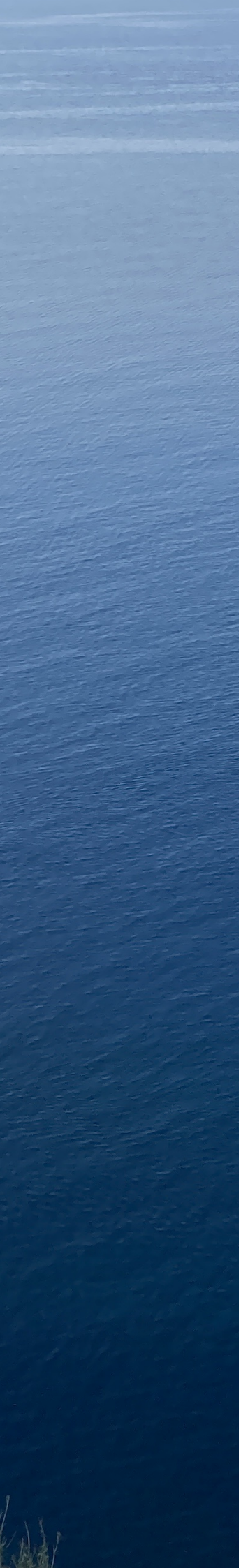
Plankton distribution at submesoscale: contributions from artificial intelligence to plankton ecology

Thelma Panaïotis



École Doctorale des Sciences de l'Environnement d'Île de France

Laboratoire d'Océanographie de Villefranche – UMR 7093

DISTRIBUTION DU PLANCTON À SUBMÉSOÉCHELLE: APPORT DE L'INTELLIGENCE ARTIFICIELLE POUR L'ÉCOLOGIE PLANCTONIQUE

Thelma Panaïotis

Sous la direction de
Jean-Olivier Irisson

Thèse pour l'obtention du grade de
Docteur de Sorbonne Université
Spécialité : Écologie

Soutenue le 06 avril 2023

Devant un jury composé de

Emmanuel Boss	Reviewer
Francesco Pomati	Reviewer
Lee Karp-Boss	Examinator
Louis Legendre	Examinator
Marina Lévy	Examinator
Jean-Olivier Irisson	PhD Supervisor

À mon être humain préféré.

I am just a child who has never grown up. I still keep asking these 'how' and 'why' questions. Occasionally, I find an answer.

— Stephen Hawking, A Brief History of Time

We have seen that computer programming is an art, because it applies accumulated knowledge to the world, because it requires skill and ingenuity, and especially because it produces objects of beauty.

— Donald E. Knuth

Sorbonne Université

École Doctorale des Sciences de l'Environnement d'Île de France
Laboratoire d'Océanographie de Villefranche - UMR 7093

Distribution du plancton à sub-mésoéchelle : apport de l'intelligence artificielle pour l'écologie planctonique

Thelma Panaïotis

Sous la direction de
Jean-Olivier Irisson

Thèse pour l'obtention du grade de
Docteur de Sorbonne Université

Spécialité : Écologie

Soutenue le 06 avril 2023

Devant un jury composé de

Emmanuel Boss	Rapporteur
Francesco Pomati	Rapporteur
Lee Karp-Boss	Examinatrice
Louis Legendre	Examinateur
Marina Lévy	Examinatrice
Jean-Olivier Irisson	Directeur de thèse

Résumé En tant que base des réseaux trophiques océaniques et élément clé de la pompe à carbone biologique, les organismes planctoniques jouent un rôle majeur dans les océans. Cependant, leur distribution à petite échelle, régie par les interactions biotiques entre organismes et les interactions avec les propriétés physico-chimiques des masses d'eau de leur environnement immédiat, est mal décrite *in situ*, en raison du manque d'outils d'observation adaptés. De nouveaux instruments d'imagerie *in situ* à haute résolution, combinés à des algorithmes d'apprentissage automatique pour traiter la grande quantité de données collectées, nous permettent aujourd'hui d'aborder ces échelles.

La première partie de ce travail se concentre sur le développement méthodologique de deux pipelines automatisés basés sur l'intelligence artificielle. Ces pipelines ont permis de détecter efficacement les organismes planctoniques au sein des images brutes, et de les classer en catégories taxonomiques ou morphologiques. Dans une deuxième partie, des outils d'écologie numérique ont été appliqués pour étudier la distribution du plancton à différentes échelles, en utilisant trois jeux de données d'imagerie *in situ*. Tout d'abord, nous avons mis en évidence un lien entre les communautés planctoniques et les conditions environnementales à l'échelle globale. Ensuite, nous avons décrit la distribution du plancton et des particules à travers un front de mésoéchelle, et mis en évidence des périodes contrastées pendant le bloom de printemps. Enfin, grâce aux données d'imagerie *in situ* à haute fréquence, nous avons étudié la distribution à fine échelle et la position préférentielle d'organismes appartement au groupe des Rhizaria, des protistes fragiles et peu étudiés, dont certains sont mixotrophes.

Dans l'ensemble, ce travail démontre l'efficacité de l'imagerie *in situ* combinée à des approches d'intelligence artificielle pour comprendre les interactions biophysiques dans le plancton et les conséquences sur sa distribution à petite échelle.

Sorbonne Université

École Doctorale des Sciences de l'Environnement d'Île de France
Laboratoire d'Océanographie de Villefranche - UMR 7093

Plankton distribution at sub-mesoscale: contributions from artificial intelligence to plankton ecology

Thelma Panaïotis

Under the supervision of
Jean-Olivier Irisson

PhD thesis in Ecology

Defended on April 6th 2023

The jury comprised

Emmanuel Boss	Reviewer
Francesco Pomati	Reviewer
Lee Karp-Boss	Examinator
Louis Legendre	Examinator
Marina Lévy	Examinator
Jean-Olivier Irisson	PhD Supervisor

Abstract As the basis of oceanic food webs and a key component of the biological carbon pump, planktonic organisms play major roles in the oceans. However, their small-scale distribution – governed by biotic interactions between organisms and interactions with the physico-chemical properties of the water masses in their immediate environment – are poorly described *in situ* due to the lack of suitable observation tools. New instruments performing high resolution imaging *in situ* in combination with machine learning algorithms to process the large amount of collected data now allows us to address these scales.

The first part of this work focuses on the methodological development of two automated pipelines based on artificial intelligence. These pipelines allowed to efficiently detect planktonic organisms within raw images, and classify them into taxonomical or morphological categories. Then, in a second part, numerical ecology tools have been applied to study plankton distribution at different scales, using three different *in situ* imaging datasets. First, we investigated the link between plankton community and environmental conditions at the global scale. Then, we resolved plankton and particle distribution across a mesoscale front, and highlighted contrasted periods during the spring bloom. Finally, leveraging high frequency *in situ* imaging data, we investigated the fine-scale distribution and preferential position of Rhizaria, a group of understudied, fragile protists, some of which are mixotrophic.

Overall, these studies demonstrate the effectiveness of *in situ* imaging combined with artificial intelligence to understand biophysical interactions in plankton and distribution patterns at small-scale.

Remerciements

Ce travail est le fruit d'une belle aventure de presque 4 années. Il n'aurait pas été le même sans la présence et le soutien de nombreuses personnes autour de moi.

J'addresse mes premiers remerciements à mes encadrants : **Sakina-Dorothée Ayata** (pour les deux premières années de cette thèse) et **Jean-Olivier Irisson**. Sakina, merci à toi d'avoir soutenu ce projet et d'avoir accepté de m'encadrer au début de cette thèse, le temps que Jean-Olivier obtienne son Habilitation. Jean-Olivier, je ne peux qu'être reconnaissante devant tout ce que tu m'as apporté au cours de ce travail. Merci de m'avoir fait confiance et de m'avoir guidée, merci pour ta rigueur, ton esprit critique, mais avant tout merci pour ton côté humain ! Malgré ton emploi du temps surchargé et tes multiples casquettes (directeur de thèse, chef d'équipe, *chargé de café et de chocolat*), tu as toujours su te rendre disponible quand cela était nécessaire. Merci aussi de m'avoir donné l'opportunité de participer à des conférences et d'y rencontrer des personnes qui ont apporté leur pierre à l'édifice.

Pour continuer, je tiens à remercier l'ensemble de cette chère **équipe COMPLEx**, tant pour la bonne humeur qui y règne que pour le côté scientifique. J'y ai toujours trouvé des oreilles attentives, des conseils pertinents, et bien souvent des petites gourmandises à grignoter. Plus particulièrement, merci à **Lars Stemmann** d'avoir co-encadré mon travail de Master 2 avec Jean-Olivier, travail dans la continuité duquel s'inscrit cette thèse. J'adresse également mes remerciements à la **PIQv**, qui est derrière de nombreux aspects de mes travaux. Merci à tous les taxonomistes qui ont participé à la construction des différents jeux de données avec lesquels j'ai travaillé. **Laëtitia** et **Lucas**, merci pour votre travail sur le tri des vignettes UVP6. **Louis**, cette thèse n'aurait pas été possible sans le travail méticuleux (et parfois rébarbatif, j'en conviens) que tu as fourni, mille merci ! Je remercie également les petites mains derrière EcoTaxa, que j'ai parfois maltraité (involontairement, cela va de soi) : **les Laurents**, **Julie** et **Béatrice**. Je tiens aussi à remercier **Rodolphe Lemée**, directeur du LOV, pour son accueil dans le laboratoire, ainsi que toutes les personnes qui participent à son bon fonctionnement.

Enfin, je remercie également les personnes qui ont contribué à faire de ma mission d'enseignement une expérience positive, plaisante et formatrice : **Laure Mousseau**, **Véronique Gourbaud-Stevens** et **Didier Jonas**.

Cette thèse repose sur beaucoup (trop ?) de données, mais n'aurait évidemment pas été possible sans, je remercie donc toutes les personnes qui ont fait de la campagne VISUFRONT un succès : **Robert Cowen**, **Robin Failletaz**, **Cedric Guigand**, **Jean-Olivier Irisson**, **Martin Lilley**, **Fabien Lombard** et **Jessica Luo**. Je souhaite également remercier les spécialises respectifs des copépodes – **Denys Altukhov** et **Stéphane Gasparini** – et des rhizaires – **Tristan Biard** – qui m'ont permis d'y voir plus clair dans les données.

This thesis is based on a lot (too much?) of data, but obviously would not have been possible without it, so I thank all the people who made the VISUFRONT campaign a success: ***Robert Cowen***, ***Robin Failletaz***, ***Cedric Guigand***, ***Jean-Olivier Irisson***, ***Martin Lilley***, ***Fabien Lombard*** *and* ***Jessica Luo***. *I also wish to thank the respective specialists of copepods –* ***Denys Altukhov*** *and* ***Stéphane Gasparini*** *– and of rhizarians –* ***Tristan Biard*** *– who enabled me to have a better understanding of the data.*

Et puisqu'il n'y avait pas assez de données, d'autres remerciements vont à des membres de l'équipe OMTAB, notamment **Antoine Poteau**, **Laurent Coppola** et **Émilie Diamond Riquier** pour leur implication dans la campagne de déploiement du glider Seaexplorer équipé de l'UVP6. **Marc Picheral** (*merci d'avoir réparé la machine à café !*), **Camille Catalo** et les membres d'**Alseamar** (souvent dérangés le week-end, parfois la nuit) viennent compléter la liste des personnes qui ont rendu cela possible. Merci d'avoir accordé votre confiance à mes qualités de pilote, tant pour la *Pelagia* que pour *Sea002*. Merci également à toutes les personnes qui ont participé aux sorties bateau de déploiement ou récupération du glider. Suite à cette mission, **Rainer Kiko** et **Martin Schröder** m'ont apporté leur aide pour le tri des images.

Durant ce travail, j'ai eu la chance de collaborer de près ou de loin avec des personnes extérieures au laboratoire et qui ont su m'apporter leur expertise. Encore un très grand merci à **Ben Woodward** pour ton aide indispensable dans la mise en place des algorithmes de traitement des données ! **Jessica Luo**, **Adam Greer**, **Moritz Schmid**, **Robin Failletaz**, **Tristan Biard** et **Robert Cowen**, merci à vous tous pour ces collaborations enrichissantes et vos apports à mes travaux.

During these years, I had the chance to collaborate with people from outside the laboratory who brought me their expertise. A very big thank you again to ***Ben Woodward*** *for your indispensable help in setting up the data processing algorithms!* ***Jessica Luo***, ***Adam Greer***, ***Moritz Schmid***, ***Robin Failletaz***, ***Tristan Biard*** *and* ***Robert Cowen***, *thank you all for these enriching collaborations and your contributions to my work.*

Au-delà des humains qui ont rendu ce travail possible, un certain nombre de machines y ont également contribué : **Niko** et **Marie** au LOV, mais également Jean-Zay géré par l'**IDRIS**. Des ressources ont également été allouées par la plateforem **ABiMS** située à Roscoff.

Beyond the humans who made this work possible, several machines also contributed: ***Niko*** *and* ***Marie*** *at the LOV, but also Jean-Zay managed by* ***IDRIS****. Computing resources have also been allocated by the* ***ABiMS*** *platform located in Roscoff.*

Je souhaite également remercier les membres de mon jury : **Emmanuel Boss**, **Lee Karp-Boss**, **Louis Legendre**, **Marina Lévy** et **Francesco Pomati** pour avoir bien voulu prendre le temps d'évaluer mon travail ; ainsi que les membres de mon comité de thèse : **Laurent Coppola**, **Marina Lévy** (à nouveau) et **Ketil Malde**. Merci pour votre bienveillance et vos judicieux conseils.

I would also like to thank the members of my jury: ***Emmanuel Boss***, ***Lee Karp-Boss***, ***Louis Legendre***, ***Marina Lévy*** *et* ***Francesco Pomati*** *for taking the time to evaluate my work; as well as the members of my thesis committee:* ***Laurent Coppola***, ***Marina Lévy*** *(again) and* ***Ketil Malde****. Thank you for your kindness and your wise advice.*

Il me reste à remercier toutes les personnes avec qui j'ai partagé un bureau au cours de ces années et qui ont contribué à ces petites choses qui m'ont fait sourire. **Laure**, membre fondatrice de la zone dauphin avec **Salomé**, merci pour ces moments de décompression intense, que ce soit avec **George** ou pour la production de gifs. **Salomé**, je pense toujours régulièrement à **Barnabé** et **Vuemer**. **Florian**, merci pour ces sorties vélo dans lesquelles j'ai souffert pour que personne ne nous rattrape (ouf !). **Ophélie**, merci pour tes cours de natation (ou de barbotage, c'est selon). **Laetitia D**, merci pour ta main verte et la transformation de notre rebord de fenêtre en canopée. Merci aussi à tous les doctorants et post-doctorants avec lesquels j'ai pu échanger et passer de bon moments : **Alberto**, **Alexandre**, **Anaïs**, **Aurélie**, **Chloé**, **Dodji**, **Flavien**, **Louis**, **Marine** et **Zoé**.

Je me dois également de saluer un groupe de personnes qui a pris une place un peu particulière dupuis la prepa agreg : **Baya**, **Chloé**, **Enzo**, **Estelle**, **Lucille**, **Maxime** et **Romane** (aka les minous), merci à vous.

Enfin, mes derniers remerciements vont à ma famille. Tout d'abord merci à mes parents de m'avoir soutenue dans mes études, d'abord en prépa puis à l'ENS. Merci de m'avoir donné le goût de la nature, et plus particulièrement celui des océans. Merci à tous les membres de ma famille d'avoir toujours été présents, c'est grâce à vous que je suis devenue la personne que je suis aujourd'hui. Merci aussi à ma belle-famille, pour votre accueil chaleureux. Je suis fière de pouvoir *enfin* vous présenter ce travail qui a du vous paraître si long et obscure. Et finalement, merci à toi Julie, pour tes conseils, ton soutien et ta bienveillance. Merci de partager ma vie et de me rendre si heureuse.

Scientific communications

Published papers

T. Panaïotis, L. Caray–Counil, B. Woodward, M. S. Schmid, D. Daprano, S. T. Tsai, C. M. Sullivan, R. K. Cowen, and J.-O. Irisson. "Content-Aware Segmentation of Objects Spanning a Large Size Range: Application to Plankton Images". In: *Frontiers in Marine Science* 9 (2022). DOI: 10.3389/fmars.2022.870005

Conference talks

T. Panaïotis et al. "Typology of Plankton Communities Seen by In Situ Imaging, from the Epi to the Mesopelagic Layers of the Global Ocean." Ocean Sciences Meeting (San Diego). 2020

T. Panaïotis et al. "Benchmark of Image Classification Using Several Large Plankton Datasets: Convolutional Neural Networks Improve Detection of Rare Taxa." Aquatic Sciences Meeting (Virtual Meeting). 2021

T. Panaïotis, L. Caray-Counil, R. Faillettaz, J. Y. Luo, C. M. Guigand, R. K. Cowen, and J.-O. Irisson. "Meter-Scale Plankton Distribution across a Mesoscale Front." Ocean Sciences Meeting (Virtual Meeting). 2022

T. Panaïotis, L. Caray–Counil, B. Woodward, M. S. Schmid, D. Daprano, S. T. Tsai, C. M. Sullivan, R. K. Cowen, and J.-O. Irisson. "Content-Aware Segmentation of Plankton Images". ICES Annual Science Conference (Dublin, Ireland). 2022

Contents

List of Figures

List of Tables

List of Acronyms

AI	Artificial Intelligence
ADCP	acoustic Doppler current profiler
ANN	Artificial Neural Network
AUV	Autonomous Underwater Vehicle
BRT	Boosted Regression Trees
CNN	Convolutional Neural Network
CTD	Conductivity, Temperature, Depth
DCM	Deep Chlorophyll Maximum
DVM	Diel Vertical Migration
DL	Deep Learning
ENSO	El Niño-Southern Oscillation
ESD	Equivalent Spherical Diameter
GPU	Graphic Processing Unit
HAC	Hierarchical Agglomerative Clustering
IFCB	Imaging FlowCytobot
ISIIS	*In Situ* Ichthyoplankton Imaging System
OMZ	Oxygen Minimun Zone
ML	Machine Learning
MLP	Multi-Layer Perceptron
MSER	Maximally Stable Extremal Region
NASS	Normalized Abundance Size Spectra
PCA	Principal Component Analysis
RDA	Redundancy Analysis
RF	Random Forest
SNR	Signal to Noise Ratio
SPC	Scripps Plankton Camera
SVM	Support Vector Machine

UVP Underwater Vision Profiler

VPR Video Plankton Recorder

Part I

General introduction

Plankton (from ancient greek πλαγκτός, "drifter") – organisms that live in a large body of water and are unable to swim against currents.

1

Context and state of the art

This chapter introduces the main notions underpinning this work. Starting with space and time scales in oceanic processes, I then address plankton ecology and its study methods and describe a few computational methods to process and analyse large ecological datasets. Finally, I present the main ecological questions tackled in this work, as well as the datasets used to attempt to address them.

1.1 Scales in oceanic processes

The oceans are in motion at a variety of scales:

The 1.3 billion cubic kilometres of water of the world ocean are in perpetual motion across a large range of spatio-temporal scales (Figure 1.1), from kilometres to centimetres and from centuries to seconds, length and time scales being roughly correlated.

1.1.1 Physical processes, from larger to smaller scales

global circulation,

Currents flow throughout all the oceans, they cover great distances and their combination creates a permanent global-scale circulation sometimes referred to as the "global conveyor belt". At this scale, the oceans are put in motion by two processes: a thermohaline circulation and a wind-driven circulation [328]. Wind-driven currents are typically restricted to the upper layers of the oceans (a few hundred meters), while interior mixing is driven by gradients of density mediated by changes in temperature and salinity. Tides also play an important role as a source of turbulent mixing in conjunction with winds. These currents transport huge amounts of heat, and thus have a major effect on climate, so that changes in currents are today the best hypothesis to explain the abrupt changes in climate observed during the Earth's history [329], and in a positive feedback loop, the currents could in turn be impacted by global changes. Likewise, global currents are responsible for large temperature differences between ocean basins at a given latitude [329].

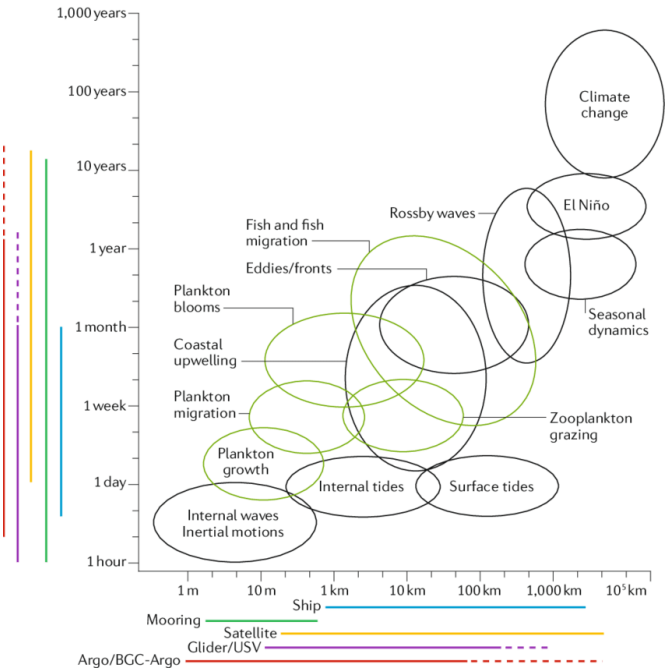


Figure 1.1: Physical (black) and biological (green) oceanic processes cover large spatio-temporal scales. From Chai et al. [72].

Nowadays, large-scale climate variations occur periodically such as El Niño–Southern Oscillation (ENSO). This phenomenon caused by a variation in winds and sea surface temperatures in the eastern tropical Pacific affects climatic conditions across tropical and subtropical zones.

Mesoscale – $\mathcal{O}$(10 - 100 km) – motions mainly consist of eddies and fronts. Eddies are rotating columns of water, spanning from 10 to 500 km in diameter, over periods of days to months [334]. Mesoscale eddies can be static when caused by the presence of a fixed obstacle, or mobile when generated by a baroclinic instability. For example, growing meanders associated with large currents, such as the Gulf Stream or the Antarctic Circumpolar Current, end up separating from the current, generating eddies [256, 267]. As the water mass inside the eddy is trapped at eddy formation, its properties differ more and more from the surrounding water as the eddy moves away from its origin location, which can result in a biological hotspot [92]. Due to the Coriolis effect, water is pushed away from the centre in cyclonic eddies, resulting in an upwelling of nutrient-rich, cold water. The opposite applies to anticyclonic eddies, resulting in a downwelling [267]. Fronts are zones where two bodies of water with different properties (e.g. temperature, salinity, turbidity, oxygen) meet [26]. They occur at a wide variety of spatial and temporal scales but some of them are permanent [26]. Frontal zones are often associated with a surface convergence flow, resulting in an increased diversity and biomass across all trophic levels. Other zones of important productivity are coastal upwellings. In such zones, cold and nutrient-rich waters are brought from the depth to the ocean surface – on a permanent or seasonal basis – in response to wind-driven offshore displacement of surface water [197]. In addition, coastal upwellings are often associated with a geostrophic current flowing parallel to the coast, possibly paired with an upwelling front.

mesoscale eddies,

fronts,

coastal upwellings,

Mesoscale features described above carry a lot a kinetic energy and significantly impact the global oceanic circulation [209]. However, the satellite observations do not quite fit the classical geostrophic turbulence theories, and smaller-scale structures have been identified as shadow players [209], potentially involved in the global heat budget [389]. Overlooked for a long time due to their short-lived and spatially restricted nature, submesoscale dynamics – $\mathcal{O}$(1 - 10 km) – now become better resolved thanks to new observation tools (e.g. floats,

submesoscale dynamics

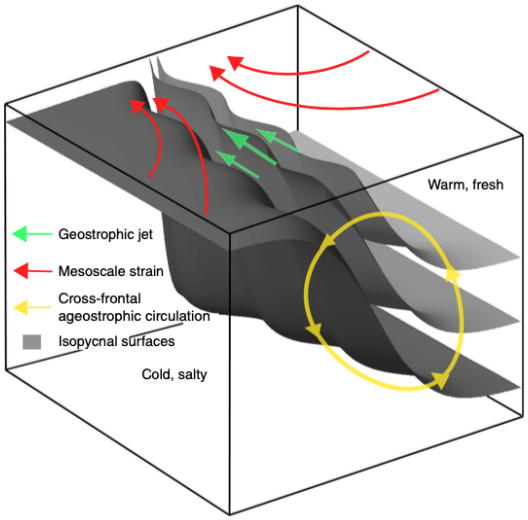


Figure 1.2: Submesoscale frontal dynamics. The front is represented by the oblique isopycnals, delimiting cold and salty waters offshore from warm and fresh waters inshore. A cross-frontal ageostrophic circulation (yellow arrows) takes place in the direction of flattening the isopycnals. From Lévy, Franks, and Smith [230].

gliders, high-resolution satellite imagery) [229]. In addition, numerical models have confirmed their importance for the intensity of the vertical velocity field [228]. Submesoscale dynamics are commonly associated with mesoscale fronts (Figure 1.2), where a cross-frontal ageostrophic circulation can drive vertical displacements of waters.

and small-scale motions.

At slightly smaller scales, internal waves [138] vertically disturb isopycnals, inducing vertical displacement of waters and organisms within [139]. Finally, multifaceted and complex microscale motions occur at even smaller scales – $\mathcal{O}$(1 mm). Their contribution to larger processes is not fully understood, even though progress has been made in understanding these turbulent dynamics [281]. As the processes described above cover 9 to 10 orders of magnitude, multiple observation tools are required to monitor them (Figure 1.1). In addition, all these processes affect the life of all ocean inhabitants, from the largest to the smallest.

1.1.2 Effects on marine life

The effects are particularly visible on human-exploited resources, such as fisheries periodically affected by ENSO [2]. Similarly, fronts have long been targeted by fishermen [297] because they aggregate harvestable fish. Top predators including tuna, elephant seals and birds [352] seem to also exploit mesoscale and submesoscale oceanic features. Such biological hotspots could result from an enrichment at all trophic levels [304]. Indeed, submesoscale dynamics associated with mesoscale fronts generate vertical motions affecting the phytoplankton growth rate by redistributing phytoplankton cells and nutrients [230, 252]. Increase in phytoplankton at fronts can propagate to the zooplankton community [275, 294] and even forage fish [32]. However, we still lack information regarding submesoscale distribution of intermediate trophic levels organisms (forage fish, zooplankton) in order to better understand how the whole ecosystem is affected by mesoscale and submesoscale processes [230]. Because they can modulate phytoplankton growth rates and occur at comparable time scales, submesoscale dynamics are particularly relevant to phytoplankton productivity and planktonic ecosystems, even though they are restricted in time and space [230, 252]. The importance of these effects for planktonic organisms is further discussed below at paragraph 1.3.1.

Life is shaped by physical processes at all scales,

but submesoscale processes might be particularly critical.

Thus, if all inhabitants of the oceans seem to be affected by these processes, this is all the more true for those who drift and cannot swim efficiently against currents and thus cannot choose their habitat.

1.2 Plankton: drifters of the oceans

1.2.1 Plankton diversity

The niche-based definition of plankton encompasses a very large diversity of organisms, not only in terms of taxonomy (Figure 1.3) but also in terms of size [69, 186] (Figure 1.4). Indeed, plankton ranges from infra-micrometer virioplankton to meter-long Cnidarians. As a consequence, various trophic modes exist in the planktonic world. A typical distinction occurs between autotrophic phytoplankton and heterotrophic zooplankton, but some planktonic organisms are also mixotrophic.

Plankton is very diverse both in size and taxonomy.

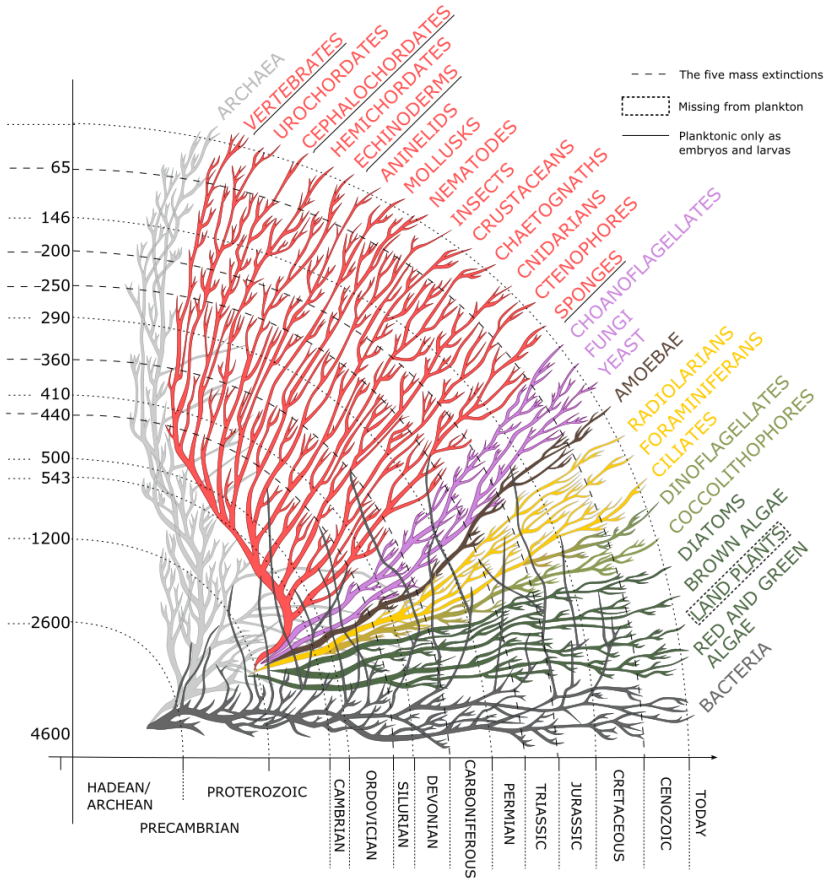


Figure 1.3: Tree of life highlighting the taxonomic diversity of plankton. Planktonic organisms can be found in all major taxonomic groups except land plants. The animal clade (red) is represented with more details than others. Adapted from Sardet [349].

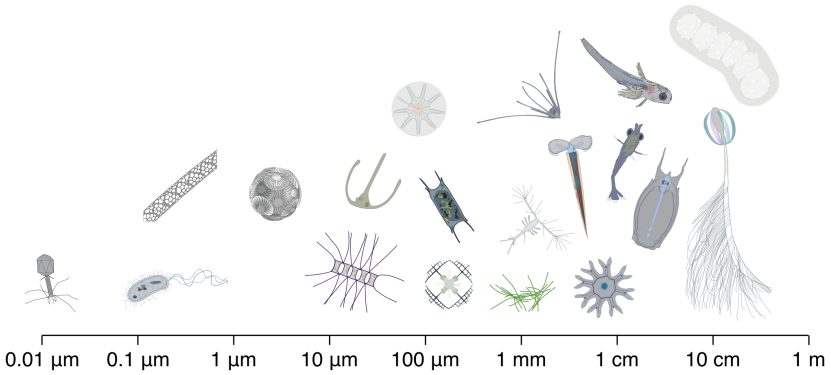


Figure 1.4: Size diversity of plankton. Plankton covers a size range from 0.01 µm to 1 m. Drawings by J. Descamps.

1.2.2 Ecological importance of plankton

Planktonic organisms play key roles in the oceans...

Photosynthetic phytoplanktonic organisms are primary producers at the basis of marine food webs [125], and zooplankton is a major trophic link between these and higher trophic levels such as fish, marine mammals and birds [413] (Figure 1.5). Phytoplankton is responsible for about half of the primary production on Earth and captures CO_2 from the atmosphere [131]. The produced organic carbon is then consumed by zooplankton and partly exported at depth through the biological carbon pump, making plankton a key link in the biogeochemical carbon cycle [246].

... and are good indicators of ecosystems' health.

Planktonic organisms are very sensitive to the environmental conditions they experience in the water masses they are embedded in, and are thus good indicators of environmental changes [172]. For these reasons, plankton biomass and diversity were endorsed as essential oceanic variables [273, 284], essential biodiversity variables [312] and essential climate variables [44].

1.2.3 Global patterns of plankton distribution

Large-scale patterns of plankton distribution are well known...

Because of their sensibility to their environment, their distribution and diversity are largely driven by environmental conditions, including temperature, oxygen, nutrients and light [172]. As these conditions vary greatly with latitude, the distribution of planktonic organisms

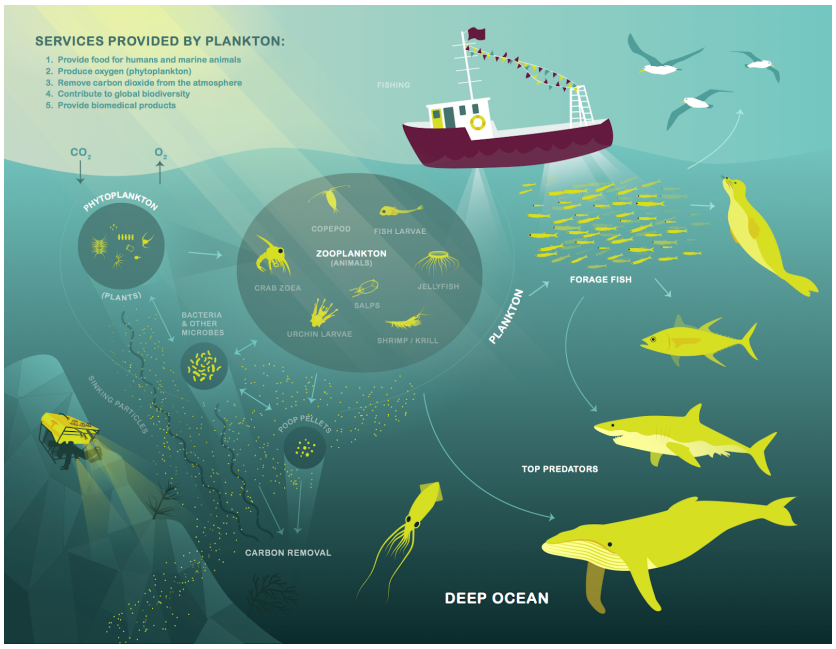
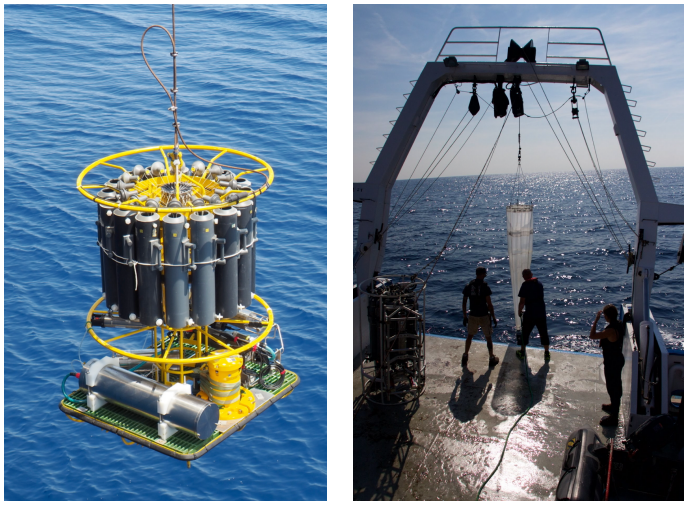


Figure 1.5: Planktonic organisms play key roles in oceanic ecosystems. Source: National Data Science Bowl.

is also related to latitude: diversity is larger in warm and nutrient-poor environments located at low latitudes [188, 340, 345, 399], while biomass is higher in nutrient-rich high latitudes environments [190].

Thus, while large-scale distribution patterns of plankton are resolved to a certain extent, much remains to be known regarding fine-scale distribution, especially for zooplankton. Knowledge gaps regarding the fine-scale distribution of plankton partly stems from the difficulty to adequately sample it at such a small scale. Indeed, traditional plankton collection methods such as pumps, nets, and bottles (Figure 1.6) lack spatio-temporal resolution as they typically integrate organisms over some vertical and/or horizontal distance and make it difficult to associate organism concentrations with their immediate environmental context [30, 243, 330]. Finally, resolving fine-scale distribution associated to small and short-lived spatial features requires fast and repeated sampling of these features. Such an approach would be costly, generate a lot of data and thus pose challenges for data analysis.

. . . but gaps persist in fine-scale distribution knowledge.



(A) **(B)**

Figure 1.6: Traditional plankton sampling tools. **(A)** Niskin bottles fixed on a rosette, **(B)** plankton net deployed from a ship.

1.3 Distribution of plankton at fine-scale

1.3.1 Why should we study plankton distribution at submesoscale?

Relevant processes to explain plankton distribution take place at fine-scale.

Submesoscale dynamics are known to influence the distribution of phytoplanktonic organisms: vertical currents may affect nutrients and cell distribution relative to the euphotic zone where photosynthesis occurs, thus affecting growth rates, while horizontal currents can stir patches into filaments. These changes are expected to propagate to higher trophic levels (zooplankton, fish, etc.) [230]. Indeed, the trophic and reproductive interactions of zooplankton occur at the scale of organisms (µm to cm). Therefore, a local concentration of phytoplankton, in a thin layer for example, has more immediate consequences on the survival and development of zooplanktonic grazers than the average chlorophyll a concentration in the region. Thus, studying zooplankton distribution at fine-scales, in relation with submesoscale dynamics, becomes relevant to understanding the processes driving its distribution at regional scale.

1.3.2 How to study plankton distribution at submesoscale?

As explained above, traditional plankton tools are not adapted to resolve fine-scale distribution. In addition, most of them damage fragile organisms and then fail to adequately estimate their abundance [330].

In situ imaging enables fine-scale studies...

The development of *in situ* imaging tools partially overcame some of these limitations: not only they can resolve the exact *in situ* position of organisms and can sample the environmental conditions in the immediate vicinity of organisms, but they also allow investigating fragile planktonic objects such as Rhizaria [42, 104], gelatinous plankton (e.g. Cnidaria, Ctenophora) [250] or even marine snow particles [161, 162, 401].

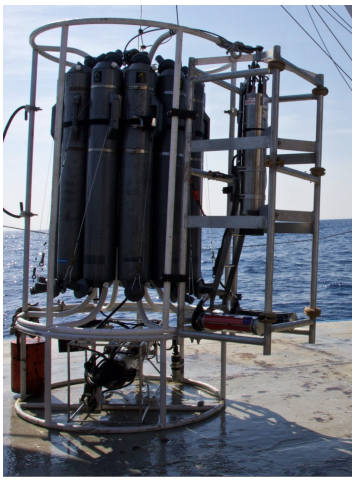
...thanks to a wide range of instruments.

Diverse *in situ* imagers were developed over time, with varying specifications (Figure 1.7) (see [243] for a detailed list). Some of them – the Imaging FlowCytoBot (IFCB) [298] and the Underwater Vision Profiler 6 (UVP6) [318] – can be deployed on fixed, long-term moorings. Others such as the In Situ Ichthyoplankton Imaging System (ISIIS) [85] and the Video Plankton Recorder (VPR) [95] perform a sawtooth-like profile while being towed by a ship (i.e. a "tow-yo" pattern). The UVP5 [317], the UVP6 [318] and the Lightframe On-sight Keyspecies Investigation

(LOKI) [359] can also be deployed from a ship, along vertical profiles. Finally, recent advances have been made towards the integration of *in situ* cameras on remotely operated or autonomous vehicles (gliders, floats...). This concerns the Zooglider [293] or the UVP6. Taken together, they cover a substantial part of the planktonic size range (Figure 1.8) and allow us to consider many sampling strategies.

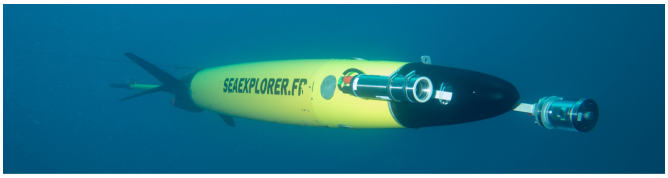
The high spatio-temporal resolution data generated by these instruments enables tackling questions that used to be out of reach. Indeed, *in situ* imaging can resolve fine-scale plankton distribution in relation to environmental conditions: McClatchie et al. [265] detected increased primary and secondary production in a frontal zone in the Southern California Bight, Greer et al. [154] and Briseño-Avena et al. [59] observed contrasted distribution patterns of planktonic organisms on both sides of frontal zones at George Bank (NE Atlantic) and in the Central Oregon Coast respectively. Luo et al. [248] investigated the distribution of gelatinous plankton across a front in the Southern California Bight and reported low effect of the front. Similarly, *in situ* imaging data collected across the Ligurian Front (NW Mediterranean) revealed plankton distributions constrained by the front [124]. Furthermore, Christiansen et al. [77] reported the abundance of a polychaete annelid in a mesoscale eddy in the tropical Atlantic Ocean, in association with very low particle concentrations. *In situ* imaging also revealed a variation of patch properties (frequency, density, size) across planktonic groups and water masses [337]. In addition, most detected patches were small (10 - 30 m), highlighting the necessity of high resolution imaging to detect such features. Several studies tackled interactions between zooplankton and phytoplankton thin layers: Greer et al. [156] demonstrated that copepods and appendicularians had contrasted distributions in relation with a phytoplankton thin layer, while ctenophores aggregated inside the layer. Similarly, doliolids and copepods were found to be distributed differently with respect to a phytoplankton thin layer [153]. In the Arctic, Schmid and Fortier [355] detected dissimilar distributions of two copepod species around a subsurface chlorophyll maximum. Moreover, larval fish interactions with their preys or predators were shown to be affected by physical features such as an eddy front [354], tidal plumes [395] or internal waves [155]. Finally, interactions (competition, parasitism, predation, commensalism) were reported directly in *in situ* images [152].

They provide insight on ecological questions that could not be resolved from nets or bottles...

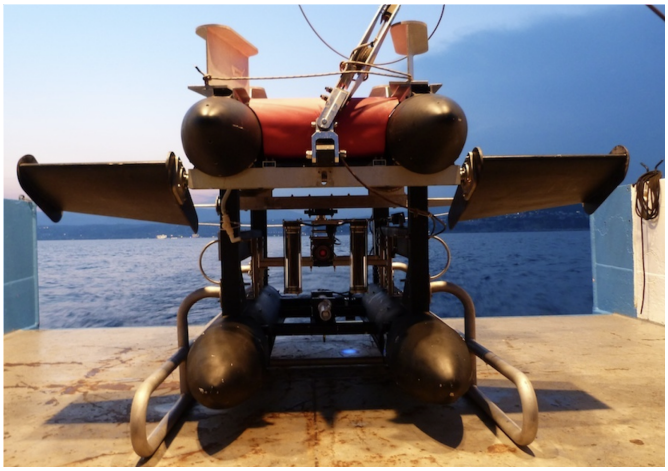


(A)

(B)



(C)



(D)

Figure 1.7: Examples of *in situ* plankton imaging instruments. **(A)** Underwater Vision Profiler (UVP5) mounted on a rosette, **(B)** UVP6 deployed on its own, **(C)** UVP6 mounted on a glider, **(D)** In Situ Ichthyoplankton Imaging System (ISIIS).

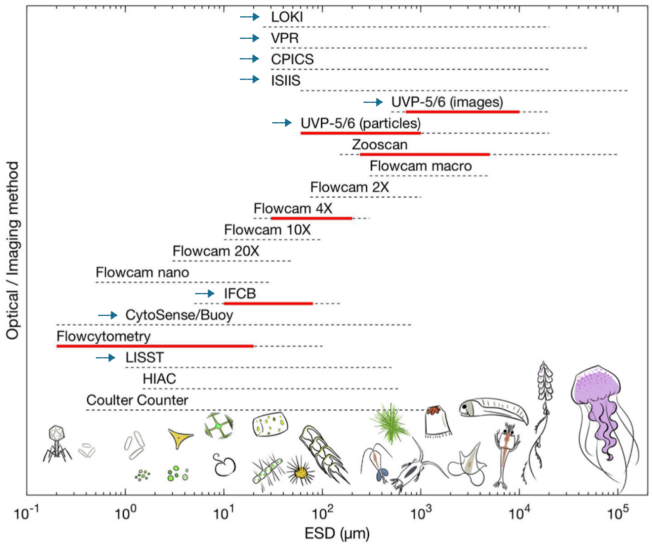


Figure 1.8: Size range in Equivalent Spherical Diameter (ESD) covered by common plankton imaging instruments. Dashed lines represent the total operational size range from commercial information while the red line represents the practical size range which is efficient to obtain quantitative information. Blue arrows indicate *in situ* imagers. From Lombard et al. [243].

. . . and even shed light on plankton behaviour.

In addition, some of these cameras (e.g. ISIIS, VPR, UVP, zooglider) can capture images of planktonic organisms without disturbance, thus potentially revealing *in situ* position, behaviour or interactions with other organisms [299]. Indeed, multiple studies reported the feeding behaviour of copepods [276, 291, 293]. Similarly, preferential orientation [140] and apparent feeding behaviour [258] were also detected in Rhizaria (unicellular Eukaryotes). Beyond behaviour, individual traits (e.g. size, opacity, apparent activity) can also be captured from *in situ* imaging: Vilgrain et al. [409] reported contrasted size and activity patterns around an Arctic ice melt zone; Sonnet et al. [370] recorded both seasonal and inter-annual variations in phytoplankton morphology from a 2 years time series of IFCB data collected in Narragansett Bay (NW Atlantic). Yet, trait investigations are not limited to planktonic organisms and revealed morphology changes in marine snow particles during the phytoplankton bloom [401].

Face the data flood.

Although *in situ* imagers typically sample smaller volumes than nets [243] (with the exception of the ISIIS, generally > 100 L s^{-1}), their increasing number and ease of use generate an increasing volume of data, resulting in a data processing bottleneck [253]. For example, a UVP images about 1.5 million objects per year (~8.6 billion pixels per year), while an hour of ISIIS deployment generates 100 billion pixels (~11 million objects) [192]. To efficiently process such a large amount of data, ecologists must now turn to computational methods.

1.4 Numerical plankton ecology

Numerical ecology: the application of computational methods to answer ecological questions.

Numerical ecology comprises the development and application of statistical methods and tools to describe and interpret datasets in order to answer ecological questions [226]. With the perpetual development of new means of observation (e.g. satellites, imagery, genomics) leading to an even increasing amount of observational data, these approaches are becoming more and more relevant [315]. Indeed, gaining knowledge from observational data is very different from analysing experimental data: observational data requires a thorough exploration and analyses before conclusions can be drawn. One main difference compared to experimental science is that neither the analyses to be conducted nor the nature of the results are necessarily known in advance. Actually, in this era of Big Data, data-driven science could even constitute a new science paradigm based on large datasets and statistical exploration [179, 208].

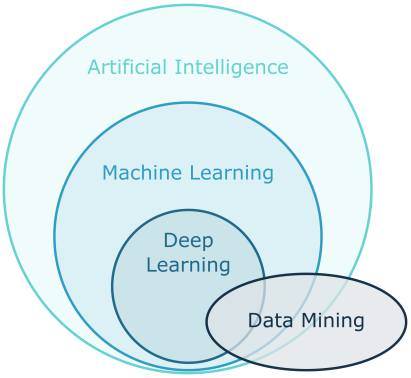


Figure 1.9: Overview of artificial intelligence and related fields: machine learning, deep learning and data mining.

As computing power becomes more and more accessible, new methods are developed and computational ecology becomes even more cost- and time-effective [323]. Despite this progress, data processing and analysis rates remain superseded by the rate of data acquisition [253].

Numerical ecology encompasses a wide variety of methods (e.g. modeling, machine learning, deep learning), some of which are described below and their relations presented in Figure 1.9.

1.4.1 Methods

1.4.1.1 *Data mining*

Data mining: finding knowledge in a flood of data.

Data mining consists of extracting fundamental insights and knowledge from large amounts of data [54, 424]. It relies on various fields, including database systems, statistics, machine learning (addressed hereafter) and pattern recognition. Data mining is just a step in the larger data analysis process: it is typically preceded by a pre-processing phase including data extraction and cleaning, and eventually data fusion, reduction and features construction. Data mining is then followed by post-processing steps such as visualisation, pattern and model interpretation, and finally hypothesis confirmation or refutation [424]. This process is highly iterative and interactive, taking the form of a succession of trial and error. In the field of plankton research, data mining approaches such as ordination methods were used to investigate morphological traits of copepods [409] and marine snow particles [401].

1.4.1.2 *Artificial intelligence*

AI: intelligence demonstrated by machines.

Artificial Intelligence (AI) is intelligence – such as perceiving, synthesizing, and inferring information – demonstrated by machines. The first AI challenge was to establish an appropriate definition to evaluate the intelligence of a machine. In 1950, Turing introduced a test in order to evaluate whether a machine was able to exhibit intelligent behaviour [403], establishing a base concept of AI. The term "artificial intelligence" as a scientific field was coined a few years later at the Dartmouth workshop [264]. After a period of optimism supported by limited success – only trivial problems could be solved – AI was heavily criticized, funds were cut and the entire field was put aside [344]. Indeed, expectations had been set too high and the difficulties had been underestimated. At the time, the main obstacles were the limited computer power, as well as the lack of large databases. AI was reborn only in the 90s, thanks to technological unlocking, opening the way to new approaches. Nowadays, AI is omnipresent in our everyday life, from our phones to hoovers.

Early stages of AI took place in the 50s, but the theory could not be applied because of computational constraints.

1.4.1.3 *Machine learning*

ML finds patterns by itself in deluges of data.

Within AI, Machine Learning (ML) algorithms identify patterns in training data and eventually predict outcomes for new data. Here, the algorithm has the ability to learn without being explicitly programmed for, by finding generalizable patterns. Many types of models can be used in ML: from the simple linear regression to more elaborated models such as support vector machines (SVM) [81] or random forests (RF) [169] to name only the most famous.

Plankton ecology already benefits from ML,...

Many applications of ML to plankton ecology have been identified [192]. ML regression models such as Boosted Regression Trees (BRT) or RF are widely used in species distribution models, which consist of relating the distribution of a taxonomic group to environmental or geographical data [116]. Such models can be used to estimate the continuous distribution of organisms from discrete sampling: Pinkerton et al. [320] and Pinkerton et al. [321] respectively estimated the distribution of Oithona and of six planktonic groups in the Southern Ocean. These models also allow to understand the drivers of such distribution, by identifying the most contributing variables in the model. Indeed, current speed and direction were identified as main drivers of Oithona and larval fish distribution in the Straits of Florida [354]; while

temperature and light explained most of picophytoplankton abundance in the South China Sea [74]. In the Southern California Bight, Luo et al. [248] linked gelatinous zooplankton abundances to environmental conditions and found that drivers were taxon-specific. Yet, predictions are not limited to abundances: Leathwick et al. [221] linked species richness to the environment in the SW Pacific and identified depth as the main driving variable. Similarly, Drago et al. [111] predicted the global plankton biomass from ~2,500 UVP5 profiles.

In addition, ML can significantly improve the speed of image data processing, e.g. for the identification of planktonic organisms from images [192] (Figure 1.10). Indeed, ML models are widely used to automatically classify images of planktonic organisms, including Support Vector Machines (SVM) [185, 250, 372] or RF [149]. However, these classic ML models cannot process raw images, either for learning or prediction. Instead, features describing image properties (size, gray levels...) have to be manually extracted before being fed to the model in the form of a vector of numbers. Of course, the quality and number of these features will strongly impact model performance, highlighting the importance of data pre-processing. Nowadays, ML algorithms are widely used because they are easy to use, flexible and not too demanding in terms of computational power, which is a very desirable property for models running in energy constrained environments such as the embedded classification model integrated into UVP6 sensors [78]. However, classic ML classification algorithms are now outperformed by more recent and more complex models.

...and ML is very useful to accelerate data processing.

1.4.1.4 *Deep Learning*

Deep Learning (DL) is a subset of ML based on artificial neural networks (ANN) with multiple layers: multilayer perceptrons (MLP) [309]. The term "deep" refers to the hidden layers. While the theory was developed in the 50s [341], the first application dates back from 1971 [193].

DL is ML using neural networks with multiple layers.

The architecture of ANN was inspired by the functioning of the animal brain, in which neurons are the base units, connected together through synaptic connections. Hence, the base unit of a MLP is the neuron. Similarly to biological neurons, each neuron receives inputs and transmits outputs depending on the inputs received (Figure 1.11) as well as the activation function. Neurons are then combined together in layers, every neuron of one layer is connected to all the neurons of the adjacent layers by weights, determining the strength of the

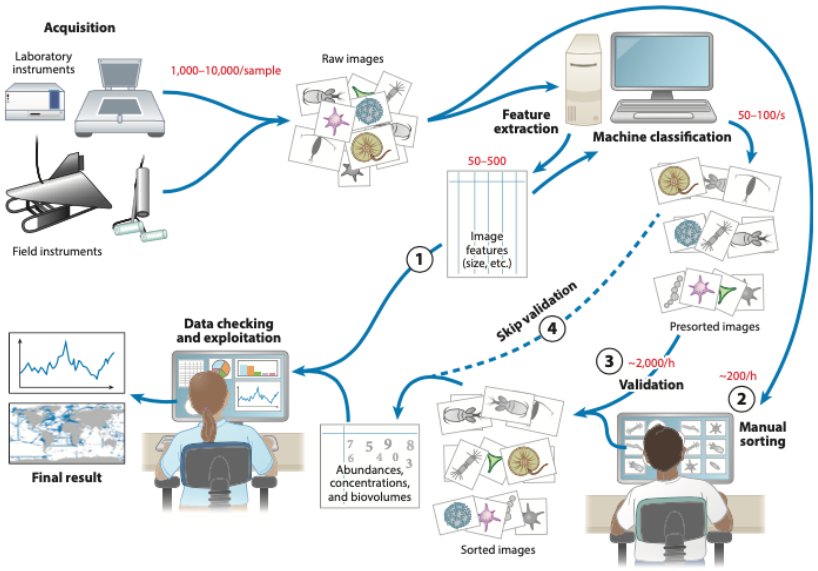


Figure 1.10: Machine learning applications for plankton images classification. (1) use of the features without classification (2) manual classification (no ML), (3) manual validation after machine classification, (4) machine classification without human validation. From Irisson et al. [192].

connection. The first layer corresponds the input layer, i.e. data to be fed to the model. Then come the hidden layers, whose number and size can be changed as desired: the more layers and neurons, the more complex the model. This likely results in better performance but also requires more data to ensure an efficient training. Finally, the last layer corresponds to the output of the model, which can be a single value in case of a regression, or a vector in the case of classification (the length of the vector being equal to the number of classes among which to distinguish). Although longer to train than classical ML models, DL models are particularly appropriate to process large amounts of unstructured data (e.g. images, text, audio...).

Many neurons connected together form a network.

MLP are very versatile and were applied to various tasks in plankton ecology: Sauzède et al. [351] were able to retrieve the vertical distribution of phytoplankton class sizes from *in situ* vertical profiles of chlorophyll fluorescence. MLP were also applied to plankton image classification tasks after extracting hand-crafted features [89, 119, 367, 416]. However, contrary to classic ML models, MLP can use the raw images as inputs. Below is a brief example of application.

DL has various applications in plankton ecology.

Let us place ourselves in the case of the application of a MLP to tackle the recognition of hand-written digits (0-9) on 28×28 pixels images of the MNIST dataset [102], widely used to benchmark image classifiers. Because the MLP takes vectors as inputs and not arrays, the 28×28 image must first be flattened into a 784 element long vector. By adding two hidden layers of size 600, and an output layer on size 10 (number of digit classes), the number of parameters reaches 837,610 ($784 \times 600 + 600 \times 600 + 600 \times 10$ weights and $600 + 600 + 10$ biases). Such a model is easily trained on a relatively recent personal computer. However, a 28×28 image is ridiculously small compared to today's images. If using the same model as previously on a 400×400 image (still small compared to typical images), the input layer would be 160,000 long, for ~96 million parameters. Here we reach the limits of what is possible with a MLP, but more recent architectures now allow to deal with this kind of input.

DL models can process raw images.

Convolutional Neural Networks Convolutional Neural Networks (CNN) are a specific type of artificial NN, mostly used in pattern recognition tasks (e.g. image classification). The architecture of a CNN is inspired from the animal visual cortex: each neuron responds to stimuli from a restricted region of the image and the number of parameters

CNN: a problem shared is a problem halved.

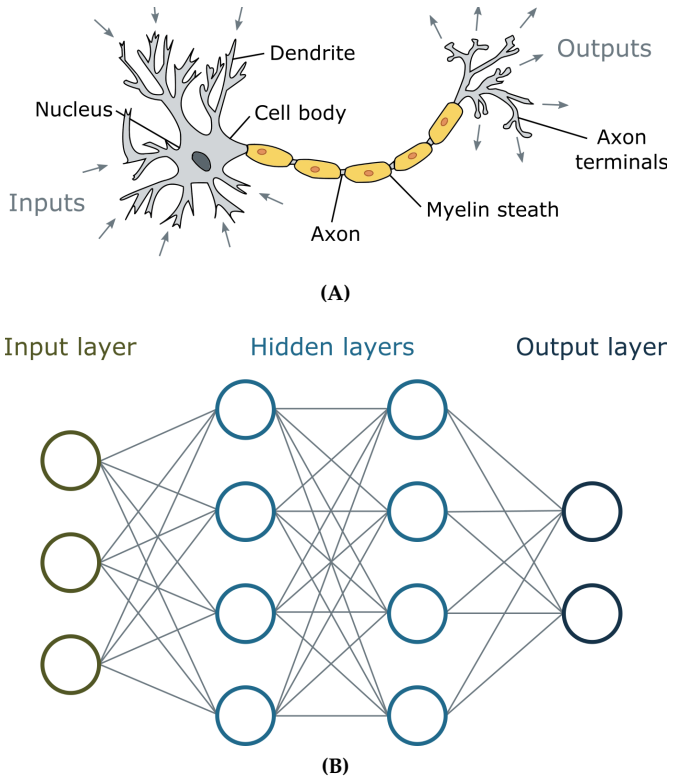


Figure 1.11: Architecture of a deep neural network. **(A)** Schematic representation of a biological neuron. **(B)** A multilayer perceptron with two hidden layers: the circles represent neurons and gray lines are the connections between the output of one neuron to the input of another (i.e. weights). This model takes vectors of three values as input and outputs a vector of size two. 36 weights (3×4 + 4×4 + 4×2) are represented by the gray lines, but each neuron receiving inputs has it owns bias, adding 10 parameters to the network, for a total of 46 parameters.

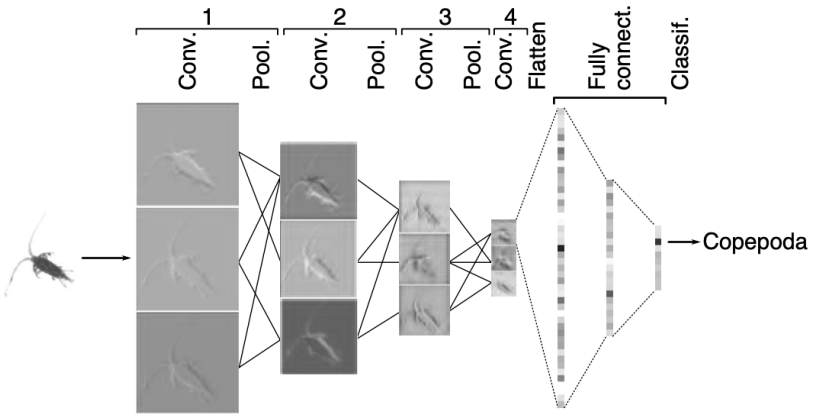


Figure 1.12: Schematic representation of a CNN performing plankton image classification. Conv = convolutional layer, Pool = pooling layer, Fully connect. = fully connected layer, Classif. = classification layer. For visualisation purposes, only a few connections are represented. Credits: JO Irisson.

is drastically reduced by taking advantage of spatial autocorrelations in images and applying given filters to all receptive fields. The typical architecture of a CNN consists of two main parts (Figure 1.12). The first part – also referred as the feature extractor – consists of a set of convolutional layers and pooling layers. As its name suggests, this part extracts relevant features from the image and stores them in a vector. The second part consists of fully connected layers, i.e. a MLP. Finally, the output layer corresponds to the classification or regression output. A great advantage of CNNs is that there is no need to manually extract features anymore: the model itself finds the most appropriate features to distinguish between the classes.

