Assimilation de données pour les problèmes non-Gaussiens : méthodologie et applications à la biogéochimie marine

Sammy Metref

UNIVERSITÉ GRENOBLE ALPES

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

Spécialité : **Océan, Atmosphère et Hydrologie**

Arrêté ministériel : 7 août 2006

Présentée par

Sammy Metref

Thèse dirigée par **Emmanuel Cosme**
et codirigée par **Pierre Brasseur**

préparée au sein **du Laboratoire de Glaciologie et de Géophysique de l'Environnement (LGGE, équipe MEOM, CNRS)**
et de **l'école doctorale Terre, Univers, Environnement**

Assimilation de données pour les problèmes non-Gaussiens

Méthodologie et applications à la biogéochimie marine

Thèse soutenue publiquement le **27 novembre 2015**,
devant le jury composé de :

Mme Clémentine Prieur
Professeur UJF (LJK, France), Présidente
M. Mark Asch
Professeur UPJV (LAMFA, France), Rapporteur
M. Laurent Bertino
Research Director (NERSC, Norvège), Rapporteur
M. Marc Bocquet
Professeur Ponts ParisTech (CEREA, France)), Examinateur
M. Ehouarn Simon
Maître de Conférence INP-Toulouse (IRIT, France), Examinateur
M. Emmanuel Cosme
Maître de Conférence UJF (LGGE, France), Directeur de thèse
M. Pierre Brasseur
Directeur de Recherche CNRS (LGGE, France), Co-Directeur de thèse

Résumé

L'assimilation de données pour les géosciences est une discipline cherchant à améliorer notre connaissance d'un système physique en se basant sur l'information issue de modèles numériques simulant ce système et sur l'information issue des mesures observant ce système. Les méthodes d'assimilation de données traditionnellement utilisées (e.g. le 4DVar ou les filtres de Kalman d'ensemble) reposent sur des hypothèses de Gaussianité des probabilités en jeu et de linéarité des modèles. Avec la complexification des modèles et des réseaux d'observations, ces hypothèses sont de plus en plus injustifiées et donc pénalisantes. Cette complexification est particulièrement forte en océanographie couplée à la biogéochimie marine.

Les objectifs de cette thèse sont de mieux comprendre l'apparition des non-Gaussianités dans un problème d'estimation, d'envisager une méthode d'assimilation de données adaptée aux problèmes fortement non-Gaussiens et, dans le cadre du couplage de la dynamique océanique et de la biogéochimie marine, d'explorer la pertinence de l'utilisation de méthodes non-Gaussiennes.

Dans un premier temps, une étude méthodologique est conduite. Cette étude, appuyé par des illustrations avec le modèle de Lorenz à trois variables, permet de mettre en évidence les limitations des méthodes traditionnellement utilisées, face à des problèmes non-Gaussiens. Cette étude aboutit sur le développement d'un filtre d'assimilation de données d'ensemble entièrement non-Gaussien : le Multivariate Rank Histogram Filter (MRHF).

Il est montré que le MRHF est performant dans des régimes fortement non-Gaussiens (notamment dans un régime bimodal) pour un nombre de membres relativement faible.

Dans un second temps, une étude numérique est conduite. Cette étude est réalisée aux travers d'expériences jumelles basées sur un modèle vertical 1D, ModECOGeL, couplant la dynamique et la biogéochimie en mer Ligure. Nous simulons différents réseaux d'observations combinant des profils *in situ* et des données satellites. Plusieurs méthodes d'assimilation sont alors comparées à l'aide de diagnostics d'évaluation d'ensemble avancés.

Nos expériences montrent l'impact du réseau d'observations et des variables de contrôle, sur le degré de non-Gaussianité d'un problème d'estimation. Le contrôle de la partie dynamique du modèle par des observations de la dynamique à différentes fréquences est un problème quasi-Gaussien, qu'un filtre aux moindres carrés, tel l'Ensemble Transform Kalman Filter, résout bien. En revanche pour ces mêmes observations, le contrôle de la biogéochimie s'avère être un problème non-Gaussien et nécessite l'utilisation d'un filtre non-Gaussien.

Enfin, il est montré que l'assimilation de la couleur de l'eau, pour le contrôle mixte de la dynamique et de la biogéochimie, est améliorée par des méthodes adaptées aux non-Gaussianités, tel l'Ensemble Kalman Filter anamorphosé. De plus, l'augmentation de la fréquence d'observation de la couleur de l'eau rend incontournable l'utilisation de filtres fondamentalement non-Gaussiens comme le MRHF.

Abstract

Data assimilation for Geosciences is a discipline seeking to improve our knowledge of a physical system based on the information from numerical models simulating this system and the information from the measures observing this system. The data assimilation methods traditionally used (eg the 4DVAR or the ensemble Kalman filters) are based on assumptions of Gaussianity of the probabilities involved and linearity of the models. With the increasing complexity of models and observation networks, these assumptions are increasingly unjustified and therefore penalizing. This complexity is particularly strong in oceanography coupled with marine biogeochemistry.

The objectives of this thesis are to understand the appearance of non Gaussianity in an estimation problem, to think out a data assimilation method adapted to highly non Gaussian problems and, in the coupling of ocean dynamics and marine biogeochemistry, to explore the relevance of the use of non Gaussian methods.

At first, a methodological study is conducted. This study, supported by illustrations with the three variable Lorenz model, allows to highlight the limitations of traditional methods when facing non Gaussian problems. This study led to the development of a fully non Gaussian data assimilation filter : the Multivariate Rank Histogram Filter (MRHF).

It is shown that the MRHF is efficient in highly non Gaussian regimes (including in a bimodal regime) for a relatively small number of members.

Secondly, a numerical study is conducted. This study is conducted with twin experiments based on a 1D vertical model, ModECOGeL, coupling dynamics and biogeochemistry in the Ligurian Sea. We simulate different observation networks combining *in situ* profiles and satellite data. Several data assimilation methods are then compared using advanced ensemble evaluation diagnoses.

Our experiments show the impact of observation networks and controled variables on the degree of non Gaussianity in an estimation problem. The control of the dynamic part of the model by observations of the dynamics at different frequencies is a quasi Gaussian problem, which a least squared filter such as the Ensemble Transform Kalman Filter solves well. In contrast, for the same observations, the control of biogeochemistry proves to be a non Gaussian problem and requires the use of a non Gaussian filter.

Finally, it is shown that assimilation of ocean color data, for the joint control of the dynamic and the biogeochemistry, is improved by methods adapted for non Gaussianities such as the Anamorphosed Ensemble Kalman Filter. In addition, increasing the ocean color observation frequency makes unavoidable the use of fundamentally non Gaussian filters such as the MRHF.

Table des matières

Préambule

Depuis de nombreuses années, la communauté scientifique cherche à mieux comprendre et à prévoir l'évolution du système Terre. À cet égard, le climat, la météorologie et plus récemment l'océanographie sont des sources de nombreuses recherches. L'intérêt pour ces disciplines a des origines aussi bien pratiques (avec l'opérationnel, prévision du temps, du climat ou des courants marins) que scientifiques (comprendre les moteurs et les enjeux des phénomènes physiques impliqués). Climat, météorologie, océanographie sont des disciplines étroitement liées. Les systèmes qu'elles étudient entretiennent de forts échanges.

En particulier, l'océan joue un rôle clef dans l'évolution du système Terre. Ainsi, l'équilibre thermique et l'évolution du système climatique sont fortement régulés par les intéractions océan-atmosphère. L'océan, avec sa forte capacité calorifique, est souvent considéré comme le "thermostat" ou "régulateur" du système climatique. En océanographie, qui est l'étude des océans et des mers de la planète Terre, la dynamique des courants et les échanges thermiques sont décrits. Le domaine d'étude de l'océanographie ne s'arrête pourtant pas là. D'autres éléments entrent également dans le fonctionnement de ce système, comme par exemple la dynamique des glaces (de l'Arctique et de l'Antarctique) ou encore la biogéochimie marine. L'importance de la biogéochimie marine vient notamment du fait que l'océan possède d'immenses capacités d'absorption de composants chimiques, comme le dioxyde de carbone à l'origine du fameux "effet de serre". Il apparaît donc crucial de mieux comprendre les océans et leurs qualités dynamiques, thermiques, glaciaires et biogéochimiques, afin de mieux appréhender le système Terre et son évolution.

L'océanographie, tout comme les géosciences en général, se base principalement sur deux grandes approches pour progresser : *l'observation* des phénomènes en jeu pour les décrire et les comprendre ; *la modélisation* de ces phénomènes pour les expliquer et les prévoir. Il va de soi que ces deux approches dépendent l'une de l'autre.

À la frontière entre ces deux approches se situe ce que l'on appelle l'assimilation de données (AD), qui va être au centre des questionnements abordés dans cette thèse. L'assimilation de données a pour but de combiner de manière optimale (ou au moins heuristique) les observations d'un système et le modèle simulant ce système, et ce, afin d'obtenir une meilleure estimation d'une partie ou de la totalité du système. L'assimilation de données se base sur des outils mathématiques, plus ou moins avancés, en se heurtant surtout aux problèmes mathématiques (contrôlabilité, malédiction de la dimensionalité ...) et numériques (faible échantillonage statistique, temps calculs ...) engendrés par les grandes dimensions des systèmes considérés.

L'équipe *Modélisation Ensembliste de l'Océan Multi-échelle* (MEOM) du *Laboratoire*

de Glaciologie et de Géophysique de l'Environnement (LGGE), au sein de laquelle s'est déroulée cette thèse, est un acteur important de la recherche en assimilation de données. Dans son histoire, l'équipe MEOM a participé à de nombreux travaux méthodologiques et algorithmiques en assimilation de données océanographique. Notamment, l'équipe a travaillé, dans les années 2000, au développement et à la mise en place algorithmique du filtre de Kalman de rang réduit, le SEEK (Pham et al., 1998), et de ses applications (Brankart, 2009; Brasseur and Verron, 2006; Cosme et al., 2010). De plus, au travers d'applications en biogéochimie marine, l'équipe MEOM a été confrontée à des problématiques d'assimilation de données non-Gaussienne. Ainsi, les travaux de Claire Lauvernet ont abouti au développement d'un filtre Gaussien tronqué (Lauvernet et al., 2009). Est également à noter le travail de l'équipe, via les projets *Mercator Vert*, *MERSEA* et *MyOcean*, à la mise en place et à l'évaluation d'un filtre de Kalman d'ensemble avec anamorphose sur un modèle couplé physique / biogéochime en Atlantique Nord (Béal et al., 2009, 2010; Brankart et al., 2012).

Dans cette lignée thématique, les problématiques d'assimilation de données non-Gaussienne pour des applications à la biogéochimie marine sont au centre des travaux de thèse présentés ici.

Pour travailler sur des problématiques d'assimilation de données océanographique, un projet Européen a vu le jour en 2012, le projet *SANGOMA* : Stochastic Assimilation for the Next Generation Ocean Model Applications. L'objectif de ce projet est de fournir de nouveaux développements en assimilation de données pour assurer les systèmes océaniques opérationnels futurs de l'utilisation de méthodes et autres outils d'analyse à la pointe de l'état de l'art en assimilation.

Cette thèse, en grande partie financée par le projet *SANGOMA*, se déroule dans cette optique, avec des considérations et des questionnements toutefois très en amont de l'opérationnel mais avec une vraie volonté de développement et de compréhension méthodologique de l'assimilation de données. Par ailleurs, la particularité de cette thèse est d'avoir pris le parti de concentrer ses travaux sur le domaine d'application précis qu'est le couplage océanographie physique et biogéochimie marine. Cette application fournit plusieurs cas problématiques de l'assimilation d'aujourd'hui et donc les challenges de l'assimilation de demain. En effet, à travers les non-linéarités et les non-Gaussianités régissant les systèmes considérés ou encore les difficultés de l'assimilation face au couplage, la biogéochimie marine met à disposition un bon nombre de défis. Cette thèse se veut d'apporter quelques éléments de réponse à ces défis.

Le premier chapitre présente le contexte entourant cette thèse, en développant brièvement les grands principes des domaines de la modélisation océanique, des observations de l'océan et de l'assimilation de données en océanographie. Le premier chapitre se termine en exposant clairement les problématiques que nous nous posons ici et auxquelles cette thèse essaye de répondre. Le deuxième chapitre introduit de manière formelle le problème de l'assimilation de données en contexte non-Gaussien. Après être revenu sur la notion de

non-Gaussianités et leurs origines, le chapitre 2 fournit un cadre mathématique pour le problème d'estimation afin d'étudier la méthodologie de l'assimilation de données. Dans le troisième chapitre, les méthodes d'assimilation nous servant tout au long de la thèse, sont décrites et illustrées à l'aide d'un modèle simple dit modèle jouet. Le quatrième chapitre (constitué d'un article en anglais paru dans *Nonlinear Processes in Geophysics*) présente le développement et l'évaluation d'une nouvelle méthode d'assimilation de données : le Multivariate Rank Histogram Filter (MRHF). Une application au couplage dynamique/biogéochimie marine est utilisée par la suite. Ainsi, le chapitre 5 décrit et valide le modèle ModECOGel (Lacroix, 1998) qui sera à la base des travaux des chapitres suivants. À travers le chapitre 6, on propose une description des relations dynamico-biogéochimiques et on en évalue l'impact sur l'utilisation d'assimilation de données pour le contrôle de la biogéochimie par observation de la dynamique. Le chapitre 7 présente les résultats de travaux dans ce même contexte pour un type d'observations particulier : la couleur de l'eau. Dans une dernière partie, sont présentées les conclusions fournies par notre étude et les perspectives qu'il sera possible de lui donner. Ce manuscrit se termine par une description personnelle des perspectives générales de l'assimilation de données future.

Chapitre 1

Du contexte à la problématique

Sommaire

L'océanographie a pour objectif la compréhension des phénomènes physiques intervenant dans les mers et les océans. Pour faciliter cette compréhension deux approches sont utilisées : la modélisation, qui met en place des modèles numériques simulant la physique océanique, et l'observation, qui utilise différents types de mesures évaluant les variables entrant en jeu dans la physique océanique.

Chacune de ces approches présente des avantages et des inconvénients que ce soit dans sa mise en place ou dans les résultats qu'elle produit.

La modélisation apporte une réponse homogène (en espace et en temps) et consistante avec les équations dont elle est issue. Malheureusement la mise en place et la résolution des modèles océaniques nécessitent bon nombre d'approximations. Ces approximations conduisent fatalement à des erreurs dont l'ampleur est en soi difficile à évaluer.

L'observation apporte également une description (plus ou moins) précise sur les phénomènes océaniques. Les observations disponibles sont par contre éparses et ne permettent pas à elles toutes seules de reconstruire la globalité de la circulation océanique.

À la croisée de ces deux approches, est née l'assimilation de données. Son objectif est de combiner les avantages de la modélisation et de l'observation afin d'obtenir une meilleure estimation de l'océan. La nature complexe de ces deux sources d'information, rend la tâche de l'assimilation difficile.

Pour mieux aborder ces travaux de thèse, nous revenons donc d'abord, en Section 1.1, sur les principes des modèles océaniques en s'attardant sur la notion de couplage océan dynamique et biogéochimie marine. Nous décrivons ensuite de manière non exhaustive les observations disponibles en océanographie.

Nous introduisons, en Section 1.2, le rôle de l'assimilation de données en océanographie ainsi que son évolution au cours du temps. Nous évoquons ensuite certains nouveaux défis qui se présentent à l'assimilation de données actuelle.

La Section 1.3 fait ressortir les points importants des ces deux premières sections. Cette synthèse permet de formuler clairement les problématiques qui dirigent nos réflexions et nos efforts dans la suite de cette thèse.

1.1 L'océanographie, entre modèles et observations

1.1.1 L'océan modélisé

En océanographie actuelle, la modélisation se décline en plusieurs facettes visant à représenter les principales composantes de la physique océanique.

On appelle communément **océan bleu**, l'aspect dynamique de l'océan. L'océan bleu regroupe les phénomènes de mouvements des masses d'eau et d'échanges thermo-halins (échanges de température et de salinité). Ces phénomènes sont décrits par les **équations primitives** que l'on présente plus en détail par la suite.

Plus récemment, de nombreux efforts sont faits pour modéliser d'autres phénomènes

océaniques. L'**océan vert**, par exemple, regroupe l'activité biogéochimique des océans. L'océan vert est l'ensemble des chaînes alimentaires au sein de l'écosystème océanique autrement appelé **réseaux trophiques**.

Les intéractions océan-glace (ou océan-cryosphère) peuvent également être modélisées. On nomme ces intéractions l'**océan blanc**.

Dans les chapitre 5 à 7, nous utilisons un modèle couplé de dynamique et de biogéochimie i.e. modélisant l'océan bleu-vert. Dans les deux sous parties suivantes nous évoquons les principes de la modélisation de ces deux "océans" et leurs éventuelles lacunes.

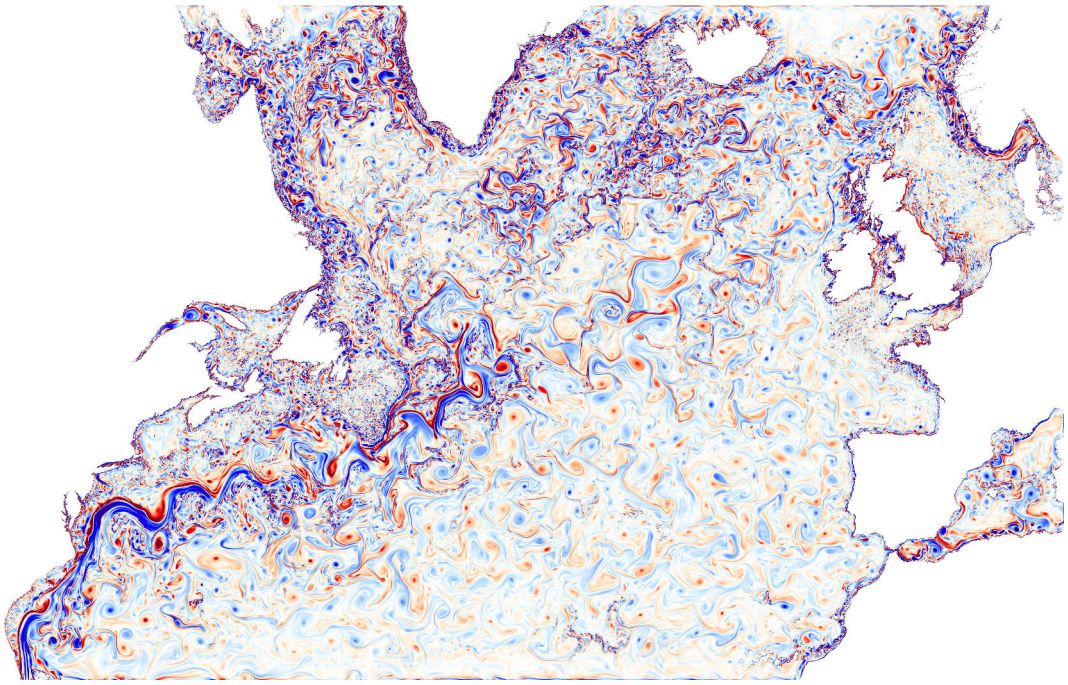

Figure *1.1* – *Champs de vorticité issu de la simulation NATL60 du modèle NEMO en configuration Atlantique Nord au $60^{ème}$ de degré.*

Les modèles de circulation (dynamique) générale

Les modèles de circulation générale sont une discrétisation sur des grilles relativement complexes des équations primitives océaniques. En océanographie, ce que l'on appelle équations primitives est la combinaison de :

– la **conservation de la masse** (l'approximation de Boussinesq permet l'équivalence à la conservation du volume),

– les **équations de mouvements horizontaux**,

– l'**équilibre hydrostatique** sur la verticale,

– les **équations de diffusion/advection de la température et de la salinité**.

L'augmentation de la résolution des modèles fournit une description détaillée en espace et en temps. Cette description de l'océan laisse apparaître une répartition hétérogène de phénomènes plus ou moins énergétiques à des échelles plus ou moins grandes.

On peut se rendre compte de la complexité et de l'inhomogéneité des phénomènes océaniques en regardant une simulation haute résolution. Un exemple de simulation haute résolution est la simulation NATL60, produite avec le modèle NEMO [1] en configuration Atlantique Nord avec une résolution horizontale au $60^{\text{ème}}$ de degré. Cette simulation est à la pointe de ce qui se fait en modélisation de l'océan à l'heure actuelle. La Figure 1.1 représente le champ de vorticité (champ de pseudo-vecteurs décrivant les mouvements de rotation locaux dans un fluide). La zone de forts tourbillons démarrant de la côte ouest étasunienne, se propageant le long de la côte vers le nord puis dans la partie nord du bassin, correspond au *Gulf Stream*. On constate la grande variété (spatiale et énergétique) des phénomènes physiques simulés.

La forte disparité des phénomènes aux différentes échelles dégrade fortement la prédictabilité de l'état du système.

Le couplage dynamique/biogéochimie marine

La biogéochimie marine décrit les relations alimentaires entre les organismes et les éléments chimiques présents dans un écosystème. Les échanges d'énergie et de biomasse au sein des réseaux trophiques s'écrivent à travers des flux de carbone (Figure 1.2) et d'azote parcourant les différents niveaux des réseaux.

Dans ce système, le phytoplancton occupe une place importante. Le phytoplancton est constitué de plusieurs espèces d'organismes végétaux. Ces organismes sont presque totalement passifs et sont donc principalement mus par la dynamique de l'océan. Dans les bonnes conditions de température et de luminosité, ils assimilent une grande variété de nutriments. En particulier, ils sont considérés comme une pompe à carbone, ce qui leur confère un rôle important dans l'évolution du climat. De plus, la compréhension des mécanismes biogéochimiques a de fortes implications environnementales et économiques (notamment pour la pêche) de part la place qu'occupe le phytoplancton à l'origine de la chaîne alimentaire océanique.

Ces organismes vivants sont plongés dans un milieu en mouvement. Ainsi, l'évolution de l'écosystème océanique dépend fortement de la dynamique océanique. C'est pourquoi,

1. Nucleus for European Modelling of the Ocean (NEMO) : *www.nemo-ocean.eu*

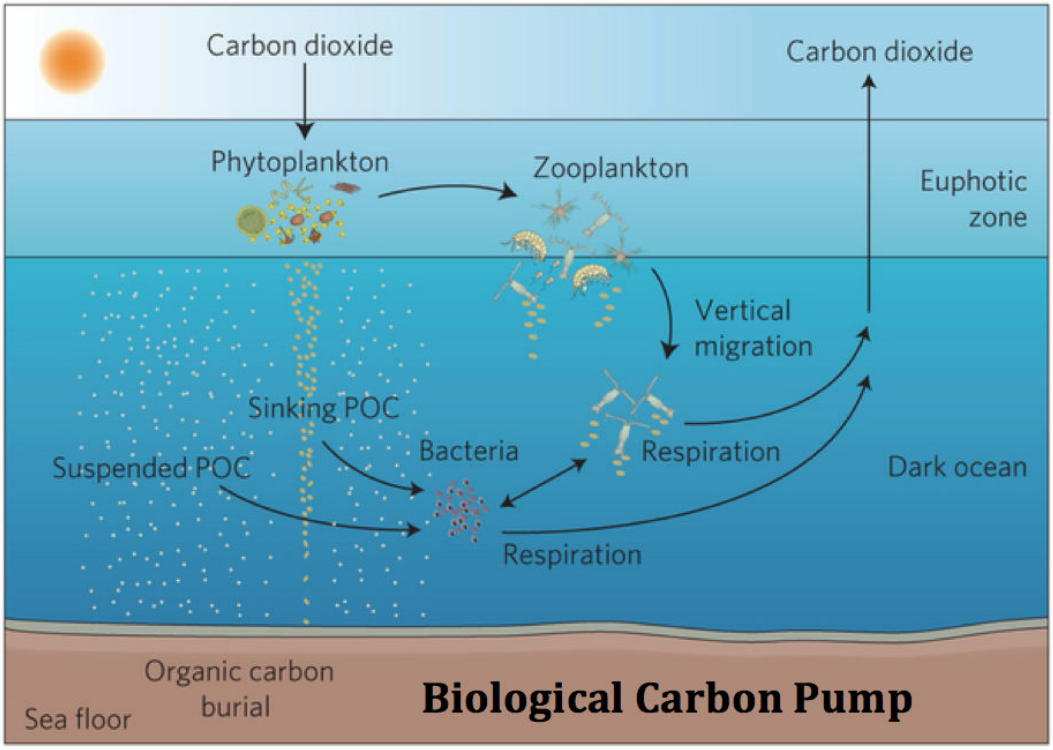

Figure *1.2* – *Représentation schématique du cycle du carbone dans l'océan.*
Schèma issu de Capone and Hutchins (2013) - La pompe biologique :
Dans la couche euphotique, le phytoplancton fixe le dioxyde de carbone en utilisant l'énergie solaire. Le carbone organique particulaire (POC) produit est "brouté" (grazing) par le zooplancton herbivore ou consommé, directement ou indirectement, par des microbes hétérotrophes se nourrissant des restes de phytoplancton dissout. Entre 1 et 40% de la production primaire est exportée hors de la couche euphotique et s'atténue de façon exponentielle vers la base de la couche mesopelagique à environ 1000m de profondeur. La reminéralisation de la matière organique dans la colonne d'eau océanique reconvertit le carbone organique en dioxyde de carbone. Seulement environ 1% de la production de surface atteint le plancher océanique.

l'océan vert et l'océan bleu sont souvent couplés dans les modèles. Ces modèles font cependant face à plusieurs difficultés.

Comme les travaux dans ce domaine sont récents, les modèles biogéochimiques présentent plusieurs lacunes. Le développement de ces modèles est bien moins avancé que les modèles de la dynamique océanique et repose sur des fondements théoriques moins solides. Les modèles mettent en jeu des paramétrisations biogéochimiques très incertaines ne prenant pas ou peu en compte la localisation géographique alors qu'elles le devraient. De plus, la biogéochimie est dirigée par des processus dynamiques de subméso-échelle ou d'échanges verticaux encore souvent mal représentés. Enfin, les processus biogéochimiques sont souvent

sporadiques (comme les *blooms* de phytoplancton) qui sont difficiles à reproduire par les modèles.

Les difficultés présentées par les modèles révélent la nécessité de faire appel à d'autres sources d'information. Dans la sous section suivante (Sec. 1.1.2), nous évoquons les moyens d'observation de l'océan (bleu-vert).

1.1.2 Observations de l'océan

Originellement, les géosciences – et l'océanographie en particulier – ont accumulé leurs savoirs en se basant sur l'observation des phénomènes physiques. Encore aujourd'hui les observations sont intensivement étudiées pour comprendre les mécanismes de l'océan et pour améliorer les modèles. La nature et la qualité des observations disponibles ont bien évolué au cours du temps. Sans être exhaustif, nous décrivons dans cette sous partie le principe et les caractéristiques des observations *in situ*, satellites et de la couleur de l'eau.

Les observations *in situ*

La locution latine *in situ* signifie sur place. Naturellement, les observations *in situ* sont donc composées des mesures effectuées sur place, ce qui leur permet de fournir une grande précision de mesure. L'utilisation systématique de données *in situ* date d'une soixantaine d'années. Leur nature est variée. Elles peuvent provenir :

- de **bateaux** (campagnes océanographiques et navires d'opportunités) évaluant diverses propriétés de l'océan à des stations de mesures ou le long des routes commerciales,

- de **plateformes *offshore*** ou de **mouillages fixes** fournissant des mesures à haute-fréquence temporelle,

- de **flotteurs lagrangiens** (flotteurs profileurs ou flotteurs de surface) mesurant la dynamique de l'océan (programme ARGO [2]) ou la biogéochimie marine (programme Bio-ARGO) en suivant les courants horizontaux de manière passive,

- ou de **gliders** et d'**AUVs** (Autonomous Underwater Vehicules) programmés pour parcourir l'océan en effectuant des profils de la dynamique, de biogéochimie et de propriétés optiques.

Il est à noter que les observations *in situ* sont, encore à l'heure actuelle, le seul moyen d'obtenir une information sur les couches profondes de l'océan.

2. http ://www.argo.net/

Nous simulons dans la suite des observations de flotteur-profileurs de type ARGO (Chapitre 6).

Le réseau d'observation ARGO, déployé au début des années 2000, est constitué de plus de 3000 flotteurs à travers les océans du globe. Il s'agit de la campagne de mesures *in situ* avec la plus importante couverture en temps et en espace.

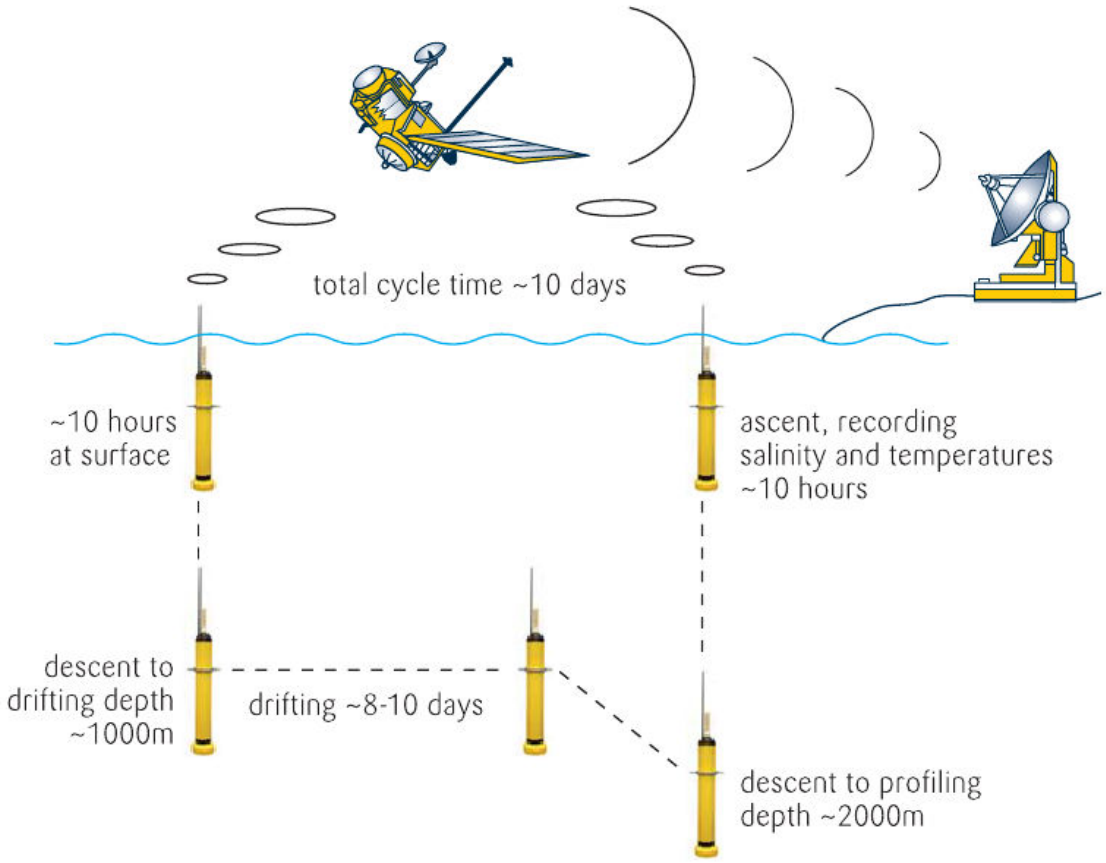

Figure *1.3* – *Cycle de fonctionnement d'un flotteur profileur ARGO. Credits : Argo Information Centre.*

Les flotteurs ARGO et Bio-ARGO prélèvent des mesures entre 2000m de profondeur et la surface effectuant la remontée en 10h environ (Figure 1.3). La fréquence de mesure était initialement entre 8 et 10 jours. Aujourd'hui, une technologie de communication (*IRIDIUM*) entre le flotteur et l'opérateur à terre permet d'accroître la fréquence de mesure à 2-3 jours lorsque l'on souhaite étudier des phénomènes à courtes échelles de temps.

Les flotteurs ARGO fournissent des observations de la température, de la salinité et des courants. Il est à noter que les flotteurs Bio-ARGO (pas étudiés ici) fournissent des observations de la biomasse ainsi que de la concentration de plusieurs éléments chimiques.

Le principal inconvénient des observations *in situ* est le caractère très local et l'irrégularité des informations qu'elles fournissent. En effet, les moyens de mesure déployés ne peuvent explorer qu'une infime partie des océans. Depuis une trentaine d'années, ces observations sont complétées par des observations plus globales : les observations satellites.

Les observations satellites traditionnelles

Avant l'avènement des satellites, l'océan était observé de manière très locale. Le lancement de plusieurs satellites scientifiques dans les années 1970-1980 a révolutionné notre compréhension des océans.

Les satellites permettent, entre autres, d'avoir une vision quasi-continue et quasi-globale des phénomènes océaniques (de surface). Les échelles spatiales couvertes par un satellite à la surface de l'océan varient de 10m à 10 000km. Les satellites peuvent observer des phénomènes allant de la journée à une dizaine d'années. Cette couverture spatio-temporelle permet l'étude des variations intra-saisonnières et saisonnières de phénomènes allant de la grande échelle à la (sous-) mésoéchelle.

Les satellites fournissent des observations sur plusieurs grandeurs océaniques.

L'altimétrie : Des radars altimétriques sont embarqués sur certains satellites (Skylab, GEOS-3, Seasat, Geosat, Topex/Poseidon, ERS-1, ERS-2, GFO, Jason-1, Envisat, Jason-2, Cryosat-2, Saral-Altika ...). Aujourd'hui, ces radars permettent de mesurer la hauteur de la surface océanique (SSH) avec une précision de l'ordre du centimètre. On peut déduire de ces mesures la hauteur de l'océan (la topographie dynamique), la hauteur des vagues et de la glace de mer. Ce type d'observations permet une caractérisation des phénomènes océaniques de surface de la moyenne échelle jusqu'à la circulation océanique globale.

la température de surface (SST) : La SST peut être mesurée par satellite suivant différents procédés de mesure : la radiomètrie spectrale (sur Aqua et Envisat), la radiomètrie infra-rouge (sur METOP, Envisat, Aqua, Terra, MeteoSat et DMSP) ou encore la radiomètrie micro-onde (sur DMSP, TRMM, Aqua, Envisat, Jason-1 et Jason-2). Les techniques (passives) de radiométrie micro-onde produisent la plus précise mesure de SST mais de moindre résolution spatiale. Les micro-ondes ne sont pas affectées par la nébulosité, ce qui permet d'avoir des observations temporellement homogènes et régulières.

la salinité de surface (SSS) : La salinité modifie les fréquences d'émissions thermiques micro-ondes de l'océan. Ces modifications sont quantifiables par radiométrie. Après inversion, une concentration de salinité est estimée avec une précision d'environ 0.2 PSS (*Prac-*

tical Salinity Scale). La salinité est une quantité moins souvent observée par satellite mais récemment le progamme SMOS (2009) et l'appareil de mesure AQUARIUS (2011) ont été lancés avec pour objectif commun de mesurer la salinité à la surface des océans.

Outre les variables océaniques observées, les satellites améliorent l'estimation des variables atmosphériques influant sur l'océan à travers les paramètres de forçages (e.g. les vents de surface, la température de l'air, les flux de chaleur radiatif ...).

Une autre observation satellitaire est la couleur de l'eau. L'utilisation de ce type d'observations est très récente et reste encore au cœur de bon nombre de travaux de recherche. Elle occupe également une place particulière dans notre étude. Nous la décrivons, donc, plus en détails dans la sous partie suivante.

La couleur de l'eau

L'observation de la couleur de l'océan a commencé par une mission expérimentale pionnière en 1978 : le *Coastal Zone Color Scanner* (CZCS). Cependant l'utilisation systématique de ce type d'observation (comme moyen d'étude de phénomènes et d'évaluation des modèles et non comme donnée assimilée) s'est répandu plus tardivement. Dans cet esprit, la mission américaine SeaWIFS [3] a été lancée en 1997. L'instrument de mesure américain MODIS [4] et européen MERIS [5] (Figure 1.4) ont suivi.

Le principe d'observation de la couleur de l'eau est simple. Il s'agit de mesurer la lumière solaire renvoyée par l'océan. La distribution spectrale (i.e. la couleur) mesurée apporte des informations sur la composition biogéochimique de l'océan. Si le principe est simple, la mise en pratique n'en reste pas moins complexe. En effet, la lumière solaire traverse différents milieux sur son trajet aller (soleil-océan) et sur son trajet retour (océan-satellite). Ces interférences perturbent la mesure et donc la caractérisation de la composition de l'océan.

La quantité biogéochimique mesurée par l'observation de la couleur de l'eau est la chlorophylle-a. La chlorophylle-a est le pigment principal du phytoplancton. En se désintéressant d'un problème (non-trivial) d'inversion, on peut considérer une observation satellite de couleur de l'eau comme une information sur la biomasse phytoplanctonique de surface. Les erreurs de mesure de chlorophylle par les satellites actuels sont de l'ordre de 20 à 100% en fonction des régions. De plus, le passage de nuages sur les images obtenues par les satellites réduit fortement le nombre d'images utilisables. Par exemple, "l'instrument couleur de l'eau Meris sur Envisat n'a que 17% de chances de bénéficier d'un ciel dégagé lorsqu'il survole

3. *Sea-viewing Wide Field-of-view sensor*
4. *Moderate-resolution Imaging Spectroradiometer*
5. *MEdium Resolution Imaging Spectrometer*

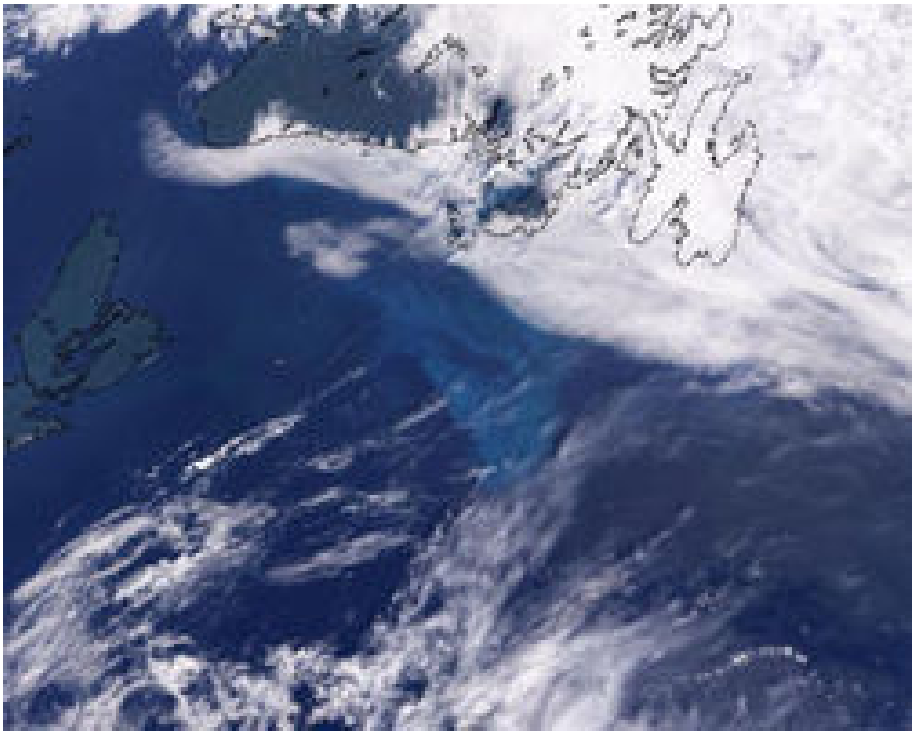

Figure *1.4* – *Phytoplancton (en turquoise) observé par l'instrument Meris à bord du satellite européen Envisat. Crédits : ESA 2002.*

la Bretagne" (E. Thouvenot, CNES, Conversations spatiales - Juin 2009). La fréquence d'observation est donc souvent bien plus faible qu'annoncée.

Il est à noter que les missions "couleur de l'eau" sus-mentionnées utilisent des satellites défilants (à orbites basses). Plus récemment, la mission coréenne GOCI[6] a été la première mission "couleur de l'eau" lancée sur un satellite géostationnaire. Les satellites géostationnaires circulent sur des orbites hautes (36 000km) à une vitesse égale à la rotation terrestre, ce qui leur permet d'observer de manière quasi-constante une région fixe à la surface du globe. Ces satellites autorisent une plus grande fréquence d'observations et peuvent donc palier au faible nombre d'images utilisables. Dans cette optique, le projet français Geo-OCAPI[7] étudie la faisabilité et l'apport scientifique du lancement d'un satellite géostationnaire pour les observations de la couleur de l'eau.

6. *Geostationary Ocean Color Imager*
7. *Geostationary Ocean Colour Advanced Permanent Imager*

Dans les expériences idéalisées d'assimilation de données que nous menons dans le chapitre 7, nous simulons des observations de la couleur de l'eau par des observations du phytoplancton total de surface en évaluant notamment l'intérêt des satellites géostationnaires.

1.2 L'assimilation de données en océanographie

Les définitions de l'assimilation de données varient d'un auteur à l'autre. La définition la plus complète que l'on a trouvé est peut-être celle proposée par Wikle and Berliner (2007) : "L'assimilation de données est une approche fusionnant les données (les observations) avec une connaissance *a priori* (e.g. des représentations mathématiques des lois physiques ; les sorties modèles) afin d'obtenir une estimation de la distribution de l'état vrai d'un processus."

En pratique, l'assimilation de données est un domaine pragmatique issu des lacunes des modèles et des observations évoquées dans les deux premières parties de cette section. Originellement, l'objectif de l'assimilation de données était d'améliorer l'estimation de l'état d'un modèle à des fins prévisionnelles. Depuis lors les objectifs et les applications de l'assimilation de données se sont multipliés.

L'idée générale est d'utiliser les différentes sources d'information disponibles, afin de produire de meilleurs estimés. Pour ce faire, des méthodes empiriques (e.g. les méthodes de rappel) ou mathématiques (issues de la théorie statistique ou de la théorie du contrôle optimal) ont été développées.

Dans un premier temps (Sec. 1.2.1), nous décrivons les motivations scientifiques à utiliser l'assimilation de données. La présentation de ces objectifs nous permet dans un deuxième temps (Sec. 1.2.2) de dérouler l'histoire et l'évolution de l'assimilation de données en océanographie. Au bout de cette évolution, nous faisons face aux nouveaux défis que l'assimilation rencontre (Sec. 1.2.3). Parmi ces défis, un retient notre attention et sera le fil conducteur de ces travaux de thèse : gérer la non-Gaussianité dans les problèmes d'assimilation.

1.2.1 Les objectifs de l'assimilation de données

Historiquement, l'assimilation de données a été développée pour améliorer les prévisions en météorologie puis un peu plus tard en océanographie. L'objectif était alors d'obtenir un état initial plus proche de la réalité pour pouvoir relancer les modèles en évitant ainsi une trop grande divergence des simulations.

De nos jours, les domaines d'application de l'assimilation de données sont nombreux. La glaciologie, la sismologie ou encore l'étude de la fusion nucléaire font toutes appel à de l'assimilation. Mais chacun de ces domaines pose de nouveaux problèmes et étend les enjeux de l'assimilation de données. En effet, estimer une condition initiale plus réaliste n'est plus le seul objectif. L'assimilation peut être utilisée pour la calibration et la validation de modèles, la mise en place de systèmes d'observations plus performants, la création de

produits de réanalyse ou encore l'estimation de paramètres.

En océanographie, également, les enjeux et les objectifs de l'assimilation de données sont variés. Nous donnons quelques exemples.

L'estimation des conditions initiales : Comme en météorologie, le premier objectif historique de l'assimilation de données en océanographie était l'estimation des conditions initiales d'un système. Ce problème s'inscrit directement dans une démarche d'amélioration des prévisions. Les systèmes océanographiques sont des systèmes chaotiques (bien que les forçages ont tendance à contenir l'évolution des systèmes). Ce caractère chaotique rend l'estimation des conditions initiales aussi cruciale, pour de bonnes prévisions, que difficile.

L'estimation de paramètres : Dans un modèle, de nombreux paramètres entrent en jeu. Ces paramètres peuvent représenter les forçages, des caractéristiques physiques ou encore la dynamique sous-maille non décrite par les équations du modèle. L'incertitude sur les paramètres d'un modèle peut être grande et n'est souvent pas prise en compte. Il est cependant possible de chercher à estimer certains paramètres d'un modèle avec l'assimilation de données.

L'estimation des paramètres sous-mailles est un exemple illustrant bien l'importance de l'estimation de paramètres. En effet, la diversité et l'intéraction de phénomènes multi-échelles dans l'océan rendent la modélisation incertaine. Il est difficile de correctement représenter les phénomènes sous-mailles (i.e. plus petits que la résolution des modèles) par des paramétrisations. Alors que, d'un point de vue dynamique, ces phénomènes ont des effets importants sur la répartition de l'énergie d'un système (par cascade énergétique directe et cascade énergétique inverse). De plus, d'un point de vue du couplage à la biogéochimie, de nombreux éléments régissants les écosystèmes sont également des phénomènes dynamiques sous-mailles (et/ou très hautes fréquences e.g. le vent). L'information manquante pour pouvoir correctement estimer ces paramétrisations peut venir des observations à travers l'assimilation de données.

La création de produits de réanalyse : Une réanalyse est une trajectoire reconstruite sur une période de temps à partir d'un modèle et de toutes les observations sur cette période de temps. On différencie un problème de filtrage qui produit une analyse en utilisant les données passées et présentes ; et un problème de lissage qui produit une réanalyse en utilisant les données passées, présentes et futures. Dans la suite nous étudions les filtres, nous n'abordons donc pas le sujet des réanalyses et du principe de lissage. Il est toutefois à noter que la demande pour des réanalyses de qualité est grande. Que ce soit pour des évaluations de modèles océaniques ou pour forcer des modèles atmosphériques, la création de réanalyses précises est devenue indispensable en océanographie.

L'évaluation de systèmes d'observations : L'évaluation de systèmes d'observations a pour but de déterminer la qualité et l'impact de l'information fournie par un système (ou réseau) d'observations. Cette évaluation peut se faire en réalisant et en comparant des assimilations sur des expériences simplifiées avec des observations artificielles (expériences jumelles) simulant certains systèmes d'observations. Il est également possible de réaliser et de comparer des assimilations sur des expériences réelles (i.e. avec observations réelles) provenant de différents systèmes d'observations. Certaines techniques d'évaluation de systèmes d'observations existent sans utiliser d'assimilation. Ces techniques sont tout de même basées sur les théories fondatrices de l'assimilation.

La liste des objectifs et des applications de l'assimilation de données est longue et ne cesse de grandir. En parallèle de la diversification de ces objectifs, la méthodologie de l'assimilation de données a beaucoup évolué. Nous évoquons maintenant les grandes lignes de cette évolution.

1.2.2 L'évolution de l'assimilation de données en océanographie

Le problème d'assimilation de données en océanographie (comme en météorologie) se pose souvent en ces termes. En disposant :

- d'un vecteur de contrôle *a priori* (soit un ou plusieurs vecteurs d'état entiers ou partiels, soit des paramètres) et son évolution dans le temps (à l'aide d'un modèle) ;
- des observations (dispersées en temps et en espace) d'une certaine quantité reliée au vecteur de contrôle par un opérateur d'observation ;

l'objectif est d'améliorer l'estimation du vecteur de contrôle à l'aide des observations.

Pour résoudre ce problème, la plupart des méthodes d'assimilation s'apparente à la méthode des moindres carrés dans le sens où elles proposent une correction moyenne entre l'*a priori* et les observations pondérées par leurs erreurs respectives.

Certaines méthodes considérent les observations les unes après les autres dans le temps. Ces méthodes sont constituées d'une étape de prévision (entre deux pas de temps observés) et d'une étape d'analyse (correction). Ce sont les méthodes de filtrage séquentiel e.g. les méthodes de rappel, les méthodes d'interpolation optimale (OI) ou encore les méthodes de Kalman (le filtre de Kalman, le filtre de Kalman étendu, les filtres de Kalman de rang réduit, les filtres de Kalman d'ensemble ...).

D'autres méthodes considèrent un ensemble d'observations réparties sur une fenêtre temporelle. Ce sont les méthodes de lissage e.g. le 4D-Var.

Les premières difficultés de l'assimilation en océanographie

L'assimilation de données en océanographie a toujours eu du retard sur l'assimilation de données en météorologie (Ghil, 1989). Ce retard est toutefois en train de diminuer. Les premières tentatives d'assimilation de données en océanographie avaient également pour

objectif l'estimation des conditions initiales. Cependant, les objectifs de l'assimilation en océanographie, que l'on vient de voir, évoluent rapidement. Alors que la météorologie a toujours comme but premier l'amélioration des prévisions à court et moyen termes.

Outre ces différences d'objectifs, l'assimilation de données en océanographie présente ses propres challenges et difficultés. Les deux premiers grands problèmes qui se sont présentés à l'assimilation en océanographie dynamique sont : la question de la propagation (ou advection) de l'information (définie en météorologie par Thompson, 1961) et la question du *trade-off* (l'échange d'information) entre les différentes variables mesurées (définie en météorologie par Charney et al., 1969).

La question de la propagation verticale de l'information de surface a été adressée très tôt avec des expériences simplifiées (expériences jumelles) dans les travaux de Hurlburt (1986) et Thompson (1986). De même, les travaux pionniers de Holland (1989), réalisés en préparation de l'expérience altimétrique TOPEX, traitaient de la propagation verticale de l'information altimétrique. Par la suite de nombreuses expériences d'assimilation altimétrique ont été réalisées (Holland, 1989; Holland and Malanotte-Rizzoli, 1989; Verron and Holland, 1989; Berry and Marshall, 1989). Ces expériences mettaient principalement en place des méthodes de rappel et des méthodes d'insertion directe (des observations dans le modèle). Toujours en essayant de traiter le problème de propagation verticale de l'information, cette fois dans un cadre d'assimilation d'images infra-rouges de température de surface, de nombreuses études ont été menées par Robinson et ses collaborateurs (Robinson and Leslie, 1985; Robinson et al., 1986; Robinson and Walstad, 1987; Robinson et al., 1988, 1989; Robinson, 1987; Mooers et al., 1987; Robinson et al., 1987; DeMey and Robinson, 1987). À la suite de ces études, une méthode de type moindres carrés a été mise en place par Robinson et al. (1989). Il s'agit de la méthode d'interpolation optimale (OI) sur l'espace réduit du modèle (utilisant les composantes principales ou *Empirical Orthogonal Functions*, EOF). L'OI peut être vue comme un rappel aux observations en prenant en compte les erreurs. Le résultat produit par l'OI est le meilleur estimé linéaire et non-biaisé (BLUE).

La question de la propagation horizontale de l'information localisée dans l'espace a été abordée en océanographie tropicale par Miller (1989). Miller (1989) a mis en place le filtre de Kalman pour assimiler des données de marégraphe dans un modèle linéaire simple du Pacifique Tropical. Le filtre de Kalman est une interpolation séquentielle effectuant une mise à jour de l'état estimé et de la matrice de covariances d'erreurs sous l'hypothèse de linéarité du modèle (entre autres hypothèses). Traitant de la même question en océanographie de moyenne latitude, Malanotte-Rizzoli and Holland (1986, 1988) ont utilisé une méthode d'insertion directe pondérant les observations en fonction de leur distance.

La question du *trade-off* entre les différentes variables mesurées a été principalement traitée, dans les premières études, pour l'océan tropical. Sont à noter les travaux de Moore and Anderson (1989), qui utilisent une forme spéciale de l'OI, et de Miller (1990), qui utilise le filtre de Kalman. En moyenne latitude, cette question a été abordée par Malanotte-Rizzoli et al. (1989) dans un modèle aux équations primitives.

Plusieurs de ces difficultés sont encore d'actualité, par exemple : la propagation verticale de l'information, avec l'augmentation des données satellitaires, ou encore la propagation horizontale de l'information, avec la question de la localisation. Néanmoins, ces travaux précurseurs ont lancé la dynamique d'évolution des méthodes d'assimilation de données.

Par la suite, avec la complexification rapide des modèles océaniques, l'hypothèse de linéarité des modèles était de moins en moins raisonnable. De plus, le coût numérique pour traiter des problèmes d'aussi grande dimension devenait trop lourd. De ces difficultés théoriques et numériques, sont nées des méthodes non-linéaires c'est à dire contournant l'hypothèse de linéarité des modèles.

L'assimilation de données non-linéaire

La plupart de ces méthodes pionnières mettent à jour les covariances d'erreurs (quand elles le font) sous l'hypothèse de linéarité des opérateurs (opérateur d'observation et modèle). Or les modèles océaniques comportent de nombreuses non-linéarités. Le besoin d'adapter les méthodes d'assimilation linéaires existantes (e.g. l'interpolation optimale ou le filtre de Kalman) s'est donc vite imposé.

La première méthode proposée, le plus directe, est le filtre de Kalman étendu (EKF). Ce filtre réalise une analyse de Kalman après avoir linéarisé le modèle. Ce filtre étendu fonctionne bien dans des cas de faibles non-linéarités et lorsque la taille des matrices à inverser est raisonnable. Il n'est donc pas applicable en océanographie réaliste où la matrice d'un modèle linéarisé peut atteindre une taille de $10^6 \times 10^6$. Une solution possible pour palier à ce problème numérique est de réduire le rang des matrices à inverser. Ces méthodes sont appelées filtres réduits (ou de rang réduit). Dans cet esprit, le filtre Singular Evolutive Extended Kalman filter (SEEK) décompose la matrice du modèle linéarisé en composantes principales (ou EOF) pour réduire son rang (Pham, 2001). Le filtre SEEK a été extensivement développé au sein de l'équipe MEOM (Brankart, 2009; Brasseur and Verron, 2006; Cosme et al., 2010). D'autres filtres réduits sont à noter : le filtre RRSQRT (Verlaan and Heemink, 2000), le filtre SEIK (Pham, 1996; Pham et al., 1998) ou encore le filtre d'estimation statistique des sous-espaces d'erreurs (Lermusiaux and Robinson, 1999a,b).

Une autre approche permettant de propager et de mettre à jour les états et les covariances d'erreurs, a été développée à la fin des années 1990. Cette approche se base sur un échantillonnage de Monte-Carlo. Il s'agit non plus de considérer les matrices de covariances d'erreurs entières mais de les représenter par un ensemble d'états du modèle. Ainsi, l'information statistique contenue dans l'ensemble peut être propagée directement par le modèle (non-linéaire) pendant l'étape de prévision et mise à jour pendant l'étape d'analyse. Ces méthodes sont appelées méthodes d'ensemble. Le filtre de Kalman d'ensemble stochastique (EnKFs) a été la première méthode d'ensemble (Evensen, 1994; Burgers et al., 1998; Evensen, 2003). De nombreuses variantes de l'EnKFs existent.

Par ailleurs, l'approche variationnelle (dont on ne parlera que peu au cours de cette thèse) est issue du contrôle optimal. Elle effectue une analyse moindres carrés en minimi-

sant une fonctionnelle. Le 4D-var, qui est la version quadridimensionnelle de l'approche variationnelle, est une méthode faisant également l'hypothèse de linéarité du modèle. L'application du 4D-Var à des problèmes non-linéaires a été rendue possible par la méthode des équations adjointes (Le Dimet and Talagrand, 1986; Talagrand and Courtier, 1987; Courtier and Talagrand, 1987). Cette hypothèse peut être réduite par l'utilisation de la version incrémentale du 4D-Var.

L'évolution de l'assimilation de données en océanographie est alimentée par deux moteurs : i) les objectifs de l'assimilation et les nouvelles applications que peut offrir l'océanographie ; ii) les difficultés et les défis que l'assimilation rencontre tant sur le plan théorique que sur le plan numérique.

Dans la sous partie suivante nous décrivons certains nouveaux défis pouvant potentiellement mettre à mal l'assimilation actuelle. Notre intérêt s'arrête sur la difficulté de l'assimilation moindres carrés à gérer les probabilités non-Gaussiennes de plus en plus fréquentes en géosciences et en biogéochimie marine en particulier.

1.2.3 Les nouveaux défis

Dans diverses disciplines des géosciences, la complexité grandissante des systèmes apporte de nouveaux problèmes en assimilation de données. De plus, les nombreux couplages inter-géosciences accentuent fortement la difficulté du problème d'estimation. Au cours de cette étude, nous nous focalisons notamment sur le couplage dynamique océanique/biogéochimie marine.

La principale difficulté de l'assimilation de données est de correctement décrire les incertitudes (ou les erreurs) de l'*a priori* et des observations (et du modèle). Comme il est détaillé en Section 2.2.2 de ce chapitre, les méthodes traditionnellement employées pour résoudre les problèmes d'estimation en géosciences considérent ces incertitudes dans un cadre statistique Gaussien. Ce cadre présente deux justifications :

- Les densités Gaussiennes sont décrites seulement par leur moyenne et leur matrice de covariances d'erreur, ce qui rend les méthodes numériquement viables (même dans leurs formes multivariées).
- La présence de statistiques Gaussiennes dans les modèles physiques (surtout dans les problèmes de grande taille) est appuyée par le théorème central limite (TCL).

Ces deux justifications ont permis de faire avancer l'assimilation de données. L'hypothèse de Gaussianité n'en reste pas moins erronée.

Avec la grande complexité et les non-linéarités des modèles en géosciences, cette hypothèse est de plus en plus mise à mal. De plus, il est à noter que dans un cas linéaire, Gaussien et bien observé, l'augmentation du nombre d'observations ne peut qu'améliorer les performances de l'assimilation aux moindre carrés puisque l'erreur *a posteriori* (i.e. après assimilation) est, par construction, inférieure ou égale à l'erreur *a priori* (i.e. avant assimilation). Ceci n'est pas vrai dans des cas non-linéaires et non-Gaussiens. Ainsi, le nombre

grandissant de données de surface disponibles à l'assimilation peut également rendre le problème de contrôle difficile.

L'exemple du couplage dynamique et biogéochimie permet d'apprécier la présence de non-Gaussianités et les difficultés qu'elles engendrent pour l'assimilation. Nous avons évoqué en Section 1.1.1 la complexité des systèmes biogéochimiques avec notamment de nombreux phénomènes à seuil. Cette complexité peut rendre la distribution des variables biogéochimiques très non-Gaussienne. Pour illustrer ce propos, la Figure 1.5 présente le nuage de points (*scatterplot*) d'un ensemble de simulations et d'une simulation de référence issus de Béal et al. (2010) représentés sur le plan Détritus (DET, en $mmolNm^{-3}$) / Chlorophylle (CHL, en $mmolNm^{-3}$) à la station du *GulfStream* (47°W,40°N). La ligne bleu représente l'observation en chlorophylle au jour 1 à cette même station. La concentration de Détritus croît avec l'augmentation de Chlorophylle jusqu'à un certain seuil ($\simeq 0.8 mmolNm^{-3}$), puis décroît lorsque la Chlorophylle continue d'augmenter. Un simple examen visuel indique que les deux variables ne sont clairement pas conjointement Gaussiennes.

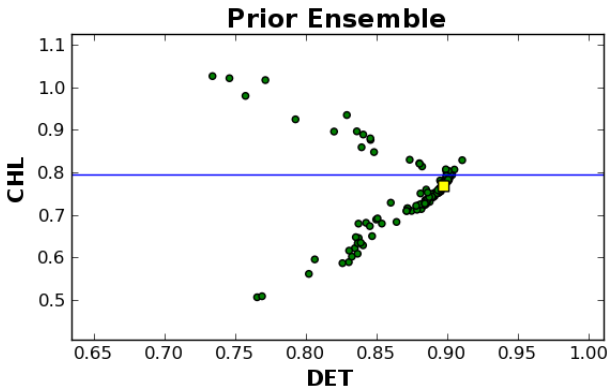

Figure *1.5* – *Scatterplot d'un ensemble de simulations (points verts) et d'une simulation de référence (carré jaune) issus de Béal et al. (2010) représentés sur le plan Détritus (DET, en $mmolNm^{-3}$) / Chlorophylle (CHL, en $mmolNm^{-3}$) à la station du GulfStream (47°W,40°N). La ligne bleu représente l'observation en chlorophylle au jour 1 à cette même station.*

L'assimilation de données moindres carrés telle qu'introduite à la Section 1.2.2 et décrite en détails à la Section 2.2.2 du chapitre suivant, gére très mal ce type de distribution. Le filtre de Kalman d'ensemble, par exemple, effectue une régression linéaire entre les variables pour propager les corrections (Sec. 2.2.2). Or, dans ce cas, la relation entre les détritus et la chlorophylle n'est clairement pas linéaire. Ainsi, les corrections que proposent le filtre de Kalman d'ensemble ne sont pas en accord avec la dynamique du système telle que décrite par l'ensemble *a priori*.

De plus, les réseaux d'observations (Sec. 1.1.2) sont de plus en plus complexes. Ces réseaux modifient les problèmes d'estimation à résoudre en augmentant le nombre et la

nature des contraintes à respecter. Ce qui rend la tâche de l'assimilation d'autant plus difficile.

Enfin, certaines variables sont intrinséquement non-Gaussiennes. Pour reprendre un exemple de biogéochimie marine, la concentration de chlorophylle est toujours positive ou nulle. Supposer une distribution Gaussienne de cette concentration n'a donc que peu de sens puisqu'une distribution Gaussienne a pour support tout $\mathbb{R}$.

L'étude et le développement de méthodes dépassant l'hypothèse de Gaussianité sont donc des sujets importants de la recherche en assimilation de données. C'est aussi ce qui constitue l'un des grands questionnements motivant les travaux présentés dans cette thèse.

1.3 Problématique de la thèse

Résumé du contexte

La section 1.1 instaure le cadre scientifique de ce travail, en évoquant les problématiques de la modélisation océanographique en général et océanographique couplée à la biogéochimie marine en particulier. Ces difficultés sont, entre autres, la complexité grandissante des systèmes et leurs formulations numériques générant de nombreuses erreurs. De plus, les systèmes de biogéochimie marine sont particulièrement non-linéaires, combinant des variables positives et des phénomènes de seuil, avec une part importante de paramétrisations générant de nombreuses incertitudes. Une autre partie du cadre scientifique évoquée dans cette section est l'observation de l'océan encore inhomogène aussi bien en temps qu'en espace. L'arrivée des observations satellites, apporte une quantité importante d'informations en surface de l'océan. Cette avancée ne vient pas sans son lot de nouveaux défis. En effet, l'assimilation de données de surface engendre des problèmes de propagation de l'information au travers de la colonne d'eau.

Entre modèles et observations, la deuxième section (Sec. 1.2) nous fait pénétrer dans le sujet qui nous concerne : l'assimilation de données en océanographie avec ses objectifs et ses difficultés. À travers les motivations de l'assimilation de données et à travers l'évolution des méthodes, cette section fait ressortir les efforts mis en place pour résoudre les nombreux problèmes d'estimation en océanographie. Parmi les nouveaux défis de l'assimilation de données actuelle en océanographie notre attention s'arrête sur les problèmes de nature non-Gaussienne.

En résumé, la compréhension des phénomènes en océanographie physique et couplée (notamment à la biogéochimie marine) est un enjeu scientifique majeur. Un outil indispensable à cette compréhension est l'assimilation de données. Cet outil est mis à mal par la complexification des systèmes introduisant des non-Gaussianités dans le cadre probabiliste

des problèmes d'estimation. L'anticipation et l'impact de ces non-Gaussianités sont souvent mal appréciés. De plus, les méthodes d'assimilation existantes sont soit inadaptées à ce type de problèmes ; soit uniquement adaptées à des cas très particuliers de non-Gaussianité ; soit encore trop coûteuses pour la grande dimension.

Problématique

De ces deux premières sections émerge une problématique déclinée en plusieurs questionnements qui alimentent les réflexions entourant cette thèse. Ces questionnements concernent la méthodologie de l'assimilation de données dans des contextes non-Gaussiens. Contextes que l'on rencontre de plus en plus fréquemment notamment dans des problèmes d'estimation en couplage de dynamique océanique et de biogéochimie marine. Ces questionnements sont :

- *Comment déceler les non-Gaussianités dans un système ?*
- *Comment envisager une assimilation de données adaptée à un contexte non-Gaussien ?*
- *Dans le cadre du couplage dynamique/ biogéochimie, que peut apporter l'assimilation de données non-Gaussiennes et comment ?*

Dans la suite de ce manuscrit, nous nous donnons comme objectif de répondre à ces questionnements. Nous essayons, aux travers d'études simples, de considérations théoriques et d'applications idéalisées ou réalistes, d'élaborer nos réflexions et d'aboutir à des conclusions pertinentes.

Plan de thèse

Le chapitre 2 fournit les concepts et les outils permettant d'aborder les questionnements méthodologiques de l'assimilation de données non-Gaussienne. Au vue du flou qui peut entourer la notion de non-Gaussianité, la Section 2.1 propose un retour bref sur la notion de non-Gaussianité. Il est par la suite fait mention des outils diagnostiques permettant de constater ou d'anticiper la présence de non-Gaussianités dans un système. Pour revenir au problème qui nous intéresse, nous discutons de l'apparition de ces non-Gaussianités dans un problème d'assimilation de données. Ces apparitions sont notamment appuyées par des exemples fréquemment rencontrés en géosciences. La section 2.2 instaure un cadre formel pour énoncer clairement un problème d'estimation. Ce cadre permet également de percevoir les limitations des méthodes de type moindres carrés. La section 2.3 aborde enfin le défi de l'assimilation de données non-Gaussienne. Allant de pair avec ce nouveau défi, sont discutées la question du résultat produit par une assimilation et la manière dont on peut l'évaluer. Puis en redéfinissant les objectifs d'une assimilation ensembliste, l'échec de l'assimilation moindres carrés face aux non-Gaussianités est présenté au travers d'une revue d'articles de la littérature. Enfin, les récents travaux portant un intérêt aux difficultés de l'assimilation en contexte non-Gaussien sont évoqués. Ces travaux présentent toutefois

plusieurs limitations.

Le chapitre 3 est un court chapitre présentant les méthodes d'assimilation de données. L'écriture et la mise en place de progammes numériques en langage *python*, qui seront utilisés tout au long de la thèse, sont décrites. De plus, une série d'illustrations sur des expériences jumelles d'assimilation de données utilisant un modèle de petite dimension (Lorenz 63 à trois variables), est présentée. Ces illustrations permettent de vérifier le bon fonctionnement des méthodes et d'apercevoir certaines de leurs difficultés notamment en contexte non-linéaire et non-Gaussien.

Le chapitre 4 reprend les difficultés des méthodes existantes à gérer certains problèmes d'estimation. Une nouvelle méthode, le *Multivariate Rank Histogram Filter* (MRHF), est développée en élargissant à toutes les variables le principe des histogrammes de rangs tels qu'utilisés par le *Rank Histogram Filter* (Anderson, 2010). Le MRHF est ensuite testé sur un modèle jouet (Lorenz 63) dans des configurations à la complexité variable.

Après avoir focalisé notre attention sur des systèmes de petites dimensions, nous nous acheminons vers des problèmes plus réalistes.

Le chapitre 5 décrit un modèle couplé de dynamique océanique et de biogéochimie marine sur une dimension verticale (ModECOGeL). Ce modèle a l'avantage de présenter un comportement non-linéaire et non-Gaussien propice à notre étude. De plus, l'aspect réaliste de ce modèle apporte de nouvelles difficultés. De ce modèle, sont conçus deux jeux d'expériences jumelles étudiés respectivement dans les deux chapitres suivants.

Le chapitre 6 présente des expériences d'assimilation de données de température et de salinité en profils verticaux et de température de surface. Le vecteur de contrôle est dans un premier temps composé seulement de la température et de la salinité puis dans un second temps est ajouté le phytoplancton. Ce premier jeu d'expériences jumelles permet de mettre en évidence le comportement d'une méthode d'assimilation moindres carrés et d'une méthode non-Gaussienne dans des contextes de non-Gaussianité différents. De plus, l'utilisation de plusieurs réseaux d'observations ajoute à l'étude un intérêt opérationnel.

Enfin le chapitre 7 se concentre sur le problème difficile de l'assimilation de la couleur de l'eau. Depuis peu, l'observation par satellites fournit une information sur la biomasse de surface (nous ne traitons pas le problème inverse qui permet de recouvrer cette information à partir d'une image). L'assimilation de ces données de biomasse en surface engendre un problème d'estimation complexe. Ce second jeu d'expériences jumelles nous offre une base simple mais réaliste pour comprendre et savoir aborder ce problème sur le plan méthodologique.

Chapitre 2

Méthodologie de l'assimilation de données

Sommaire

L'objectif de ce chapitre est de se munir des éléments théoriques et conceptuels nécessaires pour aborder d'un point de vue méthodologique le problème de l'assimilation de données non-Gaussienne.

Dans la Section 2.1, nous essayons de mieux comprendre la notion de non-Gaussianité en soi et dans le contexte de l'assimilation de données.

Dans la Section 2.2, nous abordons la méthodologie générale de l'assimilation de données en instaurant un cadre formel au problème d'estimation. Nous décrivons ensuite les approches moindres carrés de l'assimilation de données dans un formalisme probabiliste commun. Ceci nous permet d'anticiper leurs limites face à des problèmes d'estimation de plus en plus complexes.

Dans la Section 2.3, nous discutons des différents efforts qui ont été menés dans la littérature pour développer des techniques et des méthodes afin de gérer au mieux les problèmes de non-Gaussianité.

2.1 La non-Gaussianité et le problème d'estimation

Au cours des vingts dernières années, l'assimilation de données dite non-Gaussienne est devenue un sujet important au sein de la communauté de recherche en géosciences. La littérature compte déjà bon nombre de méthodes et autres techniques essayant de traiter efficacement les difficultés engendrées par la non-Gaussianité. Cependant, la notion même de non-Gaussianité est utilisée dans des applications si différentes qu'elle peut apparaître comme floue voir mal définie.

Il convient donc, avant de se plonger dans la méthodologie de l'assimilation non-Gaussienne, de préciser certains concepts. Notamment, nous essayons dans la première sous partie (Sec. 2.1.1) de répondre à des questions simples en apparence mais sous lesquelles se cachent de nombreuses subtilités : Qu'est ce qui caractérise et qu'est ce qui engendre un problème non-Gaussien ? Quels sont les liens entre non-linéarité, chaos et non-Gaussianité ? Ou encore, quels outils peuvent nous permettre d'évaluer le degré de non-Gaussianité d'un problème ? Dans la deuxième sous partie (Sec. 2.1.2), nous énonçons les difficultés qu'apportent les non-Gaussianités aux problèmes d'estimation, en particulier, en géosciences. Enfin, nous discutons de certaines tentatives présentes dans la littérature pour gérer la non-Gaussianité en assimilation de données (Sec. 2.3.2). Dans le reste de cette étude nous nous focalisons principalement sur l'assimilation séquentielle. Ce qui signifie que nous ne considérons aucune méthode variationnelle dans nos travaux bien que de nombreux efforts sont également faits dans ce cadre pour traiter la non-Gaussianité.

2.1.1 La non-Gaussianité

Nous allons chercher dans un premier temps à caractériser la non-Gaussianité. Nous commençons par faire des liens simples entre un système déterministe (non-linéaire et chaotique) et la vision probabiliste que l'on en a. Avec cette vision probabiliste du problème

d'estimation, on s'aperçoit que la complexité du système fait apparaitre des probabilités non-Gaussiennes. Nous présentons enfin, de manière plus pragmatique, certaines méthodes pour anticiper ou diagnostiquer la non-Gaussianité.

D'un système non-linéaire et chaotique à la non-Gaussianité

La présence de non-linéarités dans un système engendre sous certaines conditions un comportement très particulier : le chaos. Un système dynamique non-linéaire, s'écrivant sous la forme d'équations d'état et faisant intervenir des variables et des paramètres (de contrôle), peut évoluer selon plusieurs régimes. En faisant varier les paramètres de contrôle différents modes sont mis en jeu. Selon les modes en jeu, le système effectue des transitions d'un régime à un autre. Ces transitions s'appellent des bifurcations. Les régimes possibles peuvent être stationnaire, périodique, bipériodique ... Le nombre de modes augmentant, le comportement du système devient de plus en plus complexe jusqu'à atteindre un régime chaotique. Les régimes chaotiques sont associés à des *attracteurs étranges* (Ruelle and Takens, 1971). Les attracteurs étranges sont des domaines de l'espace dits robustes au sens où des perturbations (raisonnables) dans ce domaine ne font pas sortir la dynamique de son régime chaotique.

Les trajectoires au sein d'un attracteur étrange (une orbite) ont une forte sensibilité aux conditions initiales. C'est à dire qu'une petite erreur initiale va rapidement s'amplifier avec le temps. Une conséquence contraignante de ce phénomène est qu'il est difficile de prévoir un comportement chaotique. Par exemple, un pendule simple sans friction a un comportement périodique. Son évolution est facilement prévisible. Un double pendule est très sensible aux conditions initiales et présente un comportement chaotique. Prévoir l'évolution d'un double pendule est beaucoup plus difficile sans avoir la condition initiale et les équations d'évolution (déterministes) exactes. Par manque d'informations, le comportement du chaos "semble" aléatoire (ou stochastique).

Une manière d'essayer de prévoir l'évolution d'un système dynamique à tendance chaotique se base sur cette dernière remarque. En effet, nous nous trouvons rapidement dans une situation de manque d'informations. Il peut être donc intéressant de considérer le problème de prévision comme un problème aléatoire. Cela consiste à voir les erreurs commises dans notre estimation des conditions initiales, dans notre approximation des équations d'évolution et dans toutes autres informations disponibles comme des variables aléatoires.

En géosciences, cette démarche est déjà répandue en assimilation de données puisque le problème d'estimation prend en compte les erreurs (au sens probabiliste). Dans le problème direct de prévision, la démarche déterministe est encore employée. Cependant la prise en compte de processus stochastiques au sein des modèles est de plus en plus fréquente, notamment en océanographie (Brankart et al., 2010; Brankart, 2013).

L'assimilation de données prend depuis longtemps en compte le caractère stochastique des quantités considérées. Toutefois, ces quantités sont souvent supposées Gaussiennes. Comme on l'a vu (Sec. 1.2.3), plusieurs justifications de cette hypothèse existent. Seulement, aucune raison ne permet de savoir *a priori* la loi de probabilité d'un état d'un système dynamique non-linéaire. Au contraire, la loi d'un état dépend entre autre de la forme de l'attracteur du système qui a peu de chance d'être un domaine infini et donc propice à des lois Gaussiennes. De plus, une variable aléatoire Gaussienne propagée par un système non-linéaire ne reste pas Gaussienne (Bengtsson et al., 2003).

Ainsi, le manque d'information sur un système non-linéaire en régime chaotique, et donc difficilement prévisible, nous oblige à nous confronter à un problème d'estimation stochastique. La complexité de l'évolution du système rend l'hypothèse de Gaussianité peu appropriée pour décrire les variables aléatoires en jeu. Pour vérifier cette dernière assertion, il est bon de se munir d'outils pour diagnostiquer la présence de non-Gaussianités dans un problème d'estimation. C'est ce que nous faisons dans la sous partie suivante.

Comment diagnostiquer la non-Gaussianité ?

Evaluer la non-linéarité du modèle On vient de le voir, une distribution initialement Gaussienne est altérée lorsque propagée par un modèle non-linéaire. Ainsi, on peut se faire une première idée de la quantité des non-Gaussianités dans un système en évaluant la non-linéarité du modèle qu'il met en jeu. Nous n'effectuons pas de tels tests dans les travaux présentés ici puisque nous évaluons directement les non-Gaussianités engendrées.

Evaluer la non-linéarité inter-variables Il existe peu de diagnostics jugeant de la Gaussianité multivariée. Le peu existant est très difficile à mettre en place dans des problèmes à grande dimension comme on en rencontre souvent en océanographie. Cependant, une condition nécessaire à la Gaussianité multidimensionnelle est la linéarité entre les variables marginales. Cette notion revêt de plus une importance particulière puisque comme on l'a vu en présentant le formalisme d'Anderson (Sec. 2.2.2), les filtres de Kalman mettent en jeu une régression linéaire entre la variable observée et les variables non-observées. Il est donc important de voir si cette condition nécessaire est vérifiée.

Pour ce faire, la première quantité statistique que l'on peut regarder ici est la linéarité au sens de Pearson (Pearson, 1895) entre deux variables. Le coefficient de corrélation de Pearson, $r_{X,Y}$, entre deux variables (X, Y) est défini comme la covariance des deux variables, $cov(X, Y)$, divisée par le produit de leurs écarts-types respectifs, σ_X et σ_Y :

$$r_{X,Y} = \frac{cov(X, Y)}{\sigma_X \sigma_Y} \tag{2.1}$$

La corrélation $r_{X,Y}$ varie entre -1 et 1 avec une totale anti-corrélation quand $r_{X,Y} = -1$, une totale corrélation quand $r_{X,Y} = 1$ et aucune corrélation entre les variables quand

$r_{X,Y} = 0$. Puisque l'on souhaite savoir si une corrélation existe entre les variables évaluées nous pouvons prendre la valeur absolue du coefficient de corrélation de Pearson qui va donc de 0 à 1 pour respectivement aucune corrélation linéaire et une corrélation ou anti-corrélation linéaire totale.

Une deuxième quantité statistique à évaluer est la linéarité au sens de Spearman. Le coefficient de corrélation de Spearman ρ est une mesure non paramétrique de dépendance statistique entre deux variables. Il s'agit en réalité d'étudier la corrélation au sens de Pearson, non pas sur les variables elles-mêmes mais sur les rangs des variables.Ainsi ce coefficient évalue une corrélation dite de rang alors que le coefficient de Pearson évalue une corrélation de type affine. Pour des échantillons de taille n, X_i et Y_i, de deux variables X et Y, que l'on convertit en rangs x_i et y_i, on peut calculer $\rho_{X,Y}$ par :

$$\rho_{X,Y} = 1 - \frac{6\sum(x_i - y_i)^2}{n(n^2-1)} \tag{2.2}$$

En valeur absolue, ce coefficient vaudra 1 pour une forte corrélation de rang et 0 pour aucune corrélation de rang.

Visualisation de la non-Gaussianité En première approche, il est possible de se faire une idée visuelle de la forme d'une densité (unidimensionnelle). En rangeant les éléments d'un échantillon dans un histogramme, c'est à dire dans un certain nombre (nombre arbitraire et discutable) d'intervalles découpant l'espace de la variable, il est possible de visualiser une approximation de la densité décrite par l'échantillon. Cet histogramme s'appelle un histogramme de densité. Bien que qualitative, cette méthode permet rapidement d'observer des écarts à la non-Gaussianité d'un échantillon.

Evaluations statistiques de la non-Gaussianité Il existe de nombreux tests statistiques évaluant la Gaussianité d'un ensemble, citons par exemple : le test de D'Agostino-Pearson (D'Agostino and Pearson, 1973), le test de Kolmogorov-Smirnov (Lilliefors, 1967), le test de Anderson-Darling (Anderson and Darling, 1952), le test de Shapiro-Wilk (Shapiro and Wilk, 1965). La plupart de ces tests, évaluent une distance entre la fonction de répartition (ou *cumulative distribution function*, cdf) de l'ensemble considéré et celle d'une loi Gaussienne. Des distances fréquemment utilisées sont l'asymétrie (*skewness*) et le *kurtosis* de l'ensemble, respectivement, les moments d'ordre trois et quatre de la distribution.

Nous utilisons à plusieurs reprises dans la suite, le test de D'Agostino-Pearson. Comme beaucoup d'autres tests, il combine deux coefficients mesurant l'asymétrie et le kurtosis de l'ensemble. L'idée de base est de normaliser les mesures de l'asymétrie et du kurtosis (en se basant sur la taille de l'ensemble) et d'additionner ces deux valeurs. Ce test peut évaluer de fortes déviations à la Gaussianité pour des ensembles de taille supérieure à 20 membres indépendants et identiquement distribués. Statistiquement, il s'agit d'un test de normalité omnibus ayant comme hypothèse nulle que l'échantillon évalué provient d'une

distribution Gaussienne. La p-valeur produite par ce test est une probabilité χ^2 à deux côtés pour l'hypothèse nulle. Cette hypothèse est rejetée avec une significativité de 95% pour une p-valeur inférieure à 0.05.

Nous avons également mis en place un test évaluant la distance entre l'histogramme de rangs produit par notre ensemble et l'histogramme de rangs produit par une Gaussienne ayant même moyenne et même écart-type. Un histogramme de rangs (présenté plus en détails lors de la présentation du filtre d'histogrammes de rangs en Sec. 3.2.2) est un diagramme permettant de reconstruire une densité de probabilité (pdf). Il se construit en affiliant un poids équivalent aux intégrales sur chaque intervalle entre deux particules adjacentes de la pdf. Un schéma illustratif de ce processus est proposé dans la Figure 4.3. La distance que l'on regarde est l'intégrale de la valeur absolue entre les deux histogrammes. Cette métrique a pour avantage de prendre en compte l'importance de l'hypothèse de Gaussianité de l'assimilation (e.g. par un filtre de Kalman d'ensemble) par rapport aux filtres utilisant des histogrammes de rangs (décrits en Sec. 2.3.2 de ce chapitre et au Chap. 4).

Remarque statistique importante Il est important de préciser à ce stade que ces scores sont des scores statistiques dont la validité est soumise à bon nombre de règles. La significativité (en fonction de la taille de l'échantillon notamment) est importante pour parler de deux échantillons linéairement corrélés ou d'un échantillon Gaussien. Faire parler des statistiques est quelque chose de complexe et à manier avec précaution. Cependant, dans les travaux présentés ici, l'utilisation de ces scores est uniquement comparative. Il est donc possible que par abus de langage nous parlions de variables linéaires ou Gaussiennes, il s'agira en réalité d'une affirmation relative aux autres variables. Tel ou tel échantillon est plus linéaire ou plus Gaussien qu'un autre selon un score. Cette démarche est justifiée par l'objectif de vouloir prévoir si une méthode de moindes carrés sera *a priori* plus performante sur une variable (plus généralement, sur un problème) ou sur une autre.

