisées pour mettre à jour les paramètres de manière récursive. Il fut décrit pour la première fois pour des modèles de chaînes de Markov cachées par [2], sous le nom d'algorithme de Baum Welch puis étendu aux problèmes de variables cachées plus généraux par [10]. Il a été prouvé que la vraisemblance croît (ou n'évolue plus) à chaque itération de l'algorithme EM (cf. [10] et [58]). On peut montrer, sous des conditions générales (cf. [58]), qu'à partir de valeurs initiales $\hat{\theta}_1$ des paramètres du modèle, l'algorithme EM garantit, au fur et à mesure des itérations, la convergence vers le maximum local de la fonction de vraisemblance. Une version simplifiée de EM est donnée dans l'algorithme 2.3.



(a)



(b)



(c)

Figure 2.4: Séries temporelles $(y_{t_1}, ..., y_{t_n})$ (points noirs) et $(x_{t_1}, ..., x_{t_n})$ simulées (trait continu rouge) suivant le modèle espace d'état présenté en 2.1.3 avec $n = 200$ pas de temps réguliers et des paramètres θ fixés. L'estimation *a posteriori* de l'état $(x_{t_1}, ..., x_{t_n})$ en trait continu gras, est obtenue par lissage de Kalman en faisant varier les paramètres θ ($M = 0.5$, $Q = 0.1$ puis $R = 0.1$) respectivement en (a), (b) et (c). Elle est accompagnée de son intervalle de confiance à 95% grisé.

Algorithm 2.3 Algorithme EM "par lot".

Initialisation à $k = 1$:

$\hat{\theta}_1$ est fixé par l'utilisateur.

Tant que la vraisemblance croît :

– **Etape Expectation**

$$U\left(\theta|\hat{\theta}_k\right) = E\left(\log\left(p\left(X_{t_1}^{t_n}, y_{t_1}^{t_n}; \theta\right)\right)|y_{t_1}^{t_n}; \hat{\theta}_k\right)$$

– **Etape Maximisation**

$$\hat{\theta}_{k+1} = \arg\max_{\theta}\left(U\left(\theta|\hat{\theta}_k\right)\right)$$

Nous travaillerons sur l'estimation des paramètres du modèle espace-état linéaire et Gaussien présenté en 2.1.3, où l'échantillonnage temporel est régulier. Nous commencerons par définir la fonction de vraisemblance en 2.2.2.2.1. Ensuite, nous décrirons en 2.2.2.2.2 l'étape Expectation (notée E) de l'algorithme EM où une fonction auxiliaire, liée à cette vraisemblance, est estimée. Lors de l'étape Maximisation (notée M), présentée en 2.2.2.2.3, nous chercherons les paramètres maximisant la fonction auxiliaire de l'étape E. En 2.2.2.2.4, nous aborderons le calcul des espérances conditionnelles nécessaires à l'étape E et en 2.2.2.2.5 nous proposerons une condition d'arrêt de l'algorithme par le biais du calcul de la fonction de vraisemblance incomplète. Enfin, en 2.2.2.2.6, nous évoquerons les limites connues de cet algorithme. Notons que la version associée au cas des données échantillonnées irrégulièrement sera donnée dans le chapitre suivant, en 3.2.2.2.

2.2.2.2.1 Fonctions de vraisemblance D'après les équations (2.4) et (2.3) du modèle espace d'état étudié, on peut écrire les densités conditionnelles multivariées de lois normales $p\left(x_{t_i}|x_{t_{i-1}}; \theta\right)$ et $p\left(y_{t_i}|x_{t_i}; \theta\right)$, $\forall i \in \{1, ..., n\}$:

$$p\left(x_{t_i}|x_{t_{i-1}}; \theta\right) = \exp\left(-\frac{1}{2}\eta'_{t_i}Q^{-1}\eta_{t_i}\right)(2\pi)^{-r/2}\left(\det\left(Q\right)\right)^{-1/2} \tag{2.23}$$

$$p\left(y_{t_i}|x_{t_i}; \theta\right) = \exp\left(-\frac{1}{2}\varepsilon'_{t_i}R^{-1}\varepsilon_{t_i}\right)(2\pi)^{-p/2}\left(\det\left(R\right)\right)^{-1/2} \tag{2.24}$$

avec r la taille du vecteur d'état et p la taille des observations. De plus, nous rappelons que η_{t_i}, ε_{t_i} et $\eta^{(b)}$ s'écrivent :

$$\eta_{t_i} = x_{t_i} - Mx_{t_{i-1}}$$
$$\varepsilon_{t_i} = y_{t_i} - Hx_{t_i}$$
$$\eta^{(b)} = x_{t_1} - x^{(b)}$$

Du fait de la structure Markovienne et en rappelant que $x_{t_1}^{t_n}$ et $y_{t_1}^{t_n}$ sont respectivement les séquences de vecteurs $(x_{t_1}, ..., x_{t_n})$ et $(y_{t_1}, ..., y_{t_n})$, on obtient une fonction de vraisemblance complète :

$$p\left(x_{t_1}^{t_n}, y_{t_1}^{t_n}; \theta\right) = p\left(x_{t_1}\right)\prod_{i=2}^{n} p\left(x_{t_i}|x_{t_{i-1}}; \theta\right)\prod_{i=1}^{n} p\left(y_{t_i}|x_{t_i}; \theta\right) \tag{2.25}$$

avec l'état *a priori* suivant une loi normale définie par :

$$p\left(x_{t_1}\right) = \exp\left(-\frac{1}{2}\eta^{(b)'}B^{-1}\eta^{(b)}\right)(2\pi)^{-r/2}\left(\det\left(B\right)\right)^{-1/2}$$

En pratique, cette vraisemblance complète (2.25) n'est pas calculable car nous ne connaissons pas l'état. Nous verrons que l'étape E de l'algorithme permet d'estimer cette vraisemblance par une fonction auxiliaire (cf. équation (2.28)).

Dans l'algorithme EM, nous allons également introduire la vraisemblance incomplète, calculée à partir de la distribution multivariée de loi normale de moyenne $d_{t_i} = y_{t_i} - Hx_{t_i}^{(f)}$ donnée en (2.17), telle que d_{t_i} est indépendante de d_{t_j} $\forall i \neq j$, et de covariance $\Sigma_{t_i} = HP_{t_i}^{(f)}H' + R$ utilisée en (2.16) dans le filtre de Kalman. D'après [10], en théorie, cette vraisemblance est strictement croissante d'une itération à une autre de l'algorithme EM : on utilisera donc cette vraisemblance pour définir un critère d'arrêt de l'algorithme (cf. 2.2.2.2.5). Par la suite, nous transformerons au log cette vraisemblance incomplète qui s'écrit :

$$\log\left(p\left(y_{t_1}^{t_n};\theta\right)\right) = -\frac{1}{2}\sum_{i=2}^{n} d_{t_i}'\left(\Sigma_{t_i}\right)^{-1}d_{t_i} - \frac{nr}{2}\log\left(2\pi\right) - \frac{1}{2}\sum_{i=2}^{n}\log\left(\det\left(\Sigma_{t_i}\right)\right) \tag{2.26}$$

2.2.2.2.2 Etape Expectation L'étape E consiste à calculer une fonction auxiliaire $U\left(\theta|\hat{\theta}_k\right)$ conditionnellement aux paramètres $\hat{\theta}_k$ estimés à la dernière itération k de l'algorithme et aux observations $y_{t_1}^{t_n}$. Cette fonction auxiliaire est définie par :

$$U\left(\theta|\hat{\theta}_k\right) = E\left(\log\left(p\left(X_{t_1}^{t_n}, y_{t_1}^{t_n};\theta\right)\right)|y_{t_1}^{t_n};\hat{\theta}_k\right) \tag{2.27}$$

Il s'avère que U est généralement une bonne approximation de la vraisemblance incomplète (2.26). Elle est aussi plus facile à calculer et à maximiser. Voici son expression :

$$\begin{aligned} U\left(\theta|\hat{\theta}_k\right) \;=\; & -\frac{1}{2}\sum_{i=1}^{n} E\left(\varepsilon_{t_i}' R^{-1}\varepsilon_{t_i}|y_{t_1}^{t_n};\hat{\theta}_k\right) - \frac{n}{2}\log\left(\det\left(R\right)\right) \\ & -\frac{1}{2}\sum_{i=2}^{n} E\left(\eta_{t_i}' Q^{-1}\eta_{t_i}|y_{t_1}^{t_n};\hat{\theta}_k\right) - \frac{n-1}{2}\log\left(\det\left(Q\right)\right) \\ & -\frac{1}{2}E\left(\eta^{(b)'}B^{-1}\eta^{(b)}|y_{t_1}^{t_n};\hat{\theta}_k\right) - \frac{1}{2}\log\left(\det\left(B\right)\right) - \frac{n(r+p)}{2}\log\left(2\pi\right) \end{aligned} \tag{2.28}$$

où R, Q et B sont, par hypothèse, des matrices symétriques définies positives. En pratique, nous verrons (cf. 2.2.2.2.4) que cette fonction sera calculée à l'aide du lisseur de Kalman.

2.2.2.2.3 Etape Maximisation L'étape M de l'algorithme consiste à maximiser la fonction auxiliaire $U\left(\theta|\hat{\theta}_k\right)$, définie en (2.28). Nous cherchons les paramètres pour lesquels les

dérivées partielles de $U\left(\theta|\hat{\theta}_k\right)$ par rapport à M, H, Q et R s'annulent. Ces estimateurs seront notés $\hat{M}_{k+1}$, $\hat{H}_{k+1}$, $\hat{Q}_{k+1}$, $\hat{R}_{k+1}$ et sont donnés respectivement en (2.29), (2.30), (2.31), (2.32). Notons que plusieurs auteurs ont obtenus ces mêmes résultats sans donner de démonstrations correctement rédigées (cf. [13] et [11]).

Proposition 1. *La différentielle de la fonction définie par* $(\mathbb{R}^{r\times r}\times\mathbb{R}^{p\times r}\times\mathbb{R}^{r\times r}\times\mathbb{R}^{p\times p})\to\mathbb{R}$ $(M,H,Q,R)\mapsto U\left(\theta|\hat{\theta}_k\right)$ *s'annule au point* $\left(\hat{M}_{k+1},\hat{H}_{k+1},\hat{Q}_{k+1},\hat{R}_{k+1}\right)$ *avec :*

$$\hat{M}_{k+1} = \sum_{i=2}^{n} E\left(x_{t_i}x'_{t_{i-1}}|y_{t_1}^{t_n};\hat{\theta}_k\right)\left(\sum_{i=2}^{n} E\left(x_{t_{i-1}}x'_{t_{i-1}}|y_{t_1}^{t_n};\hat{\theta}_k\right)\right)^{-1} \tag{2.29}$$

$$\hat{H}_{k+1} = \sum_{i=1}^{n} E\left(y_{t_i}x'_{t_i}|y_{t_1}^{t_n};\hat{\theta}_k\right)\left(\sum_{i=1}^{n} E\left(x_{t_i}x'_{t_i}|y_{t_1}^{t_n};\hat{\theta}_k\right)\right)^{-1} \tag{2.30}$$

$$\hat{Q}_{k+1} = \frac{1}{n-1}\left(\sum_{i=2}^{n} E\left(x_{t_i}x'_{t_i}|y_{t_1}^{t_n};\hat{\theta}_k\right) - \hat{M}_{k+1}\sum_{i=2}^{n} E\left(x_{t_{i-1}}x'_{t_i}|y_{t_1}^{t_n};\hat{\theta}_k\right)\right) \tag{2.31}$$

$$\hat{R}_{k+1} = \frac{1}{n}\left(\sum_{i=1}^{n} E\left(y_{t_i}y'_{t_i}|y_{t_1}^{t_n};\hat{\theta}_k\right) - \hat{H}_{k+1}\sum_{i=1}^{n} E\left(x_{t_i}y'_{t_i}|y_{t_1}^{t_n};\hat{\theta}_k\right)\right) \tag{2.32}$$

Preuve de la proposition 1. On utilise les lemmes ci-dessous. Après ceux-ci, viendra la suite et fin de cette preuve de proposition. □

Lemma 2. *Si* $A\in\mathbb{R}^{n\times n}$ *une matrice inversible, le développement limité à l'ordre 1 de la fonction* $\log(\det(.))$ *au voisinage de* A *est donné par :*

$$\log(\det(A+dA)) = \log(\det(A)) + Tr\left(A^{-1}dA\right) + o(\|dA\|)$$

lorsque $\|dA\|$ *tend vers* 0*, avec* $(dA)\in\mathbb{R}^{n\times n}$*.*

Preuve du lemme 2. D'après [24] (p. 310), pour toute matrice inversible $A\in\mathbb{R}^{n\times n}$, le développement limité à l'ordre 1 du déterminant au voisinage de A s'écrit :

$$\det(A+dA) - \det(A) = \det(A)\,Tr\left(A^{-1}dA\right) + o(\|dA\|)$$

Soit,

$$\det(A+dA) = \det(A)\left(1+Tr\left(A^{-1}dA\right)\right) + o(\|dA\|)$$

Ensuite,

$$\log(\det(A+dA)) = \log(\det(A)) + \log\left(1+Tr\left(A^{-1}dA\right)\right) + o(\|dA\|)$$

En utilisant le développement limité à l'ordre 1 de $\log(1+x)$ soit $\log(1+x) = x+o(x)$. On obtient donc :

$$\log(\det(A+dA)) = \log(\det(A)) + Tr\left(A^{-1}dA\right) + o(\|dA\|)$$

□

Lemma 3. *Si la relation*

$$Q = \frac{1}{n-1}\sum_{i=2}^{n} E\left(\left(x_{t_i} - Mx_{t_{i-1}}\right)\left(x_{t_i} - Mx_{t_{i-1}}\right)' | y_{t_1}^{t_n}; \hat{\theta}_k\right) \tag{2.33}$$

est satisfaite, alors la différentielle de la fonction définie par $\mathbb{R}^{r\times r} \to \mathbb{R} \quad Q \mapsto U\left(\theta|\hat{\theta}_k\right)$ *est nulle.*

Preuve du lemme 3. Effectuons le changement de variable $\Psi = Q^{-1}$ et étudions les variations de la fonction définie par $\mathbb{R}^{r\times r} \to \mathbb{R} \quad \Psi \mapsto U\left(\theta|\hat{\theta}_k\right)$ qui sera notée abusivement $U\left(\Psi|\hat{\theta}_k\right)$. Commençons par étudier les variations locales de la fonction $U\left(\Psi|\hat{\theta}_k\right)$ en effectuant un développement limité à l'ordre 1 :

$$\begin{aligned} U\left(\Psi + d\Psi|\hat{\theta}_k\right) - U\left(\Psi|\hat{\theta}_k\right) &= -\frac{1}{2}\sum_{i=2}^{n} E\left(\eta'_{t_i}\left(\Psi + d\Psi\right)\eta_{t_i}|y_{t_1}^{t_n};\hat{\theta}_k\right) \\ &\quad + \frac{n-1}{2}\log\left(\det\left(\Psi + d\Psi\right)\right) \\ &\quad + \frac{1}{2}\sum_{i=2}^{n} E\left(\eta'_{t_i}\Psi\eta_{t_i}|y_{t_1}^{t_n};\hat{\theta}_k\right) \\ &\quad - \frac{n-1}{2}\log\left(\det\left(\Psi\right)\right) \end{aligned} \tag{2.34}$$

Ainsi, d'après le lemme 2, l'équation (2.34) s'écrit :

$$\begin{aligned} U\left(\Psi + d\Psi|\hat{\theta}_k\right) - U\left(\Psi|\hat{\theta}_k\right) &= -\frac{1}{2}\sum_{i=2}^{n} E\left(\eta'_{t_i} d\Psi \eta_{t_i}|y_{t_1}^{t_n};\hat{\theta}_k\right) \\ &\quad + \frac{n-1}{2}\left(\log\left(\det\left(\Psi\right)\right) + Tr\left(\Psi^{-1}d\Psi\right) + o\left(\|d\Psi\|\right)\right) \\ &\quad - \frac{n-1}{2}\log\left(\det\left(\Psi\right)\right) \end{aligned} \tag{2.35}$$

On sait que pour tout vecteur $a \in \mathbb{R}^n$ et toute matrice carrée $V \in \mathbb{R}^{n\times n}$, $a'Va = Tr\left(Vaa'\right)$. L'équation (2.35) devient :

$$\begin{aligned} U\left(\Psi + d\Psi|\hat{\theta}_k\right) - U\left(\Psi|\hat{\theta}_k\right) &= -\frac{1}{2}Tr\left(d\Psi\sum_{i=2}^{n} E\left(\eta_{t_i}\eta'_{t_i}|y_{t_1}^{t_n};\hat{\theta}_k\right)\right) \\ &\quad + \frac{n-1}{2}\left(Tr\left(\Psi^{-1}d\Psi\right) + o\left(\|d\Psi\|\right)\right) \end{aligned} \tag{2.36}$$

Or, pour toutes matrices carrées $A \in \mathbb{R}^{n\times n}$ et $B \in \mathbb{R}^{n\times n}$, $Tr\left(AB\right) = Tr\left(BA\right)$ et $Tr\left(A+B\right) = Tr\left(A\right) + Tr\left(B\right)$. D'où, (2.36) s'écrit :

$$U\left(\Psi + d\Psi|\hat{\theta}_k\right) - U\left(\Psi|\hat{\theta}_k\right) = -\frac{1}{2}Tr\left(d\Psi\left(\sum_{i=2}^{n} E\left(\eta_{t_i}\eta'_{t_i}|y_{t_1}^{t_n};\hat{\theta}_k\right) - \left(n-1\right)\Psi^{-1}\right) + o\left(\|d\Psi\|\right)\right)$$

Ainsi, en rappelant que $\Psi = Q^{-1}$, le premier terme du développement limité de $U\left(\Psi|\hat{\theta}_k\right)$ est nul pour :

$$\begin{aligned} Q &= \frac{1}{n-1}\sum_{i=2}^{n} E\left(\eta_{t_i}\eta_{t_i}'|y_{t_1}^{t_n};\hat{\theta}_k\right) \\ &= \frac{1}{n-1}\sum_{i=2}^{n} E\left(\left(x_{t_i}-Mx_{t_{i-1}}\right)\left(x_{t_i}-Mx_{t_{i-1}}\right)'|y_{t_1}^{t_n};\hat{\theta}_k\right) \end{aligned}$$

□

Lemma 4. *Si la relation*

$$R = \frac{1}{n}\sum_{i=1}^{n} E\left(\left(y_{t_i}-Hx_{t_i}\right)\left(y_{t_i}-Hx_{t_i}\right)'|y_{t_1}^{t_n};\hat{\theta}_k\right) \tag{2.37}$$

est satisfaite, alors la différentielle de la fonction définie par $\mathbb{R}^{p\times p}\to\mathbb{R} \quad R\mapsto U\left(\theta|\hat{\theta}_k\right)$ *est nulle.*

Preuve du lemme 4. La démonstration est similaire à celle du lemme 3. □

Lemma 5. *Si les relations (2.33) et (2.37) sont satisfaites, alors* :

$$\begin{aligned} U\left(\theta|\hat{\theta}_k\right) &= -\frac{n-1}{2}\left(U_X\left(M|\hat{\theta}_k\right)\right) \\ &\quad -\frac{n}{2}\left(U_{Y|X}\left(H|\hat{\theta}_k\right)\right) \\ &\quad -\frac{1}{2}E\left(\eta^{(b)'}B^{-1}\eta^{(b)}|y_{t_1}^{t_n};\hat{\theta}_k\right)-\frac{1}{2}\log\left(\det\left(B\right)\right) \\ &\quad -\frac{np}{2}-\frac{(n-1)\,r}{2}-\frac{n(r+p)}{2}\log\left(2\pi\right) \end{aligned}$$

avec

$$U_X\left(M|\hat{\theta}_k\right)=\log\left(\det\left(\frac{1}{n-1}\sum_{i=2}^{n}E\left(\left(x_{t_i}-Mx_{t_{i-1}}\right)\left(x_{t_i}-Mx_{t_{i-1}}\right)'|y_{t_1}^{t_n};\hat{\theta}_k\right)\right)\right)$$

$$U_{Y|X}\left(H|\hat{\theta}_k\right)=\log\left(\det\left(\frac{1}{n}\sum_{i=1}^{n}E\left(\left(y_{t_i}-Hx_{t_i}\right)\left(y_{t_i}-Hx_{t_i}\right)'|y_{t_1}^{t_n};\hat{\theta}_k\right)\right)\right)$$

Preuve du lemme 5. Supposons que la relation (2.33) soit satisfaite. En rappelant que pour

tout vecteur $a \in \mathbb{R}^n$ et toute matrice carrée $V \in \mathbb{R}^{n\times n}$, $a'Va = Tr\,(Vaa')$,

$$\begin{aligned}
\sum_{i=2}^{n} E\left(\eta'_{t_i} Q^{-1} \eta_{t_i} | y_{t_1}^{t_n}; \hat{\theta}_k\right) &= \sum_{i=2}^{n} E\left(\eta'_{t_i}\left(\frac{1}{n-1}\sum_{i=2}^{n} E\left(\eta_{t_i}\eta'_{t_i}|y_{t_1}^{t_n};\hat{\theta}_k\right)\right)^{-1}\eta_{t_i}|y_{t_1}^{t_n};\hat{\theta}_k\right) \\
&= Tr\left(\left(\frac{1}{n-1}\sum_{i=2}^{n} E\left(\eta_{t_i}\eta'_{t_i}|y_{t_1}^{t_n};\hat{\theta}_k\right)\right)^{-1}\sum_{i=2}^{n} E\left(\eta_{t_i}\eta'_{t_i}|y_{t_1}^{t_n};\hat{\theta}_k\right)\right) \\
&= (n-1)\,Tr\,(I_r) \\
&= (n-1)\,r \qquad (2.38)
\end{aligned}$$

avec I_r la matrice identité de taille r (taille du vecteur d'état). De la même manière, si (2.37) est satisfaite alors :

$$\begin{aligned}
\sum_{i=1}^{n} E\left(\varepsilon'_{t_i} R^{-1} \varepsilon_{t_i} | y_{t_1}^{t_n}; \hat{\theta}_k\right) &= \sum_{i=1}^{n} E\left(\varepsilon'_{t_i}\left(\frac{1}{n}\sum_{i=1}^{n} E\left(\varepsilon_{t_i}\varepsilon'_{t_i}|y_{t_1}^{t_n};\hat{\theta}_k\right)\right)^{-1}\varepsilon_{t_i}|y_{t_1}^{t_n};\hat{\theta}_k\right) \\
&= Tr\left(\left(\frac{1}{n}\sum_{i=1}^{n} E\left(\varepsilon_{t_i}\varepsilon'_{t_i}|y_{t_1}^{t_n};\hat{\theta}_k\right)\right)^{-1}\sum_{i=1}^{n} E\left(\varepsilon_{t_i}\varepsilon'_{t_i}|y_{t_1}^{t_n};\hat{\theta}_k\right)\right) \\
&= nTr\,(I_p) \\
&= np \qquad (2.39)
\end{aligned}$$

Ainsi, d'après (2.28), (2.38) et (2.39), $U\left(\theta|\hat{\theta}_k\right)$ s'écrit :

$$\begin{aligned}
U\left(\theta|\hat{\theta}_k\right) &= -\frac{n-1}{2}\log\left(\det\left(\frac{1}{n-1}\sum_{i=2}^{n} E\left(\eta_{t_i}\eta'_{t_i}|y_{t_1}^{t_n};\hat{\theta}_k\right)\right)\right) \\
&\quad -\frac{n}{2}\log\left(\det\left(\frac{1}{n}\sum_{i=1}^{n} E\left(\varepsilon_{t_i}\varepsilon'_{t_i}|y_{t_1}^{t_n};\hat{\theta}_k\right)\right)\right) \\
&\quad -\frac{1}{2}E\left(\eta^{(b)'}B^{-1}\eta^{(b)}|y_{t_1}^{t_n};\hat{\theta}_k\right) - \frac{1}{2}\log\left(\det\left(B\right)\right) \\
&\quad -\frac{np}{2} - \frac{(n-1)\,r}{2} - \frac{n(r+p)}{2}\log\left(2\pi\right)
\end{aligned}$$

Au final, on obtient bien le résultat énoncé dans ce lemme 5. □

Lemma 6. *La différentielle de la fonction définie par* $\mathbb{R}^{r\times r} \to \mathbb{R} \quad M \mapsto U_X\left(M|\hat{\theta}_k\right)$ *s'annule pour* $M = \hat{M}_{k+1}$*, avec* $\hat{M}_{k+1}$ *défini en (2.29).*

Preuve du lemme 6. Nous utilisons un développement limité à l'ordre 1 de $U_X\left(M|\hat{\theta}_k\right)$. Com-

mençons par développer l'expression $U_X\left(M+dM|\hat{\theta}_k\right)$:

$$\begin{aligned}
U_X\left(M+dM|\hat{\theta}_k\right) &= \log\left(\det\left(\frac{1}{n-1}\sum_{i=2}^{n}E\left(\left(x_{t_i}-Mx_{t_{i-1}}\right)\left(x_{t_i}-Mx_{t_{i-1}}\right)'|y_{t_1}^{t_n};\hat{\theta}_k\right)\right.\right.\\
&\quad -\frac{1}{n-1}\sum_{i=2}^{n}E\left(dMx_{t_{i-1}}\left(x_{t_i}-Mx_{t_{i-1}}\right)'|y_{t_1}^{t_n};\hat{\theta}_k\right)\\
&\quad -\frac{1}{n-1}\sum_{i=2}^{n}E\left(\left(x_{t_i}-Mx_{t_{i-1}}\right)\left(dMx_{t_{i-1}}\right)'|y_{t_1}^{t_n};\hat{\theta}_k\right)\\
&\quad \left.\left.+\frac{1}{n-1}\sum_{i=2}^{n}E\left(\left(dMx_{t_{i-1}}\right)\left(dMx_{t_{i-1}}\right)'|y_{t_1}^{t_n};\hat{\theta}_k\right)\right)\right)\\
&= \log\left(\det\left(\frac{1}{n-1}\sum_{i=2}^{n}E\left(\left(x_{t_i}-Mx_{t_{i-1}}\right)\left(x_{t_i}-Mx_{t_{i-1}}\right)'|y_{t_1}^{t_n};\hat{\theta}_k\right)\right.\right.\\
&\quad -\frac{1}{n-1}\sum_{i=2}^{n}E\left(dMx_{t_{i-1}}\left(x_{t_i}-Mx_{t_{i-1}}\right)'|y_{t_1}^{t_n};\hat{\theta}_k\right)\\
&\quad -\frac{1}{n-1}\sum_{i=2}^{n}E\left(\left(x_{t_i}-Mx_{t_{i-1}}\right)\left(dMx_{t_{i-1}}\right)'|y_{t_1}^{t_n};\hat{\theta}_k\right)\\
&\quad \left.\left.+o\left(\|dM\|\right)\right)\right) \qquad (2.40)
\end{aligned}$$

Pour simplifier les calculs, nous choisissons d'utiliser les notations suivantes :

$$\Xi=\frac{1}{n-1}\sum_{i=2}^{n}E\left(\eta_{t_i}\eta_{t_i}'|y_{t_1}^{t_n};\hat{\theta}_k\right) \qquad (2.41)$$

$$\Upsilon=-\frac{1}{n-1}\sum_{i=2}^{n}E\left(x_{t_{i-1}}\left(x_{t_i}-Mx_{t_{i-1}}\right)'|y_{t_1}^{t_n};\hat{\theta}_k\right) \qquad (2.42)$$

Ainsi, (2.40) s'écrit :

$$U_X\left(M+dM|\hat{\theta}_k\right)=\log\left(\det\left(\Xi+dM\Upsilon+\Upsilon'dM'+o\left(\|dM\|\right)\right)\right)$$

D'après le lemme 2, on obtient alors :

$$\begin{aligned}
U_X\left(M+dM|\hat{\theta}_k\right)-U_X\left(M|\hat{\theta}_k\right) &= Tr\left(\Xi^{-1}\left(dM\Upsilon+\Upsilon'dM'\right)\right)+o\left(\|dM\|\right) \qquad (2.43)\\
&= Tr\left(\Xi^{-1}dM\Upsilon\right)+Tr\left(\Xi^{-1}\Upsilon'dM'\right)+o\left(\|dM\|\right)
\end{aligned}$$

Intéressons-nous au terme $Tr\left(\Xi^{-1}\Upsilon'dM'\right)$ de (2.43). Nous savons que pour toute matrice carrée $A\in\mathbb{R}^{n\times n}$, $Tr\left(A\right)=Tr\left(A'\right)$. Ainsi,

$$\begin{aligned}
Tr\left(\Xi^{-1}\Upsilon'dM'\right) &= Tr\left(dM\Upsilon\left(\Xi^{-1}\right)'\right)\\
&= Tr\left(dM\Upsilon\left(\Xi'\right)^{-1}\right)
\end{aligned}$$

De plus, par construction, Ξ est symétrique d'où $\Xi = \Xi'$, et

$$Tr\left(\Xi^{-1}\Upsilon' dM'\right) = Tr\left(dM\Upsilon\Xi^{-1}\right)$$

Enfin, pour toutes matrices carrées $A \in \mathbb{R}^{n\times n}$ et $B \in \mathbb{R}^{n\times n}$, $Tr(AB) = Tr(BA)$, soit

$$Tr\left(\Xi^{-1}\Upsilon' dM'\right) = Tr\left(\Xi^{-1} dM\Upsilon\right)$$

On obtient donc :

$$U_X\left(M + dM|\hat{\theta}_k\right) - U_X\left(M|\hat{\theta}_k\right) = 2Tr\left(\Xi^{-1} dM\Upsilon\right) + +o\left(\|dM\|\right)$$

Finalement, en remplaçant Ξ et Υ par leurs valeurs définies en (2.41) et (2.42), l'équation (2.43) devient :

$$\begin{aligned} U_X\left(M + dM|\hat{\theta}_k\right) - U_X\left(M|\hat{\theta}_k\right) &= -2Tr\left(\left(\frac{1}{n-1}\sum_{i=2}^{n} E\left(\eta_{t_i}\eta'_{t_i}|y_{t_1}^{t_n};\hat{\theta}_k\right)\right)^{-1}\right. \\ &\quad \left.\times dM\left(\frac{1}{n-1}\sum_{i=2}^{n} E\left(x_{t_{i-1}}\left(x_{t_i} - Mx_{t_{i-1}}\right)'|y_{t_1}^{t_n};\hat{\theta}_k\right)\right)\right) \\ &\quad +o\left(\|dM\|\right) \end{aligned} \tag{2.44}$$

On vérifie alors que si :

$$M = \sum_{i=2}^{n} E\left(x_{t_i}x'_{t_{i-1}}|y_{t_1}^{t_n};\hat{\theta}_k\right)\left(\sum_{i=2}^{n} E\left(x_{t_{i-1}}x'_{t_{i-1}}|y_{t_1}^{t_n};\hat{\theta}_k\right)\right)^{-1}$$

alors le terme d'ordre 1 du développement limité de $U_X\left(M|\hat{\theta}_k\right)$ est nul. □

Lemma 7. *La différentielle de la fonction définie par* $\mathbb{R}^{p\times r} \to \mathbb{R} \quad H \mapsto U_{Y|X}\left(H|\hat{\theta}_k\right)$ *s'annule pour* $H = \hat{H}_{k+1}$*, avec* $\hat{H}_{k+1}$ *défini en (2.30).*

Preuve du lemme 7. La démonstration est similaire à celle du lemme 6. □

Preuve de la proposition 1 (suite et fin). Une fois obtenue $\hat{M}_{k+1}$ dans le lemme 6, nous complétons l'estimation de Q partiellement établie dans le lemme 3. $\hat{Q}_{k+1}$ s'écrit alors :

$$\begin{aligned}
\hat{Q}_{k+1} &= \frac{1}{n-1}\sum_{i=2}^{n} E\left(\left(x_{t_i}-\hat{M}_{k+1}x_{t_{i-1}}\right)\left(x_{t_i}-\hat{M}_{k+1}x_{t_{i-1}}\right)' |y_{t_1}^{t_n};\hat{\theta}_k\right) \\
&= \frac{1}{n-1}\sum_{i=2}^{n} E\left(x_{t_i}x'_{t_i}|y_{t_1}^{t_n};\hat{\theta}_k\right) \\
&\quad -\frac{1}{n-1}\sum_{i=2}^{n} E\left(x_{t_i}x'_{t_{i-1}}|y_{t_1}^{t_n};\hat{\theta}_k\right)\hat{M}'_{k+1} \\
&\quad -\frac{1}{n-1}\sum_{i=2}^{n} \hat{M}_{k+1}E\left(x_{t_{i-1}}x'_{t_i}|y_{t_1}^{t_n};\hat{\theta}_k\right) \\
&\quad +\frac{1}{n-1}\sum_{i=2}^{n} \hat{M}_{k+1}E\left(x_{t_{i-1}}x'_{t_{i-1}}|y_{t_1}^{t_n};\hat{\theta}_k\right)\hat{M}'_{k+1} \qquad (2.45)
\end{aligned}$$

Or, en remplaçant $\hat{M}_{k+1}$ par son expression (2.29), le dernier terme de (2.45) devient :

$$\begin{aligned}
\sum_{i=2}^{n} \hat{M}_{k+1}E\left(x_{t_{i-1}}x'_{t_{i-1}}|y_{t_1}^{t_n};\hat{\theta}_k\right)\hat{M}'_{k+1} &= \hat{M}_{k+1}\sum_{i=2}^{n} E\left(x_{t_{i-1}}x'_{t_{i-1}}|y_{t_1}^{t_n};\hat{\theta}_k\right)\hat{M}'_{k+1} \\
&= \sum_{i=2}^{n} E\left(x_{t_i}x'_{t_{i-1}}|y_{t_1}^{t_n};\hat{\theta}_k\right)\hat{M}'_{k+1}
\end{aligned}$$

Au final, l'équation (2.45) se simplifie et on obtient :

$$\hat{Q}_{k+1} = \frac{1}{n-1}\left(\sum_{i=2}^{n} E\left(x_{t_i}x'_{t_i}|y_{t_1}^{t_n};\hat{\theta}_k\right) - \hat{M}_{k+1}\sum_{i=2}^{n} E\left(x_{t_{i-1}}x'_{t_i}|y_{t_1}^{t_n};\hat{\theta}_k\right)\right)$$

Pour estimer $\hat{R}_{k+1}$, le raisonnement est le même que pour $\hat{Q}_{k+1}$: on remplace H par $\hat{H}_{k+1}$ (2.30) dans l'estimation de R partiellement établie dans le lemme 4. On obtient ainsi *:*

$$\hat{R}_{k+1} = \frac{1}{n}\left(\sum_{i=1}^{n} E\left(y_{t_i}y'_{t_i}|y_{t_1}^{t_n};\hat{\theta}_k\right) - \hat{H}_{k+1}\sum_{i=1}^{n} E\left(x_{t_i}y'_{t_i}|y_{t_1}^{t_n};\hat{\theta}_k\right)\right)$$

□

2.2.2.2.4 Espérances conditionnelles Afin de pouvoir calculer $\hat{\theta}_{k+1}$ dans l'étape M, nous avons besoin de déterminer, dans l'étape E, les espérances conditionnelles notées $x_{t_i}^{(s)} = E\left(X_{t_i}|y_{t_1}^{t_n};\hat{\theta}_k\right)$, $P_{t_i}^{(s)} = Var\left(X_{t_i}|y_{t_1}^{t_n};\hat{\theta}_k\right)$ et $P_{t_{i-1},t_i}^{(s)} = Cov\left(X_{t_{i-1}}, X'_{t_i}|y_{t_1}^{t_n};\hat{\theta}_k\right)$ données par le lisseur de Kalman en (2.20), (2.21) et (2.22).

On rappelle que $Var\left(X_{t_i}|y_{t_1}^{t_n};\hat{\theta}_k\right)$ vaut

$$\begin{aligned} P_{t_i}^{(s)} &= E\left(x_{t_i}x'_{t_i}|y_{t_1}^{t_n};\hat{\theta}_k\right) - E\left(x_{t_i}|y_{t_1}^{t_n};\hat{\theta}_k\right)E\left(x'_{t_i}|y_{t_1}^{t_n};\hat{\theta}_k\right) \\ &= E\left(x_{t_i}x'_{t_i}|y_{t_1}^{t_n};\hat{\theta}_k\right) - x_{t_i}^{(s)}x_{t_i}^{(s)'} \end{aligned}$$

Soit,

$$E\left(x_{t_i}x'_{t_i}|y_{t_1}^{t_n};\hat{\theta}_k\right) = x_{t_i}^{(s)}x_{t_i}^{(s)'} + P_{t_i}^{(s)} \tag{2.46}$$

De même, $Cov\left(X_{t_{i-1}}, X'_{t_i}|y_{t_1}^{t_n};\hat{\theta}_k\right)$ vaut

$$\begin{aligned} P_{t_{i-1},t_i}^{(s)} &= E\left(x_{t_{i-1}}x'_{t_i}|y_{t_1}^{t_n};\hat{\theta}_k\right) - E\left(x_{t_{i-1}}|y_{t_1}^{t_n};\hat{\theta}_k\right)E\left(x'_{t_i}|y_{t_1}^{t_n};\hat{\theta}_k\right) \\ &= E\left(x_{t_{i-1}}x'_{t_i}|y_{t_1}^{t_n};\hat{\theta}_k\right) - x_{t_{i-1}}^{(s)}x_{t_i}^{(s)'} \end{aligned}$$

Soit,

$$E\left(x_{t_{i-1}}x'_{t_i}|y_{t_1}^{t_n};\hat{\theta}_k\right) = x_{t_{i-1}}^{(s)}x_{t_i}^{(s)'} + P_{t_{i-1},t_i}^{(s)} \tag{2.47}$$

De plus,

$$E\left(y_{t_i}y'_{t_i}|y_{t_1}^{t_n};\hat{\theta}_k\right) = y_{t_i}y'_{t_i} \tag{2.48}$$

et

$$E\left(x_{t_i}y'_{t_i}|y_{t_1}^{t_n};\hat{\theta}_k\right) = x_{t_i}^{(s)}y'_{t_i} \tag{2.49}$$

2.2.2.2.5 Convergence Les paramètres $\hat{\theta}_{k+1}$ fraîchement estimés dans l'étape M servent ensuite de point de départ d'une nouvelle phase d'évaluation de l'espérance dans l'étape E. On va utiliser la log-vraisemblance incomplète (2.26), mesurant l'adéquation entre les observations et le modèle ([33]) en calculant son accroissement entre deux itérations. On itère l'algorithme EM jusqu'à ce que l'accroissement soit négligeable. Voici une condition d'arrêt possible (cf. [45]) :

$$\frac{2\left|\log\left(p\left(y_{t_1}^{t_n};\hat{\theta}_{k+1}\right)\right) - \log\left(p\left(y_{t_1}^{t_n};\hat{\theta}_k\right)\right)\right|}{\left|\log\left(p\left(y_{t_1}^{t_n};\hat{\theta}_{k+1}\right)\right)\right| - \left|\log\left(p\left(y_{t_1}^{t_n};\hat{\theta}_k\right)\right)\right|} < \xi$$

où ξ est fixé à $\xi = 10^{-4}$ dans notre cas de figure.

2.2.2.2.6 Limites Les algorithmes EM de type "par lot", comme celui présenté dans cette section, ont des inconvénients bien connus qui sont listés ci-dessous. Premièrement, la vitesse de convergence est lente au voisinage des extrema locaux. Il est souvent nécessaire d'utiliser une méthode d'optimisation numérique de type quasi-Newton pour finaliser rapidement cette convergence. Deuxièmement, du fait de la multimodalité de la fonction de vraisemblance (comme c'est le cas ici), cet algorithme peut converger vers un extremum local et

non global, surtout lors de mauvais choix de valeurs initiales $\hat{\theta}_1$ des paramètres. Plusieurs méthodes permettent d'éviter ce problème, soit en passant par une étape d'expertise (basée sur des observations réelles) afin d'avoir un ensemble de paramètres initiaux plausibles, soit en considérant une collection de paramètres initiaux qui peuvent être fixés par l'utilisateur ou aléatoirement (suivant les paramètres initiaux, il est possible que les différentes itérations de l'algorithme EM aient convergé vers des minima locaux et un minimal global). Troisièmement, lorsque la taille de l'état et des observations r et p sont grands, la maximisation sur tout le lot d'observations devient lente et couteuse en ressource informatique. Enfin, les paramètres estimés sur un lot de données sont invariants. Or, dans des problèmes de dynamiques environnementales comme le nôtre, il arrive que les paramètres évoluent au cours du temps et de l'espace. Ainsi, le côté statique des paramètres devient gênant. Une façon de prendre en compte cette considération est d'estimer les paramètres au fur et à mesure des nouvelles observations : on parlera d'estimation "en ligne". Notons que dans cette thèse, nous n'évoquerons pas cette méthode. Pour plus d'informations à ce sujet, consultez [3].

2.3 Discussion

Les modèles espace d'état permettent de prendre en compte plusieurs sources d'information qui sont pondérées par un terme d'erreur. L'objectif est d'estimer, de façon optimale au sens statistique, une variable d'état cachée. Dans le cas linéaire où les erreurs sont Gaussiennes, cette estimation *a posteriori* de l'état est possible grâce à deux approches rendant des résultats identiques. Dans cette thèse, nous nous focalisons uniquement sur l'approche dite stochastique qui estime une densité de probabilité de loi normale *a posteriori* de l'état à partir des deux premiers moments de la distribution. Nous avons notamment exposé le point de vue des probabilités conditionnelles ainsi que celui des moindre carrés qui concluent aux mêmes résultats. Ainsi, pour conclure sur l'estimation séquentielle de l'état au cours du temps, nous avons présenté les récursions du filtre et du lisseur de Kalman.

Une fois cette étape réalisée, nous nous sommes intéressés à l'estimation des paramètres régissant ce modèle espace d'état linéaire à erreurs Gaussiennes. En utilisant les estimations lissées par Kalman de l'espérance et la variance des états successifs, nous avons utilisé une méthode classique, l'algorithme EM, basée sur la maximisation de la vraisemblance du modèle.

Le modèle espace d'état proposé dans ce chapitre va être le point central de cette thèse. La résolution par Kalman ainsi que l'estimation des paramètres par maximum de vraisemblance seront utilisés tout au long de l'étude. Dans les prochains chapitres, nous allons appliquer les méthodes présentées ici pour répondre au problème de l'interpolation de données SST présentées en 1. Nous travaillerons sur des données de nature différente : séries temporelles (cf. chapitre 3) puis champs spatiaux/temporels (cf. chapitre 4). La forme même des observations conduira à la généralisation du modèle (au temps irrégulier) et des méthodes de résolution et d'estimation étudiés dans ce chapitre.

Notons finalement que les cas non-linéaires ou non-Gaussiens ne seront pas étudiés dans cette thèse. Ceux-ci nécessitent l'utilisation de filtres particulaires et d'algorithmes EM spé-

cifiques, augmentant les temps de calculs. Ceci devient d'autant plus problématique lorsque, comme c'est le cas avec des champs de SST observés à des résolutions élevées, la dimension de l'état est grande. Tout ceci justifie le fait qu'on s'intéresse uniquement au cas linéaire et Gaussien dans cette thèse.

