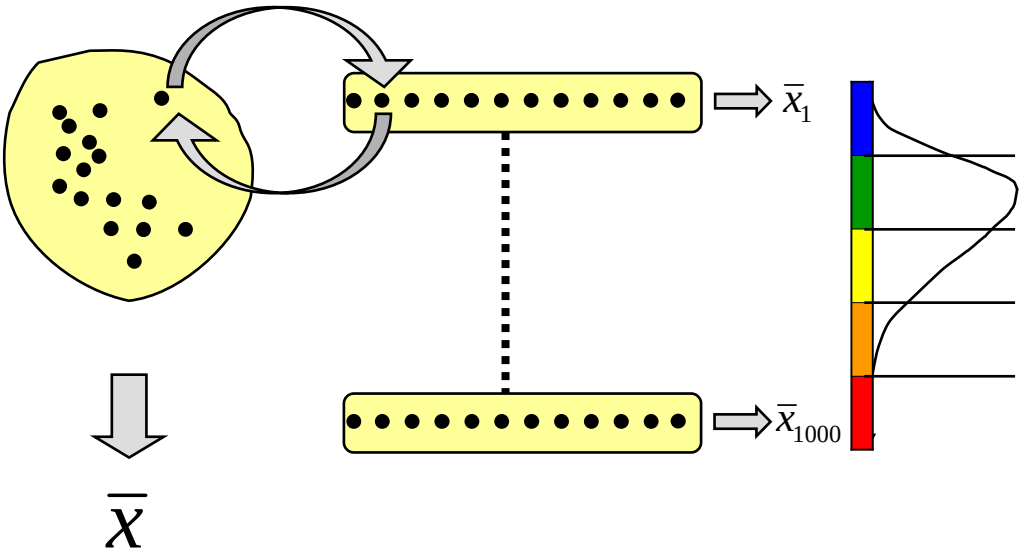
Direction des opérations
Service VIGIES

Alex Soudant
Dominique Soudant
Alain Lefebvre

Décembre 2008 - R.INT.DOP/DYNECO/VIGIES 2008-17

Influence de la période et de la fréquence d'échantillonnage sur le percentile 90 de la fluorescence



Influence de la période et de la fréquence d'échantillonnage sur le percentile 90 de la fluorescence

Fiche documentaire

Numéro d'identification du rapport : R.INT.DOP/DYNECO/VIGIES 2008-17 **Diffusion** : libre : ☑ restreinte : ☐ interdite : ☐ **Validé par** : **Y.-H. De Roeck** Adresse électronique :	**date de publication :** DEC 08 **nombre de pages :** 36 **bibliographie :** oui **illustration(s) :** oui **langue du rapport :** F

Titre de l'article

Influence de la période et de la fréquence d'échantillonnage sur le percentile 90 de la fluorescence

Contrat n° Rapport intermédiaire ☐ Rapport définitif ☑

Auteur(s) principal(aux) :	**Organisme / Direction / Service, laboratoire**
Alex Soudant	Ifremer / DYNECO / VIGIES
Dominique Soudant	Ifremer / LER / LERBL
Alain Lefebvre	

Résumé

Le percentile 90 (P90) de la chlorophylle *a* mesurée mensuellement de mars à octobre est un des indicateurs utilisé pour définir le bon état écologique des masses d'eaux pour application de la Directive Cadre sur l'Eau (DCE). En l'absence de séries temporelles de mesures haute fréquence de ce paramètre, la sensibilité du P90 à la période et la fréquence d'échantillonnage est mal connue. Mais comme il existe une relation chlorophylle *a* – fluorescence et des séries de mesure haute fréquence de la fluorescence, une première approche de l'étude de l'influence de la fréquence et de la période d'échantillonnage sur le percentile 90 de la fluorescence est possible.

Cette étude a porté sur trois années de mesures (2005, 2006, 2007) de la bouée Marel-Carnot. Quatre périodes d'échantillonnage ont été définies et 5 fréquences d'échantillonnage (mensuelle à hebdomadaire) ont été envisagées. La méthode de rééchantillonnage choisie est le *bootstrap*.
Les résultats montrent que l'échantillonnage régulier sur 12 mois permet d'obtenir les P90 les moins biaisés. L'augmentation de la fréquence d'échantillonnage induit une réduction de la variabilité de l'estimation. Ces résultats restent liés à la fluorescence et au site géographique particulier de la bouée Marel-Carnot. Compte tenu du plan d'échantillonnage actuel de la chlorophylle *a*, l'ajout d'une seule mesure en hiver (i.e. novembre, décembre, janvier, février) dotée d'un poids de 4 permettrait une meilleure estimation.

Mots-clés

Fluorescence, percentile 90, *bootstrap*, chlorophylle *a*, Directive Cadre sur l'eau (DCE), échantillonnage

Rédacteur	**Vérificateur**	**Approbateur**
Nom : D. Soudant **Date :02/12/2008** **Visa**	**Nom : Y.-H. De Roeck** **Date : 14/11/2008** **Visa**	**Nom : A. Daniel** **Date : 02/12/2008** **Visa**

Sommaire

Avant-propos

Ce travail a été réalisé pour l'obtention du Diplôme Universitaire Technique (DUT) de Statistiques et Traitements Informatiques des Données (STID) de l'université de Paris. Pour les besoins de la publication en rapport interne, les figures et tableaux ont été complétés, le texte modifié en conséquence et parfois restructuré. Nous remercions, Catherine Belin, Anne Daniel-Scuiller, Yann-Hervé De Roeck et Anne Pellouin-Grouhel pour leurs commentaires.

1 Introduction

La Directive Cadre sur l'Eau (DCE, directive 2000/60/CE) demande aux États de l'Union Européenne d'établir l'état des milieux aquatiques de manière à qualifier les masses d'eau, les classifier et éventuellement prévoir des actions de restauration. Ces évaluations sont réalisées, entre autres critères, à partir de paramètres hydrologiques et biologiques. Parmi ceux-ci, la concentration en chlorophylle *a* dans l'eau de mer permet d'apprécier la biomasse phytoplanctonique. La métrique retenue est le percentile 90. La circulaire DCE 2007/20 du MEDDAT impose une surveillance durant un plan de gestion de six ans entre les mois de mars et d'octobre à raison d'une mesure par mois dans les masses d'eaux côtières et de transition de l'Atlantique, de la Manche et de la mer du Nord.

Les fréquences de suivi de la chlorophylle-a dans les divers réseaux de surveillance opérés à l'Ifremer sont mensuelles, bimensuelles ou encore hebdomadaires selon la période de l'année et la zone d'étude. Il a été montré que des fréquences de prélèvement bimensuelles et hebdomadaires peuvent engendrer une appréciation de la biomasse phytoplanctonique différente que celle obtenue avec une fréquence mensuelle (Le Goff *et al.,* 2004). Une limitation de l'étude de la fréquence d'échantillonnage de la chlorophylle *a* provient de l'absence de séries temporelles mesurées à haute fréquence. Par contre, de telles séries sont disponibles pour la fluorescence grâce au système de mesure MAREL. Comme la fluorescence est un traceur de la chlorophylle *a*, il est possible de se baser sur ce paramètre pour avoir une première approche de sa variabilité temporelle.

Le but du présent travail est d'étudier l'influence de la fréquence et de la période d'échantillonnage sur le percentile 90 de la fluorescence. Le P90[1] de l'ensemble des données valides étant considéré comme la référence, il s'agit donc de rééchantillonner à différentes fréquences les données du système de mesures MAREL sur des années pleines ou bien sur la fenêtre mars-octobre.

2 Chlorophylle *a* et fluorescence

2.1 Chlorophylle *a*

La chlorophylle *a* est le principal pigment des organismes végétaux. En convertissant l'énergie lumineuse en énergie chimique, il permet la photosynthèse, c'est à dire la fixation de carbone induite par la lumière. La chlorophylle *a* est un indicateur universel (mais non strictement proportionnel) de la biomasse phytoplanctonique qui est à la base de la chaîne alimentaire.

[1] Ici et par la suite, les expressions « percentile 90 » et P90 sont utilisées comme synonymes.

Les pigments chlorophylliens émettent une fluorescence lorsqu'ils sont irradiés à des longueurs d'ondes très proches de leurs longueurs d'ondes d'absorption.

Sous nos latitudes tempérées, le cycle du phytoplancton présente classiquement une forte floraison de vie végétale printanière et une floraison automnale d'intensité plus faible. Ces deux floraisons sont séparées par la période estivale qui ne fournit pas les conditions nutritives nécessaires à ces fortes croissances de population végétale. L'hiver est également une période de faible production planctonique en raison d'une température et d'un ensoleillement trop faible. Cette période permet la reconstitution des stocks nutritifs. Sauf cas particulier, la stratégie de prélèvement devra donc intégrer ce facteur, avec un resserrement des mesures au printemps et allègement en hiver (Aminot 2002).

2.2 Fluorescence

La mesure de la concentration de la chlorophylle *a* par fluorimétrie *in vivo*, présente deux avantages incontestables : simplicité et rapidité. Cependant la qualité de cette méthode est soumise à la variabilité de la relation fluorescence-chlorophylle *a*. Ce rapport peut en effet varier en fonction de paramètres tels que : la diversité des populations phytoplanctoniques en présence, l'état physiologique des cellules, la solubilité de la fluorescence, l'alternance jour/nuit, l'âge de la population phytoplanctonique et la température (Herbland & Voituriez 1977). Néanmoins, elle permet d'obtenir rapidement une indication de l'état de la biomasse phytoplanctonique.

