Analyse statistique des facteurs environnementaux favorisant les efflorescences de cyanobactéries dans la retenue de Rophémel

Philippe Le Noac'H

AGROCAMPUS OUEST

☐ **CFR Angers**

☑ **CFR Rennes**

Année universitaire : 2017 - 2018 **Spécialité** : *Agronomie* **Spécialisation (et option éventuelle)** : *Data Science pour la Biologie*	**Mémoire de fin d'études** ☐ d'Ingénieur de l'Institut Supérieur des Sciences agronomiques, agroalimentaires, horticoles et du paysage ☑ de Master de l'Institut Supérieur des Sciences agronomiques, agroalimentaires, horticoles et du paysage ☐ d'un autre établissement (étudiant arrivé en M2)

Analyse statistique des facteurs environnementaux favorisant les efflorescences de cyanobactéries dans la retenue de Rophémel

Philippe LE NOAC'H

Soutenu à Rennes le 3 septembre 2018

Devant le jury composé de :

Président : **David CAUSEUR** (Pr Agrocampus Ouest)
Maîtres de stage : **Yvan LAGADEUC** (Pr Université de Rennes 1)
Alexandrine PANNARD (MC Université de Rennes 1)
Enseignant référent : **David CAUSEUR**

Autres membres du jury (Nom, Qualité)
Mathieu EMILY (MC Agrocampus Ouest)

Les analyses et les conclusions de ce travail d'étudiant n'engagent que la responsabilité de son auteur et non celle d'AGROCAMPUS OUEST

Fiche de confidentialité et de diffusion du mémoire

Confidentialité

☐ Non ☐ Oui si oui : ☐ 1 an ☐ 5 ans ☐ 10 ans

Pendant toute la durée de confidentialité, aucune diffusion du mémoire n'est possible [1].

Date et signature du **maître de stage** [2] :
(ou de l'étudiant-entrepreneur)

A la fin de la période de confidentialité, sa diffusion est soumise aux règles ci-dessous (droits d'auteur et autorisation de diffusion par l'enseignant à renseigner).

Droits d'auteur

L'auteur[3] **LE NOAC'H Philippe**
autorise la diffusion de son travail (immédiatement ou à la fin de la période de confidentialité)

☐ Oui ☐ Non

Si oui, il autorise

☐ la diffusion papier du mémoire uniquement(4)

☐ la diffusion papier du mémoire et la diffusion électronique du résumé

☐ la diffusion papier et électronique du mémoire (joindre dans ce cas la fiche de conformité du mémoire numérique et le contrat de diffusion)

(Facultatif) ☐ accepte de placer son mémoire sous licence Creative commons CC-By-Nc-Nd (voir Guide du mémoire Chap 1.4 page 6)

Date et signature de l'**auteur** :

Autorisation de diffusion par le responsable de spécialisation ou son représentant

L'enseignant juge le mémoire de qualité suffisante pour être diffusé (immédiatement ou à la fin de la période de confidentialité)

☐ Oui ☐ Non

Si non, seul le titre du mémoire apparaîtra dans les bases de données.
Si oui, il autorise

☐ la diffusion papier du mémoire uniquement(4)

☐ la diffusion papier du mémoire et la diffusion électronique du résumé

☐ la diffusion papier et électronique du mémoire

Date et signature de l'**enseignant** :

Sommaire

Introduction 1

1. Contexte écologique : phytoplancton et traits fonctionnels 1

2. Blooms 2

3. Phytoplancton et production d'eau potable 3

4. Cadre et problématique du stage 4

Matériel & Méthodes 6

1. Présentation et mise en forme des jeux de données 6

1.a. Données de comptage du phytoplancton 6

1.b. Variables physico-chimiques 7

1.c. Variables météorologiques et hydrologiques 8

2. Présentation des analyses statistiques 9

2.a. Analyse multivariée à deux tableaux : RDA 9

2.b. Analyses à 3 tableaux : fourth-corner analysis 10

2.c. Modèles de prédiction : Machine Learning 13

Résultats 14

1. Caractérisation des événements de blooms 14

2. Abondance spécifique et variables environnementales 15

3. Passage de la classification linnéenne à la notion de traits fonctionnels 16

4. Prédiction des évènements de blooms 19

Discussion 22

1. Identification des facteurs susceptibles de favoriser l'apparition des blooms 22

2. Evaluation des effets potentiels d'une stratégie de gestion 24

3. Difficultés à mettre en place des modèles de prédiction (Machine Learning) valides 25

4. Pertinence de l'utilisation des données publiques d'agence pour une l'étude d'une communauté phytoplanctonique et suggestions d'amélioration de l'échantillonnage 26

Bibliographie 29

Annexe I A1

Annexe II A2

Annexe III A3

Annexe IV A5

Annexe V A7

Annexe VI A8

Introduction

1. Contexte écologique : phytoplancton et traits fonctionnels

Le terme phytoplancton désigne l'ensemble des micro-organismes aquatiques photo-autotrophes vivant en suspension dans la colonne d'eau des océans, des cours d'eau et des réservoirs d'eau douce de la planète. Ces organismes sont parfois désignés sous le terme de microalgues. Pour se développer, le phytoplancton absorbe les nutriments dissous dans leur milieu (azote, phosphore, silice, etc.). Le phytoplancton utilise aussi l'énergie lumineuse pour sa croissance, grâce au processus de la photosynthèse. Il existe un grand nombre de genres et d'espèces de phytoplancton, de tailles et de formes très diverses (fig. 1).

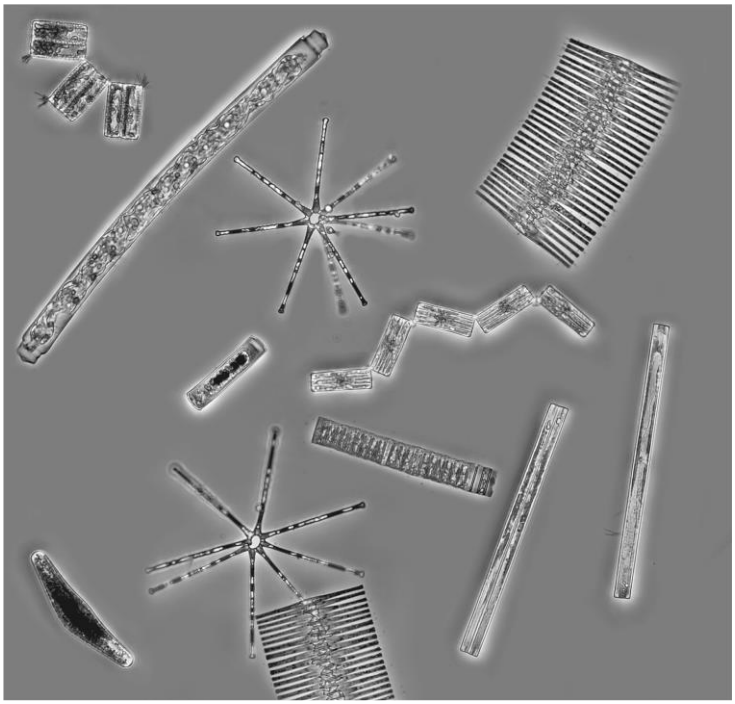

Figure 1 : Illustration de la diversité des espèces de diatomées, un des groupes dominants du phytoplancton (tiré de Edwards et al., 2013a).

De manière générale, toutes ces espèces ont des besoins physiologiques similaires, car elles appartiennent au même groupe fonctionnel (Reynolds, 2002). Par conséquence, elles partagent une même niche écologique et sont en compétition pour l'exploitation des mêmes ressources (Tilman, 1977). En théorie, le nombre d'espèces capables de coexister dans un plan d'eau devrait donc être relativement restreint. Mais malgré l'intensité de la compétition pour la ressource, les milieux aquatiques naturels abritent simultanément un grand nombre d'espèces phytoplanctoniques qui forment alors une communauté très diversifiée. Ce décalage entre la théorie et la réalité du terrain est connu sous le nom de Paradoxe du Plancton (Hutchinson, 1961). De nombreuses hypothèses ont été proposées au fil des années pour résoudre ce paradoxe, notamment l'hétérogénéité spatial et temporel de la disponibilité en ressources (Bracco et al., 2000).

L'étude des communautés, et notamment l'étude des communautés phytoplanctoniques, se fait de plus en plus par le prisme des traits fonctionnels (McGill et al., 2006 ; Litchman et al., 2007). Un trait fonctionnel est défini comme une caractéristique physio-morphologique qui va impacter de manière indirecte la *valeur adaptative* d'un organisme (c'est-à-dire sa capacité à transmettre ses gènes) par ses effets sur la croissance, la reproduction ou la survie (Violle et al., 2007). Dans le cas du phytoplancton, les valeurs des traits peuvent refléter différentes stratégies de consommation de la ressource. Par exemple, dans le cas des nutriments, certaines espèces vont plutôt maximiser la quantité de ressource absorbée par unité de temps (*Affinité).* Au contraire, d'autres espèces vont plutôt maximiser leur capacité de stockage des nutriments (*Réserve*) (Edwards et al., 2013b). D'autres traits, notamment morphologiques, affectent la capacité du phytoplancton à résister à la prédation ou encore à capter la lumière (Litchman & Klausmeier, 2008).
L'étude des traits fonctionnels fournit des éléments de réponse pour expliquer la diversité des communautés présentent dans les milieux naturels. Les traits fonctionnels fournissent un cadre d'étude plus adapté pour comprendre le lien entre composition de la communauté phytoplanctonique et variables environnementales que la classification linnéenne classique.

2. Blooms

A la suite d'un changement environnemental comme des apports récurrents d'origine anthropique de nutriments dans le milieu, une espèce plus compétitive que les autres peut être avantagée par rapport au reste de la communauté. Si la quantité de ressources est suffisante, la biomasse de cette espèce peut augmenter de façon exponentielle, formant une efflorescence (fig. 2). Selon l'OMS, on peut considérer qu'un plan d'eau est touché par une efflorescence lorsque la concentration en cellules phytoplanctoniques dépasse les 10^6 cellules par mL. Le phytoplancton appartenant au groupe taxonomique des cyanobactéries est particulièrement susceptible de former ces efflorescences (aussi appelées blooms), notamment dans les systèmes dulcicoles (Paerl et al., 2001).
Ce phénomène a des répercussions importantes sur le reste de l'écosystème : les blooms de cyanobactéries sont associés à une importante dégradation de la qualité de l'eau, notamment à cause de la production de cyanotoxines (Merel et al., 2013).
Les blooms algaux sont un enjeu de société de plus en plus important, au centre des politiques de gestion environnementale des écosystèmes aquatiques. Lors d'une efflorescence, la concentration en cyanotoxines peut augmenter dramatiquement dans le milieu (Sivonen, 1996). La production accrue de matière organique peut aussi se révéler problématique (Paerl et al., 1998). Cela a un impact direct sur les compartiments trophiques supérieur, les blooms pouvant causer une hausse de la mortalité piscicole (Marie et al., 2012). Les usages anthropiques sont aussi concernés : la réglementation impose une interdiction de la baignade dans les plans d'eau dont la concentration en cyanotoxines dépasse un certain seuil. Enfin, les blooms constituent une problématique majeure pour la production d'eau potable : la ressource doit subir des traitements supplémentaires pour gérer mitiger les effets des cyanotoxines et de la quantité accrue de matière organique (Hitzfeld et al., 2000 ; Quin et al., 2010).
La gestion des efflorescences de cyanobactéries devrait continuer à gagner en importance à l'avenir, leur fréquence ayant augmenté ces dernières décennies (Hallegraeff, 1993 ; Michalak et al., 2013). L'eutrophication des cours d'eau, liée à l'intensification des activités agricoles et industrielles, est l'un des principaux facteurs explicatifs de cette tendance (O'Neil et al., 2012).

Cette tendance devrait se poursuivre à l'avenir, notamment à cause du changement climatique (Paerl & Huisman, 2009).

Figure 2 : image satellite du lac Erie durant un bloom de *Microcystis*, en 2011 (image tirée de Steffen et al., 2014).

3. Phytoplancton et production d'eau potable

L'eau brute destinée à la consommation humaine peut être prélevée dans les nappes souterraines ou dans les eaux de surface (cours d'eau, lacs, réservoirs…). Dans les deux cas, la ressource est soumise à un processus de potabilisation. Grâce aux avancées technologiques et à la mise en place de nouvelles réglementations, les filières de potabilisation des eaux ont gagné en complexité ces dernières années afin de pouvoir traiter une gamme importante de molécules et d'éléments potentiellement dangereux pour la santé humaine : métaux lourds, pesticides, matières organiques, etc. (Legube & Mouchet, 2010).

Lorsque la ressource est captée en surface, les efflorescences phytoplanctoniques peuvent constituer un problème majeur pour les producteurs d'eau potables. La surcharge de matière organique est problématique lors de certaines étapes de la filière de potabilisation (en saturant les filtres par exemple). Plus gênant encore, l'augmentation de la concentration en cyanotoxines mais aussi de la quantité de matière organique dans le milieu oblige l'exploitant à procéder à des contrôles supplémentaires de la qualité de la ressource, voir à interrompre complétement le captage de l'eau brute jusqu'à disparition du bloom. Dans le meilleur des cas, cela engendre des surcoûts dans le processus d'exploitation. Dans le pire des cas, un bloom est susceptible de causer une crise de santé publique majeure (Falconer & Humpage, 2005).

Dans le contexte actuel de changement climatique et de raréfaction de la disponibilité de la ressource en eau potable (Hoque et al., 2016), la gestion des phénomènes de blooms constitue donc un enjeu financier et sanitaire important.

4. Cadre et problématique du stage

Notre étude porte sur la communauté phytoplanctonique de la retenue de Rophémel, dans les Côtes-d'Armor, à environ 30 km au Nord-Ouest de la ville de Rennes (Ille-Et-Vilaine).
La retenue de Rophémel est un réservoir d'eau brute alimenté par deux cours d'eau : le *Néal* et la *Rance amont* (fig. 3). Les superficies des bassins versants de ces deux affluents sont respectivement de 26 943 ha et 9 254 ha. L'existence du réservoir est dû à la présence d'un barrage (le barrage de Rophémel). Celui-ci, construit en 1930, est destiné à la production d'énergie (bien que la production hydro-électrique soit actuellement à l'arrêt). Le réservoir à une capacité maximale d'environ 5 millions de m^3. Actuellement, la principale utilité de la retenue de Rophémel est la production d'eau potable, grâce à la présence d'une usine de traitement de l'eau brute (construite en 1963 et modernisée en 2005). Cette usine fournit environ 40% des besoins de l'agglomération rennaise en eau potable, en traitant 7 à 10 millions de m^3 par an. La gestion de la retenue, du barrage et de l'usine de traitement est assurée par le syndicat de production et de distribution d'eau potable ***Collectivité Eau du Bassin Rennais***.

Figure 3 : Carte de la retenue de Rophémel (croix rouge) et des deux affluents alimentant le réservoir : la *Rance amont* (flèche bleue) et le *Néal* (flèche verte). La localisation de l'usine de traitement correspond au rond noir.

La retenue de Rophémel est régulièrement touchée par des événements d'efflorescences phytoplanctoniques depuis plus d'une décennie. Cela représente une nuisance importante pour l'exploitation de la retenue et la production d'eau potable. L'exploitant cherche donc à identifier les mécanismes écologiques et environnementaux responsables de l'apparition des blooms, afin de mettre en place des stratégies de gestion destinées à réduire leur fréquence, ou au moins de pouvoir anticiper de quelques jours les événements d'efflorescences de manière à mitiger les coûts d'exploitation.

Un suivi règlementaire de la communauté phytoplanctonique de la retenue est en place depuis plusieurs années déjà, avec identification des espèces présentes et mesures de leurs abondances respectives. De plus, dans le cadre du processus de production d'eau potable, l'exploitant est dans l'obligation légale de mesurer les variations de nombreuses variables physiques (turbidité, conductivité, température de l'eau, etc.) ainsi que les concentrations d'une grande variété de paramètres chimiques (nutriments, pesticides, métaux lourds, etc.) dans l'eau à traiter, c'est-à-dire l'eau de la retenue. Ces données biologiques et environnementales sont disponibles depuis 2006, formant ainsi plus de 10 ans de chroniques. Ces jeux de données peuvent être complétés par des chroniques de bases de données publiques, avec des variables hydrologiques (les débits des affluents alimentant la retenue) et météorologiques (insolation, précipitation, vent, etc.).

Toutes ces données nous ont été fournis par la *Collectivité Eau du Bassin Rennais*, le gestionnaire de la retenue de Rophémel. Nous nous proposons d'effectuer une analyse statistique poussée de ces données, pour d'une part étudier la relation entre composition de la communauté phytoplanctonique et variables environnementales, et d'autre part mettre en place un système de prédiction des événements de blooms. Plus globalement, nous souhaitons juger du potentiel d'utilisation des données à notre disposition, récolté dans le cadre de suivis réglementaires, pour une étude scientifique.
Après une première étape de traitement et de mise en forme des données, nous avons analysé les chroniques de comptages afin de caractériser les événements d'efflorescences. Nous avons ensuite effectué une canonique multivariée de type *RDA*, afin de mettre en évidence les liens entre variables biologiques et variables environnementales. Nous avons porté une attention particulière aux taxons influençant les espèces formant des efflorescences.
Dans un deuxième temps, nous avons voulu dépasser le cadre de la classification linnéenne classique, en étudiant les liens entre les traits fonctionnels de la communauté et les variables environnementales, pour comprendre plus précisément les mécanismes d'apparition des blooms. Pour cela, nous avons mis en place des méthodes de type *fourth-corner analysis*. Cette étape a nécessité une analyse bibliographique poussée afin de construire une matrice des traits fonctionnels pertinentes des taxons de la communauté.
Enfin, nous avons tenté de développer un modèle efficace de prédiction des événements de blooms par une approche de type *Machine Learning*.