CHAPITRE 3

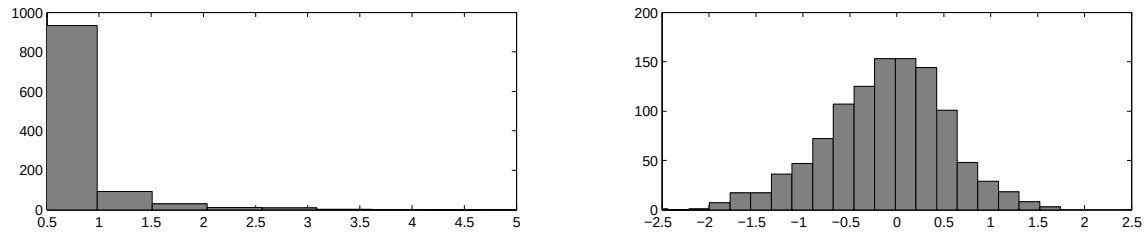
Modélisation temporelle de la SST

La première étape de notre étude, avant toute étude globale, consiste à étudier la SST de façon locale. Plus particulièrement, dans ce chapitre, nous nous intéressons à la variabilité temporelle de la SST, discussion que nous avions introduite en 1.1.1. L'étude se fera sur des séries temporelles d'observations satellitaires présentées en 1.1.2. L'objectif de ce chapitre est d'identifier les paramètres modélisant la dynamique de notre signal et d'interpoler au mieux les observations satellitaires. La modélisation espace d'état étudiée dans le chapitre 2, dans le cas de séries temporelles, répond à ces attentes (cf. [15]).

Dans ce chapitre, nous commencerons en 3.1 par introduire les données satellitaires utilisées. D'après 1.1.2, celles-ci ont un échantillonnage temporel irrégulier. Le modèle espace d'état que nous utiliserons en 3.2.1 ainsi que l'estimation des paramètres en 3.2.2 prendront en compte cette particularité des observations. Les résultats seront exposés en 3.3 : après avoir validé la qualité de l'estimation des paramètres du modèle sur des simulations en 3.3.1, nous étudierons en détail les résultats obtenus sur des observations réelles en un point particulier de l'océan en 3.3.2 puis nous généraliserons les résultats à l'Atlantique en 3.3.3. Enfin, en 3.4, une discussion sur les conclusions de cette étude et les perspectives à venir viendra clore ce chapitre.

3.1 Données

Dans ce chapitre, nous utilisons les données METOP décrites en 1.1.2, fournissant au plus deux données de SST par jour à une résolution spatiale de 0.05°. Nous nous focalisons sur la région d'étude de l'Atlantique tropical où l'occurrence des observations est importante. Plus particulièrement, nous nous intéressons au point 0°N 23°W où des données *in situ* sont également enregistrées par une bouée fixe du réseau PIRATA, présentées en 1.1.1. La période d'étude s'échelonne sur deux ans, du 11 juillet 2007 au 18 juin 2009, ce qui est représentatif d'une grande partie de la variabilité de la SST à cet endroit. Notons $(t_1, ..., t_n)$ les temps pour lesquels les données METOP sont accessibles : ceci correspond à $n = 1087$ observations. Ces



(a) (b)

Figure 3.1: Histogrammes des intervalles de temps Δ_{t_i} en jours (a) et des anomalies de SST METOP en °C (b).

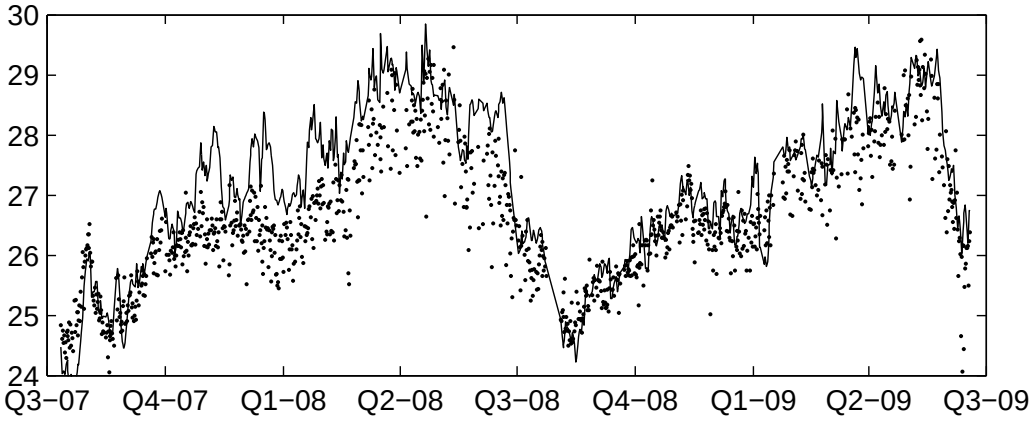
données satellitaires peuvent être contaminées par des conditions atmosphériques comme la vapeur d'eau, les aérosols ou la présence de nuages. Ainsi, certaines données sont considérées comme manquantes. L'échantillonnage temporel n'est donc pas constant et nous notons $\Delta_{t_i} = t_i - t_{i-1}$ l'intervalle de temps entre deux observations successives. Celui-ci peut varier d'une demi-journée à plusieurs jours comme le montre l'histogramme de la figure 3.1a.

La série temporelle METOP est clairement non stationnaire (cf. figure 3.2a) à cause notamment d'une importante composante saisonnière. De plus, nous observons des chutes brutales de températures entre la fin du $2^{ème}$ semestre et le $3^{ème}$ semestre des données deux années étudiées, correspondant à des vents forts, mélangeant la couche de surface. Notons également des chutes spontanées de la SST à chaque 1^{er} semestre des deux années étudiées : ceci correspond au passage de la langue d'eau froide de l'Atlantique équatorial (cf. discussion en 3.3.3.1). Afin de supprimer ces composantes complexes ne pouvant pas être modélisées par un modèle paramétrique simple (de type sinusoïdal par exemple), nous décidons d'utiliser les données de l'analyse SST OIV2 présentées en 1.1.3. Celles-ci sont indépendantes des données METOP (pour la période considérée) et sont une bonne estimation des variations moyennes de la SST.

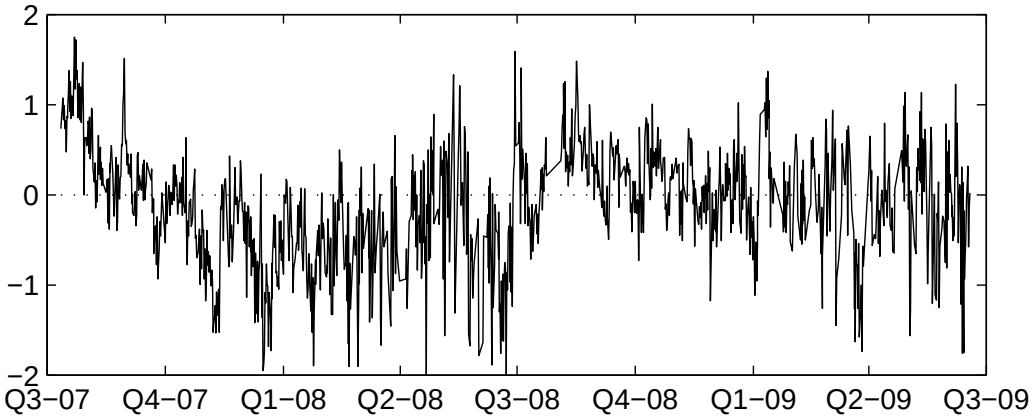
Par la suite, nous considérons la série temporelle $y_{t_1}^{t_n} = (y_{t_1}, ..., y_{t_n})$ obtenue en soustrayant les données METOP à la tendance moyenne OIV2 (cf. figure 3.2b). Notons que la distribution de $y_{t_1}^{t_n}$, donnée par l'histogramme de la figure 3.1b, semble suivre une loi normale sauf pour la queue de distribution négative. Nous dirons par la suite que $y_{t_1}^{t_n}$ est une anomalie de SST, une réalisation à temps discret d'un processus stationnaire à temps continu $\{Y_t\}$.

3.2 Méthode

Les données que nous utilisons sont issues du METOP. Elles sont donc, d'après 1.1.2, des observations imparfaites de la véritable anomalie de SST latente. D'un autre côté, nous utiliserons un modèle simple et local de dynamique de l'anomalie de SST, proposé dans la littérature et imparfait lui aussi. Au final, la formulation espace d'état présentée en 2 sera adoptée en 3.2.1 et nous inférerons sur les paramètres du modèle en 3.2.2, en utilisant des méthodes statistiques classiques (moments et maximum de vraisemblance).



(a)



(b)

FIG. 3.2: (a) : Séries temporelles SST METOP (points) et OIV2 (trait plein) en °C sur deux ans. (b) : Anomalie de SST Metop en °C après avoir retiré la tendance moyenne donnée par l'analyse OIV2.

3.2.1 Modèle

Nous utilisons la modélisation espace état présentée en Chapitre 2. Nous supposons que l'anomalie de SST latente $\{X_t\}$ est observée au temps t par $\{Y_t\}$. Ces deux variables aléatoires sont reliées par l'équation d'observation suivante :

$$Y_{t_i} = HX_{t_i} + \sqrt{R}\varepsilon_{t_i} \tag{3.1}$$

où $\{\varepsilon_{t_i}\}_{i\in\{1,...,n\}}$ est un bruit blanc Gaussien d'espérance nulle et de variance unité. R et H représentent respectivement la variance de l'erreur des observations et une éventuelle transformation de l'état. Ce sont des scalaires qui, par abus de notation, seront notés en majuscule dans ce chapitre. Notons que l'écart-type de l'erreur des observations METOP considérées ici (en gardant les données de jour et de nuit) est globalement de 0.5°C. Cette valeur a été obtenue empiriquement en comparant la SST de ce capteur à des données de bouées dérivantes et peut varier selon l'algorithme utilisé (jour ou nuit), la région et la saison étudiée (cf. [32]).

La variable latente $\{X_t\}$ est supposée suivre un processus d'Ornstein-Uhlenbeck stationnaire qui est la solution de l'équation différentielle stochastique suivante :

$$dX_t = -\lambda X_t dt + \Gamma dW_t \tag{3.2}$$

où $\{W_t\}$ représente un bruit Brownien. Cette modélisation locale de l'anomalie de SST a été proposée initialement par [18] et appliqué à la SST par [47]. Celle-ci néglige les transports horizontaux de chaleur et les échanges de température. $\lambda > 0$ est la corrélation temporelle ou "feedback parameter" qui représente l'évolution lente du transfert de chaleur et $\Gamma > 0$ est la variabilité venant des fluctuations de temps.

Notons $\sigma^2 = Var\,(X_t) = \frac{\Gamma^2}{2\lambda}$. $\{X_t\}$ est un processus Markovien qui vérifie l'équation d'état, $\forall\, i \in \{2, ..., n\}$,

$$X_{t_i} = M_{\Delta_{t_i}} X_{t_{i-1}} + \sqrt{Q_{\Delta_{t_i}}}\eta_{t_i} \tag{3.3}$$

avec $M_{\Delta_{t_i}} = \exp\left(-\lambda\Delta_{t_i}\right)$, $Q_{\Delta_{t_i}} = \sigma^2\left(1 - M^2_{\Delta_{t_i}}\right)$ et $\{\eta_{t_i}\}_{i\in\{1,...,n\}}$ un bruit blanc Gaussien d'espérance nulle et de variance unité, indépendant de $\{\varepsilon_t\}_{i\in\{1,...,n\}}$. De même que pour H et R, nous notons $M_{\Delta_{t_i}}$ et $Q_{\Delta_{t_i}}$ en majuscule alors que ce sont des scalaires. L'équation d'état (3.3) est une généralisation d'un processus AR(1) (cas où l'échantillonnage temporel des données est régulier, cf. 2.1.3). Au final, la structure Markovienne du modèle espace d'état (3.1-3.3) est résumée par un graphique acyclique directionnel (cf. figure 3.3).

3.2.2 Estimation des paramètres

Nous nous intéressons maintenant à l'estimation des paramètres du modèle présenté en 3.2.1, que nous notons par la suite $\theta = (\lambda, \sigma^2, H, R)$.

Dans 3.2.2.1, nous étudierons la structure d'ordre 2 de $\{Y_t\}$ et estimerons θ par la méthode des moments. Nous donnerons également les conditions nécessaires à l'identifiabilité

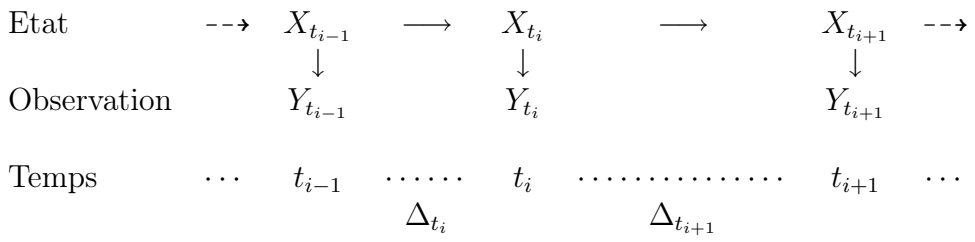


Figure 3.3: Graphique acyclique directionnel pour un modèle espace d'état avec un échantillonnage temporel irrégulier.

des paramètres. Ensuite, dans 3.2.2.2, nous généraliserons l'estimation par maximum de vraisemblance présentée en 2.2.2.2 dans le cas particulier d'un échantillonnage temporel irrégulier des données en implémentant l'algorithme EM. Enfin, dans 3.2.2.3, nous discuterons du schéma d'estimation en pratique, combinant les avantages des deux approches et utilisant une méthode d'optimisation numérique.

3.2.2.1 Méthode des moments

La méthode des moments consiste à ajuster les paramètres de telle sorte que les moments théoriques coincident avec les moments empiriques. Or, d'après les hypothèses, $\{Y_t\}$ est un processus stationnaire de loi normale d'espérance nulle et, $\forall t$, il y a indépendance d'une part entre les $\{X_t\}$ et $\{\varepsilon_t\}$ et d'autre part entre les $\{X_t\}$ et $\{\eta_t\}$. Le moment d'ordre 1 ne permet pas d'identifier les paramètres θ. Ainsi, nous nous intéressons à son moment d'ordre 2. D'après (3.1), la variance de Y_t est donnée par :

$$Var\left(Y_t\right) = H^2 Var\left(X_t\right) + R$$

Pour deux temps distincts t_i et t_j, on obtient la covariance suivante :

$$\begin{aligned} Cov\left(Y_{t_i}, Y_{t_j}\right) &= Cov\left(HX_{t_i} + \sqrt{R}\varepsilon_{t_i}, HX_{t_j} + \sqrt{R}\varepsilon_{t_j}\right) \\ &= H^2 Cov\left(X_{t_i}, X_{t_j}\right) \end{aligned}$$

De façon similaire, d'après (3.3) et étant donné que $\{X_t\}$ est un processus stationnaire,

$$Var\left(X_t\right) = \sigma^2 \tag{3.4}$$

Pour deux temps distincts t_i et t_j, on obtient la covariance suivante :

$$\begin{aligned} Cov\left(X_{t_i}, X_{t_j}\right) &= Cov\left(X_{t_i}, M_{\Delta_{t_{i,j}}} X_{t_i} + \sqrt{Q_{\Delta_{t_{i,j}}}}\eta_{t_i}\right) \\ &= \sigma^2 \exp\left(-\lambda \Delta_{t_{i,j}}\right) \end{aligned} \tag{3.5}$$

avec $\Delta_{t_{i,j}} = |t_i - t_j|$.

Finalement, d'après (4.3) et (4.4), la structure d'ordre 2 de $\{Y_t\}$ est donnée par :

$$\begin{cases} Var\left(Y_t\right) = H^2\sigma^2 + R \\ Cov\left(Y_{t_i}, Y_{t_j}\right) = H^2\sigma^2 \exp\left(-\lambda\Delta_{t_{i,j}}\right) \end{cases} \tag{3.6}$$

La covariance de $\{Y_t\}$ donnée en (4.5) dépend du produit $H^2\sigma^2$ et nous devons ajouter une contrainte pour assurer l'identifiabilité des paramètres. Ainsi, nous fixons $H = 1$ ce qui signifie qu'aucune transformation n'est nécessaire entre l'état et les observations. Nous sommes ramenés à estimer seulement trois paramètres notés $\theta = (\lambda, \sigma^2, R)$. Cette covariance correspond à un modèle classique en statistique spatiale. En effet, on retrouve un variogramme exponentiel de palier $\sigma^2 + R$, de portée temporelle $1/\lambda$ avec un effet pépite R. Par la suite, nous travaillerons sur γ, la fonction variogramme définie par :

$$\begin{aligned} \gamma\left(\Delta_{t_{i,j}}\right) &= \frac{1}{2} Var\left(Y_{t_i} - Y_{t_j}\right) \\ &= \frac{1}{2}\left(Var\left(Y_{t_i}\right) + Var\left(Y_{t_j}\right) - 2Cov\left(Y_{t_i}, Y_{t_j}\right)\right) \\ &= \sigma^2\left(1 - \exp\left(-\lambda\Delta_{t_{i,j}}\right)\right) + R1_{\{0\}}\left(\Delta_{t_{i,j}}\right) \end{aligned} \tag{3.7}$$

La méthode des moments consiste à suivre les méthodes usuelles de la géostatistique. Après avoir estimé un variogramme empirique, nous ajusterons un variogramme théorique exponentiel en utilisant les moindres carrés pondérés (cf. [7]), ce qui permettra de vérifier le réalisme du modèle et d'estimer les paramètres θ. Par la suite, nous utiliserons l'abréviation EMM pour parler des estimateurs obtenus par la méthode des moments.

3.2.2.2 Maximum de vraisemblance

La façon alternative d'estimer les paramètres θ consiste à utiliser la méthode du maximum de vraisemblance. D'après 2.2.2.2, la fonction de log-vraisemblance complète est basée sur les séquences latentes, observées et un *a priori* :

$$\log\left(p\left(x_{t_1}^{t_n}, y_{t_1}^{t_n}; \theta\right)\right) = \log\left(p\left(x_{t_1}\right)\right) + \sum_{i=2}^{n} \log\left(p\left(x_{t_i}|x_{t_{i-1}}; \theta\right)\right) + \sum_{i=1}^{n} \log\left(p\left(y_{t_i}|x_{t_i}; \theta\right)\right)$$

où les distributions conditionnelles $p\left(x_{t_i}|x_{t_{i-1}}; \theta\right)$ et $p\left(y_{t_i}|x_{t_i}; \theta\right)$ suivent des lois normales respectivement données par (3.3) et (3.1). Nous supposons que $p\left(x_{t_1}\right)$ suit une distribution normale de moyenne $x^{(b)}$ et de variance B correspondant à une connaissance *a priori* du

système. Ainsi, la log-vraisemblance du modèle s'écrit (à une constante près) :

$$\begin{aligned}\log\left(p\left(x_{t_1}^{t_n}, y_{t_1}^{t_n}; \theta\right)\right) &= -(n-1)\log(\sigma) - \frac{1}{2}\sum_{i=2}^{n}\log\left(1-\exp\left(-2\lambda\Delta_{t_i}\right)\right) \\ &\quad -\frac{1}{2\sigma^2}\sum_{i=2}^{n}\frac{\left(x_{t_i}-\exp\left(-\lambda\Delta_{t_i}\right)x_{t_{i-1}}\right)^2}{\left(1-\exp\left(-2\lambda\Delta_{t_i}\right)\right)} \\ &\quad -\frac{n}{2}\log(R) - \frac{1}{2R}\sum_{i=1}^{n}\left(y_{t_i}-x_{t_i}\right)^2 \end{aligned} \tag{3.8}$$

On cherche les paramètres θ maximisant la fonction de log-vraisemblance complète (3.8). Nous utilisons pour cela l'algorithme itératif EM (cf. algorithme 2.3). Ici, nous travaillons dans le cas particulier d'un échantillonnage temporel irrégulier. L'étape M, qui changera du cas général présenté en 2.2.2.2, sera développée ici. Nous noterons $\theta_k = (\lambda_k, \sigma_k^2, R_k)$ les approximations successives des estimateurs au long des différents cycles k de l'algorithme.

3.2.2.2.1 Etape E Afin de déterminer $U\left(\theta|\theta_k\right)$, fonction de θ, nous avons besoin des probabilités de lissage conditionnelles données en 2.2.2.2.4. Notons que le filtre et lisseur de Kalman se généralisent simplement au cas de l'échantillonnage irrégulier des données et les termes $x_{t_i}^{(s)}$, $x_{t_{i-1},t_i}^{(s)}$ et $x_{t_{i-1},t_{i-1}}^{(s)}$, que nous allons rencontrer dans l'étape M, sont détaillés et donnés en section 3.2 de l'annexe B.

3.2.2.2.2 Etape M Dans cette étape, on cherche les paramètres θ_{k+1} maximisant $U\left(\theta|\theta_k\right)$, l'espérance conditionnelle de la fonction de log-vraisemblance. Cette étape varie du cas général (cf. 2.2.2.2.3) des modèles espace d'état linéaire et Gaussien avec un échantillonnage régulier. Commençons par décomposer $U\left(\theta|\theta_k\right)$ comme suit :

$$U\left(\theta|\theta_k\right) = U_X\left(\lambda, \sigma^2|\theta_k\right) + U_{Y|X}\left(R|\theta_k\right)$$

avec,

$$\begin{aligned} U_X\left(\lambda, \sigma^2|\theta_k\right) &= -(n-1)\log(\sigma) - \frac{1}{2}\sum_{i=2}^{n}\log\left(1-\exp\left(-2\lambda\Delta_{t_i}\right)\right) \\ &\quad -\frac{1}{2\sigma^2}\sum_{i=2}^{n}\frac{x_{t_i,t_i}^{(s)} - 2\exp\left(-\lambda\Delta_{t_i}\right)x_{t_{i-1},t_i}^{(s)} + \exp\left(-2\lambda\Delta_{t_i}\right)x_{t_{i-1},t_{i-1}}^{(s)}}{1-\exp\left(-2\lambda\Delta_{t_i}\right)} \end{aligned}$$

et

$$U_{Y|X}\left(R|\theta_k\right) = -\frac{n}{2}\log(R) - \frac{1}{2R}\sum_{i=1}^{n}\left(y_{t_i}^2 - 2y_{t_i}x_{t_i}^{(s)} + x_{t_i,t_i}^{(s)}\right)$$

On trouve de façon analytique R_{k+1} qui maximise $U_{Y|X}\left(R|\theta_k\right)$:

$$\hat{R}_{k+1} = \frac{1}{n}\sum_{i=1}^{n}\left(y_{t_i}^2 - 2y_{t_i}x_{t_i}^{(s)} + x_{t_i,t_i}^{(s)}\right)$$

Ce n'est pas le cas pour la maximisation de $U_X(\lambda, \sigma^2|\theta_k)$ qui nécessite l'utilisation d'une procédure d'optimisation numérique de type quasi-Newton. Pour simplifier le problème et se ramener à l'estimation d'un seul paramètre, nous écrivons la relation suivante entre $\hat{\lambda}_{k+1}$ et $\hat{\sigma}^2_{k+1}$:

$$\hat{\sigma}^2_{k+1} = \frac{1}{n-1}\sum_{i=2}^{n} \frac{x^{(s)}_{t_i,t_i} - 2\exp\left(-\hat{\lambda}_{k+1}\Delta_{t_i}\right) x^{(s)}_{t_{i-1},t_i} + \exp\left(-2\hat{\lambda}_{k+1}\Delta_{t_i}\right) x^{(s)}_{t_{i-1},t_{i-1}}}{1 - \exp\left(-2\hat{\lambda}_{k+1}\Delta_{t_i}\right)}$$

3.2.2.3 Mise en oeuvre

L'algorithme EM a certaines limites connues dicutées en 2.2.2.2.6. Notamment, il peut converger vers un maximum local de la fonction de vraisemblance si les paramètres initiaux sont mal choisis. Il est donc important d'initialiser $\hat{\theta}_1$ à une valeur réaliste. Pour cela, nous utilisons les estimateurs robustes donnés par la méthode des moments présentée en 3.2.2.1.

Une autre limite de l'algorithme EM, comme nous l'avions déjà signalé en 2.2.2.2.6, est la convergence lente vers le maximum global de la fonction de vraisemblance. L'utilisation d'une méthode d'optimisation numérique sur la fonction de log-vraisemblance incomplète (donnée en 2.2.2.2.5) permet de palier à ce problème. Le fait d'utiliser une méthode de type quasi-Newton après l'algorithme EM permet également d'extraire une estimation de la matrice Hessienne et donc d'obtenir la matrice d'information de Fisher. Ainsi, on en déduit la variance des estimateurs du maximum de vraisembance (on utilisera l'abréviation EMV par la suite).

3.3 Résultats

Les résultats des estimations (des paramètres du modèle espace d'état présenté en 3.2.1) données par la méthode des moments (cf. 3.2.2.1) et par maximum de vraisemblance (cf. 3.2.2.2) sont discutés dans cette section.

Nous commencerons en 3.3.1 par valider la qualité des estimateurs sur des données simulées par notre modèle dont nous connaissons les vrais paramètres. Ensuite, en 3.3.2, nous travaillerons sur des données réelles de SST en un point particulier de l'océan. Nous estimerons les paramètres du modèle puis nous interpolerons les observations satellitaires d'anomalies de SST. Les résultats seront validés par des données *in situ* indépendantes. Enfin, l'estimation des paramètres sera généralisée à l'océan Atlantique et les résultats seront commentés en 3.3.3.

3.3.1 Simulations

Nous testons ici l'efficacité des deux estimateurs EMM et EMV présentés en 3.2.2.1 et 3.2.2.2 sur des simulations. Pour cela, nous étudions le comportement de ceux-ci pour différentes tailles de série temporelle $n \in \{200, 300, ..., 2000\}$. Ceci correspond à des tailles d'échan-

tillons que nous pouvons observer sur quelques années de données satellitaires. Pour chaque n, nous procédons à $N = 1000$ réalisations suivant ce schéma :

1. on simule l'échantillonnage irrégulier temporel $(\Delta_{t_i})_{i\in\{2,\dots,n\}}$ de façon iid[1] à partir de la distribution empirique de l'échantillonnage du satellite METOP (cf. figure 3.1a)
2. on simule un état initial x_{t_1} de moyenne $x^{(b)}$ et de variance B puis, récursivement et en accord avec (3.3), $(x_{t_i})_{i\in\{2,\dots,n\}}$
3. on simule les observations $(y_{t_i})_{i\in\{1,\dots,n\}}$ en utilisant (3.1).

Lors de ces différentes simulations, nous fixons la valeur des paramètres à des valeurs réalistes, proches de celles qu'on peut trouver lorsqu'on travaille sur des données d'anomalies de SST. Ainsi, $x^{(b)} = 0$, $B = 0.05$ et les paramètres θ valent $\lambda = 0.5$, $\sigma^2 = 0.05$ et $R = 0.5$. Ensuite, pour chaque taille de séquence n, le biais et la variance des estimateurs EMM et EMV sont calculés. En pratique, nous avons estimé les estimateurs EMV à partir d'un algorithme de type quasi-Newton sur la base de la vraisemblance incomplète (cf. 3.2.2.3) en prenant comme valeurs initiales les vrais paramètres afin de réduire les temps de calcul. Les résultats sont données en figure 3.4. Comme attendu, les estimateurs EMV (trait plein) sont généralement meilleurs que ceux de EMM (trait tiret-pointillé) en terme de biais et de variance. Cependant, remarquons que les estimateurs EMM donnent des résultats satisfaisants pour les différentes tailles de séquence n. De plus, ils ont l'avantage de nécessiter peu de temps de calcul et ne sont pas sensibles au choix des valeurs initiales comme pour les estimateurs EMV. Notons également que nous avons calculé les variances des trois paramètres (traits pointillés) à partir de l'information de Fisher moyenne donnée par la méthode de quasi-Newton. On remarque que ces variances et celles obtenues empiriquement par l'estimateur EMV (traits pleins) sont généralement en accord, surtout pour des grandes tailles de n.

Les résultats obtenus par simulations montrent la qualité des différents estimateurs proposés ici. Testons maintenant le schéma d'estimation proposé en 3.2.2.3 sur des données réelles d'anomalies de SST.

3.3.2 Point particulier 0°N 23°W

Travaillons maintenant sur les données satellites METOP au point 0°N 23°W décrites en 3.1. La série temporelle des anomalies de SST $y_{t_1}^{t_n}$ est divisée en deux sous-échantillons : $\left(y_{t_1}, ..., y_{t_{n_1}}\right)$ servira à estimer les paramètres θ du modèle et $\left(y_{t_{n_{1+1}}}, ..., y_{t_n}\right)$ sera utilisé comme moyen de validation. Nous choisissons un nombre d'observations $n_1 = 725$ suffisant d'après les résultats obtenus sur les simulations (cf. 3.3.1) et correspondant à environ $^2/_3$ des observations totales (plus d'une année de données).

En 3.3.2.1, nous discuterons de l'estimation des paramètres du modèle sur l'échantillon d'apprentissage. Ensuite, nous validerons le modèle par validation croisée sur les données satellitaires restantes en 3.3.2.2 et sur des données *in situ* en 3.3.2.3.

[1] indépendant et identiquement distribué

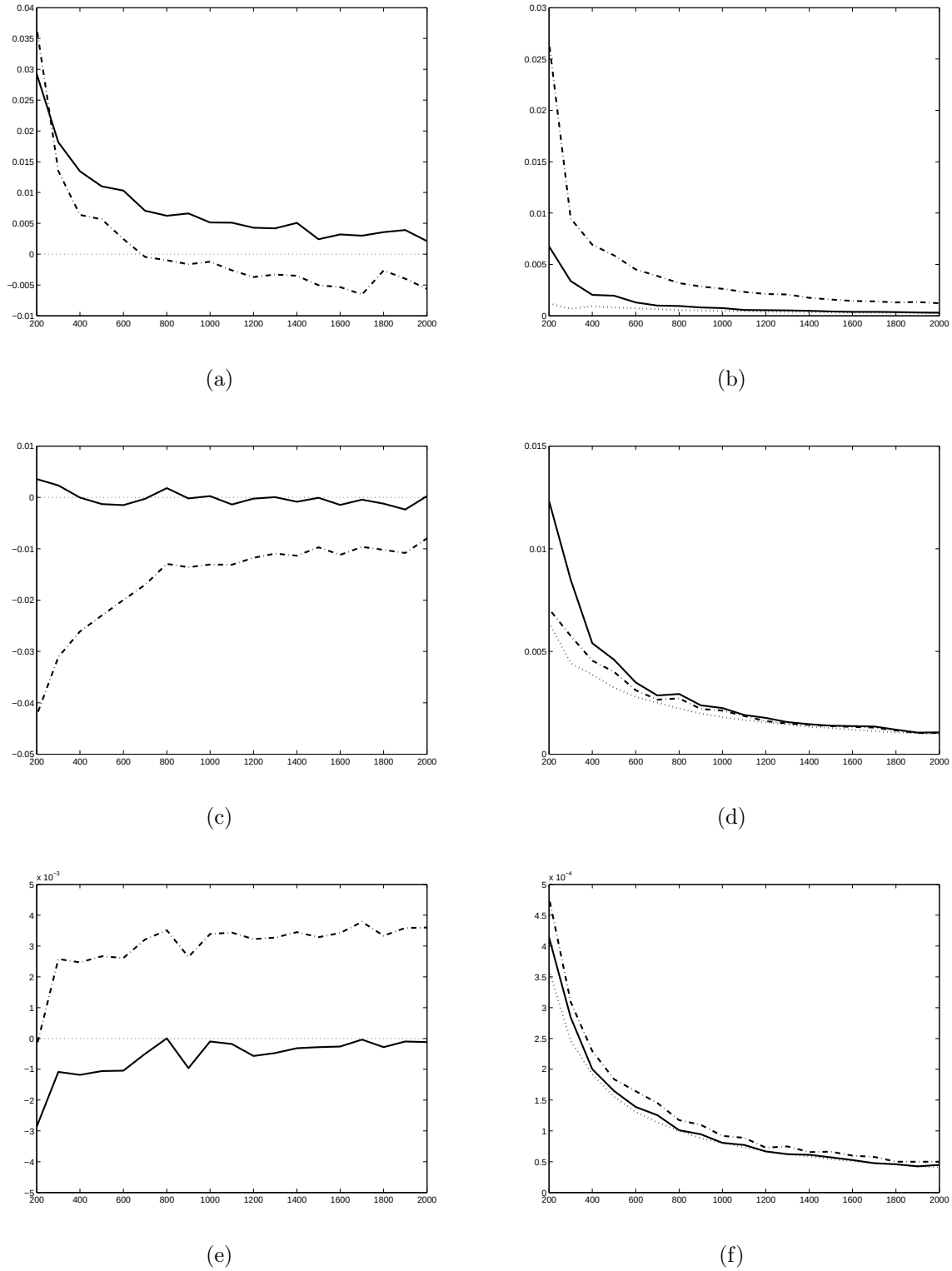


Figure 3.4: Graphique des biais (gauche) et variances simulées (droite) des estimateurs EMM (trait tiret-pointillé) et EMV (trait plein) pour différentes tailles de séquences n (axe des abcisses) et basées sur $N = 1000$ replications : λ en (a) et (b), σ^2 en (c) et (d) ainsi que R en (e) et (f). Les traits pointillés des graphiques des variances correspondent à la variance calculée à partir de l'information de Fisher.

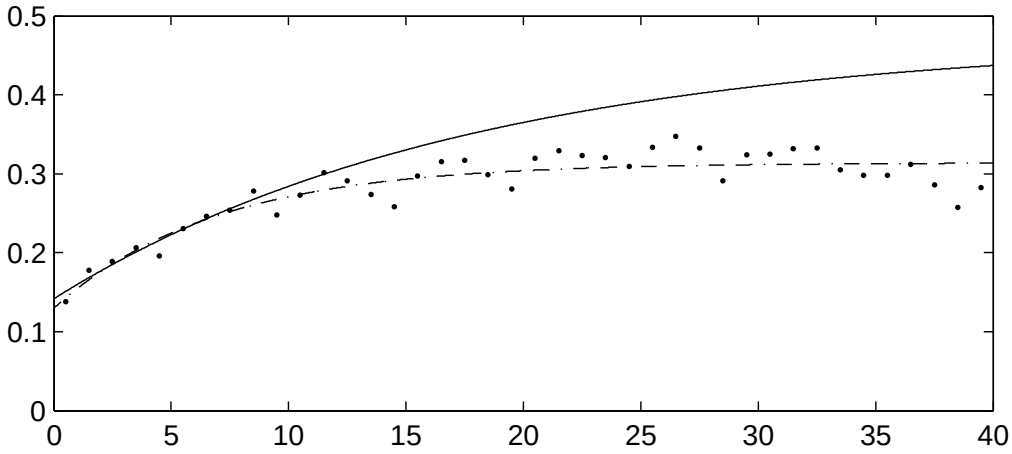


FIG. 3.5: Variogrammes empirique (points) et théorique des estimateurs EMM (trait tiret-pointillé) et EMV (trait plein) obtenus sur le jeu de données d'apprentissage. L'axe des abcisses représente la différence de temps entre deux observations successives.

3.3.2.1 Estimation des paramètres

Nous suivons le schéma d'estimation proposé en 3.2.2.3. La covariance empirique déduite de l'échantillon d'apprentissage $\left(y_{t_1}, ..., y_{t_{n_1}}\right)$ est modélisée par une covariance exponentielle théorique par la méthode des moindres carrés pondérés. Pour plus de lisibilité des résultats, notament pour mieux visualiser la discontinuité de la covariance à l'origine, nous choisissons de travailler sur le variogramme par la transformation suivante : $\gamma\left(\Delta_{t_{i,j}}\right) = Var\left(Y_t\right) - Cov\left(Y_{t_i}, Y_{t_j}\right)$. Les résultats sont visibles sur la figure 3.5. La modélisation obtenue par les estimateurs EMM (trait tiré-pointillé) est en accord avec le variogramme empirique. De plus, on retrouve une forme exponentielle, ce qui conforte les résultats théoriques obtenus sur la structure de la covariance de $\{Y_t\}$ et qui indique que la forme de la structure d'ordre 2 est réaliste.

En partant des estimations EMM obtenues en modélisant la covariance empirique, nous implémentons l'algorithme EM. Les premières itérations montrent une convergence rapide de la log-vraisemblance (cf. figure 3.6d) suivie d'une stagnation lente vers un palier. Nous choisissons alors d'accélérer la convergence en appliquant la procédure d'optimisation numérique de type quasi-Newton proposée en 3.2.2.3.

Les résultats des différentes estimations sont donnés dans le tableau 3.1. L'estimation EMV de la portée temporelle $1/\lambda$ est plus élevée (de l'ordre de 18 jours) que celle obtenue par l'estimateur EMM (environ 7 jours). De plus, les estimations de σ^2 et R varient d'une méthode à une autre. Celles données par le maximum de vraisemblance vérifient l'équation $Var\left(Y_t\right) = \sigma^2 + R$ car, empiriquement, $Var\left(y_{t_1}^{t_{n_1}}\right) = 0.47$. On remarque que ceci n'est pas vrai pour l'estimation EMM car, sur la période de 40 jours sur laquelle on construit le variogramme empirique, le palier empirique n'est pas atteint. Cependant, la différence des estimateurs EMM et EMV de σ^2 n'est pas significative du fait du fort écrat-type 0.094 (cf. tableau

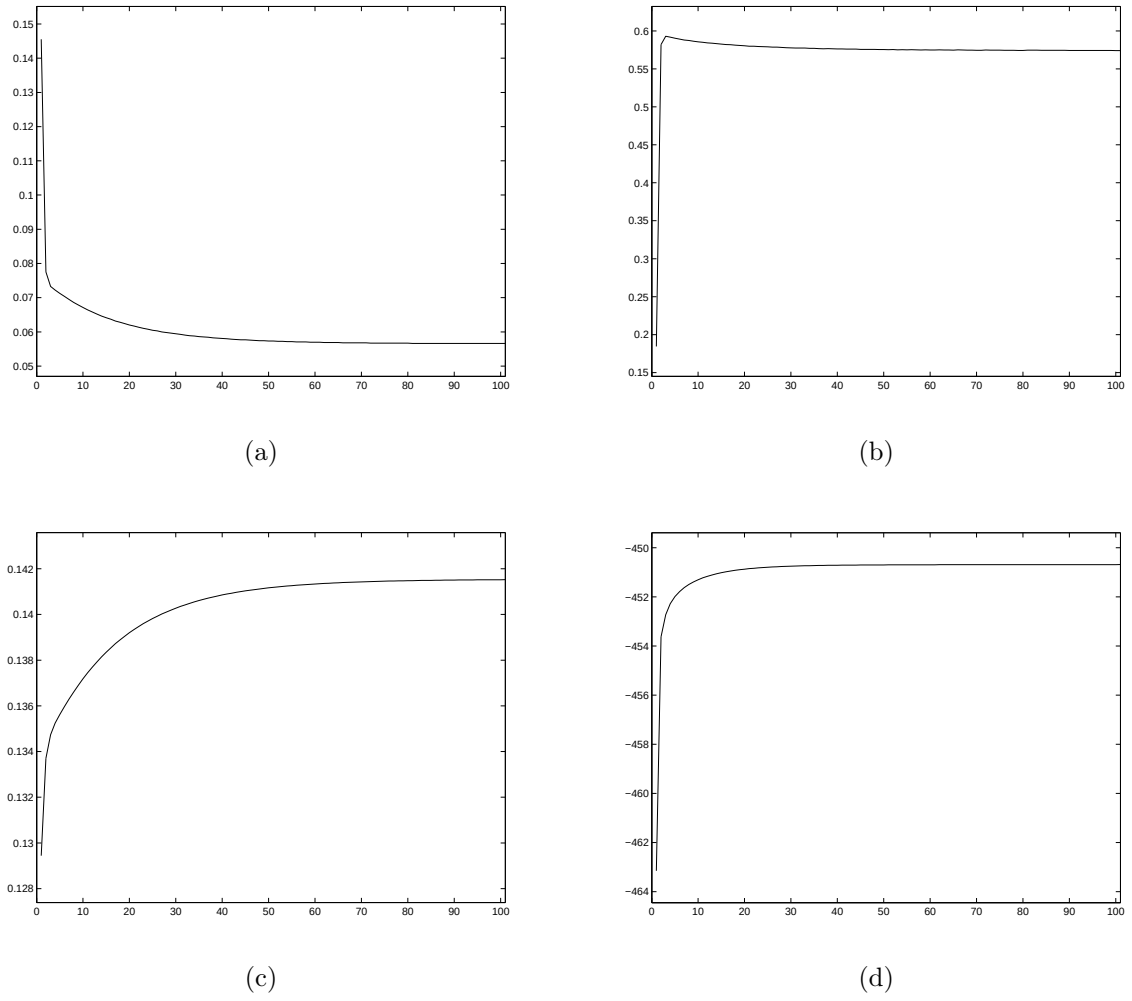


Figure 3.6: Evolution des paramètres EMV de λ (a), σ^2 (b) puis R (c) ainsi que de la log-vraisemblance (d) au cours des 100 itérations (axe des abscisses) de l'algorithme EM.

	Méthode des moments	Maximum de vraisemblance		Ecart type
		Algorithme EM	Quasi-Newton	
$\hat{\lambda}$	0.145	0.057	0.056	0.019
$\hat{\sigma}^2$	0.184	0.329	0.330	0.094
$\hat{R}$	0.129	0.141	0.141	0.010
Log-vraisemblance	-463.15	-450.69	-450.68	

TAB. 3.1: Valeurs des paramètres estimés par la méthode des moments, l'algorithme EM après 100 itérations et l'optimisation numérique de type quasi-Newton. La dernière colonne donne une estimation de l'écart-type des estimateurs EMV à partir de l'information de Fisher. Les résultats sont obtenus sur le jeu de données d'apprentissage.

3.1). Au final, le variogramme théorique obtenu par maximum de vraisemblance surestime le variogramme empirique (cf. figure 3.5). Cependant, pour des Δ_{t_i} faibles, de l'ordre d'une journée (variations qui nous intéressent principalement), les covariances théoriques obtenues par les deux méthodes d'estimation donnent des résultats similaires.

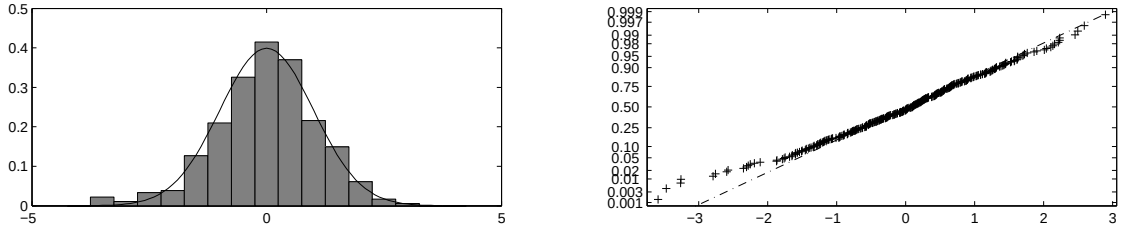
Finalement, les paramètres estimés sont en accord avec notre connaissance physique de la dynamique de l'anomalie de SST. On retrouve en particulier, d'après [32], un écart-type de l'erreur de mesure du satellite METOP en ce point de l'océan pouvant varier de 0.33 à 0.51 °C selon les conditions (algorithme de jour ou de nuit et saison d'étude). Cette fourchette coïncide avec l'intervalle de confiance à 95% autour de notre estimation de $\sqrt{R}$, qui est compris entre 0.35 et 0.40 °C. Ensuite, la faible valeur de notre estimation de λ indique une importante mémoire temporelle de l'anomalie de SST qui est justifiée par le fait qu'on se trouve dans la région équatoriale où la corrélation temporelle de la SST est élevée. Finalement, en comparant la variance de l'innovation de la dynamique temporelle au bout d'un jour ($\hat{Q}_1 = 0.04$) à celle des erreurs d'observations $\hat{R}$, on remarque que, dans le filtrage de Kalman, le poids donné à la précédente analyse est plus important que celui donné aux observations (cf. discussion à ce sujet en 2.2.2.1).