The combination of a feature extractor and a classifier.

While the first application of a CNN to image recognition dates back to 1989 [220] (on handwritten digits), they became very popular after the first and significant success of a CNN at the 2012 ImageNet Large Scale Visual Recognition Challenge [213, 343]. CNNs have now become the state of the art method for image classification [222], but they could soon be superseded by transformers [408].

The current state of the art for image analysis tasks...

Hence, CNNs can improve data processing in plankton ecology, and were used in numerous studies to sort plankton images [59, 77, 93, 118, 124, 224, 247, 249, 327, 337, 354, 395], but CNN applications go beyond simple classification. Indeed, a wide variety of tasks exist in computer vision and can be applied to plankton imaging (Figure 1.13). Among

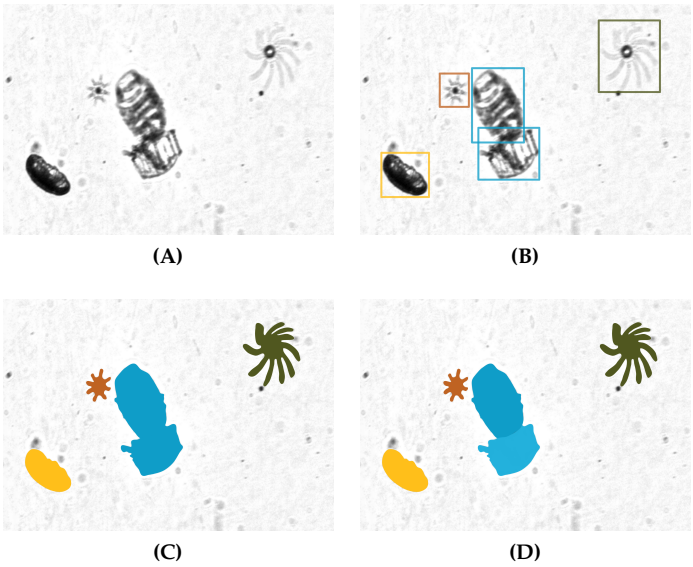


Figure 1.13: Different detection and segmentation tasks achievable by CNNs on a plankton image from the ISIIS. **(A)** raw input, **(B)** object detection, **(C)** semantic segmentation, **(D)** instance segmentation. In **(C)** and **(D)**, the background was intentionally left blank although it constitutes a class on its own. Four plankton classes are represented: Scyphozoa ephyra (Cnidaria) in yellow, Acantharia (Rhizaria) in brown, Doliolida in blue and Rhopalonematidae (Cnidaria) in green.

...and many possible applications for plankton imaging.

the most common, object detection consists in detecting instances of objects of a certain class (Figure 1.13B), while image segmentation consists in delimiting regions by assigning each pixel of an image to a class (e.g. background vs. object). In semantic segmentation, every pixel of the same class belongs to the same region (Figure 1.13C), whereas instance segmentation detects distinct instances of the same class (Figure 1.13D). These approaches are rather novative in plankton images processes, and only a few studies targeted the topic [75, 199, 232, 299] Nonetheless, object detection and/or segmentation methods are already widely used in fish ecology [8, 109, 277].

Thus, these computational approaches now enable to process large, complex datasets, but this progress would not have been possible without the evolution of computational tools.

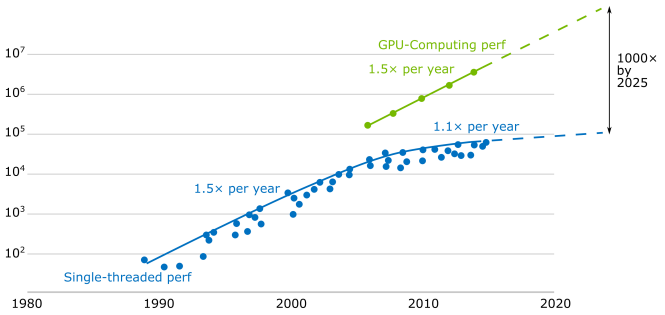


Figure 1.14: Progress in computational power over the last 30 years. Note that the y axis is log-transformed. GPU: graphical processing unit. Source: Nvidia.

1.4.2 Tools

1.4.2.1 *Computational power*

Computational power: to infinity and beyond?

According to Moore's law [278], computational power doubles approximately every two years while the cost of computers is halved in the same time. Since the 1970s, the power of computers has doubled every year or every year and a half: today's computers are millions of times more powerful than their ancestors of 50 years ago. While the trend continues (Figure 1.14), the development of Graphic Processing Units (GPU) – and more recently of Tensor Processing Units (TPU) – was particularly important for the success of CNNs [73]. Indeed, CNNs rely on a lot of simple unitary computations that can easily be run in parallel, a task for which GPUs are particularly adapted.

1.4.2.2 *Large, public available datasets*

No data, no knowledge...

...including for plankton.

The success of DL also relies on the availability of large datasets of several million annotated images such as ImageNet [101], Coco [238] or Pascal VOC [122]. These large datasets enable training complex models, but are also used to benchmark model performance [343]. The availability of such datasets is the *sine qua none* condition for the application of deep learning to a given field. In plankton imaging, only three datasets were publicly released so far (Table 1.1) and used in several studies. Thus, although we still lack systematic comparison for ML application to plankton ecology [192], the move towards standardization and intercomparison is ongoing.

Table 1.1: Plankton images benchmark datasets.

Name	References	Imaging instrument	Composition	
			Images	Classes
WHOI-plankton	[301, 373]	IFCB	3.5 M	103
ZooScanNet	[115]	ZooScan	1.4 M	93
PlanktonSet 1.0	[86]	ISIIS	30,336	121

1.4.2.3 *Open source libraries*

DL made as easy as assembling bricks.

Beyond computational power and data, the third pillar of DL is the availability of open source turnkey libraries such as Tensorflow [1], Pytorch [308] or ScikitLearn [310]. These libraries allow to design and to train conventional or custom ML/DL models, and also enable data preprocessing, visualisation and model evaluation.

1.4.2.4 *A fast evolving field*

Tools are evolving very quickly.

DL is an actively developing field, with new and more powerful tools constantly released. To provide some context, one of the tools at the core of this work was not released yet when the PhD started in September 2019. When it was published in February 2020, our then 3-year-old computation server (12 (24 logical) CPU cores, 126 GB of RAM, Tesla K20 GPU) was outdated to run it. Computations were partly performed on a new computation server (36 (72 logical) CPU cores, 192 GB of RAM, Ampere Quadro RTX8000 GPU), as well as on the Jean-Zay supercomputer (Figure 1.15) and a computation server belonging to the Roscoff Bioinformatics platform ABiMS.

1.5 Aim of this work

1.5.1 Ecological questions

Investigating the drivers of plankton distribution from global to submesoscale...

By focusing on plankton distribution across a large range of scales, from the global to the submesoscale, this work aims to advance our knowledge mostly of the drivers of plankton distribution. (i) First, I investigate the large types of plankton communities in the open ocean, as well as their difference between the epipelagic and upper-mesopelagic layers, and their relation with environmental factors. (ii)

Figure 1.15: The Jean-Zay supercomputer. A substantial part of computations performed in the context of this work were run on this computer. © Photothèque CNRS / Cyril Frésillon.

Then, I explore how plankton and particles distribution vary during the spring bloom across a mesoscale front, and how these variations relate to changes in the environment. (iii) Finally, at submesoscale, I strive to link plankton distribution to local conditions and to assess biological interactions between planktonic organisms, including taxa co-occurences as well as prey-predator relationships. To answer these questions, I leverage *in situ* plankton imaging data from three datasets, at three different scales.

. . . using three different datasets.

1.5.2 Datasets

1.5.2.1 *The global UVP5 dataset*

This global dataset encompasses oceanographic campaigns that deployed the UVP5 on vertical profiles down to 6000 m, in combination with environmental sensors (temperature, salinity, oxygen, and fluorescence). This dataset was recently published [205], but only focusing on particles. In this work I used a slightly modified version containing 2517 profiles, performed between 2008 and 2019 and spanning the world ocean. A total of 6.8 million objects were imaged with the UVP5 during these profiles, of which 330,000 were identified as planktonic

A worldwide, global dataset.

Table 1.2: Size of the VISUFRONT dataset.

	Transects			
	Along-current	Cross-current	Lagrangian	Total
Duration (H:m:s)	16:29:33	51:26:10	22:26:54	90:22:37
Image length ($\times 10^9$ px)	1.7	5.2	2.3	9.1
Image length (km)*	8.5	26.4	11.5	46.5
Flattened image length ($\times 10^{12}$ px)**	3.4	10.6	4.6	18.7
Flattened image length ($\times 10^3$ km)	17.4	54.2	23.6	95.2†
Observed volume (m^3)	6,412.3	19,998.4	8,728.9	35,138.6‡
Storage size (TB)	2.5	9.2	4.1	15.8

*Given that 1 px = 51 µm.
**This is also the total number of pixels.
†This corresponds to one fourth of the lunar distance.
‡This represents the volume of water contained in about 14 Olympic-size swimming pools.

organisms. This data allowed us to resolve the global typology of zooplankton communities in the upper 500 m of the ocean (i).

1.5.2.2 *The VISUFRONT dataset*

A fine-scale, summer snapshot dataset.

At a much restricted scale, the VISUFRONT campaign took place during one week in July 2013 and targeted the Ligurian front (NW Mediterranean Sea). Sampling consisted of transects across and along the front, as well as "Lagrangian" transects following a water mass. This campaign made use of the ISIIS, a high resolution *in situ* imaging instrument with a very high sampling rate (> 100 L s^{-1}), originally designed to image fish larvae, but even more effective to image other kinds of plankton. The dataset was partially analysed in a previous PhD thesis [123] (2012-2015) focusing on fish larvae, but tools were not powerful enough to unlock its huge potential. Developing such tools constituted a substantial part of this PhD (presented in Part ii), and allowed a thorough exploitation of the very large and dense VISUFRONT dataset (Table 1.2). Data analysis revealed the fine-scale distribution as well as *in situ* behaviour of Rhizaria (iii). However, this dataset only presents a summer snapshot of plankton distribution, without any insight regarding the temporal aspects.

1.5.2.3 *The glider + UVP6 dataset*

To fill the gap between the global and fine-scale dataset, a third, intermediate dataset was collected. Indeed, with the aim to capture temporal dynamics of plankton and particles distribution across the Ligurian front during the spring bloom (ii), we conducted a 5-month campaign involving a SeaExplorer glider equipped with a UVP6, for a sampling rate between 0.35 and 0.9 L s^{-1}. The campaign took place between January 28th and June 28th, and consisted of repeated transects across the front. Ten missions were conducted, each mission (12 to 14 days) consisting of two round trips each, separated by a few days dedicated to maintenance but also dictated by weather conditions for deployment. During these 5 months, the glider performed more than 5,000 profiles, for a total of 1,123,123 images captured by the UVP6.

An intermediate dataset, both in time and space.

However, data science tools to thoroughly analyse the VISUFRONT dataset were not available yet and had to be developed. These methodological developments constituted a substantial part of this work.

1.5.3 Methodological developments

To make the most of the very large VISUFRONT dataset, it was first required to be able to detect and identify planktonic organisms from images. The optical properties that give the ISIIS its qualities as a plankton imager also tend to complicate data processing: huge amounts of images are collected, with a constantly changing background. This prevents both human-based detection of organisms – except at the cost of a severe subsampling – and the application of a classic threshold detecting dark objects on a white background. Thus, a pipeline was developed to tackle this challenging detection task and allowed us to extract several millions of planktonic organisms from the raw ISIIS images. However, these millions of objects then had to be sorted into taxonomic or morphological classes, a daunting task which once again cannot be completed by a human being. Hence, a classification model had to be trained to recognise between plankton classes and sort collected images.

Analysing these datasets required methodological developments in data science.

1.5.4 Work structure

This work is divided into four main parts: this introduction, methodological developments, ecological results and discussion of our findings in regards to existing knowledge. Part ii addresses the implementation of an AI-based two steps pipeline to process the huge amount of data collected during the VISUFRONT campaign. Chapter 2 tackles the detection of planktonic organisms in raw ISIIS images and Chapter 3 presents a comparison of plankton image classifiers, including the classifier that was used to sort planktonic organisms detected in ISIIS images. Part iii covers ecological results that emerged from this work, from the largest to the smallest scale. Chapter 4 focuses on the worldwide distribution of plankton community types. Chapter 5 describes the spring dynamics of plankton and particles across a mesoscale front. Chapter 6 highlights the very fine-scale distribution and behaviour of planktonic organisms. Finally, Part iv is dedicated to the discussion of our results, both methodological and ecological.

Part II

Artificial intelligence for ISIIS data processing

This section discusses the processing of ISIIS imagery data. Segmentation, i.e. the detection of planktonic organisms in the raw images, is the first step of this fully automated processing. The second step consists of the taxonomic identification of the previously detected planktonic organisms, using a classification model. The methods are described in the two papers included in this chapter.

2

Content-aware segmentation of objects spanning a large size range: application to plankton images

This chapter tackles the detection of planktonic organisms in images captured by the ISIIS during the VISUFRONT campaign. This task, far from being trivial, due to the imaging properties of the ISIIS, required the development of `apeep`[1], an ISIIS-specific and intelligent segmentation pipeline. In this paper, we describe this pipeline and perform a comparison with two other segmentation methods. Overall, although our pipeline was not the fastest and required quite a heavy set-up, it achieved the best compromise between accurate detection of planktonic organisms and relatively low pollution from non planktonic objects. The segmentation of the whole VISUFRONT data required ~2600 hours of computation on GPU and was achieved thanks to the deployment of `apeep` on three supercomputers. Indeed, besides the computation server available at the LOV, computing resources were provided by the IDRIS on the Jean-Zay supercomputer and by the Roscoff Bioinformatics platform ABiMS. In the end, a total of ~160 million objects were detected by the segmentation pipeline.

Thelma Panaïotis, Louis Caray–Counil, Ben Woodward, Moritz S Schmid, Dominic Daprano, Sheng Tse Tsai, Chris Sullivan, Robert K Cowen and Jean-Olivier Irisson

Modified version of the article Panaïotis et al. 2022, Content-Aware Segmentation of Objects Spanning a Large Size Range: Application to Plankton Images, *Frontiers in Marine Science* 9

1 https://github.com/jiho/apeep

DOI: 10.3389/fmars.2022.870005

Abstract

As the basis of oceanic food webs and a key component of the biological carbon pump, planktonic organisms play major roles in the oceans. Their study benefited from the development of *in situ* imaging instruments, which provide higher spatio-temporal resolution than previous tools. But these instruments collect huge quantities of images, the vast majority of which are of marine snow particles or imaging artefacts. Among them, the In Situ Ichthyoplankton Imaging System (ISIIS) samples the largest water volumes (> 100 L s^{-1}) and thus produces particularly large datasets. To extract manageable amounts of ecological information from *in situ* images, we propose to focus on planktonic organisms early in the data processing pipeline: at the segmentation stage. We compared three segmentation methods, particularly for smaller targets, in which plankton represents less than 1% of the objects: (i) a traditional thresholding over the background, (ii) an object detector based on maximally stable extremal regions (MSER), and (iii) a content-aware object detector, based on a Convolutional Neural Network (CNN). These methods were assessed on a subset of ISIIS data collected in the Mediterranean Sea, from which a ground truth dataset of > 3,000 manually delineated organisms is extracted. The naive thresholding method captured 97.3% of those but produced ~340,000 segments, 99.1% of which were therefore not plankton (i.e. recall = 97.3%, precision = 0.9%). Combining thresholding with a CNN missed a few more planktonic organisms (recall = 91.8%) but the number of segments decreased 18-fold (precision increased to 16.3%). The MSER detector produced four times fewer segments than thresholding (precision = 3.5%), missed more organisms (recall = 85.4%), but was considerably faster. Because naive thresholding produces ~525,000 objects from 1 minute of ISIIS deployment, the more advanced segmentation methods significantly improve ISIIS data handling and ease the subsequent taxonomic classification of segmented objects. The cost in terms of recall is limited, particularly for the CNN object detector. These approaches are now standard in computer vision and could be applicable to other plankton imaging devices, the majority of which pose a data management problem.

Résumé

En tant que base des réseaux trophiques océaniques et élément clé de la pompe à carbone biologique, les organismes planctoniques jouent un rôle majeur dans les océans. Leur étude a fortement bénéficié du développement d'instruments d'imagerie *in situ*, qui offrent une résolution spatio-temporelle plus élevée que les outils précédents. Néanmoins, ces instruments collectent d'énormes quantités d'images, dont la grande majorité sont des images de particules de neige marine ou des artefacts d'imagerie. Parmi eux, l'*In Situ* Ichthyoplankton Imaging System (ISIIS) échantillonne possède le plus grand taux d'échantillonnage (> 100 L s^{-1}) et génère donc de très grandes quantités de données. Pour extraire des quantités raisonnables d'informations écologiques à partir de ces images *in situ*, nous proposons de nous concentrer sur les organismes planctoniques dès le début du processus de traitement des données, c'est-à-dire à l'étape de la segmentation. Nous avons comparé trois méthodes de segmentation, en nous focalisant sur les cibles les plus petites, pour lesquelles le plancton représente moins de 1% des objets : (i) un seuillage naïf d'image, (ii) un détecteur d'objets basé sur les régions extrémales maximalement stables (maximally stable extremal regions, MSER), et (iii) un détecteur d'objets sensible au contenu, basé sur des réseaux de neurones à convolutions (convolutional neural net ork, CNN). Ces méthodes ont été évaluées sur un sous-ensemble de données ISIIS collectées dans la mer Méditerranée, dont est extrait un ensemble de données de vérité terrain de plus de 3 000 organismes manuellement détourés. La méthode naïve de seuillage a détecté 97,3% de ces organismes, mais a produit environ 340 000 segments, dont 99,1% n'étaient donc pas du plancton (rappel = 97,3%, précision = 0,9%). En combinant le seuillage avec un CNN, quelques organismes planctoniques supplémentaires ont été manqués (rappel = 91,8%) mais le nombre de segments a été divisé par 18 (la précision passant à 16,3%). Le détecteur MSER a produit quatre fois moins de segments que le seuillage (précision = 3,5%) mais a manqué plus d'organismes (rappel = 85,4%), en étant toutefois considérablement plus rapide. Étant donné que le seuillage naïf produit ~525 000 objets à partir d'une minute de déploiement ISIIS, les méthodes de segmentation intelligentes améliorent considérablement le traitement des données ISIIS et facilitent la future classification taxonomique objets segmentés, pour un coût limité en termes de rappel, en particulier pour la méthode CNN.

Ces approches sont désormais standard en vision par ordinateur et pourraient être applicables à d'autres dispositifs d'imagerie du plancton, dont la majorité partagent un problème de gestion et traitement d'une grande quantité de données.

2.1 Introduction

2.1.1 Plankton imaging enables fine scale studies

Plankton is diverse both in terms of taxonomy and size.

As detailed in the general introduction (section 1.2.2), planktonic organisms play crucial roles in the ocean: photosynthetic phytoplankton is responsible for about half of the primary production of the biosphere [131] and is the basis of oceanic food webs [125]; zooplankton acts as a trophic link between phytoplankton and higher trophic levels [136, 413] and is a key component of the biological carbon pump, sequestering organic carbon at depth [246]. Plankton comprises organisms from very diverse taxonomic groups [105] that span from micrometer scale picoplankton to meter-long Cnidarians [243]. Given this very wide size range, plankton sampling instruments cannot tackle all organisms at once and typically target a reduced size range instead [243].

The planktonic world is dominated by small organisms.

The power law underlying plankton or marine snow particle size spectra means that concentration drastically increases when size decreases: the relationship is linear in log-log form [243, 362, 363, 380]. The larger organisms, which each contribute significantly to biomass, are rare but easy to detect. Yet, it is critical to also focus on the smaller objects, to avoid artificially cutting the effective size range of any instrument, thus potentially discarding the most numerous objects in the sample [243]. Moreover, as marine snow particles cannot grow past a few centimetres because of disaggregation [6, 7], the ratio of particles to plankton also decreases with increasing size. Therefore, while targeting small planktonic organisms is desirable, it comes with the difficulty of separating them from the largely dominant particles within the same size range.

While large scale plankton distribution patterns are resolved to a certain extent [55, 188, 340, 345, 399], much remains to be discovered regarding fine scale distribution, in particular for zooplankton. For phytoplankton, submesoscale dynamics are known to influence their distribution and concentration: vertical currents may affect nutrient

and cell distribution relative to the euphotic zone, thus affecting growth rates, horizontal currents can stir patches into filaments. These changes are expected to propagate to higher trophic levels (zooplankton, fish, etc.) [230]. Indeed, the trophic and reproductive interactions of zooplankton occur at the scale of organisms (µm to cm). Therefore, a local concentration of phytoplankton, in a thin layer for example, has more immediate consequences on the survival and development of zooplanktonic grazers than the average chlorophyll *a* concentration in the region. Thus, studying zooplankton distribution at fine scales, in relation with submesoscale dynamics, becomes relevant to understand the processes driving its distribution at regional scale.

Understanding plankton distribution at submesoscale is essential.

Our lack of knowledge regarding the fine-scale distribution of plankton partly stems from the difficulty to adequately sample it at such a small scale. Traditional plankton collection methods such as pumps, nets, and bottles typically integrate organisms over some vertical and/or horizontal distance and make it difficult to associate organism concentrations with their immediate environmental context [30, 243, 330]. Moreover, most damage fragile organisms and fail to sample some of them properly [330].

Nets and pumps do not have sufficient resolution and damage fragile organisms.

As an alternative, *in situ* pelagic imaging instruments such as the Imaging FlowCytoBot (IFCB) [298], the *In Situ* Ichthyoplankton Imaging System (ISIIS) [85], the Underwater Vision Profiler (UVP) [317], and the Scripps Plankton Camera (SPC) [302] (see [243] for a detailed list) allow studying plankton distribution at all scales: from the fine ones they resolve in each sample to long time scales and global spatial coverage through the accumulation of individual samples [133, 192, 337, 382]. As a non-destructive sampling approach, these instruments allow investigating fragile planktonic objects, such as Rhizaria [40, 42, 104], Cnidaria and Ctenophora [248], or marine snow aggregates [161, 162]. Still, *in situ* imaging systems typically sample smaller volumes than plankton nets [243], limiting their quantitative application to abundant taxa. To quantify rarer planktonic groups, sampling effort has to be increased to improve the chances of detection. For example the ISIIS was initially developed with a very high sampling volume to study the very sparsely distributed fish larvae. Because of this, all *in situ* imaging instruments collect vast amounts of data, although the acquisition rate varies from one instrument to the next. ISIIS, for instance, collects up to 11 million objects per hour of sampling, while IFCB collects images at a

Many plankton imaging instruments exist, they allow sampling fragile organisms...

rate of ~10,000 per hour [372]. Thus all these systems need efficient and automated data processing approaches, albeit with different stringency.

. . . and resolving fine scale plankton distribution.

In addition, high resolution sampling is required to tackle questions that used to be out of reach, such as fine-scale plankton distribution in relation to environmental conditions [59, 154, 265], plankton patch structure [337], interactions between zooplankton and phytoplankton fine layers [153, 156, 355] or co-occurrences revealing biological interactions such as predation [152, 155, 354, 395].

2.1.2 Objects need to be extracted automatically from pelagic images

The main processing step is segmentation of organisms from raw images.

The first data processing step is separating relevant organisms and particles from the background in raw images, i.e. image segmentation. Various segmentation methods have been applied for images collected by commonly used *in situ* imaging devices: the UVP relies on a fixed grey level threshold [317], the IFCB uses an algorithm based on edge detection [298], the SPC [302] runs a canny edge detector to initialise the segmentation of its dark-field microscopy images. To segment images generated by the Zooglider, a glider equipped with a shadowgraph, Ohman et al. [293] also applied a canny edge detector. Finally, to segment shadowgrams from the ISIIS, Tsechpenakis, Guigand, and Cowen [402] and Iyer [194] used statistical modelling of the background of the image and identified anomalies over this background as objects of interest.

ISIIS requires an adapted processing because it uses shadowgraphy and generates more data than other instruments.

The ISIIS is deployed in an undulating manner, between the surface and a given depth [85]. It targets organisms in the range 250 µm - 10 cm. Together with greyscale images, it continually records environmental variables (temperature, salinity, fluorescence, dissolved oxygen and irradiance). The use of shadowgraphy combined with a specific lens and lighting system provide a large depth of field and allow a high sampling rate (28 kHz line scan camera). Therefore, the ISIIS is capable of sampling volumes of waters larger than all other *in situ* imaging instruments (> 100 L s^{-1}; [243]). This optical design also ensures that the organism's size is not affected by its position within the depth of field. Shadowgraphs are also able to detect heterogeneities in the medium that is traversed by the light, which makes them excellent to image transparent organisms such as plankton, gelatinous organisms in particular. But it also makes them sensitive to other sources of heterogeneity, such as suspended particles or water density changes. ISIIS may thus

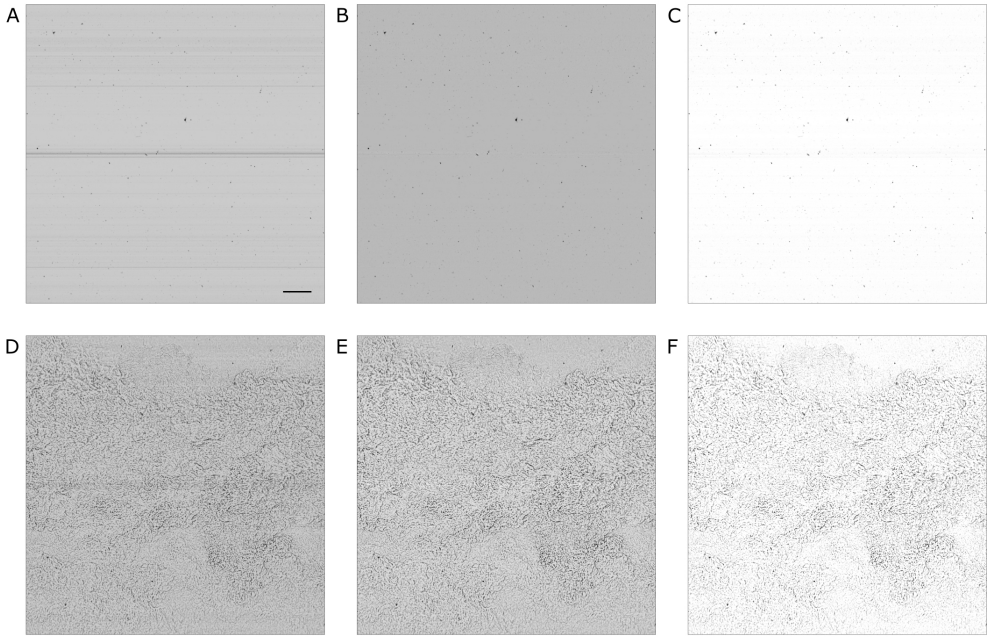


Figure 2.1: ISIIS frames in clean waters **(A-C)** and across a density change **(D-F)**. The signature of this density change is similar to what a shadowgraph would image in air, above a burning candle. The panels are: **(A, D)** raw output; **(B, E)** after flat-fielding; **(C, F)** after contrasting. The camera scans vertically and the image is acquired from the right edge, as ISIIS moves through the water. In panel A, the scale bar represents 1 cm and is applicable to other panels.

generate noisy images when deployed in turbid waters [151, 249] or across strong density gradients (Figure 2.1D-F) [124]. Furthermore, the use of a line scan camera means that marks or dust on the lens cause continuous streaks in the generated images (the line continuously scans the same speckle; Figure 2.1A, D). Those can be partially removed by applying a flat-fielding procedure, whereby the average grey value computed per row over a few thousand scanned lines is subtracted from the incoming new values (Figure 2.1B, E) [124, 151, 249].

The very characteristics that give the ISIIS its qualities as a plankton imager (large sampling volume, high speed, ability to detect transparent objects) also mean that it creates a huge amount of images, the background of which is often non-uniform. This makes segmentation of planktonic objects from raw images far from trivial. To perform this segmentation, the processing pipeline was initially based on anomalies

Table 2.1: Threshold in object area (number of pixels considered as part of the object) in studies exploiting ISIIS data. The conversion factor from area (px) to Equivalent Spherical Diameter (ESD, mm) depends on the ISIIS configuration.

Reference	Area threshold (px)	ESD (mm)
[354]	7	0.2
[249]	50	0.53
[124]	250	0.92
[158]	400	0.95
[153]	900	1.4
[152]	2000	3.0
[151]	5000	5.4

from a gaussian mixture model of the background grey levels without flat-fielding [402] and later on k-harmonic means clustering on flat-fielded images [194]. This latter method was used in several studies [151, 249, 354] and the full pipeline was open sourced in order to make plankton imaging more accessible and lower entry barriers [353]. Other studies relied on flat-fielding followed by segmentation above a fixed grey level [124, 153, 158]. However, most of these studies focused on the larger end of the size range targeted by the ISIIS, by considering only objects above a given size threshold (Table 2.1), often because those were desirable targets, not noise. Similarly, for their canny edge detector applied to ZooGlider images, Ohman et al. [293] considered objects larger than 100 pixels (Equivalent Spherical Diameter, or ESD of 0.45 mm). However, the algorithm failed when too many particles were present and had to fall back to a less sensitive (i.e. higher) grey threshold. As shown above, both planktonic organisms and particles are much more abundant towards the smaller end of the spectrum, meaning that such methods had to ignore a non-negligible part of planktonic organisms and marine snow in order to discard the background noise.

Several segmentation methods were applied to ISIIS data but they all required discarding a non-negligible part of the data.

2.1.3 Marine snow and imaging artefacts dominate *in situ* images and complicate plankton detection

Marine snow particles are much more abundant than plankton in the ocean [243], which means that the vast majority (often > 85%) of images captured by *in situ* plankton imaging instruments are actually

of various marine snow items (faecal pellets, large aggregates, small organism pieces, etc.; [317, 380, 383]). Therefore, for plankton ecology studies, the bottleneck has often become the processing and filtering of collected images [192]. To reduce the proportion of detrital particles and focus on photosynthetic plankton, the IFCB and the FlowCam can use fluorescence image triggering, hence imaging only items that contain chlorophyll [365, 372]. This is not possible over the large volumes and for the non-photosynthetic organisms that ISIIS or other zooplankton imagers target. Furthermore, density anomalies lead to the characteristically noisy shadowgrams presented above (Figure 2.1D-F), from which numerous artefactual "particles" are detected by the usual image processing pipelines. Those artefacts or noise, together with marine snow, can constitute 99% of the objects detected. Such an extreme class imbalance makes the automatic classification of these objects through machine learning a very arduous task [224].

ISIIS images are strongly dominated by noise and marine snow...

Even for a trained human operator, the differentiation of some planktonic classes from the proteiform marine snow aggregates and noise, as well as distinction between marine snow and noise themselves, can be very challenging. Towards the smaller end of the size spectrum it becomes virtually impossible. Indeed, once these small objects are segmented out, the low pixel count combined with the lack of information regarding their context in the image makes their identification very difficult, for humans and computers alike [307]. Hence, one solution could be to focus solely on planktonic organisms from the segmentation step already and try to avoid segmenting non-planktonic objects, thanks to their broader context in the image, still accessible at this step. This should result in a much more manageable amount of data to classify and a lesser class imbalance. This approach requires the development of specific and "intelligent" segmentation methods that target specific objects only. The purpose of this work was (i) to develop such "intelligent" segmentation approaches and (ii) to compare them with classic methods to test whether they significantly improve the data processing pipeline. With this in mind, we benchmarked three segmentation methods against a ground-truth human segmentation using a dataset collected by the ISIIS in the North-Western Mediterranean Sea.

...which can be difficult to distinguish.

Thus, one solution is to detect plankton only.

2.2 Materials and methods

2.2.1 Image segmentation methods

2.2.1.1 *Threshold-based segmentation*

A classic segmentation method consists of detecting dark objects.

The simplest segmentation method is to threshold pixels below a given grey level: adjoining pixels darker than the threshold are considered as segments. This threshold can be a value fixed *a priori* or dynamically computed from the properties of each image. For example, the classic method of Otsu [303] is to examine the histogram of intensity levels and define the threshold so that it separates pixels into two relatively homogeneous intensity classes. Here either a fixed threshold was set or the threshold was defined based on a quantile of the histogram of grey levels. This quantile-based approach resulted in a darker segmentation threshold on noisy images, such as those captured around the strong density gradient induced by the thermocline (Figure 2.1 D-F), which were richer in dark pixels. It was well adapted to limit the number of artefact segments generated from these images. Moreover, the first quartile is barely affected by the presence of relatively large dark objects such as jellyfish tentacles, making the segmentation threshold robust to these natural occurrences. After thresholding, segments defined by connected components were dilated by 3 pixels and eroded by 2 pixels to fill potential holes in transparent organisms and reconnect thin appendages to the organisms bodies. Finally, only segments larger than 50 pixels (400 µm in ESD) were retained, because it was the minimum size at which taxonomists could recognise organisms.

2.2.1.2 *Threshold-MSER (T-MSER) segmentation*

Alternatively, a blob-detection algorithm detects regions that differ in properties.

This approach uses a signal-to-noise ratio (SNR) cutoff, calculated on images after flat-fielding, to determine whether the frame should be processed using a Maximally Stable Extremal Region approach (MSER, [261]), or if areas of high noise should first filtered out using a naive thresholding approach before applying MSER. MSER was successfully applied to the segmentation of ZOOVIS imagery [37, 76]. SNR can be used to determine the relative noise level in an image and was computed as

$$\text{SNR} = 20 \times \log\left(\frac{\text{S}}{\text{N}}\right)$$

where S is the signal, defined as the mean of the input data, and N is the noise, computed as the standard deviation around that mean. Here, flat-fielded frames with low SNR (i.e. high noise) were binarised using a fixed thresholding in order to extract continuous regions of interest with darker pixel values. The regions identified in this way were then extracted using a mask and subsequently re-segmented using the MSER approach. MSER detects stable connected regions in images, which are areas that stay nearly unchanged over a wide range of greyscale thresholds. MSER can be tuned to allow for varying degrees of stable region area and the range of pixel grey values tested in the dynamic thresholding. High SNR frames are directly segmented using the MSER approach (Figure 2.2 skip from step B to step D). Going from a pure MSER approach to the threshold+MSER (T-MSER) on low SNR (< 50) frames increased the recall on the test data from 65% to 85%, while also substantially increasing precision. This SNR and MSER method is written in C++17. The OpenCV and OpenMP Python packages were used for general computer vision and parallel processing for high processing efficiency, respectively.

2.2.1.3 *Threshold-CNN (T-CNN) segmentation*

Another method is to use an intelligent detector to target planktonic organisms only...

Another solution is to use Convolutional Neural Networks to either detect (i.e. define bounding boxes around) or segment (i.e. define a pixel mask of) objects of interest. Such approaches open the possibility to focus the detection on some types of objects (here, plankton) and ignore others (here, marine snow and artefacts); this is also called content-aware object detection or segmentation. However, CNNs tend to underperform at detecting objects across a large size range, especially for objects starting from a few dozen pixels [65]. They work best when the target objects are of the same size as the receptive field of the model [114]. Thus, the development of detectors implementing receptive fields of various sizes constituted a major improvement, as they allowed detecting objects across a larger size range [65]. In particular, we chose the Detectron2 library [420] developed by Facebook AI Research, which provides state-of-the-art object detection and segmentation algorithms, as well as pre-trained models for such tasks.

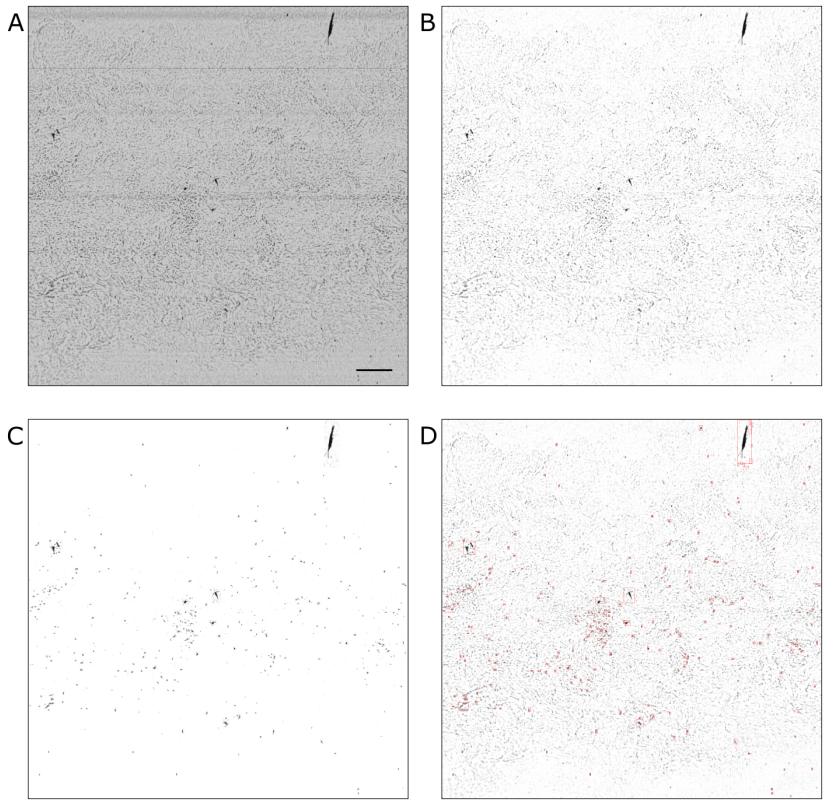


Figure 2.2: Example MSER segmentation of a noisy raw frame (with low SNR). **(A)** Raw output; **(B)** after flat-fielding; **(C)** regions of interest created through naive thresholding; **(D)** regions of interest and their bounding boxes created by applying MSER to **(C)**. In a low SNR frame such as the one above the processing steps are **A-B-C-D**, while in a high SNR frame the processing steps are **A-B-D**. In panel A, the scale bar represents 1 cm and is applicable to other panels.

Detectron2 includes a feature pyramid network [236] backbone that extracts feature maps across multiple scales to enable the detection of objects of various sizes, which was critical in our application to plankton images. Yet, this was not enough to cover the very large size range of organisms imaged by the ISIIS (from 50 to hundreds of thousands of pixels in area).

As explained above, marine snow particles and density-induced imaging artefacts are especially dominant compared to plankton in the smaller size classes. Therefore, our CNN pipeline was set up to segment the smaller objects, from 50 to 400 pixels in area, where the ability to specifically segment plankton makes the most difference. Above 400 pixels, the quantile-based threshold approach, with dilation and erosion, was used because it was simple and did not generate too many non-plankton segments.

. . . in the size range dominated by marine snow and noise.

In Detectron2, we used Mask R-CNN [173], which allows simultaneous bounding box detection and instance segmentation. The model was initialised with weights trained on the COCO reference dataset[2] but, for it to detect planktonic organisms on ISIIS images, it had to be fine-tuned on a dataset of ground truth bounding boxes and masks of such organisms. This dataset was generated by manually delineating all recognisable planktonic organisms in a set of ISIIS images, using a digital pen on a tablet computer. This produced 23,197 ground truth masks, from which bounding boxes were computed. Among those, 10,878 objects were in the 50-400 pixels area range and usable. A 524×524 pixels crop was generated around every ground truth object (pushing the crop back inside the image when it crossed the edges). The choice of this particular size is a trade-off between the maximum size of planktonic organisms that can be detected and the memory available on the graphics card. Moreover, it is in line with common input sizes for segmentation models and was convenient to generate a tiling on ISIIS images. Several objects could be present in a crop. The crops were then split into 70% for training, 15% for validation, and 15% for testing. This split was stratified by the average grey level of the crop to ensure that both noisy (darker) and clean (lighter) images were present in each split, so that the model was presented with all kinds of images during training. Indeed, a model trained on clean images only would have performed poorly on noisy ones.

A training dataset was generated by manually delineating planktonic organisms.

2 https://github.com/facebookresearch/detectron2/blob/main/configs/COCO-InstanceSegmentation/mask_rcnn_R_50_FPN_3x.yaml

One-class object detection was chosen over multiclass detection.

Detectron2 can perform both multiclass object detection or segmentation, meaning that objects are both detected/segmented and classified in a single step. However, it requires sufficient examples in each class for training. This condition could not be satisfied here, given how time-consuming it was to obtain pixel-level masks for every object and because plankton samples are usually dominated by a few abundant taxa while most others are very rare [361]. Since the focus of this study is on segmentation, we decided to perform one-class object detection/segmentation, thus training the model to recognise planktonic organisms of any taxon. This implies that classification needs to be done after segmentation. Once an object is detected, this sequential, rather than concurrent, approach does not affect the result of the classification, since the same information is available to the subsequent classifier as to the concurrent one. Furthermore, focusing on segmentation only is also more comparable with the two other methods described above.

The CNN segmenter was first fine-tuned on the training set...

The model was trained for 30,000 iterations, and evaluation was run on the validation set every 1,000 iterations to ensure that the validation loss reached a plateau. The learning rate was set to 0.0005 initially and decreased 10 fold after 10,000 and 20,000 iterations. To increase the generality of the detector, data augmentation was used in the form of random resizing of the 524 pixels crops (to 640, 672, 704, 736, 768 or 800 pixels) and random horizontal flipping. The test set was used to assess theoretical performance after training and guide the choice of model settings; the actual performance was assessed on a separate, real-world dataset (presented below).

...before it could be applied to new images.

To apply the trained model to new images, a tiling of 524×524 pixels crops (the size used during model training) was generated over each input image, resulting in an overlap of 143 pixels vertically and 135 pixels horizontally. The overlap ensured that detectable objects spread over two crops were not missed. Crops were upscaled to 900×900 pixels to improve detection of small objects [114]. For each crop, the model predicted the bounding boxes of objects and their masks. We only considered the boxes, resolved overlaps in detections caused by overlapping crops, and submitted each box to exactly the same quantile-based thresholding as what was used above 400 pixels. This was preferred over using Detectron's mask proposals because their outline was not as detailed or replicable as the threshold-based ones. Furthermore, it also ensured that morphometric measurements performed on the masks (area in particular) were exactly comparable

between the objects that went through the CNN and those above 400 pixels that were defined by simple thresholding. For each bounding box proposal, the model computes a confidence score. We retained all boxes with a score over 0.1, which is a quite low confidence threshold designed to increase the chance of detecting all objects of interest (i.e. favour recall) at the cost of some false positive detections (i.e. lower precision). Those false positives (i.e. segmented objects that are not plankton) will have the opportunity to be eliminated later, when segments are classified taxonomically.

The CNN was coded in Python with PyTorch, the original implementation library for Detectron2. Training was conducted on a Nvidia Quadro RTX 8000 GPU. The code is available at `https://github.com/ThelmaPana/Detectron2_plankton_training`. The combined CNN and threshold segmentation pipeline is implemented in `https://github.com/jiho/apeep` and this was run in several Linux-based environments, using various Nvidia GPUs.

2.2.2 Application to ISIIS data from VISUFRONT campaign

The segmentation methods are applied to ISIIS data from the VISUFRONT campaign.

We evaluated these segmentation methods on ISIIS data collected during the VISUFRONT campaign, which sampled the Ligurian current front (NW Mediterranean Sea), in the 0-100 m depth range, during July 2013. Towed at a speed of 2 m s^{-1} (4 kts) and set for a 28 kHz scanning rate, the ISIIS sampled 108 L per second. The 2048 pixels high continuous image strip created by the line scan camera moving in the water was cut in 2048×2048 pixels frames for storage. The ISIIS captured marked volutes caused by water density variations (Figure 2.1 D-F), mostly driven by temperature changes around the thermocline, previously described by Faillettaz et al. [124].

Images were pre-processed before segmentation.

The continuous image strip was reassembled from the stored frames (2048×2048 pixels). Each line of pixels was flat-fielded by subtracting the row-wise average over a 8000 pixels moving window, hence removing streaks (Figure 2.1 A to B, D to E). The cleaned image was cut into 10,240 pixels long images (5 frames, instead of 1) to reduce the probability of cutting objects across images while keeping the memory footprint of each image manageable. Finally, the image was contrasted by stretching the intensity range between percentiles 0 and 40 (Figure 2.1 B to C, E to F). These values were chosen by iteration, through discussions with the taxonomist in charge of delineating planktonic

organisms from raw images, as to achieve the highest distinguishability for those.

An independant benchmark dataset was generated by manually detouring planktonic organisms.

A ground truth dataset was generated by manually delineating all planktonic organisms (using a digital pen and tablet) in 106 10,240×2048 pixels images, regularly spread across a full transect, hence representative of different environments. This resulted in 3,356 objects that were later taxonomically sorted into 24 taxa (Figure 2.3), in the Ecotaxa web application [316]. This dataset was completely independent from the one that was used to train, validate and test the Detectron2 model. Some images were checked by two independent operators to check their consistency; when this was done, no differences were found.

Outputs of segmentation methods are compared to the human ground-truth segmentation...

Segments from each of the three automated methods were matched with ground truth segments of the same image. A bounding box intersection over union (IoU) score higher than 10% was considered as a match between segments. This threshold was set after manually inspecting a set of potential matches with various IoU values and was found to be the best value to discriminate between true and false matches. In case a ground truth segment matched multiple automatic segments, only one match was retained, to avoid inflating artificially the number of matches from the automated pipelines. In case an automatic segment matched multiple ground truth segments, the match was not counted either because it corresponded to a large segment that encompassed several organisms likely belonging to different taxa, which would make it unexploitable ecologically. Both choices made the match metrics conservative.

... and metrics were computed to compare performance.

From these matches, global precision and recall were computed to summarise performance. Precision was computed as the proportion of automatic segments that matched ground truth segments. A 100% precision means that the algorithm only extracted ground truth segments. Recall was computed as the proportion of ground truth segments detected by the automated segmentation algorithm. A 100% recall means the algorithm did segment every manually delineated organism. Precision and recall scores were also computed per size class, where size was defined as the length of the diagonal of the bounding box; size classes were defined as intervals of 10 pixels, from 10 to 100 pixels, plus a class > 100 pixels. These size classes do not aim at reflecting any ecological groups but were designed to split segments into roughly balanced classes. Recall was also computed for each taxonomic group defined in the ground truth segments. Preci-

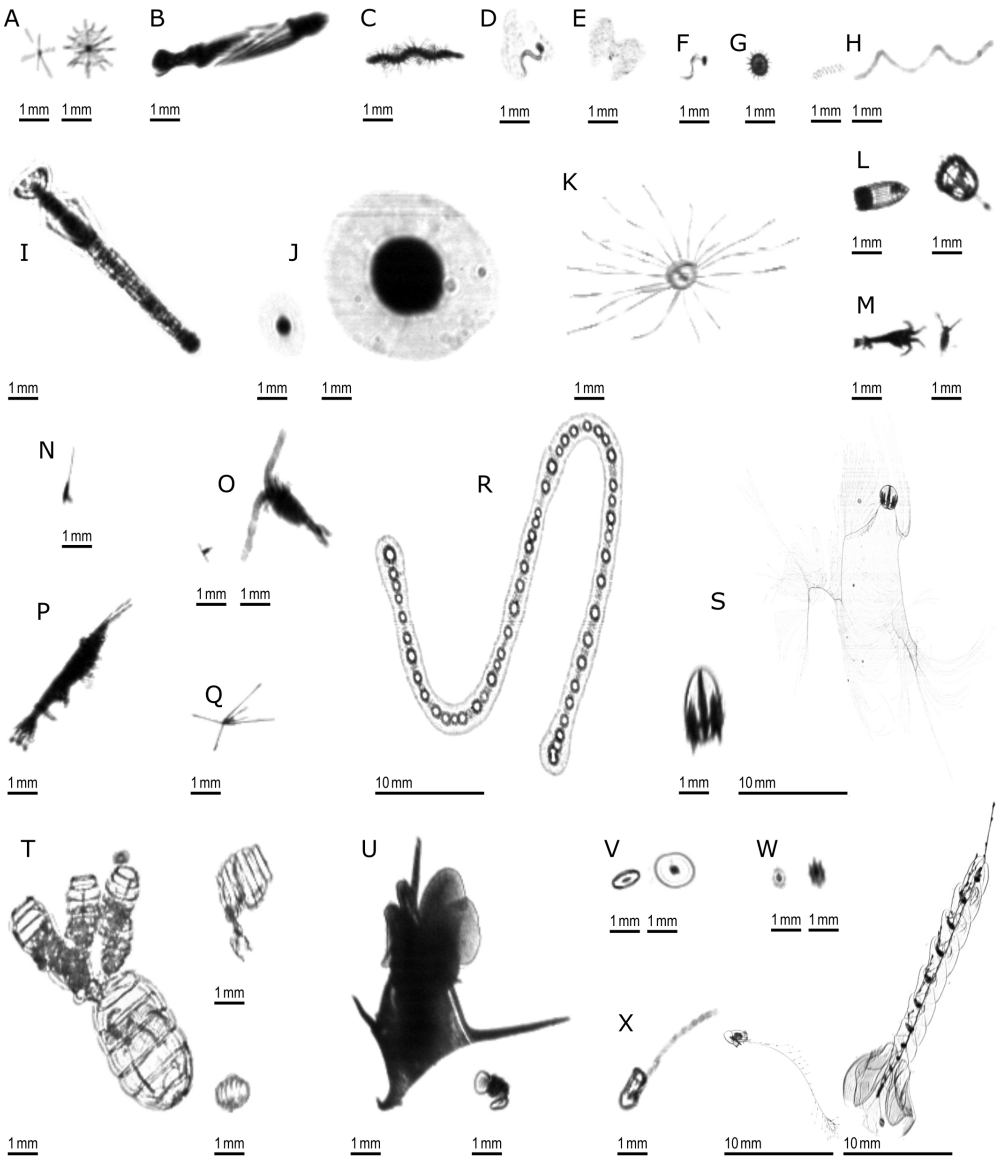


Figure 2.3: *(Caption on next page)*

Figure 2.3: Examples of planktonic organisms imaged by the ISIIS. **(A)** Acantharea; **(B)** Actinopterygii; **(C)** Annelida; **(D)** Appendicularia; **(E)** Appendicularia (house only); **(F)** Appendicularia (body only); **(G)** Aulacanthidae; **(H)** Bacillariophyceae; **(I)** Chaetognatha; **(J)** solitary Collodaria; **(K-L)** Hydrozoa; **(M)** Crustacea (other than Harpacticoida, Copepoda and Eumalacostraca); **(N)** Harpacticoida; **(O)** Copepoda (other than Harpacticoida); **(P)** Eumalacostraca; **(Q)** Echinodermata (pluteus larva); **(R)** colonial Collodaria; **(S)** Ctenophora; **(T)** Doliolida; **(U)** Mollusca; **(V)** Pyrocystis; **(W)** Rhizaria (other than Acantharea; Aulacanthidae and Collodaria); **(X)** Siphonophorae.

Table 2.2: Number of segments generated by each pipeline on the 106 benchmark images and estimation of the amount of segments they would produce on one minute of ISIIS data.

Segmentation pipeline	Number of segments on benchmark images	Average number of segments per minute of ISIIS deployment
Ground truth	3,356	~5,000
Threshold	339,907	~525,000
Threshold-MSER	82,731	~130,000
Threshold-CNN	19,048	~30,000

sion does not make sense for taxonomic groups since it would only reflect the performance of the classification, not of the segmentation. The particle matching and metric computation code is available at `https://github.com/ThelmaPana/segmentation_benchmark`.

2.3 Results

2.3.1 Number and size distribution of segments

On the 106 images of the segmentation benchmark dataset, 3,356 organisms were manually segmented, whereas the automated pipelines generated many more segments, especially the threshold-based one (Table 2.2).

Automated pipelines generated more segments than manually detected...

The normalised abundance size spectra (NASS) (Figure 2.4) displays the expected linear decrease of abundance with size in log-log scale. For the ground truth segments, the curve dips below this linear relationship for objects of 25 pixels in diagonal and smaller (dotted vertical line on Figure 2.4). Since this dataset specifically targeted recognisable planktonic organisms, this dip highlights that not all organisms below

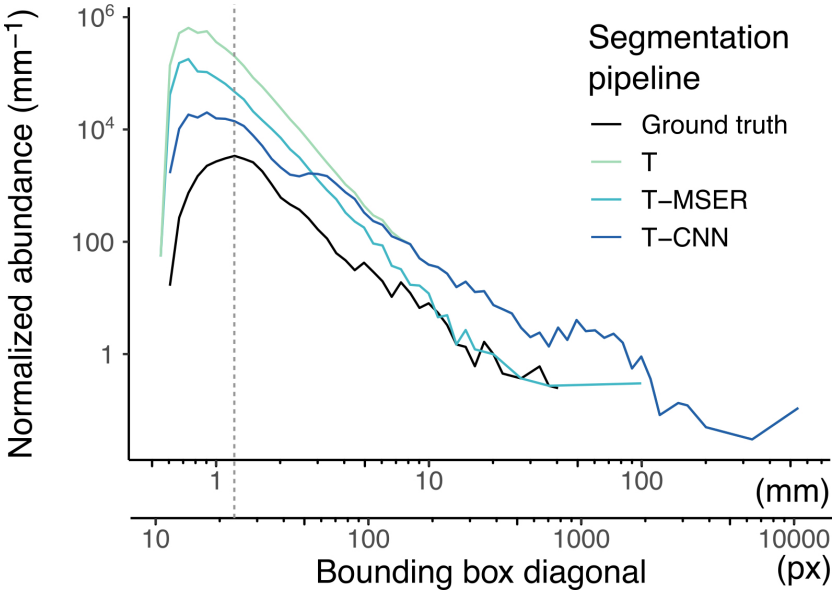


Figure 2.4: Normalised abundance size spectra (NASS) of all segments generated by the benchmarked pipelines and ground truth segmentation. To compute the NASS, segments were grouped into size classes on a log2 scale, each class size being two times wider than the previous one. Normalised abundance was computed by dividing the number of segments in each class by the size class width, resulting in an adimensional quantity (number of segments) divided by a length (mm here). The double x-axis is the length of the diagonal bounding box displayed both in pixels and after conversion in mm. The dotted vertical line highlights the slope discontinuity in the size spectrum of ground truth segments. Note that both axes use log10 scaling. T = threshold-based, T-MSER = threshold-MSER, T-CNN = threshold-CNN.

this size could be detected by a human taxonomist upon detailed examination of the images [243]. The discontinuity is towards smaller diagonal sizes in the automated pipelines, but likely because many of the small segments are of non-plankton objects.

All automated pipelines have NASS curves above the ground truth, which highlights the fact that they segmented non-plankton objects. This was true over the entire size range but was particularly pronounced for the smaller size classes. Above 10 mm/200 pixels in diagonal, the T-MSER pipeline produced a number of segments comparable to the ground truth, which is satisfying, although it does not guarantee that those are of the same objects (it might have missed some plankton and segmented marine snow/artefacts in the same size range; see

. . . especially at the bottom edge of the segmentation size range.

precision and recall performance for the largest size class in Figure 2.5 below). From the maximal size down to ~70 pixels in diagonal, the T and T-CNN pipelines produced the same segments. This coincides with the critical size of 400 pixels in area at which the segmentation method switched from threshold-based to content-aware. Indeed, the conversion from area to bounding box diagonal is not linear because it depends on the shape of the objects. For an object of 400 pixels in area, the bounding box diagonal is between 30 and 70 pixels. This shows that the T-CNN pipeline was effective in reducing the number of segments compared to naive thresholding, because the NASS diverges below that size.

NASS slopes are different.

A linear regression performed on the linear portion of the NASS (diagonal values between 30 and 500 pixels) followed by an analysis of covariance demonstrated significant difference in slopes between the segmentation methods: $F(3,105) = 133.07$; $p < 0.001$ (Table S2.1). Post hoc analysis showed a significant difference between all segmentation methods ($p < 0.001$ for all pairs) (Table S2.2).

2.3.2 Global performance statistics

The T-CNN pipeline had a better global precision.

Overall, the three pipelines demonstrate good recall: when looking at the total number of segments, they all captured over 85% of the ground truth organisms. The T-CNN pipeline largely outperformed both the threshold-based and T-MSER pipelines in terms of precision (Table 2.3). In other words, although it segmented almost all planktonic objects, the threshold-based pipeline generated mostly non-plankton segments (~99%), composed of both marine snow and density volutes artefacts. The T-CNN pipeline also produced non-planktonic segments but they "only" represented 84% of segments, while still segmenting a good proportion of planktonic objects. The T-MSER performed somewhere in between those two extremes.

2.3.3 Performance per size class

Because the behaviour of the pipelines seems to vary with size (Figure 2.4), it seems relevant to break down the matching statistics per size class. With the threshold-based pipeline, precision decreased with size: smaller segments included a lower proportion of planktonic organisms than larger ones (Figure 2.5A). The T-CNN pipeline had better precision

Table 2.3: Precision and recall values of the automated pipelines evaluated against the 3,356 ground truth organisms.

Pipeline	Precision	Recall
Threshold	0.9%	97.3%
Threshold-MSER	3.5%	85.4%
Threshold-CNN	16.3%	91.9%

than the others for small segments while T-MSER had a better precision for larger segments. In terms of recall, the threshold-based pipeline always performed better than the others, regardless of size class (Figure 2.5B). The T-MSER pipeline performed as well as the T-CNN pipeline on middle size classes, but achieved a lower recall for both very small and very large segments.

The T-CNN pipeline improved precision for small segments, but recall performance were independent of class size.

2.3.4 Performance per taxonomic group

In the ground truth dataset, half of the 24 detected taxa were represented by fewer than 18 individuals (median is 18.5), hence inducing little resolution and large variance in the performance statistics of segmentation pipelines. Among the other half of the taxa, the recall of the T-CNN pipeline was lower than that of the threshold pipeline by more than 10% for only two taxa (Bacillaryophycea and Doliolida) and for only four in the case of the T-MSER pipeline (Bacillariophyceae, Ctenophora, Acantharea, and other Rhizaria; Figure 2.6). The lowest recall values were reached for Bacillariophyceae and Ctenophora, for all pipelines. In concordance with the consistent recall performance across size classes, taxa-wise recall performance of the T-CNN pipeline do not seem linked to organism size: small organisms (e.g. Acantharea, Pyrocystis) were accurately detected.

Taxonomic-wise performance do not seem linked to organisms size.

2.4 Discussion

2.4.1 Summary of results

The threshold-based pipeline performed an exhaustive segmentation: planktonic organisms were almost all properly detected, yet they were drowned in the overwhelming majority of non-planktonic objects (Table

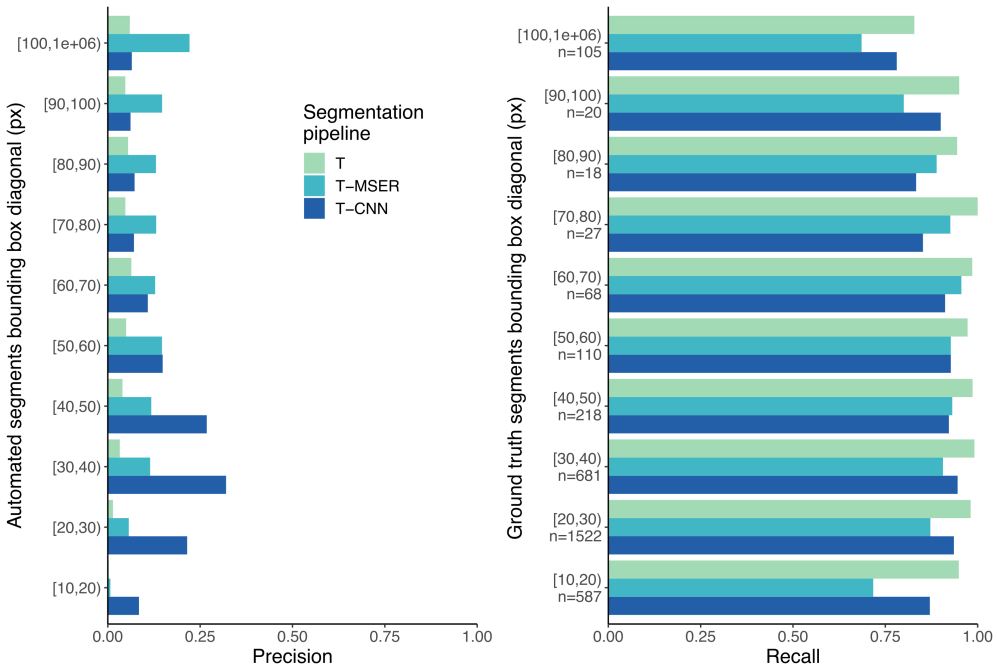


Figure 2.5: Precision **(A)** and recall **(B)** scores per size class. In **B**, n indicates the number of segments per size class for the ground truth dataset. T = threshold-based, T-MSER = threshold-MSER, T-CNN = threshold-CNN.