Matériel & Méthodes

Toutes les analyses statistiques, ainsi que la phase de mise en forme des données, ont été réalisées à l'aide du logiciel libre R (R Core Team, 2017).

1. Présentation et mise en forme des jeux de données

Les données utilisées pour cette étude ont été récoltées entre janvier 2006 et décembre 2016 par le gestionnaire de la retenue de Rophémel. Ces dernières années, l'acquisition de ce type de données par des instances de gestion de l'environnement a fortement augmentée. Ce sont des données publiques, dont la récolte gérée par différents acteurs n'a pas forcément été pensée et planifiée dans un but de recherche scientifique. Cette étude est donc une occasion de voir s'il est possible de valoriser ces données, ou si elles comportent trop de biais pour être exploitées correctement.

1.a. Données de comptage du phytoplancton

Les comptages ont été effectués par des techniciens du laboratoire de recherche ECOBIO (Université de Rennes 1) jusqu'au 4 avril 2011, puis par un cabinet d'étude privé à partir de 2011 (LIMNOLOGIE SARL, Rennes). Au total, seul deux compteurs expérimentés différents ont produit l'ensemble des données de comptage du phytoplancton. Cela garantie une régularité de la qualité de ces données sur l'ensemble de la série. Les actes de comptage ont été réalisés conformément à la norme NF EN 15.204 (2006). Selon ce protocole, le compteur identifie au minimum les 400 premiers objets algaux de l'échantillon. Lorsque l'échantillon est particulièrement riche en phytoplancton (situation typique lors d'un bloom), le compteur peut choisir d'identifier un nombre plus important d'objets algaux (jusqu'à 4000). Malgré cette précaution, il faut être conscient que le comptage d'un échantillon récolté lors d'un bloom est fortement biaisé par l'espèce dominante, et risque de sous-évaluer la diversité réelle de la communauté.
On obtient finalement, pour chaque taxon phytoplanctonique identifié, une abondance mesurée en cellules par mL. Dans le meilleur des cas, l'identification du phytoplancton est réalisée jusqu'à l'espèce (comme pour *Pediastrum boryanum* par exemple). Très souvent néanmoins, on ne dispose que du genre de la classification linnéenne (ce qui explique la présence d'une variable *Pediastrum* dans le jeu de données). Si on constate dans le jeu de donnée la présence d'un taxon au niveau de l'espèce et un taxon au niveau du genre de cette espèce, on considère que les deux variables ne se recoupent pas. Par exemple, on considère que le taxon *Pediastrum* ne comprend aucun individu du taxon *Pediastrum boryanum*, qui est compté séparément.

Nous avons appliqué un filtre destiner à supprimer tous les taxons apparaissant à moins de 10 dates différentes. Cela permet de supprimer les espèces les plus rares, susceptibles d'avoir un poids disproportionné dans les analyses statistiques au regard de leur importance écologique. Il sera également difficile de relier leur abondance aux paramètres environnementaux lors des analyses factorielles.

Lorsqu'un taxon algal n'est pas mesuré à une date donnée, on considère qu'elle est absente du milieu à cette date et on lui attribue une abondance égale à 0. Nous nous sommes assurés qu'il n'y avait pas des doublons accidentels de variables liés à d'éventuelles erreurs dans la manière d'orthographier les noms des taxon phytoplanctoniques.

Finalement, le jeu de donnée de comptage utilisé dans la suite des analyses comprend 75 variables (taxons phytoplanctoniques, espèces et genres), mesurées à 279 dates. On constate que la fréquence d'échantillonnage de la communauté est très irrégulière (fig. 4). La majorité des échantillons comptés sont prélevés entre juin et novembre, ce qui correspond à l'époque de l'année durant laquelle les blooms sont susceptibles de se produire.

Figure 4 : illustration de la fréquence de comptage du phytoplancton. Chaque tiret vertical correspond à un événement de comptage. La fréquence d'échantillonnage est nettement plus importante entre juin et novembre.

La liste des taxons de la communauté est présentée en Annexe I. Selon les recommandations de l'OMS, nous avons considéré que la retenue est en état de bloom lorsque l'abondance d'une seule espèce dépasse 10^5 cell.mL^{-1}.

1.b. Variables physico-chimiques

Dans le cadre du contrôle de la qualité de l'eau, le gestionnaire de la retenue mesure un grand nombre de variables physico-chimiques lors des opérations de captage de la ressource. Cela inclue entre autres les concentrations en pesticides (Glyphosate, Atrazine…), en éléments chimiques potentiellement nocifs (aluminium, mercure, arsenic…) ou encore les concentrations en nutriments assimilables par le phytoplancton pour sa croissance (nitrates, ammonium, orthophosphates).

La fréquence de mesures varie selon les variables. Partant de ce constat, nous avons décidé de reconstituer des séries présentant une valeur par semaine pour obtenir un jeu de données homogène. Certaines variables peuvent facilement être mesurées en continue (température de l'eau, conductivité, turbidité…), et on dispose alors d'une mesure par jour sur l'ensemble de la série. Dans ce cas, on calcul des moyenne hebdomadaires (en agrégeant la série sur des plages successives de 7 jours).
Pour d'autres paramètres, des analyses poussées en laboratoire sont nécessaires, et l'on ne dispose alors que d'une mesure par mois. Dans ce cas, on considère que cette mesure est représentative de la semaine durant laquelle elle a été effectuée. Il devient alors nécessaire d'interpoler les valeurs manquantes pour les semaines où aucune mesure n'est réalisée. Deux méthodes d'interpolation ont été testées : l'interpolation *linéaire* et l'interpolation par *splines*. C'est finalement la méthode d'interpolation *linéaire* qui a été retenue, après avoir constaté que la méthode par *splines* produisait des artefacts d'interpolation aberrants.
Cette phase d'interpolation a pu être réalisée grâce au package *pastecs* (Grosjean et al., 2018).

Au terme de l'étape d'interpolation, on constate dans le jeu de données la présence de variables qui ne sont pas mesurées sur l'ensemble de la série. Cela est généralement dû à des changements de législation qui obligent le gestionnaire à mesurer une variable spécifique à compter d'une année, alors que cette mesure n'était pas obligatoire avant cela et n'était donc pas effectuée. On cherche néanmoins à obtenir des chroniques aussi complètes que possible. Par conséquence, nous avons fait le choix de supprimer les variables comportant des « trous » trop importants dans les chroniques. La taille de ces plages de données manquantes a été fixé à 50 mesures manquantes consécutives (ce qui correspond à une absence totale de mesure sur un une période d'environ 1 an).

Ce premier jeu de données environnementales comprend finalement 34 variables et 279 dates.

1.c. Variables météorologiques et hydrologiques

Pour compléter notre jeu de données de variables environnementales, nous avons acquis des données hydrologiques et météorologiques accessibles à partir de bases de données publiques.

A partir de la banque HYDRO, une base de données du ministère de l'Ecologie portée par la DREAL, nous avons obtenu des débits hydriques quotidiens moyens pour les deux principaux affluents du réservoir. Les mesures proviennent de deux stations en amont de la retenue de Rophémel : une station située sur la Rance (station DIREN N°J0611610, St Jouan de l'Isle), et une autre sur le Néal (station DIREN N°J0626610, Médréac).
Le gestionnaire du réservoir nous a aussi transmis les chroniques quotidiennes du volume d'eau stocké dans la retenue. On dispose aussi des données relatives aux volumes journaliers prélevés pour la production d'eau potable, ainsi que des volume relargué au niveau du barrage. En divisant le volume total stocké quotidiennement par le volume totale sortant du réservoir, on peut obtenir une estimation du temps de séjour de l'eau dans la retenue de Rophémel. Il faut néanmoins être conscient que ce calcul surestime le temps de séjour réel de l'eau dans la retenue : nous ne sommes en effet pas parvenus à estimer les volumes sortant de la retenue dans leur totalité. Nous n'avons notamment pas d'informations sur les volumes sortants du réservoir lorsque la retenue est en situation de trop-plein, ni les quantités perdues par évaporation. Cette variable composite est néanmoins ajoutée au jeu de variables environnementales.

Nous avons aussi obtenu des données météorologiques auprès de Météo France. On dispose ainsi de relevés quotidiens de températures moyennes et de précipitations mesurées à la station météorologique *Caulnes EDF*, située à une dizaine de kilomètres du barrage. A cela s'ajoute des données quotidiennes de rayonnement mesurées à l'Aéroport de Dinard, ainsi que des données quotidiennes de vitesse moyenne du vent mesurées à l'Aéroport de Rennes-St-Jacques.
On constate qu'une partie de ces données météorologiques provient de stations de mesure relativement éloignées de la retenue (entre autres les mesures relatives au vent provenant de l'aéroport de Rennes). En effet, il n'y a pas de station météorologique à proximité directe de la retenue, or le fort encaissement du barrage est susceptible d'entrainer un effet tunnel sur le vent. Cela pose question sur la pertinence de ces données météorologiques et constituera un élément de discussion dans la suite de ce rapport.

Ce second jeu de données environnementales comprend huit variables. En le combinant avec le jeu de données de variables physico-chimiques établie précédemment, on obtient une matrice de données environnementales de 42 variables (voir Annexe II).

2. Présentation des analyses statistiques

2.a. Analyse multivariée à deux tableaux : RDA

Dans un premier temps, on souhaite identifier des variables environnementales susceptibles de contrôler l'apparition et la dynamique des blooms de phytoplancton. Les analyses canoniques sous contrainte, de type RDA (*Redundancy Analysis*) ou CCA (*Canonical Contraint Analysis*), sont des outils adaptés pour répondre à cette problématique. Ce type d'analyses multivariés cherche à expliquer un premier tableau de données **Y**, typiquement des données d'abondances biologiques, par un second tableau de données **X**, typiquement des variables environnementales. On fait l'hypothèse que le second tableau explique le premier tableau, d'où le caractère *sous contrainte* de l'analyse (Legendre & Legendre, 1998). Les variables environnementales sont mesurées au même endroit et en même temps que les relevés des espèces. Les deux tableaux ont donc le même nombre de lignes, correspondant aux mêmes individus statistiques (fig. 5).

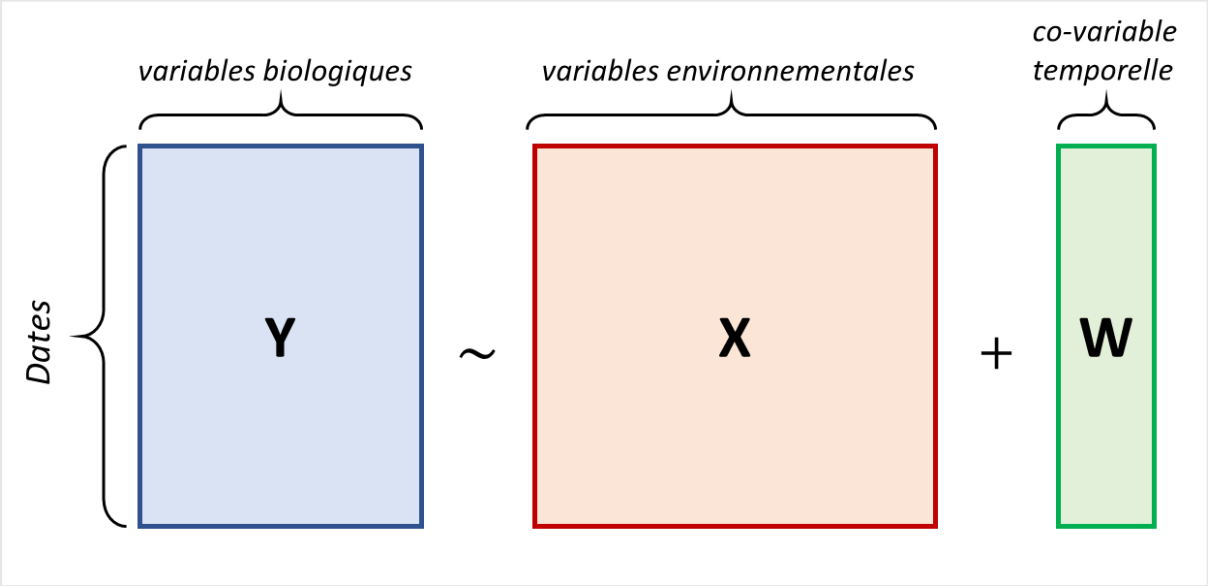

Figure 5 : Illustration du principe de l'analyse canonique partielle. On cherche à expliquer une matrice de données d'abondances **Y** par un tableau de variables environnementales **X**, tout en prenant en compte l'effet d'une (ou plusieurs) co-variable(s) (**W**).

Les analyses canoniques sont une extension du principe des analyses factoriels classiques (ACP dans le cas de la RDA et AFC dans le cas de la CCA), en incluant des éléments de régression multiple. L'objectif est de produire un nouveau plan, contraint par les variables environnementales explicatives, permettant de projeter à la fois les variables d'abondances et les variables environnementales. Les nouveaux axes de ce plan, analogues aux composantes produites en ACP ou en AFC, sont des combinaisons linéaires des variables de **X**.

Il est possible de calculer la contribution de chaque variable de **X** à l'analyse, et d'estimer si cette contribution est significative pour le modèle. On peut donc sélectionner un sous-ensemble de variables de **X** qui permet d'obtenir le modèle le plus parcimonieux. On fait l'hypothèse que les variables environnementales sélectionnées sont celles qui structurent le plus la communauté phytoplanctonique. L'analyse canonique inclue donc une étape préalable de sélection des variables de **X** qui rentreront dans la composition du plan factoriel. Nous avons utilisé une méthode de sélection de variables dite *backward* (fonction *ordistep*).

Dans beaucoup d'études statistiques, les analyses canoniques sont appliquées à des données spatialisées (Cottenie, 2005 ; Dray et al., 2012). Les individus statistiques sont donc des localisations spatiales, et non des dates comme c'est le cas dans nos données. Legendre & Legendre (1998) évoquent néanmoins la possibilité de travailler sur des dates d'échantillonnage. Pour contrôler la dépendance des dates entre elles, ils préconisent d'intégrer une co-variable **W** dans le modèle d'analyse canonique. Au lieu d'appliquer l'analyse canonique sur la matrice brute des variables environnementales, celle-ci est effectuée sur la matrice des résidus issue de la régression de chaque variable environnementale (prise séparément) sur la co-variable.
Ici, nous avons choisis comme co-variable temporelle le *jour julien* correspondant à chaque date de nos deux tableaux de données, c'est-à-dire le nombre de jours séparant une date donnée de la première date de la série.

Ce type d'analyse peut être réalisé sous R avec le package *vegan* (Oksanen et al., 2018).

2.b. Analyses à 3 tableaux : fourth-corner analysis

Les analyses canoniques permettent de mettre en évidence des liens entre des variables environnementales et l'abondance des espèces de la communauté. On souhaite maintenant s'affranchir de la notion d'espèces et faire le lien entre les variable environnementales et les traits fonctionnels des taxons présents dans la retenue. Cela doit permettre de dégager des pistes de réflexion sur les mécanismes écologiques conduisant à l'apparition des blooms.
Un type particulier d'analyse statistique multivarié a été pensé pour répondre à ce type de question : les analyse à 3 tableaux. Ces analyses cherchent une solution à ce qui est parfois appelé 'le problème du quatrième coin' (*fourth-corner*). On cherche à étudier la relation entre une matrice **R** des variables environnementale et une matrice **Q** des traits fonctionnels de la communauté écologique étudiée. La matrice **L** des abondances des taxons algaux permet de faire le lien entre les deux matrices précédentes. Ainsi, **R** et **L** on le même nombre de lignes (les dates d'échantillonnage), tandis que **L** et **Q** ont le même nombre de colonnes (les espèces de la communauté). La figure 6 permet de visualiser l'agencement des trois tableaux.

Dans la suite de l'étude, nous nous concentrerons sur la procédure d'analyse à 3 tableaux appelée *fourth-corner analysis* (Legendre et al., 1997). Celle-ci permet de tester une relation de significativité entre chaque variable environnementale et chaque trait pris séparément (ter Braak et al., 2012). Cette relation est établie en comparant une statistique du χ^2 calculée sur les données « observées » à une distribution obtenue après permutation des lignes et des colonnes de la matrice d'abondance **L**. Ce type d'analyse peut être réalisé avec le package *ade4* (Dray et al., 2018).