3 Données et méthodes

3.1 Données

Les données proviennent du système MAREL Carnot opérationnel depuis septembre 2004. C'est une station de mesure haute fréquence de la qualité de l'eau en zone côtière. L'ensemble des données est transmis deux fois par jour au Centre Ifremer Manche Mer du Nord de Boulogne-sur-Mer par liaison GSM ; ensuite débute la validation et le traitement des données. Dès cette phase de transmission, les données sont soumises à un ensemble de procédures de contrôle de qualité. Les données sont caractérisées par un niveau de traitement et par un niveau de qualité.
Une partie de ce contrôle est fait automatiquement (contrôle du format des fichiers, de la gamme de valeurs observées en référence à des valeurs de références) ; les données sont alors dans un niveau de traitement T0,5. Un contrôle visuel est également réalisé afin d'identifier « à dire d'experts » le niveau de qualité de la donnée. Après cette étape, les données passent en niveau de traitement T1,0 et sont accessibles par l'internet via http://www.ifremer.fr/difMarelCarnot/. Les accès sont possibles via trois domaines : public, scientifique et technique en fonction du profil de l'utilisateur (Lefebvre 2008).

Les données pour les années 2004 et 2008 étant incomplètes, elles n'ont pas été retenues. Les données utilisées concernent les années 2005, 2006, et 2007. Les variables utilisées pour cette étude sont la date et l'heure de mesure, la fluorescence et la hauteur d'eau. Le nombre de données de fluorescence pour l'année 2005 est de 21 378, 22 896 pour l'année 2006 et 24 163 pour l'année 2007. En regroupant les trois années, 68 437 données sont disponibles. La station mesure la fluorescence toutes les 20 minutes. Le paramètre fluorescence peut être considéré comme l'un de ceux présentant un maximum de résultats « bons » au sens de la procédure de validation / qualification mise en œuvre. Les résultats dits « faux », « hors statistiques » et « douteux » sont minoritaires et principalement localisés en dehors des périodes d'efflorescences (signal très faible) ou lors des phases de maintenance (arrêt du système). Les données de hauteur d'eau étant mesurées deux fois plus souvent, seules celles correspondant à un même moment de mesure ont été retenues.

3.2 Méthodes

3.2.1 Filtrages

Les données brutes sont filtrées pour correspondre à différentes exigences et hypothèses. Le premier filtre concerne la sélection des données diurnes. Il est primordial car, pour une même concentration de chlorophylle *a*, la fluorescence varie entre le jour et la nuit (Herbland & Voituriez 1977). Arbitrairement, la plage horaire de 7 heures à 19 heures, *Greenwich mean time* (GMT/UTC), est choisie afin de correspondre à des mesures en journée quelque soit la saison. En complément, les données de fluorescence dont la valeur est égale à 0 sont supprimées car elles ne correspondent à aucune réalité biologique. Ce traitement constitue une base sur laquelle d'autres filtres peuvent être appliqués. Le second filtre correspond aux données opérationnellement observables par un réseau de surveillance « classique ». Il s'agit en premier lieu de ne garder que les données mesurées du lundi au vendredi. De plus, une interrogation de la base de données de la surveillance environnementale a montré que 90% des relevés se déroulent entre 9 heures et 16 heures. Enfin, les prélèvements sont effectués à plus ou moins deux heures autour de la marée haute, ce qui permet d'avoir les mêmes conditions de relevés pour les mesures. Ces trois conditions forment le filtre opérationnel. Le troisième filtre répond aux exigences de la DCE d'effectuer les prélèvements de mars à octobre (Pellouin-Grouhel, 2006). Il peut être combiné soit avec le premier filtre, soit avec le second. Ceux ci forment en tout quatre filtrages différents :

- diurne ;
- diurne et opérationnel ;

- diurne et mars-octobre ;
- diurne, opérationnel et mars-octobre.

Les quatre ensembles de données issus de l'application de ces filtrages sont soumis aux rééchantillonnages.

3.2.2 Méthode de rééchantillonnage

La méthode utilisée est le *bootstrap*. Cette méthode se présente de la manière suivante (Davidson & Hinkley 1997). A partir d'un échantillon, on calcule une estimation du paramètre qui nous intéresse. L'étape suivante consiste à créer un nouvel échantillon de même taille que l'échantillon observé. Pour ce faire, on effectue un tirage aléatoire avec remise. Cette étape de rééchantillonnage est réalisée un grand nombre de fois. A partir de chaque échantillon ainsi nouvellement créé, on calcule une estimation du paramètre d'intérêt. L'ensemble de ces résultats permet d'apprécier la distribution des estimations du paramètre. Un intérêt tout particulier de cette méthode, est de pouvoir apprécier la variabilité d'un paramètre à partir de sa seule définition.

Il est possible de définir des strates structurant l'échantillon de base. Le rééchantillonnage est alors effectué dans chacune d'elles préservant ainsi l'organisation des données. Cette option est utilisée ici afin de prendre en compte la saisonnalité de la fluorescence.

3.2.3 Simulations

Les données diurnes des années 2005, 2006 et 2007 sont d'abord considérées séparément afin d'appréhender l'effet des stratégies de mesures sur une seule année. Ensuite, ces trois années sont considérées conjointement pour simuler 6 années de mesures. Les quatre filtres définis sont utilisés pour tout à la fois s'approcher pas à pas des conditions de la DCE et évaluer l'effet de chacun d'eux, les données diurnes constituant la référence.

La DCE impose une fréquence mensuelle. Dans le cadre de la surveillance, l'Ifremer utilise les fréquences mensuelle, bimensuelle et parfois hebdomadaire. Ces trois fréquences sont toujours simulées. Deux autres fréquences sont testées lorsque le filtre mars-octobre est appliqué. Elles correspondent à un effort d'échantillonnage intermédiaire entre le mensuel et le bimensuel, en rapprochant les mesures pendant les périodes d'efflorescence. Par rapport à la fréquence mensuelle, la première, désignée par « F1 », consiste à doubler les mesures des mois de mars-avril et septembre-octobre. La seconde, désignée par « F2 », double les mesures des mois de mars à mai et de août à octobre. Le récapitulatif des fréquences d'échantillonnage, par ordre croissant du nombre mesurées effectuées, est le suivant :

- mensuelle ;
- F1 ;

- F2 ;
- bimensuelle ;
- hebdomadaire.

Ces fréquences nécessitent la définition des unités temporelles « mois », « quinzaine » et « semaine ». La première est le mois calendaire. La seconde est le demi-mois calendaire tel que la première quinzaine est constituée des jours 1 à 14 et la seconde des autres jours. La semaine est une demi-quinzaine, les jours 7, 14, 21 constituant les bornes supérieures des intervalles définissant les trois premières d'un mois. Il s'ensuit de ces définitions qu'une année est constituée de 12 mois, 24 quinzaines et 48 semaines.

Les fréquences testées définissent des strates au sein desquelles une mesure est tirée lorsque l'on considère les années séparément, 6 mesures pour la simulation de six années de suivi. Le processus de simulation ainsi défini suppose des rééchantillonnages de taille différente de l'échantillon d'origine. Cette modification du *bootstrap* n'est pas conseillée lorsque la taille des nouveaux échantillons est supérieure à celle de l'échantillon initial (Davison & Hinkley 1997). Mais ici, d'une part la taille est inférieure, et d'autre part notre approche tient plus de la simulation que de l'estimation. En revanche, cette caractéristique a une incidence sur le nombre d'échantillons créés. Le nombre de 1 000 est généralement considéré comme satisfaisant pour une précision des intervalles de confiance à cinq pourcents. Ainsi, avec un échantillon de taille *n*, le bootstrap conduit à (*n* x 1 000) tirages aléatoires sans remise. Mais dans notre cas, si l'on considère l'échantillonnage mensuel des données diurnes, seuls 72 000 tirages aléatoires seront effectués contre un peu moins de 40 millions au maximum dans un *bootstrap* normal[2], soit 0,18 %. Comme dans notre simulation, le nombre de tirages aléatoires ne tient pas compte de la taille de l'échantillon initial, plus cette dernière est petite, plus le rapport des tirages aléatoires est élevé. Ainsi, dans le cas des données diurnes opérationnelles de mars à octobre échantillonnées une fois par semaine, la simulation entraine 192 000 tirages contre 5 760 000 pour un *bootstrap* normal, soit un rapport de 3,33 %[3]. Ce que souligne la faiblesse de ces pourcentages, c'est que l'examen des possibles par le tirage aléatoire sur lequel s'appuie le *bootstrap* est ici très faible. En particulier, il peut en résulter une instabilité des résultats obtenus. En première approche, de manière à palier ce problème tout en assurant la réalisation des opérations avec les processeurs disponibles, le nombre d'échantillons tirés a été fixé arbitrairement à 10 000.

---

2 Calcul effectué en supposant 3 années de 365 jours, fluorescence mesurée de manière tri-horaire 12 heures par jour : 3 x 365 x 12 x 3 x 1 000 = 39 420 000.

3 D'une part, 6 années de 8 mois de 4 semaines : 6 x 8 x 4 x 1 000 = 192 000. D'autre part, 3 années de 8 mois de 4 semaines de 5 jours à raison de 4 heures par jour à fréquence tri-horaire : 3 x 8 x 4 x 5 x 4 x 3 x 1 000 = 5 760 000, en admettant que la condition « à plus ou moins deux heures de la marée haute » soit satisfaite par la prise en compte de 4 heures par jour.

Les distributions des valeurs des percentiles 90 sont représentées par des histogrammes. Pour les simulations des plans de gestion de 6 ans, les axes et découpages en classe sont uniformisés à travers les filtres et les fréquences. Moyennes et médianes des distributions sont indiquées sur ces figures. Les percentiles 90 calculés avec l'ensemble des valeurs diurnes pour une année ou les trois années conjointes sont également portés.

De manière à apprécier la variabilité des estimations, les intervalles de confiance sont calculés selon la « méthode du percentile » (Davison & Hinkley 1997). Cette méthode est équivalente aux autres méthodes existantes dans le cas de grands échantillons. Les étendues de ces intervalles de confiance sont données.