3.3.2.2 Validation croisée

Nous procédons maintenant à une validation croisée des résultats obtenus en 3.3.2.1 sur notre échantillon de test $\left(y_{t_{n_{1+1}}}, ..., y_{t_n}\right)$. Ainsi, pour chaque $i \in \{n_{1+1}, ..., n\}$, les observations y_{t_i} sont retirées du lissage de Kalman et nous calculons

$$x^{(s)}_{t_i|i}\left(\hat{\theta}\right) = E\left(X_{t_i}|y^{t_{i-1}}_{t_{n_1+1}}, y^{t_n}_{t_{i+1}}; \hat{\theta}\right),\ P^{(s)}_{t_i|i}\left(\hat{\theta}\right) = Var\left(X_{t_i}|y^{t_{i-1}}_{t_{n_1+1}}, y^{t_n}_{t_{i+1}}; \hat{\theta}\right)$$

où $\hat{\theta}$ représente les estimateurs EMV du tableau 3.1 (en suivant le schéma suivant : méthode des moments, algorithme EM puis méthode de Quasi-Newton). Les hypothèses faites en 3.2.1 seront vérifiées si la distribution de Y_{t_i} conditionnellement aux observations passées $y^{t_{i-1}}_{t_{n_1+1}}$ et futures $y^{t_n}_{t_{i+1}}$ suit une loi normale de moyenne $x^{(s)}_{t_i|i}\left(\hat{\theta}\right)$ et de variance $P^{(s)}_{t_i|i}\left(\hat{\theta}\right) + \hat{R}$. On calcule



(a) (b)

Figure 3.7: (a) : Histogramme des résidus standardisés obtenus par validation croisée sur le jeu de données test ainsi que la distribution théorique normale centrée réduite (trait plein). (b) : QQ-plot Normal des résidus standardisés.

les résidus standardisés suivants :

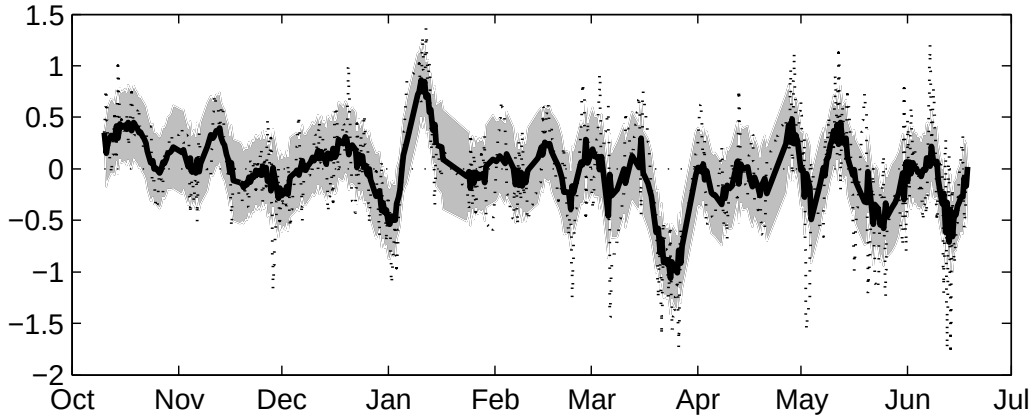
$$\frac{y_{t_i} - x^{(s)}_{t_i|i}\left(\hat{\theta}\right)}{\sqrt{P^{(s)}_{t_i|i}\left(\hat{\theta}\right) + \hat{R}}}$$

L'histogramme de ces résidus ainsi que celui de la distribution théorique de la loi normale centrée réduite sont donnés en figure 3.7a. Dans la figure 3.7b, on vérifie la normalité des résidus en donnant le QQ-plot Normal. On s'aperçoit que la distribution suit bien la distribution normale théorique sauf pour la queue de distribution inférieure qui indique un nombre élevé de résidus fortement négatifs. Cette constatation est confirmée par la figure 3.8a. On y observe des ruptures lorsque les anomalies de SST observées chutent brusquement. Le modèle proposé en 3.2.1 ne semble donc pas prendre en compte cette dynamique non linéaire de l'état caché.

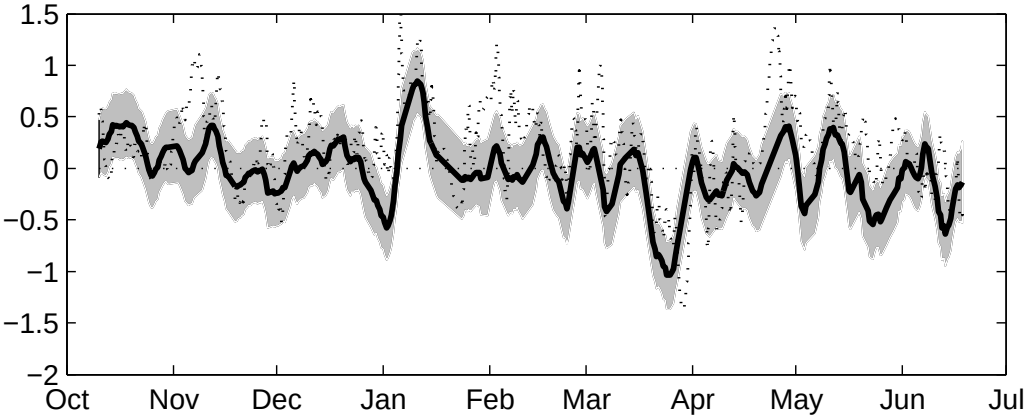
Notons que nous avons également calculé par validation croisée $x^{(s)}_{t_i|i}\left(\hat{\theta}\right)$ et $P^{(s)}_{t_i|i}\left(\hat{\theta}\right)$ (avec $\hat{\theta}$ l'EMM donné en tableau 3.1) dans l'idée de comparer les erreurs quadratiques moyennes (lorsque $\hat{\theta}$ est l'EMM puis l'EMV) données par :

$$\frac{1}{(n - n_{1+1})} \sum_{i=n_{1+1}}^{n} \left(x^{(s)}_{t_i|i}\left(\hat{\theta}\right) - y_{t_i}\right)^2$$

Les résultats montrent que les estimations de $x^{(s)}_{t_i|i}\left(\hat{\theta}\right)$ avec $\hat{\theta}$ étant l'EMV sont plus proches de y_{t_i}. En effet, l'erreur quadratique moyenne passe de 0.21 lorsque $\hat{\theta}$ est l'EMM à 0.20 lorsque $\hat{\theta}$ est l'EMV. Ceci confirme le fait que le schéma d'estimation proposé en 3.2.2.3, basé sur le couplage d'une méthode des moments et du maximum de vraisemblance, améliore la qualité des estimateurs, même si les données ne sont pas forcément Gaussiennes comme c'est le cas dans notre exemple où la distribution est clairement dissymétrique (cf. figure 3.1b).



(a)



(b)

Figure 3.8: (a) Anomalies de SST en °C observées (trait discontinu) et interpolées par lissage de Kalman pour lequel $\hat{\theta}$ est l'EMV (trait plein) avec l'intervalle de confiance à 95% correspondant (gris). (b) Anomalies de SST en °C provenant de la bouée (trait discontinu) et des données satellites interpolées (trait plain) avec l'intervalle de confiance à 95% (gris). Les résultats des interpolations sont obtenus par validation croisée sur le jeu de données test.

	Biais	Ecart-type	RMSE
Données satellites brutes	-0.22	0.47	0.52
Données satellites interpolées	-0.22	0.31	0.38

TAB. 3.2: Différence entre les anomalies de SST (brutes et interpolées par lisseur de Kalman) et les données *in situ*. Les biais, écart-types et RMSE sont obtenus sur le jeu de données test.

3.3.2.3 Comparaison aux données *in situ*

Après avoir validé nos résultats sur un jeu de données test provenant du satellite METOP, nous comparons les interpolations d'anomalies de SST par lissage de Kalman à des données *in situ* indépendantes, provenant de mesures faites toutes les 10 minutes par la bouée fixe du réseau PIRATA au point 0°N 23°W (cf. 1.1.1). Après avoir retranché l'analyse SST OIV2 (présentée en 1.1.3) aux données *in situ*, nous les comparons aux interpolations $x_{t_i}^{(s)}$ données par le lissage de Kalman aux mêmes temps $(t_{n_1+1}, ..., t_n)$. L'idée est ici d'émuler une bouée virtuelle à partir de données satellitaires. D'après la figure 3.8b, les variations des mesures *in situ* sont, en général, bien restituées par le lisseur de Kalman mais seulement 63% des observations de la bouée se trouvent dans l'intervalle de confiance à 95% des $x_{t_i}^{(s)}$. Cependant, le tableau 3.2 indique que l'utilisation du modèle avec les paramètres $\hat{\theta}$ estimés en 3.3.2.1 permet d'améliorer la qualité des données brutes du satellite METOP. En effet, on réduit de façon significative l'écart-type de l'erreur sans changer le biais moyen. Etant donné que les résultats obtenus en 3.3.2.2 indiquent que le modèle espace d'état proposé est réaliste pour des données satellitaires, on peut supposer que les différences entre la bouée virtuelle et les données *in situ* soulignées ici proviennent de la différence physique des deux sources de données. La première raison est le biais bien connu "depth-to-skin" discuté en [32] : le capteur satellite METOP observe la "skin SST" (température des premiers μm) alors que les bouées fixes mesurent la "depth SST" (température jusqu'à 1 m de profondeur) où le gradient de température évolue fortement. La seconde raison est la différence d'échelles : les bouées apportent des mesures locales tandis que le capteur METOP renseigne sur la température moyennée sur une surface de 5×5 km^2.

3.3.3 Généralisation sur l'Atlantique

Le modèle local de variation temporelle de l'anomalie SST défini en 3.2.1 semble adapté aux données satellitaires METOP. L'estimation des paramètres θ faite au point 0°N 23°W en 3.3.2 rend des résultats physiquement réalistes. Disposant de séries temporelles complètes METOP et OIV2 sur deux années complètes (2008 et 2009) sur tous les océans, nous décidons d'étendre le schéma d'estimation proposé en 3.2.2.3 sur une grande partie de l'Atlantique. La zone d'étude s'étend donc de 80°W à 30°E en longitude et 50°S à 60°N en latitude avec un pas d'échantillonnage spatial de 1°. Le nombre de points n des séries temporelles ayant servi à estimer les paramètres est donné par la figure 3.9a. On remarque que n varie selon la région d'étude. Certaines zones ont une occurrence de points très forte avec un échantillonnage temporel Δ_{t_i} des données quasiment régulier d'environ une demi-journée. D'autres, soumises

à une couverture nuageuse importante, montrent une différence de temps entre deux observations successives beaucoup plus hétérogène. De manière générale, n varie entre 100 et 1000 observations. D'après les résultats obtenus sur simulations (cf. 3.3.1), on peut donc s'attendre à des estimations biaisées de θ avec une variabilité forte pour les régions où n est faible.

Nous présenterons successivement les résultats des estimations de λ en 3.3.3.1, σ^2 en 3.3.3.2 et enfin R en 3.3.3.3. Notons que pour certains zones étudiées, les estimateurs convergent vers des valeurs aberrantes. Celles-ci ne seront pas prises en compte dans les résultats suivants.

3.3.3.1 Paramètre λ

Le paramètre $\lambda > 0$ est la corrélation temporelle ou feedback parameter qui représente l'évolution lente du transfert de chaleur à la surface des océans. L'équation (3.3) montre la relation qu'il existe entre λ et $M_{\Delta_{t_i}}$ qui, dans le cas d'un échantillonnage temporel régulier ($\Delta_{t_1} = ... = \Delta_{t_n}$), n'est autre que le coefficient autorégressif d'un modèle AR(1) à un jour. Ainsi, plus λ est grand et plus la dépendance temporelle de l'état est faible.

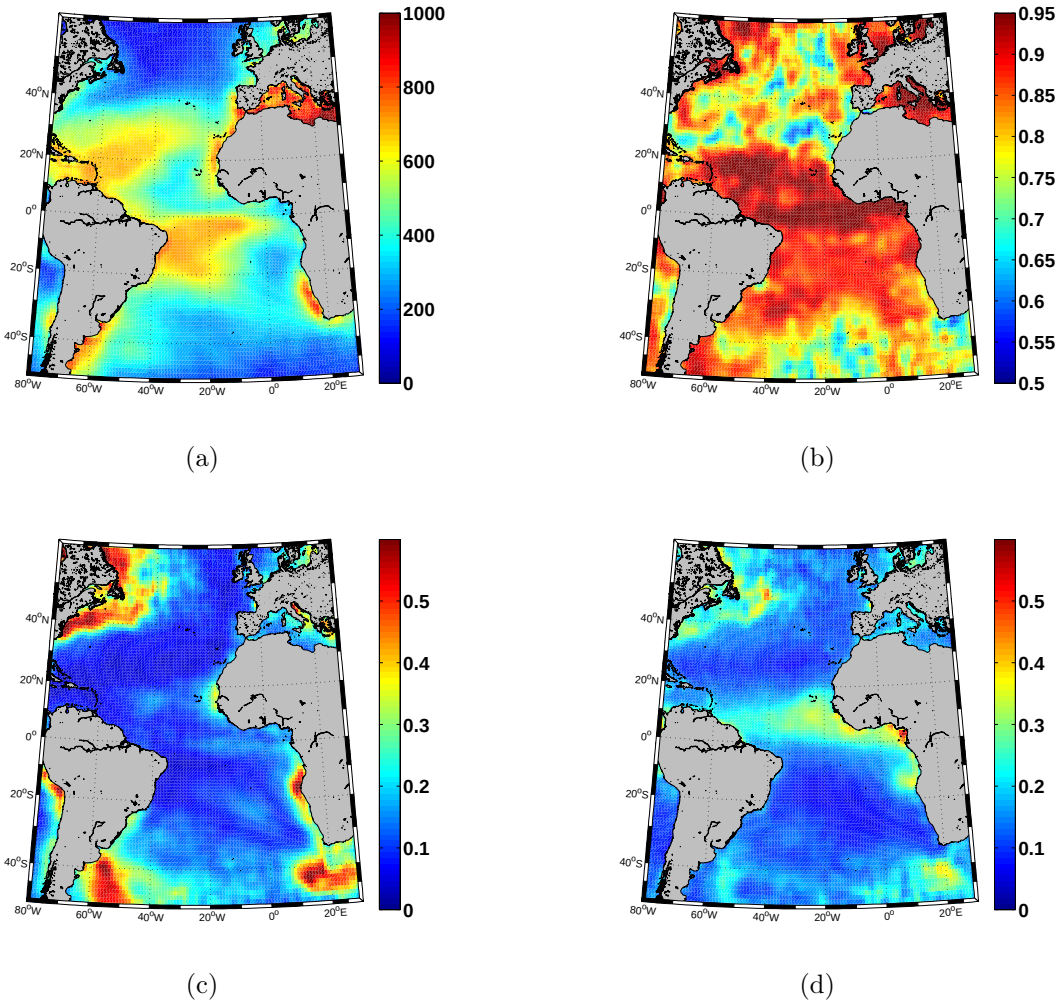
Les résultats de la répartition spatiale des estimateurs $\hat{M}_1 = \exp\left(-\hat{\lambda}\right)$ sont donnés en figure 3.9b. La distribution de $\hat{M}_1$ semble être purement latitudinale. On remarque une forte corrélation temporelle de l'anomalie de SST dans la Méditerranée. On identifie également une zone autour de l'équateur qui relie les continents africain et américain. C'est une zone de convergence intertropicale bien connue (notée ITCZ[2]) où les vents de l'Atlantique Nord et Sud convergent, formant ainsi une langue d'eau froide se déplaçant d'Est en Ouest : on parle alors du phénomène d'upwelling équatorial. D'après [34], cette région varie au cours de l'année, se déplaçant entre 15°N et 8°S. En moyenne, comme nous l'observons sur la figure 3.9b, cette langue d'eau froide, où le signal de l'anomalie de SST est fortement persistante, est légèrement au-dessus de l'équateur. Au contraire, dans les parties tempérées de l'Atlantique, λ croît de façon significative et indique une diminution du coefficient autorégressif. Enfin, dans les zones de latitudes élevées, les résultats sont plus aléatoires, du fait du faible nombre d'observations.

3.3.3.2 Paramètre σ^2

Le paramètre $\sigma^2 = Var\left(X_t\right)$ est la variance de la variable d'état ou l'anomalie de SST latente. Il dépend de la variabilité venant des fluctuations de temps notée τ, qu'on retrouve dans l'équation (3.2), ainsi que le paramètre de corrélation temporelle λ.

Les résultats de la répartition spatiale des estimateurs $\hat{\sigma}^2$ sont donnés en figure 3.9c. On retrouve une variabilité forte de l'anomalie de SST dans des zones fortement dynamiques telles que le courant du Gulf Stream, des aiguilles ainsi que celui du bord Ouest respectivement au large du Canada, de l'Afrique de Sud et de l'Argentine. D'autres régions plus particulières indiquent également une variablité élevée de notre variable d'état. Parmi elles, on identifie des régions de fort upwelling bien connues en Atlantique. La plus importante est

[2]Inter Tropical Convergence Zone



(a) (b)

(c) (d)

Figure 3.9: (a) : Taille n des différentes séries temporelles stationnaires obtenues à partir de deux années de données du capteur satellite METOP sur une partie de l'Atlantique et de la Méditerranée. Répartition spatiale des estimateurs de $M_1 = \exp(-\lambda)$ en j^{-1} (b), σ^2 (c) et R (d) obtenus sur une partie de l'Atlantique et de la Méditerranée.

le système du Benguela au Sud-Ouest de l'Afrique. Cette zone, riche en phytoplancton et donc en poisson, est une région de forte concentration démographique du fait de l'abondance des pêches. Plus au Nord, au large du Sénégal et des îles Canaries, on retrouve également un système d'upwelling. Celui-ci est identifié en figure 3.9c, par une variabilité importante de l'anomalie de SST, surtout concentrée au large du Sénégal où les vents sont forts, favorisant le mélange des eaux peu profondes. Dans le reste de l'Atlantique et de la Méditerranée, σ^2 est relativement stationnaire et faible, de l'ordre de 0.1.

3.3.3.3 Paramètre R

Le paramètre R représente la variance de l'erreur de mesure du capteur satellite METOP. Une estimation empirique de cette variance a été réalisée dans [32] en comparant les observations METOP à des données de bouées dérivant dans l'océan : on parle de variance d'erreur SSES, déjà présentée en 1.1.2 et 1.1.4. Ces résultats étaient basés sur un nombre restreint de colocalisations et ne donnaient donc pas d'estimations robustes. Par contre, dans notre cas, les longueurs des séries temporelles utilisées sont satisfaisantes (cf. figure 3.9a). Ainsi, le paramètre R est un estimateur robuste de la variance d'erreur SSES du capteur METOP.

Les résultats de la répartition spatiale des estimateurs $\hat{R}$ sont donnés en figure 3.9d. La principale source d'erreur de mesure du capteur METOP se situe au large du Sahara et s'étend jusqu'aux côtes de l'Amérique du Sud. Celle-ci est causée par les aérosols et le contenu en vapeur d'eau dans l'atmosphère qui perturbent les estimations de SST des capteurs infra-rouges. Ces aérosols sont des poussières de sable venant du Sahara qui sont déplacées vers l'Ouest le long de l'équateur (cf. [17] pour plus d'informations). Plus ponctuellement, la variance de l'erreur de METOP est forte à la sortie du fleuve Amazone, zone où une importante quantité d'eau douce pénètre dans l'Atlantique, changeant la densité et la salinité de l'eau de mer, perturbant ainsi la restitution de SST du capteur.

3.4 Discussion

Dans ce chapitre, nous nous sommes focalisés sur la dynamique temporelle de l'anomalie de SST. Pour cela, nous avons travaillé sur des séries temporelles provenant du capteur infra-rouge METOP ainsi que des données analysées OIV2 afin de stationnariser les signaux. Or, les séries temporelles satellitaires ont comme particularité d'être irrégulièrement échantillonnées dans le temps. Ainsi, l'utilisation d'un modèle espace d'état particulier, avec l'utilisation d'une équation d'état différentielle stochastique, est nécessaire. Les paramètres de ce modèle donnent l'information de la dynamique temporelle, la variance stationnaire de la variable d'état ainsi que la variance de l'erreur du capteur satellite en question : METOP. Nous avons utilisé les techniques classiques d'estimation des paramètres, la méthode des moments et celle du maximum de vraisemblance, en proposant une combinaison des deux approches. La qualité des estimateurs a été validée par simulations de séries temporelles (état et observations), connaissant les vrais paramètres d'un modèle espace d'état linéaire à erreurs Gaussiennes, avec un échantillonnage irrégulier des données. Etant donné la qualité des estimateurs obtenus par simulations, nous avons testé notre schéma d'inférence (sur un point particulier de l'océan

Atlantique au niveau de l'équateur) par validation croisée puis par comparaison des résultats de l'interpolation obtenue par notre méthode avec des données *in situ* provenant d'une bouée fixe. Les résultats indiquent que nous restituons correctement le signal d'anomalie de SST. Une fois ce travail réalisé sur un point, nous avons généralisé l'estimation des paramètres de notre modèle à l'océan Atlantique. Ainsi, nous avons identifié des zones où la dynamique temporelle de l'état est particulière. Nous avons notamment mis en évidence les zones bien connues d'upwelling, régions très riches en poisson, qui ont un grand intérêt scientifique pour les années à venir. De plus, nous avons remarqué une large zone dans l'Atlantique tropical où l'erreur commise par le capteur satellite infra-rouge METOP est importante ; c'est une région de fortes concentrations d'aérosols dans l'atmosphère dues aux poussières de sable du Sahara. Le signal est également contaminé par le contenu en vapeur d'eau.

Les résultats des estimations des paramètres du modèle espace d'état temporel proposé va servir de prémisse au chapitre 4 suivant. Nous travaillerons en particulier sur l'extension spatiale de ce modèle espace d'état où l'information des pixels voisins viendra compléter l'estimation de l'état au point étudié.

CHAPITRE 4

Modélisation spatio/temporelle de la SST

Dans les chapitres 2 et 3 précédents, nous avons présenté le modèle espace d'état linéaire et Gaussien et une version locale, temporelle, de celui-ci. Jusqu'ici, la composante spatiale qui avait été introduite en 1.1.2 n'était pas prise en compte : c'est maintenant ce que nous cherchons à faire. Afin de caractériser la variabilité spatiale, nous utiliserons des outils classiques de géostatistique.

Nous commencerons en 4.1 par présenter en détail les différentes données utilisées dans ce chapitre. Nous insisterons sur la nature des champs de SST fournis par différents capteurs infra-rouges et micro-ondes. Puis, en 4.2, nous généraliserons notre modèle en y ajoutant la dimension spatiale et en prenant en compte les différentes observations satellitaires. Viendra ensuite, en section 4.3, l'exposition des résultats de l'estimation des paramètres de ce modèle spatio/temporel sur l'Atlantique ainsi que l'estimation de notre variable d'intérêt sur différentes zones de cet océan. Enfin, en 4.4, une dicussion sur les limites et les perspectives du modèle clorera cette étude.

4.1 Données

Dans ce chapitre, nous souhaitons utiliser les données des trois capteurs présentés en 1.1.2 avec des résolutions spatiales spécifiques à chaque source d'information (cf. tableau 1.1). Dans ce chapitre, les résolutions initiales des capteurs SEVIRI (0.10°) et AMSRE (0.25°) sont ramenées à la même résolution que METOP (0.05°) par interpolation par plus proche voisin comme le montre les figures 1.5. Une fois cette transformation réalisée, on peut exploiter les caractéristiques spécifiques des trois capteurs :

- METOP est capable, grâce à sa haute résolution, de détecter des structures spatiales petite échelle et ses variances d'erreurs de mesure sont très faibles
- SEVIRI est une bon compromis entre l'échantillonnage spatial, temporel et la variance d'erreur des mesures

– AMSRE, le seul capteur micro-onde de l'étude, de plus faible résolution spatial permet, en cas de couverture nuageuse, de disposer d'informations sur la SST contrairement aux deux autres capteurs infra-rouges.

Sachant que nous prenons maintenant en compte la composante spatiale du signal de SST, nous travaillons sur deux zones $3° \times 3°$ (environ 300 km×300 km) dans l'Atlantique dont les points 49°S 59°W et 30°N 30°W sont les centres (cf. figure 1.5). Nous verrons par la suite en 4.3.1.2 que ces deux régions sont particulières du fait de la variabilité spatiale de leur SST. Concernant la période d'étude, nous nous focaliserons uniquement sur l'année 2008, les données de nuit et les pixels de qualité 4 et 5. Enfin, comme pour le chapitre 3, nous ne travaillerons pas directement sur la SST mais sur son anomalie. Ainsi, la stationnarisation temporelle des champs de SST se fera encore une fois par rapport à la source indépendante OIV2 présentée en 1.1.3. Nous signalons que, dans la zone d'étude 49°S 59°W, seuls les capteurs METOP et AMSRE fournissent des données alors que, pour 30°N 30°W, les trois sources de données sont présentes.

Dans ce chapitre, les observations et l'état méritent une indexation particulière. Premièrement, étant donné que nous raisonnons de façon spatiale et temporelle, nous indicerons par s pour l'espace et t pour le temps. Deuxièmement, nous disposons de trois capteurs satellitaires et l'exposant $k \in \{1, 2, 3\}$ permettra de faire le distinguo entre les trois sources de données. Nous noterons $y^{(k)}_{s_j,t_i}$ une observation au temps t_i, $i \in \{1, ..., n\}$ du capteur k au point particulier s_j de coordonnées $\begin{pmatrix} lon_j \\ lat_j \end{pmatrix}$, avec $j \in \{1, ..., p\}$; comme le montrent les figures 1.5 et 1.6 le nombre de pixels p définis sur l'image varie selon le satellite k et le temps t_i. L'état correspondant sera noté x_{s_j,t_i} avec cette fois-ci $j \in \{1, ..., r\}$ où $r = r_{lon} \times r_{lat}$ (valeur fixe au cours du temps et telle que $r \geq p$) désignera le nombre de pixels de la grille d'analyse (r_{lon} et r_{lat} sont les nombres respectifs de pixels en longitude et latitude). Par exemple, si on se place sur une grille d'analyse de résolution 0.05° sur une zone $3° \times 3°$, cela correspond à $r_{lon} = r_{lat} = 60$ soit un vecteur d'état de taille $r = 3600$.

Par la suite, nous ne travaillerons pas sur un pixel ponctuel, au site s_j et au temps t_i, mais sur un ensemble de points qui formeront nos vecteurs d'observation et d'état aux différents temps $(t_1, ..., t_n)$. Pour l'observation satellitaire k au temps t_i, le domaine spatial formé par les pixels référencés (non manquants) sera noté $D_{p(i,k)} = (s_1, ..., s_p)$ alors que celui de l'état sera fixe et noté $D_r = (s_1, ..., s_r)$. Ainsi, nous noterons respectivement $y^{(k)}_{D_{p(i,k)},t_i}$ et x_{D_r,t_i} les vecteurs d'observation et d'état. Une schématisation de ceux-ci est donnée en figure 4.1.

4.2 Méthode

Les données que nous utilisons dans ce chapitre proviennent cette fois-ci de trois sources satellitaires distinctes. Nous allons donc proposer en 4.2.1 un modèle espace d'état linéaire et Gaussien comportant plusieurs équations d'observations. De plus, nous prendrons en compte l'information spatiale dans l'équation d'état. L'étude de la structure d'ordre 2 du modèle conduira à l'étude, en 4.2.2, des paramètres régissant la variabilité spatiale. Au final, du fait de la prise en compte de l'information des pixels voisins, la dimension de l'état va croître

$$\begin{bmatrix} x_{s_1,t_i} & \cdots & \cdots & \cdots & \cdots \\ x_{s_2,t_i} & \ddots & & & \vdots \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \vdots \\ x_{s_{r_{lat}},t_i} & \cdots & \cdots & \cdots & x_{s_r,t_i} \end{bmatrix} \longrightarrow \begin{bmatrix} x_{s_1,t_i} \\ x_{s_2,t_i} \\ \vdots \\ \vdots \\ x_{s_r,t_i} \end{bmatrix} = x_{D_r,t_i}$$

(a)

$$\begin{bmatrix} y^{(k)}_{s_1,t_i} & \cdots & \cdots & \cdots & NaN \\ NaN & \ddots & & & \vdots \\ y^{(k)}_{s_2,t_i} & & \ddots & & \vdots \\ \vdots & & & \ddots & y^{(k)}_{s_p,t_i} \\ NaN & \cdots & \cdots & \cdots & NaN \end{bmatrix} \longrightarrow \begin{bmatrix} y^{(k)}_{s_1,t_i} \\ y^{(k)}_{s_2,t_i} \\ \vdots \\ y^{(k)}_{s_p,t_i} \end{bmatrix} = y^{(k)}_{D_{p(i,k)},t_i}$$

(b)

FIG. 4.1: Schématisation de l'obtention des vecteurs d'état (a) et d'observation (b) au temps t_i et pour le capteur satellitaire k à partir des champs d'anomalie de SST représentés par les matrices de tailles $r_{lon} \times r_{lat}$. Notons que les NaN correspondent à des observations manquantes.

largement. Nous discuterons alors en 4.2.3 de la résolution par filtre et lisseur de Kalman dans ce genre de problème à haute dimension.

4.2.1 Modèle

Nous utilisons la modélisation espace d'état présentée en chapitre 2. On suppose que les vecteurs d'état et d'observation sont reliés par l'équation suivante :

$$Y^{(k)}_{D_{p(i,k)},t_i} = H^{(k)}_{t_i} X_{D_r,t_i} + \left(R^{(k)}_{t_i}\right)^{1/2} \varepsilon^{(k)}_{D_{p(i,k)},t_i} \tag{4.1}$$

où, pour tout capteur satellite k, $\left\{\varepsilon^{(k)}_{D_{p(i,k)},t_i}\right\}_{i\in\{1,\dots,n\}}$ est un bruit blanc Gaussien en espace et en temps d'espérance nulle et de variance unité. De plus, nous supposons que les erreurs $\left\{\varepsilon^{(k)}_{D_{p(i,k)},t_i}\right\}_{i\in\{1,\dots,n\}}$ et $\left\{\varepsilon^{(k')}_{D_{p(i,k')},t_i}\right\}_{i\in\{1,\dots,n\}}$, $\forall k \neq k'$, sont indépendantes : en effet, les satellites sont de nature différente (certains sont infra-rouges et d'autres micro-ondes), leurs erreurs de mesures sont donc supposées indépendantes. $H^{(k)}_{t_i} \in \mathbb{R}^{p(i,k)\times r}$, l'opérateur d'observation, est une matrice de changement de domaine D_r vers $D_{p(i,k)}$, permettant de passer le vecteur d'état de dimension r à la dimension $p(i,k)$ du vecteur d'observation au temps t_i pour le capteur k (avec $p(i,k) \leq r$). En supposant qu'il n'existe pas de coefficient multiplicateur entre les observations et l'état, $H^{(k)}_{t_i}$ est une matrice identité pour laquelle on ne garde que les indices des lignes où nous disposons d'observations. Pour ce qui est de la covariance d'erreur d'observation, elle est supposée non corrélée dans l'espace. Cette hypothèse est souvent

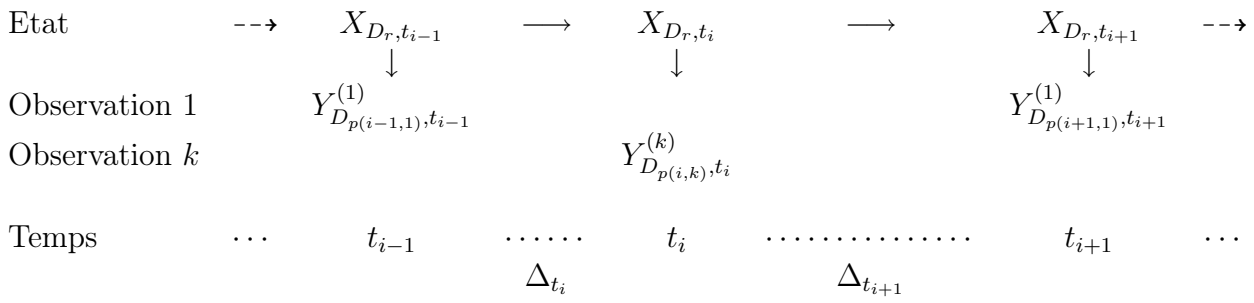


Figure 4.2: Graphique acyclique directionnel pour un modèle espace d'état avec un échantillonnage temporel irrégulier et k sources d'observation.

faite en assimilation de données satellitaires (cf. [8]). Nous obtenons alors une matrice diagonale définie positive $R^{(k)}_{t_i} \in \mathbb{R}^{p(i,k)\times p(i,k)}$, indicée par la source k et le temps t_i, où nous gardons uniquement les indices des lignes et colonnes pour lesquels nous disposons d'observations. D'après le chapitre 3, les valeurs d'erreurs d'observations peuvent être estimées à partir d'observations locales (cf. 3.3.3.3 pour le capteur METOP). Or, d'après 3.3.1, des séries temporelles ayant peu d'observations ne permettent pas d'estimer de façon précise les paramètres du modèle. Ainsi, nous préférons utiliser les variances d'erreurs fournies par les SSES pour chaque pixel de données satellitaires (cf. 1.1.2 et annexe A). Ces statistiques sont basées sur un nombre important de colocalisations.

Comme pour le modèle espace d'état à temps irrégulier présenté en 3.2.1, la variable latente $\{X_{D_r,t}\}$ est supposée suivre un processus d'Ornstein-Uhlenbeck particulier, cette fois-ci multivarié (cf. [42]), solution de l'équation différentielle stochastique (3.2). Au final, l'équation d'état est donnée par :

$$X_{D_r,t_i} = M_{\Delta_{t_i}} X_{D_r,t_{i-1}} + \left(Q_{\Delta_{lon},\Delta_{lat},\Delta_{t_i}}\right)^{1/2} \eta_{D_r,t_i} \tag{4.2}$$

avec $M_{\Delta_{t_i}} = \exp\left(-\lambda\Delta_{t_i}\right) I_r$, $Q_{\Delta_{lon},\Delta_{lat},\Delta_{t_i}} = \Sigma^2_{\Delta_{lon},\Delta_{lat}} - M_{\Delta_{t_i}} \Sigma^2_{\Delta_{lon},\Delta_{lat}} M'_{\Delta_{t_i}}$ et $\{\eta_{D_r,t_i}\}_{i\in\{1,\dots,n\}}$ un bruit blanc Gaussien d'espérance nulle et de variance unité, indépendant, pour tout k, de $\left\{\varepsilon^{(k)}_{D_{p(i,k)},t_i}\right\}_{i\in\{1,\dots,n\}}$. Notons que I_r correspond à la matrice carrée identité de taille $r \times r$. La matrice de la dynamique temporelle, $M_{\Delta_{t_i}} \in \mathbb{R}^{r\times r}$, est ici supposée diagonale, ne prennant en compte aucune diffusion de l'état au cours du temps ; pour le moment, dans un souci de simplification du modèle, toute la structure spatiale sera contenue dans $\Sigma^2_{\Delta_{lon},\Delta_{lat}}$ (nous discuterons d'une modélisation plus élaborées de $M_{\Delta_{t_i}}$ en 4.5). Nous reprenons les estimations de λ définies en 3.3.3.1 pour modéliser $M_{\Delta_{t_i}}$. La matrice de covariance d'erreur de l'état $Q_{\Delta_{lon},\Delta_{lat},\Delta_{t_i}} \in \mathbb{R}^{r\times r}$ dépend de $\Sigma^2_{\Delta_{lon},\Delta_{lat}} \in \mathbb{R}^{r\times r}$ qui représente la covariance spatiale stationnaire de l'état pour $\Delta_{t_i} = t_i - t_{i-1} = 0$. La modélisation de $\Sigma^2_{\Delta_{lon},\Delta_{lat}}$ sera discutée en 4.2.2.1.

La structure Markovienne du modèle espace d'état (4.1-4.2) est résumée par un graphique acyclique directionnel (cf. figure 4.2).

4.2.1.1 Structure d'ordre 2

Reprenons la méthode des moments présentée en 3.2.2.1 : cette fois-ci, nous sommes dans un cas multivarié. Encore une fois, nous nous intéressons uniquement au moment d'ordre 2 de $\left\{Y^{(k)}_{D_{p(i,k)},t}\right\}$ pour une source satellitaire k fixée. On rappelle que les hypothèses sur le modèle sont les suivantes : $\forall t_i$, il y a indépendance d'une part entre $\{X_{D_r,t_i}\}$ et $\left\{\varepsilon^{(k)}_{D_{p(i,k)},t_i}\right\}$ et d'autre part entre $\{X_{D_r,t_i}\}$ et $\{\eta_{D_r,t_i}\}$. Pour deux temps distincts t_i et t_j et en se plaçant sur une zone d'étude D_r, on obtient la covariance[1] suivante :

$$\begin{aligned} Cov\left(Y^{(k)}_{D_{p(i,k)},t_i}, Y^{(k)}_{D_{p(j,k)},t_j}\right) &= Cov\left(H^{(k)}_{t_i}X_{D_r,t_i} + \left(R^{(k)}_{t_i}\right)^{1/2}\varepsilon_{D_{p(i,k)},t_i}, H^{(k)}_{t_j}X_{D_r,t_j} + \left(R^{(k)}_{t_j}\right)^{1/2}\varepsilon^{(k)}_{D_{p(j,k)},t_j}\right) \\ &= H^{(k)}_{t_i}Cov\left(X_{D_r,t_i}, X_{D_r,t_j}\right)H^{(k)'}_{t_j} \end{aligned}$$

De plus, d'après (4.1), sa covariance stationnaire dans le temps, $\forall t_i$, est donnée par :

$$Cov\left(Y^{(k)}_{D_{p(i,k)},t_i}, Y^{(k)}_{D_{p(i,k)},t_i}\right) = H^{(k)}_{t_i}Cov\left(X_{D_r,t_i}, X_{D_r,t_i}\right)H^{(k)'}_{t_i} + R^{(k)}_{t_i}$$

Les notations introduites en (4.2) nous indiquent que la covariance stationnaire dans le temps de $\{X_{D_r,t_i}\}$, $\forall t_i$ est notée :

$$Cov\left(X_{D_r,t_i}, X_{D_r,t_i}\right) = \Sigma^2_{\Delta_{lon},\Delta_{lat}} \tag{4.3}$$

Pour deux temps distincts t_i et t_j, on obtient la covariance suivante :

$$\begin{aligned} Cov\left(X_{D_r,t_i}, X_{D_r,t_j}\right) &= Cov\left(X_{D_r,t_i}, M_{\Delta_{t_{i,j}}}X_{D_r,t_i} + \left(Q_{\Delta_{lon},\Delta_{lat},\Delta_{t_{i,j}}}\right)^{1/2}\eta_{D_r,t_i}\right) \\ &= \Sigma^2_{\Delta_{lon},\Delta_{lat}}\exp\left(-\lambda\Delta_{t_{i,j}}\right) \end{aligned} \tag{4.4}$$

Remarquons que nous sommes dans le cas d'un modèle de covariance spatio/temporelle séparable (cf. [21]), produit d'une covariance temporelle $\exp\left(-\lambda\Delta_{t_{i,j}}\right)$ et d'une covariance spatiale $\Sigma^2_{\Delta_{lon},\Delta_{lat}}$.

Finalement, d'après (4.3) et (4.4), la structure d'ordre 2 de $\left\{Y^{(k)}_{D_{p(i,k)},t_i}\right\}$ est donnée par :

$$\begin{cases} Cov\left(Y^{(k)}_{D_{p(i,k)},t_i}, Y^{(k)}_{D_{p(i,k)},t_i}\right) = H^{(k)}_{t_i}\Sigma^2_{\Delta_{lon},\Delta_{lat}}H^{(k)'}_{t_i} + R^{(k)}_{t_i} \\ Cov\left(Y^{(k)}_{D_{p(i,k)},t_i}, Y^{(k)}_{D_{p(i,k)},t_j}\right) = H^{(k)}_{t_i}\Sigma^2_{\Delta_{lon},\Delta_{lat}}\exp\left(-\lambda\Delta_{t_{i,j}}\right)H^{(k)'}_{t_j} \quad , t_i \neq t_j \end{cases} \tag{4.5}$$

Pour deux observations du même capteur satellite k, de coordonnées spatiales et temporelles respectives $s_{i'}, t_i$ et $s_{j'}, t_j$ la covariance s'écrit :

$$Cov\left(y^{(k)}_{s_{i'},t_i}, y^{(k)}_{s_{j'},t_j}\right) = \Sigma^2_{\Delta_{lon_{i',j'}},\Delta_{lat_{i',j'}}}\exp\left(-\lambda\Delta_{t_{i,j}}\right)$$

[1] notée Cov

Ainsi, en utilisant une approche espace-d'état avec une matrice $M_{\Delta_{t_i}}$ diagonale, on obtient une structure d'ordre 2 de $\left\{Y^{(k)}_{D_{p(i,k)},t}\right\}$ séparable en temps et en espace. Cette approche, basée sur la modélisation espace-d'état est différente de celle du krigeage spatio-temporel. En effet, pour cette dernière méthode, on souhaite modéliser au mieux la variabilité des données : ceci peut conduire à une structure d'ordre 2 non séparable avec une intéraction entre l'espace et le temps. Dans notre cas, nous cherchons avant tout à privilégier les enjeux de la thèse en gardant un modèle temporel simplement explicable sur l'état et en utilisant la structure Markovienne du modèle afin de réduire les temps de calcul (l'état au temps t_i est estimé seulement à partir de l'état précédent t_{i-1} et des observations au temps t_i).

Au final, afin d'estimer la structure d'ordre 2, il est nécessaire d'estimer $\Sigma^2_{\Delta_{lon},\Delta_{lat}}$, la covariance stationnaire de l'état (pour $\Delta_{t_i} = 0$). Ceci constitura notre sujet d'étude lors de la prochaine section.

4.2.2 Estimation des paramètres

Nous nous intéressons maintenant à l'estimation des paramètres du modèle présenté en 4.2.1 et notament, en 4.2.2.1, d'après l'étude de la structure d'ordre 2 de l'état et de l'observation (cf. 4.2.1.1), à la modélisation de la variabilité spatiale de l'anomalie de SST. Pour cela, nous utiliserons les outils classiques de géostatistique. Enfin, en 4.2.2.2, nous présenterons rapidement l'estimation *a posteriori* par maximum de vraisemblance de quelques-uns de ces paramètres spatiaux.

4.2.2.1 Modélisation de la variabilité spatiale

La variabilité spatiale de l'anomalie de SST constitue un intérêt pour les modèles d'interpolation de SST ne prenant en compte que les observations satellitaires de surface comme celui présenté dans cette thèse. Plusieurs auteurs ont proposé des méthodes pour étudier cette variabilité spatiale : [55] et [31] pour l'océan Atlantique nord, [22] pour le golfe de Gascogne ou encore [41] pour la Mer du Japon. Ici, notre objectif est de modéliser $\Sigma^2_{\Delta_{lon},\Delta_{lat}}$ définie en (4.3), la matrice de covariance spatiale stationnaire de l'anomalie de SST (pour $\Delta_{t_i} = 0$) afin de connaître au mieux, dans un but explicatif, les paramètres de cette variabilité spatiale. En conséquence, une bonne paramétrisation de celle-ci permettra d'obtenir des résultats d'interpolations spatiales réalistes.

Pour estimer au mieux $\Sigma^2_{\Delta_{lon},\Delta_{lat}}$, la meilleure approche aurait été d'utiliser les données *in situ*, considérées comme la mesure de référence de la SST latente. Or, ces données de bouées ont un échantillonnage spatial très faible (cf. figure 1.2). Il n'est donc pas possible d'avoir une estimation précise de la variabilité spatiale. Nous décidons alors d'utiliser le capteur satellite METOP (ayant une précision et un échantillonnage spatial satisfaisant) sur lequel nous retrancherons les données analysées OIV2 afin de sationnariser le processus et de ne travailler que sur la variabilité locale de la SST.

