Analyse des méthodes basées sur le contenu en information et étude de sensibilité

Sommaire

5.1 Introduction

Au cœur d'un système d'analyse et de prévision océaniques se trouve un modèle mathématique décrivant l'océan, basé sur des équations complexes tenant compte des phénomènes physiques et de leur évolution dans le temps. Afin que la simulation numérique de l'océan soit le plus réaliste possible, les modèles numériques océaniques doivent recevoir des données indispensables dont la plupart sont issues de sources satellitaires et in-situ (température, salinité, niveau de surface et vitesses de courants...). Ces mesures sont souvent de qualité variable, et disposées de manière irrégulière dans l'espace et/ou le temps. La représentation spatio-temporelle des processus océaniques faite par le modèle est aussi imparfaite. Le modèle souffre de plusieurs sources d'erreur notamment celle liée aux équations du modèle (discrétisation, paramétrisations des équations de la physique), aux forcages et aux conditions aux limites. Il est donc crucial de traiter chacune des informations provenant du modèle ou des observations au sein d'un système d'assimilation en tenant compte de l'erreur dont l'information est entachée. Pour cela l'assimilation de données reposant sur un processus mathématique permet d'obtenir la représentation la plus probable de l'état d'un système à partir de toutes les sources d'informations disponibles et des contraintes connues (*Bouttier* [2004]). En effet l'analyse résultant d'un système d'assimilation est influencée à la fois par les observations récemment assimilées et par l'information portée par l'ébauche. Il existe différentes approches pour mesurer l'impact des observations assimilées dont l'une d'elles repose sur le contenu en information, noté par la suite DFS pour Degrees of Freedom of the System. Des méthodes d'estimation du DFS ont déjà été développé pour des applications liées à l'atmosphère (*Chapnik et al.* [2006], *Cardinali et al.* [2004]) puis plus récemment lors d'applications océanographiques (*Oke et al.* [2009], *Dibarboure et al.* [2011]). Le DFS permet d'évaluer la contribution relative des observations dans n'importe quel système d'assimilation de données. Les méthodes d'estimation du contenu en information dans un système d'analyse et de prévisions réalistes s'articulent autour de l'expression de matrices de grandes dimensions. Les calculs matriciels peuvent s'avérer longs et coûteux. Pour éviter l'explicitation directe de matrices si denses, *Girard* [1987], *Desroziers and Ivanov* [2001] puis *Chapnik et al.* [2006] ont proposé de nouveaux schémas d'implémentation afin d'estimer la trace de ces matrices lors des calculs pratiques du DFS. Enfin *Lupu and Gauthier* [2010] décrit une méthode pratique afin d'estimer le DFS de manière plus directe. Nous allons comparer ces différentes approches d'estimation du contenu en information afin de déterminer laquelle conviendrait le mieux à une implémentation dans le système d'assimilation de Mercator Océan. Ainsi dans ce chapitre nous présentons la définition analytique du DFS et ses formulations dans le cadre du BLUE et de l'analyse objective en section 5.2, puis en sections 5.3 et 5.4 par le biais d'applications nous mènerons une étude de sensibilité sur le DFS,

et la mise en pratique de plusieurs méthodes alternatives (définies en section 5.2.4) afin d'évaluer approximativement le DFS dans un cadre d'analyse objective.

5.2 Mesure d'impact des observations sur l'analyse

5.2.1 Définition analytique

On considère un schéma d'assimilation de donnée, soit x_a le vecteur d'analyse optimale résultant, alors on définit le DFS comme étant la trace de la jacobienne du vecteur d'analyse dans l'espace des observations par rapport aux observations :

$$\mathsf{DFS} = \mathsf{Tr}\{\frac{\partial(\mathbf{H}x_a)}{\partial \mathbf{y}}\} \tag{5.1}$$

Où $\mathbf{H}$ est l'opérateur d'observations linéarisé au voisinage de l'ébauche, et $\mathbf{y}$ le vecteur d'observations. Le DFS ou degrès de liberté du signal, permet de quantifier l'information utile contenue dans les observations en ce sens qu'il caractérise comment le système d'assimilation utilise les observations pour construire le signal à partir de l'ébauche (*Rabier* [2005]).

5.2.2 Définition dans le cadre du BLUE

Rappellons brièvement l'approche du BLUE qui permet d'obtenir le meilleur estimateur non biaisé du problème linéaire de l'assimilation. En pratique, il s'agit de trouver la meilleure combinaison entre un état à priori du système que l'on appelle l'ébauche, et les observations (*Gelb* [1974], *Bouttier and Courtier* [1999]). Le critère d'optimalité pour obtenir le meilleur estimateur statistique de cette combinaison est la détermination du minimum de variance d'erreur d'estimation. Les informations issues de l'ébauche et des observations sont entachées d'erreur. La combinaison des deux par le biais de divers algorithmes d'assimilation (BLUE, filtre de Kalman notamment) permet de se rapprocher de l'état réel du système. Ainsi, l'algorithme d'assimilation identifie à un instant t donné, un état analysé $\mathbf{x}_a$ de la variable aléatoire $\mathbf{x}$ dépendant linéairement de l'ébauche $\mathbf{x}_b$ et des observations $\mathbf{y}^o$ (les étapes de calculs sont détaillées en annexe C). On écrit l'état analysé $\mathbf{x}_a$ tel que :

$$\mathbf{x}_a = \mathbf{x}_b + \mathbf{K}\left(\mathbf{y}^o - \mathbf{H}\mathbf{x}_b\right) = \mathbf{x}_b + \mathbf{B}\mathbf{H}^T(\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})^{-1}\left(\mathbf{y}^o - \mathbf{H}\mathbf{x}_b\right) \tag{5.2}$$

La matrice $\mathbf{K}$, appelée matrice de gain, permet de définir l'état assimilé comme la somme de l'ébauche et de l'incrément d'analyse (terme correctif)

$$\delta \mathbf{x}_a = \mathbf{K}\left(\mathbf{y}^o - \mathbf{H}\mathbf{x}_b\right) = \mathbf{K}d \tag{5.3}$$

avec d le vecteur d'innovation qui représente la mesure résiduelle dans l'espace des observations.

Ainsi la définition analytique du DFS s'obtient en dérivant (5.2) dans l'espace des observations, et (5.1) devient :

$$\frac{\partial(\mathbf{H}x_a)}{\partial \mathbf{y}} = \frac{\partial\left(\mathbf{H}x_b + \mathbf{H}\mathbf{K}\left(\mathbf{y}^o - \mathbf{H}x_b\right)\right)}{\partial \mathbf{y}} \tag{5.4}$$

$$\mathsf{DFS} = \mathsf{Tr}\{(\mathbf{HK})\} \tag{5.5}$$

L'expression (5.5) relie le DFS à la trace de la matrice de gain (Kalman).