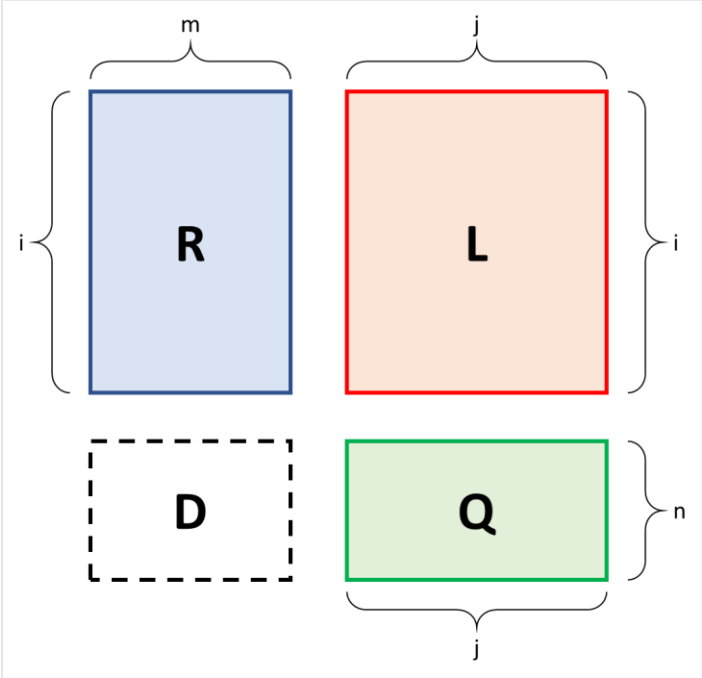

Figure 6 : Représentation graphique du problème du quatrième coin. On cherche à combiner une matrice environnementale **R** (i dates × m variables environnementales), une matrice d'abondance spécifique **L** (i dates × j taxons) et une matrice de traits **Q** (n traits fonctionnels × j taxons), de manière à obtenir une matrice **D** (le quatrième coin) résumant la relation entre variables environnementales et traits des espèces (Brown et al., 2014).

Pour réaliser ce type d'analyse, il est nécessaire de déterminer la matrice **Q** des traits fonctionnels des taxons de la communauté. Les traits considérés sont au nombre de 12, et peuvent être classés en deux catégories : les traits morphologiques et les traits de nutrition. Ils sont présentés dans le tableau 1.

Tableau 1 : Traits fonctionnels considérés dans l'étude.

Groupes	Traits fonctionnels	Modalités
Traits morphologique	Coefficient de forme (vitesse de sédimentation)	High ; Medium ; Low
	Forme de l'individu	Unicellulaire ; Petite colonie ; Grande colonie
	Production de mucilage	Présence ; Absence
	Présence de flagelle	Présence ; Absence
	Capacité de flottaison	Présence ; Absence
	Squelette siliceux externe	Présence ; Absence
	Rapport Surface/Volume	*Quantitatif*
Traits de nutrition	Taux de croissance maximal	*Quantitatif*
	Affinité pour l'Azote	*Quantitatif*
	Affinité pour le Phosphore	*Quantitatif*
	Capacité de Réserve pour l'Azote	*Quantitatif*
	Capacité de Réserve pour le Phosphore	*Quantitatif*

Les traits morphologiques reflètent les caractéristiques physiques des espèces. Nous en avons considéré sept.
Les traits de nutrition sont liés aux stratégies de consommation des nutriments par une espèce. Nous en avons considéré cinq. Les valeurs de ces traits sont calculées à partir des valeurs des paramètres du modèle mécaniste de croissance du phytoplancton de Droop (Droop, 1968 ; Klausmeier et al., 2004). L'*Affinité* pour un nutriment, c'est-à-dire la quantité de ce nutriment absorbée par unité de temps, reflète la capacité d'une espèce à absorber ce nutriment le plus rapidement possible (Edwards et al. 2013 b). La *Réserve* pour un nutriment, c'est-à-dire la capacité de stockage de cet élément, reflète la capacité d'une espèce à résister à un épuisement de ce nutriment dans le milieu (Ducobu et al., 1998). Nous avons choisi de nous limiter aux traits de nutrition relatifs à l'azote et au phosphore, car ce sont les nutriments qui sont susceptibles de limiter la croissance du phytoplancton dans les systèmes aquatiques.

Pour chaque taxon, nous avons effectué une recherche bibliographique de manière à déterminer les valeurs des 13 traits fonctionnels.
Dans le cas des traits morphologiques, nous avons pu nous appuyer sur des ouvrages d'aide à l'identification, qui détaillant précisément les caractéristiques morphologiques de la majorité des espèces de phytoplancton (John et al., 2002 ; Reynolds, 2006). Dans le cas du rapport Surface/Volume, il est nécessaire d'avoir les valeurs de surface et de volume du taxon considéré. Pour cela, nous avons utilisé des bases de données compilant les valeurs de biovolumes et de surfaces cellulaires d'un grand nombre d'espèces (Olenina et al., 2006). Pour les taxons de notre communauté correspondant à des genres, nous avons donc calculé des moyennes des valeurs de surface et de volume cellulaire des espèces de ce genre.
Dans le cas des traits de nutrition, nous nous sommes appuyés sur des bases de données compilant les valeurs de traits pour des espèce étudiée en laboratoire à l'occasion d'autres travaux de recherches (Edwards et al., 2015). Néanmoins, ces bases de données sont très incomplètes, et ne nous ont pas permis de déterminer les valeurs des traits pour une grande partie des taxons de notre communauté. En effet, il est nécessaire de de procéder à des expériences complexes pour déterminer les valeurs des paramètres du modèle de Droop spécifiques à une espèce. Lorsque nous ne disposions pas des données pour une espèce de la communauté du réservoir, nous lui avons attribué les valeurs de traits d'une espèce appartenant au même genre (en calculant une moyenne si nous disposions de valeurs pour plusieurs espèces du même genre). Lorsque le taxon de notre communauté était un genre phytoplanctonique (et non une espèce), nous avons moyenné les valeurs de traits de toutes les espèces de ce genre pour lesquelles nous avions des informations.
Des auteurs ont par ailleurs montré qu'il existait une relation entre le volume cellulaire d'une espèce et les valeurs des traits de nutrition (Edwards et al., 2012). Ces mêmes auteurs ont déterminé des relations allométriques permettant de lier le biovolume cellulaire aux paramètres du modèle de Droop. En dernier recours, nous avons donc estimés les valeurs des traits de nutrition des taxons de la communauté à partir de leurs biovolumes à l'aide de ces relations allométriques.

Finalement, on obtient une matrice **Q** de 75 colonnes (les taxons) et 13 lignes (les traits fonctionnels). La matrice des traits est présentée en Annexe III. La matrice des traits de nutrition est présentée en Annexe IV. Les relations allométriques utilisées pour l'estimation des paramètres du modèle de Droop sont présentées en Annexe V.

2.c. Modèles de prédiction : Machine Learning

Pour mettre au point des stratégies de gestion du plan d'eau permettant de diminuer efficacement la fréquence des efflorescences, il est important de comprendre précisément la nature des blooms et les mécanismes écologiques expliquant leur apparition. C'est l'objectif recherché avec les analyses décrites précédemment. Mais un modèle statistique de prédiction des blooms peut déjà s'avérer être un outil très utile pour le gestionnaire, pour anticiper ces événements et mitiger les coûts d'exploitation.
Le problème peut être résumé de la manière suivante : on veut prédire les valeurs prises par une variable qualitatives (déclenchement ou non d'une efflorescence) à l'aide d'un jeu de données de variables quantitatives (les variables environnementales). On cherche donc un modèle statistique qui utiliserait les valeurs des variables environnementales physico-chimiques, météorologiques et hydrologiques au temps $t - i$ pour prédire l'apparition d'un bloom au temps t (avec i le nombre de semaines voulu avant la date à prédire). Une approche de classification par *Machine Learning* semble tout indiquée pour répondre à cette problématique.

La variable à prédire est le déclenchement d'un bloom. Elle peut prendre une des deux modalités : déclanchement (B) ou non déclenchement (NB). On se restreint aux dates susceptibles de voir l'apparition d'un bloom, c'est-à-dire les dates entre mai et novembre. Les dates B suivent obligatoirement une date NB. Les dates de comptage comportant une abondance phytoplanctonique suffisante pour être considéré comme des dates de bloom mais qui suivent déjà une date B sont supprimées, car nous considérons qu'elles caractérisent un bloom déjà déclenché.
Nous constatons que le nombre d'événements 'bloom' B (14 dates) est bien plus faible que le nombre d'événements 'non bloom' NB (165 dates). Ce déséquilibre dans les classes à prédire risque de poser des difficultés pour estimer un modèle de prédiction efficace. Pour mitiger les effets de ce déséquilibre, nous avons choisi l'aire sous la courbe ROC comme critère de sélection lors de l'étape d'estimation des hyperparamètres (Maloof, 2003 ; Fawcett, 2006). Nous avons aussi affecté des poids aux classes de manière à pénaliser plus fortement, lors de la procédure de sélection, les modèles qui échouent à prédire correctement les dates B.
Lorsque l'on met en place un modèle de Machine Learning, on effectue une partition du jeu de données : le modèle est estimé sur un jeu de données *train* (2/3 de la partition) et son efficacité de prédiction est testée sur le jeu de données *test* (1/3 de la partition).

Nous avons testé et comparé sept algorithmes de *Machine Learning* :

- K-plus-proches-voisins (*knn*)
- Support Vector Machine (*SVM*) linear
- Support Vector Machine (*SVM*) polynomial
- Régression pénalisée *RIDGE*
- Régression pénalisée *LASSO*
- Random Forest
- Réseaux de neurones (*Neural Networks*)

Les algorithmes ont tous été implémenté à l'aide du package *caret* (Kuhn, 2018). Nous avons mis en place une procédure de cross-validation dite *n-folds repeated* ($n = 10$, 5 répétitions) pour estimer les valeurs optimales des hyperparamètres de chaque algorithme (à l'aide de la fonction *traincontrol* du package *caret*). L'Annexe VI liste les valeurs d'hyperparamètres testées pour chaque méthode.

Résultats

1. Caractérisation des événements de blooms

La figure 7 permet de visualiser l'occurrence des blooms sur l'ensemble de la série temporelle. Au total, on identifie 35 dates de blooms distinctes sur les 279 dates de comptage examinées. L'ensemble des blooms sont produits par cinq taxons appartenant tous au groupe des cyanobactéries : *Aphanizomeon flos-aquae* (une date), *Microcystis botrys* (quatre dates), *Microcystis viridis* (huit dates), *Phormidium* (quatre dates), *Planktothrix agardhii* (20 dates).
En terme d'abondances brutes, ce sont les blooms de *Microcystis* qui sont les plus intenses, avec des concentrations cellulaires dépassant les 10^6 cell.mL^{-1} en 2006 et 2010. *Planktothrix agardhii* est l'espèce la plus récurrente, et est responsable d'efflorescences sur quatre années de la série : 2011, 2013, 2015 et 2016 (fig.7).

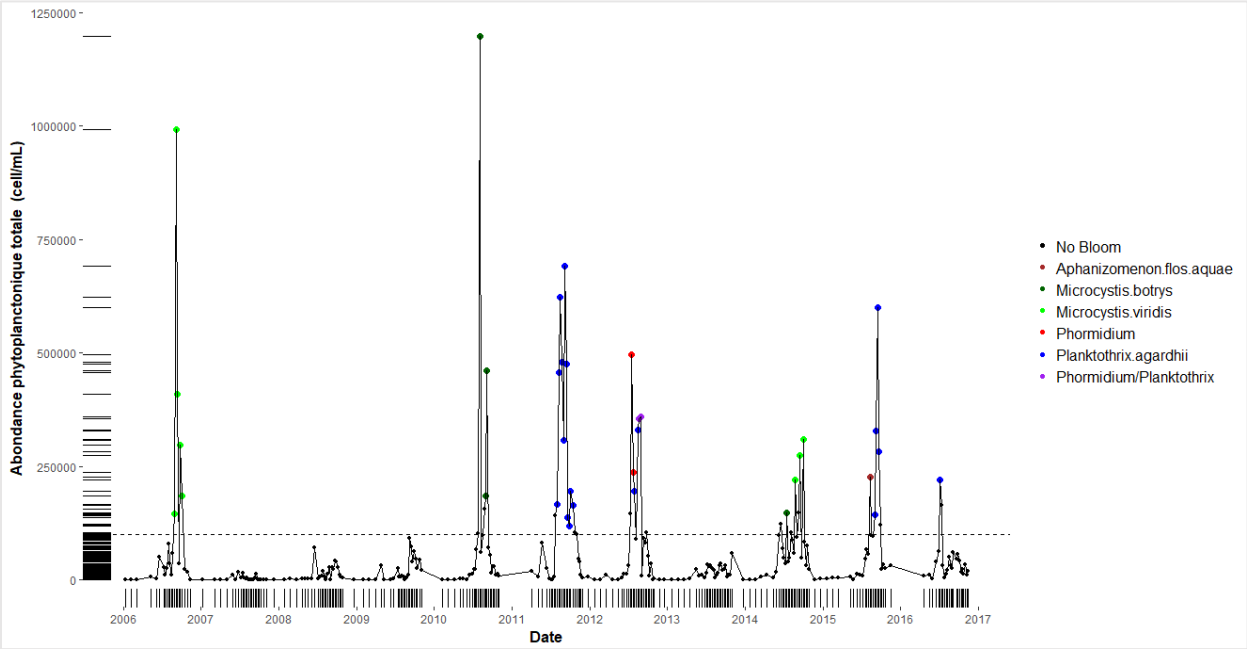

Figure 7 : Evolution de l'abondance du phytoplancton (tous taxons confondues) dans la retenue de Rophémel entre 2006 et 2016. La ligne pointillée noire correspond au seuil de 10^5 cell/mL, l'abondance au-dessus duquel on considère qu'une efflorescence a lieu. Les points de colorés indiquent les dates de blooms, chaque couleur correspondant à une espèce différente. On remarque quelques points non-coloré au-dessus du seuil : ils correspondent à des dates ou l'abondance totale de la communauté, mais pas l'abondance d'une seule espèce, dépasse le seuil de bloom.

Les événements d'efflorescences se produisent toujours pendant l'été et le début de l'automne, entre juillet et octobre. Sur les 11 années de comptages considérées, la retenue a été touchée par des efflorescence sept années. Depuis 2010, la retenue est touchée au moins une fois par an par un événement de bloom (sauf en 2013).
En général, sur une année, une seule espèce de cyanobactérie produit un bloom. Ce n'est pas le cas en 2012, où pour deux dates de comptage (le 21 août 2012 et le 28 août 2012), on constate un bloom simultané de deux espèces (*Phormidium* et *Planktothrix agardhii*). On note aussi qu'en 2016, le bloom de *Planktothrix agardhii* est précédé par un bloom de *d'Aphanizomeon flos-aquae* de moindre amplitude et d'une durée plus restreinte.

2. Abondance spécifique et variables environnementales

Avant la mise en place de la procédure d'analyse canonique partielle, nous avons appliqué une transformation de *Hellinger* au jeu de données de comptages, afin de normaliser les abondances des taxons de la communauté. Cette transformation est recommandée par Legendre & Gallagher (2001), et permet de réduire le poids donné aux espèces les plus rares. Ces mêmes auteurs, qui traitent de la mise en œuvre d'analyses factorielles simples dans l'article cité, préconisent ensuite d'appliquer une ACP sur les données transformées. Par conséquence, nous avons choisie d'utiliser une analyse canonique de type RDA (basée sur une ACP) plutôt qu'une CCA (basée sur une AFC).

On constate que les résultats de la procédure de sélection *backward* des variables environnementales du modèle de RDA présentent une certaine instabilité : si la procédure est lancée deux fois, ce n'est pas forcément exactement le même modèle qui est retenu. L'augmentation du nombre de permutation dans la fonction de sélection *ordistep* ne permet pas de résoudre ce problème. La procédure a donc été répétée 50 fois, et nous avons sélectionné les variables présentent dans au moins 48 des modèles finaux (soit au moins 95% des itérations).
A la suite de cette étape de sélection *backward*, nous avons testé la colinéarité des variables du modèle en calculant les scores VIF (*Variance Inflated Factors* ; Craney & Surles, 2002). Si une variable présente un score supérieur à 10, elle est éliminée du modèle final de RDA partielle. Cette sélection se fait variable par variable, en recalculant les scores VIF après chaque suppression.

Le modèle final de RDA partielle comporte 12 variables. Nous avons ajouté dans ce modèle une variable écartée à l'issue des précédentes étapes de sélection : le *Volume d'eau dans la retenue*. La mitigation des blooms par la baisse du niveau d'eau dans la retenue est en effet une piste de gestion envisagée par le gestionnaire de la retenue.
Le modèle analysé comporte donc 13 variables qui contribuent à la construction du plan factoriel et une co-variable pour contrôler l'effet du temps

Le modèle final de RDA partielle explique 25.2 % de l'inertie du jeu de données de comptage. La co-variable temporelle explique 4.4 % de l'inertie. La figure 8 montre le plan factoriel formé par les deux premiers axes générés par l'analyses et permet de visualiser les relations entre les variables environnementale et les taxons biologiques (les colonnes du tableau d'abondance).
Nous avons choisi de dessiner des flèches uniquement pour les variables environnementales, afin de ne pas surcharger le graphique. La longueur d'une flèche est proportionnelle à la contribution de la variable environnementale associée à la formation des axes du plan : plus la longueur d'une flèche est grande, plus la représentation de cette variable sur le plan est bonne.
Une flèche symbolise le gradient des valeurs prises par une variable environnementale. La proximité de deux barycentres indique la force (et le sens) de la relation entre deux variables (biologiques ou environnementales). Par exemple, plus un nom d'un taxon est proche de l'extrémité d'une flèche correspondant à une variable environnementale, plus l'abondance de ce taxon est associée à de fortes valeurs de cette variable. Au contraire, une opposition forte sur le plan indique qu'un taxon est associé à de faibles valeurs de la variable environnementale.