Or, d'après 3.3.3.3, un effet pépite $\hat{R}$ donné par la figure 3.9d, correspondant à une variabilité petite échelle non détectée (temporelle) ou une erreur d'instrumentation, est observée sur ce capteur METOP. Nous retrancherons par la suite ces estimations $\hat{R}$ déterminées dans la

section 3.9d à la covariance des observations définie par (4.5). Ainsi, lors de l'estimation des variogrammes théoriques (cf. 4.2.2.1.3) et par conséquent des paramètres de la covariance spatiale $\Sigma^2_{\Delta_{lon},\Delta_{lat}}$ de l'état, nous ne chercherons pas à estimer la pépite spatiale. Lorsque nous essayons de l'estimer automatiquement en balayant des zones $1° \times 1°$ sur tout l'Atlantique (comme nous le verrons en 4.3.1.3), on observe des problèmes d'instabilité numérique (divergence des paramètres) dans l'ajustement du variogramme. Malgré tout, nous sommes conscient que cette supposition faite sur l'égalité des pépites spatiales et temporelles de METOP n'est pas idéale.

Dans cette section, en 4.2.2.1.1, nous poserons les hypothèses nécessaires à la modélisation géostatistique de la variabilité spatiale d'un processus. Nous aborderons ensuite en 4.2.2.1.2 la construction du variogramme empirique permettant de prendre en compte cette variabilité. Enfin, nous parlerons de l'ajustement d'un modèle paramétrique en 4.2.2.1.3.

4.2.2.1.1 Définitions et hypothèses Il est important de s'assurer que la variabilité locale du processus étudié est la même sur le domaine de définition. Cette propriété d'invariance est appelée stationnarité spatiale. En géostatistique, on s'assure uniquement de la stationnarité des deux premiers moments du processus. On dira que l'espérance du processus ne dépend pas du site de l'étude si, $\forall i \neq j$, $E\left(X_{s_i,t} - X_{s_j,t}\right) = 0$. On doit ensuite s'assurer de la stationnarité du moment d'ordre 2. Ainsi, nous introduisons la fonction semi-variogramme γ, définie par :

$$\begin{aligned}
\gamma\left(\Delta_{s_{i,j}}\right) &= \frac{1}{2} Var\left(X_{s_i,t} - X_{s_j,t}\right) \\
&= \frac{1}{2}\left(Var\left(X_{s_i,t}\right) + Var\left(X_{s_j,t}\right) - 2Cov\left(X_{s_i,t}, X_{s_j,t}\right)\right) \\
&= \sigma^2 - Cov\left(X_{s_i,t}, X_{s_j,t}\right) \\
&= \sigma^2 - \Sigma^2_{\Delta_{lon},\Delta_{lat}}\left(\Delta_{s_{i,j}}\right) \qquad (4.6)
\end{aligned}$$

avec $\sigma^2 = Var\left(X_{s,t}\right)$, $\forall s, t$, la variance stationnaire dans le temps de $\{X_{s,t}\}$. De plus, $\Delta^2_{s_{i,j}}$ représente la distance entre deux pixels s_i et s_j d'une même image. lorsque ceux-ci sont proches, du fait que nous projettons les points sur un plan au lieu d'une sphère, on supposera que l'égalité suivante est vraie :

$$\begin{aligned}
\Delta^2_{s_{i,j}} &= \Delta^2_{lon_{i,j}} + \Delta^2_{lat_{i,j}} \\
&= \left(lon_i - lon_j\right)^2 + \left(lat_i - lat_j\right)^2
\end{aligned}$$

En pratique (cf. [7]), pour estimer la covariance $Cov\left(X_{s_i,t}, X_{s_j,t}\right)$ (qui est un élément de la matrice $\Sigma^2_{\Delta_{lon},\Delta_{lat}}$), on préfère travailler sur la fonction semi-variogramme γ car elle ne nécessite pas le calcul de $E\left(X_{s,t}\right)$. La modélisation du variogramme permettra ainsi de déterminer $\Sigma^2_{\Delta_{lon},\Delta_{lat}}$ grâce à la relation (4.6). Notons que, par abus de langage, γ sera appelé fonction variogramme (et non semi-variogramme).

4.2.2.1.2 Variogramme empirique Nous allons évaluer empiriquement la variabilité spatiale du processus stationnaire : nous noterons $\hat{\gamma}$ l'estimateur empirique du variogramme

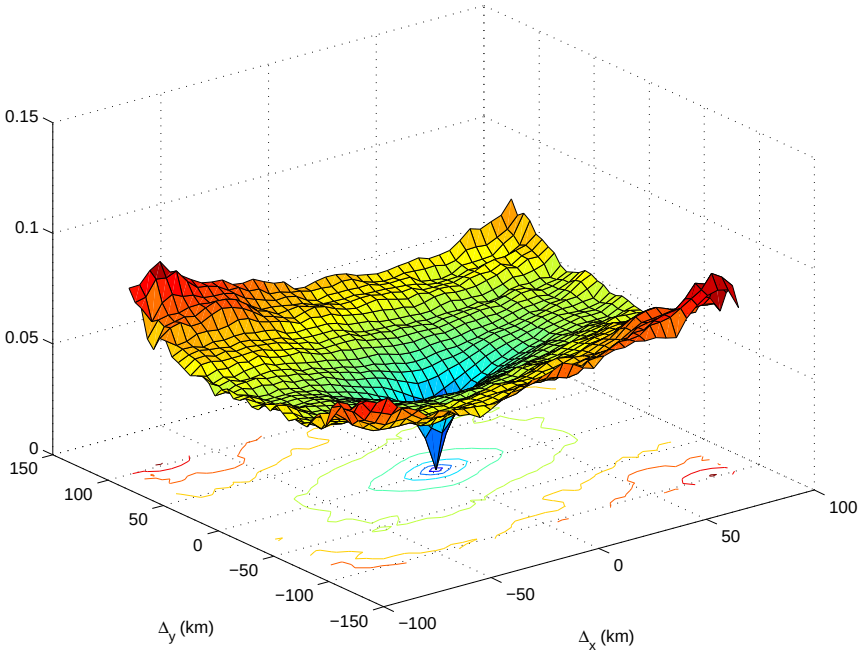
spatial. Dans le cas d'un processus spatial dit isotrope, le comportement de la variabilité est supposé identique suivant toutes les directions. Ainsi, la différenciation entre Δ_{lon} et Δ_{lat} n'a pas lieu d'être et seule la distance Euclidienne entre les points sera retenue. On construira alors une "nuée variographique" définie sur un plan 2D. Dans notre cas de figure, on peut s'attendre à un déplacement des particules suivant une direction donnée et donc observer une variabilité différente selon la direction étudiée. Dans ce cas, nous parlerons d'anisotropie et le variogramme sera fonction de Δ_{lon} et Δ_{lat}. On construira ainsi une "carte variographique" représentée sur un graphique 3D (cf. [23]). Ce genre de variogrammes est généralement estimable empiriquement lorsque le nombre d'observations est très grand, comme c'est le cas pour les données satellitaires (cf. [57] ou [9]). Notons également qu'une étude de ce type, distingant les distances longitudinales et latitudinales, a été réalisée par [41] sur des données d'anomalie de SST au large de la Mer du Japon. Nous aimerions vérifier l'anisotropie de notre variable d'intérêt ; le variogramme empirique $\hat{\gamma}$ sera défini par ses deux dimensions spatiales Δ_{lon} et Δ_{lat}. Une façon usuelle de l'estimer est donnée par (cf. [7]) :

$$\hat{\gamma}\left(\Delta_{lon_{i,j}}, \Delta_{lat_{i,j}}\right) = \frac{1}{2\left|N\left(\Delta_{lon_{i,j}}, \Delta_{lat_{i,j}}\right)\right|} \sum_{(s_i,s_j)\in N\left(\Delta_{lon_{i,j}}, \Delta_{lat_{i,j}}\right)} \sum_{t\in(t_1,\ldots,t_n)} \left(X_{s_i,t} - X_{s_j,t}\right)^2 \quad (4.7)$$

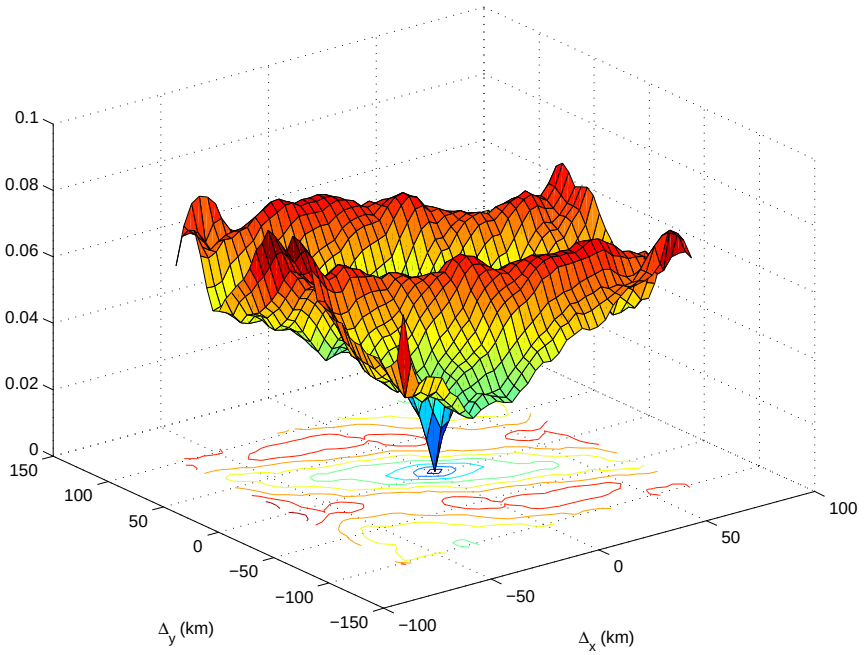
où $N\left(\Delta_{lon_{i,j}}, \Delta_{lat_{i,j}}\right)$ représente les paires de points (s_i, s_j) pour lesquelles $\Delta_{lon_{i,j}} = |lon_i - lon_j|$ et $\Delta_{lat_{i,j}} = |lat_i - lat_j|$. Pour des raisons de temps de calcul et étant donné l'importance de la modélisation du variogramme pour des courtes distances, nous choisissons de travailler sur des zones de $1° \times 1°$. Le variogramme empirique est construit à partir d'une année d'observations METOP (2008) où seuls les pixels de qualité supérieure (4 et 5) et les données de nuit sont utilisées. Quelques exemples de cartes variographiques de l'anomalie de SST sont présentés en figure 4.3. Ces cartes, estimées empiriquement sur des régions dynamiques ou non, forment un échantillon représentatif de la variabilité de notre variable d'intérêt. Remarquons que le variogramme empirique du courant des Malouines (cf. figure 4.3b) présente une oscillation dans la direction de plus petite portée. La distance spatiale de cette périodicité correspond à la taille des structures physiques (tourbillons et fronts) qu'on peut observer sur l'anomalie de SST. Par contre, cette périodicité ne se retrouve pas partout dans l'Atlantique : on la remarque seulement dans quelques zones fortement dynamiques. De plus, on peut également s'interroger sur le fait que les données sont bien stationnaires en espace. Par la suite, nous ne modéliserons pas ces oscillations.

4.2.2.1.3 Variogramme théorique La modélisation de la variabilité spatiale de la SST consiste à chercher une expression analytique simple ajustant au mieux les variogrammes empiriques. Pour cela, appuyons-nous sur les exemples de variogrammes présentés en figure 4.3 où la variabilité spatiale de l'anomalie de SST montre des comportements différents. Ceux-ci mettent en évidence plusieurs enjeux à prendre en compte dans la modélisation du variogramme. Parmi eux, on aimerait prendre en compte :

- l'anisotropie qui montre une variabilité différente selon la direction étudiée. Nous remarquons notamment une forme géométrique particulière dans notre carte variographique : une ellipse, comme pour la figure 4.3b. On parlera d'anisotropie géométrique



(a)



(b)

FIG. 4.3: Exemples de cartes variographiques empiriques moyennes sur l'année 2008 obtenues dans l'Atlantique nord autour de 30°N 30°W (a) et dans le courant des Malouines autour de 59°S 59°W (b) à partir de données METOP et OIV2 sur des zones 1° × 1°.

– la forme du variogramme. Nous remarquons que les variogrammes empiriques montrent une non-dérivabilité au voisinage de 0 suggérant une forme de type exponentielle.

Une fois les enjeux identifiés, nous cherchons à modéliser au mieux nos variogrammes par un modèle théorique unique. Or, celui-ci ne peut pas être choisi arbitrairement : la fonction variogramme doit être définie positive. Au cours de ce paragraphe, nous allons décomposer les étapes de la construction de notre modèle variographique en prenant en compte les remarques soulignés en figure 4.3.

Nous commençons par prendre en compte la forme du variogramme et notamment son comportement au voisinage de 0. Pour cela, nous travaillons sur un problème simple dit isotropique où la variabilité est la même dans toutes les directions. Nous avons testé lors de cette étude plusieurs fonctions variogrammes comme Matern ou power-exponentielle. Au final, une forme simple de type exponentielle permet de modéliser correctement la variabilité spatiale de l'anomalie de SST. Le variogramme exponentiel isotrope s'écrit :

$$\gamma\left(\Delta_{lon}, \Delta_{lat}\right) = \sigma^2 \left(1 - \exp\left(-\frac{\sqrt{\Delta_{lon}^2 + \Delta_{lat}^2}}{L}\right)\right) \tag{4.8}$$

avec la portée, notée L, représentant les longueurs caractéristiques du phénomène.

Prenons maintenant en compte la forme elliptique des variogrammes en supposant que le palier est uniforme selon toutes les directions (cas de l'anisotropie géométrique). Nous travaillons à partir des lignes de niveau du variogramme empirique et nous représentons schématiquement la modélisation de l'ellipse correspondante en figure 4.4. Notons respectivement L_{max} et L_{min} les portées maximales et minimales de l'ellipse qui s'observent selon deux directions orthogonales. L'équation de l'ellipse est donnée par :

$$\left(\frac{L_\theta \cos\left(\theta - \phi\right)}{L_{max}}\right)^2 + \left(\frac{L_\theta \sin\left(\theta - \phi\right)}{L_{min}}\right)^2 = 1 \tag{4.9}$$

où L_θ, définie par $L_{min} \geq L_\theta \geq L_{max}$, représente la portée de l'ellipse selon l'angle θ par rapport au nord, conformément à l'usage en océanographie. Nous noterons ϕ l'angle particulier indiquant la direction de portée maximale. De (4.9), nous déduisons l'expression de L_θ :

$$L_\theta = \frac{L_{min} L_{max}}{\sqrt{L_{max}^2 \cos^2\left(\theta - \phi\right) + L_{min}^2 \sin^2\left(\theta - \phi\right)}} \tag{4.10}$$

où l'angle θ est obtenu à partir des distances en longitude et latitude :

$$\theta = \arctan\left(\frac{\Delta_{lat}}{\Delta_{lon}}\right) \tag{4.11}$$

On peut donc, à partir de (4.10) et (4.11), généraliser le variogramme exponentiel défini en (4.8) dans le cas d'une anisotropie géométrique. Le modèle variographique s'écrit alors :

$$\gamma\left(\Delta_{lon}, \Delta_{lat}\right) = \sigma^2 \left(1 - \exp\left(-\frac{\sqrt{\Delta_{lon}^2 + \Delta_{lat}^2}}{L_\theta}\right)\right) \tag{4.12}$$

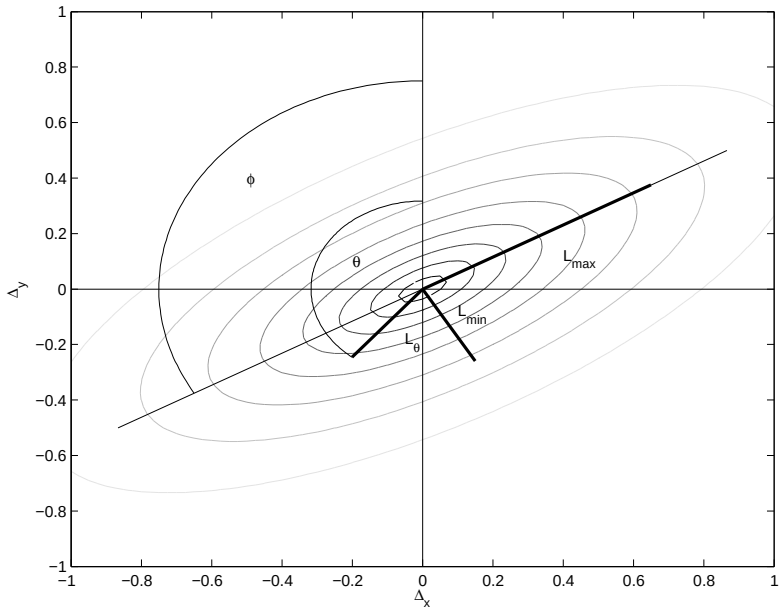


FIG. 4.4: Schématisation de la modélisation d'une ellipse et des paramètres L_{min}, L_{max}, ϕ du variogramme par lignes de niveau.

avec L_θ défini en (4.10). Enfin, d'après (4.6), la matrice de covariance de l'état stationnaire pour $\Delta_{t_i} = 0$, $\Sigma^2_{\Delta_{lon},\Delta_{lat}} \in \mathbb{R}^{r\times r}$, est obtenue à partir de l'expression du variogramme théorique (4.12) :

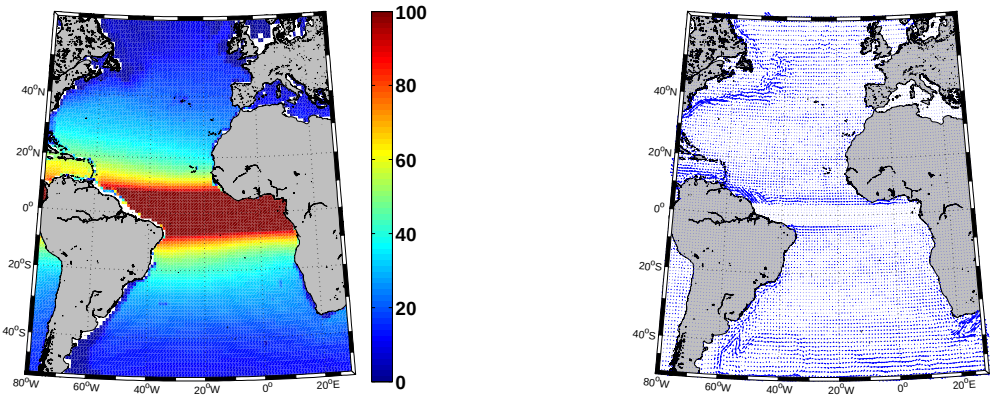
$$\Sigma^2_{\Delta_{lon},\Delta_{lat}} = \sigma^2 \exp\left(-\frac{\sqrt{\Delta^2_{lon} + \Delta^2_{lat}}}{L_\theta}\right) \tag{4.13}$$

Notons que la matrice diagonale $M_{\Delta_{t_i}}$ et le modèle exponentiel de $\Sigma^2_{\Delta_{lon},\Delta_{lat}}$ assurent le caractère défini positif de la covariance d'erreur de l'état $Q_{\Delta_{lon},\Delta_{lat},\Delta_{t_i}}$.

Nous ajusterons le variogramme théorique d'anisotropie géométrique (4.12) au variogramme empirique (4.7) selon le critère des moindres carrés pondérés (cf. [7]) en utilisant un algorithme d'optimisation numérique de type Quasi-Newton. Mathématiquement, cela consiste à chercher le vecteur $\hat{\Theta} = \left(\hat{\sigma}^2, \hat{L}_{min}, \hat{L}_{max}, \hat{\phi}\right) \in \mathbb{R}^+ \times \mathbb{R}^+ \times \mathbb{R}^+ \times [0;\pi[$ vérifiant :

$$\hat{\Theta} = \arg\min_{\Theta}\left(\frac{\sum_{(s_i,s_j)\in N\left(\Delta_{lon_{i,j}},\Delta_{lat_{i,j}}\right)}\left(\hat{\gamma}\left(\Delta_{lon_{i,j}},\Delta_{lat_{i,j}}\right) - \gamma\left(\Delta_{lon_{i,j}},\Delta_{lat_{i,j}}\right)\right)^2}{\left|N\left(\Delta_{lon_{i,j}},\Delta_{lat_{i,j}}\right)\right|}\right)$$

Afin d'éviter une convergence vers un minimum local de notre fonction, les valeurs initiales des paramètres de cette descente de gradient seront déterminées par une connaissance *a priori*. En effet, pour σ^2, nous utiliserons les estimations obtenues en 3.3.3.2. Ensuite, d'après les travaux de [5], nous initialiserons les portées L_{min} et L_{max} par le rayon de Rossby (cf. figure



(a) (b)

Figure 4.5: Rayon de Rossby (a) en km et topographie dynamique moyenne (b) servant d'indicateurs des portées et de l'angle d'anisotropie des variogrammes.

4.5a) qui donne une idée des échelles de corrélation spatiales de la SST. Enfin, un *a priori* sur ϕ sera donné par la topographie dynamique moyenne (cf. figure 4.5b, voir [49] pour plus de détails) représentée par un champ de vecteurs dont la direction est un indicateur des courants moyens pluri-annuels et donc de l'angle d'anisotropie. Les principaux courants forts de l'Atlantique (Falkland, Aiguilles et Gulf Stream) y sont identifiés. Remarquons que, dans la zone de convergence intertropicale, ces vecteurs ne sont pas renseignés.

Nous estimerons les paramètres inconnus Θ et deux exemples de modélisation du variogramme empirique seront étudiés en 4.3.1.1 et 4.3.1.2. Enfin, nous généraliserons l'estimation de la variabilité spatiale à tout l'Atlantique en 4.3.1.3.

4.2.2.2 Maximum de vraisemblance *a posteriori*

Une fois l'estimation des paramètres spatiaux faite par l'étude du moment d'ordre 2 des observations METOP, il est possible, comme pour le modèle uniquement temporel, de généraliser l'estimation des paramètres par maximum de vraisemblance. On peut implémenter par exemple un algorithme de type quasi-Newton en minimisant la vraisemblance incomplète (cf. 2.2.2.2.5) du modèle spatio/temporel présenté en 4.2.1. Nous verrons par la suite que la portée maximale L_{max} est susceptible, dans certains cas de forte anisotropie, d'être ré-estimée par maximum de vraisemblance.

4.2.3 Résolution

Le modèle espace d'état linéaire à erreurs Gaussiennes défini en 4.2.1, prenant en compte le temps d'échantillonnage irrégulier et l'information spatiale, est résolu par filtrage et lissage de Kalman respectivement donnés par les algorithmes 2.1 et 2.2. Or, dans le filtre de Kalman,

d'après (2.16), on calcule une matrice de gain $K_{t_i} \in \mathbb{R}^{r\times p}$ nécessitant l'inversion d'une matrice $\mathbb{R}^{p\times p}$. De plus, à chaque étape i du filtre nous devons stocker les matrices de covariance $P_{t_i}^{(f)}$ et $P_{t_i}^{(a)}$. Il est donc judicieux de choisir une zone d'étude de taille raisonnable permettant de prendre en compte ces contraintes. C'est pour cela que nous choisissons de travailler sur des zones relativement petites ($3° \times 3°$) avec une grille d'interpolation fine de 0.05° (identique à celle du capteur METOP). Dans ce cas de figure, $r_{lon} = r_{lat} = 60$ d'où, on obtient une taille du vecteur d'état $r = 3600$. Quant à elle, la taille p du vecteur d'observation est telle que $p \leq 3600$ (dans le cas parfait où tous les pixels sont référencés, $p = r$).

Notons tout de même que si l'on souhaite augmenter la taille de l'état et ainsi interpoler les données d'anomalies de SST sur des zones plus grandes, il est possible d'implémenter un filtre de Kalman de rang réduit, grâce à l'approximation des matrices de covariance $P_{t_i}^{(f)}$ et $P_{t_i}^{(a)}$ par décompositions en valeurs singulières. Aucun des résultats de ce chapitre n'utilisera ce genre de décompositions mais, pour plus de renseignements, consultez par exemple [25] ou [4].

4.3 Résultats

Dans les sections précédentes, nous avons présenté le schéma d'estimation des différents paramètres et matrices du modèle. Cette méthodologie va maintenant être appliquée en 4.3.1 sur des données de l'océan Atlantique. Enfin, en 4.3.2.1, nous critiquerons les résultats des interpolations spatio/temporelles des anomalies de SST.

4.3.1 Paramètres du modèle

Nous suivons la démarche présentée en 4.2.2 en nous basant uniquement sur des données du capteur METOP. Nous allons discuter des valeurs des paramètres estimés pour chaque région : 4.3.1.1 pour 49°S 59°W et 4.3.1.2 pour 30°N 30°W. Enfin, dans un souci de généralisation du modèle à une plus grande échelle, nous présenterons en 4.3.1.3 la distribution des paramètres de la variabilité spatiale dans l'océan Atlantique.

4.3.1.1 Zone particulière 49°S 59°W

La première zone d'étude se trouve dans le courant des Malouines, autour de 49°S 59°W, où la variabilité spatiale de l'anomalie de SST est fortement anisotropique (cf. figure 4.3b). L'ajustement du variogramme empirique correspondant (cf. 4.6b) se fait par le modèle théorique défini en (4.12) en utilisant les moindres carrés pondérés. Le nombre de couples de points $N(\Delta_{lon}, \Delta_{lat})$ ayant servi à estimer le variogramme empirique est donné en figure 4.6a. On remarque une forte concentration de couples de points pour des distances Δ_{lon} et Δ_{lat} faibles où l'ajustement sera privilégié. Au final, le variogramme empirique ainsi que le modèle exponentiel anisotropique géométrique correspondant (4.12) sont respectivement donnés en figure 4.6b et 4.6c et les estimateurs des paramètres du variogramme sont synthétisés dans le tableau 4.1. La modélisation semble particulièrement adéquate, notamment en ce qui

concerne ϕ, l'angle entre l'axe Δ_{lat} et l'axe de l'ellipse de portée maximale L_{max} (dans le sens anti-horaire), qui est estimé à 49°. Concernant les portées, L_{min} et L_{max} sont respectivement estimés à 13 km et 43 km. Les variogrammes exponentiels théoriques correspondants sont donnés en figure 4.6d. On remarque une portée maximale estimée forte. Ceci peut être dû à une mauvaise stationnarité des anomalies de SST. L'explication la plus probable semble que la variabilité spatiale de l'analyse SST provenant de OIV2 est, du fait du modèle utilisé, fortement isotropique. On peut donc s'attendre, dans des régions de forts courants, à des valeurs fortement aberrantes de ce paramètre de portée maximale L_{max}. Enfin, le palier σ^2 est estimé à 0.06, de l'ordre de grandeur de celui estimé à partir de séries temporelles (cf. 3.3.3.2). Concernant la dépendance temporelle du signal dans cette zone, d'après 3.3.3.1, $\hat{\lambda}$ vaut 0.06 ce qui équivaut à un coefficient AR(1) à un jour de $\hat{M}_1 = 0.94\, j^{-1}$ indiquant une forte corrélation temporelle de l'anomalie de SST. Enfin, les covariances d'erreur varient $R_{t_i}^{(k)}$ selon le capteur satellitaire k, le temps t_i et la qualité des pixels retenus. Pour METOP, celles-ci varient de 0.12 (indice de qualité 5) à 0.22 (indice de qualité 4) alors que pour AMSRE, capteur basse résolution, elles varient de 0.56 à 1.79. Ainsi, dans l'interpolation, les observations METOP auront un poids beaucoup plus important du fait de leur faible variance d'erreur. Dans cette zone de l'Atlantique sud-ouest, nous rappelons que les données SEVIRI ne sont pas fournies.

4.3.1.2 Zone particulière 30°N 30°W

Intéressons-nous maintenant à une zone de l'Atlantique Nord, autour de 30°N 30°W, où le variogramme de l'anomalie de SST est cette fois-ci de forme isotropique (cf. figure 4.3a). L'ajustement de ce variogramme empirique se fait par le modèle théorique exponentiel défini en (4.12) toujours par moindre carrés pondérés. Les deux variogrammes correspondant sont superposés sur la figure 4.7a et, au vu des résultats graphiques, ce modèle variographique semble également être adapté au cas particulier isotropique. Les paramètres estimés, synthétisés dans le tableau 4.1, sont les suivants : $\hat{\sigma}^2 = 0.06$, $\hat{L}_{min} = 20$ km, $\hat{L}_{max} = 28$ km et $\hat{\phi} = 118°$ (angle par rapport au nord dans le sens anti-horaire). L'isotropie se traduit par un rapport L_{max}/L_{min} proche de 1 et par un angle d'anisotropie ϕ n'ayant pas beaucoup de sens. La matrice de covariance spatiale stationnaire de l'état $\Sigma^2_{\Delta_{lon},\Delta_{lat}}$, donnée par (4.3), est tracée en figure 4.7b. Celle-ci, pour des pixels proches comme par exemple les valeurs de la diagonale, est de l'ordre de $\hat{\sigma}^2$ alors que pour des distances $\sqrt{\Delta^2_{lon} + \Delta^2_{lat}}$ élevées, elle tend vers 0. Notons que l'estimation du palier σ^2 est de l'ordre de grandeur de celui estimé en 3.3.3.2 (qui vaut 0.07) et d'après 3.3.3.1, $\hat{\lambda} = 0.11$, ce qui correspond à un coefficient AR(1) à un jour de $\hat{M}_1 = 0.90\, j^{-1}$. Ainsi, la corrélation temporelle de l'anomalie de SST est moins marquée que pour l'exemple autour de la zone 49°S 59°W. Enfin, dans cette région, les trois capteurs METOP, SEVIRI et AMSRE (indicés par k et le temps t_i) sont accessibles et leurs covariances d'erreur $R_{t_i}^{(k)}$ médianes repectives sont 0.12, 0.15 et 0.67.

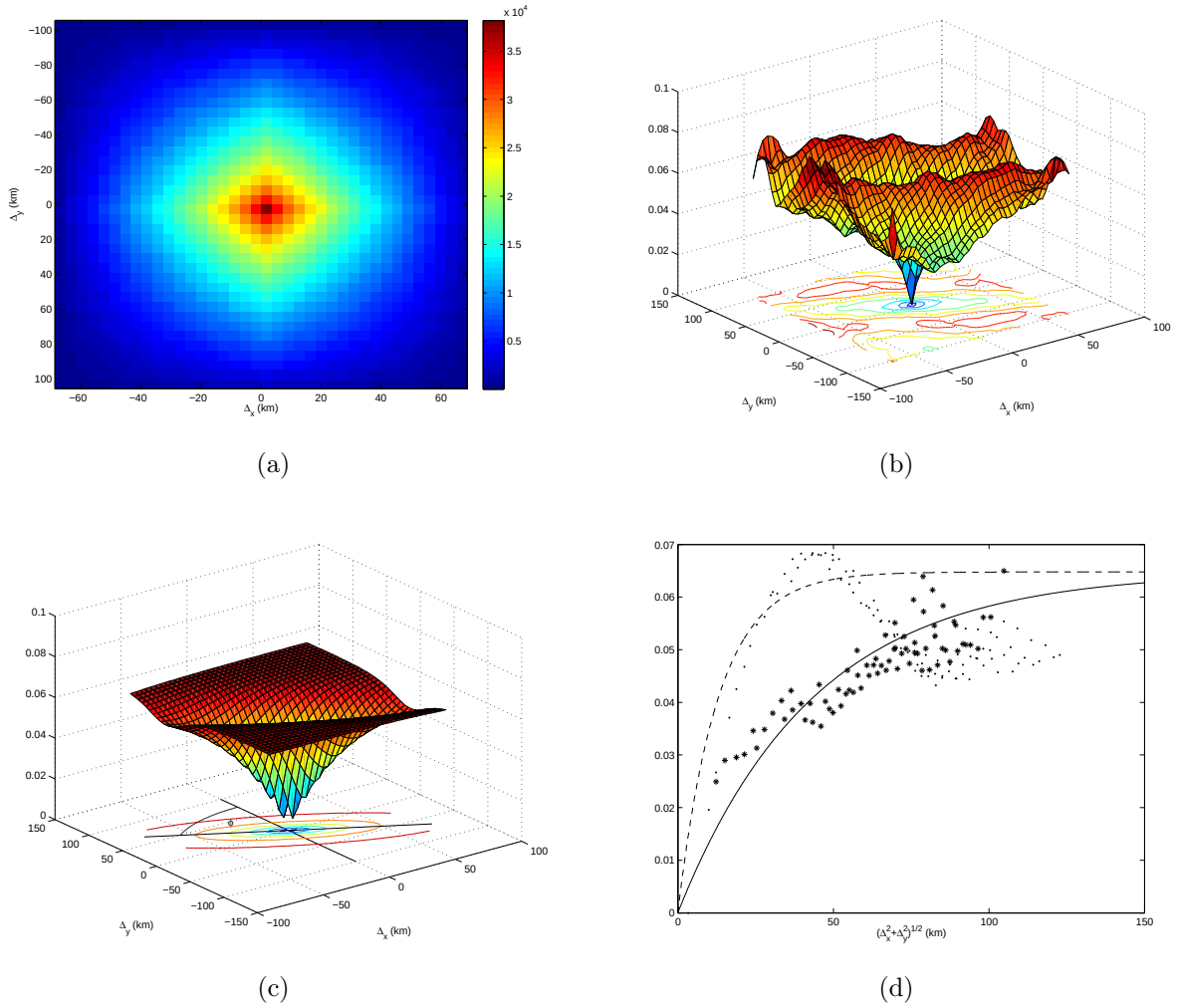
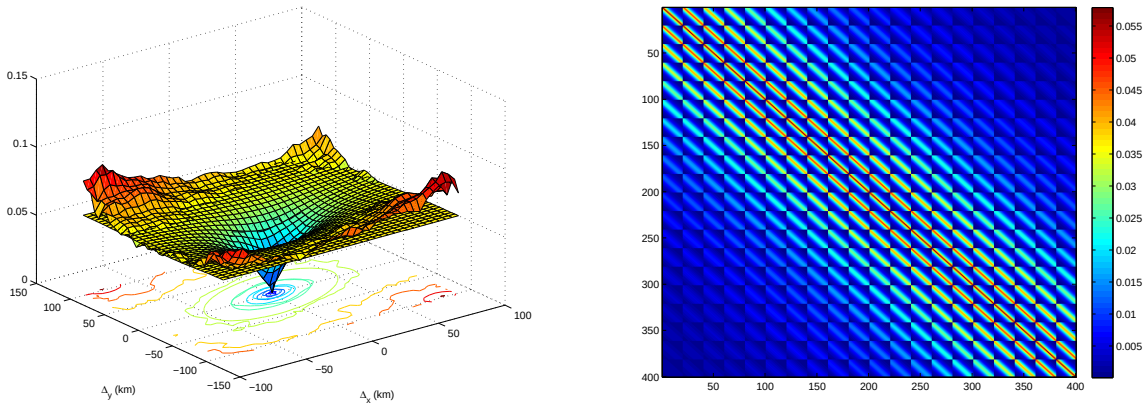


FIG. 4.6: Nombre de couples de points (a) ayant servi à ajuster un variogramme exponentiel anisotropique géométrique (c) à partir du variogramme empirique (b), calculé à partir de données METOP et OIV2 sur une zone 1° × 1° autour du point 49°S 59°W. La direction de plus grande portée de l'ellipse est répérée par l'angle $\hat{\phi} = 49°$ (c) et les variogrammes théoriques et empiriques sont représentés (d) selon $\hat{\phi}$ (trait continu et étoiles) et $\hat{\phi} + 90°$ (trait discontinu et points).



(a) (b)

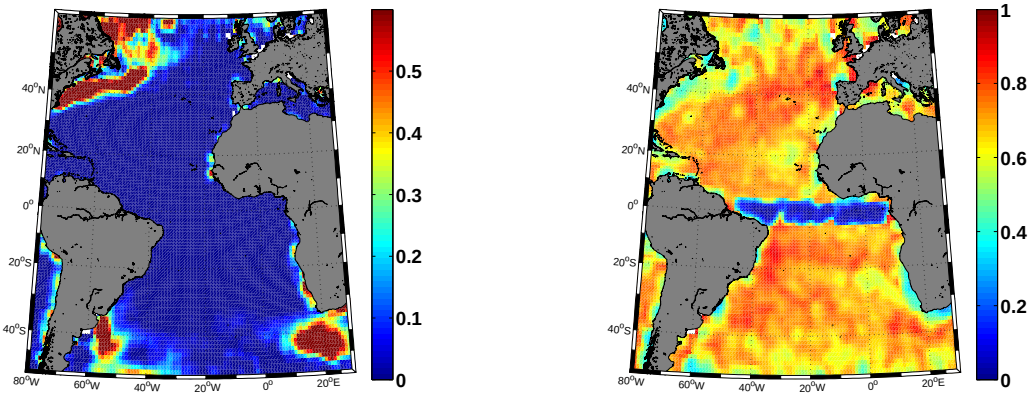
Figure 4.7: (a) : Variogramme empirique et théorique superposés, obtenus sur une zone $1° \times 1°$ autour du point 30°N 30°W à partir de données METOP et OIV2. (b) : Matrice de covariance spatiale correspondante.

	Zone 49°S 59°W	Zone 30°N 30°W
$\hat{\sigma}^2$	0.06	0.06
$\hat{L}_{min}$	13 km	20 km
$\hat{L}_{max}$	43 km	28 km
$\hat{\phi}$	49°	118°

TAB. 4.1: Tableau des résultats des estimations des paramètres de la variabilité spatiale de l'anomalie de SST pour les zones 49°S 59°W et 30°N 30°W.

4.3.1.3 Généralisation sur l'Atlantique

Le modèle variographique présenté en (4.12) est adapté aux formes de variogrammes empiriques couramment observées dans l'Atlantique. Il modélise aussi bien l'anisotropie géométrique comme celle rencontrée en 4.3.1.1 que le cas particulier de l'isotropie où $L_{max} \simeq L_{min}$ en 4.3.1.2. Reste maintenant à généraliser ces résultats à l'Atlantique, toujours à partir des données METOP et OIV2 sur l'année 2008. Pour cela, nous balayons, dans les moyennes latitudes, des zones $1° \times 1°$ pour éviter la superposition de variogrammes. Compte tenu du nombre de couples de points ayant servi à construire les variogrammes empiriques (comme par exemple la figure 4.6a), on peut s'attendre à une estimation robuste des paramètres σ^2, L_{min}, L_{max} et ϕ des variogrammes mais il peut arriver que les estimateurs (principalement L_{max} et σ^2) convergent vers des valeurs aberrantes. Dans ce cas, une estimation *a posteriori* par maximum de vraisemblance est effectuée (cf. 4.2.2.2) où, dans le cas d'une divergence, les résultats ne sont pas pris en compte. Nous avons remarqué que ces cas de non convergence sont généralement la conséquence d'une mauvaise stationnarisation des données. En



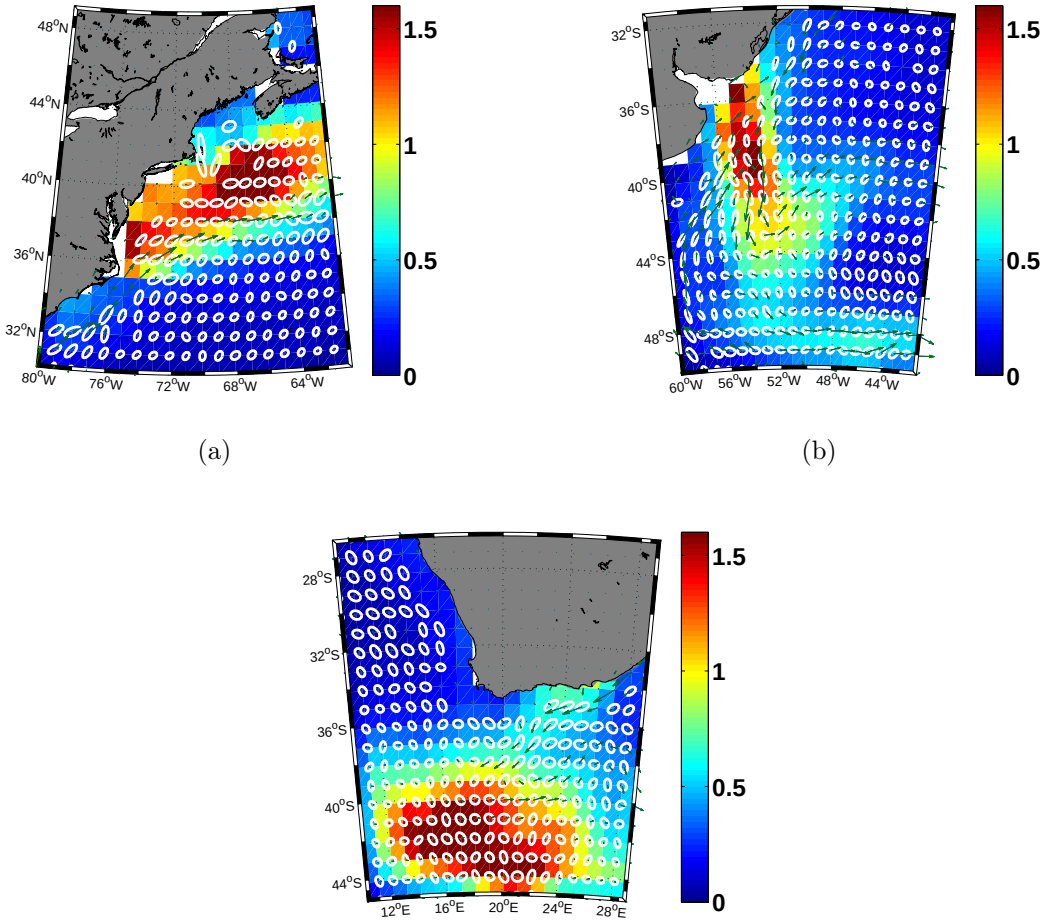
(a) (b)

FIG. 4.8: Distribution spatiale du paramètre $\hat{\sigma}^2$ (a) et du rapport $\hat{L}_{min}/\hat{L}_{max}$ (b) sur l'océan Atlantique. Les résultats sont obtenus à partir de l'ajustement de variogrammes anisotropiques moyens sur l'année 2008, sur des zones $1° \times 1°$, en utilisant des données de nuit METOP.

effet, les données issues de capteurs satellites sont contaminées par la terre le long des côtes (cf. [29] pour plus d'informations). Or, la résolution des observations utilisées par l'analyse OIV2 (servant à stationnariser la SST de METOP) est grossière, de l'ordre de 0.25°. Ainsi, la contamination par la terre s'observe dans les résultats de cette analyse et la stationnarisation des données SST le long des côtes devient alors très discutable. Le résultat de ce phénomène se traduit par des variogrammes empiriques de formes particulières. Ceux-ci ont des paliers différents selon les angles formés par L_{min} et L_{max} (σ^2 n'est pas constant selon toutes les directions) ; on parle alors de variogrammes d'anisotropie zonale, particulièrement inadaptés au modèle de variogramme proposé en (4.12). Donc, le long des côtes, la divergence de l'estimation des paramètres de la variabilité spatiale est souvent observée.