5.2.3 Definition dans le cadre de l'analyse objective

Dans le cadre simple d'une analyse objective (*Bretherton et al.* [1976]; *Gandin* [1963]), le vecteur d'ébauche peut ne pas être utilisé et être remplacé par les valeurs d'observations issues de climatologies. On cherche le meilleur estimateur linéaire au sens des moindres carrés non biaisé de $\mathbf{x}_a$, noté $\widehat{x_a}$. Le nouvel état estimé est basé seulement sur les observations de telle sorte que

$$\widehat{x_a} = \sum \alpha_{ij}\mathbf{y} \tag{5.6}$$

D'après le théorème de Gauss-Markov on obtient :

$$\widehat{x_a} = \mathbf{CdgCdd^{-1}y}$$

Enfin on applique la définition analytique du DFS dans le cadre d'une analyse objective.

$$\mathsf{DFS} = \mathsf{Tr}\{\mathbf{HCdgCdd^{-1}}\} \tag{5.7}$$

On note $\mathbf{Cdg}$ [1] la matrice d'erreurs de covariance champs-observations, $\mathbf{Cdd}$ [2] la matrice d'erreurs de covariance au sein des observations, $\mathbf{H}$ est l'opérateur d'observations.

1. data-grid
2. data-data

5.2.4 Méthodes alternatives pour l'approximation du DFS

L'information contenue dans les différents types de jeux d'observations ne possède pas le même impact sur l'analyse. Dans la littérature, *Cardinali et al.* [2004] propose une méthode permettant de distinguer l'apport respectif des différentes observations sur l'analyse. Cette approche permet également de mesurer l'influence partielle d'un sous-ensemble d'observations restreint à une zone géographique donnée. De plus *Chapnik et al.* [2006] aborde une approche similaire basée sur l'estimation précise de la trace de la matrice d'influence [3]. Cette approche évoque l'ajout de perturbations aléatoires sur les observations. L'objectif est, connaissant l'incertitude sur les observations, de mesurer la sensibilité de l'analyse par rapport aux observations perturbées (méthode de randomisation de *Girard* [1987]). Enfin une approche réalisée plus récemment repose sur l'utilisation des vecteurs d'incréments et d'innovations (notés respectivement $\mathbf{d}_b^o$ et $\mathbf{d}_b^a$). A partir de ces quantités il est possible de diagnostiquer les erreurs d'observations, d'ébauche et d'analyse à posteriori, et d'obtenir une expression du DFS (*Desroziers et al.* [2005]; *Lupu and Gauthier* [2010]). En se basant sur la formulation analytique du DFS (notre référence) nous allons confronter deux des approches citées ci-dessus : la méthode reposant sur la randomisation de Girard développée dans *Girard* [1987] et *Chapnik et al.* [2006], et la méthode dite aposteriori abordée dans *Lupu and Gauthier* [2010]. L'expression analytique du DFS ainsi que les deux approches d'estimation comparées sont détaillés ci-dessous et mises en pratique dans le cadre d'une analyse objective à la section 5.4.

Approximation du DFS selon la méthode de perturbations de Girard

En utilisant l'approche introduite dans *Desroziers and Chapnik* [2001], puis utilisée dans *Chapnik et al.* [2006], on montre alors qu'une approximation du DFS peut s'exprimer comme suit :

$$\mathsf{DFS}_{\mathsf{Girard}} = (\mathbf{y}^* - \mathbf{y})^t \mathbf{R}^{-1} (\mathbf{H}x_a^* - \mathbf{H}x_a) \tag{5.8}$$

Les observations sont obtenues en ajoutant des petites perturbations telles que

$$\mathbf{y}^* = \mathbf{y} + \mathbf{R}^{1/2}\varsigma \tag{5.9}$$

où $\varsigma \sim \mathsf{N}(0, I_p)$. Enfin $\mathbf{H}x_a^*$, $\mathbf{H}x_a$ sont les analyses dans l'espace des observations obtenues respectivement à partir des observations perturbées et non perturbées.

formulation dans le cadre de l'analyse objective De la même manière on se place dans le cadre de l'analyse objective, alors l'approximation du DFS s'écrit comme suit,

3. ou aussi matrice de gain définie à la section 5.2

$$\mathsf{DFS}_{\mathsf{Girard}} = (\mathbf{y}^* - \mathbf{y})^t \mathbf{R}^{-1} (\mathbf{H}\widehat{x_a}^* - \mathbf{H}\widehat{x_a}) \tag{5.10}$$

Approximation du DFS selon la méthode dite aposteriori

Un ensemble de diagnostics dans l'espace des observations basé sur les vecteurs d'innovation, d'incrément a été proposé par *Desroziers et al.* [2005]. A partir de ces quantités, on montre alors qu'il est possible de diagnostiquer les erreurs statistiques d'observations, d'ébauche, et d'analyse a posteriori. Par conséquent, on note $\mathbf{B}$ et $\mathbf{R}$ les matrices de covariances d'erreurs diagnostiquées respectivement d'ébauche et d'observations telles que :

$$\mathbf{E}[\mathbf{d}_b^a \mathbf{d}_b^o] = \mathbf{H}\mathbf{B}\mathbf{H}^t \tag{5.11}$$

$$\mathbf{E}[\mathbf{d}_b^o \mathbf{d}_b^o] = \mathbf{R} \tag{5.12}$$

De plus *Lupu and Gauthier* [2010] souligne aussi qu'une expression du DFS peut être obtenue via les vecteurs :

$$\mathbf{d}_b^a = \mathbf{H}(x_a) - \mathbf{H}(x_b) \approx \mathbf{H}\delta x_a = \mathbf{H}\mathbf{K}\mathbf{d}_o^b,$$
$$\mathbf{d}_a^o = \mathbf{y} - \mathbf{H}(x_b + \delta x_a) \approx (\mathbf{I} - \mathbf{H}\mathbf{K})\mathbf{d}_o^b = \mathbf{R}\mathbf{D}^{-1}\mathbf{d}_o^b$$

Où $\mathbf{K}$ est la matrice de gain de Kalman, et $\mathbf{D} = \mathbf{H}\mathbf{B}\mathbf{H}^t + \mathbf{R}$ est la matrice de covariance d'innovation. De la démonstration faite par *Lupu and Gauthier* [2010], il vient

$$\mathbf{E}[\mathbf{d}_b^{at} \mathbf{R}^{-1} \mathbf{d}_a^o] = \mathbf{E}[\mathsf{Tr}\{\mathbf{D}^{-1}\mathbf{D}^e\mathbf{K}^t\mathbf{H}^t\}] \tag{5.13}$$

Il se présente alors deux cas de figures :

– La matrice de covariance d'innovation $\mathbf{D}$ prescrite dans l'assimilation coïncide avec celle estimée $\mathbf{D}^e$, dans ce cas le DFS s'écrit :

$$\mathsf{DFS}_{\mathsf{Apost}} = \mathsf{Tr}(\mathbf{H}\mathbf{K}) = \mathbf{E}[(\mathbf{d}_b^a)^t \mathbf{R}^{-1} \mathbf{d}_a^o] \tag{5.14}$$

L'expression (5.14) donne un moyen efficace d'estimer le DFS pour un schéma d'assimilation où seuls les sous-produits de l'assimilation sont connus.

