Traitement des Incertitudes en Océanographie

Mémoire présenté en vue de l'obtention
du Diplôme d'Habilitation à Diriger des Recherches
de L'Ecole Doctorale Terre Univers Environnement
(Université de Grenoble)

Jean-Michel Brankart

Ingénieur de Recherche CNRS

2014

Rapporteurs proposés :
Frédéric Chevallier
Gérald Desroziers
Pierre Lermusiaux

Table des matières

Introduction

In research if you know
what you are doing
you should not be doing it.
In engineering if you do not know
what you are doing
you should not be doing it.
Of course, you seldom, if ever,
see either pure state.

Richard Hamming (1997)

Depuis novembre 2001, je suis *ingénieur de recherche* (chef de projet en calcul scientifique) au sein de l'équipe MEOM (au LEGI puis au LGGE), sous la direction de Jacques Verron et Pierre Brasseur. Les activités dans lesquelles je suis impliqué portent essentiellement sur le développement et l'application des *méthodes d'assimilation de données* en océanographie. En raison de leur nature même, ces activités comportent une forte composante technologique, nécessitant un fonctionnement très collectif basé sur trois piliers : (i) des étudiants en thèse et des postdoctorants, qui sont au centre de tous nos projets, (ii) un encadrement scientifique et programmatique, assuré par J. Verron, P. Brasseur et E. Cosme, et (iii) un encadrement méthodologique et technique, dont je suis responsable (pour l'assimilation de données) avec le concours récent de P.-A. Bouttier. Cette façon de travailler ensemble donne une grande liberté à l'encadrement méthodologique et technique qui peut s'organiser selon une logique et un orientation propres, avec l'objectif de satisfaire globalement et sur le long terme les besoins de nos activités de recherche (voir aussi la discussion en annexe A). C'est dans cette optique que s'inscrit ma démarche vers l'habilitation à diriger des recherches, afin d'affirmer l'intérêt d'un encadrement méthodologique et technique spécifique, avec une vision et une direction autonomes, mené par un ingénieur de recherche.

Le texte de ce mémoire est donc entièrement dédié à présenter le point de vue méthodologique que nous avons progressivement adopté, ainsi que les perspectives de développement que ce point de vue nous offre pour l'avenir. En lien direct avec le développement des méthodes, les deux éléments connexes les plus importants sont assurément : (i) la question de leur adéquation aux problèmes scientifiques qui se posent, et (ii) la question de leur implémentation concrète sur une machine de calcul. Cependant, dans ce mémoire, aucune application scientifique ne sera jamais traitée en tant que telle, mais en tant qu'illustration du fonctionnement ou de la pertinence de notre approche méthodologique. L'objectif est de rester générique et de mettre en évidence la diversité des applications possibles dans des domaines variés. Par ailleurs, la question technique de l'implémentation des algorithmes ne sera pas abordée dans le texte principal, mais séparément en annexe A, afin de rassembler en un seul endroit les principaux enseignements et perspectives en la matière. Cela permettra, je pense, de montrer plus clairement l'importance de l'orientation technique de nos travaux. Mais avant d'en arriver là, il est temps d'introduire maintenant le sujet de ce mémoire, en montrant comment il s'est peu

à peu imposé au fur et à mesure de l'évolution de mon parcours professionnel et des projets auxquels j'ai participé.

Le sujet de ce mémoire

A l'entame de ma thèse de doctorat (en septembre 1992), le problème qui m'a été posé était de calculer une barre d'erreur pour les champs climatologiques de température et salinité en Méditerranée. L'équipe dans laquelle je travaillais (le GHER à l'université de Liège) disposait en effet d'un outil original et efficace pour calculer une climatologie à partir de données hydrographiques, basé sur la minimisation d'un principes variationnel (splines) et une discrétisation par éléments finis (Brasseur, 1991, 1994; Brasseur et al., 1996). Mais il lui était régulièrement reproché de ne pas associer de barre d'erreur aux climatologies ainsi calculées, et ce fut mon travail de résoudre ce problème (Brankart, 1996; Brankart and Brasseur, 1996, 1998). Pour cela, il a fallu associer un modèle statistique au principe variationnel utilisé jusque là, et trouver le moyen d'estimer les paramètres libres de ce modèle à partir des observations (notamment par validation croisée). Après mon départ de Liège (fin 1996), cette méthode a continué d'être utilisée et développée, et j'ai eu le plaisir d'y être parfois associé, jusqu'encore très récemment (Rixen et al., 2000; Beckers et al., 2002; Troupin et al., 2012). C'est de ce travail initiatique que provient ma conviction *qu'un modèle statistique des incertitudes est une composante essentielle de toute description vraiment réaliste de l'océan.*

Par la suite, au fil de mon parcours de postdoctorant, puis d'ingénieur de recherche, cette conviction n'a fait que se renforcer, notamment au contact des nombreux étudiants en thèse et postdoctorants dont j'ai eu la chance de partager le travail. Souvent, les incertitudes étaient perçues comme quelque-chose de gênant, dont il faut bien tenir compte d'une façon ou d'une autre, mais de préférence comme une superstructure aussi distincte que possible des résultats scientifiques proprement dits. Pourtant, au cours de nos travaux, il m'est apparu de plus en plus clairement que cette distinction est le plus souvent artificielle et impossible à réaliser en océanographie, et que l'incertitude ne peut pas être traitée de façon séparée. C'est un concept physique parmi les autres qui doit être intégré à toutes nos lois et à toutes nos théories. Par ailleurs, je me rends compte rétrospectivement que la plupart des développements méthodologiques et techniques que nous avons réalisés pendant toutes ces années découlent très naturellement de ce simple constat. C'est pourquoi j'ai décidé d'en faire le fil conducteur de ce mémoire, dont l'objectif est de montrer que *la modélisation des incertitudes est indubitablement appelée à jouer un rôle de plus en plus central en océanographie.* C'est un sujet qui est rarement considéré en tant que tel par les océanographes, mais qui me paraît riche d'enseignements et de perspectives, surtout s'il est abordé de façon cohérente et transverse en modélisation et en assimilation de données.

Comment ce sujet s'est peu à peu imposé

Historiquement, en océanographie, la question de la modélisation des incertitudes s'est d'abord posée aux scientifiques qui cherchaient à résoudre un problème inverse, d'abord pour l'analyse objective de données hydrographiques (Bretherton et al., 1976), et plus tard pour donner une 'formulation objective' aux problèmes d'assimilation de données (e.g. Bennett, 1992). Et c'est un peu selon cette même logique que la question des incertitudes s'est réinvitée dans mon parcours personnel quand, après une initiation à la modélisation océanique lors d'un postodctorat (1997–1988) à Bologne (Brankart and Pinardi, 2001), j'ai commencé un postdoctorat (en janvier 1999) dont le but était le contrôle de la circulation de l'Atlantique Nord par assimilation de données (dans le cadre

du projet européen pré-opérationnel DIADEM). L'équipe que j'ai rejointe alors (l'équipe MEOM du LEGI à Grenoble) venait en effet de développer une méthode originale et efficace pour résoudre ce genre de problème, basée sur un filtre de Kalman de rang réduit (le filtre SEEK) et une description des incertitudes par quelques composantes principales de la variabilité naturelle du modèle d'océan (Pham et al., 1998; Verron et al., 1999; Brasseur et al., 1999). Mais il est apparu assez rapidement que cette description réduite des incertitudes ne pourrait pas être assez riche et réaliste pour permettre le contrôle de l'écoulement à mésoéchelle en Atlantique Nord, notamment dans la région du Gulf Stream. Et l'essentiel de mon travail à ce moment-là fut de contribuer à résoudre ce problème (Brankart et al., 2003).

Pour cela, les deux éléments principaux ont été de compléter la description des incertitudes du filtre SEEK par un algorithme de localisation de leur covariance (afin d'éliminer les corrélations à grande distance irréalistes issues de la réduction d'ordre), et par un algorithme adaptatif pour leur variance (afin qu'elle s'ajuste automatiquement à l'écart aux observations). Heureusement, ces travaux n'ont pas été menés seul, mais en bonne synergie au sein du projet DIADEM (Carmillet et al., 2001; Brusdal et al., 2003), et surtout en très étroite collaboration avec deux étudiants en thèse travaillant sur des thématiques voisines (Testut et al., 2003; Parent et al., 2003). C'est grâce à cette collaboration que ces nouveaux développements —nettement plus lourds techniquement que le filtre SEEK original— ont pu être implémentés de façon cohérente dans un nouveau logiciel (SESAM, voir annexe A), que nous ne cessons de développer depuis lors. En terme d'application, cet outil nous a permis de construire un prototype de système opérationnel pour l'Atlantique Nord dans le cadre du projet DIADEM (Brankart et al., 2001). Et c'est d'ailleurs ce même algorithme qui a été ré-implémenté (par C.-E. Testut) pour construire la deuxième version du système d'assimilation de Mercator-Océan (SAM2), encore utilisé aujourd'hui tant pour le calcul de réanalyses que pour le système de prévision en temps réel.

C'est après ce second postdoctorat que j'ai été recruté au CNRS comme ingénieur de recherche (en novembre 2001), avec la responsabilité de poursuivre le développement des méthodes et des outils d'assimilation de données au sein de l'équipe MEOM (voir annexe A). Dès ce moment-là, les attracteurs les plus puissants de notre activité en assimilation de données ont sans aucun doute été les divers projets européens (TOPAZ, MERSEA, MyOcean, MyOcean2, SANGOMA) dans lesquels l'équipe s'est impliquée, ainsi que notre participation aux programmes d'observation altimétrique de l'océan (OSTST, SARAL/AltiKa). Ces projets ont ancré notre activité dans la réalité programmatique du moment (surtout orientée vers l'océanographie opérationnelle), et nous ont prémunis contre la tentation de dériver trop loin des applications océanographiques réelles de l'assimilation de données. Mais ma trajectoire a aussi suivi une dynamique propre. En raison de mon histoire personnelle, je crois que c'est beaucoup à cause de moi si, pour le meilleur ou pour le pire, la question de la modélisation des incertitudes est restée au cœur de nos préoccupations pendant toutes ces années.

En route vers l'océanographie opérationnelle. Au début, mon activité a d'abord été orientée presqu'exclusivement vers l'application, par la mise en œuvre des méthodes et des outils précédemment développés. Ce fut la période du projet européen TOPAZ (2002–2004), qui a fait suite au projet DIADEM, au cours duquel j'ai poursuivi la mise en œuvre du prototype de système opérationnel en temps réel (avec un autre modèle). Ce projet fut aussi l'occasion d'approfondir l'exploration du fonctionnement de la méthode d'assimilation (Rozier et al., 2007) [1] et d'examiner l'impact de données gravimétriques

1. C'est d'ailleurs ce même système qui a été récemment transféré au projet HYCOM et testé sur un modèle à haute résolution de Golfe du Mexique (Srinivasan et al., 2011).

précises sur l'assimilation d'observations altimétriques en Atlantique Nord (Birol et al., 2004, 2005). Un travail de thèse fut également mené par la suite pour étudier cette même question pour un modèle du Pacifique Tropical (Castruccio et al., 2006, 2008). Parallèlement au projet TOPAZ, nous avons aussi gardé des liens forts avec Mercator-Océan, en développant un algorithme d'incorporation progressive de l'incrément d'assimilation au modèle (Ourmières et al., 2006), et en commençant à examiner l'impact de l'assimilation de données dans un modèle couplé circulation/écosystème de l'Atlantique Nord (Berline et al., 2007; Ourmières et al., 2009). Durant cette première phase, l'attraction vers les études de nature totalement applicative fut clairement la plus puissante, et mon rôle d'encadrement des étudiants en thèse et postdoctorants fut essentiellement de faire le lien entre leur sujet d'étude et les méthodes utilisées.

Ensuite, nous nous sommes rendu compte que pour améliorer le système d'assimilation, il nous fallait mieux identifier et décrire les sources d'incertitudes présentes dans le modèle lui-même. C'est pour cette raison qu'au cours du travail de thèse de Grégoire Broquet, nous nous sommes pour la première fois orientés vers des simulations d'ensemble pour simuler explicitement l'incertitude sur le forçage atmosphérique dans un modèle à haute résolution du Golfe de Gascogne. Ce faisant, nous modifiions déjà la nature du modèle, en y incluant une dimension stochastique, dont nous nous sommes aperçus qu'elle peut affecter considérablement le comportement moyen des simulations (Broquet et al., 2008). Ce fut aussi la période du projet européen d'océanographie opérationnelle MERSEA (2004-2008), quand me fut déléguée la responsabilité scientifique d'une des tâches de recherche et développement du projet (en assimilation de données), et donc du travail de deux chercheurs postdoctorants recrutés pour cela (Sergey Skachko et Chafih Skandrani). Ces travaux nous ont conduits à généraliser le filtre SEEK —par la méthode du vecteur d'état augmenté— afin d'estimer des paramètres incertains du modèle à partir des observations. La méthode a été appliquée à l'estimation des paramètres du forçage atmosphérique, d'abord par des expériences idéalisées (Skachko et al., 2009), puis dans un contexte réaliste (Skandrani et al., 2009). Grâce à sa simplicité et à son efficacité (malgré le recours à des simulations d'ensemble), cette même approche a aussi été au cœur d'un travail de thèse récent en modélisation, afin de gagner de l'information sur la fonction de forçage du modèle à partir d'observations de température de surface (Meinvielle et al., 2013). Et c'est à peu près le même genre d'approche que nous avons utilisée pour résoudre le problème de l'assimilation de données dans des modèles emboîtés (dans le travail de Melet et al., 2012), grâce à un vecteur augmenté rassemblant simplement l'état du système sur les deux grilles de calcul.

Parallèlement, nous avons aussi appliqué nos méthodes d'assimilation de données à l'évaluation de systèmes d'observation altimétrique de l'océan, dans le cadre de travaux scientifiques d'accompagnement aux mission satellitaires d'observation altimétrique. Cette activité, initiée par Verron (1990), a été poursuivie par un travail de thèse (Debost, 2004), afin d'intercomparer différents scénarios d'observation altimétrique d'un écoulement idéalisé à mésoechelle. Cette même question fut ensuite examinée dans des contextes plus réalistes par deux autres travaux de thèse. Le premier s'est orienté vers l'observation altimétrique des ondes tropicales d'instabilité et des tourbillons du Nord Brésil dans un modèle de l'Atlantique Tropical (Ubelmann et al., 2009, 2012) ; et le second, vers l'observation altimétriques de courants côtiers (rendue possible par la mission SARAL/AltiKa) dans un modèle du Golfe de Gascogne (Duchez et al., 2012). Ce sont ces études qui nous ont conduits à imaginer une méthode simple et efficace pour tenir compte des corrélations d'erreur d'observation avec le filtre SEEK (Brankart et al., 2009). Et c'est à la suite de cette publication que j'ai pris l'initiative de rationaliser et de publier les algorithmes de paramétrisation locale et adaptative des incertitudes (Brankart et al., 2010, 2011), que nous avions développés précédemment.

Vers une modélisation de plus en plus réaliste de l'incertitude. Dans toutes ces études cependant, c'est toujours le modèle gaussien qui était utilisé pour décrire les incertitudes. Ce modèle possède l'avantage de conduire à des mathématiques simples (algèbre linéaire), mais il est loin de toujours donner une représentation réaliste des incertitudes que l'on rencontre en océanographie. Nombre de variables ou de paramètres sont souvent contraints par des contraintes d'inégalité, ou bien dépendent les uns des autres de façon extrêmement non-linéaire. Pour améliorer la description des incertitudes, nous avons donc été amenés à généraliser le modèle statistique à la base du filtre SEEK, d'une part en utilisant des distributions de probabilité gaussiennes tronquées en lieu et place des distributions gaussiennes (Lauvernet et al., 2009), et d'autre part, en introduisant des changements de variable non-linéaires (transformations anamorphiques, diagnostiquées à partir d'une prévision d'ensemble) pour améliorer la validité de l'hypothèse gaussienne (Béal et al., 2010). Ce fut la période des projets d'océanographie opérationnelle MyOcean (2008–2011), puis MyOcean2 (2011–2014), qui nous ont permis de mesurer à quel point une description réaliste des incertitudes est indispensable aux applications concrètes (surtout en biogéochimie marine). Ce fut un ingrédient nécessaire au succès des expériences d'estimation de paramètres biogéochimiques (par la méthode du vecteur d'état augmenté) pour un modèle couplé circulation/écosystème de l'Atlantique Nord (Doron et al., 2011, 2013). Et ce fut aussi d'une grande utilité lors de la réanalyse de l'écosystème de l'Atlantique Nord calculée par Fontana et al. (2013). En particulier, l'algorithme des transformations anamorphiques que nous avons développé m'a paru suffisamment intéressant et original pour justifier une publication séparée (Brankart et al., 2012), en illustrant son intérêt pour l'ensemble des problèmes auxquels nous l'avons appliqué.

Simultanément, la richesse grandissante de l'observation spatiale nous a aussi conduits à explorer dans quelle mesure les images à haute résolution de température de surface ou de couleur de l'océan (résolvant la sous-mésoéchelle) ne pouvait détenir une information précieuse sur la circulation superficielle (du moins à mésoéchelle). L'advection à mésoéchelle détermine en effet de façon souvent contraignante la structure fine des champs de traceur qui y sont soumis. C'est donc la structure des images observées qui contient l'information pertinente, et c'est elle que nous avons tenté d'exploiter pour gagner autant d'information que possible sur le champ de vitesse de surface (Titaud et al., 2011; Gaultier et al., 2013). Cette étude sera d'un intérêt particulier pour ce mémoire car l'incertitude que le champ de vitesse produit sur la structure du traceur ne peut se conformer à aucun des modèles utilisés jusque-là, ce qui nous a emmenés vers l'exploration et l'application de méthodes plus générales.

Voilà donc comment, au cours de nos travaux d'assimilation de données, s'est peu à peu imposée l'idée que le *réalisme* d'une prévision exige d'admettre explicitement qu'elle est *incertaine*; et d'associer un modèle statistique aussi fin que possible des différentes sources d'incertitude, qu'il s'agisse de la condition initiale, des paramètres ou du forçage. Mais il restait à mes yeux un élément encore trop rigide sur le tableau que nous avions élaboré. Les lois dynamiques du modèle de circulation ou de l'écosystème demeuraient un carcan parfaitement déterministe pour nos simulations. C'est donc pour nous libérer de cette contrainte excessive que j'ai commencé à imaginer un moyen générique (inspiré d'études existantes en météorologie, voir chapitre 3) pour simuler explicitement dans le modèle les sources d'incertitude qui nous paraissent dominantes. Une première tentative d'application, simulant l'incertitude que les échelles non-résolues produisent sur le calcul de la densité, montre que ce genre de paramétrisation stochastique peut produire un effet moyen déterminant sur la structure de la circulation (Brankart, 2013).

Encadré 1 : Problème direct (extrait de Brankart, 1996)

Un dé qui rebondit sur un tapis de jeu, une bille qui tourbillonne dans le tambour d'une roulette, voilà deux systèmes mécaniques qu'il pourrait être intéressant de modéliser. A la suite de Newton, on peut décrire ces deux systèmes par un modèle mathématique, *qui permet d'exprimer la réalité physique par une structure mathématique adéquate. L'approche newtonienne consiste à pouvoir représenter l'état du système par une ensemble de variables habilement choisies, pour lesquelles on écrit des équations d'évolution. (. . .) Il s'agit là d'une démarche fructueuse. Elle a pu être appliquée utilement à un très grand nombre de systèmes. Parmi eux, les* systèmes marins, *comme la plupart des systèmes naturels, se distinguent par leur très grande complexité. (. . .) Ainsi, pour appréhender pas à pas l'origine de leur comportement, n'a-t-on d'abord envisagé que des modèles partiels et imparfaits. (. . .) Puis, au fur et à mesure des progrès de ces recherches, des modèles capables de tenir compte correctement d'un vecteur d'état de plus en plus représentatif sur des géométries de plus en plus réalistes ont peu à peu vu le jour. L'ambition des modèles est devenue prognostique, ils se sont voulus capables de s'identifier à l'essentiel de la complexité naturelle pour pouvoir en reproduire et en prédire l'évolution.*

Hélas, quels que soient leur complexité ou leur objet, de tels modèles sont frappés de deux types d'indétermination, qui restent inéluctables tant que l'on étudie un problème réel. Il y a d'abord les indéterminations structurelles, liées aux hypothèses simplificatrices sur lesquelles le modèle est basé. Ces approximations sont d'autant plus nombreuses, indispensables et difficilement justifiables que les phénomènes qui entrent en jeu sont le résultat de combinaisons complexes d'événements plus simples, c'est-à-dire loin de la physique fondamentale dans la hiérarchie des phénomènes de la nature. Un modèle économique, par exemple, en nécessitera beaucoup plus qu'un modèle écologique, qui se situe encore bien loin d'un modèle hydrodynamique. (. . .) D'autre part, dans le cas d'un problème réel, un nouveau type d'approximation apparaît inévitablement. Il s'agit des indéterminations statistiques liées à l'impossibilité d'avoir accès de façon exacte aux conditions initiales et aux limites.

Le poids de ces deux types d'imprécision serait néanmoins bien léger si nous pouvions être sûrs, qu'au fil du temps et à travers l'espace, propagées par la dynamique d'un modèle réaliste, leur effet perturbateur ne s'amplifiait pas. Ceci n'est cependant généralement pas garanti. Même un système non-linéaire dissipatif apparemment simple peut devenir rapidement chaotique, avec comme conséquences principales l'amplification des erreurs initiales et une limitation intrinsèque à la prédictibilité des modèles. Pour Kalman (1994), "un objet est aléatoire s'il est impossible d'éliminer la 'non-unicité' de son comportement quand toutes les régularités qui le caractérisent ont été prises en compte." Bien sûr, il est impossible de dire si le comportement du dé ou de la mer résulte fondamentalement du hasard ; par contre, devant l'impasse à laquelle conduit inévitablement l'approche déterministe, nous sommes en droit de les traiter comme tels. La mer est comme un dé dont on ne peut saisir la subtilité des mouvements. Que sa dynamique ou les conditions de son évolution soient quelque peu inaccessibles et c'est à tout son comportement que nous pouvons donner une dimension aléatoire. (. . .)

Le problème direct *a pour objectif de reproduire ou de prédire des faits observables à partir d'un modèle théorique. (. . .) Une fois que le modèle est construit, dès qu'il a été identifié au système naturel, (. . .) ou mieux, que l'on dispose d'une description statistique ou probabiliste de la qualité de la solution, il doit être confronté aux observations. (. . .) C'est l'approche scientifique de base dans sa justification épistémologique la plus claire. Cette méthode n'est cependant pas toujours suffisante pour résoudre tous les problèmes pratiques.*

Organisation de ce mémoire

Chronologiquement, c'est donc pour résoudre un problème inverse que le concept d'incertitude a d'abord été nécessaire. Et ce n'est que très récemment que le même concept a été repris pour résoudre un problème direct, principalement jusqu'ici pour améliorer le réalisme de prévisions d'ensemble météorologiques ou climatiques (Buizza et al., 1999; Palmer et al., 2005). On voit d'ailleurs bien que ce transfert de concept a été fortement favorisé là où le développement de l'assimilation de données était le plus solidement ancré, et là où le souci d'une description réaliste de l'atmosphère ou de l'océan était le plus fort. Il me semble cependant que, pour mieux tirer les enseignements de mon parcours, et montrer l'importance de la modélisation des incertitudes, cet ordre chronologique doive être renversé. Je commencerai donc par le problème direct, ce qui correspond à nos études les plus récentes (paramétrisations stochastiques, simulations d'ensemble). Et je terminerai par le problème inverse, ce qui correspond à notre cadre d'activité historique (assimilation de données, filtre SEEK,...). C'est dans cet ordre que l'aspect transverse de la question, entre modélisation et assimilation de données, apparaîtra le plus clairement, et que les idées se succéderont le plus logiquement. Car j'écris aussi ce mémoire en espérant qu'un texte synthétique, replaçant nos travaux dans un canevas un peu différent, pourra également être utile aux étudiants ou postdoctorants avec qui nous travaillons.

Ce mémoire se divise donc en deux parties : la première traite du problème direct et la seconde du problème inverse. Afin d'introduire brièvement ce que sont le problème direct et le problème inverse en océanographie, j'ai trouvé utile de reproduire ici partiellement (voir encadrés 1 et 2) deux petits textes extraits de l'introduction de ma thèse de doctorat. Dans les deux cas, c'est le rôle de l'incertitude et l'aspect statistique de chaque problème qui étaient déjà placés au centre de la description. Ils me paraissent donc être encore une introduction tout-à-fait pertinente au contenu des deux parties de ce mémoire.

La première partie (problème direct) se divise en quatre chapitres. Le premier chapitre présente le modèle océanique utilisé dans ce travail, en insistant particulièrement sur les sources d'incertitude et sur la signification statistique qu'il conviendrait de donner au modèle. Le deuxième chapitre traite des simulations d'ensemble, permettant de décrire explicitement l'évolution de la distribution de probabilité des incertitudes. Cette méthode est illustrée par des exemples issus de nos travaux impliquant des simulations d'ensemble (Broquet et al., 2008; Skachko et al., 2009; Skandrani et al., 2009; Béal et al., 2010; Doron et al., 2011, 2013; Meinvielle et al., 2013; Gaultier et al., 2013). Le troisième chapitre traite des paramétrisations stochastiques, permettant de simuler explicitement l'incertitude sur le modèle. Dans ce chapitre, nous nous concentrerons surtout sur l'effet moyen qu'elle produit sur les simulations, en l'illustrant en particulier par l'exemple de l'incertitude sur le calcul de la densité (Brankart, 2013). Le quatrième chapitre présente brièvement les observations océaniques, en insistant sur les sources d'incertitude et la manière de les décrire.

La deuxième partie (problème inverse) se divise en trois chapitres. Le cinquième chapitre introduit l'approche bayésienne, qui impose de disposer d'un modèle a priori de nature probabiliste, c'est-à-dire incluant une description des incertitudes. Ce chapitre met surtout l'accent sur la complexité algorithmique du problème inverse en océanographie, et sur la nécessité de recourir à des hypothèses simplificatrices. La sixième chapitre traite du modèle gaussien, comme l'une des hypothèses permettant de rendre le problème traitable. Il explique aussi pourquoi il est en outre nécessaire de réduire la dimension du problème (filtre SEEK), et comment cela conduit aux algorithmes que nous avons développés (en résumant les méthodes présentées dans Brankart et al., 2009, 2010, 2011).

Encadré 2 : Problème inverse (extrait de Brankart, 1996)

*Quand il s'agit d'*appliquer *les connaissances nouvelles pour un objectif pratique bien défini, si nous voulons gagner à la roulette ou savoir s'il pleuvra demain, l'approche directe ne suffit généralement pas. (...) Un élément supplémentaire de réponse à ces questions est lié à la possibilité de pouvoir résoudre le* problème inverse *: pouvoir calculer les paramètres du modèle à partir de variables observées. En règle générale, un tel problème sera non seulement fondamentalement indéterminé, il sera aussi mal posé au sens de Hadamard (Tarantola, 1987, d'après). Pour lui, de tels problèmes n'ont simplement pas de sens physique. Aujourd'hui, cependant, il est généralement admis que ces problèmes mal posés admettent des extensions acceptables (Tarantola, 1987). Pour celles-ci, néanmoins, il faut toujours postuler une certaine connaissance a priori des inconnues que l'on recherche.*

De nouveau, une façon particulièrement efficace de l'exprimer dans le cas d'un système naturel est d'utiliser une formulation statistique ou probabiliste. La solution provient alors de la 'meilleure combinaison' possible entre l'information a priori sur le modèle et l'information qui provient des données. Il s'ensuit un 'contrôle optimal' des paramètres du modèle combinant au mieux les contraintes physiques sur la valeur qu'ils peuvent prendre et la représentation du sytème tel qu'il est observé. Ce que l'on a coutume d'appeler 'assimilation de données' *en météorologie et en océanographie peut également se fonder sur ce type de considération (...). Cette technique conduit surtout à réduire les deux types d'incertitude. D'une part, les observations sont partielles et imprécises : on ne peut jamais savoir exactement sur quelle case de la roulette la bille se stabilise (indéterminations statistiques). D'autre part, elle permet d'utiliser au mieux les atouts de l'approche théorique, tout en en lissant les déficiences (indéterminations structurelles). Comme en météorologie, elle constitue certainement la clef de voûte de l'océanographie opérationnelle. Dans cette optique, néanmoins, l'approche inverse est indissociable du problème direct. Ce sont les progrès des modèles mathématiques qui sous-tendent les succès opérationnels. Et c'est le réalisme des résulats de l'approche inverse qui permet de les jauger correctement tout en stimulant l'apparition de nouvelles idées.*

Le septième chapitre décrit les difficultés du modèle gaussien, et montre par quels moyens il est parfois possible de s'en affranchir, tout en s'assurant que le problème reste traitable numériquement en situation réaliste. Ces moyens sont illustrés par des exemples issus de nos travaux en assimilation de données (Lauvernet et al., 2009; Béal et al., 2010; Doron et al., 2011, 2013; Brankart et al., 2012; Fontana et al., 2013; Gaultier et al., 2013).

La plupart des chapitres sont écrits de façon à donner une vision d'ensemble de l'approche méthodologique, en amenant progressivement les perspectives de développement futur (notamment dans les chapitres 2, 3, 4 et 7) qui seront synthétisées dans la conclusion de ce mémoire.

Première partie

Problème direct

Chapitre 1

Incertitudes sur les modèles

> Since important decisions must rely
> on simulation, it is essential
> that its validity be tested, and
> that its advocates be able
> to describe the level of authentic
> representation which they achieved.
>
> Summer Computer
> Simulation Conference (1975)
> (d'après Richard Hamming, 1997)

Le problème direct consiste à prédire le comportement d'un système océanique en utilisant un modèle mathématique du système étudié. Un des aspects les plus importants de cette démarche est de concevoir un modèle qui soit à la fois suffisamment simple pour que le problème reste traitable numériquement, et suffisamment proche de l'océan réel pour reproduire les quantités observables avec assez de précision. L'objectif de ce premier chapitre est de présenter les modèles mathématiques et numériques de l'océan qui seront utilisés pour réaliser les études présentées dans ce mémoire. Nous nous attarderons en particulier sur les approximations qui sont faites et sur l'incertitude qui en résulte.

1.1 Le modèle NEMO

Le modèle numérique mis en œuvre dans nos travaux est le modèle NEMO (Nucleus for European Modelling of the Ocean), tel que décrit par Madec and the NEMO team (2008). NEMO regroupe plusieurs modèles décrivant les différentes composantes du système océanique, en particulier un modèle de circulation océanique (OPA), un modèle de glace de mer (LIM), ainsi que des modèles d'écosystème de complexités diverses. Décrivons d'abord le modèle mathématique sur lequel chacune des composantes du modèle numérique repose.

Equations primitives. Tout d'abord, le modèle de circulation océanique (OPA) se fonde sur les équations primitives :

– l'équation de conservation de la quantité de mouvement :

$$\frac{\partial \mathbf{U}_h}{\partial t} = -\left[(\nabla \times \mathbf{U}) \times \mathbf{U} + \frac{1}{2}\nabla(\mathbf{U}^2)\right] - f\,\mathbf{k} \times \mathbf{U}_h - \frac{1}{\rho_0}\nabla_h p + \mathbf{D}^U + \mathbf{F}^U \quad (1.1)$$

où t est le temps ; $\mathbf{k}$, le vecteur unitaire vertical (dirigé vers le haut) ; $\mathbf{U}$, le vecteur vitesse ($\mathbf{U}_h$ est la composante horizontale, orthogonale à $\mathbf{k}$, et w, la compo-

sante verticale) ; p est la pression ; ρ_0, une densité de référence ; et $f = 2\boldsymbol{\Omega} \times \mathbf{k}$, l'accélération de Coriolis ($\boldsymbol{\Omega}$ est la vitesse angulaire terrestre) ;

– l'équilibre hydrostatique :

$$\frac{\partial p}{\partial z} = -\rho g \tag{1.2}$$

où z est la coordonnée verticale (dans la direction de $\mathbf{k}$) ; ρ est la densité *in situ* ; et g, l'accélération gravitationnelle ;

– l'incompressibilité :

$$\nabla \cdot \mathbf{U} = 0 \tag{1.3}$$

– les équations d'évolution pour la température et la salinité :

$$\frac{\partial T}{\partial t} = -\nabla \cdot (T\,\mathbf{U}) + D^T + F^T \tag{1.4}$$

$$\frac{\partial S}{\partial t} = -\nabla \cdot (S\,\mathbf{U}) + D^S + F^S \tag{1.5}$$

où T est la température potentielle et S, la salinité ;

– l'équation d'état :

$$\rho = \rho\,[T, S, p_0(z)] \tag{1.6}$$

où $p_0(z) = \rho_0 g z$ est la pression de référence en fonction de la profondeur. Dans les applications réalistes de NEMO, l'équation d'état est l'équation standard empirique définie par le 'Joint Panel on Oceanographic Tables and Standards', dans une version qui a été reformulée par Jackett and McDougall (1995) pour permettre le calcul direct de la densité *in situ* à partir de la température potentielle (au lieu de la température *in situ*).

Dans ces équations, $\mathbf{D}^U$, D^T, D^S représentent la paramétrisation des échelles non-résolues pour la quantité de mouvement, la température et la salinité, et $\mathbf{F}^U$, F^T, F^S sont des termes de forçages. La grande variété des formulations qui existent dans NEMO pour paramétrer ces différents termes montre déjà à quel point il est difficile de les décrire par des lois de comportement qui soient à la fois générales, simples et fiables.

Ces équations sont complétées par des conditions aux limites, appliquées au fond de l'océan et à l'interface avec l'atmosphère. Les conditions cinématiques conduisent à la condition d'imperméabilité du fond de l'océan :

$$w = -\mathbf{U}_h \cdot \nabla_h H \tag{1.7}$$

où H est la profondeur de l'océan, et à une équation pronostique pour l'élévation de surface η :

$$\frac{\partial \eta}{\partial t} = -\nabla \cdot \left[(H + \eta)\,\overline{\mathbf{U}}_h \right] + P - E \tag{1.8}$$

où $\overline{\mathbf{U}}_h$ est la vitesse horizontale moyenne entre le fond et la surface ; P, les precipitations ; et E, l'évaporation. Les conditions dynamiques conduisent à une paramétrisation des échanges de quantité de mouvement et de chaleur à travers le fond et à travers la surface. Ces conditions dépendent de la configuration du modèle (voir section 1.3).

Modèle de glace de mer. Ensuite, le modèle de glace de mer (LIM, Louvain-la-Neuve sea Ice Model) se fonde sur une description discrète de la distribution (sous-maille) des épaisseurs de glace en fonction des coordonnées spatiales et du temps. Pour chaque catégorie d'épaisseur (une seule dans LIM2, plusieurs dans LIM3), l'état du système est décrit par (i) la fraction de la surface couverte par de la glace de cette catégorie, (ii) le volume de glace que cela représente (relié à l'épaisseur moyenne pour cette catégorie), (iii) le volume de neige, (iv) l'enthalpie de la glace, (v) l'enthalpie de la neige, (vi) le contenu en sel, et (vii) l'âge de la glace. Chacune de ces 7 variables est advectées par la vitesse de déplacement de la glace (supposée identique pour toutes les catégories d'épaisseur) qui résulte de l'effet combiné des vents et des courants, et qui est déterminée à partir d'une équation de conservation de la quantité de mouvement (en paramétrant le comportement de la glace comme un milieu continu visco-plastique ou élasto-visco-plastique). Ensuite, ces variables sont continuellement modifiées par de nombreux processus thermodynamiques : création de glace nouvelle en eau libre, croissance et fonte de la glace, conversion de neige en glace, désalinisation de la glace,... Ces processus dépendent principalement du bilan énergétique aux interfaces avec l'océan et l'atmosphère et des flux radiatifs et conductifs à travers les couches de neige et de glace ; et ils produisent en retour des échanges d'eau douce et de sel avec le modèle de circulation océanique. Enfin, quand le modèle résout plus d'une catégorie d'épaisseur, il faut également tenir compte de processus complexes de redistribution des ces 7 quantités (ridging, rafting) entre les différentes catégories.

Modèle d'écosystème. Quant au modèle d'écosystème, il se fonde sur l'écriture d'équations d'évolution pour un certain nombre de concentrations $C_i,\ i = 1, \ldots, n$ donnant une description synthétique de l'état local de l'écosystème marin. Ces constituants de l'écoystème sont d'une part advectés et diffusés comme des traceurs passifs par les champs de vitesse et de diffusivité turbulente produits par le modèle de circulation, et d'autre part produits et détruits selon des lois empiriques décrivant le fonctionnement de l'écosystème :

$$\frac{\partial C_i}{\partial t} = -\nabla \cdot (C_i\, \mathbf{U}) + D^{C_i} + F^{C_i} + SMS(C_i) \tag{1.9}$$

où D^{C_i} représente la diffusion par les échelles non-résolues, F^{C_i} est un terme de forçage, et $SMS(C_i)$ est le terme de production/destruction ('source minus sink') de chaque consituant.

Le modèle pourra être plus ou moins simple et synthétique ou complexe et détaillé selon le nombre n de concentrations utilisées pour décrire l'état de l'écosystème. Dans tous les exemples de ce travail, ce sera toujours le modèle LOBSTER (développé par Lévy et al., 2005) qui sera utilisé. C'est un modèle très synthétique qui décrit l'état de l'écosystème par $n = 6$ variables d'état : la concentration C_P en phytoplancton (quelles que soient l'espèce ou la taille), la concentration C_Z en zooplancton (quelles que soient l'espèce ou la taille), les concentration C_{NO_3} et C_{NH_4} en nitrate et en ammonium dissous, la concentration C_{MOD} en matière organique dissoute, et la concentration C_B en bactéries (toutes espèces confondues). Toutes ces concentrations représentent la quantité d'azote (exprimée en mmolN/m^3) que chacun de ce compartiment renferme et dont le total est conservé. Les termes de production/destruction $SMS(C_i)$ pour chacune de ces concentrations décrivent les échanges de matières entre les compartiments qui résultent du fonctionnement de l'écosystème (production primaire, broutage, mortalité, reminéralisation,...). La complexité de leur formulation (choix des variables pertinentes, nonlinéarités fonctionnelles, phénomènes à seuil) et le grand nombre de paramètres empiriques dont elles dépendent réflètent la difficulté qu'il y a à décrire le monde vivant

par des lois de comportements qui soient à la fois suffisamment simples et fiables.

1.2 Approximations et incertitudes

Approximations. Les équations primitives utilisées dans NEMO résultent d'une simplification des équations de la mécanique des fluides par essentiellement trois types d'approximation :

1. *Sur la géométrie de l'écoulement*, afin de simplifier le système de coordonnées et l'expression des principaux termes des équations. Parmi celles-ci, il y a principalement : (i) l'approximation de la terre sphérique, qui consiste à supposer que les surfaces géopotentielles sont des sphères ; et (ii) l'approximation de la couche mince, qui consiste à négliger la profondeur de l'océan par rapport au rayon terrestre.
2. *Sur la dynamique de l'écoulement*, afin d'éliminer a priori un certain nombre de processus rapides, qu'il serait coûteux de résoudre explicitement. Il y a : (i) l'approximation de Boussinesq, qui consiste à ne prendre en compte les variations de densité que dans la force de gravité, et qui élimine en particulier la propagation des ondes sonores ; et (ii) l'approximation hydrostatique, qui consiste à réduire la composante verticale de l'équation de conservation de la quantité de mouvement à l'équilibre hydrostatique (équation 1.2), et qui élimine a priori tous les processus de convection (qui doivent donc être paramétrés). En plus de cela, comme les équations sont non-linéaires (advection, équation d'état), les échelles non-résolues produisent sur l'écoulement à grande échelle un effet important qui doit être paramétré (voir plus loin).
3. *Sur la thermodynamique de l'écoulement*, afin de simplifier la description du système et des lois de comportement qui le régissent. Il y a (i) l'hypothèse d'équilibre local, qui consiste à supposer que les définitions et les principes de la thermodynamique de l'équilibre sont localement valides (à l'échelle d'un petit élément macroscopique du fluide) ; (ii) le fait de négliger les variations relatives de concentrations des différents sels dissous (en suivant le principe de Marcet ou loi de Dittmar), pour ne garder que la concentration totale (la salinité S) ; et (iii) l'utilisation de lois de comportement simplifiées pour l'eau de mer (équation d'état empirique, capacité calorifique constante).

Par ailleurs, il faut aussi mentionner les approximations dans le modèle de glace de mer et le modèle d'écosystème, qui sont principalement liées aux échelles et aux processus que le modèle ne résout pas et qui doivent être paramétrés (voir paragraphe suivant). Toutes ces approximations font que le modèle mathématique décrit en section 1.1 ne peut s'identifier exactement au système océanique réel (ou du moins à la partie du système qu'on cherche à résoudre).

Outre cela, le passage du modèle mathématique au modèle numérique introduit encore de nombreuses approximations supplémentaires. Par exemple, la discrétisation des différents termes des équations (advection, dérivée temporelle,...) produit une erreur de troncature dont les caractéristiques dépendent des schémas numériques utilisés. Le passage à un modèle discret impose auss souvent d'approximer la géométrie de l'écoulement, comme une bathymétrie en marches d'escaliers (en coordonnée z) ou dépourvue de seuils trop raides (en coordonnée sigma). Les approximations et incertitudes liées aux algorithmes et schémas numériques utilisés pour calculer une solution des équations forment une problématique en soi, qu'il n'est pas question de traiter dans ce mémoire, et dont l'aspect peut-être le plus important est de faire en sorte que le modèle numérique ne génère pas de comportements artificiels, complètement étrangers au modèle mathématique dont il est issu.

Quoi qu'il en soit, face à un écart de comportement entre le modèle et le monde réel, il n'y a que deux approches possibles. (i) Soit il faut faire évoluer le modèle jusqu'à réduire cet écart en deça de ce qui est observable. On comprend bien néanmoins que cette approche corrective ne peut pas être menée à son terme, et que des approximations importantes resteront pour longtemps nécessaires à la modélisation océanique. (ii) Soit il faut admettre explicitement que cet écart existe, et inclure dans le modèle (mathématique et numérique) une description de ce que cet écart peut être. Cela revient juste à dire qu'il convient de décrire le degré de représentation authentique que le modèle permet (voir citation en entête de chapitre).

Paramétrisations déterministes. L'une des caractéristiques les plus marquantes des modèles qui viennent d'être décrits est qu'il s'agit de modèles *déterministes*, c'est-à-dire que la connaissance de la condition initiale et du forçage détermine de façon univoque l'évolution future de l'état du système. [1] C'est de loin l'approche la plus communément adoptée pour construire les modèles d'océan actuels. Par ailleurs, ces modèles n'ont jamais vocation à résoudre toutes les échelles ou toute la diversité chimique et biologique présentes dans le système océanique réel. Pour construire un modèle déterministe, il faut donc représenter l'effet des échelles ou des espèces chimiques et biologiques non-résolues par une paramétrisation qui ne dépende que de ce que le modèle résout (voir l'encadré 3, page 16). Cette nécessité de "fermer" le système conduit en général aux hypothèses et aux approximations les plus fortes sur lesquelles le modèle se fonde.

Dans le modèle de circulation, l'effet des échelles non-résolues est en général représenté par une diffusion supplémentaire (diffusion turbulente) dans les opérateurs $\mathbf{D}^U$, D^T et D^S. Cette paramétrisation admet de nombreuses variantes dans NEMO et devra donc être précisée pour chacune des configurations du modèle. Par exemple, pour représenter la diffusion verticale, l'une des paramétrisations les plus utilisées consiste à ajouter l'énergie cinétique turbulente (k) dans la description de l'état du système (c'est-à-dire qu'on inclut k dans le système $\mathcal{A}$ de l'encadré 3, page 16) et à en déduire la diffusivité et la viscosité verticale (avec éventuellement quelques aménagements supplémentaire pour décrire la convection verticale, la double diffusion, l'effet de marées internes,...).

Dans le modèle d'écosystème, l'effet de la diversité non-résolue est intégré dans un ensemble de fonctions génériques (production primaire, broutage,...), dont la formulation et la paramétrisation ne devraient pas dépendre du détail (non-résolu) de l'état de l'écosystème. On voit bien néanmoins qu'il s'agit là d'une difficulté importante qui pousse de nombreux modélisateurs à s'orienter vers des modèles de plus en plus complexes en terme de nombre de variables d'état et de paramètres. Bref, la meilleure façon de fermer un modèle d'écosystème déterministe demeure l'objet de très vifs débats, au point qu'il paraît légitime de se demander si une approche alternative ne serait pas plus fructueuse (voir chapitre 3).

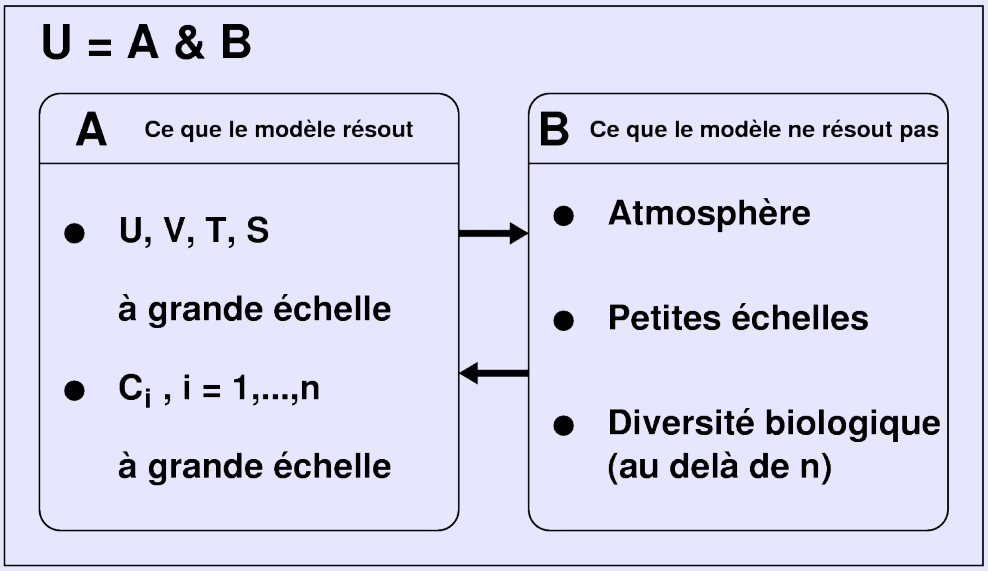
Erreurs de modélisation. Pour formaliser la discussion des incertitudes, on écrira le modèle sous la forme simplifiée :

$$\frac{d\mathbf{x}}{dt} = \mathcal{M}\left[\mathbf{x}, \mathbf{u}(t), \mathbf{p}(t)\right] \tag{1.10}$$

où $\mathbf{x}$ décrit l'état du système, $\mathbf{p}$, les paramètres du modèle et $\mathbf{u}$, le forçage. Pour NEMO, $\mathbf{x}$ contient donc les variables suivantes : $\mathbf{U}_h$, T, S et η pour le modèle de circulation, et $C_i,\ i = 1, n$ pour le modèle d'écosystème ; $\mathbf{u}$ contient la description du forçage

1. C'est du moins le cas du modèle mathématique, mais pas du modèle numérique, qui introduit des erreurs d'arrondi, dont les caractéristiques peuvent varier selon la précision numérique adoptée ou le calculateur utilisé. Ce phénomène est particulièrement important pour des systèmes chaotiques, dont l'évolution dépend de façon très sensible des conditions initiales.

Encadré 3 : Découpage entre système ($\mathcal{A}$) et environnement ($\mathcal{B}$)



Sur le schéma ci-dessus, le système $\mathcal{A}$ est le système dont on cherche à décrire la dynamique par un modèle $\mathcal{M}$. Il contient par exemple la circulation océanique à grande échelle, ainsi qu'une description synthétique de l'écosystème marin. Le système $\mathcal{A}$ interagit avec l'environnement $\mathcal{B}$ qui contient tout ce que le modèle $\mathcal{M}$ ne résout pas, par exemple l'atmosphère, les échelles non-résolues de l'écoulement océanique et la diversité non-résolue de l'écosystème.

Même quand la dynamique de l'union de ces deux systèmes $\mathcal{U} = \mathcal{A} \cup \mathcal{B}$ peut être supposée déterministe, ce ne sera en général pas le cas du système $\mathcal{A}$ considéré séparément. L'évolution future du système $\mathcal{A}$ n'est pas univoquement déterminée par sa condition initiale et par une dynamique qui lui est propre ; elle dépend aussi de ce qui se passe dans $\mathcal{B}$. Pour obtenir un modèle déterministe pour le système $\mathcal{A}$, il faut donc faire l'une des deux hypothèses suivantes. Soit il faut supposer que l'évolution du système $\mathcal{B}$ est connue. C'est ce qui est fait en général pour l'atmosphère afin d'imposer un forçage atmosphérique au modèle d'océan (à travers les opérateurs $\mathbf{F}^U$, F^T, F^S dans les équations 1.1, 1.4 et 1.5). Soit il faut supposer que l'effet de ce qui se passe dans $\mathcal{B}$ peut être paramétré en fonction de la seule description du système $\mathcal{A}$. C'est ce qui est fait en général pour paramétrer l'effet des échelles non-résolues sur l'écoulement à grande échelle (à travers les opérateurs $\mathbf{D}^U$, D^T, D^S dans les équations 1.1, 1.4 et 1.5).

Il est néanmoins important de garder à l'esprit qu'il s'agit toujours d'une approximation, plus ou moins forte selon le cas. En réalité, la dynamique du système $\mathcal{A}$ est non-déterministe. La traiter comme telle, en simulant explicitement les incertitudes provenant de $\mathcal{B}$, est à la base des paramétrisations stochastiques proposées au chapitre 3 et des simulations d'ensemble décrites au chapitre 2.

extérieur, c'est-à-dire essentiellement le forçage atmosphérique, le forçage géothermique et éventuellement les conditions aux frontières d'un domaine limité ; et $\mathbf{p}$ contient les paramètres physiques ou biologiques, comme par exemple la fréquence de Coriolis f, les paramètres de l'équation d'état (1.6) ou les paramètres gouvernant le fonctionnement de l'écosystème.

Sous l'hypothèse que le modèle décrive un sytsème réellement déterministe, il existe parallèlement un modèle idéal $\mathcal{M}^t$, un forçage idéal $\mathbf{u}^t$ et des paramètres idéaux $\mathbf{p}^t$ permettant de décrire parfaitement la trajectoire du système réel :

$$\frac{d\mathbf{x}^t}{dt} = \mathcal{M}^t \left[\mathbf{x}^t, \mathbf{u}^t(t), \mathbf{p}^t(t)\right] \tag{1.11}$$

On peut alors définit l'erreur de modélisation par l'écart :

$$\boldsymbol{\eta} = \mathcal{M} \left[\mathbf{x}^t, \mathbf{u}(t), \mathbf{p}(t)\right] - \mathcal{M}^t \left[\mathbf{x}^t, \mathbf{u}^t(t), \mathbf{p}^t(t)\right] \tag{1.12}$$

Ceci inclut donc l'erreur sur les lois du modèle $\mathcal{M} \left[\mathbf{x}^t, \mathbf{u}^t(t), \mathbf{p}^t(t)\right] - \mathcal{M}^t \left[\mathbf{x}^t, \mathbf{u}^t(t), \mathbf{p}^t(t)\right]$, l'erreur sur le forçage $\mathcal{M}^t \left[\mathbf{x}^t, \mathbf{u}(t), \mathbf{p}^t(t)\right] - \mathcal{M}^t \left[\mathbf{x}^t, \mathbf{u}^t(t), \mathbf{p}^t(t)\right]$, l'erreur sur les paramètres $\mathcal{M}^t \left[\mathbf{x}^t, \mathbf{u}^t(t), \mathbf{p}(t)\right] - \mathcal{M}^t \left[\mathbf{x}^t, \mathbf{u}^t(t), \mathbf{p}^t(t)\right]$, mais exclut l'erreur sur la condition initiale. Il vaut mieux d'ores-et-déjà les considérer toutes ensemble car, en pratique, ces différentes contributions ne peuvent que rarement être supposées agir indépendamment les unes des autres.

Par exemple, pour NEMO, l'erreur de modélisation résultera toujours de la combinaison d'effets multiples. Pour le modèle de circulation, on retiendra en particulier (i) l'ensemble des approximations inhérentes aux équations primitives (discutées en début de section), (ii) l'erreur sur le forçage atmosphérique dans $\mathbf{F}^U$, F^T et F^S (discutée au chapitre 2), et (iii) l'erreur sur la paramétrisation des échelles non-résolues, notamment dans $\mathbf{D}^U$, D^T et D^S (dont il sera question en section 3.1.3). Pour le modèle d'écosystème, on retiendra (i) l'erreur induite par le modèle de circulation lors de l'advection et de la diffusion des traceurs biogéochimiques (voir section 3.1.3), et (ii) les approximations liées à la formulation et à la paramétrisation des échanges de matières entre les compartiments du modèle d'écosystème (voir section 3.1.4).

Bien sûr, le modèle idéal de l'équation (1.11) reste toujours inconnu [2], et l'erreur de modélisation ne peut être décrite que par sa distribution de probabilité $p(\boldsymbol{\eta}; \mathbf{x}, t)$ qui représente l'incertitude du modélisateur quant à la validité de son modèle. Elle dépend en général de l'endroit où l'on se trouve dans l'espace de phase ($\mathbf{x}$) et du temps (t). Même pour un système et un modèle parfaitement déterministes, la modélisation de cette distribution de probabilité fait partie de la solution du problème direct. Cette distribution de probabilité est en effet nécessaire pour s'assurer de la compatibilité entre le modèle et les observations, et le cas échéant, pour invalider le modèle par l'observation (voir chapitre 4).

Incertitudes. Cependant, les systèmes océaniques que l'on étudie sont parfois très loin d'être déterministes [3] (voir l'encadré 3, page 16). Par exemple, lorsqu'on exclut les petites échelles du système étudié, l'état futur des échelles résolues (qui sont dans le système) peut ne pas seulement dépendre de leur état courant, mais aussi de l'état courant des échelle non-résolues (qui sont hors du système). Il s'ensuit que, même si

2. Sauf dans le cadre de problèmes idéalisés pour lesquels la solution réelle est supposée connue et où l'erreur est introduite artificiellement. C'est le cas notamment des expériences jumelles en assimilation de données (voir chapitre 6).

3. Même lorsqu'un système est vraiment déterministe (du moins en bonne approximation), il peut être utile de le décrire par un modèle non-déterministe pour modéliser les distributions de probabilité de l'erreur commise (voir chapitre 3).

l'état du système est connu de façon parfaite, son évolution future ne peut être prévue de façon univoque. Un phénomène semblable se produit quand on réduit la description d'un écosystème à un nombre restreint de concentrations C_i : l'évolution du système n'est pas univoquement déterminée par la condition initiale car elle dépend d'une description plus détaillée de l'état du système. Dans ce genre de cas, le modèle idéal $\mathcal{M}^t$ décrit par l'équation (1.11) n'existe pas. C'est pour cette raison que le concept d'erreur de modélisation, défini par l'équation (1.12), doit être remplacé par le concept plus général d'incertitude, qui englobe à la fois les erreurs de modélisation et l'incertitude liée au caractère non-déterministe du système étudié. Et comme ces deux composantes de l'incertitude n'agissent pas indépendamment, l'objectif du modélisateur ne pourra être que de les modéliser conjointement par la distribution de probabilité $p(\boldsymbol{\eta}; \mathbf{x}, t)$, qui caractérise la distribution de probabilité de la tendance en chaque point de l'espace de phase.

Même si cette distribution de probabilité apparaîtra le plus souvent indirectement ou implicitement dans les applications, elle jouera un rôle central dans l'ensemble de ce travail. Car elle nous renseigne sur la quantité d'information que le modèle contient à propos du système étudié. Moins la distribution est dispersée (c'est-à-dire plus son entropie est faible), plus le modèle contient d'information sur le système. Ainsi, deux modèles valides, c'est-à-dire deux modèles dont l'incertitude est compatible avec les observations [4], peuvent-ils être plus ou moins informatifs. L'objectif du modélisateur sera donc de produire un modèle valide, le plus informatif possible et au moindre coût.

1.3 Configurations du modèle

Pour terminer ce chapitre, voici maintenant une brève description des trois principales configurations du modèle NEMO qui seront utilisées tout au long de ce travail. Souvent cependant, nos travaux se sont basés sur d'autres configurations de NEMO, voire sur d'autres modèles d'océan (comme HYCOM, Bleck, 2002; Chassignet et al., 2003), qui seront parfois aussi utilisés pour illustrer certains chapitres de ce mémoire.

1.3.1 La configuration SEABASS

La configuration SEABASS est une configuration idéalisée du modèle NEMO (décrite par Cosme et al., 2010), décrivant un bassin océanique carré à fond plat (5000 m de profondeur) aux moyennes latitudes (entre 25°N et 45°N). Dans ce bassin fermé, une circulation en double gyre est créée par un vent zonal constant soufflant vers l'ouest au nord et au sud du bassin, et vers l'est aux latitudes intermédiares. Ces deux gyres sont caractérisés par des courants de bord ouest intenses, qui nourrissent un jet vers l'est au centre du bassin (voir fig. 1.1). En raison de l'instabilité de ce jet, l'écoulement dans le bassin est dominé par une circulation à mésoéchelle, caractérisée par des tourbillons d'environ 100 km de diamètre, des vitesses d'environ 1 m/s, et des différences de hauteur dynamique d'environ 1 m. Cette circulation est conçue pour que la turbulence de mésoéchelle qui en résulte soit semblable à ce qui est observé dans les régions du Gulf Stream (Atlantique Nord) et du Kuroshio (Pacifique Nord).

Cette configuration nous sert principalement de premier banc d'essai pour les méthodes et outils d'assimilation de données que nous développons (voir chapitre 6). C'est d'ailleurs dans cette optique qu'elle sera prochainement incluse aux configurations de référence du modèle NEMO (Bouttier et al., 2012).

4. C'est-à-dire, plus exactement, encore jamais mis en défaut par aucune nouvelle observation.

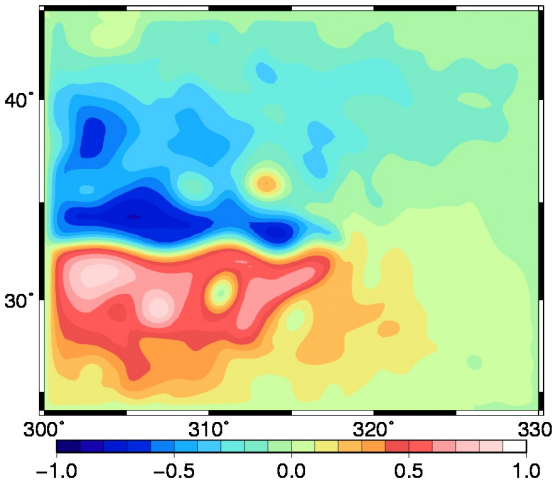


FIGURE 1.1 – Instantané de l'élévation de surface issu de la configuration SEABASS.

1.3.2 La configuration ORCA2

La configuration ORCA2 est une configuration globale à basse résolution du modèle NEMO (décrite par Madec and Imbard, 1996). Il s'agit d'une des configurations de référence de NEMO, qui discrétise les équations du modèle (éqs 1.1 à 1.6) sur une grille horizontale de type ORCA, avec une résolution horizontale de $2° \times 2°$ (réduite dans les régions tropicales), et 31 niveaux selon la verticale. La discrétisation des équations est conçue pour ne résoudre que les plus grandes échelles de la circulation océanique (voir fig. 1.2), tandis que l'effet des échelles non-résolues est paramétré dans les éqs. (1.1), (1.4) et (1.5) par les opérateurs de diffusion $\mathbf{D}^U$, D^T et D^S. La diffusion latérale est simulée par un opérateur laplacien isopycnal, avec une visosité et une diffusivité spécifiées. Et la diffusion verticale est déduite d'un schéma de fermeture turbulente, basé sur une équation pronostique pour l'énergie cinétique turbulente et une hypothèse de fermeture pour les longueurs de mélange (Blanke and Delecluse, 1993). La paramétrisation détaillée de la configuration, en particulier le forçage atmosphérique, dépend de spécifications propres à chaque étude particulière.

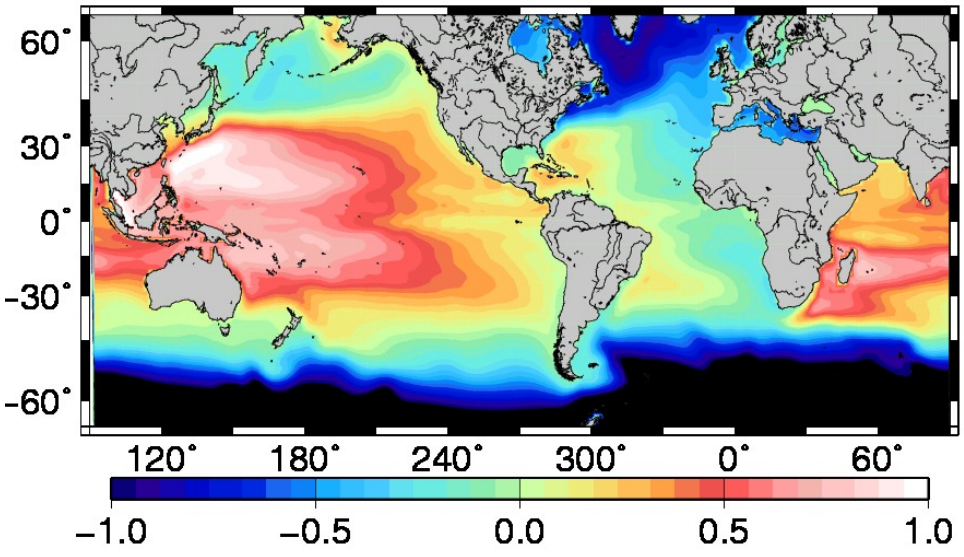


FIGURE 1.2 – Instantané de l'élévation de surface issu de la configuration ORCA2.

Dans nos travaux, cette configuration nous a servi principalement de banc d'essai pour tester des paramétrisations stochastiques de phénomènes non-résolus (voir chapitre 3), pour étudier l'effet d'incertitudes liées au forçage atmosphérique (voir chapitre 2), ou bien pour estimer des paramètres du forçage atmosphérique à partir d'observations océaniques (en utilisant les méthodes inverses décrite dans la partie II).

1.3.3 La configuration NATL025

La configuration NATL025 est une configuration régionale du modèle NEMO pour l'Atlantique Nord au 1/4° de résolution horizontale (Barnier et al., 2006), Dans cette configuration, le modèle de circulation a été couplé au modèle d'écoystème LOBSTER (par Ourmières et al., 2009) brièvement décrit plus haut. Aux moyennes latitudes, cette configuration est considérée "eddy-permitting", c'est-à-dire qu'elle autorise l'existence de tourbillons à méso-échelle (voir fig. 1.3). Une grande partie de l'écoulement à mésoéchelle reste donc non-résolue, et son effet sur les échelles résolues doit être paramétré. Pour ce faire, contrairement à ORCA2, un mélange latéral est simulé par un opérateur de diffusion biharmonique, tandis que le mélange vertical est déduit d'un schéma de fermeture turbulente (comme pour ORCA2). Ici encore, la paramétrisation détaillée de la configuration, tant pour le modèle de circulation que pour le modèle d'écosystème, dépend de spécifications propres à chaque étude particulière.

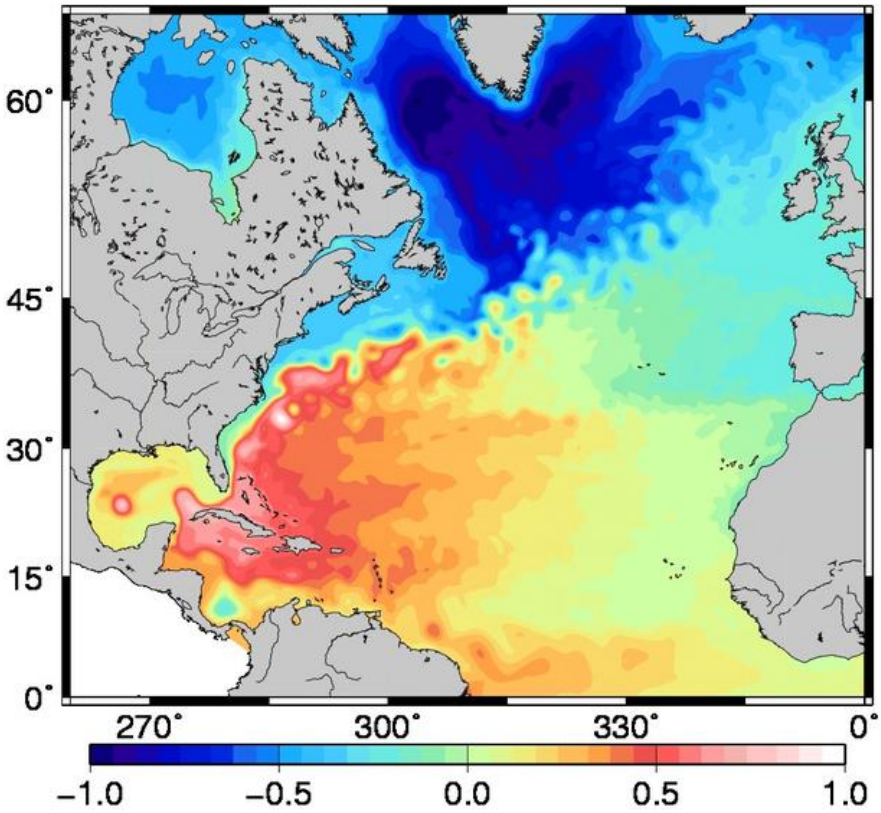


FIGURE 1.3 – Instantané de l'élévation de surface issu de la configuration NATL025.

Dans nos travaux, cette configuration nous a servi principalement de banc d'essai pour étudier l'effet d'incertitudes diverses sur le comportement du modèle d'écoystème (voir chapitre 2), ainsi que pour tester les méthodes et outils d'assimilation de données que nous développons sur un problème proche des applications réalistes (voir chapitres 6 et 7).

Chapitre 2

Simulations d'ensemble

> Without redundancy, there can be
> no check on the reliability.
>
> Richard Hamming (1997)

Quand le système étudié n'est pas déterministe, toute prévision est par nature incertaine, et doit s'exprimer sous la forme d'une distribution de probabilité. Cependant, pour un système complexe, il est en général impossible de calculer cette distribution de probabilité explicitement, et il faut se contenter de la décrire par un échantillon de taille limitée. En pratique, cela signifie qu'on doit générer un ensemble de simulations, par tirage aléatoire des différentes sources d'incertitude (lois dynamiques, paramètres, forçage, condition initiale) dans leurs distributions respectives (méthode de Monte Carlo, voir Robert and Casella, 2004). Les différents membres de l'ensemble correspondent alors à des prévisions équiprobables de l'évolution du système, échantillonnées au hasard. Le propos de ce chapitre est d'illustrer comment ce genre de simulation d'ensemble peut être utilisée pour caractériser notre incertitude sur la prévision.

La plupart des exemples donnés dans ce chapitre sont issus de simulations d'ensemble que nous avons réalisées pour résoudre des problèmes d'assimilation de données (voir chapitres 5 à 7).