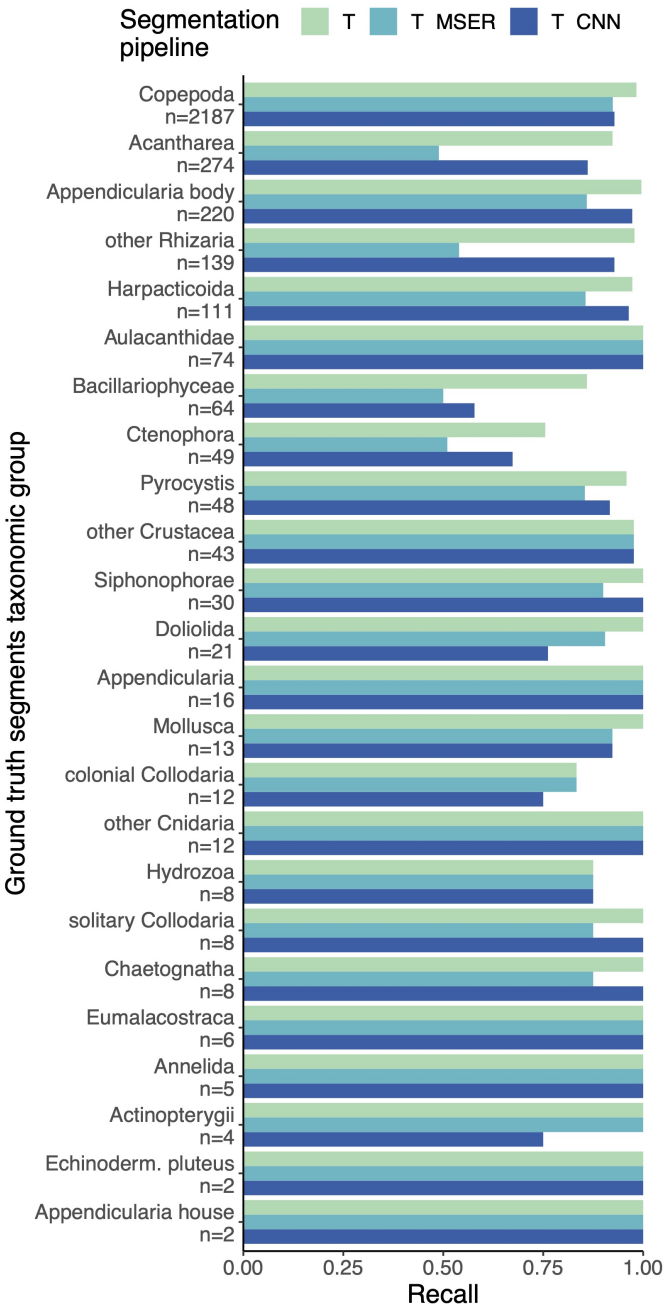


Figure 2.6: *(Caption on next page)*

Figure 2.6: Recall scores per taxon. n is the number of individuals from each taxon in the 106 benchmark images and taxa are sorted in decreasing order of abundance. T = threshold-based, T-MSER = threshold-MSER, T-CNN = threshold-CNN.

The T-CNN pipeline generated much fewer non-planktonic segments and still detected plankton accurately.

2.2). The T-CNN pipeline reduced this problem, significantly increasing precision (Table 2.3, Figure 2.5A) while still achieving a very good detection of plankton across the entire size range targeted by ISIIS. The T-MSER pipeline also reduced the segmentation of non-planktonic objects, especially at the top-end of the size range, but detected fewer planktonic organisms than the other pipelines (Figure 2.5B). Despite the large decrease in number of segmented objects, for most taxa, the MSER or CNN pipelines reduced recall by less than 10% (Figure 2.6). One explanation for these differences is that naive thresholding captured a lot of noise (i.e. density volutes) and, additionally, broke it into many small segments. The use of either MSER or a CNN allowed ignoring these noise segments and/or not breaking them apart, hence producing much fewer non-planktonic segments.

Ground-truth segmentation established a size-cutoff for detectability.

The decrease in abundance below the expected slope at the smaller end of the size spectrum of ground truth segments (Figure 2.4) suggests that identification of planktonic organisms becomes non-exhaustive below 25 pixels in bounding box diagonal. Below this size, which amounts to 600 µm in ESD on average, some organisms can still be detected. This means that relative concentrations between locations or times can likely be exploited within a taxon but that further filtering and corrections are needed to reach absolute concentrations.

Outputs of different segmentation methods should not be merged.

The statistical difference between NASS slopes (Figure 2.4) indicates that they segment different kinds and amounts of non-planktonic objects, compared to the all-plankton ground truth. This implies that the output of different segmentation approaches should not be directly compared in terms of size distribution. Segmentation methods were already shown to have an impact on the definition of particle size and shape, which propagates to subsequent analyses such as particle flux estimates [142]. This slope discrepancy as well as the vastly larger intercept of the NASS of automated pipelines compared to the ground truth means that the computation of an appropriate plankton size spectrum requires a classification step that would exclude non-planktonic objects.

2.4.2 Targeted organisms

Some taxa were systematically less often detected than others. Some of the not detected Bacillariophyceae were large, blurry, and too translucent (Figure 2.3H) to be caught by the threshold-based branch of the T-CNN pipeline or by the T-MSER method. The other, smaller, ones that were missed by the content-aware branch of T-CNN were not detected because they were quite different from the ones used during training (blurrier). Integrating more representative examples of Bacillariophyceae for CNN training could have improved performance on this taxon. Similarly, doliolids (Figure 2.3T), that were often large, should have been segmented by the threshold-based branch of T-CNN as well as by T-MSER. The ones missed, mostly by T-CNN, were also blurry and too translucent for intensity-based thresholding with a single threshold. Ctenophores (likely of the Mertensiidae family, Figure 2.3S) displayed thin, translucent tentacles that were often missed by threshold-based methods. Therefore, only the body was segmented, which resulted in a bounding box IoU value < 0.1, too low to be considered a match with the ground truth segment that included the tentacles. Still, a later CNN classifier should be able to correctly identify even such portions of organisms, as CNNs were shown to mostly rely on local shape and texture features instead of on the global shape [16, 17]. Finally, the T-MSER pipeline resulted in a lower recall for Acantharea and other Rhizaria (Figure 2.3A, W). This seems to stem from a too aggressive thresholding step in low SNR high noise frames, the preprocessing step before MSER is applied. Further fine-tuning would likely allow it to retain more or all Acantharea and other Rhizaria images.

The T-CNN and T-MSER pipelines struggled to detect some taxa.

In the present study, we aimed at performing an exhaustive detection of every planktonic organism across the size range targeted by the ISIIS. However, in general, the segmentation algorithm should be chosen according to the target organisms. For example, to focus on organisms towards the larger end of the ISIIS size range (e.g. > 10 mm), where particles – mostly marine snow aggregates – are much less abundant, a simple grey-level threshold seems sufficient.

The segmentation pipeline should be adapted to the targeted organisms.

2.4.3 Processing time and cost

The quantile-based thresholding pipeline ran on a single CPU core at a rate of 30 minutes of processing for 1 minute of ISIIS data (0.03x), on an Intel Xeon E5-2643 v3 (3.40 GHz). Its memory requirements were limited so it was easy to run simultaneous processing of multiple batches of data on a multi-core/multi-processor machine, but the treatment of ISIIS data as a continuous stream for flat-fielding prevented automatic multithreading. The T-CNN pipeline required a GPU with sufficient memory (48 GB, on a Nvidia Quadro RTX 8000 in our case) to efficiently train the CNN portion and to fit ISIIS images in at evaluation time. It processed data at the same rate as the threshold-based pipeline (30 min processing for 1 min of data, or 0.03x). The T-MSER pipeline was optimised for speed and utilised the 8 cores of an AMD Ryzen 3700, processing one minute of ISIIS data in 50 seconds (1.2x), or 6 min 40 s of processing for 1 min of ISIIS data (0.15x) when considering running on one core.

Each pipeline had its benefits but required specific hardware.

The MSER implementation followed [261] closely. The optimisation of the T-MSER approach stems from adding the SNR switch, which leads to the pre-processing of high-noise images with naive thresholding, while going straight to the MSER-based detection in low noise images. Adding these changes increased segmentation recall from 65% to 85%. Further optimisation included making the code multi-thread ready for deployment on High Performance Computing infrastructures. Using the specialised CPUs of these infrastructures, such as the AMD EPYC 7742 (64 cores, 128 threads) performance could improve well above 1.2x. At current data collection rates of 75-100 h of ISIIS data per scientific cruise, a real time or faster than real time segmentation approach constitutes a substantial benefit.

The T-MSER pipeline could be improved to be even faster.

At first glance, the T-CNN pipeline seems expensive in terms of set up and architecture: it requires a GPU with sufficient memory to operate, implies the use of relatively new deep learning coding frameworks and the preparation of a training set with manual delineation of thousands of planktonic organisms. But these costs are offset by the time gained not processing a multitude of particles in each image, resulting in a processing rate comparable to that of the pure threshold-based pipeline, as stated above. Furthermore, the fact that T-CNN produced 20 times fewer segments will also considerably reduce the classification time (often CNN based too). Finally, since recall barely decreased, the objects

A CNN may seem laborious to set up...

... but it is worth it.

ignored were mostly the dominant non-plankton objects, as per design; this will diminish the imbalance among classes that classifiers are sensitive too, further improving the classification step. Moreover, both the Detectron2 library and the baseline model on which the T-CNN pipeline relies are easily downloadable and well documented[3]. With GPU resources becoming increasingly available for scientific research and the associated frameworks becoming easier to use, such tools are poised to become more powerful and accessible.

Artificial intelligence tools and computational power are more and more accessible.

2.4.4 Detection of small objects by CNN models

The detection of objects measuring just a few pixels is still a research problem in its own right in computer sciences [113], coined very low resolution recognition problems [412]. They are characterised by targets smaller than 16×16 pixels, which can be challenging even for the perceptual abilities of human experts. They target applications for company logo detection [113, 114], face recognition from video surveillance, or text recognition [412]. The receptive fields of common object detection architectures match the target object size and range from 50×50 to 450×450 pixels which is much larger than the small objects targeted in low resolution studies [113]. Here, the smallest organisms targeted had an area of 50 pixels, which corresponded to a bounding box diagonal of 12 pixels, or an 8x8 pixels square. Thus the exhaustive detection of plankton organisms in ISIIS images, including the smaller ones, clearly falls in the domain of very low resolution recognition. A common solution is image upscaling, as highlighted by Eggert et al. [114], which we implemented in the present work. The 524×524 pixels crops were upscaled to 900×900 pixels before evaluation in the Detectron2 model. The 900 pixels size is a compromise between detection accuracy, usage of the GPU memory, and processing time. Other approaches for multi-scale object detection are described by Cai et al. [65] and include magnification of regions susceptible to contain small objects [114] or the integration of contextual information outside of regions of interest [28].

CNN have difficulties at detecting small objects...

... but a few workarounds exist.

No automated segmentation method is perfect; depending on their settings, they either avoid objects other than their targets but miss some objects of interest (high precision, low recall) or detect most objects of interest but also many others (high recall, low precision). If the

3 https://github.com/facebookresearch/detectron2

segmentation or object detection task is followed by a classification step, which is always the case for plankton imaging, we advocate in favour of recall over precision during segmentation, provided that the amount of data remains manageable. Hence, a maximum number of planktonic objects have the opportunity to be classified. The precision can be improved after classification, by filtering out low confidence, usually error prone, predictions based on the score given by the classifier [124, 249].

Favour recall over precision to extract a maximum of organisms.

To extract planktonic organisms of various taxa from ISIIS images, full instance segmentation would have been the most elegant approach, outputting classified mask instances in a single step [94]. Several obstacles still lay ahead for this approach to be applicable. First, training an instance segmentation model to recognise each taxonomic group would require hundreds to thousands of ground truth (i.e. human-produced) masks of all taxa. Given the long tailed distribution of taxa concentrations in the planktonic world, with many rare taxa, in particular the largest ones, this would require a considerable amount of searching and labelling effort. Indeed, assembling enough examples to train classifications models is already challenging [192] and manual delineation of each organism is much more time consuming than manual classification. A second obstacle is the size range of organisms imaged by ISIIS. Although Detectron2 does produce multi-scale feature maps through a Feature Pyramid Network in order to apply receptive fields of multiple size, the ratio between the largest and the smallest feature maps is only 16. Here, the ratio between the smallest and largest bounding box diagonals of manually segmented organisms is 65 and can reach > 180 in more exhaustive ISIIS datasets. To tackle this span, one could theoretically set up an ensemble of detectors, fed with crops of different sizes, each one targeting a restricted size range. Yet, this would be a particularly computationally demanding and complex set up, for a gain yet to be determined since, for larger sizes, the proportion of non-plankton objects, and therefore the advantage of a CNN-based segmentation, diminishes. Finally, masks generated by instance segmentation models currently lack both precision (their outline is smoothed, not matching the fine appendages of plankton) and reproducibility (because of the randomness included during training to avoid overfitting, two models trained on the same data will output different masks). These drawbacks are particularly critical for plankton

Full instance segmentation of ISIIS images is not for today.

application, where the size of the organisms, computed from their masks, is often of interest.

2.5 Conclusion and perspectives

We developed combined segmentation pipelines able to detect planktonic organisms spanning a broad size range. The fact that all methods comprised a deterministic, threshold-based segmentation ensured that particle shapes and measurement were consistent over the whole size range. Still, the segmentation method affected the shape of the size spectrum and additional processing steps (including classification) are needed to extract the correct size structure of living organisms. The MSER method limited over-segmentation of background noise objects and extracted more consistent segments, at a very high processing rate. This speed opens the possibility for near-real time processing, which is particularly relevant for adaptive sampling during a cruise or an early warning system in a time series context. Although at the lower limit of the detection capabilities of CNNs, our content-aware approach was able to detect planktonic organisms among an overwhelming number of marine snow and noise images, exhibiting the best recall of the three methods. Therefore, the ideal segmentation approach depends on the study objectives and operational constraints.

The T-CNN segmentation pipeline successfully detected plankton...

These approaches seem relevant for imaging studies focused on living planktonic organisms, since they reduce the number of objects from non-plankton classes that are extracted. In turn, this dampens the imbalance towards these classes, laying the foundations for easier, faster, and more accurate subsequent object classification by (i) reducing the amount of work needed to generate a training set with similar class distribution, which is essential to avoid the caveat of dataset shift [280]; (ii) decreasing the computation time because there are fewer objects; and (iii) limiting the contamination of the rare planktonic classes by the dominant, non-plankton, ones.

...and paved the way towards easier object classification.

Although CNN-based object detection may seem overwhelming at first, both in terms of set up and processing time, it actually is fast enough and within the reach of marine ecologists, particularly now that artificial intelligence frameworks and GPU computing are being made more accessible. This work constitutes a step towards the "intelligent" segmentation of ecological images, even at low resolution, which could find even wider applications such as the automated separation of

CNN-based detection is a convenient tool with many potential applications in plankton research...

objects overlapping onto each other on an image for more accurate species counts, the detection and classification in a single step for more automated surveys, or the extraction of individual-level traits to track e.g., reproductive organs development, for a richer exploitation of ecological images [299]. Such tasks are in no way limited to plankton images and are common in data collected by trawl cameras, benthic observations or surveying cameras, vessel monitoring cameras, etc.

...in an era of data abundance.

In this era of data-driven oceanography, the volume of data collected is increasing sharply, thanks to technological advances such as high frequency imagery, autonomous instruments (e.g. floats, gliders), satellite-based methods as well as environmental -omics approaches permitted by high throughput sequencing. In this context of abundant data, the development of automated and efficient data processing techniques becomes a key element in drawing a holistic understanding of oceanic ecosystems; it is needed to provide an extensive description of biodiversity, including species distributions as well as estimates of biomass and abundance).

Author contributions

J-OI and TP conceptualised the study. LC-C and TP generated and taxonomically sorted ground truth plankton segments. BW and TP developed the CNN segmentation method. J-OI and TP developed the threshold-based and the T-CNN processing pipelines. MS, DD, ST, CS, and RC set up and ran the T-MSER method. TP prepared the original draft. All co-authors proof-read the manuscript prior to submission. All authors read and approved the final manuscript.

Acknowledgements

The authors would like to thank the officers and crew of the R/V Tethys 2 who made the VISUFRONT campaign a success, as well as the additional scientists who took part in the cruise: R Faillettaz, C M Guigand, F Lombard, M Lilley, and J Luo.

Funding

This study is part of project "World Wide Web of Plankton Image Curation", funded by the Belmont Forum through the Agence Nationale de la Recherche ANR-18-BELM-0003-01 and the National Science Foundation (NSF) #ICER1927710. Funding also came from NSF #OCE2125407. The Extreme Science and Engineering Discovery Environment (XSEDE) provided computing resources to the US team through grant #TG-OCE170012. Data acquisition during the VISUFRONT cruise was funded by the Partner University Fund and supported by the French Oceanographic Fleet through ship time. TP's doctoral fellowship was granted by the French Ministry of Higher Education, Research and Innovation (#3500/2019).

Supplementary materials

Table S2.1: Effect of segmentation method on NASS slopes assessed through an ANCOVA.

Effect	DF effect	DF error	F-value	p-value
segmentation	3	105	133.075	$< 10^6$

Table S2.2: Pairwise comparisons between segmentation methods using the estimated marginal means. p-values were adjusted using the Bonferroni method.

Group 1	Group 2	DF	Statistic	Adjusted p-value
Ground truth	T-MSER	105	-8.289040	$< 10^6$
Ground truth	T	105	-18.752253	$< 10^6$
Ground truth	T-CNN	105	-14.670739	$< 10^6$
T-MSER	T	105	-10.094835	$< 10^6$
T-MSER	T-CNN	105	-6.093500	$< 10^6$
T	T-CNN	105	4.081515	$< 10^6$

3

Benchmark of image classification using several large plankton datasets: Convolutional Neural Networks improve detection of rare taxa

In this second methodological chapter, we apply and compare two classification methods for plankton images. This chapter does not only focus on ISIIS data but performs classification comparison for six widely used and commercially available *in situ* plankton imagers. Overall, we highlight the superiority of deep models in the detection of rare plankton classes, and demonstrate the importance to go beyond the often solely used global accuracy when performing classification on unbalanced datasets. In accordance to the results presented here, all the objects detected in the VISUFRONT data by the segmentation pipeline were then classified using a CNN model. This was made possible by the production of three datasets of annotated objects: a training dataset of ~120,000 objects enriched in planktonic objects (40% of plankton vs. 60% of detritus); a validation dataset of ~1 million objects with realistic proportions and a test set of ~1.3 million objects with realistic proportions too. As described in the following, the model was fitted on the training dataset and results were regularly checked on the validation dataset to limit overfitting. Finally, the model was assessed on the never-seen-before test set. To further improve the precision of the classifier, objects below a 90% probability threshold (established on the validation set) were ignored [124]. After ignoring non planktonic classes, this resulted in a fully annotated dataset of ~50 million planktonic organisms, with 86% precision and 89% recall averaged across weighted plankton classes.

Here we describe image classification methods...

...which were later applied to the VISUFRONT dataset...

...in order to produce fully annotated data ready for scientific analysis.

Thelma Panaïotis, Guillaume Boniface-Chang, Gabriel Dulac-Arnold, Benjamin Woodward and Jean-Olivier Irisson

Manuscript in preparation to be submitted to **Limnology & Oceanography: Methods**

Abstract

Plankton imaging instruments generate an ever increasing volume of data which is mostly processed through machine learning algorithms. However, classifying plankton images is a challenging computer science task in its own right: datasets are strongly unbalanced; the dominant classes are often not biologically interesting (artefacts, bubbles) and/or very heterogeneous looking (marine snow); and images span a large size range. Despite a wealth of reports on the performance of automatic plankton images classifiers, we still do not have a definitive idea regarding how methods compare with each other and where they can systematically be trusted. This is mostly because those reports rely on rather small unpublished datasets, not necessarily representative of real-life biological samples in terms of size, number of classes and proportions. Here we report the performance of a classic classification method (Random Forest on handcrafted image features) and a more recent one (a Convolutional Neural Network) on large publicly released datasets, from six widely used plankton imaging instruments. We show that using a CNN improves classification performance but only noticeably on poorly represented (a few hundred images) classes. Finally, we showcase the difference between the predictions of the two classifiers and a human-checked truth on several real-world datasets, to give insights regarding which ecological questions can or cannot be studied from computer-generated classifications only.

Résumé

Les instruments d'imagerie du plancton génèrent un volume toujours croissant de données qui sont pour la plupart traitées par des algorithmes d'apprentissage automatique. Cependant, la classification des images de plancton est une tâche informatique difficile en soi : les jeux

de données sont fortement déséquilibrés ; les classes dominantes sont souvent sans intérêt biologique (artefacts, bulles) et/ou d'aspect très hétérogène (neige marine) ; et les images couvrent une large gamme de tailles. Malgré de nombreux rapports sur les performances des classifieurs automatiques d'images de plancton, il reste difficile de savoir comment les méthodes se comparent entre elles et pour quelles tâches on peut s'y fier. Ceci est principalement dû au fait que ces rapports s'appuient sur des jeux de données non publiés et souvent petits, qui ne sont pas nécessairement représentatifs d'échantillons biologiques réels en termes de taille, de nombre de classes et de proportions. Nous présentons ici les performances d'une méthode de classification classique (Random Forest sur des propriétés extraites manuellement des images) et d'une méthode plus récente (un réseau de neurones à convolutions) sur de grands jeux de données ayant vocation à être publiés, provenant de six instruments d'imagerie du plancton largement utilisés. Nous montrons que l'utilisation d'un réseau de neurones à convolutions améliore les performances de classification, mais seulement de façon notable sur les classes peu abondantes (quelques centaines d'images). Enfin, nous montrons la différence entre les prédictions des deux classifieurs et une validation manuelle par un expert taxonomiste sur plusieurs ensembles de données du monde réel, afin de donner un aperçu des questions écologiques qui peuvent ou ne peuvent pas être étudiées à partir de classifications automatiques uniquement.

3.1 Introduction

Plankton is defined as the ensemble of organisms unable to swim against current. This definition, based on motility and ecological niche rather than phylogeny, encompasses mary taxonomic groups [397]. Furthermore, within those groups, plankton is known to be particularly diverse [186]. Thus, planktonic organisms cover a broad spectrum of size (from a few micrometers to several meters), shape, opacity, colour, etc. While a few planktonic groups are ubiquitous (e.g. copepods), many others are sparsely distributed and rare even when they are present [361]. Planktonic organisms are key elements of the oceanic system: they are the basis of oceanic food webs [125, 413], they contribute to the sequestration of organic carbon into ocean depths [246], and are responsible for half of the primary production of the biosphere [131].

Plankton plays key ecological roles.

Their diversity and ecological importance have made them the focus of scientific research for centuries [314].

Historically, plankton diversity was studied by sampling with nets and pumps followed by identification and counting by taxonomists. This very accurate but time-consuming method is now getting complemented with quantitative imaging and automated identification. Many plankton imaging instruments have been developed and now generate quantitative plankton observations [243]. Some of these instruments proceed by imaging collected samples, for example the ZooScan [149], the FlowCAM [365] or the ZooCAM [79]. Others acquire images *in situ*, for example the Underwater Vision Profiler (UVP) [317, 318], the *In Situ* Ichthyoplankton Imaging System (ISIIS) [85], the Imaging FlowCytobot (IFCB) [298] or the ZooGlider [293]. The increasing number and ease of use of instruments generates an increasing volume of plankton imaging data. This data is mostly processed through machine learning algorithms. But, often, the software pipelines did not progress as fast as the hardware, resulting in a data processing bottleneck [253].

Plankton data collection used to be time consuming but now integrates automation.

3.1.1 Plankton image classification

Typically, the automatic classification of plankton images proceeds by training machine learning models on handcrafted features extracted from the images and representative of the morphology of objects to classify (e.g. size, texture, grey levels). The classification algorithms include Support Vector Machines (SVM) [185, 250, 372], Random Forests (RF) [149] or Multi-Layer Perceptrons (MLP) [89]. A few studies performed comparisons of classifier performances [43, 117, 118, 149, 160], with varying results depending on the dataset, but in the end performance was comparable. This suggests that classifiers' performance is not only driven by the classifier but rather by the features (number, diversity...) that are fed to the classifier: models perform better when they are trained from a richer set of features [43].

Automatic identification plankton images initially involved models learning from handcrafted features...

Among these models, RFs are an ensemble learning method suitable for both classification and regression tasks. Their good performance, flexibility and ease of use contributed to their popularity [169]. RFs are based on decision trees averaging, but with the specificities of bootstrapping on the training data and random subsetting of features to compute each tree node, resulting in more robust models while reducing overfitting [57]. According to Fernández-Delgado et al. [129]

who evaluated the performances of nearly 180 classifiers on various datasets, RFs perform better than other classifiers. This conclusion was previously reached by Gorsky et al. [149], resulting in a wide use of RF classifiers to sort ZooScan data. Later on, the IFCB data processing software switched from SVM to RF [13]. Finally, EcoTaxa [316], a web application dedicated to the taxonomic annotation of images, initially implemented a RF classifier to generate classification predictions for unlabelled images.

. . . of which the random forest is a special case of decision trees.

However, since 2015, an increasing proportion of plankton image classification studies are using deep learning approaches, in particular Convolutional Neural Networks (CNN). CNNs are a specific type of artificial neural network, mainly used in pattern recognition (image classification, image segmentation, language processing, etc.). Their architecture is inspired from the animal visual cortex: each neuron responds to stimuli from a restricted region. In the case of an image classification task, a CNN directly takes an image as input, transforms it in a way similar to what the visual cortex would do, and outputs a label (i.e. a class name) for that image. Compared to the two steps approach above, here the model simultaneously extracts features from the image and uses them to classify it. Its training therefore results in both a classifier adapted to the classes in the training set and features that are optimised for this classifier to perform well.

CNNs are a different class of models. . .

The first application of a CNN to image recognition dates back to 1990 [223], but their usage strongly intensified after the first and significant success of a CNN at the 2012 ImageNet Large Scale Visual Recognition Challenge [213, 343]. This success was made possible by the availability of several million annotated images (ImageNet) [101] and the progress in computational power, especially the use of Graphic Processing Units (GPU) [73]. CNNs have now become the state of the art method for image classification [222].

. . . of which application to image identification is quite recent. . .

Their application to the classification of plankton images stems from a plankton images classification challenge hosted by the online platform Kaggle in 2015 (`https://www.kaggle.com/c/datasciencebowl/`). Since then, multiple works highlighted the success of CNNs for plankton images recognition [77, 93, 118, 124, 224, 247, 249, 354]. CNNs were shown to outperform the classic approach of feature extraction followed by classification with both RF [192, 300, 301] and SVM [118], on multiple plankton images datasets. Currently, CNNs dominate the new literature on plankton image classification [192]. However, these

. . . and even more in the case of plankton identification.

studies only compared coarse metrics such as accuracy or mean scores (precision, recall, F1-score), and focused on a single classification task. As we still lack systematic comparison regarding the application of CNNs to plankton image classification tasks [118, 192], we aim to fill this knowledge gap with an in-depth comparison on tasks of various difficulties, providing more detailed metrics in order to understand where the differences in performance lie between CNNs and RFs.

Identifying plankton from images is difficult for various reasons.

Whether with traditional or deep approaches, classifying plankton images is a challenging computer science task. First, plankton datasets are often strongly unbalanced, with a few dominating classes and many poorly represented ones [224, 254, 357]. This is actually a characteristic of planktonic communities: as mentioned above, some taxa are ubiquitous while others are scarcely found [361]. This characteristic contrasts with common benchmark datasets where classes are more evenly distributed: between 732 and 1300 images for each of the 1000 classes in ImageNet [343]. It creates a problem for rare taxa because classification performance decreases with the number of examples per class [407]. And indeed, rare planktonic taxa are often poorly predicted [249, 357]. Second, as explained above, planktonic organisms come from a wide range of taxa and constitute a morphologically heterogeneous group, of various sizes, shapes and opacities. This can result in a non negligible intragroup morphological variability [160], susceptible to induce confusion between groups, complicating plankton image classification. Finally, real-world plankton images datasets contain a large proportion of non-living objects such as marine snow aggregates or bubbles [30]; these classes often constitute the majority of the datasets [118, 192, 357]. Moreover, plankton images are typically small and often greyscale, and thus not very rich in terms of information to extract.

Assessing the evolution of performance at plankton images classification is difficult.

So far, over 175 papers addressed the topic of automated plankton image identification [192]. As shown above, a few explicitly compared models, with sometimes diverging results. But, overall, these 100+ studies used many different datasets, most of them not released publicly, of various compositions in terms of classes and number of images, while both strongly affect performance. They also reported different performance metrics and the one most commonly reported (accuracy) is flawed in the case of unbalanced datasets [192]. From a broad perspective, classification performance seems to have stayed quite stable over time, while the numbers of taxa to sort, hence the difficulty of the task, increased [192]. This would indicate that classifiers did im-

Table 3.1: Common plankton images benchmark datasets.

Name	References	Imaging instrument	Composition		Relevant publications
			Images	Classes	
WHOI-plankton	[301, 373]	IFCB	3.5 M	103	[87, 93, 224, 300]
ZooScanNet	[115]	ZooScan	1.4 M	93	[254, 357]
PlanktonSet 1.0	[86]	ISIIS	30,336	121	[107, 327, 338, 405]

prove. But the two reasons stated above make it impossible to quantify it. Nonetheless, three benchmark datasets were published and used in several studies (Table 3.1), while a few other studies focused on smaller versions of these datasets [93, 247, 429], so the move towards standardisation and intercompatibility is ongoing.

The purpose of this work is to report the performance of a classic approach, using handcrafted features and a RF classifier, and of an easy-to-train CNN on large, publicly released, datasets from six commonly used plankton imaging instruments. This study compares the two classifiers, to objectively discuss the merits of CNNs compared to the traditional approach. It also provides baseline results for the development of future plankton images classifiers.

We propose a standardized comparison of two classifiers on large, public, reference datasets.

3.2 Material and methods

3.2.1 Datasets

3.2.1.1 *Imaging tools*

Among the six widely used plankton imaging instruments from which we draw datasets, three are deployed *in situ* while the three others image plankton *ex situ*. The ISIIS [85] is a ship-towed system that undulates between the surface and a specified depth. It uses transmitted light, which allows a long depth of field and is particularly suitable for the imaging of small and transparent organisms. The size of targeted organisms ranges from less than 1 mm to several cm. A tow at 4 kts with a scanning rate of 28 kHz allows a high sampling rate of > 100 L s^{-1} [124]. The UVP6 [318] can be deployed on CTD-Rosette systems, long-term moorings or autonomous underwater vehicles, such as floats and gliders. It targets organisms between 620 µm and a few cm. The IFCB [298] is a flow imaging instrument, targeting phytoplank-

The datasets come from six commonly used plankton imaging instruments.

ton between 10 and 100 µm. It continuously operates underwater and can be deployed for months, making it adapted to long-term surveys. Image capture is usually triggered by the detection of chlorophyll fluorescence so that dead particles are not imaged. The ZooScan [149] allows operators to scan preserved plankton samples in the lab. It targets organisms larger than 200 µm. The FlowCAM [365] is a flow imaging instrument that can be used in the lab or on a ship. It targets phytoplankton and microzooplankton ranging from 20 to 200 µm. A pump flows the sample at a rate of 20 mL min^{-1} to the flow chamber where organisms are imaged one by one using transmitted light. Similarly to the FlowCAM, the ZooCAM is a flow imaging instrument but it targets larger organisms, mostly zooplankton and fish eggs larger than 300 µm. A pump drives the sample, complemented with filtered seawater, to a flow cell where objects are imaged. The ZooCAM also uses transmitted light but has a higher flow rate than the FlowCAM: from 0.28 to 1.7 L min^{-1} [79].

3.2.1.2 *Image processing*

Each imaging tool had its specific image processing and feature extraction pipeline. ISIIS data was processed with Apeep [306] and features were extracted using Scikit-image [411]. The IFCB data processing relied on multiple MatLab scripts [372] to extract various feature types. The UVPapp application [318] was developed to process UVP6 images and to extract features. Both ZooScan and FlowCAM data were processed with Zooprocess [149] that generated crops of individual objects together with a set of features. ZooCam data processing was very similar to ZooScan and FlowCAM processing [79]. Thus, for all datasets, each greyscale image was associated with a set of handcrafted features (which depended on the instrument and its processing pipeline) computed from the image, and a label.

Each dataset had its own features set.

3.2.1.3 *Datasets assembling and composition*

All datasets were created in a similar way: real-world samples were sorted by human operators; classifications were checked by one operator for each dataset; full samples particularly rich in some rare classes were added (except for IFCB and ZooCAM); classes still containing fewer than ~100 objects were merged to a taxonomically and/or morphologically neighbouring class. When no relevant merging class could

Table 3.2: Datasets composition in terms of the numbers of images, classes and handcrafted features as well as the proportion of plankton.

Instrument	Composition			
	# images [min; max per class]	Classes	Features	% plankton
FlowCAM	301,247 [74 ; 69,085]	93	47	36.2
ISIIS	408,166 [70 ; 321,335]	32	31	15.3
UVP6	634,459 [87 ; 508,817]	54	62	7.7
ZooCAM	1,286,590 [81 ; 204,132]	93	48	67.8
ZooScan	1,451,745 [90 ; 241,731]	120	48	71.2
IFCB	1,592,196 [90 ; 1,177,499]	69	72	12.6

Datasets are representative of the real-world.

be found, objects were assigned to a miscellaneous class, along with other objects in the same situation or impossible to classify. Thus, every single object was included in the classification task, ensuring that the metrics computed on those datasets are as relevant to a real-world situation as possible. While the homogeneous processing and cross-checking does not guarantee the absence of mistakes, it should still result in very consistent classes. The IFCB images were sourced from Sosik, Peacock, and Brownlee [371] (years 2011-2014); the images for other instruments were sourced from EcoTaxa [316], with the permission of their owner. The number of images in the resulting datasets ranged from 301,247 to 1,592,196, in 32 to 120 classes (Table 3.2). As expected, the datasets collected *in situ* (ISIIS, UVP6, and IFCB) were particularly rich in marine snow and other non-living objects, resulting in a low proportion of plankton.

Modification of datasets for machine learning.

To assess performance at a coarser taxonomic level, which could be sufficient in some applications scenari and is more comparable to most older papers, each class was assigned to a larger ecological group (Table S3.2). Then, each class/group was categorised as plankton or not-plankton, which allows to compute metrics for planktonic organisms only, without the, sometimes dominant, non-living objects (Table 3.2). Datasets were split into 70% for training, 15% for validation and 15% for testing. These splits were stratified by class to guarantee a good representation of each class in all datasets and they remained identical for all experiments.

3.2.2 Classification models

Both models were trained and evaluated following the same procedure.

Each dataset was classified with a classical approach (Random Forest working on the handcrafted features) and with a modern Convolutional Neural Network. The training procedure was similar for both and was replicated exactly for each dataset: (i) models were fitted on the training split, according to a loss metric, (ii) various sets of hyperparameters were assessed based on that same loss metric but computed on the independent validation split to limit overfitting, (iii) the model with optimal hyperparameters was used to predict the never-seen-before test split, once, and various performance metrics were computed.

Implementation technical details.

The RF classifier was implemented with Scikit-learn [310]. The CNN model was implemented with Tensorflow [1]. Training and evaluation were performed on a Linux machine, running Ubuntu Linux 20.04 and Python 3.8.10, sporting two 18 cores Intel Xeon Gold 6240 (72 logical cores) and a Quadro RTX 8000 GPU. The code to reproduce all results is available at `https://github.com/ThelmaPana/plankton_classif_benchmark`.

3.2.2.1 *Random Forest*

The RF learns from handcrafted features.

The RF classifier was trained on the handcrafted features extracted from images by the software dedicated to each instrument. Their number ranged from 31 to 72 depending on that software (Table 3.2). Several types of features exist: most features are global features computed on the whole image; morphological features are computed on the object silhouette, and texture features are computed as co-occurrence matrices on grey levels. The diversity of these features is determinant for classifier performance [43].

The RF has its specificities (hyperparameters, criterion).

The loss metric used during training and validation was the categorical cross-entropy, which does not optimise accuracy directly but rather the quality of the probability to be in the correct class, output by the classifier. Hyperparameters for the RF classifier were evaluated using a grid search procedure over specified values for: (i) number of trees (100, 200, 350, 500), (ii) number of features to use to compute each split (the default for classification is the square root of the number of features; here 4, 6, 8, 10 were tested) and (iii) minimum number of samples required to attempt the split of a node (the default for classification is 5; here 2, 5, 10 were tested) [169]. For each combination of values (48 in total), the RF model was fitted on the training split and evaluated

on the validation split. The model with the lowest validation loss was chosen as the best.

3.2.2.2 *Convolutional Neural Network*

The CNN learns directly from images but they have to be prepared.

Because our goal here is to assess the performance of an easy-to-use CNN, that most research teams should be able to deploy, we used transfer learning from a rather small model, pre-trained on ImageNet and then fine-tuned on each dataset.

The feature extractor portion of the CNN is from a MobileNetV2[1] [347]. The depth multiplier is 1.4, the input size is 224×224×3 and the output is a 1792-elements vector. Therefore, images were reformatted to match the input size, while preserving their aspect ratio: each image was resized so that its longest side was 224 pixels, then padded to 224×224 pixels using the median value of border pixels (to keep the background homogeneous), finally the greyscale channel was replicated to create three identical channels and reach the shape of 224×224×3 pixels. As training a CNN from scratch is usually a power and time-consuming process and requires a large volume of training data, we applied transfer learning by using a feature extractor pre-trained on the ImageNet dataset. The pre-trained feature extractor can be used as it is, as features extracted by a model trained on generic datasets are also relevant for other tasks [422], such as plankton classification [300, 338]; but can also be fine-tuned for better performance [422].

The CNN consisted of a feature extractor...

After the feature extractor, we added a dropout layer (rate = 0.5), a fully connected layer of size 680 and a classification head whose size depends on the number of classes to predict, for a total of ~5.6 million parameters. A few initial tests showed that increasing the size of the fully connected layer, or using two layers, did not improve performance while it complexified the model. The drop-out layer ensured that the model did not rely on a few key neurons only, thus reducing overfitting [378].

... followed by fully connected layers.

To improve the generalisation ability of the model and the performance, especially for rare classes, images from the training set were augmented with random vertical and horizontal flip, zoom in and out (by 20% maximum) and shearing (by 15° maximum). Images were not rotated as objects from a few classes had specific orientation (e.g. vertical-lines in the ISIIS dataset or some organisms that display a given orientation in datasets collected *in situ*). Like for the RF, the loss metric

1 https://tfhub.dev/google/imagenet/mobilenet_v2_140_224/feature_vector/4

was the categorical cross entropy. At the end of each training epoch (i.e. a complete run over all images in the training split), both loss and accuracy were computed on the validation split, to check for overfitting, and model parameters were saved. The feature extractor, fully connected and classification layers were trained for 35 to 60 epochs, depending on the dataset. This was shown *a posteriori* to be enough for the training to be exhaustive: for all experiments, the validation loss did not improve for at least the last 9 epochs. The optimiser used the Adam algorithm, with a decaying learning rate from an initial value of 0.001 and a decay rate of 0.35 per epoch. Given that training each CNN took several hours, an extensive hyperparameter search was not performed, but the number of training epochs, at least, was optimised through early stopping to reduce overfitting [368]: the parameters of the model at the epoch with minimum validation loss were selected as the final model.

CNN training specifics.

3.2.2.3 *Class weights*

In an imbalanced dataset, more importance is given to well represented classes, because examples from those classes come more often in the computation of the loss, while very small classes are almost negligible. As a result, the performance on those small classes is often very bad. An efficient workaround is to rebalance the importance of all classes by weighting the loss function in such a way that misclassification of small classes has a higher cost than that of common ones, thus forcing the model to perform better on small classes.

Class weights were used to overcome class imbalance.

A common for of weighting is by "inverse frequency" : the weight of class i is computed as

$$W_i = \frac{\max(c)}{c_i}$$

where $\max(c)$ is the number of objects in the largest class and c_i the number of objects in class i. However, this method can lead to very heavy weights for very small classes, in highly imbalanced datasets, with potential negative side effects [88]. Thus we chose a smoother distribution of weights and used the square root of the inverse frequency, which gives, for class i:

Inverse frequency weighting was too strong.

$$W_i = \sqrt{\frac{\max(c)}{c_i}}$$

We trained both non weighted and weighted CNN and RF models on the same datasets to explore the effects of weighted training.

3.2.2.4 *Model evaluation*

Classic metrics were computed to assess model performance...

After training and choosing the best model for each approach, each dataset, with and without class weight, each resulting model was evaluated on the test split, which had never been used beforehand, and the following usual metrics were computed: accuracy score (percentage of objects correctly classified), class-wise precision (percentage correct in the predicted class) and recall (percentage correct within the true class).

...but accuracy on unbalanced datasets is misleading.

However, in datasets with a strong class imbalance such as many plankton datasets, the accuracy value can be misleading. For example, in a strongly biased dataset composed of 99% of objects in the same class, a dummy model classifying every object in this class is completely useless but still has 99% accuracy. In a dataset with three classes with proportions 98%, 1% and 1%, a classifier assigning classes at random but following those proportions still has an accuracy of ~96%. Therefore, to more honestly gauge the quality of our models on imbalanced datasets, performance metrics were also computed on the output of such a random classifier. In addition, the balanced accuracy value, computed as the macro-average of per-class recall scores, was also computed since it is a better estimate of model performance in such a scenario [200].

Wwe also computed plankton-focused metrics.

Moreover, in the case of plankton datasets dominated by non planktonic classes (e.g. detritus), the accuracy value is mostly driven by this class and, therefore, does not provide information on the performance on plankton classes, which are often the topic of study. To focus on those classes, we also computed the average of per-class precisions and recalls, weighted by the number of objects in the class, and only using plankton classes. Averaged plankton recall provides a direct indication of the proportion of planktonic organisms that were correctly predicted. Averaged plankton precision gives an indication of how "pure" the predicted plankton classes are.

3.3 Results

3.3.1 Hyperparameter choices and training time

RF was faster to train than CNN.

Training and evaluation time was always shorter for RFs than for CNNs. When running on 12 CPU cores, the RF on the smallest dataset (ISIIS, ~400,000 objects) took less than one hour for gridsearch, training and evaluation, while it took a few hours on the IFCB dataset (~1.6 M objects). Regarding the CNN, it took 16 h to train the model for 40 epochs on the ISIIS dataset but 78 h for the same number of epochs on the IFCB dataset, using a Quadro RTX 8000 GPU. All CNN models were trained long enough so that the best epoch according to the validation split was reached at least 9 epochs before the end of training. The gridsearch performed to choose RF hyperparameters selected models with a high number of trees (500, Table S3.1), numerous features considered for each split (10 in most cases) but rather shallow trees as splits could only be performed in nodes containing a certain number of objects (10 in most cases).

3.3.2 Classification performance

The performance metrics of all models on all datasets are presented in Figure 3.1.

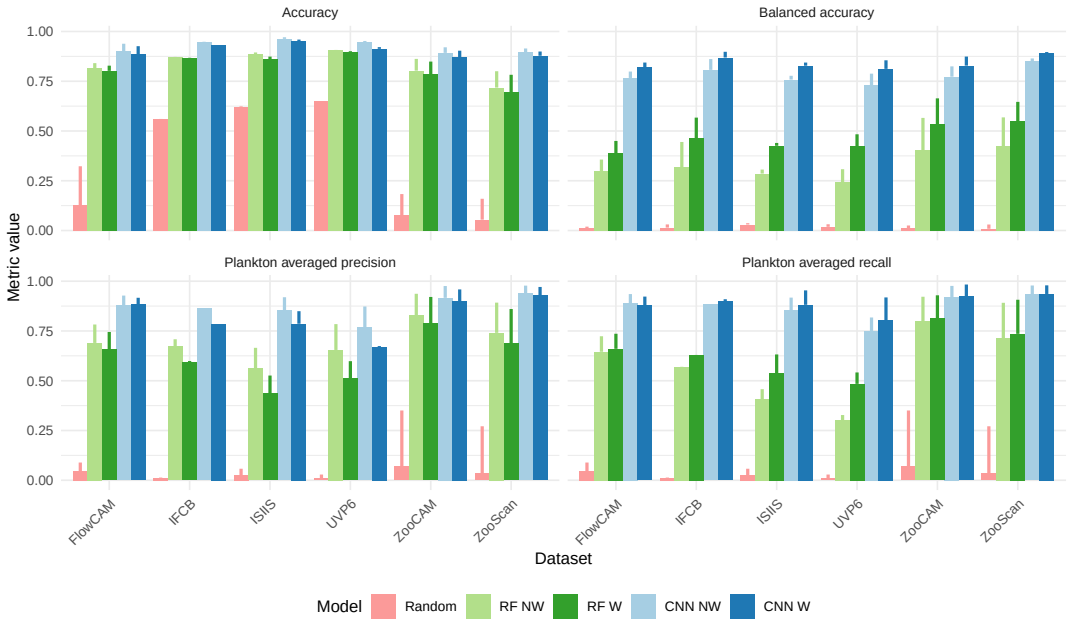


Figure 3.1: Metrics computed on the test split of each dataset for each model (random, RF, CNN; NW = non weighted, W = weighted). Thick bars show the value of each metric on the finest taxonomic level, thin bars show the value after regrouping objects in broader ecological classes.

Table 3.3: Classification report for coarse classes in the ZooScan dataset. n = number of objects per class in the test set, NW = non weighted, W = weighted, diff = performance difference between CNN and RF.

		Precision						Recall						F1					
		NW			W			NW			W			NW			W		
Class	n (test)	RF	CNN	diff	RF	CNN	diff	RF	CNN	diff	RF	CNN	diff	RF	CNN	diff	RF	CNN	diff
								Plankton											
Actinopterygii	978	74	92	19	63	89	26	60	84	24	67	89	21	66	88	22	65	89	24
Alveolata	1033	81	95	13	69	90	21	78	91	13	88	95	7	80	93	13	77	92	15
Amphipoda	492	63	96	33	44	96	52	28	89	61	40	92	52	39	93	53	41	94	52
Annelida	631	78	96	17	53	91	38	31	90	59	42	93	51	45	93	48	47	92	45
Appendicularia	8223	82	96	15	72	96	24	77	94	17	83	94	11	79	95	16	77	95	18
Chaetognatha	8604	92	99	7	92	99	8	95	99	4	95	99	4	94	99	6	93	99	6
Cirripedia	856	0	78	78	18	82	64	0	74	74	7	81	74	0	76	76	10	81	72
Cladocera	8887	75	96	20	58	92	35	65	94	29	73	97	25	70	95	25	64	95	31
Copepoda	74137	78	92	14	60	94	34	63	94	31	71	92	20	70	93	23	65	93	27
Ctenophora	137	85	95	11	56	88	31	23	92	69	41	96	55	37	94	57	47	92	44
cyphonaute	1334	75	94	18	77	96	19	78	94	16	75	91	16	77	94	17	76	93	18
Decapoda	3716	63	83	21	37	55	18	4	70	66	34	84	50	8	76	68	35	66	31
Doliolida	1461	90	98	8	71	98	26	43	96	53	63	99	36	59	97	38	67	98	32
Echinodermata	979	81	96	16	70	93	24	74	96	22	81	97	16	77	96	19	75	95	20
Eumalacostraca	3812	70	92	23	52	95	43	83	96	13	89	95	5	76	94	19	66	95	29
Harosa	244	77	94	17	66	89	23	69	91	22	75	94	19	73	92	20	70	92	22
Isopoda	83	91	97	7	84	97	13	88	97	9	90	97	6	89	97	8	87	97	10
Mollusca	6158	77	89	13	61	86	25	55	89	35	65	91	26	64	89	26	63	88	25
other_Cnidaria	1052	75	75	0	72	64	-8	88	84	-4	92	93	1	81	79	-2	81	76	-5
other_Crustacea	6566	69	64	-6	62	56	-6	37	59	22	40	64	24	49	61	13	49	60	11
Pyrosomatida	75	60	81	21	75	90	14	77	84	6	50	67	16	68	83	15	60	77	16
Rhizaria	8575	76	80	3	71	66	-4	74	66	-8	80	81	1	75	72	-3	75	73	-2

		Precision						Recall						F1					
		NW			W			NW			W			NW			W		
Class	n (test)	RF	CNN	diff	RF	CNN	diff	RF	CNN	diff	RF	CNN	diff	RF	CNN	diff	RF	CNN	diff
Salpida	2884	73	82	9	44	71	28	19	76	57	37	82	45	30	79	49	40	76	36
Siphonophorae	12232	85	87	2	84	88	3	77	94	16	75	91	16	81	90	9	80	89	10
Stramenopiles	1075	25	68	43	29	57	29	3	38	35	15	68	53	5	48	44	20	62	42
Trachylina	937	65	67	2	42	59	16	30	68	38	55	75	20	41	68	26	48	66	18
Average		***72***	***88***	***16***	***61***	***84***	***23***	***55***	***85***	***30***	***62***	***88***	***26***	***59***	***86***	***27***	***61***	***86***	***25***
								Non plankton											
artefact	7718	93	98	5	92	97	5	89	97	8	88	98	9	91	98	6	90	97	7
badfocus	6046	71	93	22	45	86	40	35	88	54	56	90	34	47	91	44	50	88	38
bubble	2432	77	98	21	71	98	27	86	97	11	88	96	8	81	97	16	78	97	18
detritus	36260	97	94	-3	50	81	32	20	80	60	34	90	56	33	87	54	40	85	45
fiber	6708	87	96	9	85	94	9	90	97	7	91	97	6	89	97	8	88	95	8
Insecta	169	81	93	12	68	88	20	71	95	24	79	97	18	76	94	18	73	92	19
other_egg	2015	80	87	8	63	88	25	63	94	31	73	94	21	70	91	20	68	91	23
other_living	40	87	99	12	80	99	19	82	98	16	84	98	14	84	98	14	82	98	16
seaweed	1272	69	81	12	50	74	23	34	81	47	46	85	38	45	81	36	48	79	30
Average		***76***	***87***	***12***	***63***	***82***	***19***	***58***	***85***	***27***	***66***	***88***	***23***	***62***	***86***	***24***	***64***	***85***	***21***

Overall, CNN performed slightly better than RF.

In terms of overall accuracy, CNNs performed only a bit better than RFs on all datasets (Figure 3.1). The use of class weights slightly decreased both CNNs and RFs accuracies, as it focused training on small classes, paying less attention to large classes that account for more in the computation of accuracy. Note that a random classifier achieved 56.2%, 62.3% and 65.0% of accuracy on the detritus-dominated IFCB, ISIIS and UVP6 datasets, respectively. While the accuracies of our models are all higher (94.9%, 96.3% and 94.6% for the non-weighted version of the CNN), they must be gauged in terms of the increase compared to the random model and not in absolute terms.

For classification of poorly represented classes, CNN strongly outperformed RF.

CNNs had much better balanced accuracy than RFs, with and without weights (Figure 3.1). The random classifier performed very poorly for all datasets, because this metric is indeed more sensitive to small classes, where RFs performed worse than CNNs. Class weights improved balanced accuracy for both CNNs (up to +8.2% for the UVP6 dataset) and RFs (up to +18.0% for the UVP6 dataset). Thus, as expected, weighting small classes more did enhance their learning, especially for RF models.

CNN classification resulted in cleaner plankton classes than RF.

CNNs outperformed RFs in terms of averaged plankton precision, on all datasets, regardless of weight use (Figure 3.1, Table 3.3). However, weighting small classes decreased precision on plankton classes for both models, in all datasets. Models paid less attention to large classes, resulting in a stronger pollution of plankton classes, that is: a lower precision.

CNN performed better at detecting planktonic organisms.

On the other hand, the use of class weights did improve the recall of plankton classes for all CNNs and RFs with the exception of the CNN on the FlowCam dataset. This improvement is not surprising since plankton classes, usually smaller than non plankton classes (e.g. detritus), are therefore weighted more and this reduces the number of false negatives, i.e. increases the recall. For all datasets, all CNN models – weighted or not – gave much better results than RF models (Figure 3.1, Table 3.3).

3.3.3 Performance on coarser groups

Regrouping classes into larger ecological groups increased all performance metrics on all datasets (Figure 3.1). Indeed, such regrouping made the classification task easier because there were fewer groups to classify but it could also induce more diversity per class, thus de-

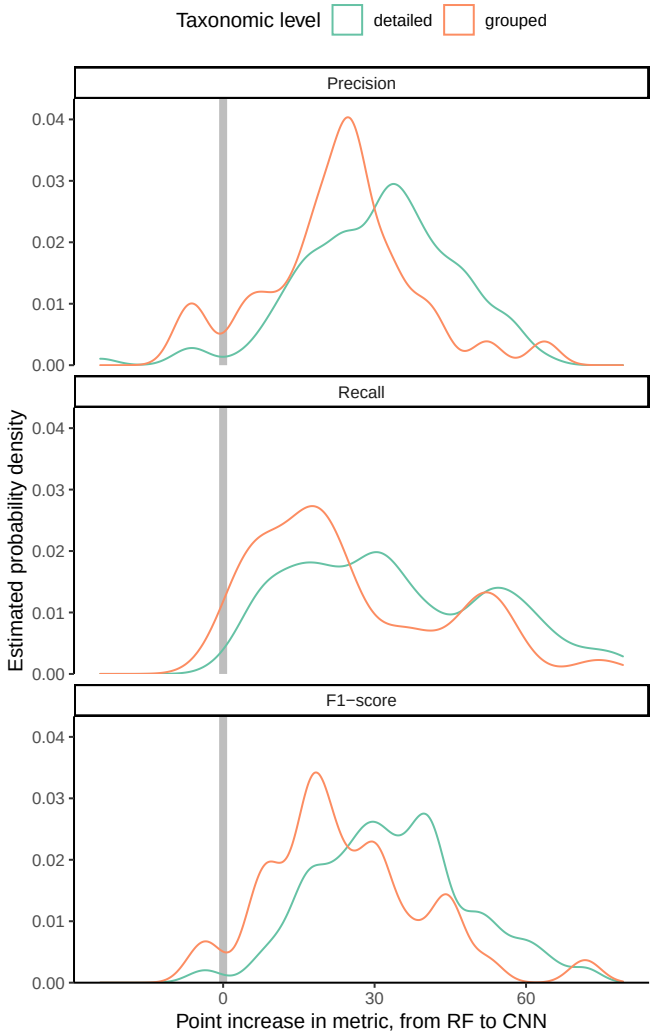


Figure 3.2: Performance increase from weighted RF to weighted CNN on the ZooScan dataset.

creasing performance of certain classes. The performance increase was stronger for RF models, particularly on the ZooCam and ZooScan datasets. This result highlights the fact that RFs make frequent confusions among the finer scale taxa within large ecological groups. A good example is Copepoda, which accounts for 22 classes in the ZooCam dataset and 20 in the ZooScan dataset, and make up 38% and 34% of images in each dataset, respectively. This is further supported by the strong increase in performance for the random classifier too, on both the ZooCam and ZooScan datasets.

Unsurprisingly, regrouping classes led to better performances.

CNN performance improvement compared to RF was stronger when the taxonomic level was more detailed (Figure 3.2), i.e. when there were more classes to predict, thus when classes were smaller and the prediction task was harder. This also suggests that CNNs are better than RFs to distinguish between rare classes.

CNN improved performance even better when the classification task was harder.

3.4 Discussion

3.4.1 Cost and benefits of using CNNs

3.4.1.1 *Main result: CNNs improve the detection of rare classes*

When looking at accuracy only, RFs performance appeared to be lower but close to CNNs performance. However, balanced accuracy and metrics on plankton classes highlighted that CNNs performed better in the classification of objects in low abundant classes, especially when class weights were used. This makes the use of CNNs particularly relevant for plankton classification tasks where datasets are often biased towards non planktonic classes.

CNNs performed better than RFs at detecting rare plankton classes.

3.4.1.2 *Considerations on computation cost*

CNNs took longer to train than RFs. However, training a RF model from images only requires first to extract features from images, which can take a non negligible time depending on the size of the dataset and the type of features. Nonetheless, the computational cost of CNNs is higher as they require a GPU to train efficiently [73]. This restricts their use to powerful enough computers and impeaches their application to onboard plankton classification tasks for now. But CNN training time heavily depends on the number of parameters of the model, here we selected a lightweight CNN model (5.5 million trainable parameters),

CNNs are more costly to train than RFs.

of which training was fast (less than 100 hours) compared to larger models [426]. As both deep-learning libraries such as Tensorflow or Pytorch become easier to use and GPU resources are becoming more available for the scientific community, these powerful tools are becoming more accessible [253].

3.4.1.3 *The model choice depends on classification goals*

The classification model should be chosen according to classification goals.

Before training a model for plankton classification, it is important to determine the goals of such a task. If one wants to maximise overall accuracy on a balanced dataset with no irrelevant classes, methods to deal with class imbalance are not relevant and should not be applied. In such a case, a classic model with shorter training time and lower computation cost might perform almost as well as a CNN. On the other hand, if the goal is to maximise the detection of rare plankton classes or to perform classification on confusing classes, a CNN will likely perform better than a classic classification model, especially in combination with class imbalance methods. Moreover, CNNs were shown to perform better than RF to predict low abundances even when training data is abundant [192].

3.4.2 Potential improvements

3.4.2.1 *CNNs do not account for size but could be added*

CNNs do not account for object size but a few workarounds exist.

While CNNs performed better than RFs, a well known drawback of CNNs is that they do not account for object size, as all input images must be resized to the same dimension. A solution to avoid every object to have the same size is to proceed to resizing for images larger than input dimension and to use padding instead of resizing for smaller images. However, this may become a risky practice in the presence of very small objects: they might just be reduced to 1 pixel after a few convolutions and all information would be lost. Another and more robust solution is to use mixed models: size information (e.g. area, feret) can be concatenated with the fully connected layers to produce a model accounting both for images and objects size properties. However, this does not necessarily provide a strong improvement in classification performance: Kerr et al. [202] report a small improvement when concatenating geometric features, while Kyathanahally et al. [214] report negligible gain. Ellen, Graff, and Ohman [118] assessed the effect of

concatenating various types of context metadata (geometric, geotemporal and hydrographic) into fully connected layers: geometric features did not improve model performance while geotemporal and hydrographic (and combination of them) did improve model performance. However, the use of such metadata for organisms classification forbids any later analysis linking these organisms with their environment as a correlation between organisms and environment is induced at the classification step. The weak to negligible improvement induced by the concatenation of size features could be explained by the fact that CNN classifiers mostly rely on local texture and shape features [16], thus they should be able to correctly predict an object regardless of its resizing.