2.1.2 Sources de non-Gaussianité en assimilation de données

Les sources éventuelles de non-Gaussianité sont la non-Gaussianité intrinsèque des variables physiques, la non-linéarité des modèles et la non-linéarité des opérateurs d'observations. Ces sources de non-Gaussianité affectent l'assimilation de données à plusieurs niveaux.

Erreurs d'observations non-Gaussiennes

Les erreurs d'observations, contrairement aux hypothèses des méthodes moindres carrés, peuvent ne pas être Gaussiennes. D'après Pires et al. (2010), " Une des sources possibles de non-Gaussianité est l'asymétrie statistique et la positivité intrinsèques à certaines variables physiques (comme la pluviométrie, l'humidité dans l'atmosphère, les concentra-

tions chimiques d'aérosols atmosphériques ou des organismes en biogéochimie marine). " La présence de non-Gaussianité à ce niveau peut être donc due à ces caractéristiques intrinsèques des variables observées (e.g. les variables contraintes par des seuils). Typiquement, l'humidité de l'air (Dee and Da Silva, 2003) ou des concentrations en glace et en phytoplancton (Brankart et al., 2012) ne sont jamais des quantités négatives. Ainsi, le domaine de définition de la distribution de la variable aléatoire "erreur d'observation" est borné donc la variable aléatoire ne peut pas être Gaussienne. Par exemple, l'erreur en concentration de chlorophylle est souvent associée à une variable log-normale.

La seconde source d'erreurs d'observations non-Gaussienne est la non-linéarité des opérateurs d'observations. Si une quantité mesurée est Gaussienne mais que l'opérateur d'observation est non-linéaire, la Gaussianité sera détériorée lors de l'assimilation.

Vecteurs d'état *a priori* non-Gaussiens

Certains articles différencient, comme sources de non-Gaussianité, les non-linéarités du modèle et les non-Gaussianités intrinsèques aux *a priori*. Nos propos dans cette thèse ne feront pas directement la distinction entre ces deux sources puisque, comme souvent en assimilation séquentielle, les *a priori* que nous utilisons sont le produit de la dynamique non-linéaire des modèles. Ainsi nous parlons tout au long de ce manuscrit d'*a priori* non-Gaussien en gardant à l'esprit que ces non-Gaussianités proviennent en partie des non-linéarités des modèles utilisés.

Nous utilisons également le fait que la linéarité inter-variables est une condition nécessaire à la Gaussianité (Sec. 2.1.1). Cette notion prend de l'importance lors de la mise en place d'une assimilation moindres carrés classique (e.g. les filtres de Kalman) qui propage ses corrections à travers les variables non-observées par régression linéaire. Dans ce cas là, les performances de ce type de méthode pour un vecteur de contrôle non-Gaussien ne sont plus assurées.

Les non-Gaussianités en géosciences

Les applications de l'assimilation de données sont en train de se développer au delà de la météorologie et de l'océanographie. Dans ces nouveaux domaines, les systèmes dynamiques sont souvent de dimensions plus modestes mais présentent de très fortes non-linéarités. De manière non exhaustive, nous citons quelques travaux d'assimilation dans ces domaines.

À la croisée entre météorologie et nivologie, un système d'assimilation des réflectances (MODIS) de surfaces visible et infrarouge dans un modèle de manteau neigeux sur les zones de relief en France a été mis en place pour la thèse de doctorat de Luc Charrois (Charrois et al., tted). Au vu des fortes non-Gaussianités de ce système, la méthode d'assimilation mise en place est un filtre particulaire. En géomagnétisme, les travaux de Fournier et al. (2010) discutent de l'application de l'assimilation de données aux problèmes d'estimation non-linéaires et non-Gaussiens dans les modèles du noyau terrestre dynamique.

Le géomagnétisme devient même un terrain expérimental pour l'assimilation de données de pointe (Morzfeld and Chorin, 2012). Nejadi et al. (2014) reconstruit les distributions non-Gaussiennes des faciés géologiques en utilisant un filtre de Kalman d'ensemble avec un ré-échantillonnage pondéré des probabilités.

Des travaux récents sur d'autres applications fortement non-Gaussiennes sortent même du domaine des géosciences comme par exemple la gestion des systèmes éléctriques (Gogonel et al., 2014).

2.2 Le formalisme de l'assimilation de données

2.2.1 Le cadre du problème d'estimation et ses notations

En géosciences, Les incertitudes sur les modèles et sur les observations engendrent des problèmes d'estimation complexes. Ces problèmes ont la caractéristique d'être de très grandes dimensions ce qui ajoute une difficulté supplémentaire en rendant leur résolution numérique délicate.

Pour aborder au mieux ce type de problèmes nous instaurons, ici, un cadre mathématique de travail.

Le vecteur d'état et le vecteur d'observations On désigne par $\mathbb{O}$ l'espace des observations et par $\mathbb{E}$ l'espace d'état. Ici, $\mathbb{O} = \mathbb{R}^m$ et $\mathbb{E} = \mathbb{R}^n$ avec m et n respectivement la dimension du vecteur d'observations et du vecteur d'état. Dans les domaines d'application que l'on considère $m << n$.

Dans ce qui suit, le vecteur d'état du modèle et le vecteur d'observations sont respectivement notés $\mathbf{x} \in \mathbb{E}$ et $\mathbf{y} \in \mathbb{O}$. Le vecteur d'état et le vecteur d'observations correspondant au pas de temps k sont donc notés $\mathbf{x}_k$ et $\mathbf{y}_k$. On fera également la différence entre le vecteur d'état avant assimilation appelé l'ébauche, l'*a priori* ou le *background*, au temps k, $\mathbf{x}_k^b$ et après assimilation le vecteur d'analyse sera noté $\mathbf{x}_k^a$.

Le modèle et l'opérateur d'observations Le système d'équations régissant la dynamique du modèle est une application $\mathcal{M} : \mathbb{E} \times \mathbb{E} \times T \to \mathbb{E}$ (avec T le temps) telle que :

$$\mathbf{x}_k = \mathcal{M}_{k-1,k}(\mathbf{x}_{k-1}, \eta_{k-1}) \tag{2.3}$$

où $\mathcal{M}_{k-1,k}$ est l'application qui associe à un vecteur d'état au temps t_{k-1} un vecteur d'état au temps t_k. La variable aléatoire η_{k-1} représente l'erreur modèle au temps t_{k-1}. Elle a pour moyenne $E[\eta_{k-1}]$ et pour matrice de covariances $Q_{k-1} = E[(\eta_{k-1} - E[\eta_{k-1}]) \cdot (\eta_{k-1} - E[\eta_{k-1}])^T]$.

Supposer le modèle parfait (ou faire l'hypothèse de contrainte forte dans le jargon variationnel) revient à supposer $\eta_k = 0$ pour tout k. Par ailleurs, si le modèle est linéaire (ou si l'on parle du modèle linéarisé) on le notera M.

L'opérateur d'observation est l'application $\mathcal{H} : \mathbb{E} \times \mathbb{O} \times T \to \mathbb{O}$ qui associe à un vecteur d'état son projeté dans l'espace des observations :

$$\mathbf{y}_k = \mathcal{H}_k(\mathbf{x}_k, \epsilon_k) \tag{2.4}$$

où ϵ_k est la représentation de l'erreur d'observation au temps t_k par une variable aléatoire de moyenne $E[\epsilon_k]$ et de matrice de covariances $R_k = E[(\epsilon_k - E[\epsilon_k]) \cdot (\epsilon_k - E[\epsilon_k])^T]$.

De la même manière, si l'opérateur d'observations est linéaire (ou si l'on parle de l'opérateur linéarisé) on le notera H.

Les erreurs On appelle erreur l'écart d'une quantité à l'état vrai $\mathbf{x}^t$ (état conceptuel généralement inconnu et donc considéré comme un vecteur aléatoire). Ainsi, l'erreur *a priori* (ou erreur de *background*) est définie comme $e^b = \mathbf{x}^b - \mathbf{x}^t$. L'erreur d'analyse est définie comme $e^a = \mathbf{x}^a - \mathbf{x}^t$. Et l'erreur d'observation est définie telle que $\mathbf{y} = \mathcal{H}(\mathbf{x}^t, e^o)$. Ces erreurs sont des vecteurs aléatoires supposés sans biais (i.e. $E[e^b] = E[e^a] = E[e^o] = 0$) et leurs matrices de covariances d'erreurs sont respectivement $P^b = E[e^b \cdot (e^b)^T]$, $P^a = E[e^a \cdot (e^a)^T]$ et $R = E[e^o \cdot (e^o)^T]$.

Le problème d'estimation Le problème d'estimation que l'assimilation se propose de résoudre consiste à déterminer la probabilité $P(\mathbf{x}_{0:K}|\mathbf{y}_{0:K})$ avec $\mathbf{x}_{0:K} = [\mathbf{x}_0, ..., \mathbf{x}_K]$ la collection des vecteurs d'états du modèle du temps 0 au temps K et $\mathbf{y}_{0:K} = [\mathbf{y}_0, ..., \mathbf{y}_K]$ la collection des vecteurs d'observation du temps 0 au temps K, sous la contrainte :

$$\begin{cases} \mathbf{x}_k = \mathcal{M}(\mathbf{x}_{k-1}, \eta_{k-1}), \\ \mathbf{y}_k = \mathcal{H}(\mathbf{x}_k, \epsilon_k), \qquad\qquad \forall k = 0, ..., K \end{cases}$$

Muni de ce cadre formel, nous pourrons par la suite présenter les méthodes d'assimilation de données proposant des solutions, plus ou moins optimales et sous certaines hypothèses, aux différents problèmes d'estimation rencontrés en géosciences.

2.2.2 L'assimilation de données moindres carrés

Dans cette sous partie nous décrivons les grandes approches de l'assimilation de données, leurs formalismes et l'esprit dans lequel elles ont été développées. Chacune de ces approches est également présentée avec une vision probabiliste afin de mettre en évidence les hypothèses respectives des méthodes et les éventuels impacts de ces hypothèses sur les performances de l'assimilation. L'objectif est de réunir les grandes approches de l'assimilation de données moindres carrés sous un même formalisme probabiliste plus général qu'est la théorie Bayesienne pour mettre en évidence leurs hypothèses communes. Dans un premier temps, nous posons le problème d'assimilation dans le cadre Bayesien.Les deux sous parties qui suivent montrent que l'approche variationnelle et l'approche séquentielle sont, sous certaines hypothèses, des applications dérivées du théorème de Bayes.

Le théorème de Bayes

L'un des premiers auteurs à considérer le problème d'estimation (en météorologie) comme un problème Bayesien à résoudre avec des méthodes Bayesiennes, fût Epstein (1962). Les travaux de Lorenc (1986) et de Tarantola (1987) ont été les premiers à rapprocher ces considérations à l'assimilation de données. Cette vision du problème présente l'avantage de regrouper les méthodes d'assimilation provenant de théories différentes sous un même formalisme.

Comme vu précédemment l'un des défis de l'assimilation est de déterminer la probabilité $P(\mathbf{x}_{0:K}|\mathbf{y}_{0:K})$ sous la contrainte :

$$\begin{cases} \mathbf{x}_k = \mathcal{M}(\mathbf{x}_{k-1}, \eta_{k-1}) \\ \mathbf{y}_k = \mathcal{H}(\mathbf{x}_k, \epsilon_k) \end{cases}$$

Ce problème peut être vu comme la recherche de la probabilité d'état du modèle conditionnelle aux observations. En oubliant temporairement les indices de temps, le problème peut se formuler de la manière suivante.

Considérons X la variable aléatoire d'état du modèle, sa densité de probabilité pour un argument $\mathbf{x}$ s'écrit $p_X(\mathbf{x})$. De même, on peut voir une observation comme une réalisation de la variable aléatoire observation Y, sa densité de probabilité pour un argument $\mathbf{y}$ s'écrit $p_Y(\mathbf{y})$. Ainsi le problème se résume à trouver $p_{X|Y}(\mathbf{x})$, ce que l'on peut décomposer avec le théorème de Bayes par

$$p_{X|Y}(\mathbf{x}|\mathbf{y}) = \frac{p_{Y|X}(\mathbf{y}|\mathbf{x})p_X(\mathbf{x})}{p_Y(\mathbf{y})} \tag{2.5}$$

Par la suite et en vue d'alléger les notations nous omettons les indices des pdf indiquant les variables aléatoires considérées ainsi $p_V(\mathbf{v})$ devient $p(\mathbf{v})$ pour la variable aléatoire V, l'argument $\mathbf{v}$ étant suffisant à la compréhension.

L'un des grands challenges de l'assimilation est de développer des méthodologies applicables aux systèmes de grandes dimensions. Ainsi une des premières difficultés à surmonter est de limiter le coût numérique des méthodes. Dans cet esprit l'une des premières approches développées est une technique empirique de rappel du modèle aux observations (ou *nudging*).

Le rappel aux observations

Les méthodes de rappel ne se basent sur aucun résultat d'optimalité de la solution produite. Ceci étant dit, elles ont été extensivement étudiées en géosciences principalement pour la facilité de leur mise en place et pour leur faible coût numérique. Le principe consiste à introduire dans les équations du système un terme dit de relaxation pour rappeler le modèle vers les observations. Cette méthode prend donc en considération l'état du modèle

et les observations mais en les supposant parfaits puisque leurs erreurs respectives ne sont pas prises en compte.

Malgré la simplicité de ce type de méthodes, les résultats obtenus sont bons sur certains problèmes au vu du coût qu'ils requièrent. De nombreuses études s'y sont intéressées, particuliérement en océanographie dans les années quatre-vingt dix (Verron and Holland (1989), Verron (1992) et Blayo et al. (1994)).

Depuis, des variantes du simple rappel, plus complexes, ont été développées, comme par exemple le nudging d'ensemble par Lei et al. (2012) ou le *Back and Forth Nudging* (BFN) par Auroux and Blum (2005). Ces nouvelles variantes sont plus chères qu'un simple rappel mais se sont montrées efficaces dans certaines applications. Des travaux supplémentaires semblent tout de même nécessaires avant l'utilisation de ces variantes dans une application réaliste de grandes dimensions.

L'approche variationnelle

Issu de la théorie du contrôle optimal, le formalisme variationnel (e.g. 4D-Var) cherche à obtenir (ou approcher) la solution du problème d'estimation en minimisant une certaine fonction dite *fonctionnelle* ou *fonction coût*.

Formulation On décrit ici la formulation variationnelle quadridimensionnelle appelée 4D-Var en contrainte forte, c'est à dire en considérant le modèle comme parfait. La formulation variationnelle prenant en compte l'erreur modèle est dite en contrainte faible et ne sera pas discutée.

Le 4D-Var peut se comprendre comme un compromis entre un a priori, une trajectoire du modèle et la totalité des observations sur une fenêtre temporelle définie (Figure 2.1).

Ainsi, en définissant une fonction coût J pour une fenêtre temporelle $[0, K]$ telle que :

$$J(\mathbf{x}_0) = \frac{1}{2}||\mathbf{x}_0 - \mathbf{x}_0^b||_{P_0^b} + \frac{1}{2}\sum_{k=0}^{K}||\mathbf{y}_k - \mathcal{H}_k(\mathbf{x}_k)||_{R_k} \tag{2.6}$$

avec $||\cdot||_A =< \cdot, A^{-1}\cdot >$ la norme de Mahalanobi associée à la matrice inversible A et $\mathbf{x}_k = \mathcal{M}_{0,k}(\mathbf{x}_0)$ la propagation au temps t_k de la condition initiale $\mathbf{x}_0$.

Sous les hypothèses de linéarité du modèle et des opérateurs d'observation, et en sachant que les matrices P^b et R_k sont définies positives, on a la convexité de la fonction coût J et donc existence et unicité d'un minimum. On peut également montrer analytiquement que sous ces hypothèses, l'état issu de la propagation du minimum de J est identique à la solution optimale de l'analyse BLUE.

En sortant du cadre de ces hypothèses, la minimisation de J peut s'avérer difficile voire inconcluante.

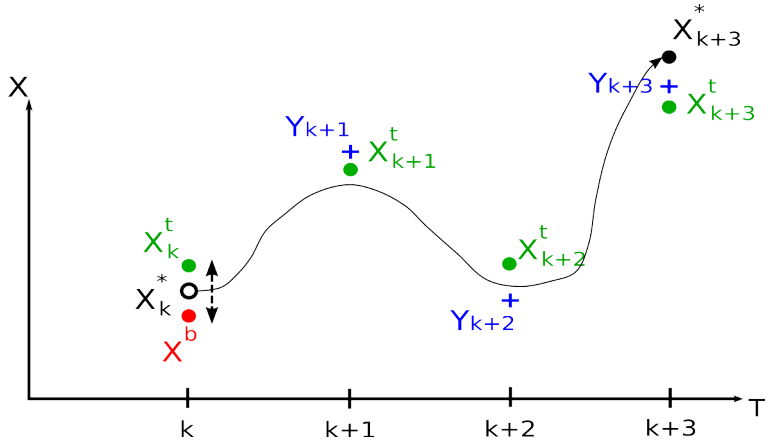

Figure *2.1* – *Description schématique du processus de l'assimilation variationnelle 4D-Var. L'état initial de la fenêtre d'assimilation x^*_k est calculé par minimisation de la fonction coût J (Eq. 2.6). La fonction J est la somme des écarts de la trajectoire du modèle (initialisé par x^*_k) avec le background et avec les observations, pondérés par l'inverse de leurs incertitudes respectives.*

Interprétation probabiliste L'assimilation de données variationnelle en contrainte forte décrite en Section 1.2, cherche à estimer l'état initial $\mathbf{x}_0$ à partir d'un a priori initial $\mathbf{x}^b_0$ et des observations $(\mathbf{y}_0, ..., \mathbf{y}_k)$ sur une fenêtre $[t_0, t_K]$. En d'autres termes, on cherche l'état initial maximisant la probabilité $p(\mathbf{x}_0|\mathbf{y}_{0:K})$. Le théorème de Bayes nous permet d'écrire :

$$p(\mathbf{x}_0|\mathbf{y}_{0:K}, \mathbf{x}^b_0) = \frac{p(\mathbf{x}_0|\mathbf{x}^b_0)p(\mathbf{y}_{0:K}|\mathbf{x}_0, \mathbf{x}^b_0)}{p(\mathbf{y}_{0:K}|\mathbf{x}^b_0)}, \tag{2.7}$$

ou la relation de proportionnalité

$$p(\mathbf{x}_0|\mathbf{y}_{0:K}, \mathbf{x}^b_0) \propto p(\mathbf{x}_0|\mathbf{x}^b_0)p(\mathbf{y}_{0:K}|\mathbf{x}_0, \mathbf{x}^b_0). \tag{2.8}$$

Choisir pour estimateur le maximum de la densité a posteriori, revient donc à chercher

$$\mathbf{x}^*_0 = arg \max_{\mathbf{x}^0 \in \mathbb{E}} [p(\mathbf{x}_0|\mathbf{y}_{0:K}, \mathbf{x}^b_0)] \tag{2.9}$$

avec $\mathbb{E}$ l'espace des états du modèle.

En supposant le bruit a priori $\mathbf{e}^b_0$ et les erreurs d'observations $\mathbf{e}^o_k$ Gaussiens et blancs de lois respectives $\mathcal{N}(0, P^b)$ et $\mathcal{N}(0, R_k)$, on peut écrire, par le théorème de Bayes, le

développement suivant.

$$
\begin{aligned}
\mathbf{x}_0^* &= arg\min_{\mathbf{x}_0\in\mathbb{E}}[ln(p(\mathbf{x}_0|\mathbf{y}_{0:K},\mathbf{x}_0^b))] \\
&= arg\min_{\mathbf{x}_0\in\mathbb{E}}[-ln(p(\mathbf{x}_0|\mathbf{y}_{0:K},\mathbf{x}_0^b))] \\
&= arg\min_{\mathbf{x}_0\in\mathbb{E}}[-ln(p(\mathbf{x}_0|\mathbf{x}_0^b)p(\mathbf{y}_{0:K}|\mathbf{x}_0,\mathbf{x}_0^b))] \\
&= arg\min_{\mathbf{x}_0\in\mathbb{E}}[-ln(e^{-(\mathbf{x}_0-\mathbf{x}_0^b){P^b}^{-1}(\mathbf{x}_0-\mathbf{x}_0^b)}e^{\sum_{k=0}^{K}-(\mathbf{y}_k-\mathcal{H}_k(\mathbf{x}_k))R_k^{-1}(\mathbf{y}_k-\mathcal{H}_k(\mathbf{x}_k))})] \\
&= arg\min_{\mathbf{x}_0\in\mathbb{E}}[(\mathbf{x}_0-\mathbf{x}_0^b){P^b}^{-1}(\mathbf{x}_0-\mathbf{x}_0^b)+\sum_{k=0}^{K}(\mathbf{y}_k-\mathcal{H}_k(\mathbf{x}_k))R_k^{-1}(\mathbf{y}_k-\mathcal{H}_k(\mathbf{x}_k))] \\
&= arg\min_{\mathbf{x}_0\in\mathbb{E}}[\frac{1}{2}(\mathbf{x}_0-\mathbf{x}_0^b){P^b}^{-1}(\mathbf{x}_0-\mathbf{x}_0^b)+\sum_{k=0}^{K}(\mathbf{y}_k-\frac{1}{2}\mathcal{H}_k(\mathbf{x}_k))R_k^{-1}(\mathbf{y}_k-\mathcal{H}_k(\mathbf{x}_k))] \\
\mathbf{x}_0^* &= arg\min_{\mathbf{x}_0\in\mathbb{E}}[J(\mathbf{x}_0)]
\end{aligned}
$$

où J est la fonction coût du 4D-Var 2.6 introduite dans la Section 1.2 et avec $\mathbf{x}_k = \mathcal{M}_{0,k}(\mathbf{x}_0)$. Lorsque l'hypothèse de Gaussianité des erreurs n'est pas vérifiée, l'optimalité de la solution (au sens du meilleur estimé linéaire et sans biais) n'est pas garantie.

L'approche du filtre de Kalman

L'interpolation statistique L'interpolation statistique est une des méthodes d'assimilation de données les plus simples ayant un fondement théorique. Il s'agit d'une méthode dite séquentielle atemporelle, c'est à dire en considérant les observations les unes après les autres dans le temps.

L'interpolation statistique réalise une régression linéaire en supposant que le vecteur d'état corrigé peut s'écrire comme une combinaison linéaire entre l'état a priori et l'observation. Ainsi la solution recherchée s'écrit :

$$\mathbf{x}^a = C_1\mathbf{x}^b + C_2\mathbf{y} \tag{2.10}$$

avec C_1 et C_2 deux matrices, coefficients de la combinaison linéaire.

En faisant, de plus, l'hypothèse de linéarité de l'opérateur d'observation $\mathcal{H}$, i.e. $\mathcal{H} \equiv H$, on trouve que :

$$\mathbf{x}^a = \mathbf{x}^b + K(\mathbf{y} - H\mathbf{x}^b), \tag{2.11}$$

avec K un opérateur linéaire ($\mathbb{R}^m \rightarrow \mathbb{R}^n$) appelé matrice de gain et $(\mathbf{y} - H\mathbf{x}^b)$ est un vecteur appelé l'innovation.

Afin de déterminer la matrice de gain la plus adaptée, on utilise la théorie de l'estimation optimale. Si l'on cherche un estimateur $\mathbf{x}^a$ sans biais et minimisant la trace de la matrice

des covariances d'erreur d'analyse P^a on obtient ce que l'on appelle l'analyse BLUE (*Best Linear Unbiased Estimator*). Et sa matrice de gain associée nous est donnée par :

$$K_{BLUE} = P^b H^T (R + H P^b H^T)^{-1} \tag{2.12}$$

où P^b est la matrice des covariances d'erreurs *a priori* et R est la matrice des covariances d'erreur d'observations. On montre également que les covariances d'erreurs *a posteriori* sont réduites à :

$$P^a = (Id - K_{BLUE} H) P^b \tag{2.13}$$

Nous verrons par la suite que ce résultat, bien que son application directe soit impossible à réaliser dans des systèmes de grandes dimensions, est équivalent sous certaines hypothèses aux méthodes d'assimilation opérationnelles.

Le filtre de Kalman Le filtre de Kalman est une interpolation séquentielle linéaire qui contrairement à l'interpolation statistique tâche de résoudre le problème d'estimation 4D, c'est à dire en prenant en compte la dimension temporelle des statistiques d'erreurs. Cette approche a été développée par Kalman (1960). Elle effectue comme précedemment une estimation à partir d'un a priori et d'une observation mais elle utilise l'information supplémentaire venant de la succession d'analyses reliées par l'évolution temporelle du modèle (ici du modèle linéaire ou linéarisé). Pour prendre en compte cette nouvelle source d'information, nous changeons un peu les notations en appelant l'a priori à un temps k : $\mathbf{x}_k^b = \mathbf{x}_k^f$ où $\mathbf{x}^f$ est l'état analysé au temps précédent propagé (*forecast*) par le modèle (linéaire ou linéarisé) i.e. $\mathbf{x}_k^f = M_{k,k-1}\mathbf{x}_{k-1}^a + \eta_{k-1}$. Ce processus est schématisé en Figure 2.2. De plus, l'opérateur d'observation est également supposé linéaire (ou linéarisé), i.e. $\mathcal{H} \equiv H$.

L'erreur modèle η_k et l'erreur d'observation ϵ_k sont supposées additives, sans biais, décorrelées dans le temps et indépendantes l'une de l'autre, c'est à dire que les contraintes du problème d'estimation deviennent :

$$\begin{cases} \mathbf{x}_k = M\mathbf{x}_{k-1} + \eta_{k-1} \\ \mathbf{y}_k = H\mathbf{x}_k + \epsilon_k \end{cases}$$

et que l'on suppose

$$\begin{aligned} E[\eta_k] &= 0, && (2.14)\\ E[\eta_k \eta_l^T] &= 0 \quad \text{si } k \neq l, && (2.15)\\ E[\epsilon_k] &= 0, && (2.16)\\ E[\epsilon_k \epsilon_l^T] &= 0 \quad \text{si } k \neq l, && (2.17)\\ E[\eta_k \epsilon_l^T] &= 0 \quad \forall k, l \in \mathbb{N}^*. && (2.18) \end{aligned}$$

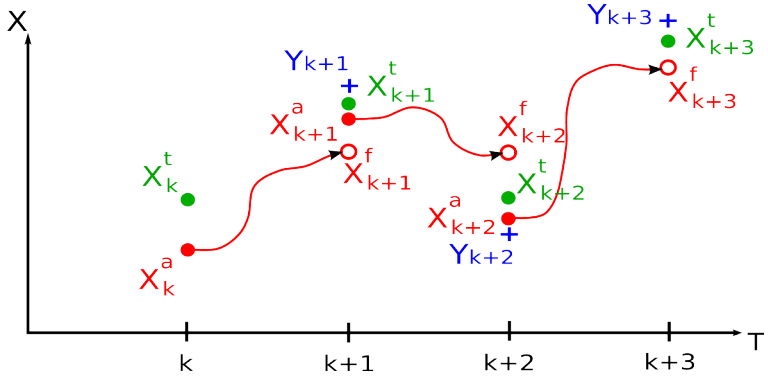

Figure 2.2 – *Description schématique du processus de l'assimilation séquentielle. À chaque temps observé (e.g. $k+1$), l'assimilation calcule l'analyse $\boldsymbol{x}^a_{k+1}$ comme la somme de l'a priori $\boldsymbol{x}^f_{k+1}$ et de l'écart entre l'a priori et l'observation dans l'espace des observations $\boldsymbol{y}_k - H\boldsymbol{x}^f_{k+1}$ (l'innovation) pondéré par la matrice de gain de Kalman (Eq. 2.30). L'analyse au temps $k+1$, $\boldsymbol{x}^a_{k+1}$,, est ensuite propagée par le modèle au temps $k+2$ et devient l'a priori $\boldsymbol{x}^f_{k+2}$. Le procédé est ainsi répété.*

Sous couvert des hypothèses de linéarité et d'estimateur sans biais, l'étape d'analyse est identique à celle du BLUE. On obtient donc

$$\mathbf{x}^a_k = \mathbf{x}^f_k + K_k(\mathbf{y}_k - H_k\mathbf{x}^f_k), \tag{2.19}$$

avec K_k la matrice de gain (similaire à K_{BLUE}) nommée gain de Kalman avec $K_k = P^f_k H^T_k (H_k P^f_k H^T_k + R_k)^{-1}$. Il convient également de mettre à jour la matrice de covariances d'erreurs a priori pour obtenir la matrice de covariances d'erreurs d'analyse, par un calcul algébrique direct on obtient $P^a_k = (Id - K_k H_k)P^f_k$ avec Id la matrice identité.

L'étape de prévision qui s'en suit se résume à la propagation par le modèle de l'état analysé et de la matrice de covariances des erreurs d'analyse :

$$\mathbf{x}^f_{k+1} = M_{k,k+1} \cdot \mathbf{x}^a_k + \eta_k \tag{2.20}$$

$$P^f_{k+1} = M_{k,k+1} P^a_k M^T_{k,k+1} + Q_k. \tag{2.21}$$

Interprétation probabiliste Pour la totalité des observations passées du temps t_0 au temps t_k, $(\mathbf{y}_0, ...\mathbf{y}_k)$, ceci se traduit par $p(\mathbf{x}_k|\mathbf{y}_{0:k})$ avec la notation $\mathbf{y}_{0:k} = (\mathbf{y}_0, ...\mathbf{y}_k)$.

La solution apportée par le théorème de Bayes est

$$p(\mathbf{x}_k|\mathbf{y}_{0:k}) = \frac{p(\mathbf{x}_k|\mathbf{y}_{0:k-1})p(\mathbf{y}_k|\mathbf{y}_{0:k-1}, \mathbf{x}_k)}{p(\mathbf{y}_k|\mathbf{y}_{0:k-1})}, \tag{2.22}$$

où $p(\mathbf{x}_k|\mathbf{y}_{0:k-1})$ est la densité de l'état a priori, $p(\mathbf{y}_k|\mathbf{y}_{0:k-1}, \mathbf{x}_k)$ est la fonction de vraisemblance des observations et $p(\mathbf{y}_k|\mathbf{y}_{0:k-1})$ est la densité d'observations. En supposant que l'a

priori $\mathbf{x}_k$ est issu d'une analyse précédente ayant vu les observations $\mathbf{y}_{0:k-1}$ et en faisant également l'hypothèse *memoryless* de Markov (justifiant les méthodes séquentielles), on obtient :

$$p(\mathbf{x}_k|\mathbf{y}_k) = \frac{p(\mathbf{x}_k)p(\mathbf{y}_k|\mathbf{x}_k)}{p(\mathbf{y}_k)}. \tag{2.23}$$

Avec pour hypothèse que l'opérateur $\mathcal{H}$ est linéaire (que l'on écrit donc H) et que les distributions des variables $\mathbf{x}_k$ et $\mathbf{y}_k|\mathbf{x}_k$ sont respectivement de lois normales $\mathcal{N}(\mathbf{x}_k^f, P_k^f)$ et $\mathcal{N}(H\mathbf{x}_k, R_k)$, on peut écrire leurs densités :

$$p(\mathbf{x}_k) = \frac{1}{(2\pi)^{n/2}|P_k^f|^{1/2}} exp(-\frac{1}{2}(\mathbf{x}_k - \mathbf{x}_k^f)^T(P_k^f)^{-1}(\mathbf{x}_k - \mathbf{x}_k^f)), \tag{2.24}$$

et

$$p(\mathbf{y}_k|\mathbf{x}_k) = \frac{1}{(2\pi)^{m/2}|R_k|^{1/2}} exp(-\frac{1}{2}(\mathbf{y}_k - H\mathbf{x}_k)^T R_k^{-1}(\mathbf{y}_k - H\mathbf{x}_k)). \tag{2.25}$$

Par le théorème de Bayes et après renormalisation nous obtenons la densité a posteriori :

$$p(\mathbf{x}_k|\mathbf{y}_k) = \frac{|HP_k^f H^T + R_k|^{1/2}}{(2\pi)^{m/2}|R_k|^{1/2}(2\pi)^{n/2}|P_k^f|^{1/2}} exp(-\frac{1}{2}(\mathbf{x}_k - \tilde{\mathbf{x}}_k)^T(P_k^a)^{-1}(\mathbf{x}_k - \tilde{\mathbf{x}}_k)), \tag{2.26}$$

avec $(P_k^a)^{-1} = (P_k^f)^{-1} + H^T R_k^{-1} H$ et $\tilde{\mathbf{x}}_k = P_k^a(H^T R_k^{-1}\mathbf{y}_k + (P_k^f)^{-1}\mathbf{x}_k^f)$.

Si l'on cherche un estimateur sans biais maximisant la densité de probabilité, on trouve :

$$\mathbf{x}_k^* = ((P_k^f)^{-1} + H^T R_k^{-1} H)^{-1}(H^T R_k^{-1}\mathbf{y}_k + (P_k^f)^{-1}\mathbf{x}_k^f) \tag{2.27}$$

ce qui s'écrit en reprenant le gain de Kalman $K_k = ((P_k^f)^{-1} + H^T R_k^{-1} H)^{-1} H^T R_k^{-1}$ introduit dans l'Équation 2.19 (après reformulation utilisant la formule de Sherman-Morrison-Woodbury) :

$$\mathbf{x}_k^* = K_k\mathbf{y}_k + ((P_k^f)^{-1} + H^T R_k^{-1} H)^{-1}(P_k^f)^{-1}\mathbf{x}_k^f) \tag{2.28}$$

et comme $(P_k^a)^{-1} = (P_k^f)^{-1} + H^T R_k^{-1} H$ et que l'équation 2.13 nous donne $P_k^a(P_k^f)^{-1} = (Id - K_k H)$ on obtient

$$\mathbf{x}_k^* = K_k\mathbf{y}_k + (Id - K_k H)\mathbf{x}_k^f \tag{2.29}$$

ou en reformulant

$$\mathbf{x}_k^* = \mathbf{x}_k^f + K_k(\mathbf{y}_k - H\mathbf{x}_k^f) \tag{2.30}$$

Pour résumer, en dérivant le théorème de Bayes, sous les hypothèses de linéarité des opérateurs et de Gaussianité des erreurs, on retrouve bien la formulation du filtre de Kalman (eq. 2.30).

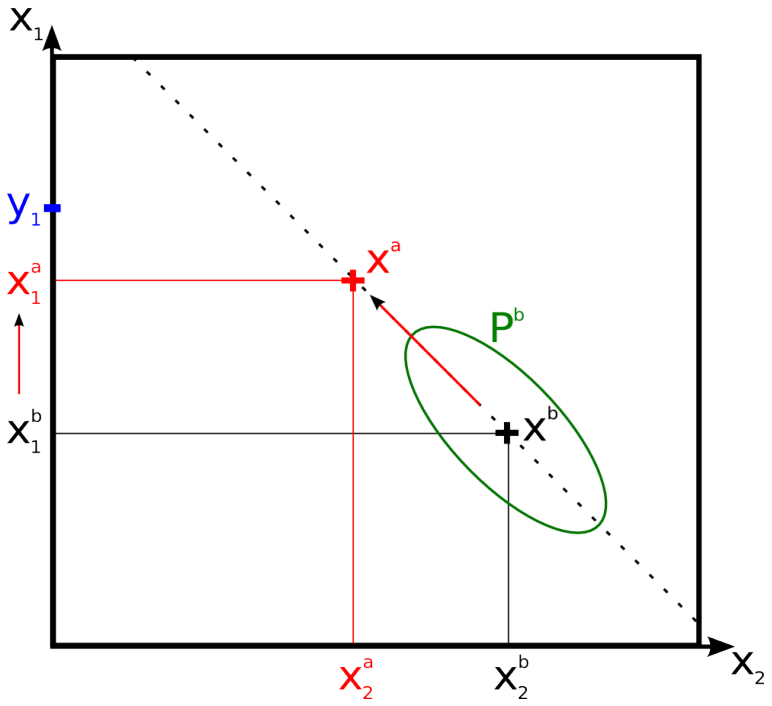

Figure ***2.3*** – *Représentation idéalisée du filtre de Kalman sous le formalisme d'Anderson. La variable x_1^b est corrigée en x_1^a avec l'observation y_1 (pas d'opérateur d'observation h, ici). L'ellipse P^b représente la matrice de covariances d'erreurs a priori dont on déduit la droite de régression linéaire (en pointillé). La variable x_2^b est corrigée en x_2^a par régression le long de cette droite.*

Le filtre de Kalman et le formalisme d'Anderson On peut également voir le filtre de Kalman non dans sa forme matricielle mais dans une forme multivariée. Ce formalisme est proposé par Anderson (2003). Ce formalisme est également en lien direct avec l'implémentation du filtre de Kalman d'ensemble avec traitement des observations en série proposée par Houtekamer and Mitchell (2001).

Il s'agit de considérer séparemment la variable observée x_1 (observée par y_1 *via* l'opérateur h) et les variables non-observées x_i. Le filtre de Kalman effectue alors les corrections en série avec une première correction (scalaire) de la variable observée :

$$x_1^a = x_1^b + K(h(x_1^b) - y_1), \tag{2.31}$$

avec $K = \frac{\sigma_{y_1}^2}{\sigma_{y_1}^2 + \sigma_{x_1^b}^2}$ où $\sigma_{y_1}^2$ et $\sigma_{x_1^b}^2$ sont respectivement les variances d'erreurs de l'observation y_1 et de l'état *a priori* x_1^b. Il s'agit d'une correction unidimensionnelle linéaire de Kalman. Cette correction est optimale pour la variable observée dans le cas où le modèle est linéaire et la variable (aléatoire) x_1 est gaussienne.

La correction des variables non-observées n'est alors qu'une régression linéaire propa-

geant la correction de x_1 sur les autres variables :

$$x_i^a = x_i^b + C(x_1^a - x_1^b). \tag{2.32}$$

Le coefficient de la régression linéaire C est, dans le cadre du filtre de Kalman, donné par $\frac{Cov(x_1^b, x_i^b)}{\sigma_{y_1}^2}$.Il ressort de cette représentation que la relation entre x_1^b et x_i^b doit être linéaire pour que la correction des variables non-observées soit optimale.

En résumé Réécrites dans un cadre probabiliste, les méthodes moindres carrés apparaissent comme équivalentes et effectuent les mêmes hypothèses. Il a été montré, à de nombreuses reprises, que ces méthodes peuvent produire des résultats satisfaisants dans de nombreux contextes. De plus, leurs utilisations dans les centres opérationnels témoignent de leurs bons rapports faisabilité-résultats. Cependant nous sommes assurés de l'optimalité des résultats qu'elles produisent uniquement sous les hypothèses de linéarité des opérateurs (le modèle et l'opérateur d'observations) et de Gaussianité des erreurs de l'*a priori* et des observations. Les conditions nécessaires de cette dernière hypothèse peuvent être décomposées d'un point de vue univarié (selon le formalisme d'Anderson) comme la Gaussianité de la variable observée et la co-linéarité entre la variable observée et les variables non-observées.

Les méthodes moindres carrés ont produit de bons résultats au cours des années. Aujourd'hui la complexification des problèmes d'estimation met à mal leurs hypothèses. Nous y consacrons maintenant quelques mots.

2.3 L'assimilation de données non-Gaussienne

2.3.1 Des scores adaptés ?

Originellement, l'objectif de l'assimilation de données était de produire un estimé unique du vecteur d'état considéré. Les objets à gérer étaient alors des éléments de $\mathbb{R}^n$, ce qui simplifie le probléme.

Dans le cadre d'évaluation d'estimé moyen, nous renvoyons le lecteur au livre de Jolliffe and Stephenson (2012) qui répertorie les scores utilisés dans la littérature pour évaluer les prévisions (atmosphériques). Dans ce même cadre, la revue proposée par Gregg et al. (2009) présente un tour d'horizon complet des scores pour l'assimilation de données réelles en océanographie et en biogéochimie marine.

Toutefois avec l'utilisation de méthodes ensemblistes, comme les filtres de Kalman d'ensemble, nous avons accés à plus d'information. L'ensemble permet de propager de nombreuses informations statistiques et de donner un sens probabiliste à la prévision. Les objets à gérer sont alors des fonctions $p(\mathbf{x})$ des n variables.

Sous l'hypothèse de Gaussianité de la variable aléatoire état du système, cette dernière se décrit par son estimé moyen et ses covariances d'erreurs, c'est à dire l'incertitude sur

l'estimé. Connaître l'estimé et ses covariances d'erreurs revient à décrire entiérement la densité de probabilité de l'état.

Lorsque cette hypothèse n'est pas faite on peut voir l'ensemble comme un échantillonnage de la densité de probabilité de l'état. Ainsi, les qualités de l'ensemble produit (qualités décrites dans cette sous partie) sont également à évaluer.

Evaluations graphiques des résultats d'assimilation La visualisation des résultats d'assimilation n'a aucune valeur quantitative. Elle permet pourtant de se faire une idée, parfois poussée, du comportement des méthodes.

- Les **diagrammes de Hovmöller** sont des représentations graphiques d'une variable dans l'espace en fonction du temps. L'avantage de ces représentations est de pouvoir suivre au cours du temps des phénomènes évoluant dans l'espace. Dans les travaux sur le modèle 1D verticale ModECOGeL (Chap. 5), ces diagrammes permettent de voir la totalité du comportement d'une variable.
- On peut également regarder des **séries temporelles** (variable à une profondeur donnée en fonction du temps) et des **profiles verticaux** (variable à un temps donné en fonction de la profondeur). Conceptuellement, ces représentations sont des sous-sections des diagrammes de Hovmöller. Ces représentations présentent tout de même l'avantage de comparer plusieurs simulations sur un même graphique. Par exemple, il est courant de représenter la *vérité*, l'ensemble estimé et l'état moyen estimé sur un même graphique.

Evaluations quantitatives de l'estimé moyen (l'analyse) La moyenne n'est pas le meilleur moyen de tester la validité d'un ensemble. Qui plus est, lorsque l'on traite des distributions non-Gaussiennes la moyenne, la médiane et l'état le plus vraisemblable ne sont plus confondus. Par exemple, si la pdf est bimodale, l'estimé moyen sera entre les deux pics de la pdf où la probabilité est très faible.

Cependant, en première approximation, il s'agit d'un critère valable tant que l'on n'oublie pas que les conclusions obtenues concernent la moyenne et non l'ensemble.

- Le **score de biais** (*Bias score*) est simplement le ratio entre la fréquence d'événements prédits par l'estimé moyen sur la fréquence des événements de la vérification. Un biais inférieur à 1 indique une "sous-prédiction" et un biais supérieur à 1, une "sur-prédiction" des événements. Bien que souvent utilisé en océanographie (Gregg et al., 2009) ce score présente certains désavantages. Il ne rend pas compte de la bonne correspondance de la prédiction par rapport à la vérification. Il ne juge que de la prédiction de l'estimé moyen.
- La racine de la moyenne des erreurs au carré ou *Root Mean Square Error* (**RMSE**) temporelle ou spatiale, est certainement le score le plus utilisé pour valider une assimilation. Ce score calcule une distance entre l'estimé moyen et la vérification pour

une variable physique du système. Cette distance est donnée par :

$$RMSE(\overline{\mathbf{x}^a}) = \sqrt{\frac{1}{|\Omega|}\int_{\Omega}(\overline{\mathbf{x}^a} - v)^2 d\Omega} \tag{2.33}$$

où l'on évalue les performances de l'estimé moyen $\overline{\mathbf{x}^a}$ par rapport à la vérification v sur l'espace Ω. L'espace Ω est le temps dans le cas d'une RMSE temporelle et les trois dimensions d'espace (x, y, z) (ou un sous-espace) dans le cas d'une RMSE spatiale. La RMSE (non-normalisée) est donnée dans l'unité de la variable évaluée.

Evaluations de la qualité de l'ensemble estimé Connaître et savoir évaluer la qualité d'un ensemble apparaît en effet primordial, dès que l'on cesse de voir la prévision ou l'assimilation comme des outils déterministes produisant une unique réalisation. Il semble important de remarquer à ce stade, que si l'on s'est donné comme objectif d'estimer la densité de probabilité *a posteriori* de l'état du système, l'évaluation de l'ensemble *a posteriori* n'est qu'une partie du travail. On pourrait notamment imaginer, des diagnostics prenant en compte les erreurs d'échantillonnage de la densité *a posteriori* afin de mesurer non pas la qualité de l'ensemble mais bien la qualité de la densité elle-même. Ceci étant dit, ces considérations n'étant pas au centre des travaux présentés ici, nous nous contentons d'évaluer la qualité des ensembles.

Ainsi, la première question à se poser est : qu'est ce qu'un bon ensemble ? Et comment évaluer ses qualités ?

Nous souhaitons obtenir, après analyse, des ensembles qui échantillonnent correctement la pdf *a posteriori*. C'est à dire que l'ensemble doit être cohérent avec une pdf solution du théorème de Bayes. Ce n'est pas toujours le cas. Cette notion s'appelle la Bayesiannité de l'ensemble. Une condition nécessaire à la Bayesianité d'un ensemble est sa fiabilité (*reliability*).Par la suite, sont présentés deux tests évaluant cette condition nécessaire. Ces tests étant assez complexes et numériquement lourds à mettre en place, on peut d'abord évaluer la bonne dispersion de l'ensemble. En sachant que la bonne dispersion de l'ensemble est une condition nécessaire à la fiabilité. Une notion indépendante de la Bayesianité mais toutefois importante pour un ensemble est la quantité d'information nouvelle qu'il apporte. Cette notion s'appelle la résolution de l'ensemble. Une autre qualité que l'on peut attendre d'un ensemble est l'équilibre physique de chacun de ses membres.

Nous évoquons à présent certains scores permettant d'évaluer les notions ensemblistes requises. Ces scores ne sont pas tous utilisés par la suite mais permettent d'avoir une idée des scores classiques (encore peu utilisés en océanographie pour certains). Pour une description plus détaillée, nous renvoyons le lecteur intéressé à de très bonnes revues de ces notions et de ces scores dans les articles de Candille and Talagrand (2005), Bröcker (2011) ou plus récemment Gogonel et al. (2014). Pour une revue de travaux d'évaluation d'assimilations en biologie océanique, il est également intéressant de consulter Gregg et al. (2009).

La plupart de ces scores se base sur un élément de vérification. Cet élément peut être la *vérité* directement si l'on travaille en expériences jumelles (ce qui est notre cas), une climatologie ou un jeu d'observations (indépendant du jeu d'observations utilisé pour l'assimilation). Les qualités attendues d'un ensemble et les évaluations associées sont présentées de manière non-exhaustive ci-dessous.

La **dispersion d'un ensemble** est le principe que la vérification doit être statistiquement indiscernable des membres de l'ensemble.

– Le score de la variable aléatoire centrée réduite (**RCRV**) est un score permettant de détailler les caractéristiques de la fiabilité en terme de biais et de dispersion. Ce score calcule l'écart de la moyenne de l'ensemble m à la vérification v pondérée par la racine de la somme des carrés de l'erreur de vérification σ_v et de la dispersion d'ensemble σ :

$$E = \frac{v - m}{\sqrt{\sigma^2 + \sigma_v^2}} \tag{2.34}$$

Le RCRV est alors égal au couple (B, D) avec B et D respectivement la moyenne et la variance de E. Comme son nom l'indique un bon RCRV doit avoir un biais B proche de 0 et une dispersion D proche de 1. Un RCRV $= (0, 1)$ est une condition nécessaire pour pouvoir considérer la vérification comme une réalisation indépendante de la loi de probabilité décrite par l'ensemble.

– Un **histogramme de rangs** est une représentation graphique (en diagramme) de la dispersion de l'ensemble. Il comptabilise le nombre d'occurence de la vérification dans chaque intervalle entre les membres. Il s'agit donc d'un diagramme constitué de $N_e + 1$ intervalles qui dans le cas d'une bonne dispersion est plat. Si l'histogramme est concave l'ensemble est sous-dispersif, si l'histogramme est convexe l'ensemble est sur-dispersif. Le principe de l'histogramme de rangs est, en réalité, d'effectuer une anamorphose de l'ensemble vers une loi uniforme. On évalue la pdf de cette loi uniforme en regardant le diagramme.

La **fiabilité** et la **résolution** d'un ensemble sont deux notions indépendantes. La fiabilité est la cohérence statistique entre les probabilités prédites par l'ensemble et la vérification correspondante. En d'autres termes, un ensemble – prédisant la densité de probabilité $\mathcal{F}$ d'un futur état de notre système – est dit fiable si l'état vrai (la vérification) est distribué selon la densité $\mathcal{F}$. La fiabilité est une condition nécessaire à la Bayesianité de l'ensemble. La résolution est la propension de la distribution des probabilités prédites à varier significativement d'une prédiction à l'autre. En d'autres termes, la résolution donne la quantité d'information qu'ajoute l'ensemble à l'information déjà contenue dans la vérification.

– Le **score de Brier** a été défini par Brier (1950). Il évalue la qualité de prévision probabiliste de l'occurence d'un événement binaire ϵ que l'on se donne. Ce score calcul l'espérance du carré de la différence entre la probabilité prévue de l'événement par l'ensemble p et la fréquence d'occurence observée $p'(p)$ de l'événement. Si l'on note $dg(p)$ la fréquence relative avec laquelle la probabilité p est prédite par l'ensemble, $p'(p)$ est la fréquence d'occurence observée de ϵ quand la probabilité p est prédite

et p_c la fréquence d'occurence climatologique (la vérification) de ϵ. il est possible de décomposer le score de Brier $\mathcal{B}$ en :

$$\mathcal{B} = \int (p - p'(p))^2 dg(p) - \int (p'(p) - p_c)^2 dg(p) + p_c(1 - p_c) \tag{2.35}$$

respectivement, un terme de fiabilité, un terme de résolution et un terme indépendant du système de prédiction dit terme d'*incertitude* (Murphy, 1973).

- La notion de score de Brier peut être étendue à plusieurs événements ce qui donne le score de probabilité de rangs discret (**DRPS**). L'idée est de découper l'espace de la variable que l'on évalue en un certain nombre d'intervalles. On applique alors le score de Brier sur chacun des intervalles et on regarde la moyenne des scores de Brier.
- Le score de probabilité de rangs continu (**CRPS**) est une extension au monde continu du DRPS. On peut le voir comme une métrique de la distance entre la fonction de répartition (cdf) de l'ensemble $F_j(\xi)$ (en une j-ème réalisation) et celle de la vérification. En faisant l'hypothèse d'une vérification parfaite (sans erreur), la fonction de répartition de la vérification est un Heaviside $\mathcal{H}(\xi - v_j))$ valant 1 ou 0 si la vérification v_j est inférieure ou supérieure à la valeur ξ. Le CRPS s'écrit alors :

$$CRPS = \frac{1}{M} \sum_{j=1}^{M} \int_{\mathbb{R}} (F_j(\xi) - \mathcal{H}(\xi - v_j))^2 d\mu(\xi) \tag{2.36}$$

À partir d'une fonction de répartition prédite $F(\xi)$, nous notons $F'_F(\xi)$ la fonction de répartition de la vérification v quand $F(\xi)$ a été prédite, $dg(F)$ est la fréquence avec laquelle F a été prédite et $F_c(\xi)$ est la fonction de répartition de la vérification. La décomposition de Hersbach (2000) permet alors de séparer le score CRPS en trois termes :

$$\begin{aligned} CRPS = & \int dg(F) \int_{\mathbb{R}} (F(\xi) - F'_F(\xi))^2 d\mu(\xi) \\ & - \int dg(F) \int_{\mathbb{R}} (F'_F(\xi) - F_c(\xi))^2 d\mu(\xi) + \int_{\mathbb{R}} F_c(\xi)(1 - F_c(\xi)) d\mu(\xi) \end{aligned} \tag{2.37}$$

respectivement un terme de fiabilité, de résolution et d'incertitude. On peut également écrire le CRPS comme le terme de fiabilité plus un terme de *CRPS potentiel* (résolution + incertitude) qui serait la résolution de l'ensemble si la fiabilité était parfaite.

Équilibre des états Le problème de l'équilibre d'un état est équivalent au problème de la condition initiale dans un modèle. Si l'on initialise un modèle avec un état qui ne respecte pas la dynamique du modèle (i.e. hors de l'attracteur du modèle), il se produit alors ce que l'on appelle un temps d'ajustement ou *spin up*. Il s'agit du temps nécéssaire au modèle pour ramener l'état dans l'attracteur du modèle. Il est important qu'une analyse

produise des états équilibrés au sens du modèle (Bertino et al., 2003). Une analyse qui présente des résultats performants selon les scores précédents perd tout intérêt si ces scores se dégradent dans les pas de temps qui suivent.

- La méthode la plus simple conceptuellement pour évaluer l'équilibre des états d'une analyse est de laisser, à la fin d'une période d'assimilation, l'ensemble se propager librement. On peut ainsi voir la qualité de l'ensemble à maintenir sa bonne estimation sur une longue période de temps sans observation.
- Lorsque l'on dispose d'un ensemble de référence (et non pas seulement d'une vérification comme précédemment) il est possible de mesurer la distance entre l'ensemble que l'on évalue et l'ensemble de référence. Pour ce faire, une métrique possible est la **divergence de Kullback-Leibler** (Kullback, 1959) autrement appelée entropie relative. Si l'ensemble de référence est bien équilibré, nous pouvons indirectement mesurer l'équilibre de notre ensemble en utilisant la divergence de Kullback-Leibler. Dans cet esprit là, ce score est décrit plus en détails en Section 4.4.1 du Chapitre 4. Les applications de la divergence de Kullback-Leibler ne se limitent pas aux diagnostics d'équilibre. Ce score a notamment été utilisé en grande dimension pour évaluer la prédictabilité d'un système par Kleeman (2002).

2.3.2 L'évolution des méthodes non-Gaussiennes

Gérer la non-linéarité des modèles

La première des difficultés pour l'assimilation de données ayant été étudiée, est la non-linéarité des modèles.

Comme on l'a vu, aussi bien le filtre de Kalman que le 4D-var dans leurs versions originelles font l'hypothèse d'un modèle linéaire. Cette hypothèse limitant fortement toutes applications, des versions contournant ce problème ont été développées :

- le filtre de Kalman étendu (EKF) est une approche simple qui consiste à considérer le modèle linéarisé (Jazwinski, 1970). L'équivalence entre le modèle non-linéaire et le modèle linéarisé est une hypothèse très forte dans bon nombre d'applications. Hypothèse qui conduit souvent à des performances limitées.
- les filtres de Kalman de rang réduit ont été développés pour diminuer le coût calcul trop grand du KF et de l'EKF. Ces méthodes sont basées sur le principe (l'hypothèse) que la dynamique d'un système à un moment précis dépend d'un contrôle vivant dans un sous-espace de l'espace des solutions. Le nombre de degrés de liberté d'un problème d'estimation serait donc bien inférieur à la taille du système entier. L'idée est donc de réduire le problème à ce sous-espace pour simplifier les calculs. Certaines méthodes se ramènent aux modes du système (e.g. le filtre RRSQRT), d'autres se ramènent aux composantes principales (EOF) du modèle linéarisé (e.g. le filtre SEEK ; Pham et al., 1998; Brasseur and Verron, 2006). Il est également possible d'approximer le modèle linéarisé par un développement de Taylor à des ordres plus élevés (e.g. le filtre SEIK ;

Pham, 2001).
- le filtre de Kalman d'ensemble (EnKF) a été développé par Evensen (1994),Evensen and Leeuwen (1996) puis par Houtekamer and Mitchell (1998). Ce filtre diffère du KF et de l'EKF dans son étape de prévision. Au lieu de propager les matrices de covariances d'erreurs, les statistiques d'erreurs sont représentées par un échantillon (ensemble) d'états du modèle. En ce sens il s'agit d'un filtre de type Monte-Carlo. L'information est propagée au cours du temps en propageant cet ensemble d'états par les équations d'évolution (non-linéaire). Dans la suite nous utilisons deux versions de filtres de Kalman d'ensemble : le filtre de Kalman d'ensemble stochastique (EnKFs) et le filtre de Kalman transformé (ETKF) qui sont décrits en détails au chapitre suivant (Sec. 3.2.1).

Également, du côté variationnel, la version incrémentale du 4D-var permet des applications aux modèles non-linéaires.

Dans ces travaux de thèse, nous nous consacrons uniquement à l'étude des méthodes séquentielles ensemblistes.

Pour les filtres de Kalman d'ensemble, l'impact direct de la non-linéarité des modèles sur l'assimilation (par direct on entend l'impact lié à la transgression de l'hypothèse de linéarité) est contourné par l'utilisation d'un ensemble. En effet, les matrices de covariances d'erreurs analysées (P_k^a) sont approximées via l'ensemble propagé de manière non-linéaire. Il faut faire attention qu'en réalité l'utilisation d'un ensemble, certes, résout le problème de non-linéarité des modèles mais le remplace en pratique par des problèmes d'échantillonnage lié à la petite taille des ensembles.

Comme il est mentionné à plusieurs reprises dans ce chapitre, l'autre grande hypothèse de l'assimilation moindres carrés est la Gaussianité des distributions. Nous estimons l'ampleur des difficultés engendrées par cette hypothèse dans la suite et nous revenons sur certaines solutions proposées dans la littérature.

Les difficultés de l'assimilation moindres carrés

Comme l'énonce Miller et al. (1999) : "Alors que l'on peut toujours effectuer une analyse moindres carrés sur tous problèmes, l'application directe des moindres carrés peut conduire à des résultats insatisfaisants dans des cas où les distributions sous-jacentes sont significativement non-Gaussiennes." Dans la littérature, de nombreuses études se sont consacrées à vérifier cette assertion.

Nous faisons ici une brève revue de certains articles discutant des difficultés de l'assimilation de données moindres carrés en contexte non-Gaussien.

Les travaux de Miller et al. (1999) présentent justement les limitations du filtre de Kalman étendu et du filtre de Kalman d'ensemble sur trois différents modèles : un modèle 1D

bimodal, le modèle de Lorenz 63 et un modèle QG en canal sur le plan β. Ils ont constaté des mauvaises estimations de la part de ces deux méthodes sans pouvoir déterminer si les mauvais résultats de l'EnKF étaient dus à l'utilisation de l'estimé moyen ou aux hypothèses de moindres carrés (Gaussianité). Bertino et al. (2003) proposent une revue et une comparaison théorique de différents filtres issus de Kalman. Ils concluent que l'application séquentielle d'un estimateur linéaire à un modèle non-linéaire nécessite deux conditions sur le vecteur d'état : i) le vecteur d'état analysé doit être physiquement équilibré (au sens de respecter les propriétés physiques du modèle) ; ii) le vecteur d'état propagé doit présenter une distribution multivariée Gaussienne afin d'optimiser l'utilisation de l'analyse linéaire statistique. Ils testent également l'EnKF sur un modèle d'écosystème simplifié. Plus récemment, Lei et al. (2011) arrivent à des conclusions similaires. Avec une étude théorique et des expériences sur deux modèles simples, ils évaluent les performances d'un filtre de Kalman stochastique et d'un filtre de Kalman déterministe. Ils observent que ces deux EnKF sont sensibles à la violation de l'hypothèse Gaussienne.

Les méthodes d'assimilation (d'ensemble) moindres carrés ont été largement utilisées par la communauté. Pour des systèmes océanographiques, ces méthodes ne sont, en général, pas optimales. Les résultats qu'elles produisent sont néanmoins très bons pour des problèmes peu non-linéaires et peu non-Gaussiens. Par exemple, sur des modèles dynamiques globaux à faible résolution, les filtres de Kalman d'ensemble restent assez performants. Cependant, la littérature des vingt dernières années tend à montrer les limites de ces méthodes Gaussiennes sur des problèmes plus complexes. Ces problèmes plus complexes sont de plus en plus fréquents notamment avec le raffinement des modèles, le couplage ou encore la non-Gaussianité intrinsèque de certaines variables (e.g. le phytoplancton). Il apparait donc nécessaire de trouver des moyens de gérer les non-Gaussianités de manière plus efficace. Pour ce faire, il est possible de mettre en place des stratégies afin d'adapter les méthodes moindres carrés aux non-Gaussianité. Une autre démarche est de repenser le problème comme étant non-Gaussien.

De l'assimilation moindres carrés au monde non-Gaussien

Il existe plusieurs manières d'accomoder les méthodes moindres carrés à des déviations légères à la Gaussianité dans le système d'assimilation. Le travail de Kalnay et al. (2007) reprend certaines des stratégies employées pour améliorer les résultats du filtre de Kalman étendu, EKF, et du filtre de Kalman d'ensemble, EnKF (et du 4D-Var) dans des contextes non-linéaires et non-Gaussiens. Depuis, de nombreuses autres stratégies, visant à renforcer le bon comportement de l'EnKF dans ces contextes, ont vu le jour. Par exemple, les différentes techniques de libre-inflation par *finite size* (Bocquet, 2011) et de localisation (Sakov and Bertino, 2010; Greybush et al., 2011) ; les stratégies de rééchantillonnage (Pham, 2001; Anderson, 2012) ; le ciblage (*targeting*) aux observations (Bishop et al., 2000) ont été et sont encore très étudiés. Bocquet et al. (2010) propose une très bonne revue de

ces stratégies plus récentes.

Dans des cas de non-Gaussianité plus importante, ces méthodes basées sur l'EnKF restent sensibles à la violation de l'hypothèse de Gaussianité (Sakov and Oke, 2008; Lei and Bickel, 2011).

Sans pour autant rejeter certains principes issus des méthodes moindres carrés, de nouvelles méthodes ont vu le jour. Une bonne revue de quelques unes des méthodes évoquées ici se trouve au Chapitre 7 du mémoire d'*Habilitation à Diriger la Recherche* (HDR) de Jean-Michel Brankart (Brankart, 2014).

On l'a vu précedemment, la présence de seuils sur certaines variables traitées (e.g. le phytoplancton) met à mal l'hypothèse Gaussienne. Lauvernet et al. (2009) proposent une solution : les Gaussiennes tronquées, qui consiste à maintenir l'hypothèse Gaussienne sur le vecteur de contrôle mais en ajoutant des contraintes d'inégalitées (e.g. variables positives).

D'autres méthodes plus complètes supposent la densité *a priori* comme une superposition de Gaussienne. L'idée est alors d'appliquer un filtre de Kalman d'ensemble sur chaque Gaussienne locale. De calculer des poids pour chacune de ces Gaussiennes locales *a posteriori*. La densité *a posteriori* est alors recomposée comme somme pondérée de ces Gaussiennes locales *a posteriori*. Plusieurs travaux proposent des méthodes dans cet esprit de mélange de Gaussiennes (Anderson and Anderson, 1999; Bengtsson et al., 2003; Hoteit et al., 2008; Sondergaard and Lermusiaux, 2013).

La non-Gaussianité univariée (sous entendu sur la variable observée) peut déjà avoir de fortes conséquences sur la qualité de l'assimilation. En reprenant le formalisme d'Anderson, présenté en Section 1.2, Anderson (2010) propose une nouvelle méthode : le *Rank Histogram Filter*. La variable observée est corrigée en faisant une application directe du théorème de Bayes (1D) à partir des densités *a priori* construites à l'aide d'histogrammes de rangs (sans hypothèse Gaussienne). Les variables non-observées sont ensuite corrigées par régression linéaire comme le ferait un filtre de Kalman. Cette méthode est décrite plus en détails par la suite.

Une autre difficulté mettant à mal l'hypothèse Gaussienne est la non-linéarité inter-variables (ou des lignes de régression). Un moyen (peu coûteux) de remédier à cette difficulté est l'anamorphose (Bertino et al., 2003). Il s'agit de pratiquer une transformation (changement de variables) unidimensionnelle qui projette chaque distribution marginale en une distribution Gaussienne. Plusieurs anamorphoses existent. Pour une comparaison de différentes anamorphoses (hybride, statique, dynamique et logarithmique), nous renvoyons le lecteur aux travaux de Simon and Bertino (2012). En pratique, nous utilisons l'anamorphose qui transforme les quantiles de l'ensemble en des quantiles correspondant d'une distribution Gaussienne (Béal et al., 2010; Brankart et al., 2012).

Un élan plus récent dans la communauté mais qui motive de nombreux travaux de recherche ainsi que des applications opérationnelles pionnières, est l'hybridation des méthodes du types EnKF et des méthodes variationnelles. Nous évoquons ces nouvelles méthodes mais nous ne les étudions pas d'avantage dans cette thèse.

Les premiers essais d'hybridation ont commencé dans les années 2000. Une hybridation d'un filtre de Kalman d'ensemble et d'un 3D-Var a été proposée par Hamill and Snyder (2002). La possibilité d'une hybridation entre le filtre de rang réduit SEEK et le 4D-Var a été étudiée par Robert et al. (2006). Par la suite le 4DEnVar a été développé (Liu et al., 2008, 2009) et, sous différentes versions, étudié sur des systèmes réalistes par Environnement Canada [1] (Buehner et al., 010a,b), par le MetOffice [2] (Clayton et al., 2013; Lorenc et al., 2015) et par MétéoFrance [3] (Desroziers et al., 2014). Ces méthodes 4DEnVar sont constituées d'une propagation d'ensemble et d'une analyse en deux parties : une analyse de type 3DVar se basant sur la climatologie de la matrice de covariances d'erreurs de background et une analyse 4D consistant en une combinaison linéaire des perturbations de l'ensemble. Il est à noter que l'utilisation directe d'un ensemble de 4DVar est opérationnelle à MétéoFrance (ensemble de 25 membres) et, bien que coûteuse, montre de très bonnes performances (Berre et al., 2015).

Repenser le problème non-Gaussien

Les mathématiques appliquées proposent depuis longtemps des solutions (analytiques ou numériques) à des problèmes non-Gaussiens. En géosciences, l'idée de penser le problème d'estimation comme un problème non-Gaussien a pourtant été rapidement écartée. Le coût numérique que représentent les éventuelles solutions n'était simplement pas envisageable dans des applications à très grandes dimensions de météorologie ou d'océanographie. Aujourd'hui, des raisons de revoir cette position émergent.

- La puissance de calculs des ordinateurs modernes devient très importante ce qui permet d'imaginer appliquer des méthodes complexes à des problèmes en grandes dimensions.
- Les travaux sur le problème de la non-Gaussianité offrent depuis plusieurs années des compromis intéressants entre non-Gaussianité et grandes dimensions.
- Les modèles en géosciences se complexifient, se raffinent et se couplent les uns aux autres. Cette inhomogénéité dans les modèles permet de traiter plusieurs (petits) problèmes en un. Il est donc possible d'imaginer des applications réduites présentant des non-Gaussianités en restant abordables pour les méthodes non-Gaussiennes (Berliner and Wikle, 2007; Hoteit et al., 2008).

Dans les récentes tentatives pour repenser entièrement le problème de manière non-Gaussienne, certaines ont retenu notre attention. Nous évoquons ici brièvement leurs principes et leurs éventuels avantages et inconvénients.

Une méthode d'assimilation séquentielle se basant sur la théorie du transport optimal (ou *mapping* optimal) a été récemment proposée par El Moselhy and Marzouk (2012), Cotter and Reich (2013), et Reich (2013). Son principe n'est pas de rechercher directement

1. Centre de prévisions météorologiques canadien : *www.ec.gc.ca*
2. Centre de prévisions météorologiques britannique : *www.metoffice.gov.uk*
3. Centre de prévisions météorologiques français : *www.meteofrance.com*

à représenter la pdf *a posteriori* mais de trouver une fonction de transfert dite *transfer map* qui redistribue l'échantillon *a priori* en un échantillon distribué selon la pdf *a posteriori*. Cette méthode, bien que prometteuse, reste une méthode de ré-échantillonnage (avec comme limitation principale le besoin d'ensembles de grande taille pour explorer au mieux l'espace des solutions) et semble encore très coûteuse dans sa gestion de la localisation.

Les filtres particulaires (Gordon et al., 1993; van Leeuwen, 2009) sont les méthodes purement non-Gaussiennes les plus populaires. Elles sont toutefois connues pour être sujettes à la malédiction de la dimensionalité ce qui rend difficile leur application aux systèmes de grandes dimensions (Snyder et al., 2008). Les recherches pour répondre à cette difficulté sont très actives (Nakano et al., 2007; van Leeuwen, 2010; Morzfeld et al., 2012; Snyder, 2012). Certaines de ces recherches reposent sur l'hybridation avec un EnKF (Bocquet et al., 2010; Lei and Bickel, 2011; Hoteit et al., 2012).

Entre la théorie du filtre particulaire et de l'hybridation, la méthode du filtre particulaire implicite (IPF) étudié par Chorin et al., 2010; Morzfeld and Chorin, 2012; Morzfeld et al., 2012 et du lisseur de Kalman d'ensemble itératif (IEnKS) développé par Bocquet and Sakov (2012) et Bocquet and Sakov (2013) sont également très prometteurs. Ces approches demandent un coût de calculs raisonnable. Leur assise théorique est solide et leurs résultats sur des modèles simples sont très bons. Elles seront certainement des méthodes d'assimilation de premier plan dans les années à venir. Toutefois, elles ne peuvent pas encore correctement traiter la question de l'erreur modèle. C'est pourquoi nous ne les mettons pas en application dans nos expériences.

2.4 Conclusions

Pour résumer ce tour d'horizon, nous avons vu dans un premier temps ce qu'était une non-Gaussianité (dans sa relation au non-linéaire et au chaos déterministe). Nous avons avancé des moyens de diagnostiquer directement ou indirectement la présence de non-Gaussianité dans un système. Nous avons discuté des sources de non-Gaussianité dans un problème d'assimilation de données en discrétisant les non-Gaussianités des erreurs d'observations et des erreurs de vecteur d'état *a priori*. La perception de ces sources nous a permis de montrer par quelques exemples la présence de non-Gaussianités dans des problèmes de géosciences.

Dans un deuxième temps, en se consacrant à l'aspect méthodologique de l'assimilation de donnée, la deuxième section introduit un cadre mathématique formel au problème d'estimation. Au travers de ce cadre, les méthodes d'assimilation moindres carrés sont réunies sous une même formulation probabiliste. Cette formulation permet de mettre en évidence les hypothèses de linéarité des opérateurs et de Gaussianités des erreurs, que font ces méthodes.

La troisième partie nous plonge dans le sujet des non-Gaussianités. En revenant sur la notion de Gaussianité des densités en jeu, il semble également approprié de réfléchir

aux scores de validation des méthodes. De manière plus générale, on peut se demander ce qu'il est attendu d'une méthode d'assimilation de données. La première sous partie de cette section fournit, après une brève discussion sur la notion de qualités d'une méthode, plusieurs scores. Ces scores seront pour la plupart utilisés dans la suite de cette thèse. Enfin, un état de l'art des réflexions et recherches consacrées aux problèmes d'assimilations dans un contexte non-linéaire et non-Gaussien, est présenté.

De ce chapitre, il ressort que les méthodes moindres carrés sont performantes et présentent de nombreux avantages dans des cas peu non-Gaussiens et de très grandes dimensions. Toutefois, de plus en plus de problèmes d'estimation en géosciences semblent mettre à mal les hypothèses des moindres carrés.

De nombreuses méthodes et techniques ont été créées pour mieux prendre en compte les non-Gaussianités. Ces méthodes présentent également des avantages pour des configurations de problèmes précis. Il est intéressant, et c'est ce qu'il sera fait dans cette thèse (Chap. 6 et 7), de mieux comprendre sur des problèmes précis les avantages de mettre en place une méthode plutôt qu'une autre.

Enfin, pour le traitement des cas fortement non-Gaussiens, les méthodes existantes comme les filtres particulaires sont très performantes sous couvert d'un nombre de particules très grand. En revanche, le coût calcul et la malédiction de la dimensionalité rendent les filtres particulaires inapplicables à des problèmes de grandes dimensions. Les recherches sur des méthodes pouvant gérer de fortes non-Gaussianités en grandes dimensions, sont donc toujours actives. Dans cet esprit, le développement et l'évaluation d'une méthode offrant une nouvelle approche aux problèmes d'assimilation sont présentés au Chapitre 4.

Chapitre 3

Mise en place des méthodes d'assimilation

Sommaire

Dans un premier temps, ce Chapitre 3 décline les principes expérimentaux des études qui constituent cette thèse. Les expériences menées par la suite se font dans le cadre d'expériences dîtes jumelles avec des modèles dits idéalisés. La première section de ce chapitre explique les raisons de ces choix.

Ensuite, les méthodes d'assimilation de données mises en place pour ces expériences sont présentées. Les caractéristiques principales de ces méthodes ainsi que la manière de les implémenter sont explicitées dans la deuxième section de ce chapitre.

Enfin, dans un troisième temps, une illustration de ces méthodes est proposée. Cette illustration a un but pédagogique mais sert aussi de banc d'essai destiné à vérifier la bonne mise en place des méthodes. Ces expériences d'illustration sont menées avec un modèle dynamique de faible dimension, le modèle de Lorenz à trois variables (L63). Plusieurs configurations sont employées. Ces configurations ont été utilisées à plusieurs reprises dans la littérature.

3.1 Principes de l'étude

3.1.1 Le principe des expériences jumelles

Afin d'évaluer une système d'assimilation de données, il est fréquent d'utiliser ce que l'on appelle des expériences jumelles. Il s'agit de définir une trajectoire, obtenue avec le modèle, comme trajectoire de référence (dite aussi trajectoire *vraie* ou *vérité*). Dans la suite, la trajectoire de référence sera appelée la *vérité*. Comme dans un cas réel d'assimilation de données, l'objectif sera alors d'estimer au mieux cette *vérité*. Pour ce faire, sont à disposition des observations créees artificiellement en perturbant la *vérité*. Ainsi qu'une information a priori (ou *background*) créee en perturbant la condition initiale de la *vérité*.

Une expérience jumelle offre un cadre méthodologique structuré et maîtrisable. On peut donc facilement augmenter la densité du réseau d'observations en temps et en espace, modifier l'erreur d'observations ou encore observer différentes variables, afin d'évaluer de manière approfondie les caractéristiques d'un système d'assimilation de données. De plus, les diagnostics utilisés pour évaluer les performances de l'assimilation seront utilisés en comparaison à la *vérité* et non à un jeu d'observations indépendant comme il est fait en expériences réelles.

Ce dernier argument est un des avantages majeurs des expériences jumelles. En effet, comme il n'est pas statistiquement correct de comparer une assimilation de données avec les mêmes observations qu'utilisées par cette dernière (particulièrement en océanographie biogéochimique ; Gregg et al. (2009)), il est donc indispensable, dans un cas réel, d'avoir à disposition un deuxième jeu de données indépendant, mais tout de même pertinent, pour effectuer les diagnostics. Un tel jeu n'est pas toujours disponible, spécialement en océanographie où les données sont souvent peu nombreuses.

3.1.2 Le principe d'un modèle jouet

Une expérience jumelle peut être générée à l'aide d'un modèle réaliste, comme il est fait au Chapitre 6 et au Chapitre 7. Une expérience utilisant des modèles réalistes peut en revanche s'avérer difficile dans sa mise en place, dans sa réalisation et dans la lecture de ses résultats.

Ainsi, pour ce travail préliminaire nous choisissons des modèles à la dynamique simple et au comportement connu. Ce type de modèles est communément appelé modèles jouets. L'utilisation de modèles jouets présente de nombreux avantages.

Nous utilisons dans ce chapitre le modèle de Lorenz à trois variables, dit Lorenz 63 (L63).Ce modèle est décrit dans la Section 3.3.

3.2 Les méthodes étudiées

L'un des objectifs de cette étude est de présenter de manière algorithmique les différentes méthodes d'assimilation qui seront étudiées par la suite. Les méthodes utilisées sont toutes écrites en langage *python*. Dans un deuxième temps, une illustration sur un modèle jouet : le Lorenz 63 à trois variables, permet de vérifier le comportement des méthodes.

3.2.1 Les méthodes aux hypothèses Gaussiennes

Nous mettons en place deux versions de filtres de Kalman d'ensemble. Les analyses produites par ces deux versions sont théoriquement équivalentes. La différence importante entre les deux versions est l'espace de résolution du problème. Nous les présentons ci-dessous.

Filtre de Kalman d'ensemble stochastique (EnKFs)

L'EnKF stochastique est un filtre de Kalman d'ensemble avec perturbation des observations (Evensen, 1994; Burgers et al., 1998; Evensen, 2003). Il s'agit d'un filtre à hypothèse Gaussienne c'est à dire qu'il ne traite que les moments des statistiques d'ordre un et deux. Contrairement au filtre de Kalman (KF) l'information statistique n'est pas transmise par une trajectoire et une matrice de covariances d'erreurs (propagée avec le modèle linéarisé) mais par un ensemble (propagé avec le modèle non-linéaire). À l'aide de l'ensemble on peut approximer la moyenne et la matrice des covariances d'erreurs (à peu de frais). Après avoir recouvré ces deux moments statistiques, l'analyse qui suit est identique à celle du filtre de Kalman présentée en Section 2.2.2. Il est à noter que l'EnKFs résout le problème d'estimation directement dans l'espace des solutions (l'espace d'état).

Pour nos expériences, cette méthode est implémentée de manière scalaire, i.e. les observations sont traitées en série, d'après le formalisme de Anderson (2003). Le procédé algorithmique a déjà été décrit en Section 2.2.2 et illustré en Figure 2.3.

Algorithme - EnKFs

Ensemble *a priori* au temps 0 : $[\mathbf{x}_0^{f,1}, ..., \mathbf{x}_0^{f,N_e}]$;
Observations au temps k : $\mathbf{y}_k$ de covariances R_k

Prévision
– Propagation des N_e membres de l'ensemble du temps $k-1$ au temps k :

$$\mathbf{x}_k^{f,i} = \mathcal{M}(\mathbf{x}_{k-1}^{f,i}), \quad i = 1, ..., N_e$$

Analyse
Pour chaque observation (scalaire) $y_k \in \mathbf{y}_k$ avec $y_k = h(x_{k,o})$ où $\mathbf{x}_k = [x_{k,o}, \mathbf{x}_{k,u}]$, $x_{k,o}$ est la variable observée et $\mathbf{x}_{k,u}$ est le vecteur des variables non-observées.
Variable observée
– Perturbation de l'observation :

$$y_k^i = y_k + \sigma_o g, \text{ avec } \sigma_o^2 \text{ la variance d'erreurs d'observation et } g \sim \mathcal{N}(0,1)$$

– Calcul de la variance d'erreurs *a priori* $\sigma^2_{x_{k,o}} = Var(x^f_{k,o})$
– Correction sur la variable observée

$$x_{k,o}^{a,i} = x_{k,o}^{f,i} + \frac{\sigma_o^2}{\sigma_o^2 + \sigma^2_{x_{k,o}}}(y_k^i - x_{k,o}^{f,i}), \quad i = 1, ..., N_e$$

Variables non-observées
– Correction sur les variables non-observées

$$\mathbf{x}_{k,u}^{a,i} = \mathbf{x}_{k,u}^{f,i} + \frac{cov(\mathbf{x}_{k,u}^{f,i}, x_{k,o}^{f,i})}{\sigma_o^2}(x_{k,o}^{a,i} - x_{k,o}^{f,i}), \quad i = 1, ..., N_e$$

Filtre de Kalman d'ensemble transformé (ETKF)

Le filtre de Kalman d'ensemble transformé est un filtre de Kalman d'ensemble résolu dans l'espace réduit de l'ensemble, à l'aide d'une transformation (Bishop et al., 2000).

L'étape de prévision est similaire à celle de l'EnKFs. L'information est transmise via la propagation de l'ensemble par le modèle non-linéaire. En revanche, contrairement à l'EnKFs, on effectue un changement de variables pour travailler non pas avec un ensemble d'états $X_k^f = [\mathbf{x}_k^{f,1}, ..., \mathbf{x}_k^{f,N_e}]$ mais avec un ensemble d'anomalies $Z_k^f = \frac{1}{\sqrt{N_e - 1}}(X_k^f - \overline{X}_k^f)$ avec $\overline{X}_k^f = [\mathbf{x}_k^f, ..., \mathbf{x}_k^f]$ la moyenne d'ensemble. La correction de ces anomalies se fait par le

calcul de la matrice $\mathbf{A} = (HZ_k^f)^T R_k^{-1} HZ_k^f$. On en déduit la transformation $Z_k^a = Z_k^f C(Id + \Gamma)^{-1/2}$ avec C la matrice des vecteurs propres et Γ la matrice diagonale des valeurs propres de $\mathbf{A}$. Une fois les anomalies corrigées, on corrige l'ensemble d'états en se basant toujours sur la théorie de Kalman et après un peu d'algèbre :

$$X_k^a = \overline{X}_k^f + Z_k^f C(Id + \Gamma)^{-1} C^T (HZ_k^f)^T R_k^{-1} (\mathbf{y}_k - \mathbf{x}_k^f) + Z_k^a \sqrt{N_e - 1}$$

Cette méthode peut être vue comme une pondération corrigée des anomalies par application de la théorie de Kalman. Il existe plusieurs façons d'implémenter cette méthode nous avons choisi la suivante.

Algorithme - ETKF

Ensemble *a priori* au temps 0 : $[\mathbf{x}_0^{f,1}, ..., \mathbf{x}_0^{f,N_e}]$;
Observations au temps k : $\mathbf{y}_k$ de covariances R_k

Prévision

– Propagation des N_e membres de l'ensemble du temps $k-1$ au temps k :

$$\mathbf{x}_k^{f,i} = \mathcal{M}(\mathbf{x}_{k-1}^{f,i}), \quad i = 1, ..., N_e$$

– Calcul de l'état moyen *a priori* :

$$\mathbf{x}_k^f = \frac{1}{N_e} \sum_{i=1}^{N_e} \mathbf{x}_k^{f,i};$$

Analyse

– Calcul des perturbations : on note $X_k^f = [\mathbf{x}_k^{f,1}, ..., \mathbf{x}_k^{f,N_e}]$ et $\overline{X}_k^f = [\mathbf{x}_k^f, ..., \mathbf{x}_k^f]$

$$Z_k^f = \frac{1}{\sqrt{N_e - 1}} (X_k^f - \overline{X}_k^f)$$

– Calcul de la matrice :

$$\mathbf{A} = (HZ_k^f)^T R_k^{-1} HZ_k^f$$

C : matrice des vecteurs propres et Γ : matrice diagonale des valeurs propres de $\mathbf{A}$

– Transformation des perturbations :

$$Z_k^a = Z_k^f C(Id + \Gamma)^{-1/2}$$

– Correction de la moyenne :

$$\overline{X}_k^a = \overline{X}_k^f + Z_k^f C(Id + \Gamma)^{-1} C^T (HZ_k^f)^T R_k^{-1} (\mathbf{y}_k - H\mathbf{x}_k^f)$$

– Correction de l'ensemble :

$$X_k^a = \overline{X}_k^a + Z_k^a \sqrt{N_e - 1}$$

3.2.2 Des méthodes non-Gaussiennes à partir d'un cadre classique

Filtre de Kalman d'ensemble avec anamorphose dynamique (EnKF-Anam)

Le filtre de Kalman d'ensemble suppose la gaussianité des densités de probabilité, notamment de la densité a priori. Cette hypothèse est la plupart du temps erronée.