2.1 Comportement central et dispersion

Il existe d'abord de nombreux cas de prévisions d'ensemble pour lesquelles on peut distinguer un "comportement central" du système, au voisinage duquel se concentre l'essentiel de la probabilité, dont la densité décroît au fur et à mesure qu'on s'éloigne. Dans ce genre de situation, la façon la plus efficace de décrire la distribution de probabilité est de caractériser ce comportement central (qui synthétise au mieux la prévision probabiliste) et la dispersion des probabilités autour de lui (l'incertitude associée). Il existe essentiellement trois façons de décrire le comportement central du système, toutes plus ou moins pertinentes selon le problème posé : la moyenne, le mode (maximum de probabilité) et la médiane. Quant à la dispersion, elle se caractérise en général par la variance, l'entropie (étalement des probabilités), ou bien certains quantiles de la distribution (intervalles de confiance).

La figure 2.1 illustre par exemple une prévision d'ensemble de la température de surface de l'océan, traduisant l'effet d'incertitudes sur le forçage atmosphérique. Cette simulation a été réalisée dans le cadre du travail de Broquet et al. (2008) à l'aide d'un modèle à couches (HYCOM) au 1/15° de résolution du Golfe de Gascogne. Le premier élément qui ressort de cette figure est que le comportement central de la prévision probabiliste, représenté ici par sa moyenne (trait en pointillé), ne correspond pas du tout à

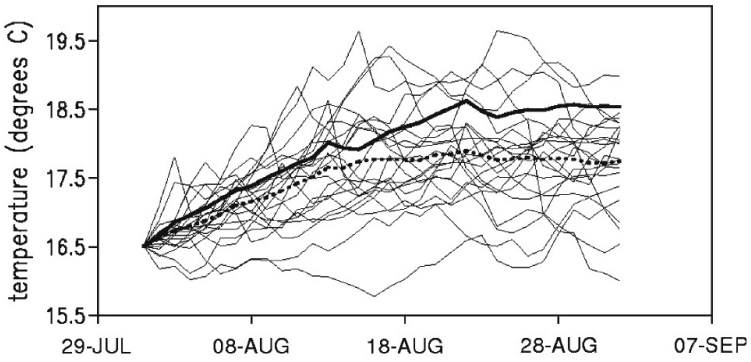


FIGURE 2.1 – Prévision d'ensemble de la température de surface traduisant l'effet d'incertitudes sur le forçage atmosphérique (d'après Broquet et al., 2008). Le trait en pointillé correspond à la moyenne d'ensemble, et le trait gras continu correspond à la prévision déterministe, réalisée avec l'espérance mathématique du forçage.

la prévision déterministe (trait gras continu) réalisée en utilisant le forçage optimal (ici l'espérance mathématique du forçage). Quand le modèle ne peut être supposé approximativement linéaire sur la plage des incertitudes, ce n'est donc pas le meilleur forçage qui produit la meilleure prévision. Cette observation (que nous retrouverons au chapitre suivant pour l'incertitude sur les lois dynamiques) est à mon sens une raison incontournable de recourir à une approche probabiliste plutôt que déterministe.

En ce qui concerne la dispersion, on voit que le comportement de l'ensemble est ici assez symétrique (autour de la moyenne) et d'allure gaussienne. Il peut donc être décrit assez simplement, par exemple par son écart-type. C'est ce qui a été fait par exemple dans le travail mené par Meinvielle et al. (2013), afin de caractériser l'impact sur la température de surface d'incertitudes sur le forçage atmosphérique (dans l'optique de réduire cette incertitude par assimilation de données) dans la configuration ORCA2 de NEMO (voir section 1.3). La figure 2.2 illustre le résultat obtenu par l'écart-type de température de surface au bout d'un mois de prévision d'ensemble (pour janvier 2004). Il y a quelque-chose de dangereusement simplificateur à réduire la richesse d'un ensemble de 200 simulations de l'océan global à l'examen de son seul écart-type. Mais on voit bien que cette représentation réduite est déjà complexe (très structurée régionalement, et variant fortement selon la saison), et qu'elle est donc souvent nécessaire. Il convient néanmoins d'être vigilant afin que les moyens utilisés pour synthétiser le résultat de la prévision probabiliste ne trahisse pas la richesse d'information contenue dans la simulation d'ensemble.

Afin d'illustrer ce point, examinons l'effet d'une incertitude sur le forçage par le vent (supposée gaussienne) dans un modèle couplé circulation/écosystème de l'Atlantique Nord (la configuration NATL025/LOBSTER de NEMO, voir section 1.3). Cette simulation d'ensemble a été réalisée dans le cadre du travail de Béal et al. (2010), avec pour objectif d'étudier de quelle manière cette incertitude peut être contrôlée par assimilation de données (voir pour cela la section 7.2). Dans ce modèle couplé, l'incertitude sur le vent se transmet principalement selon le schéma conceptuel de la figure 2.3. Et l'incertitude qui en résulte sur le comportement du système (en surface, à la station BATS) est illustrée par la figure 2.4. Les 3 graphes caractérisent les distributions conjointes de la norme du vent et respectivement 3 des variables du schéma 2.3 : la profondeur de couche de mélange, la température de surface, et la concentration en phytoplancton. Ici déjà, la notion de "comportement central" commence à prendre un sens quelque peu subjectif (ou du moins à définir selon l'application envisagée) puisque la moyenne de l'ensemble (carré vert) est très différent de la médiane (intersection des lignes interrompues) ou

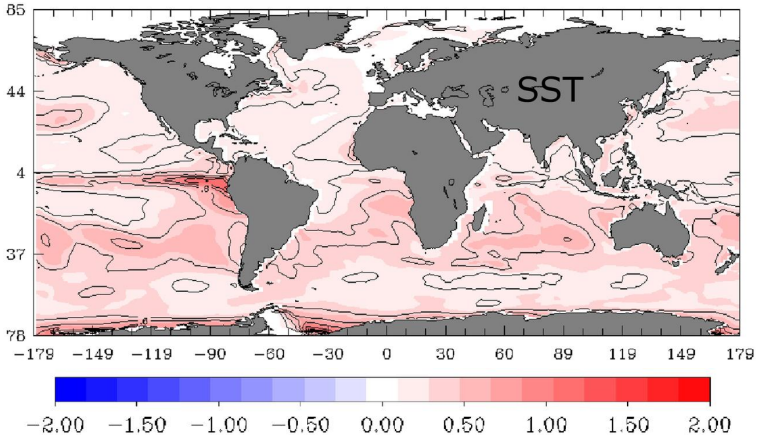


FIGURE 2.2 – Ecart-type de la température de surface résultant d'incertitudes sur le forçage atmosphérique, au bout d'un mois de prévision d'ensemble en janvier 2004 (calculé par Meinvielle, 2012).

du mode (plus grande concentration des points rouges). D'autre part, la dispersion se laisse difficilement réduire à une caractéristique aussi synthétique que l'écart-type de l'ensemble, dans la mesure où elle est très fortement asymétrique (voir les quartiles des distributions marginales, en pointillé) et étirée par la présence de quelques événements extrêmes (quelques rares cas de vent plus fort, produisant un réaction "hors norme" de la couche de mélange et de l'écosystème).

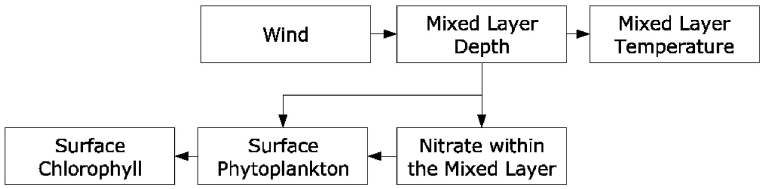


FIGURE 2.3 – Schéma conceptuel de transmission de l'incertitude sur le vent vers les autres variables du modèle.

2.2 Dépendance

Un deuxième genre de caractéristique du système qu'une simulation d'ensemble peut permettre d'acquérir est le degré de *dépendance statistique* entre diverses variables aléatoires. Quand on introduit cette notion, il importe toujours de préciser avant toute chose qu'une dépendance statistique signifie seulement que la connaissance de ce qui s'est passé pour l'une des variables influence la distribution de probabilité de l'autre, et que cela n'implique en aucune façon qu'il existe une relation de cause ou de conséquence entre elles. Deux variables aléatoires, sans lien causal entre elles, peuvent par exemple subir conjointement l'effet d'un même facteur extérieur. La connaissance des dépendances qui existent entre les différentes variables du système est d'une grande importance pour la solution des problèmes inverses, c'est-à-dire quand on cherche à gagner de l'information sur certaines variables à partir de l'observation d'autres variables. Dans cette optique, une mesure très générale de la dépendance entre deux variables aléatoire X_1 et X_2, est l'information mutuelle (e.g. Cover and Thomas, 2006) :

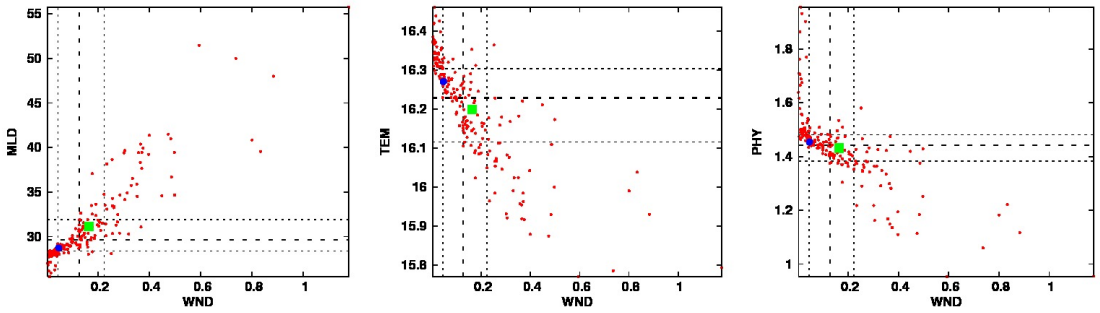


FIGURE 2.4 – Prévision d'ensemble du comportement d'un modèle couplé circulation/écosystème (profondeur de couche de mélange, température de surface, et concentration en phytoplancton) en réponse à une incertitude sur le forçage par le vent (d'après Béal et al., 2010). On observe notamment que la moyenne d'ensemble (carré vert) est très différente de la simulation non-perturbée (point bleu).

$$I(X_1, X_2) = \int\int p(X_1, X_2) \log \frac{p(X_1, X_2)}{p(X_1), p(X_2)} \, dX_1 dX_2 \tag{2.1}$$

qui représente la réduction de l'incertitude sur l'une des variables, obtenue par la connaissance de l'autre variable.

En pratique, cependant, le moyen le plus couramment utilisé pour mesurer la dépendance est le *coefficient de corrélation linéaire*. Ce moyen est entièrement suffisant pour caractériser la dépendance tant que la distribution peut être supposée gaussienne, et le coefficient de corrélation linéaire $\rho_{X_1X_2}$ détermine alors complètement l'information mutuelle (e.g. Cover and Thomas, 2006, chapitre 8) :

$$I(X_1, X_2) = -\frac{1}{2} \log \left(1 - \rho^2_{X_1X_2}\right) \tag{2.2}$$

Le modèle gaussien ne peut en effet exprimer qu'une relation de dépendance *linéaire* entre les variables (voir chapitre 6), dont la qualité est entièrement décrite par $|\rho_{X_1X_2}|$. Dans tous les autres cas, l'utilisation de $\rho_{X_1X_2}$ doit toujours se faire avec la plus grande prudence, car il peut souvent donner une très mauvaise idée de la dépendance entre variables aléatoires. C'est d'ailleurs pour cela que de nombreuses autres mesures de dépendance ont été proposées dans la littérature (tables de contingence, mesures non-paramétriques de corrélation, tau de Kendall,. . .), mais nous ne les introduirons ici qu'en cas de besoin.

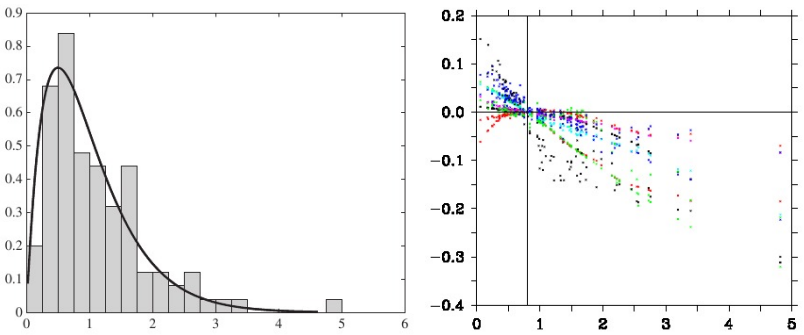


FIGURE 2.5 – Incertitude sur la salinité à 112 m de profondeur en différents points de la mer des Salomon (à droite), en réponse à un paramètre incertain (à gauche) du mélange par la marée interne (d'après Melet et al., 2012).

La figure 2.5 illustre par exemple une simulation d'ensemble de la circulation de la mer des Salomon (réalisée par Melet et al., 2012) traduisant l'effet d'incertitude sur la paramétrisation de la marée interne. A partir d'une hypothèse sur la distribution de probabilité du paramètre incertain (fig. 2.5, graphe de gauche), la simulation déduit l'incertitude qui en découle sur le comportement du système, par exemple sur la salinité à 112 m de profondeur en différents endroits (fig. 2.5, graphe de droite). L'objectif de cette étude était d'examiner dans quelle mesure le paramètre incertain pourrait être estimé à partir de données de température et salinité (mesurées par gliders, voir Melet et al., 2012), la réponse à cette question étant bien sûr complètement conditionnée par la dépendance qui existe entre le paramètre et les quantités observées. De ce que montre la figure, il ressort qu'il existe une dépendance très forte entre le paramètre incertain et la salinité, mais que cette dépendance est quelquefois assez loin d'être linéaire, et donc, dans ce cas, assez mal décrite par le coefficient de corrélation linéaire.

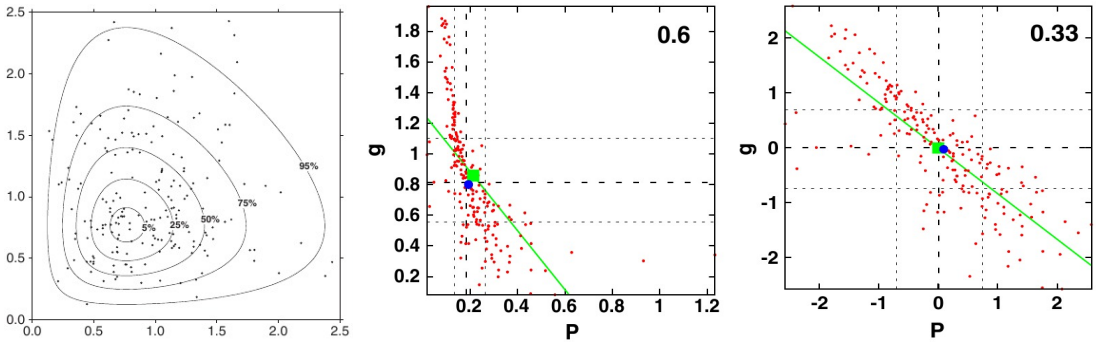


FIGURE 2.6 – Simulation d'ensemble (menée par Doron et al., 2011) de l'effet d'incertitudes sur quelques paramètres-clés du modèle d'écosystème (distribués selon une loi gamma, voir graphe de gauche). La figure illustre la dépendance entre un de ces paramètres (g) et la concentration de surface en phytoplancton (P), sans (au centre) ou avec (à droite) transformation non-linéaire de g et de P.

Dans une étude similaire menée par Doron et al. (2011), une simulation d'ensemble a été réalisée avec un modèle couplé circulation/écosystème (NATL025/LOBSTER, voir section 1.3), afin d'étudier l'effet d'incertitude sur quelques paramètres-clés du modèle d'écosystème (toujours dans le but de les estimer grâce à des observations de l'état du système). De nouveau, à partir d'une hypothèse sur la distribution de probabilité des paramètres incertains (fig. 2.6, graphe de gauche), la simulation déduit l'incertitude qui en découle sur le comportement du système, par exemple ici sur la concentration de surface en phytoplancton (indirectement reliée aux observations de couleur de l'eau) à la station NABE 19°W 36°N (fig. 2.6, graphe du centre). Ici encore, on observe que la dépendance entre le paramètre et la concentration en phytoplancton est assez loin d'être linéaire (voir le positionnement des membres de l'ensemble par rapport à la droite de régression, en vert) ; mais ici, le modèle de dépendance linéaire redevient plutôt bon (fig. 2.6, graphe de droite), à condition d'appliquer à chaque variable un changement de variable non-linéaire transformant certains quantiles de l'ensemble en les quantiles correspondants d'une distribution gaussienne normalisée (voir le détail de la méthode en section 7.2).

Cette transformation signifie que l'on adopte une description synthétique de l'ensemble composée des deux éléments suivants : le *comportement central* et la *dispersion* sont décrits par certains quantiles (par ex. les déciles) des distributions marginales de chaque variable, et la *dépendance* entre variables est décrite par la structure de corrélation linéaire entre variables transformées. Les intérêts de cette description synthétique des

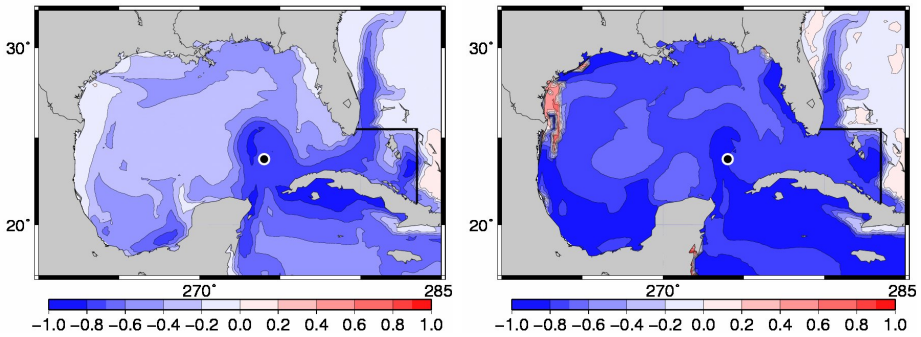


FIGURE 2.7 – Structure de corrélation linéaire entre le phytoplancton (au point de référence en noir) et le champ de nitrate, telle qu'obtenue à partir de l'ensemble de Doron et al. (2011) pour les variables non-transformées (à gauche) et les variables transformées (à droite).

incertitudes (discutés en détail dans Brankart et al., 2012, voir annexe B) sont principalement (i) qu'elle conduit à une mesure non-paramétrique de la corrélation (semblable à la corrélation de rang), (ii) qu'elle améliore ainsi la description des dépendances entre variables pour nombre d'applications océanographiques (par rapport à la corrélation linéaire entre variables non-transformées), et (iii) qu'elle est suffisamment concise et efficace pour permettre la résolution de problèmes inverses de grande taille (voir section 7.2). La figure 2.7 compare par exemple la structure de corrélation linéaire entre le phytoplancton (au point de référence en noir) et le champ de nitrate, telle qu'obtenue à partir de l'ensemble de Doron et al. (2011) pour les variables non-transformées (à gauche) et les variables transformées (à droite). La forte augmentation de la corrélation indique que la forte dépendance (non-linéaire) entre les variables ne peut être bien perçue par le coefficient de corrélation linéaire, et que la méthode de transformation ici proposée constitue un moyen assez radical de résoudre la difficulté.

Jusqu'ici, je me suis contenté d'illustrer les moyens que nous avons utilisés pour caractériser nos simulations d'ensemble, en essayant de mettre en évidence leurs limitations et leurs dangers. Un tout autre problème est de savoir dans quelle mesure ces caractéristiques correspondent à celles de la distribution de probabilité dont la simulation d'ensemble est un échantillon. Cette question du jugement sur échantillon (test de sgnificativité, intervalle de confiance,...) repose sur une littérature vaste qu'il n'y a pas la place ici d'évoquer. Mais au cours de nos travaux en assimilation de données, cette question s'est posée de façon particulièrement aiguë pour le coefficient de corrélation linéaire ; c'est pourquoi nous y reviendrons en section 6.3 en présentant l'algorithme que nous avons développé pour contourner la difficulté (en localisant les covariances).

2.3 Perspectives d'application

L'utilisation de simulations d'ensemble est encore relativement marginal en océanographie (excepté pour l'assimilation de données, voir chapitre 5 à 7). Et leur utilisation concrète ne va pas souvent au delà de ce qui a été décrit dans les deux sections précédentes. Mais une simulation d'ensemble contient souvent une information bien plus riche, qui peut être d'un intérêt pratique direct (voir l'encadré 4, page 27). L'objectif est ici simplement d'ouvrir quelques perspectives méthodologiques en partant des travaux que nous avons réalisés.

ENCADRÉ 4 : VALEUR ÉCONOMIQUE D'UNE PRÉVISION D'ENSEMBLE

En météorologie, le gain économique potentiel d'une prévision probabiliste par rapport à une prévision déterministe a depuis longtemps été mis en évidence (Murphy, 1977; Richardson, 2000; Palmer, 2002). La raison fondamentale de ce gain est que les pertes (P) occasionnées par un phénomène météorologique donné, et le coût (C) nécessaire pour s'en prémunir (en tout ou en partie) sont toujours des fonctions très non-linéaires (souvent à seuil) de l'intensité du phénomène. Cela implique en effet directement qu'un processus décisionnel basé sur une prévision probabiliste (si elle est réaliste) induira toujours un coût plus faible qu'une décision basée sur une prévision déterministe. La valeur économique de ce gain peut être importante pour de nombreuses applications commerciales des prévisions d'ensemble (Palmer, 2002), telles que la production d'électricité (par les éoliennes), le routage des bateaux, la gestion de la pollution atmosphérique, le rendement des récoltes agricoles,...

a. Probabilité d'événements. Tout processus décisionnel devrait pouvoir se fonder sur un examen des probabilités. Par exemple, la décision d'établir une route maritime à travers un détroit exige de connaître la probabilité qu'il soit libre de glace (ou bien que la fraction et l'épaisseur de glace soient inférieures à un certain seuil). A partir d'un ensemble, la probabilité d'un événement i s'estime simplement comme la fraction du nombre de membres où cet événement apparaît : $p_i = m_i/m$ (avec bien sûr une précision d'autant meilleure que m et m_i sont grands) [1]. La figure 2.8 montre un exemple de probabilité que l'océan soit libre de glace ($f = 0$) calculée à partir d'un ensemble (malheureusement encore non-stochastique) produit par Mercator-Ocean (voir Brankart et al., 2012, pour plus de détail). Dans ce travail, cet exemple était utilisé pour montrer la difficulté de se ramener à une distribution marginale gaussienne (par changement de variable non-linéaire, voir section 2.2 et 7.2) dans le cas d'une concentration de probabilité (ici en $f = 0$). Mais, plus basiquement, il me semble également bien illustrer l'intérêt d'une simulation d'ensemble pour diagnostiquer ce genre de probabilité.

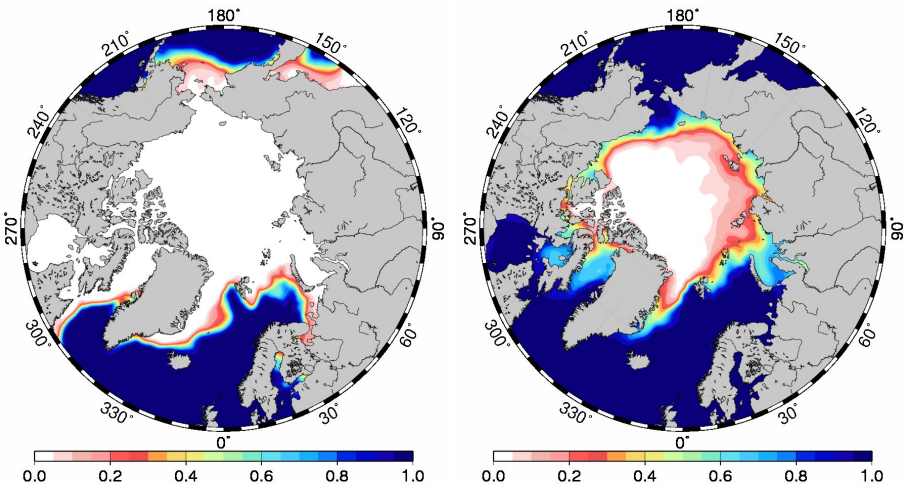


FIGURE 2.8 – Probabilité que l'océan soit libre de glace ($f = 0$) en mars (à gauche) et en septembre (à droite) calculée à partir d'un ensemble produit par Mercator-Ocean (d'après Brankart et al., 2012).

1. Une probabilité conditionnelle se calculerait de manière similaire : $p_{i|j} = m_{ij}/m_i$, à condition que m_i et m_{ij} ne soient pas trop faibles.

b. Evénements extrêmes. Un cas particulier important est celui d'événements de faible probabilité mais de grande conséquence. On sait par exemple que l'essentiel de l'érosion d'une vallée (et des dégâts en cas d'implantation humaine) se produit au moment des plus grandes crues. Et l'essentiel de l'apport hivernal de nutriments dans les couches de surface de l'océan se produit au moment des plus grandes tempêtes. La figure 2.9 (issue du travail de Lauvernet et al., 2009) montre par exemple la réponse d'un modèle de couche de mélange océanique à une perturbation stochastique (gaussienne) du forçage atmosphérique, et en particulier du vent. Dans cette expérience, ce sont bien les rares événements de vent extrême qui produisent les valeurs hors-normes de l'énergie cinétique turbulente (graphe de gauche), et donc les quelques plus grandes profondeurs de la couche de mélange (graphe de droite) qui permettent l'érosion de la nitrocline. Pour comprendre le fonctionnement réel du système, il peut donc être essentiel que la probabilité de tels événements soit explicitement décrite, même si cela exige une prévision d'ensemble de taille conséquente.

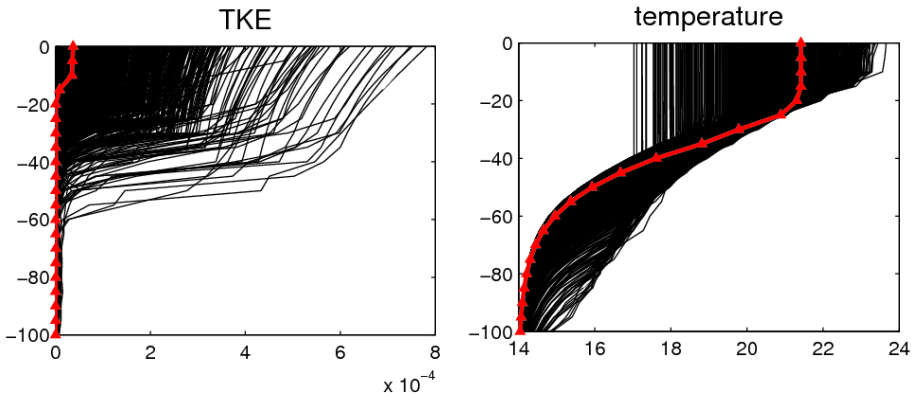


FIGURE 2.9 – Prévision d'ensemble (1000 membres) d'un profil d'énergie cinétique turbulente (en m^2/s^2) et de temperature potentielle (en $°C$), en réponse à une incertitude sur le forçage atmosphérique (d'après Lauvernet et al., 2009). La ligne rouge représente la prévision déterministe, réalisée avec l'espérance mathématique du forçage.

c. Catégories d'événements. Quelquefois, un système naturel complexe peut se comporter selon différentes modalités, chacune plus ou moins probable et avec des caractéristiques propres de dispersion et de dépendance entre les variables. Dans ce cas, la distribution de probabilité se caractérise en général par différents modes (maxima de probabilité) autour desquels se concentre une certaine fraction de la probabilité totale. Et une interprétation réaliste d'une simulation d'ensemble suppose de regrouper les membres de l'ensemble selon leur apparentement, afin d'identifier les différentes modalités de fonctionnement du système. Il existe pour cela toute une littérature (par ex. Izenman, 2008), notamment développés en génétique des populations afin de regrouper les individus ou les espèces selon leur proximité génétique. C'est ce genre de méthode qui sera envisagée en section 7.3 pour décrire une prévision d'ensemble sous la forme d'une superposition de distributions gaussiennes élémentaires, représentant chacune l'une des modalités du système. Quelquefois aussi, il peut être utile de regrouper les événements en fonction de critères spécifiques à une application particulière. La figure 2.10 illustre par exemple une simulation d'ensemble de structures frontales possibles pour les champs de traceur (par exemple, température de surface ou couleur de l'eau) traduisant une incertitude (gaussienne) sur le champ de vitesse à mésoéchelle (voir Titaud et al., 2011; Gaultier et al., 2013). Ici encore, on pourrait imaginer d'organiser les structures selon la présence ou l'absence de tel ou tel caractère de plus ou moins grande importance selon

l'application. Mais cet exemple suggère aussi qu'il n'existe pas toujours de façon simple et unique de regrouper les événements et de donner une vision synthétique de la variété des possibilités. C'est ce genre de difficulté qui nous amènera en section 7.4 à utiliser des méthodes plus générales pour tenter d'exploiter valablement l'information contenue dans nos simulations d'ensemble.

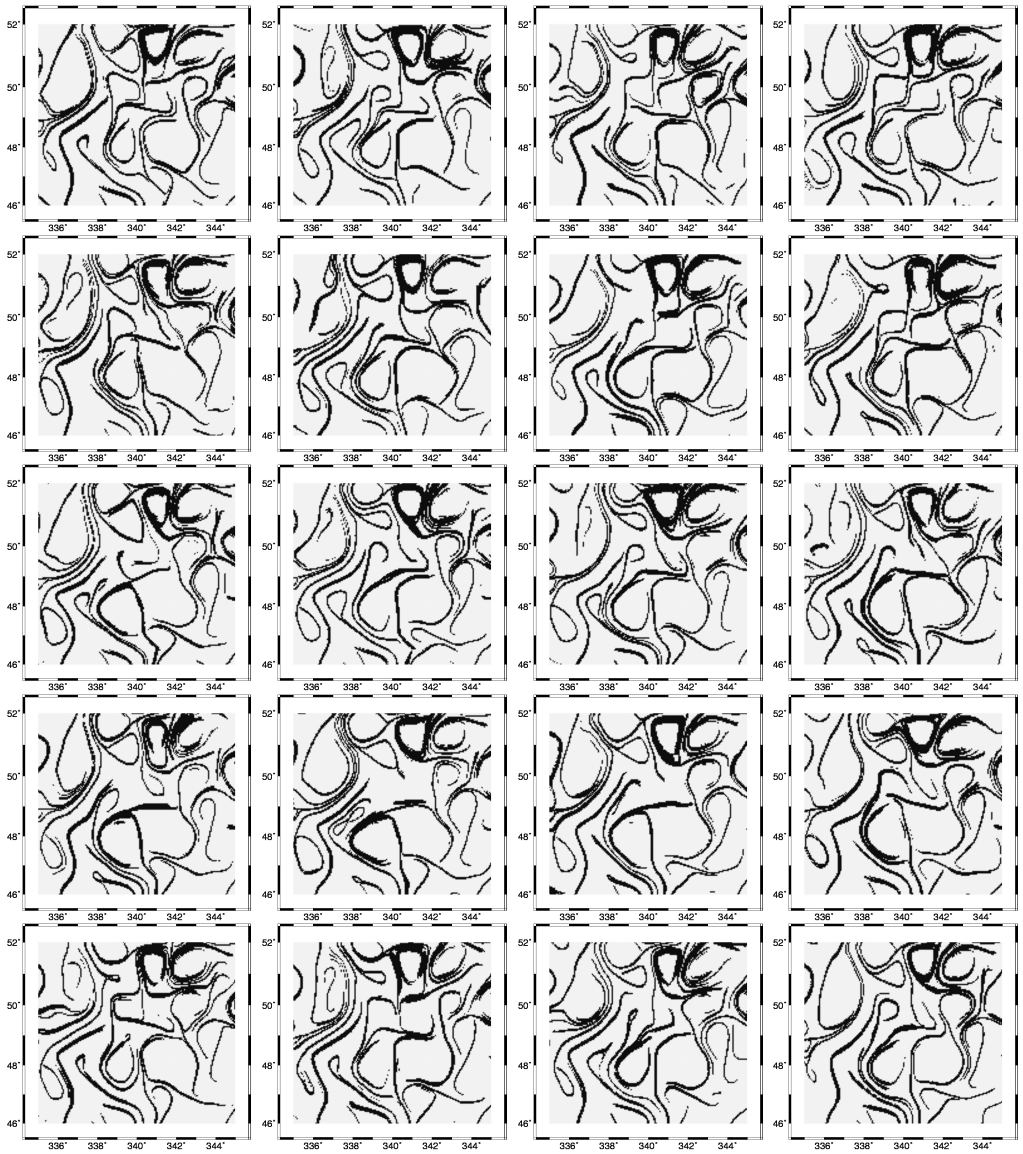


FIGURE 2.10 – Echantillon de structures frontales possibles auxquelles on peut s'attendre pour un champ de traceur (supposé passif), traduisant une incertitude (gaussienne) sur le champ de vitesse à mésoéchelle. L'échantillon des structures est déduit d'un échantillon gaussien de champs de vitesse par le calcul des exposants de Lyapounov associés à l'advection rétrograde de particules (d'Ovidio et al., 2008; Titaud et al., 2011; Gaultier et al., 2013).

Chapitre 3

Paramétrisations stochastiques

> I believe that the ultimate climate models (...) will be stochastic, i.e. random numbers will appear somewhere in the time derivatives.
>
> Edward Lorenz (1975)

Tout modèle d'océan comporte inévitablement de nombreuses sources d'incertitude. De plus, même si l'incertitudes sur les tendances est de moyenne nulle (modèle non-biaisé), son effet moyen sur une prévision à échéance finie ne peut être supposé nul en général dès que le modèle est non-linéaire. Dans ce cas, la prévision de l'état du système et de l'incertitude associée ne peuvent pas être découplées. Le propos de ce chapitre est d'examiner par quel moyen il est possible de modéliser les incertitudes sur les lois dynamiques qui sous-tendent le modèle, en particulier les incertitudes qui découlent de ce que le modèle ne résout pas et qui rendent le système essentiellement non-déterministe.

L'idée est d'abord de présenter une approche technique simple et générique (section 3.1) permettant de transformer un modèle déterministe (en particulier NEMO) en un modèle probabiliste, en simulant explicitement les différentes sources d'incertitude par une paramétrisation stochastique appropriée. Cette méthode est ensuite illustrée par différents exemples relatifs au modèle de circulation (section 3.2.1), au modèle d'écosystème (section 3.2.2), et au modèle de glace (section 3.2.3). Et ce n'est que plus loin dans ce mémoire que nous verrons les perspectives que cela procure en terme de comparaison du modèle aux observations (chapitre 4) et pour l'assimilation de données (chapitres 5 à 7).

Tous les exemples présentés dans ce chapitre sont basés sur des simulations que nous avons réalisées avec NEMO au cours des deux dernières années. Pour un point de vue plus général sur la question, se référer aux travaux de Palmer et al. (2005); Lermusiaux (2006); Frederiksen et al. (2012), qui permettent aussi de trouver de nombreuses références bibliographiques utiles.

3.1 Formulation stochastique de NEMO

Très récemment, j'ai commencé à m'intéresser au développement de paramétrisations stochastiques pour le modèle NEMO (Brankart, 2013), surtout dans la perspective de la paramétrisation des incertitudes sur le modèle pour l'assimilation de données (voir chapitres 5 à 7). Le propos de cette première section est de décrire la méthode qui a été implémentée pour transformer NEMO en un modèle probabiliste. Plus de détails sur les aspects techniques de ce développement sont donnés en annexe A.

3.1.1 Processus autorégressifs

Le point de départ de notre implémentation de paramétrisations stochastiques dans NEMO a été d'observer que la plupart des paramétrisations existantes sont basées sur des processus autorégressifs. Pour obtenir le système le plus flexible possible, une approche assez générique était donc d'ajouter un seul nouveau module à NEMO capable de produire un nombre arbitraire de processus autorégressifs. Ces processus peuvent ensuite être utilisés dans n'importe quel composant de NEMO (forçage atmosphérique, modèle de circulation, modèle d'écosystème, modèle de glace,...) pour simuler les incertitudes de façons très variées (voir sections 3.1.2, 3.1.3 et 3.1.4). Il suffit pour cela que l'utilisateur spécifie, à travers un fichier de paramètres, le nombre de processus stochastiques dont il a besoin pour simuler chaque source d'incertitude, ainsi que les caractéristiques statistiques de chacun d'eux.

En chaque point de grille du modèle, m processus autorégressifs gaussiens indépendants $\xi^{(i)}$, $i = 1, \ldots, m$ sont ainsi générés en utilisant la même équations de base :

$$\xi^{(i)}_{k+1} = a^{(i)} \xi^{(i)}_k + b^{(i)} w^{(i)} + c^{(i)} \tag{3.1}$$

où k est le numéro du pas de temps du modèle, et $a^{(i)}$, $b^{(i)}$, $c^{(i)}$ sont des paramètres définissant la moyenne ($\mu^{(i)}$), l'écart-type ($\sigma^{(i)}$) et l'échelle de corrélation temporelle ($\tau^{(i)}$) de chaque processus :

- pour les processus d'ordre 1, $w^{(i)}$ est un bruit blanc gaussien de moyenne nulle et d'écart-type égal à 1, et les paramètres $a^{(i)}$, $b^{(i)}$, $c^{(i)}$ sont donnés par :

$$\left\{ \begin{array}{l} a^{(i)} = \varphi \\ b^{(i)} = \sigma^{(i)} \sqrt{1-\varphi^2} \\ c^{(i)} = \mu^{(i)} \left(1-\varphi\right) \end{array} \right. \qquad \text{avec} \qquad \varphi = \exp\left(-1/\tau^{(i)}\right) \tag{3.2}$$

- pour les processus d'ordre n, $w^{(i)}$ est un processus autorégressif d'ordre $n-1$, de moyenne nulle, d'écart-type égal à $\sigma^{(i)}$, et d'échelle de corrélation temporelle égale à $\tau^{(i)}$; et les paramètres $a^{(i)}$, $b^{(i)}$, $c^{(i)}$ sont donnés par :

$$\left\{ \begin{array}{l} a^{(i)} = \varphi \\ b^{(i)} = \frac{n-1}{2(4n-3)} \sqrt{1-\varphi^2} \\ c^{(i)} = \mu^{(i)} \left(1-\varphi\right) \end{array} \right. \qquad \text{avec} \qquad \varphi = \exp\left(-1/\tau^{(i)}\right) \tag{3.3}$$

  De cette façon, des processus autorégressifs d'ordre arbitraire peuvent facilement être générés en utilisant le même morceau de code pour implémenter l'équation (3.1).

Par exemple, deux ou trois de ces processus peuvent être utilisés pour simuler les composantes de marches aléatoires (en 2D ou en 3D) en chaque point de grille du modèle. Trois exemples de marches aléatoires ainsi générées (pour $n = 1$, 2 or 3) sont illustrées en fig. 3.1 (voir application en section 3.1.3 ci-dessous).

Par ailleurs, une dépendance spatiale entre les processus stochastiques peut facilement être introduite en appliquant un filtre spatial aux $\xi^{(i)}$. Cela peut se faire soit par convolution des matrices 2D ou 3D contenant les $\xi^{(i)}$: $\tilde{\xi}^{(i)} = \mathcal{F}[\xi^{(i)}]$, soit en résolvant une équation elliptique : $\mathcal{L}[\tilde{\xi}^{(i)}] = \xi^{(i)}$. Dans les deux cas, l'opérateur de filtrage peut être rendu dépendant de l'écoulement simulé par le modèle, ou plus généralement de n'importe quel propriété représentée par le modèle (c'est-à-dire dans le système $\mathcal{A}$ de l'encadré 3, page 16). Techniquement, cela ne requiert que de rendre la description de l'état du modèle disponible pour les routines de filtrage. Un exemple d'application de cette option de filtrage (par un simple filtre laplacien) est donné en section 3.2.3 ci-dessous.

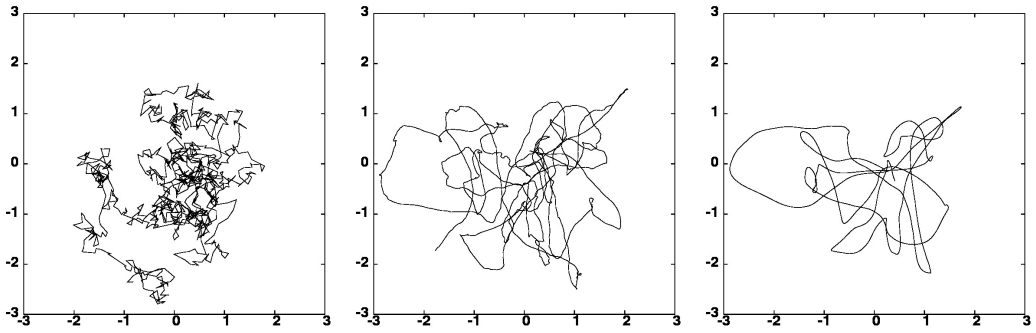


FIGURE 3.1 – Exemples de marches aléatoires obtenues à partir de processus autorégressifs d'ordre n (de moyenne nulle et d'écart-type égal à 1). Plus n est grand (1, 2 ou 3, de gauche à droite), plus lisse est la marche aléatoire (en raison du choix particulier des paramètres dans l'éq. 3.3).

En outre, la distribution marginale de chaque processus stochastique peut aussi être facilement modifiée en appliquant un changement de variable non-linéaire (transformation anamorphique) aux $\xi^{(i)}$ au moment de les utiliser dans le modèle : $\hat{\xi}^{(i)} = \mathcal{T}[\xi^{(i)}]$. L'idée est similaire à ce qui est fait dans les méthodes d'assimilation d'ensemble pour rendre gaussienne la distribution marginale de variables non-gaussiennes (voir section 7.2). Cette méthode peut par exemple être très utile si la description des incertitudes dans le modèle requiert des nombres aléatoires positifs. Dans ce cas, une transformation anamorphique peut être appliquée pour transformer les $\xi^{(i)}$ gaussiens en $\hat{\xi}^{(i)}$ non-gaussiens, par exemple avec une distribution log-normale ou gamma. Un exemple d'application de cette option d'anamorphose (vers une distribution gamma) est donné en section 3.2.3 ci-dessous.

En résumé, cette formulation procure un moyen assez simple et générique de produire une large classe de processus stochastiques. Elle est très simple à implémenter, et peut être directement utilisée pour introduire différentes formes de paramétrisation stochastique dans n'importe quel composant du modèle. Trois exemples de paramétrisation stochastique pouvant être directement implémentée au moyen de processus autorégressifs d'ordre n sont décrits dans les sections 3.1.2, 3.1.3 and 3.1.4 ci-dessous.

3.1.2 Incertitudes sur les tendances du modèle

Un premier moyen de simuler explicitement les incertitudes dans les prévisions météorologiques a été introduit il y a environ 15 ans dans le système de prévision d'ensemble d'ECMWF (Buizza et al., 1999). L'idée de base était de séparer la tendance du modèle ($\mathcal{M}$) en une tendance non-paramétrée ($\mathcal{NP}$) et une tendance paramétrée ($\mathcal{P}$) : $\mathcal{M} = \mathcal{NP} + \mathcal{P}$. La tendance non-paramétrée ($\mathcal{NP}$) contient tous les processus qui sont complètement résolus par le modèle, et peut donc être supposée exempte d'incertitude. La tendance paramétrée ($\mathcal{P}$) contient la paramétrisation de l'effet de processus non-résolus (au sens de l'encadré 3, page 16), qui est supposée être seule porteuse d'incertitude. La paramétrisation stochastique est alors introduite dans le modèle en multipliant la tendance paramétrée ($\mathcal{P}$) par un bruit aléatoire, sensé simulé explicitement l'incertitude associée à $\mathcal{P}$. La motivation originale de cette paramétrisation était de produire une prévision d'ensemble plus dispersée, et d'améliorer ainsi sa fiabilité (sa cohérence avec les observations, voir chapitre suivant). Cette technique est toujours utilisée aujourd'hui dans le système opérationnel d'ECMWF.

Ce genre de paramétrisation stochastique est tout aussi valide pour les modèles d'océan, et peut être directement implémentée avec la méthode générale décrite en sec-

tion 3.1.1. Par exemple, cela peut se faire en utilisant l'un ou plusieurs des $\xi^{(i)}$ donnés par l'éq. (3.1) comme bruit multiplicatif pour les différents termes de la tendance paramétrée :

$$\frac{d\mathbf{x}}{dt} = \mathcal{NP}(\mathbf{x}, \mathbf{u}, \mathbf{p}, t) + \sum_{i=1}^{m} \mathcal{P}^{(i)}(\mathbf{x}, \mathbf{u}, \mathbf{p}, t)\, \xi^{(i)}(t) \tag{3.4}$$

où t est le temps ; $\mathbf{x}$, le vecteur d'état du modèle ; $\mathbf{u}$, le forçage ; et $\mathbf{p}$, les paramètres. Dans ce cas, la moyenne des $\xi^{(i)}$ doit être choisie égale à 1, en supposant que les tendances paramétrées sont non-biaisées, et les autres paramètres statistiques (écart-type, corrélation temporelle et spatiale, distribution marginale) sont libres d'être ajustés en fonction de toute hypothèse raisonnable concernant les incertitudes. Dans les modèles d'océan, cette paramétrusation pourrait être appliquée à n'importe quelle paramétrisation des processus non-résolus, comme par exemple les opérateurs de diffusion, qui simulent l'effet des échelles non-résolues, les flux turbulents à l'interface air-mer, la paramétrisation des différentes fonctions présentes dans le modèle d'écosystème (souvent influencées par la diversité biologique non-résolue),... Une application de cette paramétrisation (appelée SPPT pour 'stochastic perturbed parameterized tendency') à un modèle d'écosystème sera décrite en section 3.2.2.

3.1.3 Incertitudes liées aux échelles non-résolues

Un autre moyen de simuler explicitement l'incertitude dans les modèles d'océan est de représenter directement l'effet des échelles non-résolues dans les équations du modèle en utilisant des processus stochastiques. Les échelles non-résolues peuvent en effet produire un effet à grande échelle en raison de la non-linéarité des équations. Dans les modèles d'océan, des termes non-linéaires importants sont par exemple le terme d'advection, l'équation d'état de l'eau de mer, les fonctions décrivant le comportment de l'écosystème,... En ce qui concerne le terme d'advection, l'effet des échelles non-résolues est en général paramétré par une diffusion additionnelle, tandis que pour les autres termes, il est en général purement et simplement ignoré.

Néanmoins, un moyen direct de simuler cet effet serait de générer un ensemble de fluctuations aléatoires $\delta\mathbf{x}^{(i)}$, possédant les mêmes caractéristiques que les échelles non-résolues, et de prendre la moyenne de l'opérateur modèle sur tout cet ensemble :

$$\frac{d\mathbf{x}}{dt} = \frac{1}{m} \sum_{i=1}^{m} \mathcal{M}(\mathbf{x} + \delta\mathbf{x}^{(i)}, \mathbf{u}, \mathbf{p}, t) \quad \text{avec} \quad \sum_{i=1}^{m} \delta\mathbf{x}^{(i)} = 0 \tag{3.5}$$

Ceci correspond à prendre une moyenne des équations du modèle, à travers laquelle les fluctuations $\delta\mathbf{x}^{(i)}$ (de moyenne nulle) peuvent produire un effet moyen (correspondant à l'interaction entre les systèmes $\mathcal{A}$ and $\mathcal{B}$ de l'encadré 3), dès que le modèle $\mathcal{M}$ est non-linéaire.

Manifestement, la difficulté principale de cette méthode est de générer des fluctuations $\delta\mathbf{x}^{(i)}$ possédant les bonnes caractéristiques statistiques. En première approche, cela peut être tenté en utilisant l'un ou plusieurs des $\xi^{(i)}$ donnés par l'éq. (3.1), soit en supposant que les statistiques des $\delta\mathbf{x}^{(i)}$ sont connues et peuvent être directement approximées par les $\xi^{(i)}$, ou bien en supposant que les $\delta\mathbf{x}^{(i)}$ peuvent être évalués par une fonction conjointe de l'état du modèle $\mathbf{x}$ et des processus aléatoires $\xi^{(i)}$. Par exemple, si les fluctuations $\delta\mathbf{x}^{(i)}$ peuvent être supposées proportionnelles au gradient à grande échelle $\nabla\mathbf{x}$ de l'état du système, alors il est possible de les calculer directement par le produit scalaire de $\nabla\mathbf{x}$ par des marches aléatoires $\boldsymbol{\xi}^{(i)}$ (telles qu'illustrées par la fig. 3.1) :

$$\delta\mathbf{x}^{(i)} = \boldsymbol{\xi}^{(i)} \cdot \nabla\mathbf{x} \tag{3.6}$$

Cet exemple particulier correspond à la paramétrisation stochastique proposée par Brankart (2013) pour simuler l'effet des échelles non-résolues dans le calcul du gradient de densité horizontal, à cause de la non-linéarité de l'équation d'état de l'eau de mer. Une application du schéma qui vient d'être décrit (que l'on appelera SPUF pour 'stochastic parameterization of unresolved fluctuations') à l'équation d'état sera décrite en section 3.2.1, et une application de la même méthode à un modèle d'écosystème sera présentée en section 3.2.2.

Encadré 5 : Coût énergétique du calcul déterministe

Tous les caclulateurs actuels sont construits avec l'objectif de rendre le taux d'erreur dans les opérations et les transmissions numériques aussi faible que possible. Cependant, certains auteurs (par ex. Sartori and Kumar, 2011; Lingamneni et al., 2013) soulignent que c'est précisément cette recherche d'un comportement parfaitement déterministe qui est responsable des besoins énergétiques exorbitants des supercalculateurs (en imposant une différence de potentiel éléctrique minimale entre les bits 0 et 1), et qu'une grande partie de la puissance consommée pourrait donc être économisée en admettant un petit taux d'erreur dans le calcul. Bien sûr, cela impose de moduler le taux d'apparition des erreurs selon l'importance des bits calculés, et de modifier les algorithmes de calcul pour augmenter leur tolérance à la présence de ces erreurs (Sloan et al., 2012), mais de tels processeurs inexacts ont été élaborés et construits et sont actuellement en cours d'expérimentation (http ://passat.crhc.illinois.edu/).

FFT sur un processeur stochastique (Lingamneni et al., 2012)

Avec ce genre d'approche, l'optique de l'ingénieur numéricien devra aussi changer (Shanbhag et al., 2008) : au lieu de concevoir un algorithme déterministe qui produise un résultat aussi exact que possible, il devra le concevoir stochastique en minimisant le coût énergétique nécessaire pour obtenir la précision qu'il recherche. Ce nouveau point de vue (ou ce retour aux sources, voir von Neuman, 1956) serait aussi plutôt cohérent avec l'idée d'un modèle océanique ou climatique formé de composants incertains : au lieu de concevoir un modèle déterministe qui oblige à résoudre explicitement de plus en plus de détails du système (au risque que son coût énergétique finisse par influencer le système climatique lui-même), il faudra le concevoir explicitement imparfait en maximisant la quantité d'information qu'il contient sur le système pour un coût énergétique donné.

3.1.4 Incertitudes liées à la diversité non-résolue

Une autre source générale d'incertitude dans les modèles d'océan est la présence dans le système de comportements dynamiques variés, qui ne peuvent tous être distingués par le modèle. Par exemple, les écoystèmes marins contiennent toujours une large diversité

d'espèces qui ne peuvent être toutes décrites séparément par le modèle, et qui doivent être agrégées en un nombre réduit de variables d'état. D'une façon semblable, la glace de mer peut adopter une large variété de comportements dynamiques différents, qui ne peuvent tous être explicitement traités par le modèle. Comme les échelles non-résolues, cette diversité non-résolue génère des incertitudes sur l'évolution du système, qui peuvent être explicitement simulées en utilisant une approche analogue :

$$\frac{d\mathbf{x}}{dt} = \frac{1}{m}\sum_{i=1}^{m} \mathcal{M}(\mathbf{x}, \mathbf{u}, \mathbf{p} + \delta\mathbf{p}^{(i)}, t) \tag{3.7}$$

où $\delta\mathbf{p}^{(i)}$ sont des fluctuations aléatoires des paramètres, représentant la variété des comportements dynamiques qui peuvent être simultanément présents dans le système.

L'application de cette méthode requiert une description statistique de l'incertitude sur les paramètres ; et de nouveau, en première approche, celle-ci peut être paramétrée en utilisant l'un ou plusieurs des $\xi^{(i)}$ donnés par l'éq. (3.1). En particulier, cette méthode inclut la paramétrisation stochastique proposée par Juricke et al. (2013) pour simuler explicitement les incertitudes sur la résistance de la glace dans un modèle océanique à éléments finis. Il était donc assez facile de reproduire cette paramétrisation (que l'on appelera SPUD pour 'stochastic parameterization of unresolved diversity') dans la composante 'glace de mer' du modèle NEMO. Les résultats obtenus seront résumés en section 3.2.3.

3.2 Impact sur les simulations

Le propos de cette section est maintenant d'illustrer l'impact des paramétrisations stochastiques présentées en section 3.1 dans les différents composants de NEMO : le modèle de circulation en section 3.2.1, le modèle d'écosystème en section 3.2.2, et le modèle de glace de mer en section 3.2.3. Dans cette section, l'accent sera mis sur le comportement probabiliste du système ($\mathcal{A}$) en réponse aux incertitudes (interactions avec $\mathcal{B}$). Toutes les simulations ont été réalisées à l'aide du même code générique de la formulation stochastique de NEMO décrite en section 3.1.

3.2.1 Equation d'état stochastique

En raison de la non-linéarité de l'équation d'état de l'eau de mer, les fluctuations non-résolues de température potentielle (T) et de salinité (S) (dans le système $\mathcal{B}$) peuvent produire un impact direct sur le gradient de densité à grande échelle (dans le système $\mathcal{A}$). Comme suggéré par Brankart (2013), cet effet peut être simulé par le schéma SPUF décrit en section 3.1.3, en appliquant l'éq. (3.5) à l'équation d'état :

$$\rho^{\text{stoch}}(T,S) = \frac{1}{m}\sum_{i=1}^{m} \rho\left(T + \delta T^{(i)}, S + \delta S^{(i)}\right) \quad \text{avec} \quad \sum_{i=1}^{m} \delta T^{(i)} = 0\,,\ \sum_{i=1}^{m} \delta S^{(i)} = 0 \tag{3.8}$$

où $\delta T^{(i)}$ et $\delta S^{(i)}$ simulent explicitement les fluctuations non-résolues de température potentielle et de salinité. Ces fluctuations sont générées par des marches aléatoires (voir fig. 3.1) selon l'éq. (3.6), avec les paramètres pour les $\xi^{(i)}$ donnés par le tableau 3.1.

Il est intéressant de noter (en complément à ce qui est expliqué par Brankart, 2013) qu'il existe une grande proximité entre cette correction stochastique de la densité à grande échelle et la méthode semi-pronostique proposée par Greatbatch et al. (2004); Greatbatch and Zhai (2006). Dans les deux cas, en effet, la seule correction au modèle est appliquée dans l'équation du vent thermique à travers une correction directe de la

	Equation d'état		Ecosystème		Glace de mer
	ORCA2	NATL025	SPPT	SPUF	SPUD
Nombre de processus	6×3	1×3	6	1×3	1
Ordre des processus	1	1	1	1	2
Moyenne	0.	0.	1.	0.	0.
Ecart-type	$\sigma_{xy} = 4.2$ $\sigma_z = 1$	$\sigma_{xy} = 1.4$ $\sigma_z = 0.7$	0.5	$\sigma_{xy} = 3$ $\sigma_z = 1$	1.
Corrélation temporelle	12 jours	10 jours	3 jours	12 jours	30 jours
Filtrage spatial	non	non	non	non	laplacien
Anamorphosis	non	non	non	non	gamma

TABLE 3.1 – Paramètres des processus autorégressifs utilisés pour toutes les applications décrites dans ce chapitre. En ce qui concerne la paramétrisation stochastique de l'équation d'état, l'écart-type est multiplié par sin ϕ pour ORCA2, et par sin 2ϕ pour NATL025, où ϕ est la latitude.

densité, tandis que les équations de conservation qui règlent l'évolution de la température potentielle, de la salinité et de la vitesse horizontale ne sont absolument pas modifiées. On peut donc être certain que la paramétrisation stochastique possède les mêmes propriétés de conservation que la méthode semi-pronostique ; en particulier, il n'y a aucune modification directe des propriétés T, S des masses d'eau, aucune augmentation du mélange diapycnal, et donc aucun compromis sur le fait que l'océan intérieur doive s'écouler essentiellement dans la direction du plan neutre tangent. Toute modification de la structure thermohaline de l'océan est donc uniquement produite de façon indirecte à travers la réorientation des principaux courants.

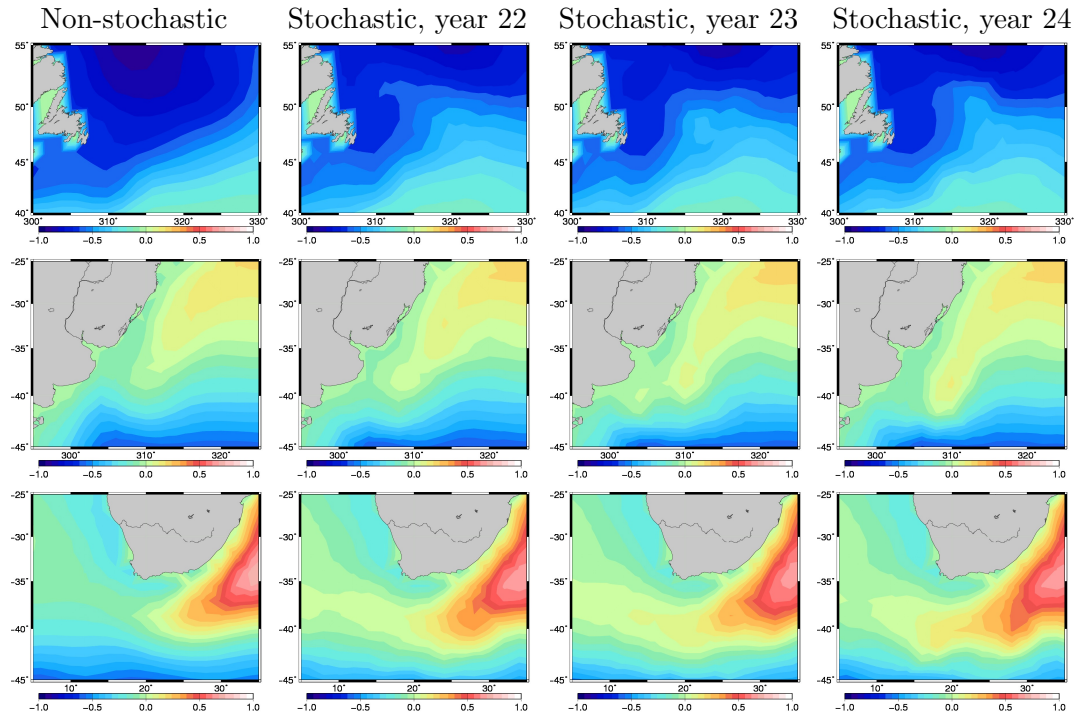


FIGURE 3.2 – Variabilité interannuelle intrinsèque de la topographie dynamique (en mètres) produite par la formulation stochastique de l'équation d'état dans une configuration globale à basse résolution (ORCA2) : Northwest corner (en haut), zone de confluence (au milieu) et courant d'Agulhas (en bas).

L'impact des fluctuations stochastiques de température et salinité se porte en effet

d'abord sur la circulation moyenne simulée par le modèle. Cet effet moyen dans une configuration à basse résolution de NEMO (ORCA2) a été discuté en détail par Brankart (2013). En résumé, la correction au champ de densité est surtout importante (et assez systématiquement négative en raison de la convexité de l'équation d'état) le long des principaux fronts qui séparent les gyres subtropicaux et subpolaires. La trajectoire des courants moyens est donc modifiée, en réduisant de façon très importante les biais du modèle déterministe. En particulier, la trajectoire du Gulf Stream est améliorée (se détachant de la côte américaine à une latitude plus correcte) et la structure du 'Northwest corner' devient plus réaliste. L'impact sur la circulation moyenne est d'une nature similaire à ce qui peut être obtenu par la méthode semi-pronostique (Greatbatch et al., 2004), mais au lieu de diagnostiquer la correction de densité à partir d'observations, le modèle stochastique possède l'avantage de se comporter comme un système dynamique autonome.

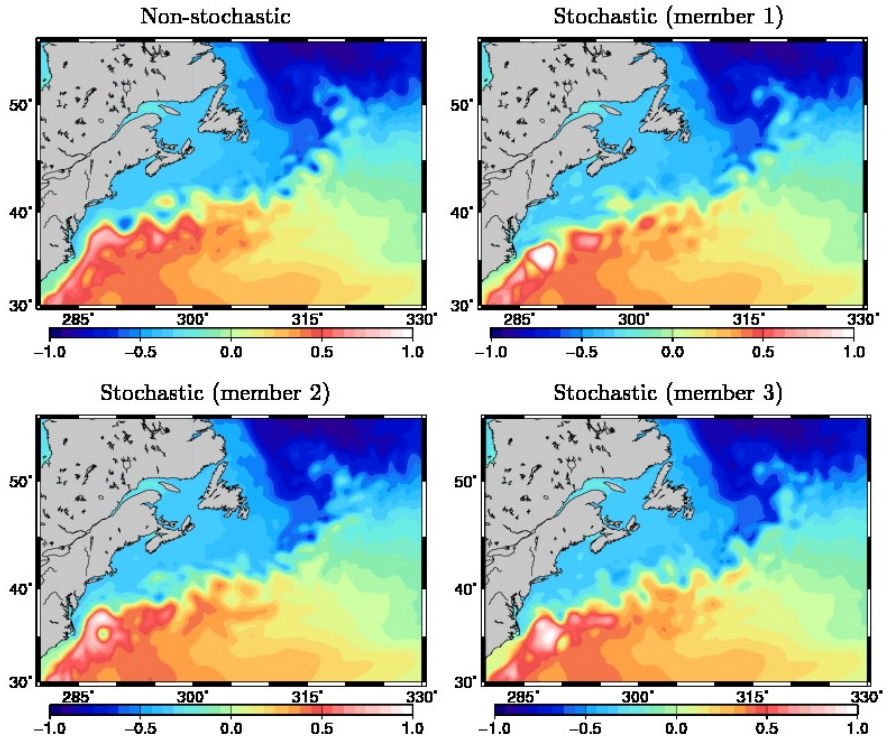


FIGURE 3.3 – Dispersion d'ensemble de la topographie dynamique (en mètres) produite par la formulation stochastique de l'équation d'état dans une configuration au 1/4° de résolution de l'Atlantique Nord (NATL025). Ce résultat a été obtenu dans le cadre du projet SANGOMA (travail de recherche de Guillem Candille).

Le deuxième effet des fluctuations stochastiques de température et salinité est de générer de la variabilité aléatoire dans le système. A cause de la non-linéarité de l'équation d'état, les petites échelles modifient sans cesse la structure à grande échelle de la densité, et donc la trajectoire des courants à grande échelle. Ceci implique qu'il existe un flux d'information continuel depuis le système $\mathcal{B}$ (petites échelles) vers le système $\mathcal{A}$ (grandes échelles), qui est simulé dans le modèle stochastique par les processus aléatoires $\xi^{(i)}$, et qui est totalement ignoré dans le modèle déterministe. Cet effet est illustré par la fig. 3.2, qui représente la structure de la topographie dynamique dans quelques régions clés de l'Atlantique : le 'Norhwest corner' (graphes du haut), la zone de confluence (graphes du milieu), et la rétroflexion du courant d'Agulhas (graphes du bas). Dans la simulation non-stochastique, en absence de variabilité interannuelle du forçage atmosphérique (comme

dans Brankart, 2013), la variabilité interannuelle est extrêmement faible : c'est pourquoi seule une année typique est illustrée sur la figure, car toute autre année apparaîtrait identique. Dans la simulation stochastique au contraire, non seulement la structure moyenne est modifiée (comme indiqué par Brankart, 2013), mais la variabilité interannuelle est aussi très fortement amplifiée. Elle devient ainsi plus compatible avec la variabilité intrinsèque à grande échelle que l'on obtient par un modèle à plus haute résolution, ou bien par les mesures issues de l'atlimétrie spatiale. Cette variabilité intrinsèque (produite en absence de toute variabilité interannuelle du forçage atmosphérique) donne une bonne idée de ce que serait la dispersion d'une prévision d'ensemble réellement probabiliste. Dans un modèle à haute résolution, cette dispersion du comportement à grande échelle ne peut résulter que de l'interaction avec la mésoéchelle (comme expliqué par Penduff et al., 2011). Dans la configuration à basse résolution (ORCA2), cette variabilité intrinsèque et donc essentiellement imprévisible (non-déterministe) de la grande échelle est ici (au moins partiellement) restaurée par une paramétrisation stochastique de l'effet de la mésoéchelle (qui est dans le système $\mathcal{B}$) sur la densité à grande échelle.

Pour continuer à explorer l'effet de ces incertitudes, la même paramétrisation stochastique commence à être appliquée à la configuration au 1/4° de l'Atlantique Nord (NATL025) dans le cadre du projet SANGOMA (travail de recherche de Guillem Candille). Comme premier exemple des résultats obtenus, la figure 3.3 compare des instantanés de topographie dynamique après deux ans de simulation, correspondants au modèle déterministe (en haut à gauche) et à trois membres d'un ensemble de simulations stochastiques. Les résultats (obtenus avec les paramètres donnés par le tableau 3.1) indiquent que la paramétrisation stochastique tend ici encore à produire un effet moyen sur la trajectoire du Gulf Stream, et à décorreler les structures à mésoéchelle produites par les différents membres de l'ensemble. Les premières questions que nous voudrions aborder avec ce genre de simulation seraient de savoir dans quelle mesure la dispersion de l'ensemble produite par ces incertitudes est capable d'expliquer une part substantielle de l'écart aux observations altimétriques, et donc dans quelle mesure ce type d'ensemble peut être utilisé pour assimiler des observations altimétriques dans NATL025. Ensuite, dans une perspective à plus long terme, il se pourrait peut-être que les processus stochastiques $\xi^{(i)}$ puisse être avantageusement utilisés comme vecteur de contrôle pour l'assimilation de données, qui posséderait dès lors automatiquement les même propriétés de conservation que la méthode semi-pronostique de Greatbatch et al. (2004).