3.2.4 Programmation

Ce projet a nécessité une étape de programmation visant à comprendre et adapter les procédures disponibles sous le logiciel R. Le package *boot* propose une mise en œuvre du *bootstrap* ainsi que le calcul d'intervalles de confiance. La principale difficulté est la création d'échantillons de taille inférieure à celle de l'échantillon d'origine. La première fonction créée (cf. Annexe 1.1) s'appuie sur ce package mais ne peut aboutir pour 10 000 échantillons. Une seconde implémentation (cf. Annexe 1.2) s'appuie sur la fonction `sample()` mais nécessite la duplication de deux fonctions du package *boot*, pour calculer les intervalles de confiance selon la méthode du percentile. Ce dernier ensemble permet la création de 10 000 échantillons.

Le code écrit pour ce projet s'inscrit dans l'architecture de travail mis en place au sein du service DYNECO/VIGIES (cf. Annexe 2). L'intérêt de cette architecture est de structurer le travail de programmation pour pouvoir reprendre facilement un travail réalisé précédemment et l'adapter selon des nouveaux besoins ou pour une étude plus récente. Le partitionnement du travail permet également à plusieurs personnes de travailler autour d'un même projet avec une plus grande facilité de communication et sans se gêner dans l'utilisation des ressources du projet.

4 Résultats

Le tableau 1 présente le nombre de données par filtre et pour les années 2005, 2006 et 2007.

Tableau 1 : Nombre de données de fluorescence par filtre et pour les années 2005, 2006 et 2007.

Filtres	2005	2006	2007	Total
Diurne	10 521	10 853	11 581	32 955
Diurne et opérationnel	2 279	2 432	2 659	7 370
Diurne et mars octobre	7 245	7 468	7 710	22 423
Diurne, opérationnel et mars octobre	1 575	1 684	1 788	5 047

Certaines données sont manquantes. La moyenne de données manquantes par mois est de 18,1%. Le maximum pour l'année 2005 atteint 45,9% pour le mois d'août, 32,9% en septembre pour l'année 2006, et enfin pour l'année 2007 il est de 42.3% au mois de juin.

Le tableau 2 présente des statistiques descriptives des données diurnes de fluorescence des années 2005 à 2007.

Tableau 2 : Statistiques descriptives concernant la fluorescence (FFU) mesurée par la bouée MAREL pour les années 2005, 2006 et 2007.

	Années		
	2005	2006	2007
Minimum	0.02	0.01	0.01
Q1	0.43	0.27	0.39
Médiane	0.85	0.67	0.58
Moyenne	1.35	2.54	1.07
Q3	1.81	1.81	1.16
Maximum	9.89	58.31	15.31

La moyenne est maximale pour l'année 2006 du fait de la présence d'une valeur très élevée par rapport aux deux autres années. Toute proportion gardée, la répartition selon les quartiles est peu affectée par cette valeur exceptionnelle

4.1 Séries temporelles

La figure 1 présente les séries temporelles des percentiles 90 par semaine des données diurnes des années 2005, 2006 et 2007. Les floraisons printanières se traduisent par les pics de fluorescence les plus importants. Leurs dates d'apparition et intensités sont différentes d'une année à l'autre. En 2005 et 2007, quelques pics de plus faibles intensités sont toutefois observés en période estivale. La variabilité interannuelle est importante.

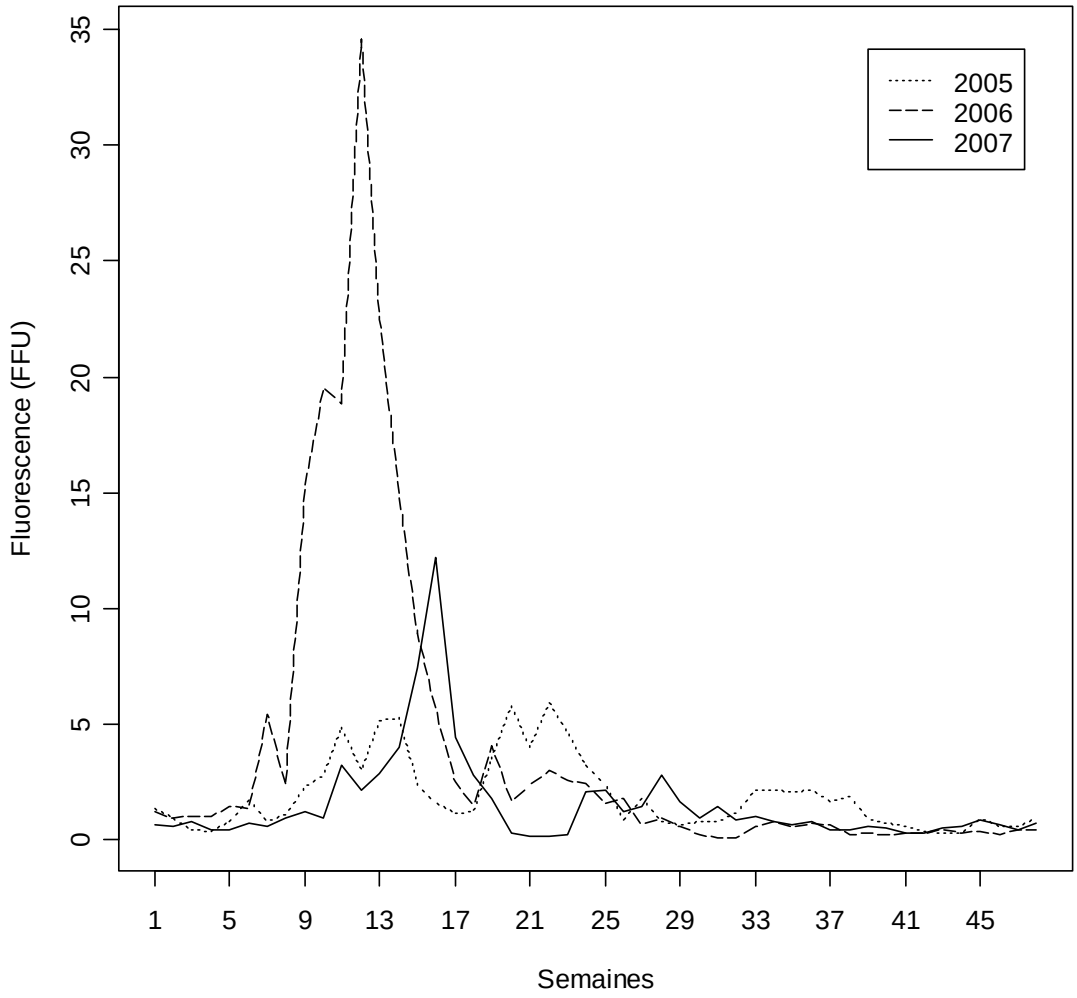


Figure 1 : Séries temporelles des percentiles 90 hebdomadaire des données diurnes de fluorescence pour les années 2005, 2006 et 2007.

4.2 Simulations annuelles

Le tableau 3 présente des indices de position et de variabilité des distributions simulées des percentiles 90.

Tableau 3 : Valeurs (FFU) des « vrais » percentiles 90 des données diurnes des années 2005, 2006 et 2007, ainsi que les moyennes, médianes, intervalles de confiance (IC) et leurs étendues des percentiles 90 simulés selon les fréquences d'échantillonnage mensuelle, bi-mensuelle et hebdomadaire pour chacune de ces années.

		Années		
		2005	2006	2007
« Vrai P90 »		3,17	7,19	2,16
Fréquence mensuelle	Moyenne	2,64	5,07	1,85
	Médiane	2,5	4,1	1,7
	IC (étendue)	[1,4 ;4,5] (3,1)	[1,5 ;14,5] (13)	[0,9 ;3,5] (2,6)
bimensuelle	Moyenne	2,79	5,31	1,92
	Médiane	2,7	4,8	1,9
	IC (étendue)	[1,9 ;4] (2.1)	[2,2 ;11,1] (8,9)	[1,2 ;2,9] (1,7)
hebdomadaire	Moyenne	2,87	5,60	1,99
	Médiane	2,8	5,4	1,9
	IC (étendue)	[2,2 ;3,6] (1.4)	[3 ;9,6] (6,6)	[1,5 ;2,7] (1,2)

Les figures 2, 3 et 4 montrent les distributions des simulations pour, respectivement, les années 2005, 2006 et 2007.

Dans les trois cas, moyennes, médianes et modes des distributions des percentiles 90 sont inférieurs aux vrais P90 calculés sur l'ensemble des données de chacune des années. L'écart semble se réduire avec le resserrement des prélèvements. Par ailleurs, les intervalles de confiance montrent toujours une diminution d'étendue avec l'augmentation de la fréquence d'échantillonnage, c'est-à-dire une diminution de la variabilité. Les figures 2 et 4 sont assez similaires. Les étendues des valeurs sont comparables et les distributions relativement symétriques. En revanche, la figure de l'année 2006 (figure 3) est asymétrique, présente une étendue de valeurs plus importante et les intervalles de confiance sont plus larges.