3.4.2.2 *Changes to the datasets*

Imbalance in plankton datasets can be overcome...

Plankton datasets are often imbalanced with plankton classes being the smallest ones, while largest classes are often made of non living objects such as detritus. The datasets used in this study are no exception. Both "algorithm-level" methods and "data-level" methods exist to deal with class imbalance and avoid training being dominated by larger classes [212].

...with algorithms...

Algorithm-level methods include the use of class weights to artificially give more importance to poorly represented classes in the loss computation [88]. This is the method we implemented in our work. Another algorithm-level method is to use a loss function such as sigmoid focal cross entropy [237] which penalises more hard examples (small classes) than easy ones (large classes). We tested the implementation of focal cross entropy instead of a categorical cross entropy for our CNNs but it did not significantly improve performance.

...or at the data level.

Data-level methods include oversampling small classes and undersampling large classes, thus modifying the distribution of classes in the training set [212]. This practice can lead to bad performances when evaluating the model on an imbalanced test set because the model also learned the class distribution. Thus, when using a model trained on an idealised training set to classify objects from a new dataset, prediction may be of poor quality [145], a problem known as dataset shift [280]. Algorithm-level and data-level methods can be used concurrently to alleviate the effect of imbalanced datasets.

Another difficulty in plankton classification tasks is to generate a training set with all potential objects that can be detected with the

instrument. Indeed, the model will not be able to predict a class that has never been seen before. As a consequence, when training a classifier with a fixed list of classes, all objects are inevitably predicted in one of these classes, impeding the discovery of new types of objects [254, 357].

We need exhaustive datasets.

3.5 Conclusion and perspectives

In the end, we show that CNN models perform slightly better than RF models at the global scale. Furthermore, the use of a class-weighted CNN model remarkably improves detection of poorly represented (a few hundred images), where a class-weighted RF model fails to overcome dataset imbalance. Our results show that both RF and CNN predictions can be trusted to answer ecological questions regarding abundant plankton, as long as the model was adequately trained and tested. However, one should be careful when looking at rare plankton with smaller-scale differences in concentrations (whether in time or space): if CNN can be a good indicator, they cannot be fully trusted on poorly represented classes, where manual validation by an operator is still required [192].

CNN performed better than RF on rare classes.

Finally, our work highlights the importance of not only considering the global accuracy when assessing model performances, especially in the case of an unbalanced dataset biased towards classes outside of the main topic area; but rather to consider metrics focusing on classes of interest. The results presented here are in line with the shift towards the use of deep learning models for plankton classification tasks [189], which was made possible thanks to the advances in computational power, an easier access to dedicated hardware, the release of large enough datasets and the development of deep learning turnkey libraries such as Tensorflow [1] or Pytorch [308]. All the datasets used in this study which are not released yet will be made publicly available to facilitate benchmark of new classification methods.

Consider other metrics than global accuracy.

Author contributions

JOI and TP conceived the study; GBC and GDA developed a first CNN classifier; TP implemented the RF classifier and the final CNN classifier from GBC's and GDA's initial work, with guidance from BW; TP conducted the experiments under the supervision of JOI; TP

wrote the original draft; all authors reviewed and approved the final manuscript.

Acknowledgements

We would like to acknowledge scientists, crew members and technicians who contributed to data collection and the taxonomist experts who sorted the images to build the datasets. This work falls within the project "World Wide Web of Plankton Image Curation", funded by the Belmont Forum through the Agence Nationale de la Recherche #ANR-18-BELM-0003-01 and the National Science Foundation (NSF) #ICER1927710. TP's doctoral fellowship was granted by the French Ministry of Higher Education, Research and Innovation (#3500/2019).

Supplementary materials

Table S3.1: Selected hyperparameters by RF gridsearch for each RF training

Dataset	Class weights	Hyperparameters		
		Number of trees	Max features	Min sample split
IFCB	NW	500	10	5
	W	500	8	10
ISIIS	NW	500	10	10
	W	500	6	10
UVP6	NW	500	10	10
	W	500	6	10
ZooScan	NW	500	10	5
	W	500	10	10
FlowCAM	NW	500	10	5
	W	500	10	10
ZooCAM	NW	500	10	5
	W	500	10	10

Table S3.2: Classification report for detailed classes in the ZooScan dataset. n = number of objects per class in the test set, NW = non weighted, W = weighted, diff = performance difference between CNN and RF.

			Precision						Recall						F1					
			NW			W			NW			W			NW			W		
Larger group	Class	n	RF	CNN	diff	RF	CNN	diff	RF	CNN	diff	RF	CNN	diff	RF	CNN	diff	RF	CNN	diff
								Plankton												
Actinopterygii	Actinopterygii	289	77	96	19	60	94	34	29	90	61	42	91	49	42	93	51	49	93	43
	egg Actino.	689	63	74	11	42	71	28	21	75	53	31	75	45	32	74	43	36	73	37
Alveolata	Neoceratium	53	63	90	27	40	84	44	19	88	68	36	90	53	29	89	59	38	87	49
	Noctiluca	980	76	80	3	71	66	-4	74	66	-8	80	81	1	75	72	-3	75	73	-2
Amphipoda	Amphipoda	125	75	90	15	50	85	35	21	86	65	28	86	58	32	88	56	35	85	50
	Cumacea	78	70	86	16	53	89	35	57	92	35	69	90	21	63	89	26	60	89	29
	Hyperiidea	289	73	86	13	63	79	15	58	89	31	63	89	26	65	87	23	63	84	20
Annelida	Annelida	349	75	87	12	33	80	48	4	85	81	19	88	69	8	86	78	24	84	60
	larvae Annel.	50	59	97	38	55	97	42	56	95	39	60	95	36	58	96	39	57	96	39
	part Annel.	149	82	100	18	58	97	39	80	95	15	88	98	10	81	97	16	70	98	28
	Tomopteridae	83	65	88	23	41	72	31	17	68	51	29	82	53	27	77	50	34	77	43
Appendicularia	Fritillariidae	1820	81	98	17	80	97	18	89	96	7	89	96	7	85	97	12	84	96	13
	Oikopleuridae	4967	50	75	25	39	50	11	1	67	66	4	76	73	1	71	70	7	60	54
	tail App.	1243	65	67	2	42	59	16	30	68	38	55	75	20	41	68	26	48	66	18
	trunk App.	193	73	72	-1	34	58	24	6	69	64	22	75	53	11	71	60	27	65	39
Chaetognatha	Chaetognatha	7859	79	96	17	59	96	37	47	90	42	63	92	29	59	93	34	61	94	33
	head Chaeto.	190	75	99	25	65	99	34	74	96	22	79	97	18	74	98	23	72	98	26
	tail Chaeto.	555	64	90	26	37	82	45	42	91	49	63	93	30	51	91	40	47	87	40
Cirripedia	cirrus	60	0	80	80	19	74	55	0	86	86	36	100	64	0	83	83	24	85	60
	cypris	147	63	83	21	37	55	18	4	70	66	34	84	50	8	76	68	35	66	31
Cirripedia	nauplii Cir.	649	60	81	21	75	90	14	77	84	6	50	67	16	68	83	15	60	77	16
	Evadne	5003	81	95	14	69	89	20	81	91	10	88	95	7	81	93	12	77	92	15

			Precision					Recall					F1				
			NW			W		NW			W		NW			W	
Larger group	Class	n	RF	CNN	diff	RF	CNN	diff	RF	CNN	diff	RF	CNN	diff	RF	CNN	diff
Cladocera	Penilia	3592	62	91	28	42	88	46	38	90	52	58	92	34	47	90	41
	Podon	292	64	80	17	38	62	24	21	61	40	44	81	37	32	69	29
	Acartiidae	8853	75	96	21	73	96	24	73	95	22	74	95	21	74	95	22
	Calanidae	6190	90	98	8	91	99	7	92	98	5	91	97	6	91	98	7
	Calanoida	22713	90	94	4	38	87	49	9	95	86	48	98	49	17	94	78
	C. pavo	71	68	97	30	59	96	37	57	97	40	62	97	36	62	97	35
	Candaciidae	1767	74	93	19	61	92	31	66	96	30	73	96	23	70	94	25
	Centropagidae	6890	0	78	78	18	82	64	0	74	74	7	81	74	0	76	76
	Copilia	99	78	92	14	60	94	34	63	94	31	71	92	20	70	93	23
	Corycaeidae	3576	76	90	14	42	88	47	15	89	74	27	96	68	25	89	64
	Eucalanidae	183	71	96	26	63	97	34	71	96	25	76	95	20	71	96	26
	Euchaetidae	1019	0	58	58	42	75	33	0	65	65	32	62	29	0	61	61
Copepoda	Haloptilus	407	87	96	9	86	94	8	92	97	5	92	97	5	90	97	7
	Harpacticoida	832	66	87	22	36	80	44	23	88	65	49	89	41	34	88	54
	Heterorhabdidae	355	70	92	23	52	95	43	83	96	13	89	95	5	76	94	19
	Metridinidae	2439	87	99	12	80	99	19	82	98	16	84	98	14	84	98	14
	Oithonidae	9847	100	75	-25	28	76	48	9	90	81	39	93	54	16	82	65
	Oncaeidae	3070	100	87	-13	55	63	9	7	80	74	24	86	62	12	84	71
	Pontellidae	1080	***85***	***87***	***2***	***84***	***88***	***3***	***77***	***94***	***16***	***75***	***91***	***16***	***81***	***90***	***9***
	Rhincalanidae	35	71	69	-2	35	47	12	5	57	52	20	75	55	9	62	53
	Sapphirinidae	162	79	96	17	52	91	39	36	91	56	50	93	43	49	93	44
	Temoridae	4549	70	89	19	55	88	33	64	82	18	74	84	9	67	85	18
Ctenophora	Ctenophora	137	71	92	21	52	85	33	33	84	51	49	88	39	45	88	43
cyphonaute	cyphonaute	1334	74	96	22	43	97	54	58	94	36	77	96	19	65	95	30
	Iv. Luciferidae	98	0	83	83	67	77	11	0	79	79	11	89	79	0	81	81
	Iv. Porcellanidae	748	69	64	-6	62	56	-6	37	59	22	40	64	24	49	61	13
	megalopa	213	64	92	28	48	91	43	49	95	46	61	94	33	55	94	38

				Precision				Recall				F1			
				NW		W		NW		W		NW		W	
Larger group	Class	n	RF	CNN	diff	RF	CNN	diff	RF	CNN	diff	RF	CNN	diff	RF	CNN	diff
Decapoda	pr. Penaeidae	59	78	89	11	59	82	23	51	84	33	60	88	28	62	86	25	59	85	26
	pr. Sergestidae	89	0	53	53	11	30	19	0	25	25	1	58	58	0	34	34	1	40	38
	zo. Brachyura	1750	68	98	30	58	96	38	60	95	35	61	97	36	63	96	33	60	97	37
	zo. Galatheidae	759	58	90	33	39	82	43	51	85	34	61	96	35	54	87	34	48	88	41
Doliolida	Doliolida	1461	66	87	21	47	81	34	33	88	56	52	94	42	44	88	44	50	87	38
Echinodermata	lv. Echinodermata	76	73	93	20	43	93	50	52	93	41	70	95	25	61	93	32	53	94	41
	plu. Echinoidea	361	0	76	76	36	79	43	0	94	94	29	88	59	0	84	84	32	83	51
	plu. Ophiuroidea	542	67	84	17	54	72	18	74	90	16	82	94	12	70	87	17	66	82	16
Eumalacostraca	Eumalacostraca	3453	87	96	9	53	100	47	25	89	64	74	89	15	38	92	54	62	94	32
	lv. Mysida	14	75	75	0	72	64	-8	88	84	-4	92	93	1	81	79	-2	81	76	-5
	Mysida	120	85	94	8	50	88	38	44	88	45	61	91	31	58	91	33	55	90	35
	pr. Eumalacostraca	225	25	68	43	29	57	29	3	38	35	15	68	53	5	48	44	20	62	42
Harosa	Harosa	244	50	42	-8	11	60	49	6	50	44	25	56	31	11	46	35	16	58	42
Isopoda	Isopoda	83	97	96	-1	35	94	59	20	90	70	43	93	50	33	93	60	38	93	55
Mollusca	Atlanta	68	73	96	22	66	94	27	77	96	19	82	97	15	75	96	21	74	95	22
	Bivalvia	777	48	87	39	41	74	34	14	83	69	37	93	56	22	85	63	39	83	44
	C. inflexa	662	87	96	9	79	95	17	82	94	12	87	95	8	84	95	11	82	95	12
	Creseidae	767	71	92	21	75	95	20	77	93	15	72	87	16	74	92	18	73	91	18
	C. acicula	1294	76	98	21	74	98	24	87	96	9	87	95	7	82	97	15	80	96	16
	C. peroni	14	89	94	4	59	94	35	63	91	28	78	95	16	74	92	18	67	94	27
	egg Mollusca	129	52	94	41	36	95	59	31	90	59	40	87	47	39	92	53	38	91	53
	Gymnosomata	79	75	98	23	66	97	30	85	96	11	89	97	7	80	97	17	76	97	21
	Limacinidae	2113	65	85	20	51	79	29	40	91	50	50	90	40	50	88	38	50	85	34
	part Moll.	255	0	97	97	47	86	39	0	81	81	24	86	62	0	88	88	32	86	54
	Actiniaria	22	93	76	-17	61	92	30	59	100	41	86	100	14	72	86	14	72	96	24
	ephyra	179	90	98	8	71	98	26	43	96	53	63	99	36	59	97	38	67	98	32
	Hydrozoa	579	100	86	-14	51	85	34	11	71	60	54	83	29	21	78	58	53	84	31

			Precision						Recall						F1					
			NW			W			NW			W			NW			W		
Larger group	Class	n	RF	CNN	diff	RF	CNN	diff	RF	CNN	diff	RF	CNN	diff	RF	CNN	diff	RF	CNN	diff
other Cnidaria	Obelia	147	71	94	23	66	93	28	72	93	20	74	94	19	72	93	22	70	93	24
	part Cnidaria	125	0	82	82	25	77	52	0	62	62	10	66	56	0	70	70	14	71	57
other Crustacea	calyptopsis	1205	91	97	7	84	97	13	88	97	9	90	97	6	89	97	8	87	97	10
	lv. Stomatopoda	245	68	85	17	48	86	37	64	95	31	68	93	25	66	90	24	57	89	33
	meta. Crustacea	37	100	69	-31	29	42	13	2	62	60	23	87	63	3	65	62	26	57	31
	np. Crust.	845	73	90	17	64	87	23	73	81	9	78	87	9	73	86	13	70	87	17
	Ostracoda	1169	**92**	**96**	**4**	**56**	**95**	**39**	**58**	**97**	**39**	**72**	**96**	**23**	**71**	**96**	**25**	**63**	**95**	**32**
	part Crust.	3065	59	94	34	38	92	54	41	95	54	55	97	42	49	94	46	45	94	49
Pyrosomatida	Pyrosomatida	75	66	65	-1	39	56	17	10	65	55	21	72	51	17	65	48	27	63	36
Rhizaria	Foraminifera	469	77	97	20	74	97	23	80	92	13	79	92	13	78	95	16	76	94	18
	Phaeodaria	8106	73	86	13	43	84	41	48	94	46	62	97	34	58	90	32	51	90	39
Salpida	endostyle	135	97	94	-3	50	81	32	20	80	60	34	90	56	33	87	54	40	85	45
	jv. Salpida	67	100	78	-22	43	68	26	2	70	68	21	70	48	4	74	70	29	69	40
	nucleus	222	76	95	18	52	90	39	49	89	41	62	92	31	60	92	32	56	91	35
	Salpida	2460	67	90	23	47	83	36	4	90	86	20	96	75	7	90	83	28	89	60
Siphonophorae	Bassia	15	73	93	20	78	95	16	84	95	11	78	94	16	78	94	16	78	94	16
	bract A. tetragona	185	66	96	29	61	93	32	58	97	39	61	97	36	62	96	34	61	95	34
	bract Diphyidae	2185	82	95	14	64	96	32	83	98	15	85	98	13	82	97	14	73	97	24
	eud. A. tetragona	98	68	95	27	64	94	30	66	95	29	66	96	29	67	95	28	65	95	29
	eud. Diphyidae	525	92	97	5	69	93	24	37	84	48	61	89	28	52	90	38	65	91	26
	gono. A. tetragona	199	72	95	23	64	95	30	66	94	28	70	93	22	69	94	26	67	94	26
	gono. Diphyidae	2460	71	98	27	63	96	34	66	95	29	72	96	24	68	97	28	67	96	29
	necto. A. tetragona	173	67	73	7	45	58	12	11	60	49	43	74	32	19	66	47	44	65	21
	necto. Diphyidae	4417	66	53	-13	32	47	15	17	68	51	50	91	41	27	60	33	39	62	23
	necto. Hippopodiidae	17	85	93	8	31	88	57	20	85	66	59	89	30	32	89	57	41	89	48
	necto. Physonectae	1386	72	94	23	47	87	39	44	85	41	60	93	33	54	89	35	53	90	37
	part Siphonophorae	412	67	95	28	45	94	49	36	91	55	58	90	32	47	93	46	51	92	41

				Precision				Recall				F1			
				NW		W		NW		W		NW		W	
Siphonophorae Larger group	Class	n	RF	CNN	diff	RF	CNN	diff	RF	CNN	diff	RF	CNN	diff	RF	CNN	diff
	Physonectae	16	78	92	14	72	92	20	81	94	13	84	93	9	80	93	13	77	93	15
	siphonula	144	85	88	3	34	83	49	29	86	58	59	88	29	43	87	44	43	85	42
Stramenopiles	Coscinodiscus	1075	68	93	24	53	91	38	33	96	63	43	96	53	45	94	50	47	93	46
	actinula	19	75	93	17	57	92	35	61	90	29	71	91	20	68	92	24	63	92	28
	Aglaura	455	80	92	12	28	85	57	3	78	75	10	90	79	6	84	78	15	87	72
Trachylina	Liriope	34	80	87	8	63	88	25	63	94	31	73	94	21	70	91	20	68	91	23
	R. velatum	373	57	66	9	25	49	23	1	72	71	15	77	62	2	69	67	19	60	41
	S. bitentaculata	56	78	68	-10	23	60	38	8	84	76	48	91	43	14	75	61	31	73	42
	Average		***69***	***87***	***18***	***52***	***83***	***31***	***42***	***85***	***43***	***55***	***89***	***34***	***47***	***86***	***39***	***52***	***85***	***33***
Non plankton																				
artefact	artefact	7718	0	71	71	67	42	-24	0	67	67	13	73	60	0	69	69	22	54	31
badfocus	badfocus	6046	73	97	25	77	98	21	86	96	10	82	96	13	79	97	18	79	97	17
bubble	bubble	2432	65	95	30	65	94	29	68	95	27	68	95	27	67	95	28	66	95	28
detritus	detritus	36260	71	94	23	50	92	42	18	93	74	33	92	59	29	93	64	40	92	52
fiber	fiber	6708	79	92	13	33	88	56	10	84	74	20	86	66	18	88	70	25	87	62
Insecta	Insecta	169	76	94	19	69	93	24	76	92	16	77	92	15	76	93	17	73	93	20
other egg	other egg	2015	0	100	100	50	100	50	0	57	57	7	64	57	0	73	73	13	78	66
other living	other living	40	86	98	11	72	97	25	86	97	11	89	98	9	86	97	11	80	98	18
seaweed	seaweed	1272	76	82	6	49	68	19	44	90	46	65	96	31	55	86	30	56	80	24
	Average		***58***	***91***	***33***	***59***	***86***	***27***	***43***	***86***	***42***	***51***	***88***	***38***	***46***	***88***	***42***	***50***	***86***	***35***

Part III

Plankton distribution at various scales

The distribution of planktonic organisms was investigated at different spatial and temporal scales. Each chapter in this section focuses on the distribution of plankton at a given scale, starting from the largest – the global scale – to finer scales – meso and submesoscale.

4

Three mesoplanktonic worlds resolved by *in situ* imaging in the upper 500 m of the global ocean

In this chapter, I leverage a global dataset of UVP5 profiles performed worldwide to investigate large types of plankton community throughout the world's ocean. Such dataset was obtained by assembling and homogenising *in situ* imaging data collected during various oceanographic campaigns, deploying the UVP5 along vertical profiles. The 6.8 million imaged objects were pre-classified with the assistance of a machine learning algorithm and manually reviewed by a human operator. A total 330,000 objects were identified as planktonic organisms, either large zooplanktonic organisms and phytoplankton colonies. Then, data mining approaches were applied to characterise plankton community types and understand how they relate to environmental conditions. This work brings to light the importance of unexpected groups such as *Trichodesmium* (cyanobacteria) and Rhizaria (unicellular eukaryote) as structuring elements of plankton communities. Moreover, the distribution of these plankton communities seems to be mostly driven by basin-scale environmental conditions. These results call for studying the distribution of plankton communities on smaller scales too in order to obtain a thorough understanding of their distribution.

Thelma Panaïotis, Marcel Babin, Tristan Biard, François Carlotti, Laurent Coppola, Lionel Guidi, Helena Hauss, Lee Karp-Boss, Rainer Kiko, Fabien Lombard, Andrew MP McDonnell, Marc Picheral, Andreas Rogge, Anya M Waite, Lars Stemmann and Jean-Olivier Irisson

Manuscript in preparation to be submitted to **Global Ecology and Biogeography**

Abstract

Aim Ocean biogeographies are mainly based on biogeochemical signatures combining *in situ* optical data, remote sensing, and sometimes biogeochemical model output. However, the consistency between these regionalisations and the distribution of planktonic organisms remains an open question.

Location Global ocean, 0-500 m depth.

Time period 2008-2019

Major taxa studied 28 groups of planktonic organisms, covering Metazoa, Rhizaria and cyanobacteria.

Methods Using *in situ* imaging, we studied the global distribution of meso- and macro-planktonic organisms (> 600 µm Equivalent Spherical Diameter). We used a global data set of 2500 vertical profiles making use of an Underwater Vision Profiler 5 (UVP5). Among the 6.8 million imaged objects, 330,000 were large zooplanktonic organisms and phytoplankton colonies, while the rest were mainly marine snow particles. Multivariate statistical ordination and regression methods were used to describe patterns in community composition and their correlation with environmental variables in the epipelagic and upper mesopelagic layers.

Results Epipelagic plankton communities were dominated by *Trichodesmium* in the intertropical Atlantic, by Copepoda at high latitudes and in upwelling areas, and by Rhizaria in oligotrophic areas. In the mesopelagic layer, Copepoda-dominated communities were also found at high latitudes and in the Atlantic Ocean, Rhizaria-dominated communities prevailed in the Peruvian upwelling system and a few mixed communities were found elsewhere. The comparison between the distribution of these communities and a set of existing regionalisations of the ocean suggested that the structure of plankton communities described above is mostly driven by basin-level environmental conditions rather than the conditions in the immediate vicinity of the sampling site.

Main conclusions In both layers, three types of plankton communities emerged and seemed to be mostly driven by regional environmental conditions. This work sheds light on the role not only of metazoans, but also of unexpected large protists and cyanobacteria in structuring plankton communities.

Keywords biogeography, global ocean, *in situ* imagery, plankton communities, spatial distribution

Résumé

But Les biogéographies océaniques sont principalement basées sur des signatures biogéochimiques combinant *in situ* des données optiques, la télédétection et parfois les résultats de modèles biogéochimiques. Cependant, la cohérence entre ces régionalisations et la distribution des organismes planctoniques reste une question ouverte.

Location Océan mondial, profondeur de 0 à 500 m.

Time period 2008-2019

Principaux taxons étudiés 28 groupes d'organismes planctoniques, comprenant des Metazoa, des Rhizaria et des cyanobactéries.

Méthodes En utilisant l'imagerie *in situ*, nous avons étudié la distribution globale des organismes méso- et macro-planctoniques (> 600 µm de diamètre sphérique équivalent). Nous avons utilisé un jeu de données global de 2500 profils verticaux en utilisant un profileur de vision sous-marine 5 (UVP5). Parmi les 6,8 millions d'objets imagés, 330 000 étaient des organismes zooplanctoniques ou des colonies de phytoplancton, tandis que le reste était principalement des particules de neige marine. Des méthodes statistiques multivariées d'ordination et de régression ont été utilisées pour décrire les modèles de composition des communautés et leur corrélation avec les variables environnementales dans les couches épipélagiques et mésopélagiques supérieures.

Résultats Les communautés planctoniques épipélagiques étaient dominées par des *Trichodesmium* dans l'Atlantique intertropical, par les copépodes aux hautes latitudes et dans les zones d'upwelling, et

par Rhizaria dans les zones oligotrophes. Dans la couche mésopélagique, les communautés dominées par les copépodes ont également été trouvées à des latitudes élevées et dans l'océan Atlantique, les communautés dominées par rhizaires ont prévalu dans le système d'upwelling péruvien et quelques communautés mixtes ont été trouvées ailleurs. La comparaison entre la distribution de ces communautés et un ensemble de régionalisations existantes de l'océan suggère que la structure des communautés planctoniques décrites ci-dessus est principalement déterminée par les conditions environnementales au niveau du bassin plutôt que par les conditions à proximité immédiate du site d'échantillonnage.

Principales conclusions Dans les deux couches, trois types de communautés planctoniques sont apparus et semblent être principalement déterminés par les conditions environnementales régionales. Ce travail met en lumière le rôle non seulement des métazoaires, mais aussi des grands protistes inattendus et des cyanobactéries dans la structuration des communautés planctoniques.

Mots clés biogéographie, océan global, imagerie *in situ*, communautés planctoniques, distribution spatiale

4.1 Introduction

Biogeography aims to describe spatial biodiversity patterns.

Biological communities are heterogeneously distributed: this is key for ecosystem functioning [226]. Biogeography is a science describing and trying to understand these distributions and how they aggregate in distinct ecosystems [60, 245]. This produces continuous distribution maps or delimitation of regions homogeneous in composition (i.e., regionalisations; [226]). Although global species distributions have been described for over two centuries [61], biogeography remains a relevant topic. Beyond species distribution, it includes the traits distribution (dispersal, polyploidy...); [10, 63, 332], providing new insights into organisms' ecology and evolution. In the marine realm, recent studies suggested new regionalisations based on the environment alone [428] but also regarding species distribution, either for phytoplankton [180] or for 65,000 species across phyla [83], highlighting great endemicity in marine phyla.

Plankton – organisms drifting with currents – are incredibly diverse and cover a large size range [69, 186]. It supports oceanic food webs [125, 413], plays a major role in biogeochemical cycles through the biological pump [246]. Phytoplankton contributes to primary production [131]. As drifters, planktonic organisms are distributed worldwide but their distribution is shaped by the conditions of the water mass they are embedded in [172]. Because these conditions vary with latitude, the corresponding variations in plankton distribution are well-known: higher diversity towards low latitudes [188, 340, 345, 399] and higher biomass towards higher latitudes [190]. They are sensitive to environmental conditions: planktonic organisms are also global change sentinels [20, 22, 172]. Studying plankton biography is relevant to understand anthropocene pelagic ecosystems.

Plankton is diversified and sensitive to its environment, thus well appropriate for biogeography.

In terrestrial biogeography, biomes rest on vegetation types, but also coincide with climatic zones and soil type distribution: they constrain plant growth [60]. Compared to the terrestrial realm, assessing oceanic biogeography presents inherent difficulties: costly global scale offshore sampling; observing distribution varying in time and space in a three-dimensional and opaque environment... Early ocean biogeographies considered various taxa's distribution, including copepods, euphausiids, Rhizaria or phytoplankton [375]. Simultaneously, non-biological regionalisations were based on the physical environment: ocean currents, temperature, salinity, ice conditions [375]. Novel technologies, such as satellites, fostered ecological ocean geography: using surface chlorophyll *a* concentration computed from ocean colour – proxy for phytoplankton concentration – new regionalisations emerged. However, most new approaches ignore organisms distribution: the 56 Longhurst Provinces [244, 245] considered physical forcing (sea surface temperature, mixed layer depth...) as phytoplankton distribution regulators. A widely used global synthetic regionalisation is based on latitudinal bands [375]. As explained above, it correlates with major environmental variables (temperature, light intensity...). Other regionalisations, the Word Marine Ecoregions [376] or the Large Marine Ecosystems [364] include biotic data, but focus on coastal areas only. In contrast, Costello et al. [83] delineated marine biogeographic realms using the distribution of marine animals, plants, and Hofmann Elizondo et al. [180] defined biomes using phytoplankton distribution. Furthermore, these regionalisations focus on the epipelagic layer: not suitable for less described deeper ocean, harder to sample and not necessarily linked

Marine regionalisations can be based on species distribution and on biogeochemical conditions.

to surface characteristics [84, 375]. Few regionalisations targeted the mesopelagic: Reygondeau et al. [331] suggested dividing it into 13 provinces, based on environmental variables, while introducing the definition of the mesopelagic layer dynamic top and bottom boundaries, based on environmental conditions (light, density, carbon flux...).

Are these regionalisations coherent with plankton distribution?

Briefly, ocean biogeography was described through regionalisations, but about the epipelagic layer or coastal areas; delineation is based on physical and biogeochemical variables. Widespread and undersampled offshore areas, however, are crucial for biogeochemical cycles [120] and target conservation through developing protected marine areas. Although a few organisms' groups' spatial distribution – copepods [23, 340, 418], microorganisms [137, 183, 233] or larger species assemblages [55, 345, 374, 399] – were described previously, consistency between these and biogeochemistry-based regionalisations remains unexplored.

In situ imaging is particularly appropriate to study plankton distribution.

Limited quantitative and basin-scale data about offshore planktonic organism distribution is available. Because plankton was traditionally sampled using nets, pumps... These methods require lengthy taxonomic identification [30], partly subjective [90], therefore not scaling well to large spatiotemporal scales biogeography. Besides, they may damage fragile organisms [330]. New *in situ* cameras now image planktonic organisms in their natural environment and resolve their fine scale vertical distribution [385], while generating large datasets, homogenised by reviewing images [192, 205]. These tools also allow studying fragile taxonomic groups: Rhizaria, whose contribution to global planktonic biomass was underestimated [42, 104]. These approaches lack in taxonomic identification fineness, but compensate with identification and data quantity consistency.

The UVP5: an in situ imager and particle counter deployed along vertical profiles.

Among these imaging systems, the Underwater Vision Profiler 5 (UVP5) images planktonic organisms and marine snow particles larger than 600 µm Equivalent Spherical Diameter (ESD) along vertical profiles [317], therefore sampling large meso- and small macro-plankton, mostly comprising animals and some large phytoplankton colonies. Later, we call our study assemblage "plankton", for simplicity. Concentration and biovolume estimates from the UVP5 proved coherent with those from net samples for large (> 1 mm ESD) Arctic copepods, while smaller organisms were underestimated [133]. Data from UVP5 was already used to estimate organic carbon vertical particle flux [162] or study zooplankton distribution [42, 133, 385]. Leveraging net data

from the Tara Oceans expedition Soviadan et al. [374] described a sharp decrease in zooplankton concentration with depth, but lower particle flux attenuation in Oxygen Minimum Zones (OMZ). The opposite occurs in other world regions. UVP5 can also provide planktonic organisms' individual traits information: different size and activity patterns around Arctic ice melt zones [409]; or decrease in organisms' sizes at low latitudes [55], and in marine snow particles, highlighting consistent particle types changes along two Arctic blooms [401]. Using a UVP5 biomass estimates regression on environmental climatologies, Drago et al. [111] estimated global scale plankton biomass: it was dominated by copepods peaks at high latitudes.

We study global scale plankton biogeography, leveraging a dataset assemblage collected by UVP5. We address these questions: what are plankton communities large types in the open ocean; what differences between epipelagic and upper-mesopelagic layers; what drivers behind these community types distribution? We first (i) describe plankton communities structures; their relation to their immediate physical and biogeochemical environment. We then (ii) assess the ability of various physics and biogeochemistry-based regionalisations to describe these planktonic communities distribution and evaluate their ecological description relevance.

This study provides new insights on plankton community distribution.

4.2 Material and methods

4.2.1 Data collection

Data from multiple oceanographic campaigns (Figure S4.1, Table S4.1) (2008 - 2019) – when UVP5 vertical profiles were performed – was aggregated, creating a large dataset covering world's oceans. This *in situ* imaging system captures objects within an approximately 1 L volume, illuminated by two led beams - up to 20 Hz frequency during a CTD cast descending part (Conductivity, Temperature, Depth sensor) [317]. All objects larger than 100 µm ESD were measured for area and grey level. Images were saved for objects larger than ~600 µm ESD; this paper focuses on the latter part. The CTD provided temperature, salinity and, for most of the campaigns, also chlorophyll *a* fluorescence and oxygen profiles.

The dataset encompasses the world ocean and consists of plankton and marine snow concentrations, CTD...

. . . and satellite data.

Satellite GlobColour[1] products completed the environmental dataset: averaged over one month on a 100×100 km area centred on UVP5's sampling site. Although this data resolution is low in terms of time and space, using higher resolution data (8 days average, 20×20 km area) provided a too great missing data proportion. Satellite data provided: surface chlorophyll a concentration; particulate backscattering coefficient (bbp); photosynthetically active radiation (PAR); diffuse attenuation coefficient (K_{dPAR}); particulate organic carbon (POC); particulate inorganic carbon (PIC). With these data, organism concentrations were associated with specific environmental conditions.

4.2.2 Data processing

Images were sorted into a predefined set of groups: plankton, marine snow, artefact, unidentified.

All UVP5 images entered the EcoTaxa web application [316]. They were classified as marine snow, artefact, badfocus, unidentified; or into several taxonomic groups according to the UniEuk taxonomic tree [34]. All objects were manually validated or corrected. Striving for consistency, a taxonomic sorting guide was published and circulated; difficult groups were reviewed by a single operator across all cruises. Human operators fully checked the resulting dataset, often several for each image. Data from 2500 fully validated profiles was retained (6.8 M objects, 300,000 classified as plankton). Differences in classification taxonomic depth among cruises caused some groups to merge, obtaining a lower common denominator. Then, other groups were merged, for exhibiting similar patterns in preliminary analysis and were first not well differentiated (e.g., Copepoda and Copepoda-like). The final list contained 28 taxa (Figure 4.1).

Concentrations were computed from individual planktonic organisms and marine snow objects.

Object counts per class and imaged water volume computed concentrations (L^{-1}) per 5 m bins along each profile. Concentration and biovolume were also computed for marine snow (objects larger than 600 µm identified as aggregates) and bulk particulate matter (all objects > 100 µm ESD imaged by the UVP, including both plankton and marine snow). Marine snow and bulk concentrations stood as environmental variables, plankton organisms being related to them. They are proxies for, respectively, organic matter amounts sinking from the upper layers to ocean depths and the overall oligo- or eutrophic water mass state.

After removing abnormal values (codes for missing data, negative salinity, oxygen concentration or fluorescence), more than 20% of miss-

1 http://globcolour.info

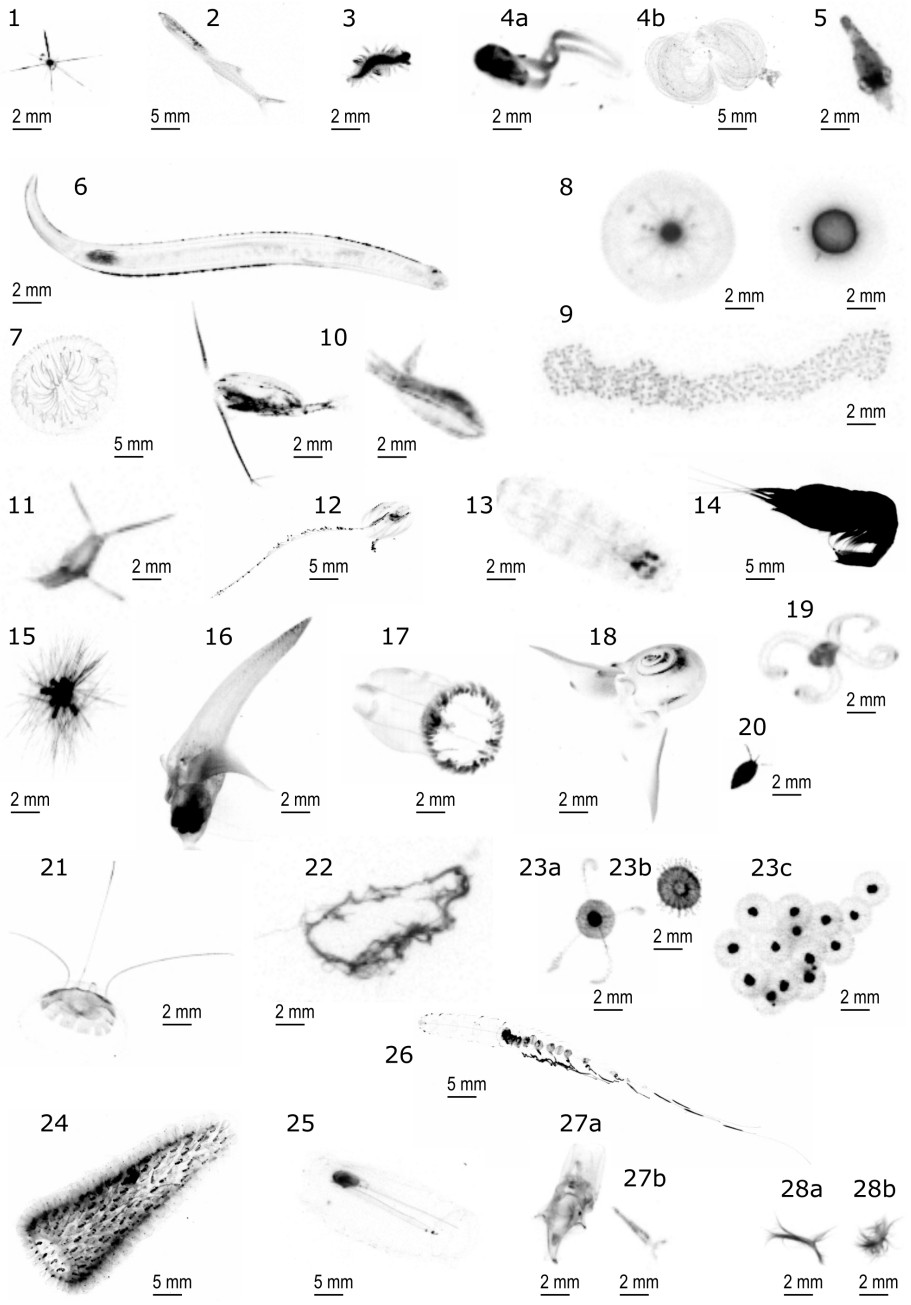


Figure 4.1: *(Caption on next page)*

Figure 4.1: Examples of UVP5 images for selected taxonomic groups. **1**: Acantharea, **2**: Actinopterygii, **3**: Annelida, **4**: Appendicularia (**4a**: Appendicularia body, **4b**: Appendicularia house), **5**: Cephalopoda, **6**: Chaetognatha, **7**: Cnidaria others, **8**: Collodaria, **9**: colonial Collodaria, **10**: Copepoda, **11**: Crustacea others, **12**: Ctenophora, **13**: Doliolida, **14**: Eumalacostraca, **15**: Foraminifera, **16**: Gymnosomata, **17**: Hydrozoa others, **18**: Limacinidae, **19**: Mollusca others, **20**: Ostracoda, **21**: Narcomedusae, **22**: Nostocales, **23**: Phaeodaria (**23a**: Coelodendridae, **23b**: Aulacantha, **23c**: colonial Aulosphaeridae), **24**: Pyrosoma, **25**: Salpida, **26**: Siphonophorae, **27**: Thecosomata (**27a**: Cavoliniidae, **27b**: Creseidae), **28**: *Trichodesmium* (**28a**: tuft, **28b**: puff).

ing data profiles for any variable were ignored. All variables were linearly interpolated at a 1 m vertical resolution. Outliers were detected by computing the absolute deviation around a moving median along the profile [231] and removed. Smoothing was performed using a moving average. Potential density and apparent oxygen utilisation (AOU) were computed from temperature, salinity and oxygen concentration. Thermocline, halocline and pycnocline depths were calculated as depth of the largest variation in the relevant variable computed in a 5 m sliding window. The mixed layer depth (MLD) was computed at depth where density differed by more than 0.03 kg m^{-3} from reference density in the 0-5 m surface layer [96]. The deep chlorophyll maximum (DCM) and euphotic zone (Z_{eu}) depths were computed from the chlorophyll profile [279]. The stratification index was computed as the difference in potential density between the surface and 250 m, deeper than pycnocline in most profiles. All 1 m precision profiles were binned over 5m to match plankton data bins. Rare instances (1.7%) of missing satellite data were replaced by the corresponding variable average value.

Additional variables were derived from environmental data.

Plankton and environmental data were averaged over two layers: epipelagic and upper mesopelagic. Instead of the commonly used fixed boundary (200 m) between these two layers [82], we applied a dynamic definition. It was modified from Reygondeau et al. [331] and meant to better represent the functional difference between the two layers: the deepest value among the mixed layer depth and the euphotic depth, hence delimiting the zone above which photosynthetic plankton activity is highest. Result: an 88-metre median value for the epi-mesopelagic boundary (quantile 25% = 52 m, 75% = 121 m, Figure S4.2). As numerous UVP5 profiles were below 500 m, we set the upper mesopelagic zone bottom at that depth. Any profile covering

A dynamic computation of the epipelagic-mesopelagic boundary was performed.

less than 80% of both layers' thickness was removed (3% and 29% of profiles for epipelagic and mesopelagic layers, respectively).

4.2.3 Global distribution of plankton communities

In each layer, the Hellinger transformation was applied to averaged plankton concentrations, to focus on community composition differences among profiles, rather than on absolute concentrations differences [45]. This helped go beyond the well-known pattern of high latitudes higher concentrations and lower ones around the equator, hence optimizing the taxonomic identification effort, while reducing very high abundances importance. To synthesise information, a principal component analysis (PCA) was performed on the Hellinger-transformed data. Environmental variables were projected into that space according to their correlation with plankton concentrations, after a log n+1 transformation for marine snow and bulk concentrations, to avoid over-representing some very high values. This helps visualise correlations directly on the PCA biplot. Each profile's scores on the first five principal components (PC) helped perform a synoptic hierarchical agglomerative clustering (HAC), using Ward's criterion [226]. Using the PC scores, not the original data, preserved most of the variance, while removing noise. The resulting dendrogram (Figure S4.3) was separated into some main branches, based on inertia jumps, identifying broad plankton community types. Composition, in terms of each taxon proportions, was computed for each plankton cluster.

Plankton community types were defined and described in terms of composition and environment.

Testing for potential diurnal and seasonal biases, we computed the variance portion in plankton community composition, explained by acquisition time or season. We used a redundancy analysis (RDA), with Hellinger-transformed concentrations as response variables, and a binary variable as explanatory, either day/night or productive/non-productive season. Seasons were defined based on latitudes and sampling months (Table S4.2) [215]. This model is based on light intensity and nutrients availability. In polar regions, light availability is often limited (namely in winter) but becomes sufficient after the summer ice breakup, allowing productivity. In mid-latitudes, both light and nutrients become available in spring and autumn, generating phytoplankton blooms. In tropical regions, productivity is limited all year by nutrients and remains low. Diurnal effects were tested in both layers; while seasonality was not tested in the mesopelagic layer since seasonal

Seasonal and circadian effects were checked for.

changes are weaker at depth [84]. In both cases, the portion of variance explained was computed as R^2.

4.2.4 Correspondence with ocean regionalisations

Multiple regionalisations were assessed for their ability to describe the distribution of planktonic communities...

To assess various regionalisations ability to capture processes driving plankton ecology, the variance part in plankton community composition explained by each was computed. Tested regionalisations included Longhurst provinces [245], 10° latitudinal bands, as well as mesopelagic provinces [331], tested for the mesopelagic layer only. Besides these regionalisations, often based on climatological averages, a regionalisation based on each profile's actual, immediate environment was generated with a PCA performed on environmental data, followed by an HAC on the first five PCs. The number of modalities was set to be similar to the other regionalisations, for comparison.

...by computing the explained variance in plankton community, for each regionalisation.

The evaluation of the variance proportion explained was performed analogously to the circadian or seasonal effects test: in each layer, an RDA was performed, with Hellinger transformed concentrations as response variables and a qualitative variable with regions from a given regionalisation as explanatory variable. To ensure adequate representativeness, each region in each regionalisation had to contain at least 25 profiles for inclusion. This limited the number of regions used for each regionalisation, but (positively) resulted in similar region numbers throughout all tested regionalisations: all contained 12-18 regions for the epipelagic and 9-11 regions regarding the mesopelagic zone (Table 4.1). Important for comparison: the variance portion explained by a categorical variable often increases with the number of modalities. To quantify how much variance is explainable by a categorical variable with those modalities numbers, a maximal model was built by computing a regionalisation on plankton concentrations themselves, with a similar groups number. We used the PCA on Hellinger-transformed concentrations data, followed by the HAC on the first five PCs, described in the previous section. Now, instead of cutting the tree based on inertia jumps, the groups number was set as a middle ground between modalities numbers for other regionalisations (epipelagic: 14; mesopelagic: 10). Note: in the epipelagic layer maximal model, the difference between 12 and 18 modalities is inconsequential: 12 modalities explain 52.8% of variance, while 18 modalities explain 55.8% (+3% of variance). Then, RDA was performed with this explanatory variable.

Table 4.1: Variance in community composition explained by different regionalisations for the epipelagic and mesopelagic layers. 2,203 and 1,193 profiles were included in the epipelagic and mesopelagic layers respectively. n = number of groups for each regionalisation.

Regionalisation	Epipelagic		Mesopelagic	
	n	R^2	n	R^2
Maximal model	14	0.541	12	0.452
Null model	15	0.007	11	0.010
Longhurst provinces	18	0.264	12	0.134
Latitude bands	13	0.175	11	0.102
Local environment	13	0.170	9	0.102
Mesopelagic provinces	-	-	11	0.118

Since response and explanatory variables are built with the same data, this RDA captures the maximum part of explainable variance. Finally, we also compared these regionalisations to a null model: profiles were randomly grouped into a similar number of clusters. If the variance portion explained by a given regionalisation is similar to the null model portion: this regionalisation does not capture plankton community composition variations.

All analyses were conducted with `R` version 4.0.3 and the `vegan` package version 2.5.7 [296].

4.3 Results

4.3.1 Circadian and seasonal cycles

Circadian and seasonal cycles explained very low variance.

Among profiles in the epipelagic layer analysis, 922 were performed by night and 1595 by day. The RDA performed with the day/night binary variable was significant ($p < 0.001$), probably because of numerous observations (n = 2517), but explained a very small part of variance (R^2 = 1.1%). In the mesopelagic layer, where 659 profiles were performed at night versus 1088 by day, the diel cycle explained the observed variance even less (R^2 = 0.9%, $p < 0.001$). On a global scale, diel vertical migration little impacts community composition, while concentrations changes can be significant [374]. About 30% of epipelagic profiles (592) occurred during the productive season, while 1925 during a

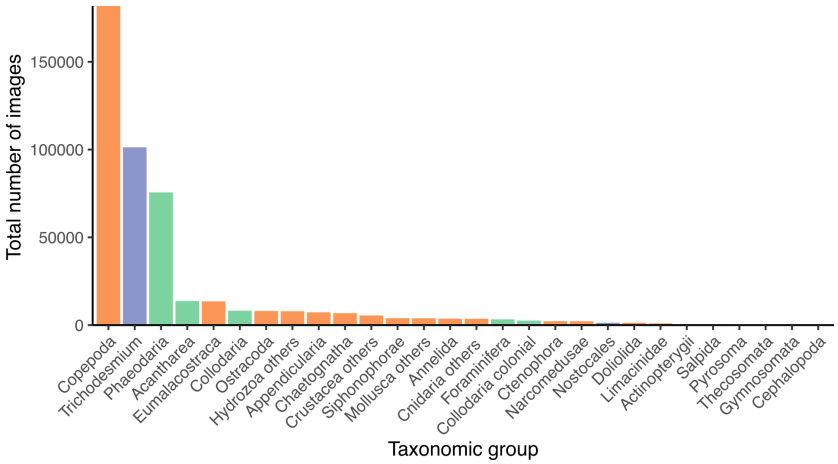


Figure 4.2: Dataset composition: total number of images per taxonomic group. Rhizaria (unicellular eukaryotes) are highlighted in green, Cyanobacteria are in purple and Metazoa in orange.

non-productive season (Table S4.2). The RDA was again significant ($p < 0.001$) but only explained 1.3% of variance. Thus, seasonal impact on studied plankton communities' structures also seems negligible compared space, at global scale.

4.3.2 Spatial distribution of plankton communities

Three plankton groups dominated in UVP5 data.

Within the size range acquired *in situ* by UVP5 (600 µm to a few cm), more than 330,000 organisms were detected. Three groups dominated in the dataset's number of individual images: Copepoda (metazoan), *Trichodesmium* (cyanobacteria) and Phaeodaria (Rhizaria subgroup, unicellular eukaryote) (Figure 4.2). Apart from Rhizaria, *Trichodesmium* and Nostocales (the latter both Cyanobacteria), all other imaged organisms belonged to the Metazoa kingdom.

4.3.2.1 *Epipelagic layer*

To describe epipelagic plankton communities, 2517 profiles were included. The first two PCs captured 41.8% of variance. The first PC distinguished between *Trichodesmium*-rich communities and copepod-rich ones. Trichosdesmium-dominated communities were associated with warm and stratified waters. Copepods dominated in cold, high

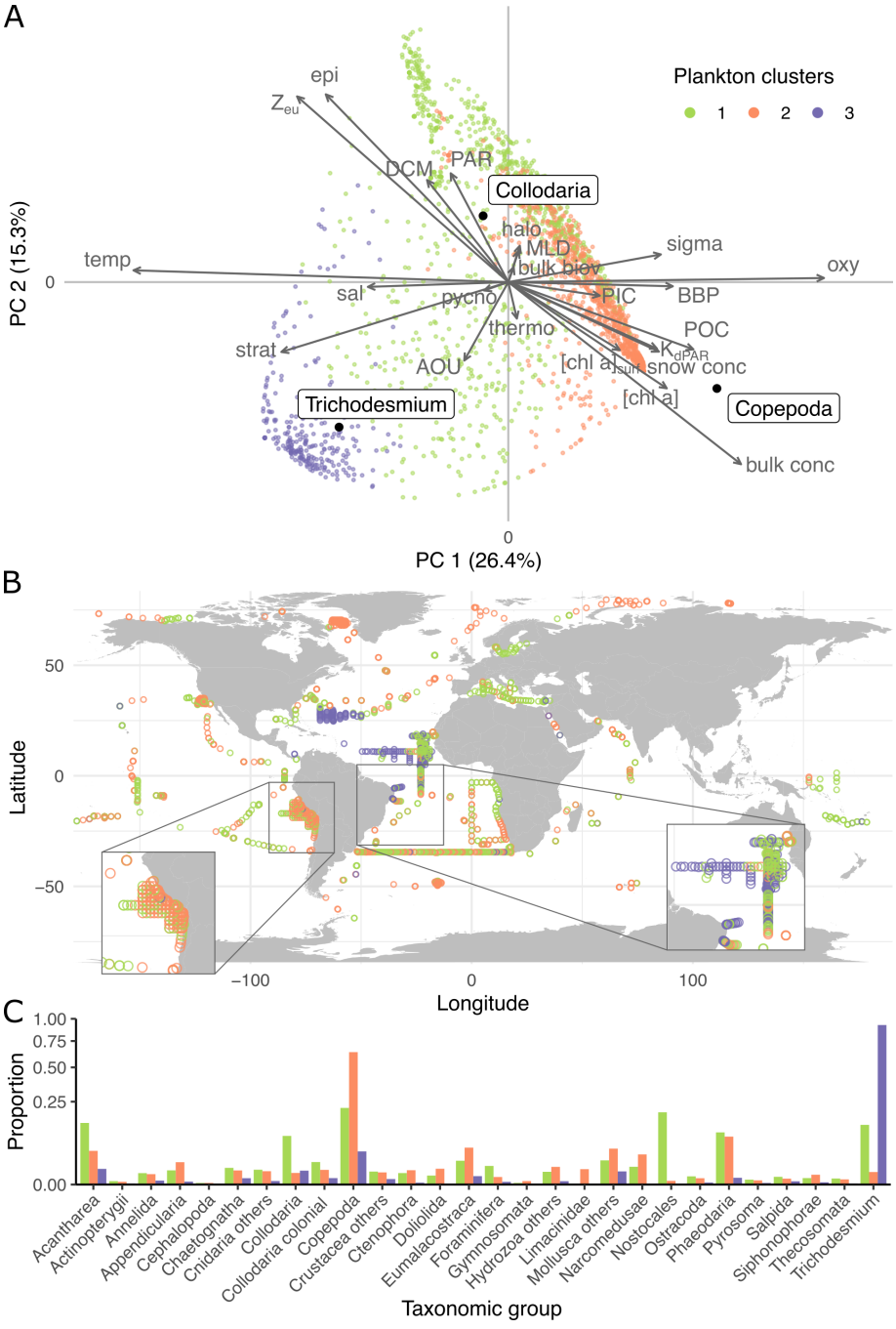


Figure 4.3: *(Caption on next page)*

Figure 4.3: Plankton clusters within the epipelagic layer. **(A)** PCA performed on Hellinger-transformed plankton concentrations, illustrated by a biplot in scaling 2. Only taxa with a contribution higher than average are represented. Environment variables are projected as supplementary variables. Points represent profiles and are coloured according to the cluster defined by the HAC. **(B)** Map of epipelagic profiles, coloured as in **(A)**. **(C)** Relative composition of epipelagic plankton clusters. Note that the y axis is square root transformed.

chlorophyll and particle-rich waters (Figure 4.3A). The second PC separated these communities from a third community, associated with collodarians (Rhizaria), extant in oligotrophic waters (with deep DCM and Z_{eu}). The HAC dendrogram was separated into 3 clusters (Figure S4.3). Cluster composition (Figure 4.3C): cluster 1 characterised by a co-dominance of multiple Rhizaria groups (Acantharea, Collodaria and Phaeodaria), copepods, nostocales and *Trichodesmium*, although Collodaria are the most structuring ones in PCA space. This community type was widely distributed in oceans but detected at lower frequencies in high latitudes (Figure 4.3B). Cluster 2, dominated by copepods, also had a widespread distribution but dominated the subpolar North Atlantic and Arctic shelf seas, and upwelling areas (California Current, Peruvian and Benguela upwellings). Cluster 3, dominated by *Trichodesmium*, exists in the Atlantic Ocean's intertropical band.

Three epipelagic communities emerged: copepod-dominated, Trichodesmium-dominated and mixed-type.

4.3.2.2 *Mesopelagic layer*

In the mesopelagic layer, 1747 profiles could describe plankton communities. The first two PCs captured a 39.6% variance. The first PC separated copepod-dominated waters from waters with Phaeodaria (Figure 4.4A). Copepod-dominated waters were oxygen-rich, while Phaeodaria-dominated waters exhibited high biological activity (high AOU) and significant stratification of the water column top. On the second PC, a third Eumalacostraca pole emerged, associated with warmer, saltier, more oligotrophic waters. The HAC dendrogram was again split into 3 clusters (Figure S4.3). The first cluster, dominated by phaeodarians, with fewer copepods (Figure 4.4C), was emblematic of the Peruvian upwelling, but also present in Mediterranean Sea and Pacific Ocean profiles. Conversely, the Pheodaria-dominated cluster 2 existed at high northern hemisphere latitudes; also present in the intertropical band and south of the Atlantic (Figure 4.4B). Finally, cluster 3 was not dominated by any taxon but had a diversified composition: Copepoda,

Three mesopelagic communities emerged: phaeodarian-dominated, copepod-dominated and mixed-type.

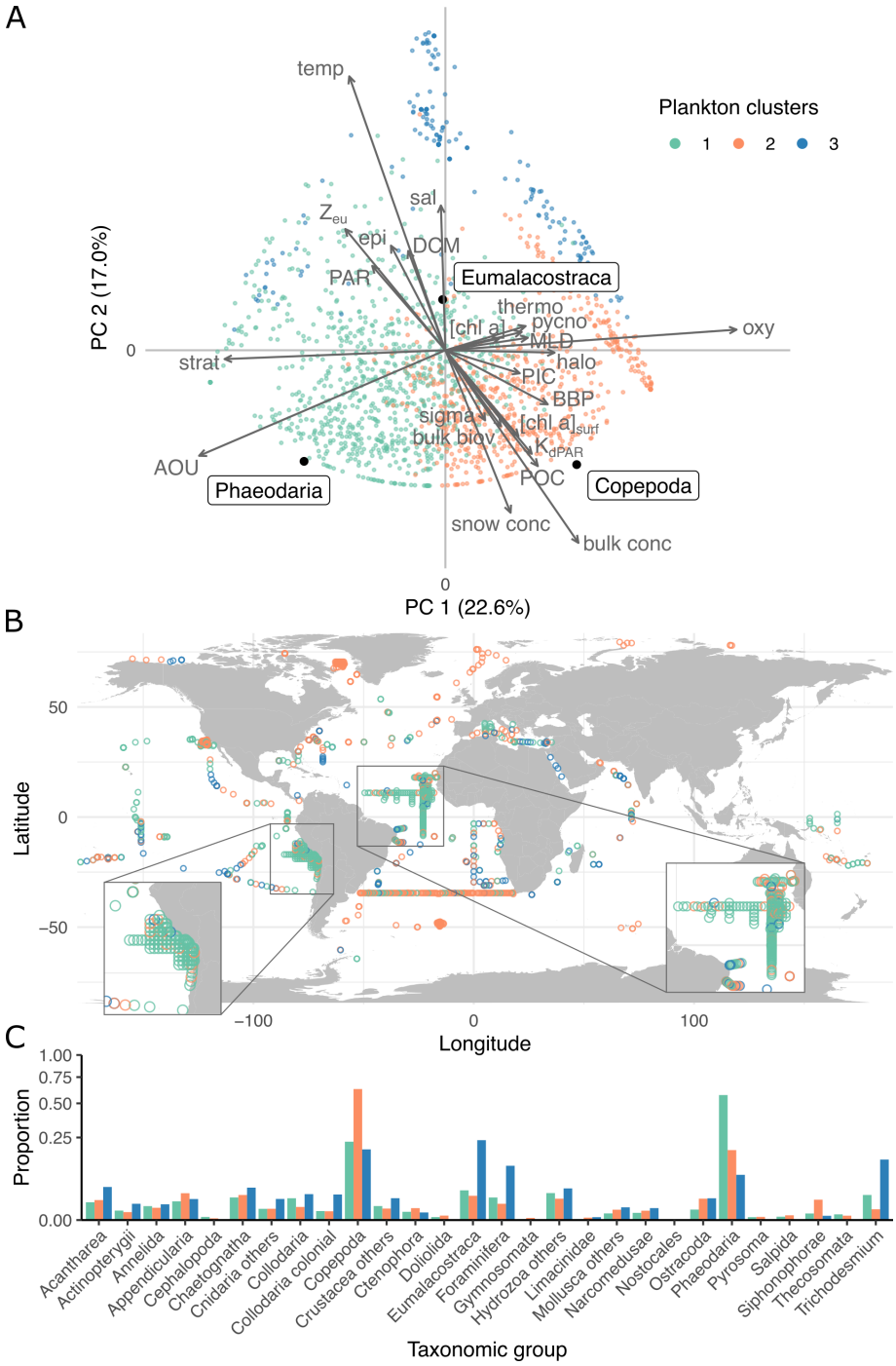


Figure 4.4: *(Caption on next page)*

Figure 4.4: Plankton clusters within the mesopelagic layer. **(A)** PCA performed on Hellinger-transformed plankton concentrations, illustrated by a biplot in scaling 2. Only taxa with a contribution higher than average are represented. Environment variables are projected as supplementary variables. Points represent profiles and are coloured according to the cluster defined by the HAC. **(B)** Map of mesopelagic profiles, coloured as in **(A)**. **(C)** Relative composition of mesopelagic plankton clusters. Note that the *y* axis is square root transformed.