3.2.2 Modèle stochastique d'écosystème

Les modèles d'écosystème marin comportent de nombreuses sources d'incertitude. Pour simplifier la discussion, seulement deux classes d'incertitude seront abordées ici : les incertitudes qui résultent de la diversité biologique non-résolue, et les incertitudes qui résultent des échelles non-résolues dans les champs de traceurs biogéochimmiques. D'une part, pour simuler les incertitudes qui résultent de la diversité non-résolue, nous utiliserons le schéma SPPT décrit en section 3.1.2 en multipliant les termes SMS ("source minus sink") du modèle d'écosystème par un bruit multiplicatif :

$$SMS_k^{\text{stoch}}(C_l) = SMS_k^{\text{ref}}(C_l) \times \xi^{(k)} \tag{3.9}$$

où C_l sont les concentrations de traceurs biogéochimmiques, et $\xi^{(k)}$ sont les processus autorégressifs obtenus par l'éq. (3.1), avec les paramètres donnés au tableau 3.1. Pour simuler la diversité non-résolue, le schéma SPUD décrit en section 3.1.4 aurait sans doute été plus naturel, mais au vu du nombre élevé de paramètres, le schéma SPPT est nettement plus simple à implémenter en première approche. D'autre part, pour simuler les incertitudes qui résultent des échelles non-résolues, nous utiliserons le schéma SPUF

décrit en section 3.1.3, en appliquant l'éq. (3.5) aux termes SMS :

$$SMS_k^{\text{stoch}}(C_l) = \frac{1}{m}\sum_{i=1}^{m} SMS_k^{\text{ref}}(C_l + \delta C_l^{(i)}) \quad \text{avec} \quad \sum_{i=1}^{m} \delta C_l^{(i)} = 0 \tag{3.10}$$

où les $\delta C_l^{(i)}$ simulent explicitement les fluctuations non-résolues des concentrations des différents traceurs. Ces fluctuations sont générées par des marches aléatoires (comme en fig. 3.1) en suivant l'éq. (3.6), avec les paramètres donnés au tableau 3.1.

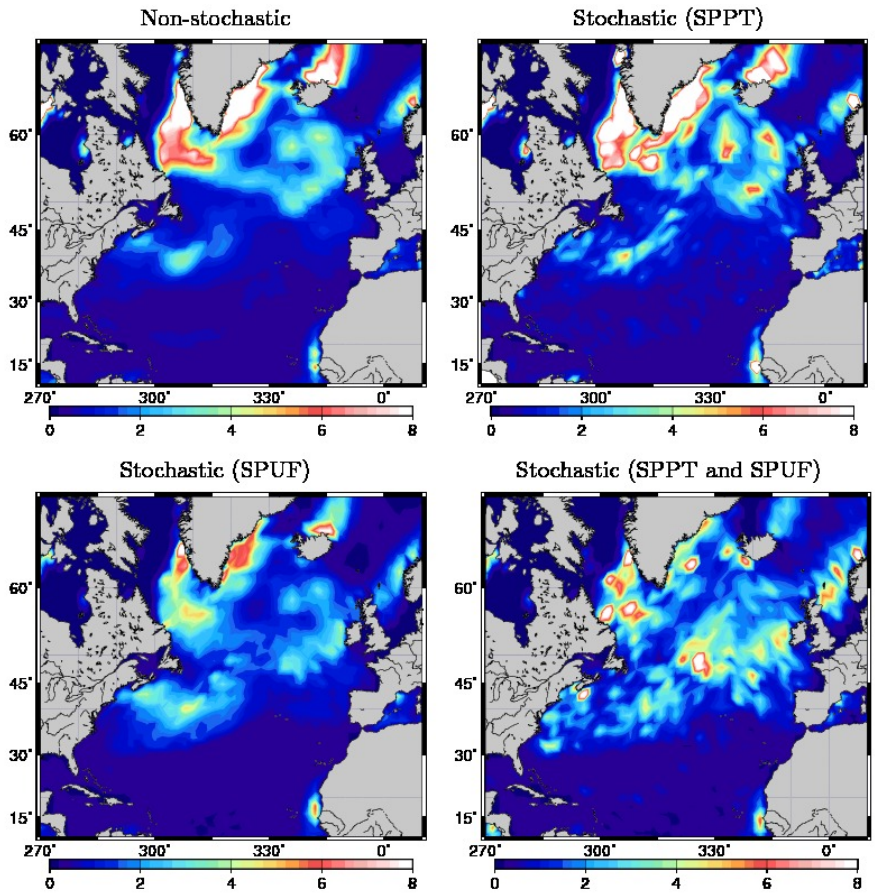


FIGURE 3.4 – Concentration de surface de phytoplankton (en mmol-N/m^3) pour le 15 juin obtenue avec différentes parametrisations stochastiques : simulation de référence, non-stochastique (en haut, à gauche), schéma SPPT seul (en haut, à droite), schéma SPUF seul (en bas, à gauche), schémas SPPT et SPUF ensemble (en bas, à droite)

En guise d'exemple, l'impact de ces deux paramétrisations stochastiques a été étudié dans la configuration globale à basse résolution ORCA2, couplée au modèle d'écosystème LOBSTER (voir section 1.1). Le comportement de ce modèle n'est ici illustré en fig. 3.4 que par la concentration de surface en phytoplancton au 15 juin de la seconde année de simulation. Par comparaison au modèle déterministe (en haut, à gauche), la simulation stochastique utilisant le schéma SPPT (en haut, à droite) ne modifie pas très fortement le comportement général du système (malgré un écart-type de 50% pour le bruit multiplicatif), mais augmente très significativement la microstructuration ou fragmentation (patchiness) de la concentration en phytoplancton. Ceci permet de conjecturer que les incertitudes (en particulier celles qui sont liées à la diversité non-résolue) peuvent partiellement expliquer la microstructuration observée sur les images satellitaire de couleur de l'océan. Inversément, le schéma SPUF (en bas, à gauche) n'augmente pas la micro-

structuration, mais peut modifier substantiellement le comportement local du système, parfois en augmentant, parfois en diminuant la production primaire (selon que la dérivée seconde du terme SMS est positive ou négative). A première vue, ces deux sources d'incertitudes sont donc insuffisantes pour expliquer l'écart considérable qui persiste entre simulation et observations de couleur de l'océan.

Une expérience supplémentaire (fig. 3.4, en bas, à droite) a ensuite été réalisée en utilisant conjointement les deux paramétrisations stochastiques (SPPT et SPUF), ce qui, techniquement, peut se faire très simplement en générant un nombre suffisant de processus autorégressifs pour nourrir simultanément les deux schémas. Le résultat indique une forte interaction entre les deux paramétrisations, qui conduit à la fois à une profonde modification du comportement du système, et à une microstructuration encore accentuée par rapport au schéma SPPT seul. De mon point de vue, ceci conduit assez directement à l'idée que les incertitudes peuvent être un ingrédient d'une importance décisive pour expliquer le comportement dynamique des écosystèmes marins, et pour assurer la cohérence (en valeur et en structure) entre le modèle et les observations de couleur de l'eau. Cette question est actuellement explorée dans le cadre de la thèse de Florent Garnier.

3.2.3 Glace de mer à resistance stochastique

L'une des difficultés principales des modèles de glace de mer est de simuler la large diversité des comportements dynamiques de la glace. Parmi les caractéristiques de la glace, le paramètre le plus sensible est certainement sa résistance $P^\star$. Dans les modèles simples (comme LIM2 dans NEMO), $P^\star$ est supposé constant, tandis que dans les modèles plus sophistiqués (comme LIM3 dans NEMO), les variations de $P^\star$ sont explicitement simulées en fonction des différents types de glace simultanément présents en chaque point de grille du modèle. L'impact d'incertitudes sur $P^\star$ a été étudié par Juricke et al. (2013) en utilisant un modèle à éléments finis (FESOM), couplé à un modèle simple de glace de mer (semblable à LIM2). Le propos de cette section est d'essayer de reproduire leur paramétrisation dans NEMO/LIM2 en utilisant l'approche technique générique décrite en section 3.1. Or, il se fait que cela peut se faire très simplement, pratiquement sans aucun nouvel effort de codage, en utilisant le schéma SPUD (eq. 3.7) avec $m = 1$ et

$$P^\star + \delta P^\star = P^\star \xi \tag{3.11}$$

où ξ est l'un des processus autorégressifs donnés par l'éq. (3.1), avec les paramètres donnés au tableau 3.1. Les paramètres ont été choisis pour se rapprocher autant que possible de la paramétrisation proposée par Juricke et al. (2013). Quelques spécificités sont : l'utilisation de processus autorégressifs d'ordre 2 au lieu de processus d'ordre 1, et l'utilisation d'une distribution marginale gamma au lieu d'un autre type de distribution positive dans le travail de Juricke et al. (2013).

Cette paramétrisation stochastique a été de nouveau appliquée à la configuration ORCA2, toujours sans variabilité interannuelle du forçage atmosphérique (selon une formulation identique à celle utilisée par Brankart, 2013). Le comportement du modèle n'est ici illustré en fig. 3.5 que par l'épaisseur de glace dans l'Arctique à la fin du mois de mars (quand l'extension de la glace est proche de son macimum). Par comparaison au modèle déterministe (en haut, à gauche), le premier impact de la paramétrisation stochastique est de systématiquement accroître l'épaisseur de glace, surtout dans les régions de glace pluriannuelle (nord du Groenland et ouest du Canada), et de légèrement diminuer son extension. Ces deux effets sont très semblables à ce qui est décrit et expliqué par Juricke et al. (2013), qui montrent aussi que cela ne peut être reproduit par une simple modification uniforme de $P^\star$.

Par ailleurs, les fluctuations stochastiques de $P^\star$ induisent également une variabilité aléatoire dans le système. Comme pour la topographie dynamique en section 3.2.1, la variabilité interannuelle de l'épaisseur de glace est extrêmement faible dans ORCA2 (du moins en l'absence de variabilité interannuelle dans le forçage atmosphérique) : c'est pourquoi une seule année typique est représentée en fig. 3.5. Dans la simulation stochastique, au contraire, non seulement l'épaisseur moyenne est modifiée (comme pour la topographie dynamique en fig. 3.2), la variabilité interannuelle est aussi très fortement amplifiée. Ce que ces premiers résultats laissent espérer, c'est d'abord que la simulation explicite des incertitudes puisse fournir une base adéquate à la comparaison probabiliste du modèle aux observations de la glace, et ensuite qu'elle puisse nous aider à produire des prévisions d'ensemble fiables pour résoudre nos problèmes d'assimilation de données. Par ailleurs, il se pourrait aussi que cette approche stochastique nous procure une alternative utile à la résolution explicite de la diversité de comportment de la glace de mer (comme dans LIM3).

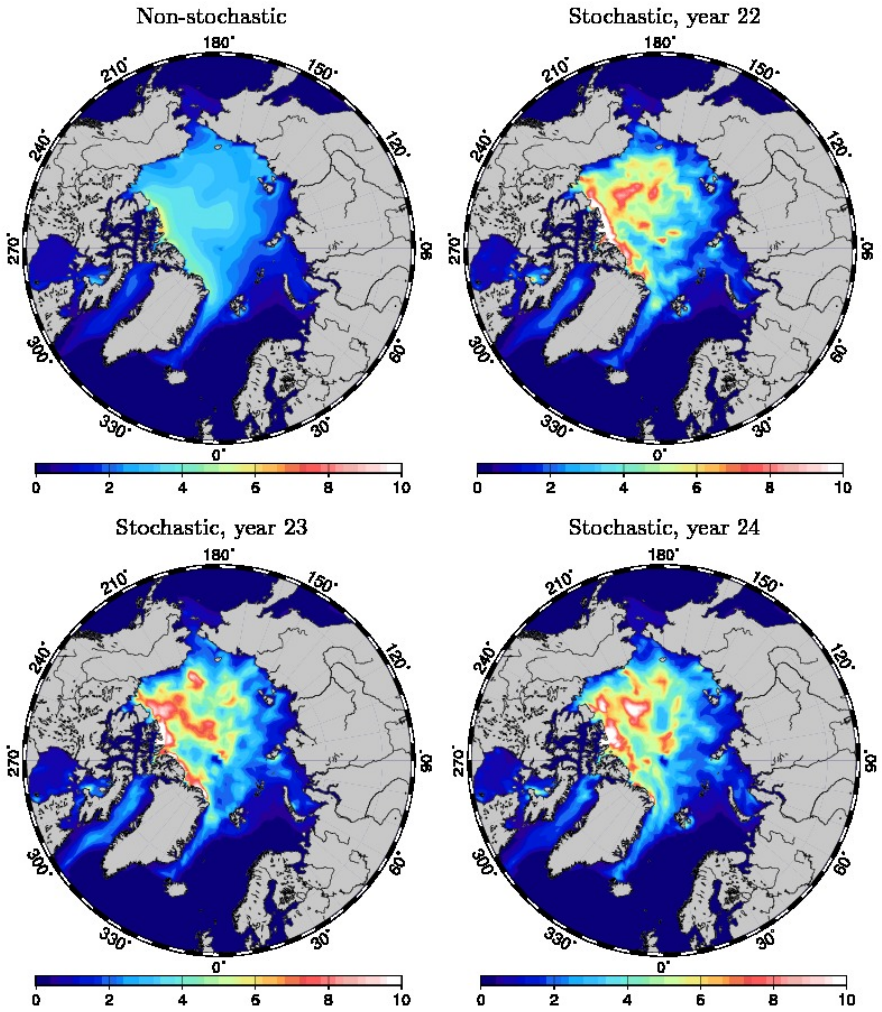


FIGURE 3.5 – Variabilité interannuelle intrinsèque de l'épaisseur de glace de mer (en mètres) produite par la paramétrisation stochastique de la résistance de la glace dans une configuration globale à basse résolution (ORCA2).

Chapitre 4

Incertitudes sur les observations

> 90% of the time the next independent measurement will fall outside the previous 90% confidence interval.
>
> Richard Hamming (1997)

L'observation consiste à acquérir de l'information à propos du comportement réel de l'océan par la mesure de quantités physiques ou biologiques variées. Cette information peut ensuite être utilisée soit pour tester une hypothèse, une théorie ou un modèle (problème direct), soit pour réduire l'incertitude inhérente à toute description simplifiée d'un système naturel complexe (problème inverse). Pour cela, il faut cependant que la nature de la relation ou de la dépendance entre l'observation et le système étudié puisse être précisée, et que l'incertitude qui lui est associée puisse être décrite. Le propos de ce chapitre est justement de nous interroger sur l'origine de cette incertitude et sur les divers moyens de la prendre en compte.

4.1 Opérateur d'observation

Supposons en premier lieu que nous disposions d'un modèle dynamique $\mathcal{M}$ pour un système océanique donné, tel que décrit en tout généralité par l'équation 1.10. A partir de lois dynamiques, éventuellement incertaines, ce modèle permet de calculer l'évolution, au fil du temps, de l'état du système $\mathbf{x}(t)$. A partir de là, *l'opérateur d'observation* peut se définir comme la relation qui permet de déduire l'équivalent simulé des quantités observées (vecteur $\mathbf{y}$) à partir de l'évolution du système $\mathbf{x}(t)$:

$$\mathbf{y} = \mathcal{H}\left[\mathbf{x}(t), \mathbf{u}(t), \mathbf{p}, \mathbf{w}\right] \tag{4.1}$$

Contrairement aux habitudes et pour des raisons qui apparaîtront ultérieurement (voir section 4.2), l'expression de $\mathcal{H}$ inclut ici un forçage $\mathbf{u}(t)$, des paramètres $\mathbf{p}$ et un forçage stochastique $\mathbf{w}$. Bien sûr, l'endroit de la séparation entre les opérateurs $\mathcal{M}$ et $\mathcal{H}$ est quelque peu arbitraire, et du point de vue de la comparaison aux observations, la seule chose qui compte vraiment, c'est la succession des deux opérateurs : $\mathcal{M}$, puis $\mathcal{H}$. Ce que nous supposerons simplement ici, c'est que $\mathcal{M}$ ne contient rien qui soit particulier à l'instrument de mesure (par ex. l'échantillonnage spatio-temporel, ou le moyen technique utilisé pour mesurer une quantité donnée), et que $\mathcal{H}$ ne contient qu'un diagnostic des quantités observées, et pas un élargissement de la définition du système que l'on résout (au sens de l'encadré 3, page 16). Cette définition n'est cependant pas complètement contraignante et nous verrons des exemples d'opérations que nous pourrions inclure indifféremment dans $\mathcal{M}$ ou $\mathcal{H}$.

Afin de rendre les choses un peu plus concrètes, voici maintenant une brève description des observations disponibles en océanographie, en essayant de les organiser selon le type d'opérateur d'observation qu'elles impliquent.

a. Interpolation spatio-temporelle. Le type le plus simple d'opérateur d'observation apparaît lorsqu'on mesure directement l'une des variables du modèle en un point et à un instant donnés. Dans ce cas, la difficulté ne provient que de l'approximation dont tout modèle numérique a besoin pour représenter des champs continus par un nombre fini de degrés de liberté. Il suffit alors de savoir comment cette approximation a été construite (éléments finis, différences finies,...) pour connaître la bonne façon d'interpoler dans l'espace et dans le temps et ainsi calculer l'équivalent simulé des observations. Cette première catégorie regroupe déjà de nombreux types d'observations "classiques" utilisés en océanographie. Les plus emblématiques sont certainement :

- les **profils de mesures *in situ*** de la pression, de la température, et de la salinité de l'eau (quel que soit le type d'instrument : CTD, flotteur dérivant,...), qui peuvent être vus comme des observations des variables T et S du modèle (indirecte pour la température potentielle) à une position verticale que l'on peut déduire de la pression. Cependant, déjà ici, on voit bien que d'autres options sont possibles. Par exemple, par un prétraitement adéquat des mesures, on peut aussi les voir comme observations de l'écart de pression entre surfaces isopycnales, de la profondeur de la couche de mélange, de la stratification de la colonne d'eau, de son contenu de chaleur, etc. Et le choix entre ces différentes options pourra dépendre de la définition du système, du type de modèle, ou bien de ce que l'on considère être l'information la plus directement pertinente pour tester le modèle (problème direct) ou pour en réduire l'incertitude (problème inverse).
- les **mesures satellitaires d'élévation du niveau de la mer** (altimétrie spatiale), qui peuvent être vues comme des observations directes de la variable η du modèle (éq. 1.8), à condition de connaître précisément la forme du géoïde (voir Birol et al., 2005, pour plus de détail). Ces observations se présentent alors sous la forme de mesures successives de η le long de la trace au sol de l'orbite, prises au moment du passage du satellite. Le rôle de l'opérateur d'observation est donc de reproduire le même mouvement dans le modèle, pour en calculer l'équivalent simulé par interpolation spatio-temporelle.

b. Intégration spatio-temporelle. Ensuite, un opérateur d'observation un peu plus sophistiqué apparaît lorsque la mesure renvoit une quantité intégrée sur une longueur, une surface, un volume ou un intervalle de temps fini (c'est-à-dire, en pratique, qui ne peut pas être négligé par rapport aux échelles que le modèle résout). Dans ce cas, outre l'interpolation, l'opérateur d'observation doit également contenir une opération d'intégration pour calculer l'équivalent simulé de ces observations. Cette seconde catégorie contient en particulier toutes les observations prétraitées qui résultent d'un regroupement et d'une moyenne de mesures élémentaires dispersées dans le temps et dans l'espace. C'est aussi le cas des images satellitaires (de température de surface, ou plus récemment de salinité de surface, grâce du satellite SMOS), quand la taille du pixel de l'image n'est pas négligeable devant la résolution du modèle. Dans un tout autre registre, l'observation des positions successives d'un flotteur dérivant requiert aussi un opérateur d'observation qui intègre le champ de vitesse pour simuler l'advection du flotteur telle que décrite par le modèle.

c. Diagnostic de quantités dérivées. Par ailleurs, en raison de la définition même du système et de ce que le modèle ne résout pas (voir l'encadré 3, page 16), il peut aussi se

faire que ce qui est mesuré ne corresponde exactement à aucune des variables du modèle. De telles observations peuvent néanmoins être très utiles et contenir une information importante sur ce que le modèle résout, même si c'est à travers une relation approximative ou incertaine. Dans ce cas, outre l'interpolation et/ou l'intégration, l'opérateur d'observation doit également contenir une opération pour diagnostiquer la quantité observée à partir des variables du modèle. Cette troisième catégorie est plus difficile à circonscrire que les précédentes puisqu'elle dépend complètement de ce qu'on décide d'inclure dans le modèle $\mathcal{M}$. Par exemple, lorsque le modèle d'océan dispose explicitement d'un module optique, alors les observations satellitaires de couleur de l'eau n'appartiennent pas à cette catégorie, puisque le spectre des radiations émises à la surface de l'océan fait partie de ce qui est simulé par le modèle $\mathcal{M}$. Mais, à l'inverse, quand le modèle ne dispose pas de module optique (comme NEMO), il existe tout de même une relation étroite entre l'état de l'écosystème et la couleur de l'océan observée par le satellite, qui peut être incluse même approximativement dans l'opérateur d'observation $\mathcal{H}$.

d. Extraction de structures. Enfin, quand l'instrument de mesure produit une image de la surface de l'océan, il peut se faire que ce qui est important dans ce qui est observé, c'est-à-dire ce qui contient l'information, n'est pas tant la valeur numérique des mesures, que les structures spatiales qui sont visibles sur l'image. Dans ce cas-là, l'opérateur d'observation doit être d'une nature assez différente de ce qui a été décrit précédemment, car il doit extraire de la solution du modèle, l'équivalent simulé des structures que l'on peut déduire des images observées. C'est ce type de diagnostic qui a été fait par exemple en section 2.3.c pour extraire la structure du manifold instable associé à l'écoulement de mésoéchelle, que l'on observe sur les images satellitaires de température de surface ou de couleur de l'eau (d'Ovidio et al., 2008; Titaud et al., 2011; Gaultier et al., 2013). Ce problème a été examiné en particulier au cours de la thèse de Lucile Gaultier.

4.2 Erreurs et incertitudes

La discussion des incertitudes sur les observations peut suivre exactement le même chemin que celui que nous avons suivi pour discuter l'incertitude sur le modèle en section 1.2. Quand le système $\mathcal{A}$ que l'on étudie (au sens de l'encadré 3, page 16) peut être supposé parfaitement déterministe (y compris l'opérateur d'observation), alors on peut aussi imaginer qu'il existe une valeur vraie $\mathbf{y}^t$ associée au vecteur d'observation :

$$\mathbf{y}^t = \mathcal{H}^t \left[\mathbf{x}^t(t), \mathbf{u}^t(t), \mathbf{p}^t\right] \tag{4.2}$$

où $\mathcal{H}^t$ est l'opérateur d'observation vrai, et $\mathbf{x}^t(t)$, $\mathbf{u}^t(t)$, $\mathbf{p}^t$ les valeurs vraies de l'état du système, du forçage et des paramètres. Dans ce cas, l'erreur d'observation $\boldsymbol{\epsilon}^o$ peut se définir par l'écart entre la valeur mesurée $\mathbf{y}^o$ et cette valeur vraie $\mathbf{y}^t$:

$$\mathbf{y}^o = \mathbf{y}^t + \boldsymbol{\epsilon}^o \quad \text{avec} \quad \boldsymbol{\epsilon}^o = \boldsymbol{\epsilon}^m + \boldsymbol{\epsilon}^r \tag{4.3}$$

Et cette erreur d'observation peut ensuite s'interpréter comme la somme d'une erreur de mesure $\boldsymbol{\epsilon}^m$ et d'une erreur de représentativité $\boldsymbol{\epsilon}^r$ (e.g. Cohn, 1997). La première est liée à l'imperfection de la relation entre la valeur mesurée et le monde réel, à cause de l'imprécision de l'instrument de mesure lui-même, ou bien à cause d'effets parasites complètement indépendants du système étudié. La seconde est liée à tout ce qui existe dans le système réel (et qui est donc vu par l'instrument), mais que le modèle ne résout pas (c'est-à-dire le système $\mathcal{B}$ de l'encadré 3, page 16). Dans nos applications, cette erreur de représentativité est donc principalement due aux échelles non-résolues et à la diversité

biologique non-résolue. Il faut bien la distinguer de l'erreur de modélisation, qui est une erreur sur ce que le modèle résout [1] (système $\mathcal{A}$), alors que l'erreur de représentativité est un erreur sur l'observation, qui résulte directement de ce que le modèle ne résout pas (système $\mathcal{B}$), et qui n'est donc pas inclut dans $\mathbf{y}^t$ (système $\mathcal{A}$). Autrement dit, il suffit de modifier le découpage entre systèmes $\mathcal{A}$ et $\mathcal{B}$ pour que de l'erreur de représentativité se transforme en erreur de modélisation, ou vice versa.

Cependant, nous avons déjà dit que le système $\mathcal{A}$ ne pouvait en général pas être supposé détermniste. Dans ce cas, la valeur vraie $\mathbf{x}^t$ n'existe pas (voir section 1.2), et $\mathbf{y}^t$ n'existe donc pas non plus. Comme en section 1.2 pour l'erreur de modélisation, le concept d'erreur d'observation doit donc être généralisé, et remplacé par le concept plus général d'incertitude, décrite par sa distribution de probabilité $p(\mathbf{y}^o|\mathbf{y})$: pour toute valeur possible de $\mathbf{y}$, $p(\mathbf{y}^o|\mathbf{y})$ donne la distribution de probabilité de la valeur mesurée $\mathbf{y}^o$. De plus, il peut exister ici une seconde raison qui fasse que $\mathbf{y}^t$ n'existe pas : c'est que l'opérateur d'observation $\mathcal{H}$ lui-même ne puisse pas être supposé déterministe (c'est-à-dire que, comme $\mathcal{M}^t$ dans l'équation 1.11, l'opérateur $\mathcal{H}^t$ de l'équation 4.2 n'existe pas). Il existe en effet des situations telles que ce que le modèle ne résout pas (système $\mathcal{B}$) influence intimement le fonctionnement de l'opérateur $\mathcal{H}$, de sorte qu'il soit préférable d'y inclure explicitement l'incertitude qui en résulte. C'est pourquoi l'équation 4.1 incluait dès le départ un forçage stochastique $\mathbf{w}$, de façon à simuler explicitement cette source potentielle d'incertitude. Cela peut permettre par exemple d'utiliser un modèle plus sophistiqué que le modèle additif de l'équation 4.3, pour décrire l'erreur de représentativité, en tout ou en partie. Ou bien, quand une même dynamique non-résolue produit à la fois de l'erreur de modélisation et de l'erreur de représentativité, on pourrait même imaginer de les décrire toutes deux en utilisant le même forçage aléatoire (mais d'une façon différente) à la fois dans $\mathcal{M}$ et dans $\mathcal{H}$. Bien sûr, quand elle est ainsi explicitement simulée dans $\mathcal{H}$ par une paramétrisation stochastique, cette partie de l'erreur de repésentativité ne soit plus être incluse dans la distribution de probabilité $p(\mathbf{y}^o|\mathbf{y})$, et le choix entre ces deux options ne peut être guidé que par la recherche d'une description aussi réaliste que possible de ces incertitudes. Voici maintenant deux exemples où le choix d'une simulation explicite dans $\mathcal{H}$ pourrait peut-être apporter quelque bénéfice.

a. Opérateur de descente d'échelle (downscaling). Le premier exemple concerne le cas d'instruments de mesure permettant de détecter explictement des échelles beaucoup plus fines que ce que le modèle résout. Cela peut être le cas par exemple d'images satellitaires de traceur à très haute résolution (température de surface, couleur de l'océan) ou des futures données altimétriques du satellite SWOT (altimétrie à large fauchée). Avec ce genre d'instrument, on peut souvent voir dans le détail des structures que le modèle ne résout pas, et qui doivent donc être traitées comme de l'erreur de représentativité. La première façon de la décrire consisterait à en spécifier la distribution de probabilité à travers $p(\mathbf{y}^o|\mathbf{y})$. Mais cela impose en général d'utiliser un modèle relativement simple de l'erreur d'observation, le plus souvent gaussien, avec une structure de corrélation spatiale et temporelle de forme spécifiée, qui ne permet pas de caractériser de façon très précise l'information contenue dans les structures observées. Cette difficulté est particulièrement aigüe quand il existe une relation de dépendance forte entre les structures fines résolues par l'instrument et les échelles plus grossières que le modèle résout. Car il devient alors nécessaire de bien caractériser cette dépendance pour exploiter correctement l'information issue des observations. Dans un tel cas, il serait certainement intéressant de compléter l'opérateur d'observation par un opérateur de downscaling exprimant ce lien entre les système que le modèle résout (système $\mathcal{A}$) et les structures fines qu'il ne résout

1. Même si elle est souvent indirectement causée par l'effet de ce que le modèle ne résout pas (système $\mathcal{B}$)

pas (système $\mathcal{B}$), quitte à en simuler explicitement l'incertitude par une paramétrisation stochastique appropriée. Cette façon de modéliser l'erreur de représentativité par downscaling stochastique vers les échelles non-résolues n'est pas sans lien avec le diagnostic de structures d'image discuté aux sections 2.3.c et 4.1.d, et serait une manière potentiellement intéressante de le prolonger, en le dotant d'une simulation explicite des incertitudes qui lui sont associées.

b. Couleur de l'eau. Le deuxième exemple concerne les instruments de mesure satellitaire de la couleur de l'océan, permettant de détecter la répartition spectrale de la lumière visible reflétée par la mer. Ce spectre dépend de façon complexe de nombreuses substances qui peuvent être contenues dans l'eau de mer, et en particulier de l'ensemble des pigments d'origine biologique (comme la chlorophylle). C'est bien sûr en raison de cette dépendance particulière que la couleur de l'eau peut-être vue comme une observation indirecte de l'écosystème marin. Cependant, même si on néglige ici tous les effets parasites indépendants de l'écosystème, la couleur de l'eau est toujours très loin de ne dépendre que de ce que le modèle résout (système $\mathcal{A}$). En raison de la grande variation dans la nature et la concentration des pigments selon les espèces, elle dépend aussi de la diversité biologique de l'écosystème que le modèle ne peut jamais résoudre de façon parfaitement détaillée (système $\mathcal{B}$). Il s'agit donc là d'une importante source d'erreur de représentativité, qu'il est encore ici difficile de caractériser de façon précise à travers $p(\mathbf{y}^o|\mathbf{y})$. Il me semble donc qu'ici encore et pour des raisons similaires, il serait préférable de simuler explicitement cette source d'incertitude par un opérateur d'observation stochastique. Car il est certainement plus difficile de construire un modèle réaliste de cette incertitude en bout de chaîne sur la mesure [à travers $p(\mathbf{y}^o|\mathbf{y})$], qu'à la source (par un bruit aléatoire $\mathbf{w}$), puis de la laisser se propager à travers l'opérateur d'observation jusqu'à le mesure.

4.3 Test de cohérence entre modèle et observations

Le principe cardinal de toute science est de produire des hypothèses, des théories ou des modèles qui puissent être testés par des observations, et qui puissent être rejetés ou amendés en cas d'incohérence. Quand le modèle est approximatif ou incertain, cette démarche implique que le modèle inclue une description de l'incertitude qu'il génère, afin qu'une éventuelle incohérence avec l'observation puisse être constatée par des moyens statistiques rigoureux. Ainsi, en tenant compte de l'incertitude sur $\mathcal{M}$ et sur $\mathcal{H}$, la prédiction que le modèle fait des quantités observées s'exprime sous la forme d'une distribution de probabilité : $p^b(\mathbf{y})$. Cette distribution caractérise l'information que le modèle contient à propos de $\mathbf{y}$. Elle peut être par exemple mesurée par l'entropie de $p^b(\mathbf{y})$, que le modélisateur essaiera alors de rendre aussi petite que possible. Ensuite, en tenant compte de l'erreur d'observation, on peut calculer la distribution de probabilité a priori pour le résultat des mesures :

$$p^b(\mathbf{y}^o) = \int p^b(\mathbf{y})\; p(\mathbf{y}^o|\mathbf{y})\; d\mathbf{y} \tag{4.4}$$

C'est cette distribution qui permet de tester la cohérence entre le modèle et les observations (ce qu'on appelle la *fiabilité* du modèle). Toute la difficulté est bien sûr de concevoir un modèle qui apporte beaucoup d'information (entropie faible) tout en restant cohérent avec les observations (modèle fiable). Par exemple, si les mesures tombent trop fréquemment en dehors de leur intervalle de confiance [décrit par $p^b(\mathbf{y}^o)$], cela signifiera que le modèle doit être rejeté, et qu'il est nécessaire d'en proposer un nouveau,

soit en reformulant les lois dynamiques elles-mêmes, soit en révisant la modélisation des incertitudes.

Il existe toute une littérature de méthodes permettant de tester la fiabilité de prévisions probabilistes (e.g. Brier, 1950; Murphy, 1973; Toth et al., 2003; Candille and Talagrand, 2005; Gneiting et al., 2008). Le propos n'est pas ici d'en faire un résumé ou une synthèse mais plutôt d'essayer d'en identifier la difficulté principale et de discuter des moyens utilisés pour la contourner. La difficulté principale de tous ces tests de cohérence provient à mon sens du fait, qu'en dépit de la nature non-déterministe du système, nous n'en observons jamais qu'une seule des réalisations possibles [2], et qu'en plus de cela, les mesures sont souvent rares et incomplètes. C'est cette difficulté-là qui nous oblige à trouver le moyen d'accumuler l'information provenant de nombreuses mesures d'éléments divers du système avant de tirer une conclusion fiable, et d'obtenir une décision sur la cohérence entre la prévision probabiliste et les observations.

a. Histogramme de rang. Le moyen aujourd'hui le plus couramment utilisé pour contourner cette difficulté et tester la fiabilité d'une prévision probabiliste est l'*histogramme de rang*, qui se concentre sur l'examen de la cohérence des distributions de probabilité marginale pour chaque quantité observée : $p^b(y_i^o)$, $i = 1, \ldots, p$ (même si des extensions multivariées existent, voir Gneiting et al., 2008). Cette méthode consiste à positionner chaque valeur mesurée $\hat{y}_i^o$ dans la distribution marginale $p^b(y_i^o)$, et à calculer le quantile qui lui correspond : $\hat{r}_i^o$. De cette manière, si chaque distribution marginale $p^b(y_i^o)$ est cohérente avec les observations, alors chacun des $\hat{r}_i^o$ doit être distribué uniformément entre 0 et 1. Et c'est précisément le fait qu'ils doivent tous être distribués de la même façon qui permet d'accumuler des observations hétérogènes en un seul diagnostic : l'histogramme des $\hat{r}_i^o$. Avec l'accumulation des observations $\hat{y}_i^o$, cet histogramme doit tendre vers une distribution uniforme entre 0 et 1. S'il en diffère significativement, ça veut dire que le modèle est incohérent avec les observations et doit en principe être rejeté (ou amendé).

On voit bien que l'élément-clé de cette méthode est l'identification de changements de variables non-linéaires qui transforment $p^b(y_i^o)$, $i = 1, \ldots, p$ en une distribution uniforme, de façon à passer des observations originales $\hat{y}_i^o$ à des observations transformées $\hat{r}_i^o$, qui sont toutes identiquement distribuées. Cette idée est assez générale et elle est d'ailleurs exactement la même que celle que nous utiliserons plus tard pour transformer les distributions marginales de chaque variable en des distributions gaussiennes identiques (anamorphose gaussienne, voir section 7.2). En pratique cependant, cette connexion peut facilement passer inaperçue, car la méthode de l'histogramme de rang est généralement présentée comme une façon de s'assurer de la cohérence d'une prévision d'ensemble, c'est-à-dire d'un échantillon de taille m de $p^b(y_i^o)$. Chacun des $\hat{r}_i^o$ est alors simplement calculé comme le rang de $\hat{y}_i^o$ (entre 0 et m) dans la suite ordonnée des membres de l'ensemble, que l'on peut alors rassembler en un histogramme de valeurs entières, entre 0 et m (histogramme de rang) qui doit tendre vers une distribution uniforme. Le rapprochement avec la méthode d'anamorphose est cependant utile car il permet d'imaginer d'utiliser une autre distribution que la distribution uniforme pour les $\hat{r}_i^o$: il suffit que ce soit toujours la même pour toutes les observations.

En particulier, le choix d'une distribution gaussienne de moyenne nulle et d'écart-type égal à 1 pourrait être une alternative intéressante, et ce pour au moins deux raisons. D'une part, l'anamorphose gaussienne des observations, déjà réalisée dans le cadre de la résolution du problème inverse (voir section 7.2), pourrait être directement utilisée

2. La seule exception à cela est le cas d'expériences idéalisées où l'incertitude sur la réalité est générée artificiellement, et où plusieurs réalisations du système peuvent donc être simulées et observées indépendamment.

pour tester la cohérence de la prévision d'ensemble. D'autre part, elle permettrait une caractérisation plus synthétique de l'incohérence de l'ensemble par les principaux moments de la distribution des $\hat{r}_i^o$. Ainsi, une moyenne des $\hat{r}_i^o$ différente de 0 indiquerait que les observations sont trop souvent plus grandes ou plus petites que la médiane de l'ensemble ; un écart-type différent de 1 signifierait que l'ensemble est trop ou pas assez dispersif ; et une valeur non-nulle pour les moments suivants impliqueraient d'autres formes d'incohérence.

a. Vraisemblance des observations. Cependant, il est important de noter qu'en ne regardant que les distributions marginales, l'histogramme de rang (du moins dans sa version univariée) n'épuise pas les possibilités d'incohérence entre modèle et observations. Il ne dit rien en particulier sur la fiabilité des dépendances statistiques entre les différentes variables, qui ne peuvent de toute façon être testées qu'avec un ensemble de très grande taille et de nombreuses observations. Mais cela ne doit pas empêcher de rechercher l'incohérence sous des angles divers, tant que ça reste possible avec les ressources et les observations disponibles. Un autre angle possible est par exemple le calcul de la vraisemblance du vecteur des valeurs mesurées $\hat{\mathbf{y}}^o$: $L = p^b(\hat{\mathbf{y}}^o)$, car pour que le modèle soit cohérent avec les observations, il faut que l'espérance mathématique de $-\ln L$ soit égale à l'entropie S^b associée à $p^b(\mathbf{y}^o)$:

$$\Delta =< S^b - \ln L >= \int \left[p^b(\mathbf{y}^o) - p^t(\mathbf{y}^o)\right] \ln p^b(\mathbf{y}^o)\, d\mathbf{y}^o = 0 \quad \text{si} \quad p^b(\mathbf{y}^o) = p^t(\mathbf{y}^o) \quad (4.5)$$

où $p^t(\mathbf{y}^o)$ est la distribution de probabilité des mesures effectuées sur le monde réel. Cet écart est égal à zéro quand modèle et observation sont cohérents, et s'il est différent de zéro, cela signifie soit que l'on surestime l'information que le modèle contient sur le système ($\Delta > 0$), soit qu'on la sous-estime ($\Delta < 0$). Cette mesure de l'incohérence est intéressante car c'est une quantité scalaire interprétable (en terme d'entropie) qui accumule l'information provenant de nombreuses observations, et qui dépend des corrélations entre variables (contrairement à l'histogramme de rang univarié).

Par exemple, quand $p^b(\mathbf{y}^o)$ est une distribution gaussienne de moyenne $\overline{\mathbf{y}^o}$ et de covariance $\mathbf{P}$, l'équation $\Delta = 0$ se ramène à :

$$< (\mathbf{y}^o - \overline{\mathbf{y}^o})^T \mathbf{P}^{-1} (\mathbf{y}^o - \overline{\mathbf{y}^o}) >= p \quad (4.6)$$

où p est la taille du vecteur d'observation. Et nous retrouvons le diagnostic classique proposé par O. Talagrand pour tester la cohérence statistique d'un système d'assimilation de données. Ce test dépend bien de la structure de corrélation simulée dans $\mathbf{P}$, et sera donc sensible à une mauvaise modélisation de la corrélation entre les différentes variables du système. Par ailleurs, dans le cas univarié, l'équation 4.6 se ramène simplement au test de cohérence habituel entre la variance de l'écart aux observations et la somme des variances des erreurs de prévision et des erreurs d'observation.

Pour conclure, je crois qu'il est utile de répéter que tous ces tests de cohérence ne pourront toujours donner que des conditions suffisantes de rejet, et jamais des conditions suffisantes d'acceptation du modèle. Ceci est directement lié au constat logique (à la base de l'épistémologie poppérienne) qu'il est impossible de vérifier explicitement un énoncé général par un nombre fini de données empiriques (sous-détermination des théories par l'expérience). Le rôle premier des observations est donc de remettre constamment en question les idées qui semblent établies, en nous montrant que notre incertitude est souvent bien plus grande qu'on ne le présume (voir citation en entête de chapitre).

Deuxième partie

Problème inverse

Chapitre 5

Réduction des incertitudes

Beware of finding
what you are looking for.

Richard Hamming

Les chapitres précédents ont montré combien un traitement approprié des incertitudes peut être utile à la résolution d'un problème direct en océanographie : (i) pour décrire explicitement l'effet de ce que le modèle ne résout pas (chapitre 1), par exemple par des paramétrisations stochastiques (chapitre 3), (ii) pour donner un sens probabiliste aux prévisions de l'océan, via des simulations d'ensemble (chapitre 2), (iii) pour tester objectivement la cohérence entre modèle et observations par des moyens probabiliste (chapitre 4). A partir de ce chapitre et dans toute la suite de ce mémoire, nous nous concentrerons maintenant sur le problème inverse, c'est-à-dire sur les moyens d'acquérir de l'information sur le système à partir d'observations de l'océan. Nous verrons en particulier que, pour cela, disposer d'une description fiable des incertitudes sur le modèle constitue un avantage précieux.

L'objectif de ce premier chapitre un peu général est de présenter et discuter ce qui est à mon avis la difficulté essentielle des problème inverse en océanographie. Cette difficulté provient directement du nombre souvent très grand de sources indépendantes d'incertitude, et de l'impossibilité d'explorer explicitement toutes les combinaisons possibles. Ceci nous permettra de comprendre pourquoi seule une petite fraction des problèmes inverse océanographiques bien-posés peuvent être résolus en pratique, et nous amènera directement aux différents types de solutions partielles at approximations présentées dans les chapitres qui suivent.

5.1 L'approche bayésienne

Le problème direct consiste à prédire l'évolution de l'état d'un système océanique $\mathbf{x}(t)$, ainsi que de quantités observables $\mathbf{y}$ (l'équivalent d'observations) en utilisant un modèle $\mathcal{M}$ et un opérateur d'observation $\mathcal{H}$:

$$\begin{cases} \frac{d\mathbf{x}}{dt} = \mathcal{M}(\mathbf{x}, \mathbf{u}, \mathbf{p}, \mathbf{w}) \quad ; \quad \mathbf{x}(0) = \mathbf{x}^0 \\ \mathbf{y} = \mathcal{H}\left[\mathbf{x}, \mathbf{u}, \mathbf{p}, \mathbf{w}\right] \end{cases} \tag{5.1}$$

où $\mathbf{x}^0$ est la condition initiale ; $\mathbf{u}$, le forçage par le monde extérieur (principalement aux frontières) ; $\mathbf{p}$, les paramètres du modèle ; et $\mathbf{w}$, un vecteur aléatoire décrivant l'incertitude sur $\mathcal{M}$ et $\mathcal{H}$ (principalement liée à ce que le modèle ne résout pas). En raison des

incertitudes sur $\hat{\mathbf{x}} = [\mathbf{x^0}, \mathbf{u}, \mathbf{p}, \mathbf{w}]$, le problème direct ne peut donner qu'une information incomplète sur $\mathbf{x}(t)$ et $\mathbf{y}$: il s'agit donc de traduire l'incertitude sur $\hat{\mathbf{x}}$, décrite par une distribution a priori $p^b(\hat{\mathbf{x}})$, en incertitude sur $\mathbf{x}(t)$ et $\mathbf{y}$. C'est ce qui a été fait aux chapitres 2 et 3 de ce travail.

Le problème inverse consiste, quant à lui, à réduire l'incertitude (ou gagner de l'information) sur $\hat{\mathbf{x}}$ (le vecteur de contrôle, supposé contenir toutes les sources significatives d'incertitude [1]), et donc sur $\mathbf{x}(t)$ et $\mathbf{y}$, grâce à l'information supplémentaire qui provient des observations $\mathbf{y}^o$ (décrites au chapitre 4). Une approche générale pour le résoudre, qui permet de fonder la plupart des méthodes utilisées en pratique, est l'approche bayésienne, directement dérivée du théorème de Bayes :

$$p^a(\hat{\mathbf{x}}) \propto p^b(\hat{\mathbf{x}})\ p(\mathbf{y}^o|\hat{\mathbf{x}}) \tag{5.2}$$

Dans cette expression, $p(\mathbf{y}^o|\hat{\mathbf{x}})$ est la distribution de probabilité conditionnelle d'observations hypothétiques $\mathbf{y}^o$ pour une valeur fixée de $\hat{\mathbf{x}}$. Elle dépend donc des opérateurs $\mathcal{M}$ et $\mathcal{H}$ (déterministes pour $\hat{\mathbf{x}}$ fixé) pour passer de $\hat{\mathbf{x}}$ à $\mathbf{y}$, et de la distribution de probabilité $p(\mathbf{y}^o|\mathbf{y})$ de l'erreur d'observation pour passer de $\mathbf{y}$ à $\mathbf{y}^o$. Considérée comme fonction de $\hat{\mathbf{x}}$ (comme ici), $p(\mathbf{y}^o|\hat{\mathbf{x}})$ mesure la vraisemblance de chaque valeur de $\hat{\mathbf{x}}$ pour la valeur réellement obtenue pour les observations $\mathbf{y}^o$. D'autre part, $p^a(\hat{\mathbf{x}})$ est la distribution de probabilité a posteriori pour $\hat{\mathbf{x}}$, c'est-à-dire conditionnée aux observations $\mathbf{y}^o$. Par rapport à la distribution de probabilité a priori $p^b(\hat{\mathbf{x}})$, on voit que densité de probabilité est amplifiée dans les régions de l'espace de contrôle que les observations rendent plus vraisemblables [$p(\mathbf{y}^o|\hat{\mathbf{x}})$ élevé] et diminuée ailleurs. C'est cette concentration de probabilité qui matérialise le gain d'information sur $\hat{\mathbf{x}}$ apporté par les observations $\mathbf{y}^o$, c'est-à-dire la diminution de l'entropie de la distribution de probabilité pour $\hat{\mathbf{x}}$: $S^a \leq S^b$.

5.2 Méthodes de Monte Carlo

Cependant, dans les applications océanographiques réelles, il est en général impossible d'évaluer explicitement $p^a(\hat{\mathbf{x}})$ en appliquant directement l'équation (5.2), même lorsque la distribution de probabilité a priori pour le vecteur de contrôle [$p^b(\hat{\mathbf{x}})$] et la distribution de probabilité pour l'erreur d'observation [$p(\mathbf{y}^o|\mathbf{y})$] sont connues explicitement sous une forme analytique simple. La raison de cela est qu'il est en général impossible d'expliciter $p(\mathbf{y}^o|\hat{\mathbf{x}})$ en tant que fonction de $\hat{\mathbf{x}}$ (en raison de la complexité des opérateurs $\mathcal{M}$ et $\mathcal{H}$). Dans le sens direct, il est en principe facile de calculer explicitement la distribution de probabilité conditionnelle d'observations hypothétiques $p(\mathbf{y}^o|\mathbf{y})$ pour une valeur fixée de $\hat{\mathbf{x}}$. Il suffit pour cela d'appliquer une seule fois les opérateurs $\mathcal{M}$ et $\mathcal{H}$ (équation 5.1) pour obtenir $\mathbf{y}(\hat{\mathbf{x}})$ et d'en déduire directement $p(\mathbf{y}^o|\hat{\mathbf{x}}) = \mathbf{p}[\mathbf{y^o}|\mathbf{y}(\hat{\mathbf{x}})]$ en tenant compte de l'erreur d'observation [$p(\mathbf{y}^o|\mathbf{y})$]. Dans le sens inverse, par contre, le problème est en général bien plus complexe, en raison de l'impossibilité de résoudre l'équation (5.1) en une expression analytique explicite de $\mathbf{y}$ en fonction de $\hat{\mathbf{x}}$, qui livrerait directement l'expression de $p^a(\hat{\mathbf{x}})$ à travers l'équation (5.2).

Une façon générale de résoudre le problème est de produire un échantillon de $p^a(\hat{\mathbf{x}})$ et d'augmenter la taille de l'échantillon jusqu'à ce que la distribution de probabilité soit décrite avec assez de précision (méthode de Monte Carlo). Supposons donc qu'une distribution de probabilité $p(\hat{\mathbf{x}})$ pour le vecteur de contrôle soit décrite approximativement par un échantillon de taille r : $[\hat{\mathbf{x}}_\mathbf{1}, \hat{\mathbf{x}}_\mathbf{2}, \ldots, \hat{\mathbf{x}}_\mathbf{r}]$, avec des poids ω_i (de somme égale à 1) éventuellement différents pour chacun de membres de l'échantillon :

1. Ou bien une partie d'entres-elles (voir la discussion sur l'erreur modèle en section 6.1), mais il est plus simple ici pour la présentation de supposer que $\hat{\mathbf{x}}$ inclut toutes les incertitudes.

$$p(\hat{\mathbf{x}}) \simeq \sum_{\mathbf{i=1}}^{\mathbf{r}} \omega_{\mathbf{i}}\, \delta(\hat{\mathbf{x}} - \hat{\mathbf{x}}_{\mathbf{i}}) \tag{5.3}$$

L'application du théorème de Bayes (éq. 5.2) revient alors simplement à multiplier chacun des poids ω_i par la vraisemblance du membre de l'échantillon correspondant (en renormalisant leur somme à 1) :

$$\omega_i^a \propto \omega_i^b\, p(\mathbf{y}^o|\hat{\mathbf{x}}_{\mathbf{i}}) \tag{5.4}$$

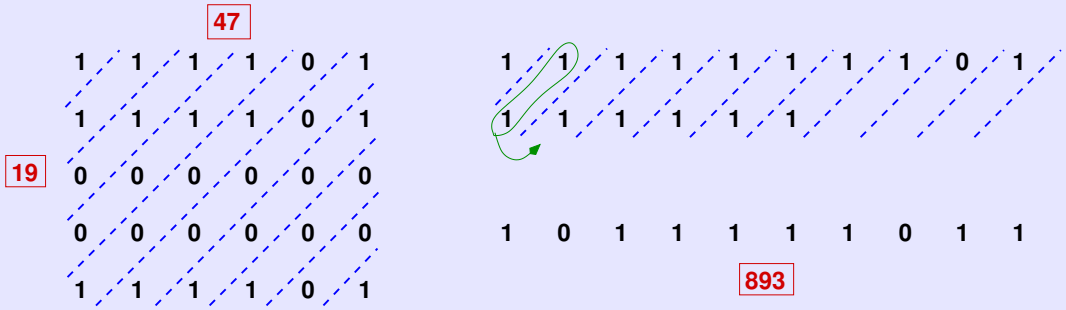
Les membres de l'échantillon que les observations rendent plus vraisemblables seront donc privilégiés par rapport aux autres, de façon exactement conforme au théorème de Bayes. Les équations (5.3) et (5.4) sont à la base des filtres particulaires (introduit par Gordon et al., 1993) qui résolvent en principe le problème inverse dans toute sa généralité (voir van Leeuwen, 2009, pour une revue de leur applications aux problèmes géophysiques).

D'un point de vue algorithmique, on peut observer (eq. 5.4) que cette méthode requiert autant d'évaluation de $p(\mathbf{y}^o|\hat{\mathbf{x}})$ que de membres dans l'ensemble, ce qui exige autant d'application du modèle $\mathcal{M}$ et de l'opérateur d'observation $\mathcal{H}$ (pour chaque $\hat{\mathbf{x}}_{\mathbf{i}}$). Elle pourrait donc s'inscrire dans le prolongement direct des simulations d'ensemble décrites au chapitre 2, en permettant d'incorporer à la prévision d'ensemble l'information issue d'observations (par une simple modification de leurs poids relatifs ω_i). Cependant, la méthode soufre d'une difficulté essentielle : le nombre de particules nécessaires pour obtenir une description correcte de la distribution de probabilité a posteriori $p^a(\hat{\mathbf{x}})$ croît exponentiellement avec la taille du vecteur $\hat{\mathbf{x}}$ c'est-à-dire le nombre de dimensions de l'espace de contrôle. Ce phénomène (connu sous le nom de "malédiction des dimensions") est intuitivement facile à comprendre. A supposer qu'il faille k particules pour explorer correctement chacune des n dimensions de l'espace de contrôle, alors il faudra k^p particules pour explorer toutes les combinaisons possibles, et détecter de façon certaine la ou les régions de l'espace de contrôle que les observations rendent vraisemblables. Réciproquement, cela signifie qu'il faudra un nombre exponentiellement grand de particules pour en détecter, ne fût-ce qu'une seule, qui soit raisonnablement proche des observations. Bien sûr, de nombreuses variantes de cette méthode ont été proposées (voir par exemple van Leeuwen, 2009), mais celles-ci ne résolvent pas le problème du coût exponentiel, à moins d'exiger des hypothèses si contraignantes qu'elles font perdre le caractère réellement général de la méthode.

A moins de circonstances particulières qui simplifient le problème, il n'existe en effet pas (encore) d'algorithme général donnant la solution d'un problème inverse quelconque en temps polynomial, même si l'application de l'opérateur direct se fait en temps polynomial. C'est donc une erreur de penser qu'un problème inverse devrait être soluble simplement parce qu'il est bien posé, qu'une solution utile existe et parce que le problème direct peut être résolu de façon extrêmement efficace. Afin d'insister sur ce point crucial, il m'a semblé utile de l'illustrer par le problème de la factorisation d'un grand nombre entier (de typiquement 250 chiffres) en ses facteurs premiers (voir encadré 6, page 56), qui ne peut encore être résolu par aucun moyen actuel, alors même que la solution existe (en vertu du théorème fondamental de l'arithmétique), et que la multiplication de facteurs (le problème direct) s'effectue en un temps dérisoire. C'est en raison de cette difficulté essentielle qu'il est nécessaire d'imaginer des approches simplifiées, assorties d'approximations adaptées à la nature des opérateurs $\mathcal{M}$ et $\mathcal{H}$ que l'on peut rencontrer en océanographie.

Encadré 6 : Multiplication et factorisation

La problème direct peut-être le plus simple que l'on puisse imaginer est la multiplication de deux nombres entiers. Il existe des algorithmes très simples et efficaces pour le résoudre. Par exemple, si nous voulons multiplier les deux nombres premiers 19 *(en binaire :* 10011*) et* 47 *(*101111*), il suffit de les disposer en matrice (figure ci-dessous, à gauche) en multipliant chaque bit de l'un des nombres par l'autre. Chaque* 1 *sur une même diagonale* i *(traits interrompus) possède alors la même valeur* (2^i)*, et on peut les faire glisser vers le haut et vers la droite le long des diagonales (figure ci-dessous, en haut à droite). Par ailleurs, deux* 1 *sur une même diagonale* $(2^i + 2^i)$ *se reportent sous la forme d'un seul* 1 *sur la diagonale suivante* (2^{i+1})*. En enchaînant ainsi glissements et reports, on obtient directement le produit (figure en bas à droite) :* 893 *(*1101111101*). Cet algorithme (correspondant grosso modo à la "multiplication longue") n'est pas le plus efficace, mais résout le problème en* n^2 *opérations binaires, si* n *est le nombre de bits des nombres à multiplier.*



Le problème inverse correspondant est la recherche des facteurs premiers d'un nombre composite. Ce problème est bien plus complexe que le problème direct, au point qu'il n'existe à l'heure actuel aucun algorithme qui le résolve en temps polynomial (en n*). Si on repart de l'algorithme précédent et que l'on recherche par quelle suite inverse de reports et de glissements le nombre final peut se décomposer (en une table de* 1 *avec des colonnes et des lignes de* 0*, comme sur la figure de gauche), on se retrouve rapidement devant une tâche impraticable de recherche de possibilités combinatoires (quand* n *devient grand). Et même avec les meilleurs algorithmes et les plus gros calculateurs actuels, aucun nombre semi-premier (produit de deux nombres premiers) de plus de 768 bits n'a jamais été factorisé (http ://en.wikipedia.org/wiki/RSA_Factoring_Challenge).*

C'est d'ailleurs l'impossibilité pratique de résoudre ce problème inverse qui est à la base du système de cryptage RSA (http ://en.wikipedia.org/wiki/RSA_algorithm) utilisé aujourd'hui pour sécuriser la plupart des communications informatiques : le produit des deux nombres premiers (clé publique) suffit pour crypter, alors que les facteurs premiers (clé privée, impossible à déduire de la clé publique) sont nécessaires pour décrypter. Si donc il existait une méthode générale de résolution de ce type de problème inverse en temps polynomial (ce qui n'a jamais non plus été démontré), cela signifierait que la sécurité du cryptage RSA ne serait plus garantie. Comme cette perspective semble encore très loin d'être réalisable, je crois que nous serons encore pour longtemps contraints d'explorer et d'élargir peu à peu la classe des problèmes inverses océanographiques que l'on peut résoudre approximativement, moyennnant des hypothèses simplificatrices fortes (comme aux chapitres 6 et 7).

5.3 Estimateurs

Une approche simplifiée pour résoudre un problème inverse consiste à rechercher une représentation réduite de la distribution de probabilité a posteriori $p^a(\hat{\mathbf{x}})$, en premier lieu en recherchant simplement un *meilleur estimé* du vecteur de contrôle $\hat{\mathbf{x}}$. Ce meilleur estimé aura bien sûr un intérêt d'autant plus grand que l'incertitude a posteriori est faible, c'est-à-dire quand le vecteur de contrôle est bien contraint par les observations disponibles, de sorte que quasiment toute la probabilité a posteriori se concentre dans un certain voisinage du meilleur estimé. C'est donc l'hypothèse que nous ferons dans cette section. Il existe néanmoins plusieurs estimateurs possibles que nous allons examiner tout à tour.

Estimateur du maximum de probabilité. Une première solution est de rechercher le mode de la distribution de probabilité a posteriori $p^a(\hat{\mathbf{x}})$:

$$\hat{\mathbf{x}}_{\text{mod}} = \arg\max_{\hat{\mathbf{x}}} p^a(\hat{\mathbf{x}}) \tag{5.5}$$

c'est-à-dire le minimum d'une fonction coût $J(\hat{\mathbf{x}}) = -\ln p^a(\hat{\mathbf{x}}) + \text{cte}$:

$$J(\hat{\mathbf{x}}) = J^b(\hat{\mathbf{x}}) + J^o(\hat{\mathbf{x}}) \quad \text{avec} \quad \begin{cases} J^b(\hat{\mathbf{x}}) = -\ln p^b(\hat{\mathbf{x}}) + \text{cte} \\ J^o(\hat{\mathbf{x}}) = -\ln p(\mathbf{y}^o|\hat{\mathbf{x}}) + \text{cte} \end{cases} \tag{5.6}$$

où $J^b(\hat{\mathbf{x}})$ est le terme d'ébauche et $J^o(\hat{\mathbf{x}})$, le terme d'observation. L'intérêt de cet estimateur $\hat{\mathbf{x}}_{\text{mod}}$ est qu'il peut être calculé efficacement (en temps polynomial) pour une large classe de problèmes inverses : il suffit pour cela (i) que les opérateurs $\mathcal{M}$ et $\mathcal{H}$ soient différentiables, et (ii) que la distribution $p^a(\hat{\mathbf{x}})$ soit unimodale (c'est-à-dire que la fonction coût ne possède pas de minima locaux secondaires, ou plus précisément que son gradient ne s'annule qu'en $\hat{\mathbf{x}}_{\text{mod}}$). Sous l'hypothèse (i), il est en effet toujours possible de calculer le gradient de la fonction coût $\nabla J(\hat{\mathbf{x}})$, en appliquant une seule fois l'opérateur direct ($\mathcal{M}$ et $\mathcal{H}$), puis l'opérateur adjoint (généralement de coût similaire, mais qu'il faut néanmoins pouvoir implémenter). En utilisant le gradient, on peut alors appliquer une méthode itérative de descente (par exemple la méthode des gradients conjugués) pour converger rapidement vers le minimum le plus proche. Or, sous l'hypothèse (ii), ce minimum est unique, et ne peut correspondre qu'à $\hat{\mathbf{x}}_{\text{mod}}$. Grâce aux hypothèses (i) et (ii), il n'est donc plus nécessaire d'explorer exhaustivement l'espace de contrôle (c'est-à-dire toutes les combinaisons possibles dans $\hat{\mathbf{x}}_{\text{mod}}$) pour détecter la "petite région" de forte probabilité a posteriori (c'est-à-dire que les observations rendent vraisemblable) : la malédiction des dimensions a été vaincue.

Cet algorithme (généralement assorti de multiples raffinements) est à la base des méthodes variationnelles d'assimilation de données (4DVAR [2]) couramment utilisées dans les systèmes de prévisions météorologiques actuels. L'algorithme peut également être complété d'une mesure de l'incertitude sur $\hat{\mathbf{x}}_{\text{mod}}$, donnée par la courbure de $J(\hat{\mathbf{x}})$ en $\hat{\mathbf{x}}_{\text{mod}}$ (matrice hessienne), mais ce calcul est plus coûteux et n'est en général pas mis en œuvre dans les applications de grande taille.

Estimateur du maximum de vraisemblance. Une variante de l'estimateur précédent, très souvent amplement décrite dans la plupart des manuels de probabilité, consiste à rechercher le valeur du vecteur de contrôle $\hat{\mathbf{x}}$ que les observations rendent la plus vraisemblable (maximum likelihood) :

2. Appelé 4DPSAS quand le vecteur de contrôle contient un vecteur aléatoire $\mathbf{w}$ décrivant une incertitude sur le modèle

$$\hat{\mathbf{x}}_{\text{lh}} = \arg\max_{\hat{\mathbf{x}}} p(\mathbf{y}^o|\hat{\mathbf{x}}) \tag{5.7}$$

ce qui revient à calculer le minimum de $J^o(\hat{\mathbf{x}})$ seul dans l'équation (5.5). Bien sûr, ce problème ne peut être bien posé que si le vecteur d'observation $\mathbf{y}^o$ est plus grand (et souvent bien plus grand) que le vecteur de contrôle $\hat{\mathbf{x}}$. C'est pourquoi le terme $J^b(\hat{\mathbf{x}})$ dans l'équation (5.5) est souvent présenté par les mathématiciens comme un terme de régularisation, qui permet de transformer un problème mathématique mal posé en un problème bien posé quand les observations manquent. Cependant, il faut garder à l'esprit qu'en pratique (et notamment dans les applications océanographiques) la distribution de probabilité a priori $p^b(\hat{\mathbf{x}}) = \exp[-J_b(\hat{\mathbf{x}})]$ peut déjà contenir une information essentielle, et que le propos du problème inverse peut n'être que de compléter cette information par de nouvelles observations. En outre, plus les observations manquent, plus la solution du problème sera sensible à la paramétrisation choisie pour $p^b(\hat{\mathbf{x}})$, qu'il convient donc de rendre aussi réaliste que possible.

Estimateur du minimum de la variance. Une autre solution est de rechercher l'espérance mathématique de la distribution de probabilité a posteriori $p^a(\hat{\mathbf{x}})$:

$$\hat{\mathbf{x}}_{\text{exp}} = \int \hat{\mathbf{x}}\, p^a(\hat{\mathbf{x}})\, d\hat{\mathbf{x}} \tag{5.8}$$

qui possède l'intérêt de minimiser la variance de l'erreur commise. L'intégrale de l'équation 5.8 peut être calculée par méthode de Monte Carlo à partir d'un échantillon de $p^a(\hat{\mathbf{x}})$. A partir d'un tel échantillon, on peut alors calculer, outre l'espérance mathématique, n'importe quelle propriété particulière de $p^a(\hat{\mathbf{x}})$, telle que la variance de l'incertitude a posteriori, son asymétrie autour de l'espérance mathématique,... Cependant, comme pour les méthodes générales présentées en section 5.2, la difficulté est alors de produire l'échantillon requis. Mais si on suppose cette fois la distribution unimodale et l'incertitude faible, cela peut se ramener à détecter d'abord la petite région de l'espace de contrôle qui concentre l'essentiel de la distribution de probabilité a posteriori au voisinage de $\hat{\mathbf{x}}_{\text{exp}}$. Sous cette hypothèse, cela peut se faire par un algorithme de minimisation, comme pour déterminer $\hat{\mathbf{x}}_{\text{mod}}$. Une fois que cela est accompli, on peut ensuite échantillonner la petite région siginifcative autour de $\hat{\mathbf{x}}_{\text{mod}}$ par une méthode d'échantillonnage efficace (échantillonneur de Gibbs ou algorithme de Metropolis/Hastings). Ce genre d'approche sera examinée plus en détail en section 7.4.

En résumé, il me semble donc que cette approche par estimateurs est particulièrement adaptée quand le problème inverse est bien contraint par les observations, de façon à concentrer la probabilité a posteriori en une seule région de l'espace de contrôle (un seul mode, sans maxima secondaires). Dans cette circonstance en effet, l'algorithme de minimisation et d'exploitation optimale des observations peut donner toute sa mesure, surtout s'il peut être complété par une méthode d'échantillonnage efficace une fois que le mode a été détecté (voir par exemple section 7.4).

5.4 Modélisation des incertitudes

Un autre genre d'approche simplifiée pour résoudre le problème inverse consiste à considérer que, malgré la complexité de $\mathcal{M}$ et $\mathcal{H}$, la distribution de probabilité a priori pour $\mathbf{x}(t)$ et $\mathbf{y}$, résultant de l'incertitude a priori sur le vecteur de contrôle décrire par $p^b(\hat{\mathbf{x}})$, garde une structure simple. Cette approche est donc basée sur un modèle a priori pour la distribution de probabilité (souvent décrit par un nombre limité de paramètres). L'objectif de l'approximation est qu'il soit possible d'estimer facilement les pa-

ramètres de la distribution conjointe $p^b(\hat{\mathbf{x}}, \mathbf{y})$, soit directement, soit à partir d'une simulation d'ensemble relativement modeste, pour ensuite appliquer directement le théorème de Bayes (eq. 5.2), sous la forme :

$$p^a(\hat{\mathbf{x}}, \mathbf{y}) \propto p^b(\hat{\mathbf{x}}, \mathbf{y})\ p(\mathbf{y}^o|\hat{\mathbf{x}}) \tag{5.9}$$

d'où la distibution marginale $p^a(\hat{\mathbf{x}})$ peut se déduire par intégration sur $\mathbf{y}$:

$$p^a(\hat{\mathbf{x}}) = \int p^a(\hat{\mathbf{x}}, \mathbf{y})\ \mathbf{dy} \tag{5.10}$$

en profitant de la forme simple de $p^a(\hat{\mathbf{x}}, \mathbf{y})$.

Pour suivre cette approche, le problème principal est donc de déterminer une forme paramétrique de la distribution conjointe a priori qui soit à la fois suffisamment générale pour être réaliste, et suffisamment simple pour rendre le problème traitable. Dans leur revue des méthodes de Monte Carlo appliquées à la résolution de problèmes inverses en géophysiques, Sambridge and Mosegaard (2002) proposent de les classer selon leur tendance à explorer la distribution de probabilité en utilisant l'échantillon disponible (c'est-à-dire, pour nous, la simulation d'ensemble), ou inversement à exploiter des hypothèses restrictives sur la forme de la distribution. Bien sûr, moins il y a d'hypothèses, plus la forme de la distribution est générale, et plus la taille de l'ensemble doit être grande pour permettre une exploration fine de la forme précise de la distribution. La figure 5.1 est une tentative de particulariser cette classification aux formes paramétriques possibles pour la distribution de probabilité a priori $p^b(\hat{\mathbf{x}}, \mathbf{y})$, en insistant particulièrement sur celles qui seront discutées dans ce travail.