– Les matrices connues à priori $\mathbf{D}$ et à posteriori $\mathbf{D}^e$ ne sont pas cohérentes. Par conséquent ces deux termes matriciels dans (5.13) ne se compensent plus. *Desroziers et al.* [2005] propose alors de considérer ces matrices de covariance estimées comme des matrices de covariance ajustées. Les équations (5.11) et (5.12) peuvent être écrites au sens de matrices de covariances ajustées, et notées $\mathbf{B}^e$, $\mathbf{R}^e$ comme suit :

$$\mathbf{E}[\mathbf{d}_b^a \mathbf{d}_b^o] = \mathbf{H}\mathbf{B}^e\mathbf{H}^t = \mathbf{H}\mathbf{B}\mathbf{H}^t(\mathbf{D}^{-1}\mathbf{D}^e) \tag{5.15}$$

$$\mathbf{E}[\mathbf{d}_b^o \mathbf{d}_b^o] = \mathbf{R}^e = \mathbf{R}(\mathbf{D}^{-1}\mathbf{D}^e) \tag{5.16}$$

Lupu and Gauthier [2010] montre ensuite que la définition du $\mathsf{DFS}_{\mathsf{Apost}}$ peut être ré-écrite avec la matrice de covariance d'erreurs d'observations $\mathbf{R}^e$

$$\mathsf{DFS}_{\mathsf{Apost}} = \mathsf{Tr}(\mathbf{H}\mathbf{K}^e) = \mathbf{E}[(\mathbf{d}_b^a)^t(\mathbf{R}^e)^{-1}\mathbf{d}_a^o] \tag{5.17}$$

formulation dans le cadre de l'analyse objective Exprimons enfin le cas de l'analyse objective

$$\mathsf{DFS}_{\mathsf{Apost}} = \mathbf{E}[\widehat{\mathbf{d}_b^{at}}\mathbf{R}^{-1}\widehat{\mathbf{d}_a^o}] = \mathbf{E}[\mathbf{H}\widehat{x_a}\mathbf{R}^{-1}(\mathbf{y} - \mathbf{H}\widehat{x_a})] \tag{5.18}$$

On précise les notations du vecteur dit de résidu et l'image du vecteur d'incrément d'analyse dans l'espace des observations en (5.19)

$$\widehat{\mathbf{d}_a^o} = \mathbf{y} - \mathbf{H}\widehat{x_a} \tag{5.19}$$

$$\widehat{\mathbf{d}_b^a} = \mathbf{H}\widehat{x_a} - \mathbf{H}\widehat{x_b}, \quad ici \quad \mathbf{H}\widehat{x_b} = 0 \tag{5.20}$$

Enfin, supposer la matrice de covariance d'erreurs d'observations diagonale implique que l'erreur entre deux observations distinctes ne soit pas corrélées. Pour des types d'observations comme les instruments in-situ, cette hypothèse est raisonnable.

5.3 Etude de sensibilité du contenu en information

Le contenu en information est très utile pour évaluer la contribution relative des observations assimilées dans un schéma d'assmilation de données. On peut alors se demander comment évolue le contenu en information si les observations sont bruitées, et/ou corrélées. Le but de cette étude est d'illustrer les facteurs influençant le DFS. Pour cela, nous utilisons un schéma d'analyse objective permettant de reconstituer un champ observé à partir de mesures issues d'un champ $\mathcal{Z}$ dont on a une connaissance statistique à priori (moyenne et covariance). Via la définition analytique du contenu en information décrite en section 5.2 nous disposons des valeurs du DFS global associé au champ reconstruit. Nous allons souligner dans la section 5.3.2, l'influence des paramètres tels que l'échelle de corrélation de la matrice de covariance du champ et la variance du bruit sur le DFS d'une part, et sur les champs vrai et estimé d'autre part.

5.3.1 Simulation de distributions gaussiennes et estimateur de Gauss-Markov

On se place sur une grille de points régulière variant de 30° à 35° en longitude et 40° à 45° en latitude. Le pas spatial est de 15km [4]. On génère un champ 2D suivant la loi de probabilité gaussienne telle que $\mathcal{Z} \sim \mathcal{N}(0, \mathbf{B})$. Soit donc B la matrice de covariance du signal, symétrique semi-définie positive de type

$$B(i,j) = \exp(\frac{-d(i,j)^2}{L^2}). \tag{5.21}$$

où **d** est la distance entre deux points de grille i et j en km, et **L** représente l'échelle de corrélation (en km aussi). Elle joue le rôle du diamètre de la bulle d'influence dans laquelle se trouvent les observations. L'analyse objective effectue une moyenne pondérée dans cette bulle d'influence afin de cibler un certain nombre d'observations. La détermination des cœfficients de pondération est liée à l'erreur des observations (ici synthétiques) et la distance au point que l'on cherche à estimer. Plus le point observé est éloigné du point à estimer, plus le coefficient poids est faible. De la même manière, le poids est d'autant plus faible que l'observation est entachée d'erreurs. Enfin l'estimation du champ 2D connaissant un nombre fini d'observations est possible grâce au théorème de Gauss-Markov. On suppose que les observations sont des mesures du champ à estimer. On cherche le meilleur estimateur linéaire au sens des moindres carrés non biaisé de $\mathcal{Z}$ (on donnera l'-expression matricielle de l'estimateur) . Le théorème de Gauss-Markov assure que parmi, tous les estimateurs linéaires non biaisés d'une combinaison linéaire des observations, l'estimateur par moindres carrés présente une variance minimale. Le réseau d'observations est régulier, et chaque observation est espacée de 80km. On obtient l'ensemble des mesures via les valeurs du champ aux points d'observations, et celles entachées d'erreurs notée $\mathbf{Y}_{obs}$ via un bruit aléatoire gaussien de moyenne nulle et de variance Br :

$$\mathbf{Y}_{obs} = \mathcal{Z}(i,j) + \varsigma_{i,j} \tag{5.22}$$

où $\varsigma_{i,j} \sim \mathcal{N}(0, Br)$ et $\mathcal{Z}$ est le champ à estimer.

En effet, soit $\theta_{est}(i,j)$ un estimateur linéaire de $\mathcal{Z}$ au point (i,j),

$$\theta_{est}(i,j) = \sum \alpha_{ij} \mathbf{Y}_{obs} \tag{5.23}$$

tel que :

- $\mathbf{E}(\varsigma_{ij}) = 0$
- $cov(\varsigma_{ii}, \varsigma_{jj}) = Br$
- $var(\varsigma_{ii}) < \infty$

4. La grille possède $25 * 25$ nœuds, sans niveau vertical.

On cherche la meilleure combinaison linéaire des observations $\mathbf{Y}_{obs}$ telle que

$$\mathbf{E}^2 =< [\theta - \theta_{est}]^2 > \tag{5.24}$$

soit minimale.
Alors sous une formulation matricielle on a

$$\theta_{est} = \mathbf{CdgCdd^{-1}Y}_{obs} \tag{5.25a}$$

$$\mathbf{Err} = diag(1 - \mathbf{CdgCdd^{-1}Cdg}^t \tag{5.25b}$$

On note encore **Cdg** la matrice d'erreurs de covariance champs-observations et $\mathbf{Cdd^{-1}}$ la matrice d'erreurs de covariance au sein des observations (t est l'opérateur de transposition).