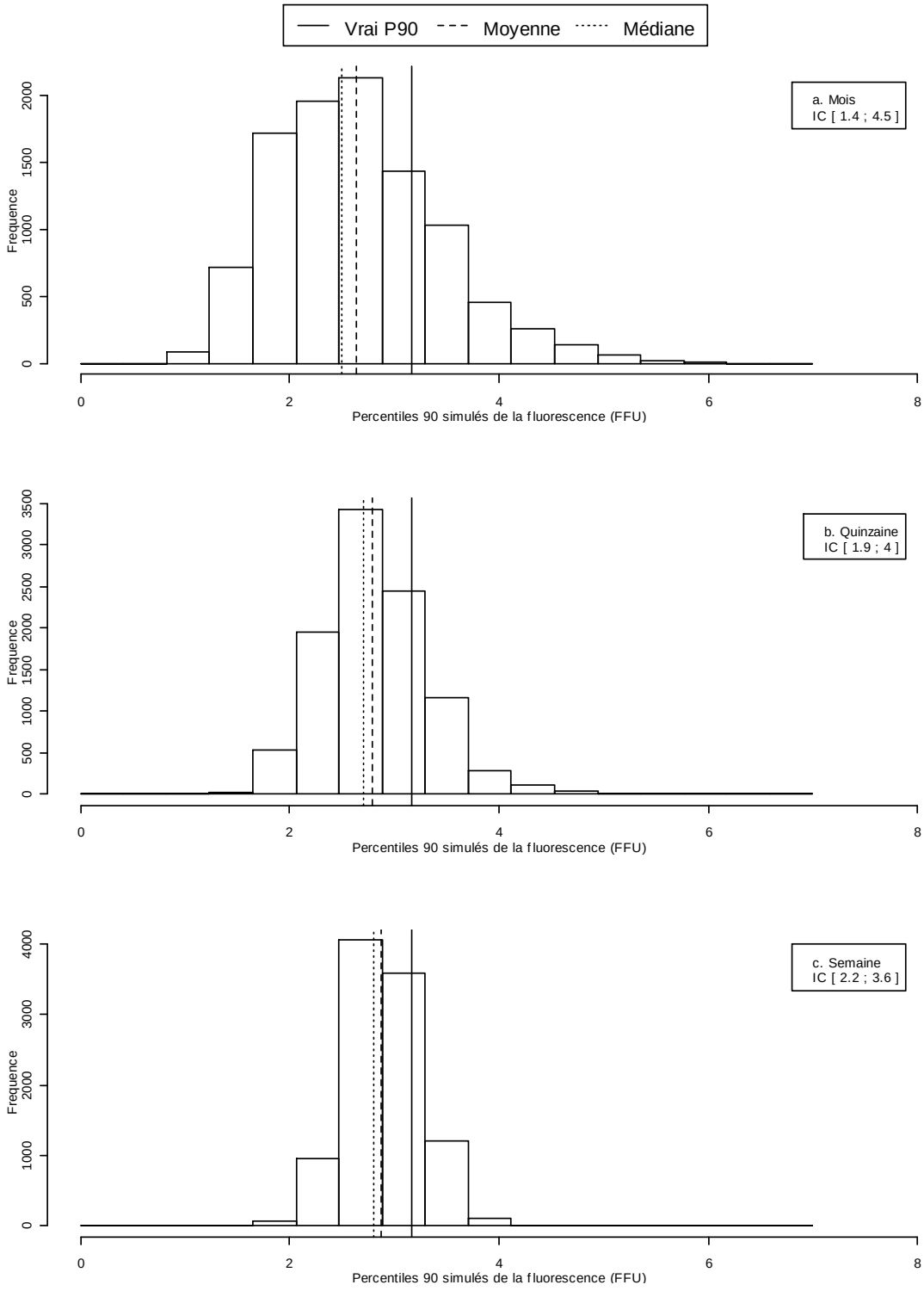


Figure 2 : Histogrammes des percentiles 90 de la fluorescence simulés pour une année d'échantillonnage à partir des données diurnes de l'année 2005 pour les fréquences a) mensuelle, b) bimensuelle et c) hebdomadaire. Moyennes, médianes et intervalles de confiance sont calculés avec les percentiles 90 simulés. Le « vrai P90 » est calculé avec les données diurnes de l'année 2005.

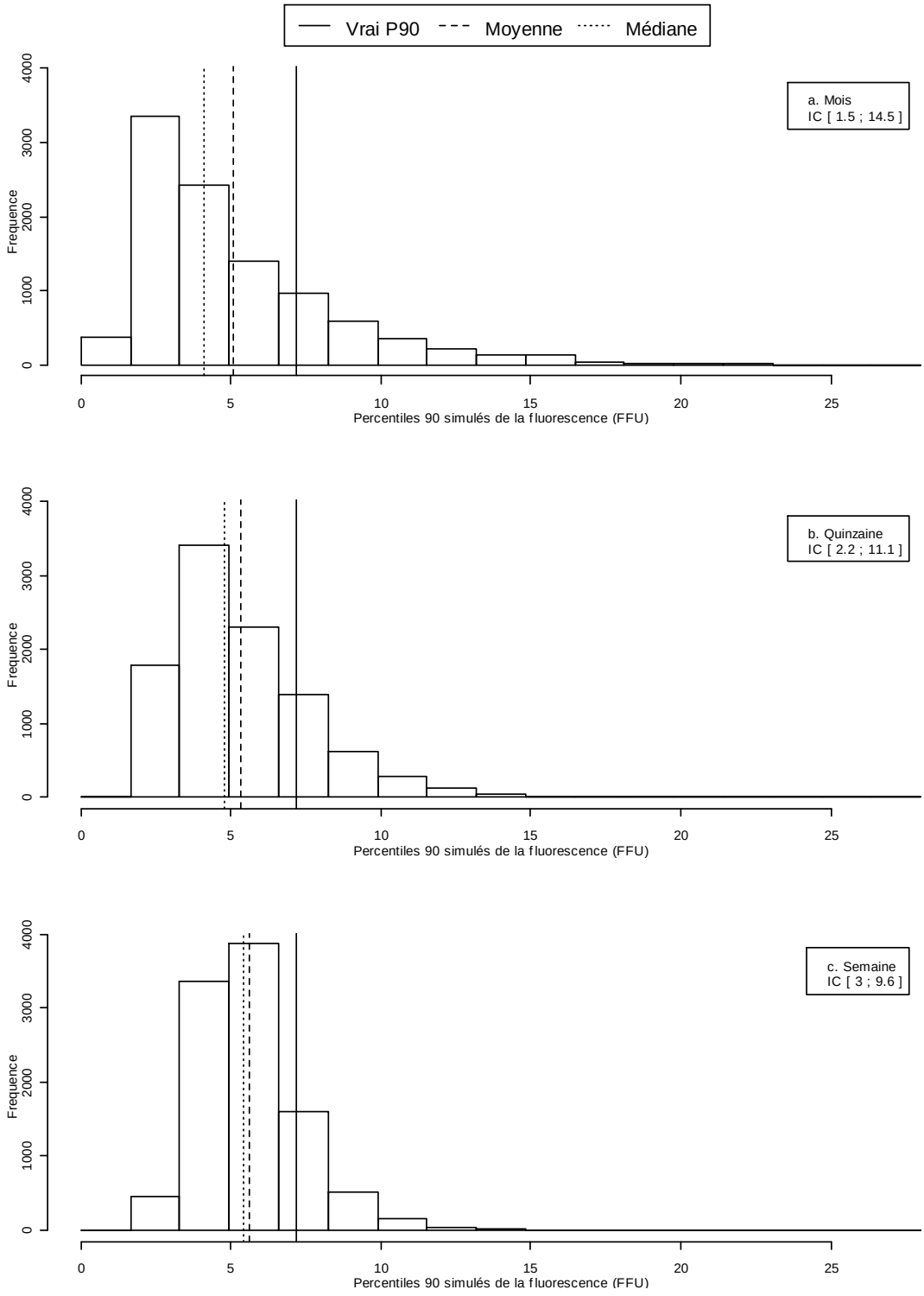


Figure 3 : Histogrammes des percentiles 90 de la fluorescence simulés pour une année d'échantillonnage à partir des données diurnes de l'année 2006 pour les fréquences a) mensuelle, b) bimensuelle et c) hebdomadaire. Moyennes, médianes et intervalles de confiance sont calculés avec les percentiles 90 simulés. Le « vrai P90 » est calculé avec les données diurnes de l'année 2006

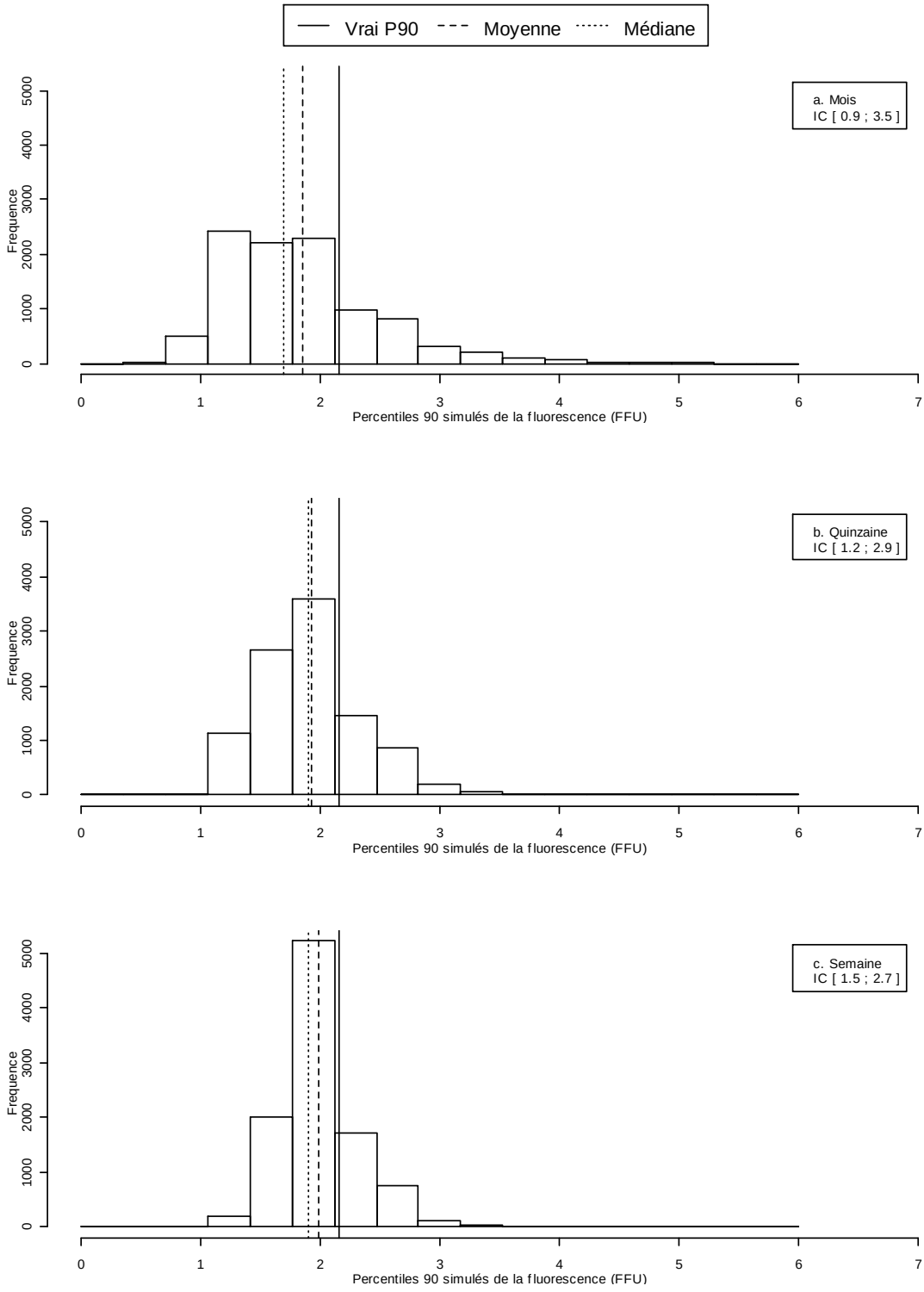


Figure 4 : Histogrammes des percentiles 90 de la fluorescence simulés pour une année d'échantillonnage à partir des données diurnes de l'année 2007 pour les fréquences a) mensuelle, b) bimensuelle et c) hebdomadaire. Moyennes, médianes et intervalles de confiance sont calculés avec les percentiles 90 simulés. Le « vrai P90 » est calculé avec les données diurnes de l'année 2007.