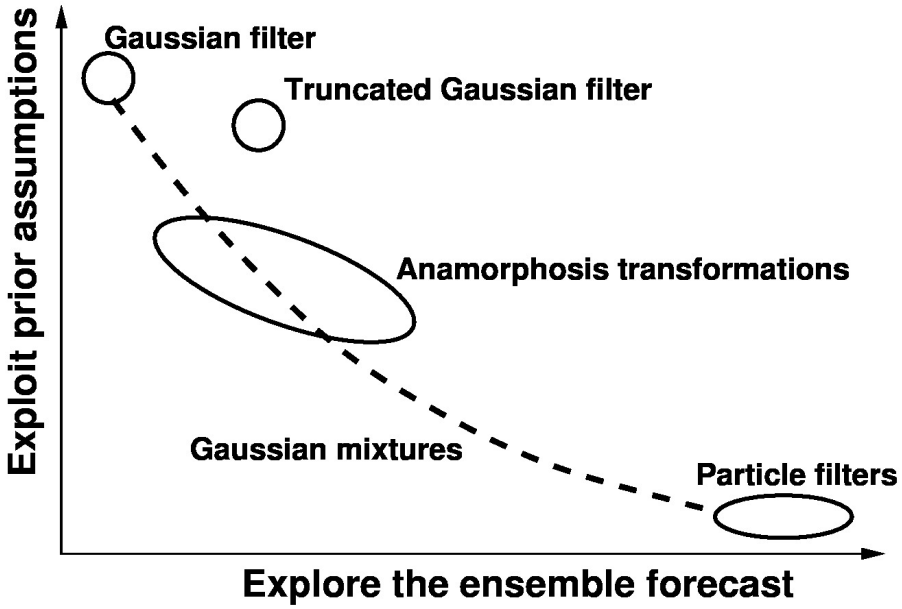


FIGURE 5.1 – Classification des modèles d'incertitude selon leur tendance à explorer la distribution de probabilité en utilisant la prévision d'ensemble (en abscisse), ou leur tendance à exploiter des hypothèses restrictives sur la forme de la distribution (en ordonnée).

Modèle gaussien. A l'extrémité la moins exploratoire du schéma, on trouve la distribution de probabilité gaussienne, dont les seuls paramètres sont la moyenne et la covariance. A partir d'une simulation d'ensemble, l'estimation optimale de ces paramètres

est immédiate, car ils correspondent simplement à la moyenne et à la covariance de l'ensemble. La propriété la plus importante de ce modèle est qu'il est préservé par une transformation linéaire. Ce modèle s'impose donc naturellement dans le cas de problèmes linéaires, mais son domaine utile d'application déborde ce cas particulier. Il sera discuté plus en détail au chapitre 6.

Gaussiennes tronquées. Pour une hypothèse a priori légèrement plus générale, on trouve le modèle des gaussiennes tronquées (Lauvernet et al., 2009), qui est approprié pour tenir compte de contraintes d'inégalité sur les variables de contrôle. Ici, l'exploration de l'ensemble n'est nécessaire que pour identifier le vecteur de localisation et la matrice d'échelle de la gaussienne tronquée (qui correspondent à la moyenne et à la covariance de la distribution gaussienne de base, mais sont plus compliqués à identifier à partir d'une simulation d'ensemble). Ce modèle sera discuté plus en détail en section 7.1.

Transformations anamorphiques. Une autre variante du modèle gaussien encore plus exploratoire consiste à rechercher, pour chaque variable du vecteur de contrôle, un changement de variable non-linéaire qui transforme sa distribution de probabilité marginale en une distribution gaussienne. Les paramètres de la transformation peuvent être diagnostiqués à partir de l'ensemble, par exemple en identifiant certains quantiles de l'ensemble aux quantiles correspondants de la distribution gaussienne (Béal et al., 2010; Brankart et al., 2012). Ce modèle sera discuté en section 7.2.

Mélange de gaussiennes. Une formulation potentiellement encore plus exploratoire que les précédentes consiste à supposer que la distribution de probabilité a priori est faite de la superposition (ou du mélange) de distributions gaussiennes élémentaires (Anderson and Anderson, 1999; Bengtsson et al., 2003; Hoteit et al., 2008; Sondergaard and Lermusiaux, 2013). En fonction du nombre q de gaussiennes superposées, le niveau d'exploration peut être rendu aussi faible qu'avec le modèle gaussien ($q = 1$), ou virtuellement aussi grand qu'avec une méthode de Monte Carlo générale ($q \to \infty$). A cette limite, la méthode devient pratiquement non-paramétrique et on parle alors parfois de méthode particulaire avec "habillage" des particules par des noyaux gaussiens. Cette approche sera discutée plus en détail en section 7.3.

Chapitre 6

Le modèle gaussien

> Least squares are popular for solving inverse problems because they lead to the easiest computations.
>
> Albert Tarantola (2005)

Le modèle le plus populaire pour décrire l'incertitude a priori sur le vecteur de contrôle est le modèle gaussien :

$$p^b(\hat{\mathbf{x}}) \;=\; \mathcal{N}_{\hat{\mathbf{x}}}(\hat{\mathbf{x}}^b_{\text{exp}}, \hat{\mathbf{P}}^b) \;\propto\; \exp\left[-\frac{1}{2}(\hat{\mathbf{x}} - \hat{\mathbf{x}}^b_{\text{exp}})^T \hat{\mathbf{P}}^{b\,-1}(\hat{\mathbf{x}} - \hat{\mathbf{x}}^b_{\text{exp}})\right] \tag{6.1}$$

où $\hat{\mathbf{x}}^b_{\text{exp}}$ est l'espérance mathématique et $\hat{\mathbf{P}}^b$ la matrice de covariance a priori. Dans ce cas, si les opérateurs $\mathcal{M}$ et $\mathcal{H}$ définissant le problème direct (eq. 5.1) sont linéaires :

$$\mathbf{y} = \mathbf{HM}\hat{\mathbf{x}} \tag{6.2}$$

alors, la distribution conjointe $p^b(\hat{\mathbf{x}}^b, \mathbf{y})$ pour $\hat{\mathbf{x}}^b$ et l'équivalent modèle des observations $\mathbf{y}$ est également une distribution gaussienne (voir l'encadré 7, page 63) :

$$p^b(\hat{\mathbf{x}}, \mathbf{y}) = \mathcal{N}_{\hat{\mathbf{x}},\mathbf{y}}\left(\begin{bmatrix} \hat{\mathbf{x}}^b_{\text{exp}} \\ (\mathbf{HM})\hat{\mathbf{x}}^b_{\text{exp}} \end{bmatrix} , \begin{bmatrix} \hat{\mathbf{P}}^b & \hat{\mathbf{P}}^b(\mathbf{HM})^T \\ (\mathbf{HM})\hat{\mathbf{P}}^b & (\mathbf{HM})\hat{\mathbf{P}}^b(\mathbf{HM})^T \end{bmatrix} \right) \tag{6.3}$$

Quand les opérateurs $\mathcal{M}$ et $\mathcal{H}$ sont non-linéaires, il est encore parfois justifiable d'utiliser un modèle gaussien pour la distribution conjointe $p^b(\hat{\mathbf{x}}^b, \mathbf{y})$. Mais dans ce cas, les caractéristiques de la gaussienne (données par l'équation 6.3 dans le cas linéaire) ne peuvent être calculées qu'approximativement, soit à partir des opérateurs linéaires tangents (quand les non-linéarités de $\mathcal{M}$ et $\mathcal{H}$ ne sont pas trop prononcées), soit à partir d'une simulation d'ensemble.

Par ailleurs, si l'incertitude sur les observations $\mathbf{y}^o$ suit également un modèle gaussien :

$$p(\mathbf{y}^o|\mathbf{y}) \;\propto\; \exp\left[-\frac{1}{2}(\mathbf{y}^o - \mathbf{y})^T \mathbf{R}^{-1}(\mathbf{y}^o - \mathbf{y})\right] \tag{6.4}$$

où $\mathbf{R}$ est la matrice de covariance des incertitudes sur les observations, alors l'incertitude a posteriori $p^a(\hat{\mathbf{x}}^b, \mathbf{y})$ donnée par le théorème de Bayes (éq. 5.9) restera une distribution gaussienne, ainsi que la distribution marginale (éq. 5.10) pour le vecteur de contrôle (voir l'encadré 7, page 63). Cette distribution est donc encore entièrement décrite par une espérance mathématique $\hat{\mathbf{x}}^a_{\text{exp}}$ et une matrice de covariance $\hat{\mathbf{P}}^a$ a posteriori, toutes deux calculables de façon générale par des opérations d'algèbre linéaire.

Le propos de ce chapitre est d'examiner par quels algorithmes $\hat{\mathbf{x}}^a_{\text{exp}}$ et $\hat{\mathbf{P}}^a$ peuvent être calculés en pratique pour résoudre les problèmes qui se posent en océanographie. Mais avant d'arriver au cadre particulier des travaux que nous avons menés (décrits à partir de la fin de la section 6.2), je crois qu'il est utile de donner d'abord un aperçu général et synthétique du cadre méthodologique dans lequel nous nous sommes placés. J'essaierai en particulier de faire apparaître le lien qu'il peut y avoir entre une description plus ou moins fine des incertitudes et la complexité algorithmique de la méthode.

6.1 Solution du problème inverse

Dans le cas du modèle gaussien, l'application du théorème de Bayes (éq. 5.9) et de la marginalisation (éq. 5.10) peuvent se faire explicitement, pour obtenir la solution générale du problème inverse, décrite par les paramètres de $p^a(\hat{\mathbf{x}})$:

$$\begin{cases} \hat{\mathbf{x}}^a_{\text{exp}} = \hat{\mathbf{x}}^b_{\text{exp}} + \mathbf{K}(\hat{\mathbf{y}}^o - \mathbf{HM}\hat{\mathbf{x}}^b_{\text{exp}}) \\ \hat{\mathbf{P}}^a = (\mathbf{I} - \mathbf{KHM})\hat{\mathbf{P}}^b \end{cases} \tag{6.5}$$

où $\mathbf{K}$ est l'opérateur de gain :

$$\mathbf{K} = (\mathbf{HM}\hat{\mathbf{P}}^b)^T \left[(\mathbf{HM})\hat{\mathbf{P}}^b(\mathbf{HM})^T + \mathbf{R}\right]^{-1} \tag{6.6}$$

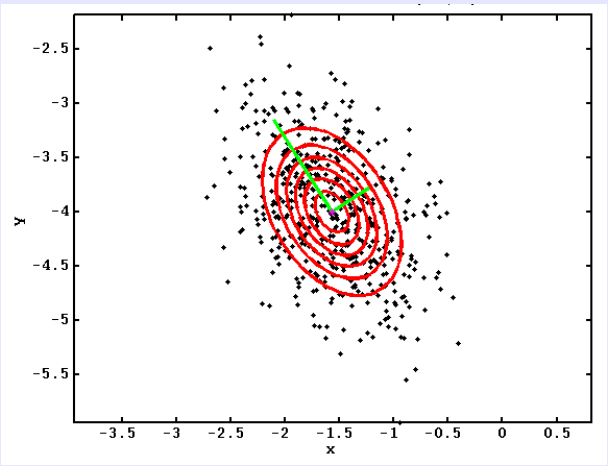
L'espérance mathématique a posteriori du vecteur de contrôle $\hat{\mathbf{x}}^a_{\text{exp}}$ est donc une combinaison linéaire de l'espérance mathématique a priori $\hat{\mathbf{x}}^b_{\text{exp}}$ et des observations $\hat{\mathbf{y}}^o$, pondérée selon la confiance qu'on leur accorde ($\hat{\mathbf{P}}^b$ et $\mathbf{R}$) par le gain $\mathbf{K}$. Et la covariance de l'incertitude a posteriori $\hat{\mathbf{P}}^a$ est réduite par rapport à la covariance de l'incertitude a priori $\hat{\mathbf{P}}^b$: $|\hat{\mathbf{P}}^a| \leq |\hat{\mathbf{P}}^b|$, de même que l'entropie ($S^a \leq S^b$) du fait de l'information apportée par les observations (voir l'encadré 7, page 63).

Complexité algorithmique. Au sujet du coût de l'algorithme, la première chose qu'il faut déduire de la solution du problème inverse donnée par les équations (6.5) à (6.6) est que le coût n'est plus exponentiel, comme pour les méthodes générales décrites en section 5.2, mais polynomial en fonction de la taille du problème. Grâce à l'hypothèse gaussienne, la malédiction des dimensions est donc également vaincue. Cependant, on observe aussi que les coûts dominants de la solution concernent (i) le calcul de la matrice $(\mathbf{HM})\hat{\mathbf{P}}^b(\mathbf{HM})^T$, qui requiert autant d'applications de l'opérateur direct que de variables dans le vecteur de contrôle, et (ii) l'inversion de la matrice $(\mathbf{HM})\hat{\mathbf{P}}^b(\mathbf{HM})^T + \mathbf{R}$, proportionnel au cube du nombre d'observations. Pour les applications océanographiques, ce coût est en général encore prohibitif et il est nécessaire de l'abaisser en réduisant la taille du problème (voir section 6.2).

Filtrage. Dans de nombreuses applications pratiques (y compris océanographiques), toutes les observations ne sont pas disponibles en même temps, mais sont acquises progressivement au fur et à mesure que le temps s'écoule. Il peut donc être nécessaire de construire un algorithme qui mette séquentiellement à jour le vecteur de contrôle chaque fois que de nouvelles informations observationnelles sont disponibles. Une façon générale de résoudre ce problème est de découper l'intervalle de temps complet $[t_0, t_n]$ en sous-intervalles successifs $[t_{i-1}, t_i],\ i = 1, \ldots, n$, et de résoudre un problème inverse pour chaque sous-intervalle l'un après l'autre, dès que les observations pour l'intervalle i sont disponibles. Pour réaliser cet enchaînement, la seule chose supplémentaire dont il faut disposer est une description de l'incertitude sur la condition initiale de l'intervalle i

Encadré 7 : La distribution de probabilité gaussienne

L'objectif de cet encadré est de résumer les propriétés de la distribution gaussienne, permettant de comprendre intuitivement les quelques formules mathématiques utilisées dans ce mémoire. Le premier guide utile à suivre est la correspondance biunivoque qui existe entre une distribution gaussienne (à n dimensions) et un ellipsoïde (à n dimensions) :



Le centre de l'ellipsoïde correspond à l'espérance mathématique du vecteur aléatoire ; ses axes principaux correspondent aux axes principaux (vecteurs propres) de la matrice de covariance ; et les demi-longueurs des ses axes principaux correspondent à l'écart-type selon les directions propres de la matrice de covariance (racine carrée des valeurs propres). La distribution possède donc la même symétrie. Moyenne, mode et médiane sont confondus, au centre de l'ellispoïde. Les lignes de régressions entre différentes variables (ligne de maximum de probabilité de l'une des variables quand l'autre est fixée) sont toujours des droites (lignes vertes sur la figure).

Ce type de symétrie est préservé par toute opération linéaire, qui transformera l'ellipsoïde en un autre ellipsoïde, et la gaussienne en une autre gaussienne : il suffira d'appliquer l'opérateur linéaire à l'espérance mathématique et à la covariance (comme dans l'éq. 6.3). Ce sera donc en particulier le cas pour de "petites perturbations" quand les lois auxquelles elles sont soumises sont dérivables. De même, le caractère gaussien est préservé par le calcul d'une distribution marginale, tout comme la projection de l'ellispoïde dans n'importe quel sous-espace reste une ellispoïde, ou par le calcul d'une distribution conditionnelle, de la même façon que toute section plane d'un ellipsoïde reste un ellipsoïde. Les éqs. 6.5 et 6.6 peuvent ainsi s'interpréter géométriquement comme le calcul des caractéristiques d'une section appropriée d'ellispoïde.

Un deuxième guide utile pour l'intuition provient de la théorie de l'information. Parmi toutes les distributions de probabilité de moyenne et de covariance données, la distribution gaussienne est celle dont l'entropie est maximale. Cela permet de comprendre intuitivement pourquoi c'est elle qui se manifeste quand l'incertitude n'est conditionnée par aucune autre source d'information que des liens de dépendance linéaire entre les variables.

(moyenne et covariance), qui correspond à l'incertitude sur la condition finale de l'intervalle précédent. Lors du calcul de la solution pour chaque sous-intervalle, il faut donc ajouter le calcul de l'incertitude sur la condition finale en plus (ou souvent à la place) du vecteur de contrôle, pour pouvoir enchaîner l'intervalle suivant. Comme toutes les distributions ont été supposées gaussiennes, cela ne pose aucune difficulté de principe supplémentaire [1] Cette solution correspond exactement au filtre de Kalman (1960), introduit initialement sur des bases sensiblement différentes.

D'un point de vue plus pratique, il faut noter que ce découpage en sous-intervalles permet souvent de diminuer le coût de calcul en réduisant le nombre d'observations à prendre en compte simultanément, et aussi parfois en réduisant la taille du vecteur de contrôle (quand on y inclut un vecteur de forçage par exemple). Quand le modèle $\mathcal{M}$ est non-linéaire, l'utilisation de fenêtres temporelles plus courtes peut aussi améliorer la validité du modèle gaussien. En raison des nombreuses non-linéarités des modèles océaniques, un réglage correct des fenêtres temporelles peut donc être d'une importance cruciale pour appliquer raisonnablement le filtre de Kalman à un problème inverse océanographique.

Lissage. Cependant, avec l'algorithme de filtrage décrit précédemment, seule la solution correspondant au dernier sous-intervalle traité a été conditionnée à l'ensemble des observations disponibles ; c'est donc la seule partie de la solution pour laquelle toute l'information observationnelle a été prise en compte. Tel quel, l'algorithme est donc surtout utile pour estimer une condition finale (et l'incertitude associée), et pour initialiser une prévision de l'état futur du système. Pour résoudre entièrement le problème inverse, il est nécessaire de compléter l'algorithme par une mise à jour de la solution des sous-intervalles précédents ($< i$) en utilisant les observations du sous-intervalle courant (i). Cet algorithme complet porte le nom de "lisseur de Kalman", car il permet de lisser la solution aux instants t_i du découpage temporel. On voit bien néanmoins que si, à chaque nouvelle étape i du filtre, il faut aussi revenir sur chacune des étapes précédentes, l'algorithme de lissage peut devenir rapidement beaucoup plus coûteux que l'algorithme de filtrage. C'est pourquoi de nombreuses variantes ont été développées, par exemple en ne revenant à chaque fois que sur un nombre donné de sous-intervalles avant i (lisseur à lag fixe). Pour une discussion plus détaillée de ces algorithmes, voir Cosme et al. (2012).

Incertitudes résiduelles sur le modèle. Conformémemnt à ce qui a été supposé en début du chapitre 5, le vecteur de contrôle a été jusqu'ici supposé contenir toutes les sources d'incertitudes dans le modèle (que ce soit à travers la condition initiale, le forçage, les paramètres, ou un forçage stochastique). Sous l'hypothèse gaussienne, il est néanmoins possible de supposer que le modèle comporte des sources d'incertitudes résiduelles (qu'on ne contrôle pas), à la condition forte que cette incertitude résiduelle sur le modèle puisse être supposée indépendante de l'incertitude sur le vecteur de contrôle $\hat{\mathbf{x}}$ (souvent la condition initiale). Pour la prendre en compte, il suffit alors d'ajouter sa moyenne b (le biais résiduel du modèle sur l'équivalent des observations) à $\mathbf{HMx}^b$ et sa covariance $\mathbf{HQH}^T$ (la covariance de l'erreur résiduelle du modèle sur l'équivalent des observations) à $(\mathbf{HM})\hat{\mathbf{P}}^b(\mathbf{HM})^T$ dans l'éq. (6.3), et ensuite d'appliquer les éqs (6.5) et (6.6) de façon parfaitement identique. Il est néanmoins important de noter que dans un problème de filtrage, quand on doit aussi estimer la condition finale $\mathbf{x}^f$ de chaque cycle d'assimilation, cette approche implique de connaître la covariance de l'erreur résiduelle

1. Plus précisément, si on effectue une mis à jour à chaque fois qu'une nouvelle observation est disponible, et si on estime à chaque fois que la condition finale, le filtre de Kalman (à l'inverse du lisseur) n'exige pour être optimal (au sens bayésien) que la gaussianité des distributions de probabilité 3D à chaque instant, et pas la gaussianité de la distribution 4D sur tout l'intervalle.

du modèle pour le vecteur $[\mathbf{y},\, \mathbf{x}^f]$, ce qui est rarement facile à paramétrer, surtout quand $\mathbf{y}$ et $\mathbf{x}^f$ ne sont pas simultanés. C'est pourquoi il me semble préférable d'éviter autant que possible de recourir à cette approche et de s'efforcer de toujours inclure explicitement les diverses sources d'incertitudes sur le modèle dans le vecteur de contrôle $\hat{\mathbf{x}}$.

6.2 Réduction d'ordre

La façon la plus directe de réduire la complexité algorithmique de la solution du problème inverse donnée par les équations (6.5) et (6.6) est de réduire la taille du problème, en supposant que l'incertitude a priori sur le vecteur de contrôle (de taille n) se cantonne dans un sous-espace de dimension faible ($r \ll n$). Dans ce cas, il est préférable de décrire la matrice de covariance d'erreur d'ébauche $\hat{\mathbf{P}}^b$ $(n \times n)$ par une racine carrée $\hat{\mathbf{S}}^b$ $(n \times r)$:

$$\hat{\mathbf{P}}^b = \hat{\mathbf{S}}^b \hat{\mathbf{S}}^{bT} \tag{6.7}$$

Cette description d'ordre réduit peut par exemple correspondre aux modes principaux dominants de $\hat{\mathbf{P}}^b$, ou bien à une description de $p^b(\hat{\mathbf{x}})$ par un échantillon de taille $r+1 \ll n$ (de moyenne $\hat{\mathbf{x}}^b_{\text{exp}}$ et de covariance $\hat{\mathbf{P}}^b$).

Dans ce cas, le nombre d'application de l'opérateur direct ($\mathbf{HM}$) dans les équations (6.5) et (6.6) est réduit de $n+1$ à $r+1$, pour calculer $\mathbf{HM}\hat{\mathbf{x}}^b_{\text{exp}}$ et $\mathbf{HM}\hat{\mathbf{S}}^b$, ce qui réduit déjà de façon conséquente la complexité de la solution. Quand cette opération est réalisée par une prévision d'ensemble [appliquée à un échantillon de taille $r+1$ de $p^b(\hat{\mathbf{x}})$], $\mathbf{HM}\hat{\mathbf{x}}^b_{\text{exp}}$ et $\mathbf{HM}\hat{\mathbf{S}}^b$ doivent simplement être remplacés par la moyenne de la prévision d'ensemble et une racine carrée de la matrice de covariance (c'est-à-dire simplement la matrice des anomalies par rapport à la moyenne, divisée par $\sqrt{r}$). Par ailleurs, l'application de l'opérateur de gain peut également être simplifiée en utilisant l'un des algorithmes suivants (voir références dans le tableau 6.1) :

Algorithme transformé. Tout d'abord, si la dimension de l'espace de contrôle devient plus faible que le nombre d'observations ($r \ll m$), il est en général moins coûteux de transformer l'expression du gain $\mathbf{K}$ (éq. 6.6) pour réaliser l'inversion dans l'espace de contrôle (réduit) plutôt que dans l'espace des observations. Une façon simple de l'obtenir (décrite dans Brankart et al., 2010, en annexe B) est d'appliquer la transformation linéaire suivante pour passer de $\hat{\mathbf{x}}$ et $\hat{\mathbf{y}}^o$ aux vecteurs réduits $\hat{\boldsymbol{\xi}}$ et $\boldsymbol{\eta}^o$:

$$\hat{\mathbf{x}} = \hat{\mathbf{x}}^b_{\text{exp}} + \hat{\mathbf{S}}^b \hat{\boldsymbol{\xi}} \quad \text{et} \quad \boldsymbol{\eta}^o = (\mathbf{HM}\hat{\mathbf{S}}^b)^T \mathbf{R}^{-1} (\hat{\mathbf{y}}^o - \mathbf{HM}\hat{\mathbf{x}}^b_{\text{exp}}) \tag{6.8}$$

Ces deux vecteurs transformés sont dans l'espace réduit (de dimension r), et leur distribution de probabilité se simplifie en :

$$p^b(\hat{\boldsymbol{\xi}}) = \mathcal{N}_{\hat{\boldsymbol{\xi}}}(0, \mathbf{I}) \qquad \text{et} \qquad p(\boldsymbol{\eta}^o | \hat{\boldsymbol{\xi}}) = \mathcal{N}_{\boldsymbol{\eta}^o}(\boldsymbol{\Gamma}\hat{\boldsymbol{\xi}}, \boldsymbol{\Gamma}) \tag{6.9}$$

avec

$$\boldsymbol{\Gamma} = (\mathbf{HM}\hat{\mathbf{S}}^b)^T \mathbf{R}^{-1} (\mathbf{HM}\hat{\mathbf{S}}^b) \tag{6.10}$$

de sorte que la distribution a posteriori pour le vecteur de contrôle réduit $\hat{\boldsymbol{\xi}}$ s'obtient directement par le théorème de Bayes :

$$p^a(\hat{\boldsymbol{\xi}}) = \mathcal{N}_{\hat{\boldsymbol{\xi}}}(\hat{\boldsymbol{\xi}}^a_{\text{exp}}, \hat{\boldsymbol{\Pi}}^a) \qquad \text{avec} \qquad \hat{\boldsymbol{\xi}}^a_{\text{exp}} = [\mathbf{I}+\boldsymbol{\Gamma}]^{-1} \boldsymbol{\eta}^o \qquad \text{et} \qquad \hat{\boldsymbol{\Pi}}^a = [\mathbf{I}+\boldsymbol{\Gamma}]^{-1} \tag{6.11}$$

De là, il est facile de retourner dans l'espace de contrôle original en inversant la transformation 6.8 :

$$\hat{\mathbf{x}}^a_{\text{exp}} = \hat{\mathbf{x}}^b_{\text{exp}} + \hat{\mathbf{S}}^b \hat{\boldsymbol{\xi}}^a_{\text{exp}} \qquad \text{et} \qquad \hat{\mathbf{P}}^a = \hat{\mathbf{S}}^a \hat{\mathbf{S}}^{aT} \qquad \text{avec} \qquad \hat{\mathbf{S}}^a = \hat{\mathbf{S}}^b \hat{\mathbf{\Pi}}^{a\,1/2} \tag{6.12}$$

Algorithme en base propre. Une variante de l'algorithme précédent consiste à travailler dans la base propre de la matrice $\mathbf{\Gamma} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ (où la matrice diagonale $\mathbf{\Lambda}$ contient les valeurs propres de $\mathbf{\Gamma}$ et la matrice unitaire $\mathbf{U}$ contient les vecteurs propres normalisés de $\mathbf{\Gamma}$). Pour cela, il suffit d'appliquer la transformation $\mathbf{U}^T$ à $\hat{\boldsymbol{\xi}}$ et $\boldsymbol{\eta}^o$: $\hat{\boldsymbol{\xi}}' = \mathbf{U}^T\hat{\boldsymbol{\xi}}$ et $\boldsymbol{\eta}^{o\prime} = \mathbf{U}^T\boldsymbol{\eta}^o$. Dans ce cas, les distributions de probabilité de l'équation (6.9) se simplifient en $p^b(\hat{\boldsymbol{\xi}}') = \mathcal{N}_{\hat{\boldsymbol{\xi}}'}(0, \mathbf{I})$ et $p(\boldsymbol{\eta}^{o\prime}|\hat{\boldsymbol{\xi}}') = \mathcal{N}_{\boldsymbol{\eta}^{o\prime}}(\mathbf{\Lambda}\hat{\boldsymbol{\xi}}', \mathbf{\Lambda})$, où toutes les matrices de covariances sont diagonales, de sorte qu'aucune inversion de matrice n'est plus nécessaire après le passage en base propre. L'intérêt de cet algorithme est qu'il permet d'éviter de "trop mélanger" les modes d'erreur (ou les membres de l'ensemble) lors de la mise à jour de la racine carrée de la matrice de covariance [2] :

$$\hat{\mathbf{S}}^a = \hat{\mathbf{S}}^b \mathbf{U} \mathbf{\Lambda}^{1/2} \mathbf{U}^T \tag{6.13}$$

Les modes d'erreurs (ou les anomalies d'ensemble) sont d'abord transformés en base propre par le facteur $\mathbf{U}$, puis réduit par le facteur $\mathbf{\Lambda}^{1/2}$ (pour tenir compte du gain d'information obtenu des observations), et enfin retransformé en arrière par le facteur $\mathbf{U}^T$. Cette propriété est indispensable pour rendre l'algorithme transformé compatible avec les méthodes de localisation de la matrice de covariance (décrites en section 6.3) sans recourir à l'algorithme d'ensemble décrit ci-après (voir aussi Brankart et al., 2010, en annexe B pour plus d'explication).

Algorithme d'ensemble. Quand la distribution de probabilité a priori $p^b(\hat{\mathbf{x}}, \mathbf{y})$, toujours supposée gaussienne, est décrite par un ensemble, c'est-à-dire par un échantillon aléatoire, plutôt que par l'espérance mathématique $\hat{\mathbf{x}}^b_{\text{exp}}$ et une racine carrée quelconque $\hat{\mathbf{S}}^b$ de la matrice de covariance, il est possible de résoudre le problème inverse en appliquant simplement la formule 6.5 (première équation) à chacun des membres $\mathbf{x}^b_i$ de l'ensemble :

$$\hat{\mathbf{x}}^a_i = \hat{\mathbf{x}}^b_i + \mathbf{K}(\hat{\mathbf{y}}^o_i - \mathbf{H}\mathbf{M}\hat{\mathbf{x}}^b_i), \quad i = 1, \ldots, r+1 \tag{6.14}$$

pour obtenir un ensemble $\mathbf{x}^a_i$ décrivant la distribution de probabilité a posteriori $p^a(\hat{\mathbf{x}})$. Il est possible de vérifier que l'ensemble $\mathbf{x}^a_i$ aura non seulement la bonne moyenne (Evensen, 1994), mais aussi la bonne covariance (c'est-à-dire celle donnée par l'éq. 6.5, seconde équation), à condition toutefois que les observations $\hat{\mathbf{y}}^o_i = \hat{\mathbf{y}}^o + \boldsymbol{\varepsilon}_i$ soient perturbées par un vecteur aléatoire gaussien $\boldsymbol{\varepsilon}_i$ de moyenne nulle et de covariance $\mathbf{R}$ (Burgers et al., 1998). Le gain $\mathbf{K}$ quant à lui est toujours calculé par l'équation (6.6), mais pourrait aussi être calculé par l'algorithme transformé à travers les équations (6.9) à (6.12).

Comparaison de ces algorithmes. Chacun de ces algorithmes a été introduit pour obtenir différentes variantes du filtre de Kalman original. Le tableau 6.1 récapitule les variantes du filtre de Kalman associées à chacun de ces algorithmes, ainsi que la première apparition de l'algorithme dans la littérature. A titre indicatif, le tableau mentionne aussi une publication plus récente dans laquelle d'autres références utiles peuvent être

2. En utilisant une racine carrée appropriée de $\hat{\mathbf{\Pi}}^a$ dans l'équation (6.12).

Algorithme	Variante du filtre	Références
classique	Filtre de Kalman	Kalman (1960); Gelb (1974)
transformé	Filtre SEEK	Pham et al. (1998); Brasseur and Verron (2006)
	Filtre SEIK	Hoteit et al. (2002); Nerger et al. (2005)
en base propre	Filtre racine carrée	Wang et al. (1992)
	Filtre RRSQRT	Verlaan and Heemink (1997); Gillijns et al. (2006)
	Filtre ESSE	Lermusiaux and Robinson (1999)
	ETKF	Bishop et al. (2001)
	Filtre SEEK	Brankart et al. (2003, 2011)
d'ensemble	EnKF	Evensen (1994); Burgers et al. (1998)

TABLE 6.1 – Références pour les algorithmes utilisés dans les différentes variantes du filtre de Kalman. Les acronymes EnKF, SEEK, SEIK, RRSQRT, ESSE and ETKF signifient respectivement 'Ensemble Kalman filter', 'Singular Evolutive Extended Kalman filter', 'Singular Evolutive Interpolated Kalman filter', 'Reduced Rank Square Root Kalman filter', 'Ensemble Subspace Statistical Estimation' and 'Ensemble Transform Kalman Filter'. Dans les algorihtme de Wang et al. (1992), Verlaan and Heemink (1997) et Lermusiaux and Robinson (1999), les matrices $\mathbf{U}$ and $\mathbf{\Lambda}$ sont directement calculées par une décomposition SVD de la matrice $(\mathbf{HM\hat{S}}^b)^T\mathbf{R}^{-1/2}$ (ce qui évite d'évaluer $\mathbf{\Gamma}$), mais cela est généralement moins efficace pour des matrices de rang très faible ($r \ll y$).

trouvées. La différence principale entre ces algorithmes réside dans leur complexité algorithmique (le nombre d'opérations à effectuer en fonction de la taille du problème) et donc dans le genre d'approximations auxquelles ils conduisent le plus naturellement. La complexité algorithmique de l'algorithme classique (comportement asymptotique dominant lorsque n et p sont grands), décrit par les équations (6.4) et (6.6), peut s'écrire :

$$C_C \sim (n+1)D + n^2p + np^2 + \alpha p^3 \tag{6.15}$$

où D est la complexité algorithmique du problème direct ; n, la taille du vecteur de contrôle ; p, la taille du vecteur d'observations ; et α, un facteur d'ordre 1 dépendant du détail de l'organisation des opérations d'algèbre linéaire. Pour l'algorithme transformé décrit par les éqs. (6.9) à (6.12) (et sa variante en base propre), la complexité algorithmique (comportement asymptotique dominant lorsque n et p sont grands) est :

$$C_T \sim (r+1)D + nr^2 + pr^2 + \alpha r^3 \tag{6.16}$$

où r est le rang de la matrice de covariance d'erreur d'ébauche (ou bien, $r+1$ la taille de l'ensemble). Dans cette expression, il a été supposé que la nécessité d'inverser la matrice $\mathbf{R}$ explicitement a été évitée (voir section 6.4). Quant à l'algorithme d'ensemble (éq. 6.14), sa complexité algorithmique (comportement asymptotique dominant) est :

$$C_E \sim (r+1)D + rnp + 2rp^2 + \alpha p^3 \tag{6.17}$$

quand l'algorithme original (eq. 6.6) est utilisé pour le calcul du gain $\mathbf{K}$ (inversion dans l'espace des observations), ou bien :

$$C_{ET} \sim (r+1)D + nr^2 + 2pr^2 + \alpha r^3 \tag{6.18}$$

quand c'est l'algorithme transformé qui est utilisé. Entre les trois dernières expressions (éqs. 6.16 à 6.18), la principale différence de complexité provient de l'espace dans lequel l'inversion de matrice est effectuée : dernier terme en p^3 quand l'inversion se fait sans l'espace des observations, ou en r^3 quand elle se fait dans l'espace de contrôle réduit. Dans le premier cas, on tendra donc à privilégier les approximations qui réduisent la quantité d'observations utilisées simultanément, alors que, dans le second cas, la linéarité en p

de la complexité algorithmique permettra de traiter plus d'observations simultanément, mais privilégiera le choix d'une matrice $\mathbf{R}$ facilement invertible (voir section 6.4). D'autre part, on voit que la seule différence entre les complexités 6.16 et 6.18 provient du doublement du troisième terme. Ceci s'explique par la nécessité, dans l'algorithme d'ensemble, d'appliquer le gain de Kalman à un ensemble d'observations différentes (c'est-à-dire de projeter l'ensemble des $\boldsymbol{\varepsilon}_i$ sur la racine carrée de la matrice de covariance, via la transformation 6.9). Ce coût supplémentaire (certes modeste) peut être évité avec l'algorithme en base propre qui peut tout aussi bien être appliqué à la mise à jour d'un ensemble (y compris avec localisation de la matrice de covariance, voir section 6.3), et qui ne requiert pas de perturber les observations.

Filtre SEEK. L'algorithme transformé (de complexité C_T) forme la base algorithmique du filtre SEEK (Singular Evolutive Extended Kalman filter) que nous utilisons dans l'équipe depuis une quinzaine d'années (depuis son développement par Pham et al., 1998). A l'origine cependant, le filtre SEEK n'incluait dans le vecteur de contrôle que la condition initiale $\hat{\mathbf{x}} = \mathbf{x^0}$ et l'incertitude a priori asscociée (paramétrée par $\hat{\mathbf{S}}^b$) était décrite par les principales EOFs (Empirical Orthogonal Functions) de la variabilité naturelle du modèle ; par ailleurs, la matrice de covariance d'erreur d'observation $\mathbf{R}$ était supposée diagonale (afin d'éviter un coût proportionnel à p^3 pour l'inverser). La méthode était également orientée vers du filtrage, c'est-à-dire destinée à n'estimer séquentiellement que la condition finale de chaque cycle d'assimilation.

Depuis lors, nous avons toujours veillé à préserver la base algorithmique originale (donnée succinctement par les éqs. 6.9 à 6.12, de complexité algorithmique C_T donnée par l'éq. 6.16), mais nous avons progressivement élargi le type de modélisation des incrtitudes qu'elle permettait (d'abord gaussienne, puis non-gaussienne, voir chapite suivant). Nous nous sommes d'abord orientés vers la formulation en base propre [3] (Brankart et al., 2003), pour permettre la localisation des covariances (voir section 6.3 ci-après). Le même noyau algorithmique a ensuite été appliqué en remplaçant les EOFs dans $\hat{\mathbf{S}}^b$ par un échantillon aléatoire (Broquet et al., 2008) initialisant une prévision d'ensemble. Cela a permis d'élargir peu à peu le vecteur de contrôle à la fonction de forçage océanique (Skachko et al., 2009; Skandrani et al., 2009; Meinvielle et al., 2013), et à des paramètres du modèle (Doron et al., 2011; Melet et al., 2012; Doron et al., 2013). Entretemps, la méthode a également été développée pour autoriser, sans modification essentielle de la complexité algorithmique, l'utilisation de matrices de covariance d'erreur $\mathbf{R}$ non diagonale (voir section 6.4), ainsi que l'inclusion de certains paramètres clés des matrices de covariance $\hat{\mathbf{P}}^b$ et $\mathbf{R}$ dans le vecteur de contrôle (voir section 6.5). Par ailleurs, un lisseur SEEK, généralisant le filtre sans coût numérique significatif a également été proposé par Cosme et al. (2010, 2012).

Grâce à ces développements, la méthode a pu être progressivement appliquée à des problèmes inverses océanographiques de natures très diverses : l'assimilation d'observations altimétriques dans un modèle du Pacifique Tropical (Verron et al., 1999; Parent et al., 2003; Castruccio et al., 2006, 2008), de l'Atlantique Nord (Testut et al., 2003; Brankart et al., 2003; Birol et al., 2004, 2005), de l'Atlantique Sud (Penduff et al., 2003), ou de l'Atlantique tropical (Freychet et al., 2012) ; l'exploration de scénarios satellitaires d'observations altimétriques dans l'Atlantique tropical (Ubelmann et al., 2009, 2012) ou le Golfe du Lion (Duchez et al., 2012) ; l'assimilation de données de température et de salinité dans un modèle global à basse résolution pour contrôler le forçage atmosphérique (Skachko et al., 2009; Skandrani et al., 2009; Meinvielle et al., 2013) ; l'étude de l'impact de l'assimilation de données physiques sur un modèle couplé biogéochimique (Berline

3. Dans l'algorithme SEEK original, la racine carrée $\hat{\mathbf{\Pi}}^{a\,1/2}$ dans l'équation (6.12) était calculée par une décomposition de Cholesky.

et al., 2007) ; ou bien l'assimilation de données de couleur de l'eau dans un modèle biogéochimique de l'Atlantique Nord (Carmillet et al., 2001; Ourmières et al., 2009; Fontana et al., 2013). La méthode et ses applications a également fait l'objet de 11 thèses à caractère applicatif (Parent, 2000; Testut, 2000; Magri, 2002; Debost, 2004; Berline, 2006; Broquet, 2006; Castruccio, 2006; Ubelmann, 2009; Duchez, 2011; Freychet, 2012; Meinvielle, 2012). Ce sont aussi ces développements, combinés à l'efficacité de l'algorithme SEEK original pour des problèmes de grande taille, qui ont permis que la technologie basée sur le filtre SEEK ait pu être transférée vers le centre français d'océanographie opérationnelle MERCATOR, et servir de base au système de prévision et de réanalyse actuel (Ferry et al., 2010).

6.3 Localisation des covariances

Un premier verrou qui rend difficile l'application de l'algorithme transformé, avec une forte réduction d'ordre ($r \ll n$), à un problème réaliste est l'impossibilité de décrire les faibles corrélations avec une précision suffisante. Dans les applications atmosphériques ou océaniques réalistes du filtre de Kalman, les dimensions horizontales du système sont en effet souvent considérablement plus grandes que l'échelle de corrélation des incertitudes, de sorte que la plupart des éléments de la matrice de corrélation (qui couplent des variables distantes) sont proches de zéro, et le rang (ou la taille de l'ensemble) requis pour les décrire avec précision dépasse toute taille numériquement raisonnable (voir par exemple Brankart et al., 2011, en annexe B). La surestimation de ces corrélations (en valeur absolue) par une trop forte réduction d'ordre (par exemple avec un ensemble de trop petite taille) conduit en général à surestimer la quantité d'information que les observations contiennent à propos du système (tout en filtrant exagérément leur contenu réel d'information), et donc à sous-estimer fortement l'incertitude a posteriori (effondrement de la matrice de covariance d'erreur).

Pour contourner cette difficulté, la méthode classique (Houtekamer and Mitchell, 1998; Ott et al., 2004; Hunt et al., 2007) est de compléter la représentation de rang réduit par l'hypothèse que les corrélations à grande distance sont négligeables, et de les forcer à zéro dans la matrice de covariance d'erreur. Pour ce faire, la méthode la mieux étayée (Houtekamer and Mitchell, 1998) consiste à effectuer le produit de Schur [4] de la matrice de covariance de rang réduit par une matrice de corrélation à support local $\mathbf{C}$ (ce qui garantit en particulier que le résultat garde les propriétés d'une matrice de covariance) : $\mathbf{P}^b_\ell = \mathbf{P}^b \circ \mathbf{C}$ (où $\mathbf{P}^b_\ell$ représente la matrice de covariance conjointe du vecteur de contrôle et des observations apparaissant dans l'éq. 6.3). Malheureusement, avec cette méthode, la forme racine carrée de la matrice de covariance n'est plus directement disponible (Bishop and Hodyss, 2009), et pour que l'algorithme transformé défini par les éqs. (6.8) à (6.12) reste applicable (au moins successivement pour toute sous-région locale i du système complet), il est nécessaire de développer une méthode permettant d'obtenir une racine carrée $\mathbf{S}^b_{\ell,i}$ pour tout bloc local $\mathbf{P}^b_{\ell,i}$ du produit de Schur $\mathbf{P}^b_\ell$:

$$\mathbf{P}^b_{\ell,i} = \mathbf{P}^b_i \circ \mathbf{C}_i = (\mathbf{S}^b_i \mathbf{S}^{bT}_i) \circ \mathbf{C}_i = \mathbf{S}^b_{\ell,i} {\mathbf{S}^b_{\ell,i}}^T \tag{6.19}$$

Méthode. La méthode que nous avons développée pour cela (voir Brankart et al., 2011, en annexe B) consiste à rechercher une factorisation locale approximative du produit de Schur : $\mathbf{P}^b_{\ell,i} \simeq \mathbf{S}^b_{\ell,i} {\mathbf{S}^b_{\ell,i}}^T$ qui garde le même rang que la matrice $\mathbf{P}^b_{\ell,i}$ originale

4. Le produit de Schur (ou de Hadamard) est une opération qui, pour deux matrices de mêmes dimensions, associe une autre matrice, de même dimension, où chaque coefficient est le produit terme à terme des deux matrices.

(c'est-à-dire que la matrice $\mathbf{S}^b_{\ell,i}$ ne comporte que r colonnes, tout comme $\mathbf{S}^b$ ou $\mathbf{S}^b_i$), en profitant du fait qu'il suffit que chaque application locale de l'algorithme transformé produise le même résultat au point central de chaque sous-région i. (Le point central est en effet la seule partie de chaque solution locale qui est utilisée pour reconstruire la solution globale pour le système complet.) Une racine carrée exacte (c'est-à-dire satisfaisant exactement l'éq. 6.19) serait facilement calculable (Bishop and Hodyss, 2009) à partir de la racine carrée de $\mathbf{P}^b_i$ et d'une racine carrée de $\mathbf{C}_i$, mais le nombre de colonnes de la racine carrée ainsi obtenue est égal au nombre de colonnes (r) de la racine carrée de $\mathbf{P}^b_i$ multiplié par le rang de $\mathbf{C}_i$. C'est ce qui a conduit déjà Bishop and Hodyss (2009) à utiliser une matrice $\mathbf{C}_i$ de rang modéré ($r_C = 126$ dans leur étude) pour pouvoir utiliser la méthode de localisation classique (éq. 6.19) avec l'algorithme transformé (en base propre). Mais malgré cela, on perd une grande part de l'intérêt de l'algorithme transformé, puisque la valeur de r dans l'expression de sa complexité algorithmique (éq. 6.16) doit être alors mulitpliée par r_C. La méthode que nous avons développée consiste donc à réduire r_C à 1, en remarquant qu'un seul mode de $\mathbf{C}_i$ est suffisant pour conserver la totalité de la covariance de $\mathbf{P}^b_{\ell,i}$ au point central de la sous-région i. Dans ce cas, le résultat produit par l'algorithme transformé en ce point central sera le même qu'avec la racine carrée exacte (éq. 6.19), à condition toutefois que la partie manquante de la covariance (par rapport au produit de Schur complet de l'éq. 6.19) soit ajoutée à la matrice de covariance d'erreur d'observation $\mathbf{R}_i$ utilisée pour la sous-région i (après projection dans l'espace des observations par l'opérateur $\mathbf{H}_i$). Ceci ne peut malheureusement se faire qu'approximativement, et nous avons proposé de prendre en compte cet effet en amplifiant simplement la variance de l'erreur d'observation avec la distance par rapport au point central de chaque sous-région.

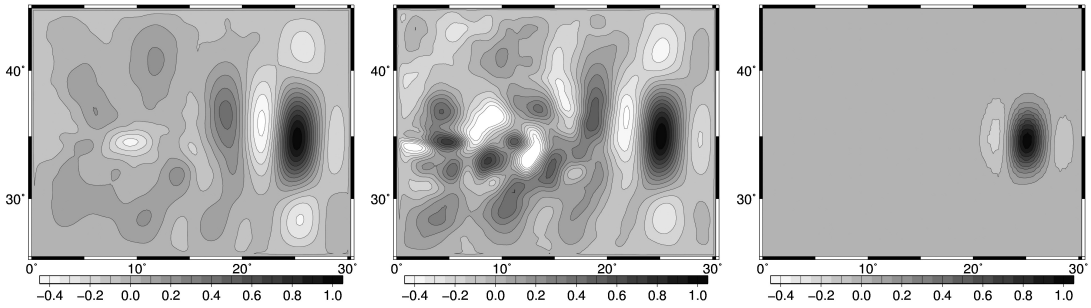


FIGURE 6.1 – Correction résultant d'une observation altimétrique parfaite située à 25°E 34°N : pour un ensemble de taille 5000 (à gauche), pour un ensemble de taille 200 (au centre), et pour un ensemble de taille 200 avec localisation des covariances (à doite).

Application. En guise d'application, imaginons par exemple que nous voulions décrire la covariance de l'incertitude sur le champ de l'élévation du niveau de la mer issu de la configuration SEABASS (présentée en section 1.3) par un ensemble de taille limitée. Afin de faire apparaître clairement l'effet d'une taille d'ensemble insuffisante, la façon la plus simple de procéder est de calculer la variation de moyenne de l'ensemble ($\hat{\mathbf{x}}^a_{\text{exp}} - \hat{\mathbf{x}}^b_{\text{exp}}$) qui résulte d'une seule observation altimétrique parfaite (située à 25°E 34°N). L'extrapolation de l'information est en effet la situation pour laquelle l'effet de la surestimation des corrélations à grande distance est le plus pénalisant. La figure 6.1 montre ainsi le résultat obtenu pour un ensemble de taille 5000 (à gauche) et pour un ensemble de taille 200 (au centre). Sur cette figure, on voit immédiatement que lorque la taille de l'ensemble est insuffisante (200 membres), l'algorithme surestime la quantité d'information qu'il est possible de tirer de l'observation, et produit du signal dans une région

(l'ouest du bassin) qui ne peut pas être reliée à la quantité observée (corrélation quasi-nulle, correctement perçue par l'ensemble de taille 5000). Pour remédier à ce problème, tout en gardant une taille d'ensemble modérée (ici 200 membres), un compromis est donc de compléter la description d'ensemble par une localisation de la covariance (via une matrice de corrélation $\mathbf{C}$ conformément à l'éq. 6.19). Le résultat (fig. 6.1, à droite) reste le même au voisinage de l'observation, mais tend progressivement vers 0 lorsque la distance à l'observation augmente. On voit bien aussi qu'il s'agit là d'une approximation plutôt fruste de la solution exacte, mais on ne saurait sous-estimer l'importance de solutions qui permettent de se contenter d'un ensemble de taille raisonnable. Au vu de la complexité algorithmique (éq. 6.16), réduire r permet en effet d'augmenter la taille du vecteur d'observation (p), d'augmenter la taille du vecteur de contrôle (n), ou bien d'appliquer la méthode à un problème direct de coût numérique (D) plus élevé.

6.4 Modélisation des incertitudes sur les observations

Un deuxième verrou qui rend difficile l'application de l'algorithme transformé à un problème réaliste est l'obligation de pouvoir inverser la matrice de covariance des erreurs d'observation $\mathbf{R}$ à un coût négligeable, c'est-à-dire en pratique de supposer qu'elle est diagonale. Cela est nécessaire pour éviter d'introduire un terme en p^3 dans la complexité algorithmique C_T (éq. 6.16), et préserver son efficacité par rapport aux algorithmes qui réalisent l'inversion dans l'espace des observations (par l'éq. 6.6). Dans les applications atmosphériques et océaniques, la présence d'un tel coût en p^3 oblige en effet toujours à réduire le nombre d'observations, soit en les agrégeant en superobservations (par une moyenne sur une petite région), soit même en écartant les observations les moins utiles ou les plus redondantes (data thinning). Or, nous verrons plus loin que des observations très voisines, aux erreurs très corrélées, et donc apparemment très redondantes, peuvent contenir une information très importante sur la structure spatiale du champ (par exemple le gradient) qu'il peut être très dommageable de négliger, ou de mal prendre en compte par une matrice de covariance inappropriée.

Méthode. Pour remédier à cela, la méthode que nous avons proposée (voir Brankart et al., 2009, en annexe B) consiste à transformer le vecteur d'observation par un opérateur linéaire $\mathbf{T}$: $\hat{\mathbf{y}}^{o+} = \mathbf{T}\hat{\mathbf{y}}^{o}$, $\mathbf{H}^{+} = \mathbf{TH}$, et à supposer diagonale la matrice de covariance d'erreur $\mathbf{R}^{+}$ pour les observations transformées. On vérifie facilement que cela revient à supposer que la matrice de covariance d'erreur $\mathbf{R}$ pour les observations originales est donnée par :

$$\mathbf{R}^{-1} = \mathbf{T}^{T}\mathbf{R}^{+^{-1}}\mathbf{T} \tag{6.20}$$

Une solution immédiate serait de choisir pour $\mathbf{R}^{+}$, la matrice des valeurs propres de $\mathbf{R}$, et pour $\mathbf{T}$, la matrice unitaire des vecteurs propres normalisés correspondants. Mais ceci n'est évidemment pas la solution recherchée, car la complexité algorithmique du calcul des valeurs propres et des vecteurs propres est de nouveau proportionnelle à p^3. (De plus, le coût de l'application de $\mathbf{T}$ serait également prohibitif.) Non, l'astuce consiste à choisir une transformation $\mathbf{T}$ (de structure simple, peu coûteuse à appliquer) qui augmente la taille du vecteur d'observation : $p^{+} > p$. De cette manière, on peut simuler une large palette de structure de corrélation dans $\mathbf{R}$, tout en préservant l'efficacité de l'algorithme transformé : il faudra simplement remplacer p par p^{+} dans l'expression (éq. 6.16) de la complexité algorithmique. Et, comme la complexité algorithmique de l'algorithme transformé est *linéaire en* p, le coût de la transformation pourra rester raisonnable.

L'exemple le plus simple d'opérateur $\mathbf{T}$ non-carré simulant une matrice de covariance $\mathbf{R}$ non-diagonale est de compléter les observations originales $\hat{\mathbf{y}}^{o}$ par leur gradient.

Il est assez facile de montrer (voir Brankart et al., 2009, en annexe B pour le détail) que cela revient à simuler une structure de corrélation de forme calculable [par exemple décroissant exponentiellement avec la distance ρ entre deux observations dans le cas unidimensionnel : $\exp(-\rho/\ell)$]. De plus, la longueur de corrélation ℓ est égale au rapport des écarts-types d'erreur d'observation sur les observations originales (σ_0) et sur leur gradient (σ_1) : $\ell = \sigma_0/\sigma_1$. Cette première étape peut ensuite être généralisée en constatant qu'en incluant dans $\mathbf{T}$ des opérateurs différentiels d'ordre de plus en plus grand, on pourra décrire avec une précision de plus en plus fine n'importe quel modèle de corrélation. Néanmoins, en augmentant ainsi la taille p^+ du vecteur d'observation, ce gain en précision finira pas aboutir à un coût numérique prohibitif. Avec cette méthode, il est donc possible de régler l'opérateur $\mathbf{T}$ de façon à obtenir le meilleur compromis entre une représentation fine de la structure de corrélation des erreurs d'observation (souvent en grande partie inconnue) et l'efficacité numérique de l'algorithme.

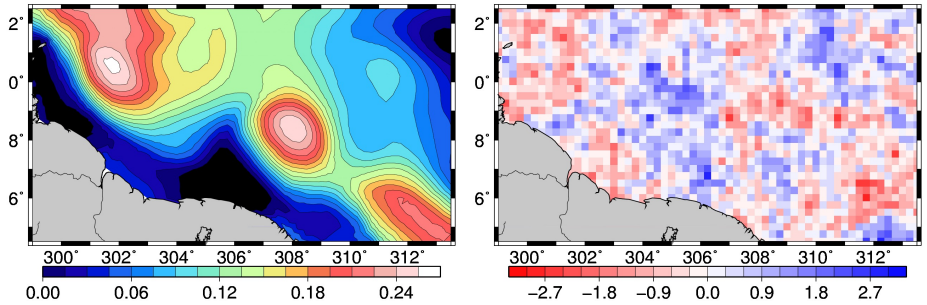


FIGURE 6.2 – Exemple d'observations altimétriques synthétiques (simulées par un modèle) dans la région des tourbillons du Brésil (à gauche), et du bruit que l'on ajoute pour simuler l'erreur d'observation (à droite).

Application. En guise d'application, supposons par exemple que nous voulions estimer le champ de vitesse dans la région des tourbillons du Brésil à partir d'observations altimétriques (fig 6.2, à gauche) perturbées par une erreur d'observations d'écart-type égal à 4 cm, et de longueur de corrélation égale à 5 points de grilles (fig. 6.2, à droite). Par souci de simplicité, la distribution de probabilité d'erreur d'ébauche (gaussienne) est décrite par la variabilité d'une simulation du modèle (échantillonnée tous les 6 jours sur une période de 5 ans). De plus, la matrice de covariance est localisée comme expliqué en section 6.3 par une fonction de corrélation $\gamma(r) = \exp(-r^2/d^2)$ avec $d = 200$ km.

La figure 6.3 (en haut, à gauche) montre l'écart-type de l'incertitude a posteriori sur la vitesse (en haut), tel qu'estimé par l'algorithme transformé quand la matrice de covariance d'erreur d'observation $\mathbf{R}$ est supposée diagonale, c'est-à-dire quand la corrélation spatiale des erreur d'observations est négligée. Cette estimation est environ 3 fois plus faible que l'écart-type de l'erreur réelle (en haut, à droite). En raison de la paramétrisation incorrecte de $\mathbf{R}$, le schéma croit les observations plus précises qu'elles ne sont en réalité, et surestime la quantité d'information réellement présente dans le système. Par ailleurs, la figure 6.3 (en bas) montre le même résultat quand le gradient des observations originales a été ajouté au vecteur d'observation, avec un poids relatif approprié pour simuler la corrélation des erreurs d'observation. Non seulement l'écart-type estimé (fig. 6.3, en bas, à gauche) est cette fois presque parfaitement cohérent avec l'écart-type de l'erreur réelle (fig. 6.3, en bas, à droite), mais l'erreur réelle est cette fois nettement inférieure à celle produite par l'estimation précédente (quand la corrélation des erreurs d'observation était négligée). Ceci est bien sûr la conséquence directe du fait que la modélisation des erreurs d'observation est cette fois cohérente avec l'erreur réelle, de sorte que le schéma est plus proche de l'optimalité.

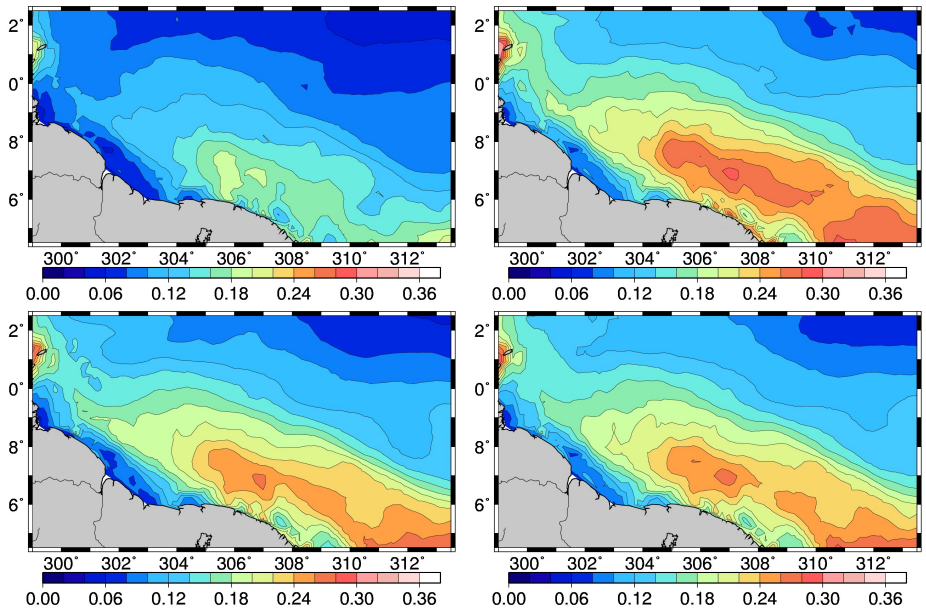


FIGURE 6.3 – Ecart-type de l'incertitude a posteriori sur la vitesse, tel qu'estimé par l'algorithme (à gauche), et écart-type de l'erreur réeelle (à droite), lorsque la corrélation spatiale des erreurs d'observations est négligée (en haut), et lorsqu'elle a été simulée en incluant le gradient des observations originales au vecteur d'observation (en bas).

Il est possible de donner de ce comportement une interprétation plus intuitive. Quand les erreurs d'observation sont corrélées spatialement, le gradient est observé avec plus de précision que si les erreurs sont décorrélées. On s'en rend facilement compte en observant le bruit corrélé horizontalement sur la fig. 6.2, qui perturbe moins le gradient que s'il s'agissait d'un bruit blanc de même variance. Ceci permet de comprendre pourquoi l'ajout d'observations du gradient (dépendantes des observations originales) dans le vecteur d'observation produit le même effet que de tenir compte explicitement de la corrélation spatiale des erreurs d'observation. Cet effet est d'autant plus important quand il s'agit d'observations d'altimétrie océanique, car c'est le gradient de l'élévation du niveau de la mer qui est directement lié à la vitesse de surface (par géostrophie). Il est donc dans ce cas particulièrement préjudiciable de négliger une information précise sur le gradient, d'autant mieux présente dans les observations que les erreurs sont spatialement corrélées.

6.5 Modélisation adaptative des incertitudes

Un troisième verrou qui rend difficile l'application du modèle gaussien, surtout avec une forte réduction d'ordre, à un problème réaliste provient d'incertitudes inévitables sur les paramètres des distributions gaussiennes a priori ($\hat{\mathbf{P}}^b$ et $\mathbf{R}$), sur leur propagation dans le temps (pour obtenir la distribution conjointe 6.3), voire même d'approximations inhérentes au modèle gaussien lui-même. La conséquence directe de ces approximations est la divergence avec le temps des matrices de covariance d'erreur (a priori ou a posteriori), qui finissent par perdre tout rapport avec l'incertitude réelle. Car, avec le modèle gaussien, le calcul des covariances d'erreur se fait indépendamment du calcul de l'espérance mathématique (voir éqs. 6.3, 6.5 et 6.6) ; il dépend donc entièrement de la validité des hypothèses le concernant, et pas du tout des observations $\hat{\mathbf{y}}^o$. Une solution assez générale à ce problème (Dee, 1995; Lermusiaux, 2007) est d'inclure aux matrices de covariance des paramètres supplémentaires décrivant ces incertitudes (par

exemple un simple facteur multiplicatif, éventuellement fonction du temps), et d'ajouter ces paramètres au vecteur de contrôle $\hat{\mathbf{x}}$. De cette manière, les matrices de covariance des incertitudes pourront *s'adapter* à l'écart aux observations, et le problème de leur divergence avec le temps sera résolu.

Méthode. Supposons par exemple que la matrice de covariance $\mathbf{P}^b$ dans la distribution conjointe de l'éq. (6.3) dépende d'un vecteur de paramètres incertains $\boldsymbol{\alpha}$, et que la matrice de covariance d'erreur $\mathbf{R}$ dépende d'un vecteur de paramètres incertains $\boldsymbol{\beta}$. Dans ce cas, la distribution conditionnelle de l'écart aux observations ($\delta\mathbf{y}^o = \mathbf{y}^o - \mathbf{H}\mathbf{M}\hat{\mathbf{x}}^b_{\text{exp}}$), pour $\boldsymbol{\alpha}$ et $\boldsymbol{\beta}$ fixés, peut s'écrire de façon générale sous la forme :

$$p(\delta\mathbf{y}^o|\boldsymbol{\alpha},\boldsymbol{\beta}) = \frac{1}{\sqrt{(2\pi)^p|\mathbf{C}(\boldsymbol{\alpha},\boldsymbol{\beta})|}}\ \exp\left[-\frac{1}{2}\,\delta\mathbf{y}^{oT}\ \mathbf{C}^{-1}(\boldsymbol{\alpha},\boldsymbol{\beta})\ \delta\mathbf{y}^o\right] \tag{6.21}$$

où la matrice $\mathbf{C}(\boldsymbol{\alpha},\boldsymbol{\beta})$ est la matrice de covariance de $\delta\mathbf{y}^o$, sommant une contribution de l'erreur d'ébauche (issue de $\mathbf{P}^b$) et la matrice de covariance d'erreur d'observation $\mathbf{R}(\boldsymbol{\beta})$. A partir de là, il devient en principe possible de déterminer une distribution de probabilité a posteriori pour $\boldsymbol{\alpha}$ et $\boldsymbol{\beta}$, conditionnée à l'écart aux observations $\delta\mathbf{y}^o$, en utilisant le théorème de Bayes. La difficulté à résoudre est bien sûr qu'au vu de la dépendance complexe de $p(\delta\mathbf{y}^o|\boldsymbol{\alpha},\boldsymbol{\beta})$ en $\boldsymbol{\alpha}$ et $\boldsymbol{\beta}$, la distribution de probabilité a posteriori $p^a(\boldsymbol{\alpha},\boldsymbol{\beta}) \propto p^b(\boldsymbol{\alpha},\boldsymbol{\beta})\ p(\delta\mathbf{y}^o|\boldsymbol{\alpha},\boldsymbol{\beta})$ ne pourra pas être supposée gaussienne. Et par ricochet, la distribution de probabilité pour le reste du vecteur de contrôle, prenant en compte l'incertitude a posteriori sur $\boldsymbol{\alpha}$ et $\boldsymbol{\beta}$, ne sera pas non plus gaussienne. Pour simplifier le problème, l'approximation qui est généralement retenue (par exemple par Dee, 1995) consiste à d'abord rechercher un meilleur estimé pour $\boldsymbol{\alpha}$ et $\boldsymbol{\beta}$ (en utilisant l'une des méthodes décrite en section 5.3), puis à utiliser ce meilleur estimé de $\boldsymbol{\alpha}$ et $\boldsymbol{\beta}$ pour résoudre le reste du problème. Cela signifie que, pour préserver le modèle gaussien, l'incertitude résiduelle sur $\boldsymbol{\alpha}$ et $\boldsymbol{\beta}$ n'est pas prise en compte.

Cependant, malgré cette approximation, le calcul d'un meilleur estimé pour $\boldsymbol{\alpha}$ et $\boldsymbol{\beta}$ reste extrêmement coûteux. Par exemple, le calcul itératif de l'estimateur au sens du maximum de vraisemblance [maximisant $p(\delta\mathbf{y}^o|\boldsymbol{\alpha},\boldsymbol{\beta})$ donné par l'éq. (6.21)] requerra le calcul de l'inverse et du déterminant de la matrice $\mathbf{C}$ (dans l'espace des observations) pour chaque itéré de $\boldsymbol{\alpha}$ et $\boldsymbol{\beta}$. Ce coût, en général prohibitif, a conduit de nombreux auteurs à rechercher une solution encore plus approximative, consistant à ajuster aux mieux les paramètres pour rendre les variances dans $\mathbf{C}$ aussi cohérentes que possible avec la variance des écarts aux observations (Testut et al., 2003; Li et al., 2009). Pour aller au delà de cela, la méthode que nous avons proposée (voir Brankart et al., 2010, en annnexe B) consiste à remarquer, qu'avec l'algorithme transformé (ou mieux encore avec l'algorithme en base propre), le coût de calcul de l'inverse et du déterminant de $\mathbf{C}(\boldsymbol{\alpha},\boldsymbol{\beta})$ est rendu négligeable, du moins pour quelques cas importants de paramètres $\boldsymbol{\alpha}$ et $\boldsymbol{\beta}$. Dans ce travail, la méthode a été appliquée au cas d'un filtre de Kalman, et les paramètres $\boldsymbol{\alpha}$ et $\boldsymbol{\beta}$ qui ont été étudiés sont : un facteur d'amplification de la matrice de covariance d'erreur sur la condition finale de chaque cycle d'assimilation, un facteur d'amplification de la matrice de covariance d'erreur d'observation, ainsi que la longueur de corrélation des erreurs d'observation (en profitant de la paramétrisation décrite en section 6.4).