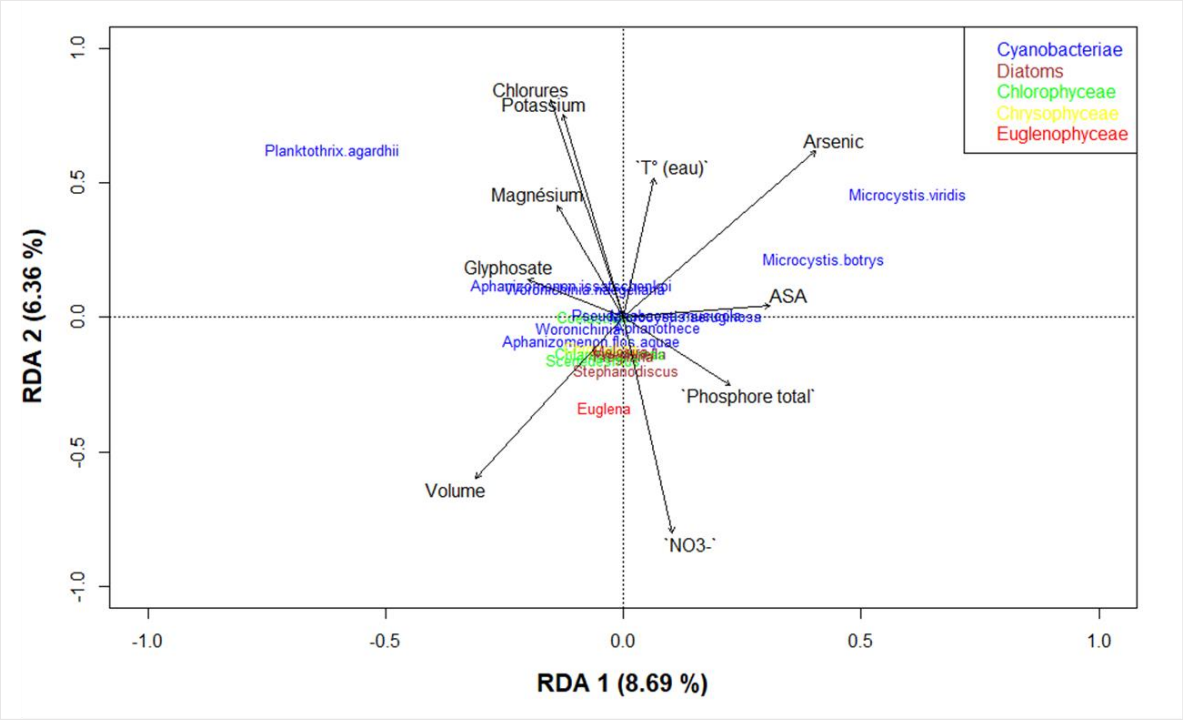

Figure 8 : Plan de RDA (Axe 1 & Axe 2). Les extrémités des flèches indiquent l'emplacement des centres de gravité des variables environnementales. Pour rendre le plan lisible, seules les espèces les mieux représentées apparaissent sur le plan. De même, trois variables mal représentées sur le plan n'apparaissent pas (Sulfates, Nickel et Aluminium). La couleur des espèces correspond à leur groupe taxonomique.

Le plan de la RDA partielle (fig. 8) montre que trois espèces ressortent nettement du nuage des taxons : *Planktothrix agardhii*, *Microcystis viridis* et *Microcystis botrys*. Ce sont trois des cinq espèces de cyanobactéries productrices de blooms, identifiées précédemment comme particulièrement problématiques, avec une séparation des *Microcystis* à droite et de *Planktothrix* à gauche de l'axe 1. Les trois principales espèces sortent en partie haute de l'axe 2 et sont donc positivement associées à la *Température de l'eau*, et négativement associés au *Volume d'eau* dans la retenue (il faut néanmoins noter que ces deux variables ne soient pas représentées de manière optimale sur le plan). Un niveau ''bas'' de la retenue est des températures élevées semblent donc favorable au bloom. Les deux espèces de *Microcystis* semblent particulièrement liées à la concentration en *Arsenic* et au faible volume du réservoir.

On remarque que les concentrations en *Chlorures*, *Potassium* et *Nitrates* forment un axe très structurant le long de l'axe 2. Les *Chlorures* et le *Potassium* sont très corrélés, et ces deux variables sont très anti-corrélées à aux *Nitrates*. Les températures faibles coïncident avec les concentrations fortes en nitrates et les concentrations faibles en chlorures et potassium.

3. Passage de la classification linnéenne à la notion de traits fonctionnels

Avant d'appliquer la procédure de *fourth-corner analysis*, nous avons décidé de transformer les traits fonctionnels quantitatifs en variables semi-quantitatives. Les traits fonctionnels morphologiques (présence de flagelle, capacité de production de mucilage…) sont déjà considérés comme qualitatifs, de même que le trait caractérisant le charactère unicellulaire ou colonial de

l'espèce. Dans le cas des traits de nutrition, nous avons créé des classes de valeurs (*low*, *medium* & *high*). Les limites de ces classes sont choisies de manière à les équilibrer en termes d'effectifs (c'est-à-dire de manière à retrouver environ le même nombre de taxons dans chaque classe pour un trait fonctionnel).

L'analyse a été réalisée en deux fois :

- Tout d'abord, nous avons testé la relation entre les cinq traits de nutrition et une sélection de variables environnementales physico-chimique, notamment les concentrations en nutriments.
- Puis nous avons testé la relation entre les huit traits de morphologiques et une sélection de variables environnementales hydrologiques et météorologiques, ainsi que la concentration en *Arsenic*, variable mise en évidence précédemment par la RDA partielle.

L'analyse est basée sur des tests de permutations. La statistique de test utilisée est le D2, ce qui permet de tester la relation des traits modalité par modalité avec les variables environnementales (Legendre et al., 1997). 10000 permutations de la matrice des abondances sont effectuées pour conduire les tests de significativité.
La *fourth-corner analysis* est une procédure qui s'apparente à des comparaisons multiples. Il est donc nécessaire de corriger les seuils de significativité des tests effectués. A une correction de Bonferroni jugée trop conservatrice, nous avons préféré une correction de Benjamini-Hochberg-Yekutieli. Le seuil de significativité (ajusté après correction) est fixé à 0,05.

L'analyse met en évidence plusieurs relations intéressantes entre les traits morphologiques et les variables environnementales (fig. 9A). Il existe des relations significatives entre des variables associées à un hydrodynamisme fort dans la retenue, comme le *Volume total* et le *Volume entrant* et des traits associés à des espèces morphologiquement caractéristiques. Ainsi, l'hydrodynamisme favorise des espèces de grandes tailles (associée à de faibles rapports S/V) incapable de réguler leur flottaison, et pénalisent les petites espèces (forts rapports S/V).
En revanche la température favoriserait les petites espèces coloniales, morphologiquement simples (pas de flagelle), capables de réguler leur position verticale dans la colonne d'eau (par la présence de vacuoles gazeuses).
D'après les relations décrites, un faible hydrodynamisme et de fortes températures de l'eau, typiques de la période estivale, favorisent les espèces productrices de blooms (tab. 2), en accord avec la littérature (Paerl & Huisman, 2009).

Tableau 2 : valeurs des traits fonctionnels morphologiques des espèces de phytoplancton productrices de blooms dans la retenue de Rophémel (Y : présence ; N : absence).

Sp_complet	Forme Ind	Coeff sédimentation	Flagelle	Mucilage	Vacuole	Hétérocyste	Silice	S/V	μ_{max}
Aphanizomenon flos-aquae	large_colonies	high	N	N	Y	Y	N	high	medium
Microcystis botrys	large_colonies	high	N	Y	Y	N	N	high	medium
Microcystis viridis	large_colonies	high	N	Y	Y	N	N	high	medium
Phormidium	large_colonies	high	N	Y	Y	N	N	high	high
Planktothrix agardhii	large_colonies	high	N	Y	Y	N	N	high	low

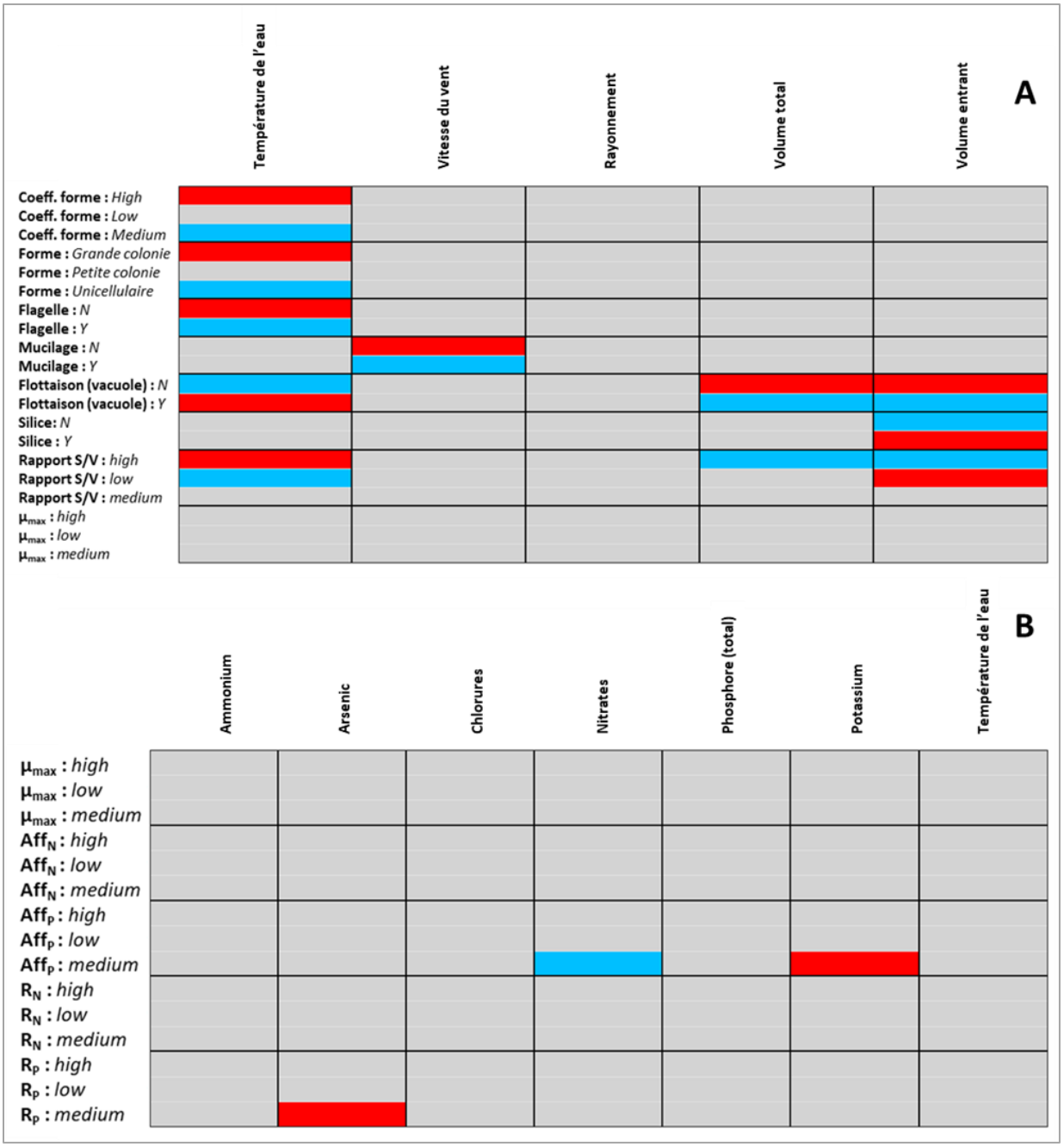

Figure 9 : Résultats de la *fourth-corner analysis*. (A) Traits de nutrition (lignes) et variables physico-chimiques (colonnes). (B) Traits morphologiques et variables hydro-météorologiques. On teste la relation entre les variables environnementales et les traits fonctionnels modalité par modalité. Une case rouge indique une relation significativement positive (la modalité du trait est associée à de fortes valeurs de la variable), une case bleu une relation significativement négative (la modalité du trait est associé à de faibles valeurs de la variable).

4. Prédiction des évènements de blooms

Nous avons tenté de mettre au point un modèle de prédiction des blooms par *Machine Learning*. Comme décrit précédemment, la variable à prédire est le déclenchement d'un bloom (qualitative, deux modalités), et les variables prédictives sont les valeurs des variables environnementales (quantitatives) dans les semaines précédant la date à prédire.

Afin de développer un modèle exploitable par les gestionnaires, nous avons réduit le jeu de données des variables explicatives à un nombre restreint de variables environnementales dont le résultat de mesure est disponible instantanément (c'est-à-dire les variables pour lesquelles il n'y a pas de délai lié à une période d'analyse après échantillonnage). Il est préférable de ne pas intégrer les autres variables dans le modèle, puisque leurs valeurs ne sont pas forcément connues lorsque le gestionnaire souhaite obtenir une prévision d'apparition d'un bloom. Ce choix méthodologique exclue une grande partie des variables physico-chimiques. Nous avons ajouté une variable supplémentaire, le rayonnement cumulé depuis le 1er mars de l'année. Cette variable est une mesure indirecte de la durée de la période d'ensoleillement qui peut s'avérer déterminante pour l'apparition d'une efflorescence (Zhang et al., 2012). Finalement, 13 variables distinctes sont utilisées pour estimer les modèles de prédiction. Les variables environnementales retenues sont listées dans le tableau 3

Tableau 3 : variables environnementales utilisée dans les modèles de prédiction par Machine Learning.

Variables environnementales
Carbone organique
pH
Température de l'eau
Titre alcalimétrique complet
Turbidité Formazine Néphélométrique
Vitesse du vent
Précipitation
Température de l'air
Débit du *Néal*
Débit de la *Rance*
Temps de résidence
Rayonnement
Rayonnement cumulé (depuis le 1er mars)

Nous avons tenté de développer trois types de modèles prédictifs distincts : des modèles prédictifs à 1 semaine, 2 semaines ou 3 semaines (au-delà, on considère que le délai est trop important et non pertinent écologiquement pour la prédiction des blooms) :

- Pour la prédiction à 3 semaines, on utilise les valeurs moyennes des mesures des variables environnementales effectuées entre les 28^{e} et 22^{e} jour avant la date à prédire.
- Pour la prédiction à 2 semaines, on utilise les valeurs moyennes des mesures des variables environnementales effectuées entre les 28^{e} et 22^{e} jours et les 21^{e} et 15^{e} jours avant la date à prédire (le nombre de variables explicatives double par rapport au modèle de prédiction à 3 semaines).
- Pour la prédiction à 1 semaine, on utilise les valeurs moyennes des mesures des variables environnementales effectuées entre les 28^{e} et 22^{e} jours, les 21^{e} et 15^{e} jours et les 14^{e} et 8^{er} jours avant la date à prédire (le nombre de variables explicatives double par rapport au modèle de prédiction à 2 semaines et triple par rapport au modèle de prédiction à 3 semaines).

Plus le délai est important, plus le gestionnaire à le temps d'anticiper le bloom et de prendre les mesures nécessaires. Il est aussi raisonnable de penser qu'une prédiction à une semaine sera plus précise, puisqu'elle utilise une quantité d'information plus importante. Nous avons donc pensé que la prédiction au lendemain pouvait présenter donc un intérêt en termes de qualité de la précision, bien que son utilité soit plus limitée pour le gestionnaire.

Pour chaque combinaison *type de modèle* × *algorithme*, nous avons réalisé 50 estimations du modèle (avec des partitions *train*/*test* différentes à chaque fois), afin de voir si les résultats sont stables.
La figure 10 montre les résultats des ''meilleurs'' modèles de prédiction pour le critère d'*Accuracy*, c'est à la dire la proportion de dates correctement prédites. Les résultats des algorithmes de régression pénalisée (LASSO & RIDGE) et les méthodes de type *Neural Networks* parviennent parfois à classer certaines date B correctement, mais toujours au prix d'un grand nombre de dates NB prédites comme B. Les résultats de ces algorithmes sont instables et ne sont finalement pas plus efficaces qu'une prédiction des classes au hasard. Les autres algorithmes ne font pas mieux, et classent presque systématiquement toutes les dates dans la catégorie NB.
Il n'y a pas de différence notable entre les trois types de modèle proposés (prédiction à 1 semaine, 2 semaines ou 3 semaines).