Une façon de gérer ces non-Gaussianités est d'appliquer une transformation, scalaire ou vectorielle, analytique ou numérique, afin de les approcher au mieux de densités Gaussiennes. Ces transformations s'appellent des anamorphoses. Après transformation, il est possible d'effectuer une analyse aux hypothèses Gaussiennes (de type BLUE) sur ces nouvelles variables quasi-Gaussiennes. Plusieurs transformations existent pour effectuer une anamorphose (Simon and Bertino, 2012).

Les anamorphoses analytiques utilisent la connaissance *a priori* que l'on peut avoir d'une variable physique. Par exemple, si l'on sait qu'une variable X est de loi log-normale (e.g. la chlorophylle), on peut lui appliquer une transformation lognormale inverse. Il s'agit donc d'un changement de variable qui produit une nouvelle variable Gaussienne $\tilde{X}$ et telle que $X = ln(\tilde{X})$. Après analyse de la variable $\tilde{X}$, on effectue la transformation inverse pour obtenir la correction de la variable X.

Lorsqu'on ne dispose pas de connaissance *a priori* de la nature des variables que l'on traite, il est possible d'effectuer une anamorphose numérique (Wackernagel (2006) en géostatistique et Bertino et al. (2003) en océanographie). Il se base sur les fonctions de répartition (cdf) des variables aléatoires. On admet que l'on possède la cdf F de notre variable non-Gaussienne X. On connaît également la cdf G de la variable aléatoire Gaussienne $\tilde{X}$ telle que $\tilde{X} \sim \mathcal{N}(0,1)$. L'anamorphose est alors définie comme la fonction de transport (*map*) $\psi = F^{-1} \circ G$ qui à une valeur $\tilde{X}$ fait correspondre une valeur X.

En possédant un ensemble qui décrit la pdf de X on peut effectuer une démarche similaire appelée l'anamorphose dynamique. La fonction de transport est une fonction définie par morceau qui fait correspondre les percentiles de la variable $\tilde{X}$ à ceux de la variable X. Elle s'écrit

$$\psi_{N_e}(x) = \begin{cases} \tilde{\xi}_1, & \text{si } x < \xi_1 \\ \tilde{\xi}_k + \frac{\tilde{\xi}_{k+1} - \tilde{\xi}_k}{\xi_{k+1} - \xi_k}(x - \xi_k), & \text{si } x \in [\xi_k, \xi_{k+1}] \\ \tilde{\xi}_{N_e}, & \text{si } x > \xi_{N_e} \end{cases} \tag{3.1}$$

avec $\xi_1, ..., \xi_{N_e}$ et $\tilde{\xi}_i, ..., \tilde{\xi}_{N_e}$ les percentiles de X et de $\tilde{X}$ respectivement. Cette anamorphose dynamique a été mise en place et a montré ses bonnes performances en océanographie biogéochimique (Béal et al., 2009, 2010). Nous utilisons cette version de l'anamorphose dans la suite.

Algorithme - EnKF-Anam

Ensemble *a priori* au temps 0 : $[\mathbf{x}_0^{f,1}, ..., \mathbf{x}_0^{f,N_e}]$;

Observations au temps k : $\mathbf{y}_k$ de covariances R_k

Prévision

– Propagation des N_e membres de l'ensemble du temps $k-1$ au temps k :

$$\mathbf{x}_k^{f,i} = \mathcal{M}(\mathbf{x}_{k-1}^{f,i}), \quad i = 1, ..., N_e$$

Analyse

Pour chaque observation (scalaire) $y_k \in \mathbf{y}_k$ avec $y_k = h(x_{k,o})$ où $\mathbf{x}_k = [x_{k,o}, \mathbf{x}_{k,u}]$

– Perturbation de l'observation :

$$y_k^i = y_k + \sigma_o g, \text{ avec } \sigma_o^2 \text{ la variance d'erreurs d'observation et } g \sim \mathcal{N}(0,1)$$

– Anamorphose :
  – Calcul des percentiles des $x_{k,o}^{f,i}$ et de la Gaussienne $\mathcal{N}(x_{k,o}^f, \sigma_{x_{k,o}^f})$
  – Changement de variables : $[\tilde{\mathbf{x}}_k^{f,i}, \tilde{y}_k^i] = \psi_{N_e}([\mathbf{x}_k^{f,i}, y_k^i])$, pour tout $i = 1, ..., N_e$
– Analyse : *idem EnKF*
  – Calcul de la variance d'erreurs *a priori* $\sigma^2_{\tilde{x}_{k,o}} = Var(\tilde{x}_{k,o}^{f,i})$
  – Calcul de la variance d'erreurs d'observation $\sigma^2_{\tilde{y}_k} = Var(\tilde{y}_k^i)$
  – Correction sur la variable observée

$$\tilde{x}_{k,o}^{a,i} = \tilde{x}_{k,o}^{f,i} + \frac{\sigma_o^2}{\sigma^2_{\tilde{y}_k} + \sigma^2_{\tilde{x}_{k,o}}}(\tilde{y}_k^i - \tilde{x}_{k,o}^{f,i}), \quad i = 1, ..., N_e$$

  – Correction sur les variables non-observées

$$\tilde{\mathbf{x}}_{k,u}^{a,i} = \tilde{\mathbf{x}}_{k,u}^{f,i} + \frac{cov(\tilde{\mathbf{x}}_{k,u}^{f,i}, \tilde{x}_{k,o}^{f,i})}{\sigma^2_{\tilde{y}_k}}(\tilde{x}_{k,o}^{a,i} - \tilde{x}_{k,o}^{f,i}), \quad i = 1, ..., N_e$$

– Anamorphose inverse : $\mathbf{x}_k^{a,i} = \psi_{N_e}^{-1}(\tilde{\mathbf{x}}_k^{a,i})$, pour tout $i = 1, ..., N_e$

Il est également important de noter que la version de l'anamorphose que nous utilisons ne simule pas de queues de densités. En revanche, une borne (inférieure) est appliquée à l'erreur d'observation dans l'espace anamorphosé afin d'éviter des corrections trop drastiques et un éventuel effondrement d'ensemble. Nous sommes conscients de l'importance que revêt les queues de densités pour les performances de l'assimilation (importance mise en évidence par Simon and Bertino, 2012) mais nous n'avons pas pu, pour des raisons de temps, les mettre en place dans notre version.

Filtre d'histogrammes de rangs (RHF)

Le RHF est une méthode traitant chaque variable en série, développée par Anderson (2010). Cette méthode est basée sur la construction de pdf à partir d'un ensemble en utilisant des histogrammes de rangs ce qui lui permet de ne pas supposer la gaussianité des densités *a priori* pour les variables observées. Les variables non-observées sont corrigées par régression linéaire.

Dans un premier temps, nous traitons la variable scalaire observée $x_{k,o}$. Dans un second temps, nous corrigeons le vecteur du reste des variables non-observées $\mathbf{x}_{k,u}$.

La pdf $P_{x_{k,o}}(x_o)$ de la variable aléatoire $x_{k,o}$ est approximée à l'aide de l'ensemble $(x_{k,o}^{f,i})_i$ en utilisant une méthode d'histogrammes de rangs. Cette méthode consiste à considérer les intervalles entre les membres (préalablement ordonnés) deux à deux. Dans ces intervalles, la pdf est constante et est d'intégrale $\frac{1}{N_e+1}$. Ainsi la pdf *a priori* s'écrit :

$$P_{x_{k,o}}(x_o) = \frac{1}{N_e+1} \sum_{j=1}^{N_e-1} \frac{1_{[x_{k,o}^{f,j}, x_{k,o}^{f,j+1}[}(x_o)}{(x_{k,o}^{f,j+1} - x_{k,o}^{f,j})} + T(x_o), \tag{3.2}$$

avec $T(x_o) = T_1(x_o)1_{]-\infty, \min_j(x_{k,o}^{f,j})[}(x_o) + T_2(x_o)1_{]\max_j(x_{k,o}^{f,j}), +\infty[}(x_o)$ deux queues (de poids $\frac{1}{N_e+1}$ également) de densités à préscrire. Sur ces mêmes intervalles est discrétisée la vraisemblance $P_{y_k|x_{k,o}}$. Ceci nous permet de faire un produit point par point (produit d'Hadamard) entre ces deux pdf qui nous donne après re-normalisation la pdf *a posteriori* $P_{x_{k,o}|y_k}$. Par échantillonnage de la pdf *a posteriori* $P_{x_{k,o}|y_k}$ nous obtenons un ensemble *a posteriori* $(x_{k,o}^{a,i})_i$.

Cet ensemble corrigé sur la variable observée $x_{k,o}$ se propage aux variables non-observées $\mathbf{x}_{k,u}$ par régression linéaire. Cette régression est similaire à celle de l'EnKFs. Elle utilise les covariances d'erreurs comme pondération. On obtient ainsi :

$$\mathbf{x}_{k,u}^{a,i} = \mathbf{x}_{k,u}^{f,i} + \frac{cov(\mathbf{x}_{k,u}^{f,i}, x_{k,o}^{f,i})}{\sigma_o^2}(x_{k,o}^{a,i} - x_{k,o}^{f,i}), \quad i = 1, ..., N_e$$

Algorithme - RHF

Ensemble *a priori* au temps 0 : $[\mathbf{x}_0^{f,1}, ..., \mathbf{x}_0^{f,N_e}]$;
Observations au temps k : $\mathbf{y}_k$ de covariances R_k

Prévision
– Propagation des N_e membres de l'ensemble du temps $k-1$ au temps k :

$$\mathbf{x}_k^{f,i} = \mathcal{M}(\mathbf{x}_{k-1}^{f,i}), \quad i = 1, ..., N_e$$

Analyse
Pour chaque observation (scalaire) $y_k \in \mathbf{y}_k$ avec $y_k = x_{k,o}$ où $\mathbf{x}_k = [x_{k,o}, \mathbf{x}_{k,u}]$
Variable observée
– Tri par ordre croissant des membres : $(x_{k,o}^{f,j})_j$ avec $x_{k,o}^{f,j} < x_{k,o}^{f,j+1}, \forall j = 1, ..., N_e$
– Création par histogramme de rangs de la pdf approchée

$$P_{x_{k,o}}(x_o) = \frac{1}{N_e + 1} \sum_{j=1}^{N_e - 1} \frac{1_{[x_{k,o}^{f,j}, x_{k,o}^{f,j+1}[}(x_o)}{(x_{k,o}^{f,j+1} - x_{k,o}^{f,j})} + T(x_o) \tag{3.3}$$

– Création sur les mêmes intervalles de la vraisemblance $P_{y_k|x_{k,o}^f}$
– Calcul de la pdf produit (et normalisation) donnant $P_{x_{k,o}|y_k}$
– Échantillonnage de $P_{x_{k,o}|y_k}$ donnant l'ensemble $(x_{k,o}^{a,i})_{i=1,...,N_e}$
Variables non-observées
– Correction sur les variables non-observées

$$\mathbf{x}_{k,u}^{a,i} = \mathbf{x}_{k,u}^{f,i} + \frac{cov(\mathbf{x}_{k,u}^{f,i}, x_{k,o}^{f,i})}{\sigma_o^2}(x_{k,o}^{a,i} - x_{k,o}^{f,i}), \quad i = 1, ..., N_e$$

3.2.3 Les méthodes non-Gaussiennes

Filtre particulaire (PF)

Les filtres particulaires font l'objet de nombreux développements notamment concernant leurs adaptations aux problèmes géophysiques de grandes dimensions (Nakano et al., 2007; van Leeuwen, 2009, 2010; Morzfeld et al., 2012). La difficulté principale dans la mise en place des filtres particulaires est de pouvoir représenter des densités de probabilités en très grandes dimensions avec des ensembles de petites tailles. Ce problème s'appelle la malédiction de la dimensionalité (Snyder et al., 2008). De nouvelles formulations de filtres particulaires faisant intervenir des *densités de transition* semblent toutefois prometteuses pour des applications océanographiques en grandes dimensions (van Leeuwen, 2003, 2009, 2010).

Par la suite, nous faisons le choix de mettre en place un filtre particulaire plus simple dit *bootstrap*. Il s'agit d'une version séquentielle d'un ré-échantillonnage d'importance (SIR) proposée par Gordon et al. (1993).

Sans faire d'hypothèse de Gaussianité (i.e. sans considérer uniquement les deux premiers moments des statistiques), le filtre *bootstrap* propage l'ensemble avec le modèle non-linéaire. Il considère alors l'ensemble propagé comme un peigne de Dirac échantillonnant la pdf *a priori*. Ces diracs sont pondérés par des poids $w_k^{f,i}$, *a priori* homogènes i.e. $w_k^{f,i} = \frac{1}{N_e}$ pour

tout i. L'étape d'analyse consiste à mettre à jour les poids en appliquant le théorème de Bayes sur ces pdf approximées :

$$P_{\mathbf{x}_k|y_k} \equiv \sum_{i=1}^{N_e} w_k^{a,i} \delta(\mathbf{x}_k - \mathbf{x}_k^{f,i})$$

avec $w_k^{a,i} \propto w_k^{f,i} P_{\mathbf{y}_k|\mathbf{x}_k^{f,i}}$ les poids mis à jour.

Cependant, il a été constaté qu'après plusieurs cycles d'assimilation un poids écrase les autres et tend vers 1. Cette dégénérescence a pour effet un effondrement de l'ensemble vers un seul état ce qui ne contient plus d'information statistique.

Pour minimiser cette dégénérescence il est commun d'effectuer un ré-échantillonnage des membres (ou particules) de l'ensemble. Un re-échantillonnage se fait en séléctionnant un sous-échantillon de taille N_{SE} des membres de l'ensemble aux poids homogènes et en démultipliant ces membres en fonction de leurs poids. Tous les poids sont alors égalisés : $w_k^{a,i} = \frac{1}{N_e}$.

Pourvu d'un ensemble de taille suffisante, le filtre *bootstrap* est adapté aux fortes non-gaussianités.

Algorithme - PF (bootstrap)

Ensemble *a priori* au temps 0 : $[\mathbf{x}_0^{f,1}, ..., \mathbf{x}_0^{f,N_e}]$;
Observations au temps k : $\mathbf{y}_k$ de covariances R_k

Prévision

– Propagation des N_e particules (membres) de l'ensemble du temps $k-1$ au temps k :

$$\mathbf{x}_k^{f,i} = \mathcal{M}(\mathbf{x}_{k-1}^{f,i}), \quad i = 1, ..., N_e$$

– Approximation de la pdf *a priori* :

$$P_{\mathbf{x}_k} \equiv \sum_{i=1}^{N_e} w_k^{f,i} \delta(\mathbf{x}_k - \mathbf{x}_k^{f,i})$$

– Les particules sont munies de poids uniformes $w_k^{f,i} = \frac{1}{N_e}$

Analyse

– Calculs des nouveaux poids en fonction des vraisemblances $P_{\mathbf{y}_k|\mathbf{x}_k^{f,i}}$

$$w_k^{a,i} \propto w_k^{f,i} P_{\mathbf{y}_k|\mathbf{x}_k^{f,i}}$$

– Approximation de la pdf *a posteriori* :

$$P_{\mathbf{x}_k|y_k}(\mathbf{x}_k) = \sum_{i=1}^{N_e} w_k^{a,i} \delta(\mathbf{x}_k - \mathbf{x}_k^{f,i})$$

– Ré-échantillonnage d'un sous-ensemble en fonction d'un seuil S :

$$(\mathbf{x}_k^{f,j})_{j(i)=1,\ldots,N_{SE}} \text{ avec } SE < NE \text{ et tels que } \forall j,\ \mathbf{x}_k^{f,j} > S$$

– Séléction d'un ensemble à N_e membres et ajoût de bruits

Filtre d'histogrammes de rangs multivarié (MRHF)

Dans nos expériences, nous utilisons aussi le filtre multivarié d'histogrammes de rangs (MRHF). Le MRHF est une extension du RHF aux variables non-observées. Cette méthode est adaptée aux fortes non-gaussianités (Metref et al., 2014). Elle a été développée au cours de cette thèse. Nous ne décrivons pas ici l'algorithme du MRHF et nous n'illustrons pas son fonctionnement car le chapitre suivant (Chapitre 4) est dédié à cette méthode.

3.3 Illustration des méthodes

3.3.1 Description du modèle et des configurations

Le modèle de Lorenz à trois variables, Lorenz 63

Le modèle de Lorenz à trois variables (Lorenz 63), est un système aux équations différentielles ordinaires décrit et étudié par Lorenz (1963). Ce modèle mathématique simplifié, simule une cellule de convection atmosphérique. Il est fréquemment utilisé comme cas test dans l'étude des systèmes dynamiques, pour ses propriétés (avec certains jeux de paramètres) fortement chaotiques (Figure 3.1). Il propose ainsi des problèmes d'estimation difficiles. L'utilisation de ce modèle dans le développement de méthodes d'assimilation de données est donc largement répandue.

Nous notons les trois variables du modèle : x, y, z. Les équations décrivant la dynamique sont les suivantes.

$$\frac{dx}{dt} = \sigma(y - x), \tag{3.4}$$

$$\frac{dy}{dt} = x(\rho - z) - y, \tag{3.5}$$

$$\frac{dz}{dt} = xy - \beta z. \tag{3.6}$$

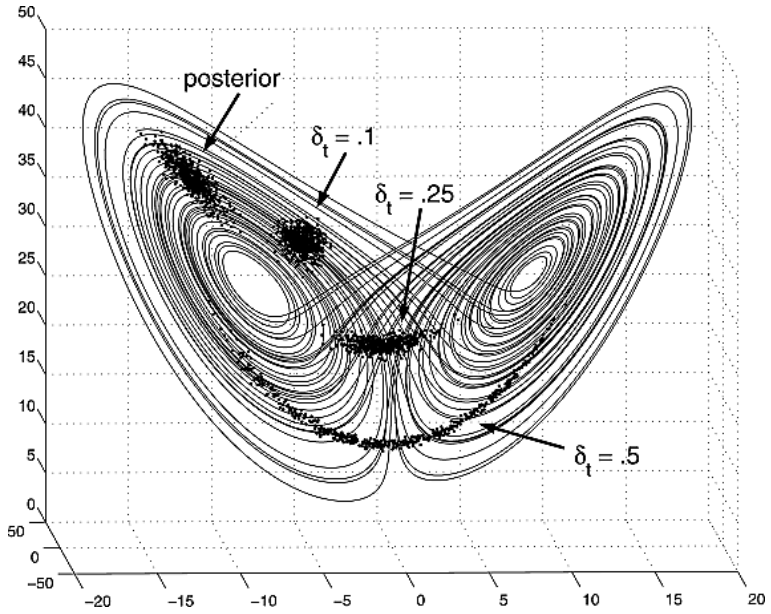

Figure *3.1* – *Figure issue de Bengtsson et al. (2003) – Illustration en trois dimensions de l'attracteur chaotique du modèle de Lorenz 63 : 400 particules ("posterior") échantillonnées d'une distribution Gaussienne et respectivement propagées à 0.1, 0.25, et 0.5 unité de temps.*

Dans ce travail, nous nous plaçons dans une configuration communément utilisée ayant pour paramètres : $\sigma = 10$, $\rho = 28$ et $\beta = 8/3$. Le pas de temps d'intégration du modèle est $\delta t = 0.01$. La condition initiale utilisée est $x = 1.508870$, $y = -1.531271$ et $z = 25.46091$ (van Leeuwen, 2010).

Les configurations des expériences jumelles

À partir du modèle Lorenz 63, une simulation de référence est réalisée et est considérée, dans les expériences qui suivent, comme la *vérité* à recouvrer. Un ensemble de taille N_e (variant selon les expériences) est crée en perturbant la condition initiale par un bruit multiplicatif Gaussien de loi $\mathcal{N}(1, 0.3)$. La Figure 3.2 représente les séries temporelles de l'ensemble libre de 100 membres (en cyan) et de la trajectoire de référence (la *vérité*, en vert) du modèle de Lorenz 63, sur 5000 pas de temps (et après 5000 pas de temps de spin up) pour les trois variables. Toutes les expériences suivantes sont réalisées sur 5000 pas de temps. De plus, une période antérieure de 5000 pas de temps est laissée afin d'éviter tout problème de stabilisation initiale (*spin up*) de l'assimilation.

Dans cette sous partie, le vecteur d'état ($[x, y, z]$) est entiérement observé. Les observations sont créees en ajoutant à la trajectoire *vérité* des perturbations indépendantes tirées

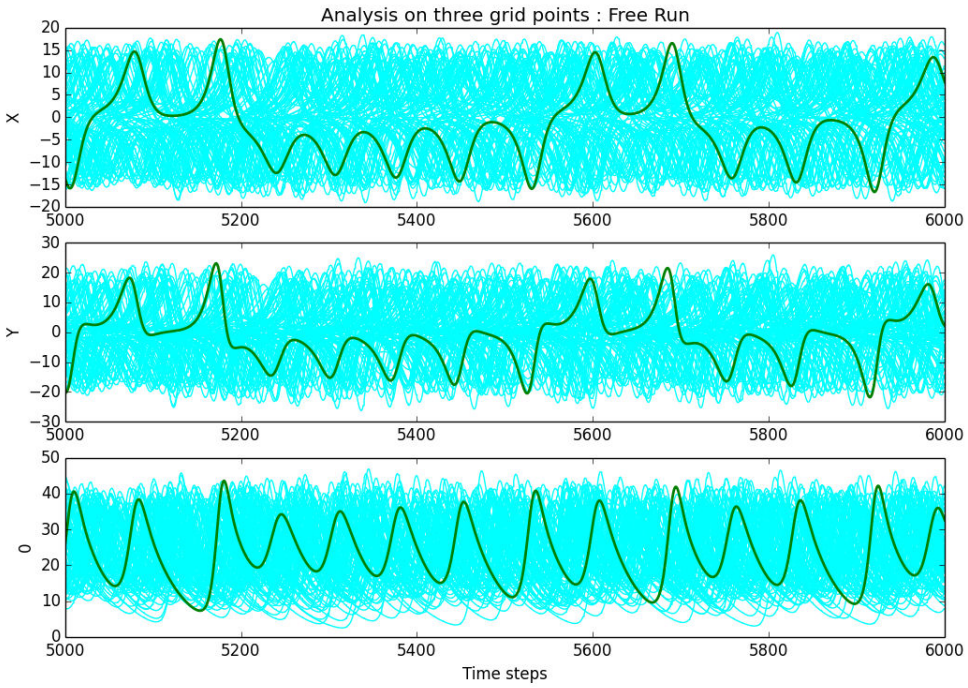

Figure ***3.2*** *– Séries temporelles de l'ensemble libre de 100 membres (en cyan) et de la trajectoire de référence (la vérité, en vert) du modèle de Lorenz 63, sur 5000 pas de temps (et après 5000 pas de temps de spin up) pour les variables x (haut), y (centre) et z (bas).*

selon un bruit blanc Gaussien de déviation standard $\sigma_o = 2$ comme dans les expériences proposées par Harlim and Hunt (2007) et Bocquet (2011). Ces trois cas d'expériences sont menées avec différentes fréquences temporelles d'observations : $\Delta t = 0.10$, $\Delta t = 0.25$ et $\Delta t = 0.50$. C'est à dire, une observation disponible tous les 10, les 25 et les 50 pas de temps respectivement. Ces trois réseaux d'observations offrent des cas tests aux faibles, moyennes et fortes non-linéarités respectivement (Fig. 3.1, issue de Bocquet (2011)).

3.3.2 Illustration

Dans cette illustration, nous regardons les méthodes présentées en Section 3.2. Chaque méthode est testée avec une taille d'ensemble pour laquelle elle est la plus performante. L'objectif n'est donc pas de comparer les méthodes entre elles mais de voir si les méthodes fonctionnent bien. On perçoit également une première idée des comportements des méthodes notamment dans l'augmentation des non-linéarités d'une configuration à l'autre.

Les filtres de Kalman d'ensemble

Les filtres de Kalman d'ensemble sont connus pour leurs bonnes performances dans un cadre quasi-linéaire et peu non-Gaussien. Ces performances sont confirmées sur cette illustration.

Le filtre de Kalman d'ensemble stochastique (EnKFs) est présenté en Figure 3.3 (graphiques de gauche, courbes rouges). Il est évalué avec un ensemble de 30 membres. Le filtre de Kalman transformé (ETKF) est présenté en Figure 3.3 (graphiques de droite, courbes bleues). Il est évalué avec un ensemble de 10 membres pour les cas $\Delta t = 0.10$ et 0.25, et avec 30 membres pour le cas $\Delta t = 0.50$. La *vérité* est représentée par la courbe verte.

Il est à noter que l'ETKF résoud le problème d'estimation dans l'espace de l'ensemble. Or comme le degré de liberté associé à ce problème est petit (3 degrés de liberté), la taille de l'ensemble n'a pas besoin d'être grande. Nous avons même constaté que l'utilisation de l'ETKF avec un ensemble de grande taille dégrade la solution (pas montré ici). Dans ces conditions, les résultats produits par ces deux filtres sont très proches. Ceci etait attendu puisque comme nous l'avons vu ces filtres sont deux versions d'un même filtre avec les mêmes hypothèses.

Dans le cas à faible non-linéarité ($\Delta t = 0.10$, graphiques supérieurs), l'EnKFs et l'ETKF produisent des ensembles dont la moyenne est très proche de la *vérité* (courbe verte). Même si nous ne regardons pas, ici, de scores de dispersion ou de résolution, l'enveloppe des ensembles semble bien capturer la *vérité*.

Plus la fréquence d'observations diminue plus les solutions se dégradent. Les ensembles produits par l'EnKFs et l'ETKF dans le cas à moyenne non-linéarité ($\Delta t = 0.25$, graphiques centraux) sont plus dispersés que dans le cas précédent. Cependant, l'enveloppe des ensembles continue d'inclure la *vérité*.

En diminuant encore la fréquence d'observations, i.e. dans le cas fortement non-linéaire ($\Delta t = 0.50$, graphiques inférieurs), les ensembles divergent par moment de la *vérité*. À partir du pas de temps 5700 et jusqu'à la fin de la fenêtre, les deux ensembles décrochent progressivement.

Le RHF et l'EnKF anamorphosé

Le filtre d'histogrammes de rangs (RHF) et le filtre de Kalman d'ensemble anamorphosé (EnKF anamorphosé) sont des filtres basés sur le cadre aux hypothèses linéaires et Gaussiennes de Kalman en prenant en compte de diverses manières les non-Gaussianités (voir Sec. 3.2). À cet égard, ils produisent des résultats proches ou moins bons que les filtres de Kalman dans un cadre quasi-linéaire et peu non-Gaussien. De plus, ces filtres nécessitent des ensembles de plus grandes tailles. Par contre, de meilleures performances sont à attendre dans un cadre plus non-linéaire. Cette illustration nous confirme ces attentes pour le RHF et nous alerte sur les limitations de l'anamorphose.

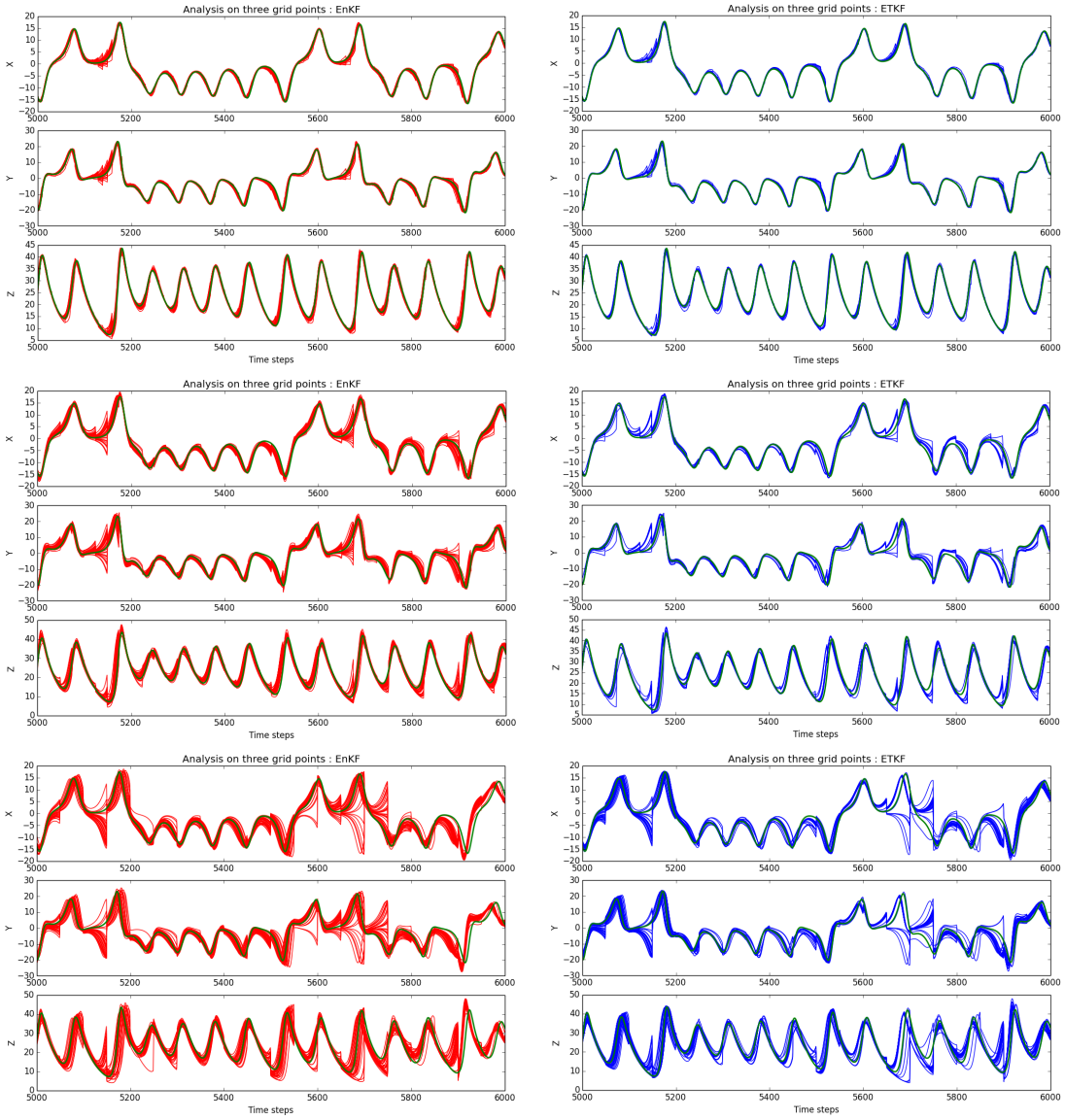

Figure ***3.3*** – *Séries temporelles des ensembles produits par l'EnKFs (graphiques de gauche, courbes rouges) et l'ETKF (graphiques de droite, courbes bleues) des pas de temps 5000 à 6000, dans les configurations* $\Delta t = 0.10$ *(graphiques supérieurs),* $\Delta t = 0.25$ *(graphiques centraux) et* $\Delta t = 0.50$ *(graphiques inférieurs). La vérité est représentée par la courbe verte.*

Le RHF est présenté en Figure 3.4 (graphiques de gauche, courbes magentas). Il est évalué avec un ensemble de 50 membres. L'EnKF anamorphosé est présenté en Figure 3.4

(graphiques de droite, courbes grises). Il est évalué avec un ensemble de 50 membres. La *vérité* est représentée par la courbe verte.

Il est apparu que l'anamorphose de l'observation, dans notre implémentation, ne fonctionne pas lorsque l'ensemble *a priori* et l'observation sont trop éloignés (problème de biais). Ce problème se produit dans les trois configurations. Il a donc fallu introduire une inflation de l'ensemble dans la première configuration ($\Delta t = 0.10$) d'un coefficient (optimisé) 1.04. Dans les trois configurations, nous avons introduit une inflation de la matrice de covariances d'erreurs d'observations $R_o = diag(4, 4, 4)$. Les coefficients (optimisés) de cette dernière inflation sont 3, 2.8 et 1.79 pour, respectivement, les configurations $\Delta t = 0.10$, 0.25 et 0.50. Nous présentons les résultats obtenus pour l'EnKF anamorphosé avec inflation malgré le fait qu'ils ne soient pas représentatifs des résultats potentiels de cette méthode dans des cas sans ces forts biais. L'apparition de cette difficulté pour l'EnKF anamorphosé, qui a par ailleurs montré ses bonnes performances (Bertino et al., 2003; Béal et al., 2010), confirme l'idée qu'aucune méthode n'est idéale pour tous les problèmes d'estimation.

Dans le cas faiblement et moyennement non-linéaire ($\Delta t = 0.10$ et $\Delta t = 0.25$), le RHF se comporte de manière identique aux filtres de Kalman d'ensemble avec un estimé moyen très proche de la *vérité* et un ensemble très peu dispersé. Le RHF semble produire un ensemble mieux dispersé que les filtres de Kalman d'ensemble, dans le cas fortement non-linéaire ($\Delta t = 0.50$). En particulier, autour du pas de temps 5700, les membres de l'ensemble produits par le RHF ne divergent pas hors du voisinage de la *vérité*.

Les résultats produits par l'EnKF anamorphosé sont plus difficiles à lire. En particulier, dans le cas très contrôlé de la configuration faiblement non-linéaire, l'inflation d'ensemble disperse énormément l'ensemble. L'ensemble est, tout de même, corrigé vers la *vérité* tous les 10 pas de temps. Dans les configurations moyennement et fortement non-linéaire, les ensembles produits par l'EnKF anamorphosé sont mieux dispersés et estiment globalement bien la *vérité*. Dans la configuration moyennement non-linéaire, avant le pas de temps 5200 et après le pas de temps 5700, l'EnKF anamorphosé ne contrôle pas bien la dispersion de l'ensemble.

Le PF-*bootstrap*

Le filtre particulaire *bootstrap* (PF-*bootstrap*) est conçu pour pouvoir gérer de fortes non-linéarités et de fortes non-Gaussianités. Comme il a été mentionné précédemment, le PF-*bootstrap* est victime de la malédiction de la dimensionnalité. Dans une application comme le Lorenz 63, la dimension est faible. Pourtant les bonnes performances du PF-*bootstrap* nécessitent un nombre de membres bien plus grand que les filtres de Kalman d'ensemble.

Le PF-*bootstrap* est présenté en Figure 3.5 (courbes noires). Il est évalué avec un ensemble de 100 membres. La *vérité* est représentée par la courbe verte.

Le PF-*bootstrap* produit un ensemble plus dispersé que les filtres de Kalman d'ensemble,

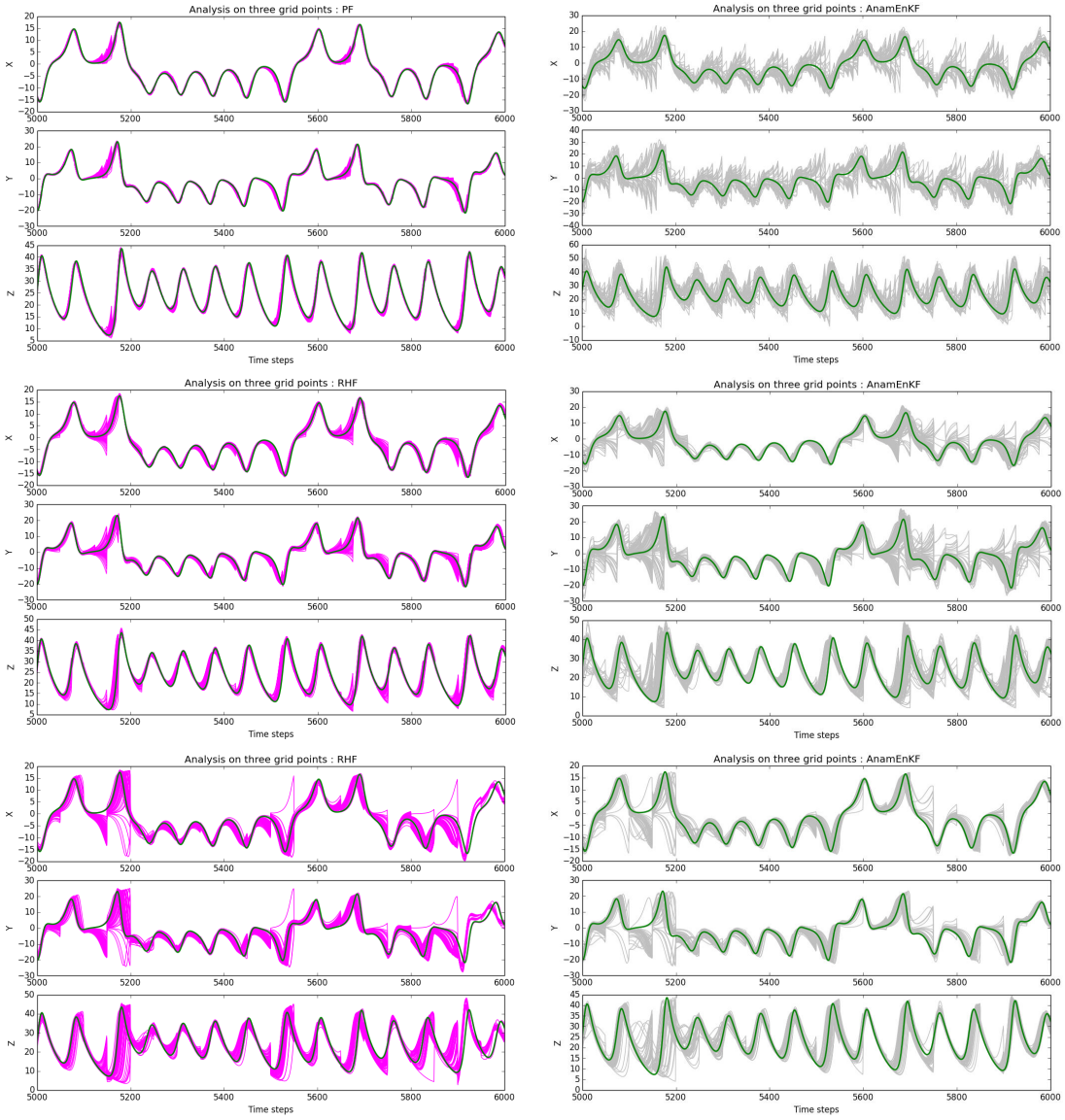

Figure *3.4* – *Idem de la Figure 3.3 pour les ensembles produits par le RHF (graphiques de gauche, courbes magentas) et l'EnKF anamorphosé (graphiques de droite, courbes grises).*

dans le cas faiblement non-linéaire ($\Delta t = 0.10$). L'ensemble semble toutefois bien couvrir la *vérité*. La qualité du PF-*bootstrap* se voit principalement sur les deux configurations plus non-linéaire ($\Delta t = 0.25$ et 0.50). En effet, les ensembles produits sont corrigés avec précision à chaque observation. De plus, le principe de re-échantillonnage du PF-*bootstrap*

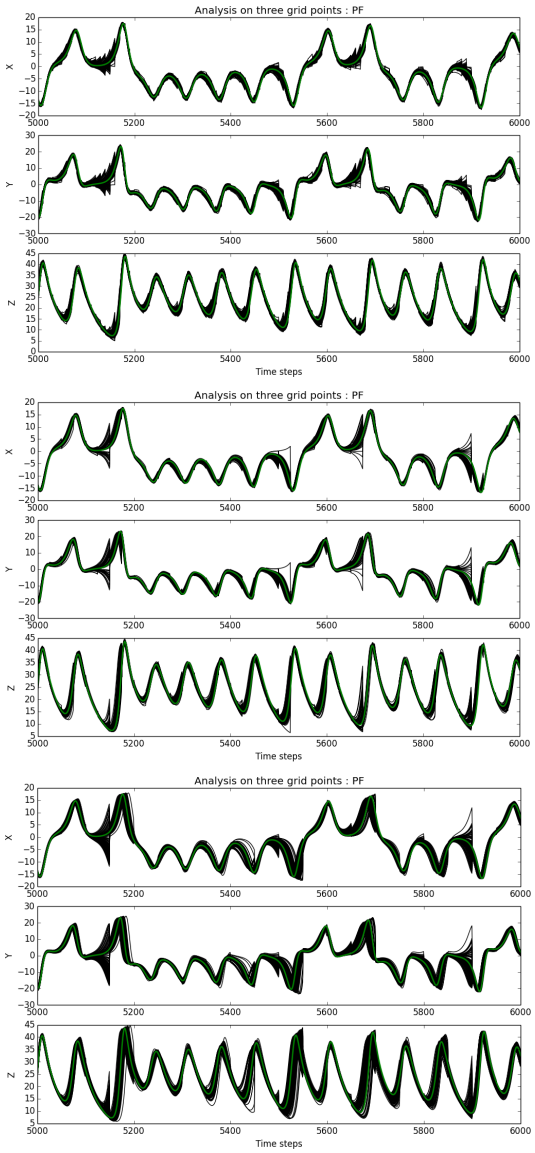

Figure *3.5* – *Idem de la Figure 3.3 pour les ensembles produits par le PF-bootstrap (courbes noires).*

produit chaque membre comme un état du modèle (sans compter l'ajoût de bruit). Ceci a pour effet de maintenir l'ensemble dans un équilibre au sein de l'attracteur du modèle et donc de stabiliser la trajectoire de chaque membre.

Les *Root Mean Square Errors*

Pour avoir une vision plus globale de la qualité des estimés moyens produits par les différentes méthodes, nous regardons à présent les erreurs RMS (*Root Mean Square Errors*). Ces RMSE sont calculées sur les trois variables (x, y, z) et sur la période temporelle entière (i.e. du pas de temps 5000 au pas de temps 10000). Elles sont calculées pour les trois configurations et pour cinq méthodes d'assimilation. Les scores RMSE sont présentés dans la Figure 3.6 en fonction des configurations $\Delta t = 0.10$, 0.25 et 0.50 pour l'EnKF (carrés bleus foncés), l'ETKF (triangles vers le haut verts), l'EnKF anamorphosé (diamants rouges), le RHF (triangles vers le bas cyans) et le PF (croix magentas).

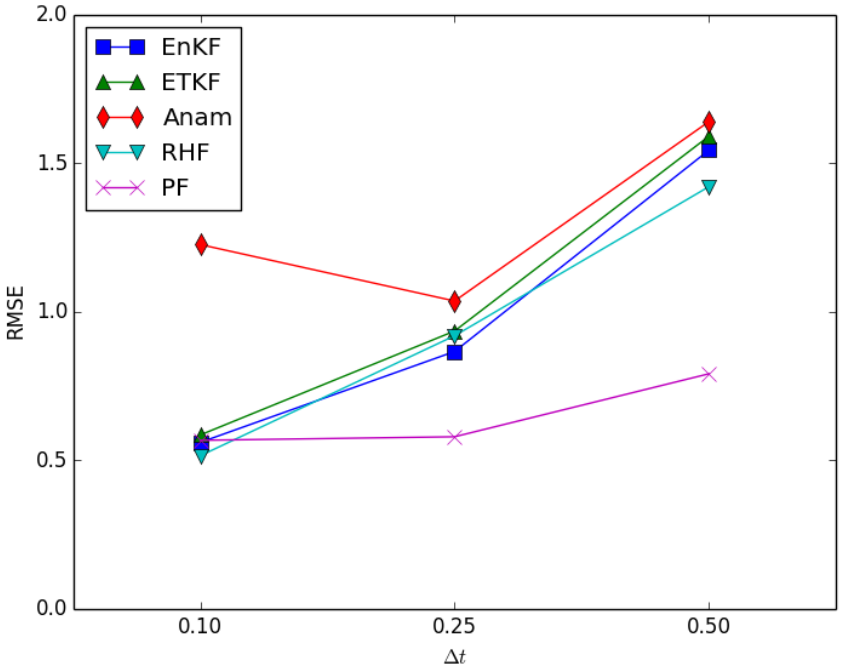

Figure *3.6* – *Scores RMSE des estimés moyens produits par l'EnKF (carrés bleus foncés), l'ETKF (triangles vers le haut verts), l'EnKF anamorphosé (diamants rouges), le RHF (triangles vers le bas turquoises) et le PF (croix magentas) ; en fonction des configurations* $\Delta t = 0.10$, 0.25 *et* 0.50. *Les RMSE sont calculées pour les trois configurations et pour cinq méthodes d'assimilation.*

Les deux filtres de Kalman d'ensemble (EnKFs et ETKF), le RHF et le PF-*bootstrap* présentent des erreurs RMS très proches dans les configurations aux faibles linéarités (cas $\Delta t = 0.10$). La qualité des estimés moyens produits par l'EnKFs, l'ETKF et le RHF se dégradent (les RMSE augmentent) avec la diminution de la fréquence d'observation. Il

est toutefois à noter que dans la configuration aux fortes non-linéarités (cas $\Delta t = 0.50$) le RHF a une RMSE plus faible, ce qui témoigne de sa capacité à mieux gérer les non-linéarités et les non-Gaussianités. De manière plus prononcée, le filtre particulaire *bootstrap* présente des erreurs RMS très faibles dans les cas de plus en plus non-linéaires. Le PF-*bootstrap* est donc très adapté aux fortes non-linéarités et non-Gaussianités lorsqu'il est pouvu d'un nombre de membres suffisant. Enfin, le cas de l'EnKF anamorphosé est plus délicat. La présence de biais entre l'ensemble *a priori* et l'observation ne permet pas de correctement anamorphoser les covariances d'erreurs d'observations. Ce problème se traduit par un effondrement de l'ensemble qui oblige à imposer de fortes inflations (sur l'ensemble et/ou sur les covariances d'erreurs d'observations). Dans ce contexte, les scores RMSE de l'ETKF anamorphosé sont tous moins bons que les autres méthodes présentées. Lors de la mise en place de l'ETKF anamorphosé, il est donc important de faire attention au bon recouvrement des observations par l'ensemble.

3.4 Conclusions

Dans ce chapitre, une description algorithmique des méthodes d'assimilation employées dans la suite de cette thèse a été proposée. Cette description fournit les outils nécessaires à la reproduction des résultats de la thèse. Elle permet également d'apercevoir le travail d'implémentation structurée qui a été mis en place. Nous nous sommes donc munis d'une plateforme python comprenant :

- un filtre de Kalman d'ensemble stochastique (EnKFs),
- un filtre de Kalman d'ensemble transformé (ETKF),
- un filtre de Kalman d'ensemble anamorphosé (Anam-EnKF),
- un filtre d'histogrammes de rangs (RHF),
- un filtre particulaire *bootstrap* (PF-*Bootstrap*),

Cette plateforme nous permet d'appliquer de manière automatique ces méthodes sur n'importe quel problème d'assimilation. C'est ce qui sera fait par la suite, pour des problèmes d'estimation en biogéochimie marine.

On peut discriminer ces cinq méthodes en trois approches pour gérer la non-Gaussianité. La première catégorie comprend les méthodes moindres carrés (donc aux hypothèses Gaussiennes) : EnKFs et ETKF. La deuxième catégorie comprend les méthodes adaptant les moindres carrés aux cas non-Gaussiens : Anam-EnKF et RHF. La troisième catégorie comprend les méthodes ne faisant aucune hypothèse de Gaussianité : PF-*Bootstrap*.

Afin, d'illustrer l'application de ces méthodes, la deuxième partie de ce chapitre présente des expériences d'assimilation de données sur un modèle de petite dimension. Le modèle employé est le modèle de Lorenz à trois variables, Lorenz 63. Ce modèle a été couramment utilisé comme cas-test de référence dans la littérature. Il présente un comportement

typiquement chaotique et ses différentes configurations fournissent des problèmes d'estimation aux non-linéarités et non-Gaussianités variables. Cette illustration nous a permis de constater les bonnes performances des filtres de type Kalman dans les configurations dites quasi-linéaires et peu non-Gaussiennes. De plus, le filtre particulaire confirme ses aptitudes à gérer les problèmes très non-Gaussiens pour un nombre de membres suffisant (ici, 100 membres). Le RHF se comporte comme les filtres aux hypothèses Gaussiennes (EnKFs et ETKF) dans les configurations aux non-linéarités faibles et moyennes. Il produit une erreur RMS légérement plus faible que ces dernières dans la configuration aux fortes non-linéarités en nécessitant moins de membres que le filtre particulaire. Enfin, les résultats produits par l'EnKF anamorphosé dénotent une limitation de notre implémentation de ce filtre. Malgré les bonnes performances de ce filtre dans d'autres contextes (Bertino et al., 2003; Béal et al., 2010), ce filtre est soumis à un effondrement d'ensemble lors de la présence de forts biais entre l'*a priori* et l'observation.

Cette illustration permet, en outre, de constater la nécessité d'une nouvelle méthode pouvant gérer de fortes non-Gaussianités en se basant sur un nombre de membres plus raisonnable que le filtre particulaire. Une telle méthode fait partie de la plateforme assimilation implémentée. Cette méthode est le filtre multivarié d'histogrammes de rangs (MRHF). Ce filtre a été développé au cours de cette thèse. Sa mise en place et son évaluation sur le modèle de Lorenz 63 sont présentées dans le chapitre-article (Metref et al., 2014) suivant.

Chapitre 4

Développement d'un filtre non-Gaussien : *le MRHF*

Sommaire

Ce chapitre est un article publié : Metref et al., 2014

Résumé

Un défi de l'assimilation de données géophysique est de traiter le problème des non-Gaussianités dans les distributions des variables physiques engendrées, dans de nombreux cas, par des modèles dynamiques non-linéaires. Les méthodes non-Gaussiennes d'analyse d'ensemble se divisent en deux catégories, celles transférant les particules de l'ensemble en approximant le meilleur estimé linéaire non-biaisé (BLUE) e.g. le filtre de Kalman d'ensemble (EnKF), et celles rééchantillonant les particules en appliquant directement la loi de Bayes, comme le filtre particulaire. Dans cet article, il est suggéré que les méthodes classiques de transfert peuvent uniquement gérer les distributions faiblement non-Gaussiennes, alors que les autres souffrent de problèmes d'échantillonnage. Entre ces deux catégories, une nouvelle méthode de transfert appliquant directement la loi de Bayes, le filtre multivarié d'histogrammes de rangs (Multivariate Rank Histogram Filter, MRHF), est présenté comme une extension du filtre d'histogrammes de rangs (Rank Histogram Filter, RHF) proposé en premier lieu par Anderson (2010). Ses performances sont évaluées et comparées à celles d'autres méthodes, sur différents niveaux de non-Gaussianité avec le modèle de Lorenz 63. Le comportement de la méthode est ensuite illustré sur un problème simple d'estimation de densités utilisant un ensemble de simulations provenant d'un modèle couplé physique bio-géochimie en Atlantique Nord. Le MRHF est performant sur des systèmes de faible dimension dans des régimes fortement non-Gaussiens.

A non-Gaussian analysis scheme using rank histograms for ensemble data assimilation

S. Metref[(1)], E. Cosme[(1)], C. Snyder[(2)] and P. Brasseur[(1)]

[(1)]CNRS, LGGE, F-38041 Grenoble, France
[(2)]National Center of Atmospheric Reaserch, Boulder Colorado, USA

Abstract

One challenge of geophysical data assimilation is to address the issue of non-Gaussianities in the distributions of the physical variables ensuing, in many cases, from nonlinear dynamical models. Non-Gaussian ensemble analysis methods fall into two categories, those remapping the ensemble particles by approximating the best linear unbiased estimate e.g. the Ensemble Kalman filter (EnKF), and those resampling the particles by directly applying Bayes' rule, like particle filters. In this article, it is suggested that the most common remapping methods can only handle weakly non-Gaussian distributions, while the others suffer from sampling issues. In between those two categories, a new remapping method directly applying Bayes' rule, the Multivariate Rank Histogram Filter (MRHF), is introduced as an extension of the Rank Histogram Filter (RHF) first introduced by Anderson (2010). Its performances are evaluated and compared with several data assimilation methods, on different levels of non-Gaussianity with the Lorenz 63 model. The method's behaviour is then illustrated on a simple density estimation problem using ensemble simulations from a coupled physical biogeochemical model of the North Atlantic ocean. The MRHF performs well with low dimensional systems in strongly non-Gaussian regimes.

4.1 Introduction

The principal goal of data assimilation is to estimate the state of a dynamical system, based on prior information and a time series of observations, while calculating probabilistic measures corresponding to the accuracy of this estimation. Kalman Filter theory (Kalman, 1960) became a reference in data assimilation as it provides the optimal solution to the linear and Gaussian filtering problem (Cohn, 1997). The

Ensemble Kalman Filter (EnKF, Evensen, 1994) and closely related methods (Lermusiaux and Robinson, 1999b; Pham, 2001; Whitaker and Hamill, 2002, to cite only a few) are different implementations of the Kalman filter relying on ensembles. For the analysis step, they transform the prior ensemble into a posterior ensemble, using a function that is optimal (an optimal map, Cotter and Reich, 2013) under the assumption of Gaussianity of the prior ensemble and the observation errors. Those methods are applicable to high dimensional systems in meteorology (Whitaker et al., 2008; Buehner et al., 010a) and oceanography (Lermusiaux, 2006; Sakov et al., 2012). This success is – in part – due to the fact that the dynamics of these systems are weakly nonlinear, that is, do not strongly deviate from a linear evolution within the space and time scales characterizing the density of available observations. Briefly speaking, a weak nonlinearity transforms a Gaussian distribution into a weakly non-Gaussian distribution, with which the EnKF still performs well. Many recipes have been developed to enforce the good behaviour of the EnKF with such systems, including localization techniques (Sakov and Bertino, 2010; Greybush et al., 2011), sampling strategies (Pham, 2001; Anderson, 2012), observational targeting (Bishop et al., 2000) ; see Bocquet et al. (2010) for more details and other examples. Nevertheless, EnKF-based methods remain sensitive to the violation of the Gaussian assumption (Lawson and Hansen, 2004; Lei et al., 2011) and may lead to unwanted phenomena such as inaccurate estimations, failure to respect non-linear physical balances, or more dramatically to instability of the filter.

Along with the developments of the EnKF, there is a growing need for non-Gaussian ensemble data assimilation methods. Data assimilation is no longer a tool solely for meteorology and oceanography. Other disciplinary fields with stronger nonlinearities and much sparser data networks (geomagnetism for instance, Fournier et al., 2010) increasingly depend upon data assimilation. Even in the traditional fields of application, models' nonlinearity tends to increase along with their complexity. Even with linear models, non-Gaussian observation error densities make the assimilation problems non-Gaussian. Intrinsically non-Gaussian variables are common in the atmosphere and the ocean, such as humidity (Dee and Da Silva, 2003), concentrations of sea ice or phytoplankton (Brankart et al., 2012).

The ensemble data assimilation methods can be sorted in two categories : those that transform the prior ensemble particles using a deterministic map (transform methods), and those that sample the posterior probability density (sampling methods). They can also be classified as parametric, non-parametric, or semi-parametric, depending on the assumptions on the shape of the probability densities they use. The EnKF falls in the parametric (with the Gaussian assumption) transform methods. The EnKF with Gaussian anamorphosis, further described in Section 4.2 of the present paper, transforms variables to make their densities Gaussian before applying

the EnKF analysis. It can be considered as a semi-parametric transform method, since it is not a fully non-parametric method able to deal with any kind of probability density, as illustrated in Section 4.2. The Truncated-Gaussian EnKF as described by Lauvernet et al. (2009) is of the parametric, sampling type. Reich (2013) introduces a sequential method of the non-parametric, transform category. The particle filter (Gordon et al., 1993; van Leeuwen, 2009) is the most popular method of the non-parametric, sampling type, but is well known to be particularly subject to the curse of dimensionality, which makes it difficult to use with high-dimensional systems (Snyder et al., 2008). Finding solutions to make the particle filter applicable to high dimensional systems is a very active topic of research (Nakano et al., 2007; van Leeuwen, 2010; Morzfeld et al., 2012; Snyder, 2012). A few of them actually rely on some hybridization with the EnKF (Bocquet et al., 2010; Lei et al., 2011; Hoteit et al., 2012).

Ensemble methods of the non-parametric, transform category have rarely been explored, although they could be less sensitive to the curse of dimensionality than the sampling methods, due to the transformation step that helps enforce a better fit of particles to the observations. In this respect, the approach proposed by Reich (2013) would deserve further examination, in particular with high-dimensional systems. Another method that could be somewhat classified as a partly non-parametric transform method is the Rank Histogram Filter (RHF, Anderson, 2010). The RHF is a hybrid between the EnKF and a fully non-Gaussian approach, named that way because it is based on a statistical processing similar to the rank histograms (Anderson, 1996; Hamill, 2001) used for ensemble forecast evaluation. The RHF corrects observed variables by representing their prior densities and the observation likelihoods as piecewise continuous functions in order to directly apply Bayes' rule. This theoretically solves the generalized problem for a single observed variable. However, the other variables are still corrected using a linear regression onto the corrections of observed variables, as in the EnKF. We believe these advantages justify a more detailed exploration of the RHF philosophy. The main objective of this paper is to present an extension of the Rank Histogram approach of Anderson (2010) to unobserved variables, yielding a fully non-parametric transform scheme for ensemble data assimilation in the spirit of the method of Reich (2013). Throughout the paper, this Multivariate RHF is referred to as MRHF.

The article is outlined as follows. In section 4.2, we present some considerations on the joint non-Gaussianity of two variables, and how ensemble analysis schemes perform with such densities. Emphasis is given to the EnKF and to the RHF of Anderson (2010). Section 4.3 develops the extension of the RHF to unobserved variables called MRHF along with an approximation of the latter. Numerical experiments are presented in Section 4.4, where the MRHF and its approximation are evaluated with

the highly nonlinear and non-Gaussian Lorenz 1963 model, and compared in different setups (corresponding to different levels of non linearities) to the EnKF, to the RHF and to a particle filter. In Section 4.5, the new schemes are finally illustrated with a density estimation problem based on a realistic ensemble from a coupled marine biogeochemical model. Even though this last experiment is not a data assimilation problem, the results give an insight into the behaviour of the method. A discussion and a conclusion are given in the last section.

4.2 Gaussian and non Gaussian analysis in ensemble filtering

The increasing popularity of ensemble filters is largely due to the relative simplicity of their implementation. They basically alternate propagation steps and analysis steps. During a propagation step, each particle of the ensemble is advanced in time using the dynamical system model, possibly including some parameterisation of the model error. An analysis step occurs after a propagation step, when an observation $\mathbf{Y}^m$ is available. The observation $\mathbf{Y}^m$ is a realization of the (random) measurement vector $\mathbf{Y}^o = h(\mathbf{X}) + \epsilon$, where h is a forward observation operator, $\mathbf{X}$ the state vector to be estimated, and ϵ the observation error. The analysis step conflates the prior ensemble $\{\mathbf{X}_i^f\}_{i=1,\dots,N_e}$, composed of N_e particles resulting from the previous forecast, and the available observation $\mathbf{Y}^m$, to provide a posterior (analysis) ensemble $\{\mathbf{X}_i^a\}_{i=1,\dots,N_e}$. Observation errors are often assumed temporally and spatially uncorrelated so that each one can be independently assimilated (Houtekamer and Mitchell, 2001; Evensen, 2003). If the spatial correlations cannot be neglected, a linear transformation of the observation vector is theoretically possible (and exact in the linear and Gaussian context), in which the observation error covariance matrix is diagonal (Anderson, 2003).

A popular implementation of the ensemble analysis is the EnKF with serial processing of observations (Houtekamer and Mitchell, 2001). Following the description given by Anderson (2003), the ensemble update by each observation is performed in two steps, namely the update of the observed variable, then the update of unobserved variables. If the observation is not a direct observation of a state variable, then the state vector $\mathbf{X}$ can be augmented with the observed function of state variables, $\mathbf{Y}^o$, to introduce a directly observed variable. In what follows, only the case of the direct observation of a variable is considered. With two scalar variables, $\mathbf{X} = (x, z)^T$, z being subject to a direct measurement z^o with realization z^m, decomposing the EnKF analysis equations shows that the correction of the observed variable z for ensemble particle i is :

$$\delta z_i = \frac{\mathrm{Var}(z)}{\mathrm{Var}(z) + \mathrm{Var}(\epsilon)}(z^m - z_i - \epsilon_i), \tag{4.1}$$

where ϵ_i is a perturbation that takes the observation error ϵ into account. The correction for the unobserved variable x, for particle i, is :

$$\delta x_i = \frac{\mathrm{Cov}(x,z)}{\mathrm{Var}(z)}\,\delta z_i. \tag{4.2}$$

It is clear from the latter equation that the correction of any unobserved variable is a function of the linear correlation between the variable and the observed variable. Such formulation must be questioned when the linear correlation is not a relevant measure of the statistical relationship between these variables, as may occur when the statistics are non-Gaussian. This issue is illustrated in Fig. 4.1 : On the left panel, a prior ensemble of a bivariate Gaussian state (X, Z) is depicted (green dots). The second variable Z is observed. The corrections for the unobserved variable, based on a relevant linear correlation with the observed variable, leads to an analysis ensemble (red dots) fitting the bivariate Gaussian pdf that would be produced by implementing Bayes' theorem. This analysis ensemble is consistent with both the physics, introduced through the prior information, and the observation. On the right panel, the two (non jointly Gaussian) variables exhibit a non linear statistical relationship, that cannot be fully captured by a linear regression. Consequently, even if the corrections of the observed variable are somewhat correct, those of the unobserved variable can be erroneous. This results in a rather poor analysis ensemble, where particles appear in unexpected parts of the phase space, in violation of the inter-variable relationship, as described by the prior ensemble.

This problem is not new ; it falls under the general designation of non Gaussian data assimilation. Solutions exist, among the "resampling" methods in particular, but their effective application in high dimensions is either impossible or requires further development ; see Bocquet et al. (2010) for a review. Some non Gaussian schemes derive from refinements to the EnKF. A promising one, mostly studied in oceanography, is the Gaussian anamorphosis (Bertino et al., 2003; Simon and Bertino, 2009; Béal et al., 2010; Brankart et al., 2012). Anamorphosis consists in transforming the initial physical variables to make them fit Gaussian distributions. The standard EnKF analysis can be applied to these transformed variables. Then, the physical analysis variables are recovered by the inverse transformation. The transformation can be either analytical or numerical (Bocquet et al., 2010). The following illustration is performed with the numerical transformation described by Brankart et al. (2012). The left panel of Fig. 4.2 shows the same non Gaussian prior ensemble as the right panel of Fig. 4.1 (green dots), along with the analysis ensemble obtained using Gaussian anamorphosis (red dots). Anamorphosis clearly improves the EnKF. But anamorphosis presents several limitations, one of which is that it is based on a one-to-one correspondence between the prior and the target (Gaussian) distributions involved in the transformation. If this is not verified, Gaussian anamorphosis fails.

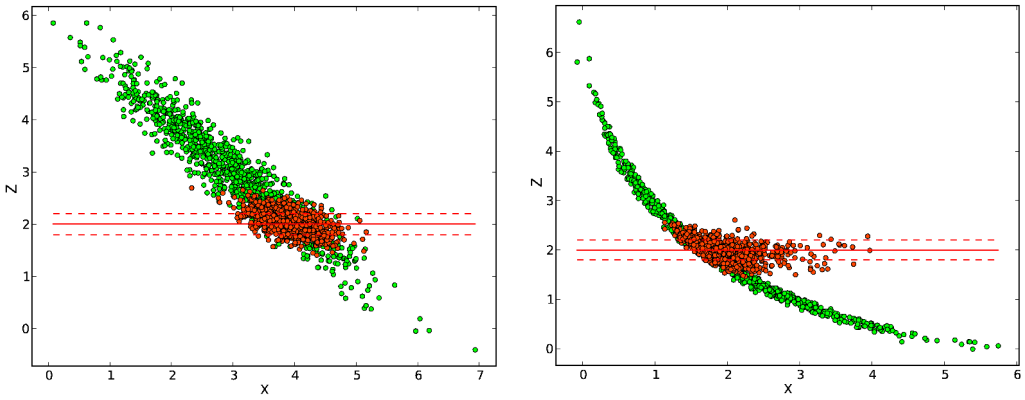

Figure ***4.1*** – *Illustration of EnKF analyses with joint Gaussian (left panel) and weakly non-Gaussian (right panel) prior distributions. The prior ensemble is represented by the green dots and the posterior ensemble is represented by the red dots. The variable corresponding to the Y axis is observed with a value shown by the red solid line. Its uncertainty (identical in all the illustrations of Sec. 4.2) is assumed Gaussian with standard deviations symbolized with the red dashed lines.*

Such failure is depicted on the right panel of Fig. 4.2, where the prior ensemble follows a strongly non Gaussian density law which exhibits bimodality under conditioning by z. The EnKF with Gaussian anamorphosis (or without ; not shown) provides a very poor analysis ensemble.

Fully non Gaussian ensemble analysis schemes, i.e. schemes derived without any assumption on the shape of the prior ensemble density, implement Bayes' rule to solve the analysis step :

$$p(\mathbf{X}|\mathbf{Y^o}) \propto p(\mathbf{X})p(\mathbf{Y^o}|\mathbf{X}), \tag{4.3}$$

where $p(\mathbf{X})$ is the prior probability density for the state vector $\mathbf{X}$ to estimate, $p(\mathbf{Y^o}|\mathbf{X})$ is the observation likelihood (identical to the observation density for Gaussian observation errors), and $p(\mathbf{X}|\mathbf{Y^o})$ the posterior density, i.e. the density of the state given the observations. A detailed Bayesian description of data assimilation is provided by Wikle and Berliner (2007) for instance. The fully non-Gaussian ensemble data assimilation problem is usually solved by resampling methods of the particle filter type. The particle filter (Gordon et al., 1993; Doucet et al., 2001) is subject to very active developments for its future application with high dimensional geophysical problems (Nakano et al., 2007; van Leeuwen, 2009, 2010; Morzfeld et al., 2012). The key point in implementing a particle filter is to beat the curse of dimensionality, and that will probably not be solved shortly for large applications (Snyder et al., 2008). One major reason, we believe, is that particle filters have yet to implement localization, in which any given observation affects the update only in a spatially local

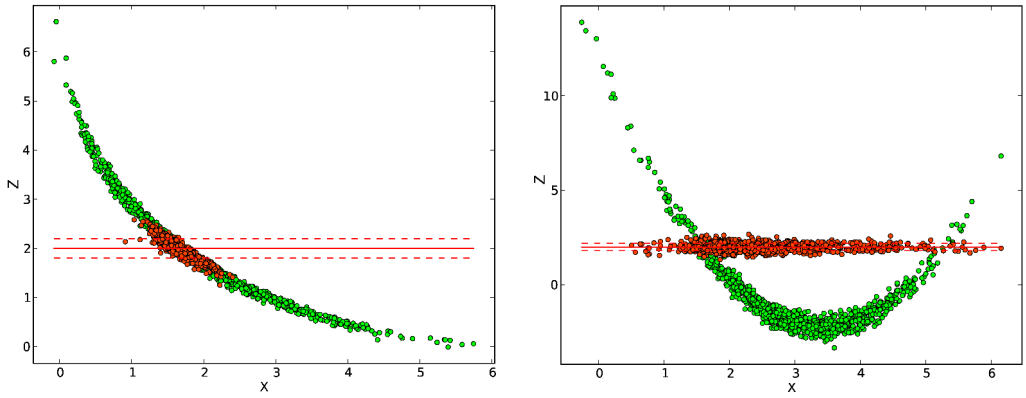

Figure ***4.2*** – *Same as Fig. 4.1, but for the EnKF with Gaussian anamorphosis, and for joint weakly non-Gaussian (left panel) and strongly non-Gaussian (right panel) prior distributions. The weakly non-Gaussian prior is the same as in Fig. 4.1, right panel.*

region near the observation location and as is common in the EnKF (Houtekamer and Mitchell, 1998; Hamill et al., 2001).

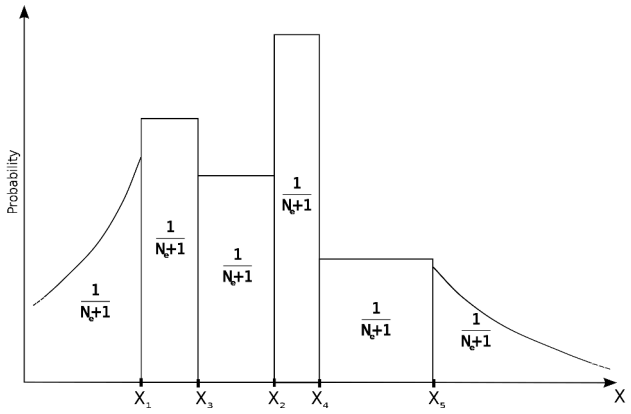

Figure ***4.3*** – *Reconstruction of a density from an ensemble using the Rank Histogram approach.*

Here, we explore a new non-Gaussian ensemble analysis scheme of the "transform" type, in which localization can be implemented. We start from the the Rank Histogram Filter (RHF), a partially non-Gaussian transform scheme, that has been proposed by Anderson (2010). The RHF processes observations serially. For the direct observation z^o of variable z, the continuous prior density for z is represented as a rank histogram (by analogy with the rank histogram diagnostic used to eva-

luate ensemble predictions, as introduced in geophysics by Anderson, 1996, and later discussed in Hamill, 2001, and Candille and Talagrand, 2005). The histogram is composed of $N_e - 1$ bounded regions partitioned by the sorted ensemble particles (the order statistics of the problem) and two unbounded regions on the tails. In each inner region, a density value is assigned so that the region contains a probability mass of $\frac{1}{N_e+1}$ (Fig. 4.3). The two outer regions are covered by tails of probability mass $\frac{1}{N_e+1}$ as well ; Their shape may be chosen freely, and this may actually be a key element for the success of the RHF (Anderson, 2010). In particular, very long tails can help to correct biases and to make the filter more resilient to divergence. More precisely, the prior density of z is written :

$$p(z) = \frac{1}{N_e + 1} \sum_{i=1}^{N_e-1} \frac{1_{[z_i, z_{i+1}[}(z)}{(z_{i+1} - z_i)} + T(z), \tag{4.4}$$

with $1_{[z_i,z_{i+1}[}(z)$ the indicator function on the interval $[z_i, z_{i+1}[$ (yielding 1 if z belongs to this interval, 0 otherwise) and $T(z)$ also a combination of indicator functions representing the tails term applied to the two outer regions. The likelihood $p(z^o|z)$ is known analytically from the observation error density. It is discretized on the same grid as $p(z)$, and the two functions are multiplied point-wise to provide a constant piecewise expression (after normalisation) of the posterior density $p(z|z^o)$. The analysis ensemble is finally obtained using a (deterministic) procedure of inversion of the cumulative distribution function. Corrections on z particles are calculated by the difference between the posterior and the prior values of z.

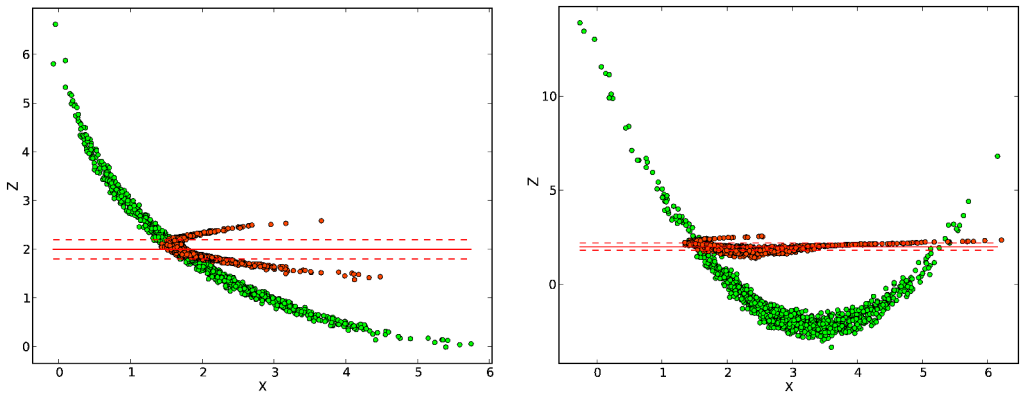

Figure 4.4 – *Same as Fig. 4.2, but for the RHF.*

These corrections are used to compute the corrections for the unobserved variables with Eq. 4.2, that is, applying a linear regression. This latter step, inherited from

the EnKF, is perhaps the main weakness of the RHF, as illustrated in Fig. 4.4 : The RHF analysis performs rather poorly in both weakly and strongly non-Gaussian cases addressed in the previous illustrations. However, considering the many positive aspects of this scheme (non Gaussian, robust, deterministic, possible to localize), it seems worth trying to correct this weakness and extend the rank histogram approach to unobserved variables.

4.3 Mutivariate Rank Histogram Filter

4.3.1 Principle

We wish to generalize the RHF to the general Bayesian framework. The RHF first addresses the analysis of the observed variable, then deals with the others. It is thus an implementation of the *sequential realization method* to sample a multivariate probability density, as presented by Tarantola (2005) for example. This method leans on the Knothe-Rosenblalt rearrangement, a decomposition of the joint probability density into a product of marginal and conditional univariate densities. With 3 scalar variables $\mathbf{X} = (x,\, y,\, z)$, this decomposition is :

$$p(x, y, z) = p(z)p(x|z)p(y|x, z). \tag{4.5}$$

A sample from the joint density is obtained by deterministically sampling $p(z)$ first, then $p(x|z)$ (using the result for $p(z)$), and $p(y|x, z)$ (using the previous two results). The purpose of data assimilation here is to condition the joint density to an observation z^o of z. Following the usual Markovian memoryless assumption for the observation process, which implies $p(z^o|x, y, z) = p(z^o|z)$, and using the decomposition (4.5) it is straightforward to find that :

$$p(x, y, z|z^o) \;\; = \;\; p(z|z^o)p(x|z)p(y|x, z). \tag{4.6}$$

Here again, the deterministic sampling of both densities conditioned on z are based on the previously sampled densities ; hence, the sampling of variables x and y depends on the observation z^o. To obtain the first factor on the right hand side, the EnKF uses Eq. 4.1 ; the RHF implements Bayes' rule for z : $p(z|z^o) \propto p(z)p(z^o|z)$. But for both methods, the second and third terms on the right hand side are computed using Eq. 4.2, which comes from a Gaussian, Kalman filtering perspective. We propose below a new non-Gaussian approach to sample scalar particles from these conditional densities to implement Eq. 4.6 with non-parametric densities. This scheme is deterministic, in the sense that no random number need be generated during the analysis process. The analyses are then reproductible and the method is of the "transform" type.

4.3.2 Implementation of the MRHF analysis

Let $\{z_i^a\}_{i=1,...,N_e}$ be the posterior ensemble of the observed variable z, i.e. a sample of $p(z|z^o)$. Consider the first unobserved variable, x in Eq. 4.6. The MRHF analysis determines x_i^a, the x analysis value for particle i, by deterministically sampling the conditional density $p_i^a(x) \equiv p(x|z = z_i^a)$. This density must first be formed. Some steps of the procedure are illustrated in Fig. 4.5. The green dots represent the prior ensemble in the $X-Z$ plane ; the blue dots at $X = 0$ represent the z analysis ensemble $\{z_i^a\}_{i=1,...,N_e}$. The red line is the observation realization. The following process is repeated for $i = 1, ..., N_e$. For a given i, a subset of particles is selected in the prior ensemble (green dots with blue circles), whose z values lie in the neighbourhood of z_i^a (blue dot with red triangle) along the z direction. The selection process is discussed later. Applying the rank histogram approach to the x values of the selected particles, a one-dimensional density is then formed to represent $p_i^a(x)$.

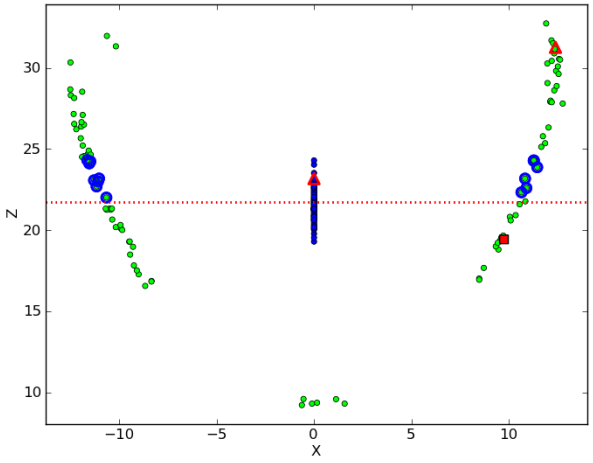

Figure 4.5 – *Illustration of the particle-by-particle MRHF analysis step for the first unobserved variable : Green dots represent the prior ensemble ; blue dots vertically aligned at $X = 0$ represent the posterior z ensemble. The red dotted line is the observation of z, the red square is the true state (not used for the analysis). The red empty triangles show the particle being processed and its corresponding z analysis value. Blue circles show the selected particles to form the posterior density $p_i^a(x)$. See text for details.*

To follow Eq. 4.6, one approach would be to draw a random realization from this density to provide x_i^a. This, however, is far from optimal from the physical viewpoint, because it can generate large corrections resulting in physical instabilities and imbalances, as previously observed by Anderson (2003) in the EnKF context. In Fig. 4.5 for instance, the prior particle (green dot with red triangle) is in the right

hand side mode of the distribution. Since the observation does not enable one to know in which mode the truth (red dot) actually is, it makes sense to try to keep this particle in its mode of origin, thus minimizing its modification.

Instead of a random draw in $p_i^a(x)$ which could arbitrarily move the particle to the left hand side mode, the following steps are proposed :

- With a similar selection and a rank histogram process, form the density of x conditioned to the *background* value of z : $p_i^b(x) \equiv p(x|z = z_i^b)$;
- Compute the cumulative distribution functions $C_i^b(x)$ and $C_i^a(x)$ from $p_i^b(x)$ and $p_i^a(x)$, respectively ;
- Compute the position of the prior particle in the prior density : $c_i = C_i^b(x_i^b)$;
- Preserve the rank of the particle in the posterior density by taking $x_i^a = C_i^{a\,(-1)}(c_i)$ as analysis value for x and particle i. This is illustrated in Fig. 4.6.

Although this method does not always prevent a particle shifting from one mode to another, two neighbouring particles (i.e. close to each other) in the prior ensemble are likely to remain neighbours in the posterior ensemble. In a multimodal case for instance, two particles in the same mode are more likely to remain in the same mode after analysis.

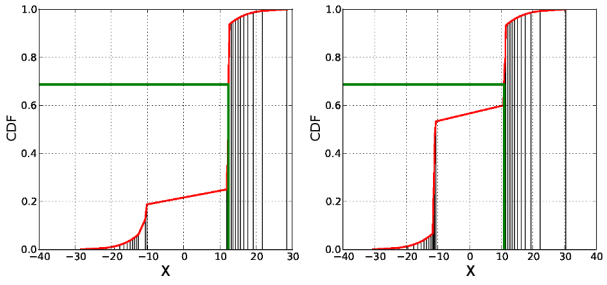

Figure *4.6* – *Illustration of the particle-by-particle MRHF analysis step for the first unobserved variable (one particle) : Red lines represent the cumulative distribution functions (cdf) of the prior density $p_i^b(x)$ (left panel) and the posterior density $p_i^a(x)$ (right panel). The vertical black lines show the selected particles used to build these densities. To compute the x analysis value for the prior x particle near 11 on the left panel, the green line must be followed : The cdf for the prior x is computed with the prior cdf (Result is near 0.68) ; from this cdf value, the x analysis value is obtained on the right panel. See text for details.*

Once the z and x analysis values are computed for each particle, the analysis values for the third variable y can be computed. The process is strictly similar to the one described above, but for the variable y, and with an additional $x = x_i^a$ term in the conditional statement. In practice, this reduces to selecting particles from the prior distribution in the neighbourhood of (x_i^a, z_i^a) in the two-dimensional plane (x, z), to

form the density $p(y|x, z, z^o)$. The other steps remain unchanged.

As a remark, one may notice that this analysis method brings some similarities with the heuristic method presented in Anderson (2003) in the EnKF context. His idea is to compute the covariance term in Eq. 4.2 using a subset of neighbouring particles. The MRHF goes one step further by considering the nonlinear relationship between those particles.

4.3.3 Selection of particles and mean-field approximation

We now come back to the selection of particles, and start with the first unobserved variable x. To represent the target density $p_i^a(x)$ accurately, the selected particles must have a z value close to z_i^a used in the conditional statement. At the same time, there may be few or no particles whose z values differ from z_i^a by less than a specified, small amount, since N_e is finite. Thus, there is a trade-off between selecting particles that are very close to z_i^a and selecting a sufficient number of particles to represent $p_i^a(x)$.

In the numerical experiments with the Lorenz 63 system, presented in Section 4.4, a maximal distance ($d_{\max}$) is prescribed, along with a minimal and a maximal number of particles. Specifications of these parameters are detailed in Section 4.3.4. No attempt has been made in this work to make the scheme independent of the variables, for example by normalizing all variable by their prior variance.

For the second unobserved variable y, the difference is computed as a distance in the two-dimensional plane (x, z). The maximal distance to consider as a threshold must be prescribed accordingly. As the algorithm proceeds to additional unobserved variables, the curse of dimensionality becomes apparent : each time a new unobserved variable is analyzed, a dimension is added and the volume of states within the maximal distance decreases so fast that each region defined by one particle and a finite radius around it has negligible probability to hold another particle.