Eumalacostraca, Foraminifera, Phaeodaria and *Trichodesmium*. Fewer profiles were in this cluster but broadly distributed.

4.3.2.3 *Comparison between epipelagic and mesopelagic layers*

An incomplete similarity between epi and mesopelagic plankton communities.

We compared each station's plankton community type in epipelagic and mesopelagic layers (Figure 4.5). Some stations, included in the epipelagic analysis, had no corresponding mesopelagic samples: the sea bottom or UVP maximum depth were too shallow. These stations are "NA" in Figure 4.5 mesopelagic row. Among the 1,122 profiles where the epipelagic part contained a mixed-type community (cluster 1 epipelagic), the mesopelagic layer contained a phaeodarian-dominated community (cluster 1 mesopelagic) in 49% of cases and a copepod-dominated community (cluster 2 mesopelagic) in 33% of cases. In copepod-dominated epipelagic community (cluster 2 epipelagic), most profiles displayed a mesopelagic copepod-dominated community (52%, cluster 2 mesopelagic). Finally, most profiles with a *Trichodesmium*-dominated epipelagic layer (cluster 3 epipelagic) had a phaeodarian-dominated mesopelagic community (75%, cluster 2 mesopelagic). Overall, this analysis highlights only an incomplete similarity between epipelagic and mesopelagic plankton communities' compositions, suggesting they are driven by different processes, which confirms they need studying separately.

4.3.3 Representativity of various ocean regionalisations

Plankton community distribution is better explained by regional rather than immediate conditions.

In both layers, the clustering built on the plankton data themselves (maximal model) explained about half of variance in community composition (epipelagic: 54.1%; mesopelagic: 45.2%; Table 4.1). The limited amount of variance explained unsurprisingly confirms plankton communities' diversity cannot be summarised in only 12 to 14 groups, but also highlights that other regionalisations should be gauged relatively

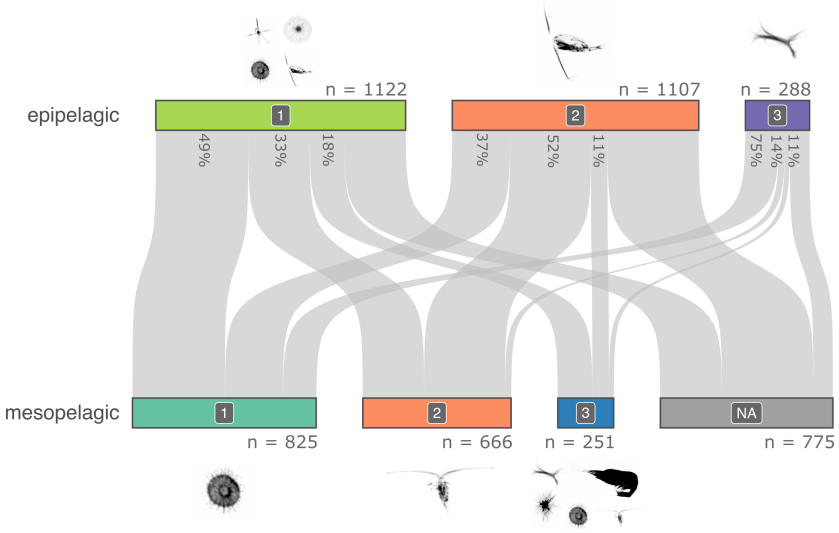


Figure 4.5: Comparison between epipelagic and mesopelagic plankton communities. Horizontal coloured bars represent profiles, split according to their plankton communities in each layer, as previously described in figures 2 and 3 (NA in the mesopelagic row represents stations with no mesopelagic portion). Grey bands show the correspondence of plankton communities between clusters in the epi and mesopelagic layer for each profile. Percentages show the repartition of epipelagic plankton communities in the mesopelagic layer, excluding profiles absent from the mesopelagic analysis (i.e., going in the NA band).

to these figures, not to 100% of variance explained. All regionalisations explained more variance than the random, null model (< 1% of variance explained in both layers). Thus, all regionalisations were relevant for explaining plankton communities' distribution. Among the tested regionalisations, Longhurst provinces explained more variance than others (epipelagic: 26.4%; mesopelagic: 13.4%), corresponding to about half of the explainable variance in the epipelagic layer and 1/3 in the mesopelagic zone. In the epipelagic layer, latitudinal bands explained similar variance to the local environment; while in the mesopelagic layer, Reygondeau's mesopelagic provinces explained some more variance than local environment and latitudinal bands. In both layers, plankton community composition was better explained by biogeochemical-based provinces than by the immediate and local environment the plankton was sampled in.

4.4 Discussion

Briefly, three meso/macro plankton communities' types were detected in both epipelagic and mesopelagic layers. Their composition was better explained by basin scale environmental conditions than by local ones. Below, we first discuss methodological aspects to assess our results' robustness and then discuss our findings' consequences in the existing knowledge context.

4.4.1 Potential Biases

4.4.1.1 *Seasonal and diel cycles effects*

DVM,...

Many – mostly pluricellular – plankton taxa conduct Diel Vertical Migrations [216]. Yet, our analysis showed no significant effect of this migration on plankton community composition (i.e., relative concentrations), in line with Soviadan et al. [374]'s findings. We performed an additional test in the California Current region, where extensive UVP sampling was conducted and compared the five most abundant taxa's day and night absolute concentrations. Copepoda, Eumalacostraca ($p < 0.001$) and Phaeodaria ($p < 0.05$) concentrations were higher at night in the epipelagic zone (Figure S4.4), showing that DVM existed in such data.

Both phytoplankton and zooplankton concentrations also vary seasonally: spring (and possibly autumn) blooms cause an increase in productivity and plankton concentration [25]. But plankton may also bloom outside seasonal blooms, due to favourable conditions following water mass displacements [268]. These sudden events, restricted spatially and temporally, are called intermittent blooms. For example, *Trichodesmium* can bloom locally in tropical and subtropical oceans [415] and high *Trichodesmium* concentrations sometimes observed in our data suggest multiple profiles in the Atlantic intertropical band occurred during such blooms. However, although seasonality affects absolute concentrations, our results suggested a negligible effect of season on community composition.

. . . seasonality,. . .

Briefly, both diel and seasonal effects are detected on absolute concentrations. Their effects' non-significance here is therefore due to Hellinger's transformation, focusing our analyses on relative rather than absolute concentrations [226]. With such focus on community composition and at the broad taxonomic level studied, the large-scale geographical effect dominated over seasonal and diel cycles.

. . . barely affected our analyses because we focused on community composition.

4.4.1.2 *Quality of taxonomic identifications*

UVP5 captures images of objects > 600 µm and, while larger than 1 mm organisms are reliably differentiated from marine snow, identifying smaller ones is trickier. Specific orientation or distinctive features are often required for performing identification confidently; otherwise, objects are classified as detritus, resulting in underestimating small organisms' concentrations. Normalised biomass size spectrum detects underestimation, showing a deviation from expected linear trends for the 600 µm - 1 mm size range [201]. Still, in our dataset, 80% of organisms were over 1 mm in ESD, therefore accurately detected by UVP5. Furthermore, provided underestimation was consistent across taxonomic groups (we checked by inspecting dominant taxa's per-taxon size spectra), the community composition is little affected by the absolute concentrations underestimation. Even for > 1 mm organisms, UVP5 images' taxonomic identification is difficult. We applied (above) different measures for taxa identification consistency (e.g., using a taxonomic guide, cross-checking among operators and regrouping organisms at coarse taxonomic level, where confusions are rarer). Finally, the same operator reviewed a random images

Taxonomic identifications were good enough.

subsample from each final taxonomic group, with error rates at < 10% for all groups and < 2.5% for most [111].

4.4.1.3 *Sampling effort heterogeneity*

The heterogeneous profile distribution did not affect plankton community patterns.

UVP5 profiles were distributed unevenly: some areas sampled intensively (California Current, Peruvian upwelling, Mediterranean Sea), others rarely visited (Indian Ocean, Southern Ocean). For results to not be solely representative of oversampled areas, we randomly downsampled the dataset to contain a maximum of 50 profiles per 2° square and conducted the same analyses again. Plankton community patterns emerging from these analyses were similar to those conducted with all profiles, thus showing they are robust. Furthermore, our analysis does not explicitly consider location or date, only each sample's community and environmental conditions; so, this relevant question: does UVP sampling cover environmental conditions representative of global scale variance? For this, we compared conditions distribution at UVP samples' locations to the same variables distribution at global scale. Of course, simultaneous worldwide *in situ* observations are not available. Instead, we used annual climatologies on a 1° grid from World Ocean Atlas (WOA) [53] for important water characteristics: temperature, salinity, and oxygen. We first checked that those climatologies were representative of the *in situ* conditions at locations sampled by UVP5, over the epipelagic and mesopelagic layers previously defined; this was the case since correlations were good (all $R^2 > 0.84$, except for AOU in the epipelagic: $R^2 = 0.35$, Figure S4.5). Then, we compared each variable distributions from the WOA data at UVP5 profiles' locations vs. worldwide (Figure S4.6), for two depth layers (0-200 m; 200-500 m), since the above dynamic boundary could not be computed from WOA data. Distributions were similar, showing UVP samples covered diverse enough environmental conditions, representative of worldwide oceans.

4.4.2 Plankton communities general structure

4.4.2.1 *Epipelagic layer*

Three types of epipelagic plankton communities.

Three plankton communities types emerged in the epipelagic layer, mostly driven by water masses' temperature and trophic statuses: copepod-dominated communities in cold and productive waters; *Tri-*

chodesmium-dominated communities in warm waters; and mixed-type communities in oligotrophic ones.

Copepods' large proportions in almost all pelagic ecosystems is already documented [55, 188, 340, 374], more so in the rich and productive Arctic waters [55, 111, 134, 374, 400]. Similarly to other zooplankton, copepods' diversity decreases with latitude [55, 188] and with size increases [63, 182]. Brandão et al. [55] found a 400-500 µm median ESD for copepods between 60°N and 60°S. UVP5 only detects copepods over 1 mm in ESD [133]. Tropical and temperate copepods' concentrations were likely underestimated but (see above) in various situations, conclusions on community composition are little affected, since all taxa concentrations are similarly underestimated near UVP target size range limits. Large changes in the biogeography, community composition and diversity of calanoid copepods in the North Atlantic Ocean are detected, as a result of global warming [20, 21]. As copepods act as a trophic link between primary producers and higher trophic levels [340], these changes might prove detrimental to marine resources, like exploited fish stocks [21]. We show that, within the 600µm range to a few cm, copepods largely dominate communities in polar waters and, to a lesser extent, in temperate ones, but clearly not in tropical ones.

Copepod-dominated in productive environments.

Epipelagic plankton communities were also shaped by *Trichodesmium*, a filamentous cyanobacteria found in (sub)tropical regions [66, 415] and previously observed with UVP5 in tropical waters [164, 346]. *Trichodesmium* contribute to primary production and fixation of atmospheric nitrogen [66] and can grow in nitrogen-limited environments, unlike other phytoplankton types [415]. Their toxicity to several zooplankton species [171] explain the quasi-exclusion of other types of planktonic organisms in *Trichodesmium*-dominated communities.

Trichodesmium-driven in oligotrophic waters.

Finally, the third cluster revealed a mixed plankton community, characterised by copepods' or *Trichodesmium*'s non-dominance. This cluster contained profiles from diverse environments with varying conditions, thus very diversified composition-wise. Its average composition (Figure 4.3) is skewed by high concentrations from a few profiles, therefore not representative of every profile' s composition aggregated in this cluster. Although Collodaria emerged as a structuring group in PCA, it did not dominate the cluster's relative composition. Actually, six groups were found, accounting for 85% of the composition: Acantharia (Rhizaria), Collodaria (Rhizaria), Copepoda, Nostocales, Phaeodaria

Rhizaria-dominated elsewhere.

(Rhizaria) and *Trichodesmium*, highlighting the importance of various rhizarian groups, recently highlighted by other *in situ* imaging-based studies [42, 104]. Acantharia and Collodaria are symbionts-bearing Rhizaria, widely distributed in oceans, but more abundant in tropical oligotrophic surface waters, where their symbionts are photosynthetically active [393], hence coherent with our observations. Collodaria contributes significantly to the total organic matter in these environments [42, 393]. They can also form colonies, with tens to thousands of cells embedded in a gelatinous matrix [393], but these colonial forms are not as structuring as the solitary individuals for community composition. Conversely, Phaeodaria are heterotrophic Rhizaria, lacking symbionts [210], flux-feeders and thus usually found below the epipelagic layer, where they feed on sinking particles [40]. However, some species are common in surface layers [286]. Because Phaeodaria's mineral skeleton is made of silica, they can act as major biogenic silica exporters [39] and their distribution could be restricted by silica availability [40, 286]. Finally, nostocales (*Aphanizomenon*, *Dolichospermum*, and *Nodularia*) were identified contextually (i.e., not just based on aspect on images), only in the Baltic Sea, and at very high concentrations. As a consequence, they account for 19% of the mixed-type community, though they were found in only a few profiles.

4.4.2.2 *Mesopelagic layer*

Three types of mesopelagic plankton communities.

In the mesopelagic layer, three types of plankton communities also emerged: Phaeodaria-dominated in cold and oxygen-depleted waters, copepod-dominated in cold and oxygenated waters, and, again, a mixed community in warmer waters.

A mixed-type community.

In the latter, four groups accounted for 65% of organisms: Copepoda, Eumalacostraca, Foraminifera, and *Trichodesmium*. Foraminifera, heterotrophic rhizarians, feed on mesopelagic plankton; typical of deep and rather poorly oxygenated waters [40]. Various organisms were identified within the broad Eumalacostraca group in our dataset, depending on the ecosystem sampled, resulting in a very heterogeneous group. Thus, Eumalacostraca are not representative of any typical environment, restricting ecological interpretations, and most likely under-sampled.

In cold waters, the plankton community type was linked to oxygen availability: copepods dominated when oxygen was available and Phaeodaria in water masses, with high AOU and low oxygen concentra-

tion, typical of Oxygen Minimum Zones. Major OMZs are found below eastern-boundary upwelling systems, with particularly high primary production. Many planktonic taxa' concentrations, such as calanoid copepods', is reduced in OMZs [14, 184, 203, 374]. Associated with the low oxygen and low plankton concentrations, and possibly caused by it, the downward flux of particulate carbon is intense and less attenuated than elsewhere, resulting in high vertical export of the organic matter photosynthesis produces at the surface [71, 121]. Conversely, Phaeodaria are typical of deep, low oxygenated waters [40, 191] and already detected in OMZs [184]. Indeed, protists, like Phaeodaria, might prove more tolerant to hypoxia, as their passive feeding mode requires less oxygen than active feeding. In OMZs, they may therefore play a disproportionate role in the regulation of the vertical flux compared to elsewhere.

Oxygen availability may constrain mesopelagic plankton community type.

Although unexpected at first glance, the detection of *Trichodesmium* in the mesopelagic layer is consistent with previous observations [29, 360, 410]. Here, the presence of *Trichodesmium* in the mesopelagic layer could be partly explained by our dynamic definition of the epi-mesopelagic boundary: it started at shallower than 50 m for 25% of profiles, a depth where the presence of *Trichodesmium* would not be surprising. However, those were found mostly at higher latitudes (Figure S4.2), where *Trichodesmium* is absent in the epilagic layer, too. Its presence at a great depth could also result from downwelling and subduction events, bringing colonies to deeper waters [164], or even simply represent dead colonies sinking down.

Trichodesmium was detected in the mesopelagic layer.

4.4.2.3 *Comparison between epipelagic and mesopelagic plankton communities*

Conditions in the epipelagic layer impact those in the mesopelagic: i.e., the epipelagic phytoplankton type influences particle sizes in the mesopelagic [163]. Similarly, mesopelagic zooplankton biomass is conditioned by the net primary production in the euphotic layer, since it feeds on its remnants [178]. However, the results above show a low similarity of plankton communities between epipelagic and mesopelagic layers.

Epipelagic communities should impact mesopelagic communities.

Where the plankton community is dominated by copepods and the environment productive in the epipelagic layer, the mesopelagic plankton community is usually copepod-dominated as well. This is consistent with high secondary production in the mesopelagic, below

Epipelagic and mesopelagic plankton communities had contrasted distribution patterns...

subpolar surface waters hosting high primary production, which is not true at subtropical latitudes [335]. Below oligotrophic Rhizaria-dominated epipelagic communities, the mesopelagic community was mostly Phaeodaria-dominated in the eastern tropical South Pacific OMZs but was copepod-dominated in the South Atlantic. The split between these two mesopelagic communities therefore seems to be driven by copepods' oxygen limitation in OMZs [121]. In the South Atlantic gyre, both epipelagic [366] and mesopelagic [392] layers are considered as oligotrophic, consistent with a Rhizaria-dominated community in the epipelagic [393]. In the Peruvian upwelling system, the epipelagic community was mostly Rhizaria-dominated; surprising, since this environment is very productive [15]. Still, the presence of Rhizaria was already reported there by Santander Bueno [348], within a diverse zooplankton community over the region. This productive upwelling area drives an OMZ in deeper waters [121, 204], which imposes a Phaeodaria-dominated community in the mesopelagic. Within *Trichodesmium*-dominated stations, only equatorial Atlantic stations were sampled deep enough to be included in the epi/mesopelagic layers analysis. As *Trichodesmium* was almost absent from the mesopelagic layer there, profiles that were *Trichodesmium*-dominated in the epipelagic had to be distributed between the other two in the mesopelagic plankton communities, and most seen as Phaeodaria-dominated.

...but this likely results from differences in forcing factors between these layers.

The forcing environmental conditions associated with communities in both layers are not the same: oxygen plays a role at depth but less near the surface; light structures live near the surface but not in the mesopelagic layer, etc. Besides, the conditions that remain structuring ones are less variable with increasing depth, leading to a more homogeneous habitat with depth [82, 84]. This also shapes the plankton community, which becomes less heterogeneous spatially. This is consistent with Soviadan et al.'s results: mesopelagic plankton communities are less spatially contrasted than epipelagic ones.

4.4.3 Plankton communities distribution was driven by regional conditions

Among the regionalisations tested, the plankton communities distribution was better explained by Longhurst provinces [245], a regionalisation based on physical forcings as phytoplankton distribution regulators, which might drive zooplankton communities, too. The

regionalisation based on 10° latitudinal bands explained less variance, though a latitudinal effect, caused by light availability and temperature, was previously demonstrated on both plankton diversity and biomass [190, 340]. In the mesopelagic, the regionalisation computed by Reygondeau et al. [331] – specifically for this layer and mainly based on annual biogeochemical variables climatologies – did not explain plankton community distribution better than Longhurst provinces, even though it was supposed to be more appropriate. Yet, more importantly, all these basin-scale regionalisations explained plankton community distribution as well, or better, than a regionalisation based on local conditions, at sampling time. This suggests the plankton communities' spatial structure worldwide for our ~30 taxonomic groups, is driven by regional environmental conditions more than by very local and immediate conditions and processes. These findings agree with those of Stemmann et al. [385], who showed that mesopelagic macro-zooplankton communities were structured by large, basin-scale processes. Genomic analyses also underlined regional scales processes' importance in structuring plankton communities [333], especially for meso-zooplankton [369].

Large-scale processes were more important than local processes to drive plankton communities.

4.5 Conclusion and perspectives

In both layers, three plankton communities types emerged and seemed mostly driven by basin-level environmental conditions. Following on studies investigating plankton distribution and diversity across life kingdoms – from viruses to metazoans [105, 188, 390] – this work highlights the role not only of metazoans, but also of unexpected large protists and cyanobacteria in structuring over 600 µm plankton organisms' communities. This confirms underwater imaging relevance to reveal the importance of otherwise overlooked plankton groups, such as Rhizaria [42, 104].

Plankton communities were structured by diverse and unexpected groups.

Wide-ranging offshore areas regionalisations are highly desired for conservation purposes, like the creation of protected marine areas. However, they should not be restricted to oceans' upper layers. Indeed, biological activity in the mesopelagic layer is key to mediate the flux of organic carbon from the surface; the deep seafloor is also heavily impacted by human activities (bottom fishing, waste disposal, oil drilling or seafloor mining) [414]. Therefore, oceans' deeper layers

Deeper regionalisations are also desirable.

biogeographies, currently rare, are also required to balance between human exploitation and ecological conservation.

Author contributions

TP, JOI and LS conceived the study; MB, TB, FC, LC, LG, HH, LKB, RK, FL, AMD, MP, AR, AW, LS contributed to data acquisition and image validation; TP conducted the analyses under the supervision of JOI and LS; TP wrote the manuscript; all authors contributed to the discussion of the results, supported manuscript preparation and approved the final submitted manuscript.

Acknowledgements

We are grateful for the ship time provided by the respective institutions and programs and we would like to thank all scientists, officers and crew of the various R/V which took part in UVP5 data collection. We are thankful to Léo Lacour for providing satellite data and to Laetitia Drago for taxonomy checks. GlobColour data (http://globcolour.info) used in this study has been developed, validated, and distributed by ACRI-ST, France. TP's doctoral fellowship was granted by the French Ministry of Higher Education, Research and Innovation (3500/2019). JOI acknowledges support by the Belmont Forum grant ANR-18-BELM-0003-01. RK and LS received support from the European Union project TRIATLAS (European Union Horizon 2020 program, grant agreement 817578). RK furthermore acknowledges support via a Make Our Planet Great Again grant from the French National Research Agency (ANR) within the Programme d'Investissements d'Avenir ANR-19-MPGA-0012 and funding from the Heisenberg Programme of the German Science Foundation KI 1387/5-1. AR was funded by the PACES II (Polar Regions and Coasts in a Changing Earth System) program of the Helmholtz Association and the INSPIRES program of the Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research. LS was supported by the CNRS/Sorbonne University Chair VISION to initiate the global observation.

Supplementary materials

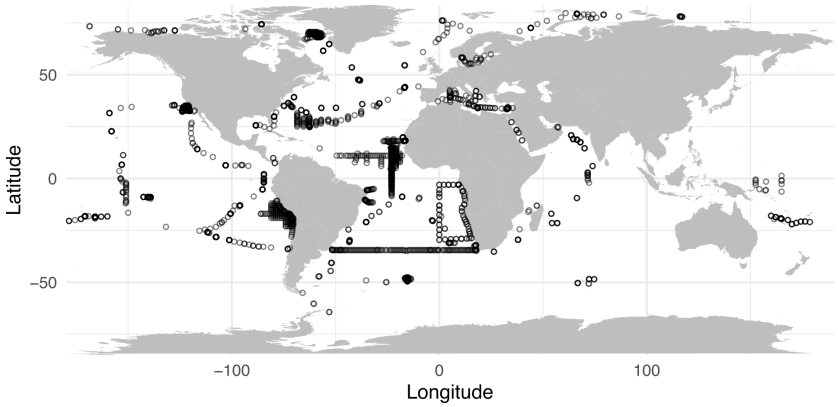


Figure S4.1: World map of included stations (whether in the epipelagic or mesopelagic layer).

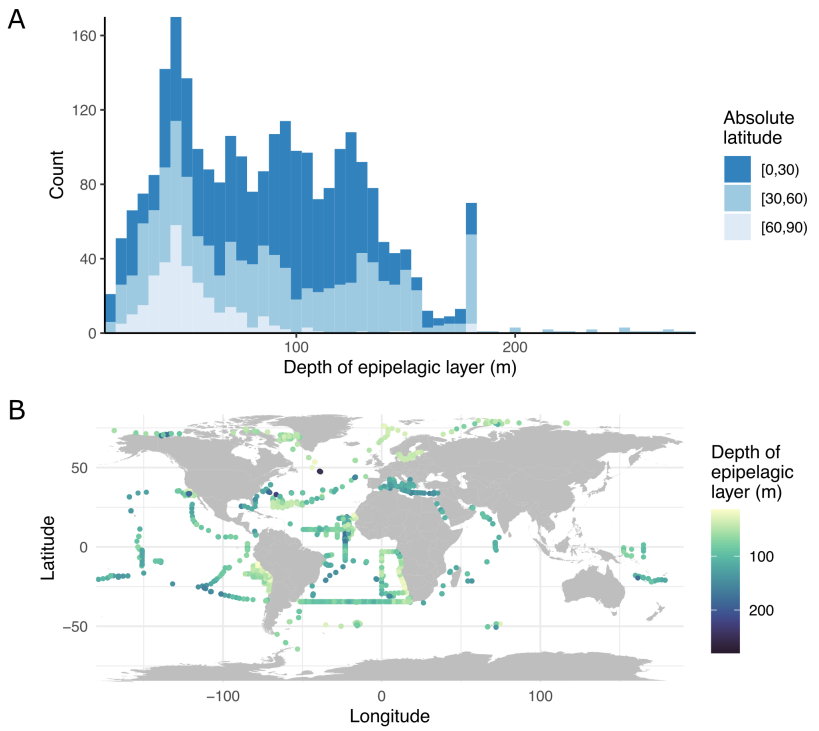


Figure S4.2: Depth of dynamic the epi-mesopelagic boundary, computed as the deepest value among the mixed layer depth and the euphotic depth. **(A)** Histogram of the epipelagic layer depth per 30° of absolute latitude bands. The peak at 180 m highlights cases of euphotic depth at 180 m and shallower mixed layer depth. **(B)** World map of the epipelagic layer depth.

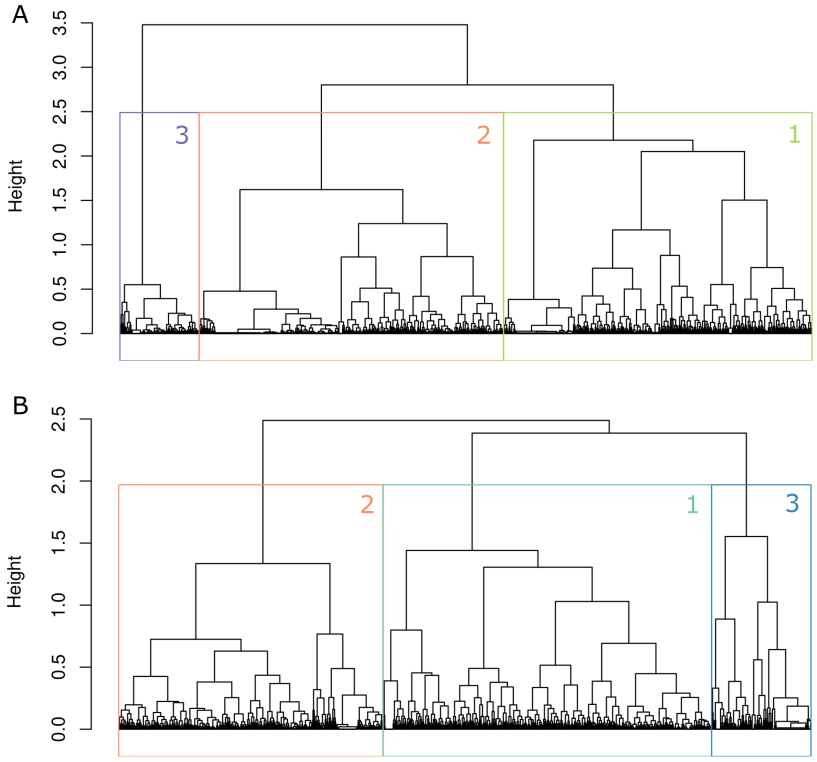


Figure S4.3: HAC dendrograms based on the first five principal components of profiles projection in the Hellinger's transformed plankton PCA data, for **(A)** epipelagic and **(B)** mesopelagic layers. Generated clusters are shown in the same colours and numbers as they appear on figures 4.3, 4.4, 4.5

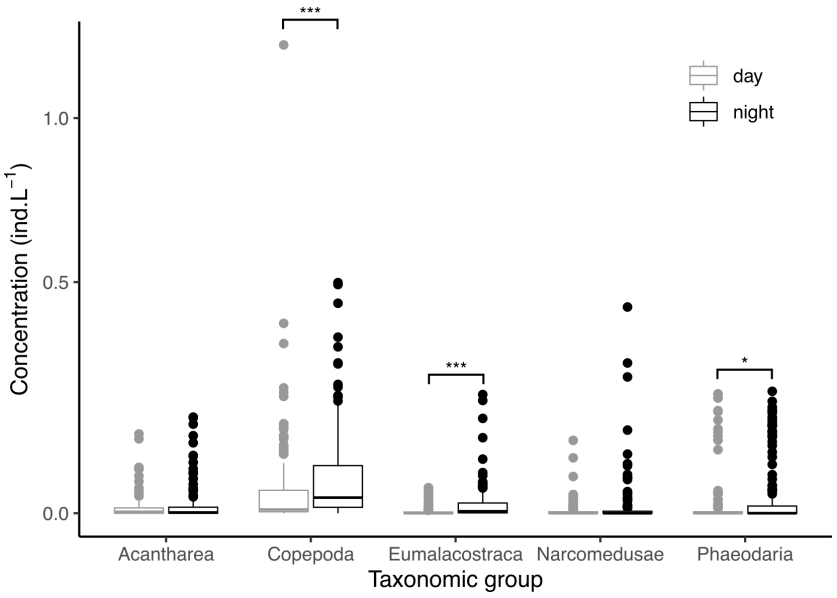


Figure S4.4: Average epipelagic concentration of the five most abundant taxa in California Current by day and by night. Note that the y axis is log-transformed. Differences were tested with a Wilcoxon-Mann-Whitney test. * = 0.05, *** = 0.001.

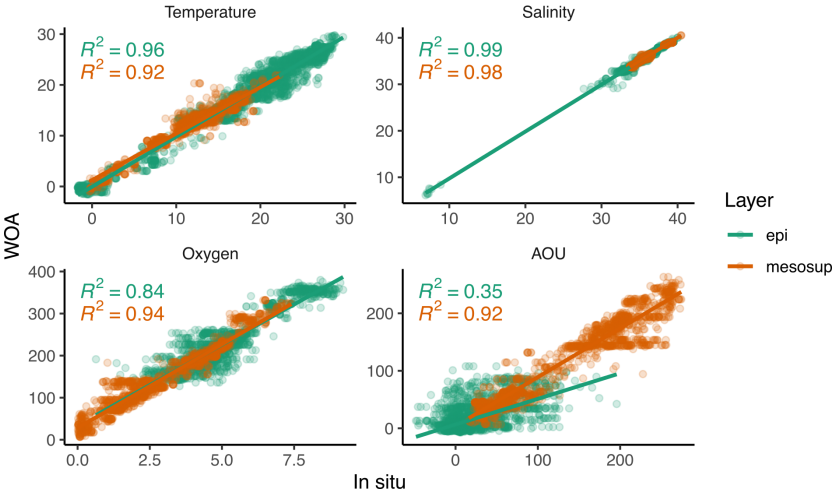


Figure S4.5: Correlation between *in situ* and annual WOA data at UVP5 profiles locations in the epipelagic and mesopelagic layers.

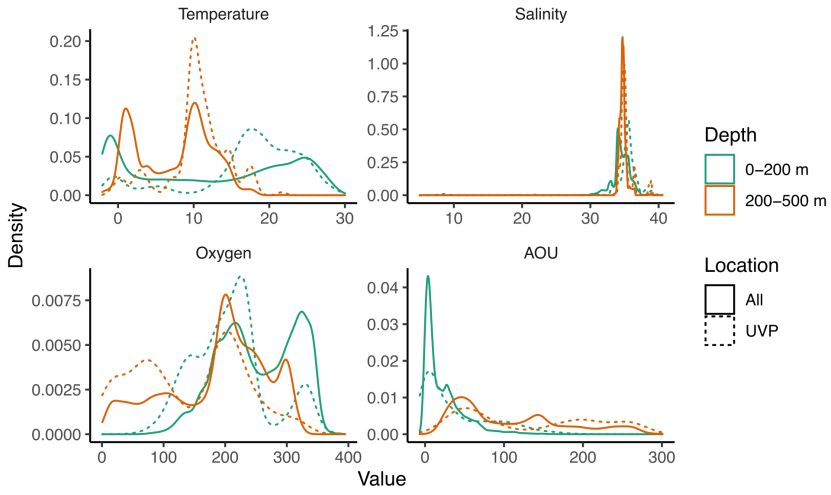


Figure S4.6: Distribution of annual WOA data all over the globe and at UVP5 profiles locations.

Table S4.1: List of oceanographic campaigns included in the study.

Campaign	Year	Nb profiles	UVP5
BOUM	2008	177	sd
CASSIOPEE	2015	13	sd
CCELTER 2008	2008	73	sd
CCELTER 2011	2011	56	zd
CCELTER 2012	2012	59	sd
CCELTER 2014	2014	60	sd
CCELTER 2017	2017	68	hd
DEWEX	2013	1	sd
MSM22	2012	101	sd
MSM23	2012	64	sd
M105	2014	8	sd
M106	2014	114	sd
M107	2014	71	sd
PS88b	2014	36	sd
M116	2015	74	sd
M121	2015	84	sd
M135	2017	138	sd
GreenEdge 2016	2016	121	hd
MSM060	2017	126	hd
IPS Amundsen 2018	2018	6	sd
JERICO 2017	2017	24	sd
KEOPS	2011	13	zd
LOHAFEX	2009	55	sd
MALINA	2009	16	sd
MooseGE	2015	3	sd
NAAMES02	2016	21	hd
OUTPACE	2015	193	sd
P16N	2015	14	sd
Sargasso	2014	84	sd
SOMBA	2014	6	sd
Tara Oceans	2009-2013	643	sd

Table S4.2: Definition of productive (1) and non-productive (0) seasons based on latitude and month.

Latitude band	Month											
	J	F	M	A	M	J	J	A	S	O	N	D
90°N - 66.5°N	0	0	0	0	0	1	1	1	0	0	0	0
66.5°N - 23.5°N	0	0	1	1	1	0	0	0	1	1	0	0
23.5°N - 23.5°S	0	0	0	0	0	0	0	0	0	0	0	0
23.5°S - 66.5°S	0	0	1	1	0	0	0	0	1	1	1	0
66.5°S - 90°S	1	1	0	0	0	0	0	0	0	0	0	1

This model is based on light intensity and nutrients availability. In polar regions, light availability is often limited (namely in winter) but becomes sufficient after the summer ice breakup, allowing productivity. In mid-latitudes, both light and nutrients become available in spring and autumn, generating phytoplankton blooms. In tropical regions, productivity is limited all year by nutrients and remains low.

5

Temporal evolution of particles and plankton distributions across a mesoscale front during the spring bloom

Following the global scale study, we investigated plankton distribution at the mesoscale during the spring bloom, across the Ligurian front. We conducted the very first continuous, multi-month deployment of a glider equipped with an imaging device, namely the UVP6, and collected 1,123,123 images. During the 5 months of the campaign, I personally contributed to glider deployment and retrieval, as well as glider piloting, in the form of on-call duty of 3 days alternating with another person, supported by on-call pilots from the Alseamer company if needed. Overall, this experience was both very demanding and rewarding.

The first success lies in the fact that we had no material damage and that we did sample the Ligurian front for several months in a row, after only one test deployment conducted a few months before. The second achievement is that we collected enough data to resolve the distribution of plankton and particles across the front, which is the focus of this chapter. Again, machine learning methods were used to facilitate image classification, while data mining approaches allowed to extract ecological knowledge from the data.

Thelma Panaïotis, Antoine Poteau, Émilie Diamond-Riquier, Lucas Courchet, Camille Catalano, Laurent Coppola, Marc Picheral and Jean-Olivier Irisson

Manuscript in preparation to be submitted to **Limnology and Oceanography**

Abstract

The effect of mesoscale features on the distribution of planktonic organisms are well documented. Yet, the interaction between these spatial features and the temporal scale, which can result in sudden increases of the planktonic biomass, is less known and not described at high resolution.

We targeted a permanent mesoscale front in the NW Mediterranean Sea that we repeatedly sampled between January and June 2021 using a glider equipped with a UVP6, a versatile *in situ* imager. We aimed to resolve mesoscale distribution of plankton and particle distribution during the spring bloom, to assess whether the front was a location of increased concentration for zooplanktonic organisms during the bloom, and if it constrained the distribution of particles. During the 5 months, the glider conducted more than 5,000 dives and the UVP6 collected 1.1 million images. We focused our analysis on shallow (300 m) transects, which gave a horizontal resolution of 900 m. Images were sorted manually, and predicted with a machine learning algorithm. In the end, about 13,000 images of planktonic organisms were retained.

Ordination methods applied to particles and plankton concentrations revealed contrasted periods during the bloom, in which changes in particle abundance and size could be explained by changes in the plankton community. The front had a strong influence on particle distribution, while the signal was not as clear for plankton, probably because of the relatively small number of organisms imaged. In addition, we also detected submesoscale features such as subduction events and submesoscale coherent vortices. This work confirms the need to sample both plankton and particles at fine scale to understand their interaction, a task for which *in situ* imaging is particularly adapted.

Résumé

L'effet des dynamiques à mésoéchelle sur la distribution des organismes planctoniques est relativement bien documenté. Cependant, l'interaction entre ces dynamiques spatiales et l'échelle temporelle, qui peut entraîner des augmentations soudaines de la biomasse planctonique, est moins connue et encore moins décrite à haute résolution.

Nous avons étudié un front permanent de méso-échelle dans le nord-ouest de la mer Méditerranée. Ce front a été échantillonné de manière

répétée entre janvier et juin 2021 en utilisant un planeur équipé d'un UVP6, un imageur *in situ* polyvalent. Nous nous sommes efforcés de décrire la distribution à méso-échelle du plancton et des particules pendant le bloom de printemps, afin d'évaluer si le front était un lieu de concentration accrue pour les organismes zooplanctoniques, et si cette structure contraint la distribution des particules.. Pendant ces 5 mois, le planeur a effectué plus de 5 000 plongées et l'UVP6 a collecté 1,1 million d'images. Nous avons concentré notre analyse sur les transects peu profonds (300 m), avec une résolution horizontale de 900 m. Certaines images ont été triées manuellement, et d'autres prédites avec un algorithme d'apprentissage automatique. Au final, environ 13 000 images d'organismes planctoniques ont été retenues.

Des méthodes statistiques d'ordination ont révélé des périodes contrastées pendant le bloom, au cours desquelles les changements dans l'abondance et la taille des particules pouvaient être expliqués par les changements dans la communauté planctonique. Le front a eu une forte influence sur la distribution des particules, alors que le signal n'était pas aussi clair pour le plancton, probablement en raison du nombre relativement faible d'organismes imagés. En outre, nous avons également détecté des structures à subméso-échelle telles que des événements de subduction et des tourbillons cohérents de subméso-échelle. Ce travail confirme la nécessité d'échantillonner à la fois le plancton et les particules à fine échelle pour comprendre leur interaction, une tâche pour laquelle l'imagerie *in situ* est particulièrement adaptée.

5.1 Introduction

5.1.1 Particles, plankton and blooms

As drifters, planktonic organisms are drifters and are thus strongly affected by the conditions of the water mass they are embedded in [172]. Therefore, both phytoplankton and zooplankton concentrations vary locally and seasonally. These strong increases in productivity and plankton concentration are called "blooming" phases [25]. Phytoplankton blooms typically occur at the end of winter, when hivernal convection, which ensures the replenishment of nutrients into the surface waters, stops and thermal stratification starts to settle [24, 417]. For zooplankton, blooms may occur through organisms aggregation or in response to favourable conditions enabling an increased growth

Blooms are temporal increases of plankton productivity and biomass.

rate [150]. In the latter case, the zooplankton bloom takes place slightly later than the phytoplankton bloom [174]. Plankton blooms also impact the concentration, composition and morphology of marine snow particles [401]. These aggregates are formed through a combination of physical coagulation and zooplankton-mediated processes [64, 207]. Physical coagulation requires collision between particles, resulting from brownian motion, differences in sinking velocity or fluid shear depending on the size of particles. Zooplankton-mediated processes include the production of faecal pellets (which are aggregates of consumed particles) and mucus feeding webs (e.g. Appendicularia houses). Thus, temporal processes shape both plankton and particle distribution, but those hemisphere-wide processes are often interacting with local physical structures at mesoscale.

5.1.2 Frontal processes

Fronts are zones of increased plankton productivity and biomass...

As explained in section 1.1.1 of the introduction, physical processes drive movements at a large number of scales in the oceans, from centimetres to thousands of kilometres [106, 225] and these processes shape the structure of planktonic ecosystems [103, 170, 225, 251]. More specifically, fronts and eddies are physical features influencing the distribution of plankton [319] and particles [404]. Fronts are oceanic features separating masses of waters with different properties. They come in a wide variation of spatio-temporal scales: from a few hundreds of metres to tens of kilometres, some are ephemeral while some are permanent [130, 230, 304]. Frontal zones are oceanic hotspots, with increased productivity and biomass across all trophic levels: phytoplankton, zooplankton and even top predators [230, 265, 304].

...due to by submesoscale dynamics.

In a frontal-jet system, i.e. a relatively steady geostrophic current associated with an along-current frontal structure and horizontal density gradients [290], the frontal structure can generate a submesoscale cross-frontal ageostrophic circulation directed in the sense of flattening the isopycnals [230]. Thus, enhanced plankton biomass at the front can result either from passive aggregation of planktonic organisms [150] or active mechanisms, such as an increased growth rate caused by the redistribution of nutrients induced by the cross-frontal recirculation [230].

5.1.3 The Ligurian frontal-jet system

5.1.3.1 *Physics*

The Ligurian frontal-jet system: a long-standing case study.

In this study, we focus on the northern frontal-jet system of the Ligurian Sea (NW Mediterranean). The northern current is a geostrophic jet flowing along the coast from NW to SE, located about 20 km offshore and stronger between the surface and 150 metres, with an average speed of 30-50 cm s^{-1} [290, 326]. A permanent mesoscale front – the Ligurian front – is associated with this current. The front, delineated by the 38.2 and 38.8 isohalines [49] and going as deep as 200 m [290], separates offshore colder and saltier waters from coastal warmer and fresher water. Three zones can thus be distinguished: coastal, frontal and central [326]. The Ligurian front and jet meander between 15 and 50 km away from the coast, moving at a speed of approximately 8 km per day [322]. Since planktonic organisms are unable to swim against currents, their distribution should be impacted by these physical features. Resolving such mesoscale distribution of planktonic organisms and physical properties during the plankton bloom requires to sample at high spatial resolution for several months, which is not achievable with ship-based sampling. Autonomous underwater vehicles, such as gliders, can be deployed and sample continuously for weeks, resolve submesoscale hydrologic features and integrate diverse miniaturised sensors [342]. They thus seem to meet the requirements of such studies.

5.1.3.2 *Plankton distribution across the front*

Plankton distribution is constrained by the front,...

Numerous studies have focused on plankton distribution across the Ligurian frontal-jet system. Early studies highlighted the relation between the spatial distribution of planktonic organisms and the physical structure of the front [48, 49]. More specifically, the front seems to act as a barrier, constraining organisms whether in coastal or offshore waters [124, 219, 275, 311]. The front also seems to shape the distribution of marine snow aggregates by acting as a barrier, constraining the distribution of aggregates on the coastal side of the front [148, 384]. Finally, regarding vertical distributions, copepod communities strongly vary between stratum [141]; and the distribution of pelagic tunicates may reflect the junction of frontal convection cells [147].

5.1.3.3 *Plankton blooms in the Ligurian sea*

and by seasonal dynamics.

Within the Mediterranean Sea, the Ligurian Sea is one of the locations hosting a phytoplankton spring bloom and sometimes of a less intense autumn bloom too [91, 262, 263]. The phytoplankton bloom later propagates to higher trophic levels with larger concentrations of zooplankton, such as copepods [110] or salps [292].

Frontal submesoscale dynamics constrain phytoplankton, . . .

Besides, evidence suggests that the phytoplankton bloom is influenced by the frontal features previously described. Goffart, Hecq, and Prieur [144] found that the bloom was more intense at the front and that phytoplankton is transported downward by the frontal convergence, following the isopycnals. As described above, submesoscale frontal features can influence phytoplankton distribution and growth [230] and these effects were detected by a glider survey of the Ligurian front, which showed the vertical transport of surface waters enriched in chlorophyll [290].

. . . thus zooplankton distribution should be investigated at this scale too.

Investigating how these effects propagate to higher trophic levels (e.g. zooplankton, larval fish) thus requires to study the distribution of these organisms at the same resolution as the one used for phytoplankton. This fine-scale is the one at which interactions between planktonic organisms occur, as well as interactions with their physico-chemical environment. A few studies already targeted zooplankton distribution at such scales: both Luo et al. [248] showed that the distribution of various groups of gelatinous organisms across a mesoscale front was driven by temperature, depth, oxygen or chlorophyll a concentration depending on the group. Similarly, Greer et al. [154] demonstrated by both zooplankton and larval fish were more abundant on the shelf side of a shelf-slope front. Finally, Greer et al. [156] revealed the interaction between phytoplankton fine layers and zooplanktonic grazers. The common denominator between these studies is that they rely on *in situ* imaging.

5.1.4 Fine-scale plankton distribution through *in situ* imaging

As explained in section 1.2.3, traditional plankton sampling instruments (nets and pumps) are not adapted to resolve fine-scale distribution of planktonic organisms in relation to their immediate environment. They lack spatio-temporal resolution because the organisms collected are integrated over the distance and/or depth of the tow. Furthermore, plankton sampling is often performed separately from the record-

ing of environmental data [85, 243]. In addition, fragile organisms can be damaged during sampling and thus their concentration underestimated [330]. The development of *in situ* imaging instruments overcame some of these limitations: they resolve the fine-scale distribution of planktonic organisms within their environment and allow to detect physico-biological relationships. As no collection *per se* is performed, organisms integrity is preserved and *in situ* behaviour can be observed [258, 299, 409]. This comes at the cost of some taxonomic resolution, since each organism cannot be manipulated and of a large data processing effort. The Underwater Vision Profiler 6 (UVP6) [318] is one of these *in situ* imagers. It counts particles from 70 µm and images marine snow and plankton larger than 1 mm (meso and megaplankton), thus mostly consisting of zooplankton and a few large phytoplankton colonies. Compared to the previous generation (UVP5) [317], the UVP6 is smaller and can be deployed on autonomous vectors (e.g. float, glider).

In situ imaging resolves fine scale plankton distribution.

5.1.5 Aim of this study

In this work, we leverage in situ imaging using a glider in order to (i) examine mesoscale and submesoscale physical and biological properties of the Ligurian front and (ii) investigate the dynamics of particles and plankton distribution across the front during the spring bloom of 2021.

5.2 Materials and Methods

5.2.1 Glider and UVP6

Sampling was performed with a Seaexplorer (Alseamar) glider, an unpropelled autonomous underwater vehicle (AUV) taking advantage of buoyancy variations to glide forward through the water in a sawtooth-like pattern, periodically coming to the surface between dives to ensure data transmission and GPS positioning. The glider was fitted with a set of sensors to record temperature, salinity, fluorescence, particles backscattering at 700 nm (BB700), colour dissolved organic matter (CDOM) and dissolved oxygen concentration. Besides environmental sensors, the glider was equipped with a UVP6-LP, consisting of a main camera and a light unit illuminating a slice of water for an

The UVP6 was deployed on a SeaExplorer glider.

image volume of 0.7 L, with an imaging rate adaptable between 0.2 and 1.3 Hz [318]. With a pixel size of 73 µm, the UVP6-LP images objects (plankton and marine snow particles) between 1 mm and 2 cm and counts particles in 28 size classes between 102 µm and 2.6 cm in equivalent spherical diameter (ESD). When embedded on a Seaexplorer glider, the UVP6-LP is deployed in supervised mode and is completely piloted by the glider, from power up to image acquisition.

5.2.2 Mission design

The mission targeted the Ligurian front during the spring bloom.

The glider was deployed outside of Villefranche Bay (43°39′18″N, 7°17′24″E, referred to as "coast" thereafter) and headed towards the Dyfamed point (43°22′02″N, 7°55′59″E). Sampling consisted of repeated transects, crossing nearly perpendicularly the Ligurian front. The glider performed round-trips between the coast and Dyfamed. From before (28 Jan 2021) to after (28 June 2021) the spring bloom, ten missions of 12 to 14 days each were conducted (Figure 5.1), each mission consisting of two round trips from the coast. On the way out, the glider's trajectory was turned slightly into the current. To avoid being too affected by this strong surface (<150 m depth) jet, it was set to dive down to 600 m. On the way back, the target depth was reduced to 300 m to increase horizontal resolution: the median distance between two surfacing events was 900 m. The UVP6 acquisition rate was set to 0.2 Hz below 220 m and to 0.5 Hz above 220 m (1.3 Hz for the last two missions), resulting in a sampling rate between 0.14 L s^{-1} and 0.91 L s^{-1}. During the 10 missions, the glider spent 2790 h at sea and the UVP6 captured a total of 1,123,123 images. However, 30% of these images were captured near the coast when the glider was in virtual mooring waiting to be recovered.

5.2.3 Data processing

5.2.3.1 *Positional data*

Transects were delimited.

First, each transect was separated into "out" and "back" parts based on surfacing position and only back parts (with higher horizontal resolution) were retained for this study. This resulted in a total of 20 transects (Figure 5.1). Note that as a transect lasted more than 24 hours, space and time (particularly daytime VS nighttime) are necessarily

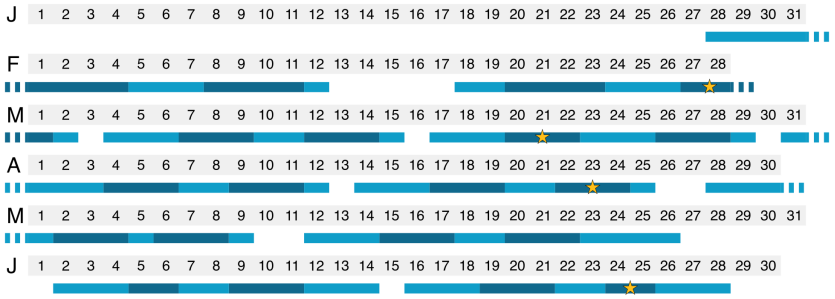


Figure 5.1: Schedule of the ten glider missions. Each mission consisted of two round trips; darker parts represent back transects on which we focused our analysis. Stars show transects selected for a synthetic visualisation in the rest of the chapter.

entangled. As GPS positioning was only available when the glider was surfacing, the geographical coordinates were linearly interpolated over each dive. Each dive consisted of a down and an upcast. Latitude and longitude were then used to compute the distance from shore (reference point at Nice cape, 43°41'9"N, 7°18'17"E).

5.2.3.2 *Plankton images*

Imaged organisms were sorted with the help of machine learning,...

Images captured by the UVP6 during cruising (n = 785,405) were first imported into the Morphocluster application [358] to quickly detect large clusters of similar objects. This allowed sorting more than 400,000 objects, mostly detritus, in a few hours. In a second step, images collected during back transects, on which we focus our study (n = 434,129), were imported onto the EcoTaxa web application [316] with their Morphocluster label in order to be sorted at a finer scale into taxonomic or morphological groups (marine snow, artefact, badfocus, reflection or unidentifiable) with the help of a supervised machine learning algorithm. Still, sorting all 400k+ images would have required a multiple months effort and we instead decided to rely on the prediction of a Random Forest classifier fed with both handcrafted and deep features generated by a MobileNet V2 feature extractor previously finetuned on UVP6 data (Chapter 3). We selected a RF classifier for the following reasons: RFs tend to produce good classification probability estimates [289], they are faster to train than a full CNN stack and, when trained with deep features, they perform as well as a full CNN (E Amblard pers. comm.).

. . . but a few groups required inspection of all images within.

Classification performance was assessed on an independent test set of 42,595 objects not used for training. This test set was built to be representative of the whole dataset by selecting 1 out of 10 profiles, equally distributed across the mission. Unsurprisingly, the test set was dominated by detritus (n > 40,000) but still contained around 100 objects or more per category for a few taxonomic groups (Copepoda, Appendicularia, Rhizaria, Salpida), allowing to compute reliable enough classification performance metrics. This revealed relatively poor precision performance for a few groups, including Appendicularia and Rhizaria (Table S5.2). To improve precision, at the expense of recall, and make sure that observed patterns were genuine, we applied a probability [124] threshold on classification scores discarding the images for which the classifier was not confident enough, in order to aim for 75% precision for all classes (Table S5.3). However, this approach strongly decreased recall for Appendicularia and Rhizaria, to such a point that it prevented the detection of any distribution pattern. We thus decided not to apply the probability threshold to these classes, instead to inspect and manually validate (n = 1500) all objects predicted as Appendicularia or Rhizaria, so that precision of these classes would reach 100% without decreasing their recall. Counts of objects per taxonomic group and per particle size class were divided by water volumes sampled to compute concentrations (# m^{-3}) within 5 m bins along the glider trajectory.

5.2.3.3 *Environmental data*

Environmental data was processed for visualisation.

After discarding abnormal values (e.g. negative fluorescence), density was computed from temperature and salinity. For each variable, outliers were detected according to a criterion based on the deviation around a moving median [231] and removed. Data was linearly interpolated at 1 m resolution, hence filling the missing values. Each variable was smoothed using a moving average within a window size of 25 m. Data was then binned over 5 m to match the plankton and particles data bins.

5.2.4 Data analysis

5.2.4.1 *Particles specificities*

Large patterns of particles...

First, particle concentrations for 13 size classes between 102 µm and 2.05 mm (classes below 102 µm were too noisy to be retained) were averaged onto 10 m × 1 km bins to reduce noise, and normalised with log-transformation to avoid very high values. A principal component analysis (PCA) was then performed to summarise these concentrations, scaled to unit variance so that each size class equally contributed to the construction of the factorial space. Supplementary variables – not used to compute the PCA space – were projected onto the PCA space to help with the interpretation. This included biogeochemical variables and metadata (coordinates, depth, distance from shore, day of year). Finally, scores of objects were visualised on transects.

5.2.4.2 *Plankton specificities*

...and plankton distribution were extracted.

A PCA was also performed to synthesise plankton concentrations. Because of the scarcity of the plankton compared to the 70 µm - 2 mm particles, plankton concentrations were first averaged across 30 m × 5 km bins, resulting in a median water volume of 373 L (Q25% = 183 L, Q75% = 573 L) per bin. Still, plankton distributions were zero-inflated and numerous bins did not contain any planktonic organisms. We thus decided to only focus on bins where plankton was present only, and assigned the average value for each variable (i.e. concentrations of plankton groups) to empty bins, so that they did not contribute to the construction of the PCA space. All concentrations were log(n+1)-transformed to normalise them and scaled to unit variance so that each taxon contributed equally to the construction of the PCA space.

Four taxonomic groups were targeted.

This PCA (Figure S5.8) highlighted the importance of a few taxonomic groups: Copepoda and Eumalacostraca emerged on PC1, Salpida, Mollusca and Appendicularia on PC2, and multiple Rhizaria subgroups on PC3. We decided to focus our analyses on four groups based on their abundance (Figure S5.7) and their importance in the PCA: Copepoda, Appendicularia, Salpida and Rhizaria (merging both Collodaria, Foraminifera and other Rhizaria since they had close projections in the PCA space).

5.2.4.3 *Visualisation*

Four transects were selected for visualisation.

Out of the 20 available transects, four were selected as representative of the dynamics and are represented in a synthetic view (Figure 5.1). All the 20 transects are available as supplementary material (Figures S5.1, S5.5 and S5.8).

All analyses were conducted with `R` version 4.1.2. Data processing and interpolations were performed with packages `dplyr` and `akima` respectively. Plots were generated with `ggplot2` using the color-blind friendly `viridis` and `cmocean` colour scales.

5.3 Results

5.3.1 Dataset composition

The dataset was dominated by Copepoda, Rhizaria, Appendicularia and Salpida.

Among the 434,129 objects imaged by the UVP6 during the back transects, 305,294 were predicted with high enough confidence to be retained after probability thresholding. After discarding marine snow particles, imaging artefacts and unidentifiable objects, 12,824 images of planktonic organisms, sorted into 10 taxonomic groups, were retained. Copepods dominated the dataset, followed by Rhizaria, Appendicularia and Salpida (Figure S5.7).

5.3.2 Environment

Environmental data successfully captured phytoplankton bloom evolution as well as submesoscale features.

During the first transect presented, which started on Feb 27th, the temperature was rather homogeneous over the water column (Figure 5.2, S5.1). At the surface, the front was located between 20 and 3 km offshore, and separated fresher waters inshore from saltier waters offshore, as expected. Fluorescence highlighted an intense bloom at the surface in offshore waters. A downwelling of chlorophyll was detected down to 300 m and seemed to follow the front. This feature was also clearly visible on salinity, oxygen and CDOM, strongly suggesting a movement of a whole water mass. Finally, inshore waters were well oxygenated down to 300 m. The second transect (March 20th) displayed similar features except for fluorescence: an important mixing event was detected, down to 200 m, where no photosynthesis can take place. This event was also visible on oxygen and CDOM concentrations. This mixing event was likely caused by strong winds

the day before (Figure S5.3). On the third transect (April 22nd), a weak stratification started to appear in coastal waters and the front became more horizontal. Fluorescence was lower at the very surface than in the waters below, hinting at the formation of a deep chlorophyll maximum (DCM). A lens of colder and fresher water was located 50 km offshore, between 150 and 300 m depth, and was also characterised by higher oxygen and lower CDOM. Finally, on June 24th, the water column was well stratified, with sharp thermocline and pycnocline. The front was nearly horizontal. The DCM was well established between 50 and 100 m (also visible on BB700), with higher chlorophyll concentration offshore. Oxygenated waters were restrained to the surface, above the DCM. These features are emblematic of the oligotrophic environment during summer in the Ligurian Sea.

5.3.3 Particles distribution

For particles, the first two axes of the PCA captured 95.5% of variance, most of it (86.3%) being on the first axis (Figure 5.3A). This axis separated high particle concentrations, associated with coastal waters characterised by high BB700, oxygen and fluorescence, from low particle concentrations, associated with offshore, deep, salty and dense waters. The second axis discriminated between size classes. The map of PC1 projections (Figure 5.3B) highlighted a decrease in particle concentration decrease over time. PC2 projections (Figure 5.3C) revealed contrasted patterns in particle size. On Feb 27th and March 20th, particles were much more abundant near the coast, and large particles, in particular, followed the downwelling event detected on fluorescence (Figure 5.2). Particles were larger offshore than inshore (Figure 5.3C). On April 22nd, particle concentration was much lower, except in very coastal waters, highlighting the near absence of particle export. Large particles were present in the 0-100 m depth layer, around the DCM that was just forming, but not below. Finally, on June 24th, the DCM was associated with abundant and small particles, likely phytoplankton cells or particles produced through biological activity; while a rain of larger particles was detected offshore, below the DCM. The circadian cycle did not seem to affect neither particle distribution nor size in any transect.

Particles concentration and size strongly varied during the bloom.