Au final, les résultats des estimations de σ^2 et du coefficient d'anisotropie L_{min}/L_{max} sont respectivement représentés par les figures 4.8a et 4.8b. Nous décidons de nous focaliser sur quelques zones intéressantes de l'Atlantique du fait de leur forte variabilité de l'anomalie de SST (comme le Gulf Stream, les Falkland et les aiguilles donnés respectivement par les figures 4.9a, 4.9b et 4.9c). Sur ces quatre dernières figures, pour représenter les paramètres de la variabilité spatiale, nous projetons les valeurs des $\hat{\sigma}^2$ avec un code couleur et nous traçons, pour chaque boîte d'analyse, l'ellipse correspondant à l'équation (4.10). Notons que cette représentation a déjà été adoptée en océanographie sur des données altimétriques (cf. [35]) : celle-ci permet de visualiser simultanément $\hat{\sigma}^2$, $\hat{L}_{min}$, $\hat{L}_{max}$ et $\hat{\phi}$. Nous discuterons successivement des résultats des estimateurs de σ^2 en 4.3.1.3.1, L_{min} et L_{max} en 4.3.1.3.2 puis ϕ en 4.3.1.3.3.

4.3.1.3.1 Paramètre σ^2 La variance stationnaire de l'anomalie de SST donnée en figure 4.8a, calculée à partir de étude de la variabilité spatiale, est sensiblement proche de celle



(a) (b)

(c)

FIG. 4.9: Visualisation des paramètres $\hat{\sigma}^2$, $\hat{L}_{min}$, $\hat{L}_{max}$ et $\hat{\phi}$ du variogramme exponentiel anisotropique géométrique dans les régions du Gulf Stream (a), les Falkland (b) et les aiguilles (c). Les courants de surface théoriques sont donnés par les flèches vertes. Les résultats sont obtenus à partir de l'ajustement de variogrammes anisotropiques moyens sur l'année 2008, sur des zones $1^\circ \times 1^\circ$, en utilisant des données de nuit METOP.

calculée à partir de l'étude de la variabilité temporelle (cf. figure 3.9c). On peut tout de même accorder plus d'importance à $\hat{\sigma}^2$ issu de l'étude des variogrammes car il se base sur une estimation plus robuste. On identifie clairement les zones de forts courants comme les aiguilles, Falkland ou le Gulf Stream ainsi que que les régions d'upwelling comme le Benguela et les Canaries. Notons, que les courants de surface moyens théoriques, représentés par les flèches vertes sur les figures 4.9a, 4.9b et 4.9c sont fortement liés aux zones de fronts de $\hat{\sigma}^2$; on le voit clairement sur les régions Gulf Stream et Falkland.

4.3.1.3.2 Paramètres L_{min} et L_{max} Intéressons-nous au rapport d'anisotropie $\hat{L}_{min}/\hat{L}_{max} \leq 1$ représenté à la figure 4.8b. Ce rapport est proche de 1 dans les zones où la variabilité de l'anomalie de SST est isotropique. Parmi celles-ci, on retrouve les régions de moyennes latitudes, loin des côtes. Plus $\hat{L}_{min}/\hat{L}_{max}$ tend vers 0 et plus nous avons à faire à de l'anisotropie. Les zones fortement anisotropes sont généralement situées sur des endroits où les courants de surface sont forts ou le long des côtes. Mais, la zone la plus anisotrope reste celle de l'ITCZ. En effet, on y trouve des portées zonales (dans l'axe Est-Ouest) fortes, pouvant aller jusqu'à 1000 km alors que les portées méridionales (dans l'axe Nord-Sud) ne dépassent pas les 500 km. Cette région de convergence des eaux de l'Atlantique Nord et Sud est donc bien particulière car les portées spatiales sont nettement plus fortes que celles observées dans les zones de latitudes moyennes dont la portée est de l'ordre de 30 km en moyenne.

4.3.1.3.3 Paramètre ϕ L'estimation de angle d'anisotropie ϕ (angle par rapport au nord dans le sens anti-horaire) est fortement lié aux courants de surfaces moyens et les figures 4.9a, 4.9b et 4.9c le confirment. En effet, sur celles-ci, nous visualisons par une flèche verte la topographie dynamique moyenne et, dans la grande majorité des cas, les angles $\hat{\phi}$ coincident avec cet indicateur de courant de surface. Cependant, ils ne sont pas en accord dans certains endroits où les courants de surface théoriques changent brutalement de direction comme c'est le cas dans les Falkland. En effet, le fait de construire les variogrammes empiriques sur des zones relativement grandes ($1^\circ \times 1^\circ$) peut conduire à se placer dans des cas où l'on trouve plusieurs régimes de variogrammes. Le résultat sera un variogramme moyen, loin de refléter les conditions réelles de variabilité spatiale.

4.3.2 Validation

Les paramètres du modèle spatio/temporel sont estimés et nous voulons maintenant valider les résultats. Dans un premier temps, en 4.3.2.1, nous présenterons différents outils permettant de qualifier les résultats d'une interpolation spatiale. Enfin, nous appliquerons ces outils aux deux zones étudiées soit 49°S 59°W en 4.3.2.2 (avec des données METOP et AMSRE) et 30°N 30°W en 4.3.2.3 (avec des données METOP, AMSRE et SEVIRI).

4.3.2.1 Moyens de validation

Plusieurs points de vue sont à prendre en compte : le modèle est-il prédictif et la structure spatiale de la SST est-elle conservée ? De plus nous aimerions connaître l'apport du spatial

par rapport au temporel : nous comparerons les résultats du modèle uniquement spatial à celui spatio/temporel.

4.3.2.1.1 Le modèle permet-il de reconstituer des champs réalistes ? Pour répondre à cette question, deux approches sont possibles. La première est d'utiliser des données de référence indépendantes des observations satellitaires comme les données *in situ* (bouées dérivantes ou fixes). Malheureusement, leur échantillonnage spatial est trop faible et nous ne pouvons pas obtenir assez de colocalisations (cf. 1.1.4). La seconde approche consiste à faire de la validation croisée en masquant artificiellement des observations satellites et en comparant celles-ci aux résultats de l'interpolation. Ainsi, pour chaque $i \in \{1, ..., n\}$, les observations, par abus de notation[2] notées y_{t_i}, sont retirées du lissage de Kalman et nous calculons

$$x^{(s)}_{D_r,t_i|i}\left(\hat{\theta}\right) = E\left(X_{D_r,t_i}|y_{t_1}^{t_{i-1}}, y_{t_{i+1}}^{t_n}; \hat{\theta}\right),\ P^{(s)}_{D_r,t_i|i}\left(\hat{\theta}\right) = Var\left(X_{D_r,t_i}|y_{t_1}^{t_{i-1}}, y_{t_{i+1}}^{t_n}; \hat{\theta}\right) \tag{4.14}$$

Nous visualiserons deux indicateurs qualitatifs : les cartes de biais $x^{(s)}_{D_r,t_i|i}\left(\hat{\theta}\right) - y_{t_i}$ et de variance d'erreur du modèle $diag\left(P^{(s)}_{D_r,t_i|i}\left(\hat{\theta}\right)\right)$. Cette dernière, permet d'avoir une carte d'erreur de notre modèle et de calculer des intervalles de confiance autour de $x^{(s)}_{D_r,t_i|i}\left(\hat{\theta}\right)$.

4.3.2.1.2 Quel est l'apport de l'approche spatiale ? Nous pouvons également nous placer en un point s_j précis et estimer $\forall i$, par lissage de Kalman sur une année comme présenté en (4.14), les espérances et variances $x^{(s)}_{s_j,t_i|i}\left(\hat{\theta}\right)$ et $P^{(s)}_{s_j,t_i|i}\left(\hat{\theta}\right)$ conditionnellement à toutes les observations sauf celle du temps t_i. Ensuite, nous confronterons visuellement les interpolations des deux modèles afin d'apprécier l'apport éventuel de l'information spatiale. Une erreur quadratique moyenne sera également calculée au point s_j pour les deux modèles. Celle-ci est définie par :

$$\frac{1}{n}\sum_{i=1}^{n}\left(x^{(s)}_{s_j,t_i|i}\left(\hat{\theta}\right) - y_{s_j,t_i}\right)^2$$

Toujours au point s_j nous calculerons également les résidus standardisés obtenus avec le modèle temporel et spatio/temporel (cf. 3.2.1 et 4.2.1) :

$$\frac{y^{(k)}_{s_j,t_i} - H^{(k)}_{t_i} x^{(s)}_{s_j,t_i|i}\left(\hat{\theta}\right)}{\sqrt{H^{(k)}_{t_i}\left(P^{(s)}_{s_j,t_i|i}\left(\hat{\theta}\right)\right)H^{(k)'}_{t_i} + R^{(k)}_{t_i}}} \tag{4.15}$$

Nous comparerons les distributions des résidus standardisés obtenus par les deux approches. L'idée est de vérifier si les hypothèses de notre modèle sont réalistes.

[2]sans indicer la source satellitaire k ni le domaine spatial de définition $D_{p(i,k)}$

4.3.2.1.3 La structure spatiale de la SST est-elle conservée ? Pour répondre à cette dernière question, nous retenons également deux approches. La première est de comparer la structure spectrale des observations brutes à celle des résultats de l'interpolation pour déterminer la puissance spectrale aux différentes échelles spatiales. Pour cela, nous estimons le spectre de densité de puissance notée S qui, pour chaque longueur d'onde associe une énergie en décibels. En pratique, on estime S à partir de la norme de la transformée de Fourier 2D normalisée de l'image et moyennée selon toutes les directions. De plus, d'après le théorème de Wiener–Khintchine (cf. [27], [30]), pour un processus stationnaire du second ordre, la relation entre le spectre de densité de puissance S et la fonction variogramme γ est donnée par :

$$S(f,\theta) = \int_{-\infty}^{\infty} \frac{\sigma^2 - \gamma(\Delta_{lon}, \Delta_{lat})}{\sigma^2} \exp(-2\pi i f L_\theta) \, \mathrm{d}L_\theta$$

avec f la fréquence étudiée et L_θ la portée selon l'angle θ.

La seconde idée consiste à étudier les gradients des images afin de voir si les structures spatiales de SST (fronts, tourbillons, etc...) sont conservées après interpolation. Pour cela, définissons une image $Z(x,y)$ de coordonnées x et y. Le gradient de Z est donné par :

$$\nabla Z = \begin{cases} \frac{\partial Z(x,y)}{\partial x} \\ \frac{\partial Z(x,y)}{\partial y} \end{cases}$$

En pratique, on approxime les dérivées partielles de Z selon x et y par différences finies. Au final, le Gradient ∇Z est caractérisé par son module et sa direction respectivement donnés par :

$$|\nabla Z| = \sqrt{\left(\frac{\partial Z(x,y)}{\partial x}\right)^2 + \left(\frac{\partial Z(x,y)}{\partial y}\right)^2} \tag{4.16}$$

$$\arg(\nabla Z) = \arctan\left(\frac{\partial Z(x,y)}{\partial y} / \frac{\partial Z(x,y)}{\partial x}\right) \tag{4.17}$$

L'idée sera de comparer ces modules et directions de gradients des observations brutes à ceux des champs interpolés par filtre de Kalman.

4.3.2.2 Zone particulière 49°S 59°W

Nous commençons par étudier la structure spatiale de l'anomalie de SST en comparant une observation de l'instrument METOP à un champ interpolé par filtre de Kalman dans la nuit du 17 décembre 2008 dans cette zone du courant Falkland. Les données utilisées pour l'interpolation sont celles de METOP et AMSRE sur une année (excepté l'image METOP de la nuit du 17 décembre 2008) et les paramètres du modèle ont été estimés à partir de METOP uniquement. Nous avons choisi d'étudier cette date précise pour plusieurs raisons :

- il s'agit d'une observation METOP (haute résolution spatiale)
- tous les pixels sont renseignés ce qui permettra, par validation croisée, de comparer l'observation aux résultats de l'interpolation

– la structure spatiale de l'anomalie de SST est intéressante avec notamment, entre les latitudes 47°S et 48.5°S, des langues et des fronts de SST marqués (cf. figure 4.10a).

On commence par étudier les modules et arguments des gradients définis respectivement en (4.16) et (4.17). On remarque notamment une zone de forts gradients au Nord-Est et au Nord-Ouest de la figure 4.10c. Après interpolation par filtre de Kalman (cf. figure 4.10b), les modules des gradients sont très estompés comme nous montre la figure 4.10d. Cependant, les arguments des gradients sont respectés dans l'ensemble. Notons également que l'angle d'anisotropie $\hat{\phi} = 49°$ par rapport au Nord, qui se caractérise comme le courant moyen de surface, se retrouve clairement sur ces deux figures. Enfin, les distributions des modules des gradients obtenus sur les observations (METOP et AMSRE) et les champs interpolés au cours de l'année 2008 sont données en figure 4.10f : ils montrent une diminution notable de la valeur des normes des gradients après interpolation. Du fait de la covariance spatiale $\Sigma^2_{\Delta_{lon},\Delta_{lat}}$, le filtre de Kalman a donc lissé les structures petites échelles visibles sur les observations satellitaires. Un autre moyen de s'en rendre compte est de comparer les spectres de densité de puissance donnés en figure 4.10e pour voir la répartition de l'énergie selon les différentes échelles. Effectivement, on observe une perte d'énergie de l'image interpolée par rapport à l'image brute METOP, caractérisée par des pentes spectrales, obtenues par régression linéaire simple, respectivement d'environ $k = -3$ et $k = -2.85$. Les comportements des deux spectres diffèrent à partir d'une résolution spatiale de 100 km soit environ 1°.

Nous discutons maintenant de la validation croisée toujours à cette date du 17 décembre 2008. Pour cela, nous visualisons les cartes de biais et de variance d'erreur du modèle à cette date précise. Concernant le biais présenté en figure 4.11a, on observe une sous-estimation forte d'une langue d'anomalie de SST dans la zone Nord-Est de l'image ainsi que des surestimations dans les zones de fronts. Pour ce qui est de la variance d'estimation du modèle présenté en 4.11b, on remarque plusieurs choses. Premièrement, la variance aux limites du domaine D_r est plus forte que dans le reste de l'image : on parle d'effets de bords, bien connus en géostatistique. En effet, les points interpolés aux bords du domaine D_r disposent de moins de voisins que ceux qui se trouvent au centre ; leur variance d'estimation est donc importante. Deuxièmement, on identifie une variance plus forte dans la direction de $\hat{\phi} = 49°$ qui correspond à la portée maximale du variogramme théorique. Troisièmement, dans cette région, le courant est orienté moyennement dans le sens Sud-Est vers Nord-Ouest. Ainsi, la variance d'erreur d'interpolation est plus forte au Sud : ceci est dû au fait que plus on se déplace vers le Nord et plus l'anomalie de SST aura eu le temps d'être assimilée faisant ainsi réduire la variance d'erreur d'interpolation. Une autre explication pourrait être que le modèle de variogramme utilisé dans tout le domaine D_r n'est pas adapté à cette région particulière.

Nous comparons ensuite les résultats des interpolations du modèle temporel et spatio/temporel. Pour cela, on se place en un point précis s_j de coordonnées 49°N 59°W et nous estimons l'état et sa variance par validation croisée, comme nous l'expliquons en (4.14), sur une année de données de nuit METOP et AMSRE (on rappelle que SEVIRI ne fournit pas de données dans cette partie de l'Atlantique). Au total, on comptabilise $n = 236$ observations sur ce point précis. Le modèle temporel défini en 3.2.1, initialement défini avec une seule équation d'observation, est généralisé pour prendre en compte les k sources d'observations satellitaires $y^{(k)}_{s_j,t_i}$, $\forall i$ ainsi que les variances d'erreurs de mesures $R^{(k)}_{t_i}$ correspondantes. La

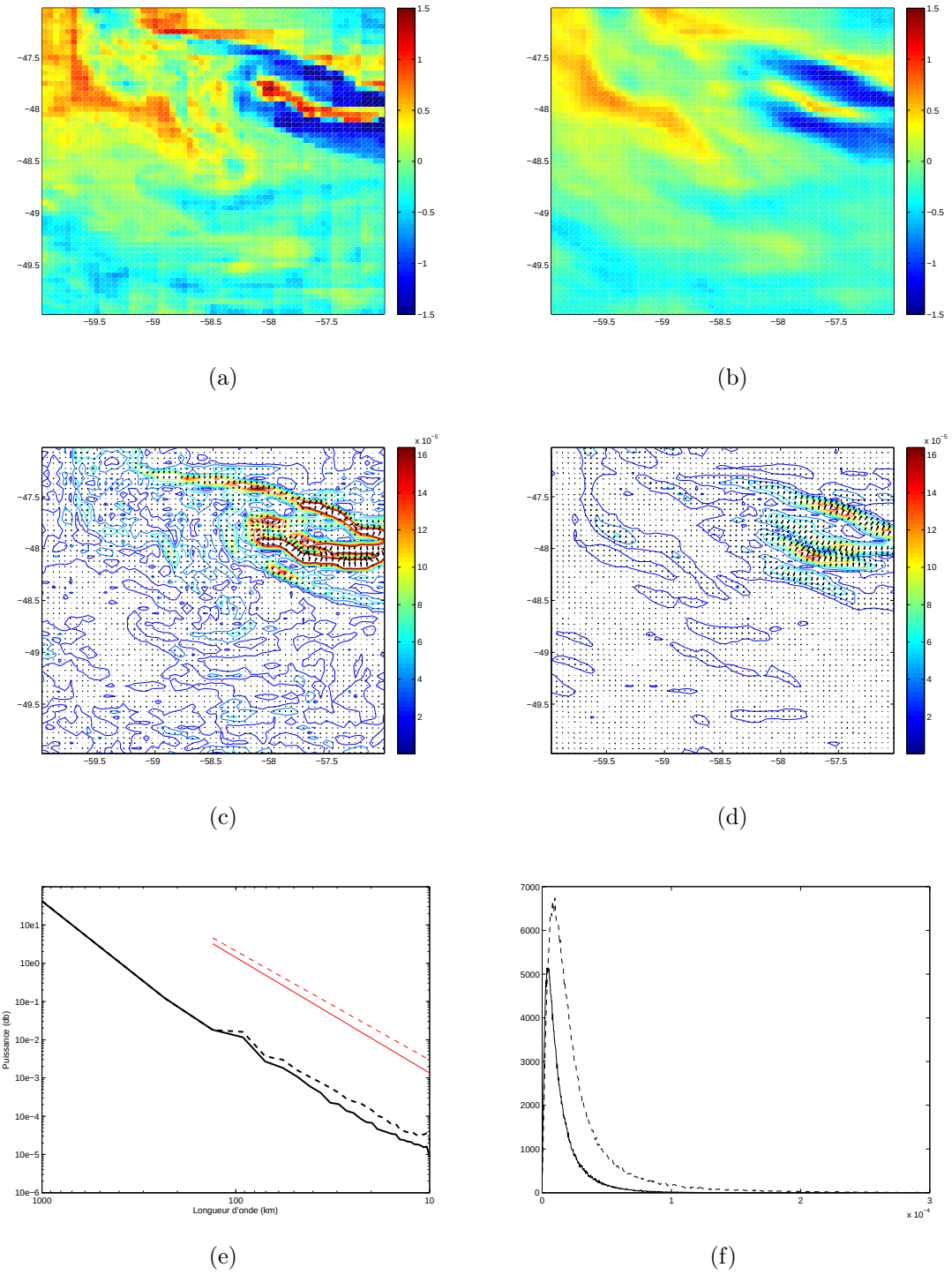
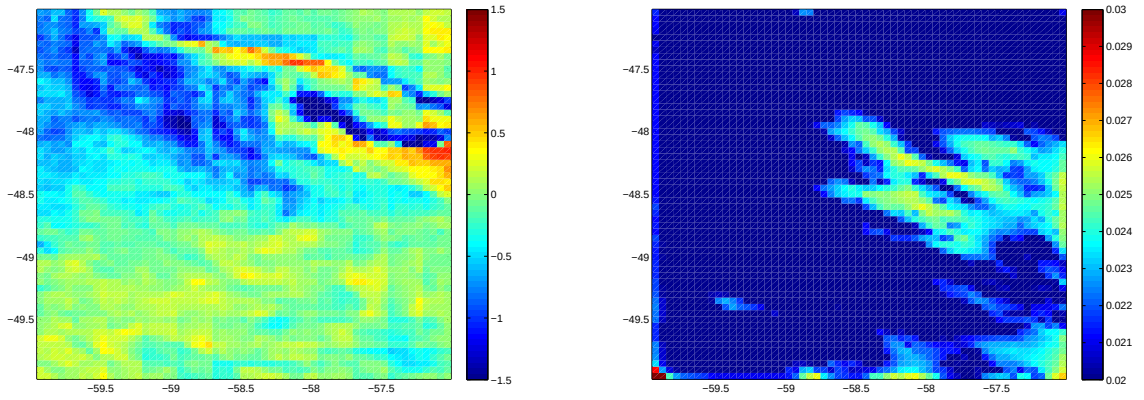


Figure 4.10: Anomalie de SST observée (a) par METOP et interpolée par filtre de Kalman (b), sans prendre en compte (a), dans la nuit du 17 décembre 2008 dans la zone $3° \times 3°$ autour du point 49°S 59°W. Les modules et directions des gradients respectifs sont données en (c) et (d). (e) : Spectres de densité de puissance de l'anomalie de SST observée (trait discontinu) et interpolée (trait continu) le même soir ainsi que les pentes des spectres respectives (en rouge). (f) : Histogrammes des modules des gradients cumulés sur l'année 2008 des observations METOP et AMSRE (trait discontinu) ainsi que ceux des champs interpolés par filtre de Kalman (trait continu) sur l'année 2008.

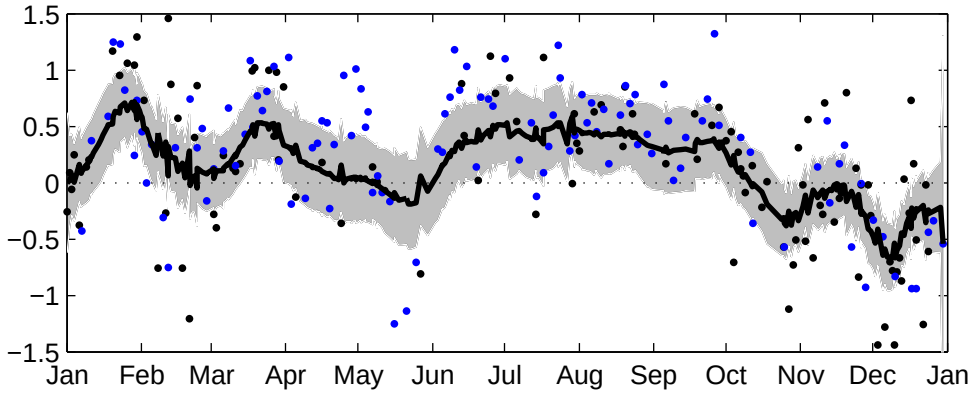


(a) (b)

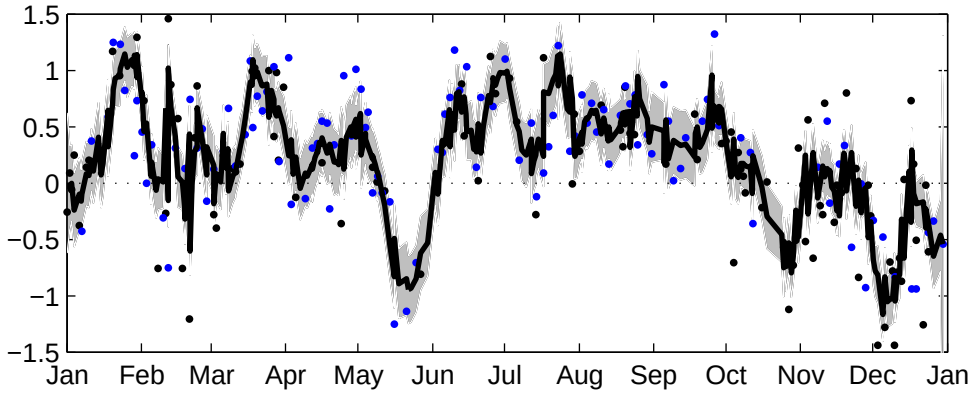
FIG. 4.11: Carte de biais (a) et de variance (b) d'estimation de l'anomalie de SST dans la nuit du 17 décembre 2008 dans la zone 49°S 59°W.

seule différence entre les deux modèles est la prise en compte de la covariance spatiale de l'état $\Sigma^2_{\Delta_{lon},\Delta_{lat}}$ (4.13) dans le modèle spatio/temporel. Les paramètres Θ utilisés sont ceux présentés en 4.3.1.1. Les résultats sont présentés dans les figures 4.12a et 4.12b. L'apport de l'information spatiale est constaté lors de ruptures de l'anomalie de SST comme en fin mai ou début décembre : à ces périodes, l'information seulement temporelle ne permet pas de suivre ce genre de phénomènes. De plus, nous remarquons une diminution des intervalles de confiance à 95% (partie grisée) montrant une diminution de la variance d'estimation après utilisation de l'information spatiale. Cette constatation est également faite par la figure 4.12c qui trace la différence entre les variances d'estimation des deux modèles. La diminution de la variance est marquée au mois de mai, période où on observe une forte rupture de l'anomalie de SST, bien prise en compte par l'apport de l'information spatiale. Enfin, d'après le tableau 4.2, l'erreur quadratique moyenne passe de 0.22 pour le modèle temporel à 0.17 pour le modèle spatio/temporel. Ceci prouve que la prise en compte de la dimension spatiale améliore significativement le modèle.

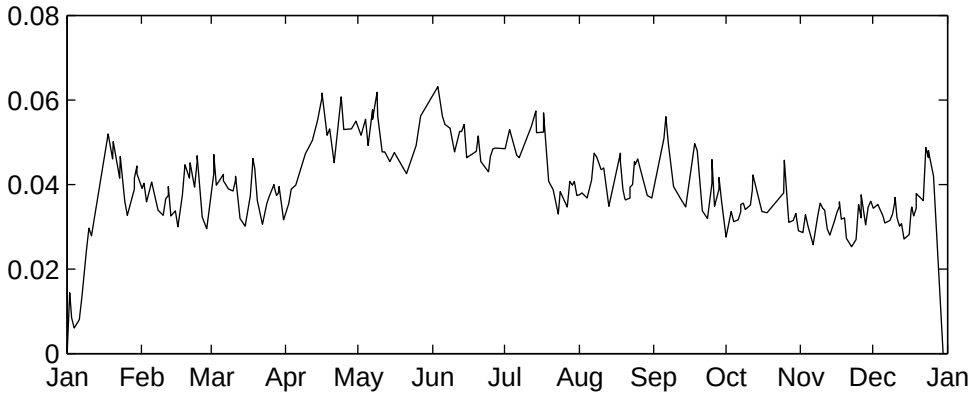
Par contre, en comparant la forme du QQ-plot Normal des résidus standardisés donnés dans l'équation (4.15) du modèle temporel à celui du modèle spatio/temporel (cf. figures 4.13a et 4.13b), on s'aperçoit que peu de choses évoluent à l'ajout de l'information spatiale. Les queues de distribution positives et négatives restent aussi éloignées de la droite d'Henry dans les deux cas. On peut donc conclure que le modèle présenté dans ce chapitre, utilisant l'information des pixels voisins, ne semble toujours pas prendre en compte une dynamique non-linéaire de l'état caché.



(a)

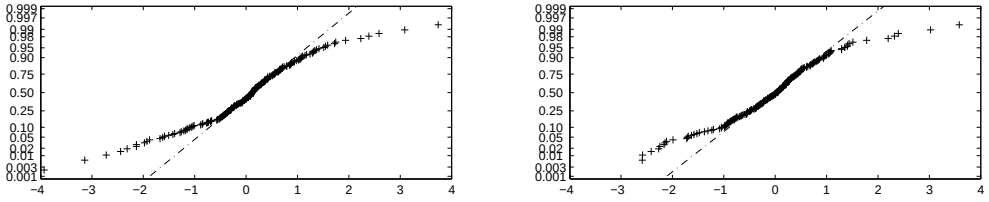


(b)



(c)

FIG. 4.12: Anomalies de SST en °C observées par les capteurs METOP (points noirs) et AMSRE (points bleus) et interpolées par lissage de Kalman avec l'intervalle de confiance à 95% (gris) pour le modèle temporel (a) et spatio/temporel (b). (c) : Différence entre les variances d'analyses $P^{(s)}_{s_j,t_i|i}$ du modèle temporel et spatio/temporel. Les résultats sont obtenus par validation croisée sur l'année 2008 dans la zone 3° × 3° autour du point 49°N 59°W.



(a) (b)

Figure 4.13: QQ-plot normal des résidus standardisés du modèle temporel (a) et spatio/temporel (b). Les résultats sont obtenus par validation croisée sur l'année 2008 dans la zone 3° × 3° autour du point 49°N 59°W.

	Point 49°S 59°W	Point 30°N 30°W
Modèle temporel	0.22	0.08
Modèle spatio/temporel	0.17	0.07

TAB. 4.2: Erreurs quadratiques moyennes calculées à partir du modèle temporel et spatio/temporel, aux points 49°S 59°W et 30°N 30°W, sur une année d'observations.

4.3.2.3 Zone particulière 30°N 30°W

Nous reprenons l'étude autour du point 30°N 30°W. Comme pour le point 49°S 59°W, on choisit un temps t_i pour lesquel on dispose d'une observation METOP quasi-complète (dans cette zone de l'Atlantique nord, les pixels non référencés des capteurs infr-rouges sont fréquents). Il s'agit de la nuit du 22 mai 2008 et la figure 4.14a donne l'observation du capteur METOP cette nuit là. On remarque que l'anomalie de SST est moins structurée que pour l'étude précédente. En effet, nous nous trouvons dans une région de l'Atlantique peu dynamique. Il semble également que cette observation satellitaire soit plus bruitée que celle donnée en figure 4.10a. Le résultat du filtre de Kalman est présenté en figure 4.14b. Les données utilisées pour l'interpolation sont celles de METOP, AMSRE et SEVIRI sur une année (excepté l'image METOP de la nuit du 22 mai 2008) et les paramètres du modèle ont été estimés à partir de METOP uniquement. La qualité de l'interpolation est d'abord discutée à partir des gradients. Ceux-ci, déterminés par différences finis, ne sont pas calculés sur les parties du champ METOP où les pixels ne sont pas référencés. D'après les figures 4.14c et 4.14d correspondant respectivement au champ observé METOP et à son interpolation, les forts modules et arguments des gradients sont bien restitués dans les parties nord et ouest du domaine spatial D_r. Mais, comme pour la zone d'étude précédente, les distributions des modules des gradients obtenus sur les observations (METOP, AMSRE et SEVIRI) et les champs interpolés au cours de l'année 2008 montrent une diminution de leur intensité après application du filtre de Kalman (cf. figure 4.14f). Par contre, l'analyse spectrale (cf. figure 4.14e) réalisée sur les images 4.14a et 4.14b ne montre pas de réelles pertes d'énergies dans les petites échelles spatiales. On peut donc supposer que, contrairement à la zone d'étude turbulente du courant des Malouines (autour de 49°S59°W), l'interpolation spatio/temporelle

est, en terme de restitution des échelles caractéristiques de l'anomalie de SST, adaptée la zone autour du point 30°N 30°W.

Nous discutons maintenant de la validation croisée toujours à cette date du 22 mai 2008. Pour cela, nous visualisons les cartes de biais (cf. figure 4.15a) et de variance d'erreur (cf. figure 4.15b) du modèle calculées à partir de (4.14). Le biais est surtout marqué dans la partie nord de l'image qui semble être une zone dynamique. Pour ce qui est de la variance d'erreur d'interpolation, elle est forte au bord du domaine D_r. Par contre, sachant que l'échelle des couleurs est la même que celle la figure 4.11b, on s'aperçoit que l'erreur commise dans cette zone 30°N 30°W est moins importante que dans la région autour de 49°S 59°W.

Ensuite, nous cherchons à comparer les résultats des deux modèles en un point précis s_j de coordonnées 30°N 30°W dans le but de mettre en évidence un apport de l'information spatiale. Pour cela, nous utilisons les mêmes observations et mêmes valeurs des paramètres (cf. 4.3.1.2) dans les deux modèles. On comptabilise $n = 375$ vecteurs d'observations au point s_j pour des données de nuit sur l'année 2008. Nous estimons les espérances et variances du lisseur de Kalman par validation croisée selon (4.14), seulement au point s_j. Les résultats des interpolations du modèle temporel et spatio/temporel sont respectivement donnés en figures 4.16a et 4.16b. Globalement, les variations des $x^{(s)}_{s_j,t_i|i}$ des deux modèles vont dans le même sens. Seules les interpolations du modèle spatio/temporel tendent à prendre en compte les anomalies de SST atypiques et voici quelques indicateurs montrant cette pertinence :

- en début mars et début aout, le capteur METOP mesure une anomalie positive de SST de l'ordre de 1°C et seul le modèle spatio/temporel, qui dispose de l'information des points voisins avant et après cet instant, prédit une anomalie fortement positive
- en mi-mai et mi-juin, les capteurs SEVIRI et AMSRE constatent une anomalie négative et, encore une fois, l'information spatiale permet d'approcher ces observations
- la taille des intervalles de confiance à 95% donnent également une information sur la qualité des estimations de l'état lissé et c'est toujours le modèle spatio/temporel qui a l'estimation la plus précise, indiquant une variance $P^{(s)}_{s_j,t_i|i}$ plus faible. Ceci est confirmé par la figure 4.16c qui trace la différence entre les variances d'estimations $P^{(s)}_{s_j,t_i|i}$ du modèle temporel et spatio/temporel au cours de la période de validation croisée. On y voit, grâce à l'apport de l'information spatiale, une réduction de la variance de l'état pouvant aller jusqu'à 0.06.

Enfin, les erreurs quadratiques moyennes au point 30°N 30°W (cf. tableau 4.2) montrent une amélioration entre le modèle temporel (0.08) et le modèle spatio/temporel (0.07).

Pour finir, le QQ-plot Normal des résidus standardisés du modèle spatio/temporel (cf. figures 4.17b) montre que les résidus fortement négatifs et positifs sont toujours éloignés de la droite d'Henry. Comme pour l'étude au point 49°S 59°W, les hypothèses de linéarité de l'état semblent donc inadaptées.

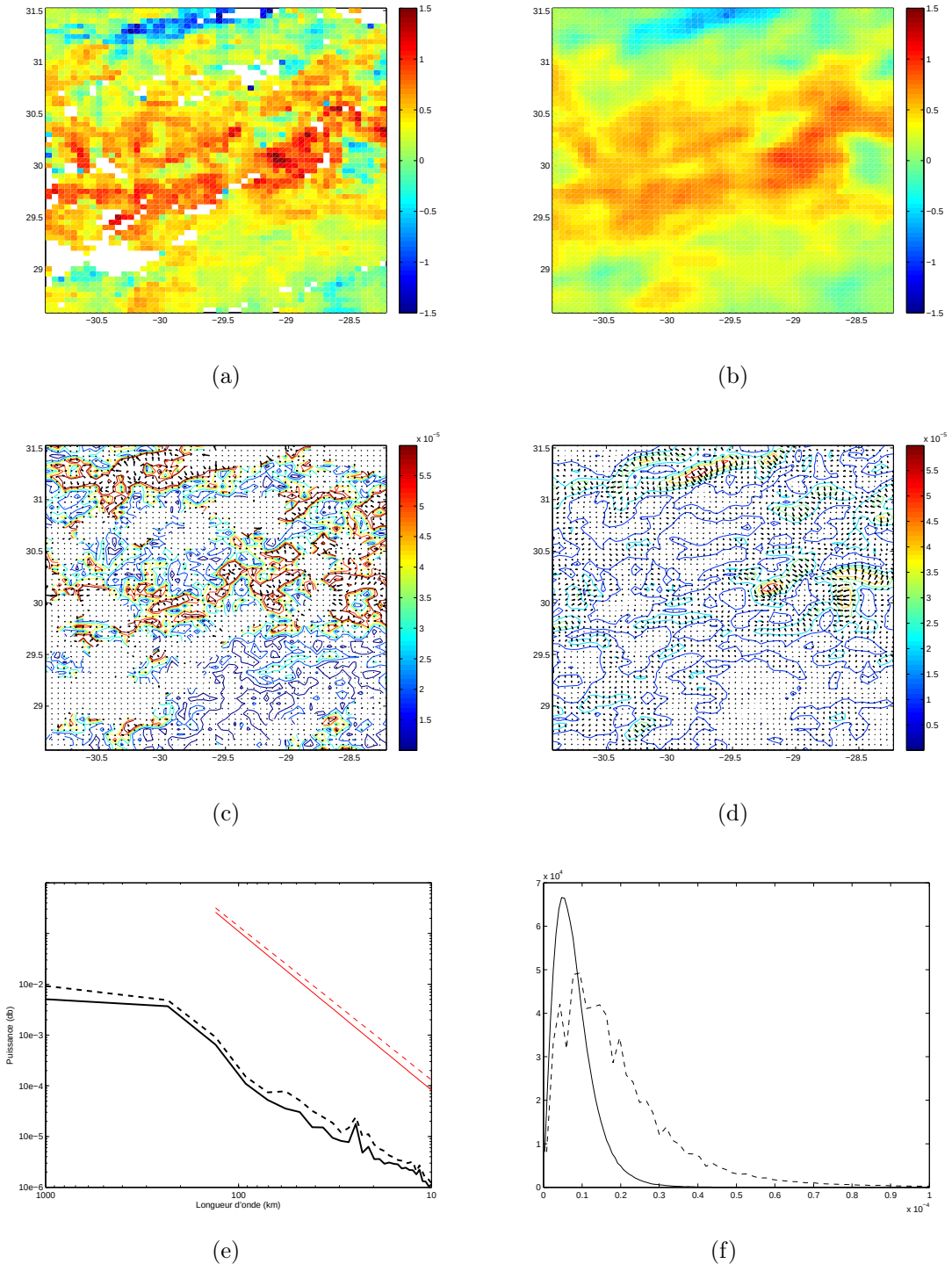
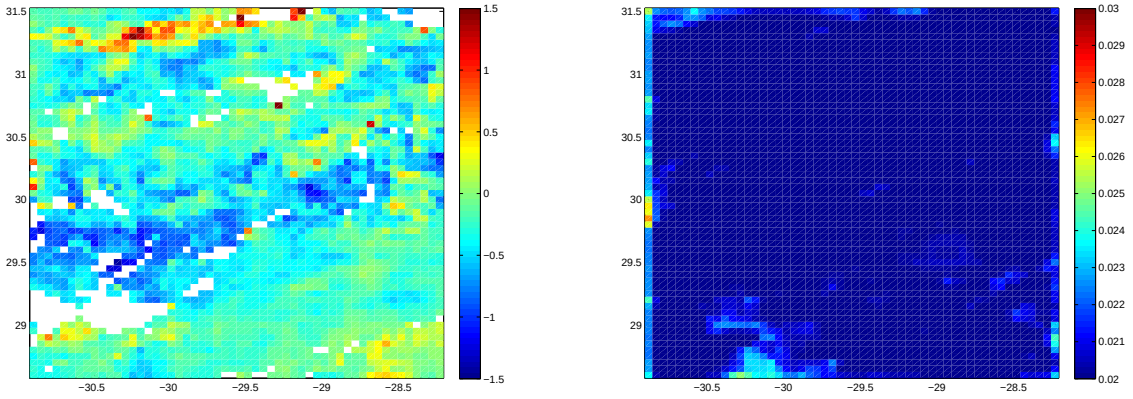


Figure 4.14: Anomalie de SST observée (a) par METOP et interpolée par filtre de Kalman (b), sans prendre en compte (a), dans la nuit du 22 mai 2008 dans la zone $3° \times 3°$ autour du point 30°N 30°W. Les modules et directions des gradients respectifs sont données en (c) et (d). (e) : Spectres de densité de puissance de l'anomalie de SST observée (trait discontinu) et interpolée (trait continu) le même soir ainsi que les pentes des spectres respectives (en rouge). (f) : Histogrammes des modules des gradients cumulés sur l'année 2008 des observations METOP, AMSRE et SEVIRI (trait discontinu) ainsi que ceux des champs interpolés par filtre de Kalman (trait continu) sur l'année 2008.



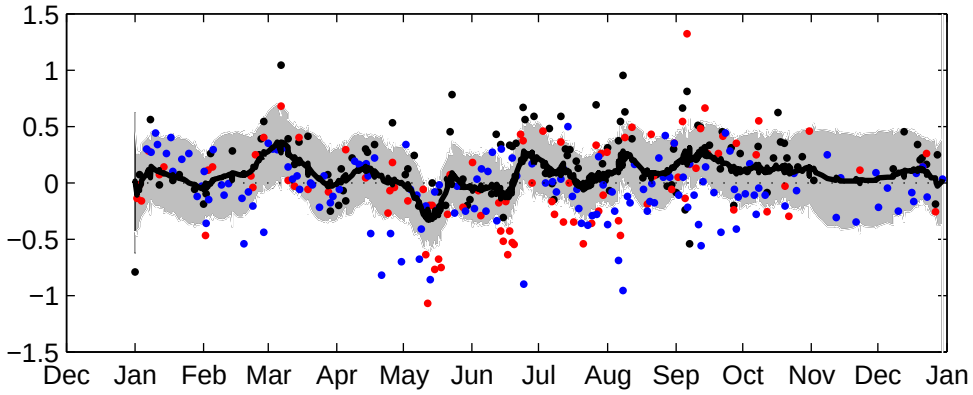
(a) (b)

FIG. 4.15: Carte de biais (a) et de variance (b) d'estimation de l'anomalie de SST dans la nuit du 22 mai 2008 dans la zone 30°N 30°W.

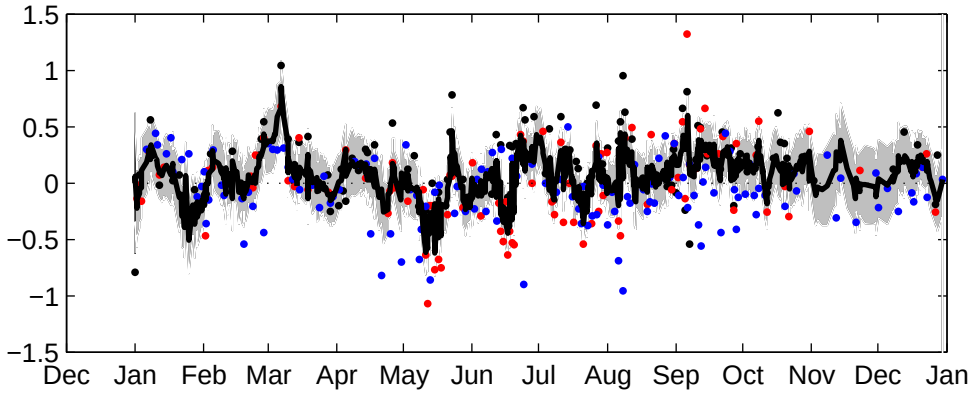
4.4 Discussion

Dans ce chapitre, nous avons généralisé l'équation d'observation de notre modèle espace d'état afin de prendre en compte plusieurs sources satellitaires, avec une pondération par la variance d'erreur de chacun des pixels assimilés. Mais l'avancée majeure réside dans l'ajout de la dimension spatiale à notre modèle qui était jusque là ponctuel. La covariance d'erreur de notre variable d'état devint alors séparable en espace et en temps. Un effort paticulier à été mis en oeuvre pour estimer les paramètres de variabilité spatiale de l'anomalie de SST et pour rendre le modèle le plus explicatif possible. Comme pour le chapitre précédent, les estimations sont faites sur l'océan Atlantique cette fois-ci à partir d'outils classiques de géostatistiques : les variogrammes. Ceux-ci ont mis en évidence une anistropie dans certaines zones de cet océan. Nous avons alors chercher à modéliser cette différence de variabilité selon l'angle d'étude et, étant donné le nombre important de couples de points utilisés pour estimer les variogrammes empiriques, les paramètres de la variabilité spatiale ont été estimés de façon robuste. Ces résultats, basés sur des données haute résolution du capteur infra-rouge METOP (qui devient une référence en terme de SST) constituent un atlas de la variabilité spatiale de la SST dans l'Atlantique, utile pour les océanographes. De plus, la méthode d'estimation des paramètres des variogrammes peut être facilement généralisée à d'autres variables océaniques observées par satellites telles que la vitesse du vent, la couleur de l'eau, la salinité, etc... Il faut néanmoins, comme c'est le cas pour la SST, disposer d'assez d'observations pour caractériser l'anisotropie de la variabilité.