As a schematic illustration, assume that $z_i^a = 30$ in Fig. 4.5, and $d_{\max} = 1$. There exist prior particles with $29 < z < 31$. Thus, the analysis for x can be conducted accurately. Assume now that the x analysis provides $x_i^a = -10$. Prior particles in the two-dimensional neighbourhood of (x_i^a, z_i^a) (for example, in a circle centered on this point with radius $\sqrt{2}$) are sparse or nonexistent. More distant particles must therefore be included and the accuracy of the analysis for y may be poor.

This obstacle leads us to introduce an approximation, termed the mean-field approximation by Cotter and Reich (2013), which consists in dropping the unobserved variables in the conditional statements in Eq. 4.6, and thus computing the posterior density as :

$$p(x, y, z|z^o) \simeq p(z|z^o)p(x|z)p(y|z). \tag{4.7}$$

This amounts to processing each unobserved variable in the same way as the first one. The approximation (4.7) limits the scheme's ability to handle complex, jointly multimodal densities, as will be illustrated with the Lorenz (1963) model in Section 4.4. An important advantage of the approximation is that it makes the analyses of the different unobserved variables independent. Thus, they can be parallelized on a computer.

4.3.4 MRHF parameters and possible tuning

Several of the MRHF parameters are related to the computation of rank histograms that is used to update both the observed and unobserved variables. Building a pdf with the rank histogram approach implies a division by the distance between two consecutive particles (Anderson, 2010). To avoid possible computational overflow, it is important to set a minimum spacing ϵ_{RHF} between two consecutive particles. This is done by moving each particle at a distance of ϵ_{RHF} from its closest neighbour when necessary. The particles are processed sequentially from the mean toward the tails of the distribution. For the following experiments with the Lorenz 63 system, a wide range of ϵ_{RHF} values are tested, from 10^{-6} to 10^{-2}, and the values that provide the smallest errors are retained. The main and expected conclusion is that experiments with the RHF and MRHF with large ensemble sizes are sensitive to ϵ_{RHF} within this range, because large ϵ_{RHF} tend to excessively diffuse peaked probability densities when those are built by a large number of particles close to each other.

As stressed by Anderson (2010), several choices are possible for the shape of the tails. With the Lorenz 63 system, sensitivity tests have shown that the MRHF with small ensembles and frequent observations ($\Delta t = 10$) is sensitive to the shape of the tails, and performs better with Gaussian tails. For larger ensembles, results are similar with different shapes of tails. Constant tails are specified because it is a bit cheaper computationally. Constant tails extend to prescribed values slightly beyond the model phase space boundaries. Tails are introduced only for the observed variables, because their probability densities are multiplied by the observation densities before resampling. No tail is introduced for unobserved variables. For the Lorenz 63 model, the minimal and maximal values are set to $[-20, -30, 0]$ and $[20, 30, 50]$, respectively.

An additional parameter controls the scheme's behaviour when the prior distribution has multiple, well separated modes. Figure 4.6 (left) shows the cumulative distribution function of a probability density made of two disjoint modes. Between the two modes, this function increases, although it should remain constant because the modes are disjoint. This is due to the rank histogram approach to build the probability density, and emphasized by the limited number of particles in the ensemble. In the analysis step, unrealistic particles may then appear in the region between the two modes. In the Lorenz 63 experiments, the probability density has been set to

zero when below a threshold of $1/6 \times 1/(N_e + 1)$.

As discussed in section 4.3.3, the particle selection depends on three additional parameters. The first is the maximal distance $d_{\max}$. The specification of $d_{\max}$ should account for the magnitude and the variability of the variables, the dimension of the space in which the difference is defined, and the ensemble size. With the Lorenz 63 system, we take $d_{\max} = d\sqrt{n}$ where n is the dimension of the space (n=1 for the first unobserved variable ; n=2 for the second unobserved variable) and d is prescribed according to the ensemble size : 1 for small ensembles ($N_e = 8, 16, 32$), 0.1 for medium ensembles ($N_e = 64, 128$), and 0.01 for the larger ensembles ($N_e = 256, 512$). To ensure a sufficient but not too large number of selected particles, it is wise to fix a minimum and a maximum number of particles. In the following experiments, those are set to 5 and 15, respectively.

Finally, like many ensemble methods, the MRHF suffers from sampling errors in the description of the densities. With the EnKF, this is usually corrected with covariance inflation. This does not make real sense with the MRHF, since the analysis does not rely on covariances. After the analysis, the particles are slightly perturbed with a white Gaussian noise, as it is often done with particle filters to avoid collapse toward one single particle. For the Lorenz 63 experiments that follow, a few values of variance have been tested in the range 0.01-0.05 for this noise, and the experiments yielding the smallest errors have been retained.

4.3.5 Localization

Localization consists in reducing the corrections to some variables, as computed during the analysis step, according to their distance from the observation. Beyond a certain distance, all corrections are set to 0. With the MRHF, localization is straightforward because it is a transform method. Similar to the EnKF with serial processing of observations, the analysis corrections are multiplied by coefficients that are functions of the distance to the observation. Therefore the correction may be restricted to elements of the state vector that are spatially close to the location of the observed variable, and changes to the state vector can then be gracefully tapered to zero as distance from the observation location increases. In the present work, no localization has been applied in the experiments since it is not needed on those assimilation problems. Localization will be studied in future works.

4.3.6 Connections with other methods

The MRHF is a non-parametric transform ensemble data assimilation method. We have presented it as a generalization of the RHF, but it also has connections with other non-parametric transform methods discussed in the Introduction, such

as the method introduced by Reich (2013). This method derives from the theory of optimal transportation (or optimal mapping), which application to sequential data assimilation has been suggested by El Moselhy and Marzouk (2012), Cotter and Reich (2013), and Reich (2013). Instead of trying to compute an approximation of the posterior pdf, the crucial idea of this theory is to find a "transfer map" f such that $f(\mathbf{X})$ is distributed according to the posterior pdf when $\mathbf{X}$ is distributed according to the prior pdf. Once such a transfer map is identified, a posterior sample can be generated from the prior sample. The transfer map is a mathematical expression of the consistency of the posterior sample depending on both the prior sample and the observation. However a transfer map is not unique (Villani, 2009). To find one, one may require the map to satisfy some additional optimality condition. Cotter and Reich (2013) and Reich (2013) propose to find the map that minimizes the expected squared distance between $\mathbf{X}$ and $f(\mathbf{X})$, so as to make the smallest possible changes to go from the prior to the posterior sample. As described in Section 4.3.1, the MRHF uses a transfer map. This map is particularly simple, since only one-dimensional probability densities are involved : we choose the map that preserves the position of the particle during its transfer from the prior to the posterior density. This also makes the smallest possible changes to go from the prior to the posterior sample.

4.4 Numerical experiments with the Lorenz 63 model

The Lorenz-63 model (L63) is a well known system of three ordinary differential equations based on a simplification of atmospheric cellular convection (Lorenz, 1963). It is also a widely used test case for developments in data assimilation. The state variables are denoted X, Y, Z. The usual configuration of the model is adopted here : The parameters are set to $\sigma = 10$, $\rho = 28$ and $\beta = 8/3$. The integration time step is $\delta t = 0.01$. In the following experiments a simulation of reference is performed and considered as the true trajectory to recover.

4.4.1 Fully observed state vector

Experimental set-up and diagnostics

In this subsection, the full state ($[X, Y, Z]$) is observed. The observations are created by adding to the true trajectory independent perturbations drawn from a white Gaussian noise with standard deviation $\sigma_o = 2$ as in Harlim and Hunt (2007) and Bocquet (2011) . Three different experiments are conducted for different observation time intervals : $\Delta t = 0.10$, $\Delta t = 0.25$ and $\Delta t = 0.50$. These observation time intervals are expected to provide mild, medium and strong nonlinear test cases (Bocquet, 2011).

Each experiment is run over 10^5 assimilation cycles. To avoid any spin-up issues, a burn-in period of 1000 analysis cycles is used. Five filters are compared : the stochastic EnKF, the RHF, a particle filter, the MRHF and the MRHF with mean-field approximation (see Section 4.3). The EnKF and the RHF are tested with a large set of inflation factors and the best in terms of root mean square error are retained for comparisons. The particle filter is implemented in its Sequential Importance Resampling (SIR particle filter) version (Gordon et al., 1993). Resampling is performed using the universal resampling method described by Whitley (1994). After resampling, the particles are perturbed with a white Gaussian noise with variance selected in the $[0.01, 0.05]$ interval to provide the smallest errors.

The filters are tested for different ensemble sizes : $N_e = [8, 16, 32, 64, 128, 256, 512]$. The filters are first evaluated by the time-averaged value of the root mean square error (RMS error) between the analysis and the simulation of reference. We also evaluate the filters' approximation of the full posterior distribution using the Kullback-Leibler (KL) divergence, or relative entropy (Kullback, 1959), which measures a distance between two probability densities P and Q according to the formula :

$$KL(P, Q) = \int \log \frac{P}{Q} dP. \tag{4.8}$$

Density P refers to the density described by the ensemble after an analysis step. Ideally, the reference density Q should be the analytical solution to the problem. Since this is not available to us, we take the SIR particle filter solution with 2048 particles as a reference. It proves to be the best of all in terms of RMS error, as it will be shown in the results section. Also, by construction, the SIR particle filter provides a physically balanced solution, assuming that the noise added to each particle during the resampling step is prescribed small enough not to affect this balance significantly. For a given assimilation method, a small KL divergence guarantees that the solution is physically balanced. For a perfect computation of the KL divergence, the joint probability densities should be used. To limit the problematic effects of subsampling, especially when the ensemble sizes are small, we stick to the marginal densities of the first Lorenz variable, X. Using Y or Z provides very similar results. The marginal densities of X are recovered from the updated ensembles of each filter by using rank histograms.

Results

When $\Delta t = 0.10$, Figure 4.7, left side, upper panel, shows that the EnKF outperforms all the other methods for $N_e \leq 128$. This is because the system is fully and accurately observed, and frequently enough to make the analysis problem close to Gaussian. However with $N_e \geq 256$, $\Delta t = 0.10$, the fully nonlinear methods perform

slightly better. In a rather similar setup, but with an Ensemble Transform Kalman filter instead of a stochastic EnKF, Bocquet (2011) also concludes that the Kalman filter, perfectly designed for such problems, is extremely hard to beat. Nonetheless, the RHF and MRHFs behave rather well, even if not as well as the EnKF. In a medium nonlinear case ($\Delta t = 0.25$, central panel of Figure 4.7, left side) and for $N_e \geq 32$, both the MRHF in its full formulation and the mean-field approximated MRHF produce a smaller RMS error than the EnKF and the RHF. The SIR particle filter needs 128 particles to perform as well as the MRHFs.

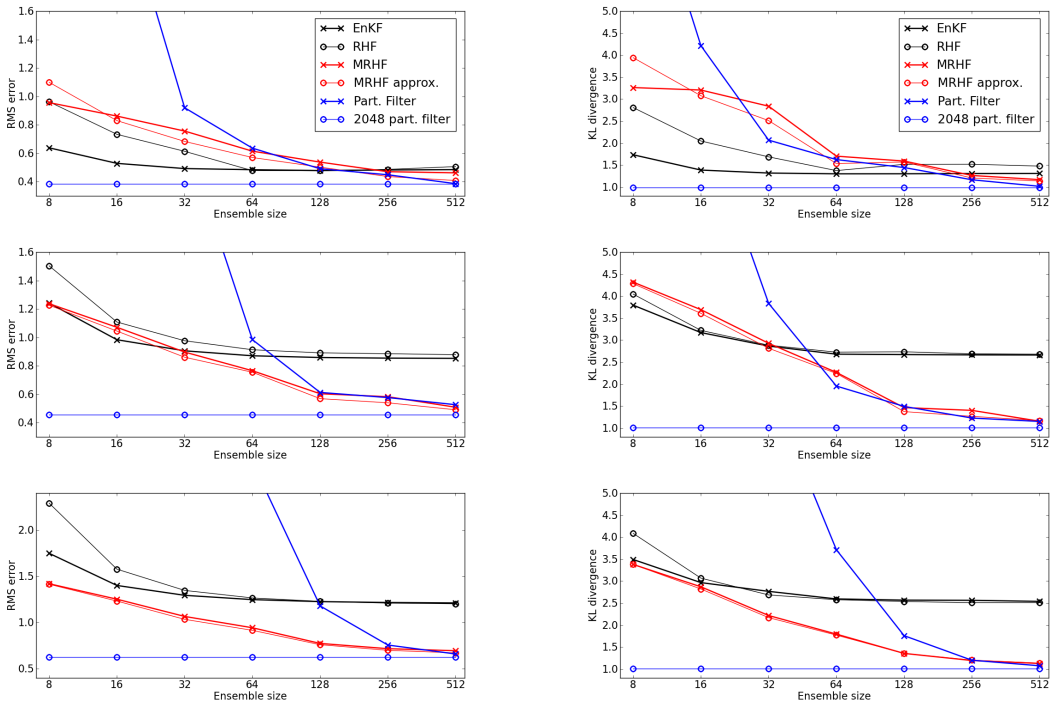

Figure 4.7 – *Time-averaged analysis rmse (left side) and mean Kullback-Leibler divergence on the X-variable (right side) for the EnKF (thick black line with crosses), the RHF (thin black line with open circles), the SIR particle filter (thick blue line with crosses), the full MRHF (thick red line with crosses) and the mean-field approximated MRHF (thin red line with open circles) ; for experiments on the fully observed Lorenz 63 with observation time intervals $\Delta t = 0.10$ (upper-panel), $\Delta t = 0.25$ (center-panel) and $\Delta t = 0.50$ (bottom-panel). The thin blue lines with open circles represents the scores for the SIR particle filter with 2048 particles, which can be considered as a target score.*

Finally, in a case of strong non-linearities (Fig. 4.7, left side, bottom-panel), the MRHFs outperforms the EnKF and the RHF for any ensemble size. The SIR particle filter needs in this case more than 256 particles in order to achieve similar performance. In all cases, the MRHF and the MRHF with mean-field approximation

behave very similarly, the latter being even slightly better most of the time. This suggests that the mean-field approximation has a small negative impact, and that the dimensionality issue that may affect the MRHF, as described in Section 4.3.3, is already present in a 3-variable system.

Figure 4.7, right side, shows the counterpart of Fig. 4.7, left side, for the KL divergence for the X-variable. The KL divergences are computed with respect to a SIR particle filter solution with 2048 particles. It is remarkable that other, independent 2048-SIR particle filter solutions do not provide null KL divergences. This is because the random perturbations introduced after resampling are different in the test and reference experiments. In all observation scenarios, this KL divergence approaches 1 ; this can then be considered as the target score for the other methods. The MRHFs perform very well, even in the mildly nonlinear case, with large ensembles. As nonlinearities go stronger, they perform increasingly well in comparison with the others. In particular, the ensemble size required by the SIR particle filter to reach the performance of the MRHFs increases dramatically. In the strongly nonlinear case (bottom panel), the MRHFs perform better than the EnKF and RHF for any ensemble size and are only outperformed by the SIR particle filter for very large ensemble sizes ($N_e \geq 256$). In any case, the SIR particle filter performs better than the others for large ensembles.

4.4.2 Bimodal case – Z observed

Experimental set-up

The L63 attractor is characterized by two lobes centered on points of attraction and connected to each other at their bottom (where the minimal values of z are encountered). Figure 4.8 displays horizontal slices through the L63 attractor represented in its phase space. The two lobes are easily identified in the region $Z > 24$ (bottom row), exhibiting 2 or 4 distinct modes. In a data assimilation framework without any prior information other than the whole attractor itself, a single observation of Z does not enable us to determine the mode where the truth actually is. The dynamics often help to determine whether it is in an ascending branch or a descending branch of the attractor so we expect bimodal posteriors. It is thus a strongly non Gaussian data assimilation problem.

In the following experiment, observations of Z are extracted from a "true" trajectory, perturbed with a white Gaussian noise of variance 1, and assimilated every 40 time steps ($\Delta t = 0.40$). The assimilation is conducted over 10^5 analysis cycles after a burn-in period of 1000 time steps.

The evaluation of the MRHF performance is strictly similar to the previous experiments, except that the reference solution to compute the KL divergence comes from

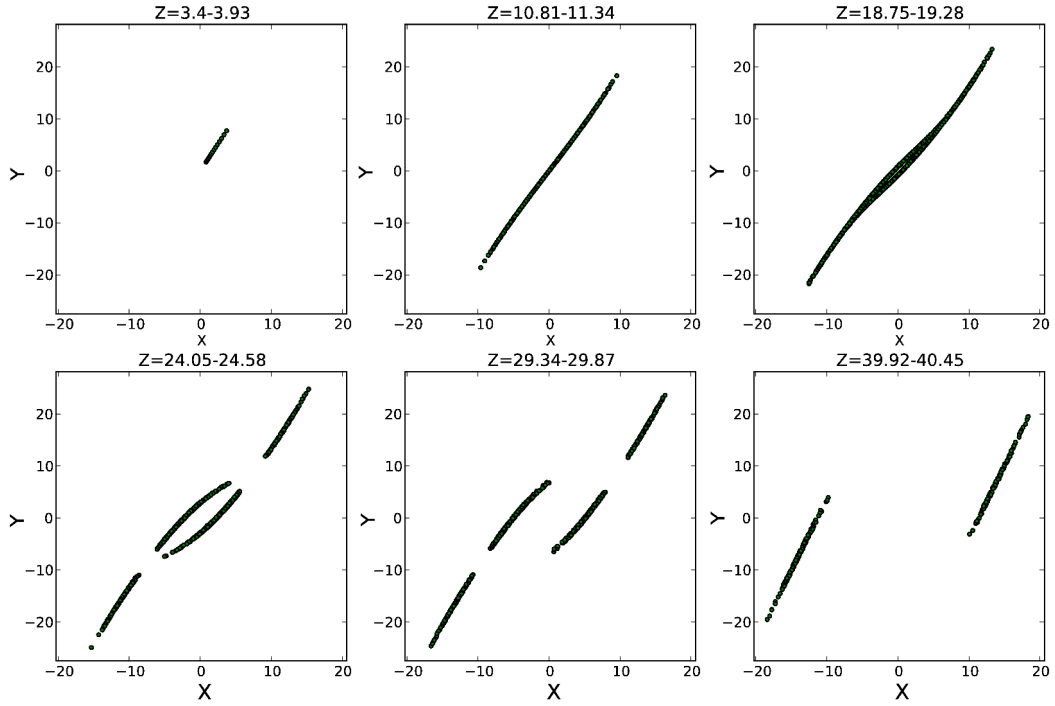

Figure ***4.8*** – *Horizontal slices (in $X-Y$ planes, Z intervals indicated on tops) through L63 attractor. In the two bottom left graphs, the two modes in the centre are parts of the descending branches of the lobes (Z decreases with time), while the modes in the corners are the ascending branches.*

the SIR particle filter with 4096 particles, instead of 2048. As it is argued in Section 4.4.2 below, the RMS error is not a meaningful diagnostic in this case, making it difficult to verify that the 4096-SIR particle filter provides an accurate solution. However, the objective of that test relies on the fact that an appropriate data assimilation method should be able to maintain the representation of the bimodality in this particular case. The SIR particle filter, given a substantial number of particles, might generate overly dispersive ensembles but does maintain the bimodality (Figure 4.9).

It has been checked that it is true for the whole integration period. Hence, a small KL divergence between the methods and the 4096-SIR particle filter will confirm that the bimodality is respected.

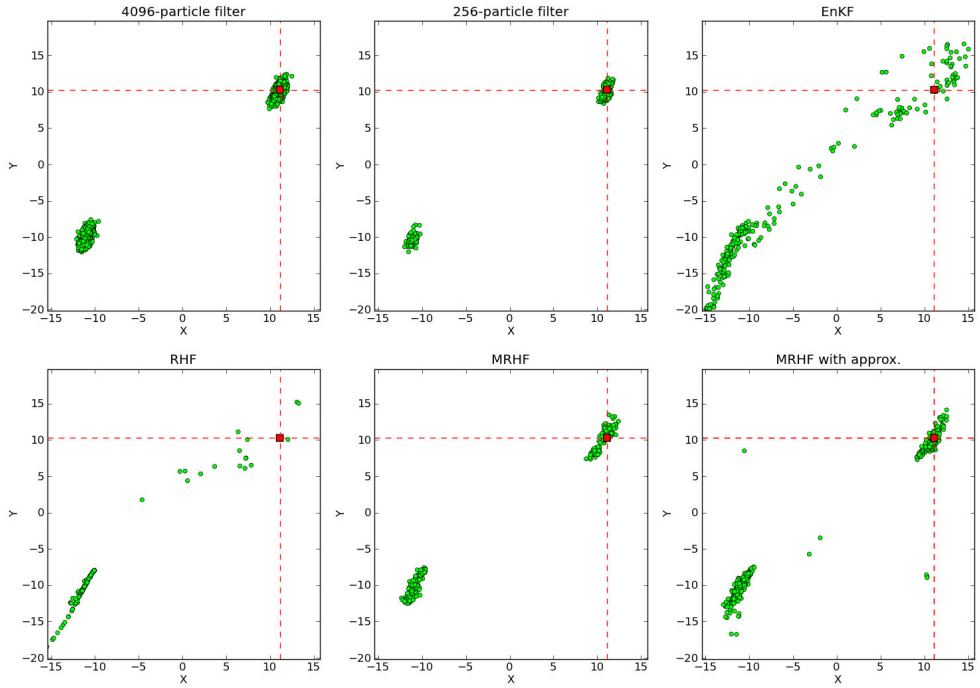

Figure *4.9* – *Scatterplots of the analysis ensembles with* 256 *particles (except upper-left) at the 389th analysis cycle, in the X-Y plane of Lorenz 63 model phase space : (Upper left)* 4096*-SIR particle filter (considered as the reference solution); (Upper center) Particle filter; (Upper right) EnKF; (Bottom left) RHF; (Bottom center) MRHF; (Bottom right) MRHF with mean-field approximation. The red dashed lines indicate the truth from which the observation of Z is produced.*

Results

The time-averaged RMS error is a classical performance diagnostic in data assimilation. However, when the solution is bimodal, the RMS error does not tell much. Here, the RMS errors vary between 3 and 7, whatever the method and the ensemble size are, and they do not decrease with increasing ensemble sizes. Since they provide no relevant information, the RMS errors are not discussed further.

The KL divergence is a much more appropriate diagnostic in this case, especially using the marginal densities of X, since bimodality is expected in the X direction. Figure 4.10 displays the KL divergences as a function of ensemble size, for the 5 methods considered in the previous setup.

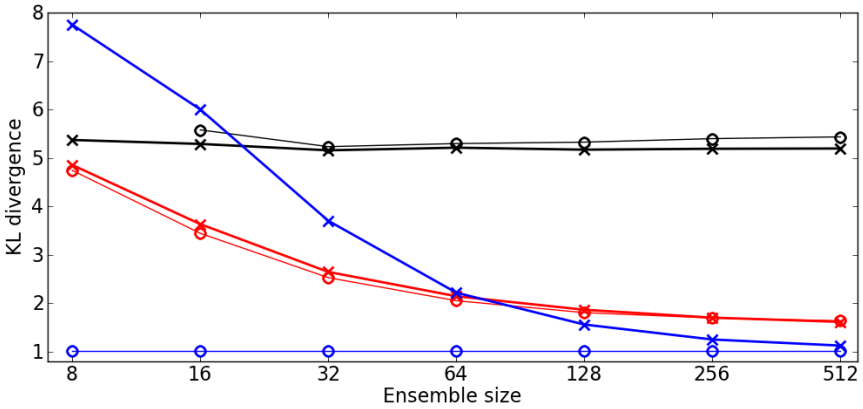

Figure *4.10* – *Time-averaged Kullback-Leibler divergence on the X-variable for the EnKF (thick black line with crosses), the RHF (thin black line with open circles), the SIR particle filter (thick blue line with crosses), the full MRHF (thick red line with crosses) and the mean-field approximated MRHF (thin red line with open circles); for the experiment on Lorenz 63 when only the third variable (Z) is observed every* 40 *time steps during* 10^5 *analysis cycles. The thin blue line with open circles represents the time-averaged Kullback-Leibler divergence for the SIR particle filter with 2048 particles, which can be considered as a target score.*

The RHF with 8 particles proved to be unstable with any inflation factor. The fully non-parametric methods (SIR particle filter, MRHFs) yield much better scores than the others (EnKF and RHF). The KL divergence of the EnKF and RHF does not depend on the ensemble size, confirming that these methods are not designed to deal with such multimodal problems. The SIR particle filter needs at least $N_e = 128$ particles to obtain a smaller KL divergence than the MRHFs.

To illustrate the differences in the behaviours of the various methods, Fig 4.9 depicts scatterplots of the analysis ensembles at the 389th cycle, given by the reference SIR particle filter (with 4096 particles), and the other 5 methods with 256 particles. The scatterplots are drawn in the $X - Y$ plane, and the true state is shown by the red squares and red dashed lines. The 389th cycle has been chosen arbitrarily for this illustration. Similar behaviours of the filters are observed throughout the entire experiments.

While the correct posterior distribution is bimodal, the EnKF and the RHF tend to create particles between the two modes. This explains their large KL divergences with respect to the reference solution. The SIR particle filter and the MRHF, on the other hand, provide visually correct, bimodal solutions in this case (but they can be subject to mode leakage too, at other cycles). The MRHF with mean-field

approximation is again broadly similar to the full MRHF. Because the correction to Y is independent of the correction to X, however, a few particles can switch from one mode to the other along the X (resp. Y) direction without switching along the Y (resp. X) direction. These particles appear in unrealistic regions of the model phase space (regions that cannot be visited by the attractor, near $X = -10$, $Y = 10$ or $X = 10$, $Y = -10$). This illustrates the limitation of the MRHF with mean-field approximation to deal with bimodal distribution. Mode leakage of this kind is significantly reduced by the deterministic resampling method used in the MRHF, in comparison with a stochastic method (result not shown). This is because the MRHF method, described in Section 4.3.2, preserves the relative rank of particles at the analysis step. Also, the Lorenz 63 system does not seem affected by a few outliers, and the KL divergence scores of the MRHF with mean-field approximation are good.

4.5 An illustration of density estimation

We next consider the analysis step for each scheme in a more complex model, with no forecast loop. The interest of this illustration is to observe their behaviour in a first realistic context.

4.5.1 The marine biogeochemical context

The interactions between ocean dynamics and biogeochemistry are complex. The variations of the Mixed Layer Depth (MLD) strongly influences the nutrient supply, hence the phytoplankton production in the euphotic layer (Dutkiewicz et al., 2001). MLD variations are themselves controlled, at least for a large part, by variations in the wind forcing. With the growing interest in understanding ocean biogeochemical cycles and thanks to the increasing amount of dedicated satellite missions (SeaWiFS, MERIS, MODIS), the assimilation of ocean color data, a proxy of chlorophyll concentration, has taken off in the last few years (Gregg et al., 2009). A better estimation of the dynamical variables, including wind forcing, appears as an interesting by-product. Because of the nonlinear relationships of biogeochemical variables between each other and with dynamical variables, ocean biogeochemical data assimilation is a fundamentally non-Gaussian problem. This is well demonstrated by Béal et al. (2010). They use a three-dimensional coupled physical-biogeochemical model of the North Atlantic with a 1/4°horizontal resolution. A 200-particle ensemble simulation is run, perturbing the wind forcing, during the 1998 North Atlantic spring bloom. The nonlinear model response, expressed in the non-Gaussianity of the bivariate distributions of the ensemble variables, is clearly demonstrated.

4.5.2 The density estimation experiment

The data used for the illustration below comes from the ensemble simulation of Béal et al. (2010). We focus on the 1-day forecast at the Gulf Stream station (47°W/40°N).

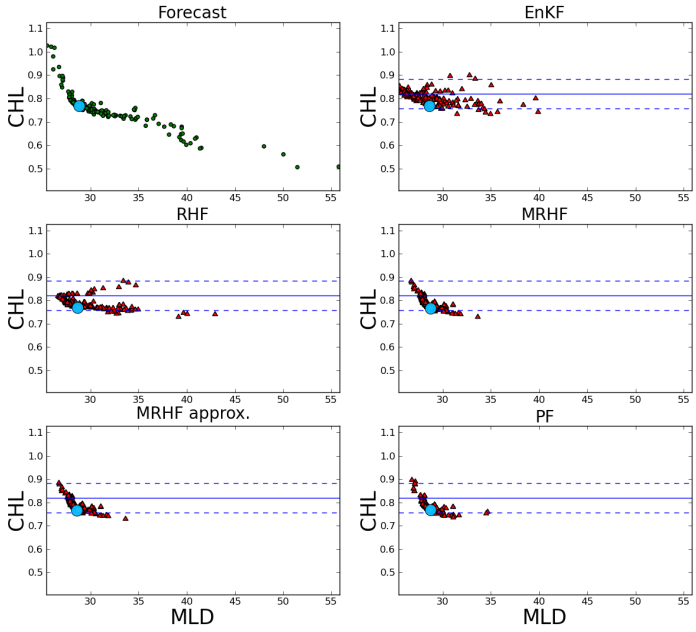

Figure *4.11* – *Prior forecast ensemble (upper-left) and posterior analysis ensembles (red triangles), on the chlorophyll (CHL)/ mixed layer depth (MLD) plane, produced by the EnKF (upper-right), the RHF (center-left), the full MRHF (center-right), the mean-field approximation MRHF (bottom-left) and the SIR particle filter (bottom-right). Only chlorophyll is observed with the value CHL_o indicated by the blue line on each panel. The two dashed blue lines represent the interval $[CHL_o - 2\sigma_o, CHL_o + 2\sigma_o]$, where $\sigma_o^2 = 0.001$ is the observation error standard deviation. The blue dot represents the true state, from which the chlorophyll value has been extracted and perturbed to form the observation CHL_o.*

The upper-left graphs of Fig. 4.11, Fig. 4.12 and Fig. 4.13 respectively show the forecast ensemble in the chlorophyll (CHL)/ mixed layer depth (MLD) plane, the chlorophyll (CHL) / biogeochemical detritus (DET) plane and the biogeochemical detritus (DET) / mixed layer depth (MLD) plane. Each green dot represents an ensemble particle and the blue dot shows the values from a reference model run. A chlorophyll observation is created by perturbing the value from this reference with a

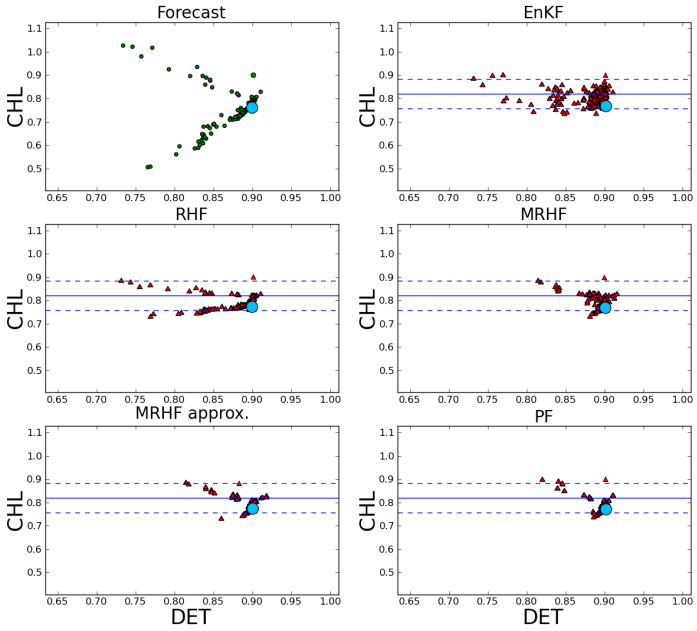

Figure *4.12* – *Same as Fig. 4.11 but on the chlorophyll (CHL)/ biogeochemical detritus (DET) plane.*

Gaussian white noise with variance $\sigma_o = 0.001$.

Each of the assimilation schemes from the previous section is applied, at a fixed time step and a fixed grid point, to the 7-variable control vector. The control vector is composed of seven prognostic variables of the model : one dynamical variable (MLD) and six biological variables (chlorophyll, detritus, dissolved organic matter, NH4, NO3 and phytoplankton). The 200-particle forecast ensemble is used as a prior distribution to test various analysis schemes using the chlorophyll observation. The other panels of Fig. 4.11, Fig. 4.12 and Fig. 4.13 display the estimated ensembles, with red triangles, on respectively the chlorophyll (CHL)/ mixed layer depth (MLD) plane, the chlorophyll (CHL) / biogeochemical detritus (DET) plane and the biogeochemical detritus (DET) / mixed layer depth (MLD) plane, obtained using the EnKF (upper-right graphs), the RHF of Anderson (center-left graphs), the full MRHF (center-right graphs), the mean-field approximation MRHF (bottom-left graphs) and the SIR particle filter (bottom-right graphs) analysis steps. The chlorophyll observation is represented by the blue full line ; the blue dashed lines are at a distance of $2 \times \sigma_o$

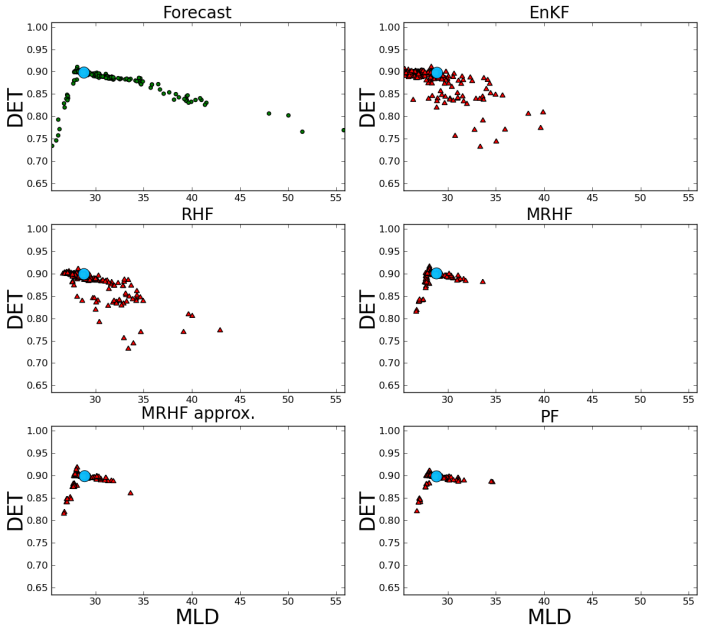

Figure *4.13* – *Same as Fig. 4.11 but on the biogeochemical detritus (DET)/ mixed layer depth (MLD) plane.*

from the blue full line.

A data assimilation method is obviously expected to move the prior particles close to the observations. However, a major requirement for those methods, especially in the non-Gaussian context, is also to maintain the information on the model attractor contained in the shape of the prior ensemble dispersion. In Fig. 4.11 and Fig. 4.12, the prior ensemble clearly says that the observed variable (CHL) and the unobserved variables (resp. MLD and DET) have an obvious statistical connection, but this connection is not linear. Thus, the methods using a linear regression in the physical space to correct the unobserved variable, i.e. EnKF and RHF, fail to maintain the non-Gaussian shape of the prior density. The SIR particle filter analysis step is a very good approximated implementation of Bayes' theorem (modulo sampling errors). Hence, with a sufficiently large ensemble, it provides a posterior ensemble consistent with the observation and the prior ensemble. In comparison the full MRHF analysis step also manages to produce a posterior distribution consistent with both the observation and the prior ensemble. Nevertheless, it clearly appears in Fig. 4.12

that the curse of dimensionality striking during the particle selection process (as discussed in Section 4.3.3) degrades the correction on the second unobserved variable. The mean-field approximation discussed in Section 4.3.3, appears to overcome this issue and produces a posterior distribution very similar to the SIR particle filter.

Fig. 4.13 displays the posterior ensembles in the mixed layer depth (MLD) / biogeochemical detritus (DET) plane. Those graphs allow us to observe the bivariate densities between two unobserved state variables. It is known that the EnKF and the RHF appropriately maintain the covariances between all pair of variables in a linear and Gaussian context. However, in a non-Gaussian case such as this one, Fig. 4.13 shows that both filters do not provide an appropriate ensemble update (in the sense of Bayes' rule). Meanwhile, the MRHF and the SIR particle filter, provide an estimated ensemble taking into accout the relationship between all pair of variables by respecting the information contained in the prior distribution.

4.6 Discussion and Conclusions

This paper has introduce the Mutlivariate Rank Histogram Filter (MRHF), a fully non-Gaussian analysis scheme for ensemble data assimilation. This filter is an extension of the Rank Histogram Filter introduced by Anderson (2010). In order to set the MRHF in the wider context of non-Gaussian analysis methods, the behaviour of some of these methods has been examined in idealized, bivariate frameworks where the variables were jointly Gaussian, weakly non-Gaussian, or strongly non-Gaussian (Section 4.2). The MRHF clearly falls into the category of methods able to deal with strong non-Gaussianity, along with the particle filters. An approximated version of the MRHF, based on the mean-field approximation (Cotter and Reich, 2013), is also proposed.

Numerical experiments with the Lorenz 1963 model, in a data assimilation problem with fully observed state vector but for different observation time intervals corresponding to different levels of nonlinearity (Section 4.4.1), confirm that the MRHF does not perform better than the EnKF in quasi-linear context with small ensemble sizes. Nevertheless, in this context, the MRHF performs slightly better with larger ensembles ($N_e \geq 256$). When nonlinearity is stronger, the MRHF considerably reduces the root mean square error and the Kullback-Leibler divergence in comparison with other methods when given a sufficiently large number of particles ($N_e \geq 64$). The MRHF with mean-field approximation exhibits very similar performance. Experiments in the most nonlinear regime, characterized by the bimodality of the state density, confirm the ability of the MRHF to handle strong non-Gaussianities (Section 4.4.2). Finally, an experiment with prior data from a coupled physical-biogeochemical model (Section 4.5) illustrates the behaviour of the MRHF analysis (in its full and

approximated forms) when facing strongly non-Gaussian densities in a more explicitly geophysical context. This illustration is not a full data assimilation problem but the posterior densities produced by the MRHF analysis show that Bayes' theorem is correctly approximated in this formulation.

Aside from documenting the performance of the MRHF, the experiments in this paper show the importance of matching the assimilation method to the level of non-Gaussianity in the problem at hand. Even though a general method such as the MRHF should perform well in any situation, the fact is that the EnKF (or other linear methods) are perfectly adequate in many applications and often much less expensive computationally. An advantage of the MRHF in this respect is that it is easily hybridized with other serial transform methods, such as the EnKF, because such schemes process observations serially and update observed and unobserved variables serially. One can then consider, for example, updating wind components with the EnKF but the more non-Gaussian humidity with the MRHF.

Relative to Kalman filters, the MRHF's deterministic scheme has assets to preserve physical balances during the analysis. Relative to particle filters, the MRHF also has advantages in addressing the curse of dimensionality. As explained in Section 4.3.5, spatial localization of the update step comes naturally with the MRHF, although it has not been tested here. Also, effects of the curse of dimensionality on the MRHF can be greatly reduced using the mean-field approximation, as illustrated in Section 4.3.3. In fact, use of this approximation is likely necessary for successful implementation of the MRHF in many realistic problems. Finally, it is relatively easy to reduce any sensitivity of the MRHF to biases in observed variables, and to decrease the associated tendency for filter divergence, because of the freedom to choose the tails of the prior density for these variables (Section 4.2). Then, contrary to the particle filters (at least in their bootstrap formulation, Gordon et al., 1993), a large number of observations helps in avoiding divergence.

According to our experiments, the MRHF is much more expensive than the EnKF in terms of computation. The MRHF analysis proves to be approximately 50, 100, and 200 times more expensive than the EnKF for typical ensemble sizes of 64, 128, and 256, respectively. The MRHF might then be best suited for rather specific problems with strong non-Gaussianity and few observations or as part of hybrid schemes where the MRHF is used only for certain variables. Nevertheless, spatial localization and a more efficient implementation of the mean field approximation have the potential to greatly reduce computation cost and expand the range of problems for which the MRHF is feasible. Both aspects need to be investigated before the application of the MRHF to realistic problems, and these are the next steps of this work.

Acknowledgements

This work was supported by the Région Rhônes-Alpes and NCAR. It is also a contribution to the SANGOMA project supported by European Comissions Seventh Framework Programme FP7/2007-2013, Grant agreement no 283580, and to the CNRS/INSU/LEFE program. The authors are grateful to Sebastian Reich and two other anonymous reviewers for their relevant and constructive comments.

Chapitre 5

Un cadre de dynamique océanique et de biogéochimie marine

Sommaire

Ce chapitre présente et valide le cadre des expériences d'assimilation mises en place dans les chapitres 6 et 7.

La première section (Sec. 5.1) décrit le modèle, après avoir évoqué les motivations entourant le choix du couplage dynamique océanique et biogéochimie marine comme cas d'étude. La section 5.2 fournit les caractéristiques des problèmes d'assimilation de données étudiés dans le chapitre 6 et le chapitre 7. Les deux sections suivantes examinent le comportement de l'ensemble sans assimilation en le comparant, dans un premier temps, à la *vérité* et aux observations avec un regard statistique (Sec. 5.3) puis en étudiant, dans un second temps, l'évolution des incertitudes dans le modèle avec un regard physique (Sec. 5.4). Enfin, les deux dernières sections proposent un bilan du chapitre et un bref préambule introduisant les Chapitres 6 et 7.

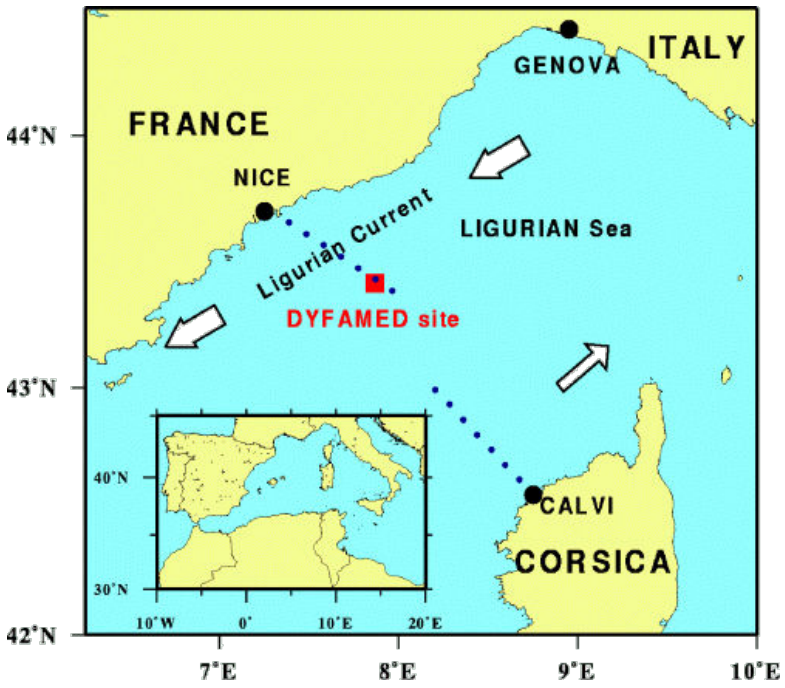

Figure ***5.1*** – *Localisation du site DYFAMED. Image courtoisie du programme DYFAMED (www.obs-vlfr.fr/cd_rom_dmtt/sodyf_main.htm).*

5.1 Le modèle ModECOGeL

Le modèle utilisé pour les expériences des chapitres 6 et 7 est un modèle à une dimension verticale, MODèle d'ECOsystème du G.H.E.R.[1] et du L.O.V.[2] (ModECOGeL), couplant un modèle d'écosystème marin à un modèle de couche de mélange océanique de la mer Ligure (site DYFAMED[3], 5.1). Ce modèle a été développé de manière à représenter les processus biogéochimiques particuliers à cette région. Le modèle ModECOGeL a été mis en oeuvre dans plusieurs études antérieures du site DYFAMED (Lacroix, 1998; Lacroix and Grégoire, 2002; Magri et al., 2005; Raick et al., 2007; Lenartz et al., 2007; Tissot, 2012).

5.1.1 Motivations

L'étude préliminaire présentée au Chapitre 3, nous a permis de mettre en évidence l'importance de la compréhension et de la prise en compte des non-linéarités et des non-Gaussianités d'un modèle pour la mise en place d'une assimilation de données performante. De plus, une méthode complexe d'assimilation non-Gaussienne, telle que le MRHF, a été décrite puis évaluée sur un modèle jouet dans le Chapitre 4.

Ces deux précédents chapitres, nous amènent à orienter nos réflexions et à étendre nos expériences vers un cadre plus réaliste. Pour la suite des travaux relatés dans cette thèse, le cas d'un système couplé de dynamique et de biogéochimie marine en une dimension verticale a été choisi. Les raisons de ce choix sont multiples :

- Le couplage de la dynamique océanique et de la biogéochimie marine présente des non-linéarités particuliérement fortes avec des effets de seuils importants. Ces non-linéarités génèrent des probabilités très non-Gaussiennes. Ce domaine d'étude a d'ailleurs souvent été le choix des travaux sur l'assimilation de données non-Gaussienne (Bertino et al., 2003; Béal et al., 2010).
- Comme mentionné dans le chapitre de contextualisation (Chap. 1), le problème d'assimilation de systèmes couplés représente en soi un vrai défi. La question principale étant la propagation des corrections d'une partie du système couplé vers l'autre. De plus, le couplage (comprenant une action de la dynamique sur la biogéochimie ainsi qu'une rétroaction de la biogéochimie sur la dynamique) produit souvent des relations inter-variables complexes et non-linéaires.
- L'utilisation d'un modèle réaliste propose des défis nouveaux par rapport aux simples modèles jouets. D'un point de vue numérique, le coût calcul est non-négligeable et donne à réfléchir différemment certaines méthodes. D'un point de vue mathématique, la dimension d'un tel modèle met en évidence des problèmes

1. GeoHydrodynamics and Environment Research
2. Laboratoire d'Océanographie de Villefranche
3. Site d'observations dans la zone centrale de la Mer Ligure

fondamentaux comme la malédiction de la dimensionalité.
– Les considérations physiques qui découlent de ce cas d'étude présentent également de nombreux intérêts en répondant à certaines questions : Une meilleure estimation du bloom de phytoplancton est-elle possible par observation de la dynamique du système ? Quel impact peuvent avoir les observations de la couleur de l'eau sur la dynamique verticale ?
– Enfin le choix de nous placer dans un cadre d'expériences jumelles nous permet une meilleure latitude quant à la variété des questions traitées.

Ainsi, le choix de la mise en place d'expériences jumelles sur un modèle couplé dynamique océanique et biogéochimie marine a été fait. Cette section (Sec. 5.1) propose une description du modèle ModECOGeL utilisé par la suite. Les deux sections suivantes décrivent la mise en place de deux types d'expériences jumelles basés sur ce modèle.

5.1.2 Descriptions physique et numérique du modèle

Description physique

Le modèle se compose d'un système décrivant la dynamique d'une colonne d'eau verticale et d'un système de biogéochime marine. Ces deux systèmes sont couplés par des actions de la dynamique sur la biogéochimie et par des rétroactions de la biogéochimie sur la dynamique. Ces relations sont décrites dans cette sous partie.

L'hydrodynamique verticale L'hydrodynamique du système est décrite par un modèle aux équations primitives, non-linéaire, barocline et utilisant un schéma de fermeture turbulente k–l (Nihoul and Djenidi, 1987). La dynamique est donc régie par cinq variables pronostiques : la vitesse zonale u $(m.s^{-1})$, la vitesse méridienne v $(m.s^{-1})$, la température T (^oC) et la salinité S (psu) ; ainsi qu'avec une cinquième variable : l'énergie cinétique turbulente k $(m.s^{-2})$ permettant la fermeture turbulente.

Les équations du modèle sont les équations primitives simplifiées par l'hypothèse d'homogénéité horizontale pour se ramener en une dimension verticale :

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial z}(\nu \frac{\partial u}{\partial z}) + fv \tag{5.1}$$
$$\frac{\partial v}{\partial t} = \frac{\partial}{\partial z}(\nu \frac{\partial v}{\partial z}) - fu \tag{5.2}$$
$$\frac{\partial T}{\partial t} = \frac{\partial}{\partial z}(\tilde{\lambda^T} \frac{\partial T}{\partial z}) + Q(T) \tag{5.3}$$
$$\frac{\partial S}{\partial t} = \frac{\partial}{\partial z}(\tilde{\lambda^S} \frac{\partial S}{\partial z}) \tag{5.4}$$
$$\frac{\partial k}{\partial t} = \nu \| \frac{\partial \mathbf{u}}{\partial z} \|^2 (1 - R_f) - \epsilon + \frac{\partial}{\partial z}(\nu \frac{\partial k}{\partial z}) \tag{5.5}$$

Les équations (5.1) et (5.2) sont les équations de mouvements horizontaux où f est le facteur de Coriolis et ν est le coefficient de viscosité turbulente ($\nu = 0.5l\sqrt{k}$ avec l la profondeur de couche de mélange). L'équation (5.3) est l'équation de bilan thermique avec $\tilde{\lambda^T}$, la diffusion turbulente de la température et $Q(T)$, le terme de forçage à la surface. L'équation (5.4) est l'équation de bilan halin où $\tilde{\lambda^S}$ est la diffusion turbulente de la salinité. Et l'équation (5.5) est l'équation d'énergie cinétique turbulente où $\mathbf{u} = (u, v)$ et avec R_f le nombre de Richardson flux (mesurant l'importance relative de la production et destruction de l'énergie cinétique turbulente) et ϵ le taux de dissipation de l'énergie cinétique turbulente ($\epsilon = \frac{k^2}{16\nu}$).

Pour ce modèle à une dimension verticale, les échanges à la surface sont réduits à leurs composantes thermiques et mécaniques. Ceci inclut le stress de vent et les flux de chaleur. Les forçages atmosphériques sont décrits avec plus de détails par la suite dans la sous-section éponyme.

La biogéochimie Le modèle biogéochimique simule un écosystème défini par le cycle de l'azote. Il est composé de 12 variables d'état biogéochimiques (concentrations) : le nitrate (NO3), l'ammonium (NH4), trois classes de tailles phytoplanctoniques (PicP, NanP, MicP), trois classes de tailles zooplanctoniques (NanZ, MicZ, MesZ), deux classes de tailles de matière organique particulaire (POM1, POM2), l'azote organique dissous (DON) et les bactéries (BAC).

L'équation générale d'évolution d'une concentration biogéochimique c_i s'écrit :

$$\frac{\partial c_i}{\partial t} = \frac{\partial}{\partial z}(\tilde{\lambda} \frac{\partial c_i}{\partial z}) + S(c_i, c_j, \alpha), \tag{5.6}$$

où l'évolution temporelle de la concentration (membre de gauche) dépend du gradient vertical de la diffusivité verticale et d'un terme $Source - Puit$ qui est fonction

de certains paramètres α, ainsi que de la relation entre la concentration c_i et les autres concentrations biogéochimiques c_j.

La figure 5.2, issue de Magri et al. (2005), schématise les relations entre les variables d'état biogéochimiques et décrit les processus impliqués. L'objectif initial de ce modèle était de permettre l'étude du rôle de l'hydrodynamique (principalement l'évolution de la couche de mélange) sur la répartition verticale des nutriments et sur la production primaire (le "bloom").

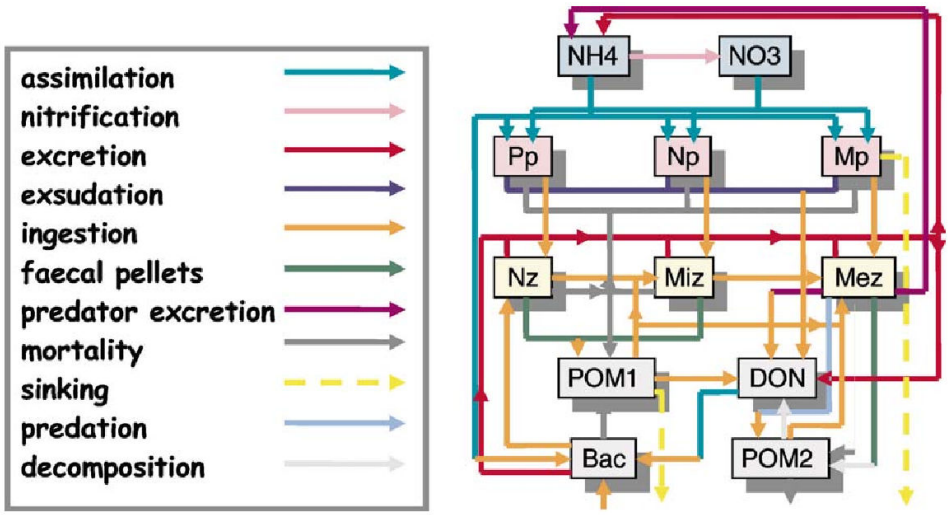

Figure 5.2 – *Schèma issu de Magri et al. (2005) : Représentation schématique du modèle biogéochimique ModECOGeL utilisé dans les simulations couplées ; les 12 compartiments représentent les variables d'état biogéochimiques suivantes : nitrate et ammonium (NO3, NH4) ; pico- (Pp), nano- (Np) et micro-phytoplancton (Mp) ; nano- (Nz), micro- (Miz) et méso-zooplancton (Mez) ; matière organique particulaire (POM1, POM2) ; azote organique dissout (DON) ; bactéries (Bac). Les processus biogéochimiques responsables des flux entre les compartiments sont décrits à gauche.*

Le couplage, actions et rétroactions Bien comprendre les actions et rétroactions du modèle nous permet, entre autres, d'anticiper les degrés de contrôle et d'observabilité intervenant lors de l'assimilation de données.

L'action de la dynamique sur la biogéochimie se fait en forçant le modèle biogéochimique par les profils verticaux de température et de coefficients de diffusivité turbulente ($\tilde{\lambda}$) obtenus avec le modèle hydrodynamique (Figure 5.3). Cette information permet notamment de présupposer qu'un bon contrôle (par l'assimilation de données)

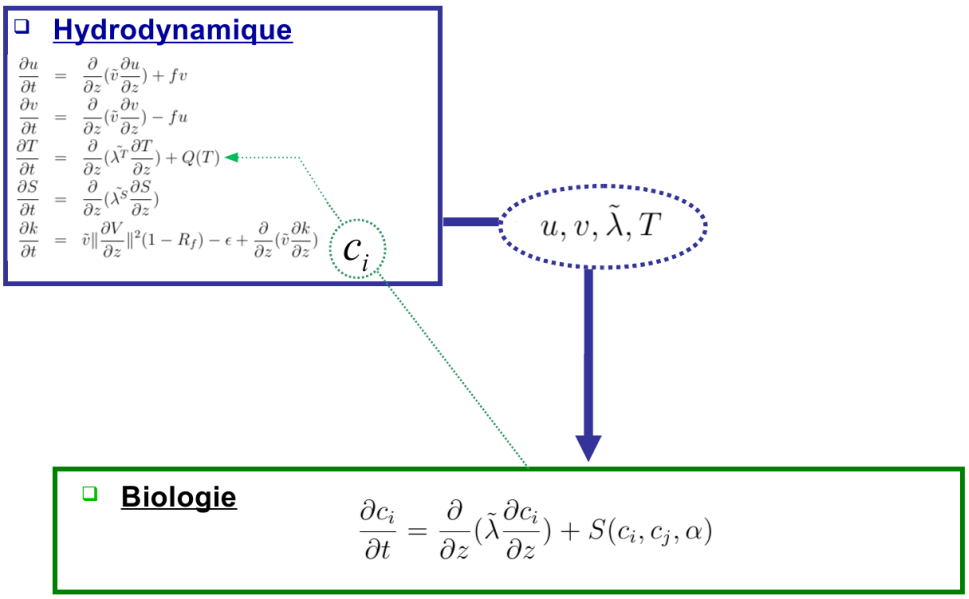

Figure *5.3* – *Les équations du modèle dynamique et du modèle biogéochimique ainsi que leurs intéractions dans MODECOGel.*

de ces deux dernières quantités permettra de maîtriser indirectement la réponse biogéochimique.

Il est également à noter qu'aucune source de nutriments externe au système n'est prise en compte. De même, l'écosystème benthique (cosytème des grandes profondeurs) n'est pas pris en compte et les flux biogéochimiques sols-océan sont mis à zéro. Ainsi, l'azote du système est conservé tout au long d'une simulation.

Un des facteurs clefs dirigeant la biogéochimie est la pénétration de la lumière dans l'eau. La profondeur de la couche d'eau à travers laquelle une quantité suffisante de lumière pénétre pour déclencher la photosynthèse du phytoplancton, va avoir un impact décisif sur le fonctionnement de l'écosystème. Cette couche s'appelle la couche euphotique. Pour rendre la diminution de la lumière avec la profondeur plus réaliste, ModECOGeL utilise le coefficient d'extinction de la lumière issu de mesures, adapté à la mer Ligure et fourni par Béthoux (Ivanoff, 1977). Il est à noter qu'il est commun de diagnostiquer la couche euphotique comme étant la couche à la surface de l'océan absorbant 99% du rayonnement solaire.

Ainsi, dans le couplage ModECOGeL, la rétroaction de la biogéochimie sur la dynamique se fait via la quantité de phytoplancton rencontrée dans les premiers mètres de la colonne d'eau. Cette quantité va modifier la pénétration de la lumière

(self-shading). Ce phénomène est décrit par (Riley (1956)) et s'observe notamment lors de la remontée de la profondeur de la couche euphotique peu de temps après l'apparition du bloom de phytoplancton. Cette modification intervient ensuite sur les flux de chaleur dus à la radiation solaire. Ce qui impacte directement la température de surface puis indirectement, par mélange et advection, les profils de température et donc *in fine* tout le système physique.

Les forçages atmosphériques

Les forçages atmosphériques jouent un rôle important dans les expériences qui suivent puisque les ensembles de simulations étudiés sont générés par perturbation du forçage (Section 5.2). Il semble donc naturel de leur consacrer quelques mots.

Le système de forçages de surface, c'est à dire à l'intéraction océan-atmosphère, est constitué de huit composantes :
- l'insolation,
- l'intensité du vent,
- la direction du vent,
- la température de l'air,
- l'humidité de l'air,
- la nébulosité,
- les précipitations,
- la pression atmosphérique.

À l'origine, le modèle a été calibré et validé pour des forçages couvrant la période allant de l'année 1984 à l'année 1988 (Lacroix and Grégoire, 2002). Pour l'année 2006 (notre période d'étude), les forçages ont été adaptés puis validés par Maëva Doron dans le cadre du projet ANR BIOCAREX.

Ces nouveaux forçages proviennent de différentes sources que l'on décrit brièvement ci-dessous.

Bouée *Météo-France*, en fréquence horaire Les forçages :
- intensité du vent,
- direction du vent,
- pression atmosphérique,
- température de l'air,
- humidité specifique,

sont issus de données directement prises de la *Bouée Météo-France* de la *Côte d'Azur*. Plus précisemment, les données ont originellement une résolution de 15 minutes, mais elles ne sont récupérées que toutes les heures.

SATMOS-MF (J.-P. Olry, *Météosat*), en fréquence horaire Le forçage :

– flux solaire descendant
est issu de données fournies par *MétéoSat*. Les données sont en $W.m^{-2}$.

MOOSE (C. Guieu, L. Coppola), en fréquence journalière Le forçage :
– précipitations
a été crée à partir de données issues de la station du Cap Ferrat. D'autres données sont également disponibles pour la station ERSA (localisée au Cap Corse, Corse). Les données que l'on considère annoncent des précipitations annuelles de 615.4 mm pour l'année 2006 et de 277.8 mm pour l'année 2007.

Climatologie 1984-1988 Le forçage :
– nébulosité (cloud)
a été calculé à partir des flux "longwaves" (ondes longues) des données *MétéoSat*.

Le rôle du vent

Le vent impacte les conditions aux limites dynamiques à travers deux forçages atmosphériques : l'intensité et la direction du vent. Le vent ($\mathbf{V}_{wind} = (u_0, v_0)$) intervient dans les conditions aux limites :
– des vitesses horizontales $\mathbf{u} = (u, v)$ par

$$\nu \frac{\partial \mathbf{u}}{\partial z}\Big|_{Surface} = C_0(1 + 0.1 \|\mathbf{V}_{wind}\|) \|\mathbf{V}_{wind}\| \mathbf{V}_{wind} \tag{5.7}$$

où $C_0 = 0.63 \times 10^{-6}$;
– de l'énergie cinétique turbulente k par

$$\nu \frac{\partial k}{\partial z}\Big|_{Surface} \approx 3C_0 10^{-3} \|\mathbf{V}_{wind}\|^3 . \tag{5.8}$$

Il est à noter qu'en surface, ces trois variables (u,v,k) dépendent du vent de manière non-linéaire (de degré 3). L'impact du vent est ensuite propagé le long de la colonne d'eau par les équations d'états (5.1-5.5).

Considérations numériques

Le modèle simule une colonne d'eau en discrétisant les équations 5.1-5.5 sur une grille verticale. Cette grille est composée d'un point de grille tous les mètres et s'étend sur une profondeur de 400m. La discrétisation temporelle est d'un pas de temps toutes les 6 min.

Les conditions aux limites sont composées des échanges à la surface avec les paramètres atmosphériques de forçage et des conditions de fond.

Les conditions initiales ont été reconstruites à partir de données issues d'une campagne en mer *Boussole*[4] pour les variables dynamiques et d'une campagne en mer *Moose*[5] pour les variables biogéochimiques. Cette initialisation a été faite au 15 décembre 2005.

La version du modèle ModECOGeL utilisée est plus amplement décrite par Lacroix (1998), Lacroix and Nival (1998) puis par Lacroix and Grégoire (2002). Le code numérique initialement en langage Fortran 77 a été réécrit en 2013 en langage Fortran 90 par Jean-Michel Brankart, avec notamment une plus grande stabilité du code et une possibilité de simulation d'ensemble en parallèle.

5.1.3 Limitations du modèle

Bien que dans les expériences qui suivent nous nous plaçons dans le cadre d'expériences jumelles, il est important de connaitre le comportement du modèle et d'identifier ses qualités et ses défauts. Dans cette sous partie, nous présentons d'abord la capacité du modèle à reproduire la réalité, en résumant les travaux de Lacroix and Grégoire (2002) et de Maéva Doron (Gregorio et al., prep). Puis, dans un second temps, nous discutons des incertitudes dans les forçages comme sources d'erreurs potentielles du modèle.

Reproduire la réalité

Dans Lacroix and Grégoire (2002), un premier travail d'évaluation du modèle, par comparaison à des observations des missions FRONTAL[6], a été réalisé sur les années 1994-1998 . Il ressort de ce travail que le modèle parvient à reproduire des phénomènes physiques importants présents dans les observations. Le déclin et l'établissement de la thermocline sont bien représentés par le modèle. La concentration de nitrate dans la couche de surface est également en accord avec les observations. Et la (faible) biomasse phytoplanctonique produite par le modèle est représentative des eaux oligotrophes mesurées par les observations.

En revanche, plusieurs limitations du modèle sont apparues dans cette étude. La température de surface en été est surestimée par rapport aux observations. De même, la profondeur de la thermocline est surestimée. Une première raison à ces surestimations est que le modèle ne simule pas l'advection verticale permanente des eaux froides. La deuxième raison est la possible surestimation des flux de chaleur de

4. Programme Boussole : Réseaux de mouillages et de campagnes mensuelles de mesures sur le site Boussole en mer Ligure

5. Programme Moose : Système d'observation intégré et multidisciplinaire en mer Méditerranée Nord occidentale

6. Programme FRONTAL (1987) : Site-ateliers étudiant les processus frontaux.

forçage issus de mesures effectuées sur le continent et non sur site. Pour la partie biogéochimique, le nitrate simulé par le modèle n'est pas en accord avec les observations dans les régions de sous-surface et dans les couches profondes, le modèle ne prenant pas compte de sources de nitrate importantes révélées par les observations. Ceci engendre une légère sous-estimation des flux annuels de nitrate qui engendrera à son tour une sous-estimation de la production primaire annuelle.

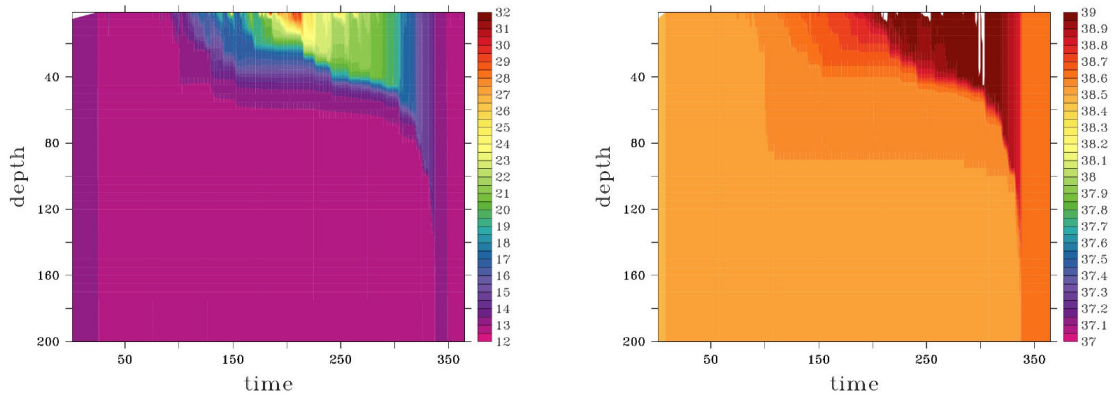

Figure *5.4* – *Graphiques de la température (à gauche, en °C) et de la salinité (à droite, en PSU) pour l'année 2006 sur la grille Temps/Profondeur obtenues avec le modèle ModECOGeL.*

Pour l'année 2006 (année qui nous intéresse), des travaux de calibration et de validation ont été réalisés par Maéva Doron. Ces travaux comprennent, notamment, une comparaison des simulations ModECOGeL avec les données *in situ* DYFAMED. Les figures ci-présentées sont extraites de ces travaux.

La Figure 5.4 présente les graphiques de la température (à gauche) et de la salinité (à droite) pour l'année 2006 sur la grille Temps/Profondeur. L'une des principales conclusions à tirer sur la température et sur la salinité est que la stratification est relativement bien représentée. On constate en revanche une dérive en sel ($\approx$ 0.1 PSU/an).

La Figure 5.5 donne les séries temporelles de la chlorophylle-a de surface des mesures *in situ* DYFAMED (à gauche) et d'une simulation libre ModECOGeL (à droite) sur l'année 2006. Le timing du bloom est bien reproduit (autour du jour 100). Quelques difficultés du modèle à reproduire la biogéochimie sont tout de même à mentionner. Le maximum de chlorophylle est fortement surestimé. On constate, de plus, une extinction totale peu réaliste de chlorophylle en été.

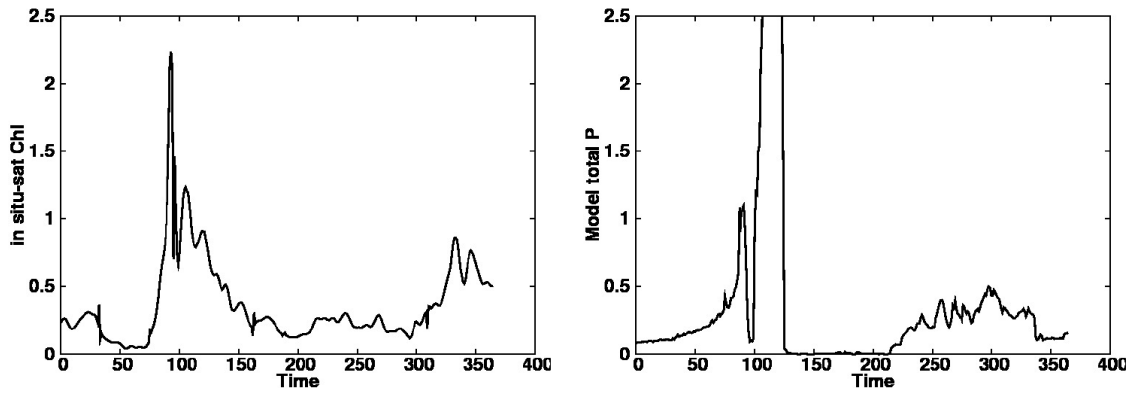

Figure *5.5* – *Séries temporelles de la chlorophylle-a de surface des mesures in situ DYFAMED (à gauche) et d'une simulation libre ModECOGeL (à droite) sur l'année 2006.*

En résumé, le modèle ModECOGeL reproduit la physique d'un écosystème marin relativement cohérente avec les observations. Il présente cependant plusieurs limitations. Dans ce contexte, il est donc intéressant de mieux comprendre et d'améliorer la mise en place d'assimilations de données.

Incertitudes dans les forçages

Le rôle originel de ce modèle est l'étude de la variabilité saisonnière et interannuelle des processus biogéochimiques en lien avec une dynamique de couche de mélange fortement variable. Pour ce modèle, une forte influence de la dynamique du système sur la variabilité saisonnière de l'écosystème a été montrée (Lacroix and Grégoire, 2002). Cette dynamique est principalement contrainte par les forçages atmosphériques réels à haute fréquence. Cette forte contrainte rend le modèle très sensible aux incertitudes sur les forçages.

Lacroix and Grégoire (2002) estiment qu'une variabilité interannuelle de 11.3% dans l'intensité du vent engendre une variabilité de 27.9% de la production primaire, de 18.5% du *ratio-f* (rapport entre la production nouvelle et la production totale primaire) et de 13.4% de la production bactérienne.

Pour illustrer cette forte dépendance à l'intensité du vent, la Figure 5.6 montre deux diagrammes de Hovmöller du phytoplancton sur le mois d'avril 2006. Ces deux diagrammes correspondent à deux simulations ModECOGeL et ne varient que par leurs intensités de vent statistiquement différentes. Ces deux simulations sont deux membres d'un ensemble décrit et étudié par la suite (Sec. 5.2.2).

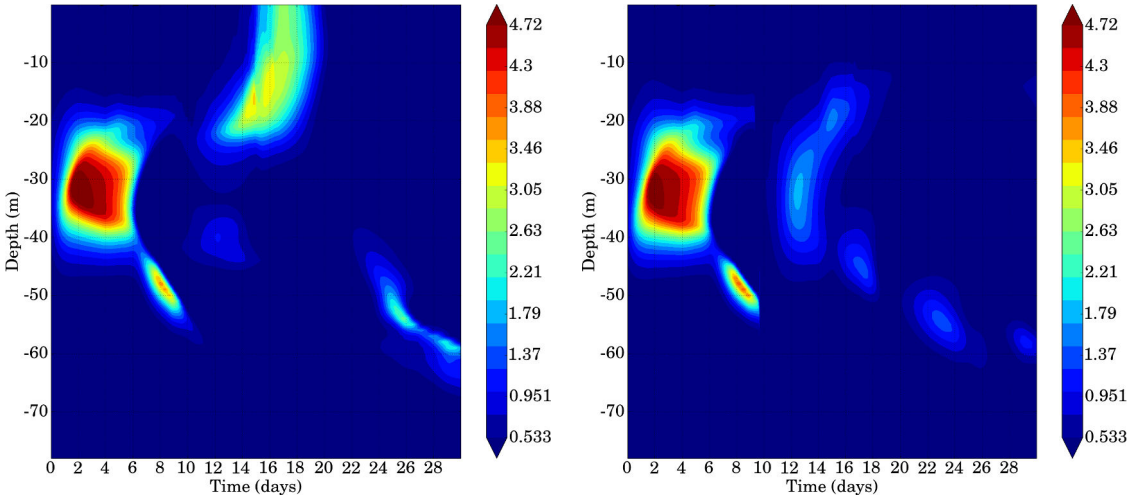

Figure *5.6* – *Diagrammes de Hovmöller, du phytoplancton total sur le mois d'avril 2006, de deux simulations ModECOGeL différant par leurs intensités de vent. Il s'agit des membres 28 et 44 de l'ensemble crée et décrit à la Section 5.2.2.*

Nous constatons déjà, en observant l'évolution du phytoplancton, la forte variabilité (spatio-temporelle) et donc la forte sensibilité du modèle à l'intensité du vent. Les grandes différences entre ces simulations se trouvent principalement dans les vingt derniers jours du mois d'avril. Nous verrons par la suite que le $10^{\text{ème}}$ jour de la simulation d'ensemble occupe une place importante dans l'évolution et la dispersion de l'ensemble.

5.2 Description des expériences d'assimilation

Les expériences décrites dans cette section sont utilisées dans les Chapitres 6 et 7.

Les principaux objectifs de ces expériences s'inscrivent sur plusieurs plans. Ces expériences permettront de comparer différentes méthodes d'assimilation de données en faisant ressortir leurs avantages et leurs inconvénients lorsqu'utilisées pour résoudre différents problèmes d'estimation. De plus, l'étude de l'impact de différents jeux de données fournira une réponse quantitative à l'apport de données de surface en assimilation de données pour l'océanographie. Finalement, d'un point de vue lié aux problématiques de biogéochimie marine, cette expérience permettra d'évaluer l'impact d'une correction des variables dynamiques sur la biogéochimie, ainsi que la possibilité de contrôler la biogéochimie en observant la dynamique.

Cette section présente, en quelque sorte, le plan d'expérience. Dans un premier

temps, nous détaillons la période d'étude et le choix de la simulation de référence (Sec. 5.2.1). Dans la deuxième sous partie (Sec. 5.2.2), nous abordons briévement le principe des paramétrisations stochastiques puis nous décrivons les perturbations d'ensemble utilisées dans nos expériences. La troisième sous partie (Sec. 5.2.3) est consacrée au choix du réseaux d'observations et du vecteur de contrôle mis en place au Chapitre 6 et au Chapitre 7. Enfin, nous étudions et validons les principales composantes des expériences (Sec. 5.3).

5.2.1 Le set-up du modèle

Le principe des expériences jumelles a été décrit dans le chapitre 3. Il est tout de même bon de préciser, que dans un contexte réaliste comme celui qui nous concerne ici, utiliser une expérience jumelle présente certains inconvénients. En effet, en supposant une trajectoire du modèle comme étant la *vérité*, une hypothèse forte est faite.

Hypothèse qu'il faut garder en mémoire et prendre en considération notamment lors de la mise en place de diagnostics. Nous avons vu notamment dans la section 5.1.3 que le modèle peut présenter de nombreuses erreurs et certains biais à la réalité. Ainsi, les diagnostics réalisés n'ont que peu de sens dans l'absolu. Leur utilisation est comparative. Par exemple, les histogrammes de rangs ne peuvent pas être totalement plats. L'objectif est pour un ensemble est donc de présenter un histogramme de rangs amélioré par rapport aux autres ensembles.

Toutefois, pour minimiser l'impact de ces erreurs et de ces biais en amont, il est important de bien calibrer le set-up du modèle pour nos expériences. C'est ce que nous faisons dans cette sous partie.

Période d'étude

Le modèle est préalablement propagé sur un temps d'ajustement (*spinup*) du 1er janvier 2006 au 31 mars 2006. L'expérience d'assimilation se déroule au mois d'avril 2006 (apparition du bloom). Elle débute le 1er avril 2006 et se termine le 30 avril 2006.

Création de la *vérité*

La *vérité* est une trajectoire de référence générée sur cette période. Cette référence est produite avec des forçages atmosphériques dits à basses fréquences (forçages horaires lissés sur une fenêtre glissante de 24h) sauf pour l'intensité du vent qui est maintenue en haute fréquence (forçage horaire). L'intérêt de maintenir les forçages à basse fréquence sauf l'intensité du vent, est de connaitre l'origine de la variabilité

haute fréquence et ainsi de pouvoir suivre la propagation des incertitudes uniquement liées au vent.

5.2.2 Création de l'ensemble

La philosophie des paramétrisations stochastiques

Traditionnellement un modèle représentant la physique océanique se veut déterministe. Cette idée implique que les phénomènes qui ne sont pas résolus par le modèle sont seulement pris en compte à travers leur effet moyen (e.g. la paramétrisation des échelles non-résolues), et que l'incertitude qu'ils génèrent est négligée.

Cette absence de représentation de certains phénomènes peut s'avérer très pénalisante. C'est pourquoi l'intérêt pour la mise en place de paramétrisations stochastiques devient de plus en plus grand. En effet, l'utilisation de paramétrisations stochastiques rend le modèle probabiliste. Ceci permet de prendre en compte les phénomènes non décrits par les équations du modèle, au travers d'une représentation aléatoire de leurs effets.

La représentation des incertitudes par paramétrisations stochastiques a un intérêt pour l'assimilation de données. Comme il a été discuté en Section 2.1.1, l'assimilation ne se contente plus de produire un estimé moyen. Par assimilation, nous souhaitons obtenir une densité de probabilité complète du système. Les paramétrisations stochastiques apportent à la probabilité *a priori*, l'incertitude non prise en compte par une description déterministe.

Pour un développement plus complet du principe des paramétrisations stochastiques nous renvoyons le lecteur aux articles de Palmer et al. (2005), Lermusiaux (2006), Frederiksen et al. (2012) et au manuscrit d'*Habilitation à Diriger la Recherche (HDR)* de Brankart (2014).

Les paramétrisations stochastiques peuvent être introduites à plusieurs niveaux du système : dans la formulation des lois de comportement (par exemple dans l'équation d'état de l'eau de mer, Brankart, 2013 ; ou dans la formulation du modèle d'écosystème, Garnier et al., 2015), dans les paramètres du modèle (Doron et al., 2011; Garnier et al., 2015) ou dans les forçages.

Dans nos expériences, nous générons notre ensemble de simulations en introduisant des paramétrisations stochastiques sur le forçage. Le paramètre de forçage que nous supposons incertain est l'intensité du vent, car c'est le paramètre que nous pensons avoir un effet important sur le comportement de notre modèle couplé (d'après la Section 5.1.3).

Perturber l'intensité du vent

L'incertitude dans le système est introduite en perturbant le forcage de l'intensité du vent à basse fréquence (24h) de la *vérité*. Les perturbations sont engendrées par un processus auto-regressif (processus temporelle aléatoire où le présent dépend du passé) Gaussien d'ordre un, de temps de corrélation $t_{cor} = 2$h (20 pas de temps) et d'écart-type $\sigma = 0.3\text{m.s}^{-}1$. Le temps de corrélation et l'écart-type ont été estimés à partir du vent haute-fréquence réel. Ainsi, cette paramétrisation stochastique permet de simuler la haute fréquence. Cette configuration permet donc d'innoculer dans le système les incertitudes liées à l'intensité du vent (déterminées comme jouant un rôle important dans les limitations du modèle, Sec. 5.1.3), tout en ayant un ensemble de vent estimant l'intensité du vent *vraie*.

Un ensemble de 50 membres est généré à partir du 28 mars 2006 et est propagé sans assimilation sur 4 jours jusqu'au 31 mars 2006. Ce spin-up d'ensemble permet aux perturbations de se propager convenablement dans le modèle (en particulier sur la verticale). Pour définir cette durée de 4 jours des tests ont été effectués en regardant la dispersion de l'ensemble sur différentes variables du modèle. Une partie de ces tests sont présentés lors de la validation de l'ensemble en Section 5.3.

En résumé, nous disposons d'un ensemble de simulations aux perturbations réalistes échantillonnant la densité de probabilité du vent haute fréquence. La génération d'ensemble par paramétrisation stochastique permet, contrairement à une simple perturbation des conditions initales, d'étudier la propagation des incertitudes au sein du système. Cette étude sera menée à la Section 5.4.

Il est également à noter que la création d'ensembles par paramétrisations stochastiques nous sort en partie du cadre d'une expérience jumelle. En effet, la *vérité* n'est pas une réalisation tirée selon la même densité de probabilité que l'ensemble. De nombreux biais pourront donc apparaître entre la *vérité* et la simulation d'ensemble. La présence de ces biais est discutée en Section 5.3. Il en découle que les diagnostics d'évaluation de l'ensemble ne peuvent pas atteindre des scores parfaits. Ces biais rendent le problème plus difficile mais nettement plus réaliste et permettent donc de faire un premier pas de l'idéalisé vers l'opérationnel.

5.2.3 Réseaux d'observations et vecteurs de contrôle

Premier jeu d'expériences jumelles

Le jeu d'expériences décrit dans cette sous partie est utilisé dans le Chapitre 6.

Pour le Chapitre 6, nous faisons le choix d'utiliser des observations de type profil de température et de salinité et des observations (satellite) de température de surface (schématisés sur la Figure 5.7). Ce choix vient de la volonté de rester au plus près

d'un problème d'assimilation réaliste tout en se laissant la possibilité d'étudier :

- l'impact d'observations dynamiques pour le contrôle de la dynamique et de la biogéochimie du système,
- la qualité de la propagation de l'information sur la verticale pour différentes méthodes d'assimilation,
- le comportement des méthodes d'assimilation face à différentes fréquences d'observations satellites.

Toutes les (pseudo-)observations sont générées à partir de la *vérité* à laquelle des erreurs d'observations sont ajoutées. Cette construction est détaillée ci-dessous pour les deux types d'observations utilisés.