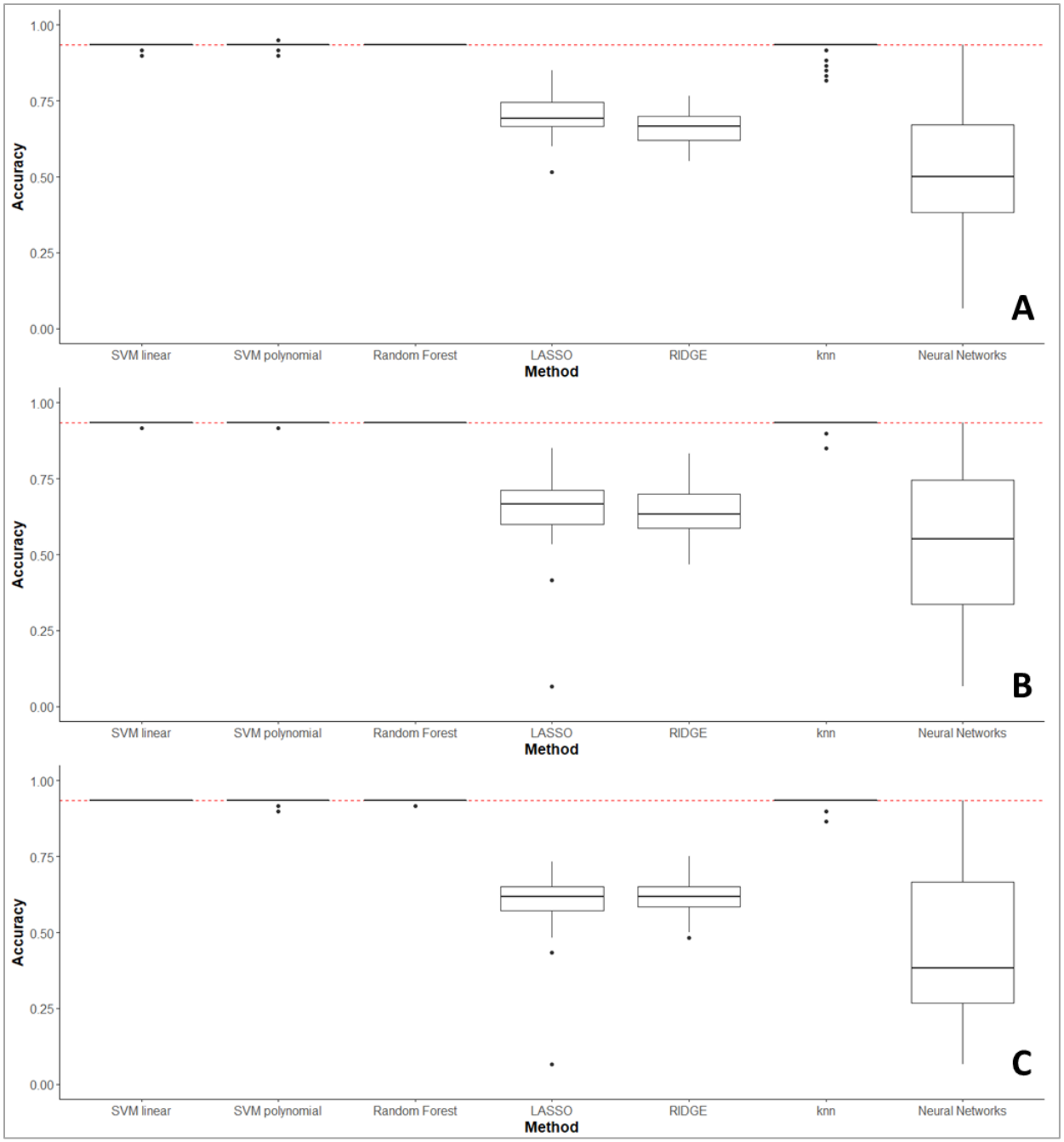

Figure 10 : Boxplots des résultats des modèles de prédiction par Machine Learning. (A) modèles de prédiction à 1 semaine ; (B) modèles de prédiction à 2 semaines ; (C) modèles de prédiction à 3 semaines. La ligne pointillée rouge indique la valeur d'*Accuracy* maximale possible pour un modèle non informatif (c'est-à-dire la valeur d'*Accuracy* lorsque toutes les dates sont prédites dans la classe la plus importante en termes d'effectifs, la classe NB).

Discussion

1. Identification des facteurs susceptibles de favoriser l'apparition des blooms

Le principal but de notre étude était de mettre en évidence des causes potentielles de l'occurrence des blooms dans la retenue Rophémel. Cet objectif, ainsi que les caractéristiques des données à notre disposition (comme la multiplicité des fréquences d'acquisition des variables), ont guidé nos choix d'analyses et la mise en place des méthodes retenues.

Nous avons identifié deux genres phytoplanctoniques responsables de l'essentiel des blooms dans la retenue : *Microcystis* (avec deux espèces : *Microcystis viridis* et *Microcystis botrys*) et *Planktothrix*. *Planktothrix agardhii* est l'espèce responsable du plus grand nombre d'événements de blooms, tandis que les efflorescences de *Microcystis* sont les plus fortes en termes de concentrations cellulaires (avec de fortes accumulations de biomasse sur quelques centimètres). Ces deux taxons de cyanobactéries coloniales diffèrent par la forme de leur colonie : les espèces de *Microcystis* forment des colonies sphériques dont la cohésion est assurée par un épais mucilage, tandis que *Planktothrix* forme des colonies filamenteuses. Une association possible entre les espèces de *Microcystis* et la concentration en *Arsenic* a aussi été mise en évidence. Cet élément chimique est toxique pour le phytoplancton. Bien qu'il n'existe pas d'étude comparative attestant d'une tolérance à l'Arsenic plus importante de *Microcystis* par rapport à d'autre taxon, plusieurs études semble indiquer que les espèces du genre *Microcystis* peuvent accumuler des quantités importantes d'Arsenic et se développer dans des milieux pollués (Gong et al., 2011 ; Wang et al., 2013). Après analyse des chroniques, les niveaux d'Arsenic ne dépassent pas les 4µg/L dans le réservoir, bien en dessous des valeurs mesurées dans des cours d'eau considérés comme pollués par cet élément (à partir de 55 µg/L selon Rosso et al., 2013). Par conséquence, il ne semble pas exister une ‘‘source’’ d'*Arsenic* sur laquelle concentrer d'éventuels efforts de gestion. Mais même à faible dose, cet élément chimique pourrait donner un avantage aux espèces de *Microcystis* par rapport au reste de la communauté et contribuer à l'apparition de blooms par ce taxon (au détriment des blooms à *Planktothrix agardhii* par exemple).
La RDA partielle suggère des pistes sur l'origine des rejets de nutriments responsables de l'eutrophisation de la retenue de Rophémel. La figure 8 montre en effet que le plan de RDA est structuré par un axe formé par trois variables environnementales : les *Chlorures* et le *Potassium*, très corrélés, et les *Nitrates*, très anti-corrélés aux deux variables précédentes. Cela pourrait constituer un signal de rejets en provenance d'un abattoir en amont de la retenue, l'abattoir de Kermené. Les *Chlorures* sont en effet utilisés comme conservateur des produits carnés, et sont très difficiles à traiter, malgré des réglementations qui s'imposent aux abattoirs. Le propos reste à nuancer : en l'état, nous ne disposons pas de suffisamment d'éléments pour rendre l'abattoir responsable des apports de nutriments au niveau de la retenue. Des sources de pollution d'origine agricole ne peuvent pas être exclues (le *chlorure de potassium* est un engrais très fréquemment utilisé pour l'amendement des terres). Nous n'expliquons pas non plus le ‘‘sens’’ de la relation des *Nitrates* avec les deux autres, à savoir une anti corrélation très forte. Il serait intéressant d'effectuer des mesures des flux de ces trois éléments dans le cours d'eau en amont et en aval de l'abattoir et sur les deux affluents du réservoir pour mieux localiser la source potentielle.

La *fourth-corner analysis*, qui prend en considération les traits fonctionnels des espèces, permet de mettre en évidence un rôle de l'hydrodynamisme dans l'apparition des blooms. En particulier, nous

avons montré que des modalités de traits morphologiques caractéristiques des espèces productrices de blooms (tableau 2) sont associées à de faibles valeurs pour des variables qui sont liées aux mouvements des masses d'eau dans la retenue, comme le *Volume total* et le *Volume entrant*. Un hydrodynamisme faible implique une stabilité du plan d'eau qui favorise logiquement les espèces de cyanobactéries productrices d'efflorescences (Zhang & Prepas, 1996). Dans ces conditions les autres espèces de la communauté, incapables de réguler leur flottabilité, tendent à sédimenter (notamment les espèces du groupe des diatomées), tandis que les cyanobactéries peuvent plus facilement se maintenir dans la zone euphotique et continuer leur croissance par photosynthèse (Dokulil & Teubner, 2000 ; Posh et al., 2012 ; Wu et al., 2013). La *fourth-corner analysis* révèle aussi que de fortes valeurs de *Température de l'eau* sont associées à des modalités de traits caractéristiques (forme coloniale, fort rapport S/V, absence de flagelle et capacité à réguler sa flottabilité) des taxons producteurs d'efflorescences. Là encore, cela pourrait suggérer un effet de l'hydrodynamisme. En effet, de fortes températures de l'eau peuvent s'accompagner d'une stratification des masses d'eau de la retenue, associée à un patron particulier de disponibilité des nutriments dont les concentrations forment un gradient décroissant du fond de la colonne d'eau vers la zone euphotique. Cela qui favorise les espèces capables de réguler leur flottabilité et de migrer dans la colonne d'eau (Jöhnk et al., 2008 ; Wagner & Adrian, 2009).
De manière général, notre analyse par les traits fonctionnels ne met pas en évidence d'effet de la disponibilité en ressource, en lien avec les traits de nutrition. Il est possible que la retenue soit tellement eutrophisée que les nutriments nécessaires à la croissance du phytoplancton ne sont jamais limitants. Les concentrations en nitrates et en phosphates fluctuent mais les concentration moyennes sur l'ensemble de la période étudiée sont de 20.5 mg/L et 64 µg/L respectivement, des niveaux élevés. Dans le cas des nitrates, les niveaux ne sont jamais inférieurs à 2.5 mg/L, soit bien au-dessus des niveaux à partir desquels un plan d'eau douce est considéré comme eutrophe (Smith et al., 1999). Néanmoins, cette absence de résultat est plus probablement liée à l'incertitude sur la valeur de ces traits pour les taxons de notre communauté. En effet, il n'existe pas de littérature sur les traits de nutrition d'une part importante des taxons de la communauté. Les traits de ces taxons donc ont été déterminés à partir de leur volume cellulaire grâce à des relations allométriques. Lors d'une précédente étude (Le Noac'h et al., 2017), nous avions mis en évidence que les espèces de phytoplancton dominantes d'un milieu peuvent être sélectionnées par la fréquence des apports de nutriments sur la base de la valeur de leurs traits de nutrition. Par conséquence, les valeurs de traits de nutrition des espèces dominantes (notamment les espèces responsables des blooms) tendent à s'écarter des valeurs prédites par les relations allométriques. Cela signifierait que notre matrice de trait est faussée dès le départ. Cette solution d'estimation des traits par les relations allométriques était la seule à notre disposition, mais elle n'apparaît finalement pas adaptée. Il parait donc nécessaire de mettre en place des expériences pour déterminer précisément les traits de nutrition de tous les taxons de la communauté. Une telle approche expérimentale se révèlerait néanmoins très coûteuse en temps et en ressources humaines et matérielles, et parait difficile à mettre en œuvre.

Sans mettre en évidence une seule variable contrôlant entièrement le déclenchement des blooms, la RDA partielle et la RLQ permettent donc d'identifier plusieurs facteurs environnementaux qui pris dans leur ensemble peuvent expliquer l'apparition des efflorescences, comme la concentration en Arsenic, la température de l'eau et la stabilité hydrodynamique du plan d'eau.

2. Evaluation des effets potentiels d'une stratégie de gestion : la mitigation des blooms par le contrôle du temps de séjour de l'eau dans la retenue

Avant la réalisation de cette étude statistique, l'une des principales pistes de gestion étudiée par le gestionnaire pour mitiger la fréquence d'occurrence des blooms est la baisse du temps de séjour de l'eau dans la retenue. Cela doit permettre d'évacuer les cyanobactéries hors de la retenue avant qu'elles ne prolifèrent, et ainsi limiter la production algale à un niveau insuffisant pour former un bloom. Des études ont également montré qu'un temps de séjour de l'eau important favorise l'abondance des cyanobactéries toxiques (Maier et al., 2004 ; Posch et al., 2012 ; Romo et al., 2013).
La principale solution envisagée pour baisser le temps de séjour est de maintenir le volume d'eau stocké dans la retenue à un niveau réduit. La RDA partielle montre cependant que les taxons identifiés comme problématiques sont plutôt associé à de faible valeur de volume (fig. 8). Une baisse du volume dans la retenue risque aussi de provoquer une hausse de la température de l'eau. Or la RDA partielle (fig. 8) et la RLQ (fig. 9) montre que de fortes températures favorisent la prolifération des espèces productrices de blooms. Les conclusions de notre étude montrent donc que la piste de gestion envisagée risque de ne pas avoir l'effet escompté, et risque même d'aggraver le problème des efflorescences.

Nous avons tout de même voulu savoir s'il existe une relation entre le temps de séjour et l'intensité d'un bloom. Pour cela, nous avons estimé des moyennes du temps de séjour quotidien au cours de chaque bloom (avec un décalage de sept jours), en considérant qu'un bloom englobe l'ensemble des dates consécutives pour lesquelles la biomasse d'une seule espèce est supérieure au seuil de 10^5 cell.mL^{-1}. Nous avons ensuite estimé l'intensité de chaque bloom, en calculant la biomasse moyenne sur l'ensemble des dates appartenant à un même bloom. La figure 11 montre la relation entre les deux variables.

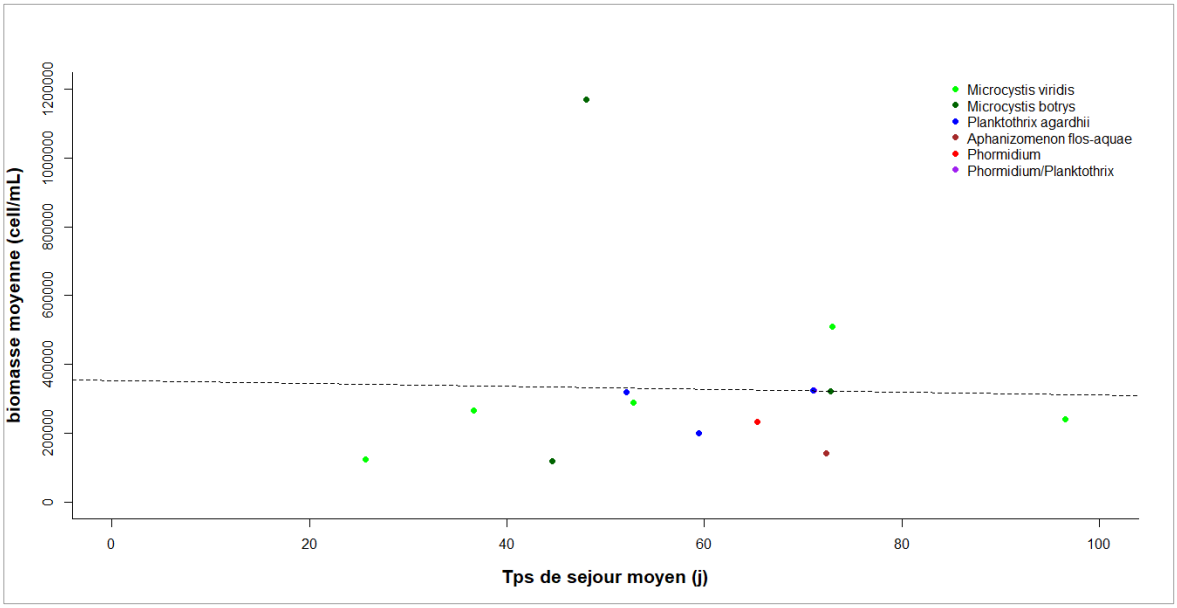

Figure 11 : Relation entre le temps le temps de séjour moyen lors d'un bloom et la biomasse moyenne du taxon dominant sur la durée de ce bloom. La ligne pointillée noir correspond à la droite de régression entre ces deux variables ($n = 14, p = 0.92$).

Nos données ne permettent pas de mettre en évidence une relation entre le temps de séjour de l'eau dans la retenue et l'intensité des blooms dans le réservoir de Rophémel. Le coefficient de la droite de régression n'est pas significativement différent de 0 ($p = 0.92$). Il faut tout de même noter que l'on s'appuie sur 14 points de données uniquement. De plus la variable du temps de séjour présente un degré d'incertitude dans son estimation, et tend à surestimer le temps de séjour réel.

On ne peut donc pas conclure qu'il existe un effet du temps de séjour sur la fréquence ou l'intensité des efflorescences. Il est donc difficile de juger du potentiel de la stratégie de gestion de réduction de la fréquence des blooms par une réduction du temps de séjour de l'eau. Au regard des éléments de discussion évoqués précédemment, nous déconseillons tout de même cette stratégie. Sa mise en œuvre par la baisse du niveau d'eau dans la retenue risque en effet de provoquer une hausse de la température et de stabiliser le plan d'eau, ce qui favoriserait les espèces productrices de blooms.

3. Difficultés à mettre en place des modèles de prédiction (Machine Learning) valides

Nous avons montré que les analyses multivariées mise en place donnent des résultats permettant d'orienter la stratégie de gestion des efflorescences. En revanche, notre tentative pour développer un modèle de prédiction des blooms a clairement échouée.