5.3.2 Tests d'influence sur le DFS

La première expérience a eu pour but de vérifier l'influence de la longueur de corrélation sur les valeurs du DFS global et à chaque point d'observation notés DFS_{obs}, la seconde celle de la variance du bruit de mesure. Le DFS global est obtenu via $\mathsf{Tr}\{\frac{\partial(\mathbf{H}x_a)}{\partial \mathbf{y}}\}$ où x_a est équivalent, ici à θ_{est} (détaillée dans la section 5.2). Le DFS global représente pour une valeur d'échelle de correlation **L** donnée, la somme des dfs_{obs}. En d'autres termes, les éléments diagonaux de la matrice $\mathbf{HCdgCdd^{-1}}$ fournissent l'apport d'information en chaque point d'observation. On rappelle aussi que le DFS est non normalisé par le nombre total d'observations $\mathsf{N_{obs}}$ et que le pas spatial entre deux observations est de 80 km. Par exemple, dans le tableau 5.1 on a :

- pour $\mathbf{L} = 40$ km , la valeur totale sur l'ensemble des 16 observations à cette échelle de corrélation est donnée par le $\mathbf{DFS} = 11.42$
- de la même manière, $\mathbf{L} = 160$ km, $\mathbf{DFS} = 7.56$

De plus, les quantitées notées $dfs_{obs_{min}}$ et $dfs_{obs_{max}}$, représentent à un point d'observation donné, l'apport d'information minimal respectivement maximal.

Variation du DFS en fonction de la longueur de corrélation pour un bruit de mesure fixé

Les variations du DFS en fonction de la longueur de corrélation sont représentées sur la figure 5.1 pour $\mathbf{L}_{min} = 40km$, à $\mathbf{L}_{max} = 240km$ ($\Delta\mathbf{L} = 20$ km). Le DFS décroit lorsque la longueur de corrélation augmente (5.1), et le champ reconstruit devient moins sensible à chacune des observations. En effet lorsque la longueur de corrélation augmente, le nombre d'observations exploitées lors la reconstruction fait de même (et pourrait fournir des observations redondantes). Observons alors les écarts entre les différents dfs_{obs} au point d'observation via le tableau 5.1.

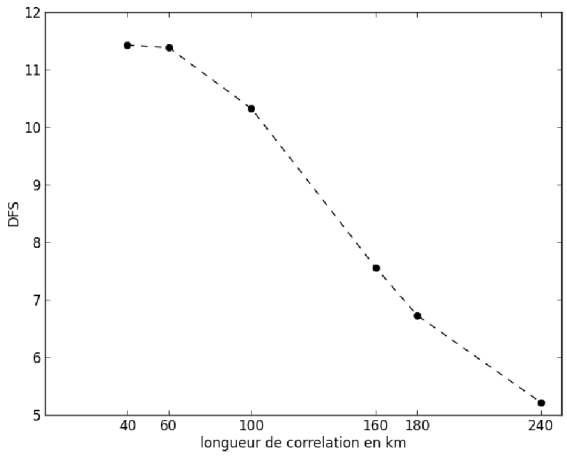


Figure 5.1 – *Variation du DFS en fonction de la longueur de corrélation pour un bruit de mesure de 40% pour les* 16 *observations*

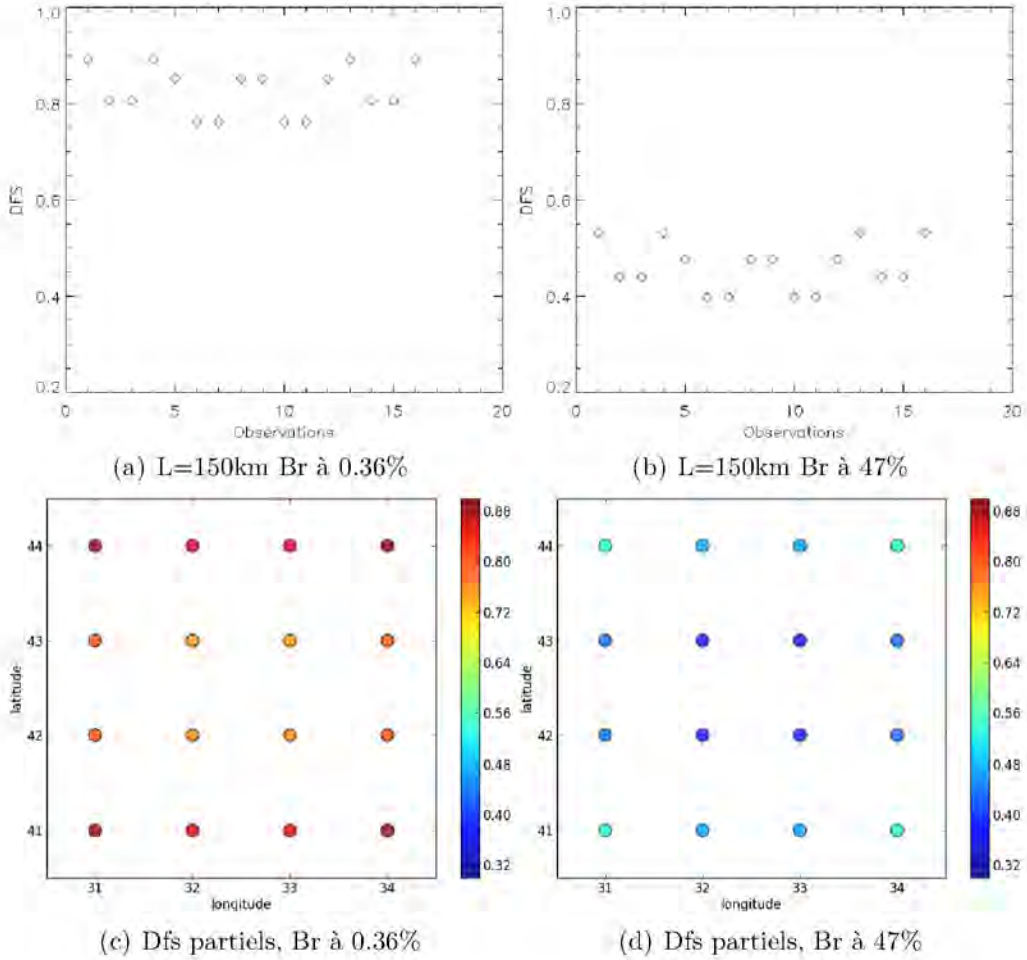
.