Application. La méthode a ensuite été testée (voir Brankart et al., 2010, en annexe B pour plus de détail) avec la configuration SEABASS (comme en section 6.3). La figure 6.4 montre par exemple le résultat de l'estimation d'un facteur d'amplification de la matrice de covariance d'erreur de prévision de l'élévation du niveau de la mer. Ce résultat a été obtenu dans le cadre d'expériences idéalisées, pour lesquelles la valeur exacte du facteur d'amplification est connue (ici $\alpha = 1$). La figure montre ainsi (à gauche) comment la

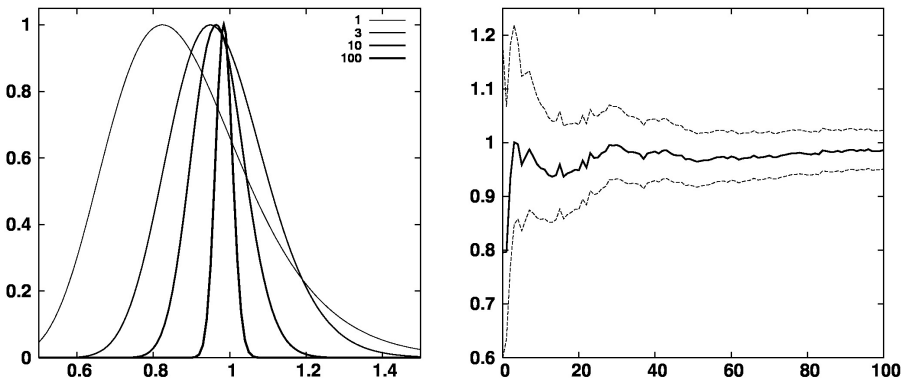


FIGURE 6.4 – Cette figure montre, en fonction de la quantité d'observations impliquées dans l'estimation, la fonction de vraisemblance pour α (à gauche), ainsi que le mode et deux quantiles (0.1 et 0.9) de sa distribution de probabilité a posteriori.

fonction de vraisemblance pour α $[p(\mathbf{y}^o|\alpha)]$ se resserre autour de la valeur exacte quand la quantité d'observations impliquées dans l'estimation augmente. Par ailleurs, si on utilise une distribution de probabilité a priori pour α [ici, $p^b(\alpha) = \exp(-\alpha)$], on peut aussi calculer une distribution de probabilité a posteriori $[p^a(\alpha) \propto p^b(\alpha)\ p(\mathbf{y}^o|\alpha)]$, illustrée sur la fig. 6.4 (à droite) par son mode, ainsi que les quantiles 0.1 et 0.9 en fonction de la quantité d'observations utilisées. Ce résultat montre que l'estimation optimale à partir des observations de certains paramètres des matrices de covariance est possible. Il convient néanmoins de rester prudent, car d'une part, l'estimation des paramètres peut être faussée par la présence d'autres approximations dans les matrices de covariance qui ne seraient pas incluses dans le vecteur de contrôle. Et d'autre part, si trop de paramètres sont inclus dans le vecteur de contrôle, ils pourraient ne pas être simultanément contrôlables par les observations disponibles. Le compromis le plus adapté à une application donnée résultera donc encore de choix partiellement subjectifs.

Chapitre 7

Au delà du modèle gaussien

> When a traveler reaches a fork
> in the road, the l1-norm tells him
> to take either one way or the other,
> but the l2-norm instructs him
> to head off into the bushes.
>
> John F. Claerbout
> and Francis Muir (1973)

Le modèle gaussien réduit la résolution du problème inverse au calcul d'une espérance mathématique et d'une covariance ($\hat{\mathbf{x}}^a_{\text{exp}}$ et $\hat{\mathbf{P}}^a$) par des opérations d'algèbre linéaire (éqs. 6.5 et 6.6). La simplicité de cette solution linéaire, ainsi que son fort pouvoir intuitif, génèrent souvent une forte tentation d'appliquer le modèle gaussien, même quand il est très approximatif, voire tout à fait inapproprié. Dans ce cas, la solution du problème peut perdre toute utilité, ou même devenir complètement paradoxale (voir citation en entête de chapitre). Le propos de ce chapitre est d'examiner s'il n'y a pas moyen d'explorer un peu plus les caractéristiques de l'ensemble (selon le schéma de la figure 5.1), en utilisant un modèle plus riche, tout en gardant une complexité algorithmique raisonnable.

7.1 Gaussiennes tronquées

Une première circonstance qui met le modèle gaussien en difficulté est la présence de *contraintes d'inégalité* sur le vecteur de contrôle $\hat{\mathbf{x}}$. De telles contraintes sont souvent associées à la définition même des variables de contrôle (concentration positive, épaisseur de glace positive, fraction de glace entre 0 et 1, précipitations positives, taux de croissance du phytoplancton positif,...), mais peuvent aussi provenir de contraintes dynamiques ou thermodynamiques (température de l'eau de mer ou de la glace de mer supérieure ou inférieure à la température de fusion, stabilité hydrostatique de la colonne d'eau,...). Dans ce cas, avec un modèle gaussien, une part importante de probabilité cumulée pourra se porter sur des valeurs du vecteur de contrôle ne satisfaisant pas les contraintes (dont la probabilité devrait être nulle), et ce d'autant plus que l'incertitude est grande et que le système est proche des contraintes. C'est également dans ces circonstances que les contraintes d'inégalité apportent l'information la plus considérable et qu'il est donc important les de prendre en compte dans la solution du problème inverse.

Méthode. Pour résoudre ce problème, la solution la plus directe (que nous avons proposée dans le travail de Lauvernet et al., 2009, voir annexe B) consiste à supposer que le vecteur de contrôle $\hat{\mathbf{x}}$ est gaussien, avec la condition supplémentaire que des contraintes

d'inégalité $I(\hat{\mathbf{x}}) \leq 0$ sont satisfaites. Cette hypothèse conduit à la distribution gaussienne tronquée :

$$\mathcal{TN}_{\hat{\mathbf{x}}}(\hat{\mathbf{x}}_{\text{loc}}, \hat{\mathbf{P}}_{\text{scal}}; I) = \begin{cases} \frac{1}{\alpha}\mathcal{N}_{\hat{\mathbf{x}}}(\hat{\mathbf{x}}_{\text{loc}}, \hat{\mathbf{P}}_{\text{scal}}) & I(\hat{\mathbf{x}}) \leq 0 \\ 0 & I(\hat{\mathbf{x}}) > 0 \end{cases} \tag{7.1}$$

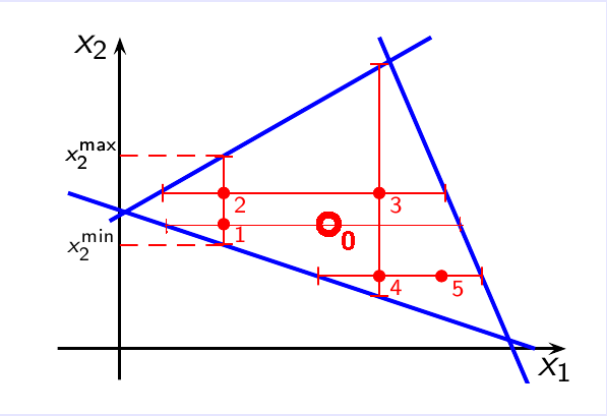
qui dépend du vecteur de localisation $\hat{\mathbf{x}}_{\text{loc}}$ et de la matrice d'échelle $\hat{\mathbf{P}}_{\text{scal}}$ (qui ne correspondent plus à l'espérance mathématique et à la covariance de $\hat{\mathbf{x}}$), et où α est un facteur de normalisation. L'intérêt de ce modèle est que l'application du théorème de Bayes [multiplication par $p(\mathbf{y}^o|\hat{\mathbf{x}})$, voir éq. (5.2)] se ramène à l'algorithme gaussien sur le domaine $I(\hat{\mathbf{x}}) \leq 0$, et donne trivialement zéro sur le domaine $I(\hat{\mathbf{x}}) > 0$. Donc, si la distribution de probabilité a priori est une gaussienne tronquée de la forme donnée par l'éq. (7.1), la distribution a posteriori le sera également. Et les valeurs a posteriori $\hat{\mathbf{x}}^a_{\text{loc}}$ et $\hat{\mathbf{P}}^a_{\text{scal}}$ pourront encore se déduire des valeurs a priori $\hat{\mathbf{x}}^b_{\text{loc}}$ et $\hat{\mathbf{P}}^b_{\text{scal}}$ par les équations (6.5) et (6.6), ou bien même par les équations (6.8) à (6.12) quand la matrice $\hat{\mathbf{P}}_{\text{scal}}$ prend la forme racine carrée (de rang réduit) de l'éq. (6.7).

Il y a néanmoins deux difficultés. La première est qu'il n'est pas facile d'estimer $\hat{\mathbf{x}}_{\text{loc}}$ et $\hat{\mathbf{P}}_{\text{scal}}$ à partir d'un échantillon (par exemple pour obtenir $\hat{\mathbf{x}}_{\text{loc}}$ et $\hat{\mathbf{P}}_{\text{scal}}$ à partir d'une prévision d'ensemble). Il existe des méthodes générales pour le faire (par exemple Griffiths, 2006), mais celles-ci sont plutôt coûteuses, de sorte qu'en pratique, nous nous sommes contentés d'utiliser la moyenne et la covariance de l'échantillon pour $\hat{\mathbf{x}}_{\text{loc}}$ et $\hat{\mathbf{P}}_{\text{scal}}$ (ce que nous avons appelé l'approximation quasi-gaussienne). Cela peut paraître une approximation sévère (et ça l'est sûrement sous bien des aspects), mais il faut remarquer que c'est ce qui serait fait avec le modèle gaussien. Ici du moins imposons-nous une probabilité nulle dans le domaine $I(\hat{\mathbf{x}}) > 0$. La deuxième difficulté est de produire un échantillon de la gaussienne tronquée elle-même (par exemple pour réaliser un nouvel ensemble à partir de $\hat{\mathbf{x}}_{\text{loc}}$ et $\hat{\mathbf{P}}_{\text{scal}}$). Pour cela, la méthode que nous avons retenue est un *échantillonneur de Gibbs* (Geman and Geman, 1984), qui consiste à construire progressivement l'échantillon par une chaîne de Monte Carlo, en ne modifiant à chaque étape qu'une des variables $\hat{x}_i$ du vecteur de contrôle $\hat{\mathbf{x}}$, et en balayant successivement toutes les valeurs de i (voir l'encadré 8, page 79). Plus précisément, à chaque étape, la nouvelle valeur de $\hat{x}_i$ est échantillonnée à partir de la distribution conditionnelle $p(\hat{x}_i|\hat{x}_1 \ldots \hat{x}_{i-1}, \hat{x}_{i+1} \ldots \hat{x}_n)$ dans laquelle les valeurs $\hat{x}_1 \ldots \hat{x}_{i-1}, \hat{x}_{i+1} \ldots \hat{x}_n$ de l'étape précédente sont utilisées pour conditionner la nouvelle valeur de $\hat{x}_i$. Or, quand la distribution conjointe $p(\hat{\mathbf{x}})$ est une gaussienne tronquée et que les contraintes d'inégalité définissent un domaine convexe [ce qui est en particulier le cas pour des contraintes linéaires : $I(\hat{\mathbf{x}}) = \mathbf{A}\mathbf{x} - \mathbf{b}$], la distribution de probabilité conditionnelle $p(\hat{x}_i|\hat{x}_1 \ldots \hat{x}_{i-1}, \hat{x}_{i+1} \ldots \hat{x}_n)$ est également une gaussienne tronquée avec $I(\hat{x}_i)$ se ramenant à $\hat{x}_i^{\min} \leq \hat{x}_i \leq \hat{x}_i^{\max}$. Et une gaussienne tronquée unidimensionnelle de cette sorte peut s'échantillonner de façon très efficace (Geweke, 1991) par une simple méthode de rejet (à partir d'une distribution gaussienne, semi-gaussienne ou bien exponentielle selon les valeurs de $\hat{x}_i^{\min}$ et $\hat{x}_i^{\max}$). Ce calcul peut être rendu particulièrement simple et efficace quand $\hat{\mathbf{x}}_{\text{loc}} = 0$ et $\hat{\mathbf{P}}_{\text{scal}} = \mathbf{I}$ (ce qui est toujours le cas avec l'algorithme transformé, voir équation 6.9), et quand les contraintes d'inégalités sont linéaires [$I(\hat{\mathbf{x}}) = \mathbf{A}\mathbf{x} - \mathbf{b}$] ; car dans ce cas, les bornes de la distribution de probabilité conditionnelle $\hat{x}_i^{\min}$ et $\hat{x}_i^{\max}$ se déduisent très directement des paramètres ($\mathbf{A}$ et $\mathbf{b}$) des contraintes d'inégalité (voir Lauvernet et al., 2009, pour plus de détail).

Applications. La contrainte d'inégalité qui nous a initialement motivés à développer cette méthode est la contrainte de stabilité hydrostatique de la colonne d'eau, qui impose (hors épisodes de convection) que la densité potentielle ρ^p (référencée à la pression in situ locale) croisse monotonément avec la profondeur : $\frac{\partial \rho^p}{\partial z} \leq 0$. Avec le modèle gaussien, cette contrainte pouvait être très fréquemment transgressée par la plus grande

ENCADRÉ 8 : L'ÉCHANTILLONNEUR DE GIBBS

L'échantillonneur de Gibbs est une méthode MCMC (Markov Chain Monte Carlo method), introduite par Geman and Geman (1984), pour échantillonner une distribution de probabilité multivariée $p(\mathbf{x})$. L'objectif de cet encadré est d'illustrer graphiquement le fonctionnement de cette méthode, utilisée en section 7.1 et 7.4 :



L'algorithme démarre d'une valeur possible $\mathbf{x}^0$ du vecteur aléatoire. Ensuite, on boucle de façon répétée sur les composantes x_i, $i = 1, \ldots, n$ du vecteur aléatoire, et on échantillonne une nouvelle valeur x_i^k de cette composante dans la distribution conditionnelle $p(x_i^k | x_1^k, \ldots, x_{i-1}^k, x_{i+1}^{k-1}, \ldots, x_n^{k-1})$, conditionnée aux $n-1$ dernières valeurs précédemment acquises pour les autres composantes. Sur le schéma (ici une gaussienne tronquée à l'intérieur du triangle bleu, à deux dimensions : $n = 2$), on passe ainsi du point 0 au point 1 en échantillonnant x_1^1 dans $p(x_1^1 | x_2^0)$; puis du point 1 au point 2 en échantillonnant x_2^1 dans $p(x_2^1 | x_1^1)$; puis du point 2 au point 3 en échantillonnant x_1^2 dans $p(x_1^2 | x_2^1)$; etc. En pratique cependant, il est parfois nécessaire d'éliminer les p premiers éléments $\mathbf{x}^k$, $k = 1, \ldots, p$, car ce n'est qu'asymptotiquement que la chaîne de Monte Carlo converge vers un échantillon de la distribution conjointe ; voire de ne retenir qu'un élément sur q : $\mathbf{x}^{p+qk}$, $k = 1, \ldots, m$, afin d'éliminer une dépendance éventuelle entre éléments successifs.

Dans le cadre de nos travaux, cette méthode peut être d'une grande utilité pour échantillonner la distribution de probabilité a posteriori pour le vecteur de contrôle $p^a(\mathbf{x})$, dans la mesure où les distributions conditionnelles (1D) s'échantillonnent beaucoup plus facilement que la distribution conjointe, et si le vecteur de contrôle ne comporte pas trop de dimensions (comme par exemple en cas de réduction d'ordre, voir section 6.2). Ce fut le cas pour les distributions gaussiennes tronquées (voir section 7.1) et pour des distributions très non-gaussiennes avec de nombreux modes secondaires (voir section 7.4). Cependant, cet algorithme ne peut absolument pas être vu comme un moyen de vaincre la malédiction des dimensions, parce qu'il ne convergera en général qu'extrêmement lentement vers les régions de forte probabilité si $\mathbf{x}^0$ n'est qu'une combinaison très peu probable des variables de contrôle, et parce qu'il ne permet de passer que très lentement d'un mode à l'autre s'ils sont séparés par des zones très peu probables. C'est pourquoi nous ne l'avons jamais considéré qu'en complément d'une autre méthode (par exemple le refroidi simulé en section 7.4) permettant de détecter une valeur initiale $\mathbf{x}^0$ qui soit suffisamment proche du mode de la distribution. En suivant la même logique, il me semble que l'échantillonneur de Gibbs serait aussi un candidat possible pour compléter les méthodes variationnelles (donnant le mode de $p^a(\mathbf{x})$, voir section 5.3) par une information sur l'incertitude a posteriori.

part des réalisations de la distribution de probabilité $p^a(\hat{\mathbf{x}})$, y compris par l'espérance mathématique $\hat{\mathbf{x}}^a_{\text{exp}}$, surtout quand la colonne d'eau n'est que très partiellement observée (par exemple, seulement la température de surface) et quand l'observation va dans le sens d'une augmentation de la densité de surface. La figure 7.1 (à gauche) montre par exemple un échantillon de $p^a(\hat{\mathbf{x}})$ obtenu dans cette circonstance quand on utilise le modèle gaussien : tous les profils de densité de l'échantillon sont hydrostatiquement instables (ainsi que l'espérance mathématique, en rouge), et ils sont tous très éloignés du profil de densité véritable (en vert). A l'inverse, avec le modèle gaussien tronqué (à droite), tous les profils de l'échantillon (générés par l'échantillonneur de Gibbs) sont hydrostatiquement stables (par construction), et l'espérance mathématique (courbe rouge continue) aussi [1]. Et on peut se rendre compte à quel point l'information contenue dans les contraintes permet de les rendre tous plus proches du profil de densité véritable (en vert). Par contre, le vecteur de localisation (courbe rouge en pointillé) de la gaussienne tronquée (identique ici à l'espérance mathématique de la gaussienne de la figure de gauche, sous l'hypothèse quasi-gaussienne) ne respecte pas les contraintes d'inégalité, et ne présente aucun intérêt pratique, à part celui d'être un paramètre de la distribution de probabilité.

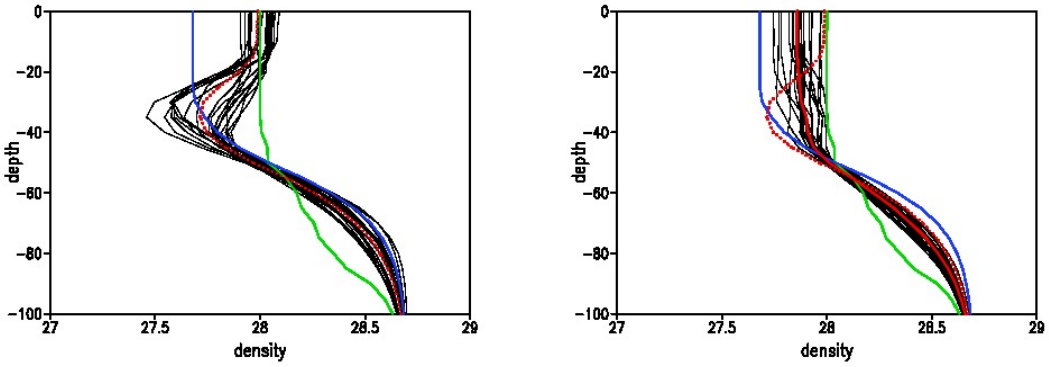


FIGURE 7.1 – Echantillon de $p^a(\hat{\mathbf{x}})$ obtenu en se basant sur le modèle gaussien (à gauche) et sur le modèle gaussien tronqué (à droite). Les courbes de couleur représente l'espérance mathématique de l'ébauche (en bleu), de l'observation (en vert) et de la distribution a posteriori (en rouge). La courbe rouge en pointillé représente le vecteur de localisation de la gaussienne tronquée.

Dans le travail de Lauvernet et al. (2009), cette méthode a également été appliquée à un modèle au 1/15° de résolution du Golfe de Gascogne (développé par Broquet et al., 2008). Il s'agissait d'un modèle à coordonnée verticale hybride (HYCOM, Hybrid COordinate Ocean Model Bleck, 2002; Chassignet et al., 2003), qui, dans l'océan intérieur (c'est-à-dire sous la couche de mélange de surface, et loin des zones côtières), décrit la stratification par une superposition de couches isopycnales (d'égale densité). Dans ce cas, la contrainte $\frac{\partial \rho^p}{\partial z} \leq 0$ se traduit simplement par une contrainte de positivité de l'épaisseur de ces couches. Avec cette application, nous avons montré que la méthode est aussi applicable à un problème de taille réaliste, et qu'ici aussi la contrainte de stabilité hydrostatique contient une information importante, qui facilite l'extrapolation verticale des observations de température de surface, tout en évitant de produire une solution non-physique ("into the bushes" pour reprendre la citation en entête de chapitre).

1. L'espérance mathématique vérifiera toujours les contraintes d'inégalité quand, comme ici, elles définissent un domaine convexe dans l'espace de contrôle.

7.2 Anamorphose

Une deuxième circonstance qui met le modèle gaussien en difficulté est la *nonlinéarité des lignes de régression* entre variables du vecteur de contrôle (et avec les grandeurs observées). Par exemple, la figure 7.2 (à gauche) montre un ensemble de valeurs de phytoplancton (variable X_1 en abscisse) et de profondeur de couche de mélange (variable X_2 en ordonnée), obtenu par perturbation gaussienne du forçage par le vent (Béal et al., 2010). Sur cet exemple, il est particulièrement clair que la ligne de régression (ligne de maximum de probabilité de X_2 pour chaque valeur fixée de X_1) n'est pas une droite. Or, sous le modèle gaussien, toutes les lignes de régressions sont toujours des droites (voir encadré 7 au chapitre 6). C'est ce qui permet d'estimer des variables non-observées en utilisant des formules linéaires (éqs. 6.5 et 6.6), ce qui se fait essentiellement en suivant les droites de régression (et aussi en donnant un poids approprié aux différentes sources d'information). Mais si, dans l'exemple de la figure 7.2 (à gauche), on cherchait à estimer X_2 à partir d'une observation de X_1 en suivant une droite de régression, même la meilleure possible, on voit bien que l'on pourrait facilement être emmené sur une valeur extrêmement improbable de X_2 ("into the bushes"). Le modèle gaussien n'est donc pas approprié.

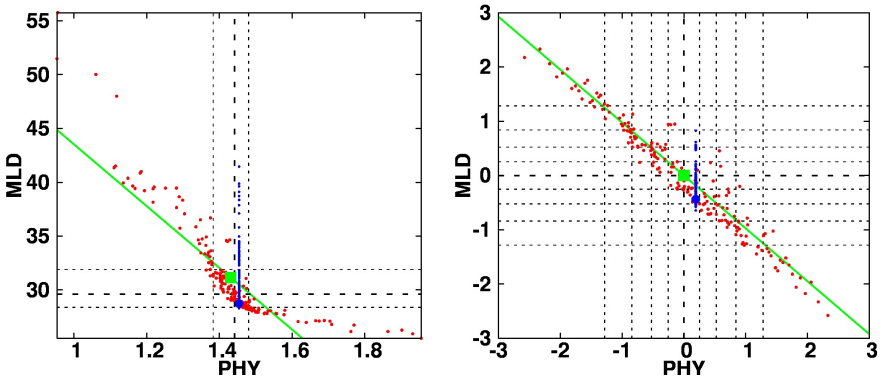


FIGURE 7.2 – Illustration de l'effet de l'anamorphose sur le conditionnement d'un ensemble à une observation parfaite de phytoplancton (65°W/32°N), en appliquant l'hypothèse gaussienne aux variables originales (à gauche), et aux variables transformées (à droite). La figure montre une prévision d'ensemble (points rouge), sa moyenne (carré vert), la droite de régression linéaire (ligne verte), la simulation de référence (gros point bleu) qui fournit l'observation de phytoplancton, et l'ensemble conditionné à l'observation (points bleus). Les lignes noires discontinues représentent la médiane (tirets) et certains autres quantiles (pointillés), à savoir les quartiles sur la figure de gauche et les déciles sur la figure de droite.

Méthode. Pour remédier à cela, un moyen potentiellement peu coûteux consiste à rechercher, pour chaque composante $\hat{x}_i,\ i = 1, \ldots, n$ du vecteur de contrôle $\hat{\mathbf{x}}$ et pour chaque composante $\hat{y}_j,\ j = 1, \ldots, p$ du vecteur d'observation, un changement de variable non-linéaire (anamorphose) qui transforme la distribution de probabilité marginale de chacune de ces composantes en une distribution gaussienne. Chacune de ces distributions unidimensionnelles peut en effet être identifiée à partir d'un ensemble relativement modeste, au contraire d'une distribution de probabilité dans un espace de dimension $m + p$ (malédiction des dimensions, voir chapitre 5). Pour identifier tous ces changements de variables non-linéaires, nous avons proposé un algorithme très simple et efficace (décrit dans les travaux de Béal et al., 2010; Brankart et al., 2012, voir annexe B), qui consiste simplement à les exprimer sous la forme de fonctions linéaires par morceaux

qui transforment certains quantiles de l'ensemble en les quantiles correspondants d'une distribution gaussienne (de moyenne nulle et d'écart-type égal à 1). C'est ce qui a été fait par exemple dans la fig. 7.2 pour obtenir le graphe de droite à partir du graphe de gauche : pour chacune des 2 variables, les déciles de l'ensemble ont été transformés en déciles d'une gaussienne (traits en pointillé, à droite). Après cette transformation, nous pouvons donc au moins être assurés (par construction) que toutes ces distributions marginales associées à la distribution conjointe $p^b(\hat{\mathbf{x}}, \mathbf{y})$ sont des gaussiennes. Il est bien sûr important de garder à l'esprit que cela ne garantit en rien la gaussianité de la distribution conjointe, comme cela a dû être supposé au chapitre 6 pour pouvoir appliquer le modèle gaussien. Mais garantir explicitement la gaussianité des distributions marginales peut néanmoins se révéler très utile. Cela résout en particulier le problème des contraintes d'inégalité n'impliquant qu'une seule variable à la fois [2] (concentrations positives, fraction de glace entre 0 et 1,...). C'est d'ailleurs principalement pour résoudre ce problème-là que cette méthode des transformations anamorphiques (développée en géostatistique) a été initialement introduite en océanographie par Bertino et al. (2003) dans le cadre du filtre de Kalman d'ensemble.

Mais il y a plus. Cette transformation modifie également le coefficient de corrélation linéaire entre les variables transformées. Cette modification est importante car, sous le modèle gaussien, seule une dépendance linéaire (décrite par le coefficient de corrélation linéaire) peut être représentée, et c'est donc seulement quand une dépendance linéaire peut être diagnostiquée qu'il est possible de gagner de l'information sur le vecteur de contrôle à partir des observations. Une façon utile de comprendre comment la corrélation linéaire est modifiée par l'anamorphose est d'observer que le coefficient de corrélation linéaire entre les variables transformées correspond à une mesure paramétrique de la corrélation (Hollander and Wolfe, 1973; Corder and Foreman, 2009) entre les variables originales. En résumé, les deux avantages principaux d'une telle mesure de la corrélation est (i) qu'elle permet de voir une dépendance non-linéaire entre variables aléatoires, et (ii) qu'elle est plus robuste à la présence de quelques points anormaux (outliers) dans les données. Or, c'est précisément dans ces deux circonstances que le coefficient de corrélation linéaire peut produire une réprésentation infidèle de la dépendance entre variables aléatoires (comme illustré par les exemples d' Anscombe, 1973). L'exemple le plus ancien et le plus simple de mesure non-paramétrique de la corrélation est la corrélation de rang (Spearman, 1904; Kendall, 1962), définie comme la corrélation linéaire entre les rangs des deux variables dans l'ensemble. Cette mesure classique est d'ailleurs très semblable à la corrélation linéaire entre variables anamorphosées, puisque le rang correspond pratiquement au résultat d'une anamorphose vers une distribution uniforme, en lieu et place d'une distribution gaussienne (voir Brankart et al., 2012, en annexe B pour plus de détail).

Applications. Cette propriété que le coefficient de corrélation linéaire entre variables transformées correspond à une mesure non-paramétrique de la corrélation entre variables originales est la raison fondamentale qui explique l'amélioration de la structure de corrélation linéaire dans tous les exemples décrits par Brankart et al. (2012). Par exemple, la figure 7.3 montre la corrélation linéaire pour la concentration de glace (obtenue à partir d'un ensemble utilisé dans le système d'assimilation de Mercator/Océan), avant et après transformation anamorphique. Aussi bien en mars (fig. 7.3, en haut) qu'en septembre (fig. 7.3, en bas), l'effet de l'anamorphose est essentiellement d'accroître la distance de corrélation horizontale. En mars, la distance de corrélation s'accroît surtout dans la direction transversale à l'écoulement, car c'est à travers le front qu'il y a le plus de dépendance non-linéaire entre concentrations de glace. Et en septembre, la distance

2. Ce qui n'était pas le cas de la contrainte $\frac{\partial \rho^p}{\partial z} \leq 0$ examinée en section 7.1.

de corrélation s'accroît surtout dans la direction du courant, car à ce moment, le point de référence est situé à proximité du bord sud de l'extension de la glace. Dans les deux cas, l'accroissement de la distance de corrélation linéaire augmente la quantité d'information que la distribution de probabilité (supposée gaussienne) contient à propos du système, et favorise donc la résolution d'un problème inverse.

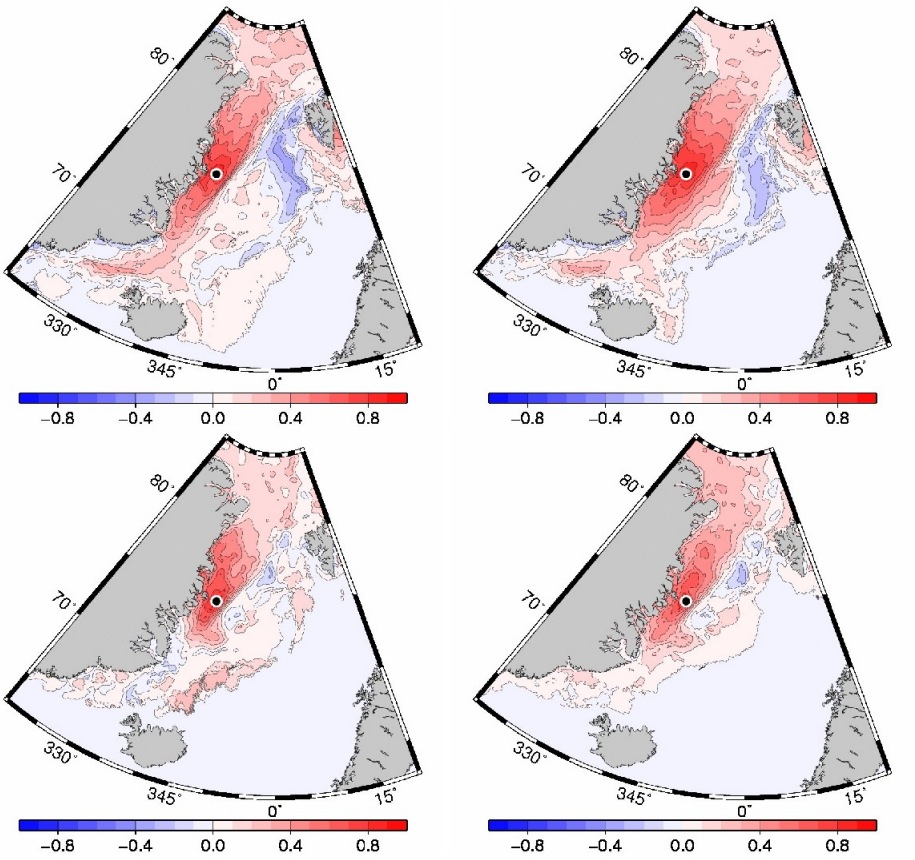


FIGURE 7.3 – Structure de corrélation horizontale pour la concentration de glace par rapport à un point de référence situé à 15°W 75°N (point noir), pour mi-mars (en haut) et mi-septembre (en bas), sans anamorphose (à gauche) et après anamorphose locale (à droite).

Et c'est toujours, à mon avis, cette même raison qui fait que l'anamorphose a été si utile au succès des travaux menés par Doron et al. (2011, 2013) et Fontana et al. (2013), pour accroître la quantité d'information à propos des paramètres ou de l'état d'un modèle biogéochimique qu'il est possible d'extraire des observations de couleur de l'océan.

7.3 Mélange de gaussiennes

Une troisième circonstance qui met le modèle gaussien en difficulté est l'existence de *plusieurs modalités* de réponse du système aux incertitudes. Dans ce cas, la distribution de probabilité a priori peut rapidement devenir multimodale, et ne peut être valablement traitée par aucune des méthodes décrites jusqu'ici. L'un des modèles proposés dans la littérature (Anderson and Anderson, 1999; Bengtsson et al., 2003; Hoteit et al., 2008; Sondergaard and Lermusiaux, 2013) pour résoudre ce problème est de décrire alors la distribution de probabilité a priori comme le mélange (ou la superposition) de distributions gaussiennes élémentaires :

$$p^b(\mathbf{x}) = \sum_{\mathbf{j=1}}^{\mathbf{q}} \mathcal{N}_{\mathbf{x}} \left[\mathbf{x^b}(\boldsymbol{\alpha}_\mathbf{j}), \mathbf{P^b}(\boldsymbol{\alpha}_\mathbf{j}) \right] \mathbf{p}(\boldsymbol{\alpha}_\mathbf{j}) \tag{7.2}$$

chacune d'elles apparaissant avec la probabilité a priori $p(\boldsymbol{\alpha}_j)$. Cependant, cette approche exige d'appliquer les formules 6.5 et 6.6 pour chacune des gaussiennes élémentaires, et elle donc soit excessivement coûteuse (Bengtsson et al., 2003), soit elle conduit à faire l'hypothèse que toutes les gaussiennes élémentaires possèdent la même matrice de covariance (Anderson and Anderson, 1999; Hoteit et al., 2008). Dans cette section, je voudrais pourtant montrer que cette méthode pourrait être appliquée à un coût additionnel assez faible à condition d'utiliser l'algorithme transformé ou l'algorithme en base propre (éqs. 6.8 à 6.12) tel qu'il a été développé pour le filtre SEEK. Ce qui est dit dans cette section présente plusieurs points de convergence avec l'étude récente (et plus générale) de Sondergaard and Lermusiaux (2013).

Mise à jour des distributions gaussiennes élémentaires. Avec l'algorithme classique (inversion dans l'espace des observations par les éqs. 6.5 et 6.6), à moins que les covariances $\mathbf{P}^b(\boldsymbol{\alpha}_j)$ ne soient toutes identiques (comme dans les travaux de Anderson and Anderson, 1999; Hoteit et al., 2008), aucun coût de calcul ne peut être économisé par la répétition des mêmes équations avec différents paramètres $\boldsymbol{\alpha}_j$, et la complexité algorithmique (généralisant l'éq. 6.15) s'écrit :

$$C_C(q) \sim (n+1)D + q(n^2p + np^2 + \alpha p^3) \tag{7.3}$$

Par contre, avec l'algorithme transformé, seule une partie de la complexité algorithmique 6.16 doit être multipliée par q :

$$C_T^I(q) \sim (r+1)D + q(nr^2 + \alpha r^3) + pr^2 \tag{7.4}$$

car aucune des opérations reliées au nombre p d'observations ne doit être répétée pour chaque $\boldsymbol{\alpha}_j$. Ceci exige cependant que le mélange de gaussiennes soit écrit dans l'espace réduit (ce que nous appelerons ici distributions de classe I) :

$$p^b(\boldsymbol{\xi}) = \sum_{j=1}^{q} \mathcal{N}_{\boldsymbol{\xi}} \left[\boldsymbol{\xi}^b(\boldsymbol{\alpha}_j), \boldsymbol{\Pi}^b(\boldsymbol{\alpha}_j) \right] p(\boldsymbol{\alpha}_j) \tag{7.5}$$

En outre, avec l'algorithme en base propre, la complexité algorithmique peut se réduire encore jusqu'à :

$$C_T^{II}(q) \sim (r+1)D + qnr + nr^2 + \alpha r^3 + pr^2 \tag{7.6}$$

à condition toutefois que la base propre de toutes les matrices de covariance $\mathbf{P}^b(\boldsymbol{\alpha}_j)$ soient identiques (ce que nous appelerons distributions de classe II), de sorte que tous les $\boldsymbol{\Pi}^b(\boldsymbol{\alpha}_j)$ se diagonalisent dans la même base.

Mise à jour des poids $p(\boldsymbol{\alpha}_j)$ des distributions élémentaires. Par ailleurs, la mise à jour du mélange de gaussiennes (éq. 7.2) requiert aussi la mise à jour des probabilités $p(\boldsymbol{\alpha}_j)$ de chaque distribution élémentaire [3], ce qui exige de calculer explicitement la probabilité des observations $\hat{\mathbf{y}}^o$ pour chaque $\boldsymbol{\alpha}_j$ donné :

3. La mise à jour des $p(\boldsymbol{\alpha}_j)$ correspond à la mise à jour des ω_i par l'équation (5.4). C'est pourquoi cette approche est parfois présentée comme une variante du filtre particulaire, avec un "habillage gaussien" de chaque particule, basé sur la dispersion des particules voisines.

$$p(\hat{\mathbf{y}}^o|\boldsymbol{\alpha}_j) = \frac{1}{\sqrt{(2\pi)^p|\mathbf{C}(\boldsymbol{\alpha}_j)|}} \exp\left[-\frac{1}{2}\mathbf{d}^T(\boldsymbol{\alpha}_j)\mathbf{C}^{-1}(\boldsymbol{\alpha}_j)\mathbf{d}(\boldsymbol{\alpha}_j)\right] \tag{7.7}$$

où $\mathbf{d}(\boldsymbol{\alpha}_j)$ et $\mathbf{C}(\boldsymbol{\alpha}_j)$ représentent le vecteur d'innovation et la covariance associée pour chaque $\boldsymbol{\alpha}_j$. Avec l'algorithme classique (utilisé par Bengtsson et al., 2003), le calcul du déterminant $D(\boldsymbol{\alpha}_j) = |\mathbf{C}(\boldsymbol{\alpha}_j)|$ et de la forme quadratique $Q(\boldsymbol{\alpha}_j) = \mathbf{d}^T(\boldsymbol{\alpha}_j)\mathbf{C}^{-1}(\boldsymbol{\alpha}_j)\mathbf{d}(\boldsymbol{\alpha}_j)$ doit être répété pour chaque $\boldsymbol{\alpha}_j$ et représente donc encore une complexité algorithmique proportionnelle à qp^3 (comme le dernier terme de l'éq. 7.3). A l'inverse, avec l'algorithme transformé, la calcul de ces deux quantités peut se faire à un coût additionnel négligeable en suivant les mêmes idées que celles développées par Brankart et al. (2011) pour le schéma adaptatif (voir section 6.5), même si les formules qu'on obtient pour $D(\boldsymbol{\alpha}_j)$ et $Q(\boldsymbol{\alpha}_j)$ sont un peu longues pour être reproduites ici.

Identification du mélange de gaussiennes à partir d'un ensemble. Finalement, cette méthode requiert aussi une approche pas trop coûteuse pour identifier le mélange de gaussiennes (éq. 7.2) à partir d'une prévision d'ensemble. Pour cela, une façon assez générale de procéder est d'utiliser une technique d'estimation non-paramétrique d'une densité de probabilité, par exemple par une méthode inspirée de la méthode des plus proches voisins (utilisée dans le travail de Bengtsson et al., 2003), mais que nous appliquerions dans l'espace réduit au lieu de l'espace original (c'est-à-dire pour obtenir directement la distribution sous la forme de l'éq. 7.5). L'idée est simplement d'utiliser certains membres de l'ensemble $\boldsymbol{\xi}_j,\ j = 1, \ldots, q$ $(q \leq m)$ comme particules pivots $\boldsymbol{\xi}^p(\boldsymbol{\alpha}_j) = \boldsymbol{\xi}_j$ des gaussiennes élémentaires, et de calculer ensuite chaque gaussienne élémentaire [c'est-à-dire la moyenne $\boldsymbol{\xi}^b(\boldsymbol{\alpha}_j)$ et la covariance $\boldsymbol{\Pi}^b(\boldsymbol{\alpha}_j)$] comme d'habitude à partir de tout l'ensemble $\boldsymbol{\xi}_i,\ i = 1, \ldots, m$, *mais avec un poids décroissant ave la distance par rapport au pivot* : $w_{ij} \propto f(d_{ij}/\delta_j)$ où d_{ij} est une distance entre chaque membre i et chaque pivot j $(d_{ij} = ||\boldsymbol{\xi}_i - \boldsymbol{\xi}^p(\boldsymbol{\alpha}_j)||)$ et δ_j est une "distance d'exploration" caractérisant la "taille" des plus fins détails que l'on cherche à détecter dans la distribution inconnue. De plus, plutôt qu'un δ_j constant, il est certainement préférable d'explorer plus finement les régions de plus haute densité, et par exemple de définir δ_j comme un certain quantile des d_{ij}, par exemple le quantile $\frac{m^\rho - 1}{m-1}$, $0 < \rho \leq 1$. L'exposant ρ caractérise alors à lui seul le "niveau d'exploration" de l'ensemble puisque m^ρ est à peu près la taille du sous-ensemble (avec les plus grands poids w_{ij} définissant chaque gaussienne élémentaire. [4] Par exemple, si $f(x)$ est définie par

$$f(x) = \begin{cases} 1 & x \leq 1 \\ 0 & x > 1 \end{cases} \tag{7.8}$$

on se ramène exactement à la méthode des plus proches voisins (Silverman, 1986) telle qu'utilisée par Bengtsson et al. (2003). L'intérêt de généraliser avec une fonction f continue (et donc de calculer chaque gaussienne élémentaire avec tout l'ensemble) est de rendre cette approche d'ores et déjà compatible avec la méthode de localisation des covariances décrite en section 6.3, en évitant l'apparition de discontinuités artificielles entre analyses locales adjacentes.

Convergence. Un des intérêts de cette approche est qu'il est possible de montrer que le mélange de gaussienne résultant (éq. 7.5) converge toujours (pour $\rho < 1$) vers la distribution exacte quand le nombre $q \leq m$ de gaussiennes superposées tend vers l'infini.

4. Ici, le niveau d'exploration ρ est donc le seul paramètre libre qu'il reste à déterminer. Une approche plus générale d'identification du mélange de gaussienne à partir d'un esnemble peut être trouvée dans le travail de Sondergaard and Lermusiaux (2013).

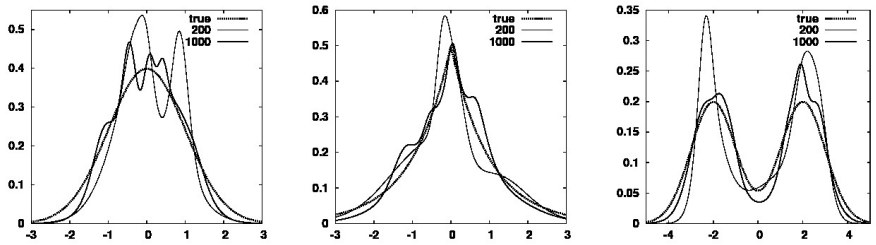


FIGURE 7.4 – Identification d'un mélange de gaussienne à partir d'un échantillon de taille 200 (ligne grasse continue) ou d'un échantillon de taille 1000 (ligne fine continue) de trois distribution unidimensionnelles : (i) la distribution gaussienne $p^b_{\text{true}}(\xi) = \mathcal{N}_\xi(0,1)$ (à gauche), (ii) la distribution de Laplace $p^b_{\text{true}}(\xi) = \frac{1}{2}\exp(-|\xi|)$ (au centre), et (iii) la distribution bimodale $p^b_{\text{true}}(\xi) = \frac{1}{2}[\mathcal{N}_\xi(-2,1) + \mathcal{N}_\xi(2,1)]$ (à droite).

Cela est vrai aussi bien pour les distributions de classe I que pour les distributions de classe II, en raison de la convergence de tous ses moments vers les moments de l'ensemble des particules pivots, qui eux-mêmes convergent vers les moments de la distribution exacte :

$$M_n = \frac{1}{q}\sum_{i=1}^{q}\int \mathcal{N}_{\boldsymbol{\xi}}[\boldsymbol{\xi}^b(\alpha_j), \boldsymbol{\Pi}^b(\alpha_j)]\boldsymbol{\xi}^n \, d\boldsymbol{\xi} \quad \rightarrow \quad \frac{1}{q}\sum_{i=1}^{q}[\boldsymbol{\xi}^p(\alpha_j)]^n \quad \rightarrow \quad \int p^b(\boldsymbol{\xi})\boldsymbol{\xi}^n \, d\boldsymbol{\xi} \tag{7.9}$$

La différence entre les distributions de classe I et de classe II est la précision de l'approximation pour q fixé. Quand q n'est pas très grand, une distribution de classe I sera plus précise, mais cet avantage tendra à disparaître avec l'augmentation de q, car si suffisamment de pièces sont disponibles, leur structure individuelle peut être laissée plus simple.

Niveau d'exploration optimal. La figure 7.4 illustre cette convergence du mélange de gaussiennes vers la distribution exacte quand $q = m$ augmente, et pour $\rho = 0.7$ fixé. Cette figure suggère aussi que pour une taille d'ensemble fixée, il vaut mieux ne pas explorer trop, afin de ne pas faire surgir de détails non-significatifs. Entre l'hypothèse gaussienne ($\rho = 1$) et le filtre particulaire ($\rho \to 0$), il doit donc exister un niveau d'exploration optimal (au sens de la figure 5.1) permettant d'extraire le maximum d'information de l'ensemble disponible. Pour estimer ce niveau d'exploration optimal, on peut imaginer l'algorithme suivant, inspiré des algorithmes de validation croisée (Wahba and Wendelberger, 1980; Brankart and Brasseur, 1996) :

1. Identifier le mélange de gaussienne $p^b_{(k)}(\boldsymbol{\xi};\rho)$ pour un ρ donné, en utilisant toutes les particules sauf la particule k, afin d'obtenir la vraisemblance de cette particule laissée indépendante : $L_k(\rho) = p^b_{(k)}(\mathbf{x}_k;\rho)$;
2. Répéter cette opération en excluant successivement chaque particule $k = 1,\ldots,m$, afin d'évaluer la "vraisemblance croisée" : $L(\rho) = \prod_{k=1}^{m} L_k(\rho)$;
3. Trouver le niveau d'exploration optimal en maximisant $L(\rho)$.

La valeur optimale de ρ ainsi obtenue permet de minimiser la divergence de Kullback-Leibler (l'entropie relative) entre le mélange de gaussienne $p^b(\boldsymbol{\xi};\rho)$ et la distribution vraie $p^t(\boldsymbol{\xi})$, puisque

$$\lambda(\rho) = \left(\frac{L(\rho)}{L^{\text{true}}}\right)^{1/m} = \left(\prod_{k=1}^{m} \frac{p^b_{(k)}(\boldsymbol{\xi}_k;\rho)}{p^t(\boldsymbol{\xi}_k)}\right)^{1/m} = \exp\left\{\frac{1}{m}\sum_{k=1}^{m} \ln\left[\frac{p^b_{(k)}(\boldsymbol{\xi}_k;\rho)}{p^t(\boldsymbol{\xi}_k)}\right]\right\}$$
$$\sim \exp\left\{-\int p^t(\boldsymbol{\xi})\ \ln\left[\frac{p^t(\boldsymbol{\xi})}{p^b(\boldsymbol{\xi},\rho)}\right]\ d\boldsymbol{\xi}\right\} = \exp\left[-D_{\text{KL}}(p^b,p^t)\right] \tag{7.10}$$

peut être interprétée comme une approximation de Monte Carlo de l'intégrale définissant $D_{\text{KL}}(p^b, p^t)$. La figure 7.5 illustre par exemple la fonction $\lambda(\rho)$ pour les trois exemples de la fig. 7.4, avec un ensemble de taille $m = 200$ et pour une dimension croissante de l'espace de contrôle ($r = 1$ à 5). La conclusion la plus importante qu'il faut tirer de cette figure est qu'avec un ensemble de taille limitée (ici $m = 200$), il est très dangereux de chercher à explorer excessivement la distribution de probabilité ($\rho < 0.25$). Par ailleurs, la malédiction des dimensions demeure (la distribution correspondant à la figure de droite contient 2^r modes), et il ne peut être bénéfique de rechercher ainsi les comportements multimodaux que dans un espace de dimension faible (par exemple les quelques premières EOFs de l'ensemble) : typiquement $r \leq 4$ pour $m = 200$ (ne croissant qu'avec le logarithme de la taille m de l'ensemble).

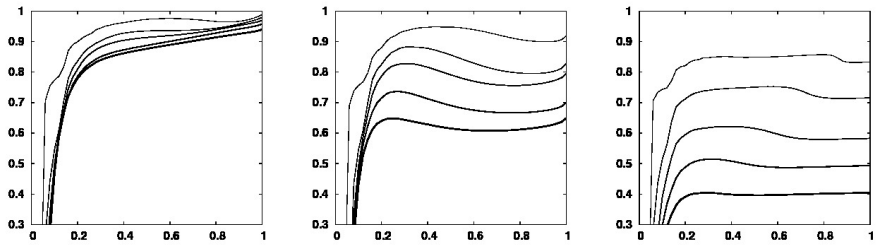


FIGURE 7.5 – Vraisemblance croisée normalisée $\lambda(\rho)$ (en ordonnée), en fonction du niveau d'exploration ρ (en abscisse). $\lambda(\rho)$ est reliée à la vraisemblance croisée par $L(\rho) = L^{\text{true}}[\lambda(\rho)]^m$ et à l'entropie relative (par rapport à la distribution vraie) par $D_{\text{KL}}(\rho) \sim -\ln[\lambda(\rho)]$ (voir eq. 7.10). Elle est montrée pour différentes dimensions de l'espace de contrôle ($r = 1$ à 5, de la ligne la plus fine à la plus épaisse), pour de mélanges de classe I et pour les mêmes trois distributions qu'en fig. 7.4 : une distribution gaussienne (à gauche), une distribution de Laplace (au centre) et une distribution mutlimodale (à droite). Pour assurer la significativité des résultats, chaque courbe représente la moyenne géométrique de 25 fonctions $\lambda(\rho)$, calculées à partir d'échantillons indépendants, chacun de taille 200.

7.4 Chaînes de Markov

Jusqu'ici, toutes les méthodes abordées aux chapitres 6 et 7 fonctionnaient toujours en deux étapes. On recherche d'abord une description explicite de l'incertitude a priori sur les quantités observées (la distribution a priori $p^b(\hat{\mathbf{x}}, \mathbf{y})$, voir éq. 6.3), par exemple par une simulation d'ensemble. Et ce n'est qu'ensuite qu'on en déduit la distribution de probabilité a posteriori, en conditionnant $p^b(\hat{\mathbf{x}}, \mathbf{y})$ aux observations. Une telle approche en deux étapes est particulièrement sensible à la malédiction des dimensions, puisqu'elle oblige à explorer d'abord explicitement tout ce qui est possible a priori, avant d'éliminer toutes le modalités de comportement du système que les observations rendent peu vraisemblables. Pour cette raison, les différentes approximations proposées aux chapitres 6 et 7 épuisent à peu près toutes les options dont le coût numérique reste abordable pour

de grands systèmes. C'est pourquoi je ne crois pas qu'il existe dans ce cadre d'autre perspective méthodologique réellement nouvelle, qui aille au delà de raffinements secondaires (tels que ceux présentés aux sections 6.3 à 6.5), et qui permette de traiter des problèmes radicalement plus complexes. Bref, ce n'est pas par cette voie que la malédiction des dimensions sera davantage circonscrite.

Une alternative générale à cette approche (déjà esquissée en section 5.3) consiste à rechercher directement la distribution de probabilité a posteriori, sans jamais produire de description explicite de la distribution de probabilité a priori (telle qu'on l'obtiendrait avec une simulation d'ensemble classique, du type de ce qui a été fait au chapitre 2). Un moyen assez générique d'y parvenir est de construire une chaîne de simulations, qui ne sont plus indépendantes les unes des autres, mais conditionnées séquentiellement les unes aux autres, ainsi qu'aux observations, et qui convergent vers une "certaine description" de la distribution de probabilité a posteriori (donnée par l'éq. 5.2). Cette idée est à la base des algorithmes MCMC (Markov Chain Monte Carlo algorithms) qui consistent à construire une chaîne de Markov dont les propriétés convergent vers la description que l'on recherche (Robert and Casella, 2004). Le plus célèbre et le plus important de ces algorithmes est certainement l'algorithme de Metropolis/Hastings (Metropolis et al., 1953; Hastings, 1970), qui procure un moyen générique et flexible d'échantillonner une distribution de probabilité (voir Robert and Casella, 2004, pour plus de détails). Afin d'introduire le sujet de façon plus progressive, on commencera cependant par illustrer deux algorithmes MCMC que nous avons commencé à utiliser dans nos travaux (et qui peuvent tous deux être vus comme une adaptation ou comme un cas particulier de l'algorithme de Metropolis/Hastings) : l'algorithme du refroidi simulé et l'échantillonneur de Gibbs. Et ce n'est qu'ensuite qu'on examinera une perspective d'application de l'algorithme de Metropolis/Hastings en repartant de sa formulation générique.

Algorithme du refroidi simulé. L'application qui a requis que nous nous tournions vers ce genre d'approche est celle qui est illustrée par l'ensemble de la figure 2.10. Le problème direct consiste à déduire les structures frontales à sous-mésoéchelle $\boldsymbol{\lambda}$ d'un traceur passif advecté par un champ de vitesse à mésoéchelle $\mathbf{u}$ (en utilisant un modèle basé sur le calcul des exposants de Lyapounov de l'advection de particules par le champ de vitesse). Et le problème inverse consiste à retrouver le champ de vitesse $\hat{\mathbf{x}} = \mathbf{u}$ à partir d'une observation des structures frontales $\mathbf{y} = \boldsymbol{\lambda}$ (en tenant compte des incertitudes sur le modèle et sur les observations). Or, un examen détaillé d'une simulation d'ensemble (telle qu'illustrée en fig. 2.10) montre assez rapidement qu'elle ne se conforme facilement à aucun des modèles proposés aux chapitres 6 et 7. Par ailleurs, le minimum de la fonction coût $J(\hat{\mathbf{x}}) = -\ln p^a(\hat{\mathbf{x}})$ ne peut pas non plus être facilement obtenu par un algorithme de descente (tel que décrit en section 5.3), en raison de la présence de nombreux minima locaux. Plusieurs minima locaux sont par exemple déjà visibles sur la coupe 2D de $J(\hat{\mathbf{x}})$ présentée en fig. 7.6 (à gauche). C'est donc pour surmonter ces difficultés que nous avons commencé à nous orienter vers des algorithmes MCMC.

Le premier algorithme que nous avons expérimenté pour cela est l'algorithme du refroidi simulé, qui consiste à construire une chaîne de Markov : $\hat{\mathbf{x}}^{(0)}$, $\hat{\mathbf{x}}^{(1)}, \ldots, \hat{\mathbf{x}}^{(k)}$ convergeant vers le minimum de $J(\hat{\mathbf{x}})$. L'algorithme [5] se résume à l'itération des étapes suivantes (Robert and Casella, 2004, section 5.2.3) :

1. Echantillonner $\hat{\mathbf{x}}$ dans la dsitribution utilitaire $q(\hat{\mathbf{x}}|\hat{\mathbf{x}}^{(k)}) = q(||\hat{\mathbf{x}} - \hat{\mathbf{x}}^{(k)}||)$;
2. Accepter $\hat{\mathbf{x}}^{(k+1)} = \hat{\mathbf{x}}$ avec la probabilité :

5. Cet algorithme du refroidi simulé peut être vu comme une application directe de l'algorithme de Metropolis/Hasting (voir page 90) pour échantillonner la distribution $p(\hat{\mathbf{x}}; T) \sim \exp[-J(\hat{\mathbf{x}})/T]$, de sorte que la chaîne de Markov convergera vers le mode de $p(\hat{\mathbf{x}}) \sim \exp[-J(\hat{\mathbf{x}})]$ quand $T \to 0$ (Robert and Casella, 2004, section 5.2.3).

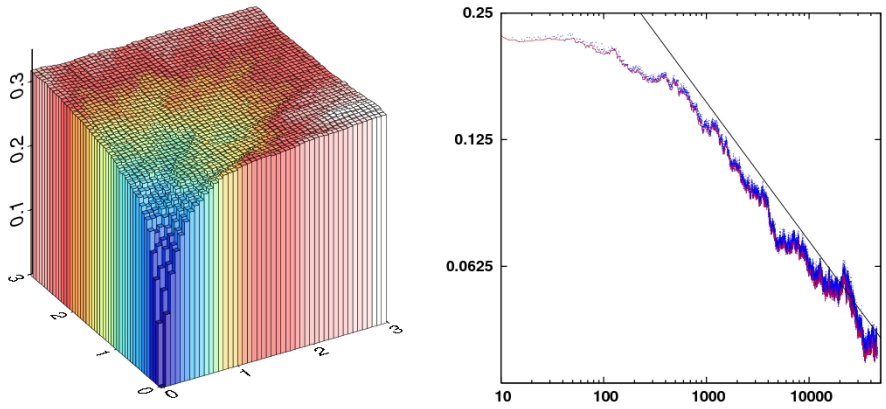


FIGURE 7.6 – Minimisation de la fonction coût $J(\hat{\mathbf{x}}) = -\ln p^a(\hat{\mathbf{x}})$ (illustrée par une coupe 2D à gauche) par l'algorithme du refroidi simulé (à droite).

$$\rho_k = \min\left[1,\ \exp\left(\frac{J(\hat{\mathbf{x}}^{(k)}) - J(\hat{\mathbf{x}})}{T^{(k)}}\right)\right] \tag{7.11}$$

Prendre $\hat{\mathbf{x}}^{(k+1)} = \hat{\mathbf{x}}^{(k)}$ sinon ;

3. Mettre à jour $T^{(k)}$ en $T^{(k+1)}$.

Pour notre application, le premier itéré $\hat{\mathbf{x}}^{(0)}$ a été choisi égal au champ de vitesse $\mathbf{u}^{(0)}$ obtenu à partir des observations altimétriques seules. Pour la distribution utilitaire $q(\hat{\mathbf{x}}|\hat{\mathbf{x}}^{(k)})$, nous avons choisi une gaussienne de moyenne $\hat{\mathbf{x}}^{(k)}$ et de covariance $\mathbf{B}$, proportionnelle à la covariance naturelle du champ de vitesse. Cette covariance a cependant été réduite à ses 50 à 100 premières EOFs (comme en section 6.2), réduisant ainsi le nombre de degré de liberté du problème de minimisation. Par ailleurs, la "température" $T^{(k)}$, réglant le taux d'acceptation des fluctuations positives de J, est diminuée progressivement au fur et à mesure que J décroît (voir Gaultier et al., 2013, pour plus de détails). Le réglage de ce taux d'acceptation des fluctuations positives de J est d'une importance cruciale dans l'algorithme car c'est lui qui permet de s'extirper des minima locaux, et de garantir la convergence de l'algorithme vers le minimum global de J (voir fig. 7.6, à droite).

Par son objectif (minimisation de J) et par la structure générale de l'algorithme (itérations sur le vecteur de contrôle), cet algorithme ressemble aux algorithmes variationnels de descente évoqués en section 5.3. Ne pourrait-on pas dès lors considérer, *à la limite*, qu'un algorithme variationnel (comme le 4DVAR) est une sorte de chaîne de Markov dont le comportement stochastique aurait été réduit jusqu'à zéro, car rendu inutile par un jeu d'hypothèses suffisantes (impliquant l'absence de minima locaux) ? Cela ne donnerait-il pas un angle d'attaque intéressant pour faire évoluer graduellement les méthodes variationnelles, en "réintroduisant" peu à peu diverses formes de comportement stochastique, par exemple d'abord dans la génération de l'itéré suivant sur la droite définie par le gradient de J, et ensuite dans la direction de descente elle-même (Robert and Casella, 2004, section 5.2.2) ? Une perspective encore plus radicale dans cette voie est esquissée plus loin (voir page 90).

Echantillonneur de Gibbs. Dans nos applications cependant, l'incertitude a posteriori est rarement négligeable et la recherche du mode de $p^a(\hat{\mathbf{x}})$ ne suffit généralement pas. Dans le cadre du problème illustré par la figure 7.6, nous avons donc cherché à caractériser cette incertitude par un échantillon de $p^a(\hat{\mathbf{x}})$, et le premier algorithme que nous

avons expériementé pour cela est l'échantillonneur de Gibbs (voir l'encadré 8, page 79). Cet algorithme [6] consiste encore à construire une chaîne de Markov : $\hat{\mathbf{x}}^{(0)}$, $\hat{\mathbf{x}}^{(1)}, \ldots, \hat{\mathbf{x}}^{(k)}$, mais qui converge cette fois vers un échantillon de $p^a(\hat{\mathbf{x}})$. Il se résumé à l'itération des n étapes suivantes (Robert and Casella, 2004, chap. 10) :

1. Echantillonner $\hat{x}_1^{(k+1)} \sim p^a(\hat{x}_1|\hat{x}_1^{(k)}, \ldots, \hat{x}_n^{(k)})$;
2. Echantillonner $\hat{x}_2^{(k+1)} \sim p^a(\hat{x}_2|\hat{x}_1^{(k+1)}, \hat{x}_2^{(k)}, \ldots, \hat{x}_n^{(k)})$;

...

n. Echantillonner $\hat{x}_n^{(k+1)} \sim p^a(\hat{x}_n|\hat{x}_1^{(k+1)}, \ldots, \hat{x}_{n-1}^{(k+1)})$.

L'intérêt de cet algorithme réside dans le fait que toutes les distributions que l'on échantillonne sont *univariées*, ce qui est en général un avantage. Pour notre application, le premier itéré $\hat{\mathbf{x}}^{(0)}$ a été choisi égal au champ de vitesse optimal obtenu par l'algorithme du refroidi simulé (voir fig. 7.6), de façon à être au cœur de la région de forte probabilité a posteriori, et chacune de distribution conditionnelle univariée est ensuite échantillonnée avec une méthode de rejet classique (voir Gaultier et al., 2013, pour plus de détails). La figure 7.7 illustre par exemple le genre d'échantillon que l'on obtient par quelques projections 2D le long de modes principaux de la matrice de covariance de l'incertitude a priori (**B**). Ces graphes montrent que le comportement central de l'échantillon est décalé par rapport à la vitesse d'ébauche (en 0), et que la dispersion de l'échantillon (c'est-à-dire l'incertitude a posteriori) n'est pas complètement négligeable par rapport à l'écart-type de l'incertitude a priori (normalisée à 1). En poursuivant le parallèle avec les méthodes variationnelles, cet algorithme pourrait aussi être utilisé exactement de la même façon pour compléter le calcul variationnel du minimum de J par un échantillon de $p^a(\hat{\mathbf{x}}) \sim \exp[-J(\hat{\mathbf{x}})]$, à condition toutefois que la taille du vecteur de contrôle ne soit pas trop grande (typiquement 50 à 100 comme dans notre problème).

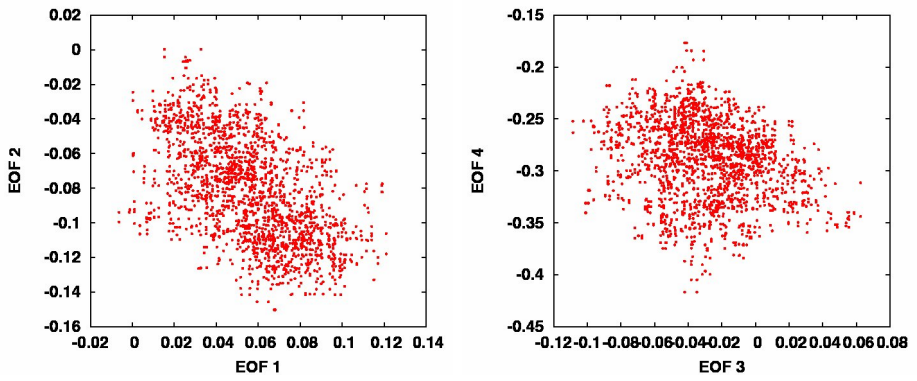


FIGURE 7.7 – Illustration de l'échantillonnage de $p^a(\hat{\mathbf{x}}) \sim \exp[-J(\hat{\mathbf{x}})]$ par échantillonneur de Gibbs. Les deux graphes représentent deux projections 2D de l'échantillon obtenu le long de modes principaux de la matrice de covariance de l'incertitude a priori (**B**).

Algorithme de Metropolis/Hastings. Pour aller plus loin, une perspective peut-être plus intéressante serait de repartir plutôt de l'algorithme de Metropolis/Hastings sous sa forme générique, qui consiste encore à construire une chaîne de Markov : $\hat{\mathbf{x}}^{(0)}$, $\hat{\mathbf{x}}^{(1)}, \ldots, \hat{\mathbf{x}}^{(k)}$, convergeant vers un échantillon d'une dsitribution cible, c'est-à-dire pour nous $p^a(\hat{\mathbf{x}})$. Cet algorithme se résume à l'itération des étapes suivantes (Robert and Casella, 2004, section 7.3) :

6. Il est possible de montrer (Robert and Casella, 2004, section 10.2.2) que l'échantillonneur de Gibbs est équivalent à la composition de n algorithmes de Metropolis/Hastings (voir page 90), dans lesquels les probabilités d'acceptation ρ sont toujours égales à 1.

Encadré 9 : Chaînes de Markov, évolution des espèces et progrès des théories

Le propos de cet encadré est de tenter un parallèle entre la solution d'un problème inverse par chaînes de Markov, l'évolution des espèces dans l'écosystème terrestre, et le progrès des théories scientifiques. On peut en effet voir la vie terrestre comme une machine de Monte Carlo, explorant stochastiquement le jeu des possibles (par mutations génétiques), en pénalisant les tirages défavorables (qui font croître la fonction coût). Par ailleurs, l'activité scientifique consiste aussi à générer des variations aux théories existantes, sans idée préconçue trop rigide, en pénalisant les moins performantes (moins simples, moins cohérentes avec les observations). En règle générale, cela conduit à un mouvement lent de l'histoire vers un écosystème plus efficace et des théories plus performantes (vers le minimum local de la fonction coût le plus proche du point de départ).

Mais cela n'est pas suffisant pour expliquer le fonctionnement de ces deux systèmes. Dans les deux cas, les processus en jeu autorisent l'apparition de fluctuations en sens contraire, qui sont perçues comme monstrueuses ou révolutionnaires. Dans l'histoire de l'évolution, ce sont les innovations provisoirement défavorables aux individus qui les portent ("monstres prometteurs"), ou les épisodes d'extinction cataclysmiques (dus à des fluctuations climatiques ou à la chute d'objets célestes). Dans l'histoire des sciences, ce sont les idées non-conventionnelles, provisoirement incapables d'expliquer aussi bien les données observées, ou les nouveaux paradigmes qui restent longtemps marginaux, avant de remplacer d'un seul coup le modèle standard. A long terme, ce sont ces événenements qui forment la trame générale de l'histoire de la vie (théorie des équilibre ponctués, Gould and Eldredge, 1977), ou de l'histoire des sciences (la structure des révolutions scientifiques, Kuhn, 1983).

En poursuivant l'analogie avec la résolution d'un problème inverse, on peut voir ces événements comme des passages de cols (à n dimensions) dans le processus de minimisation de la fonction coût, qui permettent de poursuivre la descente dans un bassin versant inexploré et d'atteindre le voisinage d'un nouveau minimum local. Au delà de l'exemple de l'eau ruisselante qui reste piégée en altitude par la loi de la plus grande pente (voir illustration ci-dessus), la Vie et la Science résolvent leur problème inverse d'une façon plus sophistiquée. Topologiquement, plus que la descente vers les minima, ce sont les passages de cols qui sont les moments clés du processus, en ouvrant de nouvelles pistes et en donnant un point de vue nouveau sur la diversité des combinaisons possibles.