Nous anticipions des difficultés liées au déséquilibre des classes à prédire. Nous avons voulu gérer ce problème en attribuant aux classes des poids de prédiction en fonction de leurs effectifs, et en choisissant un critère adéquat de sélection des hyperparamètres.
En revanche, nous n'avons pas pu tester certains algorithmes de *Machine Learning* potentiellement efficaces, comme les procédures de type *Boosting* (Schapire, 1999). Ces algorithmes nécessitent en effet d'importantes ressources de calcul informatique et sont presque toujours mis en œuvre à l'aide de clusters de calcul. Nous n'avions malheureusement pas ce type de ressource à notre disposition.
L'élément principal qui pourrait expliquer l'absence de résultats a été notre volonté de n'intégrer dans le modèle que des variables environnementales prédictives acquises en continue et dont les valeurs de mesures sont disponibles immédiatement pour le gestionnaire. Cela a drastiquement réduit le nombre de variables dans le modèle, et nous avons surement écarté des variables potentiellement essentielles pour espérer obtenir un modèle de prédiction efficace. Par exemple, le modèle n'intègre pas les concentrations des nutriments assimilés par le phytoplancton (azote, phosphore, silice…), car la mesure de ces variables implique des analyses en laboratoire, ce qui retardent la disponibilité de la mesure. Finalement, le modèle intègre principalement des variables hydrologiques et météorologiques. Or nous avons déjà émis des doutes sur la pertinence des mesures d'une partie des variables météorologiques, acquises à des stations de mesures parfois très éloignées du site d'étude.

Nous avons voulu voir si les performances des modèles étaient meilleures en intégrant les variables précédemment écartées (en utilisant leurs valeurs hebdomadaires obtenues après l'étape d'interpolation). Nous avons constaté que les résultats ne changent par rapport à ce que l'on obtenait précédemment. Cela signifie peut-être qu'il nous manque encore des variables environnementales importantes qui contrôlent la dynamique de la communauté et l'apparition des blooms.

4. Pertinence de l'utilisation des données publiques d'agence pour une l'étude d'une communauté phytoplanctonique et suggestions d'amélioration de l'échantillonnage

La qualité de l'échantillonnage des données a été une source d'interrogations constante au cours de cette étude. Une des interrogations annexes à ce travail d'analyses statistiques était de savoir si des données comme celles que nous avions à notre disposition, c'est-à-dire des données publiques collectées dans une cadre réglementaire, peuvent être utilisées dans le cadre d'une étude statistique poussée.

Nous avons déjà évoqué le cas des données météorologiques, et de l'incertitude liée au fait que les stations de mesure sont parfois très éloignées du site d'étude. Compte tenu de l'importance de l'hydrodynamisme dans le contrôle des événements d'efflorescence, il faut impérativement améliorer la qualité et la pertinence de ces données. Il semble donc essentiel d'installer une station météorologique (incluant notamment des mesures du vent et du rayonnement solaire) dans les environs immédiats de la retenue.

Il serait probablement utile, dans la mesure du possible, d'augmenter la fréquence d'échantillonnage des variables physico-chimiques, comme les nutriments ou les pesticides. En effet, la majorité des données hebdomadaires pour ces variables sont en fait des résultats d'interpolation entre des mesures réelles mensuelles. Cela introduit forcement de l'incertitude dans nos résultats. Il existe par exemple des sondes de mesure permettant des estimations en continue des concentrations de nutriments.

Enfin, il nous semble important de mettre en place un calendrier d'échantillonnage et de comptage de la communauté plus régulier. Actuellement, la fréquence des comptages est très réduite entre décembre et mai (un comptage par mois au maximum). Il n'est pas anormal de diminuer la quantité d'événements de comptages du phytoplancton en hiver, puisque la croissance des organismes est limitée et la biomasse de la communauté très réduite. Nous sommes aussi conscients des contraintes de coûts que représente un échantillonnage hebdomadaire sur toute l'année. Mais nous constatons tout de même tout de même la présence dans les chroniques de comptages de ‘‘trous’’ de plusieurs mois (par exemple entre décembre 2010 et mai 2011, ou décembre 2015 et mai 2016). Cela limite l'utilisation qui peut être faite de ces données. Dans ces conditions, il n'est par exemple pas possible d'étudier dans le détail les patrons de successions des espèces de la communauté, or il est possible que la communauté estivale soit conditionnée par la composition de la communauté printanière (Deng et al., 2014). Nous recommandons donc un échantillonnage mensuel, voir bimensuel si possible, entre décembre et mai. Il est essentiel de garder une fréquence hebdomadaire de comptage en été et en automne pour assurer un suivi convenable de la communauté lors des événements d'efflorescences. Il parait important que les changements de fréquences d'échantillonnages se fasse aux mêmes dates chaque année : cela permettrait par exemple d'étudier dans le détail des patterns de succession des espèces de manière à effectuer des comparaisons interannuelles poussées.

L'acquisition des données d'agence, prévue pour le contrôle de la qualité de l'eau, présente donc un certain nombre de limites qui rendent difficile l'exploitation de ces données dans un cadre de recherche scientifique. Des changements dans la stratégie d'échantillonnage pourraient permettre de mettre en évidence d'autres facteurs influençant la dynamique de la communauté phytoplanctonique et pourraient finalement permettre d'affiner les stratégies de gestion destinées à améliorer la qualité de l'eau.

Conclusion & Perspectives

Nous avons mené une étude statistique poussée, afin d'identifier les mécanismes environnementaux contribuant à l'apparition des efflorescences de cyanobactéries et d'orienter les stratégies de gestion destinées à réduire la fréquence d'apparition de ces efflorescences. Les données sont issues de suivis règlementaires de la qualité de l'eau dans un réservoir, dont l'enjeu est la production d'eau potable. Un des objectifs a donc été de mettre en forme les données acquises par différents acteurs et d'évaluer leur potentiel pour tester des hypothèses scientifiques. Les analyses factorielles multivariées (RDA et *fourth-corner analysis*) ont mis en évidence l'influence de l'hydrodynamisme du plan d'eau dans la formation des blooms. Nous avons aussi identifié des variables environnementales, comme l'Arsenic, favorisant potentiellement la dominance d'une espèce sur le reste de la communauté, et jouant donc probablement un rôle dans l'établissement des blooms.
Cette étude pourrait permettre d'orienter la stratégie de gestion des efflorescences par le gestionnaire de la retenue de Rophémel : la diminution du temps de séjour de l'eau par la baisse du volume de la retenue apparait comme potentiellement contre-productive, puisque qu'elle s'accompagnerait de températures de l'eau plus importantes et d'un faible hydrodynamisme, favorables à l'apparition des efflorescences. Il serait intéressant d'effectuer un travail supplémentaire de mesures, notamment pour déterminer si les rejets de l'abattoir de Kermené en amont du barrage contribuent à l'eutrophisation de la retenue.

Néanmoins, nous ne sommes pas parvenus à développer un modèle fonctionnel de prédiction du déclenchement des efflorescences. Nous avons avancé différentes explications pour expliquer cette absence de résultat probant, notamment une sélection préalable des variables trop drastiques. La prédiction des blooms de cyanobactéries toxiques reste tout de même un champ d'étude très porteur (Ahn et al., 2011 ; Xie et al., 2012 ; Lou et al., 2016), et un modèle fonctionnel constituerait un outil de gestion extrêmement utile pour le gestionnaire de la retenue. Il serait donc intéressant d'approfondir cet aspect de l'étude en intégrant des variables supplémentaires et en testant de nouveaux algorithmes de classification à l'aide de clusters de calcul.

L'une des principales limites de cette étude est le fait que nous ne prenons pas en compte la dimension spatiale de la retenue. Les mesures des variables physico-chimiques et les prélèvements du phytoplancton se font toujours au même point d'échantillonnage sur l'ensemble de la période d'acquisition des données, à l'aval du réservoir, qui correspond au point le plus intégrateur de ce qui se passe dans le lac. Mais un plan d'eau n'est pas physiquement et chimiquement homogène (Kling et al., 2000), et on ne peut pas intégrer cette complexité spatiale qui affecte la dynamique de la communauté algale en échantillonnant à un seul endroit. Par exemple, les blooms se déclenchent souvent dans les parties amont des barrages, avant de se déplacer vers l'aval sous l'effet des processus hydrologiques et météorologiques (Chen et al., 2003). Il faudrait donc refaire les analyses avec des données de variables environnementales récoltées plus en amont dans le réservoir.
Pour prendre en compte la dimension spatiale du fonctionnement de la retenue, une approche statistique comme celle que nous avons mise en place ne suffit pas. Une approche possible est le développement d'un véritable modèle mathématique spatialisé qui décrirait la compétition pour les nutriments entre les taxons de la communauté phytoplanctonique, tout en prenant en compte l'effet des processus physiques, comme la circulation des masses d'eau, la remise en suspension et le transport sédimentaire (Hillmer et al., 2008 ; Chung et al., 2014). Un tel modèle serait basé sur des

équations différentiels partielles qui décriraient la dynamique des différentes espèces de la communauté et celle des nutriments limitant la croissance du phytoplancton. Ce modèle serait basé sur les traits fonctionnels des espèces de la communauté, et devrait probablement s'accompagner d'un volet expérimental pour déterminer, pour les différents paramètres, les valeurs relatives aux différents taxons.

Bibliographie

C. Y. Ahn, H. M. Oh, Y. S. Park, Evaluation of Environmental Factors on Cyanobacterial Bloom in Eutrophic Reservoir using Artificial Neural Networks. *Journal of Phycology* **47**, 495-504 (2011).

Y. W. Chen, B. Q. Qin, K. Teubner, M. T. Dokulil, Long-term Dynamics of Phytoplankton Assemblages: Microcystis-domination in Lake Taihu, a Large Shallow Lake in China. *Journal of Plankton Research* **25**, 445-453 (2003).

S. W. Chung, J. Imberger, M. R. Hipsey, H. S. Lee, The Influence of Physical and Physiological Processes on the Spatial Heterogeneity of a Microcystis Bloom in a Stratified Reservoir. *Ecological Modelling* **289**, 133-149 (2014).

K. Cottenie, Integrating Environmental and Spatial Processes in Ecological Community Dynamics. *Ecology Letters* **8**, 1175-1182 (2005).

T. A. Craney, J. G. Surles, Model-Dependent Variance Inflation Factor Cutoff Values. *Quality Engineering* **14**, 391–403 (2002).

J. Deng, B. Qin, H. W. Paerl, Y. Zhang, J. Ma, Y. Chen, Earlier and Warmer Springs Increase Cyanobacterial (Microcystis spp.) Blooms in Subtropical Lake Taihu, China. *Freshwater Biology* **59**, 1076–1085 (2014).

M. T. Dokulil, K. Teubner, Cyanobacterial Dominance in Lakes. *Hydrobiologia* **438**, 1-12 (2000).

S. Dray, A. B. Dufour, The ade4 package: Implementing the Duality Diagram for Ecologists. *R package version 1.7-8* (2017).

S. Dray et al., Community Ecology in the Age of Multivariate Multiscale Spatial Analysis. *Ecological Monographs* **82**, 257-275 (2012).

M. R. Droop, Vitamin B12 and Marine Ecology : (4.) Kinetics of Uptake Growth and Inhibition in Monochrysis Lutheri. *Journal of the Marine Biological Association of the United Kingdom* **48**, 689-& (1968).

H. Ducobu, J. Huisman, R. R. Jonker, L. R. Mur, Competition between a Prochlorophyte and a Cyanobacterium under Various Phosphorus Regimes: Comparison with the Droop Model. Journal of *Phycology* **34**, 467-476 (1998).

K. F. Edwards, C. A. Klausmeier, E. Litchman, A Three-Way Trade-Off Maintains Functional Diversity under Variable Resource Supply. *American Naturalist* **182**, 786-800 (2013).

K. F. Edwards, E. Litchman, C. A. Klausmeier, Functional Traits Explain Phytoplankton Responses to Environmental Gradients across Lakes of the United States. *Ecology* **94**, 1626-1635 (2013).

K. F. Edwards, C. A. Klausmeier, E. Litchman, Nutrient Utilization Traits of Phytoplankton. *Ecology* **96**, 2311 (2015).

K. F. Edwards, M. K. Thomas, C. A. Klausmeier, E. Litchman, Allometric Scaling and Taxonomic Variation in Nutrient Utilization Traits and Maximum Growth rate of Phytoplankton. *Limnology and Oceanography* **57**, 554-566 (2012).

I. R. Falconer, A. R. Humpage, Health Risk Assessment of Cyanobacterial (Blue-green Algal) Toxins in Drinking Water. *International Journal of Environmental Research and Public Health* **2**, 43-50 (2005).

T. Fawcett, An Introduction to ROC Analysis. *Pattern Recognition Letters* **27**, 861–874 (2006).

Y. Gong, H. Y. Ao, B. B. Liu, S. Wen, Z. Wang, D. J. Hu, X. H. Zhang, L. R. Song, J. T. Liu, Effects of Inorganic Arsenic on Growth and Microcystin Production of a *Microcystis* Strain Isolated from an Algal Bloom in Dianchi Lake, China. *Chinese Science Bulletin* **56**, 2337-2342 (2011).

P. Grosjean, F. Ibanez, pastecs: Package for Analysis of Space-Time Ecological Series. *R package version 1.3-18* (2014).

G. M. Hallegraeff, A Review of Harmful Algal Blooms and their Apparent Global Increase. *Phycologia* **32**, 79-99 (1993).

M. O. Hill, A. J. E. Smith, Principal Component Analysis of Taxonomic Data with Multi-State Discrete Characters. *Taxon* **25**, 249-255 (1976).

I. Hillmer, P. van Reenen, J. Imberger, T. Zohary, Phytoplankton Patchiness and their Role in the Modelled Productivity of a Large, Seasonally Stratified Lake. *Ecological Modelling* **218**, 49-59 (2008).

B. C. Hitzfeld, S. J. Hoger, D. R. Dietrich, Cyanobacterial Toxins: Removal during Drinking Water Treatment, and Human Risk Assessment. *Environmental Health Perspectives* **108**, 113-122 (2000).

M. A. Hoque et al., Drinking Water Vulnerability to Climate Change and Alternatives for Adaptation in Coastal South and South East Asia. *Climatic Change* **136**, 247-263 (2016).

G. E. Hutchinson, The Paradox of the Plankton. *American Naturalist* **95**, 137-145 (1961).

D. M. John, Brian A. W., A. J. Brook, The Freshwater Algal Flora of the British Isles: An Identification Guide to Freshwater and Terrestrial Algae. *Cambridge University Press*: Cambridge, 702 pp (2002).

K. D. Jöhnk, J. Huisman, J. Sharples, B. Sommeijer, P. M. Visser, J. M. Stroom, Summer Heatwaves Promote Blooms of Harmful Cyanobacteria. *Global Change Biology* **14**, 495–512 (2008).

C. A. Klausmeier, E. Litchman, S. A. Levin, Phytoplankton growth and stoichiometry under multiple nutrient limitation. Limnology and *Oceanography* **49**, 1463-1470 (2004).

G. W. Kling, G. W. Kipphut, M. M. Miller, W. J. O'Brien, Integration of Lakes and Streams in a Landscape Perspective: the Importance of Material Processing on Spatial Patterns and Temporal Coherence. *Freshwater Biology* **43**, 477-497 (2000).

M. Kuhn, caret: Classification and Regression Training. *R package version 6.0-77* (2017).

P. Legendre, E. D. Gallagher, Ecologically Meaningful Transformations for Ordination of Species Data. *Oecologia* **129**, 271–280 (2001).

P. Legendre, R. Galzin, M. L. Harmelin-Vivien, Relating Behavior to Habitat: Solutions to the Fourth-corner Problem. *Ecology* **78**, 547-562 (1997).

P. Legendre, L. Legendre, Numerical Ecology, 2nd English Edition. *Elsevier Science B. V.*: Amsterdam, 852 pp (1998).

B. Legube, P. Mouchet, Eaux de distribution – Filières de traitement. *Environnement - Sécurité | Technologies de l'eau,* Ref. W5510 (2010).

P. Le Noac'h, Modélisation de la Compétition pour les Nutriments chez le Phytoplancton dans un Contexte d'Apports Variables de la Ressource : Approche par les Traits Fonctionnels. *Rapport de recherhce de Master 2 Modélisation en Ecologie*. Université de Rennes 1. 32 pp (2017).

T. Lindholm, P. Ohman, K. Kurki-Helasmo, B. Kincaid, J. Meriluoto, Toxic Algae and Fish Mortality in a Brackish-water Lake in Angstrom Land, SW Finland. *Hydrobiologia* **397**, 109-120 (1999).

E. Litchman, C. A. Klausmeier, Trait-Based Community Ecology of Phytoplankton. Annual *Review of Ecology Evolution and Systematics* **39**, 615-639 (2008).

E. Litchman, C. A. Klausmeier, O. M. Schofield, P. G. Falkowski, The Role of Functional Traits and Trade-offs in Structuring Phytoplankton Communities: Scaling from Cellular to Ecosystem Level. *Ecology Letters* **10**, 1170-1181 (2007).

I. C. Lou, Z. C. Xie, W. K. Ung, K. M. Mok, Freshwater Algal Bloom Prediction by Extreme Learning Machine in Macau Storage Reservoirs. *Neural Computing & Applications* **27**, 19-26 (2016).