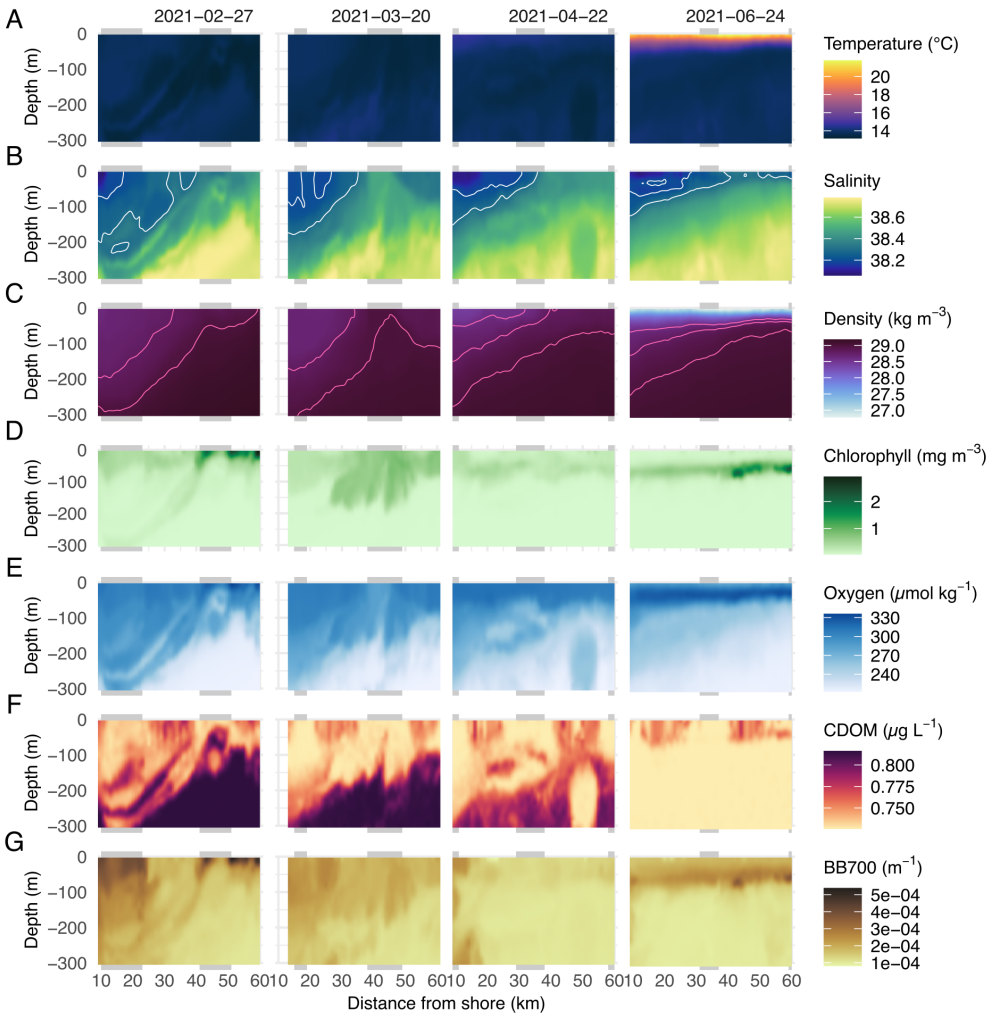


Figure 5.2: Evolution of environmental conditions across 4 transects representative of the dynamics during the bloom. Each column is labelled according to the starting date of the transect, with began offshore, on the right. Each transect lasted about 2 days and grey rectangles in the plot background represent night time. **(A)** temperature, **(B)** salinity with the 38.2 and 38.3 isohalines delimiting the front, **(C)** potential density anomaly with the 28.6, 28.8 and 29 isopycnals, **(D)** chlorophyll, **(E)** oxygen, **(F)** CDOM and **(G)** BB700. Plots for all transects are presented in Figure S5.1.

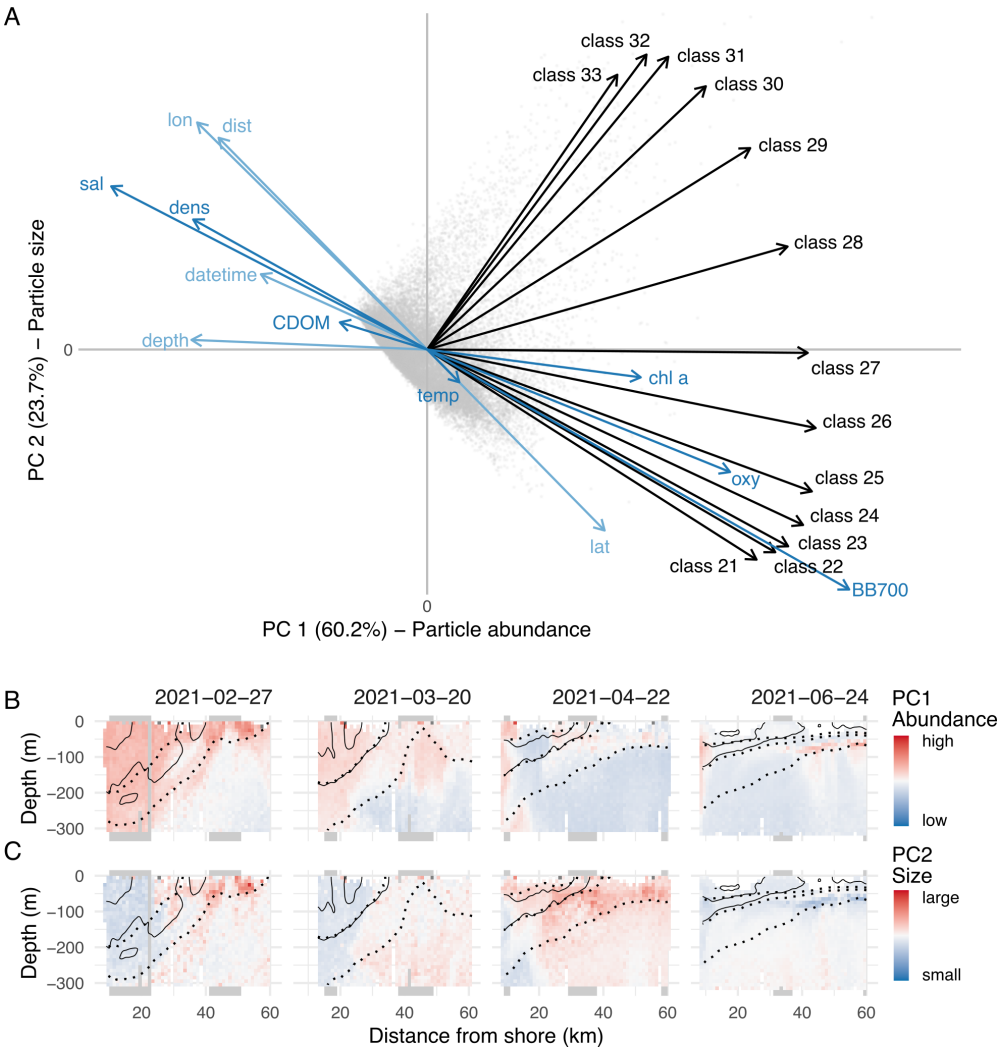


Figure 5.3: Evolution of particles distribution. **(A)** PCA performed on log-transformed particle concentrations. Definition of size classes: class 21: 102-128 µm, class 22: 128-161 µm, class 23: 161-203 µm, class 24: 203-256 µm, class 25: 256-323 µm, class 26: 323-406 µm, class 27: 406-512 µm, class 28: 512-645 µm, class 29: 645-813 µm, class 30: 813-1020 µm, class 31: 1020-1290 µm, class 32: 1290-1630 µm, class 33: 1630-2050 µm. Maps of the projections of bin scores on PC1 **(B)** and PC2 **(C)**, for the four representative transects. The 38.2 and 38.3 isohalines delimiting the front are represented as black lines. Grey rectangles in the plot background represent night time. PC1 and PC2 projections for all transects are presented in Figure S5.5.

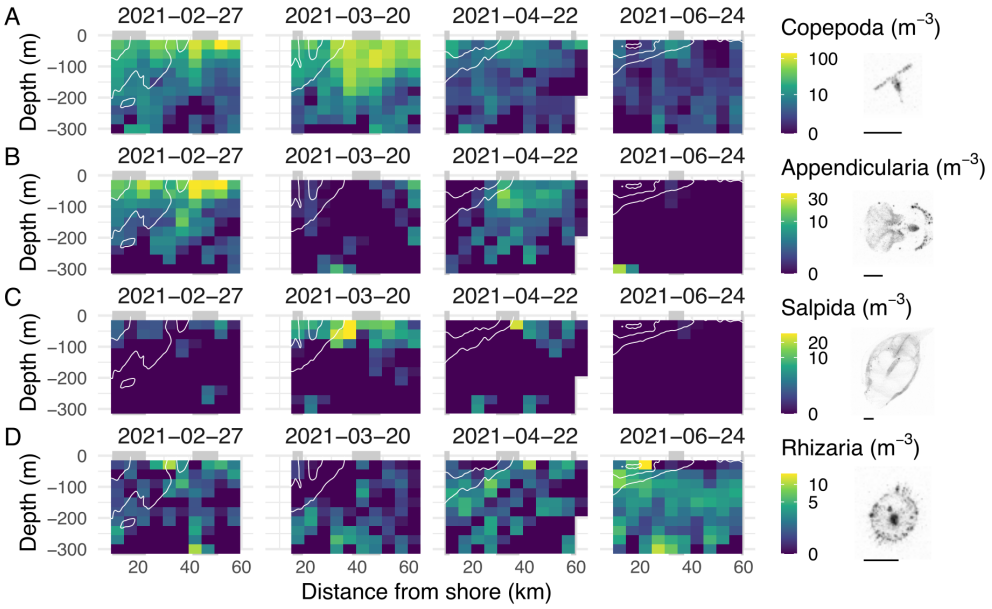


Figure 5.4: Evolution of plankton distribution for the four selected groups: **(A)** Copepoda, **(B)** Appendicularia, **(C)** Salpida and **(D)** Rhizaria. The 38.2 and 38.3 isohalines delimiting the front are represented as black lines. Grey rectangles in the plot background represent night time. Note that colour scales are log-transformed. Scale bars represent 2 mm. All transects are presented in Figure S5.9.

5.3.4 Plankton distribution

Plankton concentration revealed a succession of various zooplankton communities.

The concentration plots for the selected taxa highlighted changes in the zooplankton community (Figure 5.4). February was dominated by copepods and appendicularians close to the surface, especially offshore. In March, copepods were even more abundant, appendicularians mostly disappeared but salps were abundant at the top of the water column, particularly on the offshore side of the front. A second peak in Appendicularia abundance occurred in March, while salps were still present. In June, rhizarians became more abundant in the entire water column while copepods, appendicularians and salps were less abundant. Finally, at the considered scale, no effect of the circadian cycle could be detected.

5.4 Discussion

5.4.1 Plankton and sampled volumes

Due to the low sampling rate ($\leq$ 0.9 L s^{-1}), very few planktonic organisms were imaged and only 13,000 of them could be predicted with enough confidence to be included in our analysis. Thus, many sampled bins did not contain plankton and the signal emerging from our analyses was coarse compared to particles or environmental data. Overall, only 17,000 objects were predicted as plankton, such that manual validation of all plankton groups would have only added 4,000 organisms (+ 30%), likely not affecting the results. Although the classifier was biased towards the rare plankton classes, we cannot exclude that planktonic organisms were predicted in non-plankton classes. Checking this would have required the inspection of > 400,000 images to salvage an unknown but likely small, number of images of planktonic organisms, since the percentage of living objects is commonly < 10% in UVP datasets (Chapter 3).

We sampled too few organisms to resolve fine-scale distribution.

Although towards the end of the mission, sampling rate was increased from 0.35 to 0.91 L s^{-1} (the maximum achievable for the UVP6-LP) in the 0-220 m layer, this was still too low to investigate fine scale plankton distribution. Resolving such scales would require a UVP6-HF (high frequency) [318], but at the cost of higher energy consumption and lower glider autonomy (only one transect per mission instead of two).

5.4.2 Effects of the diel cycle

Space and time variations could not be disentangled. However, the maps of PC projections did not seem affected by day/night variations. The diel cycle seemed to have little influence; yet two tests were performed to investigate its potential effect. First, average plankton concentrations along glider yos were computed, and differences between day and night yos were investigated with a Wilcoxon test. This highlighted day/night differences in average concentration over 0-300 m for several taxa (Figure S5.10A). Then, average vertical distributions on 30 m bins were computed for day and night and compared with a Kolmogorov-Smirnov test, revealing small significant differences for Foraminifera only (Figure S5.10B). Moreover, during daytime, the UVP6

Little effect of day/night variations on our results.

was dazzled by the sunlight in the upper part of the water column (depth < 30 m) when ascending, preventing the computation of any particle or plankton concentrations in this zone. As we do not have day concentration values to compare with night values at these depths, even though many taxa are known to migrate within this depth range, we tested the relative effect of both environmental variables and a binary day / night variable on plankton concentration in the 30-60 m depth range, using a non-linear model (Boosted Regression Trees). For each of the 4 taxa we presented, the day/night effect explained less than 8% of total variance. Overall, the little effect of the diel cycle on plankton distributions likely results from the low number of detected organisms, as well as the averaging across relatively large bins.

5.4.3 Dynamics of plankton and particles during the bloom

Particle dynamics could be related to plankton variations.

According to fluorescence data, the phytoplankton bloom seems to have started around Feb 23rd, while UVP6 data shows that the zooplankton bloom began between Feb 20th and Feb 27th, so very close to the phytoplankton bloom (Figure S5.2). As described below, variations observed in particle abundance and size seems to indicate that particulate organic carbon (POC) export was affected by the composition of the plankton community, as previously reported by numerous studies [51, 52, 163, 165]. Indeed, zooplanktonic organisms are directly involved in the production of POC [379, 404]. Four phases emerged from the analysis particle and plankton distribution and are discussed below.

5.4.3.1 *Early bloom – February*

Appendicularia produced a lot of particles.

In February, the early phase of the zooplankton bloom was characterised by the presence of Appendicularia, including both inhabited and discarded houses, mostly in the 0-200 m layer. Appendicularia are filter-feeding pelagic tunicates. They grow a house made of mucopolysaccharides and cellulose used as an external mucous filter to collect food particles [4]. The house is disposable: when filters are obstructed, the house is discarded and renewed, up to several times a day [350]. These discarded houses are a major source of marine snow aggregates [7]. These large (3 mm ESD on average) particles have a relatively low sinking velocity (20-50 m d^{-1}) during the first hours, which then increases to 100 m per day after their initial deflation [242]. High Appendicularia abundance was previously linked to increased

export of large particles, mostly through discarded houses [3]. Thus, the sinking event of large particles, along the isopycnals (Figure 5.3) could correspond to these discarded houses, or faecal pellets, from Appendicularia.

5.4.3.2 *Mid bloom – March*

Copepods dominated after a mixing event induced by wind.

In a second phase, copepods were abundant in the 0-200 m layer while salps occupied the 0-100 m layer. As previously, bigger particles were found offshore but were less abundant than in February. Strong wind was recorded a few days before this transect (Figure S5.3), causing a mixing of the water column and a redistribution of phytoplankton, particles (Figures 5.2DG, 5.3BC) and possibly zooplankton (see pattern of copepods in Figure 5.4A). Such events have previously been reported during spring and can result in community changes both for phytoplankton [398] and zooplankton [339].

5.4.3.3 *Late bloom – April*

Large particles dominated in the upper part of the water column...

In April, the concentration of appendicularians increased, while that of salps of copepods decreased. This gelatinous bloom could be the result of favourable conditions in response to the gust of wind that took place two weeks before (20th March), in line with the results of Ménard et al. [270] who found that wind promoted blooms of salps, although this result could not be confirmed by Licandro, Ibañez, and Etienne [235]. At the same time, stratification began in coastal waters and the DCM started to form. The water column was mostly depleted in particles (except for very coastal waters). Particles present in the 0-100 m layer were big, a phenomenon that was found to originate from a decrease in the concentration of small particles (Figure S5.6). Further analyses showed that only 10% of objects in the 1-2 mm size range were living organisms. These large particles are therefore indeed particles and not planktonic organisms, such as appendicularians, which were more abundant in the same depth range. Moreover, the distribution of these large particles appeared to follow the isopycnals.

A first hypothesis to explain such observation could be a lack of particle aggregation, so that organic matter would remain in the form of particles to small to be detected by the UVP6 (< 70 µm). Backscattering measurements, targeting smaller particles (~10 µm), also show low

abundance in the water column on April 22nd (Figures 5.2, S5.1). Thus, this explanation does not seem to be the most appropriate.

We thus suggest that the decrease in small particles was caused by the presence of filter-feeding tunicates in the water column. Both salps and appendicularia are suspension-feeders and contribute to removing particles of various sizes (1 µm - 1 mm) from the water column [4] and aggregate them into larger, sinking particles, explaining the relative dominance of large particles. Faecal pellets produced by salps are relatively large (~5 mm) and sink at about 2000 m day^{-1} [62]. These faecal pellets are thus not likely to remain in the water column and cannot correspond to the large particles we observed in the 0-100 m range. On the other hand, recently discarded houses of appendicularians (containing faecal pellets) sink very slowly, at speeds typically below 50 m day^{-1} [242]. These particles are thus much likely to reside in the water column, by sitting on density gradients.

. . . and were likely discarded houses.

5.4.3.4 *After bloom – June*

Finally, both stratification and DCM intensified while zooplankton concentration decreased. The zooplankton community was dominated by Rhizaria. Some Rhizaria are mixotrophic and typical of oligotrophic environments [42]. In our study, we mostly detected small unidentifiable Rhizaria and solitary Collodaria. The analysis of a time series of the complete plankton assemblage from net samples collected in Villefranche Bay highlighted a peak of Rhizaria in July [339], in oligotrophic conditions that are comparable with our transect at the end of June. The fine-scale distribution of these organisms during the oligotrophic summer was studied across the Ligurian front in summer, using the *In Situ* Ichthyoplankton Imaging System (ISIIS), an in situ imager with very high sampling rate (> 100 L s^{-1}) [85]. Solitary Collodaria precisely followed the DCM, while Acantharia were mostly found on the coastal side in the upper 50 m of the water column [124]. Further results emerging from data collected with the ISIIS will be presented in Chapter 6. Such patterns could not be resolved with the UVP6, mostly because of its lower sampling rate (< 1 L s^{-1}), which required to aggregate observations into broader taxomomic groups and onto a much coarser spatial grid. Particle distribution revealed more abundant and smaller particles around the DCM, which could correspond to large phytoplanktonic cells (> 100 µm) of the accumulation of particles related to biological activity.

The end of the bloom was characterised by the presence of Rhizaria and a DCM.

5.4.4 Mesoscale features

5.4.4.1 *Particles*

Previous studies reported a strong influence of the front on the distribution of particles with small aggregates being more abundant in coastal waters while large aggregates were more abundant in the frontal zone, suggesting that these aggregates are produced in this zone through physical coagulation or biological transformation [146, 384]. In addition, the fact that large particles are more abundant in surface waters suggests that they are formed in these waters before being exported to the depths [381]. Our findings also show this strong influence of the front on the distribution of particles: small aggregates were more abundant in coastal waters and the front acted as a barrier to particles spreading offshore. Moreover, mean aggregate concentrations (0.75 L^{-1} in 512 µm - 1.02 mm; 0.09 L^{-1} in 1.02-2.05 mm) were also very close to those found by Stemmann et al. [381]. However, while large aggregates were sometimes more abundant under the frontal zone, this was not always the case (Figure S5.5).

Particle distribution was constrained by the front.

5.4.4.2 *Plankton*

Many studies detected an effect of the front on the distribution of planktonic organisms, either an increase in abundance or biomass at the front [48, 49, 234, 275] or different concentrations of certain taxa on either side of the front [124, 219, 311]. Such effects were not as clear in our data. First the coarse spatial resolution imposed by the sampling rate made it difficult to identify processes "at the front". Second the coarse taxonomic resolution may have hidden some underlying differences, within the broad group of copepods for example. Finally, for most of the study period, the current was quite close to the coast and coastal communities were little sampled, making them difficult to contrast with offshore ones. Still, the concentration of several taxa were higher in the offshore region than in the current or in the coastal one: copepods in February and March, appendicurarians in February and April, salps in March and April (Figure 5.4). This is compatible with a barrier effect of the front.

We could not show a strong effect of the front on plankton distribution.

5.4.5 Submesoscale features

5.4.5.1 *Cold lenses*

Glider sampling also allowed detecting submesoscale hydrological features such as submesoscale coherent vortices...

A lens (referred hereafter as L3) of cold, fresh water with high oxygen and low CDOM concentrations water of 9 km × 200 m was detected during the back transect on April 22nd, between 46 and 55 km offshore and from 130 to 330 m depth (Figure 5.2). It was barely visible on density and BB700. Similar structures were looked for in the whole dataset, including outgoing transects. The same lens was crossed 17 h before on the outgoing transect on April 20th, but appeared smaller (5 km width, labelled L2 on Figure S5.4). Another lens (L1), closer to the coast, was detected on the same outgoing transect but not captured before during the previous back transect which sampled the same area 25 h before. This gives us a clue to the speed of drift, which assuming the same dimension in x and y, would be in the range of 5 to 10 cm s^{-1}.

...which did not seem to affect the distribution of particles,...

Such features are likely to be submesoscale coherent vortices (SCV), previously detected in the area [46, 47]. A description of their physico-biogeochemical properties showed differences in water properties between the SCV (more oxygen, less nutrients) and surrounding waters to an extent that it affected the phytoplankton community. T his also highlighted reduced lateral exchanges between the core of the SCV and the surroundings [47]. Yet, we did not detect any effect of the SCV on the concentration or the size of particles (Figure S5.5), which suggests that it did not act as a strong barrier for particles. Finally, changes in the phytoplankton community inside the SCV are likely to propagate to zooplankton, but the sampling resolution of planktonic organisms was too coarse to detect any effect of the SCV.

5.4.5.2 *Subducting water mass*

...as well as subducting water masses...

On February 27th, a mass of high chlorophyll, high oxygen, low salinity, low temperature and low CDOM water was recorded, down to a depth of 300 m (Figures 5.2, S5.1). This water layer was about 3.8 km in width and 20 m in height, and was sinking towards the coast, following the isopycnals along the front. This is coherent with a convergence event [48], and was already observed by Niewiadomska et al. [290] from glider data collected on a similar transect across the Ligurian front in January. The water mass they observed had similar properties and was 4 km wide, subducting down to a depth of 180 m. Analo-

gous features were also detected during spring in the Corsican side of the Ligurian current: phytoplankton produced in the surface layer was carried downward along the isopycnals, resulting in a plume of chlorophyll down to 100 m [144]. Here, in addition, we were also able to demonstrate that this subductiong water mass also carried more and bigger particles from the surface towards the depth. Previously, it was only suggested that food particles could be transported downward by the frontal circulation along the isopycnals [147].

. . . which, on the contrary, brought particles downward.

5.5 Conclusion

In conclusion, repeated sections across the Ligurian front with a glider allowed us to resolve submesoscale hydrological features. A clear linked emerged between the environment, the distribution of particles and, to some extent, that of planktonic organisms, for example in a subducting water mass and during a mixing event. Moreover, the accumulation along isopycnals creates the DCM in April and July, which is reflected in the distribution of particles. The temporal evolution of the plankton community during the spring bloom was also related to changes in the abundance and size of marine snow particles. While we detected an influence of the front on the distribution of marine snow particles, the signal was coarser for planktonic organisms, probably due to the low sampling rate so that too few organisms were imaged.

A strong link emerged between environment and particles,. . .

. . . but was not crystal clear for plankton.

Overall, these results confirm the need to study physics, biogeochemistry and biology at the same scale, by sampling both the environment, particles and plankton at fine scale using *in situ* imaging. This approach should allow to better understand the biological responses to submesoscale hydrological forcings.

Author Contributions

TP designed the study under the supervision of LC, MP and JOI. TP and AP piloted the glider. All authors contributed to glider deployment and retrieval. TP, ER, MP and CC contributed to on-land maintenance of the glider and to data transfer. TP conducted the analyses and wrote the manuscript, with support from JOI. All authors contributed to the discussion of the results, supported manuscript preparation and approved the final submitted manuscript.

Acknowledgements

We thank the Alseamar Team and on-call pilots who supported us for glider piloting: Laurent Beguery, Florent Besson, Nicolas Buisson, David Diaz, Orens de Fommervault and Marion Mery. The authors also thanks everyone who contributed to glider deployment or retrieval: Ewen Ancel, Marin Cornec, Paul Dasi, Florent Hallal, Gregory Maggion, Solène Motreuil, Stéphane Renouf, Florian Ricour, Vincenzo Vellucci and Laure Vilgrain. We thank the EMBRC platform PIQv for image analysis. This work was supported by EMBRC-France, whose French state funds are managed by the ANR within the Investments of the Future program under reference ANR-10-INBS-02. This study is part of project "World Wide Web of Plankton Image Curation", funded by the Belmont Forum through the Agence Nationale de la Recherche ANR-18-BELM-0003-01. TP's doctoral fellowship was granted by the French Ministry of Higher Education, Research and Innovation (3500/2019). TP also acknowledges Fabien Lombard and Lars Stemmann for the insightful discussions regarding the interpretation of observations, as well as Martin Schröder and Rainer Kiko for their assistance to sort plankton images with the Morphocluster application.

Supplementary Materials

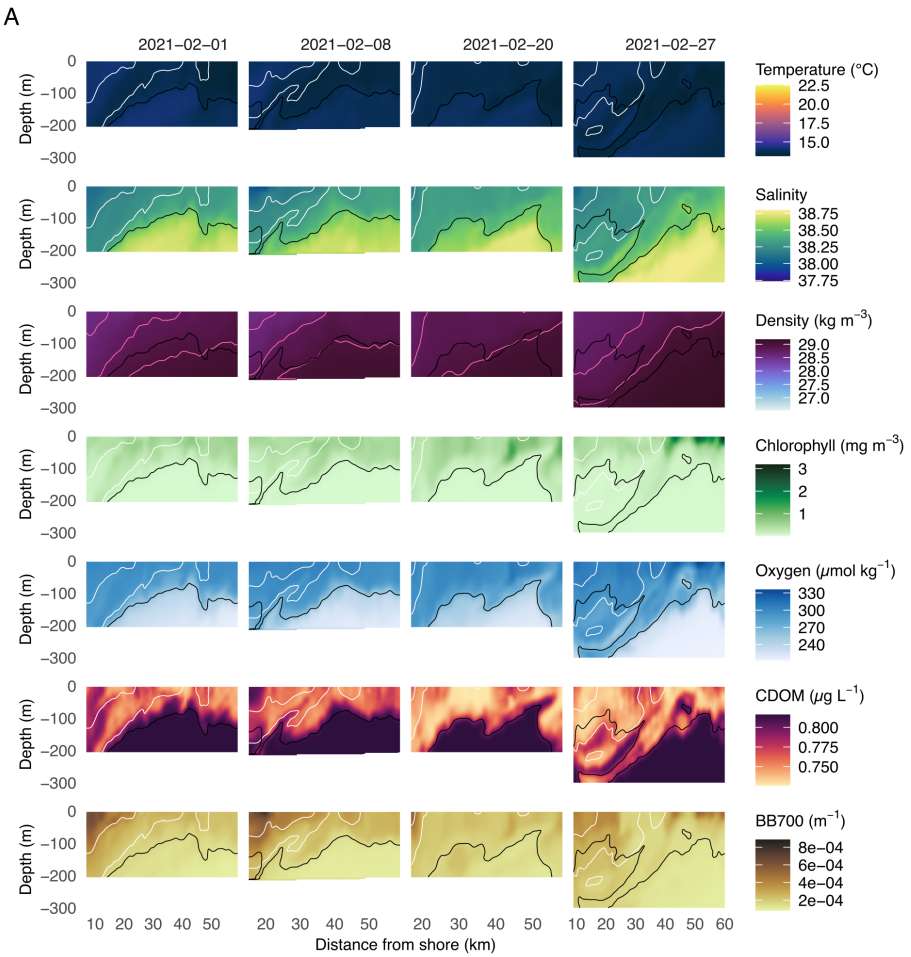


Figure S5.1: *(Figure continues)*

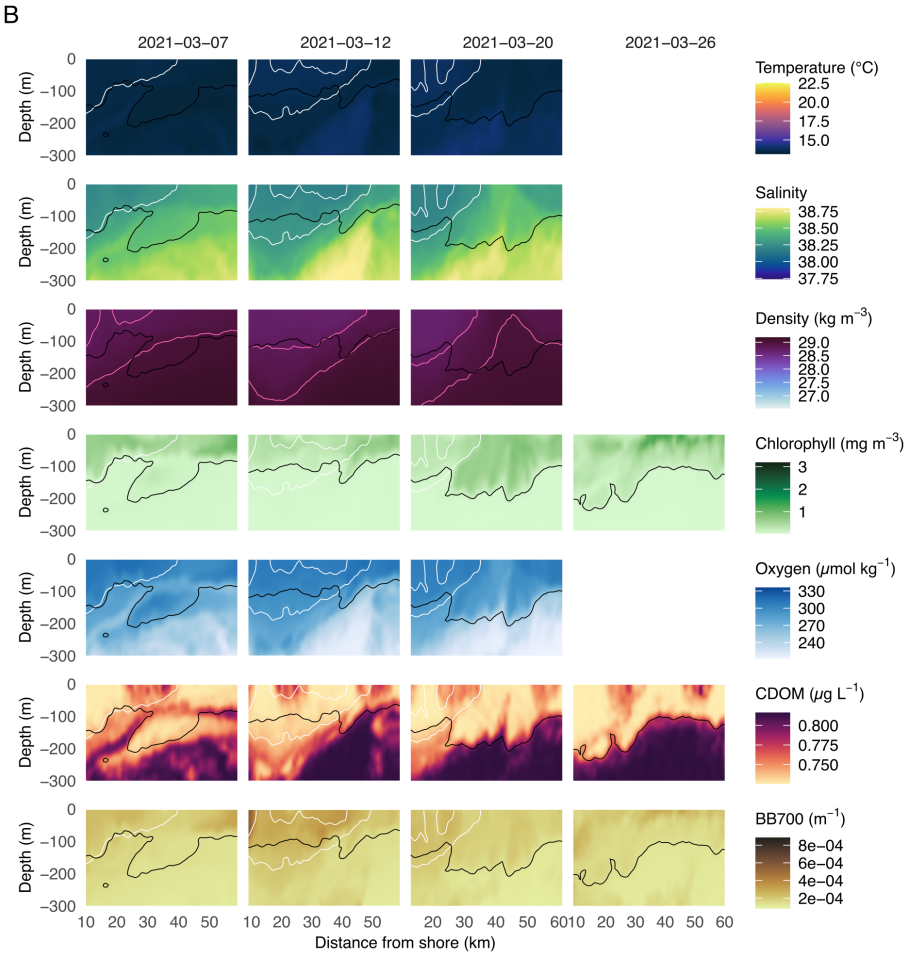


Figure S5.1: *(Figure continues)*

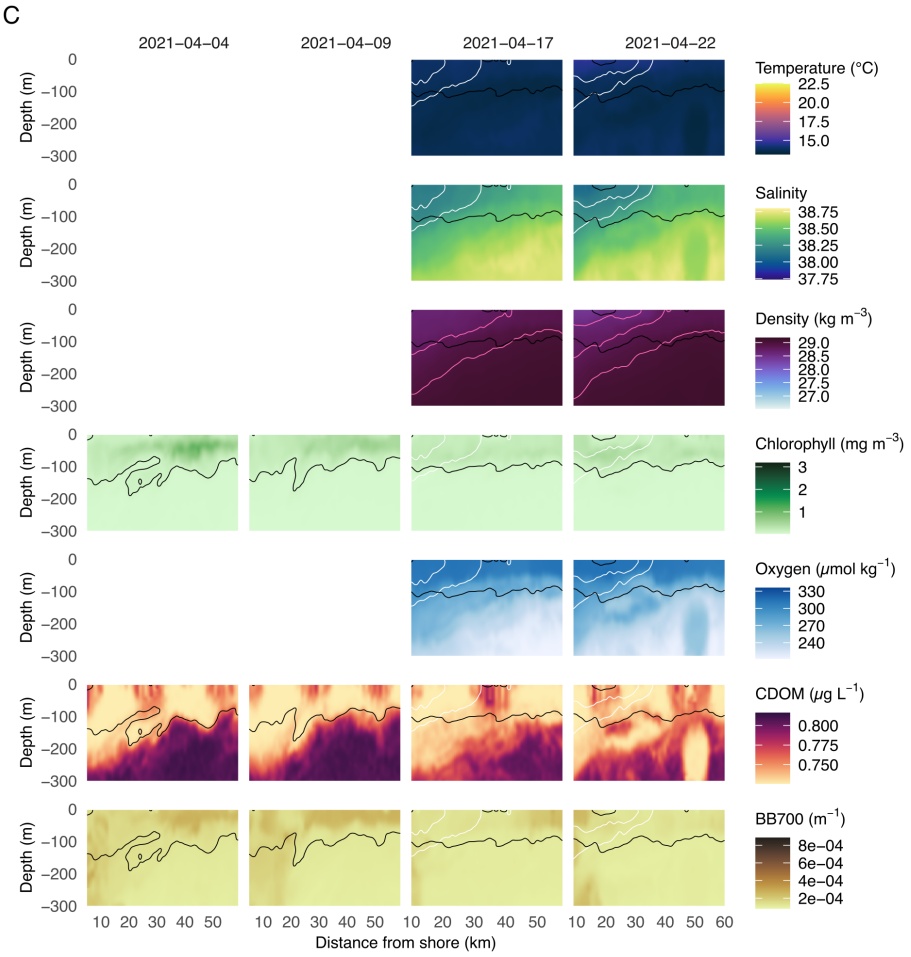


Figure S5.1: *(Figure continues)*

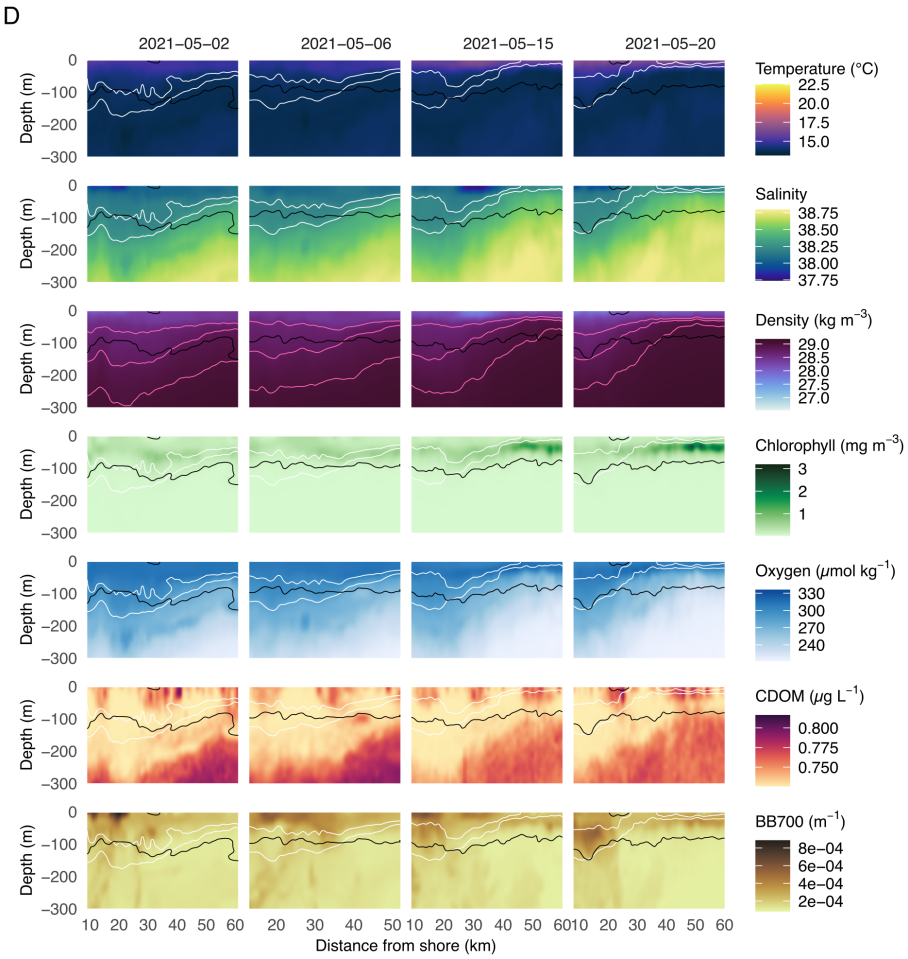


Figure S5.1: *(Figure continues)*

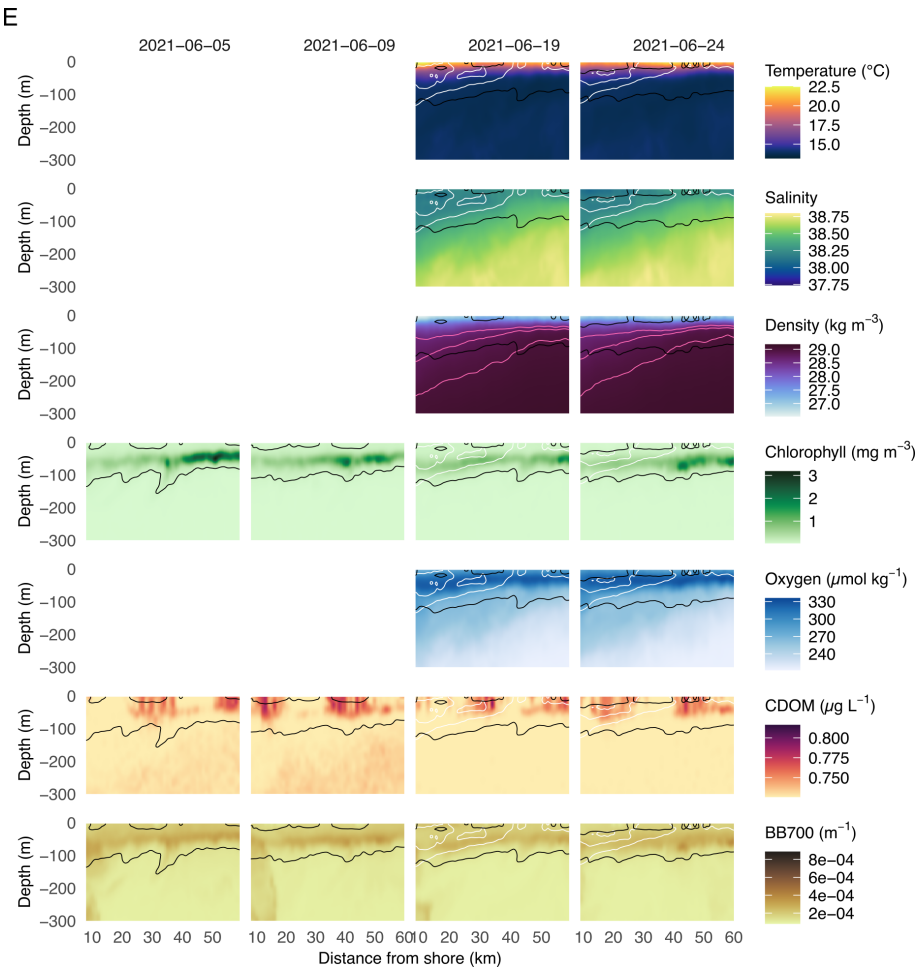


Figure S5.1: Environmental data for all transects, arranged by date. **(A)** February, **(B)** March, **(C)** April, **(D)** May and **(E)** June. Each column corresponds to a given date and rows are variables: temperature, salinity, potential density anomaly, chlorophyll, oxygen, CDOM and BB700. White lines are the 38.2 and 38.3 isohalines delimiting the front, black lines are the 0.1 chlorophyll isoline, pink lines are the 28.6, 28.8 and 29 isopycnals. Colour scales are shared among subplots to ease comparison. Missing data are due to a CTD malfunction.

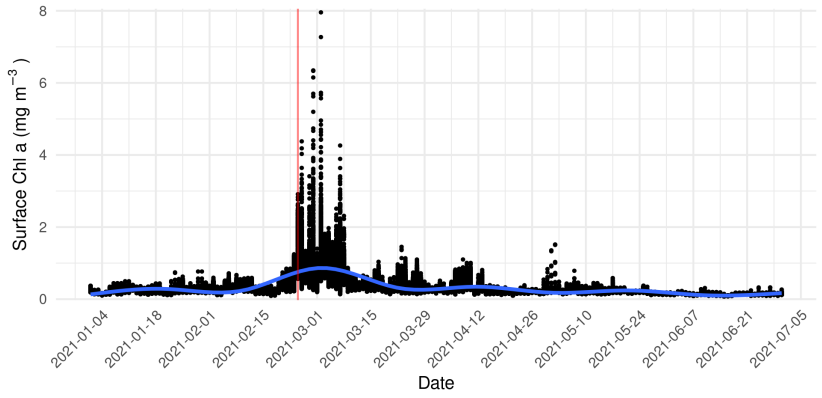


Figure S5.2: Values of surface chlorophyll concentration detected from satellite observations of ocean colour (`OCEANCOLOUR_GLO_BGC_L4_MY_009_104`) in the area covered by glider sampling. The vertical line highlights a sudden increase in chlorophyll concentration, indicating the beginning of the phytoplankton bloom, on February 23rd.

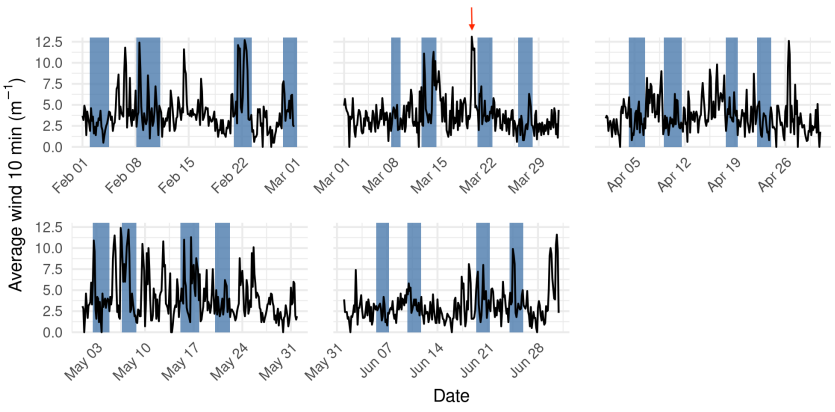


Figure S5.3: Wind averaged over 10 mins for the duration of the campaign. Blue rectangles represent the back transects. The strong wind event is highlighted by the red arrow.

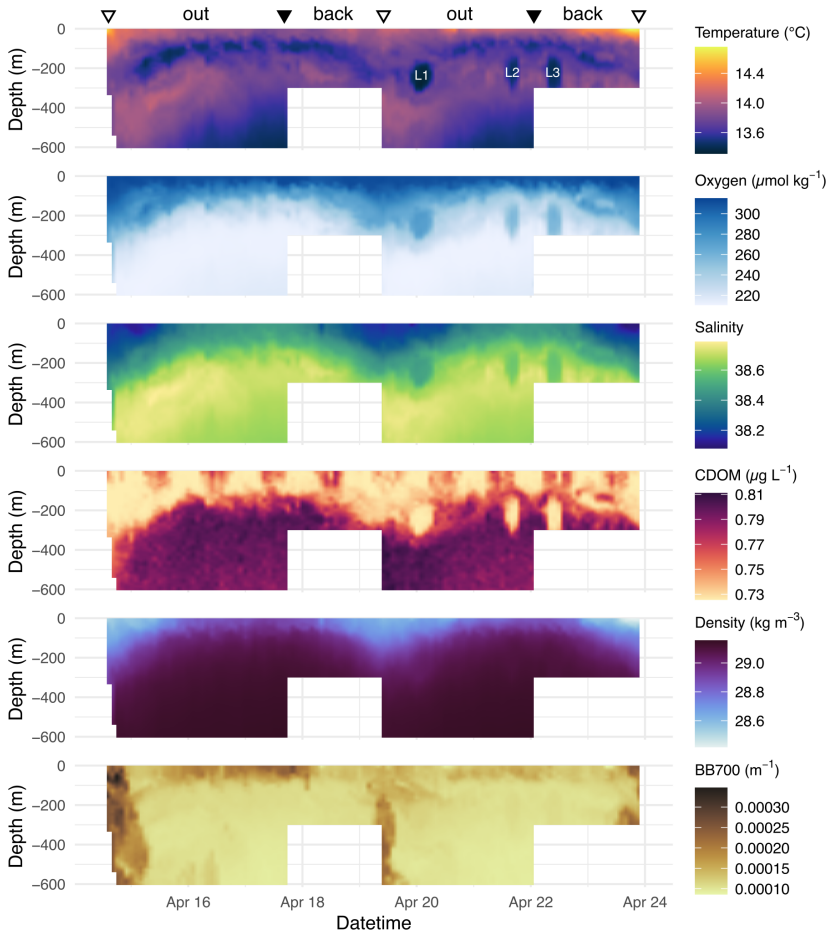


Figure S5.4: Environmental data recorded by the glider during two round trips between April 15th and 24th, showing three water lenses located around 200 to 400 m depth. **(A)** temperature, **(B)** oxygen, **(C)** salinity, **(D)** CDOM, **(E)** density, **(F)** BB700. White triangle = coast, black triangle = Dyfamed.

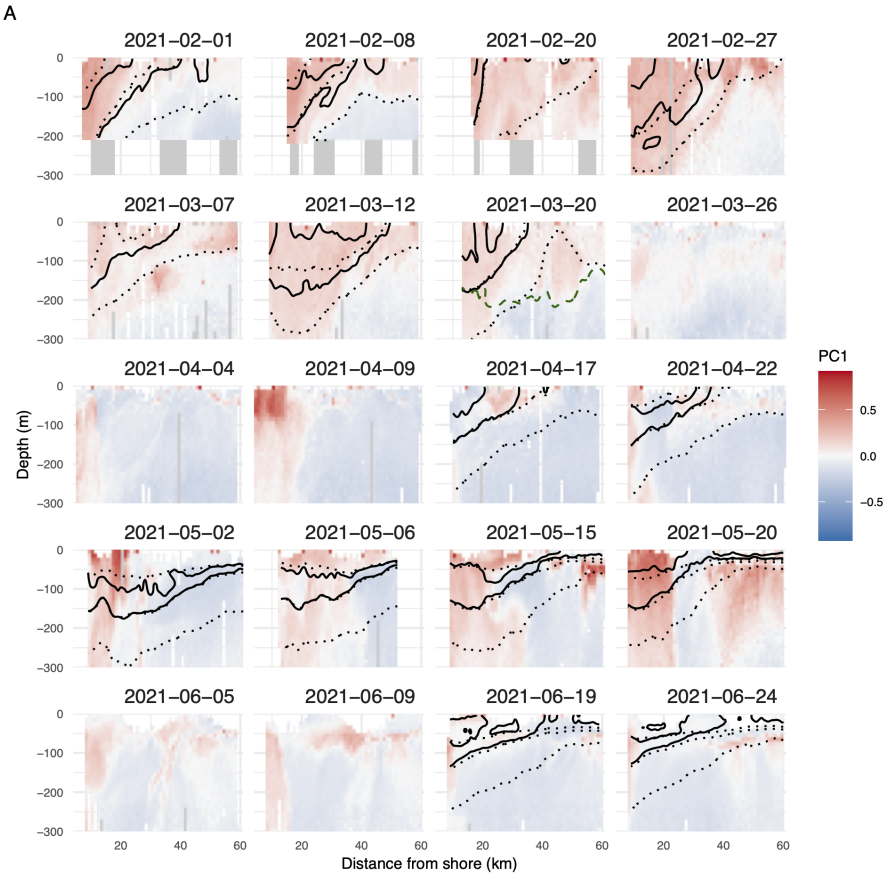


Figure S5.5: *(Figure continues)*

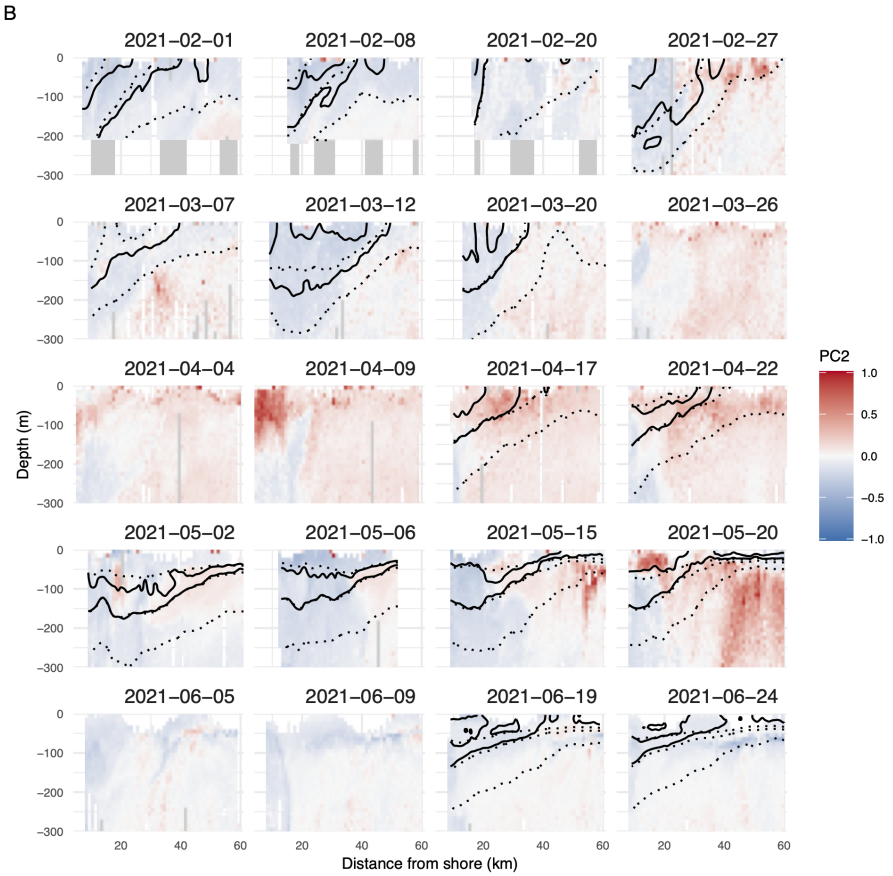


Figure S5.5: Evolution of particle distribution. PC1 **(A)** and PC2 **(B)** projections from the PCA performed on log-transformed particle data. The 38.2 and 38.3 isohalines delimiting the front are represented as black lines. Grey rectangles in the plot background represent night time.

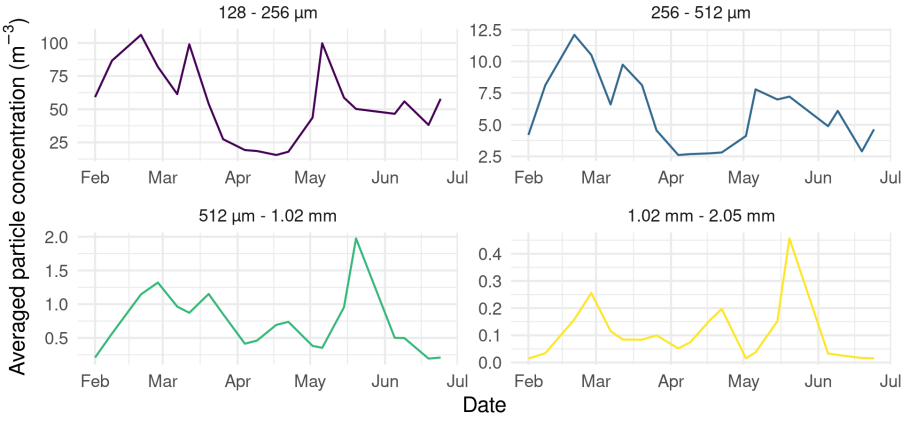


Figure S5.6: Evolution of particle concentration in the offshore (distance to coast > 20 km) top part (0 - 100 m) of the water column, for 4 size classes of particles, showing a decrease in small particles in April.

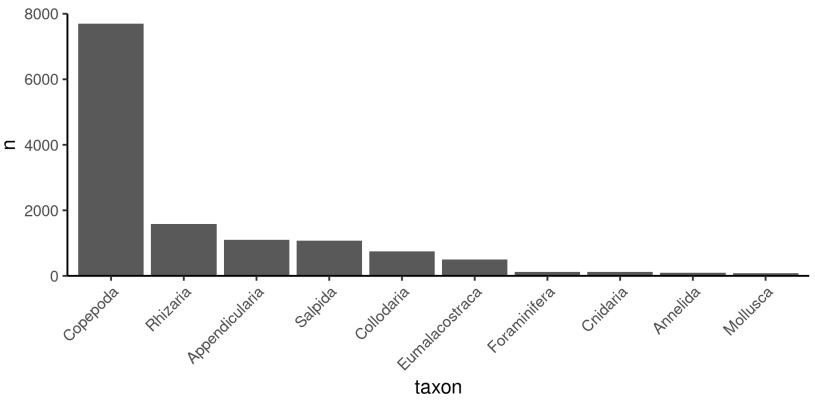


Figure S5.7: Dataset composition: total number of images per taxonomic group after probability thresholding.

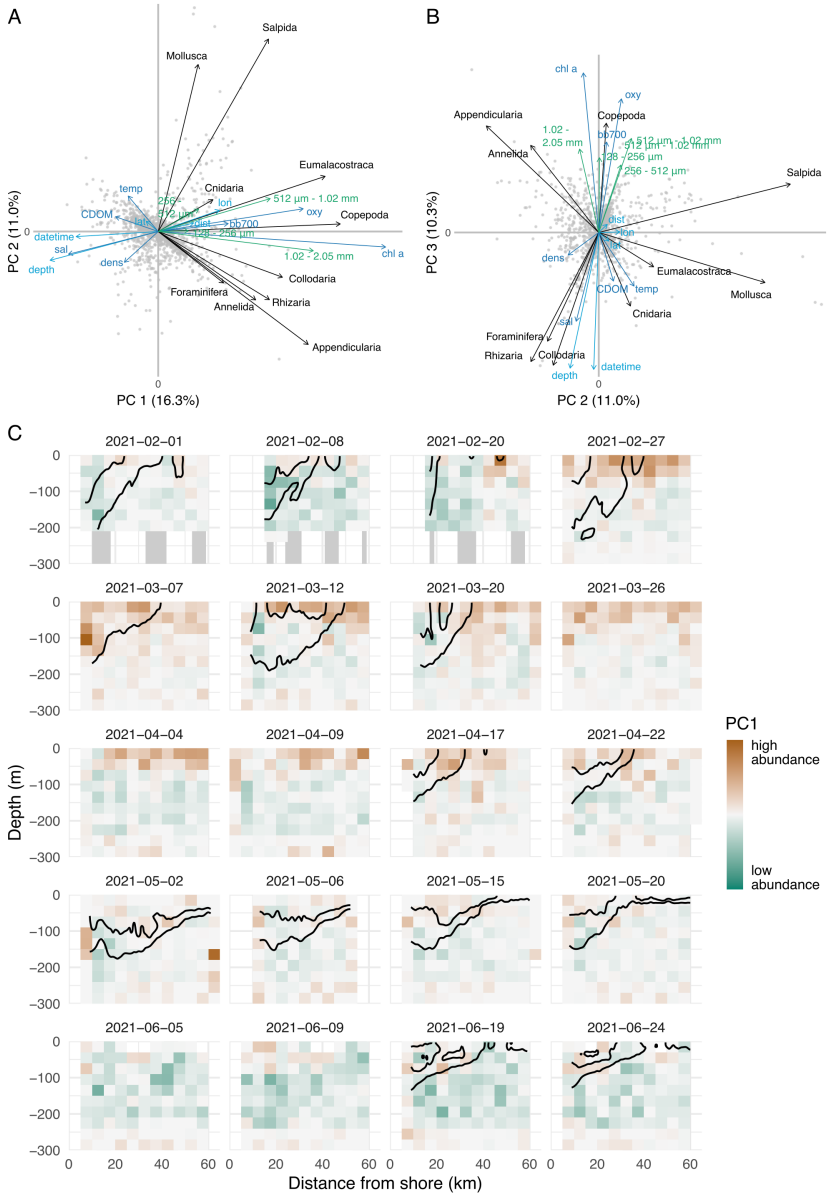


Figure S5.8: PCA performed on log-transformed plankton data: axes 1 and 2 **(A)** and axes 2 and 3 **(B)**. Projections of bin coordinates on PC1 **(C)** replaced in the four representative transects. The 38.2 and 38.3 isohalines delimiting the front are represented as black lines. Grey rectangles in the plot background represent night time.