L(km)	DFS	$DFS_{obs_{min}}$	$DFS_{obs_{max}}$
40	11.42	0.71	0.71
60	11.38	0.71	0.72
100	10,32	0.62	0.66
160	7.56	0.39	0.54
180	6.73	0.33	0.51
240	5.21	0.23	0.43

Table 5.1 – *Variation du DFS en fonction de la longueur de la corrélation minimale* 40 *km, maximale* 240 *km, pour un bruit de mesure de 40% pour les* 16 *observations*

Comme on pourrait s'y attendre la bulle d'influence délimitée par la longueur de corrélation grandit et a pour conséquence de modifier le poids de chaque observation présente. Selon la localisation du point d'observations sur la grille, lorsque la longueur de corrélation est supérieure au pas spatial d'observations (ici $dx_{obs} = 80$ km) il est alors possible de tenir compte de 3 à 8 observations voisines (si le point est au bord ou au centre). Cela entraine une différence de symétrie entre les dfs_{obs}. Pour exemple les différentes valeurs des dfs_{obs} pour 16 observations peuvent êtres observées sur les figures 5.2. Les figures 5.2 a et b soulignent 4 groupes de 4 observations comportant des valeurs de dfs_{obs} identiques. Celles-ci dépendent bien de la localisation spatiale des observations, les figures 5.2 c et d permettent de représenter la décroissance du dfs_{obs} lorsque les observations sont au cœur de la grille.

En effet les valeurs des points intérieurs sont relativement plus petites que celles des points dits extérieurs, conséquence de l'apport d'information des observations environnantes.



(a) L=150km Br à 0.36%

(b) L=150km Br à 47%

(c) Dfs partiels, Br à 0.36%

(d) Dfs partiels, Br à 47%

Figure 5.2 – *Dfs partiels pour chaque observation pour une longueur de corrélation commune $L = 150$ km et un bruit à 0.36% (gauche) et à 47% (droite).*

Variation du DFS en fonction du bruit de mesure pour une longueur de corrélation fixée

Dans un deuxième temps, nous observons l'influence du bruit de mesure sur le DFS. On a choisi deux configurations spatiales $L_1 < dx_{obs}$ et $L_2 > dx_{obs}$. Comme le montre l'ensemble des résultats présents dans les tableaux 5.2 et 5.3, lorsque le bruit de mesure augmente, le DFS diminue. En effet les observations sont moins précises et par conséquent apportent moins à la reconstruction. Notons aussi que

la sensibilité au bruit de mesure du DFS est différente selon les deux configurations :

Lorsque $L_1 < dx_{obs}$, les dfs_{obs} sont identiques, l'information aux points d'observations est utilisée de manière homogène et ce indépendamment de la valeur de bruit de mesure.

Par ailleurs, comme on peut le voir sur les figures 5.2, lorsque $L_2 \geq dx_{obs}$ les dfs_{obs} ne sont pas uniformes sur la grille et diminuent lorsque l'observation se rapproche du centre et inversement. Notons que l'interdépendance entre les dfs_{obs} en fonction des positions des observations est indépendante du bruit de mesure. En effet les figures 5.2 c et d) soulignent la décroissance des dfs_{obs} lorsque le bruit de mesure augmente, mais aussi la conservation de la même répartition du contenu en information selon la position des observations sur la grille. Les figures 5.2 a et b) le confirment.

Enfin à partir de la figure 5.3, on observe que la décroissance du DFS est moindre lorsque la longueur de corrélation est plus petite. Toujours lié à la définition du DFS, l'information utilisée est plus importante lorsque les observations sont moins corrélées.

expérience	**Br**	**DFS**	**dfs min obs**	**dfs max obs**
1	6%	15.96	0.99	0.99
2	40%	13.96	0.87	0.87
3	70%	10.74	0.67	0.67
4	100%	7.87	0.49	0.49

Table 5.2 – *Influence du bruit de mesure pour* $L_1 = 60km$

expérience	**Br**	**DFS**	**dfs min obs**	**dfs max obs**
1	6%	15.8	0.98	0.99
2	40%	12.14	0.71	0.8
3	70%	8.74	0.5	0.58
4	100%	6.39	0.37	0.43

Table 5.3 – *Influence du bruit de mesure pour* $L_2 = 120km$

Comparaison entre le champ vrai et estimé, et erreurs théoriques

Les figures 5.4 représentent les champs vrais, estimés pour différents bruits de mesure et longueurs de corrélation. Les champs estimés pour le cas $L_1 < dx_{obs}$, quelque soit le bruit de mesure imposé restent en deçà des champs vrais (figures

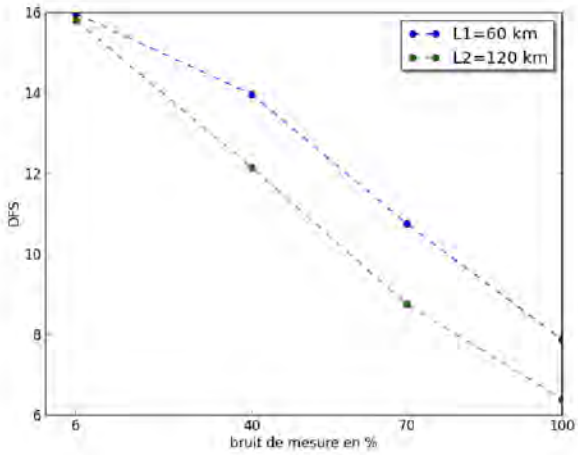
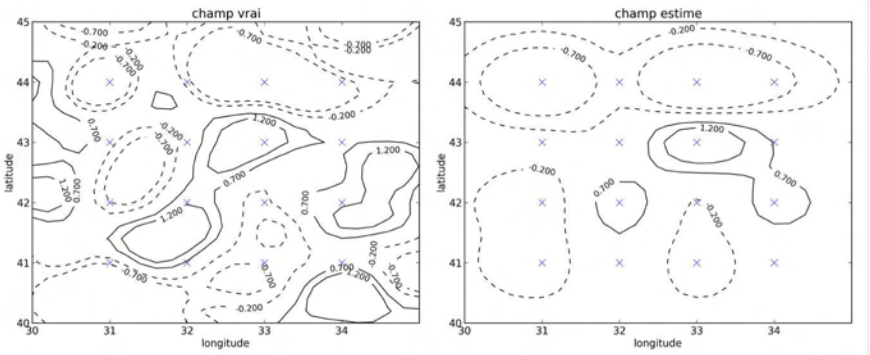


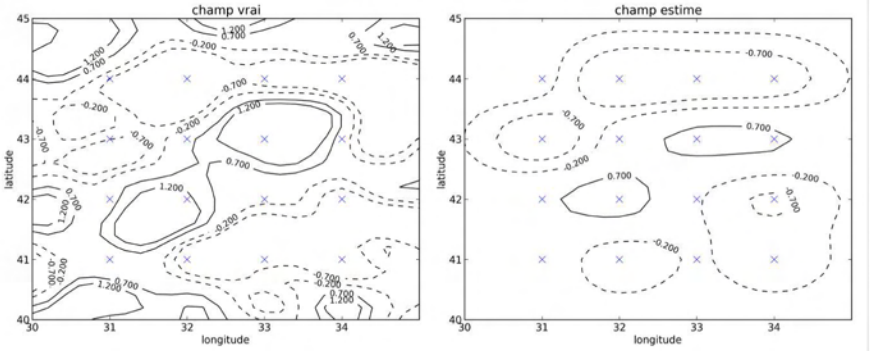
Figure 5.3 – *Variation du DFS en fonction des différentes valeurs de bruits testés (expériences 1 à 4).*

5.4 a et b). Ce constat semble être en accord avec les résultats précédents basés sur le contenu en information : la reconstruction nécessite de davantage d'information que si $L_2 > dx_{obs}$, et celle-ci sera d'autant plus dégradée si les observations sont peu corrélées et entachées d'erreurs. Aussi pour le cas $L_2 > dx_{obs}$, les champs estimés sont moins dégradés à minima de bruit de mesure (figures 5.4 c et d).