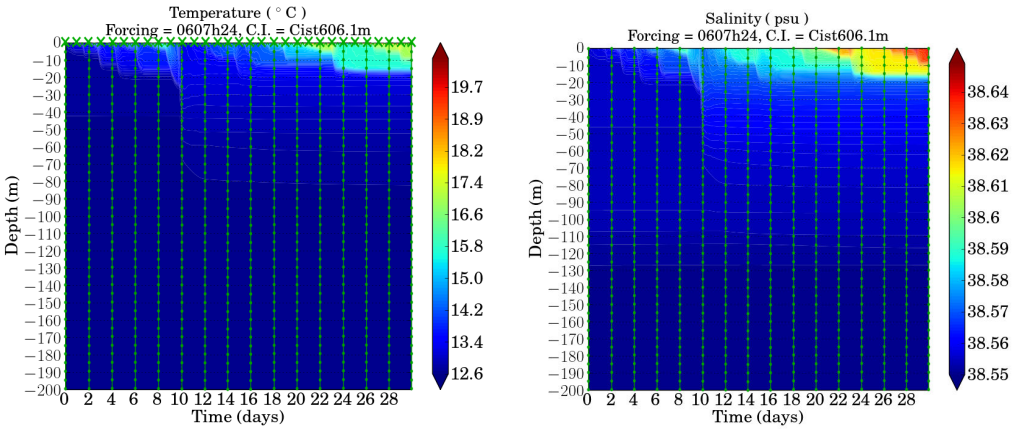

Figure *5.7* – *Réseau d'observation de l'expérience simplifiée ModECOGeL décrit sur un graphique Temps/Profondeur de température T (panneau gauche) et de salinité S (panneau droit). Les points verts représentent les observations de type profil (disponibles tous les 2 jours, tous les 5 mètres). Les croix vertes représentent les observations de type satellitaire de la SST (d'une fréquence variable).*

Observations de type profil Des profils de température (T) et de salinité (S) sont disponibles tous les 2 jours, afin de simuler un profiler du type Argo. Un profil donne des observations le long de la colonne d'eau (de la surface au fond i.e. de 0 à 400m) à raison d'un point observé tous les cinq points de grille. Les observations sont générées en perturbant la solution de référence en lui ajoutant un bruit aléatoire gaussien de moyenne 0 et d'écart-type 0.3 ° C pour la température et 0.06 PSS pour la salinité, en surface et décroissant linéairement avec la profondeur (nul à 400m).

Nom	P0S0	P2S0	P2S6	P2S3	P2S1	P2S05	P0S05
s Profils	Non	2j	2j	2j	2j	2j	Non
Satellite	Non	Non	6h	3h	1h	30min	30min

Table *5.1* – *Nomenclature des différentes expériences d'assimilation selon leurs réseaux d'observations.*

Observations de type satellite Des observations de température de surface (SST) sont disponibles avec différentes fréquences temporelles allant de 30 minutes à 6 heures. Ceci permet d'évaluer l'impact des différentes fréquences d'observation de type satellite sur les corrections. Les observations sont générées en perturbant la solution de référence par ajout d'un bruit aléatoire gaussien de moyenne 0 et d'écart-type 0.3 $^\circ$ C, simulant une erreur d'observation absolue.

Le choix d'une utilisation conjointe ou séparée de ces jeux d'observations est laissé libre et sera l'objet de tests. Les configurations des différents réseaux d'observations qui seront tour à tour testées sont répertoriées dans le Tableau 5.1.

Vecteur de contrôle L'objectif des expériences du Chapitre 6 est, dans un premier temps, de contrôler une partie de la dynamique du système (Section 6.2) et, dans un second temps, d'ajouter au vecteur de contrôle une partie de la biogéochimie du système (Section 6.3).

Ainsi pour les expériences de la Section 6.2, le vecteur de contrôle est composé des variables de température (T) et de salinité (S), deux variables d'état du modèle :

$$X = [T, S]. \tag{5.9}$$

Pour les expériences de la Section 6.3 et de la Section 6.4, le vecteur de contrôle est composé des variables T et S comme précédemment avec en plus les trois variables phytoplanctoniques :

$$X = [T, S, PicP, NanP, MicP]. \tag{5.10}$$

Le lien entre le vecteur de contrôle choisi et la méthode d'assimilation utilisé est très étroit. En effet, on constate au Chapitre 6 que le contrôle du phytoplancton, à partir d'observations de température et de salinité, change considérablement la nature du problème d'assimilation.

Second jeu d'expériences jumelles

Le jeu d'expériences décrit dans cette sous partie est utilisé dans le Chapitre 7.

Pour le Chapitre 7, nous mettons à disposition des observations de type couleur de l'eau.

Ce choix vient de la volonté d'étudier :
- l'apport des observations biogéochimiques pour le contrôle de la biogéochimie,
- l'apport des observations biogéochimiques pour le contrôle du système couplé entier,
- les éventuels bénéfices à utiliser des méthodes d'assimilation non-Gaussiennes pour la résolution d'un problème d'assimilation de données couleur de l'eau.
- la pertinence de la mise en place d'un satellite géostationnaire de couleur de l'eau (avec des mesures hautes fréquences),

Toutes les (pseudo-)observations sont générées à partir de la *vérité* à laquelle des erreurs d'observations sont ajoutées. Cette construction est détaillée ci-dessous.

Observations de type couleur de l'eau Un nouveau type d'observations est disponible depuis l'avènement des satellites : la couleur de l'eau. Il s'agit de mesurer la lumière solaire refléchie par les premiers mètres de l'océan.

Ce principe présente certaines difficultés d'inversion du signal capté par le satellite. En effet, l'océan n'est pas le seul à altérer le signal lumineux. L'atmosphère traversée par la lumière peut modifier ou refléter entièrement le signal.

Malgré ces difficultés, l'analyse du signal permet d'extraire la quantité de chlorophylle présente à la surface (les premiers mètres) de l'océan (Robinson, 2004).

Nous ne nous intéressons pas au travail d'inversion de la couleur de l'eau dans cette thèse. Ce type d'observations étant très prometteur, nous décidons d'assimiler dans le Chapitre 7 le phytoplancton total en surface. Nous faisons ainsi l'hypothèse que le travail d'inversion a déjà été fait.

Satellite	Défilant	Géostationnaire
Nom	OC3j	OC1j
Couleur de l'eau	3 j	1 j

Table *5.2* – *Nomenclature des différentes expériences d'assimilation selon leurs réseaux d'observations de la couleur de l'eau.*

Les observations de couleur de l'eau sont générées à partir de la somme des trois variables phytoplanctoniques issues de la simulation *vérité*. À cette somme est ajoutée une perturbation distribuée selon la loi normale $\mathcal{N}(0, \sigma_{Phy})$ avec $\sigma^2_{Phy} = 0.05$ $mmol.N.m^{-1}$ la variance d'erreurs d'observation du phytoplancton. La valeur $\sigma^2_{Phy} = 0.05$ $mmol.N.m^{-1}$ a été choisie pour correspondre à 20% du phytoplancton moyen (sur le mois étudié).

Dans un premier temps, les observations sont disponibles avec une fréquence (d'images exploitables i.e. sans nébulosité) de 3 jours typique d'un satellite défilant (e.g. le satellite MODIS). Dans un second temps (Sec. 7.4), nous réduisons la fréquence d'observation de la couleur de l'eau à 1 jour.La construction de ces réseaux d'observations de la couleur de l'eau permet de comparer l'intérêt d'un satellite géostationnaire (fréquence d'observations : 1 image par 1/2 heure, 1 image traitable par jour) plutôt qu'un satellite défilant (fréquence d'observations : 1 image traitable par 3 jours). Ce questionnement est d'actualité avec le projet de lancement du satellite géostationnaire Geo-OCAPI [7].

La nomenclature de ces différentes configurations est disponible dans le Tableau 5.2.

Vecteur de contrôle L'objectif du Chapitre 7 est, dans un premier temps, de contrôler la partie biogéochimique du système. Ainsi, pour les premières expériences du Chapitre 7, le vecteur de contrôle est composé des trois variables phytoplanctoniques :

$$X = [PicP, NanP, MicP]. \tag{5.11}$$

Dans un second temps, nous souhaitons contrôler les deux parties couplées du système. Ainsi, pour les expériences de la dernière section du Chapitre 7, le vecteur de contrôle est composé de la variable T et des trois variables phytoplanctoniques :

$$X = [T, PicP, NanP, MicP]. \tag{5.12}$$

5.3 Expérience d'ensemble sans assimilation

Pour toute assimilation d'ensemble, il est bon d'évaluer l'ensemble. Nous voulons savoir si l'ensemble parcourt au mieux l'espace des solutions. Évidemment cet espace est mal connu, il est donc difficile de juger quantitativement de la qualité d'un ensemble à échantillonner l'espace.

Dans un contexte d'expériences jumelles (où l'on possède la *vérité*), on peut d'abord regarder le comportement de l'ensemble par rapport à la *vérité*. La *vérité* étant le seul état connu faisant partie de l'espace des solutions. Nous estimons la bonne dispersion de l'ensemble par rapport à cette dernière dans la première sous partie de cette section.

Dans un problème d'assimilation réel, on se contente de regarder si l'ensemble n'est pas trop loin des observations. Un ensemble soumis à des corrections par des observations lointaines a de fortes chances de s'effondrer ou de générer des états instables pour la dynamique du modèle. Il s'agit donc de regarder si la dispersion de l'ensemble

7. Geostationary Ocean Colour Advanced Permanent Imager

sur les variables observées est cohérente avec la dispersion des observations. C'est ce que nous faisons dans la deuxième sous partie de cette section.

5.3.1 Ensemble et *vérité*

Afin de suivre la propagation des incertitudes au sein du modèle, nous commençons par regarder la source initiale d'erreur : l'intentité du vent.

L'intensité du vent

Comme il est décrit dans la sous partie 5.2.2, la différence entre la simulation d'ensemble et la *vérité* est l'intensité du vent. La création du vent artificiel haute fréquence de l'ensemble se fait par processus auto-régressif Gaussien d'ordre un avec les mêmes paramètres que le vent haute fréquence réel de la *vérité*.

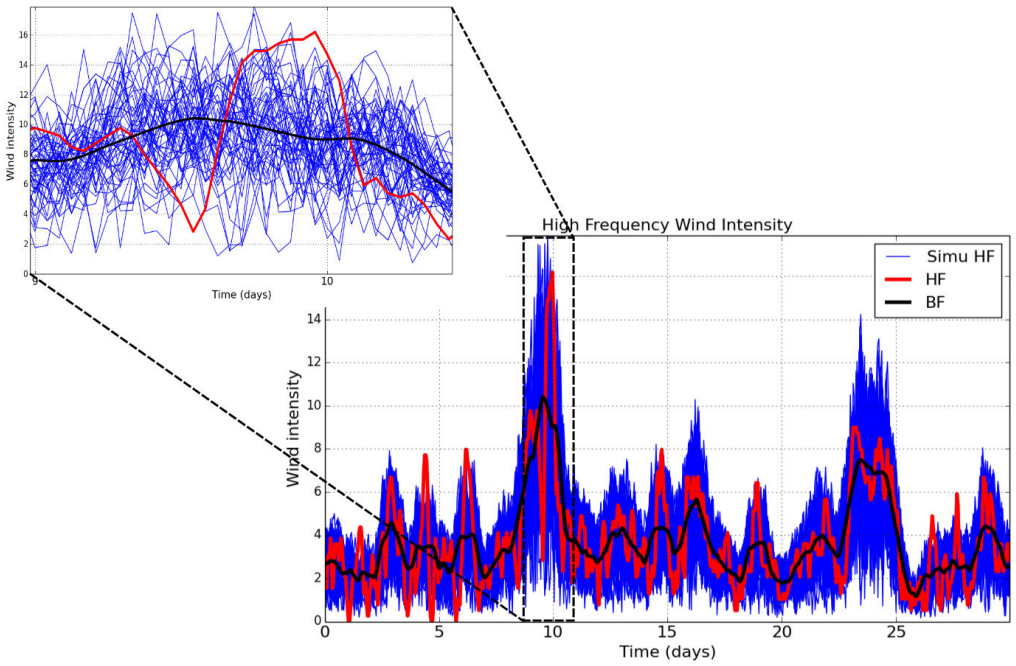

Figure *5.8* – *Série temporelle d'intensité du vent (en $m.s^{-1}$) sur le mois d'avril en haute-fréquence réaliste (HF), en basse fréquence (BF) et de l'ensemble simulant la haute-fréquence par processus stochastique (Simu HF). L'encadré de gauche est un zoom de ces séries temporelles à partir du jour 9 et juqu'au jour 10.*

D'après les séries temporelles des intensités de vent présentées en Figure 5.8 (haute fréquence réelle en rouge, basse fréquence réelle en noir et haute fréquence simulée en bleu), la représentation ensembliste du vent haute fréquence reproduit correctement, à première vue, les différents épisodes de vent. Cette représentation est correcte dans le sens où pour la plupart des événements haute fréquence réels il existe un membre de l'ensemble prescrivant la même intensité. Cependant, en regardant ponctuellement à certains pics de vent, il apparait que plusieurs fortes variations du vent réel sont sous estimées. De plus, autour du 10$^{\text{ème}}$ jour (épisode important dans ce mois d'avril comme on le verra par la suite), se produit une variation importante du vent réel allant de 2-3 $m.s^{-1}$ à 16 $m.s^{-1}$. Cette variation est non seulement importante en amplitude mais également en durée. Ainsi, l'intensité du vent se maintient à plus de 12 $m.s^{-1}$ pendant plusieurs heures. Cet événement extrême mal représenté par l'ensemble peut avoir des conséquences importantes sur l'estimation ensembliste et créer de forts biais entre l'ensemble et la *vérité* comme on le constatera par la suite.

L'énergie cinétique turbulente (TKE)

L'intensité du vent affecte directement la quantité d'énergie cinétique propagée depuis la surface vers les eaux profondes. Ainsi pour que l'ensemble reproduise bien l'évolution de la *vérité*, il faut déjà que la TKE soit bien représentée.

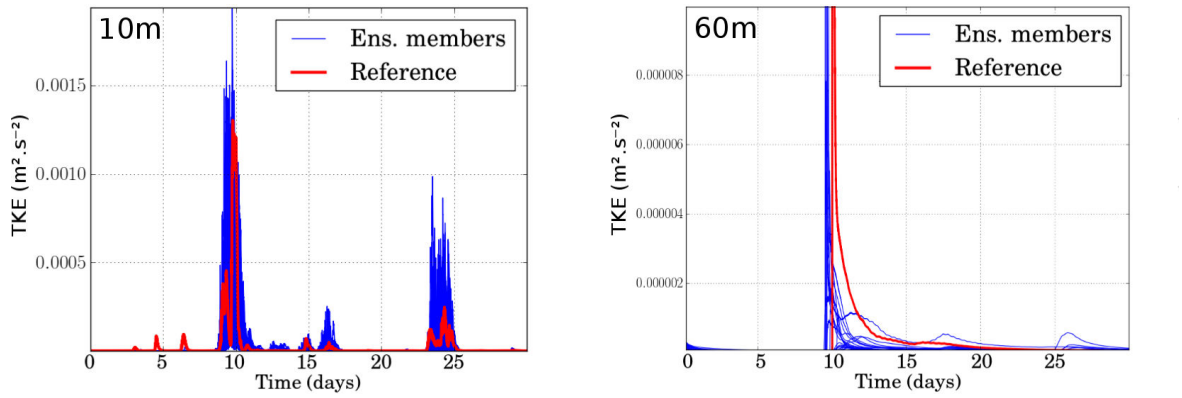

Figure *5.9* – *Série temporelle d'énergie cinétique turbulente (TKE) sur le mois d'avril à 10m (à gauche) et à 60m (à droite) de profondeur de l'ensemble libre (en bleu) et de la vérité (en rouge).*

La Figure 5.9 présente la série temporelle sur le mois d'avril à 10m (à gauche) et à 60m (à droite) de la variable TKE. À 10m, on voit que la *vérité* (en rouge) a des pics de TKE (aux 10$^{\text{ème}}$ et 25$^{\text{ème}}$ jours) correspondant à des événements éoliens forts. L'ensemble (en bleu) parvient à reproduire ces pics de TKE avec, toutefois,

une différence d'intensité. De même à 60m, le pic de TKE au 10$^{\text{ème}}$ jour est simulé par l'ensemble avec un léger décalage temporel. Cette correspondance relative de l'ensemble et de la *vérité* témoigne de la capacité du vent simulé à fournir de l'énergie cinétique turbulente au système presqu'aux mêmes moments que le vent haute-fréquence et en quantité proche. On peut cependant s'attendre à ce que les petites différences de représentation de la TKE engendrent des biais plus important sur le reste du système.

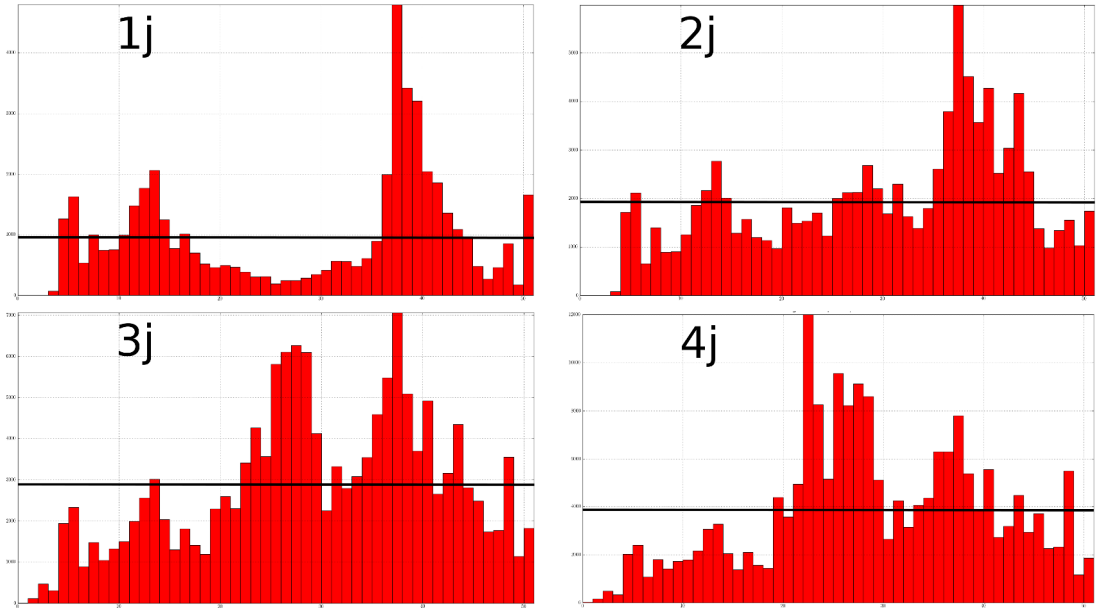

Figure *5.10* – *Histogrammes de rangs de la température sur les 100 premiers mètres et sur le premier jour de spin up (panneau supérieur gauche), sur les deux premiers jours de spin up (panneau supérieur droit), sur les trois premiers jours de spin up (panneau inférieur gauche) et sur les quatre premiers jours de spin up (panneau inférieur droit).*

Une forte dispersion d'ensemble

On attend d'un ensemble qu'il soit suffisamment dispersé pour parcourir au mieux l'espace des solutions. Bien que cette qualité d'ensemble soit importante pour l'assimilation de données, il n'existe pas de méthode généralement utilisée pour l'évaluer.

Dans notre cas, l'ensemble généré par perturbation de l'intensité du vent est intégré pendant quatre jours de spin up avant le début de l'assimilation (i.e. du 28 au 31 mars). Cette période de spin up a été estimée suffisante pour que les perturbations

se propagent à une bonne partie de la colonne d'eau. Parmi les diagnostics mis en place pour évaluer cette période, des histogrammes de rangs pour la température ont été effectués à partir du 28 mars sur 1 jour, 2 jours, 3 jours et 4 jours de spin up et sur toute la colonne d'eau (Figure 5.10). Il apparait qu'au bout de quatre jours, l'ensemble est très sur-dispersif ce qui indique que le voisinage de la *vérité* est bien exploré par l'ensemble.

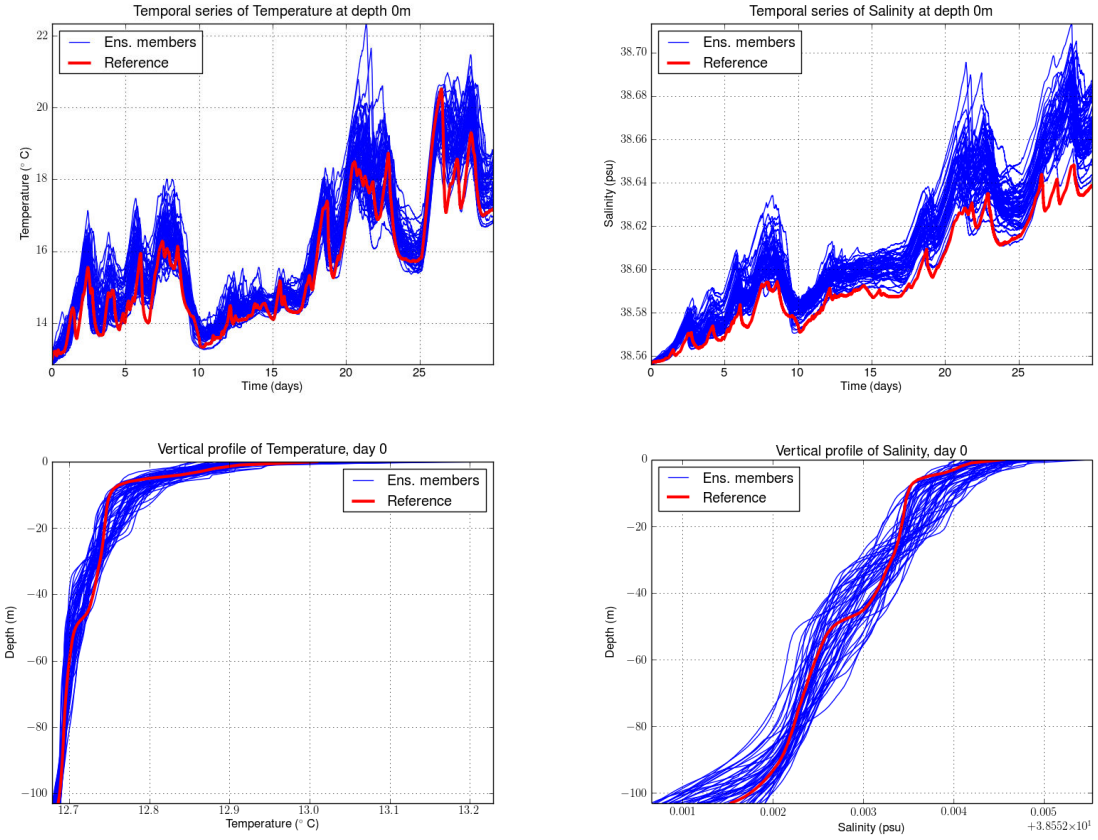

Figure *5.11* – *Séries temporelles sur le mois d'avril et à la surface (graphiques supérieurs) et profils verticaux moyennés sur le mois (graphiques inférieurs) de température (panneaux de gauche) et de salinité (panneaux de droite) pour l'ensemble libre (en bleu) et pour la vérité (en rouge).*

De plus, les profils verticaux de l'ensemble et de la *vérité*, à la fin du spin up (i.e. au jour 0 de l'assimilation) pour la température et la salinité, confirment la bonne propagation des perturbations sur la verticale (Figure 5.11, graphiques inférieurs).

Les séries temporelles de la température et de la salinité de surface sur le mois d'avril à la surface (Figure 5.11, graphiques supérieurs) indiquent que la dispersion de l'ensemble est suffisante pour appliquer une assimilation. La salinité semble cependant être systématiquement surestimée par l'ensemble. Ce type de biais existe également entre le modèle ModECOGel et les observations réelles comme l'indique la comparaison du modèle aux observations (Sec. 5). Ceci rend les assimilations qui sont menées dans les chapitres suivants plus difficiles mais aussi plus réalistes.

Nous consacrons maintenant quelques mots à la présence de ces biais.

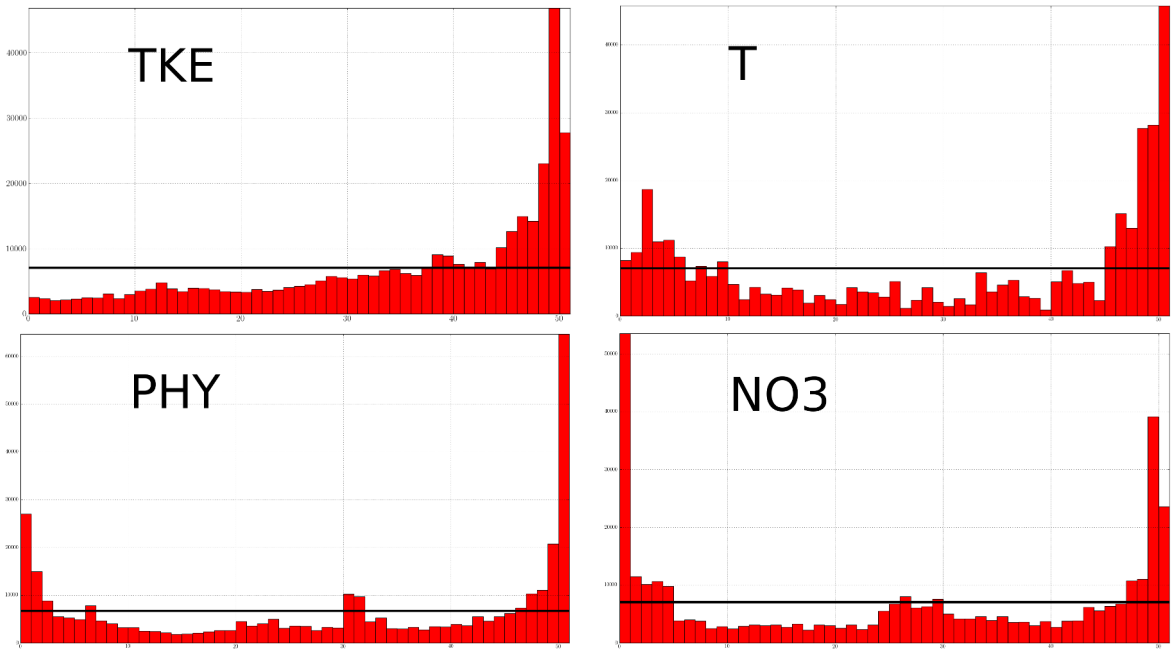

Figure *5.12* – *Histogrammes de rangs de TKE (graphique supérieur gauche), de la température (graphique supérieur droit), du NO_3 (graphique inférieur gauche) et de phytoplancton (graphique inférieur droit) sur les 50 premiers mètres et sur le mois d'avril.*

De nombreux biais

Les histogrammes de rangs que l'on utilise sont réalisés de manière globale en considérant les différents points de grille et les différents pas de temps comme différentes réalisations d'une même statistique que l'on cherche à évaluer. Cette hypothèse bien que forte nous est nécessaire au vue du coût de calcul et de stockage de chaque

expérience. Cette hypothèse présente l'inconvénient de modifier la lecture de ce diagnostic. En effet, un histogramme en forme de U n'est pas toujours significatif d'une sous dispersion de l'ensemble (au sens où la dispersion peut être améliorée). Une combinaison de plusieurs biais à différents points de grille et différents pas de temps, entraine les mêmes formes d'histogrammes. Et comme, dans nos expériences, la *vérité* et l'ensemble ne sont pas produits par le même modèle (différente nature statistique du forçage intensité de vent), augmenter les perturbations ne rendra pas forcément les histogrammes de rangs plus plats.

Si l'on observe les histogrammes de rangs sur le mois d'avril et sur les 50 premiers mètres pour l'énergie cinétique turbulente, la température, le phytoplancton et le nitrate (Figure 5.12), l'ensemble parait sous dispersif. Il s'agit en réalité, d'importants biais provoqués par des événements de vents non reproduits par l'intensité du vent simulée.

En regardant les cartes de rangs pour les mêmes variables que les histogrammes de rangs précédents (Figure 5.13), on constate des biais par régions (spatio-temporelles).

Nous rappelons que les cartes de rangs donnent la position de la vérification dans les intervalles entre les membres d'ensemble. Si la valeur donnée par la carte est 0 cela signifie que la vérification est surestimée par l'ensemble et 50 si la vérification est sous-estimée. En supposant l'homogénéité spatiale et l'ergodicité des probabilités, la carte de rangs devrait présenter chacune des couleurs (de la barre de couleurs) un nombre identique de fois (équivalent à un histogramme de rangs plat).

La variable TKE, la température et le phytoplancton sont sous estimés par l'ensemble entre 20m et 110m à partir du $10^{\text{ème}}$ jour puisque les vérifications tombent principalement au dessus du $50^{\text{ème}}$ membre (zone en rouge). Cette sous estimation coincide avec la forte variation de vent de la *vérité* au jour 10, mal représentée par l'ensemble de vent (5.8). Une forte variation de vent entraine une augmentation de la TKE (donc sous estimée par l'ensemble) qui mélange les eaux chaudes de surface avec les eaux plus profondes et augmente donc la température (également sous estimée par l'ensemble). Enfin, dans un milieu où la température augmente (vers la température optimale de 15 degré), le phytoplancton a un taux de croissance plus élevé et sa concentration augmente (augmentation sous estimée par l'ensemble). À l'inverse, le nitrate au même moment et aux mêmes profondeurs est sur-estimé (zone en violet-blanc) puisque plus consommé par la grande quantité de phytoplancton dans la vérité. Également, le phytoplancton en surface est sous-estimé avant le $10^{\text{ème}}$ jour et sur-estimé ensuite ce qui est une conséquence directe à l'approfondissement de la couche de mélange que l'on vient de décrire.

En regardant les séries temporelles en surface du phytoplancton et du nitrate, on

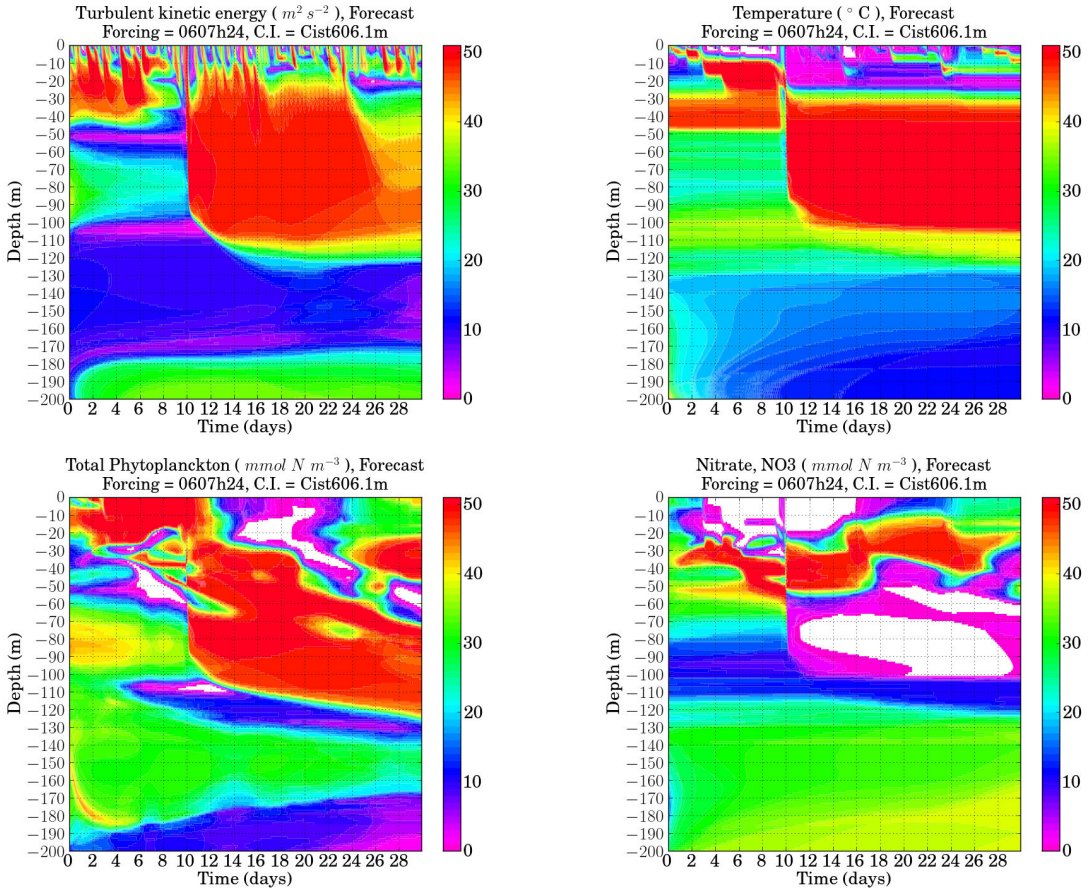

Figure *5.13* – *Cartes de rangs en fonction de la verticale et du temps de TKE (graphique supérieur gauche), de la température (graphique supérieur droit), du NO_3 (graphique inférieur gauche) et de phytoplancton (graphique inférieur droit) sur les 50 premiers mètres et sur le mois d'avril.*

constate le même phénomène (Figure 5.14). La concentration en phytoplancton en surface estimée par la *vérité* est plus forte entre les jours 5 et 10, puis elle diminue. Les concentrations de phytoplancton de l'ensemble sont, elles, faibles et peu dispersées au début du mois (puisque soumis à des intensités de vent proche) puis augmentent et se dispersent à partir du jour 10. Cet effet peut faire suite, par exemple, à une combinaison de forte intensité lumineuse et de faibles approfondissements de la couche de mélange. Le nitrate simulé par la *vérité* est plus rapidement consommé en surface Ce qui est dû à sa plus grande concentration de phytoplancton avant le $10^{\text{ème}}$ jour. Alors que les concentrations de nitrate pour l'ensemble rejoignent les mêmes valeurs seulement vers la fin du mois.

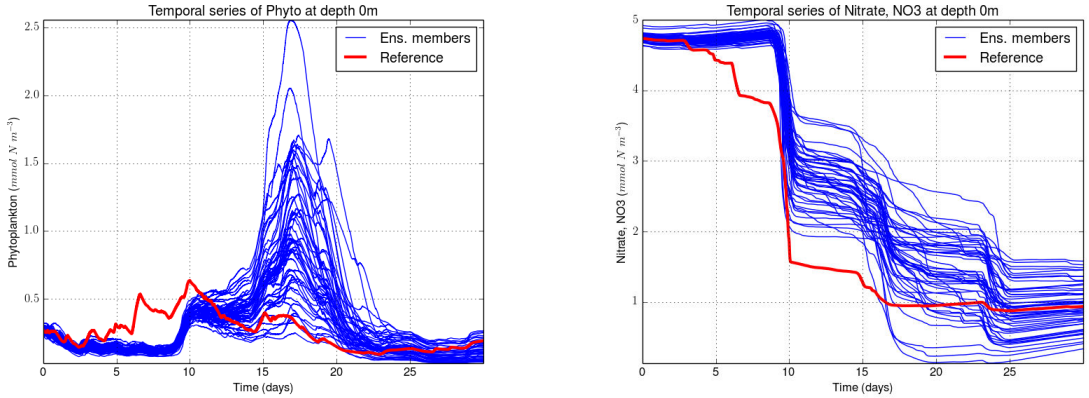

Figure *5.14* – *Séries temporelles du phytoplancton (graphique de gauche) et du NO_3 (graphique de droite), sur le mois d'avril et en surface, de l'ensemble libre (en bleu) et de la vérité (en rouge).*

Remarque L'un des objectifs de l'assimilation de données réalistes étant aussi de corriger les biais entre le modèle et la *vérité*, l'évaluation des différentes méthodes d'assimilation par histogrammes de rangs est toujours intéressante. L'histogramme de rangs contient ainsi une information combinée sur la dispersion de l'ensemble et sur la réduction des biais.

5.3.2 Ensemble et observations

Dans cette sous partie, nous comparons la dispersion de l'ensemble et les erreurs d'observations (préscrites). Une bonne cohérence de ces deux quantités est nécessaire au bon fonctionnement des filtres séquentiels d'ensemble utilisés. Cette comparaison permet de confirmer le choix de l'intensité des perturbations et le choix des erreurs d'observations préscrites.

Pour cette comparaison, nous nous focalisons principalement sur les observations de surface : la température de surface (SST) et le phytoplancton de surface (couleur de l'eau).

Observations de température de surface (SST)

Il est possible d'évaluer la pertinence des observations, en comparant la distribution de l'ensemble libre et la distribution des observations. Le graphique de gauche de la Figure 5.15 représente les séries temporelles, sur le mois d'avril, de la température

de surface pour l'ensemble libre (en bleu) et pour la *vérité* (en rouge) ainsi que les observations de SST (croix vertes). Les biais qui étaient présents entre l'ensemble et la *vérité* se retrouvent entre l'ensemble et les observations. Les températures observées sont souvent plus faibles que les températures estimées par l'ensemble. Cependant, il est important en expériences jumelles de vérifier que la dispersion de l'ensemble et celle des observations sont du même ordre de grandeur. Si ce n'est pas le cas les observations synthétiquement générées seraient systématiquement ignorées par l'assimilation puisqu'entachées par construction d'une trop forte erreur en comparaison à l'erreur *a priori* donnée par l'ensemble. Le graphique de gauche montre que la dispersion de l'ensemble et celle des observations sont proches. Ceci nous est confirmé en regardant les erreurs d'observation comparées à l'intervalle de plus ou moins deux fois l'écart-type de l'ensemble (Figure 5.15, graphique de droite). À l'exception de la période entre le jour 8 et le jour 18, où l'ensemble est beaucoup moins dispersé, l'ensemble a une dispersion du même ordre de grandeur que les erreurs d'observation.

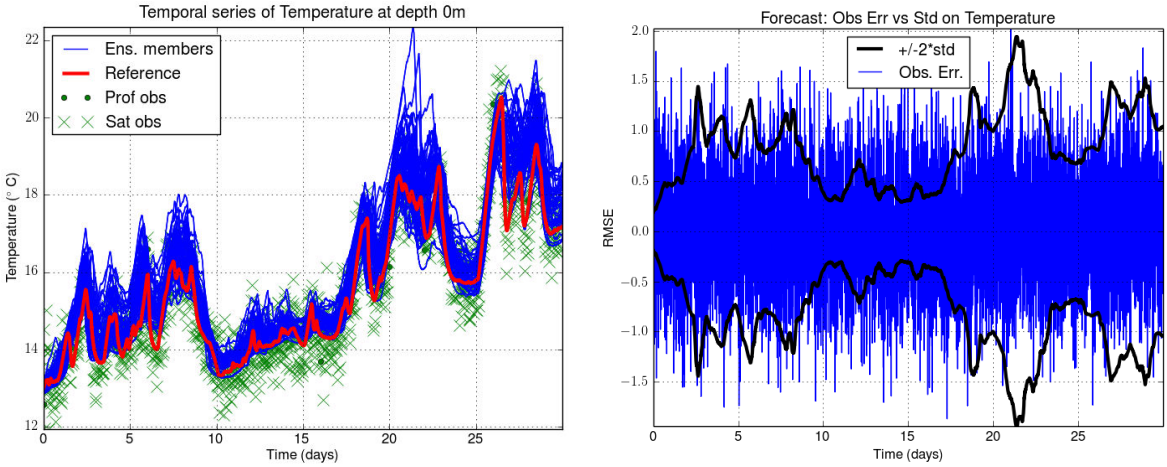

Figure *5.15* – ***Graphique de gauche :*** *Séries temporelles, sur le mois d'avril et à la surface, de la température pour l'ensemble libre (en bleu) et pour la vérité (en rouge) ainsi que les observations de SST (croix vertes).* ***Graphique de droite :*** *Erreurs d'observation SST (en bleu) et intervalle* $[-2\sigma_{ens}, 2\sigma_{ens}]$ *(en noir) avec* σ_{ens} *l'écart-type de l'ensemble, sur la température de surface en fonction du temps.*

Observations de la couleur de l'eau

Comme précédemment, nous évaluons la distribution de l'ensemble libre et la distribution des observations pour la couleur de l'eau. Les séries temporelles (Figure

5.16, graphique de gauche), sur le mois d'avril et à la surface, du phytoplancton pour l'ensemble libre (en bleu) présentent une très faible dispersion pendant les 10 premiers jours. De plus, il y a un biais important entre l'ensemble et les observations à cette période. Après plusieurs tests (non montrés ici), il est apparu qu'il n'est pas possible d'augmenter cette dispersion à partir des perturbations de l'intensité du vent. Hormis à cette période l'ensemble est largement dispersé, particuliérement entre le jour 10 et le jour 25.

Ces résultats sont confirmés par la comparaison entre les erreurs d'observation et l'intervalle de plus ou moins deux fois l'écart-type de l'ensemble (Figure 5.16, graphique de droite).

Obtenir une bonne correction, sur les 10 premiers jours, par assimilation de la couleur de l'eau semble difficile. En revanche, l'assimilation devrait permettre de fortement améliorer la dispersion de l'ensemble sur le reste du mois d'avril.

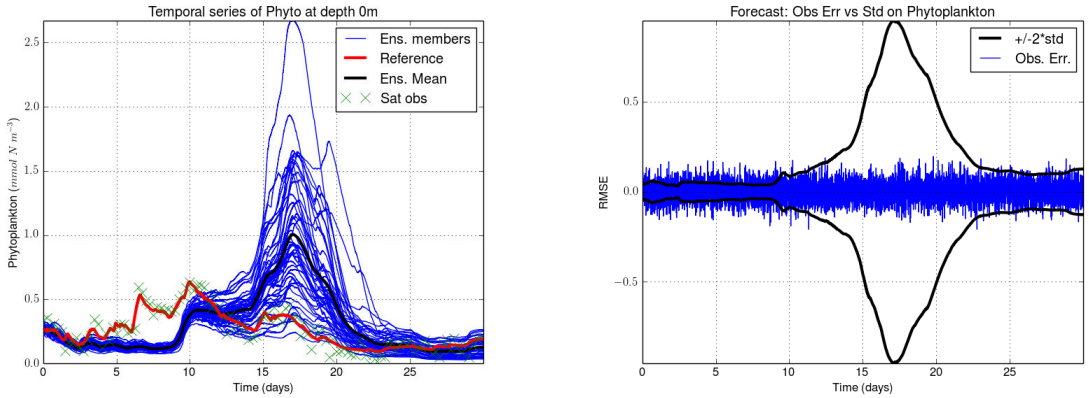

Figure *5.16* – ***Graphique de gauche :*** *Séries temporelles, sur le mois d'avril et à la surface, du phytoplancton pour l'ensemble libre (en bleu) et pour la vérité (en rouge) ainsi que les observations de couleur de l'eau (croix vertes).* ***Graphique de droite :*** *Erreurs d'observation couleur de l'eau (en bleu) et intervalle* $[2\sigma_{ens}, -2\sigma_{ens}]$ *(en noir) avec* σ_{ens} *l'écart-type de l'ensemble, sur la phytoplancton de surface en fonction du temps.*

5.4 Étude qualitative de la propagation des incertitudes

La section précédante nous a permis de constater le comportement de l'ensemble sans assimilation en le comparant de manière statistique avec la *vérité* et avec les observations. Nous cherchons à présent à comprendre, par des considérations phy-

siques, l'origine de ce comportement en étudiant la propagation des incertitudes et leurs impacts sur l'évolution de l'ensemble.

5.4.1 Évolution des incertitudes à travers la dynamique

Les premières questions, qu'il est légitime de poser en abordant un problème d'assimilation de données, concernent la nature du modèle en lui-même. En effet les systèmes considérés en géoscience font tous intervenir de nombreuses non-linéarités. En particulier, les équations d'un modèle de biogéochimie font intervenir de nombreuses et fortes non-linéarités telles que des phénomènes de seuil.

Ces non-linéarités de degrés plus ou moins importants, vont notamment jouer un rôle sur la propagation des incertitudes dans le modèle. Dans ce cas de figure, l'hypothèse de Gaussianité qui sera éventuellement faite par l'assimilation de données a de forts risques d'être erronée et donc pénalisante.

Dans cette section, nous étudions la propagation des incertitudes de manière qualitative.Nous considérons particulièrement les effets non-linéaires du système et la non-Gaussianité des incertitudes qui s'y propagent.

À l'aide d'une compréhension physique des phénomènes en jeu, nous pouvons anticiper la propagation des incertitudes depuis le forçage du vent jusqu'aux quantités qui nous intéressent (e.g. les variables de contrôle). Nous pouvons également évaluer qualitativement les degrés de non-linéarité qui accompagnent cette propagation.

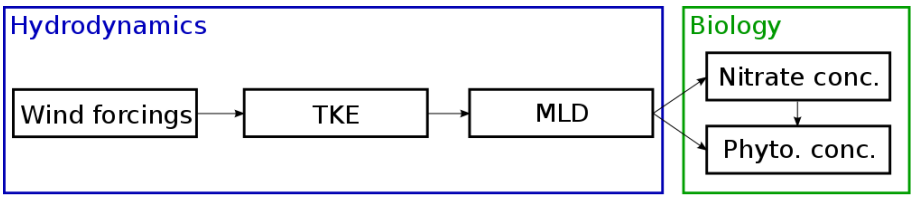

Figure *5.17* – *Fonction conceptuelle de transfert des incertitudes à travers les principales relations entre variables dans le modèle couplé dynamique-biogéochimie marine.*

La figure 5.17 schématise les principaux trajets de propagation des incertitudes dans le modèle ModECOGeL. Le forçage de l'intensité du vent (**Wind forcings**), qui est dans nos problèmes d'estimation la source des perturbations, va impacter la dynamique du modèle. Cette dynamique agira à son tour sur la biogéochimie du système, avec principalement les déplacements de la profondeur de couche de mélange (**MLD**). Les deux paragraphes qui suivent rentrent dans le détail de ce découpage conceptuel en deux étapes de la propagation des incertitudes.

Comme on l'a vu lors de la validation et l'étude de la *vérité* au chapitre précédent (Section 5.1.3), de fortes intensités de vent provoquent des incursions verticales

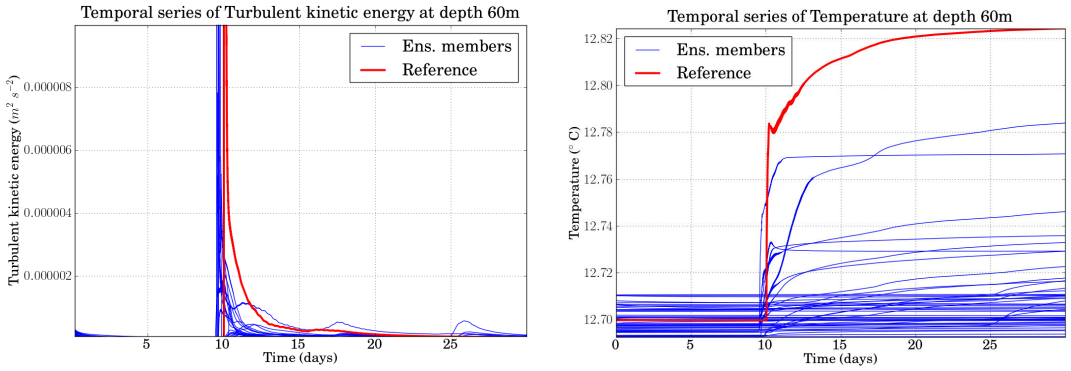

Figure *5.18* – *Séries temporelles sur le mois d'avril, à 60 mètres, de l'énergie cinétique turbulente TKE (graphique de gauche) et de la température T (graphique de droite) pour les 50 membres de l'ensemble libre en bleu et la vérité en rouge.*

d'énergie cinétique turbulente (TKE). Les perturbations relatives appliquées à l'intensité de vent auront donc une grande variabilité lors de ces épisodes de fortes intensités. L'effet principal de ce phénomène est la variation d'intensité de ces incursions de TKE d'un membre de l'ensemble à l'autre. En regardant l'équation des flux d'énergie cinétique de surface, on constate que les variations d'énergie cinétique turbulente dépendent non-linéairement de l'intensité du vent par un degré 3 :

$$\tilde{\nu}\left.\frac{\partial k}{\partial z}\right|_{Surface} = 3C_0 10^{-3}||V_{wind}||^3, \tag{5.13}$$

avec k l'énergie cinétique turbulente, $\tilde{\nu}$ la viscosité turbulente verticale et $C_0 = 0.63 \times 10^{-6}$.

Dans un milieu stratifié ou en cours de restratification, ces fortes variations de TKE (d'un membre à l'autre) provoquent des réponses non-linéaires en température. C'est ce qui se produit notamment au 10$^{\text{ème}}$ jour du mois d'avril avec une incursion de TKE jusqu'à 80m et un mélange en température à cette période.

La Figure 5.18 illustre ce dernier phénomène, en présentant une série temporelle sur le mois d'avril à 60m de profondeur de la TKE (graphique de gauche) et de la température (graphique de droite) pour la *vérité* (en rouge) et l'ensemble libre de 50 membres (en bleu). Les incursions de TKE au 10$^{\text{ème}}$ jour à 60m sont très variables selon les membres de l'ensemble (voires inexistantes pour certains membres). La réponse en température varie en conséquence. À noter que suite à cet événement, la distribution de l'ensemble en température est modifiée et n'est plus Gaussienne à partir du 10$^{\text{ème}}$ jour. Il faudra attendre une restratification progressive par la surface pour que la distribution se renormalise. Ce raisonnement sera confirmé dans la

sous partie suivante en diagnostiquant la normalité de l'ensemble sur la variable de température (Fig. 6.2).

5.4.2 Les processus dominant la concentration de phytoplancton

La biogéochimie d'un système couplé dépend de la dynamique océanique par différents phénomènes. Même dans une représentation 1D, ces dépendances sont complexes et font intervenir de fortes non-linéarités (Lévy et al., 1998; Magri et al., 2005; Perruche et al., 2010).

En particulier, les variations de la concentration du phytoplancton (Q_{phyto}) sont régies par des équations de la forme :

$$\partial_t Q_{phyto} = C - E, \tag{5.14}$$

avec C la croissance et E l'extinction du phytoplancton. La croissance C dépend de la quantité de nutriments (NO_3 et NH_4) et de la quantité de lumière disponibles. L'extinction E dépend de la mortalité et de l'ingestion par des prédateurs (e.g. le zooplancton).

Le phytoplancton est concentré dans la couche de mélange. C'est pourquoi il est important de regarder les conditions impactant la couche de mélange qui influent sur le développement du phytoplancton. La quantité de nutriment et la lumière pénétrant dans la couche de mélange modifieront donc fortement le terme de croissance C.

Lorsque la couche de mélange (MLD) est peu profonde, le phytoplancton est bien éclairé, il y a une accelération de la croissance. En revanche, une MLD profonde ne permettra pas l'éclairement du phytoplancton au bas de la couche de mélange mais permettra l'approvisionnement en nutriment par les couches inférieures. Ce qui peut se traduire également par une accelération de la croissance. Ces deux phénomènes aux causes antagonistes auront à tour de rôle un impact dominant sur la croissance du phytoplancton. Cette alternance de phénomènes dominants constitue une relation à seuils très non-linéaire entre la profondeur de la couche de mélange (MLD) et la concentration de phytoplancton. Cette relation dégradera à son tour les densités de probabilité en densités non-Gaussiennes.

5.4.3 Du vent au phytoplancton

La complexité des relations que l'on vient de décrire, allant du forçage par l'intensité du vent au phytoplancton, peut s'observer à travers les différentes évolutions des membres d'ensemble.

Par exemple, la Figure 5.19 représente les diagrammes de dispersion (nuages de points ou *scatterplots*) de l'ensemble entre la MLD et le phytoplancton total de la colonne d'eau, au jour 0, au jour 10 et au jour 15. Deux membres de l'ensemble sont mis en évidence, le membre 28 (rond bleu) et le membre 44 (rond rouge).

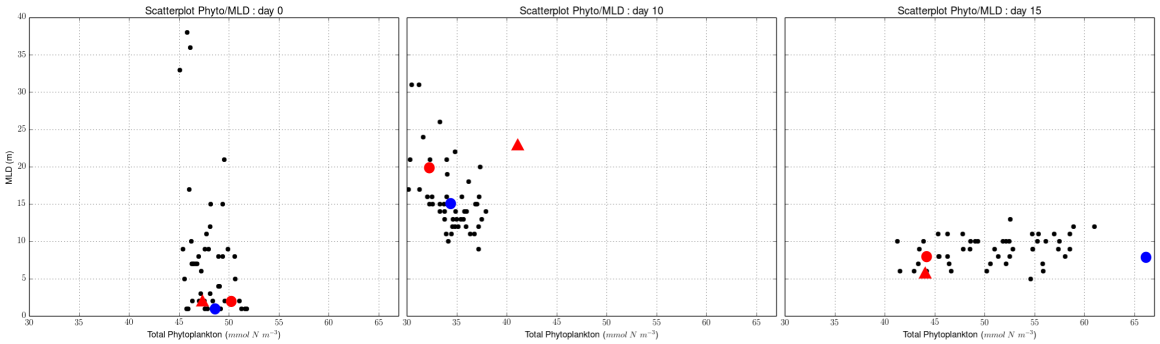

Figure *5.19* – *Diagrammes de dispersion entre la MLD et le phytoplancton (sommé sur la colonne d'eau) au jour* 0 *(diagramme de gauche), au jour* 10 *(diagramme du centre) et au jour* 15 *(diagramme de droite) des 50 membres de l'ensemble (ronds verts) et de la vérité (triangle rouge). Deux membres sont mis en évidence : le membre 28 (rond bleu) et le membre 44 (rond rouge).*

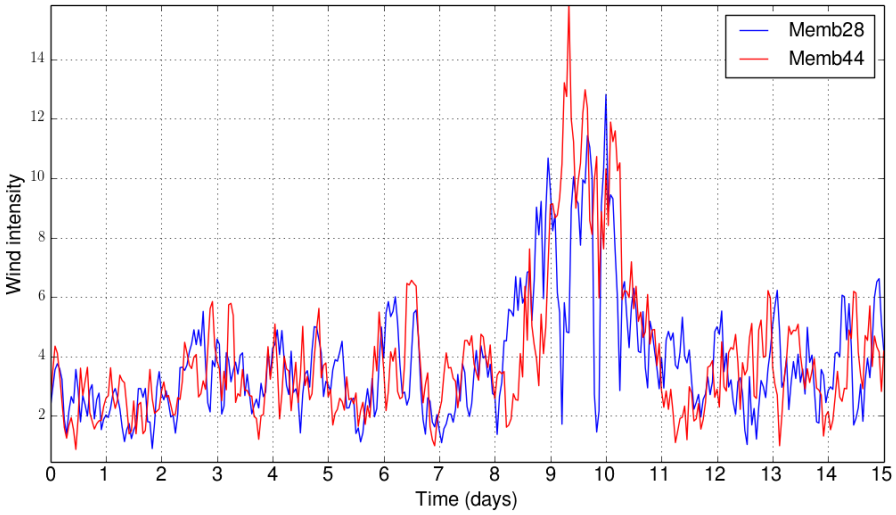

Figure *5.20* – *Séries temporelles de l'intensité du vent (haute fréquence simulée) sur la première moitié du mois d'avril pour les membres d'ensemble 28 (bleu) et 44 (rouge).*

En s'intéressant à l'évolution de ces deux membres au cours du temps, la Figure 5.19 nous permet de constater les non-linéarités régissant ce système et provoquant des bifurcations de trajectoire.

Au jour 0, les deux membres présentent une faible MLD, avec une concentration de phytoplancton légèrement plus forte du membre 44 (rond rouge). Après l'épisode

fort de vent du 10$^{\text{ème}}$ jour, les MLD des deux membres sont approfondies. Le membre 44 (avec une plus forte concentration de phytoplancton au jour 0) reçoit, par forçage, une plus grande intensité de vent autour du jour 10 (que l'on peut observer sur les séries temporelles d'intensité du vent présentées en Fig. 5.20), ce qui approfondit sa MLD de 5m de plus que le membre 28. Cette différence de 5m conduit cinq jours plus tard, au jour 15, à la présence de deux membres d'ensemble avec la même MLD mais avec une concentration de phytoplancton 1.5 fois supérieure pour le membre 28 à celle du membre 44.

5.5 Bilan

5.5.1 Bilan du chapitre

Dans ce chapitre, nous avons présenté le modèle unidimensionnel couplant dynamique et biogéochimie, ModECOGeL, et les observations, profils de température et de salinité et données satellites de la SST, qui seront utilisés par la suite.

Nous avons choisi une source d'incertitudes importante du modèle, l'intensité du vent forcée, et défini une paramétrisation stochastique qui permet de la simuler.

Une expérience d'ensemble sans assimilation a été réalisée et étudiée. Cette expérience met en évidence une bonne dispersion de l'ensemble avec toutefois la présence de nombreux biais ainsi que des bifurcations de trajectoires (impliquant de fortes non-Gaussianités).

Dans ce cas d'étude, les difficultés pour l'assimilation de données sont nombreuses. Il sera donc difficile de favoriser ou d'écarter une méthode sur la seule base de ses performances dans ce cas particulier. En revanche, se placer face à des difficultés, qui sont courantes en assimilation de données réelles, donnera un echo plus grand aux conclusions que l'on pourra tirer.

En résumé, le Chapitre 5 pose toutes les bases nécessaires à la réalisation des Chapitres 6 et 7.

5.5.2 Préambule des chapitres suivants

Nous nous sommes dotés – par l'utilisation d'un modèle couplé complexe, par la représentation des incertitudes avec des paramétrisations stochastiques, par la création de la *vérité* avec un vent réaliste haute fréquence – d'un cas d'étude allant au delà de simples expériences jumelles. Il s'agit d'un problème, miniaturisé certes, mais réaliste.

Dans les expériences des chapitres 6 et 7, nous nous plaçons donc dans un contexte réaliste (*a contrario* de l'étude du chapitre 3) de couplage de la dynamique et de la

biogéochimie marine. Ce contexte offre des problèmes d'estimation complexes, en combinant des phénomènes quasi-linéaires engendrant peu de non-Gaussianités et des phénomènes non-linéaires engendrant de fortes non-Gaussianités.

De nombreuses études ont révélé le caractère non-linéaire d'un système couplé de dynamique océanique et de biogéochimie marine que ce soit en une dimension (Lévy et al., 1998; Perruche et al., 2010) ou en trois dimensions (Béal et al., 2010). Toutefois, ces études insistent peu sur la notion de Gaussianité des probabilités en jeu.

D'autre part, l'évaluation de méthodes d'assimilation de données dans ce même contexte a également été effectuée (Magri et al., 2005; Bertino et al., 2003; Pelc et al., 2012; Béal et al., 2010). Ces évaluations ne reposent pas sur des études statistiques a priori du problème.

Il semble donc approprié d'effectuer une telle étude a priori et d'en tirer des conclusions sur le caractère non-linéaire et non-Gaussien des problèmes étudiés. On peut dans un premier temps voir si ces conclusions se confirment lors de la mise en place de l'assimilation de données. Dans un second temps, cette étude permet d'anticiper et de mieux comprendre, d'un point de vue méthodologique, les performances de méthodes d'assimilation Gaussienne et non-Gaussienne dans différentes configurations.

Les objectifs des chapitres d'expériences sont d'observer et de caractériser la complexité du problème d'assimilation dans un système de couplage de la dynamique et de la biogéochimie marine ; d'évaluer et de comprendre le comportement de méthodes d'assimilation de données Gaussienne et non-Gaussienne selon différentes configurations faisant varier le vecteur de contrôle et le réseau d'observation.

Chapitre 6

Contrôle de la biogéochimie en observant la dynamique

Sommaire

Le cadre des expériences présentées dans ce chapitre est décrit au Chapitre 5.

Nous travaillons avec le modèle couplé de dynamique et de biogéochimie marine, ModECOGeL, sur le mois d'avril 2006. La *vérité* est une simulation effectuée avec un forçage de vent réel en haute fréquence (1 heure). Un ensemble de 50 membres est propagé par le modèle avec un forçage de vent haute fréquence simulé par un processus auto-regressif Gaussien d'ordre un.

L'objectif de ce chapitre sont : de vérifier la capacité d'un filtre Gaussien (l'ETKF) à contrôler la dynamique en observant la dynamique (à différentes fréquences) ; d'évaluer l'intérêt de l'utilisation d'une méthode non-Gaussienne (le MRHF) pour le contrôle de la biogéochimie en observant la dynamique ; et enfin d'étudier l'impact des observations haute-fréquence de la SST sur ces deux types de méthodes d'assimilation.

Dans une première partie nous discriminerons différents problèmes d'estimation en fonction de leurs degrés de non-linéarité et de non-Gaussianité au moyen d'une étude a priori qualitative et statistique en utilisant certains des diagnostics présentés au chapitre 1 (Section 2.1.1). Puis nous évaluerons en section 6.2 le filtre de Kalman d'ensemble transformé (ETKF), une méthode produisant un résultat supposé proche de l'optimalité sous les hypothèses de linéarité et de Gaussianité, dans un cadre de contrôle de la dynamique (température et salinité) du système en observant une partie la dynamique (température et salinité) du système. Cette étude évaluera notamment les performances de l'ETKF en fonction du réseau d'observations fourni. Dans la section 6.3, nous comparerons l'ETKF et le filtre multivarié d'histogrammes de rangs (MRHF), une méthode non-Gaussienne, dans un cadre de contrôle mixte de la dynamique (température et salinité) et de la biogéochimie (trois classes de phytoplancton) du système en observant toujours la dynamique (température et salinité). Enfin, dans la section 6.4, nous étudierons dans ce même dernier cadre le comportement de ces deux méthodes face à des observations haute-fréquence de la SST. La dernière section (Section 6.5) synthétise les résultats obtenus dans ce chapitre.

6.1 Étude des relations dynamico-biogéochimiques

6.1.1 Étude statistique des relations dynamico-biogéochimiques

L'étude statistique que l'on effectue dans cette sous partie se déroule sur le mois d'avril 2006 considéré pour nos expériences et se focalise sur les premiers 100m de la colonne d'eau où les effets qui nous intéressent prennent place. On décrit puis on utilise une partie des diagnostics sommairement présentés dans la Section 2.1.1. Ces diagnostics sont appliqués à un ensemble de 50 simulations libres générées par

perturbation de l'intensité du vent comme il est décrit au chapitre précédent.

Non-linéarités inter-variables

La première quantité statistique que l'on peut regarder ici est la linéarité au sens de Pearson (Pearson, 1895) entre deux variables r (Sec. 2.1.1). Puisque l'on souhaite savoir si une corrélation existe entre les variables évaluées nous pouvons prendre la valeur absolue du coefficient de corrélation de Pearson qui va donc de 0 à 1 pour respectivement aucune corrélation et une corrélation ou anti-corrélation totale. La seconde quantité statistique que l'on regarde est la linéarité au sens de Spearman ρ(Sec. 2.1.1). En valeur absolue, ce coefficient vaudra 1 pour une forte corrélation de rang et 0 pour aucune corrélation de rang. Ces diagnostics nous apportent deux informations différentes :

- La première est la nature des correlations entre deux variables qui va nous permettre de prévoir le contrôle d'une variable par l'observation d'une autre. Dans ce cas là, il est à noter que pour un coefficient de Pearson (en valeur absolue) proche de 1, on peut s'attendre à de bonnes performances de la part des méthodes d'assimilation utilisant une régression linéaire pour propager les corrections des variables observées aux variables non observées (e.g. EnKF, ETKF et RHF). Par contre pour un coefficient de Pearson faible et un coefficient de Spearman proche de 1, l'anamorphose telle que décrite au chapitre 1 (Section 2.3.2) doit correctement propager les corrections à toutes les variables. L'anamorphose est une transformation adaptée aux faibles non-Gaussianités. Dans le cas des coefficients de Pearson et Spearman tous deux proches de 0, seule une méthode d'assimilation fondamentalement non-Gaussienne pourrait appliquer des corrections adaptées s'il existe des relations statistiques autres que linéaires.
- La seconde peut donner une indication sur le phénomène dominant dans un processus. Une corrélation forte entre deux variables peut signifier que l'une des deux variables joue un rôle important dans l'évolution de l'autre. Ce qui permet de confirmer l'origine des variations et le degré de non-linéarité qui les caractérise.

La Figure 6.1 représente en fonction de la profondeur (en mètres sur les 100 premiers mètres de la colonne d'eau) et du temps (en jours sur le mois), le coefficient $r_{X,Y}$ (colonne de gauche), le coefficient $\rho_{X,Y}$ (colonne du centre) et le ratio des deux $\frac{r_{X,Y}}{\rho_{X,Y}}$ (colonne de droite) pour les variables $X, Y = TKE, TEM$ (première ligne), $X, Y = TEM, NO3$ (deuxième ligne), $X, Y = NO3, Phyto$ (troisième ligne) et $X, Y = TEM, Phyto$ (quatrième ligne). La première remarque au vu des graphiques de la Fig. 6.1, est que les coefficients de Pearson et Spearman sont très proches.

Une information qu'apportent la deuxième et la quatrième ligne de la Fig. 6.1 est que la relation entre la température et la biogéochimie est très non-linéaire (au sens

de Pearson). Il sera donc difficile de contrôler la biogéochimie (e.g. le phytoplancton) avec un filtre de type ETKF ou RHF en observant la température. De plus, pour la relation température-phytoplancton, le ratio des deux coefficients (colonne de droite) est souvent proche de 1, ce qui confirme la proximité des deux coefficients. Ceci nous indique déjà que lorsqu'il y a une relation statistique entre ces deux variables il s'agit d'une linéarité au sens de Pearson. Il n'y a donc que peu de cas pour lesquels on retrouve une linéarité uniquement de rangs. Ainsi on peut déjà savoir a priori que l'anamorphose apportera peu d'amélioration au filtre de Kalman d'ensemble.

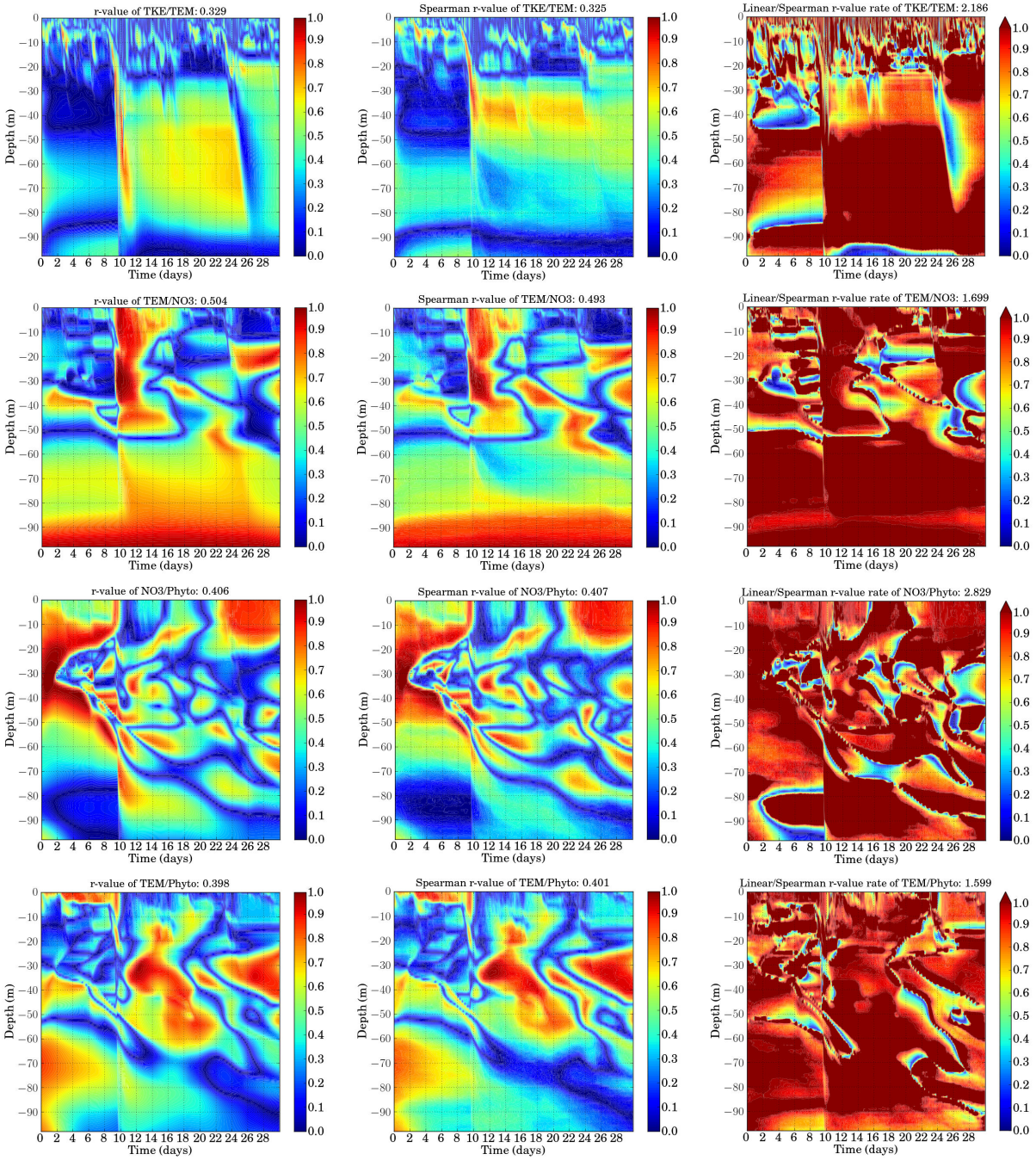

Figure *6.1* – *Graphiques temps/profondeur de la valeur absolue du coefficient $r_{X,Y}$ (colonne de gauche), du coefficient $\rho_{X,Y}$ (colonne du centre) et du ratio des deux $\frac{r_{X,Y}}{\rho_{X,Y}}$ (colonne de droite) pour les variables $X,Y = TKE,TEM$ (première ligne), $X,Y = TEM,NO3$ (deuxième ligne), $X,Y = NO3,Phyto$ (troisième ligne) et $X,Y = TEM,Phyto$ (quatrième ligne).*

La relation entre TKE et la température, nous indique le rôle dominant de l'incursion d'énergie cinétique à partir du $10^{\text{ème}}$ jour sur la dispersion de température entre 20m et 80m. Ceci confirme le raisonnement apporté à la sous partie précédente. Les variations de température dépendent donc principalement et de manière non-linéaire des variations de TKE. Une autre information qui ressort de ces graphiques est l'importante non-linéarité que traverse la propagation des incertitudes (telle que décrite qualitativement à la section précédente) depuis le vent impactant directement TKE modifiant la structure verticale de la température pour finalement influencer la biogéochimie. À travers cette succession de relations non-linéaires, les perturbations Gaussiennes sur le vent vont progressivement se dégrader au cours de leur propagation et perdre leur nature Gaussienne.

Ce dernier phénomène est constaté dans le paragraphe suivant à l'aide d'un diagnostic de normalité.

Non-Gaussianité des variables

Le coefficient de D'Agostino-Pearson (D'Agostino and Pearson, 1973) combine deux coefficients mesurant l'asymétrie (*skewness*) et le kurtosis d'un échantillon afin d'établir une distance à la Gaussianité. De ce coefficient est crée un test dit test de normalité omnibus de D'Agostino-Pearson qui a comme hypothèse nulle que l'échantillon évalué provient d'une distribution Gaussienne. La p-valeur produite par ce test est une probabilité χ^2 à deux côtés pour l'hypothèse nulle. Cette hypothèse est rejetée avec une significativité de 95% pour une p-valeur inférieure à 0.05. Plus la p-valeur est proche de 0 plus l'échantillon diffère d'un échantillon Gaussien.

La Figure 6.2 montre la p-valeur du test de normalité omnibus de D'Agostino-Pearson en fonction de la profondeur (en mètres sur les 100 premiers mètres de la colonne d'eau) et du temps (en jours sur le mois), pour TKE (panneau supérieur gauche), la température (panneau supérieur droit), le nitrate (panneau inférieur gauche) et le phytoplancton (panneau inférieur droit). Les p-valeurs sont données entre 0 et 0.05, les zones en blanc sont donc les zones où l'échantillon peut être considéré comme provenant d'une distribution Gaussienne.

La variable TKE est très non-Gaussienne sur presque toute la colonne d'eau. Ceci est dû à la faible dispersion d'énergie cinétique turbulente puisque sans événement de fort vent le terme TKE est quasiment nul pour tous les membres. On peut toutefois voir les quelques incursions de TKE en surface et leur distribution Gaussienne, sous l'effet du vent (avec un événement fort vers le $10^{\text{ème}}$ jour). Bien que la température soit très Gaussienne dans les 10 premiers jours on constate que la non-Gaussianité se propage avec l'approfondissement des couches de température de surface entre 40m et 90m dans les 20 derniers jours du mois. Ce *pattern* (motif) correspond au *pattern* de forte correlation, entre TKE et la température, observée sur le graphique

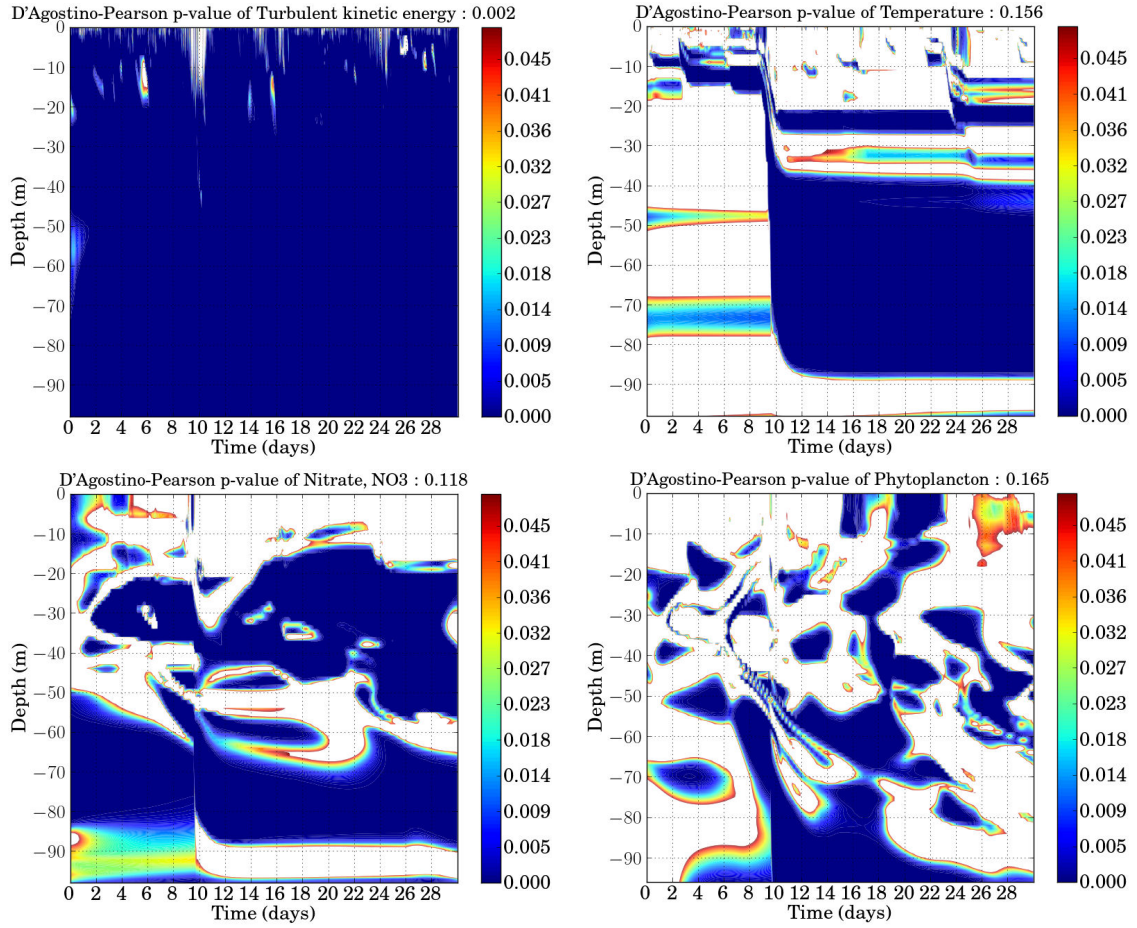

Figure *6.2* – *Graphique temps/profondeur de la p-valeur du test de normalité de D'Agostino-Pearson pour les variables TKE (panneau supérieur gauche), T (panneau supérieur droit), NO3 (panneau inférieur gauche) et le total de phytoplancton (panneau inférieur droit).*

de la Figure 6.1 (première ligne, première colonne). Ceci est en accord avec le raisonnement de la sous partie précédente sur la génération de non-Gaussianité par l'incursion de différentes intensités du mélange TKE en profondeur. Les variables biogéochimiques (le nitrate et le phytoplancton) présentent des caractères que l'on peut alternativement considérer comme Gaussiens ou très non-Gaussiens répartis de manière hétérogène sur la verticale et dans le temps. Cette répartition s'explique par la succession de phénomènes dominants décrits dans la section précédente. Il ressort également de la Figure 6.2 que l'utilisation d'une méthode d'assimilation Gaussienne pourra suffire à contrôler la température dans une grande partie du domaine. Le

contrôle du phytoplancton est plus difficile avec des distributions alternativement Gaussienne et très non-Gaussienne.

6.1.2 Bilan

La complexité anticipée des relations entre les variables a été confirmée par l'étude des non-linéarités inter-variables qui rentrent en jeu dans notre système. Ces non-linéarités vont avoir deux impacts sur l'assimilation de données. Tout d'abord, la relation non-linéaire (affine et de rangs) entre les variables peut rendre difficile la propagation des corrections de l'assimilation des variables observées aux variables non observées par des méthodes telles que l'EnKF, l'ETKF, le RHF ou encore l'EnKF anamorphosé. D'autre part, ces non-linéarités vont dégrader la Gaussianité des distributions de probabilité du système. Ce dernier point a été confirmé par l'étude de normalité de l'ensemble. Ceci mettra en défaut l'hypothèse de Gaussianité émise par les méthodes d'assimilation aux moindres carrés.

Il ressort de cette étude que le contrôle de la température et de la salinité en observant la température et la salinité s'apparente à un problème peu non-linéaire et quasi-Gaussien. Alors que le contrôle de la biogéochimie (e.g. le phytoplancton) s'avère plus compliqué et peut être qualifié de problème fortement non-Gaussien.

6.2 Contrôler la dynamique, un problème quasi-Gaussien

Pour cette première expérience d'assimilation de données avec le modèle ModECO-GeL nous nous plaçons dans le cadre décrit au chapitre précédent. Nous souhaitons contrôler sur le mois d'avril 2006, la température (T) et la salinité (S) à partir de profils de (T,S) et d'observations satellites de (T). Les profils sont disponibles tous les deux jours et la fréquence des observations satellites varie de 6h, 3h, 1h à 30min (voir le Tableau 5.1 pour la nomenclature des configurations).

Il a été constaté dans la section précédente qu'un tel problème s'apparentait à un probème peu linéaire et quasi-Gaussien. Ainsi, dans cette section, nous utilisons le filtre de Kalman d'ensemble transformé (ETKF) en s'attendant à de bonnes performances.

Après avoir réalisé de nombreuses assimilations, nous observons à l'aide de plusieurs diagnostics le comportement de l'ETKF dans ce contexte et en fonction de la fréquence d'observations satellites.

Dans une première sous partie, nous évaluons l'impact de l'assimilation sur la partie dynamique du système (température et salinité). Comme le suggère la Section 2.3.1 sur l'évaluation d'une assimilation d'ensemble, nous évaluons non seulement l'estimé moyen (l'analyse) mais aussi la qualité de l'ensemble généré par l'ETKF. Dans une deuxième sous partie, nous regardons l'impact qu'une correction des variables dyna-

miques a de manière indirecte, i.e. via l'équilibre du modèle, sur la biogéochimie.

6.2.1 Impact sur la dynamique du modèle

Dans cette sous partie, on effectue le contrôle de la température et de la salinité après observation de profil de température et de salinité ainsi que des données plus fréquentes de température de surface (SST). D'après la section précédente, on sait que la variable température, en particulier, n'est pas totalement Gaussienne (après 10 jours entre 40 et 90 mètres de profondeur). Néanmoins, le contrôle direct de la variable T observée le long des profils est à la portée d'une assimilation Gaussienne telle que l'ETKF. La question se pose toutefois pour la propagation de l'information provenant de la température de surface observée dans les couches plus profondes.

Une fréquence plus élevée d'observations satellites de la température permet-elle, à une assimilation Gaussienne (e.g. l'ETKF), d'améliorer l'estimation de la dynamique ?

Nous portons donc, ici, une attention particulière aux différences de corrections entre les configurations d'observation (Tableau 5.1).

Précision de l'état d'analyse

La température est une des quantités d'importance majeure dans un système de modélisation des océans. Il s'agit de bien l'estimer pour correctement décrire la dynamique océanique. De plus, nous avons vu dans la section précédente l'impact des variations de température sur les variables biogéochimiques, en particulier le nitrate et le phytoplancton.

Diagrammes de Hovmöller Dans un premier temps, nous observons qualitativement les états analysés produits par une assimilation ETKF en fonction du type et de la fréquence des observations disponibles.