4.3 Simulations pour un plan de gestion de 6 ans

Le tableau 4 présente les indices de position et de variabilité pour les simulations des suivis sur 6 années.

Tableau 4 : Valeurs (FFU) des moyennes, médianes et intervalles de confiances (IC) et leurs étendues des distributions des P90 de la fluorescence pour les simulations des plans de gestion DCE de 6 ans en fonction du filtre appliqué et des fréquences d'échantillonnage. La « vraie » valeur du P90 calculé sur les données diurnes des années 2005, 2006 et 2007 est : 3,28 FFU.

		Filtres			
		Diurnes	Diurnes Opérationnel	Diurnes Mars octobre	Diurnes Opérationnel Mars octobre
Fréquence					
Mensuelle	Moyenne	3,13	3,39	4,28	4,74
	Médiane	3	3,2	3,9	4,4
	IC (étendue)	[2,1 ;4,9] (2,8)	[2,2 ;5,3] (3,1)	[2,5 ;7,7] (5,2)	[2,7 ;9,6] (6,9)
F1	Moyenne			5,07	5,83
	Médiane			4,8	5,4
	IC (étendue)			[3 ;9,2] (6,2)	[3,3 ;10,5] (7,2)
F2	Moyenne			4,56	5,15
	Médiane			4,4	4,9
	IC (étendue)			[2,9 ;7,7] (4,8)	[3,2 ;8,9] (5,7)
Bimensuelle	Moyenne	3,14	3,37	4,17	4,61
	Médiane	3,1	3,3	4	4,5
	IC (étendue)	[2,3 ;4,3] (2)	[2,4 ;4,8] (2,4)	[2,9 ;6,3] (3,4)	[3 ;7,2] (4,2)
Hebdomadaire	Moyenne	3,16	3,43	4,25	4,76
	Médiane	3,1	3,4	4,2	4,6
	IC (étendue)	[2,6 ;4] (1,4)	[2,7 ;4,4] (1,7)	[3,2 ;5,7] (2,5)	[3,5 ;6,5] (3)

La distribution du percentile 90 pour le filtre sur les données diurnes est présentée par la figure 5. Les indices de positions sont très proches de la vraie valeur quelque soit la fréquence d'échantillonnage et toujours inférieurs. Comme pour les années prises séparément, l'étendue de la distribution diminue avec l'augmentation de la fréquence d'échantillonnage. Il en va de même des étendues des intervalles de confiance. En ajoutant le filtre opérationnel sur ces données, les résultats sont similaires (cf. figure 6). Les indices de position sont cette fois-ci légèrement supérieurs à la vraie valeur à une exception près.

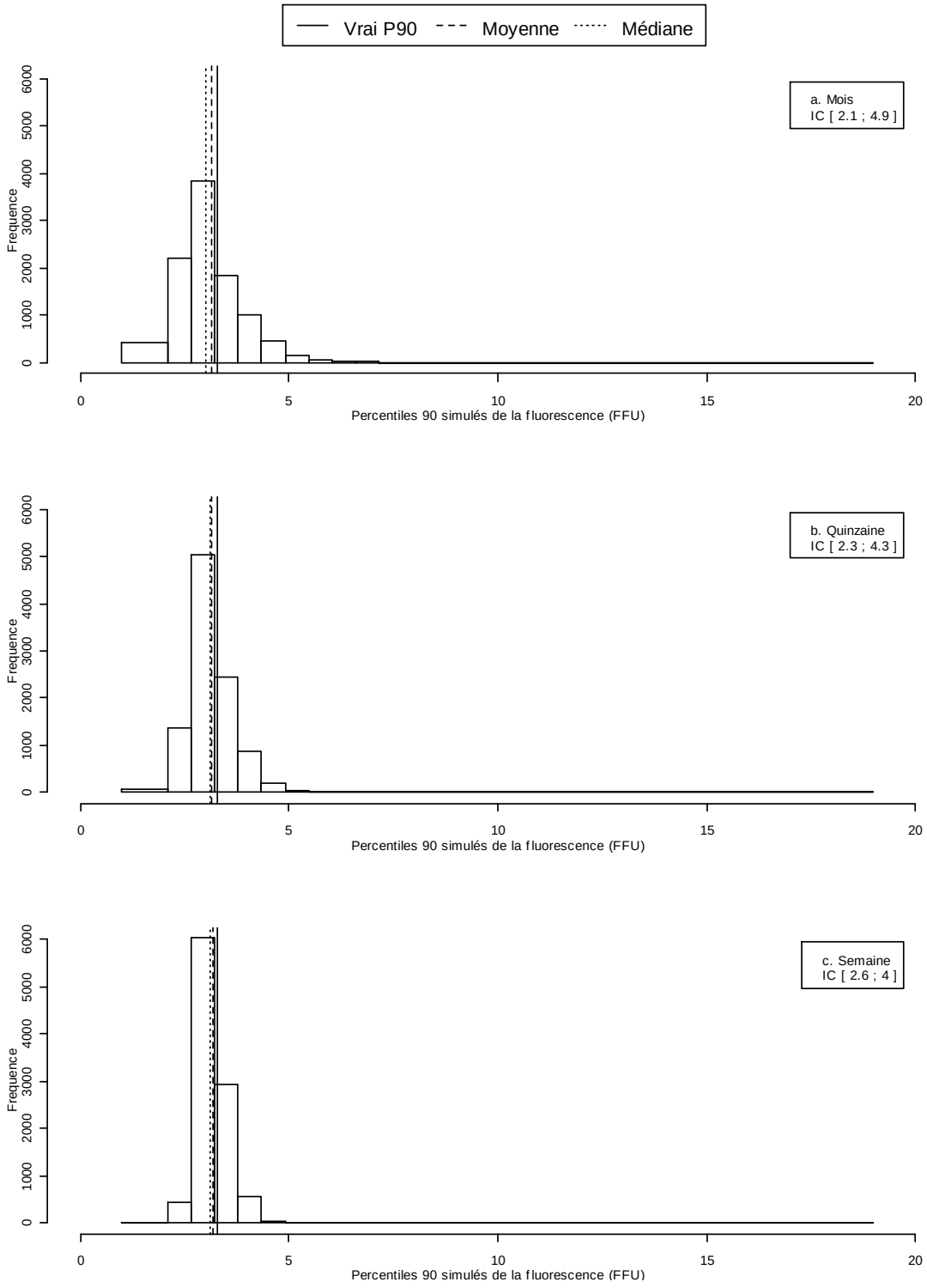


Figure 5 : Histogrammes des percentiles 90 de la fluorescence simulés pour un plan de gestion de 6 ans à partir des données diurnes des années 2005, 2006 et 2007 pour les fréquences d'échantillonnage a) mensuelle, b) bimensuelle et c) hebdomadaire. Moyennes, médianes et intervalles de confiance sont calculés avec les percentiles 90 simulés. Le « vrai P90 » est calculé avec les données diurnes des trois années considérées.