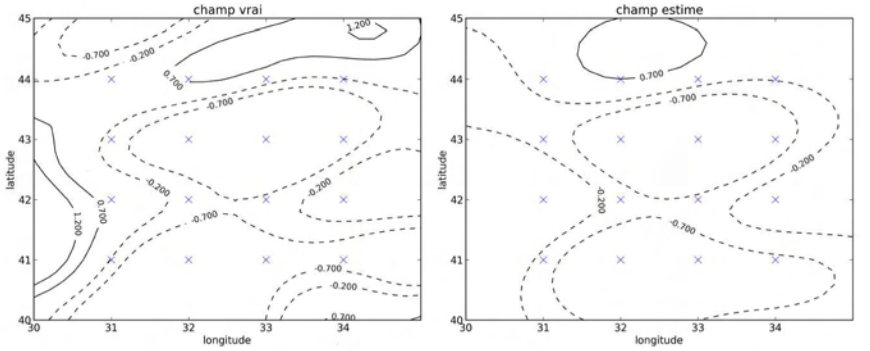
Les figures 5.5 soulignent l'influence du bruit de mesure sur les erreurs théoriques : la zone d'incertitude grandit avec le bruit de mesure. Ces quatre dernières figures reflètent aussi l'impact de la longueur de corrélation sur la zone d'incertitude et donc sur les erreurs théoriques. Lorsque la longueur de corrélation augmente, ces dernières diminuent. La reconstruction nécessite de moins d'information lorsque les observations sont corrélées et engendrerait moins d'erreurs. Ce résultat est cohérent avec la décroissance observée du DFS lorsque L augmente.



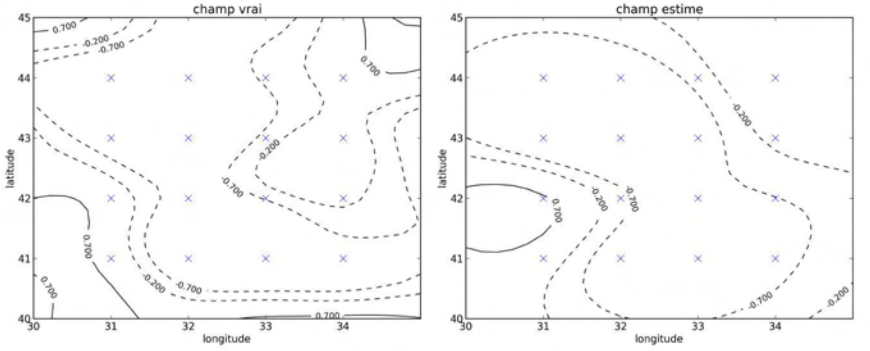
(a) $L_1 = 60$ km et Br à 10%, DFS = 15.25



(b) $L_1 = 60$ km et Br à 50%, DFS = 12.72

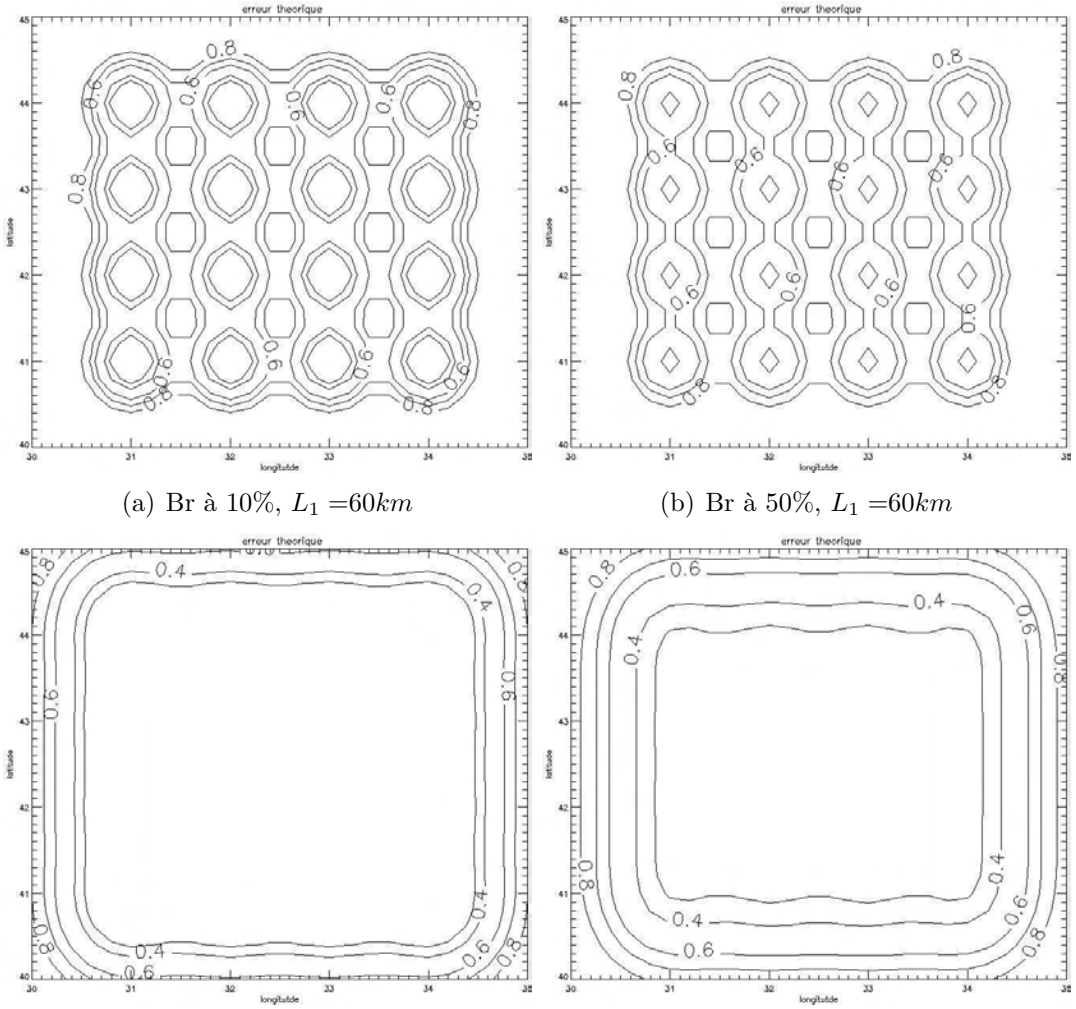


(c) $L_2 = 120$ km et Br à 10%, DFS = 15.40



(d) $L_2 = 120$ km et Br à 50%, DFS = 11.66

Figure 5.4 – *Influence du bruit sur les champs vrai et estimé pour $L_1 < dx_{obs}$ et $L_2 > dx_{obs}$. Les contours en pointillés sont pour des valeurs négatives. Les observations (croix bleue) sont espacées régulièrement de $dx_{obs} = 80$ km.*