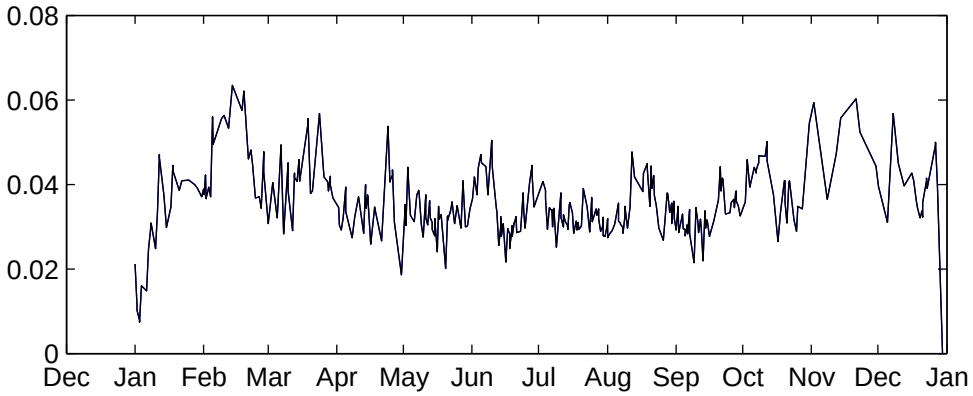
Une fois la partie explicative terminée, nous avons discuté de la partie prédictive du modèle. Or, la validation des résultats d'interpolation de variables telles que la SST est une question importante pour les applications océanographiques. Ceci est du aux différents aspects qui sont à prendre en compte : la physique de la variable est-elle conservée, quelles



(a)

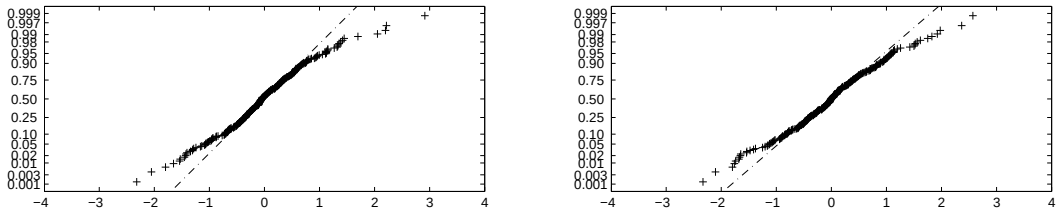


(b)



(c)

FIG. 4.16: Anomalies de SST en °C observées par les capteurs METOP (points noirs), SEVIRI (points rouges) et AMSRE (points bleus) et interpolées par lissage de Kalman avec l'intervalle de confiance à 95% (gris) pour le modèle temporel (a) et spatio/temporel (b). (c) : Différence entre les variances d'analyses $P^{(s)}_{s_j,t_i|i}$ du modèle temporel et spatio/temporel. Les résultats sont obtenus par validation croisée sur l'année 2008 dans la zone 3° × 3° autour du point 30°N 30°W.



(a) (b)

Figure 4.17: QQ-plot normal des résidus standardisés du modèle temporel (a) et spatio/temporel (b). Les résultats sont obtenus par validation croisée sur l'année 2008 dans la zone $3° \times 3°$ autour du point 30°N 30°W.

sont les échelles que le modèle arrive à restituer ou encore le modèle est-il, en moyenne, un bon prédicteur ? Nous avons cherché à prendre en compte les deux approches de validation. Pour tester notre modèle, nous avons choisi deux zones d'études : l'une, au large des îles Falkland, où l'on observe des forts courants de surface entrainant une variabilité spatiale anisotropique et l'autre, au milieu de l'Atlantique Nord, où la variabilité de l'anomalie de SST est isotropique. Les indicateurs visuels (gradients, cartes de biais et de variances d'estimation) montrent que les erreurs commises par le modèle spatio/temporel sont surtout concentrées dans les zones turbulentes comme c'est le cas dans la région autour du point 49°S 59°W. L'interpolation a tendance à lisser les petites échelles comme nous montrent le spectre et l'histogramme du module des gradients dans cette zone. Par contre, autour de 30°N 30°W, où la dynamique de l'anomalie est moins marquée, les résultats graphiques du modèle spatio/temporel sont meilleurs. Lorsque nous comparons ensuite le modèle temporel au modèle spatio/temporel en un point précis, par validation croisée, on constate que la composante spatiale apporte une information pertinente du fait de la diminution de l'erreur quadratique moyenne sur l'année 2008. De plus, la variance d'estimation diminue largement. Par contre, l'étude des résidus standardisés a, de nouveau, soulevé une problématique. En effet, il semblerait que l'approche spatio/temporelle ne parvient pas à modéliser les résidus fortements positifs ou négatifs. On pointe du doigt les limites du modèle linéaire, qui, bien qu'il restitue correctement les variations moyennes de notre variable d'intérêt, peine à modéliser les variations brusques de l'anomalie de SST. Ceci constitue une piste d'étude possible pour de futurs travaux dont voici une présentation.

4.5 Perspectives

Dans cette section, nous allons présenter les différentes idées que nous pensons intéressantes à développer. Quelques-unes d'entres elles ont été développées partiellement mais nous ne disposons pas d'assez de résultats pour en faire en chapitre complet. Les innovations proposées concernent principalement la paramétrisation des matrices du modèle espace d'état. Pour ce qui est de M et Q, nous présenterons en 4.5.1, la prise en compte de déplacement

horizontaux de l'anomalie de SST. Enfin, en 4.5.2, nous proposerons des paramétrisations de H et R.

4.5.1 Modélisation spatio/temporelle et dynamique de la SST

Pour prendre en compte une partie de cette dynamique non linéaire de l'anomalie de SST, nous utilisons le fait que la température de surface des océans est soumise à des contraintes extérieures qui lui font changer d'état. Jusque là, les modèles présentés dans cette thèse (cf. 3.2.1 et 4.2.1) ne prenaient pas en compte les contraintes suivantes :

- interaction verticale entre l'atmosphère et la surface des océans. Par exemple, lors de la formation de tempêtes tropicales, on observe des échanges de températures entre les eaux chaudes (dans les premiers mètres de profondeur) et l'atmosphériques provoquant ainsi un refroidissement de la SST
- transports horizontaux de chaleur à la surface par des courants de surface. Ces transports sont expliqués par la mécanique des fluides et deux phénomènes : l'advection (le fait qu'une particule soit déplacée par un champ de vitesse) et la diffusion (le fait qu'une particule ait tendance à s'étendre)
- interaction verticale entre la surface des océans et les eaux plus profondes. Par exemple, en été, un fort coup de vent peut mélanger les eaux de surface chaudes aux eaux plus profondes et ainsi faire chuter la SST.

Les interactions verticales, que ce soient entre l'atmosphère ou des eaux plus profondes, sont complexes et nécessitent l'information de nombreuses covariables. Les transports horizontaux de chaleur, quant à eux, peuvent directement être déterminés par la température de surface des océans. C'est donc sur cette composante que nous allons insister. Un modèle intégrant les transports horizontaux de l'anomalie de SST a été initialement proposée par [19], [37], [26] puis [38]. En reprenant les notations des chapitres précédents, l'équation de transport s'écrit :

$$dX_{D_r,t} = -U\nabla X_{D_r,t}dt + D\nabla^2 X_{D_r,t}dt - \lambda X_{D_r,t}dt + \Gamma dW_{D_r,t} \tag{4.18}$$

où $\{W_{D_r,t}\}$ représente un bruit Brownien, λ la dynamique temporelle de la SST et $\Gamma > 0$ est la variabilité venant des fluctuations de temps. Les déplacement horizontaux de chaleurs sont pris en compte par les termes d'advection U et de diffusion D. Notons qu'une version plus complète de (4.18) fut proposée plus tard par [40] : dans celle-ci, les échanges verticaux de chaleur entre la surface et les eaux plus profondes complètent l'équation d'état.

D'après [44] (chapitre 8), la solution de l'équation différentielle stochastique de l'état (4.18), dans le cas où la diffusion D est isotropique et l'advection U (décomposée en une composante zonale u et méridionale v), peut-être obtenue par différences finies et est exacte si les coefficients sont fixes dans le temps. Ainsi, la solution de (4.18) est donnée par $\{X_{D_r,t_i}\}$, un processus Markovien qui vérifie l'équation linéaire d'état, $\forall\, i \in \{2, ..., n\}$,

$$X_{D_r,t_i} = M_{\Delta_{lon},\Delta_{lat},\Delta_{t_i}} X_{D_r,t_{i-1}} + \left(Q_{\Delta_{lon},\Delta_{lat},\Delta_{t_i}}\right)^{1/2} \eta_{D_r,t_i} \tag{4.19}$$

avec $Q_{\Delta_{lon},\Delta_{lat},\Delta_{t_i}} = \Sigma^2_{\Delta_{lon},\Delta_{lat}} - M_{\Delta_{lon},\Delta_{lat},\Delta_{t_i}} \Sigma^2_{\Delta_{lon},\Delta_{lat}} M'_{\Delta_{lon},\Delta_{lat},\Delta_{t_i}}$ et $\{\eta_{D_r,t_i}\}_{i\in\{1,...,n\}}$ un bruit blanc Gaussien d'espérance nulle et de variance unité, indépendant, pour tout k, de

$X_{s_j-\Delta_{lon}+\Delta_{lat},t_i}$	$X_{s_j+\Delta_{lat},t_i}$	$X_{s_j+\Delta_{lon}+\Delta_{lat},t_i}$
$X_{s_j-\Delta_{lon},t_i}$	X_{s_j,t_i}	$X_{s_j+\Delta_{lon},t_i}$
$X_{s_j-\Delta_{lon}-\Delta_{lat},t_i}$	$X_{s_j-\Delta_{lat},t_i}$	$X_{s_j+\Delta_{lon}-\Delta_{lat},t_i}$

FIG. 4.18: Définition des pixels voisins du point d'intérêt X_{s_j,t_i}.

$\left\{\varepsilon^{(k)}_{D_{p(i,k)},t_i}\right\}_{i\in\{1,\dots,n\}}$. Les termes advectifs u et v et le terme diffusif D sont pris en compte dans la matrice opérateur d'état $M_{\Delta_{lon},\Delta_{lat},\Delta_{t_i}}$. Avant de décrire celle-ci, introduisons quelques notations permettant de prendre en compte l'information des huits pixels voisins d'un point donné X_{s_j,t_i} (cf. figure 4.18). Un lien possible entre X_{s_j,t_i} et l'état autour des huit pixels voisins du site s_j, au temps t_{i-1}, a été proposé par [44]. Ce lien, déterminé par moindres carrées, se traduit par la matrice AR(1) à un jour, $M_{\Delta_{lon},\Delta_{lat},\Delta_{t_i}} \in \mathbb{R}^{r\times r}$, déterminée de la façon suivante :

$$
\begin{aligned}
M_{0,0,\Delta_{t_i}} &= \exp\left(-\lambda\Delta_{t_i}\right)\alpha_s \\
M_{0,\Delta_{lat},\Delta_{t_i}} &= \exp\left(-\lambda\Delta_{t_i}\right)\alpha_{s+\Delta_{lat}} \\
M_{\Delta_{lon},0,\Delta_{t_i}} &= \exp\left(-\lambda\Delta_{t_i}\right)\alpha_{s+\Delta_{lon}} \\
M_{0,-\Delta_{lat},\Delta_{t_i}} &= \exp\left(-\lambda\Delta_{t_i}\right)\alpha_{s-\Delta_{lat}} \\
M_{-\Delta_{lon},0,\Delta_{t_i}} &= \exp\left(-\lambda\Delta_{t_i}\right)\alpha_{s-\Delta_{lon}} \\
M_{-\Delta_{lon},\Delta_{lat},\Delta_{t_i}} &= \exp\left(-\lambda\Delta_{t_i}\right)\alpha_{s-\Delta_{lon}+\Delta_{lat}} \\
M_{\Delta_{lon},\Delta_{lat},\Delta_{t_i}} &= \exp\left(-\lambda\Delta_{t_i}\right)\alpha_{s+\Delta_{lon}+\Delta_{lat}} \\
M_{\Delta_{lon},-\Delta_{lat},\Delta_{t_i}} &= \exp\left(-\lambda\Delta_{t_i}\right)\alpha_{s+\Delta_{lon}-\Delta_{lat}} \\
M_{-\Delta_{lon},-\Delta_{lat},\Delta_{t_i}} &= \exp\left(-\lambda\Delta_{t_i}\right)\alpha_{s-\Delta_{lon}-\Delta_{lat}}
\end{aligned}
$$

Pour des distances $|\Delta_{lon}|$ et $|\Delta_{lat}|$ supérieures à 0.05° (résolution spatiale du domaine D_r), les éléments de la matrice M seront nuls.

Les neuf coefficients α sont donnés ici :

$$
\begin{aligned}
\alpha_s &= \frac{5}{9} - \frac{1}{3}\left(\frac{u^2\Delta_{t_i}^2}{\Delta_{lon}^2} + \frac{v^2\Delta_{t_i}^2}{\Delta_{lat}^2} + \frac{2D\Delta_{t_i}}{\Delta_{lon}^2} + \frac{2D\Delta_{t_i}}{\Delta_{lat}^2}\right) \\
\alpha_{s+\Delta_{lat}} &= \frac{2}{9} - \frac{v\Delta_{t_i}}{6\Delta_{lat}} - \frac{u^2\Delta_{t_i}^2 + 2D\Delta_{t_i}}{3\Delta_{lon}^2} + \frac{v^2\Delta_{t_i}^2 + 2D\Delta_{t_i}}{6\Delta_{lat}^2} \\
\alpha_{s+\Delta_{lon}} &= \frac{2}{9} - \frac{u\Delta_{t_i}}{6\Delta_{lon}} - \frac{v^2\Delta_{t_i}^2 + 2D\Delta_{t_i}}{3\Delta_{lat}^2} + \frac{u^2\Delta_{t_i}^2 + 2D\Delta_{t_i}}{6\Delta_{lon}^2} \\
\alpha_{s-\Delta_{lat}} &= \frac{2}{9} + \frac{v\Delta_{t_i}}{6\Delta_{lat}} - \frac{u^2\Delta_{t_i}^2 + 2D\Delta_{t_i}}{3\Delta_{lon}^2} + \frac{v^2\Delta_{t_i}^2 + 2D\Delta_{t_i}}{6\Delta_{lat}^2} \\
\alpha_{s-\Delta_{lon}} &= \frac{2}{9} + \frac{u\Delta_{t_i}}{6\Delta_{lon}} - \frac{v^2\Delta_{t_i}^2 + 2D\Delta_{t_i}}{3\Delta_{lat}^2} + \frac{u^2\Delta_{t_i}^2 + 2D\Delta_{t_i}}{6\Delta_{lon}^2} \\
\alpha_{s-\Delta_{lon}+\Delta_{lat}} &= -\frac{1}{9} + \frac{u\Delta_{t_i}}{6\Delta_{lon}} - \frac{v\Delta_{t_i}}{6\Delta_{lat}} + \frac{1}{6}\left(\frac{u^2\Delta_{t_i}^2}{\Delta_{lon}^2} + \frac{v^2\Delta_{t_i}^2}{\Delta_{lat}^2} + \frac{2D\Delta_{t_i}}{\Delta_{lon}^2} + \frac{2D\Delta_{t_i}}{\Delta_{lat}^2}\right) - \frac{uv\Delta_{t_i}^2}{4\Delta_{lon}\Delta_{lat}} \\
\alpha_{s+\Delta_{lon}+\Delta_{lat}} &= -\frac{1}{9} - \frac{u\Delta_{t_i}}{6\Delta_{lon}} - \frac{v\Delta_{t_i}}{6\Delta_{lat}} + \frac{1}{6}\left(\frac{u^2\Delta_{t_i}^2}{\Delta_{lon}^2} + \frac{v^2\Delta_{t_i}^2}{\Delta_{lat}^2} + \frac{2D\Delta_{t_i}}{\Delta_{lon}^2} + \frac{2D\Delta_{t_i}}{\Delta_{lat}^2}\right) + \frac{uv\Delta_{t_i}^2}{4\Delta_{lon}\Delta_{lat}} \\
\alpha_{s+\Delta_{lon}-\Delta_{lat}} &= -\frac{1}{9} - \frac{u\Delta_{t_i}}{6\Delta_{lon}} + \frac{v\Delta_{t_i}}{6\Delta_{lat}} + \frac{1}{6}\left(\frac{u^2\Delta_{t_i}^2}{\Delta_{lon}^2} + \frac{v^2\Delta_{t_i}^2}{\Delta_{lat}^2} + \frac{2D\Delta_{t_i}}{\Delta_{lon}^2} + \frac{2D\Delta_{t_i}}{\Delta_{lat}^2}\right) - \frac{uv\Delta_{t_i}^2}{4\Delta_{lon}\Delta_{lat}} \\
\alpha_{s-\Delta_{lon}-\Delta_{lat}} &= -\frac{1}{9} + \frac{u\Delta_{t_i}}{6\Delta_{lon}} + \frac{v\Delta_{t_i}}{6\Delta_{lat}} + \frac{1}{6}\left(\frac{u^2\Delta_{t_i}^2}{\Delta_{lon}^2} + \frac{v^2\Delta_{t_i}^2}{\Delta_{lat}^2} + \frac{2D\Delta_{t_i}}{\Delta_{lon}^2} + \frac{2D\Delta_{t_i}}{\Delta_{lat}^2}\right) + \frac{uv\Delta_{t_i}^2}{4\Delta_{lon}\Delta_{lat}}
\end{aligned}
$$

où $\Delta_{lon} = \Delta_{lat} = 0.05°$ dans notre cas de figure et $\Delta t_i = t_i - t_{i-1}$, l'intervalle de temps entre deux observations successives. Cette discrétisation en espace, qui n'est pas la seule possible, est simple et facilement interprétable. Cependant, elle reste vraie dans certaines conditions que voici :

$$
\begin{aligned}
|u|\,\Delta_{t_i} &< \Delta_{lon} \\
|v|\,\Delta_{t_i} &< \Delta_{lat} \\
\sqrt{D\Delta_{t_i}} &< \Delta_{lon} \\
\sqrt{D\Delta_{t_i}} &< \Delta_{lat}
\end{aligned}
$$

Ainsi, la différence avec le modèle spatio/temporel présenté en 4.2.1, réside dans la paramétrisation de $M_{\Delta_{lon},\Delta_{lat},\Delta_{t_i}}$. En effet, dans le cas du modèle avec advection et diffusion comme ici, cette matrice autorégressive prend en compte les distances spatiales et temporelles Δ_{lon}, Δ_{lat} et Δ_{t_i}. Notons que plusieurs auteurs géostatisticiens utilisent ces matrices prenant en compte l'équation de transport d'une variable : on trouve nottament les travaux de [56] (seulement la diffusion) et ceux de [59] (advection et diffusion). Dans les deux cas cités, l'espace est discrétisé dans une seule dimension alors que dans notre cas, on différencie les déplacements en longitude et latitude. Ils abordent également le problème des conditions aux bornes du domaine de définition D_r. En effet, lorsqu'une advection intervient dans un modèle, il est nécessaire d'avoir, en dehors de la grille d'analyse, des valeurs limites à propager. Ces valeurs d'anomalies de SST seront fixées à 0 dans notre modèle (comme ce qui

est fait dans les deux articles précédemment cités). On peut donc s'attendre à des effets de bords assez conséquents. Il est donc nécessaire de travailler sur des zones D_r importantes.

Pour conclure, la principale difficulté de cette prise en compte du transport de l'anomalie de SST est, dans notre cas, la linéarisation (par écriture matricielle) des termes advectifs et de la diffusifs. De plus, cette linéarisation n'est possible que dans un cas bien particulier, lorsque la diffusion (isotropique) et l'advection sont constantes en espace et en temps. Dans notre cas de figure, la variable étudiée évolue lentement et ces hypothèses ne sont pas restrictives. En outre, étant donné que nous restons dans un cadre linéaire à erreurs Gaussiennes, les filtres et lisseurs de Kalman sont optimaux et permettent d'évaluer la vraisemblance du modèle. Ainsi, il est théoriquement possible d'appliquer des méthodes d'estimations basées sur la maximisation de la vraisemblance (ce qui est fait par [39]) pour estimer les paramètres d'avection zonale u et méridionale v ainsi que la diffusion D ; le modèle (4.19) et (4.1) constitue donc une piste intéressante à étudier pour la suite. Un récapitulatif des méthodes traitant de la modélisation stochastique de l'anomalie de SST est disponible dans [44]. Cependant, dans le cas de variables ayant une dynamique non linéaire forte (comme par exemple la vitesse du vent et la direction du vent), cette linéarisation de l'équation différentielle stochastique (4.18) est trop simpliste. On atteint alors les limites de l'approche linéaire stochastique et un moyen de résoudre des modèles non linéaires consiste à utiliser une approche variationnelle (cf. 2.2.1.2) ou des méthodes de type Kalman d'ensemble (cf. [16]).

4.5.2 Autres pistes d'étude

La section précédente concernait la paramétrisation des matrices M et Q de l'équation d'état du modèle espace d'état. Nous proposons maintenant une discussion autour des matrices H et R de l'équation d'observation. En effet, lors de cette thèse, nous avons décidé de transformer les observations AMSRE et SEVIRI afin quelles aient la même résolution spatiale que le capteur METOP soit 0.05°. Cela s'est fait par interpolation par plus proches voisins comme nous le montre la figure 1.5, ne changeant en rien les champs d'observations des capteurs AMSRE et SEVIRI. De plus, nous avons fixé notre grille d'interpolation à 0.05°. Par conséquent, H, la matrice de changement de grille entre l'état et les observations était constante et fixée à la matrice identité. Cependant, nous aurions pu traiter les observations avec leurs résolutions spatiales d'origine et décomposer la modélisation de l'anomalie de SST selon plusieurs échelles : la grande avec AMSRE, la moyenne avec SEVIRI et la fine avec METOP. De plus, grâce à des décompositions comme les ondelettes (convolution d'une image par un filtre de type fenêtre glissante), nous aurions utiliser les contextes[3] des observations METOP et SEVIRI pour, par exemple, apporter une contribution à la modélisation de la grande échelle de l'anomalie de SST. Ainsi, H aurait pu s'écrire comme une matrice de filtrage spatial.

Concernant la modélisation de la matrice de covariance d'erreur d'observation R, nous avons suivi le schéma classiquement utilisés en assimilation de données satellitaires, en la fixant diagonale et en utilisant les variances d'erreurs proposées par les fournisseurs de don-

[3]résultats des différentes convolutions (les contextes ont des résolutions de plus en plus grossières)

nées. Ces valeurs dépendent d'un indiçage de qualité des pixels : par exemple, pour un capteur infra-rouge, les données aux bords d'un nuage sont supposées plus variables. Le nombre d'indices qualité est, par convention, fixé à cinq et il y a autant de variances possibles. Celles-ci sont déterminées à partir de colocalisations avec des données *in situ* (considérées comme la SST de référence) et n'évoluent plus une fois fixées. Or, ne peut-on pas supposer une dérive possible du capteur au cours du temps ? Est-il judicieux de penser que la variance d'erreur d'un capteur est la même au cours d'une année alors que les conditions athmosphériques changent ? Pour répondre à ces questions, une première approche aurait été de laisser les éléments de la matrice évoluer librement et de les estimer au fur et à mesure ; la méthode d'estimation par maximum de vraisemblance, l'algorithme EM "en ligne", permet de répondre à ces attentes. Une seconde approche aurait consisté à prendre en compte des covariables telles que la vitesse du vent ou le taux d'aérosols contenus dans l'atmosphère, connues pour perturber l'estimation de la SST des capteurs. Une étude de ce type sur le capteur AATSR (non utilisé dans cette thèse) a fait l'objet d'une publication. Celle-ci est donnée en annexe A.

Conclusion

En océanographie, la température de surface des océans (SST) est une variable importante. Elle intervient dans de nombreuses applications comme l'évolution climatique ou la modélisation de la circulation océanique. Nous nous sommes intéressés à l'anomalie de cette température, différence entre des observations satellitaires (haute résolution) et un signal moyen (basse résolution). La présentation de notre variable d'intérêt ainsi que les problématiques associées à son estimation ont été le sujet du chapitre 1.

Les anomalies de SST, supposées stationnaires dans le temps et l'espace, ont fait l'objet d'un ajustement par un modèle espace d'état. Etant donné la dimension du vecteur d'état ainsi que le volume et la fréquence des données à traiter quotidiennement, le modèle linéaire à erreurs Gaussiennes a été retenu. En effet, ce modèle nécessite moins de ressources informatiques et de temps de calcul que des versions non-linéaires ou non-Gaussiennes qui obligent le passage par des méthodes de type filtrage particulaire ou Kalman d'ensemble. L'inférence statistique du modèle espace d'état linéaire et Gaussien a été entièrement décrite dans le chapitre 2. Nous avons commencé par étudier l'estimation de la variable d'état conditionnellement aux observations et aux paramètres du modèle. Le filtre et le lisseur de Kalman ont été présentés en détail. Nous avons décidé de ne pas aborder les méthodes variationnelles, équivalentes aux approches stochastiques dans ce cas de figure linéaire et Gaussien. En effet, ces méthodes variationnelles ne permettent pas d'estimer la variance de l'état ce qui aurait compromis l'estimation des paramètres du modèle. Dans notre thèse, à partir des densités de probabilité obtenues par le lisseur de Kalman, nous avons pu utiliser une méthode d'estimation des paramètres par maximum de vraisemblance : l'algorithme Expectation Maximisation (EM).

Une fois l'étude théorique réalisée autour du modèle espace d'état linéaire et Gaussien, nous avons décidé d'étudier la variabilité de l'anomalie de SST par étape, en ajoutant des composantes au fur et à mesure. La première dimension prise en compte fut temporelle et présentée dans le chapitre 3. Une série temporelle, en un point fixe de l'Atlantique (sans utiliser l'information des points voisins) et issue du capteur satellite infra-rouge METOP, fut l'objet de notre étude. Compte tenu du caractère irrégulier des données rendues par ce

capteur, nous avons commencé par généraliser le modèle espace d'état au temps continu. Après une présentation du filtre et du lisseur de Kalman dans ce cas particulier, nous avons proposé une méthode d'estimation des paramètres temporels et d'erreurs du modèle par un couplage de la méthode des moments avec celle du maximum de vraisemblance présentée dans le chapitre 2. Les résultats obtenus sur des séries temporelles simulées montrent que la méthode converge vers les vrais paramètres du modèle. Nous avons appliqué cette méthodologie à la série temporelle issue du capteur METOP. Une fois les paramètres estimés, nous avons comparé les résultats de l'interpolation à ceux des données *in situ* (considérées comme la SST de référence). Nous avons également procédé à une validation croisée. Les résultats indiquent que nous restituons correctement le signal d'anomalie de SST mais que les hypothèses de linéarité du modèle semblent parfois inadaptées. Enfin, nous avons généralisé le schéma d'estimation des paramètres à tout l'océan Atlantique. Nous avons mis en évidence des zones où la variabilité de l'anomalie est forte (comme les upwellings) ou des zones de fortes variances d'erreurs du capteur METOP (induites par les poussières du Sahara).

Le chapitre 4 était consacré à l'ajout de deux composantes au modèle temporel présenté en chapitre 3 : la dimension spatiale et la généralisation de l'équation d'observation. Ainsi, nous avons pris en compte deux autres capteurs satellitaires : SEVIRI et AMSRE. Nous avons commencé par étudier la structure d'ordre 2 de notre nouveau modèle spatio/temporel. Celle-ci a montré l'intérêt d'estimer la covariance d'erreur spatiale stationnaire de l'état. pour ce faire, nous avons utilisé les données METOP. Nous avons disposé d'un bon échantillonnage spatial permettant l'utilisation d'un outil classique en géostatistique : le variogramme. L'étude de la variabilité spatiale en plusieurs zones de l'Atlantique a montré une anisotropie (comportement de la variabilité différent suivant les angles étudiés) dans les zones de courants de surface forts. Par contre, dans d'autres régions moins turbulentes, les variogrammes empiriques sont isotropiques. Nous avons alors construit un variogramme théorique, ayant suffisamment de paramètres pour prendre en compte une anisotropie géométrique. Nous avons généralisé l'estimation de la variabilité spatiale de la SST à tout l'Atlantique, sur des boites $1° \times 1°$. Ce variogramme théorique a permis une estimation robuste des paramètres. Ensuite, nous nous sommes penchés sur deux zones particulières avec différents moyens de validation : certains indicateurs sont visuels et basés sur l'étude des gradients ou des spectres, alors que les autres, plus classiques en statistique, servent à vérifier les hypothèses du modèle ou à comparer les interpolations du modèle temporel à celles du modèle spatio/temporel. Les résultats montrent que dans une zone turbulente où le courant de surface est fort et où des structures petites échelles apparaissent, l'interpolation spatio/temporelle a tendance à lisser ces structures. Par contre, dans une zone plus calme, l'interpolation conserve les échelles caractéristiques de l'anomalie de SST. Dans les deux régions d'étude, l'apport de l'information spatiale a fait diminuer l'erreur quadratique moyenne. Cependant, nous arrivons à la même conclusion que celle exposée au chapitre 3 : l'étude des résidus standardisés montre que les hypothèses du modèle ne sont pas respectées. En effet, nous nous interrogeons particulièrement sur la linéarité du modèle espace d'état. Cela introduit des perspectives pour de nouveaux travaux de recherche.

Parmi les pistes à prendre en compte, nous privilégions la modélisation des déplacements horizontaux de l'anomalie de SST. Ces déplacements sont régis par une advection et une

diffusion. Nous avons cherché à linéariser ces opérateurs (un paramètre fixe pour la diffusion et deux paramètres fixes pour l'advection) afin qu'ils puissent être intégrés à notre modèle espace d'état linéaire et Gaussien spatio/temporel présenté en chapitre 4. L'objectif serait d'estimer ces trois nouveaux paramètres. Cependant, dans des régions fortement turbulentes, nous pouvons nous attendre à observer des dynamiques non-linéaires plus marquées, nécessitant le passage à des méthodes variationnelles. D'autres pistes sont aussi à explorer. Elles concernent la matrice de changement de grille entre l'état et les observations, l'estimation "en ligne" des variances d'erreurs des différents capteurs satellites ou encore l'utilisation de covariables.

Sur un plan plus personnel, ces travaux de recherche m'ont permis d'aborder des domaines variés. A l'IFREMER, dans le domaine océanographique, les échanges professionnels ont été constants pendant trois ans sur la partie opérationnelle et sur le traitement des données volumineuses. En qualité de statisticien, l'étude m'a permis de collaborer avec des chercheurs en mathématiques appliquées et en Physique (mécanique des fluides). Pour IFREMER, le bénéfice de cette thèse est double. Au niveau explicatif, le comportement de la variabilité spatiale et temporelle de la SST est disponible grâce à l'estimation de ces différents paramètres sur l'océan Atlantique. Au niveau prédictif, ils disposent d'un modèle pertinent permettant d'améliorer les résultats de l'analyse ODYSSEA qui fournit des cartes quotidiennes de SST. Enfin, ces méthodes d'inférence statistique sont applicables à d'autres variables étudiées dans le laboratoire comme par exemple la salinité des océans ou les vents à la surface des océans.

Ces travaux de recherche ont été l'occasion de présentations orales dans plusieurs rassemblements scientifiques internationaux comme celui du GHRSST[4] à Perros-Guirec en juin 2008 ou l'EGU[5] à Vienne en avril 2009. Deux articles, consultables en annexes A et B, ont été accepté et un d'entre eux est publié dans la revue "Geoscience and Remote Sensing Letter, IEEE". Une troisième publication, faisant le bilan de ce travail de thèse, est en cours de rédaction.

[4]Group for High-Resolution SST
[5]European Geosciences Union

BIBLIOGRAPHIE

[1] E. Autret and JF Piollé. Deliverable D2. 2.10 Implementation of a global SST analysis WP 02 Task 2.2. 2007.

[2] L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The Annals of Mathematical Statistics*, 41(1) :164–171, 1970.

[3] Albert Benveniste, Pierre Priouret, and Michel Métivier. *Adaptive algorithms and stochastic approximations*. Springer-Verlag New York, Inc., New York, NY, USA, 1990.

[4] L. Bertino, G. Evensen, and H. Wackernagel. Sequential Data Assimilation Techniques in Oceanography. *International Statistical Review*, 71 :223–241, 2003.

[5] D.B. Chelton, R.A. Deszoeke, M.G. Schlax, K. El Naggar, and N. Siwertz. Geographical variability of the first baroclinic Rossby radius of deformation. *Journal of Physical Oceanography*, 28 :433–460, 1998.

[6] G. K. Corlett and D. J. Poulter. An Updated Confidence Flagging Scheme for Medspiration AATSR L2P Files. Technical report, Medspiration, 2008.

[7] N.A.C. Cressie. *Statistics for spatial data*. John Wiley & Sons, New York, 1993.

[8] J. Cummings, L. Bertino, P. Brasseur, I. Fukumori, M. Kamachi, M.J. Martin, K. Mogensen, P. Oke, C.E. Testut, J. Verron, et al. Ocean data assimilation systems for GODAE. *Oceanography*, 22(3) :96–109, 2009.

[9] P.J. Curran and P.M. Atkinson. Geostatistics and remote sensing. *Progress in Physical Geography*, 22(1) :61, 1998.

[10] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1) :1–38, 1977.

[11] L. Deng and X. Shen. Maximum likelihood in statistical estimation of dynamic systems : Decomposition algorithm and simulation results. *Signal Processing*, 57(1) :65 – 79, 1997.

[12] J. Derber and A. Rosati. A Global Oceanic Data Assimilation System. *Journal of Physical Oceanography*, 19(9) :1333–1347, 1989.

[13] V. Digalakis, J. R. Rohlicek, and M. Ostendorf. Ml estimation of a stochastic linear system with the em algorithm and its application to speech recognition. 1(4) :431–442, Oct. 1993.

[14] C. Donlon, K. Casey, C. Gentemann, P. LeBorgne, I. Robinson, R. Reynolds, C. Merchant, D. Llewellyn-Jones, P. Minnett, J. Piolle, et al. Successes and Challenges for the Modern Sea Surface Temperature Observing System. *Community White Paper for OceanObs*, 9, 2009.

[15] J. Durbin and S.J. Koopman. *Time series analysis by state space methods*. Oxford University Press, 2001.

[16] G. Evensen. The ensemble Kalman filter : Theoretical formulation and practical implementation. *Ocean dynamics*, 53(4) :343–367, 2003.

[17] G.R. Foltz and M.J. McPhaden. Impact of Saharan dust on tropical North Atlantic SST. *J. Climate*, 21 :5048–5060, 2008.

[18] C. Frankignoul and K. Hasselmann. Stochastic climate models. II- Application to sea-surface temperature anomalies and thermocline variability. *Tellus*, 29 :289–305, 1977.

[19] C. Frankignoul and R.W. Reynolds. Testing a dynamical model for mid-latitude sea surface temperature anomalies. *Journal of Physical Oceanography*, 13(7) :1131–1145, 1983.

[20] M. Ghil and P. Malanotte-Rizzoli. Data assimilation in meteorology and oceanography. *Advances in geophysics*, 33 :141–266, 1991.

[21] T. Gneiting, M.G. Genton, and P. Guttorp. Geostatistical space-time models, stationarity, separability and full symmetry. *Statistical Methods for Spatio-Temporal Systems*, pages 151–175, 2007.

[22] F. Gohin and G. Langlois. Using geostatistics to merge in situ measurements and remotely-sensed observations of sea surface temperature. *International Journal of Remote Sensing*, 14(1) :9–19, 1993.

[23] P. Goovaerts. *Geostatistics for natural resources evaluation*. Oxford University Press, USA, 1997.

[24] X. Gourdon. *Les maths en tête : analyse*. Ellipses, 1994.

[25] AW Heemink, M. Verlaan, and AJ Segers. Variance reduced ensemble Kalman filtering. *Monthly Weather Review*, 129 :1718–1728, 2001.

[26] K. Herterich and K. Hasselmann. Extraction of mixed layer advection velocities, diffusion coefficients, feedback factors and atmospheric forcing parameters from the statistical analysis of North Pacific SST anomaly fields. *Journal of Physical Oceanography*, 17 :12, 1987.

[27] TA Hewett. Fractal distributions of reservoir heterogeneity and their influence on fluid transport. In *SPE Annual Technical Conference and Exhibition*, 1986.

[28] K. Ide, P. Courtier, M. Ghil, and A.C. Lorenc. Unified notation for data assimilation : Operational, sequential and variational. *Practice*, 75(1B) :181–189, 1997.

[29] A.B. Kara, A.J. Wallcraft, and H.E. Hurlburt. A Correction for Land Contamination of Atmospheric Variables near Land–Sea Boundaries. *Journal of Physical Oceanography*, 37 :803–818, 2007.

[30] K. Koch, J. Ohser, and K. Schladitz. Spectral theory for random closed sets and estimating the covariance via frequency space. *Advances in Applied Probability*, 35(3) :603–613, 2003.

[31] W. Krauss, R. Doscher, A. Lehmann, and T. Viehoff. On eddy scales in the eastern and northern North Atlantic Ocean as a function of latitude. *J. Geophys. Res*, 95(18) :049–18, 1990.

[32] P. Le Borgne, G. Legendre, and A. Marsouin. Operational SST Retrieval from MetOp/AVHRR. In *Proc. 2007 EUMETSAT Conf., Amsterdam, the Netherlands*, 2007.

[33] R. K. Mehra and N. K. Gupta. Computational aspects of maximum likelihood estimation and reduction in sensitivity calculations. *IEEE Trans Automatic Control*, 1 :774–783, 1974.

[34] T.P. Mitchell and J.M. Wallace. The annual cycle in equatorial convection and sea surface temperature. *Journal of Climate*, 5(10) :1140–1156, 1992.

[35] R. Morrow, R. Coleman, J. Church, and D. Chelton. Surface eddy momentum flux and velocity variances in the Southern Ocean from Geosat altimetry. *Journal of Physical Oceanography*, 24(10), 1994.

[36] E. J. Noyes, P. J. Minnett, G. K. Corlett, and et al. Validation of the AATSR Gridded Sea Surface Temperature Product with In-Situ Measurements from the M-AERI. In *Envisat ERS Symposium*, volume 572 of *ESA Special Publication*, April 2005.

[37] A.G. Ostrovskii and L.I. Piterbarg. Diagnosis of the seasonal variability of water surface temperature anomalies in the North Pacific. *Meteorology and Hydrology*, 12 :51–58, 1985.

[38] A.G. Ostrovskii and L.I. Piterbarg. Inversion for heat anomaly transport from sea surface temperature time series in the northwest Pacific. *Journal of Geophysical Research*, 100(C3) :4845–4865, 1995.

[39] A.G. Ostrovskii and L.I. Piterbarg. A New Method for Obtaining Velocity and Diffusivity from Time-Dependent Distributions of a Tracer via the Maximum Likelihood Estimator for the Advection-Diffusion Equation. *Journal of Computational Physics*, 133(2) :340–360, 1997.

[40] A.G. Ostrovskii and L.I. Piterbarg. Inversion of upper ocean temperature time series for entrainment, advection, and diffusivity. *Journal of Physical Oceanography*, 30 :201–214, 2000.

[41] K.A. Park and J.Y. Chung. Spatial and temporal scale variations of sea surface temperature in the East Sea using NOAA/AVHRR data. *Journal of Oceanography*, 55(2) :271–288, 1999.

[42] C. Penland. A stochastic model of IndoPacific sea surface temperature anomalies. *Physica D : Nonlinear Phenomena*, 98(2-4) :534–558, 1996.

[43] J. F. Piolle and Prevest C. Match-up database user manual. Technical report, IFREMER, http ://www.medspiration.org/tools/mdb/, 2006.

[44] L.I. Piterbarg and A.G. Ostrovskii. *Advection and diffusion in random media : implications for sea surface temperature anomalies*. Kluwer Academic Publishers, 1997.

[45] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes 3rd Edition : The Art of Scientific Computing*. Cambridge University Press, New York, NY, USA, 2007.

[46] R. W. Reynolds, N. A. Rayner, T. M. Smith, D. C. Stokes, and W. Wang. An Improved In Situ and Satellite SST Analysis for Climate. *Journal of Climate*, 15 :1609–1625, July 2002.

[47] RW Reynolds. Sea surface temperature anomalies in the North Pacific Ocean. *Tellus*, 30 :97–103, 1978.

[48] R.W. Reynolds, T.M. Smith, C. Liu, D.B. Chelton, K.S. Casey, and M.G. Schlax. Daily high-resolution-blended analyses for sea surface temperature. *Journal of Climate*, 20(22) :5473–5496, 2007.

[49] M.H. Rio and F. Hernandez. A mean dynamic topography computed over the world ocean from altimetry, in situ measurements, and a geoid model. *Journal of Geophysical Research*, 109(C12) :C12032, 2004.

[50] S. Roweis and Z. Ghahramani. A Unifying Review of Linear Gaussian Models. *Neural Computation*, 11 :305–345, 1999.

[51] J. Servain, A. J. Busalacchi, M. J. McPhaden, A. D. Moura, G. Reverdin, M. Vianna, and S. E. Zebiak. A Pilot Research Moored Array in the Tropical Atlantic (PIRATA). *Bulletin of the American Meteorological Society*, 79 :2019–2032, October 1998.

[52] J. Sherman and W.J. Morrison. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics*, 21(1) :124–127, 1950.

[53] A.C. Stuart-Menteth, I.S. Robinson, and P.G. Challenor. A global study of diurnal warming using satellite-derived sea surface temperature. *J. Geophys. Res*, 108 :3155, 2003.

[54] P. Tandeo, E. Autret, J. F. Piolle, J. Tournadre, and P. Ailliot. A multivariate regression approach to adjust aatsr sea surface temperature to in situ measurements. *Geoscience and Remote Sensing Letters, IEEE*, 6(1) :8–12, Jan. 2009.

[55] T. Viehoff. Mesoscale variability of sea surface temperature in the North Atlantic. *International Journal of Remote Sensing*, 10(4) :771–785, 1989.

[56] C.K. Wikle. Hierarchical Bayesian models for predicting the spread of ecological processes. *Ecology*, 84(6) :1382–1394, 2003.

[57] C.E. Woodcock, A.H. Strahler, and D.L.B. Jupp. The use of variograms in remote sensing : II. Real digital images. *Remote Sensing of Environment*, 25(3) :349–379, 1988.

[58] C. F. Jeff Wu. On the convergence properties of the em algorithm. *The Annals of Statistics*, 11(1) :95–103, 1983.

[59] K. Xu and C.K. Wikle. Estimation of parameterized spatio-temporal dynamic models. *Journal of Statistical Planning and Inference*, 137(2) :567–588, 2007.

Annexe A

Cet article a été réalisé au cours de cette thèse. Il a été publié dans la revue "Geoscience and Remote Sensing Letter, IEEE", volume 6, n°1 de Janvier 2009.

Résumé Le capteur satellite Along-Track Scanning Radiometer (AATSR), à bord d'Envisat, est réputé pour fournir des mesures précisent de températures de surface des océans (SST). Une base de donnée colocalisée (MDB) a été mise en place, utilisant des données *in situ* de bouées dérivantes, afin de calibrer les données AATSR. Pour le moment, les biais de SST de l'AATSR Medspiration sont des valeurs discrètes et peuvent introduire des discontinuités dans les champs de produits L2. Les nouveaux biais de SST de l'AATSR présentés dans cette lettre sont continus. Ils sont calculés à partir de données de nuit et des pixels ayant un indice de qualité fort, par régression linéaire sur différentes covariables de la MDB (vitesse du vent, différence entre la SST dual et nadir, aérosols, etc...). Les résultats montrent que la différences entre la SST dual et nadir explique 26% de la variabilité.