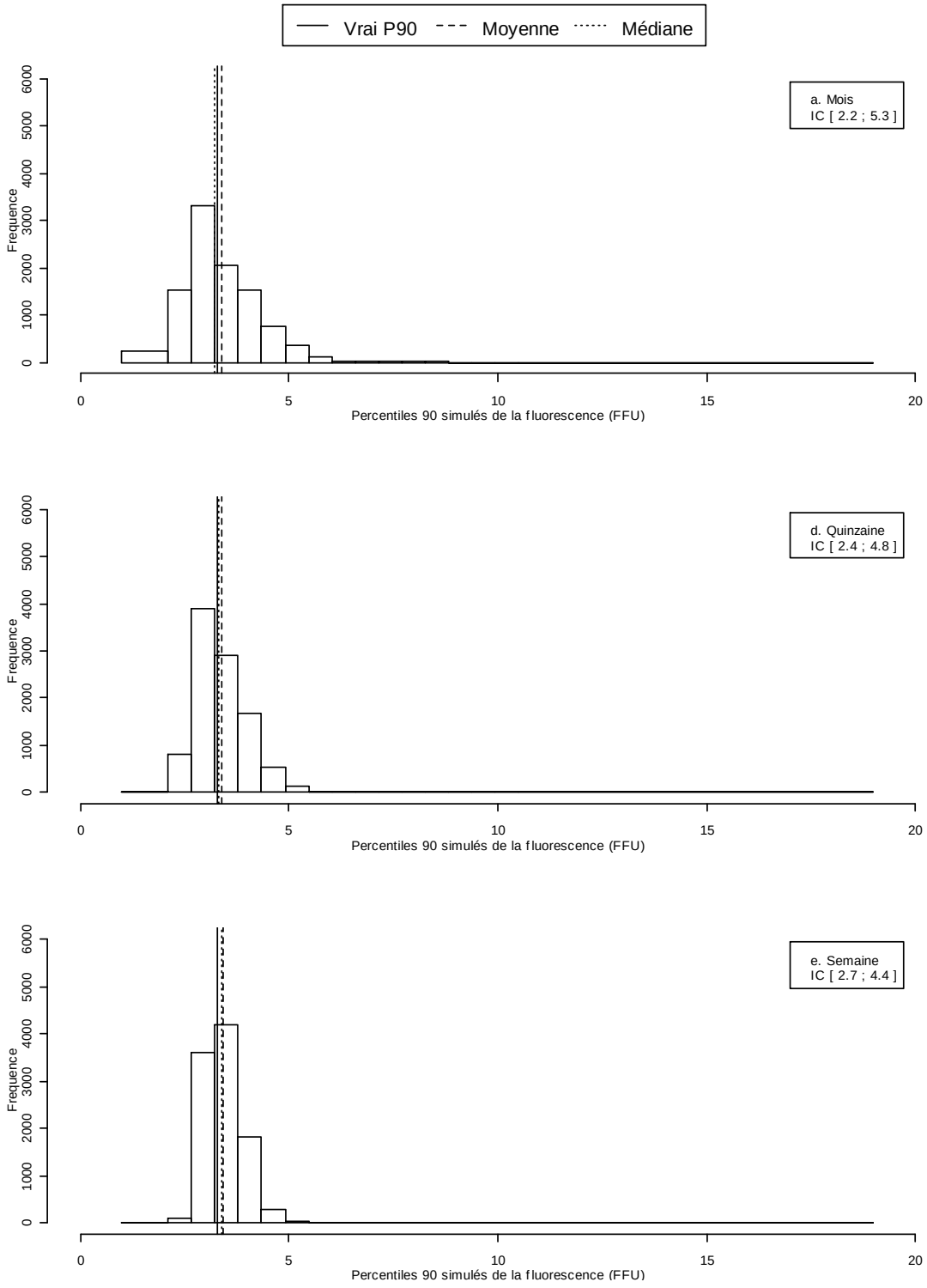


Figure 6 : Histogrammes des percentiles 90 de la fluorescence simulés pour un plan de gestion de 6 ans à partir des données des années 2005, 2006 et 2007, diurnes et opérationnelles pour les fréquences d'échantillonnage a) mensuelle, b) bimensuelle et c) hebdomadaire. Moyennes, médianes et intervalles de confiance sont calculés avec les percentiles 90 simulés. Le « vrai P90 » est calculé avec les données diurnes des trois années considérées.

La figure 7 présente la distribution du percentile 90 avec les données diurnes de mars à octobre. Quelque soit la fréquence d'échantillonnage, le « vrai » percentile 90 est inférieur aux moyennes, médianes et modes des distributions. L'importance de l'écart au vrai P90 est encore plus marqué avec les fréquences F1 et F2, intermédiaires entre le mensuel et l'hebdomadaire. L'étendue semble suivre le même schéma qu'avec le filtre diurne : elle diminue avec l'augmentation de la fréquence d'échantillonnage. Il faut toutefois noter que l'étendue de l'intervalle de confiance pour la fréquence F1 est supérieure à celle de la fréquence mensuelle. Les mêmes observations peuvent être faites avec l'ajout du filtre opérationnel (cf. figure 8).

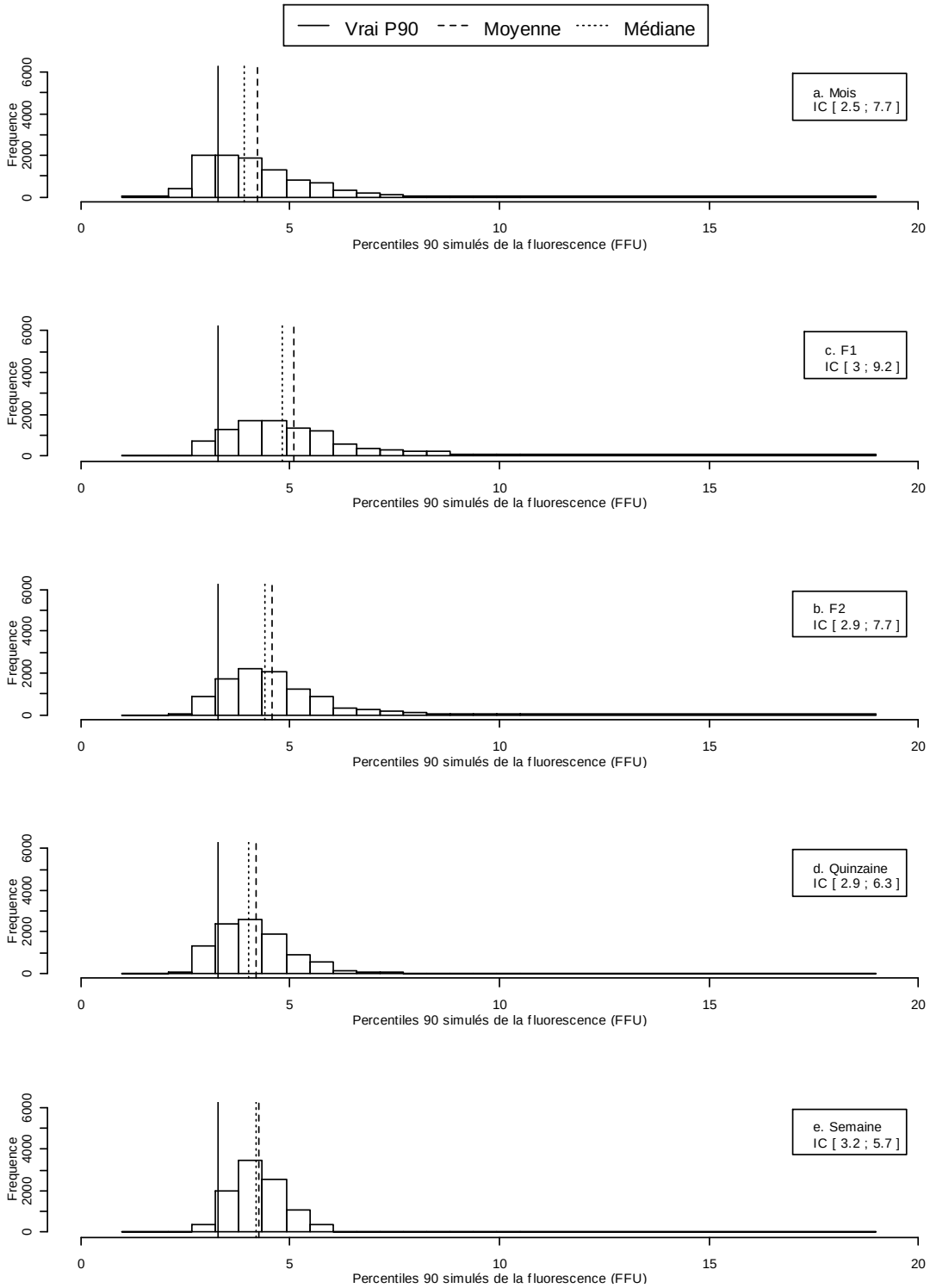


Figure 7 : Histogrammes des percentiles 90 de la fluorescence simulés pour un plan de gestion de 6 ans à partir des données des années 2005, 2006 et 2007, diurnes, et de mars à octobre pour les fréquences d'échantillonnage a) mensuelle, b) F1, c) F2 d) bimensuelle et e) hebdomadaire. Moyennes, médianes et intervalles de confiance sont calculés avec les percentiles 90 simulés. Le « vrai P90 » est calculé avec les données diurnes des trois années considérées.