(a) Br à 10%, L_1 =60km (b) Br à 50%, L_1 =60km

(c) Br à 10%, L_2 =120km (d) Br à 50%, L_2 =120km

Figure 5.5 – *Influence du bruit de mesure sur les erreurs théoriques, pour* $L_1 < dx_{obs}$ *et* $L_2 > dx_{obs}$*. Les contours non nommés sont à* 0.3 *ou* 0.7*. Les observations sont espacées régulièrement de* $dx_{obs} = 80$ *km.*

5.4 Comparaison des estimations du contenu en information

5.4.1 Illustration à partir d'une application au système d'assimilation de Mercator Océan

Le DFS est difficilement estimable de manière exacte dans les systèmes de grande dimension et avec inversion locale tels que ceux opérés par Mercator Océan. C'est pourquoi des méthodes alternatives sont recherchées. Deux approches présentées et définies en section 5.2 ont été testées dans le système global au 1/4° : celle utilisant les différences aux points d'observation entre deux analyses obtenues avec et sans bruitage des observations et celle proposée par Lupu et al [2010]. Alors que la première méthode requiert la réalisation de deux expériences et la perturbation des observations suivant les variances d'erreurs prescrites sur les observations, la seconde méthode est plus attractive car elle ne requiert aucune simulation supplémentaire. Elle est donc beaucoup plus facile à mettre en œuvre dans le cadre d'un système opérationnel : les variables nécessaires à son calcul sont accessibles "offline". Cependant lorsque les deux approches ont été testées pour estimer le DFS de la température de surface (SST) dans une analyse du système global au 1/4°, les résultats sont apparus très différents. La figure (5.6) montre les DFS moyens par bin de 5° x 5° divisés par le nombre d'observations estimés par la méthode dite ici de Girard sur une semaine de août 2010. Alors que les résultats obtenus via la première méthode semblent pouvoir être expliqués physiquement, via la profondeur de la couche de mélange, cela n'est pas le cas pour l'estimation proposée par Lupu (carte non montrée). L'absence d'une estimation exacte du DFS nous empêche de valider ces résultats, et a motivé les tests réalisés dans ce chapitre et présentés en section 5.4.2.

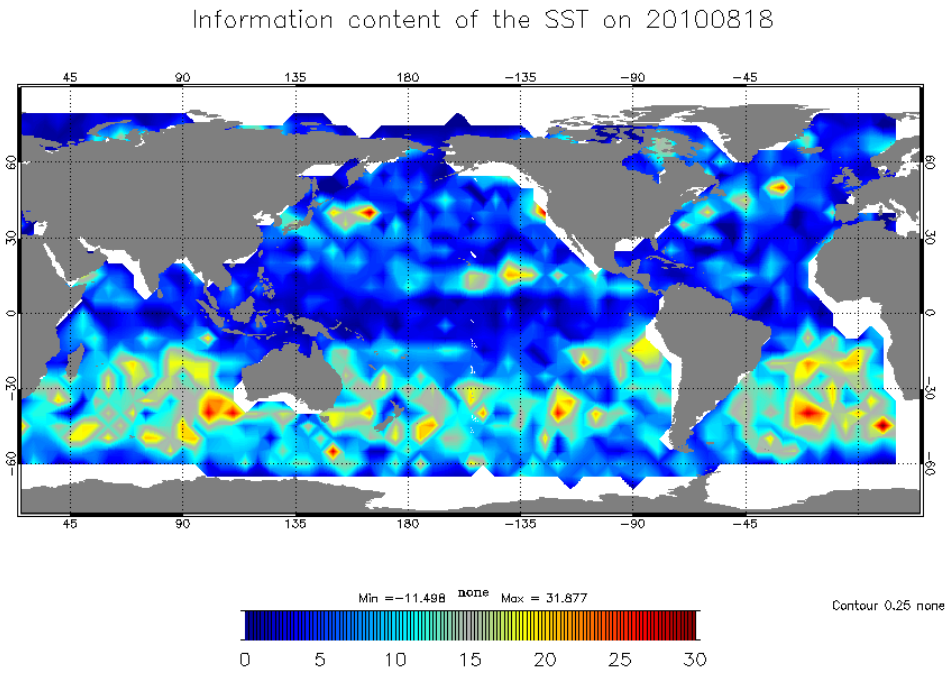


Figure 5.6 – *Contenu en information des observations de température de surface (SST) en pourcentage pour une analyse du système global au* $1/4°$ *en août* 2010*, figure réalisée par E. Remy.*

5.4.2 Mise en pratique dans le cadre d'analyse objective

Ainsi dans cette section on cherche à confronter différentes méthodes de calculs du contenu en information : celle issue de la méthode de randomisation de Girard, celle obtenue via la matrice de covariance d'erreurs des observations supposée diagonale et estimée a posteriori et celle issue de la formulation analytique (définitions section 5.2).

On se place dans le même domaine $2D$ décrit en section 5.3 (grille régulière, pas spatial de 15 km). On note Δx le pas spatial entre deux observations, égal à $1°$ (i.e 80 km). Le vecteur d'état vrai noté x_t est de dimension $ns = 26 > p = 16$, où p est la dimension du vecteur d'observations. Le nombre de degrès de liberté du système ns est supérieur, comme souvent dans la réalité, au nombre d'observations, ce qui en fait un système sous-déterminé. Le système d'observations est donc fixé à 16 observations réparties régulièrement tous les degrès et centrées loin des bords du domaine (même configuration que pour la section 5.3). Les observations sont simulées en ajoutant un bruit normal au vecteur d'état vrai (notation 5.26). On rapelle brièvement que le vecteur d'état suit une loi normale centrée. La matrice

B de covariance du vecteur d'état a été définie à la section 5.3.1.

$$x_t \sim \mathcal{N}(0, \mathbf{B})$$

$$\mathbf{Y}^0 = \mathbf{H}(x_t) + \epsilon^0 \tag{5.26}$$

$$\epsilon^0 \sim \mathcal{N}(0, \sigma_{0t})$$

$$\mathbf{R} = \mathbf{E}[\epsilon^0(\epsilon^0)^t] = \sigma_0^2 \mathbf{I}_p$$

Où σ_{0t} est la variance d'erreurs d'observations spécifiée (true) et σ_0 celle que l'on cherche à ajuster via la matrice de covariance d'erreurs d'observations **R**. Les erreurs d'observations sont décorrélées spatialement ce qui justifie la forme diagonale de la matrice de covariance d'erreurs d'observations.