A Multivariate Regression Approach to Adjust AATSR Sea Surface Temperature to *In Situ* Measurements

Pierre Tandeo, Emmanuelle Autret, Jean François Piollé, Jean Tournadre, and Pierre Ailliot

***Abstract*—The Advanced Along-Track Scanning Radiometer (AATSR) onboard Envisat is designed to provide very accurate measurements of sea surface temperature (SST). Using colocated *in situ* drifting buoys, a dynamical matchup database (MDB) is used to assess the AATSR-derived SST products more precisely. SST biases are then computed. Currently, Medspiration AATSR SST biases are discrete values and can introduce artificial discontinuities in AATSR level-2 SST fields. The new AATSR SST biases presented in this letter are continuous. They are computed, for nighttime and best proximity confidence data, by linear regression with different MDB covariables (wind speed, latitude, aerosol optical depth, etc.). As found, the difference between dual-view and nadir-only SST products explains most of the variability (26%).**

***Index Terms*—Advanced Along-Track Scanning Radiometer (AATSR), remote sensing, sea surface temperature (SST), validation.**

I. Introduction

THE ALONG-TRACK Scanning Radiometer (ATSR) series [including the Advanced ATSR (AATSR)] is a generation of infrared radiometers which were built specifically to retrieve accurate measurements of sea surface temperature (SST) for environmental and climate applications. The AATSR onboard Envisat is a fundamental data set used within the Godae High Resolution Sea Surface Temperature Pilot Project (GHRSST-PP) analysis systems (see [1]) as the reference data set to estimate and correct bias with other satellite data prior to merging and objective analysis processes.

An optimal assignment of error and confidence for each measurement is indeed essential to merging and objective analysis. In the case of infrared SST data sets, data are categorized according to wind speed (used to identify favorable conditions to sea surface stratification) and their proximity to clouds. Bias and standard deviation errors [also called sensor-specific error statistics (SSES)] estimated by statistical analysis of *in situ* satellite matchups are then assigned to the SST samples. A proximity confidence value is then eventually assigned based on its "proximity" to several criteria known to have effects on the final SST estimate.

The current AATSR SSES (see [2]) given by Medspiration (the European node to GHRSST system; http://www.medspiration.org) are stratified in a 12-fold clustering of data, corresponding to different values of covariates. Among them, moderate to high wind speed ($> 6\ \mathrm{m} \cdot \mathrm{s}^{-1}$) is identified to have an impact on the mixed layer and to reduce SST variability. Moreover, the difference between the two possible AATSR SST measurements, i.e., the dual-view (combination of nadir and 55° incidence view; see [3] for more details) and nadir-only SST retrievals, is split into three categories. The dual–nadir SST difference allows the detection of possible contamination by clouds, aerosols, or other contaminants. Finally, the well-known problem of diurnal cycle is solved by splitting the data into nighttime and daytime measurements. The corresponding 12 biases and standard deviations of the SSES have been computed from the colocated data stored in the matchup database (MDB) presented in Section II. Applying such discrete biases can introduce artificial discontinuities in AATSR level-2 (L2) SST fields which can cause instabilities in the interpolated analysis.

Our objective is to define and determine continuous AATSR SSES based on the same MDB as the segmented Medspiration AATSR biases. We focus on the SST measurements currently used by most of the AATSR SST users, i.e., nighttime and best proximity confidence data. The regression method used to compute continuous bias values is described in Section III. Section IV presents the results of the MDB analysis in terms of SSES, as well as a comparison between the derived AATSR SST biases. Section V gives some concluding remarks and perspectives for future work.

II. Data

The development of the GHRSST-PP system has strengthened the need for the validation and intercomparison of satellite SST products. *In situ* data are a reliable independent source of measurements widely used to validate these products (see [4]), and the Institut Français de Recherche pour l'Exploitation de la Mer (IFREMER), Brest, France, operationally produces an MDB (http://www.medspiration.org/tools/mdb/) that provides colocated satellite and *in situ* measurements. The system relies on the Coriolis system (for the worldwide collection and

Manuscript received August 5, 2008; revised September 10, 2008. First published November 11, 2008; current version published January 14, 2009. This work was supported by the Medspiration Project by the European Space Agency.

P. Tandeo, E. Autret, J. F. Piollé, and J. Tournadre are with the Laboratoire d'Océanographie Spatiale, Institut Français de Recherche pour l'Exploitation de la Mer (IFREMER), 29280 Plouzané, France (e-mail: pierre.tandeo@ifremer.fr; emmanuelle.autret@ifremer.fr; jfpiolle@ifremer.fr; jean.tournadre@ifremer.fr).

P. Ailliot is with the Département de Mathématiques, Université de Bretagne Occidentale, 29200 Brest, France (e-mail: pierre.ailliot@univ-brest.fr).

Digital Object Identifier 10.1109/LGRS.2008.2006568

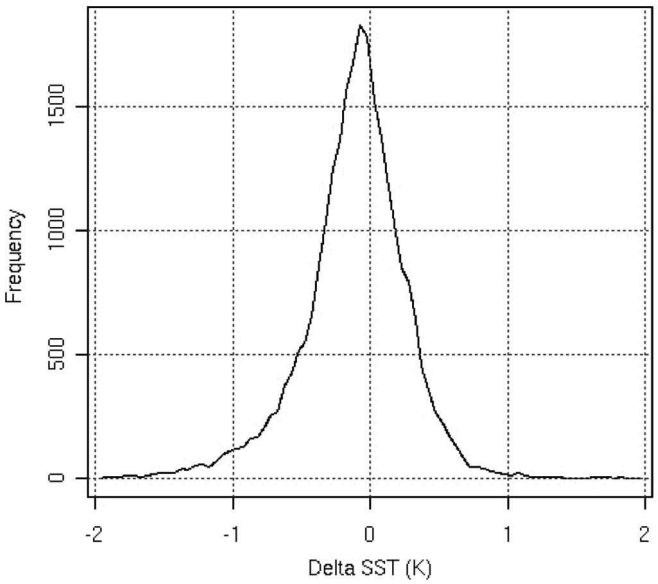


Fig. 1. Distribution of the AATSR ΔSST for nighttime best quality pixels. The mean bias is −0.11 K with an std of 0.39 K.

archiving of *in situ* sea data; http://www.coriolis.eu.org) for *in situ* measurements and on the Medspiration archive for satellite data, both hosted at IFREMER.

Assuming that the drifting-buoy measurements constitute the sea truth, the accuracy of AATSR SST, ΔSST, is defined as

$$\Delta\mathrm{SST} = \mathrm{SST}_{\mathrm{dual}} - \mathrm{SST}_{\mathrm{buoys}} \tag{1}$$

where $\mathrm{SST}_{\mathrm{dual}}$ is the dual-view skin SST estimation given in the AATSR L2 products and $\mathrm{SST}_{\mathrm{buoys}}$ is the temperature measured by drifting buoys at a depth that depends on the buoy type (is generally close to 1 m) that is called subskin SST. They are provided by the Coriolis system where they are quality controlled and archived. The skin-to-subskin adjustment presented by [5] is not used here as it will be directly integrated into our model.

The best quality pixels, i.e., nighttime colocated data with the highest proximity confidence value (i.e., $\mathrm{SST}_{\mathrm{dual}} - \mathrm{SST}_{\mathrm{nadir}} \in \{-0.51, 0.51\}$ K), are taken into account. Colocation criteria are $\Delta x < 10$ km, $\Delta t < 2$ h, and $|\Delta\mathrm{SST}| < 2$ K. The latter criterion is used to eliminate obvious outliers that can strongly modify the standard deviation. The data set contains 25 384 matchups. The corresponding distribution of ΔSST is shown in Fig. 1.

In addition to ΔSST, the MDB also provides covariates from different sources such as model outputs or ancillary fields about data quality. Nine covariates are available, and their descriptions are given in Table I. Among them, the two covariates providing statistics on the satellite pixels surrounding the *in situ* measurement are computed within a 25×25 km^2 box. They give a contextual information on the satellite data quality and variability and/or possible cloud contamination.

The MDB used in this letter covers two and a half years (from April 2005 to September 2007) and most of the global ocean. It should, however, be noted that the Atlantic Ocean is overrepresented because of a higher availability of *in situ* data. Moreover, Fig. 2 shows that the sampling in latitude is irregular with a relatively small number of colocated data in the tropical

TABLE I
DESCRIPTION OF THE MDB COVARIATES USED IN THIS LETTER

Name	Description	Source
Longitude	Measurement longitude	L2
Latitude	Measurement latitude	L2
Aerosol optical depth	Aerosol atmosphere quantity	NAAPS
Wind speed	Near surface wind speed	ECMWF
Zenithal solar angle	Angle between zenith and sun	L2
Valid pixel in box	Valid pixel by $25 \times 25\,\mathrm{km}^2$ box	Ifremer
Sigma SST in box	Std SST by $25 \times 25\,\mathrm{km}^2$ box	Ifremer
SST_{dual}	SST dual view	L2
$SST_{dual} - SST_{nadir}$	Dual-nadir SST difference	L2

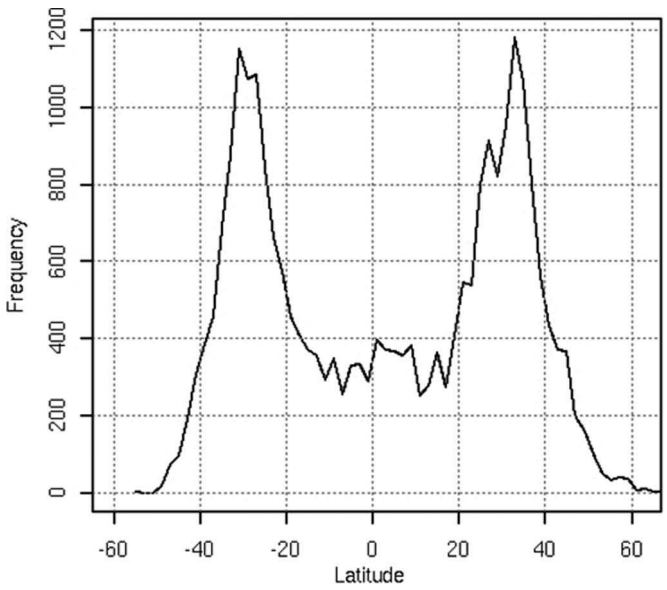


Fig. 2. Number of colocated samples by band of 2° latitude.

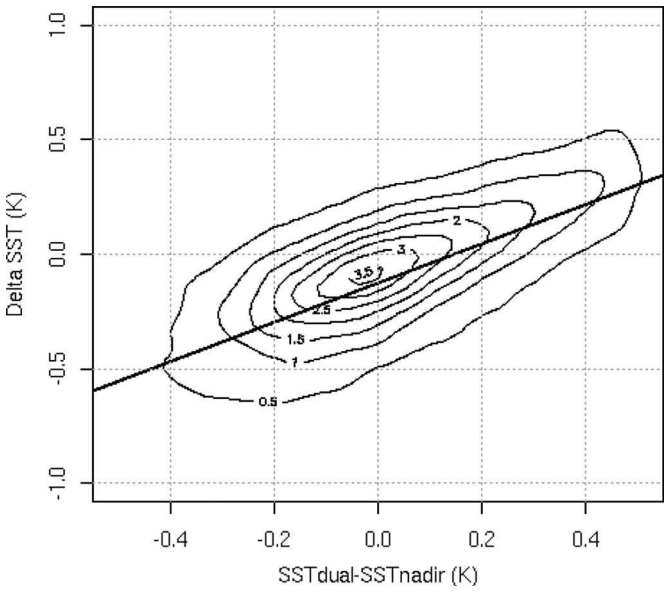


Fig. 3. Bidimensional pdf of ΔSST and AATSR dual–nadir SST difference and the corresponding regression line (intercept $= -0.13$, slope $= 0.86$, and $R^2 = 26\%$).

TABLE II
COEFFICIENTS OF THE MODEL FOR THE SELECTED COVARIATE (NIGHTTIME MEASUREMENTS ONLY)

Covariate	Associated parameter value	Confidence interval (95 %)
Intercept $\alpha^{(0)}$	0.006	$[-0.005; 0.016]$
Sigma SST in box $\alpha^{(1)}$	-0.422	$[-0.452; -0.392]$
Zenithal solar angle * $SST_{dual} - SST_{nadir}$ $\beta^{(1)}$	0.007	$[0.0068; 0.0071]$

TABLE III
COEFFICIENTS OF THE MODEL FOR THE SELECTED COVARIATE (DAYTIME MEASUREMENTS ONLY)

Covariate	Associated parameter value	Confidence interval (95 %)
Intercept $\alpha'^{(0)}$	0.039	$[0.018; 0.061]$
Sigma SST in box * SST_{dual} $\beta'^{(1)}$	0.030	$[0.027; 0.032]$
Sigma SST in box * $SST_{dual} - SST_{nadir}$ $\beta'^{(2)}$	1.112	$[0.972; 1.252]$

band and a large number at midlatitude. The model that can be estimated from the MDB cannot be mainly representative of high-latitude conditions ($> 40^\circ$).

III. METHOD

Our aim is to estimate ΔSST in real time in order to adjust $\mathrm{SST_{dual}}$ measurements. In statistical terms, it is equivalent to finding the effects of simultaneous p covariates $\{X^{(i)}, i = 1, \ldots, p\}$ on a response variable ΔSST. The analysis of the bidimensional probability density function (pdf) of ΔSST and $\mathrm{SST_{dual}} - \mathrm{SST_{nadir}}$, shown in Fig. 3, shows that the relationship between the two variables is quite linear. The pdfs of ΔSST and the other covariates (not shown here) present very similar features. Therefore, a simple linear model, which will thus implicitly include a skin-to-subskin adjustment, is used to model ΔSST as a function of the covariates. The simplest linear model can be expressed as

$$\Delta\mathrm{SST} = \alpha^{(0)} + \sum_{i=1}^{p} \alpha^{(i)} X^{(i)} + \varepsilon \tag{2}$$

where ε is a Gaussian white noise and $X^{(i)}$ denotes the p covariates. The coefficients $\alpha^{(0)}$ and $\alpha^{(i)}$ are the intercept and principal effects on covariates, respectively. However, some covariate effects might be linked, and their possible interaction can be taken into account by including quadratic terms in the model in the form

$$\Delta\mathrm{SST} = \alpha^{(0)} + \sum_{i=1}^{p} \alpha^{(i)} X^{(i)} + \sum_{i=1}^{p} \sum_{j=1}^{p} \beta^{(i:j)} X^{(i)} X^{(j)} + \varepsilon \tag{3}$$

where $\beta^{(i:j)}$ corresponds to the interaction between the covariates $X^{(i)}$ and $X^{(j)}$. All the parameters in (3) are estimated by the least squares method. In order to select the best fit model with fewer parameters, we conducted an exhaustive search based on the best fit criterion R^2_{adj} defined as

$$R^2_{\mathrm{adj}} = 1 - \frac{\mathrm{Var_{error}}}{\mathrm{Var_{total}}} \tag{4}$$

where

$$\mathrm{Var_{error}} = \frac{\sum(\Delta\mathrm{SST} - \widehat{\Delta\mathrm{SST}})^2}{n - p - 1} \tag{5}$$

$$\mathrm{Var_{total}} = \frac{\sum(\Delta\mathrm{SST} - \overline{\Delta\mathrm{SST}})^2}{n - 1} \tag{6}$$

with $\widehat{\Delta\mathrm{SST}}$ being the ΔSST estimated by the model, $\overline{\Delta\mathrm{SST}}$ being the mean of ΔSST, and n being the number of observations. R^2_{adj} gives an information about the quality of the fit with a penalization on p (the complexity), the number of parameters estimated in the model.

IV. RESULTS

An exhaustive search of the model with all covariate combinations is done in accordance with the R^2_{adj} introduced earlier. Model selection shows that the interaction between the zenithal solar angle and $\mathrm{SST_{dual}} - \mathrm{SST_{nadir}}$ is the most important effect on ΔSST. The addition of the standard deviation of SST by validation box of 25×25 km^2 around the pixel improves the adjustment ($R^2_{\mathrm{adj}} = 28\%$). Owing to the small increase of R^2_{adj} in adding new covariates ($\mathrm{SST_{dual}}$, aerosol optical depth, etc.), the simple model with three parameters (including intercept) is chosen. The α and β parameters and their confidence interval corresponding to those variables are given in Table II. This model identifies the dual–nadir difference retrieval as the most important covariate. It explains 26% of the residual SST variability (see Fig. 3).

Although daytime data are rarely used for objective analysis purposes (mainly because the correction of diurnal warming is not fully solved), the method has also been applied to best quality flag daytime data. The data set contains 7931 matchups. Three parameters are also retained and are given in Table III. The residual SST variability ($R^2_{\mathrm{adj}} = 15.3\%$) is less explained certainly because, even for the best quality sample, diurnal effect cannot be completely eliminated.

The SSES model has been tested on real-time data, i.e., AATSR L2 files corresponding to one day (May 24, 2008) with nighttime measurements and best quality pixel in North Atlantic, corresponding to a small difference between dual-view and nadir-only SST retrievals ($|\mathrm{SST_{dual}} - \mathrm{SST_{nadir}}| < 0.51$ K) and, therefore, small cloud or aerosol contaminants

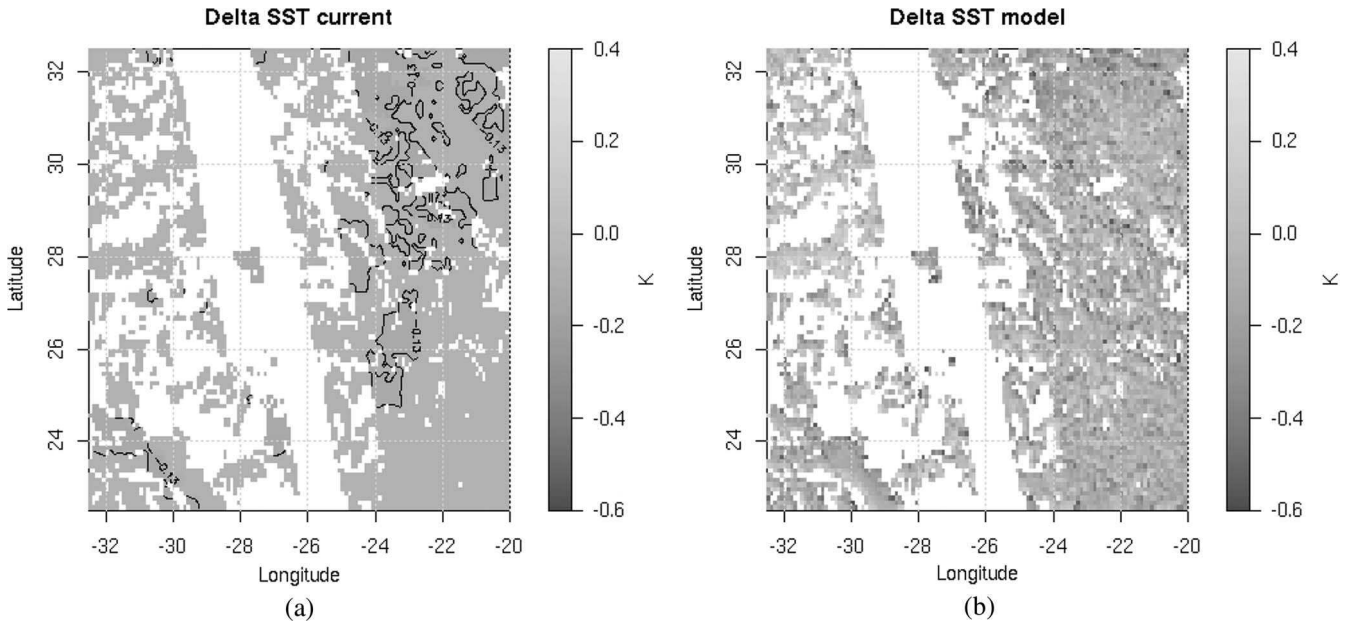


Fig. 4. (a) $\Delta\widehat{\mathrm{SST}}_{\mathrm{current}}$ given in the AATSR L2 files in a part of North Atlantic for nighttime measurements and best quality pixels (May 26, 2008; contour plot identifies the two values, namely, -0.13 K and -0.07 K). (b) $\Delta\widehat{\mathrm{SST}}_{\mathrm{model}}$ computed by model (3) on the same data.

of the $\mathrm{SST}_{\mathrm{dual}}$. Fig. 4 compares the original $\Delta\widehat{\mathrm{SST}}_{\mathrm{current}}$ and $\Delta\widehat{\mathrm{SST}}_{\mathrm{model}}$ computed by the model proposed in Table II. For these particular data, $\Delta\widehat{\mathrm{SST}}_{\mathrm{current}}$ takes only two values, namely, -0.13 and -0.07, corresponding to low- and high-wind-speed $(> 6\ \mathrm{m}\cdot\mathrm{s}^{-1})$ conditions, respectively. Both values include the -0.17-K-depth-to-skin adjustment proposed by [5]. The figure also shows that discrete biases can create artificial SST gradients, while the new model gives continuous $\Delta\widehat{\mathrm{SST}}_{\mathrm{model}}$, as shown in Fig. 4(b).

The current and model SST biases are also validated by comparison to an independent data set constituted by the SST measured by the infrared Advanced Very High Resolution Radiometer (AVHRR) sensor onboard METOP-A. The adjusted SSTs are defined by

$$\mathrm{SST}_{\mathrm{adj}} = \mathrm{SST}_{\mathrm{dual}} - \Delta\widehat{\mathrm{SST}} \tag{7}$$

where ΔSST is the current or model bias. Only nighttime, best confidence value, and $\Delta t < 3$ h colocated data are studied. We test the model on three days in the year under different conditions representing 197 688 pixels. Fig. 5 shows the pdfs of the difference between AVHRR SST and the nonadjusted (without any bias) and adjusted (current and model) AATSR dual SST. The distribution of the AVHRR and dual SST AATSR is, as expected, positively biased (0.09 K) because no subskin-to-skin correction is applied. The adjustment of AATSR SST by the current SSES strongly reduces the mean bias to -0.004 K, but the distribution has the same standard deviation (0.42 K). The model adjustment further reduces the standard deviation (0.36 K).

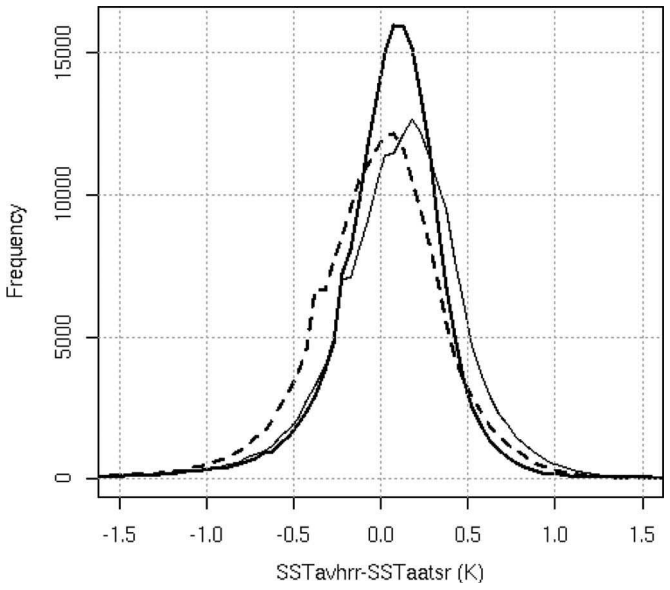


Fig. 5. Histograms of the difference between the AVHRR SST and (thick solid line) $\widehat{\mathrm{SST}}_{\mathrm{model}}$, (dashed line) $\widehat{\mathrm{SST}}_{\mathrm{current}}$, and (thin solid points) $\mathrm{SST}_{\mathrm{dual}}$ for the 197 688 colocated pixels in the Atlantic Ocean on several days (January 20, May 26, and August 17, 2008). Only nighttime best quality pixels are considered.

V. Conclusion and Perspective

The MDB maintained at IFREMER within the Medspiration project to colocate satellite and *in situ* SST measurements has been used to estimate a continuous model of SST bias $(\Delta\mathrm{SST}_{\mathrm{model}})$ for the AATSR sensor. $\Delta\mathrm{SST}_{\mathrm{model}}$ has been fitted by linear regression on nine covariates. Among them, the $\mathrm{SST}_{\mathrm{dual}} - \mathrm{SST}_{\mathrm{nadir}}$ explains a large part of the ΔSST variability. The continuous nature of the model biases means that it is well adapted for merging and objective analysis processes while being similar to the categorized $\Delta\mathrm{SST}_{\mathrm{current}}$ given by Medspiration.

Nevertheless, the use of the MDB in order to fit a model on SST bias is questionable. First, we use drifting-buoy

measurements as reference which measure the subskin SST. Second, we point the lack of sampling in high latitude. An upcoming study will evaluate the latitude distribution of AATSR L2 files in order to reproduce a similar distribution in a new MDB. This artificial variability in low sampled latitudes can be created by bootstrap methods. Finally, we need a more elaborate study on the colocation parameters (distance and time difference) between the AATSR sensor and drifting buoys in order to specify the data sampling.

As derived, the method presented in this letter can also be applied to instruments such as the AVHRR and the Advanced Microwave Scanning Radiometer which are also available in the MDB.

ACKNOWLEDGMENT

The authors would like to thank Dr. G. Corlett for his help during this work.

REFERENCES

[1] The GHRSST-PP Science Team and ISDI-TAG Working Group, *The Recommended GHRSST-PP Data Processing Specification GDS. No. GHRSST/17*, Mar. 2004. [Online]. Available: http://ghrsst-pp.metoffice.com/

[2] G. K. Corlett and D. J. Poulter, "An updated confidence flagging scheme for Medspiration AATSR L2P files," Univ. Leicester, U.K., Tech. Rep., MED-UL-REP-002-1D, Medspiration, 2008. [Online]. Available: http://medspiration.org

[3] D. Llewellyn-Jones, M. C. Edwards, C. T. Mutlow, A. R. Birks, I. J. Barton, and H. Tait, "AATSR: Global-change and surface-temperature measurements from Envisat," *ESA Bull.*, vol. 105, pp. 11–21, Feb. 2001.

[4] G. K. Corlett, I. J. Barton, C. J. Donlon, M. C. Edwards, S. A. Good, L. A. Horrocks, D. T. Llewellyn-Jones, C. J. Merchant, P. J. Minnett, T. J. Nightingale, E. J. Noyes, A. G. O'Carroll, J. J. Remedios, I. S. Robinson, R. W. Saunders, and J. G. Watts, "The accuracy of SST retrievals from AATSR: An initial assessment through geophysical validation against *in situ* radiometers, buoys and other SST data sets," *Adv. Space Res.*, vol. 37, no. 4, pp. 764–769, 2006.

[5] C. J. Donlon, P. J. Minnett, C. Gentemann, T. J. Nightingale, I. J. Barton, B. Ward, and M. J. Murray, "Toward improved validation of satellite sea surface skin temperature measurements for climate research," *J. Clim.*, vol. 15, no. 4, pp. 353–369, Feb. 2002.

Annexe B

Cet article a été réalisé au cours de cette thèse. Il a été accepté dans la revue "Stochastic Environmental Research and Risk Assessment, Springer Link".

Résumé Les satellites fournissent des informations sur de nombreuses variables météorologiques et océanographiques. Les modèles espace d'état sont utilisés pour analyser ce genre de données qui disposent d'une variance d'erreur. Dans ce travail, nous proposons de généraliser le modèle espace d'état linéaire et Gaussien au cas des données irrégulièrement échantillonnées dans de le temps, comme c'est le cas pour des données satellitaires en un point fixe.

Nous discutons de l'estimation des paramètres en utilisant la méthode des moments et celle du maximum de vraisemblance. Des résultats simulés indiquent que la méthode des moments rend des estimateurs robustes, idéales pour initialiser l'algorithme EM, qui est combiné à une méthode d'optimisation numérique dans le but de maximiser la fonction de vraisemblance.

Le modèle est validé sur des données de températures de surface de l'eau (SST) observées par un capteur satellite. Les résultats indiquent que la méthodologie proposée peut être utilisée pour reconstruire des séries temporelles réalistes de SST en un point précis. Cette méthodologie permet également d'obtenir des informations sur la qualité des mesures satellitaires ainsi que sur la dynamique temporelle de la SST.

Stoch Environ Res Risk Assess
DOI 10.1007/s00477-010-0442-8

ORIGINAL PAPER

Linear Gaussian state-space model with irregular sampling: application to sea surface temperature

Pierre Tandeo · Pierre Ailliot · Emmanuelle Autret

Abstract Satellites provide important information on many meteorological and oceanographic variables. State-space models are commonly used to analyse such data sets with measurement errors. In this work, we propose to extend the usual linear and Gaussian state-space to analyse time series with irregular time sampling, such as the one obtained when keeping all the satellite observations available at some specific location. We discuss the parameter estimation using a method of moment and the method of maximum likelihood. Simulation results indicate that the method of moment leads to a computationally efficient and numerically robust estimation procedure suitable for initializing the Expectation–Maximisation algorithm, which is combined with a standard numerical optimization procedure to maximize the likelihood function. The model is validated on sea surface temperature (SST) data from a particular satellite. The results indicate that the proposed methodology can be used to reconstruct realistic SST time series at a specific location and also give useful information on the quality of satellite measurement and the dynamics of the SST.

P. Tandeo (✉) · E. Autret
Laboratoire d'Océanographie Spatiale, IFREMER, Plouzané, France
e-mail: pierre.tandeo@ifremer.fr

E. Autret
e-mail: emmanuelle.autret@ifremer.fr

P. Ailliot
Laboratoire de Mathématiques, UMR 6205, Université Européenne de Bretagne, Brest, France
e-mail: pierre.ailliot@univ-brest.fr

Keywords State-space model · Irregular sampling · Ornstein–Uhlenbeck process · EM algorithm · Sea surface temperature

1 Introduction

Sea surface temperature (SST) is an important oceanographic variable for many applications (see e.g. [7] and references therein). Several satellites and buoy networks provide continuous observations of this variable, leading to a huge amount of data. Statistical methods are then needed to combine all this information and provide realistic SST analysis at any date and any location in the ocean.

State-space models provide a flexible methodology for analysing such complex environmental data sets, and they have already been used in a wide range of problems (see e.g. [13]) including meteorological and oceanographic applications (see e.g. [1, 11, 17, 26]). The basic idea of these models consists in introducing the "true" value of the physical variable of interest as a hidden variable (the "state"). Then, stochastic models are used both to describe the dynamics of the state and to relate the observations to the state. When linear Gaussian models are used, we get the so-called linear Gaussian state-space model which has been extensively studied in the literature (see e.g. [8] and references therein). Note that [14] proposed unified notations for state-space models and data assimilation in oceanography and meteorology which are partially adopted here.

In this work we analyse satellite SST data at a single location, where buoy data is available for comparison, and we consider the time series obtained by keeping all the satellite data available nearby this location. It leads to a time series with irregular time-step, with generally several data each day but also sometimes gaps of several days with

Published online: 11 November 2010

no data. We adopt a continuous-time state-space model to analyse this time series in which the state is supposed to be an Ornstein–Uhlenbeck process. It leads to a simple generalization of the usual linear Gaussian state-space model with regular time-step.

The most usual method for estimating the parameter in models with latent variable consists in computing the maximum likelihood estimates using the Expectation-Maximisation (EM) algorithm. In this work, we propose to improve the numerical efficiency of the EM algorithm by combining it with a method of moment and a standard numerical optimization procedure. The method of moment is used to provide realistic starting values to the EM algorithm with the extra benefit of providing graphical tools which permit to asses the realism of the model. The standard numerical optimization procedure is used to accelerate the convergence of the EM algorithm near the maxima and provide estimates of the observed information matrix and thus important information on the variance of the estimates.

The paper is organised as follows. The SST data and the model are introduced in Sect. 2 Then, the parameter estimation is discussed in Sect. 3: after describing the practical implementation of the various methods, we asses the efficiency of the whole procedure through simulations. In Sect. 4, we discuss the results obtained on the data with the proposed methodology. Conclusions are drawn in Sect. 5.

2 Data and model

Several instruments on-board satellites provide measurements of SST over the entire surface of the ocean with different spatial and temporal resolutions. In this work, we focus on the data provided by the infrared Advanced Very High Resolution Radiometer instrument on-board the METOP satellite (see [18] for more details). This satellite covers the global ocean with a spatial resolution of 0.05° and provides two SST observations per day at the most in optimal conditions. In this paper, we first consider the data available at a given location, with geographical coordinates (0°N, 23°W), in the tropical region of the Atlantic Ocean. More precisely, we consider 2 years of data, from 11-Jul-2007 to 18-Jun-2009, which are representative of the variability of the SST conditions at this location. Hereafter, $(t_1, \ldots, t_n)$ denotes the times at which the METOP satellite data are available, with $n = 1087$ the total number of observations. Since satellite observations may be contaminated by atmospheric conditions (e.g. cloud coverage), some data are missing and the time difference $\Delta_i = t_i - t_{i-1}$ between two consecutive observations may vary from a half day to a few days (see Fig. 1).

The resulting time-series is clearly non-stationary (see Fig. 2) with in particular important seasonal components. The non-stationary components have complex features and we could not find any appropriate parametric model to describe them. We have thus decided to use the SST analysis produced by the National Climatic Data Center (NCDC) (daily "OIV2 analysis" with 0.25° spatial resolution) to remove these components. These analysis are derived from different satellite sources independent of METOP data (see [22]) and we assume that they provide a good estimate of the low-variations of the SST conditions. Both data sources METOP and OIV2 are available at the URL http://www.hrdds.net.

Then we consider the time series $y_{t_1}^{t_n} = (y_{t_1}, \ldots, y_{t_n})$ obtained by removing the OIV2 analysis from the METOP data (see Fig. 3). We assume that this new time-series, referred as the SST anomaly hereafter, is a discrete-time realization of a continuous-time stationary process $\{Y_t\}$. Modelling the time series $\{y_{t_1}^{t_n}\}$ may provide important information on the small scale variability of SST and also on the quality of METOP measurements and OIV2 analysis as it will be shown in Sect. 4 and finally lead to a better assimilation of these data into numerical models.

The model that we consider for $\{Y_t\}$ is introduced below. First, we assume that the observed SST anomaly at

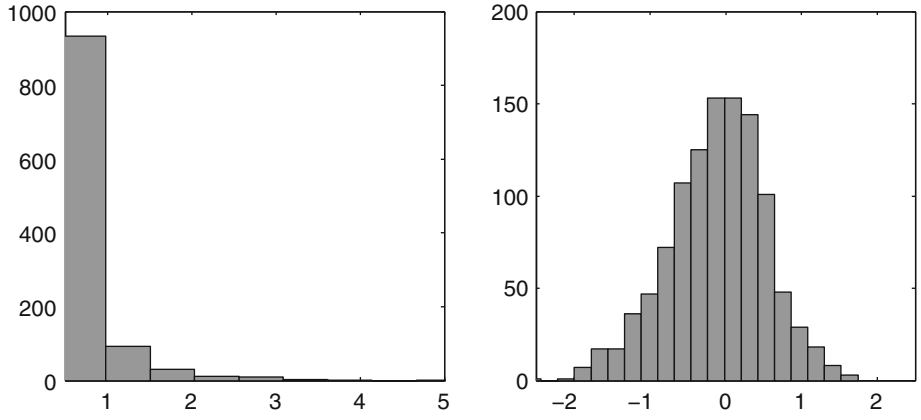


Fig. 1 Histogram of the time lags Δ_i in days (*left*) and of the SST anomalies $\{y_{t_i}\}$ in °C (*right*)

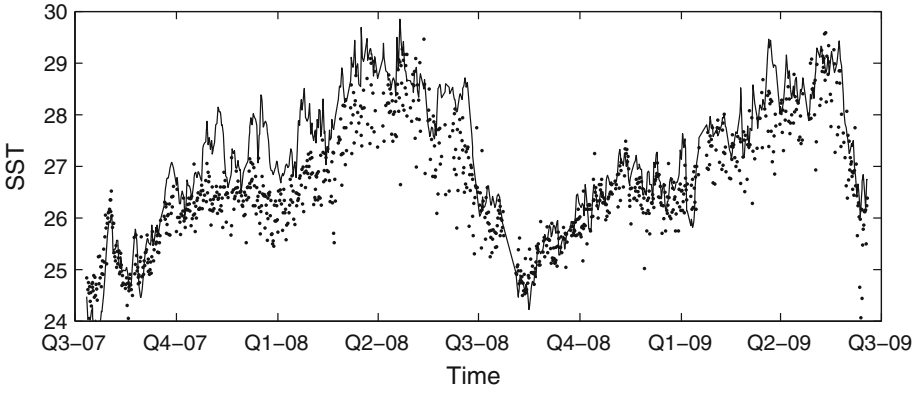


Fig. 2 Raw METOP SST (in °C) time series (*dotted line*) and OIV2 SST analysis (*full line*)

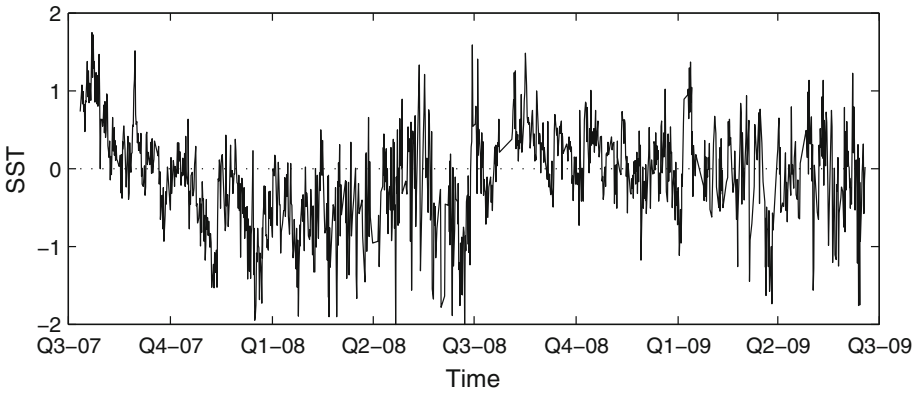


Fig. 3 SST anomalies (in °C) obtained by removing the OIV2 analysis from METOP data

time t, Y_t, is related to the "true" SST anomaly at time t, denoted X_t, by the measurement equation below:

$$Y_t = HX_t + \sqrt{R}\epsilon_t \tag{1}$$

where $\{\epsilon_t\}$ is a Gaussian white noise sequence with zero mean and unit variance. In practice R represents the variance of the observation error and H allows a transformation between the state and the observations. For the particular METOP measurements considered in this paper (we keep only the best quality data), the standard deviation of the observation error has been estimated globally to 0.5°C, but it is known that it may vary according to the retrieval algorithm (day-time and night-time), the region and the season (see [18] for more details). The observation equation (1) could be modified to take into account these fluctuations in the accuracy of the data. In the same way, we could include the various covariates which alter the quality of the satellite measurements (see [25]) or assume that the parameters H and R depend on the satellite if the observed time series was obtained by mixing data from different satellites.

Then we assume that the latent process $\{X_t\}$ is a simple Ornstein–Uhlenbeck process, that is a stationary solution of the following stochastic differential equation:

$$dX_t = -\lambda X_t \mathrm{d}t + \tau \mathrm{d}W_t \tag{2}$$

where $\{W_t\}$ denotes a standard Brownian motion. A physical justification of using this model to describe the local dynamics of the SST, when neglecting horizontal transport and heat exchange, is given in [10]: $\lambda > 0$ is the time correlation (in day) or feedback parameter which represents the slowly evolving transfer of heat and $\tau > 0$ the variability coming from weather fluctuations (see also [19, 21]).

Hereafter, we denote $\sigma^2 = \mathrm{Var}(X_t) = \frac{\tau^2}{2\lambda}$ the variance of the stationary distribution. $\{X_t\}$ is a Markov process which satisfies, for $i \in \{2, \ldots, n\}$,

$$X_{t_i} = M_{\Delta_i} X_{t_{i-1}} + \sqrt{Q_{\Delta_i}}\eta_{t_i} \tag{3}$$

with $M_{\Delta_i} = \exp(-\lambda\Delta_i)$, $Q_{\Delta_i} = \sigma^2\left(1 - {M_{\Delta_i}}^2\right)$ and $\{\eta_{t_i}\}_{i\in\{2,\ldots,n\}}$ a Gaussian white noise sequence with zero mean and unit variance independent of $\{\epsilon_{t_i}\}_{i\in\{1,\ldots,n\}}$. In the particular case when the temporal sampling is regular, i.e. when $\Delta_1 = \cdots = \Delta_n$, we retrieve a standard $AR(1)$ process and the usual linear Gaussian state-space model. Here again, more complicated models could be considered, with for example non-linear dynamics, but this would complicate the statistical inference methods discussed in the next section.

Finally, the various conditional independence assumptions which imply the particular Markovian structure of the

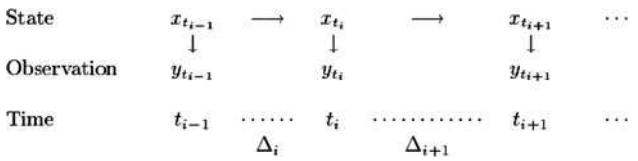


Fig. 4 Directed acyclic graph for the linear Gaussian state-space model with irregular time step

state-space model, when observed at discrete time $t_1, \ldots, t_n$, are summarized on the directed acyclic graph shown on Fig. 4.

3 Parameter estimation

The estimation of the unknown parameters in Gaussian linear state-space models observed at regular time step has been addressed by many authors and the most usual method consists probably in computing the maximum likelihood (ML) estimates using the EM algorithm (see e.g. [8]).

However, before computing the ML estimates, it is important to check the identifiability of the parameters. For the particular model under consideration, it is possible to show that the observations follow a multivariate Gaussian distribution with an explicit covariance function. Using this result, we can give conditions on the parameters which ensure identifiability and also propose a first method based on the moments to estimate the parameters. The corresponding estimates will be denoted MOM estimates hereafter. This is discussed in Sect. 3.1 Then, in Sect. 3.2, we detail the practical implementation of the EM algorithm for the Gaussian linear state-space model with irregular time-step. We discuss how it can be combined with the method of moment and a more standard numerical optimization procedure proposed in [16] to get a computationally efficient and numerically robust estimation procedure. Finally, this is illustrated in Sect. 3.3 through simulations.

3.1 Covariance function

With the various assumption made in the previous section, $\{Y_t\}$ is a stationary Gaussian process with zeros mean and covariance function

$$\mathrm{Cov}(Y_t, Y_{t'}) = H^2\sigma^2 \exp(-\lambda|t - t'|) + R\mathbf{1}_{\{0\}}(t - t') \qquad (4)$$

We deduce that the distribution of the observed sequence $(y_{t_1}, \ldots, y_{t_n})$ is a multivariate Gaussian distribution with zeros mean and covariance matrix which can be expressed from the unknown parameter H, R, σ^2 and λ. According to (4), this covariance matrix depends on the parameters H and σ^2 only through the product $H^2\sigma^2$ and thus we need to add a constraint in order to ensure identifiability of the parameters. Hereafter, we fix $H = 1$ and denote $\theta = (\lambda, \sigma^2, R) \in (0, +\infty)^3$ the unknown parameters.