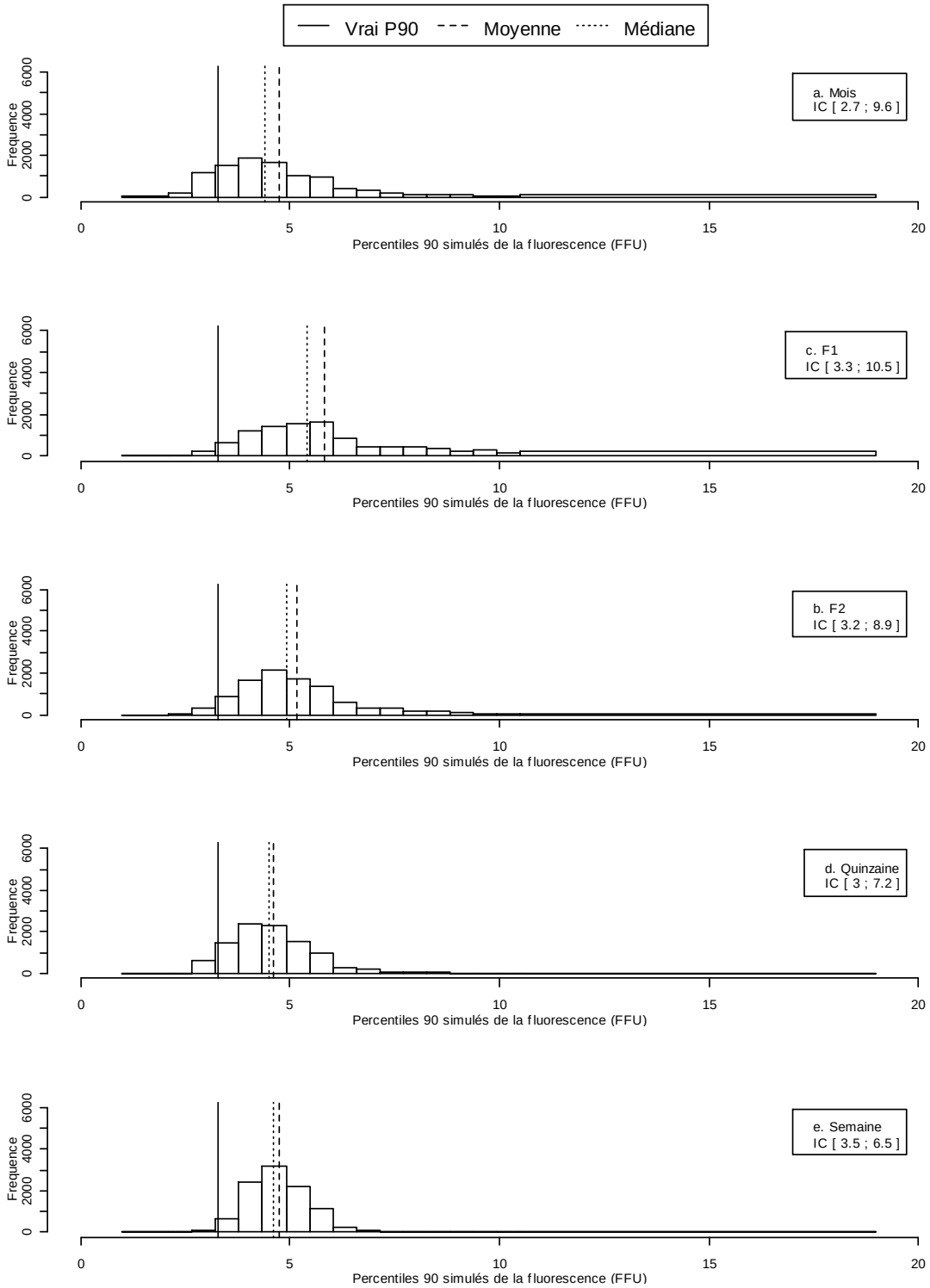


Figure 8 : Histogrammes des percentiles 90 de la fluorescence simulés pour un plan de gestion de 6 ans à partir des données des années 2005, 2006 et 2007, diurnes, opérationnelles et de mars à octobre pour les fréquences d'échantillonnage a) mensuelle, b) F1, c) F2 d) bimensuelle et e) hebdomadaire. Moyennes, médianes et intervalles de confiance sont calculés avec les percentiles 90 simulés. Le « vrai P90 » est calculé avec les données diurnes des trois années considérées.

5 Discussion

La chlorophylle *a* est un paramètre qui a été choisi dans le cadre de la DCE pour apprécier la biomasse phytoplanctonique. La métrique retenue est le percentile 90 des mesures effectuées pendant 6 années, de mars à octobre à raison d'un prélèvement par mois. Les données de chlorophylle *a* disponibles ne permettent pas de tester l'influence de la fréquence d'échantillonnage sur le calcul du percentile 90. En revanche, de telles données sont disponibles pour la fluorescence qui est un traceur de la chlorophylle *a*. Des simulations ont permis d'évaluer l'influence de la période et de la fréquence d'échantillonnage sur la distribution du percentile 90 de la fluorescence.

Les simulations sur les données diurnes des années prises séparément ont montré une sous estimation systématique du vrai P90 calculé sur l'ensemble des données (figures 2, 3 et 4). L'augmentation de la fréquence d'échantillonnage semble diminuer le biais (tableau 3). Il est très diminué lorsque l'on considère les simulations sur 6 ans (figure 5), de même que sa réduction avec l'augmentation de la fréquence d'échantillonnage (tableau 4). Ces effets sont attribuables tout à la fois au mélange des années atténuant les particularités de chacune et au nombre de mesures utilisées (*i.e.* 6 par strate contre 1 par strate) qui augmente la probabilité de tirer des données pendant les courtes périodes de floraison. L'utilisation de 6 années de mesures contribue donc à une meilleure estimation du P90.

L'utilisation des seules données opérationnellement observables conduit à un biais toujours faible (figure 6), plutôt un peu supérieur au vrai P90 et sans diminution nette avec l'augmentation de la fréquence d'échantillonnage (tableau 4). En revanche, le filtrage sur les mois de mars à octobre, c'est à dire la non prise en compte des faibles valeurs hivernales, conduit toujours à une surestimation du P90 (figure 7), renforcée avec le filtre opérationnel (figure 8 et tableau 4). Il s'ensuit que la limitation à la période mars-octobre, période pendant laquelle se produisent les blooms phytoplanctoniques, induit un biais dans l'estimation du P90 de la fluorescence.

Globalement, plus la fréquence d'échantillonnage est élevée, plus la variabilité du percentile 90 est limitée. Cette affirmation est prise en défaut pour la fréquence F1, intermédiaire entre le mensuel et l'hebdomadaire : les étendues des intervalles de confiance sont supérieurs à ceux de la fréquence mensuelle (tableau 4). De plus, F1 et F2 induisent un biais supplémentaire dans les estimations (figures 7 et 8). Ces fréquences mises à part, l'augmentation de la fréquence d'échantillonnage ne semble pas modifier les valeurs des paramètres de position des distributions des P90 simulés. Ainsi, sur la période mars-octobre, resserrer les mesures pendant les périodes théoriques d'efflorescences, ajoute un biais et peut augmenter la variabilité des P90 de la fluorescence.
Finalement, sur les simulations de 6 années de mesures, les fréquences F1 et F2 mises à part, l'estimation la plus biaisée et la plus variable est obtenue pour

les données diurnes opérationnelles limitées à la période mars-octobre avec une fréquence d'échantillonnage mensuelle. L'utilisation des données diurnes et opérationnelles toute l'année avec une fréquence mensuelle conduit à une estimation apparemment sans biais et d'une variabilité deux fois moindre au regard des étendues des intervalles de confiance. Il est à noter que ce résultat est obtenu au prix de 24 mesures supplémentaires sur un plan de gestion de 6 ans. Toutefois, la période hivernale présentant généralement des faibles valeurs de fluorescence, il serait envisageable de ne faire qu'une mesure en hiver et de lui donner un poids quatre fois supérieur dans le calcul de la métrique. On augmenterait alors que de six valeurs le total des mesures à réaliser. En revanche, la même procédure ne peut être appliquée à la période estivale : les masses d'eau présentent alors une importante variabilité et peuvent s'écarter largement du schéma théorique.

La portée de ces observations doit être limitée par les considérations suivantes. Cette étude constitue une tentative d'approcher les variations de la chlorophylle *a* à travers celles de la fluorescence. En dépit du lien qui existe entre ces deux variables, il serait dangereux de tirer ici des conclusions définitives concernant la chlorophylle *a*. Par ailleurs, les résultats obtenus à Boulogne ne peuvent être généralisables à l'ensemble du littoral français. D'autres systèmes de mesures MAREL sont en usage en métropole. Leurs données pourraient être analysées de manière identique. Il serait alors possible de comparer les résultats obtenus selon leur localisation. D'une manière plus générale, le développement des estimations de la concentration de la chlorophylle *a* par traitement d'images satellitaires permettrait la généralisation de la simulation à l'ensemble des masses d'eaux côtières.

Enfin, certains points techniques de la simulation nécessitent d'être examinés. Une des conditions d'application du *bootstrap* est l'utilisation de données indépendantes et identiquement distribuées. Cette condition n'est pas respectée dans le cas des séries temporelles. Cependant, Davison et Hinkley (1997) proposent plusieurs solutions techniques pour résoudre ce problème. Les données des années présentent des profils très marqués. De la même manière que la saisonnalité du phénomène devait être prise en compte par l'utilisation de strates, on peut se demander dans quelle mesure la spécificité des années ne devrait pas être conservée. Autrement dit, dans quelle mesure le facteur année devrait être inclus dans la stratification. Dans cette hypothèse, à défaut de disposer de 6 années complètes de mesures pour simuler un plan de gestion de 6 ans, on pourrait tirer deux fois plus de données dans les trois strates d'un an. Pour finir, la période de 7 à 19 heures (GMT/UTC) choisie pour identifier les données diurnes n'est peut être pas adaptée. La plage de 9 à 16 heures des 90% des prélèvements mélange heures légales d'été et d'hiver. Il est apparu *a posteriori* qu'elles ont été traitées comme des heures en temps universel. L'identification des périodes « marée haute plus ou moins deux heures » procède par recherche du maximum de la hauteur d'eau sur la tranche 9-16 heures. Dans certains cas, cela peut conduire à la conservation de données situées hors de l'intervalle cible. Ainsi, il serait sans doute

souhaitable d'ajuster plus finement certains de ces paramètres et plus généralement de procéder à une étude de sensibilité.

6 Conclusion

La limitation à la seule période mars-octobre induit un biais dans l'estimation du P90 de la fluorescence. Plus la fréquence d'échantillonnage est élevée, plus la variabilité du percentile 90 est limitée. L'estimation la plus biaisée et la plus variable est obtenue pour les données diurnes opérationnelles limitées à la période mars-octobre avec une fréquence d'échantillonnage mensuelle. L'utilisation des données diurnes et opérationnelles toute l'année avec une fréquence mensuelle conduit à une estimation de meilleure qualité. Une seule mesure en hiver dotée d'un poids de 4 permettrait de limiter à 6 le nombre de mesures supplémentaires par rapport au cahier des charges actuel de la DCE.

Il est important de garder à l'esprit que ces conclusions concernent la fluorescence. La spécificité géographique de l'origine des données ne permet pas l'extrapolation à l'ensemble du littoral. Il existe toutefois d'autres sources de données de fluorescence (*e.g.* autres bouées MAREL) ou de chlorophylle a (*e.g.* données satellitaires) qui pourraient être traitées de manière similaire. Cette étude a permis de jeter les bases d'une approche et d'un outil réutilisables. Toutefois, ils nécessitent ajustements et compléments d'études.