Lors des deux expériences réalisées, on utilise une liste croissante de longueur de corrélation du signal L_t prescite dans la définition de la matrice de covariance des erreurs statistiques utiles à l'analyse objective (aussi définies dans la section 5.2.3). La première expérience repose sur l'hypothèse que la variance d'erreurs d'observations est parfaitement connue ($\sigma_{0t} = \sigma_0$). La deuxième expérience repose sur l'hypothèse que la variance d'erreur d'observations est sous-estimée ($\sigma_{0t} > \sigma_0$). Pour l'ensemble de ces expériences la variance d'erreurs d'observations se doit d'être faible par rapport au vecteur d'état afin de permettre une bonne estimation, car dans un cas d'analyse objective nous ne disposons pas d'ébauche pour améliorer le vecteur d'état. Par conséquent les observations seront faiblement bruitées (proches du vecteur vrai) afin de permettre une reconstruction de bonne qualité : $\sigma_0 t = 0.2$.

1^e **expérience : variance d'erreurs d'observations parfaitement connue et observations faiblement bruitées** Lorsque la variance d'erreurs d'observations est connue parfaitement, les deux formulations estimées (5.18) et (5.10) du DFS coïncident avec la formulation objective théorique. De plus l'expression du DFS issue des statistiques d'erreurs à posteriori et celle issue de la méthode de Girard présentent des résultats similaires, tableau (5.4) : leur différence relative pour l'ensemble des longueurs de corrélation testées est en moyenne inférieure à 2%. Enfin lorsque la longueur de corrélation augmente le contenu en information tends à diminuer : cette tendance est en accord avec les tests des facteurs d'influence sur le DFS réalisés à la section (5.2.3).

variance d'erreur des observations parfaitement connue : $\sigma_{0t} = \sigma_0 = 0.2$			
L(km)	$\mathsf{DFS}_{\mathsf{Analytique}}$	$\mathsf{DFS}_{\mathsf{Girard}}$	$\mathsf{DFS}_{\mathsf{Apost}}$
$60 < \Delta x$	15,36	15,26	15,92
100	14,84	14,94	15,04
180	10,98	10,76	10,43
240	8,60	8,76	8,81
300	7,09	6,81	7,01

Table 5.4 – 1^e *expérience : variance d'erreurs d'observations parfaitement connue et observations faiblement bruitées*

2^e **expérience : variance d'erreurs d'observations sous-estimée et observations faiblement bruitées** Dans cette série de résultats (tableau 5.5), la variance d'erreur d'observations est choisie inférieure à celle connue . Par comparaison numérique avec l'expérience précédente, à moindre bruit le DFS est supérieur. Cette tendance est encore en accord avec celle observée dans la section 5.2.3. De plus, les résultats issus des expressions du $\mathsf{DFS}_{\mathsf{Girard}}$ et du $\mathsf{DFS}_{\mathsf{Apost}}$ sont similaires, avec une différence relative de 1% à 3% pour une longueur de corrélation, $L_t = 60$ km inférieure à la résolution spatiale des observations. Par ailleurs dès que L_t est $> \Delta x$ les résultats issus de l'expression $\mathsf{DFS}_{\mathsf{Apost}}$ augmentent plus fortement que pour le cas issu du $\mathsf{DFS}_{\mathsf{Girard}}$, avec une différence relative atteignant les 35%. De plus lorsque la longueur de corrélation augmente, les valeurs du $\mathsf{DFS}_{\mathsf{Apost}}$ ne diminuent pas, contrairement à ce qui serait attendu (tendance soulignée en section 5.2.3). La tendance ainsi que les valeurs numériques issues de l'expression $\mathsf{DFS}_{\mathsf{Apost}}$ ne sont pas en accord avec ceux de la formulation analytique.

variance d'erreur des observations sous-estimée : σ_{0t}=0.2 et σ_0=0.1			
L(km)	$\mathsf{DFS}_{\mathsf{Analytique}}$	$\mathsf{DFS}_{\mathsf{Girard}}$	$\mathsf{DFS}_{\mathsf{Apost}}$
60	15,83	15,43	15.93
100	15,67	15,93	16,74
180	13,15	12,99	17,14
240	10,71	10,65	15,41
300	8,92	9,00	12,42

Table 5.5 – 2^e *expérience : variance d'erreurs d'observations sous-estimées et observations faiblement bruitées*

L'hypothèse de variance d'erreurs d'observations sous-estimée faite en expérience 2, permet de remarquer que le DFS est dans ce cas sur-estimé. Lors de ces deux expériences nous nous sommes placés dans un cadre d'analyse objective afin d'avoir accès à la formulation analytique du DFS. Par comparaison elle permet

donc de conclure quant à la robustesse de des formulations présentées en section 5.2. La méthode d'approximation du DFS basée sur la randomisation de Girard présente des résultats cohérents avec la formulation analytique du DFS, ce qui n'est pas le cas pour la seconde expérience réalisée avec la méthode dite aposteriori. En effet lorsque l'hypothèse de variance d'erreurs d'observations sous-estimée est faite, on se place alors dans le deuxième cas de figure décrit en section 5.2. Lorsque les matrices de covariance d'innovation connues à priori et a posteriori ne sont plus cohérentes, il semble que la formulation $\mathsf{DFS}_{\mathsf{Apost}}$ ne soit plus valable. Enfin la méthode d'approximation du DFS basée sur la randomisation de Girard, et utilisant deux jeux d'observations perturbées (suivant les variances d'erreurs prescrites sur les observations) semble plus robuste pour estimer le DFS.

5.5 Conclusion

Au sein de ce chapitre, nous avons mis en évidence les relations inhérentes entre le contenu en information, les échelles de corrélation et le bruit de mesure. Il a été mit en avant que le DFS permet de quantifier l'apport d'information et d'analyser en quelle proportion les observations sont utilisées (proportion quantitative et localisée selon la grille). Il paraît donc essentiel d'évaluer précisément et de manière optimale le contenu en information. De ce fait, plusieurs formulations ont été évaluées dans un cas simple d'analyse objective dans le but d'élaborer une première approche d'estimation du DFS applicable par la suite dans le système d'assimilation de données de Mercator Océan. Le DFS peut être calculé de manière exacte et comparé à ces différentes estimations. Les résultats obtenus semblent indiquer que seule la première appoche (celle dite de Girard) se montre robuste. La seconde alternative sera donc écartée pour le calcul des DFS dans les systèmes mis en œuvre par Mercator Océan. Des travaux et tests complémentaires sont actuellement effectués par les équipes *R&D* d'assimilation de Mercator Océan.