H. R. Maier, G. B. Kingston, T. Clark, A. Frazer, A. Sanderson, Risk-Based Approach for Assessing the Effectiveness of Flow Management in Controlling Cyanobacterial Blooms in Rivers. *River Research and Applications* **20**, 459–471 (2004).

M.A. Maloof, P. Langley, T.O. Binford, R. Nevatia, S. Sage, Improved Rooftop Detection in Aerial Images with Machine Learning. *Machine Learning* **53**, 157–191 (2003).

B. Marie et al., Effects of a Toxic Cyanobacterial Bloom (*Planktothrix agardhii*) on Fish: Insights from Histopathological and Quantitative Proteomic Assessments Following the Oral Exposure of Medaka Fish (*Oryzias latipes*). *Aquatic Toxicology* **114**, 39-48 (2012).

B. J. McGill, B. J. Enquist, E. Weiher, M. Westoby, Rebuilding Community Ecology from Functional Traits. *Trends in Ecology & Evolution* **21**, 178-185 (2006).

S. Merel et al., State of Knowledge and Concerns on Cyanobacterial Blooms and Cyanotoxins. *Environment International* **59**, 303-327 (2013).

A. M. Michalak et al., Record-setting Algal Bloom in Lake Erie Caused by Agricultural and Meteorological Trends Consistent with Expected Future Conditions. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 6448-6452 (2013).

J. Oksanen, F. G. Blanchet, M. Friendly, R. Kindt, P. Legendre, D. McGlinn, P. R. Minchin, R. B. O'Hara, G. L. Simpson, P. Solymos, M. H. H. Stevens, E. Szoecs and H. Wagner, vegan: Community Ecology Package. R package version 2.4-5. (2017).

I. Olenina, S. Hajdu, L. Edler, A. Andersson, Biovolumes and Size-classes of Phytoplankton in the Baltic Sea. *Baltic Sea Environnemental Proceedings* **106** (2006).

J. M. O'Neil, T. W. Davis, M. A. Burford, C. J. Gobler, The Rise of Harmful Cyanobacteria Blooms: The Potential Roles of Eutrophication and Climate Change. *Harmful Algae* **14**, 313-334 (2012).

H. W. Paerl, R. S. Fulton, P. H. Moisander, J. Dyble, Harmful Freshwater Algal Blooms, with an Emphasis on Cyanobacteria. *The Scientific World* **1**, 76-113 (2001).

H. W. Paerl, J. Huisman, Climate Change: a Catalyst for Global Expansion of Harmful Cyanobacterial Blooms. *Environmental Microbiology Reports* **1**, 27-37 (2009).

H. W. Paerl, J. L. Pinckney, J. M. Fear, B. L. Peierls, Ecosystem Responses to Internal and Watershed Organic Matter Loading: Consequences for Hypoxia in the Eutrophying Neuse River Estuary, North Carolina, USA. *Marine Ecology Progress Series* **166**, 17-25 (1998).

T. Posch, O. Koster, M. M. Salcher, J. Pernthaler, Harmful Filamentous Cyanobacteria Favoured by Reduced Water Turnover with Lake Warming. *Nature Climate Change* **2**, 809-813 (2012).

B. Q. Qin et al., A Drinking Water Crisis in Lake Taihu, China: Linkage to Climatic Variability and Lake Management. *Environmental Management* **45**, 105-112 (2010).

R Core Team, R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*, Vienna, Austria (2017)

C. S. Reynolds, V. Huszar, C. Kruk, L. Naselli-Flores, S. Melo, Towards a Functional Classification of the Freshwater Phytoplankton. *Journal of Plankton Reasearch* **24**, 417-728 (2002).

C. S. Reynolds, 2006. The Ecology of Phytoplankton. *Cambridge University Press*: Cambridge, 551 pp (2006).

S. Romo, J. Soria, F. Fernandez, Y. Ouahid, A. Baron-Sola, Water Residence Time and the Dynamics of Toxic Cyanobacteria. *Freshwater Biology* **58**, 513-522 (2013).

J. J. Rosso, N. F. Schenone, A. P. Carrera, A. F. Cirelli, Concentration of Arsenic in Water, Sediments and Fish Species from Naturally Contaminated Rivers. *Environmental Geochemistry and Health* **35**, 201–214 (2013).

R. E. Schapire, A Brief Introduction to Boosting. Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, 1-6 (1999).

K. Sivonen, Cyanobacterial toxins and toxin production. *Phycologia* **35**, 12-24 (1996).

V. H. Smith, G. D. Tilman, J. C. Nekola, Eutrophication: Impacts of Excess Nutrient Inputs on Freshwater, Marine, and Terrestrial Ecosystems. Environmental Pollution **100**, 179–196 (1999).

D. Tilman, Resource Competition between Planktonic Algae - Experimental and Theoretical Approach. *Ecology* **58**, 338-348 (1977).

C. Violle et al., Let the Concept of Trait be Functional! *Oikos* **116**, 882-892 (2007).

C. Wagner, R. Adrian, Cyanobacteria Dominance: Quantifying the Effects of Climate Change. *Limnology and Oceanography* **54**, 2460–2468 (2009).

T. F. Wu et al., Dynamics of Cyanobacterial Bloom Formation During Short-term Hydrodynamic Fluctuation in a Large Shallow, Eutrophic, and Wind-exposed Lake Taihu, China. *Environmental Science and Pollution Research* **20**, 8546-8556 (2013).

Z. C. Xie, I. Lou, W. K. Ung, K. M. Mok, Freshwater Algal Bloom Prediction by Support Vector Machine in Macau Storage Reservoirs. *Mathematical Problems in Engineering* (2012).

Y. Zhang, E. E. Prepas, Regulation of the Dominance of Planktonic Diatoms and Cyanobacteria in Four Eutrophic Hardwater Lakes by Nutrients, Water Column Stability, and Temperature. *Canadian Journal of Fisheries and Aquatic Sciences* **53**, 621-633 (1996).

M. Zhang, H. Duan, X. Shi, Y. Yu, F. Kong, Contributions of Meteorology to the Phenology of Cyanobacterial Blooms: Implications for Future Climate Change. *Water Research* **46**, 442-452 (2012).

Annexe I : Liste des taxons retenus pour les analyses statistiques. Le groupe taxonomique d'appartenance est aussi précisé.

Taxons	Groupes
Achnanthes	Diatoms
Anabaena	Cyanobacteriae
Ankistrodesmus	Chlorophyceae
Ankyra	Chlorophyceae
Aphanizomenon flos-aquae	Cyanobacteriae
Aphanizomenon gracile	Cyanobacteriae
Aphanizomenon issatschenkoi	Cyanobacteriae
Aphanocapsa	Cyanobacteriae
Aphanothece	Cyanobacteriae
Asterionella	Diatoms
Asterionella formosa	Diatoms
Aulacoseira	Diatoms
Aulacoseira granulata	Diatoms
Chlamydomonas	Chlorophyceae
Chlorella	Chlorophyceae
Chromulina	Chrysophyceae
Closteriopsis	Trebouxiophyceae
Closterium	Chlorophyceae
Coelastrum	Chlorophyceae
Coelomoron	Cyanobacteriae
Cosmarium	Chlorophyceae
Crucigenia	Chlorophyceae
Cryptomonas	Cryptophyceae
Cyanodictyon	Cyanobacteriae
Cyclotella	Diatoms
Cymbella	Diatoms
Diatoma elongatum	Diatoms
Dictyospherium	Chlorophyceae
Eudorina	Chlorophyceae
Euglena	Euglenophyceae
Fragilaria	Diatoms
Fragilaria crotonensis	Diatoms
Kephyrion	Chrysophyceae
Kirchneriella	Chlorophyceae
Limnothrix redekei	Cyanobacteriae
Mallomonas	Chrysophyceae
Melosira	Diatoms

Taxons	Groupes
Micractinium	Chlorophyceae
Microcystis aeruginosa	Cyanobacteriae
Microcystis botrys	Cyanobacteriae
Microcystis flos-aquae	Cyanobacteriae
Microcystis viridis	Cyanobacteriae
Monoraphidium	Chlorophyceae
Navicula	Diatoms
Nitzschia	Diatoms
Oocystis	Trebouxiophyceae
Pediastrum	Chlorophyceae
Pediastrum boryanum	Chlorophyceae
Pediastrum clathratum	Chlorophyceae
Pediastrum tetras	Dinophyceae
Phacotus lenticularis	Chlorophyceae
Phormidium	Cyanobacteriae
Planktolyngbya	Cyanobacteriae
Planktothrix agardhii	Cyanobacteriae
Pseudanabaena catenata	Cyanobacteriae
Pseudanabaena limnetica	Cyanobacteriae
Pseudanabaena mucicola	Cyanobacteriae
Rhizosolenia longiseta	Diatoms
Scenedesmus	Chlorophyceae
Selenastrum	Chlorophyceae
Snowella	Cyanobacteriae
Snowella lacustris	Cyanobacteriae
Sphaerocystis	Chlorophyceae
Staurastrum	Charophyta
Stephanodiscus	Diatoms
Stephanodiscus astrea	Diatoms
Synedra	Diatoms
Synura	Synurophyceae
Tetraedron	Chlorophyceae
Tetrastrum	Chlorophyceae
Trachelomonas	Euglenophyceae
Woronichinia	Cyanobacteriae
Woronichinia compacta	Cyanobacteriae
Woronichinia naegeliana	Cyanobacteriae

Annexe II : Liste des variables environnementales utilisées dans les analyses statistiques.

Variable environnementale	Unité
Agents de surface anioniques	µg/L
Aluminium	µg/L
Ammonium	mg(NH4)/L
Arsenic	µg/L
Atrazine	µg/L
Carbone organique	mg(C)/L
Chlorophylle a	µg/L
Chlorures	mg/L
Conductivité	µS/cm
Demande chimique en oxygène (DCO)	mg(O2)/L
Demande biochimique en oxygène (DBO5)	mg(O2)/L
Dureté totale	°f
Fer	mg(Fe)/L
Glyphosate	µg/L
Isoproturon	µg/L
Magnésium	mg/L
Manganèse	µg(Mn)/L
Matières en suspension	mg/L
Nickel	µg/L
Nitrates	mg(NO3)/L
Nitrites	mg(NO2)/L
Oxygène dissous	mg(O2)/L
Phosphore total	mg(P)/L
Potassium	mg/L
pH	
Silice	mg/L
Sodium	mg/L
Somme des pesticides analysés	µg/L
Sulfates	mg(SO4)/L
Température de l'Eau	°C
Titre alcalimétrique complet	°f
Turbidité Néphélométrique	NFU
2-hydroxy atrazine	µg/L
Zinc	mg/L
Débit moyen quotidien (Rance)	m3/s
Débit moyen quotiden (Néal)	m3/s
Température moyenne de l'air	°C
Vitesse moyenne quotidienne du vent	m/s
Rayonnement global quotidien	Joules/cm2
Précipitations quotidiennes	mm
Volume d'eau stocké dans la retenue	m3
Temps de séjour de l'eau	jours

Annexe III : Matrice des traits fonctionnels morphologiques utilisée dans la *fourth corner analysis*. Le trait **Rapport S/V** correspond au rapport surface/volume du taxon considéré (ici présenté de manière catégorielle).

Taxons	Groupes	formeInd	Flagelle	Mucilage	Vacuole gazeuse	Silice	Biovolume	Biovolume (catégories)	Rapport S/V (catégories)	Coefficient de forme (catégories)
Achnanthes	Diatoms	Petites colonies	N	Y	N	Y	137.44	medium	low	medium
Anabaena	Cyanobacteriae	Grandes colonies	N	Y	Y	N	26.1799388	low	high	high
Ankistrodesmus	Chlorophyceae	Petites colonies	N	Y	N	N	81.3323432	medium	medium	medium
Ankyra	Chlorophyceae	Unicellulaire	N	N	N	N	183	medium	medium	medium
Aphanizomenon flos-aquae	Cyanobacteriae	Grandes colonies	N	N	Y	N	180.64	medium	high	high
Aphanizomenon gracile	Cyanobacteriae	Grandes colonies	N	N	Y	N	180.64	medium	high	high
Aphanizomenon issatschenkoi	Cyanobacteriae	Grandes colonies	N	N	Y	N	180.64	medium	high	high
Aphanocapsa	Cyanobacteriae	Grandes colonies	N	Y	N	N	4.18879021	low	high	high
Aphanothece	Cyanobacteriae	Grandes colonies	N	Y	N	N	4.18879021	low	high	high
Asterionella	Diatoms	Petites colonies	N	N	N	Y	612.5	high	medium	high
Asterionella formosa	Diatoms	Petites colonies	N	N	N	Y	245	medium	high	high
Aulacoseira	Diatoms	Grandes colonies	N	Y	N	Y	769.69	high	medium	high
Aulacoseira granulata	Diatoms	Grandes colonies	N	N	N	Y	351.86	high	medium	high
Chlamydomonas	Chlorophyceae	Unicellulaire	Y	Y	N	N	45	low	low	low
Chlorella	Chlorophyceae	Unicellulaire	N	Y	N	N	91.9	medium	low	low
Chromulina	Chrysophyceae	Unicellulaire	Y	N	N	N	26.2	low	low	low
Closteriopsis	Trebouxiophyceae	Unicellulaire	N	Y	N	N	70.69	medium	medium	medium
Closterium	Chlorophyceae	Unicellulaire	N	N	N	N	1172.86	high	medium	medium
Coelastrum	Chlorophyceae	Petites colonies	N	N	N	N	291.83	medium	low	low
Coelomoron	Cyanobacteriae	Grandes colonies	N	Y	N	N	4.18879021	low	high	high
Cosmarium	Chlorophyceae	Unicellulaire	N	Y	N	N	2356.19	high	low	low
Crucigenia	Chlorophyceae	Petites colonies	Y	N	N	N	447.17	high	low	low
Cryptomonas	Cryptophyceae	Unicellulaire	Y	Y	N	N	1112.125	high	medium	medium
Cyanodictyon	Cyanobacteriae	Unicellulaire	N	N	Y	N	4.18879021	low	high	low
Cyclotella	Diatoms	Unicellulaire	N	N	N	Y	2513.27	high	low	medium
Cymbella	Diatoms	Unicellulaire	N	Y	N	Y	250	medium	medium	medium
Diatoma elongatum	Diatoms	Unicellulaire	N	Y	N	Y	960	high	medium	medium
Dictyospherium	Chlorophyceae	Grandes colonies	N	Y	N	N	292.56	medium	low	low
Eudorina	Chlorophyceae	Grandes colonies	Y	Y	N	N	590.14	high	low	low
Euglena	Euglenophyceae	Unicellulaire	Y	Y	N	N	21903.8833	high	low	medium
Fragilaria	Diatoms	Grandes colonies	N	N	N	Y	942.48	high	low	high
Fragilaria crotonensis	Diatoms	Grandes colonies	N	N	N	Y	245.5	medium	low	high
Kephyrion	Chrysophyceae	Unicellulaire	N	N	N	N	63.62	medium	low	low
Kirchneriella	Chlorophyceae	Grandes colonies	N	Y	N	N	335.1	medium	low	high
Limnothrix redekei	Cyanobacteriae	Grandes colonies	N	N	Y	N	6.5449847	low	high	high
Mallomonas	Chrysophyceae	Unicellulaire	Y	N	N	N	3161.84	high	low	medium
Melosira	Diatoms	Grandes colonies	N	Y	N	Y	1178.1	high	low	high