Annexe 1 Programmation de la simulation par booststrap

1.1 Package *boot*

Cette approche ne permet pas dans notre cas de simuler 10 000 échantillons.

```
# Nom               : IFR.boot.r
# Type              : Macros
# Objet             : Bootstrap avec un nombre d'échantillonnage différent de
#                     l'échantillon initial
# Input             :
#
#
# Output            : Aucun
# Auteur            : ASoudant
# R version         : 2.4.1
# Date de création  : 16 JUIL 2008
#_____________________________________________________________


IFR.boot <- function(x,strate,n,Test=FALSE){

  Métrique <- function(x,i,n=length(x),Test=FALSE){
    x[order(strate),]
    x00 <- cbind(x,i)
    met <- by(x00,x00$mois,function(y) y[1:n,"i"])
     if (Test){
        cat("\nSummary de la métrique\n")
        print(summary(met))
        cat("\npremières lignes de la métrique\n")
        print(head(met))
     }
    # Sélection des n premières valeurs
    x01 <- x00[unlist(met),]

    # Calcul de la métrique
    P90 <- round(quantile(as.numeric(x01$Fluo)
                          ,probs=0.9
                          ,names=FALSE
                          ,type=4
                          ,na.rm= TRUE)
                 ,digits=1)

      if(Test){
        cat("\nSummary du P90\n")
        print(summary(P90))
        }

    return(P90)
  }
  if(Test) x.boot <- boot(x, Métrique, R = 10,strata=strate, n=n,Test=Test)
  else x.boot <- boot(x, Métrique, R = 1000,strata=strate, n=n,Test=Test)
  P90 <- x.boot$t

  boot_ci <- boot.ci(x.boot,type ="perc")
  boot_ci

  ci <- boot_ci$percent[1,4:5]
  Objet <- list()
  Objet$P90 <- P90
  Objet$ci <- ci
  return(Objet)
  rm(x.boot)
  rm(boot_ci)

}
```

1.2 Fonction *sample*

Cette approche permet la simulation de 10 000 échantillons et nécessite la duplication de deux fonctions du package *boot* (cf. `IFR.perc.ci()` et `IFR.norm.inter()`).

```
IFR.boot <- function(DataIn,Strate,n,Test=FALSE){
  Simulations <- numeric(10000)
  for(i in 1:length(Simulations)){
    Simulation <- unlist(by(DataIn$Fluo,Strate,sample,size=n,replace=TRUE))
    P90 <- round(quantile(Simulation
                          ,probs = 0.9
                          ,names = FALSE
                          ,type = 4
                          , na.rm = TRUE)
                 , digits = 1)
    Simulations[i] <- P90
  }

  CI <- IFR.perc.ci(Simulations)

  Objet <- list()
  Objet$P90 <- Simulations
  Objet$ci <- CI[1,4:5]
  return(Objet)
}

IFR.perc.ci <- function (t, conf = 0.95, hinv = function(t) t)
{
    alpha <- (1 + c(-conf, conf))/2
    qq <- IFR.norm.inter(t, alpha)
    out <- cbind(conf, matrix(qq[, 1], ncol = 2), matrix(hinv(qq[,
        2]), ncol = 2))
    out
}

IFR.norm.inter <- function (t, alpha)
{
    t <- t[is.finite(t)]
    R <- length(t)
    rk <- (R + 1) * alpha
    if (!all(rk > 1 & rk < R))
        warning("extreme order statistics used as endpoints")
    k <- trunc(rk)
    inds <- 1:length(k)
    out <- inds
    kvs <- k[k > 0 & k < R]
    tstar <- sort(t, partial = sort(union(c(1, R), c(kvs, kvs +
        1))))
    ints <- (k == rk)
    if (any(ints))
        out[inds[ints]] <- tstar[k[inds[ints]]]
    out[k == 0] <- tstar[1]
    out[k == R] <- tstar[R]
    not <- function(v) xor(rep(TRUE, length(v)), v)
    temp <- inds[not(ints) & k != 0 & k != R]
    temp1 <- qnorm(alpha[temp])
    temp2 <- qnorm(k[temp]/(R + 1))
    temp3 <- qnorm((k[temp] + 1)/(R + 1))
    tk <- tstar[k[temp]]
    tk1 <- tstar[k[temp] + 1]
    out[temp] <- tk + (temp1 - temp2)/(temp3 - temp2) * (tk1 -
        tk)
    cbind(round(rk, 2), out)
}
```

Annexe 2 Architecture de développement

La présentation ci-après est extraite du site intranet :

http://w3.ifremer.fr/surveillance/aurige/R_Archi_Stockage.htm

A un projet correspond un et un seul répertoire. Dans celui-ci (*cf.* modèle) nous allons trouver les dossiers suivants :

1. Documents
2. Original data sets
3. Programs
4. Macros
5. Log
6. Derived data sets
7. Out

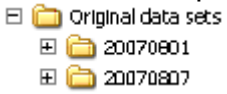
A ce premier répertoire s'ajoute un répertoire de macros globales transversales aux projets. Il ne faut pas confondre ce dernier avec celui listé au point 4 qui ne contiendra que les codes spécifiques à un projet.

Documents

On y range l'ensemble des documents relatifs à l'étude. Cela peut être des cahiers des charges, des spécifications générales et/ou détaillées, des exemples de sorties, des rapports de références, de la bibliographie, les différentes versions du rapport/article du projet etc. L'organisation interne de ce répertoire est à la discrétion de chacun.

Original data sets

Les données originales du projet sont stockées dans ce dossier. Ce sont, par exemple, les fichiers issus d'une extraction d'une base de données ou bien les fichiers résultant d'une saisie. Il est très important de conserver ces données brutes et exemptes de toutes modifications de manière à pouvoir y revenir à tout moment de l'étude. Si une nouvelle extraction venait a être réalisée, des données supplémentaires saisies ou des corrections portées aux données alors non seulement il faut conserver la trace de ces changements mais en plus il convient de garder les données originales avant ces changements et/ou ajouts. De ce fait les données sont stockées dans un sous répertoire au nom explicite. Si des changements interviennent alors il suffit de créer un nouveau sous répertoire et d'y enregistrer les nouvelles données. A titre de suggestion, le nom de ces sous répertoires peuvent être la date de création du répertoire au format "aaaammjj".



Il est souhaitable de conserver les données originales dans leur format natif (*e.g.* Excel) mais de les enregistrer également au format texte avec la tabulation comme séparateur. Ce dernier format est en

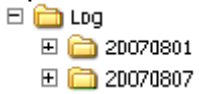
effet très portable d'un système d'exploitation à un autre et généralement reconnu par la majorité des logiciels et langages de programmation.
Il est a noter que l'archivage des données dans un sous-dossier de Original data sets est un élément structurant fort et utilisé par l'architecture de programmation et se trouve ainsi, obligatoire.

Programs et Macros

Ces deux répertoires vont accueillir l'ensemble du code relatif au projet. Dans Programs on range les programmes (*cf.* programmes principaux et programmes simples) et dans Macros on range des fichiers qui contiennent une à plusieurs fonctions de portée générale et utilisés dans plusieurs programmes (*cf.* macros). La section programmation détaille cette organisation.

Log

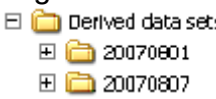
Ce répertoire contiendra les traces d'exécution des programmes. On appelle trace d'exécution d'un programme un fichier contenant le code du programme exécuté, les éventuels *warnings* signalés par R, s'il y a une erreur, le message de l'erreur rencontrée et le cas échéant des messages programmés (exemple). Les fichiers traces portent le nom du programme dont ils contiennent la trace : Data_Flores.r.log.txt est la trace de l'exécution du programme Data_Flores.r. Les fichiers traces sont rangés dans un sous répertoires dont le nom est celui des données utilisées.



Ainsi les traces d'exécution des programmes utilisant les données du répertoire Original data sets/20070801 sont dans le répertoire Log/20070801. Les sous-répertoires sont créés de manière automatique.

Derived data sets

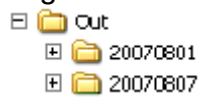
Les ensembles de données de travail dérivés des données originales viendront s'archiver dans ce répertoire. Ces data sets sont dits dérivés parce qu'ils peuvent concerner des sous-échantillons et/ou posséder des variables dérivées des variables originales et/ou présenter une structure différente (*e.g.* par la clé composée de plusieurs variables identifiant de manière unique un individu statistique). Comme ces fichiers sont directement liés aux données originales, ils sont archivés dans des sous répertoires créés automatiquement et portant le même nom que celui des données originales utilisées.



Out

The last but not the least, le répertoire Out est destiné à recevoir les résultats des programmes exploitant les derived data sets. Ces résultats peuvent être par exemple des listings de données, des tables de statistiques descriptives ou inférentielles, des graphiques.

Comme précédemment, ces résultats sont stockés dans des répertoires créés automatiquement dont le nom identifie les données originales dont ils sont issus.



Références bibliographiques

Aminot A., 2001. Chlorophyll *a* : Determination by spectroscopic methods. ICES Techniques In Marine Environmental Sciences No. 30.

Anonyme, 2008. Cahier de Procédures et de Programmation REPHY 2008. Date d'application : 25 février 2008. Document de prescription. 67 pages.

Davidson A., Hinkley D.V., 1997. Bootstrap Methods and Their Application. Cambridge University Press.

Directive Cadre sur l'Eau. Directive 2000/60/CE du 23 octobre 2000.

Herbland, Voituriez, 1977. Relations Chlorophylle *a* – Fluorescence *in vivo* dans l'atlantique tropical influence de la structure hydrologique. Cah O.R.ST.O.M., sér. Océanogr., vol. 15, n°1, 1977 : 67-77.

Lefebvre A., 2008. MAREL Carnot : Rapport n°3 : Valorisation des données d'une surveillance à haute fréquence en zone côtière sous influence anthropique (Boulogne-sur-Mer). Bilan de l'année 2007. Ifremer/RST.LER.BL/08.04, 23 pages.

Le Goff R., Nogues L., Lampert L. et Riou P.. Réseau Hydrologique Littoral Normand : Rapport 2007, volume 1. RST. LERN - 07.14.

Pellouin-Grouhel A., Belin C. & Daniel A., novembre 2006. Recommandations techniques pour le contrôle de surveillance dans le cadre de la DCE, pour le phytoplancton et les paramètres physico-chimiques (hors contaminants chimiques). Stratégies d'échantillonnage, indicateurs, et grilles de classement. 52 pages.