1. Echantillonner $\hat{\mathbf{x}}$ dans la dsitribution utilitaire $q(\hat{\mathbf{x}}|\hat{\mathbf{x}}^{(k)})$;
2. Accepter $\hat{\mathbf{x}}^{(k+1)} = \hat{\mathbf{x}}$ avec la probabilité :

$$\rho_k = \min\left[1,\ \frac{p^a(\hat{\mathbf{x}})}{p^a(\hat{\mathbf{x}}^{(k)})}\ \frac{q(\hat{\mathbf{x}}^{(k)}|\hat{\mathbf{x}})}{q(\hat{\mathbf{x}}|\hat{\mathbf{x}}^{(k)})}\right] \tag{7.12}$$

Prendre $\hat{\mathbf{x}}^{(k+1)} = \hat{\mathbf{x}}^{(k)}$ sinon.

L'intérêt de cet algorithme est sa très grande généralité et sa grande souplesse, puisqu'il converge vers un échantillon de $p^a(\hat{\mathbf{x}})$ indépendamment du choix de la distribution utilitaire q (appelée "instrumental distribution" ou "proposal distribution"), ou du moins sous des conditions très peu contraignantes. Mais dès lors, toute la difficulté est de choisir une distribution q qui rende l'algorithme efficace, et cela exige au moins (i) qu'elle puisse être échantillonnée facilement, (ii) que la probabilité d'acceptation ρ soit facilement calculable, et (iii) que le taux de rejet $(1-\rho)$ soit le plus faible possible. Considérons par exemple l'algorithme du refroidi simulé décrit plus haut, mais avec $T^{(k)} = 1\ \forall k$, pour obtenir un échantillon de $p^a(\hat{\mathbf{x}})$ (et non pas $T^{(k)} \to 0$ pour trouver le maximum). Il vérifie aisément les deux premières conditions : (i) par exemple si q est choisi gaussien, et (ii) puisque $p^a(\hat{\mathbf{x}}) \sim \exp[-J(\hat{\mathbf{x}})]$ et $q(\hat{\mathbf{x}}^{(k)}|\hat{\mathbf{x}}) = q(\hat{\mathbf{x}}|\hat{\mathbf{x}}^{(k)})$. Mais pas la troisième, car quand le nombre de dimensions est grand, le nombre de directions de montée (à distance finie) de J peut devenir localement beaucoup plus grand que le nombre de directions de descente. Dans ce cas, il arrive fréquemment que $\rho \ll 1$, et l'algorithme requerra de nombreuses tentatives infructueuses avant d'accepter l'itéré suivant. Le nombre d'itérations nécessaires est d'ailleurs la difficulté principale que nous ayons rencontrée dans notre application (voir fig. 7.6, à droite).

Il est certainement possible d'imaginer de nombreuses approches pour définir la distribution utilitaire $q(\hat{\mathbf{x}}|\hat{\mathbf{x}}^{(k)})$. Pour pouvoir faire un bon choix, une question qui me paraît importante serait de savoir dans quelle mesure le gradient de la fonction coût ∇J (déjà disponible à faible coût pour les méthodes variationnelles) ne pourrait être utile pour définir $q(\hat{\mathbf{x}}|\hat{\mathbf{x}}^{(k)})$ et rendre l'algorithme plus efficace. Par exemple, une méthode de descente classique calculerait l'itéré suivant par une formule dur type :

$$\tilde{\mathbf{x}}(\hat{\mathbf{x}}^{(k)}) = \hat{\mathbf{x}}^{(k)} - \alpha_k \nabla J(\hat{\mathbf{x}}^{(k)}) \tag{7.13}$$

Mais plutôt que de choisir simplement $\hat{\mathbf{x}}^{(k+1)} = \tilde{\mathbf{x}}(\hat{\mathbf{x}}^{(k)})$, ne pourrait-on pas utiliser $\tilde{\mathbf{x}}(\hat{\mathbf{x}}^{(k)})$ pour définir $q(\hat{\mathbf{x}}|\hat{\mathbf{x}}^{(k)})$? Une première idée serait par exemple de choisir $q(\hat{\mathbf{x}}|\hat{\mathbf{x}}^{(k)})$ gaussienne, de moyenne $\tilde{\mathbf{x}}(\hat{\mathbf{x}}^{(k)})$ et de covariance $\beta_k\mathbf{B}$, proportionnelle à la covariance d'erreur d'ébauche, et de chosir le coefficient β_k pour que la variance associée à $q(\hat{\mathbf{x}}|\hat{\mathbf{x}}^{(k)})$ dans la direction du gradient soit égale ou proportionnelle à $||\tilde{\mathbf{x}} - \hat{\mathbf{x}}^{(k)}||/||\nabla J||$. Structurellement, un tel algorithme serait très semblable aux algorithmes variationnels actuellement utilisés en océanographie (Weaver et al., 2003; Bouttier et al., 2012). Il suffirait de compléter le calcul de $\tilde{\mathbf{x}}$ (éq. 7.13) par un tirage aléatoire dans la distribution $q(\hat{\mathbf{x}}|\hat{\mathbf{x}}^{(k)})$, et de l'accepter avec la probabilité ρ. Les deux intérêts d'un tel algorithme seraient en principe (i) de produire une description de l'incertitude a posteriori (par un échantillon), et (ii) de ne plus requérir d'hypothèse de convexité de la fonction coût (en tolérant la présence de minima locaux). Cependant, sans avoir essayé, il est difficile de savoir si ce genre d'algorithme pourraient être rendu assez efficace pour les applications océanographiques. La probabilité ρ serait-elle suffisamment élevée pour éviter un trop grand nombre de rejet (nécessitant chacun l'évaluation de J et de ∇J) ? Combien d'itérations faudrait-il consentir pour donner une solution raisonnable au problème posé ?

Conclusion

> Error is viewed, therefore,
> not as an extraneous and
> misdirected or misdirecting accident,
> but as an essential part
> of the process under consideration.
>
> John von Neuman (1956)

Le fil conducteur de ce mémoire était de montrer combien un traitement approprié des incertitudes est nécessaire au bon fonctionnement de la plupart des méthodes utilisées en océanographie, et de suggérer pourquoi le concept d'incertitude devra progressivement devenir un élément central, et non plus périphérique, de la description des systèmes océaniques.

- Le chapitre 1, tout d'abord, a montré que l'incertitude est une caractéristique essentielle et inévitable de tout modèle d'océan. En raison surtout de ce que le modèle ne résout pas, toute prévision de l'océan ne peut donc être que probabiliste. Et le concept d'incertitude devrait progressivement remplacer le concept d'erreur de modélisation dans la plupart de nos méthodes et de nos applications.
- Le chapitre 2 a montré comment réaliser une prévision probabiliste de l'océan par une simulation d'ensemble, caractérisant notre incertitude sur la prévision. Cette méthode probabiliste devrait se généraliser dans l'avenir car elle est nécessaire pour donner une perception correcte de la dynamique du système (événements extrêmes, probabilité de bifurcations,...) et pour élargir le spectre des applications (diagnostic non-linéaire complexe, aide à la décision,...).
- Le chapitre 3 a montré combien la simulation explicite des incertitudes sur les lois dynamiques elles-mêmes peuvent influer sur le comportement du modèle, que ce soit dans le modèle de circulation, le modèle d'écosystème, ou le modèle de glace de mer. Comme en météorologie, la paramétrisation stochastique des incertitudes sur le modèle devrait jouer un rôle de plus en plus important en océanographie, que ce soit pour améliorer la fiabilité des prévisions d'ensemble (cohérence avec les observations), ou pour améliorer le traitement des incertitudes de modélisation dans les systèmes d'assimilation de données.
- Le chapitre 4, ensuite, a montré par quels moyens il est possible de prendre en compte l'incertitude sur les observations, qu'elle résulte de l'instrument utilisé (erreur de mesure) ou bien de ce que le modèle ne résout pas (erreur ou incertitude de représentativité). Ce chapitre a rappelé aussi que l'approche probabiliste est nécessaire à toute méthode permettant de tester la cohérence entre modèle et observations, et il suggère, en s'inspirant du chapitre 3, que des paramétrisations stochastiques spécifiques pourraient être un bon moyen de traiter les incertitudes sur les observations (dues aux échelles ou à la diversité non-résolues).
- Le chapitre 5 a introduit une formulation générale du problème inverse fondée sur l'approche bayésienne (en incluant un contrôle sur la condition initiale, les pa-

ramètres, les forçages déterministes et stochastiques), et a montré que l'obstacle le plus important à la réduction des incertitudes par l'observation est l'effet combinatoire lié au grand nombre de variables de contrôle (malédiction des dimensions). C'est cela qui nous contraint à baser nos méthodes sur des hypothèses restrictives (unimodalité,...) ou des modèles simplificateurs (gaussianité,...), et c'est cela qui limitera encore dans l'avenir la gamme des problèmes inverses que nous sommes capable de résoudre.

– Le chapitre 6 a montré comment le modèle gaussien peut être appliqué à la résolution de problèmes d'assimilation de données en océanographie, et combien pour cela une gestion fine des incertitudes est toujours nécessaire (découpage en fenêtres d'assimilation suffisamment courtes pour que le modèle gaussienn reste valide, réduction d'ordre pour limiter le coût numérique, localisation des covariances pour compenser une réduction d'ordre excessive, schéma adaptatif pour compenser les imperfections du modèle gaussien, gestion de la dépendance entre les observations,...). On voit bien néanmoins que cette approche a atteint sa phase de maturité, et que la plupart des efforts récents ne portent plus que sur des raffinements et des réglages à la marge. Ces efforts de développement spécifiques (souvent d'une très grande importance pratique) sont donc appelés à migrer des équipes de recherche vers les systèmes opérationnels.
– Le chapitre 7, finalement, a décrit nos efforts récents pour dépasser le cadre du modèle gaussien (gaussiennes tronquées, anamorphose, mélange de gaussiennes), en montrant l'importance du bénéfice que cela peut générer pour nombre d'applications océanographiques. Il a cependant été suggéré que, pour aller plus loin et circonscrire encore davantage la malédiction des dimensions, le calcul de chaînes de Markov constituait peut-être l'approche méthodologique la plus prometteuse. D'un point de vue technique et algorithmique, elle pourrait se construire progressivement comme un prolongement naturel des méthodes variationnelles existantes, tout en accentuant peu à peu l'importance d'une description probabiliste des incertitudes.

En résumé, l'incertitude est une caractéristique essentielle des systèmes que nous étudions, aussi bien des systèmes océaniques réels (au sens de l'encadré 3, page 16) que des systèmes informatiques utilisés pour les décrire. Elle joue un rôle majeur dans tous les composants du système (circulation, écosystème, glace de mer) et à toutes les échelles de temps (climatique, saisonnière, moyen terme,...). Elle doit donc modifier notre compréhension des systèmes marins, et influencer l'élaboration de nos méthodes tant en modélisation qu'en assimilation de données. Quelques pistes dans cette direction ont d'ores et déjà été proposées, mais la meilleure façon d'incorporer une description de l'incertitude dans nos théories et dans nos méthodes reste une question ouverte.

Le point de vue peut-être le plus pertinent pour approfondir cette question me semble être celui de l'ingénieur (et même plus précisément celui de l'ingénieur-sytème, au sens que lui donne Richard Hamming, voir annexe A). En soutien à ce point de vue, il convient d'abord de se rappeler que c'est pour résoudre une problème d'ingéniérie que, par deux fois, la notion d'incertitude s'est invitée dans l'histoire des sciences : (i) pour l'optimisation des machines thermiques (Carnot, 1824), ce qui a conduit au concept d'entropie en physique, et (ii) pour l'optimisation des réseaux de télécommunications (Shannon, 1948), ce qui a conduit à la théorie de l'information. Ne serait-il pas alors assez naturel que, pour des raisons similaires, la rationalisation du concept d'incertitude finisse par devenir incontournable pour optimiser les systèmes informatiques qui nous servent à anticiper l'évolution du climat, de l'atmosphère ou de l'océan ? Néanmoins, avant d'en arriver là, un ingénieur-système (voir discussion en annexe A) se doit de concevoir un programme de développement progressif et pragmatique :

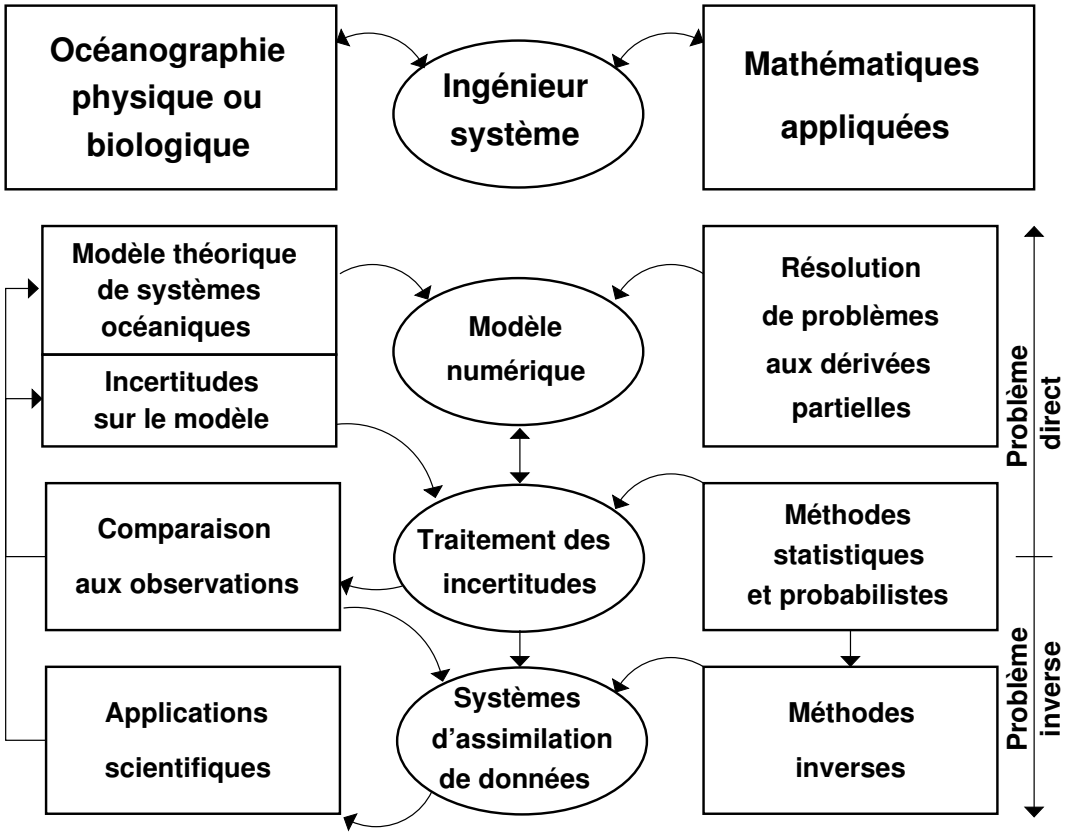


FIGURE 8.1 – Représentation schématique du rôle de l'ingénieur-système à l'interface entre recherche océanographique et mathématiques appliquées.

- Le traitement des incertitudes est tout d'abord nécessaire pour rendre le système techniquement plus performant et plus robuste. Un système de prévision océanique est un système automatique complexe, à la jonction de nombreuses disciplines scientifiques et techniques, dont le fonctionnement dépend d'un grand nombre de composants en interaction. Pour des raisons purement techniques, il ne peut être conçu comme le simple assemblage des meilleurs composants possibles, car les propriétés émergentes du système complexe (sur lesquelles reposent sa performance et sa robustesse) peut dépendre considérablement des sources résiduelles d'incertitude, qu'il est donc nécessaire de comprendre et de simuler.
- Le traitement des incertitudes doit être réalisé de façon cohérente à travers tout le système. Pour cette raison, le problème direct (modèles d'océan) et le problème inverse (assimilation de données) ne peuvent pas être envisagés séparément l'un de l'autre. Il est nécessaire que les modèles soient conçus et dimensionnés pour exprimer l'incertitude qu'ils génèrent (par exemple par une simulation d'ensemble), et donc permettre la résolution de problèmes inverses. Il est nécessaire que les méthodes d'assimilation de données soient conçues pour respecter scrupuleusement la contrainte dynamique exprimée par le modèle, par exemple en ne contrôlant que les paramètres ou le forçage (extérieur et stochastique) à l'intérieur de leur barre d'erreur. C'est une perspective certes encore lointaine et ambitieuse en océanographie, mais elle est réalisable, à condition toutefois de parvenir à un consensus suffisant sur le traitement des incertitudes (comme en thermodynamique ou en ingéniérie des télécommunications).
- Le traitement des incertitudes requiert d'être exploré par les laboratoires de recherche et appliqué par les centres opérationnels. L'exploration devrait se faire sur

des systèmes de taille et de coût modérés, qui permettent de comprendre ce qui se passe par un traitement statistique rigoureux (par exemple avec de grands ensembles), mais de complexité comparable aux applications réalistes. L'application devrait se faire en ajustant les méthodes génériques aux contraintes opérationnelles dictées par les besoins de la société (justifiant le coût du système). L'interaction entre les deux univers peut alors s'opérer si d'une part, les laboratoires de recherche s'impliquent pour tester leurs idées sur des applications opérationnelles, et si d'autre part, les centres opérationnels s'impliquent pour maintenir leur système ouvert à de nouvelles idées.

En conclusion, ce que j'ai surtout essayé de faire émerger de ce tableau, c'est la place et le rôle d'ingénieurs-systèmes, spécifiquement océanographes, y compris dans les laboratoires de recherche, pour piloter le développement des systèmes informatiques complexes dont la recherche en océanographie a désormais constamment besoin. C'est du moins ainsi que je conçois mon activité d'ingénieur de recherche au sein de notre équipe de modélisation des écoulements océaniques multi-échelles (MEOM). De par leur nature, ces développements techniques se situent également à l'interface entre océanographie et mathématiques appliquées (voir figure 8.1), et le traitement de l'incertitude me paraît y jouer un rôle charnière entre modélisation, observations et assimilation de données. C'est pourquoi il est utile que l'élaboration du système puisse s'organiser selon une logique et une direction autonomes, qui soient à la fois techniques, méthodologiques et scientifiques, afin de statisfaire au mieux les besoins de la recherche en océanographie.

Bibliographie

Anderson, J. L. and S. L. Anderson, 1999 : A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Monthly Weather Review*, **127**, 2741–2758.

Anscombe, F. J., 1973 : Graphs in statistical analysis. *American Statistician*, **27(1)**, 17–21.

Barnier, B., et al., 2006 : Impact of partial steps and momentum advection schemes in a global ocean circulation model at eddy permitting resolution. *Ocean Dynamics*, **56(5-6)**, 543–567.

Béal, D., P. Brasseur, J.-M. Brankart, Y. Ourmières, and J. Verron, 2010 : Characterization of mixing errors in a coupled physical biogeochemical model of the North Atlantic : implications for nonlinear estimation using gaussian anamorphosis. *Ocean Science*, **6**, 247–262.

Beckers, J., et al., 2002 : Model intercomparison in the mediterranean : Medmex simulations of the seasonal cycle. *Journal of Marine Systems*, **33–34**, 215–251.

Bengtsson, T., C. Snyder, and D. Nychka, 2003 : Toward a nonlinear ensemble filter for high-dimensional systems. *Journal of Geophysical Research*, **108**, 8775.

Bennett, A. F., 1992 : *Inverse Methods in Physical Oceanography*. Cambridge University Press, 346 pp.

Berline, L., 2006 : Assimilation de données dans un modèle couplé physique-biogéochimie de l'océan atlantique nord. Ph.D. thesis, Université Joseph Fourier (Grenoble).

Berline, L., J.-M. Brankart, P. Brasseur, Y. Ourmières, and J. Verron, 2007 : Improving the dynamics of a coupled physical-biogeochemical model of the north atlantic basin through data assimilation : impact on biological tracers. *Journal of Marine Systems*, **64(1-4)**, 153–172.

Bertino, L., G. Evensen, and H. Wackernagel, 2003 : Sequential data assimilation techniques in oceanography. *International Statistical Review*, **71**, 223–241.

Birol, F., J.-M. Brankart, F. Castruccio, P. Brasseur, and J. Verron, 2004 : Impact of ocean mean dynamic topography on satellite data assimilation. *Journal of Marine Geodesy*, **27**, 59–78.

Birol, F., J.-M. Brankart, J.-M. Lemoine, P. Brasseur, and J. Verron, 2005 : Assimilation of satellite altimetry referenced to the new grace geoid estimate. *Geophysical Research Letters*, **32(6)**, doi10.1029/2004GL021329.

Bishop, C. H., B. J. Etherton, and S. J. Majumdar, 2001 : Adaptive sampling with the ensemble transform Kalman filter. Part i : theoretical aspects. *Monthly Weather Review*, **129**, 420–436.

Bishop, C. H. and D. Hodyss, 2009 : Ensemble covariances adaptively localized with ECO-RAP. Part 2 : a strategy for the atmosphere. *Tellus*, **61A**, 97–111.

Blanke, B. and P. Delecluse, 1993 : Low frequency variability of the tropical atlantic ocean simulated by a general circulation model with mixed layer physics. *Journal of Physical Oceanography*, **23**, 1363–1388.

Bleck, R., 2002 : An oceanic general circulation model framed in hybrid isopycnic-cartesian coordinates. *Ocean Modelling*, **4**, 55–88.

Bouttier, P.-A., E. Blayo, J.-M. Brankart, B. P., E. Cosme, J. Verron, and A. Vidard, 2012 : Toward a data assimilation system for nemo. *Mercator Ocean Quarterly Newsletter*, **46**, 24–30.

Brankart, J.-M., 1996 : Modélisation statistique de l'hydrologie méditerranéenne. validation et contrôle de qualité d'une climatologie de référence. Ph.D. thesis, Université de Liège, 209 pp.

Brankart, J.-M., 2013 : Impact of uncertainties in the horizontal denisty gradient upon low resolution global ocean modelling. *Ocean Modeling*, **66**, 64–76.

Brankart, J.-M. and P. Brasseur, 1996 : Optimal analysis of in situ data in the Western Mediterranean using statistics and cross-validation. *Journal of Atmospheric and Oceanic Technology*, **16 (2)**, 477–491.

Brankart, J.-M. and P. Brasseur, 1998 : The general circulation in the Mediterranean Sea : a climatological approach. *Journal of Marine Systems*, **18**, 41–70.

Brankart, J.-M., P. Brasseur, and J. Verron, 2001 : Quel sera le courant jeudi prochain dans l'atlantique nord ? a l'aube de l'océanographie opérationnelle. *CNRS Info, Lettre d'information destinée aux médias*, **389**.

Brankart, J.-M., E. Cosme, C.-E. Testut, P. Brasseur, and J. Verron, 2010 : Efficient adaptive error parameterizations for square root or ensemble kalman filters : application to the control of ocean mesoscale signals. *Monthly Weather Review*, **138 (3)**, 932–950.

Brankart, J.-M., E. Cosme, C.-E. Testut, P. Brasseur, and J. Verron, 2011 : Efficient local error parameterizations for square root or ensemble kalman filters : application to a basin-scale ocean turbulent flow. *Monthly Weather Review*, **139 (2)**, 474–493.

Brankart, J.-M. and N. Pinardi, 2001 : Abrupt cooling of the Mediterranean Levantine Intermediate Water at the beginning of the 1980s : observational evidence and model simulation. *Journal of Physical Oceanography*, **31 (8)**, 2307–2320.

Brankart, J.-M., C.-E. Testut, P. Brasseur, and J. Verron, 2003 : Implementation of a multivariate data assimilation scheme for isopycnic coordinate ocean models : Application to a 1993–96 hindcast of the North Atlantic Ocean circulation. *Journal of Geophysical Research*, **108 (C3)**, 19(1–20).

Brankart, J.-M., C.-E. Testut, D. Béal, M. Doron, C. Fontana, M. Meinvielle, P. Brasseur, and J. Verron, 2012 : Towards an improved description of ocean uncertainties : effect of local anamorphic transformations on spatial correlations. *Ocean Science*, **8**, 121–142.

Brankart, J.-M., C. Ubelmann, C.-E. Testut, E. Cosme, P. Brasseur, and J. Verron, 2009 : Efficient parameterization of the observation error covariance matrix for square root or ensemble kalman filters : application to ocean altimetry. *Monthly Weather Review*, **137 (6)**, 1908–1927.

Brasseur, P., 1991 : A variational inverse method for the reconstruction of general circulation fields in the Northern Bering Sea. *Journal of Geophysical Research*, **96 (C3)**, 4891–4907.

Brasseur, P., 1994 : Reconstitution de champs d'observations océanographiques par le modèle variationnel inverse : Methodologie et applications. Ph.D. thesis, Université de Liège.

Brasseur, P., J. Ballabrera, and J. Verron, 1999 : Assimilation of altimetric data in the mid-latitude oceans using the SEEK filter with an eddy-resolving primitive equation model. *Journal of Marine Systems*, **22**, 269–294.

Brasseur, P., J.-M. Beckers, J.-M. Brankart, and R. Schoenauen, 1996 : Seasonal temperature and salinity fields in the Mediterranean Sea : Climatological analyses of an historical data set. *Deep-Sea Research*, **43 (2)**, 159–192.

Brasseur, P. and J. Verron, 2006 : The SEEK filter method for data assimilation in oceanography : a synthesis. *Ocean Dynamics*, **56 (12)**, 650–661.

Bretherton, F. P., R. E. Davis, and C. B. Fandry, 1976 : A technique for objective analysis and design of oceanographic experiment applied to MODE-73. *Deep-Sea Research*, **23**, 559–582.

Brier, G. W., 1950 : Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, **78**, 1–3.

Broquet, G., 2006 : Caractérisation des erreurs de modélisation pour l'assimilation de données dans un modèle océanique régional du golfe de gascogne. Ph.D. thesis, Université Joseph Fourier (Grenoble).

Broquet, G., P. Brasseur, G. Rozier, J.-M. Brankart, and J. Verron, 2008 : Estimation of model errors generated by atmospheric forcings for ocean data assimilation : experiments in a regional model of the Bay of Biscay. *Ocean Dynamics*, **58 (1)**, 1–17.

Brusdal, K., J. Brankart, G. Halberstadt, G. Evensen, P. Brasseur, P. van Leeuwen, E. Dombrowsky, and J. Verron, 2003 : A demonstration of ensemble-based assimilation methods with a layered OGCM from the perspective of operational ocean forecasting systems. *Journal of Marine Systems*, **40–41**, 253–289.

Buizza, R., M. Miller, and T. N. Palmer, 1999 : Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quaterly Journal of the Royal Meteorological Society*, **125**, 2887–2908.

Burgers, G., P. J. van Leeuwen, and G. Evensen, 1998 : Analysis scheme in the ensemble Kalman filter. *Monthly Weather Review*, **126**, 1719–1724.

Candille, G. and O. Talagrand, 2005 : Evaluation of probabilistic prediction systems for a scalar variable. *Quaterly Journal of the Royal Meteorological Society*, **131**, 2131–2150.

Carmillet, V., J.-M. Brankart, P. Brasseur, H. Drange, G. Evensen, and J. Verron, 2001 : A singular evolutive extended Kalman filter to assimilate ocean color data in a coupled physical-biochemical model of the North Atlantic. *Ocean Modeling*, **3**, 167–192.

Carnot, S., 1824 : *Réflexions sur la puissance motrice du feu et sur les machines propres à développer cette puissance*. Paris, Bachelier libraire, 39 pp.

Castruccio, F., 2006 : Apports des données gravimétriques grace pour l'assimilation de données altimétriques et in-situ dans un modèle de l'océan pacifique tropical. Ph.D. thesis, Université Joseph Fourier (Grenoble).

Castruccio, F., J. Verron, L. Gourdeau, J. Brankart, and P. Brasseur, 2006 : On the role of the grace mission in the joint assimilation of altimetric and tao data in a tropical pacific ocean model. *Geophysical Research Letters*, **33**, L14 616, doi :10.1029/2006GL025823.

Castruccio, F., J. Verron, L. Gourdeau, J. Brankart, and P. Brasseur, 2008 : Joint altimetric and in-situ data assimilation using the grace mean dynamic topography : a 1993-1998 hindcast experiment in the tropical pacific ocean. *Ocean dynamics*, **58 (1)**, 43–63.

Chassignet, E., L. Smith, G. Halliwell, and B. R., 2003 : North Atlantic simulations with the HYbrid Coordinate Ocean Model (HYCOM) : Impact of the vertical coordinate choice, reference pressure, and thermobaricity. *J. Phys. Oceanogr.*, **33**, 2504–2526.

Claerbout, J. F. and F. Muir, 1973 : Robust modelling with erratic data. *Geophys.*, **38**, 826–844.

Cohn, S. E., 1997 : An introduction to estimation theory. *Journal of the Meteorological Society of Japan*, **75**, 257–288.

Corder, G. W. and D. I. Foreman, 2009 : *Nonparametric Statistics for Non-Statisticians : A Step-by-Step Approach*. Wiley, 264 pp.

Cosme, E., J.-M. Brankart, J. Verron, P. Brasseur, and M. Krysta, 2010 : Implementation of a reduced rank, square-root smoother for high resolution ocean data assimilation. *Ocean Modelling*, **33**, 87–100.

Cosme, E., J. Verron, P. Brasseur, J. Blum, and D. Auroux, 2012 : Smoothing problems in a bayesian framework and their linear gaussian solutions. *Monthly Weather Review*, **140**, 683–695.

Cover, T. M. and J. A. Thomas, 2006 : *Elements of information theory*. Wiley, 748 pp.

Debost, F., 2004 : Etude de nouveaux scénarios d'altimétrie satellitaire pour la reconstruction de la circulation océanique moyenne échelle par assimilation de données altimétriques. Ph.D. thesis, Université Joseph Fourier (Grenoble).

Dee, D., 1995 : On-line estimation of error covariance parameters for atmospheric data assimilation. *Monthly Weather Review*, **123**, 1128–1145.

Doron, M., P. Brasseur, and J.-M. Brankart, 2011 : Estimation of biogeochemical parameters of a 3d ocean coupled physical-biogeochemical model with a stochastic data assimilation method : twin experiments. *Journal of Marine Systems*, **87**, 194–207.

Doron, M., P. Brasseur, J.-M. Brankart, S. Loza, and A. Melet, 2013 : Stochastic estimation of biogeochemical parameters from globcolour ocean color satellite data in a north atlantic 3d ocean coupled physical-biogeochemical model. *Journal of Marine Systems*.

d'Ovidio, F., J. Isern-Fontanet, C. López, E. Hernández-Garcia, and E. Garcí a Ladona, 2008 : Comparison between the Okubo-Weiss parameter and finite-size Lyapunov exponents computed from altimetry in the Algerian basin. *Journal of Geophysical Research*.

Duchez, A., 2011 : Contrôle du courant nord méditerranéen dans le golfe du lion : une approche par simulation du système d'observation. Ph.D. thesis, Université Joseph Fourier (Grenoble).

Duchez, A., J. Verron, J.-M. Brankart, Y. Ourmières, and P. Fraunié, 2012 : Monitoring the northern current in the gulf of lions with an observing system simulation experiment. *Scientia Marina*, **76(3)**, 441–453.

Evensen, G., 1994 : Sequential data assimilation with a non linear quasigeostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research*, **99,(C5)**, 10 143–10 162.

Ferry, N., L. Parent, G. Garric, B. Barnier, N. Jourdain, and the Mercator Ocean team, 2010 : Mercator global eddy permitting ocean reanalysis GLORYS1V1 : Description and results. *Mercator Ocean Newsletter*, **36**, 15–27.

Fontana, C., P. Brasseur, and J.-M. Brankart, 2013 : Toward a multivariate reanalysis of the north atlantic ocean biogeochemistry during 1998-2006 based on the assimilation of seawifs chlorophyll data. *Ocean Science*, **9**, 37–56.

Frederiksen, J., T. O'Kane, and M. Zidikheri, 2012 : Stochastic subgrid parameterizations for atmospheric and oceanic flows. *Physica Scripta*, **85**, 068 202, doi :10.1088/0031-8949/85/06/068202.

Freychet, N., 2012 : Assimilation rétrospective de données par lissage de rang réduit : application et évaluation dans l'atlantique tropical. Ph.D. thesis, Université Joseph Fourier (Grenoble).

Freychet, N., E. Cosme, P. Brasseur, J.-M. Brankart, and E. Kpemlie, 2012 : Obstacles and benefits of the implementation of a reduced rank smoother with a high resolution model of the atlantic ocean. *Ocean Science*, **8**, 797–811.

Gaultier, L., J. Verron, J.-M. Brankart, O. Titaud, and P. Brasseur, 2013 : On the inversion of submesoscale tracer fields to estimate the ocean surface circulation. *Journal of Marine Systems*.

Gelb, A., 1974 : *Applied optimal estimation*. The MIT press, 374 pp.

Geman, S. and D. Geman, 1984 : Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6 (6)**, 721–741.

Geweke, J., 1991 : Efficient simulation from the multivariate normal and Student t-distributions subject to linear constraints. *Computer Sciences and Statistics Proceedings of the 23d Symposium on the Interface*, 571–578.

Gillijns, S., D. Bernstein, and B. De Moor, 2006 : The reduced rank transform square root filter for data assimilation. *Proceedings of the 14th IFAC Symposium on System Identification (Newcastle, Australia)*, 6 pp.

Gneiting, T., L. I. Stanberry, E. P. Grimit, L. Held, and N. A. Johnson, 2008 : Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. Tech. Report 537, Univ. Washington, Dpt. Statistics,.

Gordon, N. J., D. J. Salmond, and A. F. M. Smith, 1993 : Novel approach to nonlinear/non-gaussian bayesian state estimation. *IEEE proc.-F*, **140**, 107–113.

Gould, S. and N. Eldredge, 1977 : Punctuated equilibria : the tempo and mode of evolution reconsidered. *Paleobiology*, **3(2)**, 115–151.

Greatbatch, R., J. Sheng, C. Eden, L. Tang, X. Zhai, and J. Zhao, 2004 : The semi-prognostic method. *Continental Shelf Research*, **24**, 2149–2165.

Greatbatch, R. and X. Zhai, 2006 : Influence of assimilated eddies on the large-scale circulation in a model of the northwest atlantic ocean. *Geophysical Research Letters*, **33**, L02 614.

Griffiths, W., 2006 : A Gibbs' sampler for the parameters of a truncated multivariate normal distribution. Ph.D. thesis, University of Melbourne.

Hamming, R. W., 1997 : *The Art of Doing Science and Engineering.* Gordon and Breach science publishers, 364 pp.

Hastings, W., 1970 : Monte carlo sampling methods using markov chains and their applications. *Biometrika*, **57(1)**, 97–109.

Hollander, M. and D. A. Wolfe, 1973 : *Nonparametric statistical methods.* Wiley, 787 pp.

Hoteit, I., D. Pham, G. Triantafyllou, and G. Korres, 2008 : A new approximate solution of the optimal nonlinear filter for data assimilation in meteorology and oceanography. *Monthly Weather Review*, **136**, 317–334.

Hoteit, I., D.-T. Pham, and J. Blum, 2002 : A simplified reduced order Kalman filtering and application to altimetric data assimilation in tropical Pacific. *Journal of Marine Systems*, **36**, 101–127.

Houtekamer, P. L. and H. L. Mitchell, 1998 : Data assimilation using an Ensemble Kalman Filter technique. *Monthly Weather Review*, **126**, 796–811.

Hunt, B., E. Kostelich, and I. Szunyogh, 2007 : Efficient data assimilation for spatiotemporal chaos : A local ensemble transform Kalman filter. *Physica D*, **230**, 112–126.

Izenman, A. J., 2008 : *Modern Multivariate Statistical Techniques : Regression, Classification, and Manifold Learning.* Springer, 734 pp.

Jackett, D. R. and T. J. McDougall, 1995 : Minimal adjustment of hydrographic data to achieve static stability. *Journal of Atmospheric and Oceanic Technology*, **12**, 381–389.

Juricke, S., P. Lemke, R. Timmermann, and T. Rackow, 2013 : Effects of stochastic ice strength perturbation on arctic finite element sea ice modeling. *Journal of Climate*, **26**, 3785–3802.

Kalman, R., 1960 : A new approach to linear filtering and prediction problems. *Trans. ASME (J. Basic Eng.)*, **82D**, 35–50.

Kalman, R. E., 1994 : Randomness reexamined. *Modeling, Identification and Control*, **15 (3)**, 141–151.

Kendall, M. G., 1962 : *Rank correlation methods.* Griffin, 210 pp.

Kuhn, T., 1983 : *La structure des révolutions scientifiques.* Champs Flammarion, 284 pp.

Lauvernet, C., J.-M. Brankart, F. Castruccio, G. Broquet, P. Brasseur, and J. Verron, 2009 : A truncated Gaussian filter for data assimilation with inequality constraints : application to the hydrostatic stability condition in ocean models. *Ocean Modeling*, **27**, 1–17.

Lermusiaux, P., 2006 : Uncertainty estimation and prediction for interdisciplinary ocean dynamics. *Journal of Computational Physics*, **217**, 176–199.

Lermusiaux, P., 2007 : Adaptive modeling, adaptive data assimilation and adaptive sampling. *Physica D*, **230**, 172–196.

Lermusiaux, P. and A. Robinson, 1999 : Data assimilation via error subspace statistical estimation. part I : Theory and schemes. *Monthly Weather Review*, **127 (7)**, 1385–1407.

Lévy, M., M. Gavart, L. Mémery, G. Caniaux, and A. Paci, 2005 : A four-dimensional mesoscale map of the spring bloom in the northeast atlantic (pomme experiment) : results of a prognostic model. *Journal of Geophysical Research*, **110**, C07S21.

Li, H., E. Kalnay, and T. Miyoshi, 2009 : Simultaneous estimation of covariance inflation and observation errors within an ensemble kalman filter. *Quaterly Journal of the Royal Meteorological Society*, **135**, 523–533.

Lingamneni, A., C. Enz, K. Palem, and C. Piguet, 2013 : Designing energy-efficient arithmetic operators using inexact computing. *Journal of Low Power Electronics*, **9(1)**.

Lingamneni, A., K. K. Muntimadugu, C. Enz, R. M. Karp, K. Palem, and C. Piguet, 2012 : Algorithmic methodologies for ultra-efficient inexact architectures for sustaining technology scaling. *Proceedings of the 9th conference on Computing Frontiers*, 3–12.

Lorenz, E., 1975 : Climate predictability. *The physical basis of climate and climate modelling*, Geneva, World Meteorological Organization, WMO GARP Publ. Ser., Vol. 16, 132–136.

Madec, G. and M. Imbard, 1996 : A global ocean mesh to overcome the north pole singularity. *Climate Dynamics*, **12**, 381–388.

Madec, G. and the NEMO team, 2008 : NEMO ocean engine. Note du Pôle de modélisation 27, Institut Pierre-Simon Laplace (IPSL), France. ISSN 1288-1619.

Magri, S., 2002 : Assimilation de données dans un modèle décosystème marin couplé à un modèle de couche de mélange océanique de la mer ligure. Ph.D. thesis, Université Joseph Fourier (Grenoble).

Meinvielle, M., 2012 : Ajustement optimal des paramètres de forçage atmosphérique par assimilation de données de température de surface pour des simulations océaniques globales. Ph.D. thesis, Université Joseph Fourier (Grenoble).

Meinvielle, M., J.-M. Brankart, P. Brasseur, B. Barnier, R. Dussin, and V. J., 2013 : Optimal adjustment of the atmospheric forcing parameters of ocean models using sea surface temperature data assimilation. *Ocean Science*, soumis.

Melet, A., J. Verron, and J.-M. Brankart, 2012 : Potential outcomes of glider data assimilation in the solomon sea : control of the water mass properties and parameter estimation. *Journal of Marine Systems*, **94**, 232–246.

Metropolis, N., M. Rosenbluth, A.W. Rosenbluth, A. Teller, and E. Teller, 1953 : Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, **21(6)**, 1087–1092.

Murphy, A. H., 1973 : A new vector partition of the probability score. *J. Appl. Meteor.*, **12**, 595–600.

Murphy, A. H., 1977 : The value of climatological, categorical and probabilistic forecasts in the cost-loss ratio situation. *Monthly Weather Review*, **105**, 803–816.

Nerger, L., W. Hiller, and J. Schröter, 2005 : A comparison of error subspace Kalman filters. *Tellus*, **57A**, 715–735.

Ott, E., et al., 2004 : A local ensemble Kalman filter for atmospheric data assimilation. *Tellus*, **56A**, 415–428.

Ourmières, Y., J.-M. Brankart, L. Berline, P. Brasseur, and J. Verron, 2006 : Incremental analysis update implementation into a sequential ocean data assimilation system. *Journal of Atmospheric and Oceanic Technology*, **23(12)**, 1729–1744.

Ourmières, Y., P. Brasseur, M. Lévy, J.-M. Brankart, and J. Verron, 2009 : On the key role of nutrient data to constrain a coupled physical-biogeochemical assimilative model of the north atlantic ocean. *Journal of Marine Systems*, **75**, 100–115.

Palmer, T., 2002 : The economic value of ensemble forecasts as a tool for risk assessment : from days to decades. *Quaterly Journal of the Royal Meteorological Society*, **128**, 747–774.

Palmer, T., G. Shutts, R. Hagedorn, F. Doblas-Reyes, T. Jung, and M. Leutbecher, 2005 : Representing model uncertainty in weather and climate prediction. *Annu. Rev. Earth Planet. Sci.*, **33**, 163–193.

Parent, L., 2000 : Assimilation de données dans l'océan pacifique tropical sur la période 1994-1998. Ph.D. thesis, Université Joseph Fourier (Grenoble).

Parent, L., C. Testut, J. Brankart, J. Verron, P. Brasseur, and L. Gourdeau, 2003 : Comparative assimilation of Topex/Poseidon and ERS altimeter data and of TAO temperature data in the Tropical Pacific Ocean during 1994–1998, and the mean sea-surface height issue. *Journal of Marine Systems*, **40–41**, 381–401.

Penduff, T., P. Brasseur, C.-E. Testut, B. Barnier, and J. Verron, 2003 : Assimilation of sea-surface temperature and altimetric data in the South Atlantic Ocean : impact on basin-scale properties. *Journal of Marine Research*, **60**, 805–833.

Penduff, T., M. Juza, B. Barnier, J. Zika, W. Dewar, A.-M. Tréguier, J.-M. Molines, and N. Audiffren, 2011 : Sea-level expression of intrinsic and forced ocean variabilities at interannual time scales. *Journal of Climate*, **24**, 5652–5670.

Pham, D. T., J. Verron, and M. C. Roubaud, 1998 : Singular evolutive extended Kalman filter with EOF initialization for data assimilation in oceanography. *Journal of Marine Systems*, **16**, 323–340.

Richardson, D. S., 2000 : Skill and relative economic value of the ecmwf ensemble prediction system. *Quaterly Journal of the Royal Meteorological Society*, **126**, 649–668.

Rixen, M., J.-M. Beckers, J.-M. Brankart, and P. Brasseur, 2000 : A numerically efficient data analysis method with error map generation. *Ocean Modeling*, **2**, 45–60.

Robert, C. and P. Casella, 2004 : *Monte Carlo Statistical Methods*. Springer, 645 pp.

Rozier, D., F. Birol, E. Cosme, P. Brasseur, J.-M. Brankart, and J. Verron, 2007 : A reduced order kalman filter for data assimilation in physical oceanography. *SIAM Reviews*, **49(3)**, 449–465.

Sambridge, M. and K. Mosegaard, 2002 : Monte Carlo methods in geophysical inverse problems. *Reviews in Geophysics*, **40 (3)**, 1009.

Sartori, J. and R. Kumar, 2011 : Architecting processors to allow voltage/reliability tradeoffs. *International Conference on Compilers, Architecture, and Synthesis of Embedded Systems*.

Shanbhag, N. R., S. Mitra, G. de Veciana, R. Orshansky, M. Marculescu, J. Roychowdhury, D. Jones, and J. M. Rabaey, 2008 : The search for alternative computational paradigms. *IEEE Design & Test of Computers*, **25(4)**, 334–343.

Shannon, C., 1948 : A mathematical theory of communication. *Bell System Technical Journal*, **27**, 379–423, 623–656.

Silverman, B. W., 1986 : *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 175 pp.

Skachko, S., J.-M. Brankart, F. Castruccio, P. Brasseur, and J. Verron, 2009 : Improved turbulent air-sea flux bulk parameters for the control of the ocean mixed layer : a sequential data assimilation approach. *Journal of Atmospheric and Oceanic Technology*, **26 (3)**, 538–555.

Skandrani, C., J.-M. Brankart, N. Ferry, J. Verron, and P. Brasseur, 2009 : Controlling atmospheric forcing parameters of global ocean models : sequential assimilation of sea surface mercator-ocean reanalysis data. *Ocean Science*, **5**, 403–419.

Sloan, J., J. Sartori, and R. Kumar, 2012 : On software design for stochastic processors. *49th Design and Automation Conference*.

Sondergaard, T. and P. Lermusiaux, 2013 : Data assimilation with gaussian mixture models using the dynamically orthogonal field equations. part I. theory and scheme. *Monthly Weather Review*, **141 (6)**, 1737–1760.

Spearman, C., 1904 : The proof and measurement of association between two things. *Amer. J. Psychol.*, **15**, 72–101.

Srinivasan, A., et al., 2011 : A comparison of sequential assimilation schemes for ocean prediction with the hybrid coordinate ocean model (hycom) : Twin experiments with static forecast error covariances. *Ocean Modeling*, **37(3-4)**, 85–111.

Tarantola, A., 1987 : *Inverse problem theory.* Elsevier, 612 pp.

Tarantola, A., 2005 : *Inverse problem theory and methods for model parameter estimation.* SIAM, Philadelphia, 342 pp.

Testut, C., P. Brasseur, J. Brankart, and J. Verron, 2003 : Assimilation of sea-surface temperature and altimetric observations during 1992–1993 into an eddy permitting primitive equation model of the North Atlantic Ocean. *Journal of Marine Systems*, **40–41**, 291–316.

Testut, C.-E., 2000 : Assimilation de données satellitales avec un filtre de kalman de rang réduit dans un modèle aux équations primitives de l'océan atlantique. Ph.D. thesis, Université Joseph Fourier (Grenoble).

Titaud, O., J.-M. Brankart, and J. Verron, 2011 : On the use of finite-time lyapunov exponents and vectors for direct assimilation of images in ocean models. *Tellus A*, **63(5)**, 1038–1051.

Toth, Z., O. Talagrand, G. Candille, and Y. Zhu, 2003 : Probability and ensemble forecasts. *Forecast Verification : A Practitioner's Guide in Atmospheric Science*, S. D. Jolliffe I., Ed., Wiley, 137–163.

Troupin, C., et al., 2012 : Generation of analysis and consistent error fields using the data interpolating variational analysis (diva). *Ocean Modeling*, **52–53**, 90–101.

Ubelmann, C., 2009 : Etude de scénarios d'altimétrie satellitaire pour le contrôle de la circulation océanique dans l'océan atlantique tropical par assimilation de données. Ph.D. thesis, Université Joseph Fourier (Grenoble).

Ubelmann, C., J. Verron, J.-M. Brankart, P. Brasseur, and E. Cosme, 2012 : Assimilating altimetric data from multi-satellite scenarios to control atlantic tropical instability waves : an observing system simulation experiments study. *Ocean Dynamics*, **62(6)**, 867–880.

Ubelmann, C., J. Verron, J.-M. Brankart, E. Cosme, and P. Brasseur, 2009 : Impact of upcoming altimetric missions on the prediction of the three-dimensional circulation in the tropical atlantic ocean. *Journal of Operational Oceanography*, **2(1)**, 3–14.

van Leeuwen, P.-J., 2009 : Particle filtering in geophysical systems. *Monthly Weather Review*, **137**, 4089–4114.

Verlaan, M. and A. W. Heemink, 1997 : Tidal flow forecasting using reduced rank square root filters. *Stoch. Hydrol. Hydraul.*, **11 (5)**, 349–368.

Verron, J., 1990 : Altimeter data assimilation into an ocean circulation model : Sensitivity to orbital parameters. *Journal of Geophysical Research*, **95 (C7)**, 11 443–11 459.

Verron, J., L. Gourdeau, D. T. Pham, R. Murtugudde, and A. J. Busalacchi, 1999 : An extended Kalman filter to assimilate satellite altimeter data into a non-linear numerical model of the Tropical Pacific : method and validation. *Journal of Geophysical Research*, **104**, 5441–5458.

von Neuman, J., 1956 : Probabilistic logics and the synthesis of reliable organisms form unreliable components. *Automata studies*, C. E. Shannon and J. McCarthy, Eds., Princeton university press, 43–98.

Wahba, G. and J. Wendelberger, 1980 : Some new mathematical methods for variational objective analysis using splines and cross validation. *Monthly Weather Review*, **108**, 1122–1143.

Wang, L., G. Libert, and P. Manneback, 1992 : Kalman filter algorithm based on singular value decomposition. *Proceedings of the 31st Conference on Decision and Control (Tucson, Arizona)*, 6 pp.

Weaver, A., J. Vialard, and D. Anderson, 2003 : Three- and four-dimensional variational assimilation with a general circulation model of the tropical pacific ocean, part 1 : formulation, internal diagnostics and consistency checks. *Monthly Weather Review*, **131**, 1360–1378.

Annexes

Annexe A

Développements techniques

> The deeper, long term understanding
> of the nature of the problem must
> be the goal of the system engineer,
> wheras the client always wants
> prompt relief from the symptoms
> of his current problem.
>
> Richard Hamming (1997)

Tous les outils de modélisation des incertitudes et de résolution de problèmes inverses utilisés pour nos travaux ont été développés et implémentés par nos soins au fur et à mesure de nos besoins et du progrès de nos idées. Et ce fut ma responsabilité d'ingénieur de recherche d'organiser ces développements, et de rendre ces outils utilisables par les étudiants ou postdoctorants de l'équipe ou d'ailleurs. L'objectif de cette annexe est d'essayer de tirer quelques enseignements généraux de cette activité, et d'examiner comment elle devrait se poursuivre dans l'avenir.

Ma façon de concevoir le rôle d'un ingénieur, au service direct d'une activité de recherche, ne s'est vraiment clarifiée qu'à la lecture, il y a quelques années, du très beau livre de Richard Hamming (1997) : *"The Art of Doing Science and Engineering"*. Dans ce livre, l'auteur propose trois règles fondamentales pour l'ingénieur système (*"systems engineering"*) :

Rule 1 : *If you optimize the components you will probably ruin the system performance.*

Rule 2 : *Part of systems engineering design is to prepare for changes so that they can be gracefully made and still not degrade the other parts.*

Rule 3 : *The closer you meet the specifications the worse the performance when overloaded.*

C'est à l'aune de ces trois règles que je voudrais discuter ici le système d'outils que nous avons élaboré, d'abord de façon générale, puis de façon spécifique pour chacun d'eux.

De façon générale, je crois qu'on peut lire les trois règles comme des moyens d'éviter les trois erreurs principales dans la construction d'un système. Premièrement, d'un point de vue pratique et technique, il est bien sûr nécessaire que chacun des multiples composants d'un système puisse être développé de façon aussi indépendante que possible. Mais, il importe aussi de réaliser qu'une fois rassemblés, les composants entrent en interaction, et le système complexe qui en résulte possède des propriétés émergentes qui ne peuvent se déduire de l'examen de chacun des composants. Dans une équipe de recherche, l'ingénieur est à mon avis celui qui connaît le fonctionnement technique du système dans sa globalité, et qui peut mesurer l'effet d'une modification locale sur les propriétés de l'ensemble. Il a donc un rôle important à jouer pour préserver l'intégrité du système,

face à des développements spécifiques qui pourraient en ruiner la performance globale (règle 1).

Le deuxième écueil qu'il faut éviter est de considérer que l'outil qu'on est en train de développer constitue une réponse définitive au problème posé. Car dans cette optique, la tendance naturelle est de rendre le système excessivement rigide, et inapte à s'adapter à des problématiques évoluant inévitablement au fil du temps. Ce problème est particulièrement aigu dans une équipe de recherche, dont le but est précisément de faire évoluer les idées sur le monde et la façon de résoudre les problèmes. Dans ce contexte, le rôle de l'ingénieur est donc de maintenir un système qui reste toujours assez *flexible* pour s'ajuster à la nouvelle compréhension des choses (règle 2). L'important n'est donc pas tant de concevoir une solution parfaitement adaptée à la situation présente, que de premettre, qu'à chaque étape de développement, une meilleur connaissance du problème puisse être progressivement acquise. Car il vaut bien mieux une solution constamment un peu fausse d'un problème que l'on comprend de mieux en mieux, qu'une solution de plus en plus exacte d'un problème que l'on aurait figé une fois pour toute.

Troisièmement, quand on élabore un système pour résoudre un problème donné, la tentation est souvent grande d'implémenter une solution qui réponde optimalement à chaque demande précise qui est formulée. Cela aurait sans doute le mérite de satisfaire pleinement les desiderata des utilisateurs de l'équipe de recherche. Mais il s'agit tout de même d'une tentation à laquelle l'ingénieur se doit de résister, car son but est de comprendre et de satisfaire globalement et sur le long terme les besoins de la recherche. Et il est souvent bien moins coûteux, en temps et en énergie, de développer une solution *générique*, certes imparfaite, pour un large spectre de problèmes possibles, plutôt qu'une solution spécifiquement optimale pour chaque problème (règle 3).

A.1 Le logiciel SESAM

SESAM (System of Sequential Assimilation Modules) est une boîte à outils qui contient la plupart des composants algorithmiques nécessaires à nos travaux en assimilation de données (présentés aux chapitres 6 et 7). Pour fixer les idées, les composants les plus emblématiques de SESAM sont : l'algorithme en base propre (éqs. 6.8 à 6.12) pour conditionner une distribution gaussienne à un jeu d'observations, l'algorithme de localisation des matrices de covariance (section 6.3), l'algorithme de modélisation adaptative des incertitudes (section 6.5), l'algorithme d'échantillonnage de distributions de probabilité gaussiennes tronquées (section 7.1), ou encore l'algorithme de transformation anamorphique (section 7.2). Mais SESAM contient aussi des composants plus standards tels que le calcul d'EOFs, l'interpolation d'une solution du modèle aux points d'observations, le calcul de statistiques d'écarts aux observations, etc. L'idée générale est en effet de rendre aussi facile que possible l'implémentation de nouveaux composants, à condition qu'ils puissent s'exprimer en fonction des différents types de données génériques que SESAM permet de traiter (vecteur de contrôle, vecteur d'observation, ou ensemble de ces vecteurs), sans jamais dépendre de leur signification particulière (physique, biologique, ou autre).

Pour fonctionner sur un problème particulier, SESAM doit donc être configuré pour définir ce que contient chacun de types de données génériques. Cela se fait via un fichier de configuration qui spécifie la liste et les caractéristiques des variables du vecteur de contrôle, ainsi que la liste et les caractéristiques des observations associées. A partir de cette information, SESAM peut alors savoir comment lire un vecteur de contrôle ou un vecteur d'observation à partir de fichiers, ou réciproquement comme les écrire sur le disque. Et pour le reste, le fonctionnement interne de SESAM restera indépendant du problème posé, isolé par cette interface assez générale entre fichiers de l'utilisateur (en

externe) et types de données génériques (en interne). Ainsi, une matrice de covariance de rang réduit pourra-t-elle être fournie à SESAM sous la forme d'un ensemble de vecteurs (dans l'espace de contrôle, dans l'espace d'observation, ou dans l'espace conjoint), et une matrice de covariance d'erreur d'observation (diagonale) sous la forme d'un vecteur d'observation. De cette manière, SESAM est complètement indépendant du sujet de l'étude, et pourrait donc être utilisé tel quel, sans aucun codage, à n'importe quelle discipline pour laquelle ses composants seraient pertinents.

La raison de cette formulation assez générique est, je crois, que dès le départ, SESAM a été conçu pour fonctionner non-seulement avec des modèles de circulation océaniques de natures diverses (Testut et al., 2003; Parent et al., 2003; Brankart et al., 2003), mais aussi avec un modèle d'écosystème (Carmillet et al., 2001). Et c'est, je pense, ce mode de fonctionnement initial qui a incité à une plus grande souplesse algorithmique, et à une plus grande flexibilité dans le type d'application que nous avons pu traiter au fil du temps. En voici les quelques exemples les plus marquants :

- Initialement, les algorithmes de SESAM n'ont pas été conçus pour tenir compte des corrélations spatiales des erreurs d'observation : la matrice $\mathbf{R}$ était toujours supposée diagonale. Pourtant, de la façon dont le code était construit, cette possibilité était déjà là sans que nous en soyons conscients. Il suffisait d'augmenter le vecteur d'observation avec des observations dépendantes des observations originales (selon la méthode décrite en section 6.4). Et pour le faire, il nous a suffi de modifier le fichier de configuration de SESAM pour inclure ces observations additionnelles, *sans modifier en quoi que ce soit* le code de SESAM lui-même, qui était déjà assez général et flexible pour cela.
- De la même manière, SESAM n'a jamais été conçu pour fonctionner avec un opérateur d'observation non-linéaire. Pourtant, a posteriori, on peut voir que SESAM offre une façon assez pratique de le faire : il suffit d'augmenter le vecteur de contrôle avec de nouvelles variables non-linéairemement reliées aux variables originales, ce qui peut encore se faire de l'extérieur en ne modifiant que le fichier de configuration, et pas le code de SESAM lui-même. D'un certain point de vue, cette façon de faire est tout à fait sous-optimale (vecteur plus long, plus de mémoire nécessaire, etc.), mais d'un autre point de vue, en décrivant les distributions de probabilité dans un espace agrandi, elle est plus générique et donc plus flexible. Elle permet par exemple d'utiliser des transformations anamorphiques différentes pour les variables originales et pour les variables nouvelles.
- Dans les premières années de l'existence de SESAM, nous n'avons non plus jamais imaginé que le vecteur de contrôle contiendrait autre chose que le vecteur d'état du modèle (c'est d'ailleurs toujours son nom dans SESAM). Pourtant, quand nous avons voulu estimer le forçage (Skachko et al., 2009; Skandrani et al., 2009; Meinvielle et al., 2013) ou les paramètres du modèle (Doron et al., 2011; Melet et al., 2012; Doron et al., 2013), rien n'a dû être modifié dans le code lui-même. Il a de nouveau suffi de modifier le fichier de configuration pour augmenter le vecteur de contrôle, ici par le forçage et là par des paramètres.
- De même, SESAM n'a pas été initialement pensé pour fonctionner avec un modèle comportant plusieurs grille emboîtées. Mais quand nous avons voulu le faire (Melet et al., 2012), nous avons constaté qu'aucun nouveau codage n'était nécessaire (excepté pour l'algorithme de localisation des covariances, qui devait connaître le positionnement relatif des grilles), et qu'il suffisait de modifier le fichier de configuration, en définissant chaque variable séparément pour chaque grille.

Toutes ces généralisations peuvent sembler anodines a posteriori, d'autant qu'elles découlent toutes très naturellement de la formulation du problème telle qu'elle est présentée aux chapitres 5, 6 et 7, où les vecteurs de contrôle et d'observations sont étendus dès

le départ à tout ce qu'ils peuvent contenir. Mais il faut se rappeler que ce n'est pas ainsi que le filtre SEEK était formulé à l'origine. Qu'est ce qui alors a été à l'origine de ces généralisations successives de la méthode, de la diversification des applications (vers la correction du forçage et des paramètres), et même des extensions récentes aux problèmes non-gaussiens ? Bien sûr, il y a le fait que cela permettait de répondre à des problèmes scientifiques nouveaux et pertinents, mais il y a à mon avis autre chose qui a influencé notre conception du problème, et qui est de nature essentiellement technique. Je crois en effet que, si nous avons suivi cette voie, c'est aussi beaucoup parce que c'était possible et cohérent techniquement, parce que ça correspondait à une évolution graduelle du système de départ, et parce qu'on préservait le côté générique de l'outil, en évitant les développements trop spécifiques, à usage unique ou sans avenir, c'est-à-dire en bref parce que c'était parfaitement cohérent avec les trois règles de l'ingénieur-système énoncées au début de cette annexe.

Mais cette approche n'est pas sans contrepartie, et nous avons dû aussi nous accomoder des suboptimalités techniques qu'elle implique. En premier lieu, le code a vieilli et, comme il n'a jamais été réécrit, il a inévitablement conservé les défauts de sa jeunesse. Sa structure générale n'est pas vraiment conforme aux standards actuels, et en dépit de nombreux commentaires, le code reste difficilement lisible pour un utilisateur non-averti. Ensuite, SESAM ne dispose pas des optimisations nécessaires pour aborder les très gros problèmes. Pour cette raison, il ne pourrait pas être appliqué à un système opérationnel de très grande taille, pour lequel un outil spécifiquement optimisé (tel que le système SAM2 de Mercator) est nécessaire. En particulier, seuls certains composants de SESAM ont été parallélisés, ce qui ne permet pas de répondre à toutes les exigences d'un problème de très grande dimension. Pourtant, malgré tout cela, je ne crois pas qu'il soit à l'heure actuelle souhaitable de reprendre le codage de SESAM sur des bases nouvelles. Car il reste à ce jour suffisant pour le type d'application qu'on peut envisager dans une équipe de recherche, et il est suffisamment souple pour permettre d'explorer et de tester facilement des pistes méthodologiques nouvelles. A charge alors pour ceux qui veulent appliquer ces développements à un système opérationnel de grande taille de récupérer le code de SESAM et de l'adpater à leurs exigences. C'est d'ailleurs par exemple ce qui a été fait récemment pour l'anamorphose (section 7.2), quand elle a été transférée de SESAM au système d'assimilation de Mercator (SAM2). La seule et vraie raison qui devra à mon avis nous inciter à changer radicalement de système est un bouleversement de notre approche méthodologique. Un premier exemple de la façon dont cela pourrait se faire est discuté dans la section suivante.

A.2 Le logiciel OSMIUM

OSMIUM (Ocean Surface Mesoscale velocity Inversion from the Unstable Manifold) est le logiciel que nous avons développé pour résoudre le problème de la reconstitution d'un champ de vitesse de surface à mésoéchelle à partir d'images de traceurs (température de surface ou couleur de l'eau) à très haute résolution, résolvant la sous-mésoéchelle (voir sections 2.3.c, 4.1 et 7.4). D'un point de vue technique, ce logiciel est construit sous la forme d'un assemblage de modules assez indépendants les uns des autres. Certains de ces modules sont très étroitement liés au problème posé :

- un module d'acquisition d'une image de traceur à haute résolution ;
- un module d'acquisition de champs de vitesse à mésoéchelle, avec interpolation spatio-temporelle pour permettre le calcul des exposants de Lyapounov ;
- un module d'acquisition d'une base d'erreur réduite pour le champ de vitesse (modes propres de la matrice de covariance d'erreur d'ébauche).

D'autres modules sont plus génériques, mais toujours assez liés au problème posé :

- un module d'advection de particules par un champ de vitesse pour le calcul des exposants de Lyapounov ;
- un module de calcul des exposants de Lyapounov à partir de champs de vitesse ;
- un module de binarisation d'une image à haute résolution (traceur ou exposants de Lyapounov associés au champ de vitesse) ;
- un module d'évaluation de la fonction coût par le calcul d'une norme de l'écart entre deux images binarisées.

Et d'autres modules enfin sont tout à fait génériques et indépendants du problème posé :

- un module de génération de nombres aléatoires ;
- un module de minimisation d'une fonction coût par la méthode du recuit simulé ;
- un module d'échantillonnage de la distribution de probabilité a posteriori par échantillonnage de Gibbs.

C'est de l'interaction de ces modules dans un système intégré qu'émerge la possibilité d'inférer de façon plus ou moins performante (voir règle 1) le champ de vitesse à partir de l'image du traceur.

Mais avec un peu plus de recul, et en ne regardant que l'aspect technique des choses, on peut aussi voir ce logiciel comme une première tentative de changer radicalement d'approche méthodologique. On peut le voir comme une première étape de développement permettant d'acquérir, à peu de frais, une première expérience du fontionnement de méthodes plus générales pour résoudre un problème inverse (basés sur des chaînes de Markov, voir section 7.4). Et forts de cette première étape qui nous a permis de mieux comprendre le problème, nous sommes désormais mieux armés pour imaginer ce que pourrait être la suite (voir règle 2). Par ailleurs, les modules de bases sont suffisamment génériques pour être réutilisables ou facilement adaptables à un large spectre de problèmes possibles (voir règle 3). Cependant, si, le cas échéant, ce genre de méthode se révèle inadapté aux problèmes que nous nous posons, eh bien, il y aura eu suffisamment peu d'effort de développement consenti pour se permettre de l'abandonner, purement et simplement !

A.3 Le modèle NEMO

NEMO (Nucleus for European Modelling of the Ocean) est le système de modélisation océanique (voir chapitre 1) que nous utilisons pour nos travaux. Dans cette section, je n'aborderai cependant que la question des paramétrisations stochastiques (voir chapitre 3) et des simulations d'ensemble (voir chapitre 2), qui sont les seuls outils que j'ai effectivement implémentés dans NEMO.

D'une part, pour initier l'implémentation de paramétrisations stochastiques dans NEMO, j'ai choisi une approche simple et générique, qui permette d'acquérir rapidement une première expérience de son effet potentiel sur le comportement du modèle. Il s'est agi, pour l'essentiel, d'inclure deux nouveaux modules à NEMO. Le premier est simplement un générateur autonome de nombres aléatoires. Cela permet de générer exactement la même séquence de nombres aléatoires, indépendamment de la machine ou du compilateur, avec une graine qui dépend de façon déterministe du membre de l'ensemble (et du sous-domaine, dans le cas où le modèle est parallélisé par découpage du bassin océanique). En cas de besoin, un des membres de l'ensemble peut donc être rejoué avec un forçage stochastique parfaitement identique (à condition d'utiliser le même découpage en sous-domaines). Le deuxième module permet quant à lui de générer, pour chaque point de grille du modèle (en 2D ou en 3D), une série de processus autorégressifs d'ordre n, dont on peut spécifier la moyenne, l'écart-type, et le temps de décorrélation via le fichier de paramètres. Une dépendance spatiale peut également être introduite en appliquant un filtre spatial aux séries temporelles ainsi créées. Par ailleurs, quand le

modèle est arrété, le module permet de créer un fichier de redémarrage spécifique qui contient l'état courant des processus autorégressifs et du générateur de nombre aléatoire, afin de pouvoir prolonger la simulation tout comme si elle n'avait pas été interrompue.

Grâce à cette implémentation générique (et donc probablement non-optimale), il est facilement possible d'explorer l'effet d'incertitudes dans n'importe quel composant de NEMO, tant dans le modèle de circulation que dans le modèle de glace ou le modèle d'écosystème (voir chapitre 3). Il suffit pour cela de faire usage du module stochastique dans la partie correspondante de NEMO, et de la perturber en utilisant l'un des processus stochastiques qui ont été calculés.

D'autre part, simuler explicitement l'incertitude conduit naturellement au besoin de simulations d'ensemble. Pour cela, il est apparu qu'une très légère modification de la parallélisation de NEMO suffisait pour lancer simultanément un ensemble de n simulations à l'aide d'*un seul* appel au modèle, simplement parallélisé sur n fois plus de processeurs. Au lieu que chaque processeur détermine seulement quel sous-domaine il doit calculer, il fallait juste qu'il détermine aussi à quel membre de l'ensemble il appartient. Et au lieu d'un communicateur unique rassemblant la totalité de processeurs, il fallait juste définir un communicateur spécifique pour chacun des membres de l'ensemble. A l'intérieur de chacun de ces communicateurs, chaque membre de l'ensemble peut vivre sa vie indépendamment des autres sans que rien d'autre dans le modèle ne doive être altéré. Bien sûr, pour que tous les membres de l'ensemble ne fassent pas la même chose, il faut que le numéro du membre puisse influencer la simulation, soit à travers le nom des fichiers de condition initiale, de forçage ou de paramètres, soit à travers la graine du générateur de nombres aléatoires utilisé pour les paramétrisations stochastiques.