La Figure 6.3 représente un diagramme de Hovmöller de la variable température pour la *vérité* (ligne 1, colonne 1), pour la moyenne de l'ensemble libre (ligne 1, colonne 2), pour les états moyens analysés par l'ETKF dans les configurations P2S0 (ligne 2, colonne 1), P2S6 (ligne 2, colonne 2), P2S05 (ligne 3, colonne 1) et P0S05 (ligne 3, colonne 2). Il s'agit de la température le long de la colonne d'eau (de 0m à 100m) au cours du mois d'avril. En comparant, la *vérité* à la moyenne d'ensemble libre, on constate que plusieurs phénomènes ne sont pas correctement représentés par

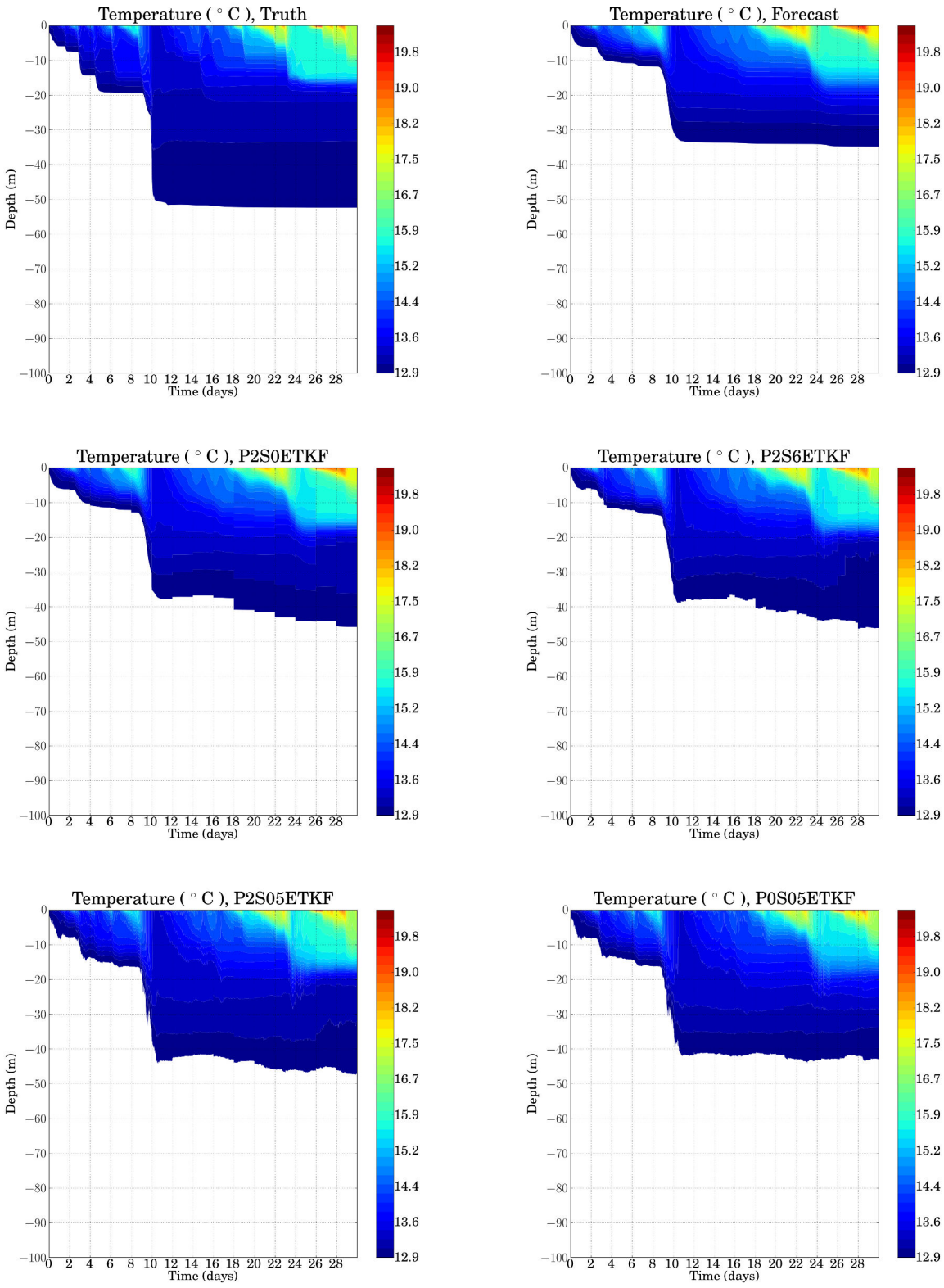

Figure *6.3* – *Diagrammes de Hovmöller : Graphiques de la variable température T pour la vérité (panneau supérieur-gauche), pour l'ensemble libre (panneau supérieur-droit), pour l'analyse de l'ETKF avec le réseau d'observations P2S0 (panneau centre-gauche), P2S6 (panneau centre-droit), P2S05 (panneau inférieur-gauche) et P0S05 (panneau inférieur-droit) sur la grille Temps/Profondeur. La partie blanche correspond au températures inférieures à 12.9 ° C.*

les processus stochastiques générant l'ensemble. Les températures sont plus basses entre 10m et 20m dans les 10 premiers jours. On observe sur le diagramme de Hovmöller de la *vérité* un approfondissement des contours d'iso-température dans le temps avec un fort approfondissement le $10^{\text{ème}}$ jour. Cet approfondissement d'eau plus chaude est bien moins important pour l'ensemble, ce qui est probablement dû à un événement éolien mal représenté. De plus, la finesse des variations temporelles de la température de surface de la *vérité* n'est pas bien traduite par la moyenne d'ensemble.

L'apport de l'assimilation de profils pour améliorer la représentation ensembliste du système, s'apprécie en étudiant la configuration P2S0 (ligne 2, colonne 1). L'approfondissement à partir du $10^{\text{ème}}$ jour (à noter qu'un profil est disponible au $10^{\text{ème}}$ jour) est nettement mieux estimé par l'analyse ETKF avec par exemple l'iso-température 12.9°C qui descend à 45m de profondeur en fin de mois. En revanche, les 10 premiers jours ainsi que la finesse des variations temporelles en surface ne sont que très peu améliorés par les données de profils.

La comparaison des trois autres configurations (P2S6, P2S05 et P0S05) nous permet d'évaluer l'impact des données satellite. Assez naturellement, les corrections en surface sont plus précises en observant la surface toutes les 30min plutôt que toutes les 6h. La propagation sur la verticale de l'information de surface par l'ETKF se fait bien, notamment lorsque l'on regarde les 10 premiers jours avec un approfondissement de 10m à 15m de l'iso-température 12.9°C. De plus, sur cette même iso-température à partir de 10 jours on constate que la correction a un impact à des profondeurs plus grandes avec un approfondissement allant jusqu'à 45m au lieu de 35m au $10^{\text{ème}}$ jour. On constate parallélement que l'assimilation sans profil (P0S05) donne de bons résultats. Bien qu'il puisse s'agir d'un épiphénomène, les profils n'ont que peu d'impact sur la qualité de l'état moyen analysé.

Erreurs RMS Pour avoir une idée plus quantitative des réductions d'erreurs permises par l'assimilation, nous regardons l'erreur Root Mean Square (erreur RMS ou RMSE) par rapport à la trajectoire *vérité*. Il s'agit de calculer l'erreur quadratique pour la variable température, soit moyennée dans le temps, soit moyennée dans l'espace.

La Figure 6.4 présente les erreurs RMS moyennées en temps en fonction de la verticale (les 4 graphiques supérieurs) et moyennées dans l'espace en fonction du temps (les 4 graphiques inférieurs). Chacune de ces erreurs est donnée pour les 4 configurations d'assimilation précédentes : P2S0 (graphique supérieur gauche), P2S6 (graphique supérieur droit), P2S05 (graphique inférieur gauche) et P0S05 (graphique inférieur droit). La courbe rouge est l'écart RMS à la *vérité* de l'ensemble libre et la courbe bleu celui de l'ensemble analysé par l'ETKF.

Les 4 graphiques inférieurs nous confirment les nettes réductions d'erreurs dues aux observations satellites hautes fréquences : dans un premier temps sur la période des 10 premiers jours ; puis dans un second temps sur la période après le $10^{\text{ème}}$ jour. Les 4 graphiques supérieurs précisent quant à eux, que ces réductions se situent en deux régions de l'espace. Le premier est sur les premiers 10-15 mètres, ce qui correspond au raffinement des variations de la température de surface. Le second est l'approfondissement des eaux plus chaudes à partir du $10^{\text{ème}}$ jour entre 25m et 100m.

Pour avoir une vision plus globale de ces erreurs, la Figure 6.5 montre les erreurs RMS moyennées en temps et en espace en fonction de toutes les configurations de réseaux d'observations. Le graphique de gauche correspond aux erreurs RMS totales de température et celui de droite de la salinité.

Cette dernière figure confirme dans un premier temps le fort apport des données de profils seules (différence entre P0S0 et P2S0). Puis les améliorations en terme d'erreurs RMS sont confirmées sur toute l'expérience pour des données satellites de plus en plus fréquentes. A noter que la configuration P2S3, bien que non aberrante, ne permet pas de réduire les erreurs par rapport à P2S6. Ceci est probablement dû à des perturbations (lors de la création des observations artificielles) anormalement fortes sur les observations 3h. Les performances de l'état analysé pour une assimilation de données satellite uniquement apparaissent très bonnes ici encore. Enfin, le graphique de droite nous permet de voir que les tendances constatées sur la variable température sont similaires sur la salinité.

Qualité de l'ensemble

Comme on l'a mentionné déjà à plusieurs reprises, obtenir une bonne estimation moyenne n'est plus la seule caractéristique requise d'une assimilation d'ensemble. L'ensemble fournit une information supplémentaire sur la densité du système considéré. Il faut donc également évaluer la qualité de l'ensemble en lui-même. Ce que l'on fait ici en utilisant des histogrammes de rangs, diagnostics présentés au Chapitre 1 (Sec. 2.3.1). Il est de bon ton de repréciser que, dans un cadre d'assimilation aussi réaliste avec notamment la présence de nombreux biais, il n'est pas possible d'obtenir des histogrammes de rangs parfaitement plats. Ce diagnostic est donc utiliser de manière comparative et permet de constater la réduction des biais et l'amélioration de la dispersion.

La Figure 6.6 présente des histogrammes de rangs réalisés sur tout le mois et sur les 200 premiers mètres. Les ensembles diagnostiqués sont les ensembles après

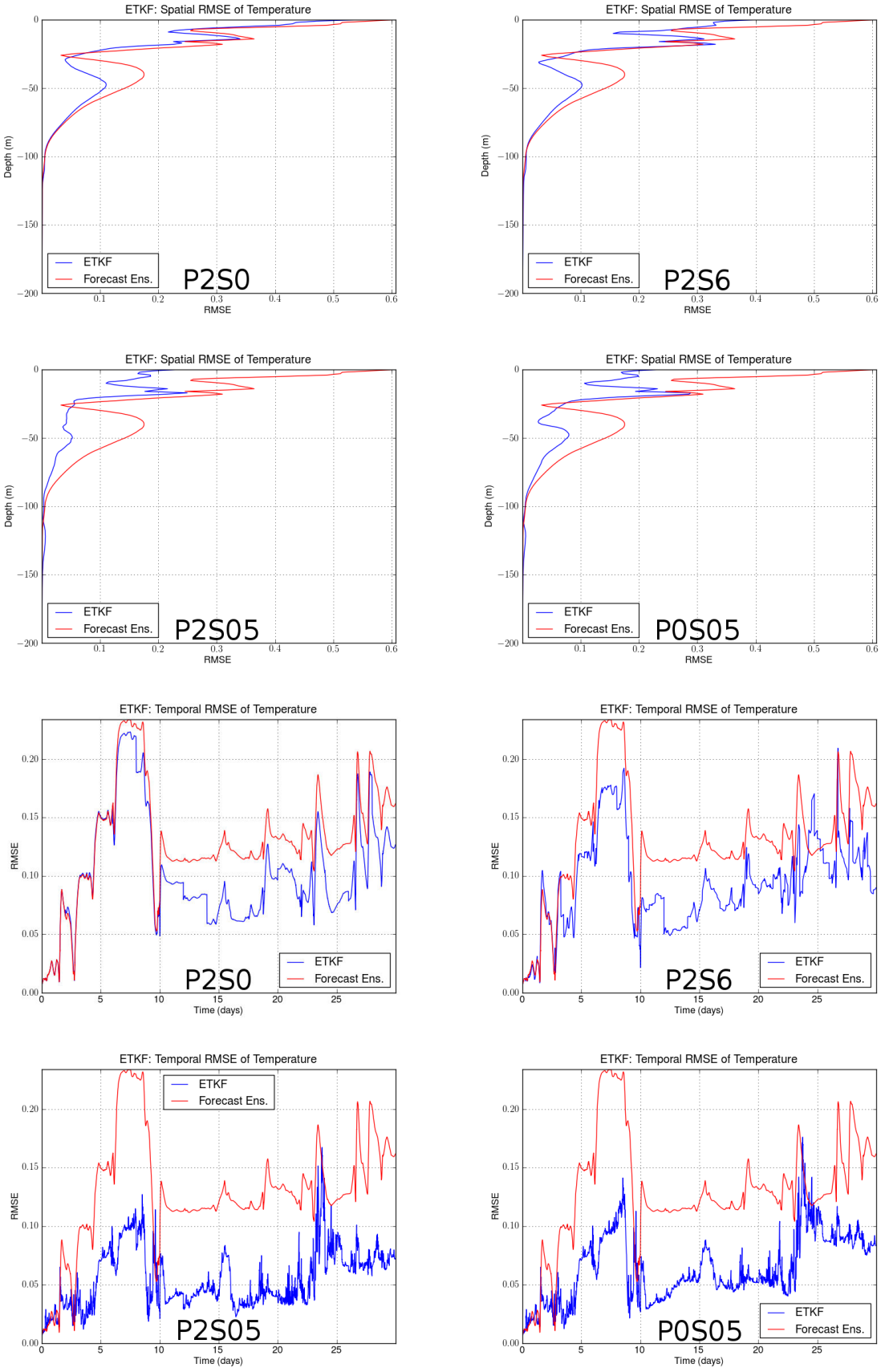

Figure *6.4* – *RMSE spatiale (panneaux supérieurs) et temporelles (panneaux inférieurs) sur le mois de la variable température T pour les réseaux d'observations P2S0 (supérieur gauche), P2S6 (supérieur droit), P2S05 (inférieur gauche) et P0S05 (inférieur droit). Les membres de l'ensemble du run libre sont les lignes bleues et la trajectoire de référence est la ligne rouge.*

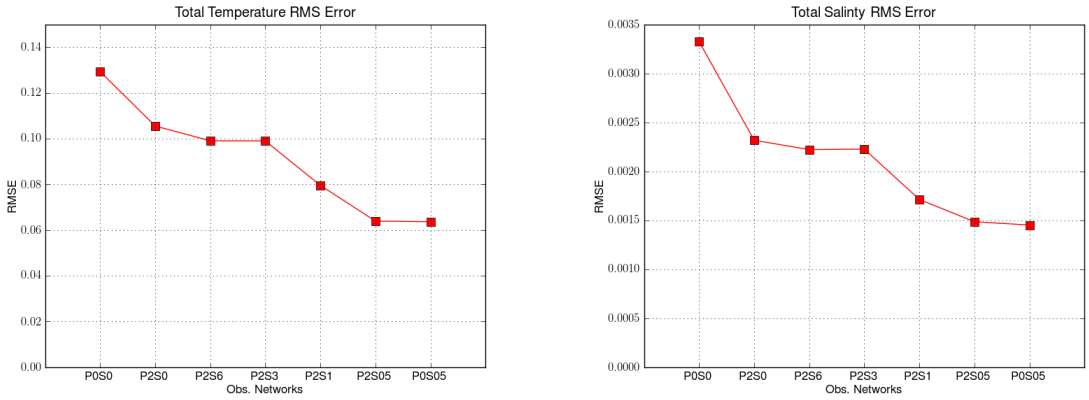

Figure *6.5* – *RMSE totale de la température T (panneau de gauche) et de la salinité S (panneau de droite) de l'analyse ETKF pour chaque configuration.*

assimilation ETKF dans les configurations : P2S0 (ligne 1, colonne 1), P2S6 (ligne 1, colonne 2), P2S3 (ligne 1, colonne 3), P2S1 (ligne 2, colonne 1), P2S05 (ligne 2, colonne 2) et P0S05 (ligne 2, colonne 3). L'ensemble a priori comprend un fort biais en sous estimant souvent la *vérité* à partir du fort épisode de vent du $10^{\text{ème}}$ jour (zone rouge sur la Fig. 5.13, Chapitre 5). Ce biais se retrouve sur l'histogramme de rangs de l'assimilation de profils seuls, ce qui témoigne que les profils ne sont pas suffisants pour débiaiser et mieux disperser l'ensemble. Plus les observations satellites sont fréquentes mieux le biais sera corrigé. En configuration P2S1 (ligne 2, colonne 1), le biais est presque entierement corrigé mais on voit apparaitre une sous dispersion (par une augmentation d'occurences dans les premières et dernières *bins*). La configuration P2S05 (ligne 2, colonne 2), permet de nettement améliorer la dispersion de l'ensemble. Enfin, la configuration sans profils (P0S05 : ligne 2, colonne 3) qui jusque là présentait un très bon état moyen, montre une très forte sous-dispersion. Cette sous-dispersion est probablement due à un effondrement de l'ensemble. Il est donc important de noter que l'utilisation des données profils permettent de débiaiser l'ensemble en profondeur ce qui engendre un histogramme de rangs aplati.

En résumé, l'utilisation de profils réduit fortement les erreurs d'estimation en profondeur et permet d'éviter un effondrement de l'ensemble. Pour obtenir une plus grande finesse dans l'estimation, les observations satellites sont nécessaires. L'ETKF réussit à bien propager l'information de surface, ce qui permet d'apporter des corrections importantes jusqu'à 100m de profondeur. Enfin, dans ce contexte quasi-Gaussien de correction de la température par observation de la température, l'aug-

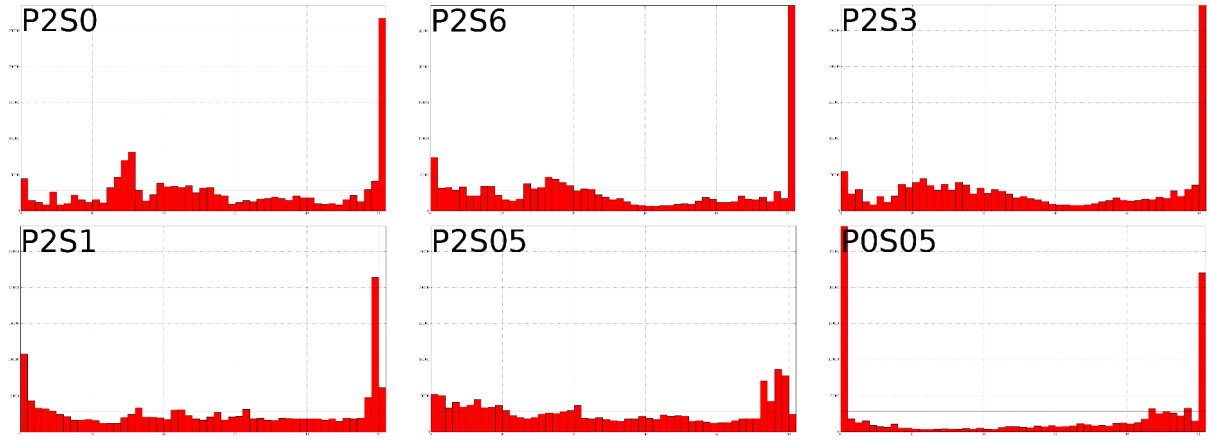

Figure *6.6* – *Histogrammes de rangs, sur le mois et les 200 premiers mètres, de la variable température T pour les réseaux d'observations P2S0 (ligne 1, colonne 1), P2S6 (ligne 1, colonne 2), P2S3 (ligne 1, colonne 3), P2S1 (ligne 2, colonne 1), P2S05 (ligne 2, colonne 2) et P0S05 (ligne 2, colonne 3).*

mentation de la fréquence d'observations satellites est bénéfique à l'assimilation aussi bien pour l'estimation de l'état que pour la représentation ensembliste.

6.2.2 Impact indirect sur la biogéochimie

L'un des intérêts des systèmes couplés est d'utiliser la meilleure représentation d'une partie du système pour améliorer la partie couplée. Ainsi dans notre cas, on est en droit de s'attendre - après corrections de variables dynamiques par assimilation de données - à une amélioration de la partie biogéochimique du système. Cette sous-partie se consacre à évaluer les effets (via la prévision) des corrections dynamiques sur la biogéochimie.

Le cas hypothétique d'un contrôle parfait de la dynamique Nous souhaitons vérifier s'il est possible d'améliorer la biogéochimie en contrôlant la dynamique. Pour ce faire, nous remplaçons sur la période d'assimilation (le mois d'avril) le vent perturbé par le vent haute-fréquence de référence. Ceci a pour effet de rapidement (quelques pas de temps) rétablir une dynamique quasiment identique à la *vérité*. Ce qui est équivalent à un contrôle parfait de la dynamique.

La Figure 6.7 montre les RMSE de cette simulation (en bleu) et de l'ensemble libre (en rouge) sur la variable NO_3 (à gauche) et sur la somme de phytoplancton (à droite) sur le mois d'avril (graphiques supérieurs) et sur la colonne d'eau (graphiques inférieurs). Les erreurs RMS totales sur la variable NO_3 et sur la somme de phyto-

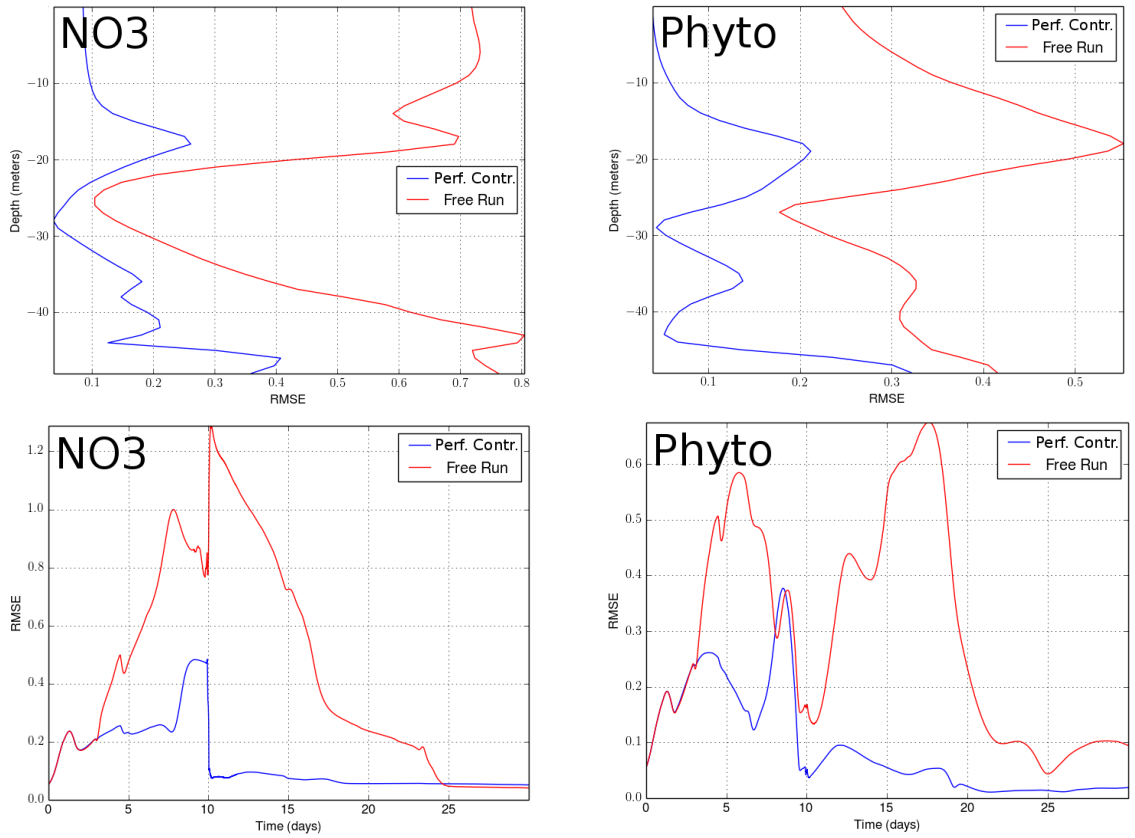

Figure *6.7* – *RMSE spatiale (panneaux supérieurs) et temporelles (panneaux inférieurs) sur le mois de la variable NO_3 (à gauche) et la somme de phytoplancton (à droite) pour l'expérience du cas hypothétique d'un contrôle parfait de la dynamique. Les membres de l'ensemble du run libre sont les lignes bleues et la trajectoire de référence est la ligne rouge.*

plancton sont respectivement 0.129 et 0.090. On observe une nette amélioration des erreurs RMS pour ces deux quantités biogéochimiques lorsque la dynamique est bien contrôlée. L'erreur résiduelle peut s'expliquer par les différentes réactions des états biogéochimiques des membres d'ensemble face à une même dynamique océanique.

Cette expérience nous apprend qu'il est possible d'améliorer la biogéochimie simplement à travers l'équilibre du modèle pour une dynamique parfaitement décrite. Bien entendu, ce cas (hypothétique) de contrôle parfait n'est pas réaliste. Dans les deux paragraphes suivants nous observons l'impact des corrections apportées par l'assimilation ETKF de la dynamique dans les différentes configurations de réseaux

d'observations.

Erreurs RMS totales Nous évaluons, ici, les erreurs sur deux variables biogéochimiques : le nitrate (NO_3) et le phytoplancton (Phyto). La Figure 6.8 est similaire à la Figure 6.5 mais pour le phytoplancton (à gauche) et le nitrate (à droite). Ont été ajoutées les valeurs RMSE (en lignes pointillées) de phytoplancton (0.129) et de nitrate (0.090) obtenues dans le cas hypothétique d'un contrôle parfait (Fig. 6.7).

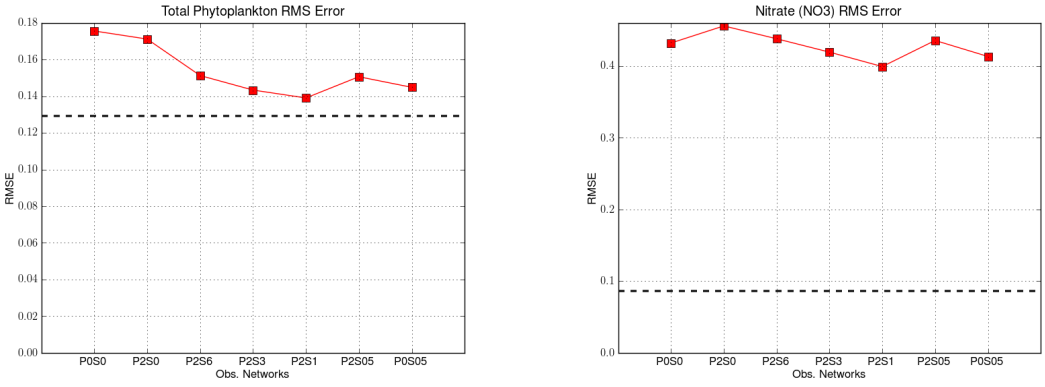

Figure *6.8* – *RMSE totale de la somme de phytoplancton (graphique gauche) et de nitrate (graphique droit) de l'analyse ETKF pour chaque configuration. Les lignes pointillées représentent les valeurs RMSE de phytoplancton (0.129) et de nitrate (0.090) obtenues dans le cas hypothétique d'un contrôle parfait (Fig. 6.7).*

Le phytoplancton décroît très légérement avec l'utilisation de profils (entre P0S0 et P2S0). Les données satellites permettent de réduire encore un peu les erreurs sur le phytoplancton. Cependant augmenter la fréquence d'observations (de P2S6 à P2S05) ne fait que peu varier les erreurs. La RMSE du phytoplancton obtenue en configuration P2S1 (0.14) est pourtant très proche de la RMSE obtenue dans le cas hypothétique d'un contrôle parfait (0.129). Les modifications de température apportées par l'assimilation n'ont donc pas assez d'impact dans l'évolution du système pour améliorer la prévision de phytoplancton.

Les erreurs de concentration de nitrate sont encore moins impactées par la correction des variables dynamiques. Ni profils ni observations satellites ne font décroître les erreurs RMS de nitrate en dessous de 0.4, ce qui est encore très loin de l'impact d'une dynamique parfaite (0.09).

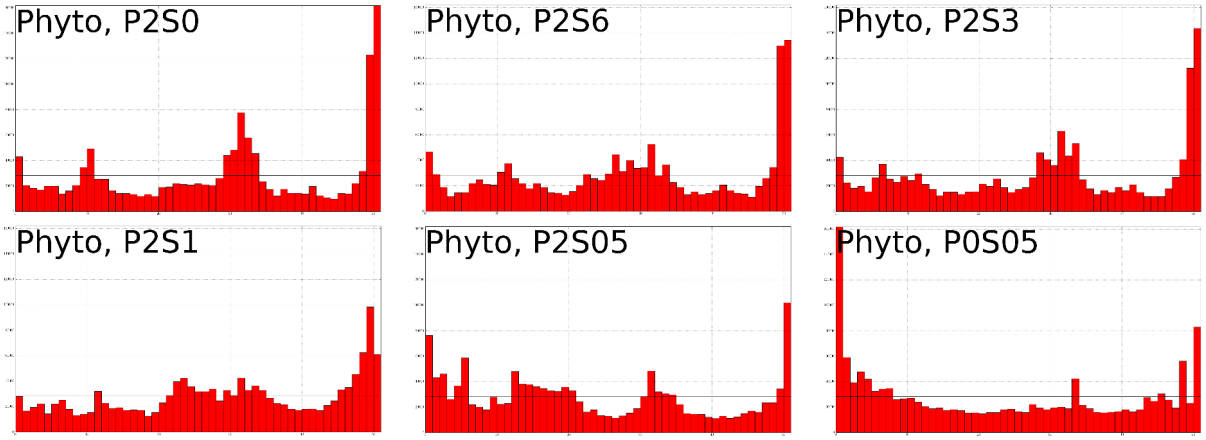

Figure *6.9* – *Histogrammes de rangs, sur le mois et les 200 premiers mètres, de la somme des variables phytoplancton (PicP,NanP,MicP) pour les réseaux d'observations P2S0 (ligne 1, colonne 1), P2S6 (ligne 1, colonne 2), P2S3 (ligne 1, colonne 3), P2S1 (ligne 2, colonne 1), P2S05 (ligne 2, colonne 2) et P0S05 (ligne 2, colonne 3).*

Histogrammes de rangs Après avoir constaté la mauvaise estimation des états biogéochimiques moyens engendrés par l'ETKF on peut toujours évaluer son impact sur l'ensemble biogéochimique. Les Figures 6.9 et 6.10 représentent les mêmes histogrammes de rangs que la Figure 6.6 pour le phytoplancton et le nitrate respectivement.

Si l'on regarde la Figure 6.9, on voit que le biais de sous-estimation du phytoplancton que présente l'ensemble n'utilisant que des profils est diminué par les données satellites hautes-fréquences (P2S1, P2S05 et P0S05). Cependant, les histogrammes obtenus ne sont pas plats. L'ensemble reste mal dispersé. Le dernier histogramme (ligne 2, colonne 3) nous confirme également la nécessité des données profils pour éviter une sous dispersion de l'ensemble.

La Figure 6.10 permet de tirer les mêmes conclusions, avec la correction d'un biais de sur-estimation du nitrate. La dispersion est encore mauvaise malgré l'utilisation des données très hautes-fréquences. Il est toutefois à noter que le nitrate semble moins dépendant des données profils puisque les histogrammes pour P2S05 et P0S05 ne diffèrent que peu.

Aussi bien sur le plan de la précision de l'estimé moyen que sur le plan de la qualité de l'ensemble, l'assimilation ETKF corrigeant la température et la salinité n'a pas ou peu d'impact sur la partie biogéochimique du système. Ceci peut soit être dû à un mauvais équilibre (*balance*) des états-membres générés par l'analyse soit

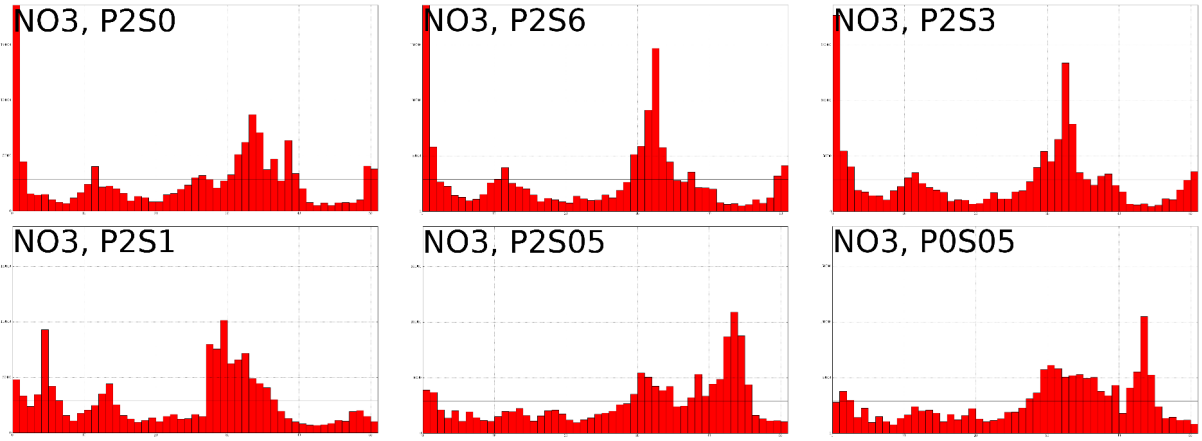

Figure *6.10* – *Histogrammes de rangs, sur le mois et les 200 premiers mètres, de la variable nitrate (NO3) pour les réseaux d'observations P2S0 (ligne 1, colonne 1), P2S6 (ligne 1, colonne 2), P2S3 (ligne 1, colonne 3), P2S1 (ligne 2, colonne 1), P2S05 (ligne 2, colonne 2) et P0S05 (ligne 2, colonne 3).*

les modifications de la température et de la salinité ne sont pas assez fortes (face au forçage et à la dynamique basse fréquence du système) pour avoir un effet sur la biogéochimie. Bien que la première possibilité ne soit pas à écarter, nous n'investigons pas davantages la question. Par ailleurs, en considérant la seconde possibilité, nous pouvons alors contrôler une variable du système biogéochimique pour corriger ce dernier. C'est ce qui sera fait dans les deux sections suivantes.

6.2.3 Bilan

La première sous partie présente les bonnes performances apportées par l'impact direct des corrections de l'ETKF sur la dynamique. Une plus grande fréquence d'observations satellites produit de meilleures corrections notamment en réduisant fortement les biais. En effet le filtre transformé ETKF, fait décroître les scores RMSE sur les variables dynamiques et améliore la dispersion de l'ensemble. Ces résultats confirment la quasi-Gaussianité du problème. Observer la température et la salinité pour contrôler la température et la salinité est un problème d'estimation assez direct (peu non-linéaire, peu non-Gaussien) qu'un filtre Gaussien peut résoudre. Cependant, la deuxième sous partie nous montre que, dans ce système, contrôler la dynamique seulement (aucune variable biogéochimique dans le vecteur de contrôle) ne permet pas d'impacter la partie biogéochimique. D'où la nécessité d'inclure dans le vecteur de contrôle (toujours en n'observant que les variables dynamiques de température et

de salinité) des variables biogéochimiques. Nous étudions ce problème dans la section suivante.

6.3 Contrôler la biogéochimie, un problème non-Gaussien

Nous venons de voir dans la section précédente qu'une méthode d'assimilation Gaussienne telle que l'ETKF produit de bons résultats sur la dynamique du modèle en contrôlant la dynamique. Cependant ces corrections n'ont que très peu d'impact via le modèle sur les variables biogéochimiques. Il apparait donc nécessaire de contrôler également une partie du système biogéochimique. Dans cette section, pour les mêmes observations nous contrôlons la température (T), la salinité (S) et les trois types de phytoplancton (PicP, NanP, MicP).

Par ailleurs, la Section 6.1 nous a fourni des éléments pour caractériser ce nouveau problème d'assimilation comme non linéaire et non-Gaussien. Dans ce cadre là, les hypothèses de linéarité inter-variables et de Gaussianité émises par l'ETKF ont de grandes chances d'être enfreintes. Nous pouvons donc craindre de mauvaises estimations par l'ETKF lors du contrôle du phytoplancton.

L'objectif de cette section est donc de mettre également en place une méthode dite non-Gaussienne : le Multivariate Rank Histogram Filter (MRHF, Chapitre 4 : Metref et al., 2014), afin d'évaluer et de comparer les performances des deux méthodes.

Puisqu'ici l'étude des différents réseaux d'observations n'est pas centrale à notre investigation, nous nous plaçons dans la configuration P2S6, c'est à dire assimilation de profils à deux jours et de données satellites à 6h. Nous ferons de nouveau varier les réseaux d'observations dans la section suivante.

Il est également à noter que pour des raisons de temps calculs, nous assimilerons toujours les profils en utilisant l'ETKF. Ce que l'on appelle par la suite assimilation MRHF est l'assimilation combinée des profils par l'ETKF et des données satellites par le MRHF.

Il s'agit donc ici d'évaluer et de comparer deux méthodes d'assimilation - une Gaussienne et une non-Gaussienne - dans le cadre d'un contrôle de variables dynamiques et biogéochimiques (phytoplancton) par des observations de températures de surface (SST).

6.3.1 Le contrôle du nitrate

Comme l'indique l'étude statistique réalisée en première section de ce chapitre (Sec. 6.1.1), les corrélations linéaires entre température et nitrate sont plus fortes qu'entre le phytoplancton et la température. L'information issue des observations de SST sera donc relativement bien propagée au nitrate. De plus le nitrate en surface est Gaussien selon le critère du test de D'Agostino-Pearson. La question que l'on

peut se poser est : pourquoi ne pas contrôler le nitrate avec un filtre aux moindres carrés comme l'ETKF ?

A priori, il est envisageable de contrôler le nitrate, pour transmettre l'information à la partie biogéochimie du système. Des tests ont été effectués (non montrés ici) et effectivement le contrôle du nitrate améliore l'estimation du nitrate. Cependant l'impact indirect (via le modèle) de la correction du nitrate n'améliore que peu l'estimation du phytoplancton.

Comme nous nous sommes donnés pour objectif principal de mieux représenter le phytoplancton, nous abandonnons cette démarche et nous nous consacrons au contrôle du phytoplancton directement.

6.3.2 Les méthodes non-Gaussiennes à écarter

L'utilisation de nombreuses méthodes d'assimilation de données non-Gaussienne, ayant chacune leurs spécificités, est possible. Nous utilisons dans cette section et la suivante la méthode MRHF décrite au chapitre 4. Plusieurs tests ont été menés dans ce cadre d'expérience utilisant les différentes méthodes décrites et présentées en Section 2.3. Ces résultats ne sont pas présentés dans le manuscrit. Il convient cependant d'évoquer pourquoi ces différentes méthodes ne sont pas pertinentes dans le cas présent et d'expliquer la raison du choix du MRHF.

Les méthodes dérivées de l'ETKF, telles que le RHF, utilisant le principe de régression linéaire pour propager les corrections sur les variables non-observées n'apporteront pas de meilleurs résultats que l'ETKF lui même. Ces méthodes n'ont d'intérêt que pour l'assimilation de variables observées non-Gaussiennes. Dans notre cas, les variables observées sont la température et la salinité qui peuvent être considérées quasi-Gaussiennes comme l'ont confirmé la Section 6.1 et les bons résultats de l'ETKF en Section 6.2.

L'utilisation d'un filtre de Kalman d'ensemble anamorphosé pourrait être intéressante dans ce cadre puisque l'anamorphose s'applique à toutes les variables de contrôle (variables observées et non-observées). Ceci dit, la comparaison entre la linéarité au sens de Pearson et la linéarité au sens de Spearman en particulier pour les variables de température et de phytoplancton, nous indique que l'anamorphose ne suffira pas à retrouver des densités Gaussiennes (Sec. 6.1). Un filtre de Kalman anamorphosé, apportera peu à l'ETKF.

Le filtre itératif de Kalman d'ensemble (IEnKF) a déjà montré ses capacités à gérer les non-Gaussianités. Cependant, l'IEnKF émet l'hypothèse contrainte forte, i.e. hypothèse d'un modèle parfait. La génération d'un ensemble par une continuelle perturbation du vent agit comme l'application d'une erreur modèle. Le filtre itératif n'arrive pas à converger dans ces circonstances.

Enfin, les filtres particulaires (PF) ne font aucune hypothèse Gaussienne. Ces filtres sont, de plus, particulièrement adaptés aux fortes non-Gaussianités. Par contre, ils sont très dépendants du nombre de membres disponibles pour l'assimilation lorsque la dimension du problème est élevée (Snyder et al., 2008; Metref et al., 2014). Plusieurs tests (non présentés ici) ont confirmé l'effondrement rapide de l'ensemble dans notre problème d'estimation.

Il faut également dire que dans la grande zoologie des méthodes d'assimilation adaptées à traiter la non-Gaussianité certaines méthodes pourraient convenir ici. Notre choix s'est arrêté au MRHF pour la simplicité de sa mise en place ainsi que pour ses résultats prometteurs présentés au Chapitre 4.

6.3.3 Filtre non-Gaussien et estimation de la dynamique

Dans la section précédente nous avons observé les bonnes performances de l'ETKF dans le contrôle de la dynamique du système. Dans un premier temps, nous souhaitons donc évaluer le MRHF qui malgré le caractère quasi-linéaire de ce problème doit en théorie produire des performances similaires.

La Figure 6.11 présente les erreurs RMS en température pour l'état moyen estimé par l'ETKF (graphiques de gauche) et par le MRHF (graphiques de droite). Il s'agit des erreurs moyennées en temps en fonction de la verticale (graphiques supérieurs) et moyennées en espace en fonction du temps (graphiques inférieurs). Les analyses sont en bleu et l'état moyen de l'ensemble libre est en rouge.

Globalement, les erreurs RMS sont très similaires entre l'ETKF et le MRHF. Sur les erreurs moyennées en temps, on peut voir une RMS plus faible de l'ETKF entre 25 et 50 mètres. Par contre l'analyse MRHF est un peu moins erronée en surface. Les erreurs moyennées en espace sont presque identiques dans les 25 premiers jours. Le MRHF produit des erreurs légérement plus faibles sur les 5 derniers jours.

En moyenne temporelle et spatiale, la RMS de l'ETKF est de 0.1295 et celle du MRHF est de 0.1279.

En salinité, la Figure 6.12 conduit aux mêmes conclusions : les erreurs RMS moyennées en temps ou en espace pour les deux méthodes sont très proches.

L'erreur moyenne en temps et en espace est de 0.0031 pour l'ETKF et 0.0030 pour le MRHF.

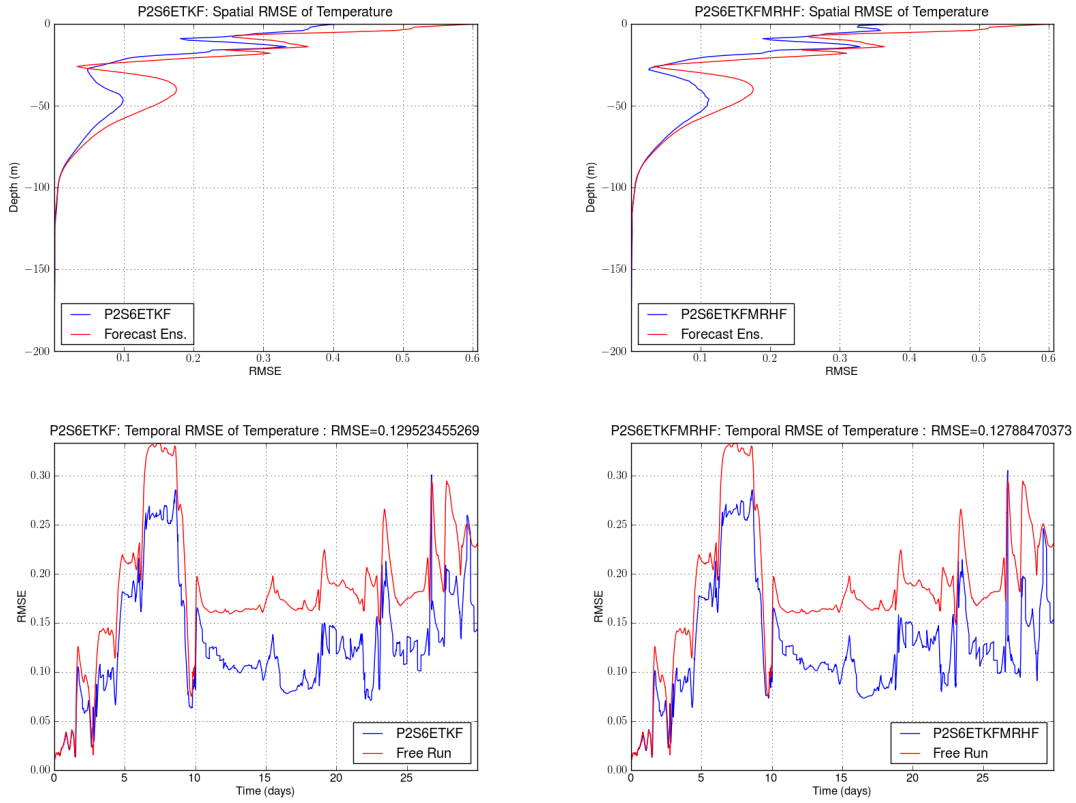

Figure *6.11* – *RMSE de la température en fonction du temps (panneaux supérieurs) et de la verticale (panneaux inférieurs), pour l'analyse de l'ETKF (panneaux de gauche) et pour l'analyse combinée de l'ETKF-MRHF (panneaux de droite) avec l'analyse en bleu et le run libre en rouge.*

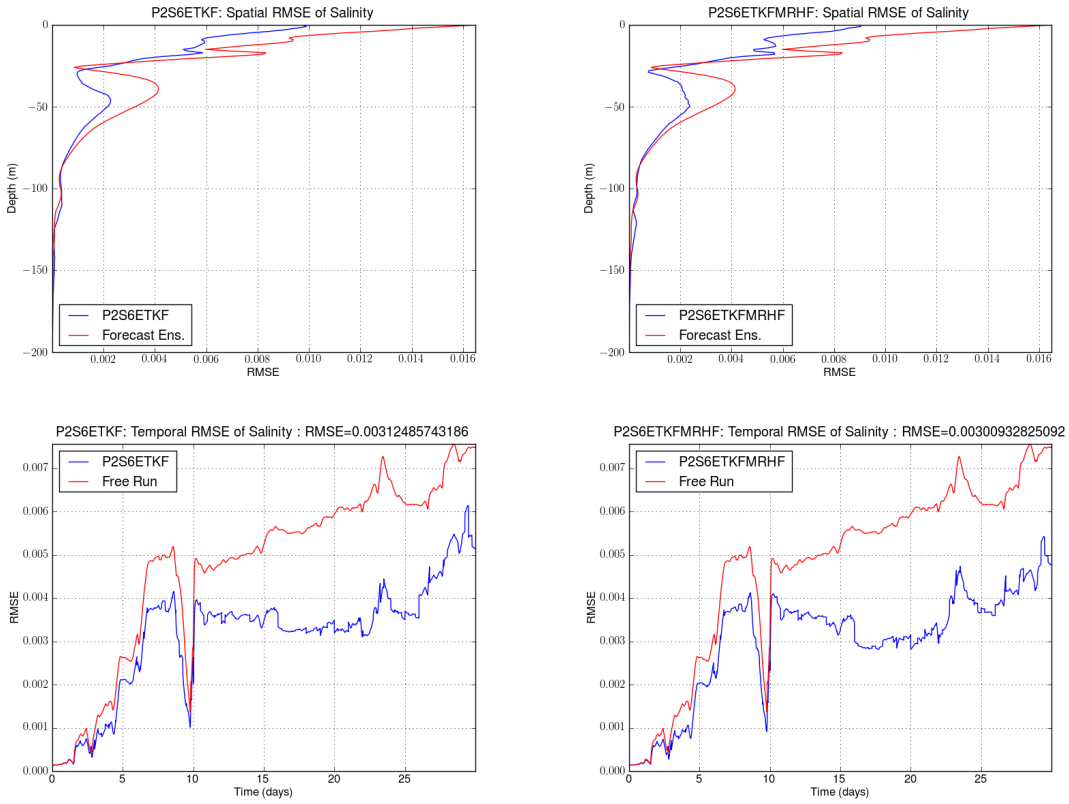

Figure *6.12* – *RMSE de la salinité en fonction du temps (panneaux supérieurs) et de la verticale (panneaux inférieurs), pour l'analyse de l'ETKF (panneaux de gauche) et pour l'analyse combinée de l'ETKF-MRHF (panneaux de droite) avec l'analyse en bleu et le run libre en rouge.*

Dans ce cadre quasi-Gaussien de contrôle de la dynamique, le MRHF et l'ETKF produisent, tous deux, des états moyens estimés de faible erreur. Il s'agit maintenant de voir ce que donne le contrôle du phytoplancton.

6.3.4 Amélioration de l'estimation des variables biogéochimiques

Nous rappelons que dans ces expériences aucune variable du système biogéochimique n'est observée. Les trois types de phytoplancton sont seulement controlés à partir des observations de la dynamique.

Nous regardons dans un premier temps les corrections directes apportées par l'ETKF et le MRHF sur le phytoplancton. Dans un second temps, la propagation

des corrections via le modèle sur le reste du système biogéochimique est regardée en évaluant les erreurs RMS sur le nitrate.

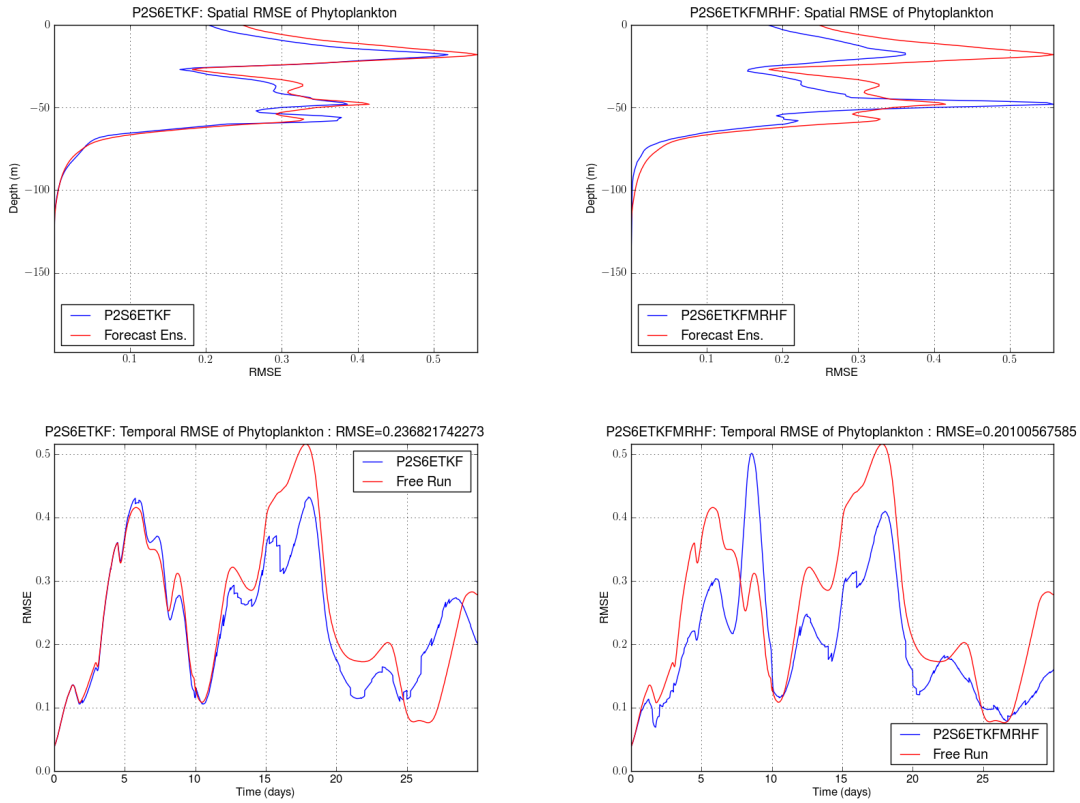

Figure *6.13* – *RMSE du phytoplancton en fonction du temps (panneaux supérieurs) et de la verticale (panneaux inférieurs), pour l'analyse de l'ETKF (panneaux de gauche) et pour l'analyse combinée de l'ETKF-MRHF (panneaux de droite) avec l'analyse en bleu et le run libre en rouge.*

Correction du phytoplancton

Nous évaluons la précision de l'état moyen à l'aide d'erreurs RMS et la qualité de l'ensemble sur la somme des trois variables phytoplanctoniques (PicP, NanP, MicP).

Erreurs RMS La Figure 6.13 est identique aux Figures 6.11 et 6.12 mais pour le phytoplancton total. L'analyse produite par l'ETKF (graphiques de gauche) a une

erreur quasi-égale à celle de l'ensemble libre sur les 10 premiers jours. Le phytoplancton est légérement corrigé les jours suivants sauf pour les 5 derniers jours où l'analyse conduit à des erreurs RMS plus fortes. Sur la verticale, le gain de l'ETKF par rapport à l'ensemble libre est minime.

Le MRHF (graphiques de droite), commet des erreurs RMS importantes entre 7 et 9 jours mais propose une réduction des erreurs pendant tout le reste du mois. Sur la verticale, il y a également un pic d'erreur à 50m alors qu'à toute autre profondeur la réduction d'erreurs est substantielle.

En moyenne spatiale et temporelle, l'ETKF a une erreur RMS de 0.2368 et le MRHF une de 0.2010.

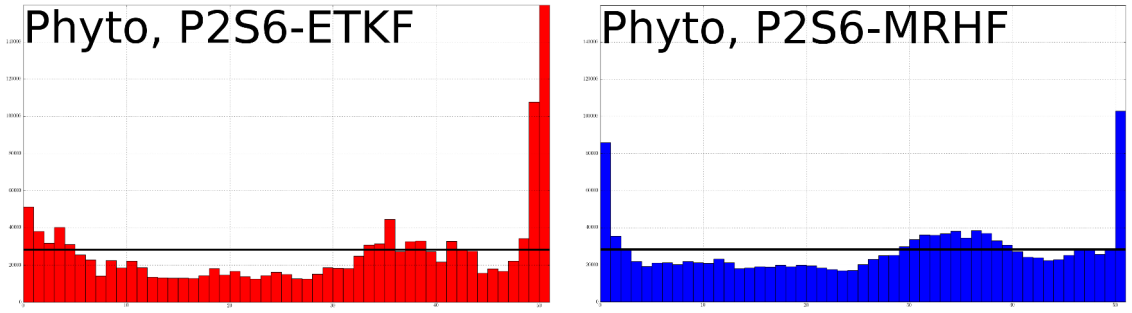

Figure *6.14* – *Histogrammes de rangs, sur le mois et les 200 premiers mètres, de la somme des variables phytoplanctoniques pour l'analyse de l'ETKF (graphique à gauche) et pour l'analyse combinée de l'ETKF-MRHF (graphique à droite) avec le réseau d'observations P2S6.*

Histogrammes de rangs Nous étudions maintenant la dispersion des ensembles que ces deux méthodes engendrent.

La Figure 6.14 est composée de deux histogrammes de rangs de la somme des variables phytoplanctoniques, sur le mois et les 200 premiers mètres de profondeur, pour l'ETKF (graphique de gauche) et pour le MRHF (graphique de droite). L'ensemble généré par l'ETKF n'a pas réussi à corriger le biais de sous-estimation et présente une légère sous dispersion.

A contrario, le MRHF a corrigé totalement ce biais. L'ensemble est toutefois sous dispersif. Ce dernier résultat laisse penser que le MRHF mériterait un réglage plus affiné.

Certaines lacunes du MRHF apparaissent (fortes erreurs RMS ponctuelles et légère sous dispersion de l'ensemble). Il ressort de ces résultats que le MRHF mériterait un réglage plus fin de ses paramètres ou un nombre de membres d'ensemble plus grand.

Malgré cela, les résultats présentés par le MRHF sont bien meilleurs que ceux de l'ETKF. Ceci confirme la qualité non-Gaussienne du contrôle du phytoplancton. La question est alors, est-ce que ces bonnes corrections se répercutent convenablement sur d'autres parties du système biogéochimique.

Le contrôle du phytoplancton par observation de température et de salinité est donc possible et ce contrôle se fait d'autant mieux avec une méthode d'assimilation de données non-Gaussiennes.

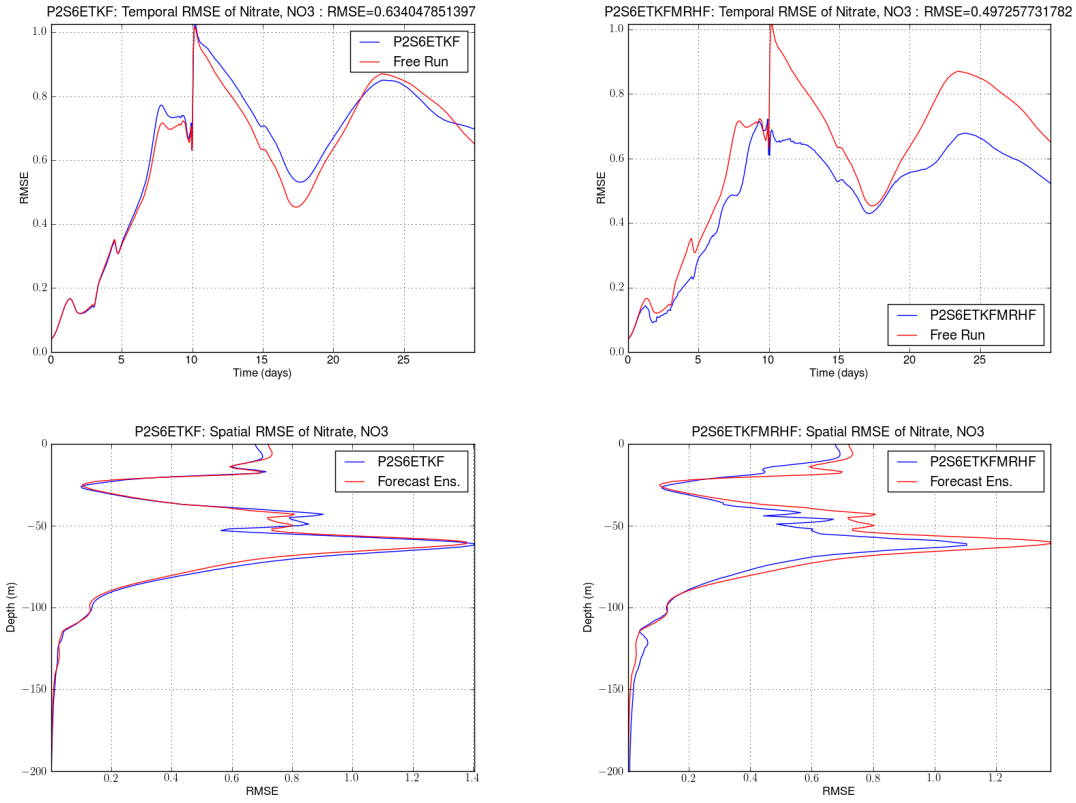

Figure *6.15* – *RMSE de nitrate (NO_3) en fonction du temps (panneaux supérieurs) et de la verticale (panneaux inférieurs), pour l'analyse de l'ETKF (panneaux de gauche) et pour l'analyse combinée de l'ETKF-MRHF (panneaux de droite) avec l'analyse en bleu et le run libre en rouge.*

Répercussions de l'assimilation sur le nitrate

Le nitrate n'est pas une variable contrôlée. L'impact de l'assimilation sur cette variable s'effectue donc par l'évolution du modèle notamment grâce à des corrections sur le phytoplancton précises et bien équilibrées. Comme précédemment, nous évaluons les ensembles engendrés par l'ETKF et le MRHF à travers leurs états analysés moyens et leurs dispersions.

Erreurs RMS La Figure 6.15 est identique aux Figures 6.11, 6.12 et 6.13. Mais pour la variable nitrate (NO_3), il apparait clairement que l'état analysé par l'ETKF (graphiques de gauche) n'apporte pas d'amélioration et donne des erreurs RMS du même ordre et parfois plus grandes que le simple ensemble libre.

Grâce à ses bonnes corrections en phytoplancton et aussi à la génération d'états mieux équilibrés, le MRHF réussit à fortement diminuer les erreurs sur tout le mois et sur toute la verticale.

Histogrammes de rangs La Figure 6.16 présente les histogrammes de rangs sur la variable de nitrate pour l'ETKF à gauche et le MRHF à droite. Il apparait comme pour le phytoplancton que l'ETKF ne parvient pas à corriger le biais (de surestimation cette fois) que présentait l'ensemble libre. La dispersion de cet ensemble est encore assez mauvaise.

Le MRHF ne présente pas un histogramme complétement plat mais réussit, malgré une fréquence satellite de 6h, à débiaiser l'ensemble et améliore tout de même considérablement la dispersion de l'ensemble.

Après avoir observé le bon comportement du MRHF sur le phytoplancton, dans ce contexte de contrôle difficile, nous pouvons constater la bonne transmission de l'information d'analyse au sein du système biogéochimique (à travers le modèle).

6.3.5 Bilan

Dans ce cadre d'expérience d'assimilation de profils et de données satellites, on constate que le contrôle de variables biogéochimiques par des observations de la dynamique n'est pas correctement assuré par un filtre aux hypothèses Gaussiennes tel que l'ETKF. Ce résultat concorde avec l'analyse de la première sous partie de ce chapitre, où l'on observait que la propagation de l'incertitude Gaussienne sur le vent vers les variables biogéochimiques s'accordait avec une perte de Gaussianité significative.

En ce sens, le filtre non-Gaussien MRHF produit de bons résultats. L'état moyen analysé présente des erreurs RMS similaires à l'ETKF sur les variables dynamiques

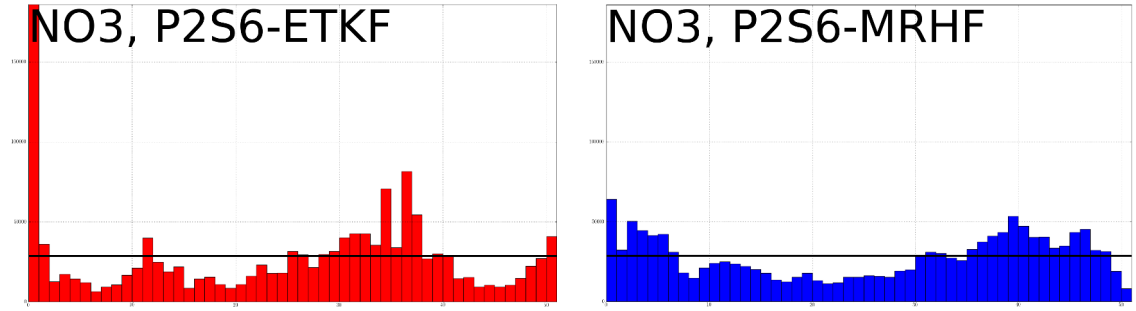

Figure *6.16* – *Histogrammes de rangs, sur le mois et les 200 premiers mètres, de la variable nitrate pour l'analyse de l'ETKF (graphique à gauche) et pour l'analyse combinée de l'ETKF-MRHF (graphique à droite) avec le réseau d'observations P2S6.*

et de plus faibles erreurs RMS sur les variables biogéochimiques. L'ensemble gagne également en dispersion.

Le temps calculs nécessaire à l'assimilation MRHF est, dans cette configuration, de l'ordre de 5 fois supérieur à celui de l'ETKF. Le contrôle quasi-inefficace de la biogéochimie par l'ETKF montre tout de même que, malgré son coût calculs important, le MRHF présente un avantage non-négligeable en contexte non-Gaussien.

On peut se demander maintenant comment se comporte le MRHF en augmentant la fréquence d'observations satellites. La section suivante se consacre à ce problème.

6.4 Problème non-Gaussien et observations hautes fréquences

La section précédente nous a permis de constater les bonnes perfomances du MRHF dans le cadre du contrôle du phytoplancton en observant T et S. Les résultats pour la configuration avec des profils à deux jours et des observations SST à 6h (P2S6) montrent une forte réduction des erreurs RMS en comparaison à l'ETKF. La qualité de l'ensemble produit par le MRHF, indique une bonne correction des biais mais indique également une légère sous-dispersion. Cette sous-dispersion peut s'amplifier avec l'augmentation de la fréquence d'observations. En effet un ensemble trop fréquemment contraint peut s'effondrer sur lui-même si l'erreur modèle n'est pas adaptée. Nous tâchons d'évaluer ce phénomène dans cette section.

Nous considérons les résultats produits par l'ETKF et le MRHF dans les différentes configurations de réseaux d'observations. La première sous partie, permet d'évaluer l'état d'analyse moyen en regardant des diagrammes de Hovmöller et des erreurs RMS. Puis nous regardons l'évolution de la dispersion, de la fiabilité et de la résolution des ensembles engendrés par les assimilations.

6.4.1 Précision de l'état d'analyse

Dans un premier temps, nous comparons visuellement les corrections apportées sur le phytoplancton (variable non-observée) par l'ETKF et le MRHF pour le réseau d'observation le plus dense, c'est à dire, dans la configuration P2S05. Nous regardons, dans un second temps, l'évolution des erreurs RMS respectives aux deux méthodes en fonction des réseaux d'observations sur la température, la salinité, le phytoplancton et le nitrate.

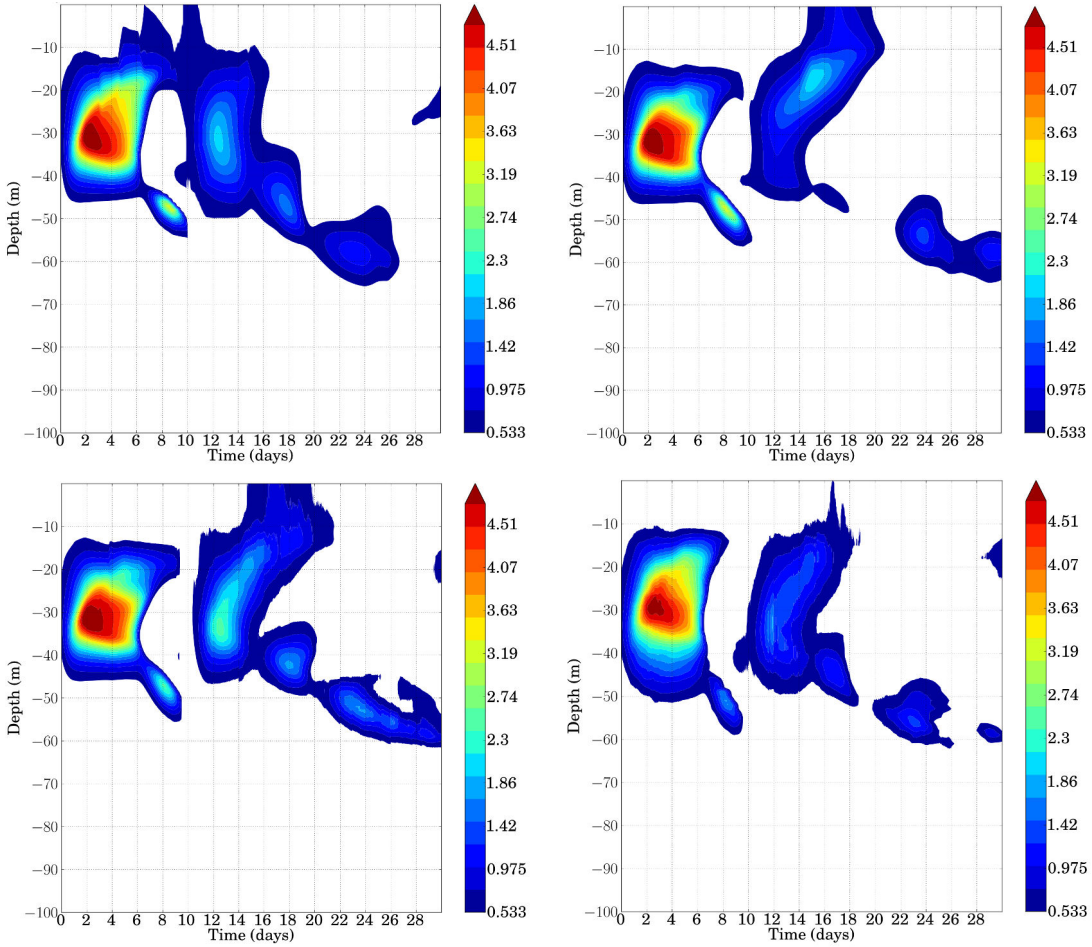

Figure ***6.17*** – *Diagrammes de Hovmöller de la somme des trois variables de phytoplancton (PicP,NanP,MicP en $mmolNm^{-3}$) pour la vérité (panneau supérieur gauche), pour l'ensemble en run libre (panneau supérieur droit), avec le réseau d'observations P2S05 pour l'analyse de l'ETKF (panneau inférieur gauche) et pour l'analyse du MRHF (panneau inférieur droit).*

Estimation du phytoplancton

Diagrammes de Hovmöller La Figure 6.17 représente les diagrammes de Hovmöller du total de phytoplancton (en $mmolNm^{-3}$) sur le mois d'avril et les 100 premiers mètres, pour la *vérité* (graphique supérieur gauche), l'ensemble libre (graphique supérieur droit), les assimilations P2S05 par l'ETKF (graphique inférieur gauche) et par le MRHF (graphique inférieur droit).

La Figure 6.17 nous permet de constater les grandes différences de corrections entre l'ETKF et le MRHF. Contrairement à l'ETKF, l'analyse MRHF réussit à réduire la surestimation du phytoplancton en surface entre le $12^{\text{ème}}$ et le $16^{\text{ème}}$ jour. L'appronfondissement du phytoplancton après le bloom est également bien représenté par le MRHF. La concentration en phytoplancton produite par l'ETKF à 60 mètres en fin de mois est encore trop importante. Pendant les 10 premiers jours, l'ETKF comme le MRHF ne modifie que peu la représentation du phytoplancton. Ceci est dû à la faible dispersion en phytoplancton de l'ensemble à cette période là (Fig. 5.14).

Visuellement, l'état moyen produit par le MRHF représente mieux la *vérité*. Ce résultat est confirmée par l'étude des erreurs RMS qui suit.

Quantités intégrées Avec le timing du bloom et la vitesse d'approfondissement du phytoplancton, il est aussi important pour une simulation de bien représenter les quantités biogéochimiques intégrées sur la verticale.

La Figure 6.18 montre les quantités intégrées sur les 50 premiers mètres du phytoplancton total (à gauche) et du nitrate (à droite) pour la *vérité*, pour l'ensemble en run libre, pour l'analyse de l'ETKF et pour l'analyse du MRHF avec le réseau d'observations P2S05. Le principal pic de phytoplancton des cinq premiers jours est sous estimé par l'ensemble libre (courbe verte) et par l'ensemble produit par l'ETKF (courbe rouge). Le MRHF surestime légérement l'intensité de ce pic mais parvient à retrouver le taux de diminution des quantités intégrées en fin de pic. Un second pic de phytoplancton se produit après le $10^{\text{ème}}$ jour (lié à un fort événement de vent). L'analyse ETKF estime correctement l'augmentation de phytoplancton en surestimant légérement son intensité. L'analyse MRHF sous estime cette augmentation. À la suite de ce pic, la diminution de la quantité intégrée de phytoplancton (entre le $15^{\text{ème}}$ et le $20^{\text{ème}}$ jour) est une fois de plus bien estimée par le MRHF.

Pour la quantité de NO_3 intégrée sur la verticale, le MRHF restitue une diminution autour du $10^{\text{ème}}$ jour proche de la *vérité*, alors que l'ETKF la sous estime fortement.

En résumé, le MRHF et l'ETKF produisent des analyses du phytoplancton proche de la *vérité*. Le MRHF parvient mieux à reproduire les quantités intégrées de phy-

toplancton lors des moments de fortes concentrations. Cette meilleure estimation se répercute sur le nitrate intégré sur la verticale, que l'ETKF surestime fortement.

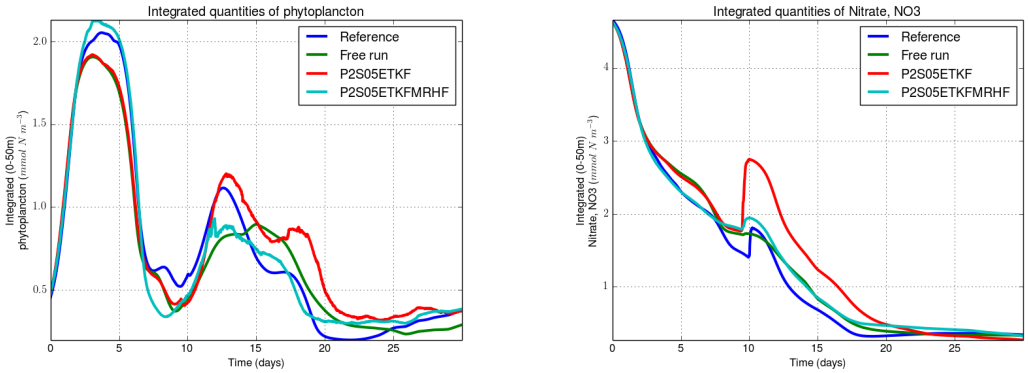

Figure *6.18* – *Quantités intégrées sur les 50 premiers mètres de la somme des trois variables de phytoplancton (à gauche, en $mmolNm^{-3}$) et du nitrate (à droite, en $mmolNm^{-3}$) pour la vérité (ligne bleu foncé), pour l'ensemble en run libre (ligne vert), pour l'analyse de l'ETKF (ligne rouge) et pour l'analyse du MRHF (ligne bleu ciel) avec le réseau d'observations P2S05.*

Estimation globale

Erreurs RMS totales De la même manière que pour les Figures 6.5 et 6.8, nous nous intéressons aux erreurs RMS moyennées sur le mois d'avril et sur les 200 premiers mètres. L'objectif est de comparer l'évolution des erreurs de l'état moyen après assimilation des deux méthodes en fonction des différentes configurations.

La Figure 6.19 représente ces erreurs sur la variable température (graphique supérieur gauche), salinité (graphique supérieur droit), phytoplancton (graphique inférieur gauche) et nitrate (graphique inférieur droit). Les erreurs de l'analyse ETKF sont en rouge et celles de l'analyse MRHF sont en bleu.

Les erreurs en température et salinité sont du même ordre de grandeur pour l'ETKF et le MRHF. Pour la configuration P2S05, le MRHF a toutefois plus de difficultés à bien corriger la dynamique.

Les résultats sur les variables biogéochimiques sont plus nets. On remarque tout d'abord que l'augmentation de la fréquence d'observations n'a pas ou peu d'impact sur le contrôle de la biogéochimie par l'ETKF. A contrario, pour le phytoplancton, le MRHF décroît globalement à chaque fois que le réseau d'observations est plus fin.

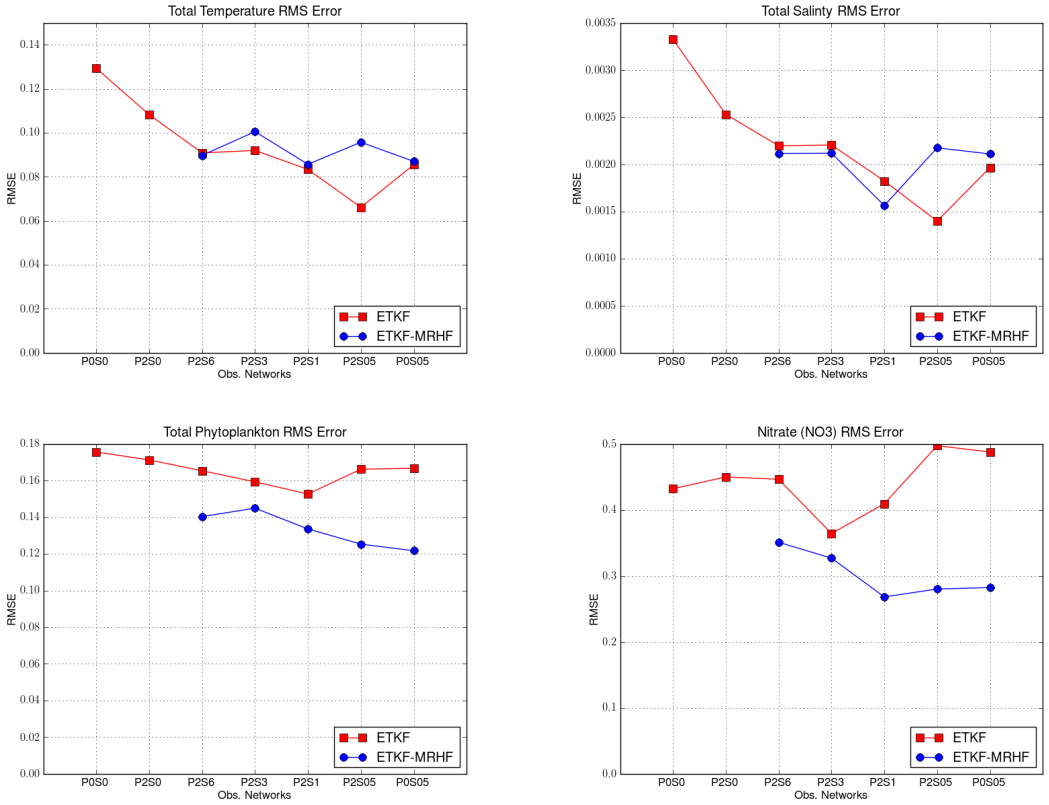

Figure *6.19* – *RMSE totale de la température (panneau supérieur gauche), de la salinité (panneau supérieur droit), du phytoplancton (panneau inférieur gauche) et du nitrate (panneau inférieur droit) pour chaque expérience avec l'ETKF (courbe rouge avec carrés) et avec l'ETKF-MRHF combinés (rond bleu).*

Pour le réseau P2S05, par exemple, l'écart entre les erreurs RMS de l'ETKF et du MRHF est considérable.

En résumé, dans ce contexte non-Gaussien l'utilisation d'une méthode dite non-Gaussienne pour des réseaux d'observations fins est pertinente au regard de l'état moyen. Il s'agit maintenant de considérer la qualité des ensembles produits.

6.4.2 Dispersion, fiabilité, résolution

Bien que le sujet des qualités requises pour un ensemble ait été discuté au Chapitre 1 (Sec. 2.3.1), nous rappelons que la fiabilité (cohérence statistique) est une condition nécessaire à la Bayesianité d'un ensemble. La dispersion d'un ensemble est, elle-même, condition nécessaire à la fiabilité. Et la résolution (qualité informative) d'un

ensemble est une notion complémentaire sur la qualité de l'ensemble.

Nous avons vu dans la section précédente que l'ensemble généré par le MRHF dans la configuration P2S6 présentait une légère sous-dispersion. Cette sous-dispersion peut être due à des corrections trop fortes sur l'ensemble à chaque observation satellite. Dans ce cas, des observations satellites fréquentes ne laisseraient pas le temps au système de correctement re-disperser l'ensemble ce qui pourrait engendrer un effondrement de ce dernier.

La question que l'on se pose est donc : comment se comportent les ensembles ETKF et MRHF face à un problème non-Gaussien pour des données hautes fréquences ?

Pour répondre à cette question, nous considérons d'abord la dispersion des ensembles visuellement et au travers d'histogrammes de rangs. Nous évaluons ensuite la fiabilité et la résolution avec le *Continuous Rank Probability Score* (CRPS), un diagnostic présenté en Section 2.3.1.

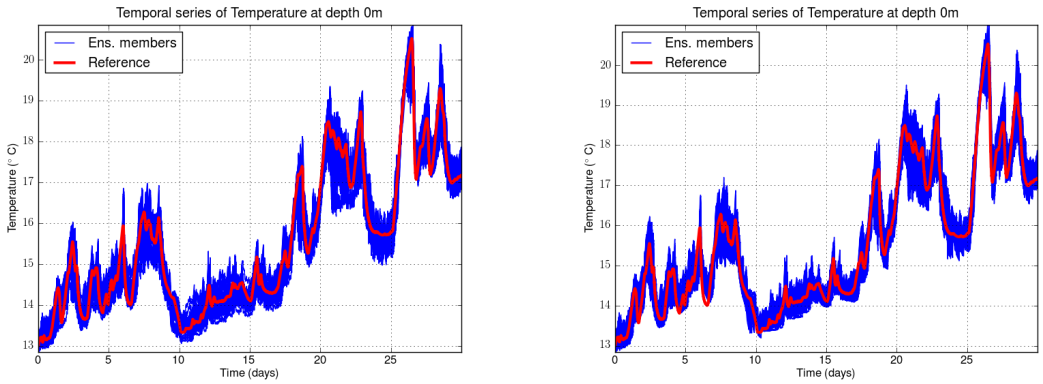

Figure *6.20* – *Séries temporelles de la température de surface pour les ensembles (en bleu) de l'ETKF (panneau de gauche) et de l'ETKF-MRHF (panneau de droite) et la vérité (en rouge) dans la configuration P2S05 (la plus observée).*

Dispersion de l'ensemble

La Figure 6.20 montre les séries temporelles de la température de surface pour les ensembles de l'ETKF (panneau de gauche) et de l'ETKF-MRHF (panneau de droite) comparées à la *vérité* dans la configuration la plus observée : un profil tous les deux jours et la SST toutes les demi-heures (P2S05).

Il s'agit donc de regarder la correction d'une variable peu non-Gaussienne en un point de grille directement observé. Pour ce type de problème direct et peu non-Gaussien, le filtre de Kalman d'ensemble est une référence. Toutefois, on constate que les deux ensembles produits par l'ETKF et le MRHF sont très similaires et estiment bien la *vérité*. À partir du $10^{\text{ème}}$ jour, on observe même une dispersion moindre de l'ensemble MRHF autour de la *vérité*.

La Figure 6.21 présente les histogrammes de rangs de phytoplancton, sur le mois d'avril et sur les 200 premiers mètres de la colonne d'eau, pour les différentes configurations : P2S6 (ligne 1), P2S3 (ligne 2), P2S1 (ligne 3), P2S05 (ligne 4) et P0S05 (ligne 5). Les histogrammes de gauche (en rouge) sont ceux des ensembles générés par l'ETKF et les histogrammes de droite (en bleu) ceux des ensembles générés par le MRHF.

Les histogrammes de l'ETKF (en rouge) montrent que l'assimilation ne parvient pour aucun réseau d'observations à corriger les biais de sous-estimation du phytoplancton. On voit également, que l'ensemble ETKF devient sous dispersif pour des observations satellites à 30 minutes.

Comme il était prévisible, la légère sous dispersion de l'ensemble MRHF en configuration P2S6 s'intensifie lorsque la fréquence d'observations augmente. Les histogrammes de toutes les configurations présentent une forte sous-dispersion.

Pour confirmer ces résultats nous utilisons à présent le CRPS qui nous donnera deux informations sur l'ensemble : la fiabilité et la résolution.

Fiabilité et résolution

Pour évaluer la fiabilité et la résolution (et l'incertitude) associées à nos ensemble nous utilisons la décomposition de Hersbach (2000) décrite en Section 2.3.1. Nous regardons deux indices : le terme de fiabilité "REL" et la somme des termes de résolution et d'incertitude appelée "CRPS_{pot}". Le CRPS_{pot} est le CRPS potentiel qui peut être vu comme la résolution d'un ensemble si sa fiabilité était parfaite.