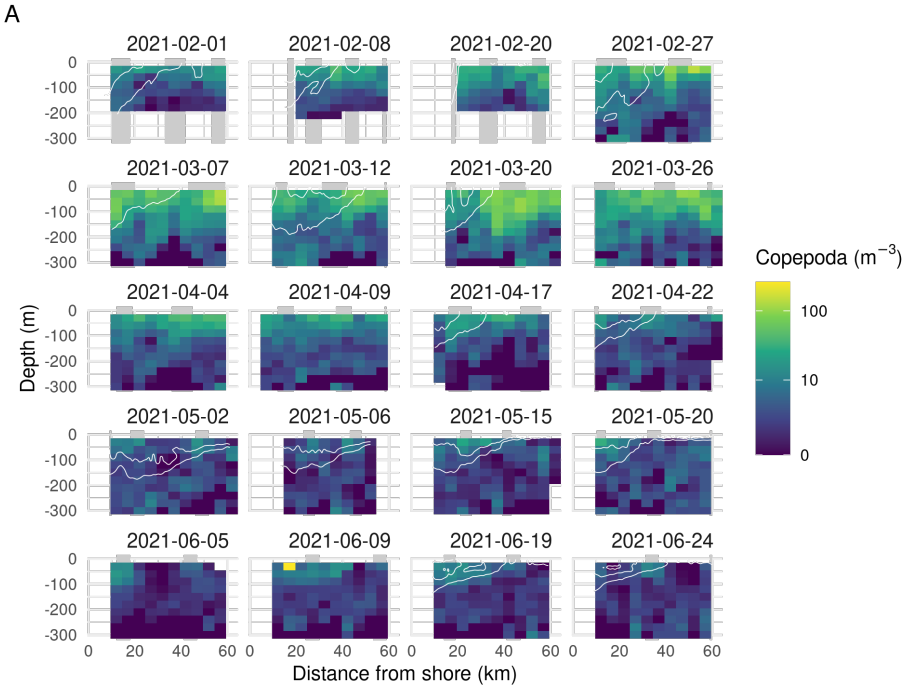


Figure S5.9: *(Figure continues)*

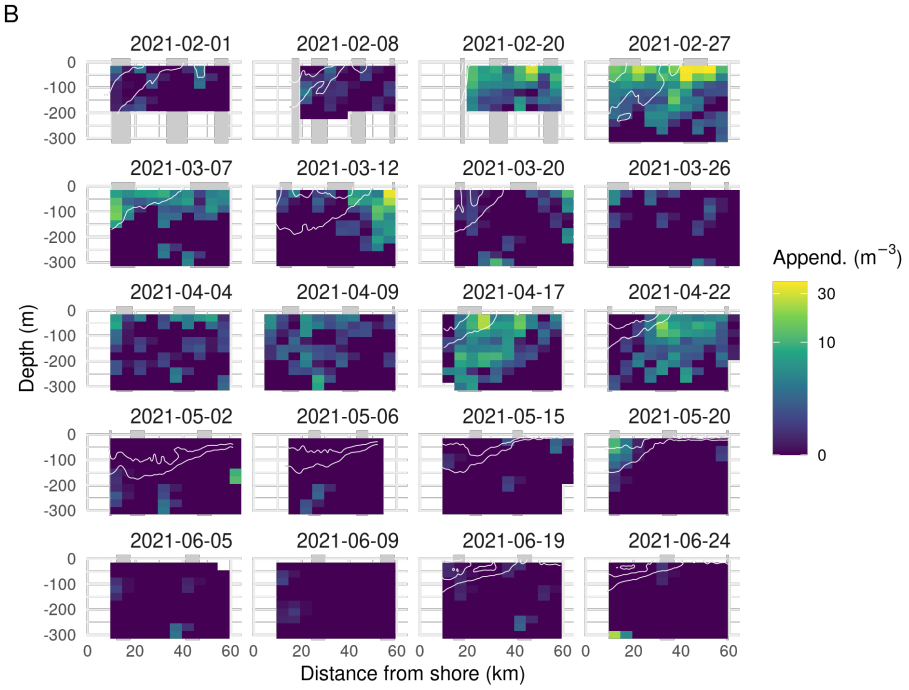


Figure S5.9: *(Figure continues)*

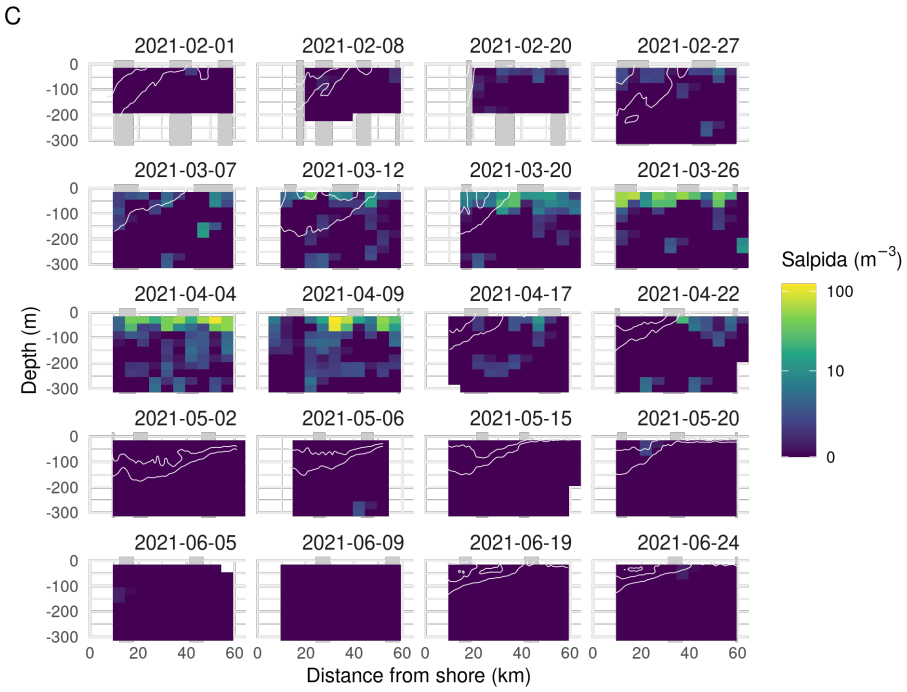


Figure S5.9: *(Figure continues)*

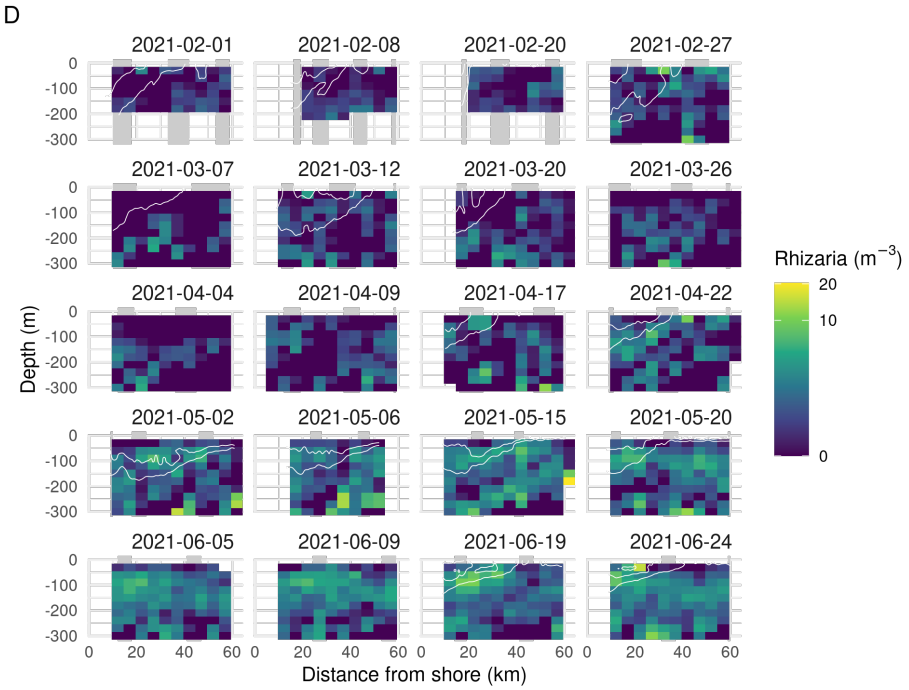


Figure S5.9: Evolution of plankton concentration on all transects. **(A)** Copepoda, **(B)** Appendicularia, **(C)** Salpida, **(D)** Rhizaria. The 38.2 and 38.3 isohalines delimiting the front are represented as white lines. Grey rectangles in the plot background represent night time.

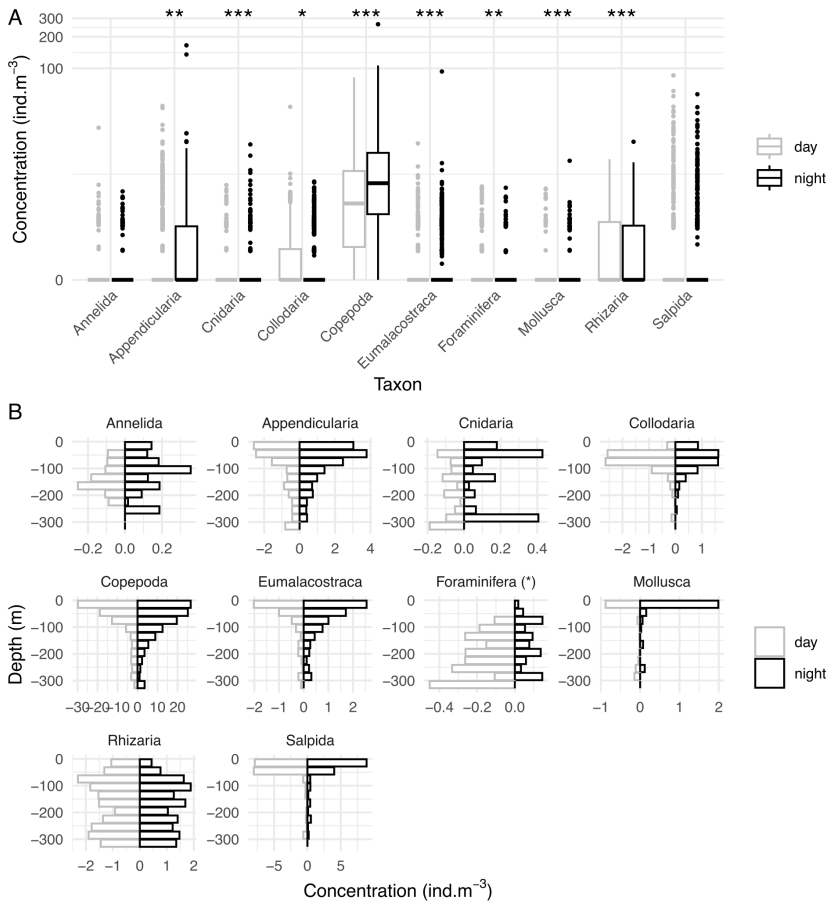


Figure S5.10: Comparison of day and night plankton concentrations. **(A)** Concentrations averaged along yos performed at day and night for each taxonomic group. Differences in median concentrations were tested with a Wilcoxon test, the significance of which is indicated by stars. **(B)** Distributions computed on 30 m vertical bins at day and night for each taxonomic group. Differences in distributions were tested with a Kolmogorov-Smirnov test. * = 0.05, ** = 0.01, *** = 0.001.

Table S5.1: Missions summary.

Deployment	Retrieval	Time at sea (h)	Number of images		Acquisition frequency (Hz)	
			Total	Back transects	0-220 m	> 220 m
2021-01-28 10:21:08	2021-02-12 10:02:11	360	121,840	51,302	0.5	0.2
2021-02-18 13:33:39	2021-03-02 12:42:13	287	154,519	65,591	0.5	0.2
2021-03-04 14:37:08	2021-03-15 14:33:16	264	107,319	31,025	0.5	0.2
2021-03-17 11:55:48	2021-03-29 09:46:25	286	82,819	38,113	0.5	0.2
2021-03-31 12:45:47	2021-04-12 10:16:14	286	118,826	51,788	0.5	0.2
2021-04-14 12:00:36	2021-04-25 07:49:46	260	70,264	29,584	0.5	0.2
2021-04-28 12:49:31	2021-05-09 07:13:08	258	40,651	26,572	0.5	0.2
2021-05-12 10:23:28	2021-05-26 10:25:21	336	260,578	80,959	0.5	0.2
2021-06-02 10:13:56	2021-06-13 02:20:46	256	92,732	39,353	1.3	0.2
2021-06-16 13:16:54	2021-06-28 09:25:35	284	73,575	19,004	1.3	0.2

Table S5.2: Classification performance before probability threshold. n test = number of objects in test set.

Class	Precision	Recall	n test
Annelida	75.0%	50.0%	12
Appendicularia	37.8%	59.3%	91
artefact	88.9%	63.9%	601
Cnidaria	50.0%	40.0%	5
Collodaria	94.8%	88.7%	62
Copepoda	60.6%	74.0%	918
detritus	98.6%	98.1%	40575
Eumalacostraca	50.0%	87.9%	33
Foraminifera	84.2%	84.2%	19
Mollusca	37.5%	75.0%	4
other_living	82.4%	31.8%	44
Rhizaria	44.3%	77.3%	141
Salpida	69.2%	92.2%	90

Table S5.3: Classification performance after probability threshold. n test = number of objects in test set. Most classes do reach 75% of precision, with the exception of Cnidaria for which precision could not reach such a score, because this class was polluted by many objects with higher prediction scores than actual Cnidaria. As Appendicularia and Rhizaria had low recall, these classes were instead manually validated.

Class	Precision	Recall	n test
Annelida	71.4%	41.7%	12
Appendicularia	75.9%	24.2%	91
artefact	89.1%	63.9%	601
Cnidaria	0.0%	0.0%	5
Collodaria	94.7%	87.1%	62
Copepoda	75.2%	56.4%	918
detritus	98.6%	98.1%	40575
Eumalacostraca	77.4%	72.7%	33
Foraminifera	83.3%	78.9%	19
Mollusca	100.0%	25.0%	4
other_living	87.5%	31.8%	44
Rhizaria	83.3%	7.1%	141
Salpida	76.2%	88.9%	90

6

High throughput *in situ* imaging reveals complex ecological behaviour of giant mixotrophic protists

To pursue our zoom towards smaller scales, the following chapter tackles the fine-scale distribution of planktonic organisms across the aforementioned Ligurian front. More specifically, we focus our study on rhizarians, a group of understudied protists. Data supporting this work was collected with the ISIIS during the VISUFRONT cruise, and fully processed thanks to the AI-based pipelines presented in Chapters 2 and 3.

Taking advantage of the fact that *in situ* imaging enables to study these fragile organisms in their undisturbed environment, we are able to resolve meter-scale vertical distribution, as well as previously unreported preferential orientation in multiple groups of Rhizaria. Finally, we also provide new observations that are consistent with the current knowledge regarding the life cycle of mixotrophic Rhizaria.

Thelma Panaïotis, Tristan Biard, Louis Caray–Counil, Robin Faillettaz, Jessica Luo, Cedric M Guigand, Robert K Cowen and Jean-Olivier Irisson

Manuscript in preparation to be submitted to **PNAS**

Abstract

Plankton play crucial roles in the oceans, both as the base of oceanic food webs and a key link in global biogeochemical cycles. Although planktonic organisms have been the topic of scientific research for cen-

turies, some organisms have fallen through the cracks, such as Rhizaria. Indeed, these unicellular eukaryotes are particularly delicate and often crushed by classical plankton sampling instruments. Despite some Rhizaria have been reported being mixotrophic and host photosynthetic symbionts, gaps persist about their trophic ecology. Knowledge regarding their reproductive cycle is even scarcer. Their substantial contribution to the planktonic biomass was recently brought to light thanks to *in situ* imaging. Such an approach allows the study of these organisms in their undisturbed environment. Leveraging high frequency *in situ* imaging data, we investigated the fine-scale distribution and *in situ* position of ~230,000 organisms belonging to three groups of Rhizaria, including taxa Acantharia, Collodaria and Phaeodaria. We brought to light differences in vertical distribution between subgroups, likely underpinned by different life strategies and contrasted abilities for buoyancy control. We also detected a previously undocumented preferential orientation of some organisms in each taxon. Finally, we try to relate some of our observations to presumptive steps of the obscure Collodaria life cycle, likely involving variations of buoyancy control to reach new environments.

Résumé

Le plancton joue des rôles cruciaux dans les océans, à la fois en tant que base des réseaux trophiques océaniques et en tant qu'élément essentiel des cycles biogéochimiques globaux. Bien que les organismes planctoniques fassent l'objet de recherches scientifiques depuis des siècles, certains organismes sont passés à travers les mailles du filet, comme les Rhizaria. En effet, ces eucaryotes unicellulaires sont particulièrement fragiles et souvent endommagés par les outils classiques d'échantillonnage. Bien que certains Rhizaria soient connus comme mixotrophes hébergeant des symbiotes photosynthétiques, des lacunes persistent quant à leur écologie trophique. Les connaissances concernant leur cycle de reproduction sont encore plus rares. Toutefois, leur contribution substantielle à la biomasse planctonique a récemment été mise en évidence grâce à l'imagerie *in situ*. En effet, cette approche permet l'étude de ces organismes dans leur environnement non perturbé. En exploitant les données d'imagerie *in situ* récoltées à haute fréquence, nous avons étudié la distribution à fine échelle et la position *in situ* de ~230 000 organismes appartenant à trois groupes de Rhizaria (Acan-

tharia, Collodaria et Phaeodaria). Nous avons mis en évidence des différences dans la distribution verticale entre les sous-groupes, probablement causées par des stratégies de vie différentes et des différences de capacités dans le contrôle de la flottabilité. Nous avons également détecté une orientation préférentielle, non documentée auparavant, de certains organismes. Enfin, nous avons essayé de relier certaines de nos observations aux étapes présumées du cycle de vie méconnu des Collodaria, révélant potentiellement des variations du contrôle de la flottabilité des organismes afin d'atteindre l'environnement dans lequel se déroule l'étape suivante de leur cycle.

6.1 Introduction

Different kinds of mixotrophy exist...

Mixotrophy – the ability to use alternate sources of nutrients – exists in plants and metazoans but is much more widespread in planktonic organisms such as protists (unicellular eukaryotes) [356]. Because mixotrophy enables the occupation of many ecological niches, mixotrophic planktonic protists, combining phototrophic and heterotrophic nutrition [67], are ubiquitous and dominate both freshwater and marine ecosystems [274, 386]. Multiple variations of mixotrophy co-exist within protists [70]. Constitutive mixotrophs are photosynthetic protists with the ability to ingest prey, sometimes referred to as "plant that eats" [227]. Non-constitutive mixotrophs become mixotrophic thanks to the acquisition of phototrophy, which can be mediated through the acquisition of specific free-living photosynthetic protists (endosymbiosis) or by the retention of functional plasts (kleptoplastidy) from ingested preys. Non-constitutive protists – representing 40 to 60% of protists [227] – are diverse (e.g. Foraminifera, Radiolaria, dinoflagellates, diatoms) [97, 387, 421] and common in all oceanic biomes. Planktonic protists are the most abundant eukaryotes in pelagic ecosystems and substantial contributors to the global plankton biomass [70]. Involved in primary production, carbon sequestration through the biological carbon pump and linking trophic levels, planktonic protists occupy critical ecological roles in the oceans [295, 419].

...and are widespread among planktonic protists...

...which play key ecological roles.

Although planktonic organisms have been the topic of scientific research for centuries [314], most research focused on larger organisms, especially Metazoa [305], leaving aside a whole part of the planktonic biodiversity. First described more than one century ago [166, 283],

Rhizaria are understudied...

Rhizaria are a diverse group of protists belonging to the SAR supergroup. Most Rhizaria bear a mineral skeleton of which composition varies between taxa: strontium sulphate for Acantharia, calcium carbonate for Foraminifera, opaline silica for Phaeodaria and polycystine Radiolaria [286, 393]. In addition, internal structures are specific to the different groups: solitary Collodaria have a large central capsule surrounded by bubble-like aveoli [12]; Phaeodaria also possess a central capsule but also carry a phaeodium which is an aggregate of food and waste vacuoles [210]. Overlooked for a long time, recent studies based on underwater imaging shed light on their ecological roles [68], in particular as key elements of carbon [165, 217], silica [36, 240] and strontium [35] global biogeochemical cycles. Rhizaria ranges from tens of micrometers to several millimetres [286, 393], but Collodaria are also found under the form of large solitary cells (> 1 mm), consisting of a central capsule and encompassed by a gelatinous matrix [12]. Despite being unicellular organisms, all collodarian species exhibit colonial forms with up to tens of thousands of single cells embedded in a gelatinous cytoplasmic matrix. The size of colonies ranges from millimetres to meters [393]. Previously considered as different clades because of obvious morphological differences, molecular analyses demonstrated that solitary and colonial forms actually share the same molecular signature, likely constituting the two phases the life cycle of the same species [41].

...small unicellular Eukaryotes.

Still much to discover about the ecology of Rhizaria.

Because Rhizaria are not cultivated in the lab – except for a few species of Foraminifera [206] – the scarce knowledge we have of this group comes from *in situ* observation and sampling only, whether of individuals or environmental DNA. Most Rhizaria are phagotrophs, i.e. they feed on particles, living or dead, including a large variety of planktonic organisms [38] or detritic material [12]. Epipelagic Rhizaria (e.g. Acantharia, Collodaria) typically host photosynthetic symbionts [97], resulting in mixotrophy. In such an association, endosymbionts provide nutritional resources to their host and benefit in return from a favourable microenvironment, rich in nutrients and protected from predators and parasites [421]. But the hosts have to preserve their symbionts and transmit them to their offspring. Endosymbiotic cells can multiplicate within the host to compensate for the loss of symbionts.

Their trophic modes are diverse, ...

Photosymbionts can be very diverse, from Prokaryotes to unicellular algae. Most Foraminifera and Radiolaria are obligate mixotrophs: adult life-stages cannot survive without their symbionts [97]. In Acantharia

the symbionts undergo severe and irreversible modifications of their morphology and cellular machinery, suggesting energetic exploitation by the host cell which is, the only one to benefit from this association, that would rather be qualified as parasitism [406]. Phaeodaria, on their side, do not host photosynthetic symbionts and are thus purely heterotroph, likely feeding on suspended matter and other planktonic organisms [286].

Knowledge regarding Rhizaria life cycle remains scattered for most groups. The release of swarmers – small bi-flagellated cells (2–5 µm) – has been reported across several taxa including Acantharia, Collodaria, Phaeodaria and Foraminifera [206, 286, 393], but whether these swarmers are gametes remains unknown. Asexual reproduction by binary fission has also been reported in Foraminifera [50, 206], Phaeodaria [286] and Collodaria [11, 56]. During asexual reproduction by mitosis of the host, symbionts can be transmitted to the daughter cells through vertical transmission. As no vertical transmission has been reported during sexual reproduction in planktonic symbioses, cells produced by fecundation are thought to acquire their symbionts *de novo* in the environment [97], but how potential symbionts are specifically recognized remains an open question.

. . . and gaps persist in their reproduction cycle.

Due to their fragility, Rhizaria are often damaged when sampled by classic plankton sampling tools such as nets or pumps. As a consequence, their abundance and biomass have been overlooked for a long time [68]. Moreover, these sampling methods do not allow to resolve fine-scale distribution in relation to environmental conditions [85, 243]. *In situ* plankton imaging overcomes some of these limitations by providing high spatio-temporal resolution and preserving the relationship between the organisms and their environment. Being non-destructive, *in situ* imaging is more appropriate to sample fragile organisms such as Rhizaria [42, 104, 285]. In the past 30 years, multiple *in situ* imagers were developed with various specificities [243]. These approaches revealed contrasted patterns in the distribution of these unicellular organisms [40], but also highlighted their important contribution to the oceanic carbon biomass [42, 111]. Some *in situ* imagers enable capturing images of organisms without disturbing them, with the potential to reveal specific position, behaviour or interactions between organisms [299]. For example, copepods feeding behaviour was reported in various environments [276, 291, 293, 409]. Similarly, Gaskell, Ohman, and Hull [140] investigated planktonic Foraminifera and revealed the

Overlooked because of damaging sampling methods. . .

. . . Rhizaria can now be investigated through in situ imaging. . .

. . . which also enables the investigation of plankton behaviour.

volume they occupied *in situ* and brought to light a preferential orientation of these organisms.

Thus, many questions remain open regarding the ecology of Rhizaria: which factors drive the fine-scale distribution of both mixotroph and non-mixotroph? Are life stages and symbionts acquisition related to the position in the water column? In this study, we leverage high resolution *in situ* imaging data of planktonic Rhizaria from the NW Mediterranean, across a front in an oligotrophic environment – a habitat where Rhizaria are important – to (i) resolve the fine-scale distribution of these organisms, especially comparing mixotrophic and non-mixotrophic ones, (ii) reveal individual aspects, and (iii) fill gaps in ecological knowledge – mostly reproduction cycle – of these organisms.

6.2 Results

6.2.1 An extensive dataset

Detected organisms were diverse in terms of taxonomy and size.

A total of ~8 million planktonic organisms were imaged by the *In Situ* Ichthyoplankton Imaging system (ISIIS) in the 0-100 m layer across the Ligurian front, a permanent mesoscale front in the Ligurian Sea (NW Mediterranean). Within these, ~230,000 Rhizaria were sorted into 14 taxonomic and/or morphological categories, belonging to four larger groups: Acantharia, Phaeodaria, solitary Collodaria and colonial Collodaria. Each group contained between 147 and 104,455 individuals, covering a size range from 0.4 mm (equivalent spherical diameter) for Acantharia to 35 mm for colonial Collodaria (Figure 6.1). Acantharia ($n \approx 150{,}000$) and Aulacanthidae ($n \approx 50{,}000$; Phaeodaria) dominated the dataset while Collodaria were less abundant ($n \approx 10{,}000$ solitary; $n \approx 15{,}000$ colonies).

Environmental conditions were emblematic of the oligotrophic summer period.

Organisms were distributed in a strongly stratified water column (Figure S6.1), with a thermocline/pycnocline around 10 m depth. The front was well marked, delimiting fresher water inshore from saltier water offshore. The deep chlorophyll maximum (DCM) was located between 50 m inshore and 75 m offshore where it was more spread out. Finally, oxygenated waters were found between the thermocline and the DCM and highlighted two tongues of downwelling waters, also visible on temperature. Overall, these features are typical of the oligotrophic summer period in the Ligurian Sea.

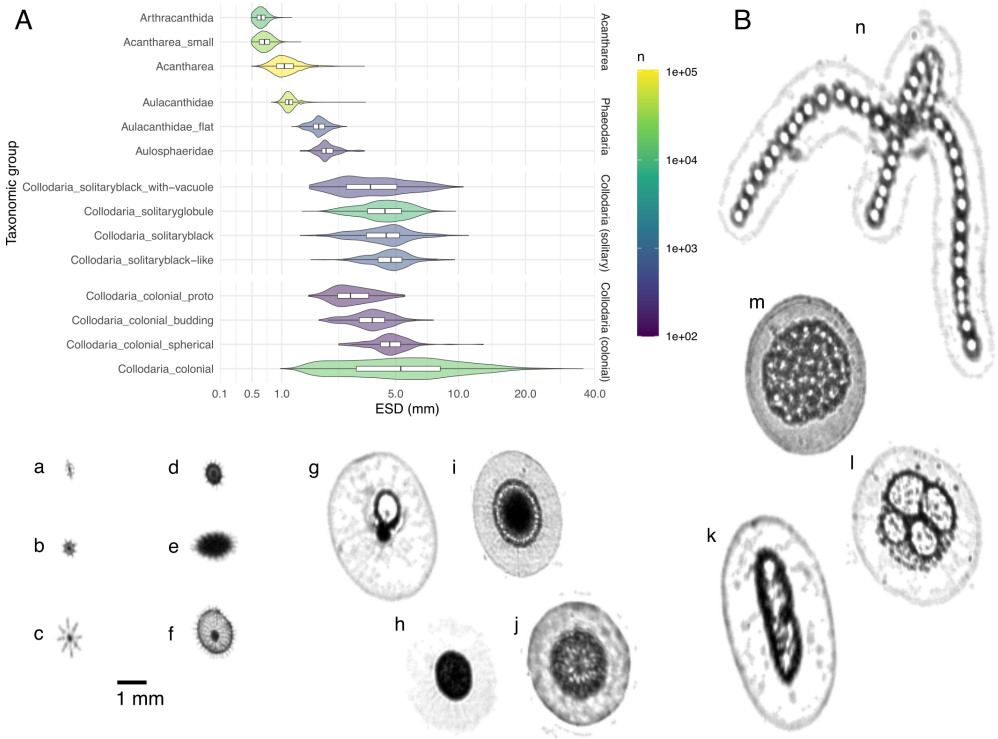
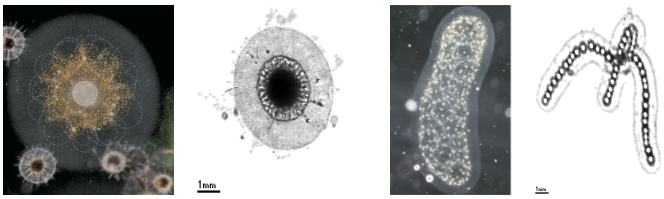


Figure 6.1: Dataset composition. **(A)** Equivalent Spherical Diameter (ESD) distribution and number of organisms per taxonomic group. **(B)** Examples of ISIIS images for each taxonomic group: (a) Arthracanthida, (b) Acantharia small, (c) Acantharia other, (d) Aulacanthidae, (e) flat Aulacanthidae, (f) Aulosphaeridae, (g) Collodaria solitary with vacuole, (h) Collodaria solitaryglobule, (i) Collodaria solitaryblack, (j) Collodaria solitaryblack like, (k) Collodaria budding, (l) Collodaria proto colony, (m, n) Collodaria colonial.

6.2.2 Vertical distribution of Collodaria depended on life stages

Collodaria

- solitary and colonial forms: two steps of the life cycle
- 100 µm - 3 m
- mixotrophic
- no shell



(A) **(B)** **(C)** **(D)**

Examples of Collodaria images in light microscope and imaged by the ISIIS. **(A, B)** Solitary Collodaria, **(C, D)** colonial Collodaria. (**A**) from Sardet [349], (**C**) from Biard and Ohman [40].

Collodaria are mixotrophic and lack a shell. Beyond the typical solitary and colonial forms, additional forms that could correspond to the transition phase between these were also detected by the ISIIS (Figure 6.2C). Overall, solitary forms were found close to the DCM, with the exception of vacuole-bearing organisms found deeper below the DCM; while colonies were more spread out in the water column although still centred on the DCM (Figure 6.2D). Lastly, the vacuoles of the 350 manually annotated solitary organisms demonstrated a clear orientation towards the surface (Figure 6.2B) and their size increased with depth (Figure 6.2A, S6.2).

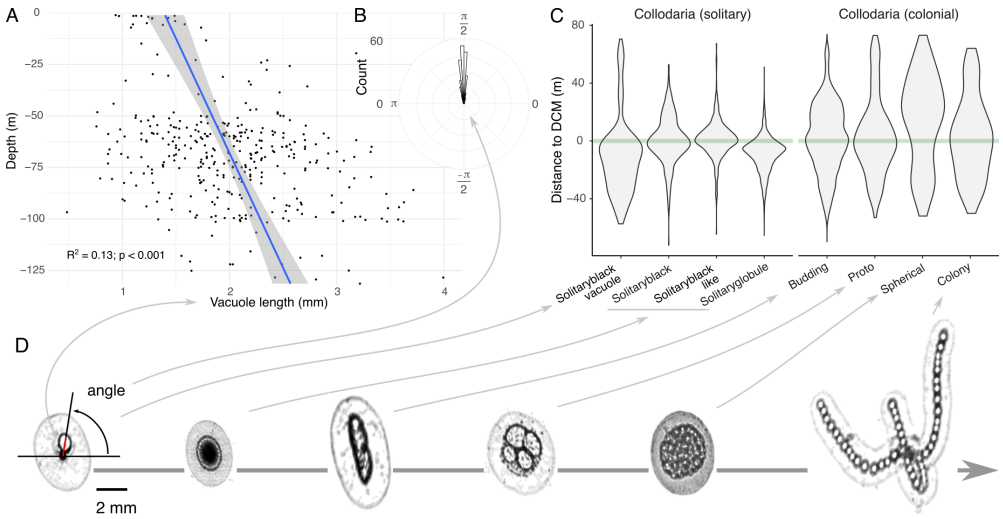
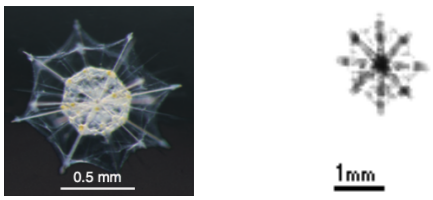


Figure 6.2: Collodaria distribution and vacuole properties. **(A)** Vacuole length (shown as a red line in **(D)**) VS depth in solitary Collodaria. **(B)** Angle of vacuoles position (shown in **(D)**) in solitary Collodaria. **(C)** Distribution relative to DCM for Collodaria groups. **(D)** Forms of Collodaria detected with the ISIIS and their putative chronological order.

6.2.3 Acantharia had disparate vertical distributions

Acantharia

- solitary only
- 50 µm - 1 mm
- mixotrophic
- strontium sulphate skeleton



(A) **(B)**

Examples of Acantharia images in light microscope and imaged by the ISIIS. (**A**) from Biard and Ohman [40].

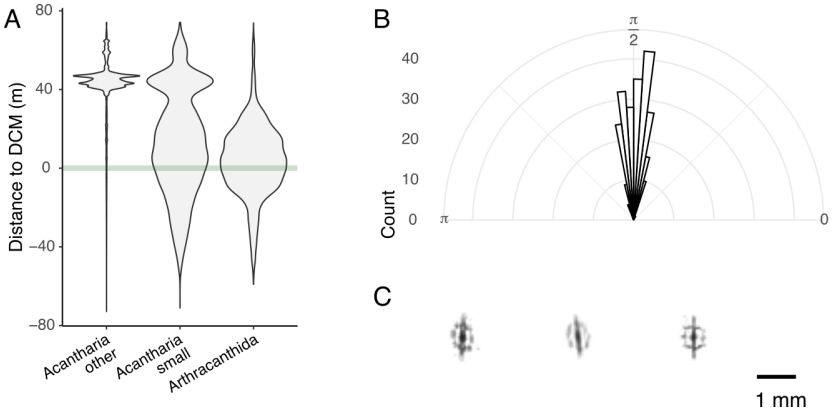


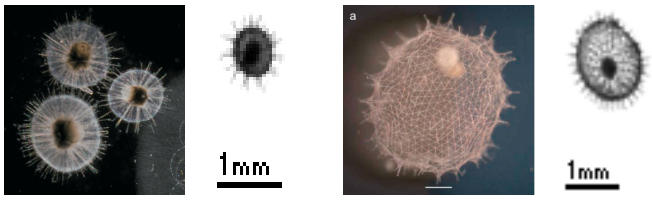
Figure 6.3: Acantharia vertical distribution and orientation. **(A)** Distance to DCM for Collodaria groups. **(B)** *In situ* orientation of Arthracanthida. **(C)** Examples of Arthracanthida images.

Acantharia are also mixotrophic but are smaller than Collodaria and bear a strontium sulphate skeleton. Three Acantharia subgroups were identified and displayed distinct vertical distributions (Figure 6.3A): Arthracanthida were found around the DCM, other Acantharia were found well above the DCM, just below the surface, while small Acantharia displayed an intermediate distribution between these, some organisms being close to the surface while others were around the DCM. Moreover, Arthracanthida displayed a preferential orientation with their largest spicule oriented at the vertical (Figure 6.3BC).

6.2.4 Phaeodaria differed from mixotrophic Rhizaria

Phaeodaria

- solitary only
- 100 µm - 3 mm
- heterotrophic
- siliceous skeleton



(A) **(B)** **(C)** **(D)**

Exemples of Phaeodaria images in light microscope and imaged by the ISIIS. **(A, B)** Aulacanthidae, **(C, D)** Aulosphaeridaea. **(A)** from Sardet [349], **(C)** from Biard et al. [39].

Unlike Collodaria and Acantharia, Aulacantha (Phaeodaria) lack photosymbionts and are heterotrophic. The two subgroups of Aulacanthidae were found around the DCM (Figure 6.4A), but seem to be brought downwards by sinking waters: when oxygen concentration is higher on the 28.7 isopycnal (corresponding to sinking water), Aulacanthidae concentration is lower (Figure 6.4B, S6.3). This was not the case for other Rhizaria dwelling in the DCM (Figure S6.4). Some organisms identified as Aulacanthidae had an ellipsoid shape and were bigger than typical spherical Aulacanthidae. These organisms presented the very same distribution as Aulacanthidae (Figure S6.3) but were all oriented horizontally (Figure 6.4C). Aulosphaeridae were found deeper, below the DCM (Figure 6.4A). In these organisms, the phaeodium was found to be typically positioned towards the bottom of the cell (Figure 6.4D).

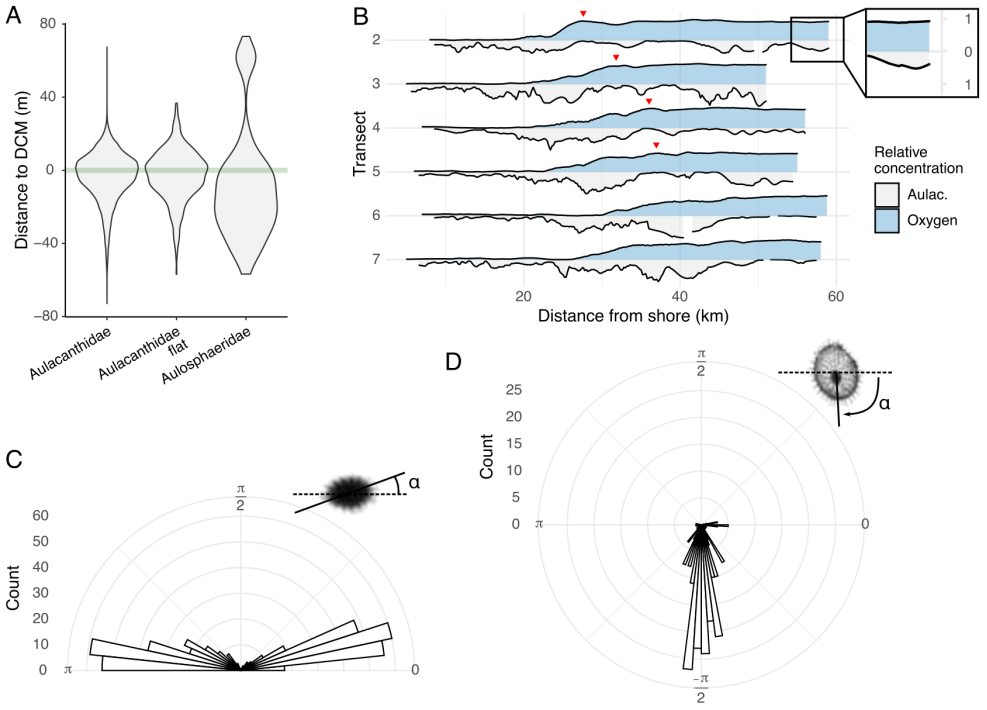


Figure 6.4: Phaeodaria distribution and orientation. **(A)** Distance to DCM for Phaeodoaria groups. **(B)** Relative Aulacanthidae and oxygen concentration along the 28.7 isopycnal. Red arrowheads highligth downwelling of high oxygenated waters corresponding to lower concentration in Aulacanthidae. **(C)** *In situ* orientation of flat Aulacanthidae. **(D)** Angle of phaeodium position in Aulosphaeridae.

6.3 Discussion

6.3.1 The complex life cycle of Collodaria

We observed Collodaria organisms with diverse morphology (solitary with vacuoles, proto colonies...). Here, we try to relate these observations to present knowledge regarding the life cycle of Collodaria, although much remains to be known. Previous studies demonstrated that solitary and colonial Collodaria are actually two steps in a complex life cycle [41, 324, 427] and that colonies can emerge from solitary cells by binary fission [56, 181]. Colonies have the ability to keep growing by a succession of binary divisions [11].

Our observations may contribute to filling gaps in the Collodaria life cycle.

Besides this ability for vegetative reproduction, multiple signs point in the direction of a sexual reproduction in Collodaria. First, both solitary and colonial forms of Collodaria have the ability to produce small biflagellated cells called swarmers [12, 181]. Yet, there is no evidence that these swarmers of unknown ploidy are gametes as there is no report of fertilisation or offspring production [12]. Swarmer production was also reported in Acantharia [100], Foraminifera [206] and Phaeodaria [210]. Second, sexual reproduction is common in other phyla of Rhizaria [12, 206, 286] but remains more hypothetical for Acantharia [98, 99]. Phaeodaria are thought to reproduce both asexually and sexually, but their life cycle has never been completed *in situ* [286]. Finally, as sexual reproduction is extremely widespread in the Eukaryotic world and was already present in the last eukaryotic common ancestor [377], it seems reasonable to suppose that Collodaria engage in sexual reproduction.

Collodaria, both solitary and colonial, produce swarmers.

In Collodaria, buoyancy loss and morphological changes occur concurrently to swarmers release [12]. Although it is not clear where swarmers are released exactly, there is indirect evidence that a hypothetical fertilisation might occur at depth. First, in both Acantharia and Foraminifera, fertilisation is thought to take place at depth [98, 206]. Second, Collodaria was previously detected at depth from metabarcoding [126, 313] while all species host photosymbionts and therefore typically dwell in the photic zone. Collodaria was also detected in the picoplankton (0.2-2 µm fraction size) [282], while all Collodaria are typically super-millimetric. These elements, very small and detected at depth, could correspond to swarmers. Finally, Collodaria swarmers were shown to contain a crystal of strontium sulphate in their cyto-

Fecundation may occur at depth.

plasm [423], which is lacking in the juvenile stages. While its precise function remains obscure, it may play a role sometime between swarmers release and the early stage of the next generation. This dense crystal could act as a ballast facilitating the swarmer's descent to depth [423].

The offspring has to ascend from the depths.

If fertilisation does occur at depth, the newly formed organism then has to reach the photic zone where adults live. Enter the vacuoles. The vacuoles in Collodaria are thought to contain lipids [12, 128]. All the vacuoles we detected in solitary Collodaria were oriented upwards, suggesting that they bring positive buoyancy to the cell. Moreover, vacuoles were smaller when approaching the surface, which is consistent with a diminishing need for buoyancy when organisms reach their target habitat. Thus, we hypothesise that vacuole-bearing cells could correspond to newly formed organisms, migrating from the fertilisation location to their next dwelling place: the DCM. Furthermore, deep (> 80 m) cells typically had one or two large vacuoles in the top half of the cell, while in cells closer to the DCM, more smaller vacuoles were present all around the cell (Figure S6.2). This could be a means to reduce ascent speed and stabilise when the cell reaches the DCM, a source of food and potential symbionts.

The solitary cell acquires symbionts de novo in the DCM.

In Collodaria, the number of symbionts per cell or in the colony varies a lot: a few hundreds in a solitary cell, several thousand for a colony [38]. No vertical transmission of symbionts is likely in Rhizaria, since swarmers are too small to host symbionts from their parent cell [99]. Moreover, Rhizaria often losing their symbionts just before gametogenesis is also in line with symbionts acquisition at the next generation [97]. This suggests a need for the *de novo* acquisition of symbionts, which in the oligotrophic waters of the Mediterranean Sea during summer, are only abundant in the DCM, where Collodaria could also feed on various planktonic organisms [41].

From a solitary cell to a colony by budding.

Thanks to these acquired energetic resources, a solitary cell can enter in a budding phase to create a new colony by vegetative multiplication [181] and the colony keeps growing through cell division [11]. In the meantime, symbionts are able to reproduce within the host [97], so that, at some point, the host does not need to acquire new symbionts. Our data highlighted that colonies are more spread out in the water column than solitary cells, and this is coherent with previous observations [124]. This might result from a lower pressure to stay close to the DCM, enabling the use of a larger habitat.

Overall, we cannot demonstrate that these steps constitute the life cycle of Collodaria, but our observations tell a coherent story. To confirm it, vacuoles bearing organisms could be sampled *in situ* to check for the absence of symbionts, since this stage is supposed to occur before symbionts acquisition. Yet, this seems infeasible given the very small number of observations: ~350 solitary Collodaria with vacuoles were detected in the $17{\times}10^6$ L sampled in 44 hours. Moreover, the low proportion of vacuoles bearing cells compared to the total number of solitary Collodaria (~10,000) suggests that the duration of vacuole stage would be very short compared to the time spent in the DCM. Another solution could be to use multispectral *in situ* imaging to detect fluorescence within organisms [135, 239, 425], but, once again, would require a very high sampling rate, of the same order of magnitude as the ISIIS (> 100 L s^{-1}).

How to turn the hypothesis into proofs?

6.3.2 Vertical distribution and buoyancy control in rhizarians

Although non-motile unicellular planktonic organisms were initially thought to be freely suspended in the water column [19], diverse buoyancy control mechanisms were later discovered, mostly consisting in the accumulation of substances modifying the cell density (e.g. lipid droplets in Rhizaria) or active change of the cell shape and volume (e.g. myoneme contraction in Acantharia [127]), but see [128] for a review. Beyond protists, cyanobacteria such as *Trichodesmium* can regulate their buoyancy through gas vacuoles [410], while density changes can also occur in Cnidaria and Ctenophora in order to reach equilibrium buoyancy [272].

Many buoyancy regulation mechanisms exist in the planktonic world.

Several Rhizaria categories had a vertical distribution centred around the DCM – a potentially favourable environment (for feeding, symbiont acquisition) – which is a probable indicator of an active buoyancy control. Such active buoyancy control was previously reported in Rhizaria [12, 128, 196, 271]. The solitary Collodaria vacuoles – systematically oriented towards the surface – and the relationship between vacuole size and depth mentioned above are other elements in favour of active buoyancy regulation.

Rhizaria actively control their buoyancy,...

The DCM is typically situated around the density gradient (Figure S6.1) that separates the nutrient depleted surface layer from a light-limited deep layer [175], thus corresponding to favourable conditions for phytoplankton growth. Still, neutral buoyancy and accumulation

. . . but passive buoyancy cannot be totally excluded for Phaeodaria.

on density gradients of phytoplankton cells can also contribute to creating DCMs [241]. Therefore, a passive accumulation on the same density gradient of other organisms also explains their distribution. This is likely for Aulacanthidae (Phaeodaria), the distribution of which seemed affected by small-scale downwellings, which could be a sign of very limited buoyancy control. Still, Aulacanthidae are thought to maintain their buoyancy thanks to a low carbon and biogenic silica density (their test is porous) [388] and Coelodendrid (Phaeodaria) can maintain neutral buoyancy *in vitro* [394]; to our knowledge, no active buoyancy regulation mechanisms were reported in the literature for Phaeodaria, contrary to other Rhizaria groups.

In contrast, those with photosymbionts should be close to the surface for a maximal exposition of photosymbionts to sunlight, but this environment is particularly depleted in nutrients, hence the importance of mixotrophy. Yet, the two mixotrophic groups we studied (Acantharia and Collodaria) had very different distributions. Three distinct vertical patterns were detected in Acantharia. However, the limited pixel resolution of the ISIIS (1 px = 51 µm) prevented a finer identification, which could have explained the bimodal distribution of small Acantharia. Larger, likely symbionts bearing, Acantharia were found very close to the surface, in concordance with literature [271].

Acantharia and Collodaria: two mixotrophs with different life strategies.

Collodaria – also mixotrophic – had a very different distribution from Acantharia: solitary cells were located around the DCM while colonies were more spread out (although colonies can accumulate close to the surface [393]). Multiple hypotheses could explain these two different strategies: Collodaria may not need as much sunlight exposition as Acantharia, Collodaria could not provide their symbionts with sufficient nutrients in such a depleted environment, or, Collodaria could be at a higher predation risk near the surface because they are much larger than Acantharia. Regarding the wider distribution of colonial Collodaria, this could emerge from the lower habitat pressure and a compromise between available nutrients at depth and exposition of symbionts close to the surface, nor can it be ruled out that colonies successively occupy these two habitats.

6.3.3 Preferential orientation of unicellular organisms

Similarly to vertical position, cell orientation in unicellular planktonic organisms was considered to be random because of small-scale turbu-

lence [132]. But cell orientation conditions various essential functions such as reproduction, sensing, metabolism or locomotion and non-uniform orientational distributions are the norm [19]. For non-motile unicellular organisms with homogeneous density, cell orientation is mediated by fluid-cell interactions at the microscale, so that cells adopt a hydrodynamically favourable orientation [19]. Otherwise, orientation can be mediated through differences of density of inner structures.

Non-random orientation is involved in various functions...

Multiple studies targeted the orientation of particles of various natures (marine snow aggregates, detrital material and phytoplankton, including colonies). This revealed a preferential orientation at the horizontal occurred in regions of low shear [288]. Moreover, time spent horizontally increased with the aspect ratio of particles, in accordance with predictions of Jeffery's theoretical model [195]. Similar results were reached on diatom chains *in situ* [255, 266, 396] and *ex situ* [198]. Moreover, the horizontal orientation of phytoplankton colonies seems ecologically beneficial as it increases the area exposed to sunlight, which could result in an increase of photosynthetic activity [58, 266]. In contrast, pennate diatoms were found to vertically reorient when sinking from surface turbulent waters [132].

...and was previously investigated in several unicellular organisms.

The use of *in situ* imaging with image captured from the side allowed us to resolve *in situ* orientation of multiple organisms. We detected preferential orientation both from the shape of organisms (oblate or prolate ellipsoid) or from the specific position of internal asymmetric structures in spherical organisms (e.g. lipid vacuoles, phaeodium). For example, positively buoyant lipid vacuoles were oriented towards the top of the spherical solitary Collodaria cells; while the phaeodium, a denser aggregate of waste and food [210], was located towards the bottom of the also spherical Phaeodaria cells. *In situ* preferential orientation was previously reported in protists, for Foraminifera that have bubble capsules positioned towards the surface [140].

In situ imaging can resolve Rhizaria orientation...

...either of internal structures...

Regarding the shape of the organisms, this literature cited above shows that a horizontal orientation is expected for non-motile plankton. This was the case for our oblate, flat Aulacanthidae. Nonetheless, Arthracanthida (Acantharia) displayed different preferential orientation, with the two longest and thickest spicules at the vertical. Their skeleton is made of strontium sulphate, the densest known oceanic biomineral (3.96×10^3 kg m^{-3}) [259]. Their vertical orientation could result from a passive equilibrium imposed by the weight of the skeleton, although the presence of an inner structure with a higher or lower density

...or of the whole organism.

We detected both horizontally and vertically oriented Rhizaria.

influencing cell orientation cannot be excluded. Moreover, Acantharia have been reported to actively deploy long cytoplasmic extensions that could be involved in predation by catching food particles, but with no certitude [258]. These structures could also be involved in other functions, including buoyancy control.

In the end, our results are the opposite of the historical view of a totally passive life of planktonic protists, even more so for the mixotrophic ones.

6.3.4 New insights on mixotrophy

Mixotrophy offers the possibility of colonising new habitats but imposes a more complex life cycle.

In 1851, Huxley was the first to describe yellow cells inside colonial Collodaria [187], also described in other polycystine Radiolaria a few years later [283]. But it was only later that these cells were identified as symbionts [56], concomitantly to the description of symbiosis in lichens. Since then, planktonic symbioses have received much less attention than others, although they play critical ecological roles in the oceans. Mixotrophy has appeared several times in the course of evolution, notably within several groups of eukaryotes [387], highlighting the benefits of this feeding strategy, although it comes with the need for the host to maintain a favourable environment for the symbionts to thrive. Here, we highlight that the transition between hetero and mixotrophy could shape the life cycle and distribution of Collodaria. We show that the mixotrophy enables the exploitation of habitat otherwise unfavourable for photosynthesis, albeit with different trade off in distribution in Acantharia and Collodaria. Finally, we also point out unexpected and often neglected behaviour of unicellular organisms.

6.4 Material and methods

Data was collected using the ISIIS during the VISUFRONT cruise in summer 2013.

The ISIIS was deployed during the VISUFRONT cruise that took place in July 2013 in the NW Mediterranean Sea, to study plankton distribution across the Ligurian Front. Sampling consisted of six transects performed across the front, perpendicularly to the coast, for a duration of 6 to 8 hours each, during which the ISIIS was deployed in a tow-yo fashion between the surface and 100 meters. The ISIIS is an imaging instrument targeting planktonic organisms from 250 µm to 10 cm in size, i.e. meso- and megazooplankton. With a sampling rate > 100 L s^{-1}. Moreover, the shadowgraphy method used in the ISIIS is particularly

appropriate to image transparent planktonic organisms, and their inner structure. However, organisms may be horizontally distorted depending on the towing speed and the acquisition rate of the line scanning camera of the ISIIS. In our case, organisms were laterally compressed so that a circle appears as a vertical ellipse and organisms were thus imaged smaller than in reality. This deformation was corrected when estimating the size of the organisms. Finally, the ISIIS also continuously records environmental data: temperature, salinity, fluorescence and oxygen.

Nearly 44 hours (equivalent to a 185 million pixels long image) of data was recorded with the ISIIS, the processing of which had to be automated. The first processing step consisted of detecting planktonic organisms in raw images. Using a content-aware segmentation pipeline based on a convolutional neural network (CNN) [306], more than 20 million potential planktonic organisms were extracted. During a second step, these organisms were automatically sorted into 24 taxonomic groups using a CNN classifier (MobileNetV2) previously trained and tested on ISIIS data (Chapter 3). The 1.8 M organisms sorted within Rhizaria were selected for a finer classification step into 14 categories, using a classification model specifically trained for this purpose (Figure 6.1). The identification of the 1.8 M images could not be manually validated. Instead, uncertain predictions – below a probability threshold computed so that 90% of mistakes occurred below this threshold – were discarded. This method decreased recall but improved precision, i.e. concentrations are underestimated but distribution patterns are preserved [124].

Imaged planktonic organisms were automatically detected and taxonomically sorted...

After classification, each organism was characterised morphologically by measuring a set of features, including proxies for size (e.g. area, perimeter), transparency (e.g. mean grey level) and *in situ* position (e.g. orientation of the major axis). Each element of the dataset consisted of an image, a taxonomy label, a set of features and associated environmental data, from which concentrations and vertical distributions were computed. Additionally, the positions of specific structures of three categories were detected manually using a keypoint annotation tool[1]. About 350 solitary Collodaria displayed asymmetric vacuoles. By recording the position of both the centre of the nucleus and the tip of the largest vacuole, vacuole orientation and size were computed for each organism. Similarly, the position of the phaeodium with respect

... and then individually measured for morphological features.

1 https://github.com/luiscarlosgph/keypoint-annotation-tool

to the centre of the organisms was recorded for 224 Aulosphaeridae (Phaeodaria). Finally, the orientation of Arthracanthida (Acantharia) was manually evaluated by recording the position of the extremities of the longest spicule in 232 organisms. All orientation values were corrected for the pitch of the ISIIS.

Environmental data was processed to highlight the main hydrographic features.

Abnormal environmental values (e.g. negative temperature and oxygen) were first removed. Density was computed from temperature and salinity. Finally, a bilinear interpolation of transects using distance from shore in x (200 m steps) and depth in y (0.5 m steps) was performed to obtain full images of transects. The DCM depth was computed on interpolated data, as the depth of highest fluorescence. Because it highlighted well the downwelling of oxygenated waters, the depth of the 28.7 isopycnal was computed from interpolated density transects, and oxygen and Rhizaria concentration values along this isopycnal were extracted to investigate whether Rhizaria distribution was affected by these vertical movements of water.

All analyses were conducted with `R` version 4.1.2. Data processing and interpolations were performed with packages `dplyr` and `akima` respectively. Plots were generated with `ggplot2` using the color-blind friendly `viridis` and `cmocean` color scales.

Acknowledgments

The authors acknowledge officers and crew of the R/V Tethys 2, as well as the additional scientists who took part in the cruise: F Lombard and M Lilley.

Funding sources

This study is part of project "World Wide Web of Plankton Image Curation", funded by the Belmont Forum through the Agence Nationale de la Recherche ANR-18-BELM-0003-01 and the National Science Foundation (NSF) ICER1927710. Data acquisition during the VISUFRONT cruise was funded by the Partner University Fund and supported by the French Oceanographic Fleet through ship time. This work was granted access to the HPC resources of IDRIS under the allocation 2021-AD011013092 made by GENCI. We are also grateful to the Roscoff Bioinformatics platform ABiMS (`http://abims.sb-roscoff.fr`), part of the Institut Français de Bioinformatique (ANR-11-INBS-0013) and

BioGenouest network, for providing computing resources. TP's doctoral fellowship was granted by the French Ministry of Higher Education, Research and Innovation (contract 3500/2019).

Supplementary materials

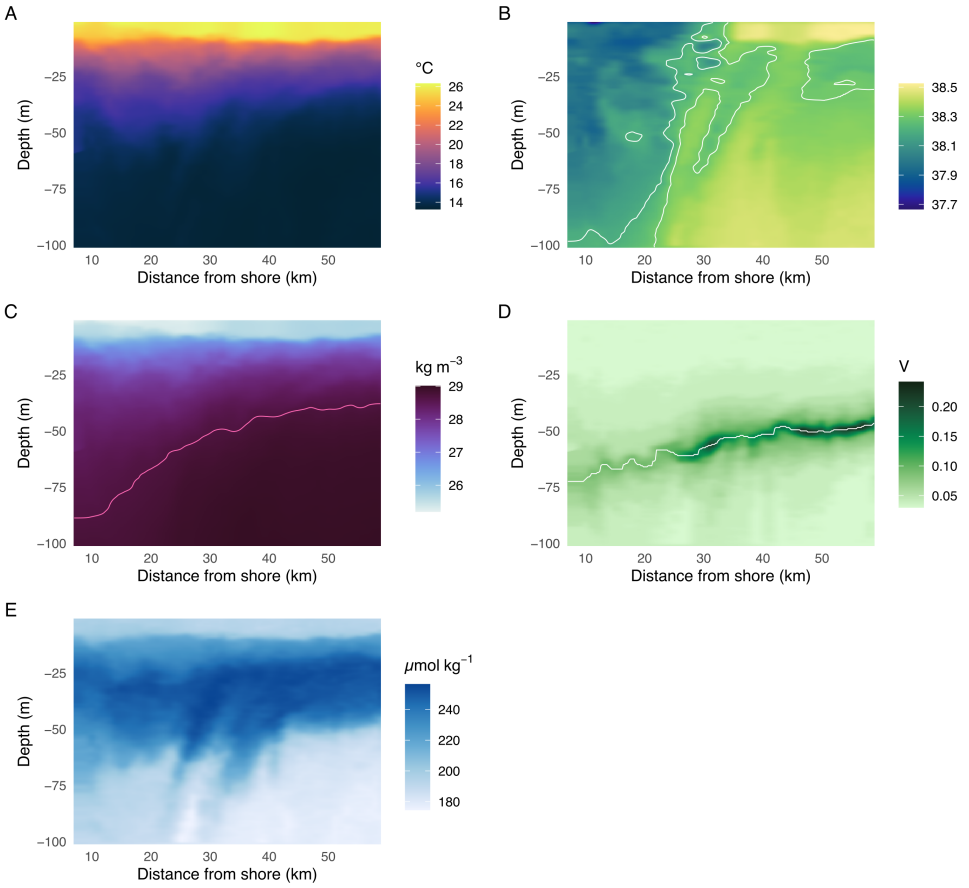


Figure S6.1: Environmental data along one transect, representative of the other transects. **(A)** temperature, **(B)** salinity with 38.2 and 38.3 isohalines delimiting the front, **(C)** density anomaly with 28.7 isopycnal in pink, **(D)** fluorescence with DCM represented as a white line, **(E)** oxygen.

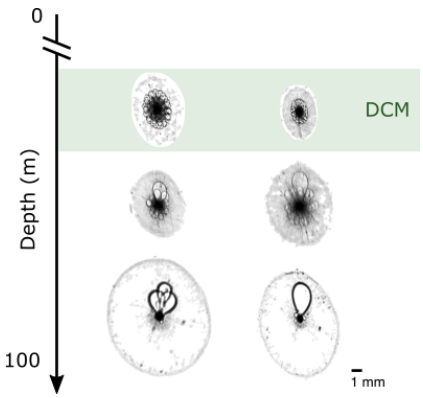


Figure S6.2: Schematic vertical distribution of size and shape of vacuoles in solitary Collodaria.

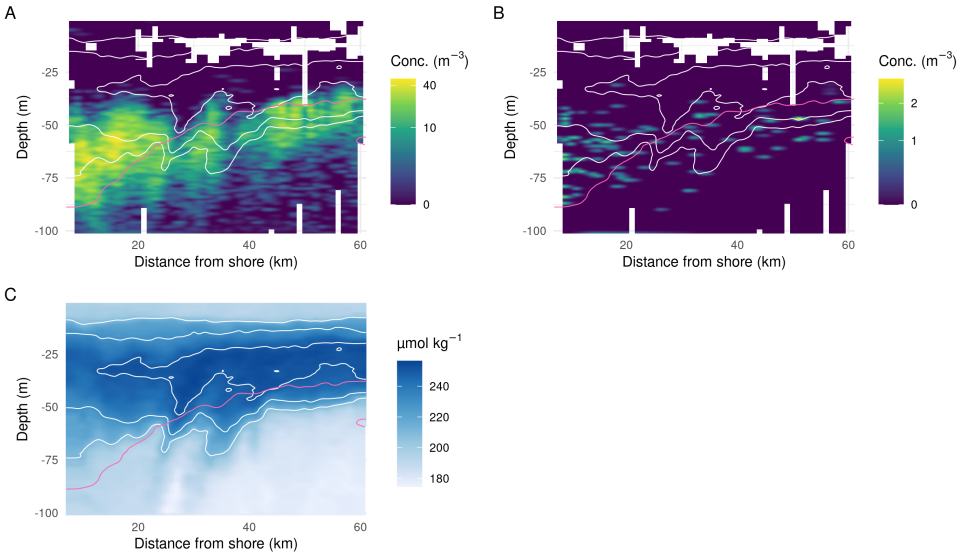


Figure S6.3: Effect of downwelling waters on Aulacanthidae distribution along a transect, representative of the other transects. Distributions of **(A)** Aulacanthidae and **(B)** Flat Aulacanthidae. **(C)** Oxygen concentration. Oxygen isolines are drawn in white and the 28.7 isopycnal is drawn in pink.