Jusque là, du point de vue des résultats produits, cette méthode pour réaliser une simulation d'ensemble est parfaitement équivalente à lancer n simulations différentes à l'aide de n appels distincts au modèle. A ce stade, l'unique différence est qu'elle est potentiellement plus fexible et ouvre des perspectives nouvelles d'évolution du système. En effet, en plus de définir un communicateur pour chaque membre, il devient possible de définir aussi un communicateur pour chaque sous-domaine, à travers tout l'ensemble, afin de calculer en cours de simulation des statistiques d'ensemble, telles que la moyenne, l'écart-type ou n'importe quelle caractéristique de la distribution de probabilité. De cette façon, il devient donc aussi possible de faire en sorte que l'évolution de chacun des membres dépende de ces statistiques d'ensemble. Une première application qui vient à l'esprit serait par exemple de relaxer la moyenne d'ensemble vers des observations tout en laissant les anomalies se développer librement.

Pour conclure cette section, il me semble important de remarquer que NEMO est, d'un point de vue technique, un système complexe dont le fonctionnement dépend d'un grand nombre de composants en interaction. Quand le système fonctionne, son comportement possède des propriétés émergentes qu'il est en pratique impossible de déduire d'une connaissance détaillée de ses composants. Dans cette optique, la simulation explicite des incertitudes me semble aussi trouver une justification purement technique. Quand on la néglige, en ne considérant que l'assemblage de composants optimaux (mais imparfaits), on peut altérer considérablement les propriétés émergentes du système complexe et donc ruiner sa performance (voir règle 1 ci-dessus). Ainsi, bien que très mineurs d'un point de vue technique (en regard de la complexité de NEMO), les développements décrits ici me semblent-ils constituer une perspective importante pour l'amélioration de la performance et de la robustesse du système, qui se situe sur une frontière floue entre recherche océanographique et ingéniérie des systèmes complexes.

Annexe B

Articles scientifiques

Dans cette annexe, j'ai choisi de présenter huit articles récents dont je suis premier auteur ou 'corresponding author', et qui me paraissent bien refléter l'aspect scientifique de mon actvité au cours des années récentes. Au vu de cette série d'article, on voit bien néanmoins que cette activité est résolument orientée vers le développement de méthodes, parfois de nature assez générale et indépendantes du sujet traité, et parfois spécifiquement adaptées aux incertitudes océanographiques.

Comme annoncé dans l'introduction, l'ordre chronologique que j'ai suivi ici ne correspond pas à l'organisation du mémoire. Voici donc le lien entre chaque article présenté ici et les chapitres de ce mémoire :

- **Skachko et al. (2009)**, *Journal of Atmospheric and Oceanic Technology*. Cet article traite des incertitudes liées au forçage atmosphérique (chapitre 2), et sur la façon de la contrôler par la méthode du vecteur d'état augmenté (chapitre 6).
- **Skandrani et al. (2009)**, *Ocean Science*. Cet article fait suite au précédent et traite du même sujet (chapitres 2 et 6)
- **Lauvernet et al. (2009)**, *Ocean Modelling*. Cet article traite d'incertitudes sujettes à des contraintes d'inégalité, de leur modélisation par une distribution gaussienne tronquée, et de la façon de résoudre un problème inverse sous cette hypothèse (section 7.1).
- **Brankart et al. (2009)**, *Monthly Weather Review*. Cet article propose un algorithme efficace pour modéliser la corrélation spatiales des incertitudes sur les observations (section 6.4).
- **Brankart et al. (2010)**, *Monthly Weather Review*. Cet article propose un algorithme efficace de modélisation adaptative des incertitudes (section 6.5).
- **Brankart et al. (2011)**, *Monthly Weather Review*. Cet article propose un algorithme efficace de localisation des matrices de covariance d'erreur (section 6.3).
- **Brankart et al. (2012)**, *Ocean Science*. Cet article décrit l'algorithme de transformation anamorphique que nous avons développé (sections 2.2 et 7.2), en l'illustrant par des exemples divers.
- **Brankart (2013)**, *Ocean Modelling*. Cette article propose une paramétrisation stochastique (chapitre 3) permettant de simuler explicitement l'incertitude liée aux échelles non-résolues dans le calcul du gradient horizontal de densité.

Par ailleurs, afin d'alléger le poids de ce mémoire, je n'ai inclus pour chaque article que les deux premières pages, ce qui est toujours suffisant pour permettre une lecture complète du résumé et de l'introduction. Comme tous ces articles sont publiés, le reste est de toute façon accessible par ailleurs au lecteur intéressé.

Improved Turbulent Air–Sea Flux Bulk Parameters for Controlling the Response of the Ocean Mixed Layer: A Sequential Data Assimilation Approach

SERGEY SKACHKO, JEAN-MICHEL BRANKART, FRÉDÉRIC CASTRUCCIO, PIERRE BRASSEUR, AND JACQUES VERRON

LEGI/CNRS, UMR 5519, Grenoble, France

(Manuscript received 1 October 2007, in final form 1 September 2008)

ABSTRACT

Bulk formulations parameterizing turbulent air–sea fluxes remain among the main sources of error in present-day ocean models. The objective of this study is to investigate the possibility of estimating the turbulent bulk exchange coefficients using sequential data assimilation. It is expected that existing ocean assimilation systems can use this method to improve the air–sea fluxes and produce more realistic forecasts of the thermohaline characteristics of the mixed layer. The method involves augmenting the control vector of the assimilation scheme using the model parameters that are to be controlled. The focus of this research is on estimating two bulk coefficients that drive the sensible heat flux, the latent heat flux, and the evaporation flux of a global ocean model, by assimilating temperature and salinity profiles using horizontal and temporal samplings similar to those to be provided by the Argo float system. The results of twin experiments show that the method is able to correctly estimate the large-scale variations in the bulk parameters, leading to a significant improvement in the atmospheric forcing applied to the ocean model.

1. Introduction

Turbulent momentum, heat, and freshwater fluxes at the air–sea interface remain among of the main sources of error in present-day ocean models (Large 2006). They strongly limit the capacity of such models to provide a realistic forecast of the thermohaline characteristics of the mixed layer and of the surface ocean currents. This is the reason why we decided to investigate ways of improving the knowledge of these fluxes through assimilation of oceanic observations.

A way to address this problem would be to use a four-dimensional variational data assimilation (4DVAR) scheme, by including the fluxes in the control parameters of the model in addition to the initial conditions (Roquet et al. 1993; Stammer et al. 2004). However, many existing ocean assimilation systems are based on sequential assimilation schemes that would greatly benefit from more precise estimates of the fluxes. Hence, it is useful to look for alternative schemes that are numerically less expensive and easy to implement in a wide range of existing systems. Our starting point is a reduced-order Kalman filter in which the background error covariance matrix is represented by a set of 3D error modes in the state space of the ocean model (Pham et al. 1998). In this paper, we will use an optimal interpolation scheme derived from the singular evolution of the extended Kalman (SEEK) filter (Brasseur and Verron 2006). However, the method that we propose using to correct the fluxes is directly applicable in any similar assimilation scheme, such as ensemble optimal interpolation schemes, ensemble Kalman filters (Evensen 2003), or any variant of reduced-order Kalman filters: the reduced-rank square root (RRSQRT) scheme (Heemink et al. 2001), the error subspace statistical estimation (ESSE) system (Lermusiaux and Robinson 1999), the SEEK filter (Pham et al. 1998), etc.

A general method used to identify model parameters (like the air–sea flux bulk parameters) by means of a Kalman filter is to augment the filter control space by including these parameters in addition to the state variables. This method, although common in the engineering literature (Cox 1964; Ho and Whalen 1963; Nelson and Stear 1976; Ljung 1979), has not often been used in atmospheric or oceanographic applications of the Kalman filter. A first demonstration of the applicability

Corresponding author address: Jean-Michel Brankart, LEGI/CNRS, BP 53X, 38041 Grenoble CEDEX, France.
E-mail: jean-michel.brankart@hmg.inpg.fr

DOI: 10.1175/2008JTECHO603.1

of the method to atmospheric problems was performed by Anderson (2001), who used it for estimating one forcing parameter of the 40-variable Lorenz model, using an ensemble-based data assimilation scheme. Second, a work by Losa et al. (2003) deals with the joint state and parameter estimation, in a zero-dimensional ecosystem model, using a sequential-importance resampling filter. In a third study, by Annan et al. (2005), the method is applied to the tuning of the surface temperature climatology in a spectral primitive equation atmospheric GCM, using the ensemble Kalman filter. Similarly, Aksoy et al. (2006) also use an ensemble Kalman filter to estimate up to six (spatially constant) model parameters in a two-dimensional sea-breeze model. These four studies show the relevance of sequential data assimilation methods for estimating model parameters, but also point up the difficulty in proving the effectiveness of the method in all applications. Within that context, the objective of this paper is to demonstrate the possibility of estimating parameters driving the turbulent air–sea fluxes of a realistic ocean GCM, using a sequential assimilation scheme to invert oceanic observations.

If the purpose of the Kalman filter is to control the air–sea fluxes, the control parameters could be the fluxes themselves or the atmospheric fields from which they are computed. But it seems preferable to include a few key parameters of the bulk formula in the control vector (i) because they are more likely to persist over time (the aim is to improve the forecast), (ii) because they are likely to be as easy to control by ocean observations (provided that we only include parameters that are linearly linked to the value of the flux), and (iii) because they can be assumed (Large 2006) to be the real source of error (even if this will probably result in compensating errors on the atmospheric parameters by correcting the bulk coefficients).

In this paper, we present an example in which only the latent heat flux coefficient (C_E) and the sensible heat flux coefficient (C_H) are included in the control vector because they can be assumed to be one of the main sources of error (although in realistic experiments, other sources of error, like the wind stress drag coefficient, precipitation, or cloud cover, need to be considered). The procedure is tested using twin assimilation experiments with a low-resolution ($2° \times 2°$) global ocean configuration. Within this context, assimilation experiments, with sequential corrections of the bulk coefficients, can be performed if we provide a background error covariance matrix in the augmented control space. This will be done using ensemble simulations characterized by various values of C_E and C_H.

The synthetic observations that will be assimilated to correct the fluxes are temperature and salinity profiles with horizontal and temporal samplings comparable to those of the Argo floats system (about 3000 free-drifting profiling floats measuring the temperature and salinity of the upper 2000 m of the ocean; information online at http://www.argo.ucsd.edu/). In particular, we will study the ability to reconstitute the bulk coefficients by assimilating these kinds of oceanic observations, and examine how correcting the parameters can help improve the quality of the air–sea fluxes and thereby the quality of the temperature and salinity forecasts. Investigations will also be conducted into the behavior of the method in the presence of systematic errors arising from inaccurate parameters.

The structure of the paper is as follows: sections 2 and 3 present the ocean model and the assimilation methodology, respectively; section 4 describes the experimental protocol; and section 5 presents the results of the experiments.

2. The ocean model

The ocean model used in this study is the ORCA2 configuration of the Océan Parallélisé (OPA) model (Madec et al. 1998). This is a global ocean configuration using a $2° \times 2°$ ORCA-type horizontal grid, with the meridional grid spacing reduced to 1/2° in the tropical regions in order to improve the representation of the equatorial dynamics. This is a free-surface configuration based on the resolution of the ocean dynamic primitive equations, with a z-coordinate vertical discretization. There are 31 levels along the vertical, and the vertical resolution varies from 10 m in the first 120 m to 500 m at the bottom. The lateral mixing for active tracers (temperature and salinity) is parameterized along isopycnal surfaces, and the model uses a turbulent kinetic energy (TKE) closure scheme to evaluate the vertical mixing of the momentum and tracers. The vertical eddy viscosity and diffusivity coefficients are computed from a 1.5 turbulent closure model based on a prognostic equation for the turbulent kinetic energy, and a closure assumption for the turbulent length scales (see Blanke and Delecluse 1993 for more details).

The model is forced at the surface boundary with heat, freshwater, and momentum fluxes. The momentum flux is derived from the European Remote Sensing Satellite (ERS) scatterometer wind stresses complemented by Tropical Atmosphere–Ocean (TAO) derived stresses (Menkes et al. 1998). The heat and freshwater turbulent fluxes are computed using bulk formulation from National Centers for Environmental Prediction (NCEP) atmospheric interannual data provided by the National Oceanic and Atmospheric Administration–Cooperative Institute for Research in Environmental Sciences

Ocean Sci., 5, 403–419, 2009
www.ocean-sci.net/5/403/2009/

Ocean Science

Controlling atmospheric forcing parameters of global ocean models: sequential assimilation of sea surface Mercator-Ocean reanalysis data

C. Skandrani[1], **J.-M. Brankart**[1], **N. Ferry**[2], **J. Verron**[1], **P. Brasseur**[1], **and B. Barnier**[1]

[1]Laboratoire des Ecoulements Géophysiques et Industriels (LEGI/CNRS), Grenoble, France
[2]Mercator-Océan, Toulouse, France

Received: 26 May 2009 – Published in Ocean Sci. Discuss.: 18 June 2009
Revised: 17 September 2009 – Accepted: 21 September 2009 – Published: 16 October 2009

Abstract. In the context of stand alone ocean models, the atmospheric forcing is generally computed using atmospheric parameters that are derived from atmospheric reanalysis data and/or satellite products. With such a forcing, the sea surface temperature that is simulated by the ocean model is usually significantly less accurate than the synoptic maps that can be obtained from the satellite observations. This not only penalizes the realism of the ocean long-term simulations, but also the accuracy of the reanalyses or the usefulness of the short-term operational forecasts (which are key GODAE and MERSEA objectives). In order to improve the situation, partly resulting from inaccuracies in the atmospheric forcing parameters, the purpose of this paper is to investigate a way of further adjusting the state of the atmosphere (within appropriate error bars), so that an explicit ocean model can produce a sea surface temperature that better fits the available observations. This is done by performing idealized assimilation experiments in which Mercator-Ocean reanalysis data are considered as a reference simulation describing the true state of the ocean. Synthetic observation datasets for sea surface temperature and salinity are extracted from the reanalysis to be assimilated in a low resolution global ocean model. The results of these experiments show that it is possible to compute piecewise constant parameter corrections, with predefined amplitude limitations, so that long-term free model simulations become much closer to the reanalysis data, with misfit variance typically divided by a factor 3. These results are obtained by applying a Monte Carlo method to simulate the joint parameter/state prior probability distribution. A truncated Gaussian assumption is used to avoid the most extreme and non-physical parameter corrections. The general lesson of our experiments is indeed that a careful specification of the prior information on the parameters and on their associated uncertainties is a key element in the computation of realistic parameter estimates, especially if the system is affected by other potential sources of model errors.

Correspondence to: J.-M. Brankart
(jean-michel.brankart@hmg.inpg.fr)

1 Introduction

One of the most accurate and ubiquitous information about the surface state of the ocean is provided by the satellite measurements of sea surface temperature. It is in particular significantly more accurate than the sea surface temperature that is simulated by any state-of-the-art general circulation ocean model. Part of this discrepancy is explained by the relative inaccuracy of the atmospheric parameters that are used to compute the air-sea momentum, heat and fresh water fluxes which determine the surface boundary condition of the ocean model (WGASF, 2000). There is thus an important potential benefit to expect from the improvement of these parameters using the available sea surface observations. In practice, the atmospheric parameters controlling the air-sea fluxes (i.e. air temperature, relative humidity, cloud fraction, precipitation or wind speed) are derived from atmospheric reanalysis data (as delivered for instance by the ECMWF or NCEP centers) and from a variety of satellite products. For instance, the atmospherically forced ocean hindcast simulations performed by The DRAKKAR Group (2007) compute their air-sea fluxes by using forcing data that merge a variety of different data sets (in situ, satellite and NWP products), with objective corrections based on observations (Large and Yeager., 2008; Brodeau et al., 2009). Hence, as long as forced

Published by Copernicus Publications on behalf of the European Geosciences Union.

models are used to simulate the ocean component alone, the control of the atmospheric parameters using ocean surface observations is certainly an appropriate way of improving the realism of model interannual simulations, the accuracy of ocean reanalyses or the usefulness of sea surface temperature operational forecasts. It is thus also an important contribution to the GODAE[1] objectives (GODAE, 2008), which is the reason why a large part of the MERSEA[2] effort in the development of data assimilation has been devoted to this problem.

In this study, which has been conducted as part of the MERSEA project, this problem is investigated using idealized experiments in which Mercator-Ocean[3] ocean reanalysis data are used as the reference simulation (i.e. the "truth" of the problem). Synthetic observation datasets (for sea surface temperature and sea surface salinity) are extracted from the reanalysis to be assimilated in a coarse resolution global ocean model. With respect to Skachko et al. (2009), who investigated a similar problem using twin assimilation experiments, the present study is thus more realistic, since the difference between model and reanalysis is now very similar in nature to the real error. It is closer to the real problem even if the experiments are still somewhat ideal in the sense that no real observations are assimilated, and that the full reference model state (the reanalysis, in three dimensions) is available for validation. Another difference with respect to Skachko et al. (2009) is that, in this paper, we extend the control vector to 6 atmospheric parameters instead of 2 turbulent exchange coefficients in their example (but we exclusively focus on the control of the parameters, while they also considered the joint optimal estimate of the ocean state vector together with the atmospheric parameters). However, in order to solve this more realistic problem, we needed to further develop the methodology towards a better specification of the prior information about the parameters and their associated uncertainty. We observe indeed that making appropriate assumptions on that respect is increasingly important as the estimation problem is becoming more realistic, because it is more and more difficult to make the distinction between forcing errors and the other potential sources of error in the system. An additional important objective is thus to find means of identifying properly the part of the observational misfit that can be interpreted as resulting from inaccurate atmospheric parameters.

In order to reach this objective, the plan is to apply sequentially a Bayesian inference method to compute piecewise constant optimal parameter corrections. A possible algorithm to solve this problem is to compute the optimal parameters by direct maximization of the posterior probability distribution for the parameters, using for instance a 4DVAR scheme (as done in Roquet et al., 1993 or Stammer et al., 2004). But, in addition to the technical difficulties that the algorithm may involve, this solution requires that the cost function resulting from the optimal probabilistic criterion be quadratic or at least differentiable everywhere in parameter space, so that it is by no way straightforward to optimally impose strict inequality constraints to the parameters (by setting zero prior probability in prohibited region of the parameter space for instance). This is why, in this study, we prefer using a Monte Carlo algorithm to simulate the ocean response to parameter uncertainty, and use the resulting ensemble representation of the prior probability distribution to infer optimal parameter corrections from the ocean surface observations. It is in the specification of this prior probability distribution that two methodological improvements are introduced with respect to Skachko et al. (2009). First, the error statistics are computed locally in time for each assimilation cycle, by performing a sequence of ensemble forecasts around the current state of the system (while they are assumed constant in their study). And second, the probability distribution is assumed to be a truncated Gaussian distribution (as proposed by Lauvernet et al., 2009, as an improvement to the classical Gaussian hypothesis), in order to avoid the most extreme and non-physical parameter corrections. These two improvements are indeed found necessary to solve the more realistic assimilation problem at stake in this paper.

However, before explaining this in more detail, we first summarize in Sect. 2 the background existing elements that are used to perform the study: the ocean model, the assimilation method for parameter estimation and the Mercator-Ocean reanalysis data. Then, in Sect. 3, we present the details of the method that is used to perform the assimilation experiments: experimental setup and statistical parameterization. And finally, in Sect. 4, we discuss and interpret the results, focusing on the accuracy of the mixed layer thermohaline characteristics and on the relevance of the parameter estimates.

2 Background

In this section, we present the three existing ingredients that are used later as a background information to set up our assimilation system (Sect. 3) and to perform the experiments (Sect. 4): (i) the ocean model, focusing on the role of the atmospheric forcing parameters, (ii) the assimilation method, in order to introduce the various approximations and parameterizations that are needed to solve the problem, and (iii) the Mercator-Ocean reanalysis, from which the synthetic observations are extracted.

2.1 Ocean model

The OGCM used in this study is a global ocean configuration (ORCA2) of the NEMO-OPA model (Madec et al., 1998), using a $2^\circ \times 2^\circ$ ORCA type horizontal grid, with a meridional grid spacing reduced to $1/2^\circ$ in the tropical regions

[1] http://www.godae.org

[2] http://www.mersea.eu.org

[3] http://www.mercator-ocean.fr

Ocean Modelling 27 (2009) 1–17

Contents lists available at ScienceDirect

Ocean Modelling

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

A truncated Gaussian filter for data assimilation with inequality constraints: Application to the hydrostatic stability condition in ocean models

Claire Lauvernet, Jean-Michel Brankart *, Frédéric Castruccio, Grégoire Broquet, Pierre Brasseur, Jacques Verron

LEGI/CNRS, BP53X, 38041, Grenoble Cedex, France

ARTICLE INFO

Article history:
Received 24 April 2008
Received in revised form 17 October 2008
Accepted 21 October 2008
Available online 24 November 2008

Keywords:
Data assimilation
Inequality constraints
Reduced-order Kalman filter
Hybrid coordinate ocean models

ABSTRACT

In many data assimilation problems, the state variables are subjected to inequality constraints. These constraints often contain valuable information that must be taken into account in the estimation process. However, with linear estimation methods (like the Kalman filter), there is no way to incorporate optimally that kind of additional information. In this study, it is shown that an optimal filter dealing with inequality constraints can be formulated under the assumption that the probability distributions are truncated Gaussian distributions. The statistical tools needed to implement this truncated Gaussian filter are described. It is also shown how the filter can be adapted to work in a reduced dimension space, and how it can be simplified following several additional hypotheses. As an application, the truncated Gaussian assumption is shown to be adequate to deal with the condition of hydrostatic stability in ocean analyses. First, a detailed evaluation of the method is made using a one-dimensional z-coordinate model of the mixed layer: particular attention is paid to the parameterization of the probability distribution, the accuracy of the estimation and the sensitivity to the observation system. In a second step, the method is applied to a three-dimensional hybrid coordinate ocean model (HYCOM) of the Bay of Biscay (at a 1/15° resolution), to show that it is efficient enough to be applied to real size problems. These examples also demonstrate that the algorithm can deal with the hydrostatic stability condition in isopycnic coordinates as well as in z-coordinates.

1. Introduction

Linear estimation methods (like the Kalman filter) are either defined as the linear estimator minimizing the variance of the estimation error or based on the stronger assumption that the error probability distributions are Gaussian. This last assumption is indeed a sufficient condition for the error variance minimizing estimator to be linear and equal to the maximum probability estimator. Problems can occur if such linear methods are used to estimate a vector of variables subjected to inequality constraints. In such cases, the Gaussian assumption is not valid, and nothing prevents the best linear estimation from violating the inequality constraints. The information contained in the inequality constraints is not taken into account in the estimation process.

In many data assimilation problems, however, the state variables are subjected to inequality constraints that must be taken into account in the assimilation process because: (i) they can be an important source of information about the system, (ii) they are often a prerequisite for a valid model initial condition. For instance, in ocean systems, gravitational stability means that the Brunt-Väisälä frequency $N^2 = -\frac{g}{\rho}\frac{\partial \rho}{\partial z}$ (where ρ is the potential density referenced to the local in situ pressure) remains positive. This condition often contains valuable information that can improve the accuracy of the estimation. The reason is that in the ocean there is most often a top layer (the mixed layer) that is marginally stable, so that the ocean state is usually close to a number of hydrostatic stability inequality constraints. Moreover, in isopycnic coordinate ocean models (representing the ocean state as a stack of layers of prescribed density, e.g. Bleck, 2002), the hydrostatic stability condition is expressed in terms of a constraint of non-negativity of the layer thicknesses. Estimating an ocean state with non-negative layer thicknesses is obviously a prerequisite for initializing a model forecast. Ocean models are also subjected to other kinds of inequality constraints: non-negativity of tracer concentration, sea-water temperature above freezing point,... For these reasons, several solutions have been proposed in the literature for including inequality constraints in existing linear data assimilation schemes.

First, inequality constraints can be introduced in variational data assimilation methods by adding non-quadratic terms (i.e. with a nonlinear gradient) in the cost function to be minimized (e.g. Fujii et al., 2005). However, this amounts to changing the

* Corresponding author.
E-mail address: Jean-Michel.Brankart@hmg.inpg.fr (J.-M. Brankart).

1463-5003/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ocemod.2008.10.007

nature of the problem, replacing the strong constraints by weak constraints. Moreover, this solution requires using a variational algorithm to determine the minimum of a non-quadratic cost function.

A second way of dealing with inequality constraints is to make a nonlinear change of variables (anamorphosis) that transforms the constrained variables x to unconstrained variables x' (Bertino et al., 2002). For instance, for a non-negative variable $x \geqslant 0$ (like a concentration or a layer thickness), one possible solution is to use a logarithmic transformation (e.g. $x' = \log_{10} x$) and assume a Gaussian distribution in the logarithmic space. The problem with this method is that it is not easy to find the change of variables that leads to a reasonable assumption. If we assume, for example, that x' is distributed like $\mathcal{N}(0,1)$, the mean $x' = 0$ corresponds to a unitary distance ($x = 1$) with respect to the constraint ($x \geqslant 0$), the left tail of the Gaussian corresponds to probability tending to zero close to the constraint (typically $x' < -3$ or $x < 10^{-3}$), and the right tail of the Gaussian corresponds to probability tending to zero far from the constraint (typically $x' > 3$ or $x > 10^3$). The resulting assumption for x is therefore zero probability close to the constraint ($x \sim 0$), and significant probability for very large values ($x \sim 10^3$). We will see that such features are not very appropriate for the application considered in this paper.

A third solution proposed in the literature for including inequality constraints in linear schemes is to apply the linear scheme as it is and to introduce an adjustment operator to restore the constraints whenever needed. In their assimilation system, Brankart et al. (2003) introduced an adjustment operator (see Section 5) to restore non-negative layer thicknesses in an isopycnic ocean model (MICOM) after the observational update. More recently, in a hybrid coordinate ocean model (HYCOM), Thacker (2007) proposed using the correlations contained in the background error covariance matrix to move the ocean state from the linear solution to an adjusted solution with non-negative layer thicknesses. His solution is similar to that proposed by Simon and Simon (2005) who solved a quadratic programming problem to compute the adjusted solution (see Section 2.6.1). However, the difficulty with these adjustment operations is that they are based on heuristic considerations, and are only used to compensate for problems resulting from the linearity assumption.

Apart from these transformed linear schemes, there is also the possibility of using more general nonlinear estimation methods like the particle filters (e.g. Van Leeuwen, 2003; Losa et al., 2003; Vossepoel and Van Leeuwen, 2007). Since there is no hypothesis about the shape of prior probability distributions (approximated by an ensemble of model states), the particle filter can potentially deal with any kind of problem and, among other things, solves the problem of dealing with inequality constraints. However, describing a general distribution with sufficient accuracy requires using large ensembles, which involves numerical costs that can become prohibitive for practical ocean data assimilation problems.

In this paper, we choose to develop an optimal nonlinear scheme like the particle filter, with the difference that one prior assumption is made about the shape of the probability distributions: we assume that they are truncated Gaussian distributions (i.e. basically Gaussian distributions truncated by the inequality constraints). Consequently, the resulting estimation will be nonlinear in the input data and the maximum probability estimator will become different from the error variance minimizing estimator. A side effect of these developments is that we will be able to interpret the solution proposed by Simon and Simon (2005) or Thacker (2007) as an approximation of the optimal scheme (in the maximum probability estimator version). In this way, this study can also be seen as an attempt to provide a more precise theoretical basis for the adjustment operations proposed in these studies. These theoretical aspects are presented in Section 2 of this paper.

With regard to applications of the method, we will seek to determine whether it is able to deal adequately with the condition of hydrostatic stability in ocean analyses, using twin experiments (i) with a one-dimensional z-coordinate model of the ocean mixed layer (Sections 3 and 4) and (ii) with a three-dimensional hybrid coordinate model of the Bay of Biscay (Section 5). With the one-dimensional model, we will be able to describe in detail the parameterization of the probability distributions (Section 3) and the observational update of truncated Gaussian distributions (Section 4), while, with the three-dimensional model, only synthetic results will be presented (Section 5).

2. Methodology

The purpose of this section is to show that an optimal filter dealing with inequality constraints can be formulated, under the assumption that the error probability distributions are truncated Gaussian. For that purpose, we first give the definition and main properties of truncated Gaussian probability distributions (Section 2.1). Then we present the truncated Gaussian filter (TG filter), with the basic hypothesis sustaining it (Section 2.2). In the two following (Sections 2.3 and 2.4), we describe the statistical tools that are needed to effectively implement the filter. And finally, we show how the filter can work in a reduced dimension space (Section 2.5), or can be approximated using additional assumptions (Section 2.6).

2.1. Truncated Gaussian probability distributions

The truncated normal probability distribution can be defined from the normal distribution as follows (Tallis, 1965; Robert, 1995). If a random vector $\mathbf{x}$ is Gaussian with probability distribution $p(\mathbf{x}) = \mathcal{N}_{\mathbf{x}}(\tilde{\mathbf{x}}, \widetilde{\mathbf{C}})$, where $\tilde{\mathbf{x}}$ and $\widetilde{\mathbf{C}}$ are the expected value and covariance of $\mathbf{x}$, then, the posterior probability distribution of $\mathbf{x}$ with the condition that a set of linear inequality constraints $I : \mathbf{Ax} \leqslant \mathbf{b}$ are satisfied is called a *truncated Gaussian probability distribution*. It is entirely described by the parameters $\tilde{\mathbf{x}}$ and $\widetilde{\mathbf{C}}$ of the original Gaussian process together with the inequality constraints I, and can be noted:

$$p(\mathbf{x}|I) = \mathcal{TN}_{\mathbf{x}}(\tilde{\mathbf{x}}, \widetilde{\mathbf{C}}; I)$$

$\tilde{\mathbf{x}}$ and $\widetilde{\mathbf{C}}$ are sometimes called the location vector and the scale matrix of the truncated Gaussian distribution. The truncated Gaussian probability density function (pdf) is proportional to the Gaussian pdf for $\mathbf{Ax} \leqslant \mathbf{b}$ and equal to zero for $\mathbf{Ax} > \mathbf{b}$:

$$\mathcal{TN}_{\mathbf{x}}(\tilde{\mathbf{x}}, \widetilde{\mathbf{C}}; I) = \begin{cases} \frac{1}{\alpha}\mathcal{N}_{\mathbf{x}}(\tilde{\mathbf{x}}, \widetilde{\mathbf{C}}) & \mathbf{Ax} \leqslant \mathbf{b} \\ 0 & \mathbf{Ax} > \mathbf{b} \end{cases} \tag{1}$$

The normalizing factor α can be determined by integrating the Gaussian pdf over the region defined by the inequality constraints I:

$$\alpha = \int_{\mathbf{Ax} \leqslant \mathbf{b}} \mathcal{N}_{\mathbf{x}}(\tilde{\mathbf{x}}, \widetilde{\mathbf{C}})\, d\mathbf{x} \tag{2}$$

The expected value $\bar{\mathbf{x}}$ and covariance $\mathbf{C}$ of the truncated Gaussian distribution $\mathcal{TN}(\tilde{\mathbf{x}}, \widetilde{\mathbf{C}}; I)$ can be determined by integrating the Gaussian distribution over the region defined by the inequality constraints I:

$$\bar{\mathbf{x}} = \frac{1}{\alpha} \int_{\mathbf{Ax} \leqslant \mathbf{b}} \mathbf{x}\mathcal{N}_{\mathbf{x}}(\tilde{\mathbf{x}}, \widetilde{\mathbf{C}})\, d\mathbf{x} \tag{3}$$

$$\mathbf{C} = \frac{1}{\alpha} \int_{\mathbf{Ax} \leqslant \mathbf{b}} (\mathbf{x} - \bar{\mathbf{x}})\mathcal{N}_{\mathbf{x}}(\tilde{\mathbf{x}}, \widetilde{\mathbf{C}})(\mathbf{x} - \bar{\mathbf{x}})^T\, d\mathbf{x} \tag{4}$$

Efficient Parameterization of the Observation Error Covariance Matrix for Square Root or Ensemble Kalman Filters: Application to Ocean Altimetry

JEAN-MICHEL BRANKART, CLÉMENT UBELMANN, CHARLES-EMMANUEL TESTUT,*
EMMANUEL COSME, PIERRE BRASSEUR, AND JACQUES VERRON

LEGI/CNRS-Grenoble Universités, Grenoble, France

(Manuscript received 20 June 2008, in final form 10 October 2008)

ABSTRACT

In the Kalman filter standard algorithm, the computational complexity of the observational update is proportional to the cube of the number y of observations (leading behavior for large y). In realistic atmospheric or oceanic applications, involving an increasing quantity of available observations, this often leads to a prohibitive cost and to the necessity of simplifying the problem by aggregating or dropping observations. If the filter error covariance matrices are in square root form, as in square root or ensemble Kalman filters, the standard algorithm can be transformed to be linear in y, providing that the observation error covariance matrix is diagonal. This is a significant drawback of this transformed algorithm and often leads to an assumption of uncorrelated observation errors for the sake of numerical efficiency. In this paper, it is shown that the linearity of the transformed algorithm in y can be preserved for other forms of the observation error covariance matrix. In particular, quite general correlation structures (with analytic asymptotic expressions) can be simulated simply by augmenting the observation vector with differences of the original observations, such as their discrete gradients. Errors in ocean altimetric observations are spatially correlated, as for instance orbit or atmospheric errors along the satellite track. Adequately parameterizing these correlations can directly improve the quality of observational updates and the accuracy of the associated error estimates. In this paper, the example of the North Brazil Current circulation is used to demonstrate the importance of this effect, which is especially significant in that region of moderate ratio between signal amplitude and observation noise, and to show that the efficient parameterization that is proposed for the observation error correlations is appropriate to take it into account. Adding explicit gradient observations also receives a physical justification. This parameterization is thus proved to be useful to ocean data assimilation systems that are based on square root or ensemble Kalman filters, as soon as the number of observations becomes penalizing, and if a sophisticated parameterization of the observation error correlations is required.

1. Introduction

In atmospheric or oceanic applications of the Kalman filters, the growing number of available observations often leads to a prohibitive cost of the observational update (analysis step), and to the necessity of simplifying the problem. Ad hoc solutions must be found to make the problem numerically tractable. One first option is to synthesize the observational information by aggregating observations in superobservations, or even by dropping the least useful or most redundant measurements (data thinning). Another option is to transform the original algorithm and reduce its computational complexity by taking advantage of prior hypotheses on the error statistics (i.e., on the shape of the state and observation error covariance matrices). Simplifications are thus applied on the error second-order statistical moments (which are anyway only approximately known) rather than on the observations themselves. Of course, these two options are not mutually exclusive; they can interact with and complement each other. As explained in Rabier (2006), the need for data thinning can also result from over simplistic assumptions in the parameterization of the observation error covariance matrix. For instance, with a suboptimal scheme neglecting observation error correlations, decreasing the observation density can help improving the accuracy of

* Current affiliation: MERCATOR-Ocean, Toulouse, France.

Corresponding author address: Jean-Michel Brankart, LEGI/CNRS, BP53X, 38041 Grenoble CEDEX, France.
E-mail: Jean-Michel.Brankart@hmg.inpg.fr

DOI: 10.1175/2008MWR2693.1

the estimation (Liu and Rabier 2002, 2003). In this paper, we propose to reduce the numerical cost of the observational update by using simplified (but rather general) parameterizations of the observation error covariance matrix. The expected consequence is that, with improved efficiency, together with sufficient accuracy and robustness in the representation of the observation error covariance matrix, this method can substantially reduce the need for data thinning.

If the forecast error covariance matrix is available in square root form, as in square root or ensemble Kalman filters, it is possible to use a modified observational update algorithm (proposed by Pham et al. 1998), whose computational complexity is linear in the number of observations (instead of being cubic in the standard formula), providing that the observation error covariance matrix can be inverted at low cost, as for instance if it is diagonal. It is the purpose of this paper to introduce specific parameterizations of the observation error correlations that preserve the numerical efficiency of that modified algorithm. This can be done (i) by expressing the observation error covariance matrix as the sum of a diagonal and a low rank matrix, or (ii) by applying a linear transformation to the observation vector (and assuming uncorrelated observations in the transformed space). It is interesting to note that, in parameterization ii, nonsquare transformation matrices are possible, which means that the observation vector can be augmented with new observations that are linear combinations of the original observations. Both parameterizations are presented in section 2 of this paper. In section 3, a specific choice of linear transformation, consisting of augmenting the observation vector with gradients of the original observations, is studied in more detail.

In section 4, the algorithm is applied to ocean altimetric observations, as simulated by a 1/4° model of the tropical Atlantic Ocean, and focusing on the North Brazil Current. It is known indeed that altimetric observation errors are spatially correlated, because, for example, of orbit errors or atmospheric correction errors. Moreover, these correlations are important to take into account, because they can directly improve the quality of the observational update (especially for the dynamic height gradient, and thus for velocities), and the accuracy of the associated error estimates. In the North Brazil Current, the ratio between signal amplitude (about 5 cm) and typical observational noise (about 4 cm) remains moderate: the signal is marginally observed; this example is thus appropriate to show the importance of accounting for error correlations to reconstruct the ocean circulation, and to check the validity of our simplified parameterizations.

2. Parameterization of the observation error covariance matrix

In data assimilation problems, the observation error $\boldsymbol{\epsilon}$ is defined as the difference between the observation vector $\mathbf{y}$ (size y) and the observation counterpart in the true state $\mathbf{x}^t$ (size x):

$$\mathbf{y} = \mathbf{H}\mathbf{x}^t + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{H}(y \times x)$ is the observation operator. The specification of the observation error statistics thus always requires defining properly the truth of the problem (Cohn 1997; Kalnay 2003), which generally amounts to identifying the exact scope of the estimation problem. In atmospheric or oceanic applications, this is usually done by restricting the range of resolved scales in space and time, using for instance a filtering or averaging operator acting on the continuous state of the atmospheric or oceanic system. Observation error thus not only includes a measurement error, but also a representation error that results from this restriction in the scope of the problem. In this paper, it is assumed that the total observation error $\boldsymbol{\epsilon}$ is characterized by a zero mean $\langle\boldsymbol{\epsilon}\rangle = 0$ (unbiased observations) and a known covariance matrix $\mathbf{R} = \langle\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathrm{T}}\rangle$. Our purpose is to introduce efficient approximate parameterizations of this known observation error covariance matrix for use in square root or ensemble Kalman filters.

a. *Observational update in square root or ensemble Kalman filters*

In Kalman filters, the standard formula to compute the observational update increment $\delta\mathbf{x}$ of the model state vector is

$$\delta\mathbf{x} = (\mathbf{H}\mathbf{P}^f)^{\mathrm{T}}(\mathbf{H}\mathbf{P}^f\mathbf{H}^{\mathrm{T}} + \mathbf{R})^{-1}\delta\mathbf{y}, \tag{2}$$

where $\delta\mathbf{y} = \mathbf{y} - \mathbf{H}\mathbf{x}^f$ is the innovation vector, representing the difference between the observation vector $\mathbf{y}$ (size y) and the model equivalent to the observation in the forecast state vector $\mathbf{x}^f$ (size x), and $\mathbf{P}^f(x \times x)$ is the forecast error covariance matrix. The computational complexity (leading behavior for large x and y) of this standard formula is

$$C_0 \sim \frac{y^3}{6} + xy. \tag{3}$$

In the computation of C_0, it is assumed that a linear system is solved to compute $(\mathbf{H}\mathbf{P}^f\mathbf{H}^{\mathrm{T}} + \mathbf{R})^{-1}\delta\mathbf{y}$, with

Efficient Adaptive Error Parameterizations for Square Root or Ensemble Kalman Filters: Application to the Control of Ocean Mesoscale Signals

JEAN-MICHEL BRANKART, EMMANUEL COSME, CHARLES-EMMANUEL TESTUT,*
PIERRE BRASSEUR, AND JACQUES VERRON

LEGI/CNRS-Grenoble Universités, Grenoble, France

(Manuscript received 4 June 2009, in final form 15 September 2009)

ABSTRACT

In Kalman filter applications, an adaptive parameterization of the error statistics is often necessary to avoid filter divergence, and prevent error estimates from becoming grossly inconsistent with the real error. With the classic formulation of the Kalman filter observational update, optimal estimates of general adaptive parameters can only be obtained at a numerical cost that is several times larger than the cost of the state observational update. In this paper, it is shown that there exists a few types of important parameters for which optimal estimates can be computed at a negligible numerical cost, as soon as the computation is performed using a transformed algorithm that works in the reduced control space defined by the square root or ensemble representation of the forecast error covariance matrix. The set of parameters that can be efficiently controlled includes scaling factors for the forecast error covariance matrix, scaling factors for the observation error covariance matrix, or even a scaling factor for the observation error correlation length scale.

As an application, the resulting adaptive filter is used to estimate the time evolution of ocean mesoscale signals using observations of the ocean dynamic topography. To check the behavior of the adaptive mechanism, this is done in the context of idealized experiments, in which model error and observation error statistics are known. This ideal framework is particularly appropriate to explore the ill-conditioned situations (inadequate prior assumptions or uncontrollability of the parameters) in which adaptivity can be misleading. Overall, the experiments show that, if used correctly, the efficient optimal adaptive algorithm proposed in this paper introduces useful supplementary degrees of freedom in the estimation problem, and that the direct control of these statistical parameters by the observations increases the robustness of the error estimates and thus the optimality of the resulting Kalman filter.

1. Introduction

In Kalman filters, the accuracy of the estimated error covariances closely depends on the quality of the assumptions about model error and observation error statistics. Inaccurate parameterization may even lead to filter divergence, with error estimates becoming grossly inconsistent with the real error (Maybeck 1979; Daley 1991). To avoid this divergence, one recognized solution (adaptive filtering) is to determine the list of uncertain parameters in the model or observation error statistics, and try to adjust them using the actual differences between forecasts and observations (Daley 1992; Dee 1995; Wahba et al. 1995; Blanchet et al. 1997; Hoang et al. 1997; Wang and Bishop 2003; Lermusiaux 2007; Li et al. 2009). In particular, if the forecast and observation error probability distributions can be assumed Gaussian, a possible solution (proposed by Dee 1995) is to compute the maximum likelihood estimate of the adaptive parameters given the current innovation vector. This strategy is used and further developed in Mitchell and Houtekamer (2000) or Anderson (2007, 2009) in the more specific context of the ensemble Kalman filter. It is also this line of thought that is followed in this study to compute optimal estimates of adaptive statistical parameters.

A major difficulty with this kind of method is that, in general, the computational complexity of the parameter estimation is several times larger than the complexity of the estimation of the system state vector (i.e., than the classic observational update of the Kalman filter). The reason is that, in Kalman filters, the optimal state estimate is linear in the observation vector (of size y),

* Current affiliation: MERCATOR-Ocean, Toulouse, France.

Corresponding author address: Jean-Michel Brankart, LEGI/CNRS, BP53X, 38041 Grenoble CEDEX, France.
E-mail: jean-michel.brankart@hmg.inpg.fr

DOI: 10.1175/2009MWR3085.1

whereas the optimal parameter estimate is intrinsically nonlinear in the observation vector, so that the optimal solution must be computed iteratively (for instance using a downhill simplex method to find the maximum of the likelihood function, as in Mitchell and Houtekamer 2000). A first objective of this paper is to show that there exists nonetheless a few types of important parameters, for which a maximum likelihood optimal estimate can be computed at a numerical cost that is asymptotically negligible (for large y) with respect to that of the standard Kalman filter observational update. Second, taking advantage of this small additional computational complexity, the method is extended to condition the current parameter estimation on the full sequence of past innovations, which amounts to solving an additional (nonlinear) filtering problem for the unknown statistical parameters.

Furthermore, in square root or ensemble Kalman filters, the forecast error covariance matrix is always available in square root form, making it possible to use a modified observational update algorithm [proposed by Pham et al. (1998), as one of the essential elements defining the singular evolutive extended Kalman (SEEK) filter algorithm], whose computational complexity is linear in the number of observations, instead of being cubic as in the standard formula. Originally, this modified algorithm requires that the observation error covariance matrix be diagonal, but solutions exist to preserve its numerical efficiency (linearity in y) in presence of observation error correlations, as shown by Brankart et al. (2009), who also give a detailed comparison of the modified versus the original algorithms. In the present paper, we first show in section 2 how the optimal adaptive filtering problem described above can be formulated in the framework of this modified square root algorithm. It is indeed in this framework that optimal parameter estimates can be computed at negligible additional numerical cost. This is shown in section 3, where the discussion focuses on the few types of parameters for which such computational efficiency is possible. These important parameters are (i) scaling factors for the forecast error covariance matrix, (ii) scaling factors for the observation error covariance matrix, and (iii) scaling factors for the observation error correlation length scale.

In section 4, this adaptive filter is applied to the problem of estimating the evolution of an ocean mesoscale signal using observations of the ocean dynamic topography. To demonstrate the behavior of the adaptive mechanism, idealized experiments are performed, in which the reference signal (the truth of the problem) is generated by a primitive equation ocean model and sampled to produce synthetic observations with known error statistics. In that way, it is possible to check that the method is able to produce accurate parameter estimates and to explore the ill-conditionned situations (inappropriate prior assumptions or uncontrollability of the parameters) in which adaptivity can be misleading.

2. Formulation of the problem

a. Nonadaptive statistics

Let us consider the problem of estimating the evolution of a system described by the state vector $\mathbf{x}(t)$, between times t_0 and t_{N+1}, given a set of observation vectors $\mathbf{y}_k$ at times t_k, $k = 1, \ldots, N (t_k < t_{k+1})$:

$$\mathbf{y}_k = \mathbf{H}_k \mathbf{x}_k + \boldsymbol{\epsilon}_k, \tag{1}$$

where $\mathbf{x}_k = \mathbf{x}(t_k)$, $\mathbf{H}_k$, and $\boldsymbol{\epsilon}_k$ are the state vector, the observation operator, and the observational error at time t_k, respectively. We also assume that we have information on the initial condition $\mathbf{x}_0 = \mathbf{x}(t_0)$ and optionally on dynamical laws governing the time evolution of $\mathbf{x}(t)$. In many situations, it is useful to solve the filtering problem, in which the estimation at time t is computed using only past information. This means that the information only needs to be propagated forward in time from t_0 to t_{N+1}, with discrete updates each time that an observation vector is available. In a probabilistic framework, this amounts to computing sequentially the following probability distributions:

$$\begin{aligned} &p_0(\mathbf{x}_0),\, p_1^f(\mathbf{x}_1),\, p_1^a(\mathbf{x}_1), \ldots, p_k^f(\mathbf{x}_k), \\ &p_k^a(\mathbf{x}_k), \ldots, p_N^f(\mathbf{x}_N),\, p_N^a(\mathbf{x}_N),\, p_{N+1}(\mathbf{x}_{N+1}), \end{aligned} \tag{2}$$

where $p_0(\mathbf{x}_0)$ and $p_{N+1}(\mathbf{x}_{N+1})$ are the initial and final probability distributions, and where $p_k^f(\mathbf{x}_k)$ and $p_k^a(\mathbf{x}_k)$ are the probability distributions at time t_k before and after that the observation vector $\mathbf{y}_k$ is taken into account (superscripts "f" and "a" stand for "forecast" and "analysis"). Since we solve a filtering problem, it is implicit that every probability distribution is conditioned on the past observations [i.e., $\mathbf{y}_{k'}$ with $k' < k$ for $p_k^f(\mathbf{x}_k)$, and $\mathbf{y}_{k'}$ with $k' \leq k$ for $p_k^a(\mathbf{x}_k)$].

In this study, it is assumed that every probability distribution of the sequence (2) is Gaussian:

$$p_k^f(\mathbf{x}_k) = \mathcal{N}(\mathbf{x}_k^f, \mathbf{P}_k^f) \quad \text{and} \quad p_k^a(\mathbf{x}_k) = \mathcal{N}(\mathbf{x}_k^a, \mathbf{P}_k^a), \tag{3}$$

where $\mathbf{x}_k^f$ and $\mathbf{x}_k^a$ are the expected forecast and analysis state vectors at time t_k, respectively, and where $\mathbf{P}_k^f$ and $\mathbf{P}_k^a$ are the corresponding forecast and analysis error covariance matrices, respectively. They are computed by repeating the following two steps in sequence from $k = 1$ to $k = N$. The forecast step computes $\mathbf{x}_k^f$ and $\mathbf{P}_k^f$

Efficient Local Error Parameterizations for Square Root or Ensemble Kalman Filters: Application to a Basin-Scale Ocean Turbulent Flow

JEAN-MICHEL BRANKART, EMMANUEL COSME, CHARLES-EMMANUEL TESTUT,* PIERRE BRASSEUR, AND JACQUES VERRON

LEGI/CNRS-Université de Grenoble, Grenoble, France

(Manuscript received 8 December 2009, in final form 18 June 2010)

ABSTRACT

In large-sized atmospheric or oceanic applications of square root or ensemble Kalman filters, it is often necessary to introduce the prior assumption that long-range correlations are negligible and force them to zero using a local parameterization, supplementing the ensemble or reduced-rank representation of the covariance. One classic algorithm to perform this operation consists of taking the Schur product of the ensemble covariance matrix by a local support correlation matrix. However, with this parameterization, the square root of the forecast error covariance matrix is no more directly available, so that any observational update algorithm requiring this square root must include an additional step to compute local square roots from the Schur product. This computation generates an additional numerical cost or produces high-rank square roots, which may deprive the observational update from its original efficiency. In this paper, it is shown how efficient local square root parameterizations can be obtained, for use with a specific square root formulation (i.e., eigenbasis algorithm) of the observational update. Comparisons with the classic algorithm are provided, mainly in terms of consistency, accuracy, and computational complexity. As an application, the resulting parameterization is used to estimate maps of dynamic topography characterizing a basin-scale ocean turbulent flow. Even with this moderate-sized system (a 2200-km-wide square basin with 100-km-wide mesoscale eddies), it is observed that more than 1000 ensemble members are necessary to faithfully represent the global correlation patterns, and that a local parameterization is needed to produce correct covariances with moderate-sized ensembles. Comparisons with the exact solution show that the use of local square roots is able to improve the accuracy of the updated ensemble mean and the consistency of the updated ensemble variance. With the eigenbasis algorithm, optimal adaptive estimates of scaling factors for the forecast and observation error covariance matrix can also be obtained locally at negligible additional numerical cost. Finally, a comparison of the overall computational cost illustrates the decisive advantage that efficient local square root parameterizations may have to deal simultaneously with a larger number of observations and avoid data thinning as much as possible.

1. Introduction

A common difficulty with ensemble Kalman filters is to obtain the required ensemble size for representing weak correlations with sufficient accuracy. Small-sized ensembles systematically overestimate the information that observations contain about the system, and give undue confidence to suboptimal posterior estimates (ensemble collapse). In realistic atmospheric or oceanic applications of the ensemble Kalman filter, the horizontal dimensions of the system are generally much larger than the correlation length scales, so that most elements of the correlation matrix (coupling distant state variables) are close to zero, and the required rank for the error covariance matrices can be significantly larger than any affordable ensemble size. To circumvent this difficulty, the usual solution (Houtekamer and Mitchell 1998, 2001; Anderson 2003; Brusdal et al. 2003; Brankart et al. 2003; Testut et al. 2003; Ott et al. 2004; Szunyogh et al. 2005; Corazza et al. 2007; Hunt et al. 2007; Bishop and Hodyss 2009; Yang et al. 2009) is to add the prior information that long range correlations are negligible and force them to zero in the forecast ensemble covariance matrix. In that way, the rank of the matrix is artificially increased, and any spurious long-range influence of the observations is removed. To produce such local

* Current affiliation: Mercator-Ocean, Toulouse, France.

Corresponding author address: Jean-Michel Brankart, LEGI/CNRS, BP53X, 38041 Grenoble CEDEX, France.
E-mail: jean-michel.brankart@hmg.inpg.fr

DOI: 10.1175/2010MWR3310.1

parameterizations of the error covariances, the accepted reference algorithm (proposed by Houtekamer and Mitchell 2001) modifies the ensemble covariance matrix by a Schur product with a local support correlation matrix (this method is commonly referred to as localization of the ensemble covariance). It can be shown indeed that this preserves the positive definiteness of the original matrix, and that it is appropriate for a valid and well-conditioned observational update (see Houtekamer and Mitchell 2001; Hamill et al. 2001, for more details).

However, with this parameterization, the square root of the forecast error covariance matrix is no more directly available, so that any observational update algorithm requiring this square root must include an additional step to compute local square roots from the Schur product. This computation generates an additional numerical cost or produces high-rank square roots, which may deprive the observational update from its remarkable efficiency. Algorithms taking benefit from the direct availability of a low-rank square root have been initially developed for square root Kalman filters (Verlaan and Heemink 1997; Pham et al. 1998; Lermusiaux and Robinson 1999), and have also been introduced in ensemble Kalman filters (Bishop et al. 2001; Heemink et al. 2001; Tippett et al. 2003, see also section 2b). The chief benefit that can be obtained from these algorithms is that the computational complexity of the observational update can be made linear in the number of observations. This is very useful to reduce the need for data thinning and allow the optimal analysis of a larger number of observations (see e.g., Brankart et al. 2009, who also show how this advantage can be preserved in presence of observation error correlations). In addition, in a recent study, Brankart et al. (2010) have shown that a variant of these algorithms (the eigenbasis observational update) can produce an optimal adaptive estimate (optimal in the sense of Dee 1995) of some important statistical parameters at negligible numerical cost. With this variant, it is also directly possible to avoid adding random perturbations to the observations (a virtue that is shared by the serial processing algorithm of Whitaker et al. 2008). There is thus a significant benefit to expect if these properties of the eigenbasis observational update could be preserved with localized error covariance matrices. Proposing possible answers to this question is the main objective of the present paper.

One possible way of computing the required local square roots is the direct factorization of the Schur product (as recently proposed by Bishop and Hodyss 2009). In this way, the computation of the local square roots is very efficient, but high-rank square roots are produced, which may be detrimental to the efficiency of the eigenbasis algorithm (especially with adaptive statistics). On the other hand, local implementations of algorithms working with local square roots have already been developed [for the singular evolutive extended Kalman filter in Brankart et al. (2003); Testut et al. (2003), or for the ensemble transform Kalman filter in Ott et al. (2004); Hunt et al. (2007); Yang et al. (2009)], but they do not rely on the Schur product by a correlation matrix to localize the ensemble covariances. By introducing low-rank local square roots producing the same effect as the Schur product, a subsidiary result of our paper (see section 3b) is also to suggest a possible link between these two approaches.

In section 2, we start by describing how the eigenbasis observational update algorithm (linear in the number of observations, with costless optimal adaptive parameter estimates) can be used to perform ensemble observational update (as an alternative to the original algorithm of the ensemble Kalman filter), and we propose a comparison of their respective merits, mainly in terms of computational complexity. In section 3, we present compatible local square root parameterizations of the forecast error covariance matrix, as an approximation to the parameterization of Houtekamer and Mitchell (2001). An academic example is used to illustrate the effect of the various schemes, which are also compared in terms of computational complexity. And finally, in section 4, we provide a more realistic application of the scheme by analyzing synthetic altimetric observations resulting from a long-term (100 yr) model simulation of a basin-scale ocean turbulent flow. We study in particular why local parameterizations are needed (in terms of accuracy and efficiency), how accurately the exact solution can be approached, and how local adaptive parameter estimates can be obtained.

2. Efficient ensemble observational update

a. Classic ensemble observational update

In the ensemble Kalman filter (Evensen 1994; Evensen and van Leeuwen 1996), the time propagation of the probability distribution for the state of the system is approximated by an ensemble model forecast, which is meant to provide a random sample of the system probability distribution at any future time. Each time that a new observation vector $\mathbf{y}$ is available, this probability distribution must be updated to take into account the new observational information. For that purpose, Gaussianity is assumed and the ensemble forecast is updated so that its mean and covariance are consistent with the result given by the standard Kalman filter observational update formula. Thus, if the ensemble forecast is noted $\mathbf{x}_i^f$, $i = 1, \ldots, m$ (where m is the size of the ensemble) with mean and covariance:

Ocean Sci., 8, 121–142, 2012
www.ocean-sci.net/8/121/2012/
doi:10.5194/os-8-121-2012

Ocean Science

Towards an improved description of ocean uncertainties: effect of local anamorphic transformations on spatial correlations

J.-M. Brankart[1], C.-E. Testut[2], D. Béal[1], M. Doron[1], C. Fontana[1], M. Meinvielle[1], P. Brasseur[1], and J. Verron[1]

[1]LEGI/CNRS, UMR5519, Grenoble, France
[2]Mercator-Océan, Toulouse, France

Correspondence to: J.-M. Brankart (jean-michel.brankart@hmg.inpg.fr)

Received: 20 September 2011 – Published in Ocean Sci. Discuss.: 28 October 2011
Revised: 23 February 2012 – Accepted: 24 February 2012 – Published: 6 March 2012

Abstract. The objective of this paper is to investigate if the description of ocean uncertainties can be significantly improved by applying a local anamorphic transformation to each model variable, and by making the assumption of joint Gaussianity for the transformed variables, rather than for the original variables. For that purpose, it is first argued that a significant improvement can already be obtained by deriving the local transformations from a simple histogram description of the marginal distributions. Two distinctive advantages of this solution for large size applications are the conciseness and the numerical efficiency of the description. Second, various oceanographic examples are used to evaluate the effect of the resulting piecewise linear local anamorphic transformations on the spatial correlation structure. These examples include (i) stochastic ensemble descriptions of the effect of atmospheric uncertainties on the ocean mixed layer, and of wind uncertainties or parameter uncertainties on the ecosystem, and (ii) non-stochastic ensemble descriptions of forecast uncertainties in current sea ice and ecosystem pre-operational developments. The results indicate that (i) the transformation is accurate enough to faithfully preserve the correlation structure if the joint distribution is already close to Gaussian, and (ii) the transformation has the general tendency of increasing the correlation radius as soon as the spatial dependence between random variables becomes nonlinear, with the important consequence of reducing the number of degrees of freedom in the uncertainties, and thus increasing the benefit that can be expected from a given observation network.

1 Introduction

As a result of inescapable inaccuracies or approximations in the observations and in the models, uncertainties are inherent to any description or simulation of the real ocean. A realistic and efficient modelling of these uncertainties is of key importance for many oceanographic applications: (i) to objectively check simulation results against independent observations, (ii) to optimally assimilate data, and thus obtain the maximum benefit from an expensive, but incomplete, observing system, and (iii) to rationaly design future observation networks. It is thus essential to the production and use of ocean operational data, as delivered for instance by the MyOcean system[1], which is the target application of this study.

Ensemble (or Monte Carlo) methods provide a good way of describing uncertainties in ocean dynamical systems, by explicitly exploring how uncertainties in the governing laws, parameters or forcings (the prior information) propagate to the observed quantities or to the operational products (Palmer et al., 2005; Lermusiaux, 2006). However, even if an explicit stochastic modelling is used to solve a practical problem, there is often a strong temptation (in large size applications) to simplify the result using a Gaussian model, because it is much more efficient (i) to describe the uncertainties (by the mean and covariance), and (ii) to assimilate observations (using linear update formulas, as in the ensemble Kalman filter, see Evensen and van Leeuwen, 1996). Without a prior assumption about the shape of the probability distribution, large size problems are indeed very complex in general (van Leeuwen, 2009, 2010; Bocquet et al., 2010), mainly because the size of the sample that is required to identify a general multivariate distribution increases exponentially with

[1]http://www.myocean.eu.org/

Published by Copernicus Publications on behalf of the European Geosciences Union.

the number of dimensions (curse of dimensionality). To circumvent this difficulty, one possible simplification is to look for univariate nonlinear changes of variables (anamorphosis transformations) transforming the marginal distribution of each random variable into a Gaussian distribution. One-dimensional probability distributions can indeed be identified with a much smaller sample, and it may well happen that such a separate transformation for each random variable also helps improving the Gaussianity of their joint distribution (although this needs to be checked in every practical application). This technique originates from geostatistics (Wackernagel, 2003) and was first introduced in oceanography by Bertino et al. (2003), in the framework of the ensemble Kalman filter.

However, the studies presented in Bertino et al. (2003) and later in Simon and Bertino (2009) were directly focused on the impact that anamorphic transformations may have on the performance of the ensemble Kalman filter, without much emphasis on the improvements in the multivariate statistics. In this context, they also preferred to apply the same transformation over the whole model domain (but different for each model variable), so that a much larger sample is available to identify the transformation function. Yet, if the objective is also to propose a generic method (beyond the Gaussian scheme) to improve the description of the uncertainties, which can be spatially inhomogeneous, any practical possibility of extending this towards local anamorphic transformations should be evaluated. In a recent paper, Béal et al. (2010) proposed a very simple algorithm to obtain such local transformations, and started evaluating its potential for describing a 30-day ensemble forecast of the North-Atlantic ecosystem (simulating the effect of wind uncertainties). However, the paper was exclusively focused on the improvement of local correlations (at given locations) between phytoplankton and the other ecosystem compartments (nutrients, zooplankton), in the perspective of ocean colour data assimilation. Yet, with an algorithm working locally (i.e. transforming each model grid point with a different anamorphosis function), it is also important to study how the spatial correlations are modified, and hopefully improved, by the transformation.

The purpose of the present paper is thus to evaluate the effect of local anamorphic transformations on spatial correlations for various kinds of ocean uncertainties. The study includes, on the one hand, the stochastic ensemble description of the ocean mixed layer response to atmospheric forcing uncertainties (Sect. 3), the ecosystem response to wind uncertainties (i.e. the same application as in Béal et al., 2010, in Sect. 4), and the ecosystem response to parameters uncertainties (Sect. 5). On the other hand, we also show examples of anamorphic transformations applied to the non-stochastic ensemble description of forecast uncertainties in current pre-operational developments for the sea ice component (Mercator system, Sect. 6) and for the ecosystem component (MyOcean project, Sect. 7). In addition, before going to the applications, the paper includes a brief summary of the algorithm (presented in a more deductive way than in Béal et al., 2010), with a quantitative discussion of the computational complexity and accuracy of the approximation (Sect. 2).

2 Anamorphosis transformations

The basic problem of the algorithm is to look for a nonlinear change of variable transforming a random variable X with known cumulative distribution function (cdf) $F(x) = p(X \leq x)$ into a new random variable Z with the target cdf $G(z) = p(Z \leq z)$. Elementary probability calculus (e.g. Von Mises, 1964) provides a general solution for the forward and backward transformations:

$$Z = G^{-1}[F(X)] \quad \text{and} \quad X = F^{-1}[G(Z)] \tag{1}$$

providing that F and G are invertible. In particular, if $Z \sim \mathcal{U}(0,1)$ is uniformly distributed on the interval $[0,1]$, with $G(z) = z$, then $x = F^{-1}(k/q)$ is the kth q-quantile of X; and if $Z \sim \mathcal{N}(0,1)$ is normally distributed, with $G(z) = \frac{1}{2}[1 + \mathrm{erf}(z/\sqrt{2})]$, then Eq. (1) defines the forward and backward Gaussian anamorphosis transformation of the random variable X (Wackernagel, 2003, chapter 33).

However, it is important to keep in mind that transforming all variables of a random vector using Eq. (1) can only ensure that the marginal distribution of each variable becomes Gaussian. This does not imply that their joint probability distribution becomes a multivariate Gaussian distribution, which is the condition required to apply linear estimation techniques. As pointed out by Wackernagel (2003), it is thus important to check in practice that at least bivariate distributions of the transformed variables become close to bi-Gaussian, so that linear inference may be close to optimal. It is the purpose of the present paper to check this in various oceanic applications, by studying how the transformation in Eq. (1), applied separately for every random variable, at every spatial location, modifies the spatial correlation structure. But before going to the applications, this section is dedicated to describing the specific algorithm that we have implemented to approximate Eq. (1) using a limited-size sample of the random variables.

2.1 Efficient approximate algorithm

In the Monte Carlo estimation methods (like the ensemble Kalman filter), the prior probability distribution for the control variables is only approximately described by a finite-size sample. The anamorphosis transformation in Eq. (1) for each control variable can thus only be approximately computed from the available sample using a nonparametric estimate $\tilde{F}(x)$ of the exact marginal cdf $F(x)$. The most simple nonparametric estimate of a probability density function (pdf) $\tilde{f}(x) = d\tilde{F}(x)/dx$ is the histogram (Izenman, 2008, chapter 4): a piecewise constant pdf $\tilde{f}(x)$, or a piecewise linear cdf $\tilde{F}(x)$. As a simple choice for the classes of the

Ocean Modelling 66 (2013) 64–76

Contents lists available at SciVerse ScienceDirect

Ocean Modelling

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

Impact of uncertainties in the horizontal density gradient upon low resolution global ocean modelling

Jean-Michel Brankart

CNRS/Univ. Grenoble 1, Laboratoire de Glaciologie et Géophysique de l'Environnement (LGGE) UMR 5183, Grenoble, F-38041, France

ARTICLE INFO

Article history:
Received 13 September 2012
Received in revised form 6 February 2013
Accepted 9 February 2013
Available online 14 March 2013

Keywords:
Model uncertainties
Stochastic parameterization
Equation of state
Global ocean model

ABSTRACT

In this study, it is shown (i) that, as a result of the nonlinearity of the seawater equation of state, unresolved scales represent a major source of uncertainties in the computation of the large-scale horizontal density gradient from the large-scale temperature and salinity fields, and (ii) that the effect of these uncertainties can be simulated using random processes to represent unresolved temperature and salinity fluctuations. The results of experiments performed with a low resolution global ocean model show that this parameterization has a considerable effect on the average large-scale circulation of the ocean, especially in the regions of intense mesoscale activity. The large-scale flow is less geostrophic, with more intense associated vertical velocities, and the average geographical position of the main temperature and salinity fronts is more consistent with observations. In particular, the simulations suggest that the stochastic effect of the unresolved temperature and salinity fluctuations on the large-scale density field may be sufficient to explain why the Gulf Stream pathway systematically overshoots in non-stochastic low resolution ocean models.

1. Introduction

One of the most salient feature of today's state-of-the-art ocean models is that they are essentially *deterministic* models, in the sense that they do not involve *random numbers* to represent uncertainties in the model equations, parameters and forcing, or to simulate the effect of unresolved processes. Yet, this deterministic model dynamics is known to become chaotic as soon as mesoscale eddies are resolved by the model, so that the simulated mesoscale flow can only be viewed as one random realization sampled from a large set of possibilities. It is thus only in a statistical sense that the mesoscale can be compared to the real world, and it is only as a stochastic process that the effect of the mesoscale in the model can be analysed. Mesoscale fluctuations indeed produce a considerable effect on the general circulation of the ocean (Zhai et al., 2004; Penduff et al., 2010), with prominent contributions to momentum, heat and salt fluxes, which cannot be easily parameterized in low resolution models.

As a general rule, the effect of uncertainties or unresolved processes (even if unbiased) does not average to zero in a nonlinear model. For instance, if the wind is fluctuating or if it is uncertain, then neglecting the fluctuations or the uncertainties systematically underestimates the air–sea momentum flux (proportional to the square of the wind speed). In the same way, the average effect of the mesoscale fluctuations does not vanish in the two nonlinear terms of the primitive equations: the advection term and the equation of state. Concerning the advection term, this effect was originally parameterized in ocean models using empirically specified horizontal diffusion (Bryan et al., 1979), and afterwards using more and more sophisticated advection/diffusion operators (see Griffies et al., 2000 for a review). Concerning the equation of state, the effect of the mesoscale temperature and salinity fluctuations on the large-scale density field is generally ignored, maybe because it cannot be easily parameterized using a deterministic formulation. However, it can easily be argued (see Section 3.1), that, in low resolution ocean models, the resulting approximation in the large-scale density is a major source of uncertainties in the horizontal pressure gradient, and thus in the horizontal momentum balance equation.