Nous avons donc deux indices, l'un décrivant la cohérence statistique de l'ensemble et l'autre donnant l'apport d'information de l'ensemble si l'ensemble était statistiquement cohérent.

Un faible indice REL correspond à une bonne fiabilité de l'ensemble. De même, un faible indice CRPS_{pot}, équivaut à un bon CRPS potentiel, c'est à dire une bonne résolution si REL était nul. De manière globale, une faible valeur CRPS indique une bonne fiabilité et une bonne résolution.

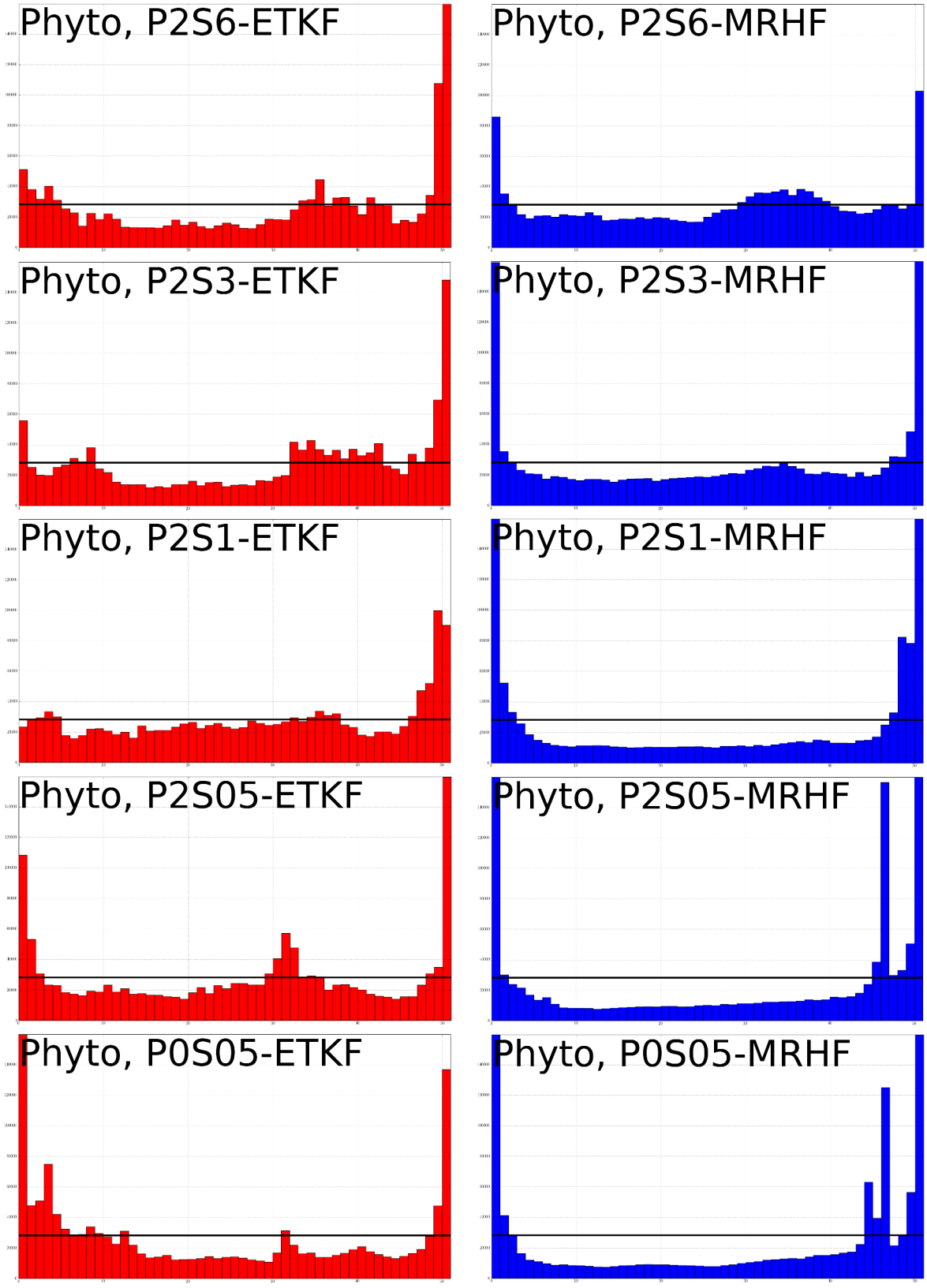

Figure *6.21* – *Histogrammes de rangs, sur le mois d'avril, du phytoplancton total pour les réseaux d'observations P2S6 (ligne 1), P2S3 (ligne 2), P2S1 (ligne 3), P2S05 (ligne 4) et P0S05 (ligne 5). Les histogrammes des ensembles ETKF sont en rouge (colonne 1) et ceux des ensembles MRHF sont en bleu (colonne 2). La barre noire symbolise l'histogramme plat (dispersion parfaite).*

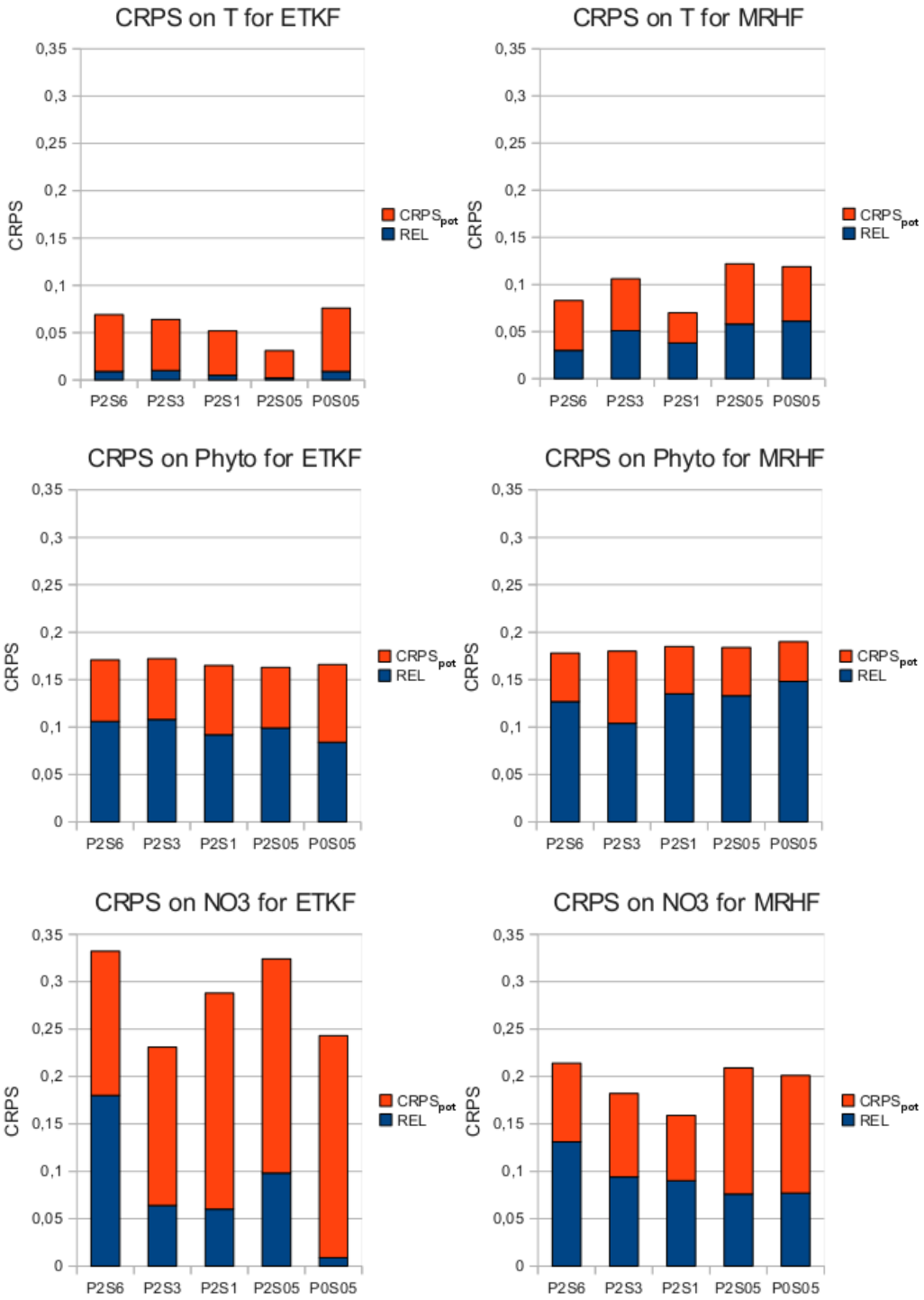

Figure *6.22* – *CRPS sur le mois de la variable température (panneaux supérieurs), du phytoplancton (panneaux centraux), du nitrate (panneaux inférieurs) des ensembles analysés produits par l'ETKF (panneaux de gauche) et par le MRHF (panneaux de droite) pour les réseaux d'observations P2S6, P2S3, P2S1, P2S05 et P0S05.*

La Figure 6.22 est constituée de diagrammes présentant le CRPS comme la somme de REL (en bleu) et de CRPS_{pot} (en orange) et ce pour les différentes configurations. Ces CRPS sont calculés pour la température (ligne 1), pour le phytoplancton (ligne 2) et pour le nitrate (ligne 3). On étudie l'ensemble issu de l'analyse ETKF (colonne 1) et l'ensemble issu de l'analyse MRHF (colonne 2).

La première observation est que, pour toutes les variables, la fiabilité des ensembles ETKF est globalement meilleure (REL plus petit) que celle des ensembles MRHF. Ceci confirme la sous-dispersion des ensembles observée précédemment.

Bien que la fiabilité soit toujours moins bonne pour le MRHF, le score CRPS total ne varie que peu entre l'ETKF et le MRHF pour la variable température. Les CRPS des deux méthodes sont presque égaux pour le phytoplancton et celui du MRHF est plus faible pour le nitrate.

Le faible CRPS potentiel des ensembles MRHF, indique qu'il serait très intéressant de travailler à améliorer la fiabilité des ensembles. En effet, l'utilisation du MRHF présente déjà l'avantage de fournir des erreurs RMS faibles sur les variables biogéochimiques mais on voit également que pour un ensemble fiable l'information ensembliste produite par le MRHF est plus pertinente que celle de l'ETKF.

En résumé, la mauvaise dispersion des ensembles produits par le MRHF pour des observations hautes fréquences est inquiétante. Cette mauvaise dispersion a été confirmée par des diagnostics CRPS indiquant une mauvaise fiabilité.

Toutefois, les bons résultats du MRHF en terme d'erreurs RMS et de résolution motivent à essayer d'améliorer la fiabilité des ensembles analysés. Il s'agit notamment d'une des perspectives importantes à ce travail.

6.4.3 Bilan

Employer une méthode dite non-Gaussienne ne suffit pas pour résoudre un problème contenant de fortes non-Gaussianités. Plusieurs précautions sont à prendre. La conclusion à tirer de cette sous partie est notamment que dans ce contexte particulier d'observations satellites si le réseau temporel d'observations est trop dense les méthodes non-Gaussiennes peuvent générer des ensembles trop sous dispersifs. En effet, certaines méthodes non-Gaussiennes sont victimes de la malédiction de la dimensionalité. Le MRHF est moins soumis à ce problème que, par exemple, les filtres particulaires puisqu'il s'agit d'une méthode de transport (i.e. déplaçant les membres vers la solution) et non de séléction. Représenter des densités de probabilités en grande dimension avec une ensemble de petite taille reste difficile et une succession de correction peut dégrader la représentation des densités par l'ensemble. Ceci se produit

lorsque l'espace des solutions est grand devant l'espace exploré par l'ensemble. Et chaque nouvelle observation va restreindre encore plus l'espace des solutions bien sûr mais également l'espace exploré par l'ensemble. Un réseau d'observation trop dense peut donc avoir pour conséquence, ce que l'on constate ici, de réduire l'espace exploré jusqu'à ce qu'on appelle l'effondrement (*collapse*) de l'ensemble. Des techniques pour corriger (temporairement) ce problème peuvent être par exemple d'augmenter la taille de l'ensemble, d'utiliser de l'inflation ou de la localisation.

6.5 Conclusions

Dans ce chapitre l'étude d'un problème d'assimilation de données avec le modèle ModECOGeL est réalisée. Cette étude combine plusieurs caractéristiques particulières : l'utilisation d'un modèle couplé dynamique/biogéochimie marine, la génération de l'ensemble par perturbations stochastiques du forçage vent, l'utilisation d'une méthode d'assimilation non-Gaussienne, les différents réseaux d'observations utilisés. Il en ressort de nombreuses conclusions que l'on énonce dans cette section.

Une étude a priori de caractérisation du problème d'assimilation, au regard de la non-Gaussianité, doit être effectuée afin de choisir la méthode d'assimilation la plus adaptée. En effet, une simple évaluation du caractère plus ou moins Gaussien des problèmes d'assimilation dans la première section de ce chapitre (Section 6.1), nous a permis d'identifier la qualité quasi-Gaussienne du problème de contrôle de la dynamique seule et la qualité clairement non-Gaussienne du contrôle du phytoplancton à l'aide d'observations de température et de salinité. Ces deux degrés de Gaussianité se sont confirmés lors de la réalisation d'expériences d'assimilation (Section 6.2 et Section 6.3) avec la méthode aux hypothèses Gaussiennes qu'est l'ETKF puis avec le MRHF, méthode non-Gaussienne.

Un filtre Gaussien produit de bons résultats dans le cadre quasi-Gaussien de contrôle de la dynamique par des observations dynamiques. Il apparait dans la Section 6.2, que l'ETKF produit sur les variables dynamiques une analyse performante, avec de faibles RMSE, et un ensemble de dispersion correcte, avec des histogrammes de rangs nettement améliorés. Par contre, ces corrections n'ont que des conséquences (via la physique du modèle) très limitées sur les variables biogéochimiques. D'un point de vue physique, ceci nous indique également, l'importance du contrôle d'au moins une variable biogéochimique pour propager l'information apportée par les observations dynamiques à la partie biogéochimique du modèle.

L'utilisation d'un filtre non-Gaussien est plus adaptée dans le cadre non-Gaussien de contrôle de la biogéochimie par des observations dynamiques. La Section 6.3, s'attache à comparer les performances de l'ETKF et du

MRHF, dans le contexte du contrôle des variables phytoplanctoniques (en plus de T et S) pour les données de SST à 6h. Il est important de noter que les profils de température et de salinité (à 2 jours) sont systématiquement assimilées par l'ETKF (pour des raisons de temps de calcul) et que cette configuration ETKF-MRHF est permise par la flexibilité d'implémentation du MRHF avancé au Chapitre 4. Dans ce nouveau cadre les hypothèses Gaussiennes de l'ETKF, ne lui permettent pas de correctement contrôler le phytoplancton et donc de réduire les erreurs d'analyse sur les variables biogéochimiques. Le choix d'une méthode non-Gaussienne, se montre bénéfique puisque l'analyse fournie par le MRHF réussit à représenter correctement les variables biogéochimiques (faibles erreurs RMS du phytoplancton total et du nitrate), de plus, l'ensemble produit présente une fiabilité et une résolution correctes.

Un réseau d'observations fin dégrade l'ensemble produit par un filtre non-Gaussien. Les résultats produits dans la Section 6.4 relativisent les bonnes performances d'un filtre non-Gaussien, en montrant que le choix d'une méthode non-Gaussienne ne doit pas seulement se faire sur le degré de non-Gaussianité du modèle. Bien que la représentation des variables biogéochimiques faite par l'estimé moyen du MRHF reste meilleure que celle de l'estimé moyen de l'ETKF, il est montré dans cette section qu'un réseau d'observation trop serré dégrade nettement les résultats du MRHF en allant jusqu'à provoquer l'effondrement de l'ensemble. Cette dégradation se produit également pour l'ETKF mais à un degré moindre. En effet, les méthodes non-Gaussiennes sont connues pour être plus demandeuses en nombre de membres d'ensemble et l'utilisation d'un réseau d'observation fin contraint trop rapidement la dispersion de l'ensemble qui ne parcourt plus qu'une partie trop réduite de l'espace des solutions. Ce problème doit être traité sur deux fronts : en appliquant des techniques simplifiant le problème (augmenter le nombre de membres, ajouter de l'inflation, appliquer une localisation ...) ; en poursuivant le développement du MRHF, encore jeune.

Chapitre 7

Assimilation de la donnée de couleur de l'eau

Sommaire

Le cadre des expériences présentées dans ce chapitre est décrit au Chapitre 5 et est similaire à celui des expériences du chapitre précédent (Chap. 6).

Nous travaillons avec le modèle couplé de dynamique et de biogéochimie marine, ModECOGeL, sur le mois d'avril 2006. La *vérité* est une simulation effectuée avec un forçage de vent réel en haute fréquence (1 heure). Un ensemble de 50 membres est propagé par le modèle avec un forçage de vent haute fréquence simulé par un processus auto-regressif Gaussien d'ordre un.

L'objectif premier de ce chapitre est d'évaluer ce que l'assimilation de données non-Gaussiennes peut apporter au problème complexe d'assimilation de la couleur de l'eau.

Pour ce faire, nous nous munissons, dans un premier temps, d'un jeu d'observations de la couleur de l'eau avec une fréquence temporelle de trois jours. Nous étudions les performances d'un filtre de Kalman d'ensemble, l'ETKF, assimilant ce jeu d'observations pour le contrôle du phytoplancton. Puis nous comparons trois filtres : l'ETKF, l'EnKF anamorphosé et le MRHF assimilant ce même jeu d'observations pour le contrôle du phytoplancton et de la température. Dans un second temps, nous regardons le comportement de ces trois mêmes méthodes face à une fréquence d'observations plus grande (fréquence temporelle d'un jour). Cette dernière étude se place dans le contexte d'observations issues d'un satellite géostationnaire de la couleur de l'eau.

Le tableau de la Figure 7.1 récapitule les différentes expériences mises en place dans les sections 7.2, 7.3 et 7.4.

Dans les travaux réalisés ici, nous cherchons à contrôler uniquement les variables d'états. Il est à noter que le contrôle des variables d'état n'est pas le seul moyen d'améliorer l'estimation de la dynamique. La source d'incertitudes impactant la dynamique du modèle étant l'intensité du vent, il est également possible de vouloir contrôler directement cette source. Par exemple, les travaux de Gregorio et al. (prep), dans une configuration similaire à celle-ci, contrôle par assimilation de données les coefficients du forçage stochastique. Nous n'étudions pas ce problème d'assimilation dans cette thèse.

Dans la première section de ce chapitre (Sec. 7.1), nous caractérisons le problème en décrivant la nature de l'observation de la couleur de l'eau puis en étudiant les relations statistiques des variables du problème. Dans la Section 7.2, nous présentons les résultats de la première expérience de contrôle du phytoplancton par l'ETKF. Dans la Section 7.3, nous comparons l'ETKF, l'EnKF anamorphosé et le MRHF dans le contrôle du phytoplancton et de la température. Enfin, la dernière section (Sec. 7.4), ouvre la perspective d'une assimilation de la couleur de l'eau haute fréquence.

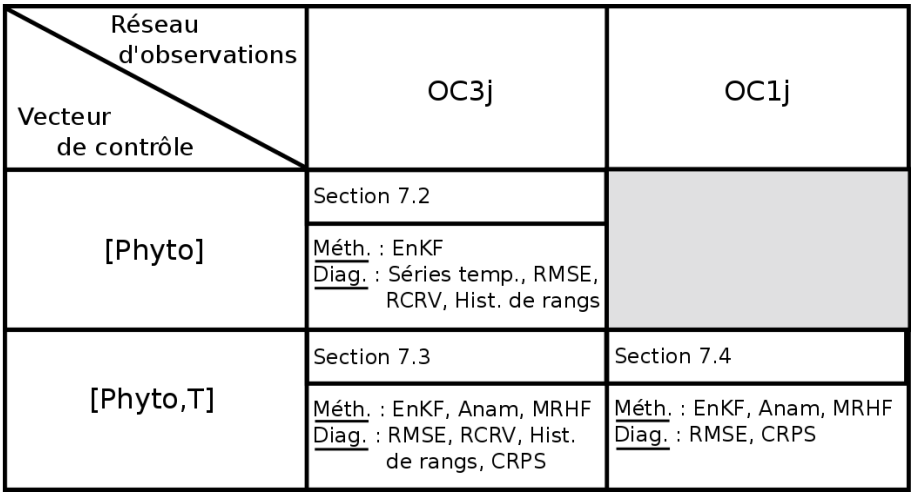

Figure *7.1* – *Récapitulatif des différentes expériences d'assimilation selon leurs réseaux d'observations de la couleur de l'eau, leurs vecteurs de contrôle, les méthodes étudiées et les diagnostics utilisés.*

7.1 Caractérisation du problème d'assimilation de couleur de l'eau

Dans cette première section, nous essayons de mieux comprendre l'observation de la couleur de l'eau dans notre étude ModECOGeL ainsi que les challenges que ce type d'observation peut apporter à l'assimilation de données. Dans la première sous partie, nous décrivons la couleur de l'eau telle que nous la simulons dans nos expériences avec le modèle ModECOGel. Dans la seconde sous partie, nous essayons de caractériser le problème de l'assimilation de couleur de l'eau en terme de non-Gaussianité et de non-linéarité et ce à l'aide d'outils statistiques. Les objectifs de cette caractérisation sont de pouvoir anticiper et de mieux comprendre le comportement des différentes méthodes d'assimilation qui seront mises en place par la suite pour résoudre ce problème d'assimilation.

7.1.1 La couleur de l'eau dans ModECOGeL

La couleur de l'eau apparait comme une nouvelle source d'information sur l'océan. Plusieurs méthodes pour exploiter l'information issue de la couleur de l'eau existent. Par exemple, de nombreux efforts sont faits pour assimiler les structures dynamiques (horizontales) observées sur les images de la couleur de l'eau en considérant la chlorophylle comme un traceur (par *Finite Size Lyapunov Exponent*, Aurell et al., 1997 ; par *Finite Time Lyapunov Exponent*, Beron-Vera et al., 2008; Titaud et al., 2011 ;

ou encore par exposants de singularité, Turiel et al., 2008).

Une méthode plus directe pour exploiter l'information fournie par la couleur de l'eau est d'en déduire une quantité de chlorophylle puis d'assimiler cette quantité dans un modèle couplé à la biogéochimie marine. L'intérêt est alors l'amélioration de la représentation des phénoménes biogéochimiques d'une part mais également du modèle dynamique au travers des interactions au sein du couplage ou au travers des correlations statistiques des différentes variables. Une très bonne revue de l'assimilation dans des modèles couplés dynamique-biogéochimie a été proposée par Robinson and Lermusiaux (2002).

Dans ce dernier chapitre, nous nous interessons à cette deuxième méthode.

Pour nos expériences ModECOGeL, nous faisons l'hypothèse de connaitre la quantité de chlorophylle associée à la couleur de l'eau. Ainsi nous ne considérons pas le problème d'inversion de l'observation couleur de l'eau en une quantité de chlorophylle. De plus, le modèle ECOGeL n'ayant pas la chlorophylle comme variable, nous faisons donc le raccourci de travailler directement avec la somme des trois variables phytoplanctoniques. Ce phytoplancton total sera considéré comme l'information obtenue en surface de l'océan après traitement de l'information de couleur de l'eau.

La création des observations artificielles de la couleur de l'eau pour nos expériences est décrite au chapitre 5.

7.1.2 Étude des relations statistiques

La non-Gaussianité du phytoplancton de surface

La variable observée par la couleur de l'eau est la concentration de phytoplancton total en surface. La concentration du phytoplancton est une variable positive. De plus, la distribution de phytoplancton dans les modèles biogéochimiques est souvent simulée par une distribution log-normale. Il est donc important, avant de mettre en place une assimilation, d'étudier l'écart à la Gaussianité du phytoplancton.

Dans un premier temps, nous cherchons à estimer le degré de non-Gaussianité du phytoplancton total de surface dans notre système. Pour cela nous effectuons le test statistique de D'Agostino-Pearson, présenté à la Section 2.1.1, à partir de l'ensemble libre. Nous rappelons que l'hypothèse de Gaussianité est rejetée avec une significativité de 95% pour une p-valeur inférieure à 0.05. Les résultats de ce test sont présentés, dans la Figure 7.2, sous forme d'une série temporelle de la p-valeur obtenue pour le phytoplancton total de surface.

Encore une fois, il apparait que le phytoplancton a une distribution statistique complexe et très variable. Dans la première moitié du mois, la distribution peut

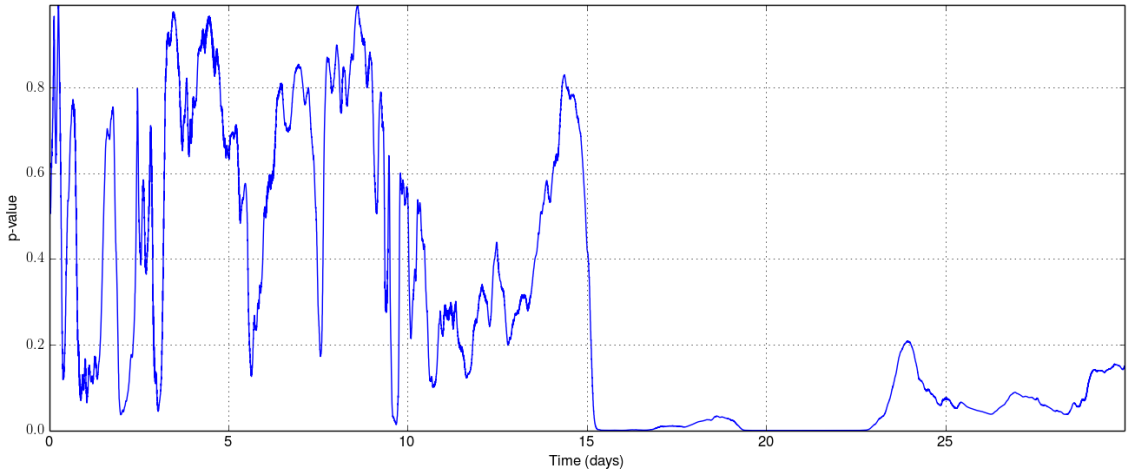

Figure 7.2 – *P-valeur du test de normalité de D'Agostino-Pearson pour le phytoplancton total de surface (variable observée) sur le mois d'avril.*

en moyenne être considérée comme Gaussienne mais des pics de non-Gaussianité apparaissent fréquemment. Ces pics peuvent notamment être dus aux perturbations de vent haute fréquence qui impactent directement la répartition du phytoplancton sur la colonne d'eau. Au cours de la seconde moitié du mois, l'hypothèse de Gaussinité est presque toujours rejetée.

Dans le cadre de l'assimilation de la couleur de l'eau, la complexité de la distribution de phytoplancton de surface, pourra favoriser les méthodes ne faisant pas l'hypothèse de Gaussianité de la variable observée (e.g. RHF, MRHF et EnKF anamorphosé).

Les non-linéarités inter-variables

Diagrammes de Hovmöller de corrélations linéaires Comme on l'a déjà vu, un moyen de constater l'absence de linéarité entre deux échantillons scalaires est de calculer leur r-valeur. C'est ce qui est fait dans la Figure 7.3. Le graphique de gauche montre le diagramme de Hovmöller de la valeur absolue de la r-valeur entre le phytoplancton de surface (la couleur de l'eau) et le phytoplancton sur le reste de la colonne d'eau. La valeur absolue de la r-valeur est proche de 1 lorsque la correlation (ou l'anticorrelation) linéaire est forte et proche de 0 quand les deux échantillons sont très peu correlés. Puisqu'il s'agit de la même variable : le phytoplancton total, il est normal d'observer des correlations linéaires très fortes près de la surface (et égales à 1 à la surface). En revanche, il apparait sur ce graphique que la r-valeur décroît très vite avec la profondeur. L'ensemble des concentrations de phytoplancton en des-

sous d'une dizaine de mètres est presque entièrement décorrelé du phytoplancton de surface. Ceci signifie qu'à l'exception de quelques zones spatio-temporelles (comme entre 30 et 60m du 12$^{\text{ème}}$ au 20$^{\text{ème}}$ jour ou entre 0 et 40m après le 24$^{\text{ème}}$ jour), la relation entre le phytoplancton de surface (la variable observée par la couleur de l'eau) et le phytoplancton sur le reste de la colonne d'eau (la variable contrôlée) est très mal représentée par une droite de régression linéaire. D'un point de vue de l'assimilation, si un lien statistique autre que linéaire existe, ceci peut se traduire par l'échec des méthodes aux moindres carrés, comme les filtres de Kalman d'ensemble qui propagent leurs corrections entre les variables par régression linéaire.

Le graphique de droite de la Figure 7.3 montre le diagramme de Hovmöller de la valeur absolue de la r-valeur entre le phytoplancton de surface (la couleur de l'eau) et la température sur le reste de la colonne d'eau. La valeur absolue de la r-valeur ne dépasse que très rarement 0.8 et est inférieure à 0.5 partout ailleurs. À partir de ce diagnostic, on peut donc s'attendre à ce que le contrôle de la température avec des observations de la couleur de l'eau soit difficile pour les filtres de type moindres carrés.

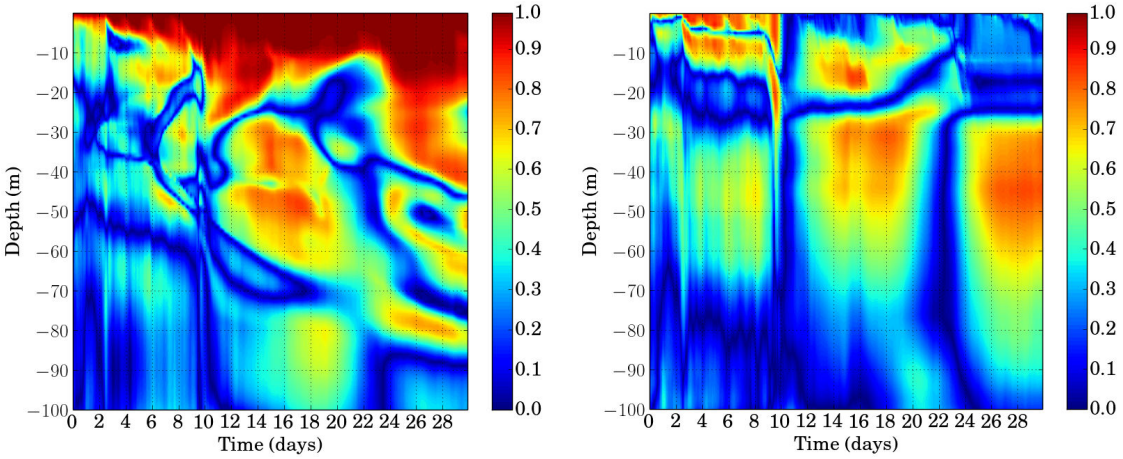

Figure 7.3 – *Valeur absolue de la r-valeur entre le phytoplancton total de surface avec le phytoplancton dans la colonne d'eau (graphique de gauche) et avec la température (graphique de droite) en fonction de la verticale et du temps sur le mois d'avril.*

Profils des corrélations linéaires Comme il a déjà été discuté dans le Chapitre 1, une des difficultés de l'assimilation de données en océanographie est la propagation de l'information sur la verticale. La couleur de l'eau, comme toutes autres observations satellites, apporte une information à la surface (ou dans les premiers mètres).

Pour mieux mettre en évidence la difficulté que peut rencontrer une assimilation aux moindres carrés propageant l'information de la surface aux eaux plus profondes par régression linéaire, nous regardons les mêmes quantités que présentées en Figure 7.3 mais moyennées sur le mois. Ainsi, des profils verticaux (de 0m à 50m) des r-valeurs moyennées sur le mois sont calculées entre le phytoplancton de surface et : la température (bleu), la salinité (vert), le nitrate NO_3 (rouge), le pico-phytoplancton PicP (bleu ciel), le nano-phytoplancton NanP (violet), le micro-phytoplancton MicP (jaune) et le phytoplancton total (noir). Ces profils sont exposés sur la Figure 7.4.

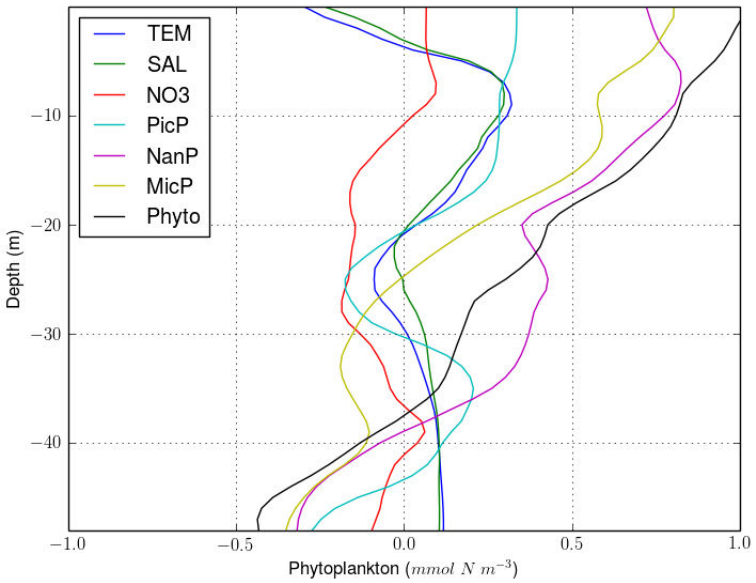

Figure 7.4 – *Profils verticaux (0m-50m) de la valeur absolue des r-valeurs moyennées sur le mois entre le phytoplancton total en surface et : T (bleu), S (vert), NO_3 (rouge), PicP (bleu ciel), NanP (violet), MicP (jaune) et le phytoplancton total (noir).*

La correlation entre le phytoplancton sur la colonne d'eau et le phytoplancton de surface décroît rapidement de 1 à 0 sur les 50 premiers mètres. Cette corrélation devient inférieure (en valeur absolue) à 0.5 à partir de 20m de profondeur et le reste sur le reste de la colonne d'eau. Ceci nous apprend qu'en profondeur : soit le lien statistique n'existe pas (au sens où il n'est pas contenu dans l'ensemble), dans ce cas appliquer aucune correction est la solution optimale ; soit le lien statistique existe mais n'est pas linéaire, dans ce cas les méthodes Gaussiennes (appliquant une correction par régression linéaire) ne sont pas adaptées. Il en va de même pour les trois types de phytoplancton.

Les coefficients de correlation linéaire du phytoplancton de surface avec le nitrate, la salinité et la température restent inférieurs à 0.3 en valeur absolue. Si relations statistiques il y a, elles ne peuvent donc pas être considérées comme des relations linéaires. Dans ce cas, le contrôle d'une de ces variables par assimilation aux moindres carrés n'est pas pertinent.

Bien que notre étude se focalise uniquement sur la couleur de l'eau, ces profils témoignent, par ailleurs, de la nécessité d'une information complémentaire sous la surface (e.g. les profils BioARGO).

Pour notre étude, au vu de la faible linéarité de ces relations, il semble que la correction du phytoplancton en profondeur et d'autres variables comme la température pourrait être plus efficace avec des méthodes non-Gaussiennes, n'effectuant pas de régression linéaire.

7.1.3 Bilan

Les tests statistiques effectués sur l'ensemble libre indiquent que le phytoplancton de surface exhibe un comportement complexe, présentant de nombreuses non-Gaussianités, principalement dans la seconde moitié du mois. De plus, il apparait que les relations entre le phytoplancton de surface et le phytoplancton le long de la colonne d'eau sont (quasi-) linéaires dans les 10 premiers mètres mais se décorrelent rapidement avec la profondeur. Les relations entre le phytoplancton de surface et les autres variables, comme la température, la salinité et le nitrate, sont également peu linéaires.

Cette étude confirme que le système couplé dynamique/biogéochimie est non-linéaire et présente de nombreuses non-Gaussianités. Ainsi, dans ce contexte, le problème de l'assimilation de la couleur de l'eau est loin de réunir les conditions d'optimalité des méthodes moindres carrés. Pour cette raison, les méthodes non-Gaussiennes peuvent présenter un intérêt dans ce contexte.

Les Sections 7.2, 7.3 et 7.4 étudient cette question de manière expérimentale.

7.1.4 Zone d'intérêt

L'objectif que l'on donne à l'assimilation dans ce chapitre est d'améliorer les représentations de la physique du système lors d'importants événements biogéochimiques. En choisissant le mois d'avril 2006, nous avions déjà pris le parti de focaliser notre attention sur les blooms de phytoplancton. Ainsi afin de ne s'intéresser aux performances de l'assimilation que dans les régions (spatio-temporelles) importantes en terme d'événements biogéochimiques, nous restreignons la zone de diagnostics.

Pour diagnostiquer les assimilations dans les expériences de ce chapitre, nous nous munissons d'un masque séléctionnant la zone où la *vérité* présente de fortes concen-

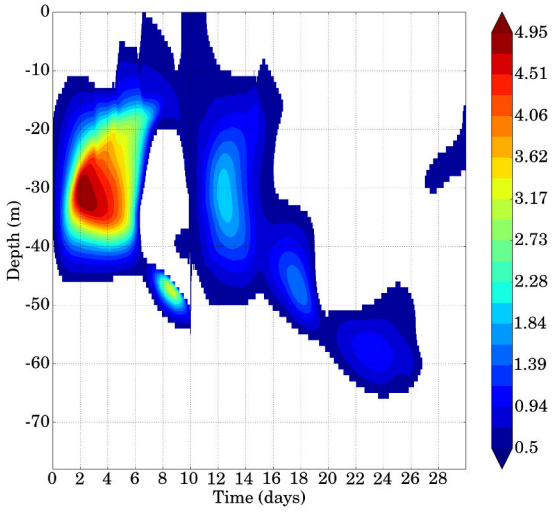

Figure 7.5 – *Diagramme de Hovmöller du phytoplancton après application du masque selectionné pour les diagnostics : Concentrations de phytoplancton supérieures à* $0.5 mmol.N.m^{-3}$*. Par la suite, nous nommons cette zone spatio-temporelle : zone d'intérêt.*

trations de phytoplancton. La raison derrière ce choix est que les zones de faibles concentrations de phytoplancton (e.g. en profondeur) sont souvent bien représentées mais les valeurs obtenues proches de l'erreur machine pénalisent quand même les diagnostics. Cette zone est donc créée en ne considérant que les régions (spatio-temporelle) possédant une concentration de phytoplancton total supérieure à $0.5 mmol.N.m^{-3}$ (valeur arbitraire mais raisonnable au vue de la Figure 7.5). Dans la suite du chapitre, nous nommons cette zone : zone d'intérêt de notre étude. La Figure 7.5 représente dans un diagramme de Hovmöller cette zone d'intérêt.

7.2 Contrôler le phytoplancton par assimilation de couleur de l'eau

Dans cette section, nous appliquons le filtre de Kalman d'ensemble transformé, l'ETKF, au problème de contrôle du phytoplancton par observation de la couleur de l'eau (phytoplancton de surface).

À l'aide des diagnostics habituels, nous évaluons la capacité de l'ETKF à corriger le phytoplancton et l'impact de ces corrections sur le nitrate (Sec. 7.2.1) puis nous évaluons l'impact de ces corrections sur la dynamique du modèle (Sec. 7.2.2).

Il est à noter que, dans cette section, aucune inflation n'est utilisée.

7.2.1 Correction de la biogéochimie par l'ETKF

Le phytoplancton total

Évaluation visuelle : La série temporelle est un outil simple pour estimer les performances d'une prévision et pour mieux comprendre son comportement.

Nous nous intéressons d'abord à l'ensemble libre. Les graphiques supérieurs de la Figure 7.6 sont respectivement (de gauche à droite) les séries temporelles à 0m, 20m et 50m de l'ensemble libre (en bleu), de la moyenne d'ensemble (en noir) et de la *vérité* (en rouge). De plus, les observations de la couleur de l'eau sont représentées (points verts) sur la série temporelle à 0m (graphique supérieur gauche).

On remarque, à chaque profondeur, la présence de biais plus ou moins importants entre l'ensemble libre et la vérité. Ces biais viennent de la nature statistiquement différente entre l'ensemble et la *vérité* (contrairement à la génération de l'ensemble par perturbation de la condition initiale de la *vérité*). Il est à noter que ce phénomène de biais ne se produit pas en expériences jumelles pures (quand l'ensemble et la *vérité* sont initialement issus de la même densité de probabilité). Ce phénomène est, en revanche, courant dans des expériences d'assimilation de données réelles. Réduire au mieux ces biais permet d'améliorer l'estimation et fait donc parti des objectifs de l'assimilation.

On voit également qu'en surface ces biais sont aussi présents entre l'ensemble libre et les observations. Ceci peut avoir un effet important sur l'assimilation. On a vu par exemple, dans l'illustration du Chapitre 2, que l'anamorphose rencontre des difficultés face à des biais entre l'ensemble et l'observation. De plus, la dynamique du modèle forcé par le vent simulé (modèle produisant l'ensemble) ne semble pas permettre une dispersion suffisante de l'ensemble dans les 10 premiers jours. Encore une fois, une faible dispersion d'ensemble (créee par la différence statistique entre l'ensemble et la *vérité*) aura aussi un impact sur l'assimilation. Cette faible dispersion engendre une grande confiance en l'ensemble (par sous-estimation des covariances d'erreurs *a priori*) qui est pourtant loin de la *vérité*.

Les graphiques inférieurs de la Figure 7.6 présentent les séries temporelles aux mêmes profondeurs que les graphiques supérieurs mais pour l'ensemble produit par l'ETKF. Les corrections effectuées par l'ETKF parviennent à corriger certains biais. Principalement dans la deuxième moitié du mois, l'ensemble et la moyenne d'ensemble, à la surface (attention à la différence d'échelles) et à 20 mètres, ne surestiment plus autant la *vérité*. À 50 mètres, la moyenne d'ensemble est améliorée. Toutefois, les corrections proposées par l'ETKF ne semblent pas réalistes pour l'évolution d'une concentration de phytoplancton. Il s'agit probablement ici des limites de la régression linéaire discutées à la section précédente.

Dans la première moitié du mois, la dynamique du modèle ne produit pas une

grande dispersion de l'ensemble dans les 10 premiers jours. Comme attendu, à cette période, l'ETKF a bien tendance à faire "trop confiance" à l'ensemble peu dispersé (i.e. faibles variances d'erreurs d'*a priori*). En conséquence, les observations n'influent presque pas sur l'assimilation.

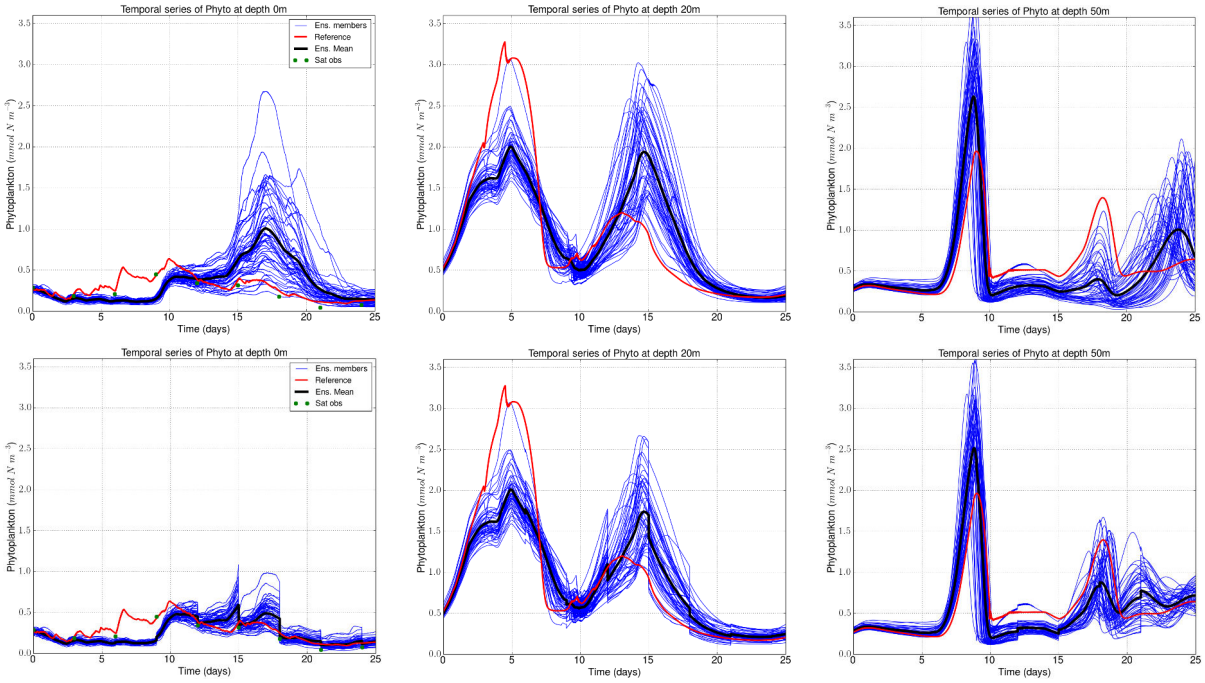

Figure *7.6* – *Séries temporelles du phytoplancton total de l'ensemble (en bleu) libre (graphiques supérieurs) et produit par l'ETKF (graphiques inférieurs) sur le mois d'avril à 0m (gauche), 20m (centre) et 50m (droite) de profondeur. La série temporelle de la moyenne d'ensemble est représentée en noir. La vérité est repésentée en rouge et les observations de la couleur de l'eau (points verts) sont représentées sur la série temporelle de l'ensemble libre (graphique supérieur gauche).*

Pour évaluer la capacité de l'ETKF à propager l'information de la couleur de l'eau le long de la colonne d'eau, nous comparons à présent les profils verticaux de l'ensemble libre et de l'ensemble assimilé. La Figure 7.7 représente ces profils verticaux de 0m à 50m, au jour 0 (à gauche), au jour 16 (au centre) et au jour 30 (à droite). À l'instar de la Figure 7.6, l'ensemble libre et l'ensemble assimilé sont représentés en bleu sur les graphiques supérieurs et les graphiques inférieurs, respectivement. La *vérité* est toujours en rouge.

Les profils verticaux de l'ensemble initial (au jour 0, graphiques de gauche) couvrent la *vérité*. Il n'y a donc pas de biais initial. Au jour 16 (graphique supérieur du centre),

l'ensemble libre surestime la concentration de phytoplancton entre 10 et 30 mètres et la sous-estime sous les 30 mètres. Ceci peut être dû à un faible mélange vertical engendré par l'ensemble de vents simulés. De même, au jour 30 (graphique supérieur de droite), les concentrations de phytoplancton de l'ensemble libre décroissent rapidement sur la verticale à partir de 20 mètres de profondeur et l'ensemble sous-estime la *vérité* entre 30 et 45 mètres. Ainsi la structure verticale du phytoplancton (moyen en temps) n'est pas correctement représentée par le modèle forcé par le vent haute-fréquence simulé. L'un des objectifs de l'assimilation de couleur de l'eau est donc de corriger ces structures à partir du phytoplancton de surface.

Les profils verticaux de l'ensemble assimilé montrent une dispersion de l'ensemble qui réduit certains biais à la *vérité*. Toutefois, l'ensemble semble se disperser de plus en plus avec le temps. À cette dispersion croissante avec le temps s'ajoute l'apparition de profils peu réalistes présentant des concentrations de phytoplancton oscillant entre 30 et 40 mètres.

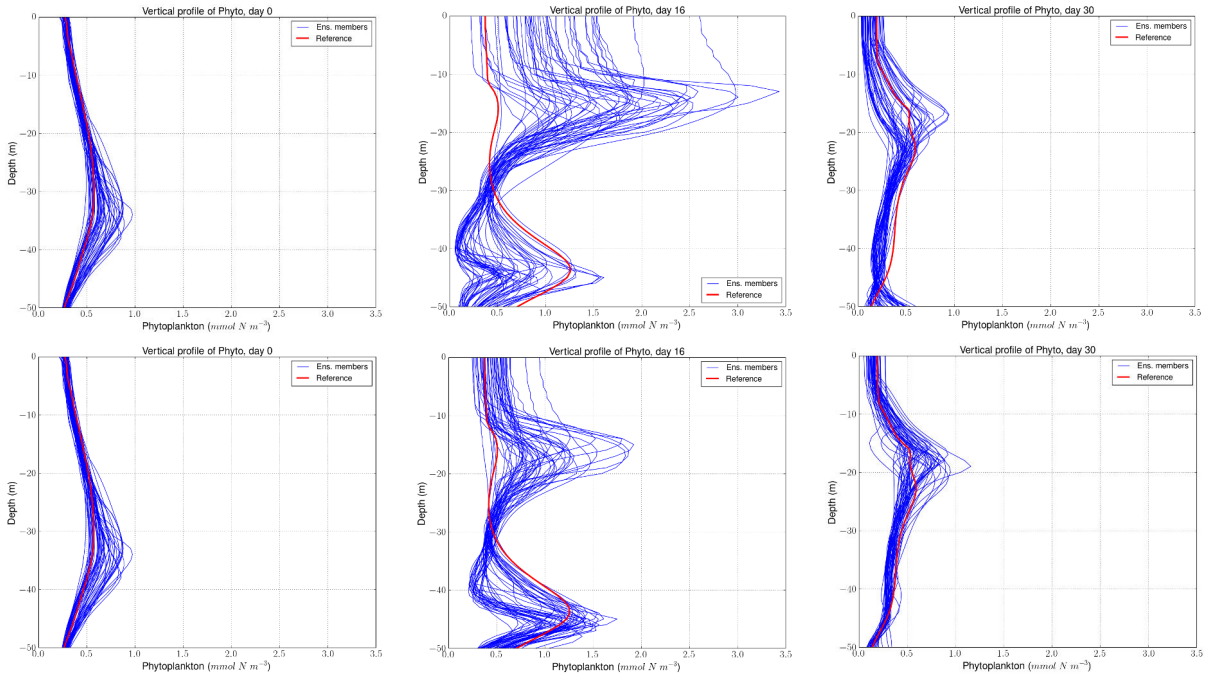

Figure 7.7 – *Profils verticaux (0m-50m) du phytoplancton total de l'ensemble (en bleu) libre (graphiques supérieurs) et produit par l'ETKF (graphiques inférieurs) sur 50m (droite) de profondeur au jour 0 (gauche), 16 (centre) et 30 (droite). La vérité est repésentée en rouge.*

Au vu des séries temporelles et des profils verticaux de l'ensemble assimilé par l'ETKF, la propagation de l'information couleur de l'eau le long de la colonne d'eau

par régression linéaire ne semble pas appropriée. Pour une évaluation quantitative de ces performances, nous évaluons dans la suite le score RMSE et l'histogramme de rangs de cet ensemble.

Qualité de l'estimé moyen : Les RMSE spatiales (en fonction du temps) et temporelles (en fonction de l'espace) sont calculées pour l'estimé moyen de l'ensemble libre et pour l'estimé moyen de l'ensemble ETKF. Ces scores sont présentés sur la Figure 7.8. Les RMSE spatiales (à gauche) et temporelles (à droite) sont en rouge pour l'ensemble libre et en bleu pour l'ensemble assimilé.

En regardant les RMSE spatiales, on observe que l'ensemble libre commet des erreurs importantes sur deux périodes : entre le jour 0 et le jour 10 puis entre le jour 15 et le jour 20. D'après la zone d'intérêt (Fig. 7.5) que l'on regarde, ces erreurs correspondent, respectivement, à la mauvaise représentation du bloom de début du mois et à la mauvaise représentation de l'approfondissement du phytoplancton à la suite du bloom du début du mois. L'ETKF n'a que peu d'impact sur l'amélioration de l'estimation du bloom. Les séries temporelles nous ont montré la faible dispersion de l'ensemble à cette période, ce qui explique le faible impact des observations et donc la faible correction. En revanche, l'ETKF parvient à diminuer son erreur RMS sur la période d'approfondissement du phytoplancton. Par ailleurs, la RMSE temporelle de l'ensemble assimilé est améliorée le long de la verticale. Ainsi l'ETKF montre qu'il parvient, malgré la présence de profils peu réalistes précédemment observés, à propager une partie de l'information sur la verticale et ainsi à diminuer les erreurs RMS de sa moyenne d'ensemble.

Dispersion de l'ensemble : Nous nous intéressons à présent à la dispersion de l'ensemble produit par l'ETKF. Pour ce faire, nous utilisons le diagnostic de l'histogramme de rangs. L'histogramme de l'ensemble libre et celui de l'ensemble assimilé ont été calculés et sont représentés, respectivement, sur le graphique de gauche et de droite de la Figure 7.9.

L'ensemble libre présente de nombreux biais supérieurs (sous-estimation de la *vérité*) et quelques biais inférieurs (sur-estimation de la *vérité*). Comme on l'a vu sur les séries temporelles, les biais supérieurs correspondent aux 10 premiers jours pendant lesquels l'ensemble est peu dispersé et a des concentrations de phytoplancton bien inférieures à la *vérité*.

Les biais sont fortement réduits par l'assimilation. Cependant, comme le montrent les séries temporelles, plusieurs biais supérieurs persistent notamment au cours des 10 premiers jours.

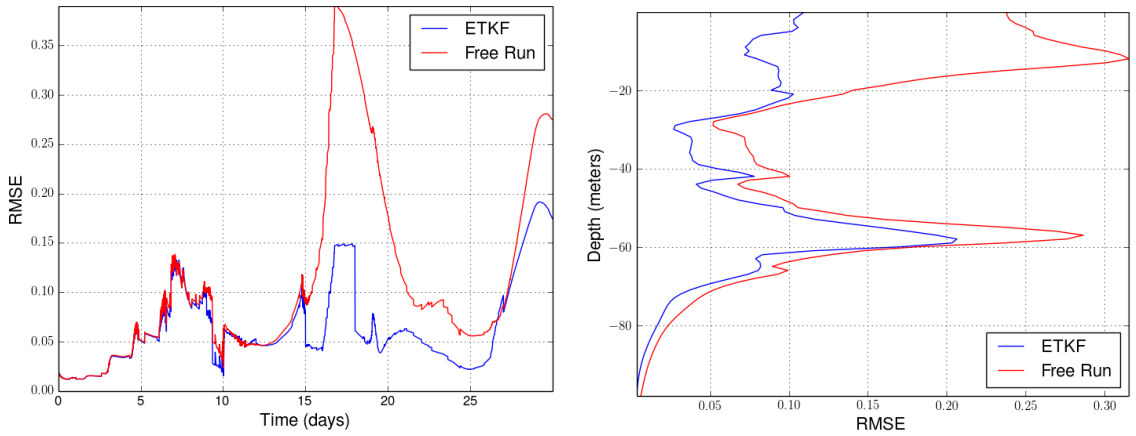

Figure 7.8 – *RMSE verticale en fonction du temps (graphique de gauche) et temporelle en fonction de la verticale (graphique de droite) sur le phytoplancton total de l'estimé moyen par l'ensemble libre (en rouge) et par l'ensemble produit par l'ETKF (en bleu). Les RMSE sont calculées sur la zone d'intérêt (voir Fig. 7.5).*

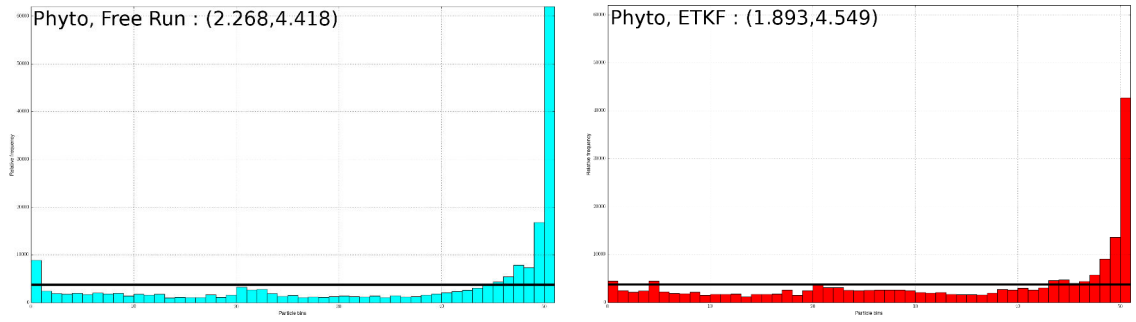

Figure 7.9 – *Histogrammes de rangs sur le phytoplancton total pour l'ensemble libre (graphique de gauche) et pour l'ensemble produit par l'ETKF (graphique de droite). Le $RCRV = (B, D)$ est donné sur l'histogramme. Les histogrammes de rangs sont calculés sur la zone d'intérêt (voir Fig. 7.5).*

Le score RCRV de l'ensemble libre vaut $(2.268, 4.418)$. L'ETKF réduit le RCRV à $(1.893, 4.549)$. Bien que le RCRV de l'ensemble assimilé ne soit pas centré réduit, l'assimilation diminue le biais de l'ensemble.

Le nitrate

Le nitrate n'est pas une variable de contrôle pour l'assimilation. Cependant les corrections de phytoplancton modifient les concentrations de nitrate via les liens étroits entre le phytoplancton et le nitrate au sein de la dynamique du modèle. Étudier les

estimations de nitrate est donc un bon indicateur pour évaluer la qualité des corrections apportées par l'assimilation. En effet, des estimations de nitrate améliorées impliquent que les estimations de phytoplancton soient améliorées et que les corrections apportées au phytoplancton génèrent des états équilibrés.

L'estimé moyen du nitrate : Comme pour le phytoplancton, les RMSE de nitrate sont très peu modifiées pendant les premiers quinze jours. La Figure 7.10 qui, de la même manière que la Figure 7.8, montre les RMSE spatiale et temporelle mais pour la variable de nitrate, témoigne de cette faible correction. Entre le jour 18 et le jour 27, l'estimation des concentrations de nitrate par l'ETKF est considérablement améliorée. De même, l'estimé moyen de la concentration de nitrate entre 40 et 60m est améliorée par l'assimilation.

En terme de son estimé moyen, l'ETKF produit des corrections du phytoplancton qui ont, via la physique du modèle, un impact bénéfique sur la représentation des concentrations de nitrate.

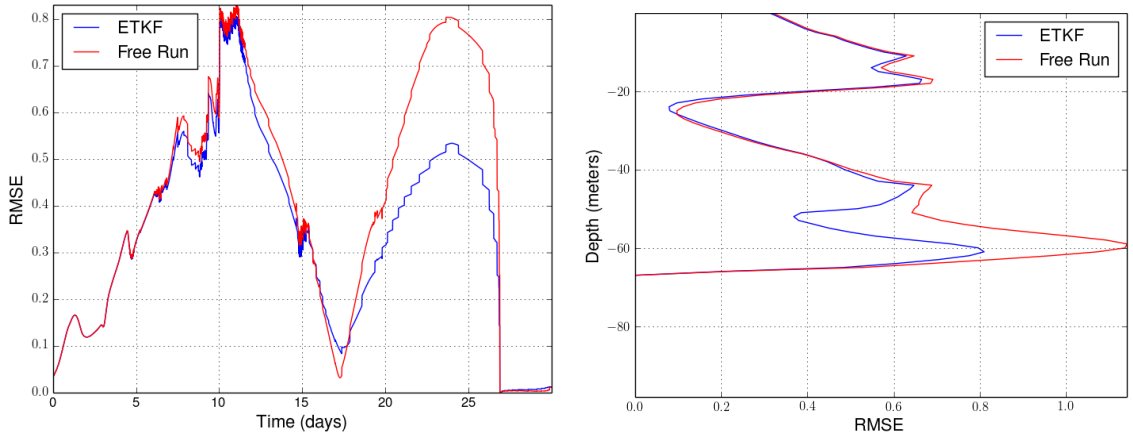

Figure *7.10* – *Idem à la Figure 7.8 sur le nitrate.*

Dispersion du nitrate : Les histogrammes de rangs des deux ensembles sur le nitrate (Figure 7.11), indiquent que l'ensemble des concentrations de nitrate engendré par l'assimilation réduit quelques biais. Il est cependant à noter, que l'histogramme de l'ensemble assimilé présente une forme légérement en W. Les côtés de l'histogramme témoignent de la présence (réduite) de biais alors que le centre de l'histogramme peut révéler un début de sur-dispersion de l'ensemble. Malgré la réduction de biais, il semble que la dispersion de l'ensemble en nitrate soit faiblement impactée.

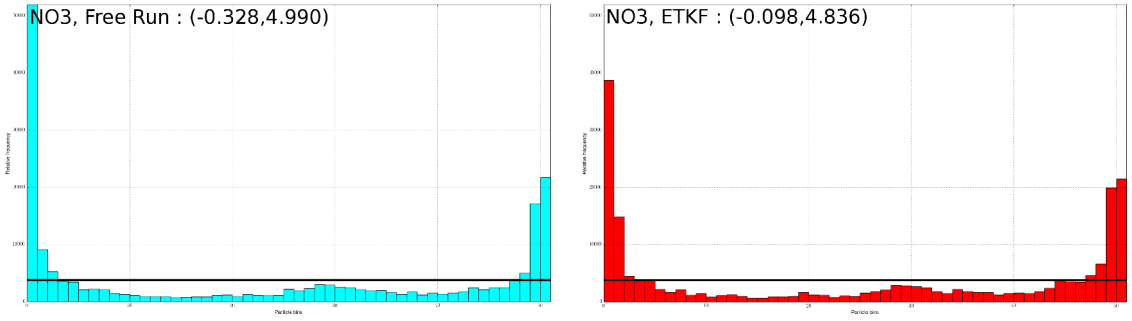

Figure *7.11* – *Idem à la Figure 7.9 sur le nitrate.*

La faible évolution du RCRV de l'ensemble libre $(-0.328, 4.990)$ au RCRV de l'ensemble après assimilation $(-0.098, 4.836)$ confirme ce faible impact.

Le contrôle du phytoplancton par le filtre ETKF a donc aussi un impact bénéfique sur le biais, même s'il est limité, sur la dispersion du nitrate.

7.2.2 Peu d'impact sur la dynamique

De même que le nitrate, les variables dynamiques du système ne sont pas des variables de contrôle pour l'assimilation. La rétroaction de la biogéochimie se réalise au travers de la dépendance des flux de chaleur dans les couches de surface à la concentration en phytoplancton. Une forte concentration de phytoplancton en surface diminue la pénétration de la lumière dans la colonne d'eau et, par conséquent, diminue les échanges de chaleur.

Ceci étant dit cette rétroaction est faible et possède une grande inertie. Dans ce sens, observer un impact conséquent des variations de phytoplancton sur la dynamique au cours d'un seul mois n'est pas possible. Comme le montre (Fig. 7.12) les scores de RMSE spatiales de l'estimé moyen de l'ensemble ETKF pour les variables de température (à gauche) et de salinité (à droite), l'impact ici constaté est nul.

Bien que prévisible, ce résultat confirme que l'amélioration de la dynamique du système par assimilation de données de la couleur de l'eau ne peut se faire qu'en incluant dans le vecteur de contrôle au moins une variable dynamique.

7.2.3 Bilan

D'après l'évaluation visuelle (les séries temporelles et les profils verticaux) plusieurs biais entre l'ensemble libre et la *vérité* sont corrigés par l'assimilation ETKF. Ce résultat est également confirmé par l'évolution des histogrammes de rangs. Cependant, les profils verticaux semblent peu réalistes, ce qui implique que l'information

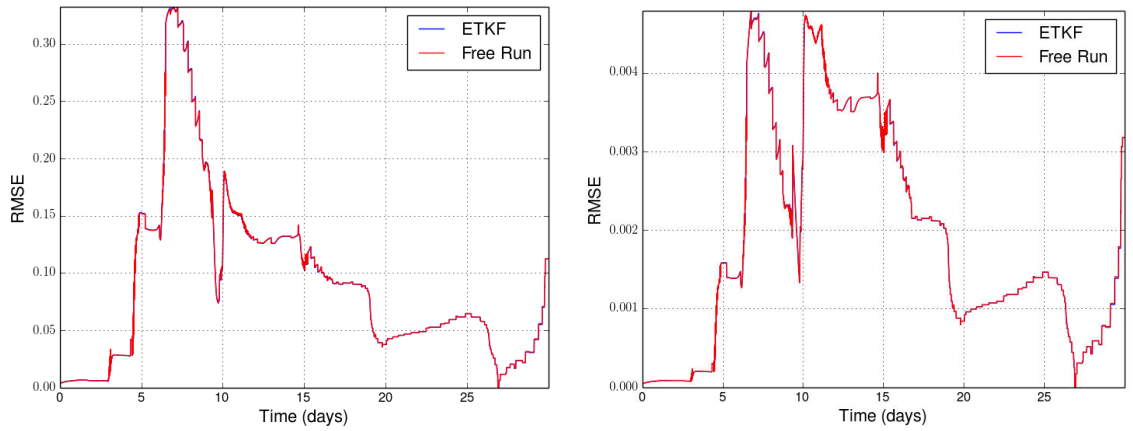

Figure ***7.12*** – *Idem au graphique de gauche de la Figure 7.8, sur la température (graphique de gauche) et la salinité (graphique de droite).*

n'est pas correctement propagée le long de la colonne d'eau. Malgré la création de profils peu réalistes l'ETKF parvient à réduire les erreurs RMS et améliore la dispersion d'ensemble (les histogrammes de rangs sont aplanis et les RCRV réduits).

En résumé, selon les scores RMSE, les histogrammes de rangs et les RCRV, l'ETKF réussit à corriger le phytoplancton. Toutefois, l'apparition de profils peu réalistes laisse penser qu'une assimilation sur une période plus longue ou avec une fréquence d'observation plus grande peut conduire à la dégradation des solutions proposées par l'ETKF.

Par ailleurs, la correction du phytoplancton par l'ETKF a un impact bénéfique sur l'estimé moyen de nitrate et quelques biais de l'ensemble sur le nitrate sont corrigés. Les histogrammes de rangs et la faible amélioration des scores RCRV indiquent, par contre, que la dispersion de l'ensemble n'est que peu modifiée. Ce qui signifie que les biais sont quantitativement réduits (la taille de l'intervalle entre l'ensemble et la *vérité* est réduite) mais les biais sont toujours présents ce qui ne modifie pas les histogrammes de rangs.

La partie dynamique du système, en revanche, ne voit pas l'information issue de l'observation de la couleur de l'eau. La température et la salinité ne sont pas impactées par les corrections sur le phytoplancton.

Afin d'améliorer la dynamique du système à partir d'observation de la couleur de l'eau, il est donc nécessaire d'inclure des variables dynamiques dans le vecteur de contrôle.

C'est ce que nous faisons dans la section suivante, en ajoutant la température au vecteur de contrôle.

7.3 Contrôler la température par assimilation de couleur de l'eau

Dans cette partie, nous examinons la possibilité de contrôler la température à partir de données de la couleur de l'eau. Dans ce nouveau problème d'estimation le vecteur de contrôle contient le phytoplancton et la température.

Avant de poursuivre ces expériences, nous justifions l'intérêt de contrôler la température à partir de la couleur de l'eau. Ensuite, nous évaluons l'estimation de la température et de la salinité par le filtre de Kalman d'ensemble transformé, l'ETKF. Puis, deux méthodes d'assimilation non-Gaussiennes, l'EnKF anamorphosé et le MRHF, sont mises en place. Leurs estimations de température, de salinité, de phytoplancton et de nitrate sont comparées à celles de l'ETKF. L'objectif est de savoir si l'assimilation de données non-Gaussiennes peut apporter une solution à ce problème d'estimation.

Les trois méthodes sont soumises à une inflation des anomalies d'ensemble après analyse avec un coefficient d'inflation de 1.2.

7.3.1 Pourquoi contrôler la température avec la couleur de l'eau ?

Plusieurs types d'observations fournissent de l'information sur la température de l'océan. La couleur de l'eau peut ne pas sembler être la source d'information la plus naturelle. Cependant elle présente certains avantages. Nous évoquons ici les caractéristiques des observations de la température issues des flotteurs, des satellites mesurant la SST et de la couleur de l'eau.

Les flotteurs : Les profils de température fournis par les flotteurs produisent des observations sur une grande partie de la colonne d'eau. De plus, ces observations sont précises avec une faible erreur de mesure. En revanche, la répartition spatiale des flotteurs est encore assez éparse et inhomogène.

La SST : La température de l'eau de surface (SST) pour contrôler la température plutôt que la couleur de l'eau semblerait plus facile et plus directe. En pratique, les observations de SST fournissent une information sur ce qui s'appelle la *température de peau* de l'océan. Il s'agit de la température de la pellicule d'eau de quelques millimètres affleurant à la surface de l'océan. Cette température n'est pas nécessairement représentative de la température des premiers mètres de la colonne d'eau. La plupart des modèles ne représentant pas cette pellicule d'eau, l'observation n'est donc pas représentative de la température de surface représentée par ces modèles.

La couleur de l'eau : La couleur de l'eau est une donnée satellite qui présente donc les avantages d'une bonne répartition spatiale et d'une répartition temporelle

qui tend à grandir avec l'apparition des satellites géostationnaires de couleur de l'eau. Bien qu'indirecte et difficile à extraire, l'information sur la température fournie par l'observation de la couleur de l'eau correspond à la température des premières couches (telle que représentée par les modèles). Il semble donc intéressant de s'armer pour résoudre ce problème inverse.

7.3.2 Échec de l'ETKF

Les corrections que peut apporter l'ETKF à la température proviennent d'une régression linéaire des corrections du phytoplancton de surface. Or la Figure 7.3 de la Section 7.1 nous a montré que la relation entre le phytoplancton de surface et la température est peu linéaire. On peut donc s'attendre *a priori* à une faible amélioration, par l'ETKF, de la température et à plus grande échelle du système dynamique.

Pour vérifier cette assertion, nous évaluons à présent la qualité de l'estimé moyen et de l'ensemble produit par l'ETKF de la température ainsi que son impact (indirect via l'équilibre du modèle) sur la salinité.

Qualité de l'estimé moyen : Nous avons vu à la section précédente que les scores RMSE de la température et de la salinité (Fig. 7.12) n'étaient pas améliorés par le contrôle du phytoplancton seul. La Figure 7.13 présente les mêmes scores RMSE que la Figure 7.12 après contrôle par l'ETKF du phytoplancton et de la température.

La RMSE de la température est légérement réduite entre le $6^{\text{ème}}$ et le $27^{\text{ème}}$ jour. En revanche les corrections apportées par l'ETKF ne semblent pas diminuer l'erreur sur les 6 premiers jours et dégrade l'estimé moyen sur les 3 derniers jours. Cette dégradation correspond à l'apparition d'une forte concentration de phytoplancton entre 20 et 30 mètres de profondeur (Fig. 7.5), mal représentée par l'ETKF.

Cette faible réduction d'erreur sur le phytoplancton n'est pas suffisante pour correctement améliorer l'estimation de la salinité. La RMSE sur la salinité indique une très faible amélioration au milieu du mois.

Ces scores de RMSE confirment l'idée *a priori* que nous avions sur les performances limitées de l'ETKF à produire un estimé moyen.

Dispersion de l'ensemble : Les histogrammes de rangs de l'ensemble libre et de l'ensemble assimilé sur la température et sur la salinité (Fig. 7.14) indiquent toutefois une réduction des biais sur ces deux variables. En effet, la présence de pics à l'extrémité droite des histogrammes de l'ensemble libre (graphiques de gauche) est symptomatique de nombreux biais entre l'ensemble libre et la *vérité*. La réduction de

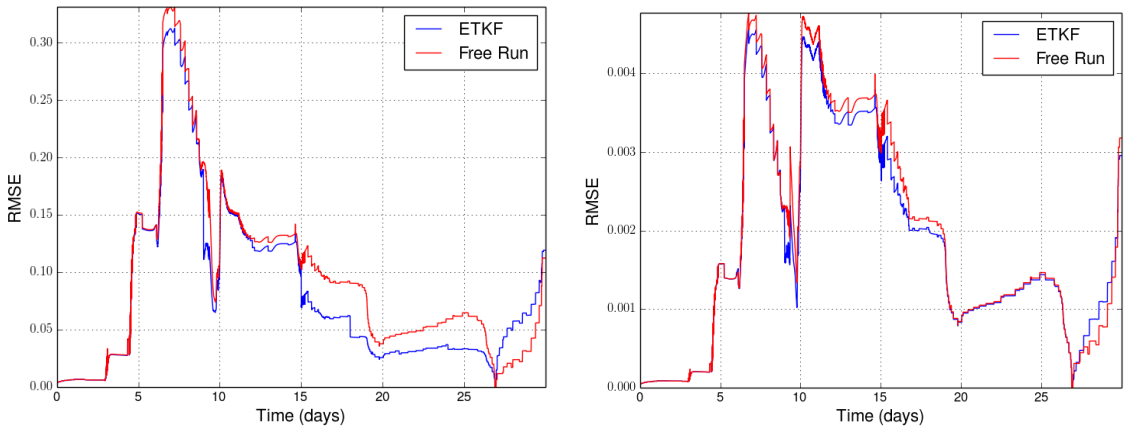

Figure *7.13* – *Idem à la Figure 7.12 pour l'ETKF contrôlant le phytoplancton et la température.*

ces pics indique que la sous-estimation importante de la *vérité* par l'ensemble libre est considérablement réduite par l'assimilation de l'ETKF.

Les valeurs de RCRV de l'ensemble libre sont de $(2.497, 4.255)$ pour la température et de $(2.204, 4.204)$ pour la salinité. Après assimilation, ces scores sont respectivement réduits à $(1.414, 3.153)$ et $(1.884, 3.766)$. Ces résultats indiquent également que l'assimilation ETKF parvient à réduire le biais de l'ensemble libre et à améliorer sa dispersion.

Bien que l'ETKF ne réduise que faiblement les erreurs RMSE de son estimé moyen, il améliore sensiblement la dispersion d'ensemble.

7.3.3 Apport de l'assimilation de données non-Gaussienne

Dans cette sous-partie nous cherchons à voir si l'utilisation d'une méthode d'assimilation prenant en compte les non-Gaussianités et les non-linéarités intervariables, telles que décrites en Section 7.1, peut améliorer l'estimation des variables dynamiques et biogéochimiques à partir de la couleur de l'eau. Pour ce faire, nous mettons en place deux méthodes d'assimilation précedemment décrites : l'EnKF anamorphosé (dont l'algorithme est décrit au Chapitre 2) et le MRHF (développé au Chapitre 4).

Ces deux méthodes sont également appliquées avec une inflation d'ensemble de coefficient $\alpha = 1.2$. L'inflation est nécessaire dans ce cas car l'augmentation de la fréquence d'observation engendre un contrôle trop fréquent de l'ensemble qui n'a pas le temps de se (re)disperser suffisamment entre deux pas de temps observés.

Nous comparons à présent les scores des trois méthodes (ETKF, AnamEnKF et MRHF) sur l'expérience d'assimilation de la couleur de l'eau contrôlant le phyto-

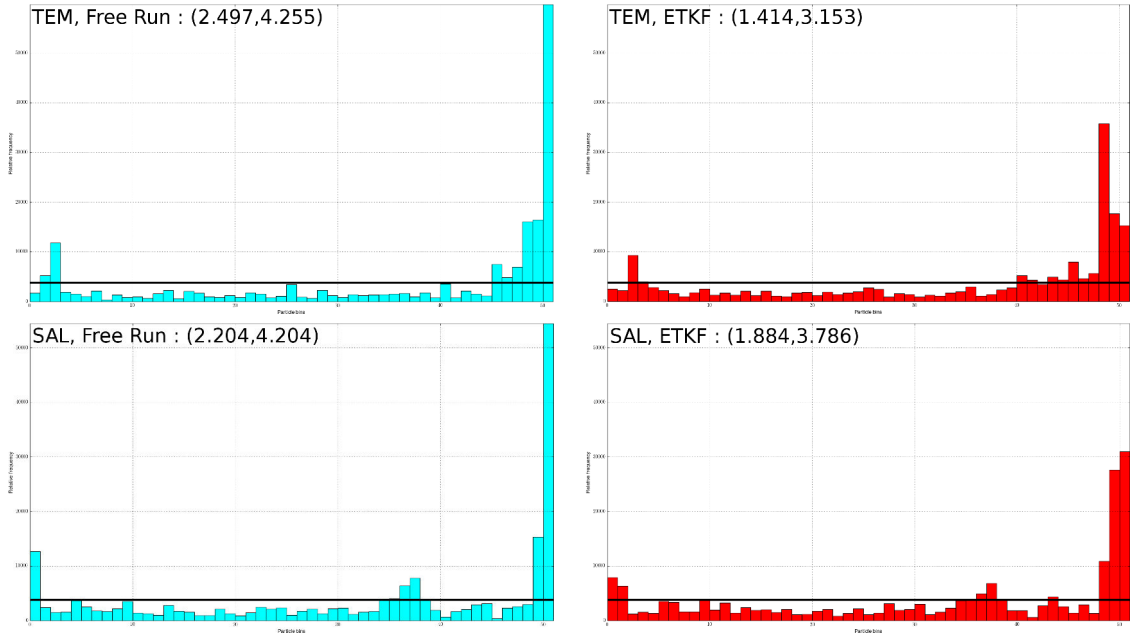

Figure 7.14 – *Histogrammes de rangs sur la température (graphiques supérieurs) et sur la salinité (graphiques inférieurs) pour l'ensemble libre (graphique de gauche) et pour l'ensemble produit par l'ETKF (graphique de droite) contrôlant le phytoplancton et la température. Le RCRV = (biais, dispersion) est donné sur l'histogramme. Les histogrammes de rangs sont calculés sur la zone d'intérêt (voir Fig. 7.5).*

plancton et la température. Les scores étudiés sont : un score sur l'estimé moyen produit : l'erreur RMS en espace moyennée en temps ; deux scores sur l'ensemble produit : les histogrammes de rangs (et les RCRV) et le CRPS calculés en espace et en temps sur la zone d'intérêt.

RMSE moyenne des différentes méthodes

Les erreurs RMS spatiales sur la température, la salinité, le phytoplancton et le nitrate ont été calculées pour les trois méthodes. La Figure 7.15 compare la moyenne sur le mois de ces RMSE. Cette figure permet d'avoir une vision globale de la qualité des estimés moyens produits par les trois méthodes.

La première information, qui ressort de ces scores, est que l'utilisation des filtres non-Gaussiens (l'EnKF anamorphosé et le MRHF) améliore globalement la précision de l'estimé moyen. En effet, à l'exception de la mauvaise représentation du nitrate par

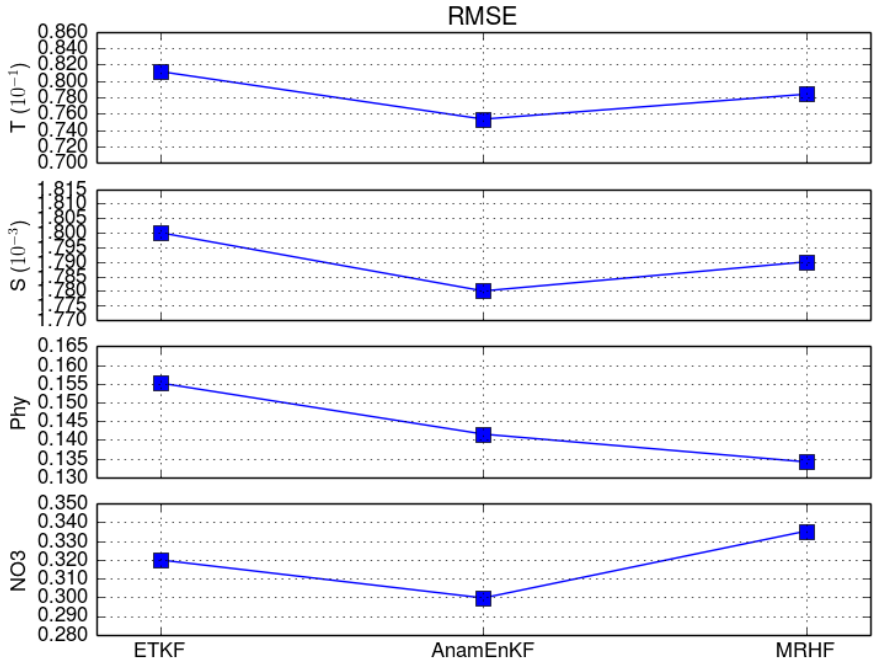

Figure 7.15 – *RMSE verticales moyennées sur le mois d'avril pour la température (ligne 1), la salinité (ligne 2), le phytoplancton total (ligne 3) et le nitrate (ligne 4) des estimés moyens produits par l'ETKF (gauche), l'EnKF anamorphosé (centre) et le MRHF (droite). Les RMSE sont calculées sur la zone d'intérêt (voir Fig. 7.5).*

le MRHF, les RMSE de l'EnKF anamorphosé et du MRHF sont systématiquement plus faibles que celles de l'ETKF.

Par ailleurs, la très bonne estimation du phytoplancton par le MRHF indique notamment que la représentation de la pdf de phytoplancton de surface par histogrammes de rangs est appropriée à ce type de distribution. En revanche, la mauvaise répercussion de ces corrections sur le nitrate pourrait être symptomatique de la génération d'états mal équilibrés. Ce qui indique que le MRHF nécessite encore du développement.

Enfin, hormis sur le phytoplancton, l'EnKF anamorphosé est le filtre le plus performant en terme de réduction d'erreurs de l'estimé moyen.

Dispersion, fiabilité et résolution

Nous nous intéressons cette fois aux ensembles produits par les assimilations. Nous utilisons comme diagnostics : l'histogramme de rangs, le RCRV et le CRPS.

Histogrammes de rangs et RCRV : Les histogrammes de rangs des trois méthodes sont présentés sur la Figure 7.16 pour la température (à gauche) et la salinité (à droite) et sur la Figure 7.17 pour le phytoplancton (à gauche) et le nitrate (à droite). Les scores RCRV sont donnés sur leurs histogrammes respectifs.

Sur les variables contôlées (i.e. le phytoplancton et la température), l'assimilation par filtres non-Gaussiens produit des ensembles mieux dispersés que l'assimilation aux moindres carrés. Les histogrammes de rangs de l'EnKF anamorphosé et du MRHF sont plus plats que ceux de l'ETKF et les biais de l'ensemble sont diminués.