Taxons	Groupes	formeInd	Flagelle	Mucilage	Vacuole gazeuse	Silice	Biovolume	Biovolume (catégories)	Rapport S/V (catégories)	Coefficient de forme (catégories)
Micractinium	Chlorophyceae	Petites colonies	N	Y	N	N	54.5	low	low	medium
Microcystis aeruginosa	Cyanobacteriae	Grandes colonies	N	Y	Y	N	26.1799388	low	high	high
Microcystis botrys	Cyanobacteriae	Grandes colonies	N	Y	Y	N	22.8056356	low	high	high
Microcystis flos-aquae	Cyanobacteriae	Grandes colonies	N	Y	Y	N	16.7551608	low	high	high
Microcystis viridis	Cyanobacteriae	Grandes colonies	N	Y	Y	N	26.1799388	low	high	high
Monoraphidium	Chlorophyceae	Unicellulaire	N	N	N	N	103.846667	medium	medium	medium
Navicula	Diatoms	Petites colonies	N	Y	N	Y	2356.19	high	low	medium
Nitzschia	Diatoms	Unicellulaire	N	Y	N	Y	225	medium	medium	medium
Oocystis	Trebouxiophyceae	Petites colonies	N	Y	N	N	148.876667	medium	medium	medium
Pediastrum	Chlorophyceae	Grandes colonies	N	N	N	N	366.8	high	medium	high
Pediastrum boryanum	Chlorophyceae	Grandes colonies	N	N	N	N	350	high	medium	high
Pediastrum clathratum	Chlorophyceae	Grandes colonies	N	N	N	N	350	high	medium	high
Pediastrum tetras	Dinophyceae	Grandes colonies	N	N	N	N	58315.81	high	medium	high
Phacotus lenticularis	Chlorophyceae	Unicellulaire	Y	Y	N	N	150.67	medium	low	low
Phormidium	Cyanobacteriae	Grandes colonies	N	Y	Y	N	31.42	low	high	high
Planktolyngbya	Cyanobacteriae	Grandes colonies	N	N	Y	N	15.71	low	high	high
Planktothrix agardhii	Cyanobacteriae	Grandes colonies	N	Y	Y	N	58.9	low	high	high
Pseudanabaena catenata	Cyanobacteriae	Grandes colonies	N	N	Y	N	35.34	low	medium	high
Pseudanabaena limnetica	Cyanobacteriae	Grandes colonies	N	N	Y	N	35.34	low	medium	high
Pseudanabaena mucicola	Cyanobacteriae	Grandes colonies	N	N	Y	N	35.34	low	high	high
Rhizosolenia longiseta	Diatoms	Unicellulaire	N	N	N	Y	5119.9779	high	low	medium
Scenedesmus	Chlorophyceae	Petites colonies	N	N	N	N	270.178	medium	medium	medium
Selenastrum	Chlorophyceae	Petites colonies	N	Y	N	N	141.37	medium	medium	medium
Snowella	Cyanobacteriae	Grandes colonies	N	Y	N	N	6.5449847	low	high	high
Snowella lacustris	Cyanobacteriae	Grandes colonies	N	Y	N	N	14.7262156	low	high	high
Sphaerocystis	Chlorophyceae	Petites colonies	N	Y	N	N	221	medium	low	low
Staurastrum	Charophyta	Unicellulaire	N	Y	N	N	9632.91	high	medium	medium
Stephanodiscus	Diatoms	Unicellulaire	N	N	N	Y	744.165	high	low	medium
Stephanodiscus astrea	Diatoms	Unicellulaire	N	N	N	Y	744.165	high	low	medium
Synedra	Diatoms	Grandes colonies	N	N	N	Y	1500	high	medium	high
Synura	Synurophyceae	Grandes colonies	Y	Y	N	N	942.48	high	medium	high
Tetraedron	Chlorophyceae	Unicellulaire	N	N	N	N	187.5	medium	medium	medium
Tetrastrum	Chlorophyceae	Petites colonies	N	Y	N	N	163.6	medium	low	medium
Trachelomonas	Euglenophyceae	Unicellulaire	Y	Y	N	N	1114	high	low	low
Woronichinia	Cyanobacteriae	Grandes colonies	N	Y	Y	N	9.42477796	low	high	high
Woronichinia compacta	Cyanobacteriae	Grandes colonies	N	Y	Y	N	22.0893234	low	high	high
Woronichinia naegeliana	Cyanobacteriae	Grandes colonies	N	Y	Y	N	35.3429174	low	high	high

Annexe IV : Matrice des traits fonctionnels de nutrition. Les valeurs imprimées en rouge correspondent aux valeurs de traits estimés à partir du biovolume cellulaire du taxon considéré à l'aide des relation allométriques (voir Annexe 5).

Taxon	Taux de croissance maximal (μ_{max})	Affinité (Azote)	Affinité (Phosphore)	Réserve (Azote)	Réserve (Phosphore)
Achnanthes	1.15	1 32E-02	1 35E-01	2 88E+00	1 14E+01
Anabaena	1.28	7 85E-01	1 85E+02	2 28E+00	5 95E+00
Ankistrodesmus	1.54	1 37E+00	2 65E+00	2 59E+00	2 36E+01
Ankyra	0.87	1 77E-02	1 19E-01	2 85E+00	1 20E+01
Aphanizomenon flos-aquae	0.87	1 74E-02	6 42E+02	2 85E+00	4 58E+03
Aphanizomenon gracile	45.00	1 74E-02	6 42E+02	2 85E+00	4 58E+03
Aphanizomenon issatschenkoi	0.87	1 74E-02	6 42E+02	2 85E+00	4 58E+03
Aphanocapsa	1.86	3 73E-04	6 30E-01	3 32E+00	5 93E+00
Aphanothece	1.86	4 85E-03	6 30E-01	4 32E+01	5 93E+00
Asterionella	0.68	7 14E+00	7 72E+01	1 68E+00	3 59E+00
Asterionella formosa	0.82	1 55E+01	7 72E+01	3 09E+00	3 59E+00
Aulacoseira	0.65	7 67E-02	6 32E-02	2 69E+00	1 57E+01
Aulacoseira granulata	0.76	3 45E-02	8 93E-02	2 78E+00	1 35E+01
Chlamydomonas	1.15	4 22E-03	8 33E+02	3 02E+00	9 95E+01
Chlorella	1.00	6 24E-01	5 02E+01	2 09E+02	1 77E+02
Chromulina	1.28	2 43E-03	2 81E-01	3 08E+00	8 35E+00
Closteriopsis	1.05	6 69E-03	1 81E-01	2 96E+00	1 00E+01
Closterium	0.59	1 18E-01	5 25E-02	2 64E+00	1 69E+01
Coelastrum	0.70	4 05E-01	1 47E+01	2 42E+00	6 11E+00
Coelomoron	1.86	3 73E-04	6 30E-01	3 32E+00	5 93E+00
Cosmarium	0.60	2 41E-01	1 90E+00	2 57E+00	8 64E+01
Crucigenia	0.72	4 41E-02	8 03E-02	2 75E+00	1 42E+01
Cryptomonas	0.60	1 12E-01	2 46E+00	2 65E+00	7 67E+02
Cyanodictyon	1.86	3 73E-04	6 30E-01	3 32E+00	5 93E+00
Cyclotella	1.06	6 32E+00	2 74E-01	6 31E+01	8 42E+02
Cymbella	0.18	2 43E-02	1 04E-01	2 82E+00	1 27E+01
Diatoma elongatum	0.62	9 61E-02	2 12E+01	2 67E+00	2 44E+03
Dictyospherium	0.63	1 08E-01	3 52E+01	1 50E+00	6 97E+00
Eudorina	0.68	5 85E-02	7 11E-02	2 72E+00	1 49E+01
Euglena	0.38	2 35E+00	5 17E+00	2 35E+00	2 41E+03
Fragilaria	0.62	6 13E+00	4 60E+01	3 28E+00	5 79E+00
Fragilaria crotonensis	0.82	2 18E+00	1 36E+01	3 17E+00	5 83E+00
Kephyrion	1.07	6 01E-03	1 90E-01	2 98E+00	9 84E+00
Kirchneriella	0.77	3 28E-02	9 12E-02	2 78E+00	1 34E+01
Limnothrix redekei	1.70	5 88E-04	2 62E+00	3 26E+00	1 21E+01
Mallomonas	0.49	3 25E-01	3 39E-02	2 54E+00	2 04E+01
Melosira	0.46	1 19E-01	5 24E-02	2 64E+00	1 69E+01

Taxon	Taux de croissance maximal (μ_{max})	Affinité (Azote)	Affinité (Phosphore)	Réserve (Azote)	Réserve (Phosphore)
Miracidium	1.11	5 13E-03	2 03E-01	2 99E+00	9 56E+00
Microcystis aeruginosa	0.67	3 76E-03	2 36E+02	4 77E+00	1 20E+01
Microcystis botrys	1.15	5 58E-03	2 50E+00	8 21E+00	9 02E+00
Microcystis flos-aquae	1.15	3 17E-03	2 50E+00	6 48E+00	9 02E+00
Microcystis viridis	1.15	7 18E-03	2 50E+00	9 13E+00	9 02E+00
Monoraphidium	0.87	2 99E-01	4 06E-01	2 33E+00	1 00E+01
Navicula	0.68	3 54E-02	1 33E+01	2 57E+00	1 12E+04
Nitzschia	0.53	5 95E+00	1 28E+02	6 74E+01	8 61E+01
Oocystis	0.52	1 43E-02	1 67E+00	2 87E+00	1 48E+02
Pediastrum	0.63	2 49E-01	1 07E+01	1 44E+00	6 18E+00
Pediastrum boryanum	0.76	7 76E+00	2 70E+02	1 44E+00	6 20E+00
Pediastrum tetras	0.63	2 49E-01	1 07E+01	1 44E+00	6 18E+00
Peridinium	0.27	6 38E+00	9 38E-03	2 26E+00	3 50E+01
Phacotus lenticularis	0.90	1 45E-02	1 30E-01	2 87E+00	1 16E+01
Phormidium	2.72	2 92E-03	2 59E-01	3 06E+00	8 63E+00
Planktolyngbya	1.42	1 44E-03	3 52E-01	3 15E+00	7 59E+00
Planktothrix agardhii	0.50	8 50E-02	1 90E+00	2 98E+00	9 70E+00
Pseudanabaena catenata	1.21	3 30E-03	2 46E-01	3 05E+00	8 82E+00
Pseudanabaena limnetica	1.21	3 30E-03	2 46E-01	3 05E+00	8 82E+00
Pseudanabaena mucicola	1.21	3 30E-03	2 46E-01	3 05E+00	8 82E+00
Rhizosolenia longiseta	0.44	5 32E-01	2 74E-02	2 49E+00	2 23E+01
Scenedesmus	0.80	1 09E+00	9 12E+01	2 17E+00	5 52E+00
Selenastrum	1.48	4 17E-01	5 22E+00	4 76E+00	1 52E+01
Snowella	1.70	5 88E-04	5 17E-01	3 26E+00	6 45E+00
Snowella lacustris	1.44	1 35E-03	3 62E-01	3 16E+00	7 50E+00
Sphaerocystis	0.48	2 14E-02	4 22E+00	2 83E+00	4 78E+02
Staurastrum	0.39	1 01E+00	1 79E+01	2 43E+00	1 54E+01
Stephanodiscus	0.65	7 41E-02	1 23E-01	2 69E+00	2 99E+01
Stephanodiscus astrea	0.65	7 41E-02	1 23E-01	2 69E+00	2 99E+01
Synedra	0.29	2 18E+00	6 99E+02	3 76E+01	6 20E+02
Synura	0.62	9 44E-02	5 78E-02	2 67E+00	1 63E+01
Tetraedron	0.86	1 81E-02	1 18E-01	2 85E+00	1 20E+01
Tetrastrum	0.89	1 58E-02	1 25E-01	2 86E+00	1 17E+01
Trachelomonas	0.60	1 12E-01	5 37E-02	2 65E+00	1 68E+01
Woronichinia	1.58	8 54E-04	4 40E-01	3 21E+00	6 90E+00
Woronichinia compacta	1.33	2 04E-03	3 02E-01	3 11E+00	8 08E+00
Woronichinia naegeliana	1.21	3 30E-03	2 46E-01	3 05E+00	8 82E+00

Annexe V : Relations allométriques liant les valeurs des paramètres du modèle de Droop au volume cellulaire V_{cell} d'une cellule (en µm^3). La relation pour un paramètre X est telle que : $log_{10}(X) = a \times log_{10}(V_{cell}) + b$.

Paramètres	Source	a	b	n	R^2	*p-value*
μ_{max}	Edwards et al. (2015)	-0.20	-0.98	60	0.20	$< 10^{-3}$
V_{maxN}^{hi}	Edwards et al. (2015)	2.7	-11.78	13	0.43	0.015
K_N	Edwards et al. (2015)	0.88	-0.07	14	0.02	0.58
V_{maxP}^{hi}	Edwards et al. (2015)	0.88	-8.43	41	0.47	$< 10^{-6}$
K_P	Edwards et al. (2015)	0.48	0.29	67	0.05	0.066
Q_{minN}	Edwards et al. (2015)	0.81	-7.66	22	0.63	$< 10^{-5}$
Q_{maxN}	Edwards et al. (2015)	0.77	-7.11	14	0.27	0.059
Q_{minP}	Edwards et al. (2015)	0.83	-8.8	52	0.72	$< 10^{-14}$
Q_{maxP}	Edwards et al. (2015)	1.02	-8.15	28	0.65	$< 10^{-6}$

Les traits de nutrition d'*Affinité* et de *Réserve* pour une ressource R sont calculés à partir des paramètres du modèle de Droop à l'aide des formules suivantes :

- $R_R = \frac{Q_{maxR}}{Q_{minR}}$ **(Ducobu et al., 1998)**
- $Aff_R = \frac{V_{maxR}^{hi}}{K_R \times Q_{minR}}$ **(Edwards et al., 2012)**

Annexe VI : Tableau récapitulatif des valeurs testées pour les hyperparamètres des différents algorithmes de *Machine Learning*.

Algorithme	Méthode (caret)	Hyperparamètre	Valeur
SVM linear	*svmLinear*	*C*	1;10;100;1000;5000;10000
SVM polynomial	*svmPoly*	*C*	1;10;100;1000;5000;10000
		degree	1;2;3
		scale	1,2
knn	*knn*	*k*	$1 \leq k \leq 40$ (20 valeurs)
RIDGE	*glmnet*	λ	$e^{-5} \leq \lambda \leq e^{1}$ (50 valeurs)
		α	0
LASSO	*glmnet*	λ	$e^{-5} \leq \lambda \leq e^{1}$ (50 valeurs)
		α	1
Random Forest	*rf*	*mtry*	2;3
		ntree	500 (par défaut)
Neural Networks	*nnet*	*size*	$1 \leq size \leq 20$ (20 valeurs)
		decay	0.5;0.1; 10^{-2}; 10^{-3}; 10^{-4}; 10^{-5}; 10^{-6}; 10^{-7}

Diplôme : **M2**

Spécialité : **Agronomie**

Spécialisation / option : **Data Science pour la Biologie**

Enseignant référent : **David CAUSEUR**

Auteur(s) : **Philippe LE NOAC'H**

Date de naissance* : **16 septembre 1994**

Nb pages : 33 Annexe(s) : 8

Année de soutenance : **2018**

Organisme d'accueil :

UMR 6553 ECOBIO

Adresse : Université de Rennes 1

Avenue du Général Leclerc

Campus de Beaulieu

35042 RENNES Cedex – France

Maîtres de stage : **Yvan LAGADEUC**

Alexandrine PANNARD

Titre français : *Analyse statistique des facteurs environnementaux favorisant les efflorescences de cyanobactéries dans la retenue de Rophémel*

Titre anglais : *Statistical analysis of the environmental factors promoting cyanobacterial blooms in Rophémel reservoir*

Résumé (1600 caractères maximum) : Une efflorescence de phytoplancton est définie comme une augmentation exponentielle de la biomasse d'une espèce phytoplanctonique dans un plan d'eau, souvent à la suite d'apport de nutriments (N et P) d'origine anthropique. Les efflorescences de cyanobactéries toxiques constituent un problème majeur pour le fonctionnement des écosystèmes aquatiques et la production d'eau potable. Au cours de notre étude, nous avons cherché à caractériser la dynamique d'apparition des efflorescences phytoplanctonique dans la retenue de Rophémel, un plan d'eau eutrophe faisant office de réservoir pour l'approvisionnement en eau potable de la ville de Rennes. Nous avons analysé des données de comptage de la communauté phytoplanctonique du plan d'eau, et nous avons étudier des chroniques de différentes variables physico-chimiques, météorologiques et hydrologiques mesurées entre 2006 et 2016. Nous avons notamment cherché à identifier les variables environnementales susceptibles de jouer un rôle dans l'apparition des efflorescences à l'aide de méthodes d'analyse à tableaux multiples (RDA, *fourth-corner analysis*). La majorité des efflorescences est causée par deux taxons de cyanobactéries, *Microcystis* et *Planktothrix*. Nous avons identifié un rôle possible de l'Arsenic qui pourrait privilégier les espèces de *Microcystis* au détriment des autres taxons, même à faibles concentrations. En nous basant sur les traits fonctionnels des taxons de la communauté étudiée, nous avons aussi mis en évidence le rôle de l'hydrodynamisme et de la température de l'eau dans la formation des blooms, ce que la littérature confirme. Nous recommandons des changements dans l'échantillonnage des variables environnementales pour la mise au point d'un modèle de type *Machine Learning* capable de prédire efficacement l'apparition des événements d'efflorescences dans la retenue.

Abstract (1600 caractères maximum): A bloom is defined as an exponential increase of the biomass of a given phytoplankton species in a water body, often linked to nutrients discharges caused by human activities. Harmful cyanobacterial blooms are a major threat for the structure of aquatic ecosystems and the production of drinkable water. In this study, we shed light on the occurrence of phytoplankton blooms in the reservoir of Rophémel, an eutrophic water body used to supply the city of Rennes (France) with drinking water. We analysed species counts of the phytoplankton community of the reservoir, and we studied records of various physical, chemical, hydrological and meteorological variables measured in and around the reservoir between 2006 ad 2016. Using different methods of analysis applied to multiple data tables (RDA, *fourth-corner* analysis), we set out to identify the variables most susceptible to favour bloom outbreaks. Most of the blooms recorded are caused by two cyanobacterial taxons, *Microcystis* and *Planktothrix*. We suspect that Arsenic levels in the reservoir might favor *Microcystis* against the rest of the community, even a low concentration. Using a functional traits-based approach, we highlighted the role of hydrodynamics and water temperature, in agreement with the literature on the subject. We recommend some modifications in the sampling of environmental variables to produce an efficient *Machine Learning* predictive model able to forecast efficiently bloom outbreaks in the reservoir.

Mots-clés : blooms, séries temporelles, analyses factorielles, Machine Learning

Key Words: blooms, time series, factorial analysis, Machine Learning

* *Elément qui permet d'enregistrer les notices auteurs dans le catalogue des bibliothèques universitaires*