The covariance function (4) corresponds to a classical model in spatial statistics since we retrieve an exponential model with nugget R, sill $\sigma^2 + R$ and range $1/\lambda$. Usual methods in geostatistics permit to compute an empirical estimate of the variogram from the data (see e.g. [3, p. 69]). The variogram is directly related to the covariance function for second order stationary processes and the empirical variogram can be used to check the realism of the parametric model 4 and also fit it using the weighted least square method. Here the weights depend on the number of pairs of time points which are available to estimate the empirical variogram as discussed in [3, p. 96]. The corresponding estimates will be denoted MOM estimates hereafter.

3.2 Maximum likelihood estimation

Alternatively, the parameters can be estimated by computing the ML estimates. According to the conditional independence assumptions shown on Fig. 4, the complete log-likelihood, based on both the latent and observed sequences, is given by

$$\log\left(p(x_{t_1}^{t_n}, y_{t_1}^{t_n}; \theta)\right) = \log(p(x_{t_1})) + \sum_{i=2}^{n} \log(p(x_{t_i}|x_{t_{i-1}}; \theta)) + \sum_{i=1}^{n} \log(p(y_{t_i}|x_{t_i}; \theta))$$

where the conditional distributions $p(x_{t_i}|x_{t_{i-1}}; \theta)$ and $p(y_{t_i}|x_{t_i}; \theta)$ are Gaussian distributions which characteristics are given respectively by (3) and (1). Hereafter we will assume that the initial distribution $p(x_{t_1})$ is a Gaussian distribution with known mean $x^{(b)}$ and variance B and in practice these values will be estimated using historical data. Thus, apart from a constant, we obtain

$$\begin{aligned}
&\log\left(p(x_{t_1}^{t_n}, y_{t_1}^{t_n}; \theta)\right) \\
&= -(n-1)\log(\sigma) - \frac{1}{2}\sum_{i=2}^{n} \log(1 - \exp(-2\lambda\Delta_i)) \\
&\quad - \frac{1}{2\sigma^2}\sum_{i=2}^{n} \frac{(x_{t_i} - \exp(-\lambda\Delta_i)x_{t_{i-1}})^2}{(1 - \exp(-2\lambda\Delta_i))} \\
&\quad - \frac{n}{2}\log(R) - \frac{1}{2R}\sum_{i=1}^{n}(y_{t_i} - x_{t_i})^2 \qquad (5)
\end{aligned}$$

The ML estimates $\hat{\theta}$ is the value of θ that maximises the (incomplete) likelihood of the observations $y_{t_1}^{t_n}$ formed by integrating the complete likelihood (5) over the missing variables.

In this paper, the EM algorithm due to [4] is used to compute $\hat{\theta}$. This recursive algorithm computes successive approximations $\hat{\theta}_k = \left(\lambda_k, \sigma_k^2, R_k\right)$ of $\hat{\theta}$ by cycling through the following steps.

E-step: Compute $U(\theta|\hat{\theta}_k) = E(\log(p(X_{t_1}^{t_n}, y_{t_1}^{t_n}; \theta))|y_{t_1}^{t_n}; \hat{\theta}_k)$ as a function of θ.
M-step: Determine the updated parameter estimate $\hat{\theta}_{k+1} = \arg\max_{\theta} U(\theta|\hat{\theta}_k)$.

Under certain general conditions it can be shown that the sequence of estimates $\hat{\theta}_n$ yields monotonically increasing values of the incomplete likelihood, and converges to a maximum of this function (see [15]). Thus the EM algorithm provides an alternative method of maximising the incomplete log-likelihood which is commonly used in models with hidden or latent variables such as the model proposed here. The EM algorithm directly utilises the hidden structure and, as a consequence, is often more robust in practice to the choice of starting values than usual numerical optimization methods. Its computational efficiency is enhanced if the E and M steps are readily evaluated. Various authors have discussed the practical implementation of these steps for linear Gaussian state-space models with regular time sampling ([2, 5, 6, 24], pp. 384–388). Hereafter, we discuss the extension to the case with irregular sampling.

E step To determine $U(\theta|\hat{\theta}_k)$ as a function of θ we need to compute the following smoothing probabilities, for $i = 1, \ldots, n$:

$$\begin{aligned} &x_{t_i}^{(s)} = E(X_{t_i}|y_{t_1}^{t_n}; \hat{\theta}_k), \quad x_{t_i,t_i}^{(s)} = E(X_{t_i}^2|y_{t_1}^{t_n}; \hat{\theta}_k), \\ &x_{t_{i-1},t_i}^{(s)} = E(X_{t_{i-1}}X_{t_i}|y_{t_1}^{t_n}; \hat{\theta}_k) \end{aligned} \tag{6}$$

These quantities can be computed using the Kalman recursions described hereafter. This is a particular case of the general Kalman recursions given for example in [24] and [2] pp. 127–147.

• **Kalman filter** Let us denote

$$x_{t_i}^{(f)} = E(X_{t_i}|y_{t_1}^{t_{i-1}}; \hat{\theta}_k), \quad P_{t_i}^{(f)} = \mathrm{Var}(X_{t_i}|y_{t_1}^{t_{i-1}}; \hat{\theta}_k)$$

the mean and the variance of the forecast probabilities and

$$x_{t_i}^{(a)} = E(X_{t_i}|y_{t_1}^{t_i}; \hat{\theta}_k), \quad P_{t_i}^{(a)} = \mathrm{Var}(X_{t_i}|y_{t_1}^{t_i}; \hat{\theta}_k)$$

the mean and the variance of the filtering probabilities. These quantities can be computed using the recursion below.

Initialization: Compute the Kalman filter gain $K_{t_1} = \frac{B}{B+R}$ and

$$x_{t_1}^{(a)} = x^{(b)} + K_{t_1}(y_{t_1} - x^{(b)}), \quad P_{t_1}^{(a)} = (1 - K_{t_1})B$$

where the parameters $x^{(b)} = E[X_{t_1}]$ and $B = \mathrm{Var}(X_{t_1})$ of the initial distribution are supposed to be known.

Recursion: for $i = 2, \ldots, n$

– **Time update:**

$$x_{t_i}^{(f)} = M_{\Delta_i} x_{t_{i-1}}^{(a)}, \quad P_{t_i}^{(f)} = M_{\Delta_i}^2 P_{t_{i-1}}^{(a)} + Q_{\Delta_i}$$

– **Observation update:** compute the Kalman filter gain $K_{t_i} = \frac{P_{t_i}^{(f)}}{P_{t_i}^{(f)} + R}$ and

$$x_{t_i}^{(a)} = x_{t_i}^{(f)} + K_{t_i}\left(y_{t_i} - x_{t_i}^{(f)}\right), \quad P_{t_i}^{(a)} = (1 - K_{t_i})P_{t_i}^{(f)}$$

• **Kalman smoother** Let us denote

$$P_{t_i}^{(s)} = \mathrm{Var}(X_{t_i}|y_{t_1}^{t_n}; \hat{\theta}_k)$$

the variance of the smoothing probabilities at time t_i. These quantities and the conditional expectation $x_{t_i}^{(s)}$ define in (6) can be computed using the backward recursions below.

Initialization:

$$x_{t_n}^{(s)} = x_{t_n}^{(a)}, \quad P_{t_n}^{(s)} = P_{t_n}^{(a)}$$

Recursion: for $i = n-1, \ldots, 1$ compute the Kalman smoother gain $K_{t_i}^{(s)} = \frac{P_{t_i}^{(a)} M}{P_{t_{i+1}}^{(f)}}$ and

$$\begin{aligned} x_{t_i}^{(s)} &= x_{t_i}^{(a)} + K_{t_i}^{(s)}\left(x_{t_{i+1}}^{(s)} - x_{t_{i+1}}^{(f)}\right), \\ P_{t_i}^{(s)} &= P_{t_i}^{(a)} + \left(K_{t_i}^{(s)}\right)^2\left(P_{t_{i+1}}^{(s)} - P_{t_{i+1}}^{(f)}\right) \end{aligned}$$

Finally $U(\theta|\hat{\theta}_k)$ can be computed from the quantities computed with the Kalman smoother above and the relations

$$\begin{aligned} x_{t_i,t_i}^{(s)} &= P_{t_i}^{(s)} + \left(x_{t_i}^{(s)}\right)^2, \\ x_{t_{i-1},t_i}^{(s)} &= \mathrm{Cov}(X_{t_{i-1}}, X_{t_i}|y_{t_1}^{t_n}; \hat{\theta}_k) + x_{t_{i-1}}^{(s)} x_{t_i}^{(s)} \end{aligned}$$

where

$$\begin{aligned} \mathrm{Cov}(X_{t_{i-1}}, X_{t_i}|y_{t_1}^{t_n}; \hat{\theta}_k) &= (1 - K_{t_i})MP_{t_{i-1}}^{(a)} \\ &\quad + \frac{P_{t_i}^{(s)} - P_{t_i}^{(a)}}{P_{t_{i-1}}^{(a)}}(1 - K_{t_i})MP_{t_i}^{(a)} \end{aligned}$$

M step

The function $U(\theta|\hat{\theta}_k)$ can be decomposed as

$$U(\theta|\hat{\theta}_k) = U_X(\lambda, \sigma^2|\hat{\theta}_k) + U_{Y|X}(R|\hat{\theta}_k)$$

where

$$\begin{aligned} &U_X(\lambda, \sigma^2|\hat{\theta}_k) \\ &= -(n-1)\log(\sigma) - \frac{1}{2}\sum_{i=2}^{n}\log(1 - \exp(-2\Delta_i\lambda)) \\ &\quad - \frac{1}{2\sigma^2}\sum_{i=2}^{n}\frac{x_{t_i,t_i}^{(s)} - 2\exp(-\Delta_i\lambda)x_{t_{i-1},t_i}^{(s)} + \exp(-2\Delta_i\lambda)x_{t_{i-1},t_{i-1}}^{(s)}}{1 - \exp(-2\Delta_i\lambda)} \end{aligned}$$

and

$$U_{Y|X}(R|\hat{\theta}_k) = -\frac{n}{2}\log(R) - \frac{1}{2R}\sum_{i=1}^{n}\left\{y_{t_i}^2 - 2y_{t_i}x_{t_i}^{(s)} + x_{t_i,t_i}^{(s)}\right\}$$

The second term $U_{Y|X}$ is similar to the case with regular sampling and the maximum is obtained for $R = R_{k+1}$ with

$$R_{k+1} = \frac{1}{n}\sum_{i=1}^{n}\left\{y_{t_i}^2 - 2y_{t_i}x_{t_i}^{(s)} + x_{t_i,t_i}^{(s)}\right\}$$

The first term U_X is specific to the case with irregular sampling and numerical optimisation procedures have been used to compute $(\lambda_{k+1}, \sigma_{k+1}^2)$ since we could not derive analytic expressions these quantities. Here the relation

$$\sigma_{k+1}^2 = \frac{1}{n-1}\sum_{i=2}^{n} \times \frac{x_{t_i,t_i}^{(s)} - 2\exp(-\Delta_i\lambda_{k+1})x_{t_{i-1},t_i}^{(s)} + \exp(-2\Delta_i\lambda_{k+1})x_{t_{i-1},t_{i-1}}^{(s)}}{1-\exp(-2\Delta_i\lambda_{k+1})}$$

has been used to transform the initial two-dimensional optimization problem into a simple one-dimensional optimisation problem and reduce computational time.

The EM algorithm has several well known limitations. First it may converge to a non-interesting local maximum of the likelihood function depending on the starting value $\hat{\theta}_0$, and thus it is important to provide realistic initial parameter values. Here we have used the estimates obtained using the method of moment described in Sect. 3.1 Indeed the various tests that we have done indicate that this method leads to robust estimates and generally provide a good starting value to the EM algorithm with low numerical cost (see Sect. 3.3). This is particularly useful to avoid numerical problem when fitting the model to a large number of data sets for regional studies such as the one performed in Sect. 4.4

Another limitation of the EM algorithm is its slow convergence near the maxima where using a standard optimization algorithm is generally far more efficient, at least when it is possible to compute the incomplete likelihood function quickly. For the model under consideration, the incomplete likelihood function is a sub-product of the Kalman filter since we have

$$p(y_{t_1}^{t_n};\theta) = \prod_{i=2}^{n} p(y_{t_i}|y_{t_1}^{t_{i-1}})$$

where the conditional distribution $p(y_{t_i}|y_{t_1}^{t_{i-1}})$ is a Gaussian distribution with mean $E(X_{t_i}|y_{t_1}^{t_{i-1}})$ and variance $\mathrm{Var}(X_{t_i}|y_{t_1}^{t_{i-1}}) + R$ and these quantities are computed recursively in the Kalman filter (see Sect. 3.2). Eventually, the gradient of the log-likelihood function could also be computed to accelerate the convergence of the numerical optimization procedure. In this work, we did not provide the gradient to the Matlab function used for the numerical optimization but we did not encounter any numerical problem and the computational efficiency was good enough.

Another advantage of switching from the EM algorithm to a quasi-Newton algorithm close to the maxima is that quasi-Newton algorithms provide an approximation of the Hessian of the log-likelihood function, and thus useful information on the variance of the ML estimates (see Sect. 3.3)

3.3 Simulations

In this section, the relative performances of ML and MOM estimates are assessed through simulations. More precisely, for various values of $n \in \{200, 300, \ldots, 2000\}$, we have simulated $N = 1{,}000$ sequences of length n using the scheme described below:

1. Simulate the time lags $(\Delta_i)_{i\in\{2,\ldots,n\}}$ as an i.i.d. sample from the empirical distribution of the time lags for satellite data (see Fig. 1).
2. Simulate the initial state x_{t_1} as a Gaussian variable with mean $x^{(b)}$ and variance B and then recursively $(x_{t_i})_{i\in\{2,\ldots,n\}}$ according to (3).
3. Simulate the observed process $(y_{t_i})_{i\in\{1,\ldots,n\}}$ using (1).

The following parameters values have been chosen for the numerical experiment: $\lambda = 0.5$, $B = \sigma^2 = 0.05$, $R = 0.5$ and $x^{(b)} = 0$. It corresponds to realistic values for the application discussed in the next section.

Then, for each simulated sequence the ML and MOM estimates have been computed. In practice, ML and MOM estimates have been computed using a quasi-Newton algorithm with the true values of the parameters as initial value. Although such initialization is not possible for practical applications, it permits to avoid convergence to non interesting local maxima of the likelihood function and a fair comparison of the two estimates. Figure 5 shows the empirical estimate of the bias and variance of the estimates computed from these simulations. As expected, the ML estimates generally outperform the MOM estimates in terms of both bias and variance. However, the MOM estimates give satisfactory results for the different values of n and have the advantage of being computed with low computational costs and less sensitive to the choice of realistic starting values than the EM algorithm. For comparison purpose, the variances computed from the inverse of the observed information matrix are also shown on Fig. 5. The agreement with the empirical variances of the ML estimates is generally good, especially for large sample size as expected from the general asymptotic theory for the ML estimates.

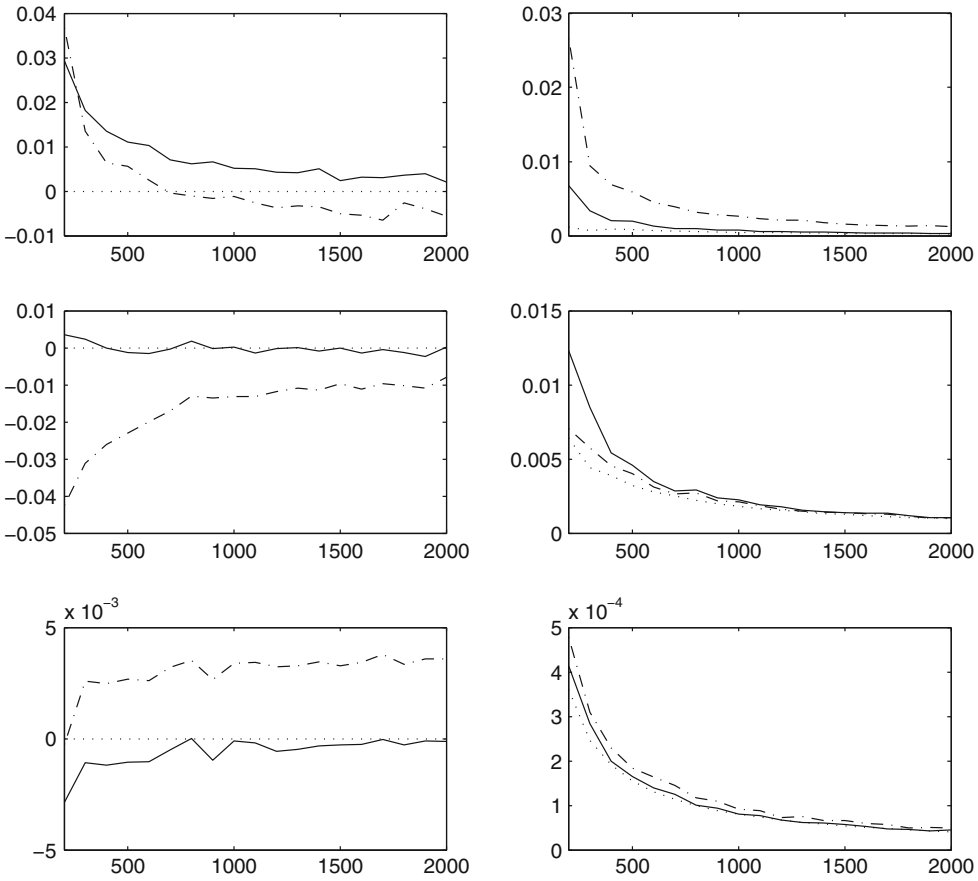


Fig. 5 Plot of the simulated bias (*left*) and variances (*right*) for the MOM (*dashed-dotted line*) and ML estimates (*full line*) for different length sequences n (x axis). Estimate of λ (*top panel*), of σ^2 (middle) and R (bottom). The *dotted lines* on the *right panel* is the variance computed from the information matrix (empirical mean over the different simulations). The simulated results are based on $N = 1{,}000$ replications

4 Application to SST data

In this section the model is first fitted and validated on the SST data introduced in Sect. 2. The original time series has been divided into two consecutive parts: the first one $(y_{t_1}, \ldots, y_{t_{n_1}})$ for estimating the parameters and second one $(y_{t_{n_1+1}}, \ldots, y_{t_n})$ for validating the model. In practice, we used $n_1 = 725$ observations to fit the model, a reasonable amount of data according to the simulation results given in Sect. 3.3 It corresponds to a proportion of about two-thirds of the data (more than 1 year).

In Sect. 4.1, we first discuss the results obtained when fitting the model on the training data set. Then the model is validated using cross-validation on the validation data set in Sect. 4.2 and by comparison to buoy data in Sect. 4.3 Finally, in Sect. 4.4, the methodology is applied to data at many locations on a regular grid covering the Atlantic ocean and the spatial behaviour of the parameter estimates is discussed.

4.1 Parameter estimation

The parametric covariance model (4) has been fitted to the empirical estimate of the autocovariance function of the SST anomaly using weighted least square method leading to the MOM estimates (see Sect. 3.1). The corresponding variograms are shown in Fig. 6. The overall agreement is good, except maybe a 5 days component which is visible on the empirical variogram function (see [12] for a discussion on the existence of peak frequencies in SST time series). This indicates that the assumptions made on the shape of the covariance function is realistic, at least when focussing to time lags up to 40 days. Let us remark that according to Fig. 1, it seems also reasonable to assume that the marginal distribution is approximately Gaussian except maybe the lower tail of the distribution.

Starting from the MOM estimates obtained by fitting the covariance function, we have run the EM algorithm. The first iterations are efficient and the likelihood function increases rapidly (see Fig. 7) but after some iterations the convergence becomes rather slow, and switching to a standard numerical optimisation procedure permits to save computational time. According to Table 1, the ML estimate of λ is significantly lower than the MOM estimate and the ML estimates of σ^2 and R^2 are higher than the corresponding MOM estimates, although the differences for σ^2 and R^2 do not seem to be statistically significant if we compare the differences in the parameter values to the

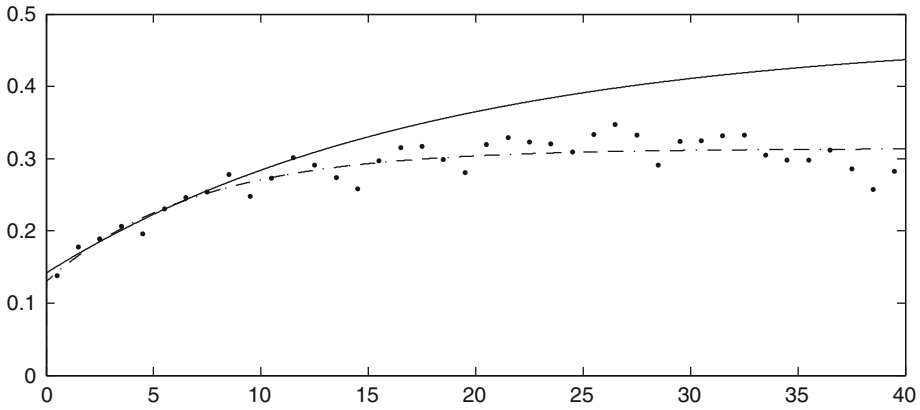


Fig. 6 Empirical (*dotted line*) and fitted theoretical variogram for the MOM (*dashed-dotted line*) and ML (*full line*) estimates. Results obtained on the training data set. The x axis is the time lag (in days)

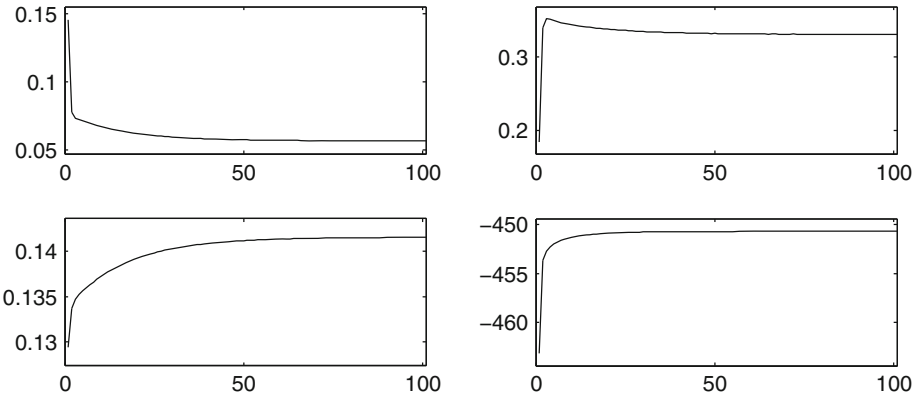


Fig. 7 Evolution of the parameters values during the 100 iterations (x axis) of the EM algorithm: $\hat{\lambda}$ (*top-left*), $\hat{\sigma^2}$ (*top-right*), $\hat{R}$ (*bottom-left*). The *bottom-right panel* shows the increase of the log-likelihood function

Table 1 Parameter value after the different steps of the fitting procedure: method of moment (first column), 100 iterations of the EM algorithm (second column) and numerical optimization of the likelihood function with a quasi-Newton algorithm (third column)

	Method of moments	Maximum likelihood		Standard deviation
		EM algorithm	Quasi-Newton	
$\hat{\lambda}$ (day^{-1})	0.145	0.057	0.056	0.019
$\hat{\sigma^2}$	0.184	0.329	0.330	0.094
$\hat{R}$	0.129	0.141	0.141	0.010
Log-likelihood	−463.15	−450.69	−450.68	

The last column gives an estimate of the standard deviation of the ML estimates computed from the information matrix. Results obtained on the training data set

standard deviations given in Table 1. ML estimates identify a second-order structure with a higher sill, which better coincides with the empirical variance of the time series (about 0.47), and also a higher range. Despite these differences in the parameters values, the agreement between the covariance functions is good for time lags less than 10 days (see Fig. 6) and thus we may expect that we would get similar results if using the model with the MOM instead of the ML estimates for estimating the true SST in Sects. 4.2 and 4.3.

The final parameter values are in good agreement with our knowledge of the physical process under consideration. In particular, according to [18], the standard deviation of the measurement error of the METOP data considered in this paper may vary between 0.33 and 0.51°C depending on the conditions. This range matches with the 95% confidence interval for the standard deviation of the observation error (we get approximately the interval between 0.35 and 0.40°C). Then, the low value of λ imply an important temporal persistence of the SST conditions and is coherent with the climatology of the place of interest were SST anomaly is known to have a strong temporal correlation. Finally, comparing the variance of the innovation of the dynamics for a time lag of 1 day ($\hat{Q}_1 = 0.04$) with the one of observation error $\hat{R}$ indicates that more weights will generally be given to the previous analysis than to the current observation in the Kalman recursions.

4.2 Cross-validation

In this section, we validate the model using cross-validation on the validation data set. For each $i \in \{n_1 + 1, \ldots, n\}$, the observation at time t_i is removed and the Kalman recursions are used to compute

$$x^{(s)}_{t_i|i}(\hat{\theta}) = E(X_{t_i}|y^{t_{i-1}}_{t_{n_1+1}}, y^{t_n}_{t_{i+1}}; \hat{\theta}), P^{(s)}_{t_i|i}(\hat{\theta}) = \mathrm{Var}(X_{t_i}|y^{t_{i-1}}_{t_{n_1+1}}, y^{t_n}_{t_{i+1}}; \hat{\theta})$$

If the various assumptions made in Sect. 2 are valid, then the conditional distribution of Y_{t_i} given the past observations $y^{t_{i-1}}_{t_{n_1+1}}$ and the future observation $y^{t_n}_{t_{i+1}}$ should be approximately Gaussian with mean $x^{(s)}_{t_i|i}(\hat{\theta})$ and variance $P^{(s)}_{t_i|i}(\hat{\theta}) + \hat{R}$. An histogram of the standardized residuals

$$\frac{y_{t_i} - x^{(s)}_{t_i|i}(\hat{\theta})}{\sqrt{P^{(s)}_{t_i|i}(\hat{\theta}) + \hat{R}}}$$

is shown on Fig. 8 together with the probability density function of the standard normal distribution and a normal probability plot (formal goodness of fit test are hard to implement since the residuals are not independent). The fit is generally good except again for the lower part of the distribution and this indicates that there are too many low residuals. According to Fig. 9, it corresponds to breaks in the observed time series at date when the SST anomaly suddenly drops. It is known that various factors (aerosol optical depth, wind speed or proximity to clouds for example) may perturb the quality of the data and a careful examination of these factors at the dates when the SST drops has been done. We could not identify anything special at these dates and thus we believe that the drops are due to non-linearities in the dynamics of the true SST anomaly. It indicates that using a non-linear model instead of (2) may be more appropriate. Let us remark that the standardized residuals may also provide useful information on outliers.

4.3 Comparison with buoy data

Using the model proposed in this work and the Kalman smoother on SST anomaly derived from satellite data, we can estimate the "true" SST anomaly at any time and thus emulate a virtual buoy. In order to check the realism of such virtual buoy, we have been compared the result with SST buoy measurements available at high temporal resolution (10 min) from the Pilot Research Moored Array in the Tropical Atlantic (PIRATA, see [23]) at the same location (0°N, 23°W). According to Fig. 10, the virtual buoy obtained by smoothing satellite data has some similarities with buoy data, but there are also important differences (only 63% of buoy measurements are contained in the 95% fluctuation intervals for the smoothing probabilities). However, Table 2 indicates that using the model proposed in this paper permits to improve the quality of the

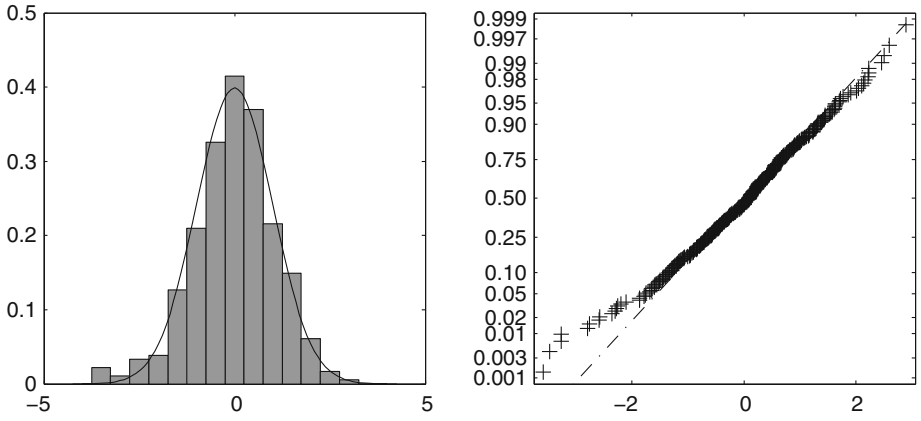


Fig. 8 *Left panel* histogram of the standardised residuals obtained by cross-validation on the validation data set and probability density function of the standard Gaussian distribution (*full line*). *Right panel* normal quantile–quantile plot of the standardised residuals

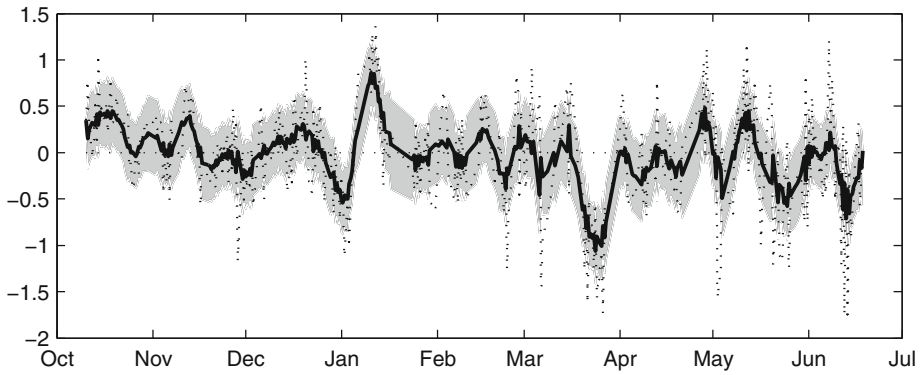


Fig. 9 Raw (*dotted line*) and interpolated (*full line*) satellite SST anomalies (in °C) together with a 95% fluctuation interval for the smoothing probabilities (*grey*). Results obtained by cross-validation on the validation data set

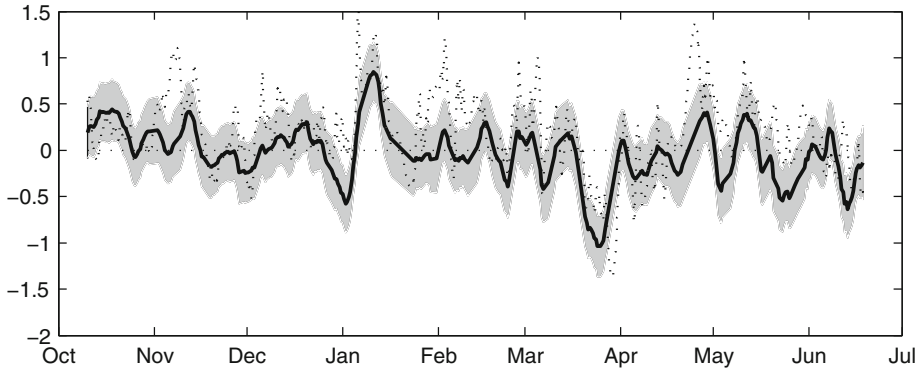


Fig. 10 Buoy SST anomalies (*dotted line*) and smoothed satellite SST anomalies (*full line*) in °C together with a 95% fluctuation interval for the smoothing probabilities (*grey*). Results obtained on the validation data set

Table 2 Difference between satellite (raw and smoothed) and buoy SST (bias, standard deviation and root mean square error) computed on the validation data set

	Bias	Standard deviation	RMSE
Raw satellite data	−0.22	0.47	0.52
Smoothed satellite data	−0.22	0.31	0.38

original satellite data and decrease the standard deviation of the error but can not correct the negative bias present in the original satellite data (underestimation of the SST measured at the buoy).

Since the results given in the previous sections indicate that the state-space model proposed in this paper is realistic for satellite data, we may conclude that the significant differences between the buoy and the virtual buoy are due to differences in the satellite and buoy data. A first reason may be the well know depth-to-skin bias discussed in [18]: METOP satellite measures the skin SST (the temperature of the sea in the first μm) whereas the buoy measures the temperature at a depth of about 1 meter and the temperature gradient evolves strongly in this surface layer. A second possible reason is the difference in the scale of the measurements: buoy data are local measurements and are able to identify small scale variation whereas METOP data describes larger scale variations since they retrieve the mean SST over a 5×5 km^2 surface.

Finally, these results highlight the difficulties of building a realistic SST time series from satellite data. Possible improvements are discussed in the conclusion.

4.4 Generalization to the Atlantic ocean

The methodology introduced above for the point with geographical coordinates (0°N, 23°W) has been applied to locations on a regular grid with 1° resolution in both latitude and longitude covering the Atlantic Ocean. The state-space model is fitted at each point on the time series of SST anomalies obtained by removing OIV2 analysis from METOP data. The length n of the time series depends on the location of interest and varies from 100 to 900 (see Fig. 11). According to the simulation results given in Sect. 3.3, this may lead to estimates with high variance at locations with poor satellite coverage.

The spatial behaviour of the parameter estimates shown on Fig. 11 gives important information on the small-scale variability of SST and also on the quality of METOP data and OIV2 analysis. First, the feedback parameter λ (expressed in day^{-1}) informs us about the heat transfer at the surface of the ocean. In order to facilitate the interpretation, we have chosen to represent the spatial evolution of $M_1 = \exp(-\lambda)$ which corresponds to the autoregression coefficient for a time lag of 1 day between two observations. The estimate of M_1 mainly depends on the latitude with longer range temporal dependence in the inter-tropical convergence zone (ITCZ) than in the mid-latitudes. Then, the variance of the stationnary distribution of the state σ^2 informs us about the variability of the SST anomaly. According to Fig. 11, the areas with high variability correspond to places, like the Falkland area off the Brazilian coast and the Gulf Stream off the Canadian coast, with strong sea-surface currents and wind conditions. Moreover, the more important upwelling systems of the Atlantic ocean can also be identified, e.g. the Canary and Benguela regions which are areas with strong winds yielding to a mixing of the ocean layer. In the rest of the Atlantic ocean, the variance is about 0.1. Finally, the value of the parameter R is the variance of the measurement errors of the METOP sensor. Estimate of this variance were provided in a previous study [18] by comparing METOP observations to data from drifting buoys. Unfortunately, the number of buoys is limited and covers a small part of ocean. The approach presented in this paper, based only on remotely sensed data, presents a global view of the spatial distribution of R. According to Fig. 11, the principal sources of contamination of METOP infra-red sensor seem to be the aerosol of the Saharan dust (see [9]) and the wildfire off the Angola coast.

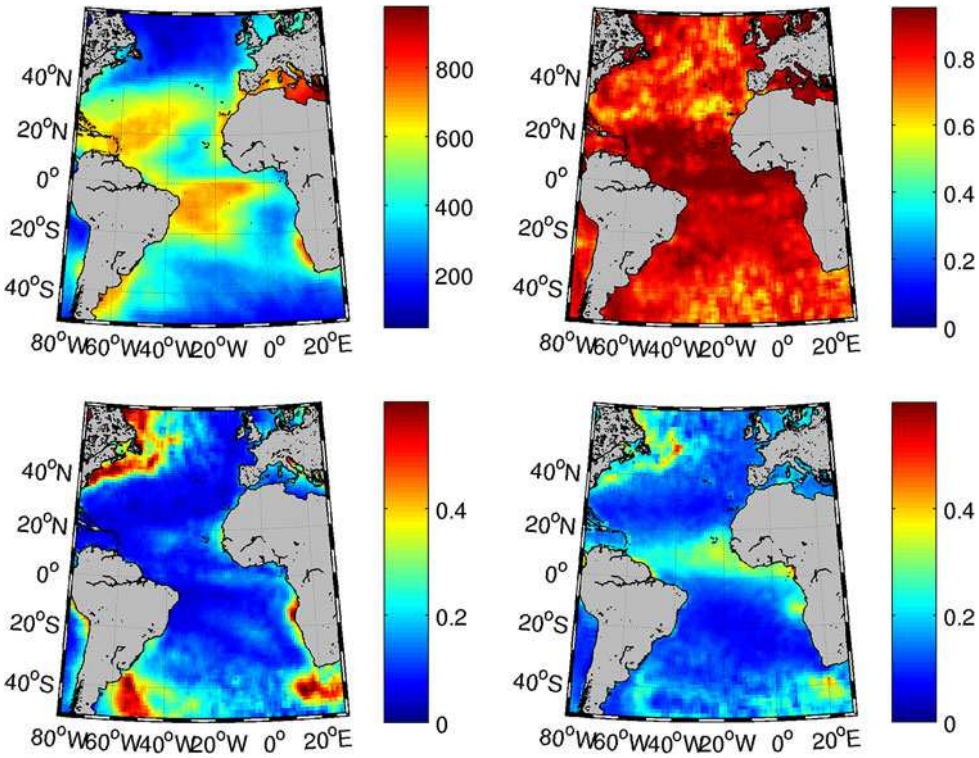


Fig. 11 *Top-left* number of METOP data at each grid point. *Top-right* spatial evolution of the estimate of the 1 day autocorrelation coefficient $\hat{M}_1 = \exp(-\hat{\lambda})$. *Bottom-left* spatial evolution of the estimate of the variance of the stationary distribution $\hat{\sigma}^2$. *Bottom-right* spatial evolution of the estimate of the variance of the measurement error R

5 Conclusion and perspectives

In this paper, we propose an extension of the usual linear and Gaussian state space model to analyse satellite data at irregular time step. We propose to combine various methods and algorithms to estimate the parameters efficiently. Indeed, simulation results indicate that the method of moment leads to a computationally efficient and numerically robust estimation procedure suitable for initializing the EM algorithm. A standard numerical optimization procedure is then used in the vicinity of the maximum of the likelihood function identified by the EM algorithm. It permits to accelerate the convergence of the EM algorithm with the extra benefit of giving as output an estimate of the information matrix which provide an estimate of the variance of the estimates.

This paper focus on SST data from the METOP satellite and the various results given in this paper indicate that the model is appropriate for describing some important properties of this data set such as the temporal structure and the measurement errors. Comparison with buoy data indicates that there is work to be done in order to estimate realistic SST conditions from METOP data. Nevertheless, we think that the state-space formulation adopted in this work is an appropriate method. In order to reconstruct realistic SST maps, we plan to extend the formulation in space and time to handle SST data from various satellites with their own accuracies and space-time resolutions. Indeed, using such formulation has several benefits. First, it allows modelling flexibility. For example, non-linear dynamics, which incorporate the effects of advection and diffusion (see [20] and references therein) or non-linear evolution in the atmospheric variability can be considered. We also plan to investigate more elaborated measurement equations and include covariates to model the changing biases and variances of the different satellites (see e.g. [25]). Then, the Markovian structure of the model leads to efficient methods for the statistical inference. In particular, it allows to compute the maximum likelihood estimates and it is shown that these estimates are more efficient than the ones obtained using the method of moment commonly used in geostatitics with kriging. The Kalman recursions used to compute the smoothing probabilities take also benefit of the Markovian properties of the model and permit to save computational time compared to space-time kriging where high dimensional linear systems need to be solved.

Acknowledgements We would like to thank the TAO Project Office of the National Oceanic and Atmospheric Administration/ Pacific Marine Environmental laboratory (NOAA/PMEL), the National Climatic Data Center (NCDC) and the Godae High Resolution Sea Surface Temperature Pilot Project (GHRSST-PP) for respectively providing in situ PIRATA, real-time OIV2 SST analysis and satellite METOP SST measurements. We are grateful to the Meteo-France Lannion Team, A. Bentamy and O. Talagrand for their expertise and valuable comments on this work.

References

1. Bertino L, Evensen G, Wackernagel H (2003) Sequential data assimilation techniques in oceanography. Int Stat Rev 71: 223–241
2. Cappé O, Moulines E, Rydén T (2005) Inference in hidden Markov models. Springer Verlag, New York
3. Cressie NAC (1993) Statistics for spatial data. Wiley, New York
4. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc B (Methodol) 39(1):1–38
5. Deng L, Shen X (1997) Maximum likelihood in statistical estimation of dynamic systems: decomposition algorithm and simulation results. Signal Process 57(1):65–79
6. Digalakis V, Rohlicek JR, Ostendorf M (1993) Ml estimation of a stochastic linear system with the em algorithm and its application to speech recognition 1(4):431–442
7. Donlon CJ, Minnett PJ, Gentemann C, Nightingale TJ, Barton IJ, Ward B, Murray MJ (2002) Toward improved validation of satellite sea surface skin temperature measurements for climate research. J Clim 15:353–369
8. Durbin J, Koopman SJ (2001) Time series analysis by state space methods. Oxford University Press, Oxford
9. Foltz GR, McPhaden MJ (2008) Impact of Saharan dust on tropical North Atlantic SST. J Clim 21:5048–5060
10. Frankignoul C, Hasselmann K (1977) Stochastic climate models. II Application to sea-surface temperature anomalies and thermocline variability. Tellus 29:289–305
11. Ghil M, Malanotte-Rizzoli P (1991) Data assimilation in meteorology and oceanography. Adv Geophys 33:141–266
12. Grodsky SA, Carton JA, Provost C, Servain J, Lorenzzetti JA, McPhaden MJ (2005) Tropical instability waves at 0 N, 23 W in the Atlantic: a case study using Pilot Research Moored Array in the Tropical Atlantic (PIRATA) mooring data. J Geophys Res 110(C8):C08010
13. Heemink AW, Segers AJ (2002) Modeling and prediction of environmental data in space and time using Kalman filtering. Stoch Environ Res Risk Assess 16(3):225–240
14. Ide K, Courtier P, Ghil M, Lorenc AC (1997) Unified notation for data assimilation: operational, sequential and variational. Practice 75(1B):181–189
15. Jeff Wu CF (1983) On the convergence properties of the em algorithm. Ann Stat 11(1):95–103
16. Jones RH, Boadi-Boateng F (1991) Unequally spaced longitudinal data with AR(1) serial correlation. Biometrics 47(1):161–175
17. Kaplan A, Cane MA, Kushnir Y, Clement AC, Blumenthal MB, Rajagopalan B (1998) Analyses of global sea surface temperature 1856–1991. J Geophys Res 103(18):567–589
18. Le Borgne P, Legendre G, Marsouin A (2007) Operational SST retrieval from MetOp/AVHRR. In: Proceedings of 2007 EUMETSAT conf., Amsterdam, The Netherlands
19. Penland C (1996) A stochastic model of IndoPacific sea surface temperature anomalies. Phys D Nonlinear Phenom 98(2–4):534–558
20. Piterbarg LI, Ostrovskii AG (1997) Advection and diffusion in random media: implications for sea surface temperature anomalies. Kluwer Academic Publishers, Dordrecht
21. Reynolds RW (1978) Sea surface temperature anomalies in the North Pacific Ocean. Tellus 30:97–103
22. Reynolds RW, Smith TM, Liu C, Chelton DB, Casey KS, Schlax MG (2007) Daily high-resolution-blended analyses for sea surface temperature. J Clim 20(22):5473–5496
23. Servain J, Busalacchi AJ, McPhaden MJ, Moura AD, Reverdin G, Vianna M, Zebiak SE (1998) A Pilot Research Moored Array in the Tropical Atlantic (PIRATA). Bull Am Meteorol Soc 79:2019–2032
24. Shumway RH, Stoffer DS (1982) An approach to time series smoothing and forecasting using the EM algorithm. J Time Ser Anal 3(4):253–264
25. Tandeo P, Autret E, Piolle JF, Tournadre J, Ailliot P (2009) A multivariate regression approach to adjust Aatsr sea surface temperature to in situ measurements. Geosci Remote Sens Lett IEEE 6(1):8–12
26. Wikle CK, Cressie N (1999) A dimension-reduced approach to space-time Kalman filtering. Biometrika 86:815–829