Les variables non-contôlées (i.e. le nitrate et la salinité) sont moins sensibles (par impact indirect) au choix du filtre. Sur ces variables, les RCRV et les histogrammes de rangs sont peu modifiés par les différents filtres à l'exception du MRHF. En effet, contrairement à ce que sa mauvaise RMSE de nitrate (Fig. 7.15) laissait penser, le MRHF améliore la dispersion de l'ensemble sur les variables non-estimées. Ce résultat est signe d'un bon équilibre des corrections du MRHF puisque les corrections apportées aux phytoplanctons (variable estimée) améliore *via* le modèle la représentation des variables non-estimées.

De manière générale, le MRHF produit l'ensemble le mieux dispersé selon les scores RCRV et les histogrammes de rangs.

Comparaison des CRPS : Nous évaluons à présent les ensembles à l'aide du diagnostic de vérification d'ensemble le plus complet pour évaluer la fiabilité et la résolution : le CRPS. Les résultats qui sont présentés sur la Figure 7.18 correspondent aux CRPS des trois méthodes pour les quatre variables précédemment étudiées. Les CRPS sont décomposés en un terme de fiabilité (en bleu) et un terme de CRPS potentiel (en orange) selon la décomposition de Hersbach (2000).

L'information majeure de ces résultats est, encore une fois, que les performances des filtres non-Gaussiens en terme de CRPS sont meilleures que celles d'un filtre moindres carrés. En effet, les CRPS (CRPS potentiel + fiabilité) sur toutes les variables examinées sont améliorés par l'utilisation des filtres non-Gaussiens.

Le terme de CRPS potentiel est également réduit par l'utilisation des filtres non-Gaussiens, ce qui signifie que l'EnKF anamorphosé et le MRHF parviennent mieux à

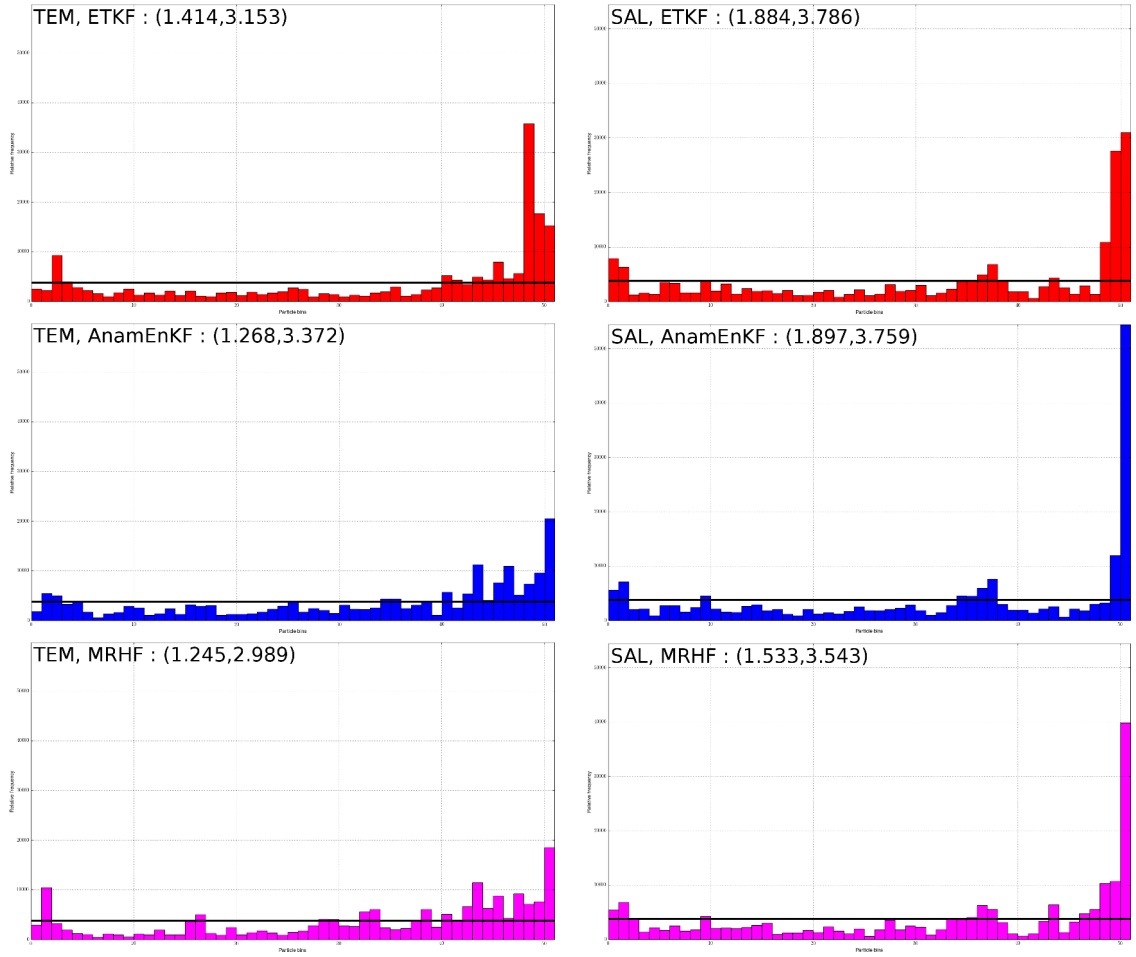

Figure 7.16 – *Histogrammes de rangs sur la température (à gauche) et la salinité (à droite) pour l'ensemble produit par l'ETKF (en rouge), l'EnKF anamorphosé (en bleu) et le MRHF (en magenta) contrôlant le phytoplancton et la température. Le RCRV* $= (biais, dispersion)$ *est donné sur l'histogramme. Les histogrammes de rangs sont calculés sur la zone d'intérêt (voir Fig. 7.5).*

extraire de l'information nouvelle des observations. À l'exception du phytoplancton, il en est de même pour le terme de fiabilité. Le terme de fiabilité du phytoplancton pour l'ensemble produit par l'ETKF est très bas et ce malgré une dispersion (selon l'histogramme de rangs) médiocre. Ce phénoméne reste à ce jour inexpliqué mais repose sans doute sur une limitation de la décomposition de Hersbach.

Les CRPS des deux méthodes non-Gaussiennes sont proches sur toutes les variables. Cependant, selon cette décomposition, l'ensemble produit par l'EnKF anamorphosé est globalement plus fiable que celui du MRHF. Encore une fois ce dernier

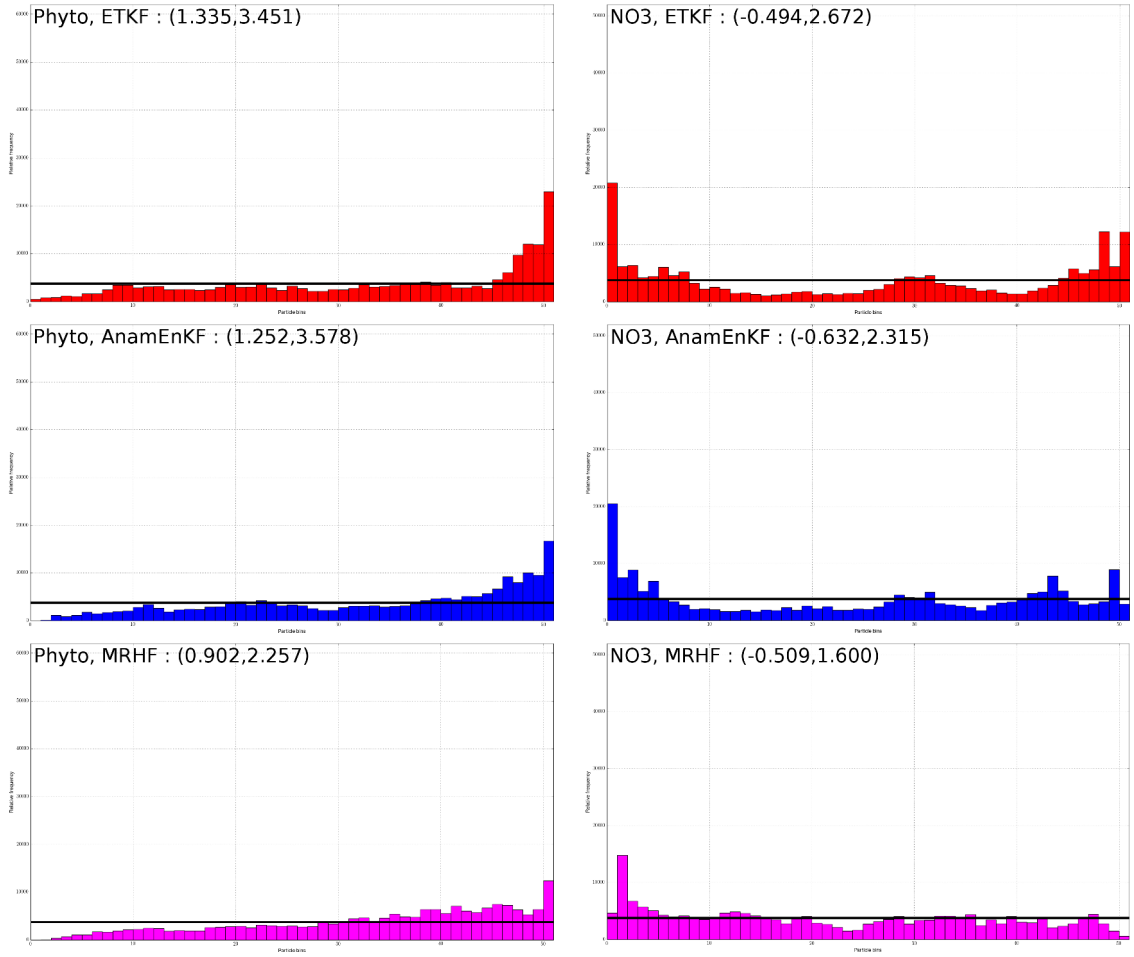

Figure 7.17 – *Histogrammes de rangs sur le phytoplancton total (à gauche) et le nitrate (à droite) pour l'ensemble produit par l'ETKF (en rouge), l'EnKF anamorphosé (en bleu) et le MRHF (en magenta) contrôlant le phytoplancton et la température. Le RCRV* $= (biais, dispersion)$ *est donné sur l'histogramme. Les histogrammes de rangs sont calculés sur la zone d'intérêt (voir Fig. 7.5).*

résultat est en désaccord avec la dispersion constatée précedemment.

7.3.4 Bilan

La première conclusion à tirer de cette section est que la dynamique peut être en partie corrigée par assimilation de la couleur de l'eau.

Cependant, l'utilisation d'un filtre Gaussien ne parait pas optimale.

La deuxième conclusion à tirer de cette section est que l'estimation de la biogéochimie et de la dynamique bénéficie de l'utilisation de méthodes d'assimilation non-Gaussiennes.

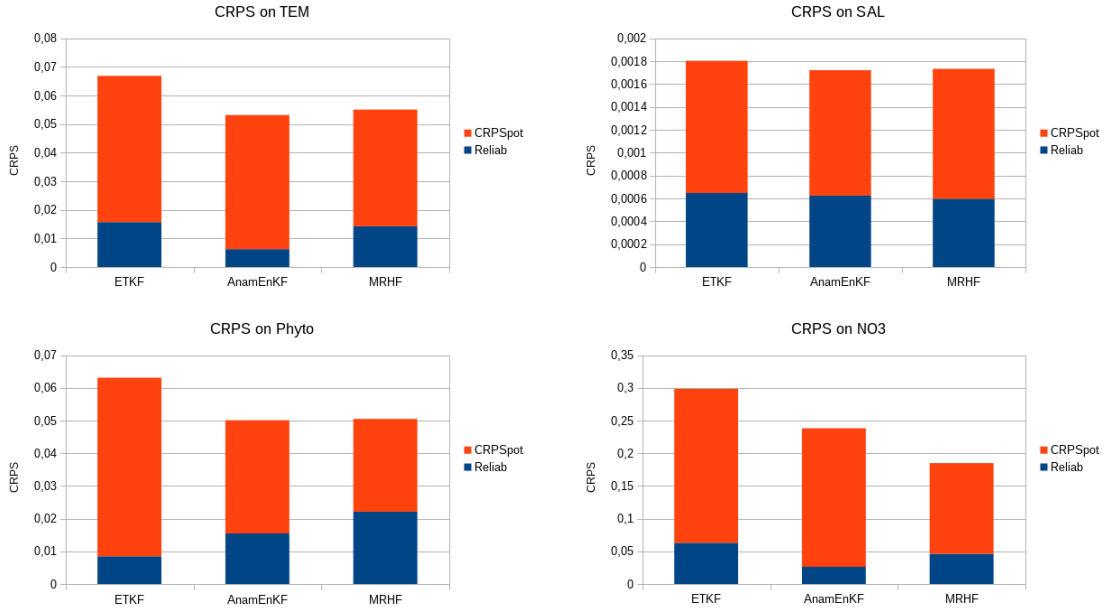

Figure 7.18 – *CRPS sur la température (graphique supérieur-gauche), la salinité (graphique supérieur-droit), le phytoplancton total (graphique inférieur-gauche) et le nitrate (graphique inférieur-droit) pour l'ensemble produit par l'ETKF, l'EnKF anamorphosé et le MRHF contrôlant le phytoplancton et la température. Le CRPS est la somme du terme de CRPS potentiel (en orange) et du terme de fiabilité (en bleu). Les CRPS sont calculés sur la zone d'intérêt (voir Fig. 7.5).*

Les erreurs RMS sont systématiquement réduites par l'EnKF anamorphosé et par le MRHF. Globalement, l'EnKF anamorphosé est le filtre le plus performant en terme de réduction d'erreurs de l'estimé moyen. De plus, ces deux méthodes produisent des ensembles mieux dispersés, plus fiables et de meilleure résolution que l'ETKF. Encore une fois, l'ensemble produit par l'EnKF anamorphosé est globalement plus fiable que celui produit par le MRHF. Les bons résultats du MRHF indiquent, tout de même, que la représentation par histogrammes de rangs des densités de probabilités est adaptée á l'observation de la couleur de l'eau.

7.4 Vers un satellite géostationnaire

Les satellites géostationnaires circulent sur des orbites hautes (36 000km) à une vitesse égale à la rotation terrestre, ce qui leur permet d'observer de manière quasi-constante une région fixe à la surface du globe. Depuis la mission coréenne GOCI [1],

1. *Geostationary Ocean Color Imager*

qui a été la première mission “couleur de l’eau” lancée sur un satellite géostationnaire, plusieurs autres sont en projet (e.g. Geo-OCAPI). L’avantage de ce type de satellites est qu’ils permettent une plus grande fréquence d’observations.

Dans le cadre de l’assimilation de données, il est bon de se demander à quel point une fréquence plus élevée d’observations peut améliorer l’estimation. De plus, nous l’avons vu, l’assimilation de données de couleur de l’eau est un problème non-Gaussien. L’augmentation de la fréquence d’observations peut, dans ce cas, accentuer la complexité du problème d’estimation.

Avec le même schéma d’expérience que précédemment, nous augmentons la fréquence d’observations de couleur de l’eau de trois jours (OC3j) à un jour (OC1j). L’objectif de cette section est de savoir : (i) quel est le gain apporté par une plus grande fréquence d’observation de la couleur de l’eau ; (ii) si l’assimilation non-Gaussienne offre une meilleure réponse à ce problème d’estimation que l’assimilation aux moindres carrés. Pour ce faire, nous comparons de nouveau les trois méthodes : l’ETKF, une méthode moindres carrés ; l’EnKF anamorphosé, une méthode non-Gaussienne à partir d’un cadre classique ; le MRHF, une méthode fondamentalement non-Gaussienne.

Les trois méthodes sont utilisées sans inflation. Contrairement au cas précédent d’une observation tous les trois jours, la fréquence d’observation est trop grande pour qu’une inflation ait un impact. En effet, vingt-quatre heures de propagation ne semblent pas suffisant pour que l’inflation est un impact notable (en terme d’histogramme de rangs) sur la dispersion de l’ensemble. Des tests ont été effectué avec différents coefficients d’inflation (non montrés ici) et mènent aux mêmes résultats que l’assimilation sans inflation.

Il est convenu que ce schéma d’expériences est encore loin du problème réel de l’assimilation opérationnelle de la couleur de l’eau. Cependant, ces expériences permettent de révéler d’éventuelles difficultés auxquelles l’assimilation opérationnelle devra se confronter.

7.4.1 Couleur de l’eau haute-fréquence

De la même manière que dans la section précédente (Fig. 7.15), les erreurs RMSE spatiales des estimés moyens produits par l’ETKF, l’EnKF anamorphosé et le MRHF ont été calculées. La Figure 7.19 présente les moyennes sur le mois d’avril de ces erreurs. Nous redonnons sur la figure, les résultats précédents produits par les trois méthodes dans la configuration couleur de l’eau à “basse fréquence”, une observation tous les trois jours (OC3j, en bleu). Les résultats produits dans la configuration couleur de l’eau à “haute fréquence”, une observation tous les jours (OC1j), sont représentés en vert.

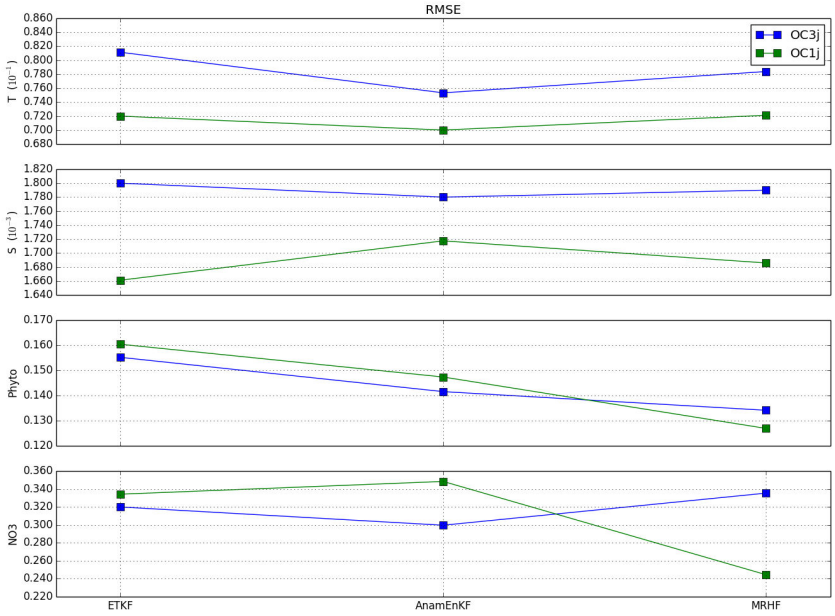

Figure *7.19* – *Idem que la Figure 7.15 pour l'ETKF (à gauche), l'EnKF anamorphosé (au centre) et le MRHF (à droite) dans les configurations une observation couleur de l'eau tous les 3 jours (en bleu) étudiée à la Section 7.3 et une observation tous les jours (en vert). Les RMSE sont calculées sur la zone d'intérêt (voir Fig. 7.5).*

Pour répondre à la question du gain obtenu par l'assimilation d'observations hautes fréquences, en terme de score RMSE, il suffit de comparer les résultats OC1j et OC3j de chaque méthode. L'ETKF réussit à réduire son erreur sur la température et la salinité. En revanche, l'augmentation de la fréquence d'observations dégrade la solution de l'ETKF sur le phytoplancton et le nitrate. Ce résultat vient confirmer le fait que le problème d'assimilation de la couleur de l'eau n'est pas un problème linéaire et Gaussien. En effet, dans un problème linéaire et Gaussien, augmenter le nombre d'observations améliore toujours l'estimation.

De plus, l'EnKF anamorphosé, qui est adapté à gérer certaines non-Gaussianités et qui présentait les meilleures performances pour des observations basse-fréquence, dégrade également son estimation des variables biogéochimiques. Ceci vient confirmer la complexité d'un tel problème d'estimation.

L'utilisation d'un filtre fondamentalement adapté pour gérer les non-Gaussianités prend alors toute son importance. En effet, le MRHF se montre être un outil capable

de résoudre ce type de problème d'estimation. Les erreurs RMS sont réduites pour chacune des variables. Les erreurs RMS en température et en salinité sont du même ordre que celles de l'ETKF (légérement inférieur) et que celles de l'EnKF anamorphosé. Les variables biogéochimiques sont, en revanche, fortement améliorées par l'assimilation MRHF.

	T	S	Phy	NO3	**% Moy**
ETKF	11.3 %	7.7 %	-3.4 %	-4.5 %	**2.8%**
AnamEnKF	13.7 %	4.6 %	5.0 %	-8.9 %	**3.6%**
MRHF	11.1 %	6.3 %	18.2 %	23.6 %	**14.8%**

Figure *7.20* – *Tableau du gain (en %) des erreurs RMS en configuration OC1j pour les trois méthodes par rapport à l'utilisation de l'ETKF en configuration OC3j (configuration étudiée en Section 7.3).*

Pour avoir une idée quantitative de ce gain, le Tableau 7.20 contient le gain (en %) des réductions d'erreurs RMS en utilisant la configuration une observation par jour (OC1j) pour les deux méthodes par rapport à l'utilisation d'une observation tous les trois jours pour l'ETKF (configuration étudiée en Section 7.3) :

$$1 - \frac{RMSE(OC1jMethod)}{RMSE(OC3jETKF)} \tag{7.1}$$

où la $Method$ est l'ETKF, l'EnKF anamorphosé (AnamEnKF) ou le MRHF.

Pour les trois méthodes, les données de couleur de l'eau en fréquence journalière permettent de réaliser des réductions d'erreurs RMS de l'ordre de 11-14% sur la correction de la température et l'impact sur la salinité est de l'ordre de 4 à 8%. En revanche, dans cette configuration, la correction du phytoplancton et son impact sur le nitrate sont dégradés par l'assimilation ETKF de 3.4% pour le phytoplancton et de 4.5 % pour le nitrate. L'EnKF anamorphosé améliore de 5% le phytoplancton mais dégrade fortement (de 8.9%) l'estimation du nitrate. Alors que le MRHF parvient à réduire considérablement les erreurs RMS du phytoplancton et du nitrate de 18.2% et de 23.6%, respectivement. Le gain moyen sur ces quatre variables est de 14.8% pour le MRHF alors qu'il est de l'ordre de 3-4% pour les autres méthodes.

Dans nos expériences, l'assimilation de données de couleur de l'eau haute fréquence (de type géostationnaire) conduit à des réductions d'erreurs RMS importantes en utilisant un filtre fondamentalement non-Gaussien tel que le MRHF. Un filtre Gaussien, tel que l'ETKF, ou un filtre adapté à la non-Gaussianité à partir d'un cadre classique, tel que l'EnKF anamorphosé, améliore l'estimation des variables dynamiques mais dégrade l'estimation des variables biogéochimiques.

Les bonnes performances du MRHF sont confirmées par les scores CRPS. Comme sur la Figure 7.18, les CRPS des ensembles produits par l'ETKF, l'EnKF anamorphosé (AnamEnKF) et le MRHF pour l'assimilation de la couleur de l'eau haute fréquence sont présentés sur la Figure 7.21.

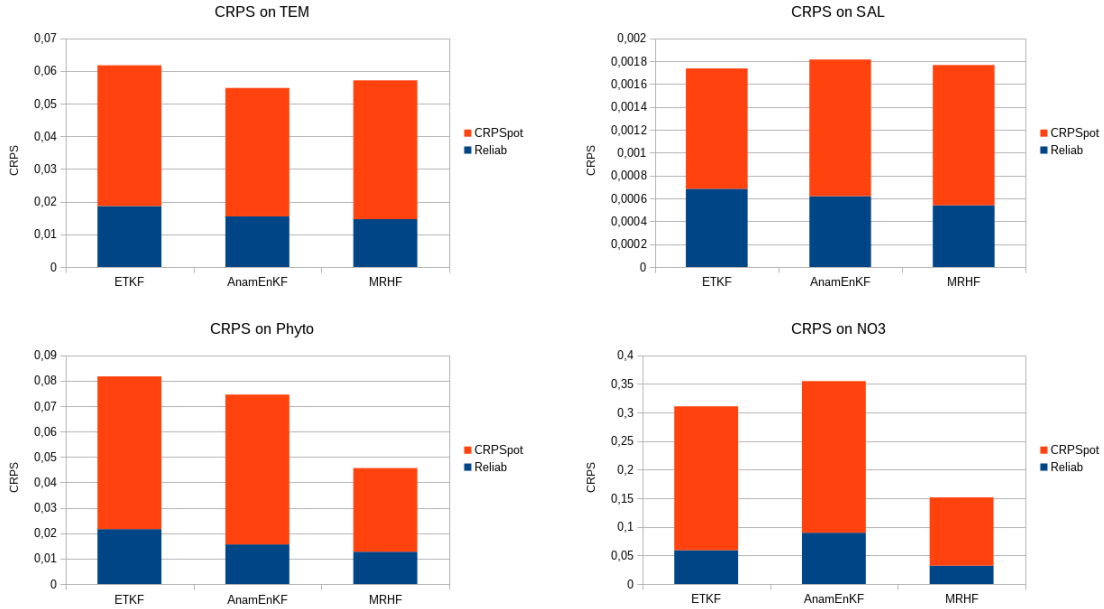

Figure 7.21 – *CRPS sur la température (graphique supérieur-gauche), la salinité (graphique supérieur-droit), le phytoplancton total (graphique inférieur-gauche) et le nitrate (graphique inférieur-droit) pour l'ensemble produit par l'ETKF, l'EnKF anamorphosé (AnamEnKF) et le MRHF dans la configuration d'une observation journalière. Le CRPS est la somme du terme de CRPS potentiel (en orange) et du terme de fiabilité (en bleu). Les CRPS sont calculés sur la zone d'intérêt (voir Fig. 7.5).*

Le CRPS des ensembles sur les variables de température et de salinité varient très peu selon les méthodes. L'ensemble produit par le MRHF est toutefois légèrement plus fiable que les deux autres ensembles.

Pour l'ETKF et l'EnKF anamorphosé, le CRPS sur la variable de nitrate n'est pas amélioré par une plus grande fréquence d'observation. De plus, l'ETKF, qui avec des observations basse fréquence présentait (Fig. 7.18) une très bonne fiabilité sur le phytoplancton, dégrade par deux son score de fiabilité.

Alors que, sur ces deux variables biogéochimiques, l'augmentation de la fréquence d'observation permet au MRHF de réduire encore son score CRPS.

7.4.2 Bilan

Les difficultés rencontrées à la section précédente deviennent plus contraignantes puisque plus fréquentes. Également, les corrections passablement équilibrées dans le problème basse fréquence (e.g. les profils verticaux de l'ensemble ETKF présentés dans la Figure 7.7) peuvent s'accroitre en haute fréquence et destabiliser l'estimation.

Ainsi, bien que l'ETKF et l'EnKF anamorphosé réussissent à réduire leurs erreurs sur la température et la salinité, l'augmentation de la fréquence d'observations dégrade leurs solutions sur le phytoplancton et le nitrate. Les erreurs RMS sont plus grandes que dans le cas basse fréquence et les CRPS sont beaucoup plus élevées que pour le MRHF. Ce résultat vient confirmer le fait que le problème d'assimilation de la couleur de l'eau est un problème non-linéaire et fortement non-Gaussien.

Dans nos expériences, utiliser un filtre fondamentalement non-Gaussien tel que le MRHF pour l'assimilation de données de couleur de l'eau haute fréquence (de type géostationnaire) conduit à des réductions d'erreurs RMS importantes (de l'ordre de 6 à 8% sur la partie dynamique et de l'ordre de 20 à 30% sur la partie biogéochimique du système) par rapport à l'utilisation de l'ETKF avec des observations basse-fréquence. De plus, contrairement aux résultats obtenus (au Chapitre 6) pour le contrôle de la biogéochimie en observant la dynamique, l'augmentation de la fréquence d'observation de la couleur de l'eau améliore la fiabilité et la résolution de l'ensemble produit par le MRHF.

Les résultats obtenus dans cette section, tendent à confirmer que le délicat problème de l'assimilation de couleur de l'eau en haute fréquence doit être traité avec des méthodes d'assimilation pouvant gérer les fortes non-Gaussianités, comme ici, avec le MRHF.

7.5 Conclusions

Dans ce chapitre, nous nous sommes intéressés au problème d'assimilation de la couleur de l'eau. Dans le même cadre que l'étude du chapitre précédent, une étude a été réalisée avec le modèle ECOGeL.

Après la caractérisation du problème par des outils statistiques, des expériences d'assimilation de la couleur de l'eau ont été mises en place. Le contrôle du phytoplancton seul puis le contrôle du phytoplancton et de la température sont examinés. Cette étude nous permet de voir les limites d'un filtre de type moindres carrés ainsi que l'apport de l'assimilation non-Gaussienne face à des problèmes non-Gaussiens. Il ressort de cette étude de nombreuses conclusions que l'on énonce dans cette section.

L'assimilation de la couleur de l'eau est un problème non-linéaire et non-Gaussien. L'étude statistique présentée en Section 7.1 confirme la non-linéarité et la non-Gaussianité du problème de l'assimilation de la couleur de l'eau. Ainsi, les

conditions d'optimalité des solutions de l'assimilation aux moindres carrés ne sont pas réunies. Pour cette raison, l'utilisation de l'assimilation de données non-Gaussienne peut être favorable.

Un filtre Gaussien tel que l'ETKF améliore l'estimation de la partie biogéochimique du système mais génère des états peu réalistes. Effectivement, les erreurs RMS sont réduites et la dispersion d'ensemble est améliorée par l'ETKF. L'évaluation révèle cependant l'apparition progressive d'états peu réalistes. Ces états peu réalistes peuvent faire dégénerer la solution si l'on poursuit l'assimilation sur une plus grande période de temps ou si on augmente le nombre d'observations. De plus, le contrôle du phytoplancton seul n'a aucun impact sur la partie dynamique du système. Il apparait donc nécessaire pour améliorer l'estimation du système global, d'inclure des variables dynamiques dans le vecteur de contrôle.

La dynamique du système peut être en partie contrôlée par assimilation de la couleur de l'eau. L'information contenue dans de la couleur de l'eau permet d'améliorer l'estimation de la dynamique. En mettant en place, le contrôle du phytoplancton et de la température par assimilation de la couleur de l'eau avec l'ETKF, les scores RMSE et les histogrammes de rangs sur la température et la salinité sont améliorés. Cependant, les résultats présentés par ce filtre Gaussien ne semblent pas optimaux.

Les filtres non-Gaussiens tels que l'EnKF anamorphosé et le MRHF améliorent les produits d'assimilation de la couleur de l'eau. Toujours dans le contexte du contrôle du phytoplancton et de la température par assimilation de la couleur de l'eau, nous avons comparé l'ETKF avec l'EnKF anamorphosé et le MRHF, deux filtres non-Gaussiens. Aussi bien sur la partie biogéochimique que dynamique du modèle, les résultats obtenus indiquent le bénéfice apporté par l'utilisation de filtres non-Gaussiens. Pour résoudre ce problème, et selon les scores évalués, l'EnKF anamorphosé est le filtre le plus performant.

Le problème complexe de l'assimilation de la couleur de l'eau haute fréquence rend incontournable l'utilisation de filtres fondamentalement non-Gaussiens. En augmentant la fréquence d'observation de la couleur de l'eau (pour mimer les données obtenues par le satellite géostationnaire Geo-OCAPI), nous avons comparé les performances de l'ETKF, de l'EnKF anamorphosé et du MRHF. Il apparait que l'ETKF et l'EnKF anamorphosé parviennent à réduire leurs erreurs sur la partie dynamique, en revanche, la grande fréquence d'observation dégrade leurs solutions sur la partie biogéochimique. Ceci confirme le caractère non-linéaire et fortement non-Gaussien du problème de la couleur de l'eau. Par ailleurs, la mise en place d'un filtre non-Gaussien permet des réductions d'erreurs RMS de l'ordre de 20% sur la partie biogéochimique. Le MRHF produit également un ensemble plus fiable et mieux résolu. Ces derniers résultats confirment que le problème de l'assimilation de

couleur de l'eau en haute fréquence doit être traité avec des méthodes d'assimilation adaptées aux fortes non-Gaussianités. Ainsi, dans la mesure de nos expériences idéalisées, les résultats affirment que le programme spatiale Geo-OCAPI permettra d'améliorer grandement notre connaissance de l'océan (bleu et vert) seulement si les observations qu'il produit sont assimilées par des méthodes non-Gaussiennes.

Conclusions et perspectives

Le sujet de cette thèse s'inscrit dans le cadre scientifique de la méthodologie de l'assimilation de données en océanographie et en biogéochimie marine.

Dans ces domaines, l'assimilation de données se confronte à une difficulté majeure : comment résoudre des problèmes d'estimation présentant de fortes non-linéarités et de fortes non-Gaussianités. Cette difficulté, qui apparait de plus en plus fréquemment avec la complexification des systèmes d'assimilation en océanographie (complexification des modèles et des réseaux d'observations, apparition de nouveaux types d'observations ...), remet en question l'assimilation de type moindres carrés traditionnellement utilisée dans ces domaines.

Les principaux objectifs de ces travaux de thèse sont de mieux comprendre, de mieux aborder et de mieux traiter cette difficulté au travers de travaux méthodologiques dans un premier temps puis au travers d'applications à des problèmes d'estimation en biogéochimie marine.

Nous rappelons les questionnements moteurs et directeurs de cette thèse, énoncés dans le Chapitre 1 :

- *Comment déceler les non-Gaussianités dans un système ?*
- *Comment envisager une assimilation de données adaptée à un contexte non-Gaussien ?*
- *Dans le cadre du couplage dynamique/ biogéochimie, que peut apporter l'assimilation de données non-Gaussiennes et comment ?*

Nous revenons, ci-dessous, sur les apports importants de ces travaux de thèse. Nous commençons par évoquer les apports personnels que ces travaux, entre théorie et ingénierie, ont permis. Puis, nous détaillons les apports scientifiques, autant sur le plan théorique et méthodologique que sur le plan de l'application à la biogéochimie marine.

Entre la théorie et l'ingénierie

Dans un premier temps, une étude méthodologique a été menée avec des expériences sur modèles "jouets" : le modèle de Lorenz 63, à trois variables (Chap. 3) et le modèle de Lorenz 96, à quarante variables (pas montré ici). Cette étude a notamment abouti au développement d'une nouvelle méthode d'assimilation d'ensemble non-Gaussienne (Chap. 4). La réalisation de cette étude m'a poussé à analyser les

théories fondatrices de bon nombre de méthodes d'assimilation de données. J'ai pu également pour cette étude mettre en place des bancs d'essais simples et efficaces pour mettre en évidence les difficultés des méthodes d'assimilation moindres carrés face à des non-Gaussianités. Enfin, le développement du *Multivariate Rank Histogram Filter* (MRHF) m'a demandé beaucoup de temps passé loin de mon ordinateur avec simplement feuille et crayon pour dériver une théorie correcte et appropriée. Ce type de travail (qui est de plus en plus rare) m'a permis de bien comprendre toutes les difficultés liées à un problème d'estimation.

Dans un second temps, une application à un système couplé de dynamique et de biogéochimie marine a été mise en place. Le modèle numérique couplé étant écrit en FORTRAN 90 et toutes les routines d'assimilation de données étant écrites en python, il a fallu développer une interface entre ces deux niveaux de calculs. Ce travail m'a permis de me former à l'utilisation des classes en python afin de rendre optimales les communications modèle-assimilation. Après avoir écrit toutes les méthodes d'assimilation et tous les diagnostics utilisés dans cette thèse en python, j'ai dû créer un cas-test complet sur la biogéochimie en python également. Bien que ce ne soit que peu mentionné dans ce manuscrit, cette étude comporte donc une forte dimension technique. De plus, l'application à la biogéochimie marine m'a fait découvrir toute l'importance et toute la complexité de la modélisation d'un écosystème marin.

Apports méthodologiques

Pour notre étude, nous nous sommes munis d'une plateforme python comprenant des méthodes moindres carrés (donc aux hypothèses Gaussiennes) : l'EnKF et l'ETKF ; des méthodes adaptant les moindres carrés aux cas non-Gaussiens : l'Anam-EnKF et le RHF ; et une méthode ne faisant aucune hypothèse de Gaussianité : le filtre particulaire PF-*Bootstrap*. Cette plateforme nous permet d'appliquer de manière quasi-automatique ces méthodes sur n'importe quel problème d'assimilation de petites dimensions.

La mise en place des méthodes d'assimilation

Une illustration sur le modèle de Lorenz à trois variables (le Lorenz 63) confirme les caractères plus ou moins non-Gaussiens des méthodes d'assimilation.

Les filtres Gaussiens, l'EnKFs et l'ETKF, produisent de très bons résultats pour un faible nombre de membres d'ensemble dans les cas faiblement et moyennement non-linéaires (et peu non-Gaussiens). Le RHF réduit l'erreur de l'estimé moyen dans le cas fortement non-linéaire. De plus, cette illustration a permis de révéler une difficulté de l'EnKF anamorphosé dynamique (sans queues de densité) à représenter les covariances d'erreurs d'observation dans l'espace anamorphosé lors de la présence

de biais. Le PF-*Bootstrap* produit les meilleurs résultats dans les cas les plus non-linéaires mais sous la condition d'un nombre important de membres d'ensemble.

Cette illustration met également en évidence la nécessité d'une nouvelle méthode pouvant gérer de fortes non-Gaussianités en se basant sur un nombre de membres plus raisonnable que le filtre particulaire. Une telle méthode a été développée et évaluée au Chapitre 4.

Le développement du MRHF

Le quatrième chapitre de cette thèse est un article publié : Metref et al. (2014). Dans cet article, est présenté le Multivariate Rank Histogram Filter (MRHF), un schéma d'analyse entièrement non-Gaussien pour l'assimilation de données d'ensemble. Des expériences numériques sont proposées sur plusieurs configurations, avec différents degrés de non-Gaussianités, du modèle de Lorenz 1963 à trois variables.

À l'aide de scores tels que l'erreur RMS – évaluant la précision de l'estimé moyen – et la divergence de Kullback-Leibler – évaluant la bonne représentation des densités par l'ensemble – il est montré que le MRHF est performant dans des régimes fortement non-Gaussiens (notamment dans un régime bimodal) pour un nombre de membres relativement faible.

Par ailleurs, le coût calcul du MRHF est encore important mais devient raisonnable pour un faible nombre d'observations.

L'application biogéochimique

L'application biogéochimique sur laquelle sont basés les Chapitres 6 et 7 est réalisée avec le modèle couplé de dynamique et de biogéochimie marine, ModECOGeL. Le modèle et les configurations des expériences jumelles d'assimilation sont décrits au Chapitre 5. Le modèle ModECOGeL est un modèle undimensionnel constitué d'une partie dynamique simulant les mouvements verticaux de la couche de mélange et d'une partie biogéochimique simulant l'activité d'un écosystème marin en mer Ligure. Les expériences jumelles sont réalisées sur le mois d'avril 2006. La *vérité* est une simulation effectuée avec un forçage de vent réel en haute fréquence (1 heure). Un ensemble de 50 membres est propagé par le modèle avec un forçage de vent haute fréquence simulé par un processus auto-regressif Gaussien d'ordre un.

L'assimilation de données dynamiques

Un premier jeu d'expériences propose un problème d'assimilation de données dynamiques. Les observations disponibles à l'assimilation sont des profils de température et de salinité tous les deux jours et des données satellites de température de surface

(SST) de fréquence d'observation variant de six heures à trente minutes. Le vecteur de contrôle est constitué de la température et de la salinité, en premier lieu, puis lui sont ajoutés les trois types de phytoplancton.

Une étude *a priori* des statistiques d'ensemble montre que le problème de contrôle de la dynamique seule est un problème quasi-Gaussien et que le problème de contrôle de la biogéochimie présente de nombreuses non-Gaussianités.

Dans le cadre quasi-Gaussien, du contrôle de la partie dynamique du système par des observations de la dynamique, les résultats de l'expérience tendent à montrer que l'utilisation d'un filtre Gaussien, tel que l'ETKF, est pertinente. L'importance des profils pour réduire les biais et par conséquent pour engendrer des ensembles fiables est également mise en évidence.

Dans le cadre non-Gaussien du contrôle de la biogéochimie par des observations de la SST, il ressort qu'un filtre non-Gaussien, le MRHF, est plus adapté que l'ETKF.

En revanche, en augmentant la fréquence d'observation de la SST, il apparait qu'un réseau d'observations fin dégrade plus fortement la dispersion de l'ensemble produit par un filtre non-Gaussien.

L'assimilation de la couleur de l'eau

Un deuxième jeu d'expériences s'attaque au problème délicat de l'assimilation de la couleur de l'eau. La couleur de l'eau artificiellement créée dans ces expériences jumelles correspond à la somme des trois variables phytoplanctoniques en surface. La fréquence d'observation de la couleur de l'eau est initialement établie à une observation tous les trois jours.

Une étude *a priori* des statistiques d'ensemble montre que l'assimilation de la couleur de l'eau est un problème non-linéaire et non-Gaussien.

Une première expérience indique que le contrôle du phytoplancton par l'ETKF améliore globalement l'estimation des variables biogéochimiques (malgré le caractère non-Gaussien du problème) mais génère des états peu réalistes.

Pour le contrôle mixte de la température et du phytoplancton par l'assimilation de la couleur de l'eau, les résultats confirment que les filtres non-Gaussiens, tels l'EnKF anamorphosé et le MRHF, améliorent l'estimation d'ensemble du système.

Lors de l'augmentation de la fréquence d'observation de la couleur de l'eau, un filtre adapté à la non-Gaussianité à partir d'un cadre classique, tel que l'EnKF anamorphosé, n'est plus suffisant pour résoudre ce problème aux très fortes non-Gaussianités. Dans le cadre de nos expériences, le problème complexe de l'assimilation de la couleur de l'eau haute fréquence rend incontournable l'utilisation de filtres fondamentalement non-Gaussiens tel que le MRHF.

Dans la mesure de nos expériences idéalisées, les résultats affirment que le programme spatiale Geo-OCAPI permettra d'améliorer grandement notre estimation

de l'océan (bleu et vert) par assimilation de données seulement si les observations qu'il produit sont assimilées par des méthodes fondamentalement non-Gaussiennes.

Les limites et perspectives de notre étude

Le développement du MRHF

Le *Multivariate Rank Histogram Filter* conduit à des coûts de calculs importants. Comme il est mentionné dans Metref et al. (2014) les temps calculs du MRHF sont de 50 à 200 fois supérieurs à ceux de l'EnKF pour les expériences sur le Lorenz 63. Cependant ces coûts sont très dépendants du nombre d'observations (en temps et en espace). Une utilisation du MRHF pour des problèmes spécifiques est possible. Par exemple, l'assimilation de la couleur de l'eau par le MRHF (Chap. 7) – avec une observation (scalaire) disponible tous les trois jours ou tous les jours – conduit à des temps de calculs seulement cinq à dix fois supérieurs à l'ETKF. Pour mieux quantifier ce coût calcul, il serait toutefois bon de réaliser une étude théorique sur l'efficacité algorithmique du MRHF et de mettre en place des tests destinés à évaluer les temps de calculs. Parallèlement, pour améliorer l'efficacité de la méthode, de nouvelles implémentations du *mean field approximation* sont à l'étude et ont la possibilité de grandement réduire les temps de calculs. Par exemple, une manière d'appliquer la correction sur la variable non-observée peut être imaginée par voisinage et non pas membre par membre. Également, l'adaptation d'une localisation spatiale au MRHF est à l'étude. La localisation permettrait, d'une part, d'étendre les applications possibles du MRHF et, d'autre part, de diminuer les coûts.

L'affinement des expériences

La période de temps de nos expériences est d'un mois, le mois d'avril 2006. Cette période peut s'avérer courte pour évaluer l'assimilation d'une part – nous avons vu que l'ETKF produisait des profils physiquement peu réalistes qui peuvent dégénérer avec le temps – et d'autre part pour faire face à des phénomènes physiques de plus grande variabilité temporelle (saisonnière, annuelle ...). La configuration du modèle ModECOGeL sur les années 2006-2007 existe et a été validée. Une expérience d'assimilation sur ces deux années serait donc une suite intéressante à nos travaux.

La génération d'ensemble est réalisée par une paramétrisation stochastique du forçage de vent. Il ne s'agit que d'une petite partie de l'incertitude du modèle sur la physique qu'il cherche à décrire. Plusieurs paramétrisations stochastiques (sur d'autres forçages, dans les équations d'états ...) peuvent être combinées pour améliorer la représentation des phénomènes physiques. De plus, la combinaison de plusieurs perturbations permettrait d'augmenter la dispersion d'ensemble qui, dans

nos expériences, était parfois trop faible pour l'assimilation (e.g. la dispersion du phytoplancton en surface des dix premiers jours du mois d'avril).

Dans nos expériences, six méthodes d'assimilation ont été comparées. Il serait bien entendu intéressant d'élargir l'étendue de cette comparaison. Les méthodes que nous n'avons pas étudié mais qui nous semblent pertinentes sont, par exemple, le lisseur itératif de Kalman d'ensemble (IEnKS, Bocquet and Sakov, 2013), les filtres particulaires implicites (IPF, Chorin et al., 2010), le filtre particulaire de poids équivalents (EWPF, van Leeuwen, 2010) ou encore plusieurs méthodes hybridant le variationnel au stochastique.

Les diagnostics d'évaluation utilisés sont des scores de performances indépendants du problème considéré. Il serait bon de mettre en place des diagnostics plus en relation à la nature du problème physique, en particulier pour la biogéochimie. Par exemple, il est envisageable d'utiliser comme critère diagnostique l'évaluation de la pente d'approfondissement du *bloom* de phytoplancton.

La mise en place de nouvelles expériences

Les expériences menées dans cette thèse sont toutes des expériences jumelles (i.e. assimilant des observations artificielles issues d'une trajectoire du modèle). L'assimilation de données réelles apporte son lot de nouvelles difficultés techniques mais aussi méthodologiques. C'est pourquoi une étape intéressante, pouvant donner suite à ce travail, est la comparaison de méthodes d'assimilation dans le contexte non-Gaussien de la biogéochimie marine, avec des données réelles. Pour ce faire, le site DYFAMED du modèle ModECOGeL présente l'avantage de disposer d'une bouée, la bouée BOUSSOLE, fournissant des observations réelles de la dynamique et de la biogéochimie.

Enfin, le modèle utilisé pour nos expériences, le modèle ModECOGeL, est un modèle simplifié unidimensionnel. Bien que la dynamique de ModECOGeL soit relativement complexe, nous sommes encore loin des modèles réalistes. Il est possible d'imaginer une étape à plus long terme, prolongeant nos travaux de compréhension méthodologique de l'assimilation en contexte non-Gaussien, à partir du modèle couplé NEMO-PISCES. Cette étape serait notamment facilitée par les travaux de thèse de Florent Garnier.

Perspectives générales

Une meilleure représentation de l'océan doit passer par l'amélioration des méthodes d'assimilation de données.

D'une part, les modèles océaniques sont de plus en plus raffinés. La physique qu'ils décrivent est de mieux en mieux représentée. Et l'augmentation de leurs coûts de calculs nécessitent un travail d'ingénierie considérable. D'autre part, les réseaux d'observations sont déjà denses. Et le coût financier pour augmenter encore le nombre d'observations est très grand. Ces deux grandes approches : la modélisation et l'observation, permettant la compréhension des phénomènes océaniques, ont encore une marge de progression importante. Cependant, le potentiel d'évolution de l'assimilation et son faible coût (relatif) font d'elle le levier nécessaire pour aller au delà des limitations actuelles à l'avancement de l'océanographie. Ainsi que le disait Andrew Lorenc dans une note de bas de page de Lorenc (2003) : "We cannot expect to have sufficient observations to define directly what we need to know, since observations are relatively expensive compared to data assimilation systems, so it is more cost effective to reduce the observations and have a sophisticated assimilation.". La méthodologie en assimilation de données est donc un champ de recherche qui doit être poursuivi et perfectionné.

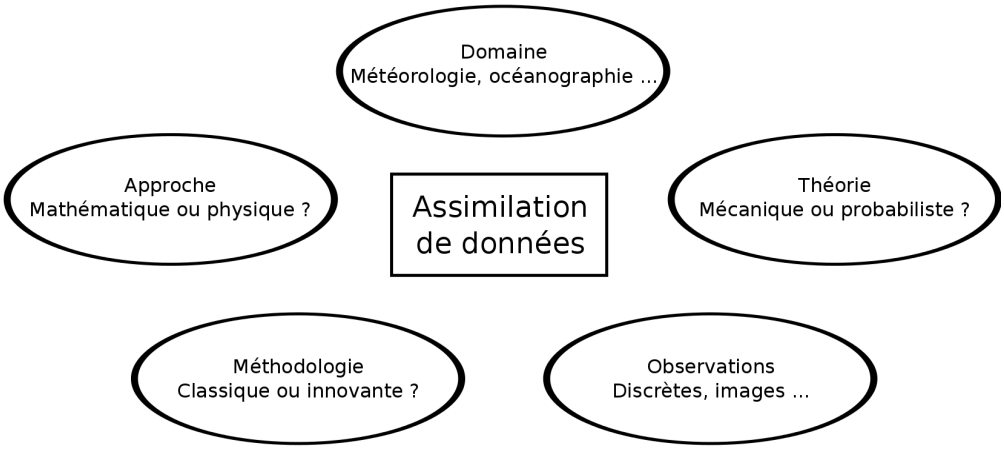

Figure 7.22 – *Les grands axes d'évolution de l'assimilation de données de demain schématisés.*

Dans les années à venir, les grands défis de l'assimilation de données vont s'appuyer sur plusieurs questions cruciales. Le schéma présenté en Figure 7.22 est une représentation non-exhaustive de ces questions qui, selon moi, seront au centre de l'évolution de l'assimilation de données.

De manière pragmatique, le domaine d'application s'imposera avec ses défis ainsi qu'avec les observations dont il dispose. Ce cadre sera clé dans l'élaboration des

nouvelles problématiques de l'assimilation de données. Ces problématiques devront-elles être abordées selon une approche mathématique (telles qu'elles le sont aujourd'hui) ou plus proche de la physique que l'on souhaite décrire ? Faut il choisir une conception déterministe ou probabiliste des événements ? Les méthodologies existantes seront-elles toujours pertinentes ou faudra-t-il poursuivre l'innovation technique et théorique ?

Ces questions commencent déjà à poindre aux travers des premiers grands défis de l'assimilation de données océanographique. Selon moi, quelques-uns de ces grands défis sont :

- l'*estimation de paramètres* dans les modèles GCM (*Global Circulation Models*). L'estimation de paramètres par assimilation de données est un sujet déjà très répandu. Cependant, il me semble important de savoir contrôler un jeu de paramètres dans des modèles réalistes, tout en maintenant la cohérence entre ces paramètres.
- l'*assimilation dans des modèles couplés*. Les travaux présentés dans cette thèse ont en partie abordé ce sujet. De manière plus générale, il est important de penser le couplage au sein de l'étape d'assimilation. En particulier en couplage océan-atmosphère, le défi est de réussir à assimiler des observations de nature et de fréquence différentes dans l'atmosphère et dans l'océan sans déstabiliser l'équilibre du couplage.
- l'*assimilation d'image*. L'assimilation d'image est un des sujets montants de ces vingt dernières années. Que ce soit avec une approche physique, mathématique ou technique d'imagerie, de nombreux travaux de recherche tentent d'adresser ce problème. Pourtant, la notion d'image est encore mal définie avec notamment la question de la pertinence de l'opérateur H, reliant une image à la physique du modèle. Cet écart conceptuel entre l'observation et le modèle est peut-être une des limitations du paradigme de l'assimilation tel qu'on le connait.
- l'*assimilation pour les problèmes non-Gaussiens*. Bien évidemment, ce défi est la motivation première des travaux de thèse qui viennent d'être présentés. Nous ne reviendrons pas ici sur l'importance de ce défi pour l'assimilation de demain mais ce sujet est vaste et encore sous-exploré. Une quantité de travail considérable est donc encore à fournir dans cette direction.

Pour faire évoluer la méthodologie de l'assimilation de données, il y a, à mes yeux, trois démarches possibles à adopter.

Faire évoluer et adapter les approches existantes. Les travaux d'hybridation de l'approche variationnelle et de l'approche ensembliste entrent dans cette démarche. Comme le montre la grande zoologie des méthodes hybrides, ce sujet est en pleine

ébullition et beaucoup de travail reste encore à faire. De même, la notion de localisation (sujet complexe) au sein des méthodes existantes d'assimilation est encore à améliorer. Enfin, pour améliorer les méthodes existantes, il me semble important de mieux représenter l'erreur modèle.

Penser des approches nouvelles. La démarche de s'affranchir des approches classiques est celle que nous avons adopté pour le développement du MRHF. Nous avons également montré qu'il existe un gain conséquent à développer de nouvelles approches pour gérer les non-Gaussianités. Une autre approche qui me semble prometteuse est de ne pas chercher à décrire explicitement la densité de probabilité *a priori* (ce qui est fait par toutes les méthodes discutées jusque là) mais de chercher directement la densité *a posteriori*. Cette approche, peu répandue en assimilation pour les géosciences, est utilisée par les méthodes *Markov Chain Monte Carlo* (MCMC).

Redéfinir le paradigme du problème d'estimation.

Cette dernière démarche est très difficile à anticiper puisque nous essayons de résoudre des problèmes déjà fortement ancrés dans le paradigme actuel. Cependant, on peut imaginer s'en détacher progressivement. Par exemple, l'utilisation de paramétrisations stochastiques au sein des modèles est un premier pas vers des modèles prenant intrinséquement en compte les incertitudes de la physique. En poursuivant cette idée, il est possible de s'imaginer des modèles probabilistes intégrant les observations en leur sein et non comme une information extérieure. Un autre paradigme peut être de considérer l'assimilation de données comme fonciérement physique. C'est à dire une assimilation qui intégre les observations par l'information physique qu'elles apportent et non par la représentation quantitative que l'on s'en fait. La transition vers ce dernier paradigme peut commencer en n'assimilant plus seulement des variables d'états mais d'autres quantités comme par exemple l'énergie spectrale du système.

Le travail restant à accomplir est important. Il me semble néanmoins que dans ces trois démarches réside la réponse aux besoins nés, naissants, ou à naitre de l'assimilation de données.

Bibliographie

Anderson, J. L. (1996). A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *Journal of Climate*, 9 :1518–1530.

Anderson, J. L. (2003). A local least squares framework for ensemble filtering. *Monthly Weather Review*, 131 :634–642.

Anderson, J. L. (2010). A non-Gaussian ensemble filter update for data assimilation. *Monthly Weather Review*, 138 :4186–4198.

Anderson, J. L. (2012). Localization and sampling error correction in ensemble Kalman filter data assimilation. *Monthly Weather Review*, 140(7) :2359–2371.

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

Anderson, T. W. and Darling, D. A. (1952). Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *Ann. Math. Stat.*, 23 :193–212.

Aurell, E., Boffetta, G., Crisanti, A., Paladin, G., and Vulpiani, A. (1997). Predictability in the large : an extension of the concept of Lyapunov exponent. *Journal of Physics A : Mathematical and General*, 30(1) :1.

Auroux, D. and Blum, J. (2005). Back and forth nudging algorithm for data assimilation problems. *C.R. Acad. Sci. Paris*, I :873–878.

Béal, D., Brasseur, P., Brankart, J.-M., Ourmières, Y., and Verron, J. (2009). Controllability of mixing errors in a coupled physical biogeochemical model of the North Atlantic : A nonlinear study using anamorphosis. *Ocean Sci. Discuss.*, 6 :1289–1332.

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 Sci.*, 6 :247–262.

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

Berliner, L. M. and Wikle, C. K. (2007). Approximate importance sampling Monte Carlo for data assimilation. *Physica D*, 230 :37–49.

Beron-Vera, F., Olascoaga, M., and Goni, G. (2008). Oceanic mesoscale eddies as revealed by Lagrangian coherent structures. *Geophysical Research Letters*, 35(12).

Berre, L., Varella, H., and Desroziers, G. (2015). Modelling of flow-dependent ensemble-based background-error correlations using a wavelet formulation in 4D-Var at Météo-France. *Quarterly Journal of the Royal Meteorological Society*.

Berry, P. and Marshall, J. (1989). Ocean modelling studies in support of altimetry. *Dyn. Atmos. Oceans*, 13(3-4) :269–300.

Bertino, L., Evensen, G., and Wackernagel, H. (2003). Sequential data assimilation techniques in oceanography. *Internat. Stat. Rev.*, 71 :223–241.

Bishop, C. H., Etherton, B. J., and Majumdar, S. J. (2000). Adaptive sampling with the ensemble transform Kalman filter. Part I : Theoretical aspects. *Monthly Weather Review*, 129(3) :420–436.

Blayo, E., Verron, J., and Molines, J.-M. (1994). Assimilation of Topex/Poseidon altimeter data into a circulation model of North Atlantic. *Quarterly Journal of the Royal Meteorological Society*, 24 :691–705.

Bocquet, M. (2011). Ensemble Kalman filtering without the intrinsic need for inflation. *Nonlin. Processes Geophys.*, 18 :735–750.

Bocquet, M., Pires, C. A., and Wu, L. (2010). Beyond Gaussian statistical modeling in geophysical data assimilation. *Monthly Weather Review*, 138 :2997–3023.

Bocquet, M. and Sakov, P. (2012). Combining inflation-free and iterative ensemble Kalman filters for strongly nonlinear systems. *Nonlin. Processes Geophys.*, 19 :383–399.

Bocquet, M. and Sakov, P. (2013). An iterative ensemble Kalman smoother. *Quarterly Journal of the Royal Meteorological Society*, 00 :2–24.

Brankart, J. M. (2009). Square root or ensemble observational update with SESAM. *LEGI/MEOM Technical report, Grenoble*, page 29pp.

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

Brankart, J. M. (2014). Traitement des incertitudes en océanographie - HDR.

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.

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.

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

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

Bröcker, J. (2011). Estimating reliability and resolution of probability forecasts through decomposition of the empirical score. *Climate Dynamics*, 39 :655–667.

Buehner, M., Houtekamer, P. L., Charette, C., Mitchell, H. L., and He, B. (22010a). Intercomparison of variational data assimilation and the ensemble Kalman filter for global deterministic NWP. Part I : Description and single-observation experiments. *Monthly Weather Review*, 138 :1550–1566.

Buehner, M., Houtekamer, P. L., Charette, C., Mitchell, H. L., and He, B. (22010b). Intercomparison of variational data assimilation and the ensemble Kalman filter for global deterministic NWP. Part II : One-month experiments with real observations. *Monthly Weather Review*, 138 :1567–1586.

Burgers, G., van Leeuwen, P., and G., E. (1998). Analysis scheme in the ensemble Kalman filter. *Am.Meteorol.Soc.*, 126 :1719–1724.

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

Capone, D. G. and Hutchins, D. A. (2013). Microbial biogeochemistry of coastal upwelling regimes in a changing ocean. *Nature Geoscience*, 6 :711–717.

Charney, J. G., Halem, M., and Jastrow, R. (1969). Use of incomplete historical data to infer the present state of the atmosphere. *Journal of Atmospheric Science*, 26 :1160–1163.

Charrois, L., Dumont, M., and Cosme, E. (submitted). Unknown. *Unknown*.

Chorin, A., Morzfeld, M., and Tu, X. (2010). Implicit particle filters for data assimilation. *Comm. App. Math. and Comp. Sci.*, 5(2) :221–240.

Clayton, A. M., Lorenc, A., and Barker, D. M. (2013). Operational implementation of a hybrid ensemble/4D-Var global data assimilation system at the Met Office. *Quarterly Journal of the Royal Meteorological Society*, 139 :1445–1461.

Cohn, S. E. (1997). An introduction to estimation theory. *J. Meteor. Soc. Japan*, 75(1B) :257–288.

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.

Cotter, C. J. and Reich, S. (2013). Ensemble filter techniques for intermittent data assimilation, in large scale inverse problems. *Radon Ser. Comput. Appl. Math.*, 13 :91–134.

Courtier, P. and Talagrand, O. (1987). Variational assimilation of meteorological observations with the adjoint vorticity equation. II : Numerical results. *Quarterly Journal of the Royal Meteorological Society*, 113 :1329–1347.

D'Agostino, R. B. and Pearson, E. S. (1973). Testing for departures from normality. *Biometrika*, 60 :613–622.

Dee, D. and Da Silva, A. M. (2003). The choice of variable for atmospheric moisture analysis. *Monthly Weather Review*, 131 :155–171.

DeMey, P. and Robinson, A. R. (1987). Assimilation of altimeter eddy fields in a limited-area quasi-geostrophic model. *Journal of Physical Oceanography*, 17 :2280–2293.

Desroziers, G., Camino, J.-T., and Berre, L. (2014). 4DEnVar : link with 4D state formulation of variational assimilation and different possible implementations. *Quarterly Journal of the Royal Meteorological Society*, 140 :2097–2110.

Doron, M., Brasseur, P., and Brankart, J.-M. (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.

Doucet, D., de Freitas, N., and Gordon, N. (2001). An introduction to sequential Monte Carlo methods. In Doucet, D., de Freitas, N., and Gordon, N., editors, *Sequential Monte Carlo Methods in Practice*, Statistics for Engineering and Information Science. Springer-Verlag.

Dutkiewicz, S., Follows, M., Marshall, J., and Gregg, W. W. (2001). Inter-annual variability of phytoplankton abundances in the North Atlantic. *Deep Sea Res., Part II*, 48 :2323–2344.

El Moselhy, T. A. and Marzouk, Y. M. (2012). Bayesian inference with optimal maps. *J. Comput. Phys.*, 231 :7815–7850.

Epstein, E. S. (1962). A Bayesian approach to decision making in applied meteorology. *J. Appl. Meteorol.*, 1 :169–177.

Evensen, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research*, 99 :10143–10162.

Evensen, G. (2003). The ensemble Kalman filter : Theoretical formulation and practical implementation. *Ocean Dyn.*, 53 :343–367.

Evensen, G. and Leeuwen, P. J. V. (1996). Assimilation of Geostat altimeter data for the Agulhas current using the ensemble Kalman filter. *Monthly Weather Review*, 124 :85–96.

Fournier, A., Hulot, G., Jault, D., Kuang, W., Tangborn, A., Gillet, N., Canet, E., Aubert, J., and Lhuillier, F. (2010). An introduction to data assimilation and predictability in geomagnetism. *Space Science Reviews*, 155(1-4) :247–291.

Frederiksen, J., Kane, T. O. ., and Zidikheri, M. (2012). Stochastic subgrid parameterizations for atmospheric and oceanic flows. *Physica Scripta*, 85.

Garnier, F., Brasseur, P., Brankart, J.-M., and Cosme, E. (2015.). Stochastic parameterizations of biogeochemical uncertainties in a 1/4° NEMO/PISCES model for probabilistic comparisons with ocean color data. *Journal of Marine Systems*.

Ghil, M. (1989). Meteorological data assimilation for oceanographers. Part I : Description and theoretical framework. *Dynamics of Atmosphere and Oceans*, 13 :171–218.

Gogonel, A., Collet, J., and Bar-Hen, A. (2014). Implementation of two statistical methods for ensemble prediction systems in the management of electrical systems. *CS-BIGS*, 5 (2) :74–87.

Gordon, N. J., Salmond, D. J., and Smith, A. F. M. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEEE Proc. F.*, 140 :107–113.

Gregg, W. W., Friedrichs, M. A., Robinson, A. R., Rose, K. A., Schlitzer, R., Thompson, K. R., and Doney, S. C. (2009). Skill assessment in ocean biological data assimilation. *Journal of Marine Systems*, 76 :16–33.

Gregorio, S., Brasseur, P., Brankart, J.-M., Metref, S., and Doron, M. (in prep.). Estimation of parameters describing forcing uncertainties in a 1D biogeochemical model of the Ligurian Sea : A twin experiment approach.

Greybush, S. J., Kalnay, E., Miyoshi, T., Ide, K., and Hunt, B. R. (2011). Balance and ensemble Kalman filter localization techniques. *Monthly Weather Review*, 139 :511–522.

Hamill, T. M. (2001). Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, 129 :550–560.

Hamill, T. M. and Snyder, C. (2002). Using improved background-error covariances from an ensemble Kalman filter for adaptive observations. *Monthly Weather Review*, 130 :1552–1572.

Hamill, T. M., Whitaker, J. S., and Snyder, C. (2001). Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Monthly Weather Review*, 129 :2776–2790.

Harlim, J. and Hunt, B. (2007). A non-Gaussian ensemble filter for assimilating infrequent noisy observations. *Tellus A*, 59 :225–237.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15 :559–570.

Holland, W. R. (1989). *Combining data and dynamics*, chapter Altimeter-data assimilation into ocean circulation models–some preliminary results in oceanic circulation models, pages 203–230. D. L. T. Anderson and J. Willebrand, Amsterdam, kluwer academic publ. edition.

Holland, W. R. and Malanotte-Rizzoli, P. (1989). Along-track assimilation of altimeter data into an ocean circulation model : Space versus time resolution studies. *Journal of Physical Oceanography*, 19 :1507–1534.

Hoteit, I., Luo, X., and Pham, D.-T. (2012). Particle Kalman filtering : A nonlinear Bayesian framework for ensemble Kalman filters. *Monthly Weather Review*, 140(2) :528–542.

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

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

Houtekamer, P. L. and Mitchell, H. L. (2001). A sequential ensemble Kalman filter for atmospheric data assimilation. *Monthly Weather Review*, 129 :123–137.

Hurlburt, H. E. (1986). Dynamic transfer of simulated altimeter data into subsurface information by a numerical ocean model. *Journal of Geophysical Research*, 91 :2372–2400.

Ivanoff, A. (1977). *Modelling and Prediction of the Upper Layers of the Ocean*, chapter Oceanic absorption of solar energy, pages 47–71. Kraus, E.B. (Ed.), Pergamon, School of Marine and Atmospheric Science, University of Miami, FL.

Jazwinski, A. H. (1970). *Stochastic processes and filtering theory*. Academic Press, New York.

Jolliffe, I. T. and Stephenson, D. B. (2012). *Forecast Verification. A practitioners guide in atmospheric science*, volume 2nd ed. Wiley-Blackwell.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Trans. Am. Soc. Mech. Mech. Eng. J. Basic Eng.*, 82(D) :35–45.

Kalnay, E., Li, H., Miyoshi, T., Yang, S.-C., and Ballabrera, J. (2007). 4D-Var or ensemble Kalman filter. *Tellus*, 59A :758–773.

Kleeman, R. (2002). Measuring dynamical prediction utility using relative entropy. *Journal of Atmospheric Science*, 59 :2057–2072.

Kullback, S. (1959). *Information Theory and Statistics*. Wiley, New York.

Lacroix, G. (1998). *Simulation de l'écosystème pélagique de la mer Ligure à l'aide d'un modèle unidimensionnel. Étude du bilan de matière et de la variabilité saisonnière, interannuelle et spatiale*. PhD thesis, Université de Liège, Belgique.

Lacroix, G. and Grégoire, M. (2002). Revisited ecosystem model (MODECOGeL) of the Ligurian Sea : seasonal and interannual variability due to atmospheric forcing. *Journal of Marine Systems*, 37, 4 :229–258.

Lacroix, G. and Nival, P. (1998). Influence of meteorological variability on primary production dynamics in the Ligurian Sea (NW Mediterranean Sea) with a 1D hydrodynamic/biological model. *J. Mar. Syst.*, 16 (1-2) :23–50.

Lauvernet, C., Brankart, J.-M., Castruccio, F., Broquet, G., Brasseur, P., and 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.

Lawson, W. G. and Hansen, J. A. (2004). Implications of stochastic and deterministic filters as ensemble-based data assimilation methods in varying regimes of error growth. *Monthly Weather Review*, 132(8) :1966–1981.

Le Dimet, F. X. and Talagrand, O. (1986). Variational algorithms for analysis and assimilation of meteorological observations : Theoretical aspects. *Tellus*, 38A :97–110.

Lei, J. and Bickel, P. (2011). A moment matching ensemble filter for nonlinear non-Gaussian data assimilation. *Monthly Weather Review*, 139 :3964–3973.

Lei, J., Bickel, P., and Snyder, C. (2011). Comparison of ensemble Kalman filters under non-Gaussianity. *Monthly Weather Review*, 138(4) :1293–1306.

Lei, L., Stauffer, D., Haupt, S. E., and Young, G. (2012). A hybrid nudging-ensemble Kalman filter approach to data assimilation. Part I : Application in the Lorenz system. *Tellus A*, 64.

Lenartz, F., Raick, C., Soetaert, K., and Grégoire, M. (2007). Application of an ensemble Kalman filter to a 1-D coupled hydrodynamic-ecosystem model of the Ligurian Sea. *68*, pages 327–348.

Lermusiaux, P. F. J. (2006). Uncertainty estimation and prediction for interdisciplinary ocean dynamics. *J. Comp. Phys.*, 217 :176–199.

Lermusiaux, P. F. J. and Robinson, A. R. (1999a). Data assimilation via error subspace statistical estimation. Part I : Theory and schemes. *Monthly Weather Review*, 127 :1408–1432.

Lermusiaux, P. F. J. and Robinson, A. R. (1999b). Data assimilation via error subspace statistical estimation. Part II : Middle Atlantic Bight shelfbreak front simulations and ESSE validation. *Monthly Weather Review*, 127 :1385–1407.

Lévy, M., Memery, L., and Andre, J.-M. (1998). Simulation of primary production and export fluxes in the Northwestern Mediterranean Sea. *Journal of Marine Research*, 56 :197–238.

Lilliefors, H. W. (1967). On the KolmogorovSmirnov test for normality with mean and variance unknown. *J. Amer. Stat. Assoc.*, 62 :399–402.

Liu, C., Xiao, Q., and Wang, B. (2008). An ensemble-based four-dimensional variational data assimilation scheme. Part I : Technical formulation and preliminary test. *Monthly Weather Review*, 136 :3363–3373.

Liu, C., Xiao, Q., and Wang, B. (2009). An ensemble-based four-dimensional variational data assimilation scheme. Part II : Observing system simulation experiments with advanced research WRF (ARW). *Monthly Weather Review*, 137 :1687–1704.

Lorenc, A. (1986). Analysis methods for numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society*, 112 :1177–1194.

Lorenc, A., Bowler, N. E., Clayton, A. M., Fairbairn, D., and Pring, S. R. (2015). Comparison of hybrid-4DEnVar and hybrid-4DVar data assimilation methods for global NWP. *Monthly Weather Review*, 143 :212–229.

Lorenc, A. C. (2003). The potential of the ensemble Kalman fi filter for NWPa comparison with 4D-Var. *QJR Meteorol. Soc*, 129 :3183–3203.

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of Atmospheric Science*, 20 :130–141.

Magri, S., Brasseur, P., and Lacroix, G. (2005). Data assimilation in a marine ecosystem model of the Ligurian Sea. *C.R. Geoscience*, 337 :1065–1074.

Malanotte-Rizzoli, P. and Holland, W. R. (1986). Data constraints applied to models of the ocean general circulation. Part I, The steady case. *Journal of Physical Oceanography*, 16 :1665–1687.

Malanotte-Rizzoli, P. and Holland, W. R. (1988). Data constraints applied to models of the ocean general circulation. Part II, The transient, eddy resolving case. *Journal of Physical Oceanography*, 18 :1093–1107.

Malanotte-Rizzoli, P., Young, R. E., and Haidvogel, D. B. (1989). Initialization and data assimilation experiments with a primitive equation model. *Dyn. Atmos. Oceans*, 13(3-4) :349–378.

Metref, S., Cosme, E., Snyder, C., and Brasseur, P. (2014). A non-Gaussian analysis scheme using rank histograms for ensemble data assimilation. *Nonlin. Processes Geophys.*, 21 :869–885.

Miller, R. N. (1989). Direct assimilation of altimetric differences using the Kalman filter. *Dyn. Atmos. Oceans*, 13(3-4) :317–334.

Miller, R. N. (1990). Tropical data assimilation experiments with simulated data : The impact of tropical ocean and global atmosphere thermal array for the ocean. *Journal of Geophysical Research*, 95 :11461–11483.

Miller, R. N., Carter, E. F., and Blue, S. T. (1999). Data assimilation into nonlinear stochastic models. *Tellus*, 51A :167–194.

Mooers, C. M. R., Robinson, A. R., and Thompson, J. D. (1987). Ocean Prediction Workshop 1986 : A status and prospect report on the scientific basis and the Navy's needs. *Institute for Naval Oceanography, NSTL, Mississippi.*

Moore, A. M. and Anderson, D. L. T. (1989). The assimilation of XBT data into a layer model of the tropical Pacific Ocean. *Dyn. Atmos. Oceans*, 13(3-4) :441–464.

Morzfeld, M. and Chorin, A. (2012). Implicit particle filtering for models with partial noise, and an application to geomagnetic data assimilation. *Nonlin. Processes Geophys.*, 19 :365–382.

Morzfeld, M., Tu, X., Atkinsa, E., and Chorin, A. (2012). A random map implementation of implicit filters. *J. Comp. Phys.*, 231 :2049–2066.

Murphy, A. H. (1973). A new vector partition of the probability score. 12(4) :595–600.

Nakano, S., Ueno, G., and Higuchi, T. (2007). Merging particule filter for sequential data assimilation. *Nonlin. Processes Geophys.*, 14 :395–408.

Nejadi, S., Leung, J., and Trivedi, J. (2014). Merging particule filter for sequential data assimilation. *Mathematical geosciences*, 47(2) :193–225.

Nihoul, J. C. J. and Djenidi, S. (1987). *Three-Dimensional Models of Marine and Estuarine Dynamics*, chapter Perspective in three-dimensional modelling of the marine system, pages 1–34. Nihoul, J. C. J., Jamart, B. (Eds.), Amsterdam, elsevier edition.

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

Pearson, K. (1895). Notes on regression and inheritance in the case of two parents. *Notes on regreProceedings of the Royal Society of London*, 58 :240–242.

Pelc, J. S., Simon, E., Bertino, L., Serafy, G. E., and Heemink, A. W. (2012). Application of model reduced 4D-Var to a 1D ecosystem model. *Ocean Modelling*, 57-58 :43–58.

Perruche, C., Rivière, P., P. Pondaven, and Carton, X. (2010). Phytoplankton competition and coexistence : Intrinsic ecosystem dynamics and impact of vertical mixing. *Journal of Marine Systems*, (81) :99–111.

Pham, D. T. (1996). A singular evolutive interpolated Kalman filter for data assimilation in oceanography. *Technical Report 163, Project IDOPT CNRS-INRIA*.

Pham, D. T. (2001). Stochastic methods for sequential data assimilation in strongly nonlinear systems. *Monthly Weather Review*, 129 :1194–1207.

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

Pires, C. A., Talagrand, O., and Bocquet, M. (2010). Diagnosis and impacts of non-Gaussianity of innovations in data assimilation. *Physica D*, (239) :1701–1717.

Raick, C., Alvera-Azcarate, A., Barth, A., Brankart, J.-M., Soetaert, 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. *J. Mar. Syst.*, 65 :561–583.

Reich, S. (2013). A nonparametric ensemble transform method for Bayesian inference. *SIAM J.Sci. Comput.*, 35 :A2013–A2024.

Riley, G. A. (1956). Oceanography of Long Island Sound. *Physical Oceanography*, Bull. Bingham Oceanogr. Collect.(15) :15–46.

Robert, C., Blayo, E., and Verron, J. (2006). Comparison of reduced-order, sequential and variational data assimilation methods in the tropical Pacific Ocean. *Ocean Dynamics*, 56(5-6) :624–633.

Robinson, A. R. (1987). *Three dimensional ocean models of marine and estuarine dynamics*, chapter Predicting open ocean currents, fronts and eddies, pages 89–112. J. C. F. Nihoul and B. M. Jamart, Amsterdam, elsevier. edition.

Robinson, A. R., Carton, J. A., Pinardi, N., and Mooers, C. M. R. (1986). Dynamical forecasting and dynamical interpolation : An experiment in the California current. *Journal of Physical Oceanography*, 16 :1561–1579.

Robinson, A. R. and Lermusiaux, P. F. J. (2002). *The Sea : BiologicalPhysical Interactions in the Sea*, volume vol. 12, chapter Data assimilation for modeling and predicting coupled physicalbiological interactions in the sea, pages 475–536. Wiley, New York.

Robinson, A. R. and Leslie, W. B. (1985). Estimation and prediction of oceanic eddy fields. *Prog. Oceanogr.*, 14 :485–510.

Robinson, A. R., Spall, M. A., Leslie, W. G., Walstad, L. J., and McGillicuddy, D. J. (1987). Gulfcasting : Dynamical forecast experiments for Gulf Stream rings and meanders, November 1985–June 1986. *Harvard Uni. Rep. in Meteor. and Oceanogr., No.22, Harvard University, Cambridge, Massachusetts*.

Robinson, A. R., Spall, M. A., and Pinardi, N. (1988). Gulf Stream simulations and the dynamics of ring and meander processes. *Journal of Physical Oceanography*, 18 :1320–1353.

Robinson, A. R., Spall, M. A., Walstad, L. J., and Leslie, W. G. (1989). Data assimilation and dynamical interpolation in gulfcast experiments. *Dyn. Atmos. Oceans*, 13(3-4) :269–300.

Robinson, A. R. and Walstad, L. J. (1987). The Harvard open ocean model : Calibration and applications to dynamical process forecasting and data assimilation studies. *J. Appl. Numer. Math.*, 3 :89–121.

Robinson, I. (2004). *Measuring the Oceans from Space : The principiles and methods of satellite oceanography*. Springer-Verlag Berlin and Heidelberg GmbH & Co. KG.

Ruelle, D. and Takens, F. (1971). On the nature of turbulence. *Commun. math. phys*, 20(3) :167–192.

Sakov, P. and Bertino, L. (2010). Relation between two common localisation methods for the EnKF. *Computational Geosciences*, 15(2) :225–237.

Sakov, P., Counillon, F., Bertino, L., Lisaeter, K. A., Oke, P. R., and Korablev, A. (2012). TOPAZ4 : An ocean-sea ice data assimilation system for the North Atlantic and Arctic. *Ocean Science*, 8(4) :633–656.

Sakov, P. and Oke, P. R. (2008). A deterministic formulation of the ensemble Kalman filter : An alternative to ensemble square root filters. *Tellus A*, 60(2) :361–371.

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52 :591–611.

Simon, E. and Bertino, L. (2009). Application of the Gaussian anamorphosis to assimilation in a 3-D coupled physical-ecosystem model of the North Atlantic with the EnKF : A twin experiment. *Ocean Science*, 5 :495–510.

Simon, E. and Bertino, L. (2012). Gaussian anamorphosis extension of the DEnKF for combined state parameter estimation : Application to a 1D ocean ecosystem model. *J. Mar. Syst.*, 89 :1–18.

Snyder, C. (2012). Particle filters, the optimal proposal and high-dimensional systems. *Proc. Seminar on Data Assimilation for Atmosphere and Ocean, ECMWF, Reading, Berkshire*, pages 161–170.

Snyder, C., Bengtsson, T., Bickel, P., and Anderson, J. (2008). Obstacles to high-dimensional particle filtering. *Monthly Weather Review*, 136 :4629–4640.

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

Talagrand, O. and Courtier, P. (1987). Variational assimilation of meteorological observations with the adjoint vorticity equation. I : Theory. *Quarterly Journal of the Royal Meteorological Society*, 113 :1311–1328.

Tarantola, A. (1987). *Inverse Problem Theory*. Elsevier, 613 pp.

Tarantola, A. (2005). *Inverse problem theory and methods for model parameter estimation*. SIAM, Philadelphia, USA.

Thompson, J. D. (1986). Altimeter data and geoid error in mesoscale ocean prediction : some results from a primitive equation model. *Journal of Geophysical Research*, 91 :2401–2417.

Thompson, P. D. (1961). A dynamical method for analysing meteorological data. *Tellus*, 13 :334–349.

Tissot, J.-Y. (2012). *Sur la décomposition ANOVA et l'estimation des indices de Sobol. Application à un modèle d'écosystème marin.* PhD thesis, Ecole doctorale Mathématiques, Sciences et Technologies de l'Information, Informatique.

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

Turiel, A., Solé, J., Nieves, V., Ballabrera-Poy, J., and García-Ladona, E. (2008). Tracking oceanic currents by singularity analysis of microwave sea surface temperature images. *Remote Sensing of Environment*, 112(5) :2246–2260.

van Leeuwen, P. J. (2003). A variance-minimizing filter for large-scale applications. *Monthly Weather Review*, 131 :2071–2084.

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

van Leeuwen, P. J. (2010). Nonlinear data assimilation in geosciences : an extremely efficient particle filter. *Quarterly Journal of the Royal Meteorological Society*, 136 :1991–1999.

Verlaan, M. and Heemink, A. W. (2000). Nonlinearity in data assimilation applications : A practical method for analysis. *Mon. Wea. Rev.*, 129 :1578–1589.

Verron, J. (1992). Nudging satellite data into quasi-geostrophic ocean models. *Journal of Geophisical Research*, 97 :7479–7491.

Verron, J. and Holland, W. R. (1989). Impacts de données d'altimétrie satellitaire sur les simulations numériques des circulations océaniques aux latitudes moyennes. *Annales Geophysicae*, 71 :31–46.

Villani, C. (2009). *Optimal transportation : Old and new.* Springer-Verlag, Berlin Heidelberg.

Wackernagel, H. (2006). *Multivariate Geostatistics*, volume 3rd ed. Springer.

Whitaker, J. S. and Hamill, T. M. (2002). Ensemble data assimilation without perturbed observations. *Monthly Weather Review*, pages 1913–1924.

Whitaker, J. S., Hamill, T. M., Wei, X., Song, Y., and Toth, Z. (2008). Ensemble data assimilation with the NCEP Global Forecast System. *Monthly Weather Review*, 136(2) :463–482.

Whitley, D. (1994). A genetic algorithm tutorial. *Statist. Comput.*, 4 :65–85.

Wikle, C. K. and Berliner, L. M. (2007). A Bayesian tutorial for data assimilation. *Physica D : Nonlin. Phenom.*, 230 :1–16.