A different point of view can also be adopted to deal with model uncertainties. Rather than parameterizing their mean effect in the model, they can be explicitly simulated by including a random forcing in the model equations. This can be done to produce ensemble forecasts (Buizza et al., 1999; Palmer et al., 2005) or to simulate model error in ensemble data assimilation methods (Evensen, 1994). In such applications, the random forcing is not only responsible for the dispersion of the ensemble; it can also produce a significant mean effect in the simulations (Berner et al., in press; Williams, 2012; Palmer, 2012). In this study, the same kind of approach is used to simulate the uncertainties that unresolved mesoscale temperature and salinity fluctuations produce on the

E-mail address: Jean-Michel.Brankart@hmg.inpg.fr

1463-5003/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ocemod.2013.02.004

large-scale horizontal density field. The objective is to propose a simple (empirically specified) stochastic parameterization of these uncertainties (in Section 3), and to evaluate the impact that this parameterization may have on the ocean circulation (in Section 4), as simulated by a low resolution global model configuration (described in Section 2).

2. A low resolution global ocean model

The purpose of this section is to present the NEMO primitive equation model and to describe the ORCA2 low resolution global ocean configuration.

2.1. The NEMO primitive equation model

The ocean general circulation model that is used in this study is the NEMO model (Nucleus for European Modelling of the Ocean), as described in Madec (2008). The model approximates the ocean circulation by the primitive equations:

- the momentum balance equation:

$$\frac{\partial \mathbf{U}_h}{\partial t} = -\left[(\nabla \times \mathbf{U}) \times \mathbf{U} + \frac{1}{2}\nabla(\mathbf{U}^2)\right] - f\mathbf{k} \times \mathbf{U}_h - \frac{1}{\rho_0}\nabla_h p + \mathbf{D}^U + \mathbf{F}^U \quad (1)$$

where t is time; $\mathbf{k}$, the local upward unit vector; $\mathbf{U}$, the velocity vector ($\mathbf{U}_h$ is the horizontal component, orthogonal to $\mathbf{k}$, and w, the vertical velocity); p is pressure; ρ_0, a reference density; and $f = 2\boldsymbol{\Omega} \times \mathbf{k}$, the Coriolis acceleration (where $\boldsymbol{\Omega}$ is the Earth angular velocity);

- the hydrostatic equilibrium equation:

$$\frac{\partial p}{\partial z} = -\rho g \quad (2)$$

where z is the vertical coordinate (in the direction of $\mathbf{k}$); ρ is *in situ* density; and g, gravitational acceleration;

- the incompressibility equation:

$$\nabla \cdot \mathbf{U} = 0 \quad (3)$$

- the heat and salt conservation equations:

$$\frac{\partial T}{\partial t} = -\nabla \cdot (T\mathbf{U}) + D^T + F^T \quad (4)$$

$$\frac{\partial S}{\partial t} = -\nabla \cdot (S\mathbf{U}) + D^S + F^S \quad (5)$$

where T is potential temperature and S, salinity;

- the equation of state:

$$\rho = \rho[T, S, p_0(z)] \quad (6)$$

where $p_0(z) = \rho_0 g z$ is the reference pressure as a function of depth.

In these equations, $\mathbf{D}^U, D^T$ and D^S represent the parameterization of small-scale physics for momentum, temperature and salinity, and $\mathbf{F}^U, F^T$ and F^S are surface forcing terms.

These equations are complemented by boundary conditions, which are applied at the ocean bottom and at the interface with the atmosphere. Kinematic conditions consist in a 'no flow' condition across the ocean bottom:

$$w = -\mathbf{U}_h \cdot \nabla_h H \quad (7)$$

where H is ocean depth, and a prognostic equation for the sea surface height η:

$$\frac{\partial \eta}{\partial t} = -\nabla \cdot \left[(H + \eta)\overline{\mathbf{U}}_\mathbf{h}\right] + \mathbf{P} - \mathbf{E} \quad (8)$$

where $\overline{\mathbf{U}}_\mathbf{h}$ is the vertical average of horizontal velocity; P, precipitation; and E, evaporation. Dynamic boundary conditions parameterize the exchange of momentum and heat across the bottom and surface boundaries. Since they depend on the parameterization used for $\mathbf{D}^U$ and D^T, they will be described later in Section 2.2.

From Eqs. (2) and (8), it results that the horizontal pressure gradient $\nabla_h p$ in Eq. (1) is given by:

$$\nabla_h p = \nabla_h p_s + \int_{\zeta=z}^{\zeta=0} g \nabla_h \rho \, d\zeta \quad (9)$$

where $p_s = \rho_s g \eta$ is the surface pressure gradient, and ρ_s is surface density. Thus the horizontal pressure gradient depends on the thermohaline structure of the ocean (T and S) through the equation of state in Eq. (6). In realistic applications of NEMO, the equation of state is the standard empirical equation defined by the Joint Panel on Oceanographic Tables and Standards (UNESCO, 1983), in a version that has been reformulated by Jackett and McDougall (1995) (by a modification of the coefficients of the K polynomial in the equation below), to allow direct computation of *in situ* density from potential temperature (rather than *in situ* temperature):

$$\rho(T, S, p) = \frac{\rho(T, S, 0)}{1 - p/K(T, S, p)} \quad (10)$$

where $\rho(T, S, 0)$ is a 15-term polynomial in T and S; and $K(T, S, p)$, a 26-term polynomial in T, S and p. One of the main characteristics of the seawater equation of state is thus to be quite nonlinear (see Fig. 1). In addition, it must be remembered that, in principle, it is only valid for a fluid parcel in thermodynamic equilibrium.

2.2. The ORCA2 configuration

The NEMO configuration used in this study is the ORCA2 configuration, as described in Madec and Imbard (1996). It is a low resolution configuration, which is provided with the model code (<http://www.nemo-ocean.eu/>), and which is used here exactly

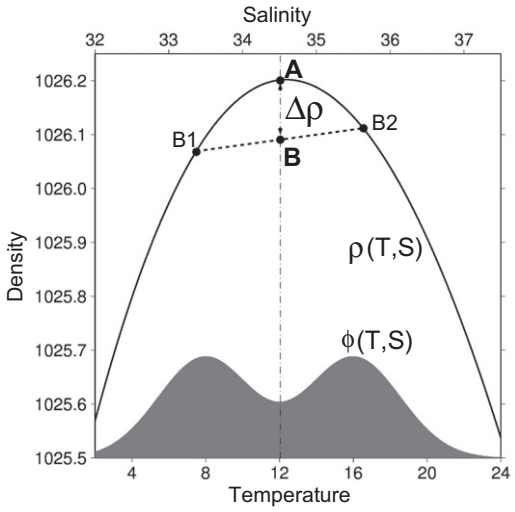


Fig. 1. Sea water equation of state (thick solid line) for joint temperature and salinity variations between 2 °C and 24 °C (bottom axis) and 32 and 37.5 (top axis) respectively. A typical distribution of unresolved temperature and salinity fluctuations is represented by the grey histogram, which superposes two Gaussian distributions with means at $T = 8$ °C, $S = 33.5$ and $T = 16$ °C, $S = 35.5$, and identical standard deviations: $\sigma_T = 2.5$ °C, $\sigma_S = 0.625$. The density at point A is computed by applying the equation of state to the mean of the distribution: $T = 12$ °C and $S = 34.5$, whereas the density at point B takes into account the distribution of unresolved temperature and salinity fluctuations. Points B1 and B2 show that the same density can be obtained as the mean of two densities obtained from opposite temperature and salinity fluctuations.

Annexe C

Curriculum Vitæ

Jean-Michel BRANKART

Né à Waremme,
le 16 septembre 1969

Nationalité : Belge

Etat civil : célibataire

Adresse professionnelle :
Laboratoire de Glaciologie et de
Géophysique de l'environnement
BP53X, F-38041 Grenoble Cedex
Tel : +33 4 76825034
Fax. +33 4 76825271
E-mail : Jean-Michel.Brankart
@legi.grenoble-inp.fr

Diplômes universitaires

1992 : Ingénieur Civil Mécanicien (Mécanique Physique), Université de Liège.

Stage au service de Thermomécanique des Phénomènes Irréversibles, sous la direction du Professeur G. Lebon (janvier à mars 2001).

1993 : DEA européen en modélisation de l'environnement marin, des Universités de Liège, Paris VI, Corse, Lisbonne, Las Palmas, et des Iles Baléares.

Stage de modélisation de l'environnment marin, Université P. & M. Curie (Paris VI), Station zoologique de Villefrance-sur-Mer, sous la direction du Professeur P. Nival (février 1993).

1996 : Docteur en Sciences appliquées (Docteur-Ingénieur), Université de Liège.

Parcours professionnel

1992–1996 : Assistant-professeur (36 mois) et **ingénieur de recherche** (16 mois) au service de Mécanique des Fluides Géophysiques de l'Université de Liège, sous la direction du Professeur J. Nihoul.

Responsabilité d'enseignement universitaire : travaux dirigés des cours de Mécanique Analytique et de Systèmes Non-linéaires (pour étudiants ingénieurs).

Co-investigateur pour les projets européens MERMAIDS, MODB et OMEGA.

1997–1998 : Postdoctorat (bourse de mobilité 'TMR' de la Commission Européenne) à l'"Istituto per lo studio delle metodologie geofisiche ambientali" au centre du CNR à Bologne, sous la direction du Dr. Nadia Pinardi.

Participation à l'école : "Ocean Modeling and Parameterization" (NATO Advanced Study Institute) au Centre de Physique des Houches (France) en janvier 1998.

1999–2001 : Postdoctorat au LEGI à Grenoble dans le cadre du projet européen DIADEM, sous la direction du Dr. Jacques Verron.

2001–2012 : Ingénieur de recherche au CNRS (IR2) au Laboratoire des Ecoulements Géophysiques et Industriels (LEGI) à Grenoble, sous la direction du Dr. Jacques Verron.

2012–2013 : Ingénieur de recherche au CNRS (IR1) au Laboratoire de Galciologie et de Géophysique de l'Environnement (LGGE) à Grenoble, sous la direction des Dr. Jacques Verron et Pierre Brasseur.

Mon activité d'ingénieur de recherche

Mon activité d'ingénieur de recherche peut se décomposer, à peu près à parts égales, selon trois composantes principales : le développement, l'encadrement et les projets scientifico-techniques. Ce sont ces trois composantes que je vais maintenant décrire successivement.

Développements méthodologiques et techniques

En premier lieu, je suis responsable du développement de la maintenance et de l'exploitation des instruments d'analyse et de prévision océanique de l'équipe MEOM.

1. Les **aspects méthodologiques** de cette activité forment l'épine dorsale de ce mémoire. Ils comportent principalement :
   - le développement de méthodes et d'outils de résolution de **problèmes inverses** et d'**assimilation de données** (chapitres 6 et 7) ;
   - le développement de méthodes et d'outils de **paramétrisations stochastiques** et de **simulations d'ensemble** (chapitres 3 et 2).
2. Les **aspects techniques** de cette activité sont discutés plus en détail dans l'annexe A. Ils comportent principalement :
   - le développement du **logiciel SESAM** (en collaboration successivement avec C.-E. Testut, L. Parent, E. Cosme, C. Lauvernet et F. Castruccio) ;
   - le développement du **logiciel OSMIUM** (en collaboration avec L. Gaultier) ;
   - le développement de modules de paramétrisation stochastique et de simulation d'ensemble pour le **modèle NEMO**.

Encadrement d'étudiants en thèse et de postdoctorants

Deuxièmement, je suis responsable du volet technique de l'encadrement des étudiants en thèse et des postdoctorants qui utilisent ces systèmes (sous la direction de Jacques Verron et/ou Pierre Brasseur). Mais cette activité a aussi impliqué de fait un co-encadremment méthodologique et scientifique de ces recherches, qui a été spécialement important durant la période des mandats administratifs de J. Verron (directeur du LEGI de 2000 à 2004) et de P. Brasseur (directeur-adjoint du LEGI de 2005 à 2009). Ce fut le cas en particulier pour :

1. Etudiants en thèse :
   - **Léo Berline (2002–2005) :** Assimilation de données dans un modèle couplé physique-biogéochimie de l'océan Atlantique Nord, sous la direction de P. Brasseur et J. Verron (voir Berline et al., 2007).
   - **Grégoire Broquet (2003–2007) :** Caractérisation des erreurs de modélisation pour l'assimilation de données dans un modèle océanique régional du Golfe de Gascogne, sous la direction de P. Brasseur et J. Verron (voir Broquet et al., 2008).

- **Frédéric Castruccio (2003–2006) :** Apports des données gravimétriques GRACE pour l'assimilation de données altimétriques et in-situ dans un modèle de l'Océan Pacifique Tropical, sous la direction de J. Verron (voir Castruccio et al., 2006, 2008).
- **Clément Ubelmann (2005–2009) :** Etude de scénarios d'altimétrie satellitaire pour le contrôle de la circulation océanique dans l'océan Atlantique tropical par assimilation de données, sous la direction de J. Verron (voir Ubelmann et al., 2009, 2012; Brankart et al., 2009).
- **Aurélie Duchez (2007–2011) :** Contrôle du Courant Nord Méditerranéen dans le golfe du Lion : une approche par simulation du système d'observation, sous la direction de J. Verron (voir Duchez et al., 2012).
- **Marion Meinvielle (2008–2012) :** Ajustement optimal des paramètres de forçage atmosphérique par assimilation de données de température de surface pour des simulations océaniques globales, sous la direction de P. Brasseur et B. Barnier (voir Meinvielle et al., 2013; Brankart et al., 2012).
- **Lucile Gaultier (2011–2013) :** sous la direction de J. Verron et P. Brasseur (voir Gaultier et al., 2013).

2. Postdoctorants :
   - **Florence Birol (2002–2004) :** assimilation de données avec le modèle HYCOM, dans le cadre du projet TOPAZ, sous la direction de P. Brasseur (voir Birol et al., 2004, 2005).
   - **David Rozier (2003–2005) :** assimilation de données avec le modèle HYCOM, dans le cadre du projet PEA/SHOM, sous la direction de P. Brasseur (voir Rozier et al., 2007).
   - **Yann Ourmières (2004–2005) :** accompagnement scientifique du développement du schéma SAM-2 de Mercator, dans le cadre d'un projet d'Ingénieur Réseau Bleu, dont j'avais las responsabilité scientifique, en collaboration avec P. Brasseur (voir Ourmières et al., 2006).
   - **Sergey Skachko (2004–2006) :** développement d'une méthode de correction du forçage atmosphérique pour le modèle NEMO, dans le cadre d'une des tâches de recherche et développement du projet MERSEA, dont j'avais la responsabilité scientifique, sous la direction de J. Verron (voir Skachko et al., 2009).
   - **Claire Lauvernet (2005–2007) :** développement d'une méthode d'assimilation de données sous l'hypothèse de distributions gaussiennes tronquées pour le modèle HYCOM, dans le cadre du projet ONR/NICOP, sous la direction de P. Brasseur (voir Lauvernet et al., 2009).
   - **Frédéric Castruccio (2006–2007) :** développement d'une méthode d'assimilation de données sous l'hypothèse de distributions gaussiennes tronquées pour le modèle HYCOM, dans le cadre du projet ONR/NICOP, sous la direction de P. Brasseur (voir Lauvernet et al., 2009).
   - **David Béal (2007–2008) :** dévloppement de l'anamorphose pour l'assimilation de données dans le modèle couplé NEMO/LOBSTER, dans le cadre du projet MERSEA, sous la responsabilité de P. Brasseur (voir Béal et al., 2010; Brankart et al., 2012).
   - **Chafih Skandrani (2007–2009) :** développement d'une méthode de correction du forçage atmosphérique pour le modèle NEMO, dans le cadre d'une des tâches de recherche et développement du projet MERSEA, dont j'avais la responsabilité scientifique, sous la direction de J. Verron (voir Skandrani et al., 2009).
   - **Maéva Doron (2008–2012) :** estimation de paramètres bio-géochimiques dans le modèle NEMO/LOBSTER, dans le cadre du projet MyOcean, sous la direction de P. Brasseur (voir Doron et al., 2011, 2013).

- **Angélique Melet (2010–2011) :** assimilation de données et estimation de paramètres dans un modèle emboîté (NEMO/AGRIF), sous la direction de J. Verron (voir Melet et al., 2012).
- **Guillem Candille (depuis 2012) :** développement de méthodes de prévision d'ensemble pour l'assimilation de données altimétriques dans le modèle NEMO, dans le cadre des projets MyOcean2/SANGOMA, sous la direction de P. Brasseur.

Contribution à des projets scientifiques et techniques

Troisièmement, à travers la participation de l'équipe MEOM, j'ai été impliqué dans de nombreux projets de recherche à des degrés divers :

1. Projets internationaux :
   - **TOPAZ (2001–2004) :** *Towards an Operational Prediction system for the North Atlantic European coastal Zones.* Mon rôle principal, sous la responsabilité scientifique de P. Brasseur et J. Verron, était le développement du système d'assimilation, et la mise en œuvre d'une expérience pré-opérationnelle en temps réel pour l'Atlantique Nord.
   - **MERSEA (2004–2008) :** *Development of a European system for operational monitoring and forecasting of the ocean physics, biogeochemistry, and ecosystems, on global and regional scales.* Dans le cadre de ce projet, j'était responsable scientifique de la tâche de recherche et développement en assimilation de données (incluant notamment les postdocorats de S. Skachko et C. Skandrani), sous la supervision de J. Verron.
   - **Projet ONR/NICOP (2005-2007) :** *Development of a sequential data assimilation system for regional ocean predictions using HYCOM (US partner : Eric Chassignet, RSMAS/MPO).* Mon rôle principal était de développer des solutions pour adpater la méthode d'assimilation au problème posé (gaussiennes tronquées), sous la responsabilité scientifique de P. Brasseur.
   - **MyOcean/MyOcean2 (2009–2014) :** *Prototype Operational Continuity for the GMES Ocean Monitoring and Forecasting Service.* Mon rôle principal était de développer des solutions pour adpater la méthode d'assimilation au problème posé (anamorphose), sous la responsabilité scientifique de P. Brasseur.
   - **SANGOMA (2011–2015) :** *Stochastic Assimilation for the Next Generation Ocean Model Applications.* Mon rôle principal est de contribuer au partage des outils d'assimilation au sein du consortium SANGOMA (logiciel SANGOMA), et de particper à l'élaboration des méthodes que nous dévelopons (paramétrisation stochastique pour les simulations d'ensemble), sous la responsabilité scientifique de P. Brasseur.
   - **Projets OSTST :** Participation à divers projets d'accompagnement scientifique au développement des plateformes staellitaires d'observation altimétrique de l'océan, sous la responsabilité scientifique de P. Brasseur et/ou J. Verron.
2. Projets nationaux :
   - **Projet PEA/SHOM (2003–2006) :** *Développements de méthodes séquentielles d'assimilation de données d'ordre réduit par filtrage SEEK. Application à la modélisation côtière.* Mon rôle principal, sous la responsabilité scientifique de P. Brasseur, était le développement du système d'assimilation, et la gestion des expériences d'assimilation de données.
   - **Ingénieur réseau bleu Mercator (2004–2005) :** Projet d'accompagnement scientifique du développement du schéma SAM-2 (développement de la méthode d'Incremental Analysis Update), dont j'étais responsable scientifique.

- **Projets du Groupe Mission Mercator-Coriolis (2001–2012) :** Participation à divers projets d'accompagnement au développement des méthodes d'assimilation de données pour le projet Mercator, sous la responsabilité scientifique de P. Brasseur et/ou J. Verron.

3. Projets de calcul :
   - **Projets IDRIS :** Il s'agit de projets annuels récurrents intitulés : *"Développements de l'assimilation de données dans le modèle NEMO" (GENCI-IDRIS, grant 011279.)*, dont je suis le responsable scientifique (depuis 8 ans) et technique (depuis 12 ans). Notre allocation pour 2013 s'élève à 200 000 heures de calcul sur IBM-x3750-M4 (ada), et 120 To de stockage.

Arbitrage scientifique

Participation au comité de lecture de publications scientifiques pour les revues suivantes :

1. Journal of Marine Systems (1995).
2. Journal of Atmospheric and Oceanic Technology (1997).
3. Journal of Marine Systems (2001).
4. Journal of Marine Systems (2002).
5. Ocean Modelling (2003).
6. Annales Geophysicae (2004).
7. Quarterly Journal of the Royal Meteorological Society (2005).
8. Tellus A (2005).
9. Ocean Science (2006).
10. Ocean Dynamics (2006).
11. Ocean Science (2009).
12. Geophysical Research Letters (2010).
13. Journal of Atmospheric and Oceanic Technology (2011).
14. Ocean Science (2012).
15. Progress in Oceanography (2012).

Publications

A – Revues internationales avec comité de lecture

1. Brankart, J.-M. and Brasseur, P. (1996). Optimal analysis of in situ data in the Western Mediterranean using statistics and cross-validation. Journal of Atmospheric and Oceanic Technology, 16(2), 477-491.
2. Brasseur, P., Beckers, J.-M., Brankart, J.-M., and Schoenauen, R. (1996). Seasonal temperature and salinity fields in the Mediterranean Sea : Climatological analyses of an historical data set. Deep-Sea Research, 43(2),159-192.
3. Uu, D.-V. and Brankart, J.-M. (1997). Seasonal variation of temperature and salinity fields and water masses in the Bieng Dong (South China) Sea. Mathematical and Computer Modelling, 26(12), 97-113.
4. Brankart, J.-M. and Brasseur, P. (1998). The general circulation in the Mediterranean Sea : a climatological approach. Journal of Marine Systems, 18, 41-70.

5. Rixen, M., Beckers, J.-M., Brankart, J.-M., and Brasseur, P. (2000). A numerically efficient data analysis method with error map generation. Ocean Modelling, 2, 45-60.
6. Brankart, J.-M. and Pinardi, N. (2001). Abrupt cooling of the Mediterranean Levantine Intermediate Water at the beginning of the 1980s : observational evidence and model simulation. Journal of Physical Oceanography, 31(8), 2307-2320.
7. Carmillet, V., Brankart, J.-M., Brasseur, P., Drange, H., Evensen, G., and Verron, J. (2001). A singular evolutive extended Kalman filter to assimilate ocean color data in a coupled physical-biochemical model of the North Atlantic. Ocean Modelling, 3,167-192.
8. Beckers J.M., Rixen M., Brasseur P., Brankart J.M., Elmoussaoui A., Crépon M., Herbaut C., Martel F., Van den Berghe F., Mortier L., Lascaratos A., Drakopoulos P., Korres G., Nittis K., Pinardi N., Masetti E., Castellari S., Carini P., Tintore J., Alvarez A., Monserrat S., Parilla D., Vautard R., Speich S. (2002). Model intercomparison in the Mediterranean : MEDMEX simulations of the seasonal cycle. Journal of Marine Systems, 33-34, 215-251.
9. Brankart, J.-M., Testut, C.-E., Brasseur, P., and Verron, J. (2003). Implementation of a multivariate data assimilation scheme for isopycnic coordinate ocean models : Application to a 1993-96 hindcast of the North Atlantic Ocean circulation. Journal of Geophysical Research, 108(C3), 19(1-20).
10. Brusdal, K., Brankart, J.-M., Halberstadt, G., Evensen, G., Brasseur, P., van Leeuwen, P.-J., Dombrowsky, E., and Verron, J. (2003). A demonstration of ensemble-based assimilation methods with a layered OGCM from the perspective of operational ocean forecasting systems. Journal of Marine Systems, 40-41, 253-289.
11. Parent, L., Testut, C.-E., Brankart, J.-M., Verron, J., Brasseur, P., and Gourdeau, L. (2003). Comparative assimilation of Topex/Poseidon and ERS altimeter data and of TAO temperature data in the Tropical Pacific Ocean during 1994-1998, and the mean sea-surface height issue. Journal of Marine Systems, 40-41, 381-401.
12. Testut, C.-E., Brasseur, P., Brankart, J.-M., and Verron, J. (2003). Assimilation of sea-surface temperature and altimetric observations during 1992-1993 into an eddy permitting primitive equation model of the North Atlantic Ocean. Journal of Marine Systems, 40-41, 291-316.
13. Birol, F., Brankart, J.M., Castruccio, F., Brasseur, P. and Verron, J. (2004). Impact of ocean mean dynamic topography on satellite data assimilation. Marine Geodesy, 27, 59-78.
14. Birol, F., Brankart, J.M., Lemoine, J.M., Brasseur, P. and Verron, J. (2005). Assimilation of satellite altimetry referenced to the new GRACE geoid estimate. Geophysical Research Letters, 32(6), doi10.1029/2004GL021329.
15. Brasseur, P., Bahurel, P., Bertino, L., Birol, F., Brankart, J.-M., Ferry, N., Losa, S., Remy, E., Schröter, J., Skachko, S., Testut, C.-E., Tranchant, B., van Leeuwen, P.J., and Verron, J. (2005). Data assimilation for marine monitoring and prediction : the MERCATOR operational assimilation systems and the MERSEA developments. Quarterly Journal of the Royal Meteorological Society, 131, 3561-3582.
16. Castruccio F., Verron, J., Gourdeau, L., Brankart, J.M., and Brasseur, P. (2006). On the role of the GRACE mission in the joint assimilation of altimetric and TAO data in a Tropical Pacific Ocean model. Geophysical Research Letters, 33, L14616.
17. Ourmières, Y., Brankart, J.M., Berline, L., Brasseur, P., and Verron, J. (2006). Incremental Analysis Update implementation into a sequential data assimilation system. Journal of Atmospheric and Oceanic technology, 23(12), 1729-1744.

18. Berline L., Brankart, J.M., Brasseur, P., Ourmières, Y., et Verron, J. (2007). Improving the physics of a coupled physical-biogeochemical model of the North Atlantic through data assimilation : impact on the ecosystem, Journal of Marine Systems, 64(1-4), 153-172.
19. Raick, C., Alvera-Azcarate, A., Barth, A., Brankart, J.M., Soetart, K. and Grégoire, M. (2007). Application of a SEEK filter to a 1D biogeochemical model of the Ligurian Sea : twin experiments and real in situ data assimilation, Journal of Marine Systems, 65(1-4), 561-583.
20. Rozier, D., Birol, F., Cosme, E., Brasseur, P., Brankart, J.M. and Verron, J. (2007). A reduced order Kalman filter for data assimilation in physical oceanography. SIAM Rev., 49(3), 449-465.
21. Broquet, G., Brasseur, P., Rozier, D., Brankart, J.M. and Verron, J. (2008). Estimation of model errors generated by atmospheric forcings for ocean data assimilation : experiments in a regional model of the Bay of Biscay. Ocean dynamics, 58(1), 1-17.
22. Castruccio, F., Verron, J., Gourdeau, L., Brankart, J.M. and Brasseur, P. (2008). Joint altimetric and in-situ data assimilation using the GRACE mean dynamic topography : a 1993-1998 hindcast experiment in the Tropical Pacific Ocean. Ocean dynamics, 58(1), 43-63.
23. Ourmières, Y., Brasseur, P., Levy, M., Brankart, J.M. and Verron, J. (2009). On the key role of nutrient data to constrain a coupled physicalbiogeochemical assimilative model of the North Atlantic Ocean. Journal of Marine Systems, 75(1-2), 100-115.
24. Lauvernet C., Brankart J.M., Castruccio F., Broquet G., Brasseur P., Verron J. (2009). A truncated Gaussian filter for data assimilation with inequality constraints : application to the hydrostatic stability condition in ocean models. Ocean Modelling, 27, 1-17.
25. Skachko S., Brankart J.-M., Castruccio F., Brasseur P., Verron J. (2009). Improved turbulent air-sea flux bulk parameters for the control of the ocean mixed layer : a sequential data assimilation approach. Journal of Atmospheric and Oceanic Technologies. 26(3), 538-555.
26. Ubelmann C., Verron J., Brankart J.M., Cosme E., Brasseur P. (2009). Impact of upcoming altimetric missions on the prediction of the three-dimensional circulation in the tropical Atlantic ocean. Journal of Operational Oceanography, 2(1), 3-14.
27. Brankart J.M., Ubelmann C., Testut C.E., Cosme E., Brasseur P. and Verron J. (2009). Efficient parameterization of the observation error covariance matrix for square root or ensemble Kalman filters : application to ocean altimetry. Monthly Weather Review, 137(6), 1908-1927.
28. Skandrani C., Brankart J.-M., Ferry N., Verron J., Brasseur P. and Barnier B. (2009). Controlling atmospheric forcing parameters of global ocean models : sequential assimilation of sea surface Mercator-Ocean reanalysis data. Ocean Science, 5, 403-419.
29. Béal D., Brasseur P., Brankart J.-M., Ourmières Y. and Verron J. (2010). Characterization of mixing errors in a coupled physical biogeochemical model of the North Atlantic : implications for nonlinear estimation using Gaussian anamorphosis. Ocean Science, 6, 247-262.
30. Cosme E., Brankart J.-M., Verron J., Brasseur P. and Krysta M. (2010). Implementation of a reduced-rank, square-root smoother for high resolution ocean data assimilation. Ocean Modelling, 33, 87-100.

31. Brankart J.-M., Cosme E., Testut C.-E., Brasseur P. and Verron J. (2010). Efficient adaptive error parameterizations for square root or ensemble Kalman filters : application to the control of ocean mesoscale signals. Monthly Weather Review, 138(3), 932-950.

32. Brankart J.-M., Cosme E., Testut C.-E., Brasseur P. and Verron J. (2011). Efficient local error parameterizations for square root or ensemble Kalman filters : application to a basin-scale ocean turbulent flow. Monthly Weather Review, 139(2), 474-493.

33. Srinivasan A., Chassignet E.P., Bertino L., Brankart J.-M., Brasseur P., Chin M., Counillon F., Cummings J.A., Mariano A.J., Smedstad O.M. and Thacker W.C. (2011). A comparison of sequential assimilation schemes for ocean prediction with the HYbrid Coordinate Ocean Model (HYCOM) : Twin Experiments with static forecast error covariances. Ocean Modeling, 37(3-4), 85-111.

34. Doron M., Brasseur P. and Brankart J.-M. (2011). Stochastic estimation of biogeochemical parameters of a 3D ocean coupled physical-biogeochemical model : twin experiments. Journal of Marine Systems, 87, 194-207.

35. Titaud O., Brankart J.-M., and Verron J. (2011). On the use of Finite-Time Lyapunov Exponents and Vectors for direct assimilation of images in ocean models. Tellus A, 63(5), 1038-1051.

36. Melet A., Verron J. and Brankart J.-M. (2012). Potential outcomes of glider data assimilation in the Solomon sea : control of the water mass properties and parameter estimation. Journal of Marine Systems, 94, 232-246.

37. Brankart J.-M., Testut C.-E., Béal D., Doron M., Fontana C., Meinvielle M., Brasseur P. and Verron J. (2012). Towards an improved description of ocean uncertainties : effect of local anamorphic transformations on spatial correlations. Ocean Science, 8, 121-142.

38. Juza M., Penduff T., Brankart J.-M. and Barnier B. (2012). Estimating the distortion of mixed layer porperty distributions by the ARGO sampling. Journal of Operational Oceanography, 5(1), 45-58.

39. Ubelmann C., Verron J., Brankart J.-M., Brasseur P., and Cosme E. (2012). Assimilating altimetric data from multi-satellite scenarios to control Atlantic tropical instability waves : an observing system simulation experiments study. Ocean Dynamics, 62(6), 867-880.

40. Troupin C., Barth A., Sirjacobs D., Ouberdous M., Brankart J.-M., Brasseur P., Rixen M., Alvera-Azcárate A., Belounis M., Capet A., Lenartz F., Toussaint M.-E., and Beckers J.-M. (2012). Generation of analysis and consistent error fields using the Data Interpolating Variational Analysis (Diva). Ocean Modelling, 52-53, 90-101.

41. Freychet N., Cosme E., Brasseur P., Brankart J.-M. and Kpemlie E. (2012). Obstacles and benefits of the implementation of a reduced rank smoother with a high resolution model of the Atlantic ocean. Ocean Science, 8, 797-811.

42. Duchez A., Verron J., Brankart J.-M., Ourmières Y. and Fraunié P. (2012). Monitoring the Northern Current in the Gulf of Lions with an observing system simulation experiment. Scientia Marina, 76(3), 441-453.

43. Fontana C., Brasseur P., and Brankart J.-M. (2013). Toward a multivariate reanalysis of the North Atlantic ocean biogeochemistry during 1998-2006 based on the assimilation of SeaWiFS chlorophyll data. Ocean Science, 9, 37-56.

44. Doron M., Brasseur, P., Brankart J.-M., Losa S. N., and Melet A. (2013). Stochastic estimation of biogeochemical parameters from Globcolour ocean colour satellite data in a North Atlantic 3D ocean coupled physical-biogeochemical model. Journal of Marine Systems, 117-118, 81-95.
45. Brankart J.-M. (2013). Impact of uncertainties in the horizontal density gradient upon low resolution global ocean modelling. Ocean Modelling, 66, 64-76.
46. Gaultier L., Verron J., Brankart J.-M., Titaud O., and Brasseur P. (2013). On the inversion of submesoscale tracer fields to estimate the surface ocean circulation. Journal of Marine Systems, sous presse.

B – Thèses et Mémoires

1. Brankart, J.-M. (1992). Méthodes variationnelle et objective pour l'inversion de données océanographiques : analyse spectrale et optimisation. Mémoire d'ingénieur, Université de Liège. 104 pp.
2. Brankart, J.-M. (1993). Etude statistique de données océanographiques. Inversion par analyse objective et par fonctions splines. Mémoire (Diplôme européen d'approfondissement en modélisation de l'environnement marin), Universités de Liège, Paris VI, Corse, Lisbonne, Las Palmas, et des îles Baléares. 85 pp.
3. Brankart, J.-M. (1996). Modélisation statistique de l'hydrologie méditerranéenne. Validation et contrôle de qualité d'une climatologie de référence. Thèse de doctorat, Université de Liège. 209 pp. (http ://www-meom.hmg.inpg.fr/Web/pages-perso/brankart/These/).

C – Autres revues, actes de colloques et chapitres d'ouvrages

1. Brasseur, P. and Brankart, J.-M. (1993). Reconstruction of oceanic data fields and data assimilation. In Progress in Belgian Oceanographic Research, pages 9-22. Royal Academy of Belgium, IRMA publ.
2. Schoenauen, R., Brasseur, P., and Brankart, J.-M. (1994). Application of an analysis tool to a historical data set of the Mediterranean Sea. In Seabra-Santos, F. and Temperville, F., editors, Modelling of Coastal and Estuarine Processes, pages 235-243, Coimbra.
3. Brasseur, P. and Brankart, J.-M. (1995). The Mediterranean Oceanic Data Base. In MAST Days Proceedings. EC Publ. 18 pp.
4. Brasseur, P., Beckers, J.-M., Brankart, J.-M., and Schoenauen, R. (1995). Seasonal temperature and salinity fields in the Mediterranean Sea : Climatological analyses of a historical data set. In Rapports et Procès Verbaux du XXVème congrès-assemblée plénière de la Commission Internationale pour l'Exploration Scientifique de la Mer Méditerranée, volume 34, page 171. CIESM.
5. Brankart, J.-M., Brasseur, P., Rixen, M., Schoenauen, R., and Walrave, S. (1996). The MODB Project. In Progress in Belgian Oceanographic Research, pages 35-38. Royal Academy of Belgium, IRMA publ.
6. Beckers, J.-M., Rixen, M., and Brankart, J.-M. (1997). Computing synoptic T,S fields by relocating data points. In Progress in Oceanography of the Mediterranean Sea., pages 305-306, Rome. Marine science and technology programme, EC Publ.
7. Brankart, J.-M. and Brasseur, P. (1998). The impact of the MODB project upon Mediterranean modeling. In Bohle-Carbonell, M., editor, Experiences in Project Data Management, pages 227-259, Ispra. Marine science and technology programme, EC Publ.

8. Brankart, J.-M. and Pinardi, N. (1998a). The Levantine Intermediate Water interannual variability : data analysis and model simulation. In Lykousis, V. and Sakellariou, D., editors, 3rd MTP-II Workshop on the variability of the Mediterranean Sea, pages 17-19, Rhodos. Marine science and technology programme, EC Publ.
9. Brankart, J.-M. and Pinardi, N. (1998b). Long term variability of the Mediterranean large scale circulation. In Rapports du 35ème Congrès de la CIESM, volume 35(1), pages 130-131, Dubrovnik. CIESM.
10. Brankart, J.-M., Carmillet, V., Brasseur, P., Verron, J., Evensen, G., Drange, H., Brusdal, K., van Leeuwen, P.-J., and Halberstadt, G. (1999). Assimilation of remote sensing data in a coupled circulation and ecosystem model for the North Atlantic : the DIADEM project. In OCEANOBS 99 : The ocean observing system for climate, volume 2, Saint-Raphaël, France.
11. Brankart, J.-M., Testut, C.-E., Debreu, L., and Brasseur, P. (2000). Assimilation d'altimétrie satellitale et de température de surface dans un modèle aux équations primitives de l'océan Atlantique Nord. In Atelier de modélisation de l'atmosphère, pages 111-114, Toulouse. METEOFRANCE.
12. Piccinali, J.-G., Birol, F., Brankart, J.-M., Brasseur, P., and Verron, J. (2002). Multivariate assimilation of remote sensing and in situ data in a high resolution layered model of the North Atlantic : a contribution to the DIADEM and TOPAZ projects. International Symposium "En route to GODAE". Biarritz.
13. Birol, F., Brankart, J.-M., Piccinali, J.-G., Brasseur, P., and Verron, J. (2002). Projet TOPAZ : Assimilation multivariée dans une configuration Atlantique Nord d'un modèle d'océan à coordonnée verticale hybride. In Atelier de modélisation de l'atmosphère, pages 83-86, Toulouse. METEOFRANCE.
14. Verron, J., Blayo, Blum, J., Brankart, J.M., Brasseur, P., Chassignet, E., Cosme, E., Evensen, G., Gourdeau, L., Le Dimet, F.X., van Leeuwen, P.J. and Schröter, J. (2006). Advanced Altimeter Data Assimilation for Physical Ocean Prediction and Ecosystem Monitoring. Proceeding of the OSTST conference, Venice, 17 pp.
15. Berline, L., Brankart, J.-M., Brasseur, P., Ourmières, Y. and Verron, J. (2006). Un défi pour le couplage physico-biogéochimique dans Mercator : améliorer la physique des modèles couplés par assimilation de données. In : La lettre trimestrielle Mercator Océan n°20, pages 14-24, Toulouse. MERCATOR.
16. Skachko, S., Brankart, J.-M., Castruccio, F., Brasseur, P., and Verron, J. (2006). Air-sea fluxes correction by sequential data assimilation. In : Mercator Ocean Quarterly Newsletter 22, pages 24-28, Toulouse. MERCATOR.
17. Broquet, G., Brasseur, P., Rozier, D., Brankart, J.M. and Verron, J. (2006). Estimation de l'erreur modèle dans une configuration océanique régionale emboîtée pour l'assimilation de données. Proceedings du Colloque National sur l'Assimilation de Données, Toulouse, 4 pp.
18. Skachko, S., Brankart, J.M., Castruccio, F., Brasseur, P., and Verron, J. (2006). Estimation de paramètres de la fonction de forçage atmosphérique d'un modèle océanique par filtrage de Kalman. Proceedings du Colloque National sur l'Assimilation de Données, Toulouse, 4 pp.
19. Castruccio, F., Verron, J., Gourdeau, L., Brankart, J.-M., and Brasseur, P. (2007). GRACE : an improved MDT reference for altimetric data assimilation. In : Mercator Ocean Quarterly Newsletter 25, pages 20-31, Toulouse. MERCATOR.
20. Skandrani C., Brankart J.M., Ferry N., Verron J., Brasseur P. and Barnier B. (2009). Contrôle des paramètres gouvernant le forçage atmosphérique des modèles

d'océan par assimilation séquentielle d'observations de surface de la mer issues de la réanalyse MERCATOR. In Atelier de modélisation de l'atmosphère, Toulouse. METEOFRANCE, 9 pp.

21. Cosme, E., Brankart, J.M., Brasseur P. and Verron J. (2009). A data assimilation scheme for oceanic reanalyses : the SEEK smoother. In : Mercator Ocean Quarterly Newsletter 34, pages 14-19, Toulouse. MERCATOR.
22. Brankart J.M., Barnier B., Béal D., Brasseur P., Brodeau L., Broquet G., Castruccio F., Cosme E., Lauvernet C., Mathiot P., Meinvielle M., Molines J.M., Ourmières Y., Penduff T., Skachko S., Skandrani C., Ubelmann C. and Verron J. (2009). Is there a simple way of controlling the forcing function of the ocean ? In : Mercator Ocean Quarterly Newsletter 34, pages 20-26, Toulouse. MERCATOR.
23. Brankart J.-M., Skachko S., Castruccio F., Brasseur P., Verron J. (2009). Using sequential data assimilation to improve ocean forcing parameters. Note in : Bulletin of the American Meteorological Society, 90(9), 1264-1265.
24. Meinvielle M., Brasseur P., Brankart J.M., Barnier B., Penduff T. and Molines J.-M. (2011). Optimally improving the atmospheric forcing function of long-term global ocean simulations with sea-surface temperature observations. In : Mercator Ocean Quarterly Newsletter 42, pages 24-32, Toulouse. MERCATOR.
25. Fontana C., Brasseur P. and Brankart J.M. (2011). A multivariate reanalysis of the North Atlantic ocean biogeochemistry during 1998-2007 based on the assimilation of SeaWiFS data. In : Mercator Ocean Quarterly Newsletter 43, pages 10-19, Toulouse. MERCATOR.
26. Bouttier P.-A., Blayo E., Brankart J.-M., Brasseur P., Cosme E., Verron J. and Vidard A. (2012). Toward a data assimilation system for NEMO. In : Mercator Ocean Quarterly Newsletter 46, pages 24-30, Toulouse. MERCATOR.

D – Communications invitées

1. Brasseur P., L. Berline, J. M. Brankart et J. Verron, 2004 : Impact of SSH, SST and SSS data assimilation in a coupled physical-biogeochemical model of the North Atlantic, 35th COSPAR Scientifc Assembly, Paris, 19 juillet 2004.
2. Brasseur P., Bahurel P., Bertino L., Birol F., Brankart J. M., N. Ferry, S. Losa, E. Remy, J. Schroter, S. Skachko ;, C. E. Testut, B. Tranchant, J. Verron, P. J. Van Leeuwen, 2005 : Data assimilation in the MERCATOR/MERSEA operational ocean forecasting systems. International Symposium on assimilation of Observations in Meteorology and Oceanography, WMO, Prague, 18-22 avril 2005.
3. Brasseur P., D. Béal, J. M. Brankart, G. Broquet, F. Castruccio, E. Cosme, M. Doron, C. Lauvernet, Y. Ourmières, J. Verron, 2008 : Extensions non-Gaussiennes du filtre SEEK pour l'assimilation de données dans les modèles couplés physico-biogéochimiques de l'océan. CNA2008, Paris, 1-2 décembre 2008.
4. Cosme E., J. M. Brankart, J. Verron, M. Krysta et P. Brasseur, 2008 : Développement d'un lisseur pour l'assimilation de données océanographiques. CNA2008, Paris, 1-2 décembre 2008.
5. Verron J., O. Titaud, J. M. Brankart et P. Brasseur, 2009 : Assimilation of submesoscale data into ocean models using Lyapunov exponents. Conference on Lyapunov analysis, from theory to geophysical applications. ISC, 26-30 October 2009, Paris
6. Verron J., J. M. Brankart, O. Titaud, P. Brasseur, 2010 : Assimilation of submesoscale observations into ocean models, Conference on Altimetry for Oceans and Hydrology, 21-22 octobre 2010, Lisbonne.

7. Verron J., L. Gaultier, J. M. Brankart et P. Brasseur, 2012 : On the use of submesoscale tracer information for the improvement of altimetry-derived velocity fields. ESA GlobCurrents Conference, Brest, 7-9 mars 2012.
8. Brankart J.-M. (2013). Gagner de l'information sur la paramétrisation d'un modèle d'océan par assimilation de données. Journée thématique LEFE-MANU : "Que peuvent attendre les modélisateurs de l'assimilation de données ?", Paris, février 2013.
9. Brankart J.-M. (2013). Impact of uncertainties in the horizontal density gradient upon low resolution global ocean modelling. Workshop on Stochastic Modelling and Computing for Weather and Climate Prediction, Oxford, March 2013.
10. Verron J., L. Gaultier, N. Djath, J. M. Brankart et P. Brasseur, 2013 : On the inversion of submesoscales. Workshop A frontier in modern oceanography : Modeling, observing and assimilating submesoscale dynamics. Center for Prototype Climate Modeling, New York University Abu Dhabi, Abu Dhabi, 2-4 avril 2013.

E – Résumés de conférences

1. Beckers, J.-M., Brasseur, P., and Brankart, J.-M. (1994a). Month-to-month variability of the Western Mediterranean circulation : a combination of historical data and mathematical simulation. In AGU Ocean Sciences Meeting, volume 75(3) of Eos Transactions, page 226, San Diego.
2. Beckers, J.-M., Brasseur, P., and Brankart, J.-M. (1994b). Simulations numériques 3D de l'hydrodynamique de la Méditerranée. In Annales du troisième Congrès National Belge de Mécanique Théorique et Appliquée, Brussels.
3. Beckers, J.-M., Brasseur, P., and Brankart, J.-M. (1994c). Water, heat and salt budget in the Western Mediterranean Sea as computed by a combination of inverse models and primitive equation models. In Fluxes through straits and passages, CEE MAST workshop, Malaga.
4. Brankart, J.-M. and Brasseur, P. (1994). Cross-Validation : a tool for the statistical study of hydrological data : application to a historical data base of the Mediterranean Sea. In EGS Proceedings, volume 12 of Annales Geophysicae, page C232, Grenoble.
5. Brankart, J.-M., Brasseur, P., and Beckers, J.-M. (1994). Hydrology of the Mediterranean Sea : results of climatological analyses. In AGU Ocean Sciences Meeting, volume 75(3) of Eos Transactions, page 226, San Diego.
6. Brasseur, P., Schoenauen, R., Beckers, J.-M., and Brankart, J.-M. (1994). A multimodular framework for analysing historical data : hydrology of the Mediterranean Sea as a case study. In EGS Proceedings, volume 12 of Annales Geophysicae, page C232, Grenoble.
7. Brankart, J.-M. and Pinardi, N. (1998). Decadal and interannual variability in the Mediterranean Sea : model simulations and observations. In EGS Proceedings, volume 16 of Annales Geophysicae, page C594, Nice.
8. Demirov, E., Pinardi, N., De Mey, P., Brankart, J.-M., Bianco, L., and Castellari, S. (1999). Fifteen years circulation variability of the Mediterranean Sea. Results from model simulation and assimilation of in situ data. In MTP Workshop Proceedings, Perpignan, France. EC Publ.
9. Brankart, J.-M., Testut, C.-E., Brasseur, P., and Verron, J. (2001). Assimilation of sea surface temperature and altimetric observations in a MICOM configuration of the North Atlantic Ocean : results of hindcast and near real time experiments. Layered Ocean Model Workshop, Miami.

10. Tranchant B., Brasseur P., Brankart J.M., Piacentini A. and Testut C.-E. (2001). A new assimilation system for the French MERCATOR Operational Oceanographic project. AGU fall meeting, San Francisco.
11. Testut, C.-E., Brasseur, P., Brankart, J.-M., and Verron, J. (2002). A reduced-order Kalman filter to assimilate SSH, SST and SSS in a primitive equation model of the North Atlantic Ocean. In EGS scientific programme, page 177, Nice.
12. Brasseur P., Birol F., Blayo E., Brankart J.M., Debost F., Faugeras B., Galmiche M., Magri S., Penduff Th., Piccinali J.G., Robert C., Testut C.E. and Verron J. (2002). Advanced Data Assimilation for the Development of Operational Oceanography. 3rd EuroGOOS conference, Athens.
13. Verron J., Birol F., Blayo E., Brankart J.M., Brasseur P., Debost F., Delcroix T., Demey P., Durand F., Durbiano S., Evensen G., Galmiche M., Gourdeau L., Magri S., Molines J.M., Parent L., Penduff T., Pham D.T., Piccinali J.G., Schroeter J. and Testut C.E. (2002). Advanced altimeter data assimilation for the development of operational oceanography. Jason-1 SWT meeting, Biarritz.
14. Barth, A., Alvera-Azcarate, A., Rixen, M., Beckers, J.-M., Testut, C.-E., Brankart, J.-M., and Brasseur, P. (2003). Assimilation of sea surface temperature in a doubly, two-way nested primitive equation model of the Ligurian Sea. In EGS-AGU-EUG Joint Assembly, scientific programme, page 397, Nice.
15. Birol, F., Brankart, J.-M., Brasseur, P., and Verron, J. (2003). Multivariate assimilation into layered ocean models of the North Atlantic using the SEEK filter. In EGS-AGU-EUG Joint Assembly, scientific programme, page 399, Nice.
16. Debost, F., Brankart, J.-M., Brasseur, P., and Verron, J. (2003). Monitoring the mesoscale ocean circulation with multi-satellite altimetric missions : a data assimilation study. In EGS-AGU-EUG Joint Assembly, scientific programme, page 399, Nice.
17. Ourmières Y. and Brankart J.M. (2004). Incremental Analysis Update implementation for the SAM-2 configuration. GMMC meeting, Toulouse.
18. Birol, F., Brankart, J.-M., Brasseur, P., and Verron, J. (2004). Sensitivity of ocean general circulation prediction systems to mean dynamic topography. In AGU Ocean Sciences Meeting, volume 84(52) of Eos Transactions, Portland.
19. Verron, J., Brankart, J.-M., Brasseur, P., Berline, L., Castruccio, F., Ourmières, Y., and Skachko, S. (2004). Development of a global OPA data assimilation system for the MERSEA project. GODAE symposium, St Petersburg (Florida).
20. Ourmières Y., Brankart J.M., Berline L. and Brasseur P. (2005). Incremental Analysis Update implementation to an intermittent data assimilation system for ocean forecast. Christian Le Provost Colloquium, Toulouse.
21. Ourmières, Y., Castruccio, F., Brankart, J.-M., Brasseur, P., and Verron, J. (2005). Incremental Analysis Update implementation into an intermittent data assimilation system for Ocean General Circulation models. In EGS-AGU-EUG Joint Assembly, scientific programme, Vienne.
22. Berline L., Ourmières Y., Brankart J.M., Brasseur P. and Verron J. (2005). Assimilation of satellite data in a coupled physical-biogeochemical model of the North Atlantic at eddy-permitting resolution. AMEMR symposium, Plymouth.
23. Broquet G., Brasseur P., Brankart J.M., Rozier D. and Verron J. (2006). Estimation of model error covariance in a nested coastal model for multivariate data assimilation system. 15 years of progress in radar altimetry Symposium, Venice.

24. Castruccio F., Verron J., Gourdeau L., Brankart J.M. and Brasseur P. (2006). On the role of GRACE for the joint assimilation of altimetry and in-situ data. 15 years of progress in radar altimetry Symposium, Venice.

25. Skachko S., Berline L., Bertino L., Brankart J.M., Brasseur P., Ourmières Y., van Leeuwen P.J. and Verron J. (2006). Recent advances in data assimilation in the MERSEA project. 15 years of progress in radar altimetry Symposium, Venice.

26. Cosme E., Verron, J., Castruccio, F., Ourmières, Y., Robert, C., Skachko, S., Blayo, E., Brasseur, P. and Brankart, J.-M. (2006). Some recent advances in ocean data assimilation with the SEEK filter. OSTST meeting, Venice.

27. Ourmières Y., Brankart J.M., Berline L., Brasseur P. and Verron J. (2006). Implementation of a coupled physical biogeochemical model with data assimilation. MERSEA 3rd plenary meeting, Londres.

28. Skachko, S., Brankart, J.M., Castruccio, F., Brasseur, P., and Verron, J. (2006). Estimating the turbulent air-sea flux bulk parameters by sequential data assimilation. In EGS-AGU-EUG Joint Assembly, scientific programme, Vienne.

29. Ourmières Y. and Brankart J.M. (2006). Incremental Analysis Update implementation for the SAM-2 configuration. GMMC meeting, Toulouse, 6-7 octobre 2004, poster.

30. Brasseur P., Broquet G., Brankart J.M., Castruccio, F., Lauvernet C. and Verron J. (2008). Improving the Parameterization of Errors Statistics for Data Assimilation in a HYCOM Bay of Biscay regional configuration. Ocean Science Meeting. Orlando.

31. Cosme E., Krysta M., Brankart J.M., Verron J. and Brasseur P. (2008) A data assimilation method for reanalyses of the ocean circulation : the SEEK smoother. Ocean Science Meeting. Orlando.

32. Skandrani, C., Skachko, S., Brankart, J.-M., Brasseur, P. and Verron, J. (2008). Controlling the air-sea fluxes in a global oceanic model by assimilation of SST and SSS data. Ocean Science Meeting. Orlando.

33. Ubelmann C., Brankart J.M., Brasseur P., Cosme E. and Verron, J. (2008). Constraining the tropical Atlantic Ocean circulation by assimilating satellite altimetric observations : insights from observing system simulation experiments. Ocean Science Meeting. Orlando.

34. Brankart J.-M., Lauvernet C., Brasseur P., Castruccio F., Broquet G. and Verron J. (2008). A truncated Gaussian filter for data assimilation with inequality constraints. In EGU Proceedings, Vienne.

35. Brasseur P., Béal D., Brankart J.-M., Broquet G., Castruccio F., Cosme E., Doron M., Lauvernet C., Ourmières Y. and Verron J. (2008). Extensions non gaussiennes du filtre SEEK pour l'assimilation de données dans les modèles couplés physico-biogéochimiques de l'océan. Colloque National sur l'Assimilation de Données. Paris

36. Cosme E., Brankart J.-M., Brasseur P., Krysta M. and Verron J. (2008). Développement d'un lisseur pour l'assimilation de données océanographiques. Colloque National sur l'Assimilation de Données. Paris

37. Brankart J.M., Ubelmann C., Testut C.E., Cosme E., Brasseur P. and Verron J. (2009). Efficient parameterization of the observation error covariance matrix for square root or ensemble Kalman filters : application to ocean altimetry. THORPEX/WWRP workshop on 4D-VAR and ensemble Kalman filter inter-comparisons. Buenos Aires.

38. Demirov E., Brankart J.-M., Zhu J. and Picke-Thackray C. (2009). On the predictive skills of North Atlantic eddy permitting ocean model.. In EGU Proceedings, Vienne.
39. Srinivasan A., Chassignet E., Smedstad O.M., Chin T.M., Counillon F., Brankart J.-M., Brasseur P., Thacker W.C. and Cummings J.A. (2009). A comparison of sequential data assimilation schemes for ocean prediction with HYCOM : Twin Experiments. Layered Ocean Model Workshop, Miami.
40. Cosme E., Brankart J.-M., Brasseur P. and Verron, J. (2009). Implementation of a reduced rank smoother for high resolution oceanography. Fifth WMO International Symposium on Data Assimilation of Observations in Meteorology, Oceanography and Hydrology. Melbourne.
41. Verron, J., Brankart, J.-M., Cosme, E., Brasseur, P. and Titaud O. (2009). Linking Altimetry and Ocean Color : A Data Assimilation approach using Lyapunov exponents. Fifth WMO International Symposium on Data Assimilation of Observations in Meteorology, Oceanography and Hydrology. Melbourne.
42. Doron M., Béal D., Brankart J.-M. and Brasseur P. (2009). Accounting for non-linear and non-Gaussian behavior of ocean coupled physical-biogeochemical models to improve sequential data assimilation and parameters estimation. Fifth WMO International Symposium on Data Assimilation of Observations in Meteorology, Oceanography and Hydrology. Melbourne.
43. Verron J., Brankart J.-M., Titaud O. and Brasseur P. (2010). Assimilation of sub-mesoscale observations into ocean models. Conference on Altimetry for Oceans and Hydrology, Lisbon.
44. Brasseur P., Béal D., Brankart J.-M., Broquet G., Castruccio F., Doron M., Fontana C., Lauvernet C., Ourmières Y. and Verron J. (2010). Extensions non-gaussiennes du filtre SEEK pour l'assimilation de données dans les modèles couplés physico-biogéochimiques de l'océan. Colloque National sur l'Assimilation de Données. Grenoble.
45. Cosme E., Verron J., Brasseur P., Brankart J.-M., Blum J. and Auroux D. (2010). Problèmes de lissage dans un cadre Bayesien et solutions linéaires gaussiennes. Colloque National sur l'Assimilation de Données. Grenoble.
46. Doron M., Brasseur P., Brankart J.-M. and Fontana C. (2010). Réduction de l'incertitude sur trois paramètres principaux d'un modèle couplé physique-biogéochimique 3D de l'Atlantique Nord. Colloque National sur l'Assimilation de Données. Grenoble.
47. Duchez A., Verron J. and Brankart J.-M. (2010). Contrôle du courant nord méditerranéen par l'altimétrie côtière d'AltiKa : une approche par simulation du système d'observation (OSSE). Colloque National sur l'Assimilation de Données. Grenoble.
48. Fontana C., Brasseur P., Brankart J.-M. and Doron M. (2010). Assimilation séquentielle de données couleur de l'eau dans un modèle couplé physique-biogéochimique de l'Atlantique Nord. Colloque National sur l'Assimilation de Données. Grenoble.
49. Freychet N., Cosme E., Kpemlie E., and Brasseur P. (2010). Assimilation dans un modèle haute résolution de l'Atlantique Tropical. Colloque National sur l'Assimilation de Données. Grenoble.
50. Kpemlie E., Brankart J.-M., Freychet N., Cosme E., Brasseur P. and Verron J. (2010). Diagnostics pour une paramétrisation efficace du filtre adaptatif. Colloque National sur l'Assimilation de Données. Grenoble.

51. Meinvielle M., Brankart J.-M., Brasseur P. and Barnier B. (2010). Ajustement optimal de paramèrtres de forçage atmosphérique dans des simulations océaniques globales. Colloque National sur l'Assimilation de Données. Grenoble.
52. Melet A., Verron J. and Brankart J.-M. (2010). Glider assimilation in the Solomon Sea. Colloque National sur l'Assimilation de Données. Grenoble.
53. Testut C.-E., Greiner E., Garric G. and Brankart J.-M. (2010). Transformation gaussienne des statistiques d'erreur pour un filtre de Kalman de rang réduit : application à un système d'analyse de la glace de mer. Colloque National sur l'Assimilation de Données. Grenoble.
54. Titaud O., Brankart J.-M. and Verron J. (2010). On the use of finite-time Lyapunov exponents and vectors for direct assimilation of tracer images in ocean models. Colloque National sur l'Assimilation de Données. Grenoble.
55. Doron M., Béal D., Fontana C., Brasseur P. and Brankart J.-M. (2011). Data assimilation in 3D ocean coupled physical-bigeochemical models : state and parameter estimation, using nonlinear extensions of the Kalman filter. 2nd SARAL/AltiKa Science Workshop. Ahmedabad.
56. Doron M., Brasseur P., Brankart J.-M. and Fontana C. (2011). A North Atlantic 3D coupled physical-biogeochemical model : a stochastic approach to estimate biogeochemical parameters from ocean color data using a nonlinear and non-Gaussian framework. EGU General Assembly. Vienna.
57. Fontana C., Brasseur P., Brankart J.-M. and Doron M. (2011). Sequential ocean color data assimilation in a coupled physical-biogeochemical model of the North Atlantic. EGU General Assembly. Vienna.
58. Freychet N., Cosme E., Kpemlie E., Brasseur P., Brankart J.-M. and Verron J. (2011). Data assimilation with a reduced rank smoother. EGU General Assembly. Vienna.
59. Kpemlie E., Cosme E., Freychet N., Brankart J.-M. and Brasseur P. (2011). Adaptive parameterisation of error statistics in ensemble or reduced order square root filters. EGU General Assembly. Vienna.
60. Meinvielle M., Brasseur P., Brankart J.-M., Barnier B., Penduff T. and Molines J.-M. (2011). Optimal adjustment of atmospheric forcing parameters for long term simulations of the global ocean circulation. EGU General Assembly. Vienna.
61. Titaud O., Verron J. and Brankart J.-M. (2011). Opérateurs d'observations basés sur des Structures Lagrangiennes Cohérentes pour l'assimilation d'images océaniques. Journée Inversion et Assimilation d'Images, Télécom ParisTech, Paris.
62. Gaultier L., Verron J., Brasseur P. and Brankart J.-M. (2011). On the use of sub-mesoscale tracer information for the control of ocean circulations. 43rd International Liège Colloquium on ocean dynamics, Liège.

F – Rapports internes et documentation de logiciels

1. Beckers, J.-M., Brasseur, P., and Brankart, J.-M. (1994). Month-to-month variability of the general circulation fields in the Western Mediterranean Sea : Inventory of simulation results. Progress report, University of Liège, Liège. 240 pp.
2. Brankart, J.-M. (1994). The MODB local quality control. Technical report, University of Liège, Liège. 5 pp.
3. Brasseur, P., Brankart, J.-M., and Beckers, J.-M. (1994). Seasonal variability of general circulation fields in the Mediterranean Sea : Inventory of climatological analyses. Progress report, University of Liège, Liège. 221 pp.

4. Beckers, J.-M., Schmitz, F., Brasseur, P., Brankart, J.-M., Crépon, M., Herbaut, C., Martel, F., Van den Berghe, F., Lascaratos, A., Drakopoulos, P., Pinardi, N., Carini, P., Tintore, J., Alvarez, A., Parrilla, D., Vautard, R., and Speich, S. (1996). Mediterranean models evaluation experiment : MEDMEX annual report, E.U. concerted action. 32pp.
5. Brankart, J.-M., Brasseur, P., and Rixen, M. (1996). Climatological Atlas of the Mediterranean Sea. GHER technical report, University of Liège. 253 pp.
6. Brankart, J.-M. (1997). GrADS interactive navigator (GIN) : a collection of GrADS graphical modules. Technical report, IMGA, Bologna. 9 pp.
7. Brankart, J.-M. (1999a). Spinup simulation of the North Atlantic circulation with MICOM (DIADEM1 grid). DIADEM technical report, MEOM-LEGI, Grenoble. 18 pp.
8. Brankart, J.-M. (1999b). Representation of the pressure field in MICOM. Implication for sequential data assimilation. DIADEM technical report, MEOM-LEGI, Grenoble. 15 pp.
9. Brankart, J.-M. (1999c). A set of reading/writing routines implementing a standard structure for ocean NetCDF files. DIADEM technical report, MEOM-LEGI, Grenoble. 12 pp.
10. Brankart, J.-M. (2000). Assimilation of SSH/SST observations in the DIADEM coarse resolution model using the SEEK filter. DIADEM technical report, MEOM-LEGI, Grenoble. 7 pp.
11. Brankart, J.-M. (2001). Les filtres SEEK et EnKF : analyse locale et évolution dynamique de l'erreur. Rapport technique, MEOM-LEGI, Grenoble. 7 pp.
12. Brankart, J.-M., Testut, C.-E., and Brasseur, P. (2001). Analysis of non-synoptic observations : Seek filter and seek smoother. Technical report, MEOM-LEGI, Grenoble. 7 pp.
13. Brankart, J.-M., Testut, C.-E., and Parent, L. (2003). An integrated system of sequential assimilation modules. SESAM3.2 reference manual. Technical report, MEOM-LEGI, Grenoble. 85 pp.
14. Brankart, J.-M (2004, 2008). A set of operators for ocean NetCDF files. Technical report, MEOM-LEGI, Grenoble. 11 pp.
15. Brankart, J.-M (2004). Développement du système d'assimilation dans OPA pour le projet MERSEA : plan d'implémentation. Technical report, MEOM-LEGI, Grenoble. 16 pp.
16. Brankart, J.-M (2004). Parallélisation du code SESAM. Technical report, MEOM-LEGI, Grenoble. 4pp.
17. Ourmières Y., Brankart J.M., Berline L. and Brasseur P. (2005). Incremental Analysis Update implementation into the intermittent data assimilation system using the SEEK filter for the ocean general circulation model OPA8.1 in the North Atlantic 1/3° configuration NATL3. Rapport du projet Réseau Bleu (contrat GIP MERCATOR-CNRS n° 2004/10005). 38 pp.
18. Ourmières Y., Brankart J.M., Berline L. and Brasseur P. (2005). A data assimilative platform for coupled physical-biogeochemical experiments. Improving the physics of a coupled physical biogeochemical model of the North Atlantic through data assimilation : impact on the ecosystem. MERSEA project, deliverable report 7.2.2. 34 pp.
19. Skachko S.. Brankart J.M., Brasseur P. an Verron J. (2005). Assimilation scheme to control the model error due to air-sea fluxes. MERSEA project, deliverable report 7.3.7. 9 pp.

20. Ourmières Y., Brasseur P., Brankart J.M. and Verron J. (2006). First upgrade of the coupled assimilative platform, including free-surface model and assimilative scheme (NATL4/SEEK) and N2P1-type biogeochemical component (LOBSTER). MERSEA project, deliverable report 7.2.7. 14 pp.
21. Brankart, J.-M (2006). Optimal nonlinear filtering and smoothing assuming truncated Gaussian probability distributions in a reduced dimension space of adaptive size. Technical report, MEOM-LEGI, Grenoble. 17 pp.
22. Skandrani C. and Brankart J.M. (2008). Assimilation scheme to control the ocean mixed layer. MERSEA project, deliverable report 7.3.11. 11 pp.
23. Brankart, J.-M (2008). Prior probability distributions for biogeochemical rate parameters. Technical report, MEOM-LEGI, Grenoble. 4 pp.
24. Brankart, J.-M (2008). Sensitivity of finite-size Lyapunov exponents to mesoscale velocity errors. Technical report, MEOM-LEGI, Grenoble. 16 pp.
25. Brankart, J.-M (2009). Square root or ensemble observational update with SESAM. Technical report, MEOM-LEGI, Grenoble. 18 pp.
26. Brankart, J.-M and Melet A. (2010). Multigrid ocean data assimilation : SESAM with AGRIF. Technical report, MEOM-LEGI, Grenoble. 12 pp.

G – Vulgarisation scientifique

1. Brankart, J.-M, Brasseur, P., et Verron, J. (2001). Quel sera le courant jeudi prochain dans l'Atlantique Nord ? A l'aube de l'océanographie opérationnelle. CNRS Info, Lettre d'information destinée aux médias, numéro 389.

H – Documents multimédias

Webmaster des sites suivants :

1. The Mediterranean Oceanic Data Base (sur la période 1994-1996) : http ://modb.oce.ulg.ac.be
2. The DIADEM project. Contribution of the MEOM group (sur la période 2000-2002) : http ://www-meom.hmg.inpg.fr/Web/Projets/DIADEM
3. Site web de l'équipe MEOM (depuis 2007) : http ://www-meom.hmg.inpg.fr/
4. Site web du logiciel SESAM (depuis 2009) : http ://www-meom.hmg.inpg.fr/SESAM