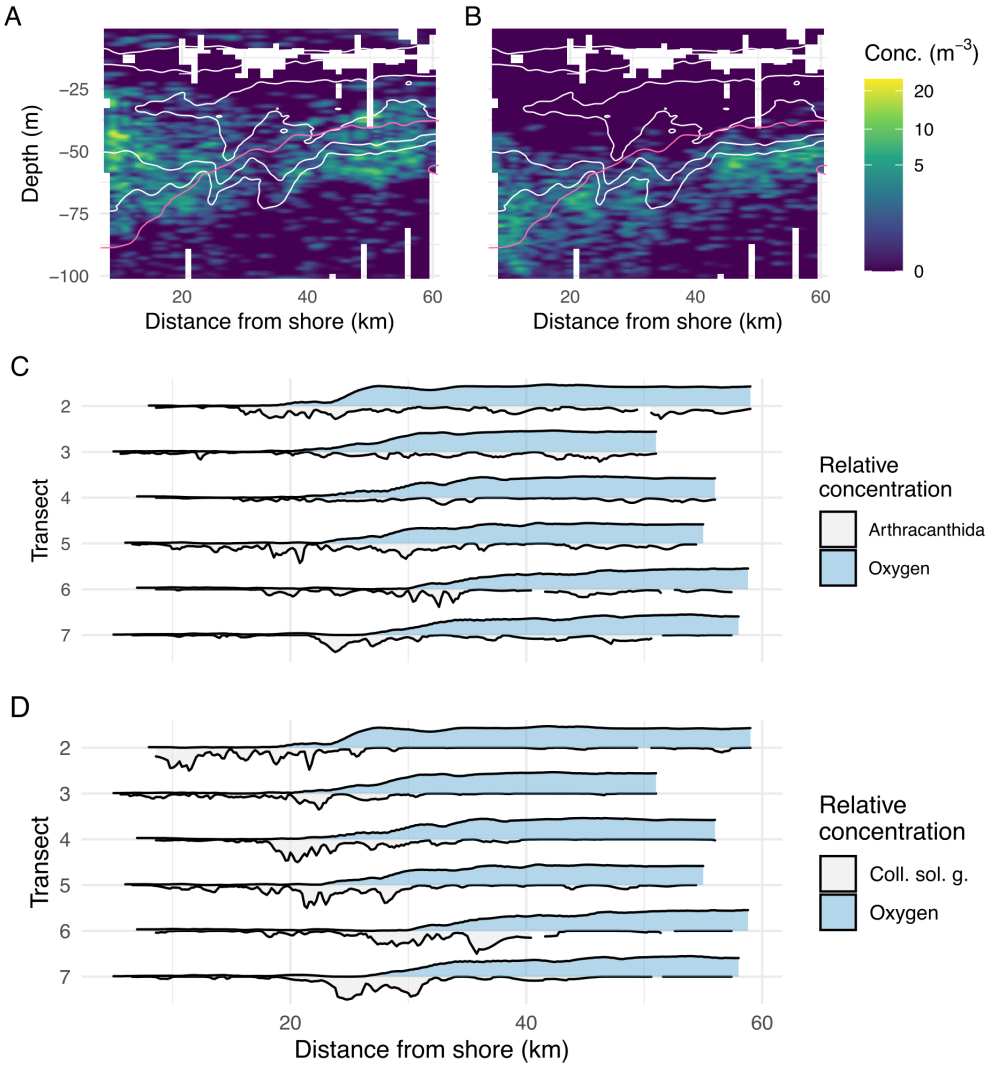


Figure S6.4: Absence of effect of downwelling waters on other Rhizaria. Distributions of **(A)** Arthracanthida and **(B)** solitaryglobule Collodaria. Oxygen isolines are drawn in white and the 28.7 isopycnal is drawn in pink. Relative **(C)** Arthracanthida, **(D)** solitaryglobule Collodaria and oxygen concentration along the 28.7 isopycnal.

Part IV

General discussion

After a brief summary of the main results, these are discussed in the light of the existing literature. Some considerations regarding the paradigm of data-driven science are addressed. Finally, a few perspectives of this work and their contributions to plankton ecology are examined.

7

Discussion and perspectives

7.1 Summary of key findings

Deep learning for...

In a first part, DL-based approaches were developed in order to ease the processing of large amounts of *in situ* imaging data. A CNN-based intelligent segmentation pipeline, based on the Detectron2 library providing state-of-the-art object detection and segmentation algorithms, was shown to be efficient to detect planktonic organisms in raw, full images. In comparison with other methods, this pipeline was not the fastest and required a decent GPU to be run, but it achieved the best compromise between accurate detection of planktonic organisms and relatively low pollution from non-planktonic objects. We successfully deployed and run this segmentation pipeline on several computing servers with various GPUs. Moreover, the pipeline was made open source[1] for anyone to use it with images from the ISIIS instrument.

...segmentation...

...and classification of plankton images.

To sort objects detected by the segmentation pipeline into taxonomical or morphological categories, we trained a classifier based on a CNN. Compared to a classic classification model, the CNN improved classification performance but only noticeably on poorly represented (a few hundred images) classes. Moreover, the comparison with a dummy random classifier highlighted the importance of assessing other metrics than the, often solely used, global accuracy when performing a classification task on unbalanced datasets, such as plankton image datasets.

Numerical ecology to...

...characterise global plankton community types,...

In a second part, numerical ecology tools were applied to study plankton distribution at different scales. From a global dataset of 2,500 vertical profiles using the UVP5, an *in situ* imaging instrument, we investigated the global distribution of large plankton community types in relation to their environment. Both in the epipelagic and the mesopelagic layer, we detected three types of community and concluded that they were driven more by basin-scale environmental

1 https://github.com/jiho/apeep

conditions than the very local conditions in which the profiles were performed.

. . . understand plankton and particles distribution across a mesoscale front. . .

Next, a 5-month campaign operating a glider equipped with an UVP6 was conducted across the Ligurian front (NW Mediterranean) to investigate plankton distribution across this mesoscale front during the spring bloom. During these five months, we detected large shifts in particles concentration and size and related them to temporal variations in the plankton community, during the bloom.

. . . and resolve fine-scale Rhizaria distribution.

Finally, leveraging high frequency *in situ* imaging data collected by the ISIIS, we investigated the fine-scale distribution as well as *in situ* position of organisms belonging to the Rhizaria clade, a group of understudied, fragile, mostly mixotrophic protists. We brought to light differences in vertical distribution between subgroups, likely underpinned by different life strategies. We also reported previously undocumented preferential orientation of some organisms, as well as observations of presumptive steps of the poorly known life cycle of Collodaria. These undescribed forms suggest the existence of a fine control of their buoyancy, even by these unicellular organisms, to reach the location of the next step in their life cycle.

In situ imaging + AI = ♡

Overall, this work highlights how the processing of large amounts of *in situ* imaging data can be made easier and faster thanks to artificial intelligence approaches; and demonstrates the effectiveness of *in situ* imaging data to understand biophysical interactions in plankton and distribution patterns at all scales.

7.2 *In situ* imaging to resolve plankton distribution across scales

7.2.1 Microscale – $\mathcal{O}$(1 mm)

At microscale, in situ imaging can resolve. . .

Interactions between individual planktonic organisms and with their environment occur at the microscale, $\mathcal{O}$(1 mm) [19]. At such a small scale, turbulence has many effects on plankton: nutrient uptake and encounter rates are increased, but duration of contact is decreased and feeding currents generated by suspension feeders are weakened, thus affecting growth rates and community composition [325]. Yet, few microscale studies were performed *in situ*, because few tools are adapted. Still, Font-Muñoz et al. [132] resolved *in situ* orientation and mating pairing of diatoms using Laser *In Situ* Scattering and Transmissometry

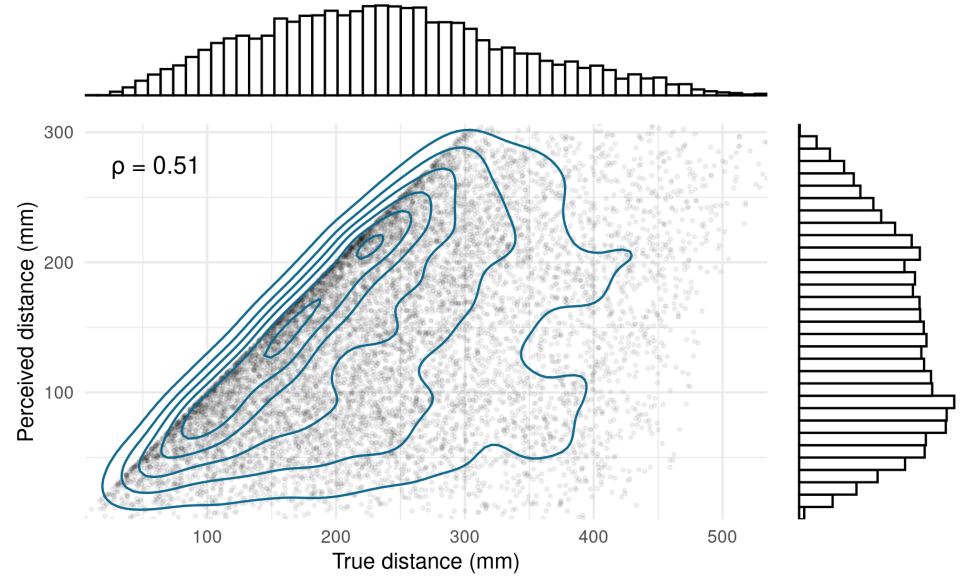
(LISST-100×) [260]. Using *in situ* holography, Talapatra et al. [396] were able to relate diatom chain orientation to mean shear rate; while Nayak et al. [288] observed a preferential horizontal cell orientation in relation with regions of low velocity shear. Indeed, holography can resolve the 3D position of planktonic organisms from a few microns to a few centimetres within the imaged volume [287]. Moreover, recent systems are now compact enough to be embedded on profiling or towed platforms, as well as on AUVs.

. . . individual distances between organisms. . .

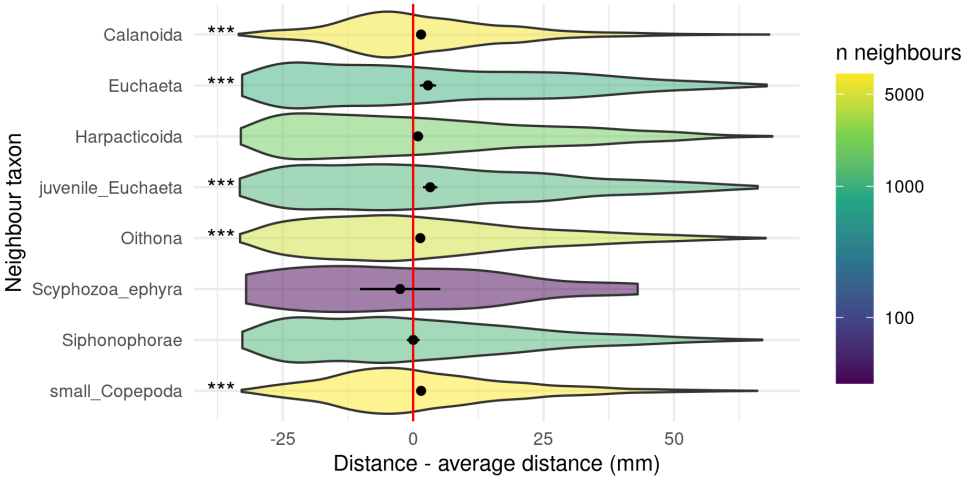
With *in situ* imaging instruments of high enough resolution, we can partially resolve microscale interactions (e.g. [152]) as well as the individual positions of planktonic organisms. Indeed, when several organisms are present in the field of view, distances between organisms can be computed and potential interactions inferred.

. . . which we investigated in ISIIS data,. . .

In ISIIS images, we considered sets of five consecutive 2048×2048 pixels frames, referred to as "images", corresponding to a volume of 52.5×10.5×50 cm^3. For each taxon (e.g. Euchaeta (Copepoda)), we selected images within which an individual was positioned in the middle frame, so that a wide array of distances to the neighbouring organisms could be computed before reaching the sides of the image. One thousand such "target" individuals were selected, per taxon, and distances from them to all others were computed. However, these distances were computed on a 2D images projected from a 3D volume, hence inducing an error. To evaluate it, 1000 points were randomly generated within the considered volume. One point, located in the central part, was considered as the target and distances to all other points were computed, either from the 3D position (true distance) or the 2D projection (perceived distance). This was repeated 10 times and revealed that perceived distances were well correlated to true distances although they underestimated them (Figure 7.1A). Note that the optical system of ISIIS is telecentric [85], which makes the projection orthographic and therefore does not induce an additional parallax error. Going back to organisms, our goal was to test whether some taxa were closer to the target taxon than others, which required a null hypothesis. We initially considered comparing to a random distribution of distances. But this was deemed inappropriate since living organisms are rarely infinitely close (they tend to "space out") while this was possible with a random distribution. We thus considered the average distance to all neighbours as the comparison point: the null hypothesis can be rejected if distances between the target and neighbours of a given taxon



(A)



(B)

Figure 7.1: *(Caption on next page)*

Figure 7.1: Individual distances computed within ISIIS images. **(A)** True vs perceived distance for 10 sets of 1000 objects randomly distributed in a space corresponding to one ISIIS image. Correlation was assessed with Spearman's rank. **(B)** Distribution of distances computed between Euchaeta (Copepoda, n = 26,578) and individuals belonging to 8 other taxonomic groups. Distances were standardised by the average distance to all objects, which is therefore represented by the red vertical line at zero. For each neighbour taxon, a Student's *t*-test was performed to assess whether the standardised distances differed from zero. * = 0.05, ** = 0.01, *** = 0.001. The colour bar shows how many neighbours of each taxon were detected and used to compute statistics, i.e. the sample size for each *t*-test. The black point represents the mean distance and the segment corresponds to the confidence interval computed from the *t*-test.

differ from distances to all neighbours, i.e. if distances minus the average distance is not equal to zero. All computed distances were thus standardised by subtracting the average distance between all targets and all their neighbours. A Student's *t*-test was then performed to assess whether these standardised distances differed from zero. Overall, for many taxa, we detected longer distances between organisms of the same taxonomic group that with other kinds of organisms, as shown for the Euchaeta in Figure 7.1B. This work was not mature enough to be included as a chapter in this manuscript but is an interesting avenue to follow up on, in the theoretical context of the ideal free distribution of organisms.

Moreover, if image acquisition is performed from the side (e.g. ISIIS, UVP6, zooglider) rather than from the top or the bottom, it gives access to the orientation of organisms with respect to the vertical plane. We took advantage of this in Chapter 6, where cell orientation of Rhizaria was investigated. Similar observations were previously reported in Foraminifera imaged with a zooglider [140]. Within ISIIS images, we also detected what seems to be a preferential orientation towards the vertical in harpacticoid copepods (Figure 7.2A), but this has not been further investigated yet. Such consistent orientation must provide an environmental benefit, whether for feeding, escaping predators or reproduction. For example, harpacticoids oriented towards the bottom could be following the plume generated by the marine snow particles they feed on [211], or could be harder to distinguish from the latter for potential predators. A vertical orientation was also detected in fibres imaged by the UVP6 (Figure 7.2B), which could correspond to the most hydrodynamically favourable position for sinking fibres.

. . . as well as in situ orientation.

(A)

(B)

Figure 7.2: *(Caption on next page)*

Figure 7.2: Vertically oriented objects captured by *in situ* imaging presented in the EcoTaxa web application [316]. **(A)** Vertically oriented harpacticoids (Copepoda) detected by the ISIIS. Most of them are oriented towards the bottom but a few of them are oriented towards the surface. **(B)** Vertically oriented fibres detected by the UVP6 mounted on the SeaExplorer glider, likely sinking (the larger part of the aggregate is located at the bottom, with a trail of matter above it).

Thus, *in situ* imaging in combination with artificial intelligence approaches can resolve some processes within microscale plankton ecology.

7.2.2 Fine-scale – $\mathcal{O}$(1-10 m)

In situ imaging can be used to investigate plankton thin layers,...

Plankton thin layers are features less than 5 m thick, in which plankton concentration is 1.5 to 3 times higher than the background concentration, and can horizontally extend across kilometres [325]. Plankton thin layers can be composed of diverse objects: phytoplankton, zooplankton, marine snow aggregates... [5, 269]. They are typically associated with vertical discontinuities in the water column (e.g. pycnocline) or found in regions of reduced flow [269, 325]. Formation can be mediated through biological mechanisms such as local growth or active swimming behaviour, but also involves physical processes such as fluid flow or accumulation on density gradients [112, 325]. Organisms of higher trophic levels (e.g. zooplankton, fish) were found to be associated with thin phytoplankton layers [31, 33], highlighting their potentially important role for the ecology of these consumers [112]. The study of the phytoplankton thin layers and their relation with zooplanktonic organisms was a potential topic of investigation in the context of this work, as the ISIIS data was of high enough spatial resolution for this purpose [156]. However, we were not able to detect such thin layers during the VISUFRONT campaign. First, as explained in Chapter 2, diatom fibres were not efficiently detected by our segmentation pipeline (Figure 2.3), notably because they are rather translucent and blend with the background, but also because they are easily confused with non-living fibres. Second, given the oligotrophic conditions that prevail in the NW Mediterranean Sea in summer, it is unlikely that any phytoplankton thin layer could form outside of the deep chlorophyll maximum (DCM).

...but such features were not detected during the VISUFRONT campaign.

Thicker than plankton thin layers [112], DCMs are also located around discontinuities, at the interface between the nutrient-depleted surface waters and the light-limited deep waters [175], where phytoplankton can grow. As an important source of food for zooplankton, DCMs are often associated with enhanced concentrations of zooplanktonic grazers [246], which may in turn attract their predators. In addition, detritivorous zooplanktonic organisms were previously reported to occur just below the pycnocline [246], a favourable location to catch sinking detritic particles formed inside the DCM. Passive accumulation of phytoplankton cells on density gradients can also contribute to creating DCMs [241] and such accumulations can also affect the distribution of zooplankton [150, 168]. In our study on the fine-scale distribution of Rhizaria across the Ligurian Front (Chapitre 6), we detected a preferential distribution of multiple Rhizaria groups centred around the DCM. Multiple clues were in favour of active buoyancy control by solitary Collodaria that allowed them to occupy the DCM; for them, it constitutes a source of food and potential symbionts. Detritivorous Aulacanthidae (Phaeodaria), on the other hand, were more likely to be passively accumulated on the same density gradient that created the DCM. Indeed, they were carried downwards by submesoscale subducting water masses.

Some Rhizaria were associated with the deep chlorophyll maximum,...

Beyond Rhizaria, Copepoda were also found to be more abundant around the DCM (Figure C.1). A total of ~8,5 million copepods were automatically extracted and identified from ISIIS images. We further separated them into five size classes between, 0.4 and 5 mm ESD (Figure 7.3A) Their distribution with respect to the DCM shows that smaller copepods tend to be above the DCM, while larger ones stay below (Figure 7.3B), and despite the existence of a diel vertical migration (DVM), this pattern was preserved at night. Such distribution patterns were previously observed from *in situ* electronic zooplankton counter data [176, 177], but high sampling-rate *in situ* imaging enables refining these observations. The asymmetric distribution of copepods is likely reflecting the distribution of phytoplancton cells, with smaller cells at the top of the DCM and larger cells at the bottom [218].

...while copepod distribution around the DCM was mediated by their size.

7.2.3 Submesoscale – $\mathcal{O}$(1-10 km)

As mentioned in the introduction, phytoplankton distribution is strongly affected by submesoscale features, especially in frontal zones [230].

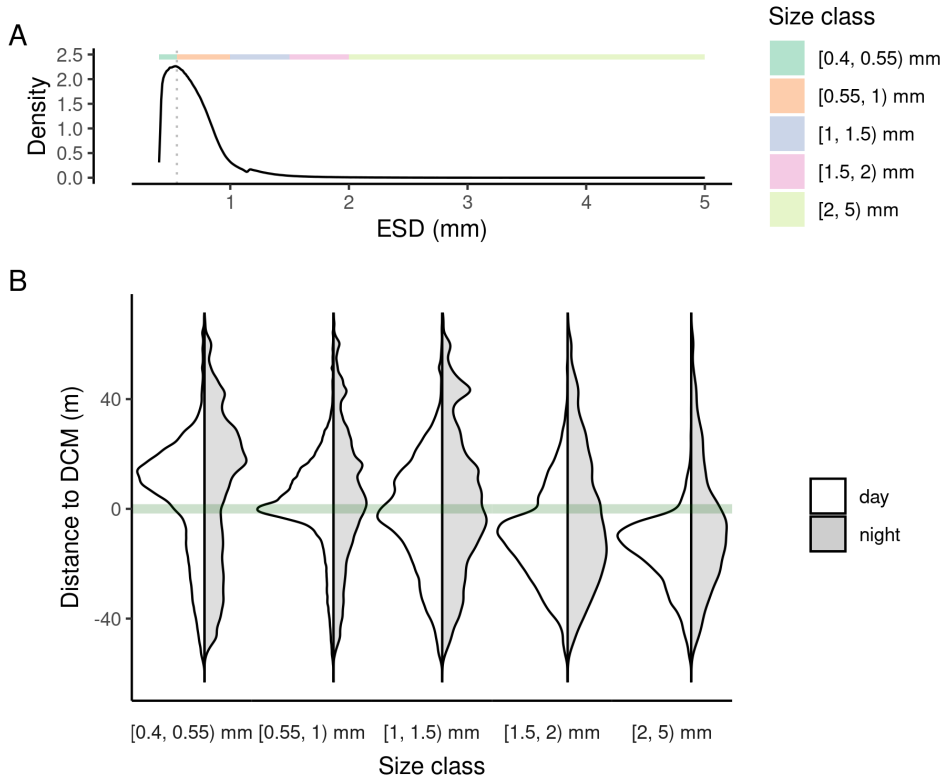


Figure 7.3: Vertical distribution of copepods with respect to the DCM. **(A)** Density estimate of copepods's ESD and representation of the five size classes chosen. The dotted vertical line at ESD = 0.55 mm highlights the size below which copepods are not quantitatively detected. **(B)** Vertical density distribution of five copepod size classes with respects to the DCM, at day and night. The green line represents the DCM.

Resolving plankton distribution across submesoscale features,...

...such as cross-frontal circulation.

Thanks to the ISIIS data collected during the VISUFRONT campaign, we were able to resolve the distribution of various plankton groups at the metre-scale. Among these, Aulacanthidae (Phaeodaria) were the only organisms whose distribution was clearly affected by submesoscale recirculation (see Figure C.1 in the appendix). While it is not surprising that swimming planktonic organisms can counter submesoscale vertical currents [269], it is much more striking for non-motile organisms (e.g. Rhizaria), and suggests active and efficient buoyancy control, even for single-cell organisms, as we hypothesised for Collodaria (Chapitre 6). Vertical velocities of downwelling waters measured

from acoustic Doppler current profiler (ADCP) would have been a great help to investigate our hypotheses. Unfortunately, two malfunctions prevented us from using this data: the ship's GPS was not connected to the ADCP and the gyroscope was not working, so that recorded data could not be corrected for the movements of the boat and was not usable for submesoscale analyses.

Here, only the data collected during the cross-front transects of the VISUFRONT campaign were analysed. Besides cross-front transects ($n = 7$), two other types of transects were performed to study other aspects of plankton distribution (Figure 7.4). Along-front transects ($n = 7$) were conducted both at dawn and dusk, in supposedly homogeneous environmental conditions, to observe the DVM as it happened. The distributions of planktonic organisms imaged during these transects are shown in Figure C.2. In addition, Lagrangian transects ($n = 14$) followed a water mass during 24h tagged with drifting surface buoys, to investigate plankton community changes inside the water mass and assess plankton swimming abilities. Distributions are presented in Figure C.3. While environmental conditions and distributions are rather homogeneous for the first transects, the last ones clearly crossed the front and this is reflected on the distribution of organisms, but this would require more investigation. Thus, there is still a lot of knowledge to extract from this extensive dataset.

But there is still much to extract from the VISUFRONT dataset.

7.2.4 Mesoscale – $\mathcal{O}$(10-100 km)

As described in Chapter 5, fronts are often the location of plankton aggregation [304]. While plankton distribution across large fronts can be investigated with net sampling (e.g. [48, 49]), *in situ* imaging can provide a more detailed view of such distributions [154, 248].

In situ imaging can resolve plankton distribution across fronts...

After performing repeated sampling across the Ligurian front during the spring bloom, we were able to detect a potential increase in zooplankton in the frontal area, and possibly a constrained distribution of planktonic organisms on one side or another of the front. Yet, the front did act as a barrier for particle distribution, as previously reported [148, 384]. However, the sampling rate of the UVP6-LP was likely too low to image enough planktonic organisms to definitely detect its effect.

...provided that a sufficient volume is sampled,...

Yet, when this very front was sampled at a much higher resolution with the ISIIS (> 100 L s^{-1}) during the VISUFRONT campaign, its effect was much clearer, although this data is only a snapshot of summer

...which was not the case for our glider-UVP6 campaign.

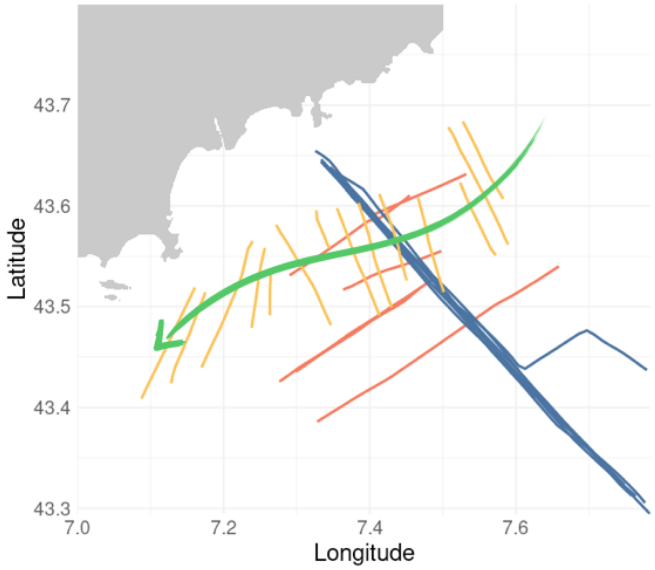


Figure 7.4: Transects performed with the ISIIS during the VISUFRONT campaign. Blue: cross-front transects, red: along-front transects, yellow: Lagrangian transects. The cross-front transect with an angular trajectory was not retained in our analyses. The green arrow represents the approximate location of the Ligurian current during the Lagrangian transects.

conditions. Some organisms were indeed constrained on one side or another of the front: Appendicularia, Doliolida, Hydrozoa, Pyrocystis and Siphonophorae were more abundant on the coastal side of the front (Figure C.1), in accordance with previous results [124]. However, this data did not highlight any accumulation of zooplankton at the front either.

When sampling is intensive enough, clear distribution patterns are detected.

7.2.5 Basin scale – $\mathcal{O}$(1000 km)

Finally, global plankton distribution can be investigated from *in situ* imaging data by aggregating coherent datasets from various locations. Such an approach requires standardised, inter-calibrated, commercially available instruments such as the UVP [317, 318]. For example, Kiko et al. [205] published a global dataset of particle size distribution from 8,805 UVP5 profiles. A similar dataset for planktonic organisms imaged by the UVP5 is nearing publication. This dataset supported the work presented in Chapter 4 and allowed computing a global estimates of carbon biomass from various taxa [111], a study in which I was also

Data aggregation can build global, coherent datasets.

Global UVP5 datasets for particles and plankton.

involved (see Appendix B). I personally contributed to the sorting effort to build this dataset, by validating ~250,000 images with the help of a taxonomic guide established specifically for UVP5.

While not as widespread as UVP, the ISIIS was nevertheless produced in several copies that have been deployed in many ecosystems around the world. This allowed the consolidation of a consistent enough dataset to carry out comparative studies, including one on doliolids [159]. For this study (Appendix B), I provided data for the Mediterranean Sea (n $\approx$ 80,000 doliolid images). This data was processed by the segmentation and classification pipelines presented in Chapters 2 and 3. Including data from six coastal ecosystems, this work revealed that the strongest driver of doliolid abundances was temperature, followed by chlorophyll *a* fluorescence and dissolved oxygen.

A global ISIIS dataset.

Thus, such datasets allow comparative studies at a much larger scale than allowed from a single oceanographic campaign, even though campaigns such as *Tara* Oceans [391] enabled studies at rather large scales [55, 374]. However, a single campaign necessarily has a strong spatio-temporal correlation, so that temporal effects are hard to disentangle from spatial effects.

In the future, the new miniaturised UVP6 [318] embedded on autonomous vectors such as gliders or floats will contribute to improving both spatial and temporal coverages of *in situ* imaging, including in harsh conditions that are difficult to sample with ships (e.g. winter months at high latitudes). However, images cannot be sent through satellite connection, along with other biogeochemical data collected because they are too big. Because floats are rarely recovered, images have to be classified onboard so that only numerical values of the concentrations of a few taxa are sent by satellite (a much smaller volume of data). Furthermore, the classification model must run in a power-limited environment. This prevents the use of CNNs, leaving only the classical approaches of extracting handcrafted features and using a simple classifier (gradient boosted trees in this case), although we showed that they were inferior, for small classes in particular (Chapter 3). Hence, only a few ($n = 20$), broad classes were retained. At this time, performance of embedded classification models on unbalanced datasets (as always in plankton imaging) are not reliable enough, except for a few groups (marine snow, copepods, *Trichodesmium*; F Ricour pers. comm.), so that such data will likely be more relevant for e.g. biogeochemistry studies but not for ecology. Both hardware and software progress will

Novative sampling methods to improve spatio-temporal coverage.

be required to improve embedded classification and enabling studies of the plankton community.

7.3 Ecology in the era of big data

Just like our daily lives, ecology, as a field of research, has also entered the era of big data.

7.3.1 High sampling rate imaging

Extensive sampling opens the door to unexpected discoveries.

Thanks to the ISIIS having the highest sampling rate of all plankton imagers (> 100 L s^{-1}, equivalent to one Olympic size swimming pool imaged every ~7h), we collected millions of plankton images. Among those, we can highlight a few remarkable examples of delicate objects (Figure 7.5). Beyond admiration and curiosity, it is, above all, an opportunity to discover new species, or organisms with special or unknown features. Indeed, given the long-tailed distribution of the abundance of living, and by extension of plankton [361], a higher sampling rate means a higher probability to sample rare organisms. Still, this requires high efficiency processing pipelines, since detecting rare objects can be like looking for a needle in a haystack. In our case, vacuole bearing solitary Collodaria (described in Chapter 6) could have past our attention if a well-informed eye (Tristan Biard's in that case) had not caught them, as they accounted for less than 3% of solitary Collodaria, themselves representing a tiny fraction of imaged organisms.

Increasing the chances of making new observations.

In the absence of such opportunistic detection, unexpected observations can be made through the meticulous examination of images. For example, Greer et al. [157] investigated batesian mimicry in fish larvae, from ISIIS images. Detecting these unusual morphologies of fish larvae required manually examining over 1 million images over hundreds of hours, to spot only a few hundreds fish larvae. Similarly, the study of Gaskell, Ohman, and Hull [140] relies on ~400 foraminifera, likely representing a tiny fraction of objects imaged in the water volume they observed(> 1000 m^3). Still, this allowed them to detect previously undocumented morphological features of these organisms.

To facilitate such observations, citizen science was explored for the detection and classification of plankton organisms, within ISIIS images, but the approach was not entirely conclusive, since most participants

were not familiar with the task of classifying plankton images [336]. Finally, automated approaches such as the use of vector embeddings to organise images in a reduced but meaningful space have the ability to detect non-standard objects [254], thus constituting an efficient way to discover previously unseen organisms.

7.3.2 Towards a data-driven ecology

A deluge of observation data

The development of more effective, versatile and cost-effective observation tools goes beyond imaging instruments, and the rate of data collection increased dramatically [208]. More and more observational data becomes accessible to ecologists [167]. This opened the era of data driven ecology, based on the exploratory analysis of large amounts of data to extract patterns and knowledge. It is considered as a fourth major scientific paradigm [179, 208]. This approach seems particularly adapted to address large-scale ecosystem questions [27]. In this context, real-time approaches, such as the T-MSER segmentation of plankton images presented in Chapter 2, are particularly appreciated. Numerical ecology methods, presented in the introduction (Section 1.4), as well as progress in computational power contribute to the ability to handle large amounts of data. Moreover, to ensure the reproducibility of studies in numerical ecology, ecologists should adhere to good practices. This involves coding according to best practices (e.g. uniform naming conventions, code description, simple writing, organised folder structure, version control), code sharing and data accessibility [9, 80].

and the same goes for oceanography.

Of course, the field of marine ecology is also affected by these changes: more and more sensors are embedded onto autonomous platforms [72] and, beyond plankton, instruments such as Baited Remote Underwater Video Station [257] or fish trawling cameras [8] are increasingly used. In plankton ecology, many imaging instruments exist [243], and databases are growing exponentially [192]. At the time of writing, 270 million objects are contained in the EcoTaxa database [316]. However, these imaging tools generate large amounts of data at the cost of some taxonomic resolution (organisms cannot be manipulated) and some objects even remain unidentified, which can lead to distorted analyses. Increasing the pixel resolution of *in situ* imagers is one solution to improve identifications. Moreover, when objects are imaged by different imagers deployed in the same area, differences in imaging techniques (e.g. reflected light for the UVP vs transmitted light for

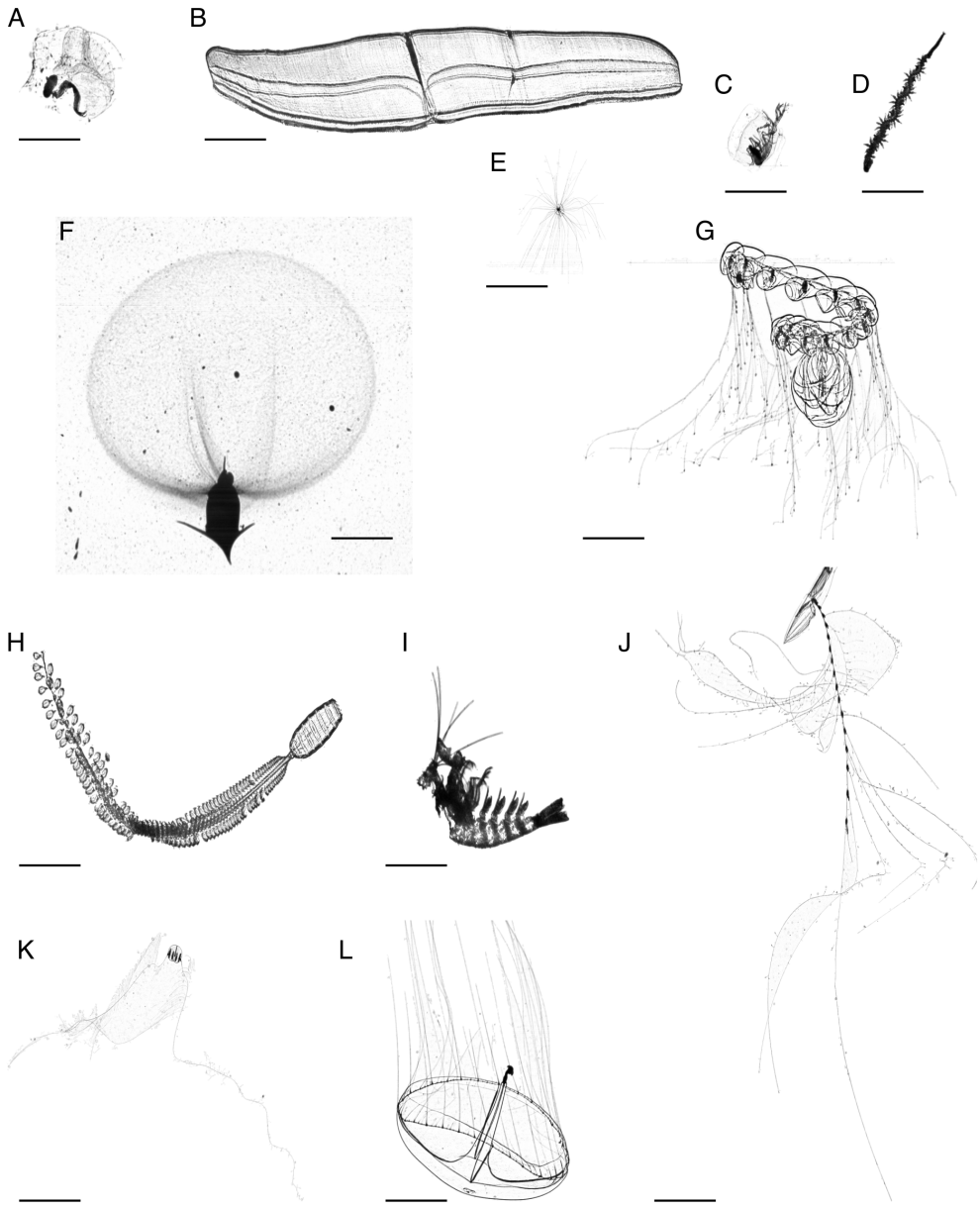


Figure 7.5: *(Caption on next page)*

Figure 7.5: A few remarkable examples of delicate objects imaged by the ISIIS. All scale bars represent 10 mm. **(A)** Appendicularia inside its house, with visible filters, **(B)** Cestidae (Ctenophora), **(C)** *Phronima* inside its barrel, **(D)** Annelida, **(E)** Foraminifera with expended pseudopods, **(F)** Thecosomata with an external feeding mucous mesh (see [143]), **(G)** Prayidae (Siphonophorae), **(H)** Doliolida nurse stage, with a cadophore carrying dozens of asexually produced gonozoids, **(I)** exuviae of Eumalacostraca, **(J)** Diphyidae (Siphonophorae), **(K)** Cydippida (Ctenophora), **(L)** *Geryonia proboscidalis* (Cnidaria).

the ISIIS) can bring complementary information to identify the objects. Finally, sharing images with other researchers can help lift the doubt on certain identifications, but is not always conclusive (Figure 7.6).

7.4 Methodological considerations

7.4.1 Efficient sorting of plankton images

UVP6 image classification could have been better handled.

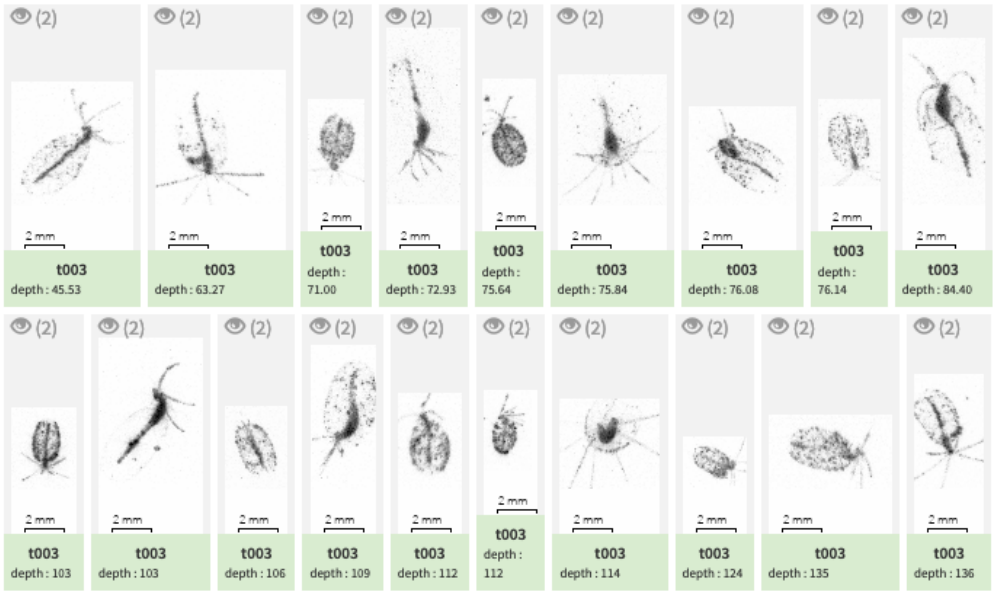
This paragraph is a recollection on how we handled the processing of UVP6 data, with the aim of highlighting what could have been done more efficiently. During the 5 months at sea, the UVP6 collected 1,123,123 images that had to be classified. Our initial plan was to classify some of these images manually, train a classification model on those, and then rely solely on model predictions for all other images. But we had not anticipated which images should be manually validated, nor how to ensure the quality of the predictions, resulting in a certain waste of time. If I were to do it again, here is how (Figure 7.7).

Choose a strategy.

First, one should determine if all images are going to be manually validated or if predictions will be used without human verification. In both cases, an automatic classification model is required, either to make manual validation easier and faster, or to use its predictions. For ~1 million images, manual validation after prediction by a classification model is achievable, but represents several months of work. Moreover, it is important to consider that this automated classified will be difficult to train because plankton imaging datasets are often strongly dominated by a few classes (marine snow particles in the case of the UVP). Some of the consequences of such an imbalance in the datasets were detailed in Chapter 3.

Then, in both cases, the second step is to generate a representative subset of the data, e.g. by selecting 1 out of n samples (profile, sampling

(A)



(B)

Figure 7.6: *(Caption on next page)*

Figure 7.6: Two sets of morphologically coherent but unidentified objects imaged *in situ*, presented in the EcoTaxa web application [316]. **(A)** Objects imaged by the UVP5 during the 2008 BOUM campaign in the Mediterranean Sea. They had a restricted distribution, in the centre of the basin, and were all around 70 m depth. These are probably some kind of larva. **(B)** Objects imaged by the UVP6 during the glider deployment conducted in across the Ligurian front. Their depth varied from ~50 to 200 m.

date, ...) equally distributed all across the data, with $n = 2^k,\ k \in \mathbb{N}$. marginparGenerate a representative subsample. This last rule allows to easily inflate the subset later if needed, by selecting $1/\frac{n}{2}$ samples, thus preserving the representativeness of the subset. At least part of this subset will be used as a training set for the automatic classifier, after being manually identified. Its size should be a compromise between the manual effort required and the need for enough examples, which depends on the selected strategy: at least ~100 objects per class of interest to guide manual validation, at least ~500-1000 objects per class of interest for robust predictions.

If the strategy is to fully validate the data, performance estimates of the classification model are not needed, but training the model on a representative subset of the data should result in better performance than with a non representative learning set, for example generated by validating "easier" (e.g. large planktonic organisms) images. This representativity is very important since automated classifiers tend to learn classes distributions as well; therefore, their predictive power diminishes when the class distribution is different between the training data and the new data to predict, a problem known as "dataset shift" [280]. For the same reason, if disproportionate attention is given to the validation of some classes, the model predictions will be biased towards them and will include objects from other classes (hence reducing precision). This is particularly significant if the lesser validated classes are the dominant marine snow particles: they will contaminate all other plankton classes.

Train a classification model...

If the strategy is to rely on predictions, we need to estimate their quality. Here, the fully validated subsample should be split into three parts, typically 70% for training, 15% for validation and 15% for testing. It is usual to stratify this random sampling by taxon/category, to ensure that, if the original subset is representative of the whole dataset, the smaller validation and test sets also are. The validation set will help to tune model parameters to get the best performance. The

...and assess its performance on new data if predictions are to be used.

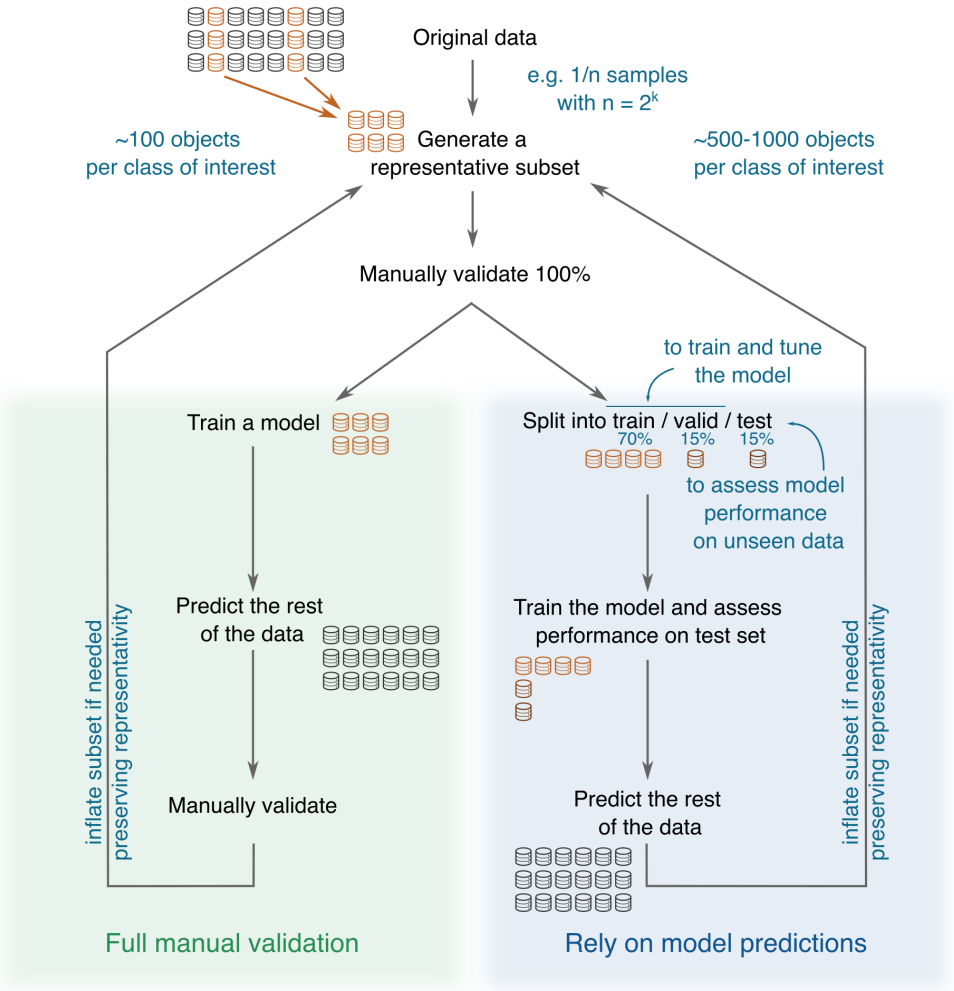


Figure 7.7: Suggestions for efficient pipelines for sorting plankton images. Two strategies are possible: manually validate all images (green) or rely on predictions generated by a classification model (blue).

test set will be used to compute the final performance estimate. As models tend to perform better on the data they were trained on, a phenomenon called "overfitting" [108], it is crucial that this test set is never seen by the model during training to avoid overestimating model performance [192]. Finally, for this performance to be estimated precisely, the test set should contain a sufficient number (~100) of examples in classes of interest, which should be the case if the original subset was large enough.

Inflate the subsample if needed.

In both cases, if the quality of predictions is not satisfying, the initial subsample can be inflated. It should continue to be representative of the whole data: if the $1/n$ (e.g. $1/16$) subsample was too small, the $1/\frac{n}{2}$ (e.g. $1/8$) subsample can be easily generated and preserves both previously validated objects as well as the representativeness of the whole data.

7.4.2 Making the most of *in situ* imaging

The striking results we obtained on Collodaria...

Among the many things I was able to explore during this PhD, the result I find the most striking is the differential distribution of the life stages of Collodaria, especially the buoyancy control it involves. Single-cell organisms, ascending from the depths, taking advantage of bubble-like alveoli that appear to function like hot-air balloons; this is *fascinating*. However, as explained in Chapter 6, we do not demonstrate such process, we just tell a coherent story based on our observations. I my opinion, refining these observations to confirm (or infirm!) our hypotheses would be a very interesting research topic. Besides, additional questions arise: where do solitary Collodaria acquire their symbionts when there is no DCM? How does their concentration vary seasonally? Is sexual reproduction seasonal? Do colonial forms prevail when new symbionts cannot be acquired?

...deserve, in my opinion, to push our analyses further.

As we have seen, the ISIIS is a great instrument to image large volumes of water and thus detect rare organisms, without disturbing them. Hence, this seems to be the perfect instrument for this kind of study, as solitary cells with vacuoles were particularly scarce. Still, an efficient detection algorithm should be developed to specifically target Collodaria, both solitary and colonial forms, to efficiently go through the large amounts of data ISIIS generates. In addition, it would be particularly interesting to be able to resolve the composition of phytoplankton community in the immediate proximity of Collodaria.

This could be done with fluorescence sensors detecting other pigments than chlorophyll *a* or with other imaging instruments targeting smaller sizes and could bring information regarding the preferred potential symbionts or food source of solitary cells. Surveys should be conducted at various seasons to describe the variations in the concentrations of the diverse phases of the life cycle of Collodaria through time. Moreover, this could guide additional sampling to potentially detect Collodaria swarmers, during the sexual reproduction phase. Finally, the icing on the cake would be to physically sample one of these vacuole-bearing cells intact and be able to look for symbionts inside. Although this seems extremely challenging, it would contribute to demonstrating that, as we suppose, these cells are newly formed and still free of symbionts.

Overall, I think that the combination of *in situ* imaging and AI-based methods is a powerful approach, with a lot of potential to address unresolved questions in plankton ecology.

Part V

Appendix

Additional elements that could not be included in the main text of the manuscript are presented here. This includes a summary of this work in French, an overview of collaborative works and the results of preliminary analyses conducted on the VISUFRONT dataset.

A

Résumé de la thèse en français

Ce chapitre présente un résumé de la thèse en français. Après avoir introduit les notions nécessaires à la compréhension du sujet, deux parties de méthodologie traitent de l'application de méthodes d'intelligence artificielle au traitement d'images de plancton. Suivent trois parties de résultats écologiques abordant la distribution des organismes planctoniques à différentes échelles, de l'échelle globale à la subméso-échelle. Enfin, les résultats sont discutés en regard de la littérature.

A.1 Introduction

A.1.1 Échelles dans les processus océaniques

A.1.1.1 *De la circulation globale à la petite échelle*

Des océans en mouvement à toutes les échelles,...

Les océans sont en mouvement à de nombreuses échelles spatiales et temporelles [72]. À l'échelle globale, les océans sont mis en mouvements par la circulation thermohaline [328]. Ces courants sont porteurs de grandes quantités de chaleur et ont ainsi un effet majeur sur le contrôle du climat [329]. À méso-échelle – $\mathcal{O}$(10 - 100 km) – les mouvements prennent principalement la forme de tourbillons [334] et de fronts [26]. Les fronts sont des zones de rencontre entre masses d'eau avec des propriétés différentes. Ils existent à des échelles spatiales et temporelles variées, mais certains sont permanents. Les zones frontales sont souvent associées à des courants convergents de surface, résultant en une augmentation de la diversité et de la biomasse à tous les niveaux trophiques [26, 304]. Longtemps négligées, les structures à subméso-échelle – $\mathcal{O}$(1 - 10 km) – sont de mieux en mieux résolues grâce à des moyens d'observations plus performants et à la modélisation [229]. Par ailleurs, les dynamiques à subméso-échelle sont souvent associées à des fronts à méso-échelle, où elles sont responsables de mouvements d'eau verticaux [228, 230] (Figure A.1). Des mouvements existent aussi à plus petite échelle – $\mathcal{O}$(1 mm) – mais leur contribution aux processus

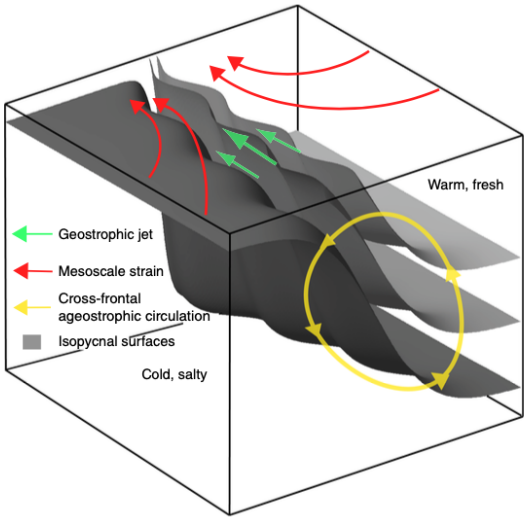


Figure A.1 : Les dynamiques frontales à subméso-échelle. Le front est représenté par les isopycnales obliques, délimitant des eaux froides et salées au large d'eaux plus chaudes et moins salées à proximité de la côte. Une circulation agéostrophique cross-frontale (flèches jaunes) a lieu dans le sens de l'aplatissement des isopycnales. De Lévy, Franks et Smith [230].

à plus grande échelle reste peu comprise [281]. Ainsi, les processus décrits ici couvrant 9 à 10 ordres de grandeur, des outils d'observations complémentaires sont donc nécessaires à leur étude. De plus, tous ces processus affectent les organismes peuplant les océans, des plus petits aux plus grands.

A.1.1.2 *Effets sur les organismes marins*

. . . ce qui se répercute sur les organismes marins,. . .

Ces effets sont particulièrement remarquables sur les ressources exploitées, telles que les pêcheries [297]. Les top prédateurs (thon, éléphant de mer. . .) semblent aussi exploiter les structures à méso-échelle et subméso-échelle pour leur prédation [352]. Les dynamiques à subméso-échelle associées aux fronts génèrent des mouvements verticaux qui affectent le taux de croissance du phytoplancton, via la redistribution des cellules et des nutriments [230, 252]. Cette augmentation de biomasse du phytoplancton peut se propager aux niveaux trophiques supérieurs dont les organismes zooplanctoniques [275, 294] ou le poisson fourrage [32]. Cependant, peu d'informations sont disponibles

quant à la distribution à fine échelle de ces organismes. Or, de telles informations permettraient de comprendre comment l'écosystème est affecté par les dynamiques à subméso-échelle, qui bien qu'étant spatialement et temporellement restreintes [230, 252], prennent place à la même échelle que les processus de croissance du phytoplancton et doivent donc tout particulièrement affecter la productivité de ce dernier et par extension l'écosystème planctonique [230]. Ainsi, si tous les organismes océaniques semblent être affectés par ces processus, cela est d'autant plus vrai pour ceux qui dérivent et ne peuvent nager efficacement contre les courants et ne peuvent ainsi choisir leur habitat.

. . . du phytoplancton aux niveaux trophiques supérieurs.

A.1.2 Le plancton, dérivant au gré des courants

A.1.2.1 *Importance écologique du plancton*

Les organismes planctoniques sont définis comme étant incapables de lutter contre les courants. Cette définition basée sur la niche écologique englobe ainsi une grande diversité taxonomique, ainsi qu'une grande diversité de taille [69, 186] (Figure A.2). Le plancton joue des rôles écologiques clés : les organismes photosynthétiques du phytoplancton produisent environ la moitié du dioxygène atmosphérique [131] et sont les producteurs primaires à la base des réseaux trophiques océaniques [125]. Les organismes du zooplancton interviennent quant à eux dans la pompe à carbone biologique qui contribue à la séquestration du carbone organique vers les profondeurs [246] et constituent également un maillon trophique entre le phytoplancton et les niveaux trophiques supérieurs [413]. Les organismes planctoniques sont particulièrement sensibles aux conditions environnementales dans les masses d'eau dans lesquelles ils sont insérés, et constituent donc de bons indicateurs de potentiels changements [172].

Une grande diversité. . .

. . . et des rôles écologiques clés.

Le plancton est un bon indicateur de la santé des écosystèmes.

A.1.2.2 *Patrons globaux de la distribution du plancton*

Ainsi, la distribution et la diversité du plancton sont largement gouvernées par les conditions environnementales (température, nutriments, lumière. . .) [172]. Ces conditions variant fortement avec la latitude, il en résulte des gradients de biomasse et de biodiversité liés à la latitude : la biomasse est plus élevée aux hautes latitudes [190] tandis que l'inverse est observé pour la diversité [188, 340, 345, 399].

Les schémas de distribution du plancton à grande échelle sont relativement connus. . .

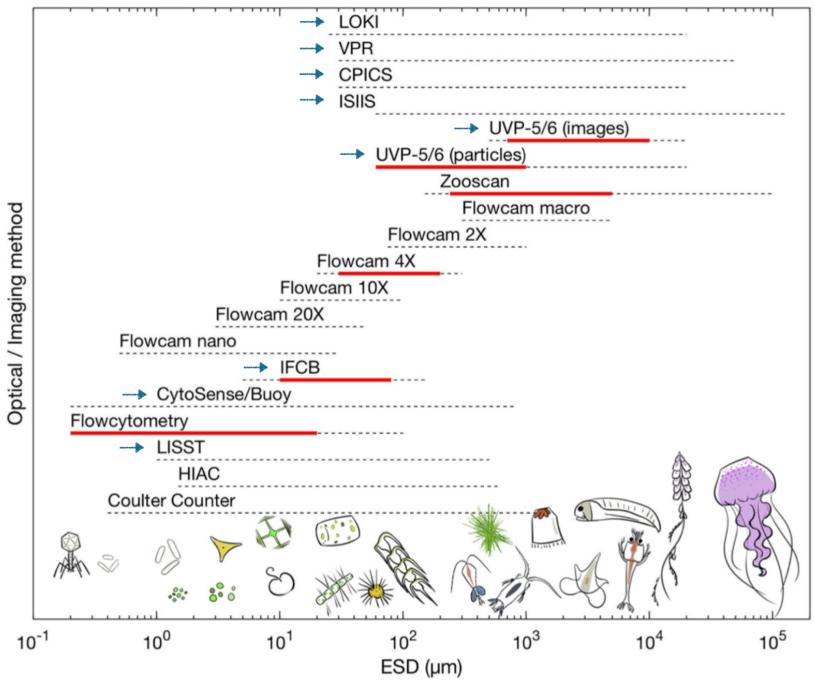


Figure A.2 : Gamme de taille en diamètre sphérique équivalent (ESD) couverte par les principaux instruments d'imagerie du plancton. Les lignes pointillées représentent la gamme de taille opérationnelle totale, tandis que les lignes rouges représentent la gamme dans laquelle peuvent être obtenues des informations quantitatives. Les flèches bleues indiquent les imageurs *in situ*. De Lombard et al. [243].

Bien que la répartition globale du plancton soit connue dans une certaine mesure, il reste encore beaucoup à découvrir concernant la distribution à fine échelle, notamment pour le zooplancton. Ce manque de connaissances provient essentiellement d'un échantillonnage non adapté à cette échelle : les outils classiques (pompes, filets...) intègrent les organismes pendant le prélèvement et ne permettent pas de connaître les conditions environnementales dans leur voisinage immédiat [30, 243, 330]. De plus, l'étude de structures à fine échelle spatiale et temporelle requerrait un échantillonnage répété, et donc coûteux.

...mais des lacunes persistent dans la connaissance de la distribution à fine échelle.

A.1.3 Distribution du plancton à fine échelle

A.1.3.1 *Intérêt écologique*

Les structures à subméso-échelle sont susceptibles d'influencer la distribution des organismes planctoniques : comme mentionné ci-dessus, les courants verticaux peuvent redistribuer les nutriments et les cellules phytoplanctoniques, les déplaçant dans ou en dehors de la zone euphotique dans laquelle se produit la photosynthèse, affectant le taux de croissance du phytoplancton, tandis que les courants horizontaux peuvent transformer les patchs en filaments. Ces changements sont susceptibles de se propager aux niveaux trophiques supérieurs (zooplancton, poissons...) [230]. En effet, les interactions trophiques et reproductives du zooplancton se font à l'échelle des organismes (µm à cm). Ainsi, une concentration locale de phytoplancton, par exemple dans une couche fine [112], a des conséquences plus immédiates sur la survie et le développement du zooplancton que la concentration moyenne en chlorophylle a dans la région. Ainsi, l'étude de la distribution du zooplancton à des échelles fines, en relation avec les dynamiques à subméso-échelle, devient pertinente pour comprendre les processus qui régissent sa distribution à l'échelle régionale.

Les processus pertinents pour expliquer la distribution du plancton se déroulent à une échelle fine.

A.1.3.2 *Les outils à disposition*

Comme mentionné auparavant, les outils classiques d'échantillonnage du plancton ne sont pas adaptés pour étudier sa distribution à fine échelle, sans compter que la plupart peuvent endommager certains organismes qui sont alors sous-estimés [330]. Le développement d'outils d'imagerie *in situ* a permis de surmonter (au moins partiellement)

L'imagerie in situ permet des études à fine échelle...

ces limitations : tout d'abord ils permettent de connaître la position exacte des organismes et peuvent échantillonner les conditions environnementales au voisinage immédiat des organismes, mais ils permettent également d'étudier des éléments fragiles comme les Rhizaria [42, 104], le plancton gélatineux (Cnidaria, Ctenophora) [250] ou même des particules de neige marine [161, 162, 401].

. . . grâce à une large gamme d'instruments.

De nombreux imageurs *in situ* ont été développés au fil du temps [243]. Ces outils sont très variés de par leur type de déploiement (long-terme, profil vertical. . .), la gamme de taille ciblée, l'éclairage ou encore leur intégration sur des plateformes autonomes. Ensemble, ils couvrent une majeure partie de la gamme de taille du plancton (Figure A.2) et permettent d'envisager des stratégies d'échantillonnage variées.

Ils permettent d'aborder de nouvelles questions écologiques. . .

Les données à haute résolution spatiale et temporelle générées par ces instruments permettent d'aborder des questions écologiques qui étaient auparavant hors de portée, telles que la distribution du plancton à fine échelle en lien avec les conditions environnementales au niveau de fronts [59, 124, 154, 248, 265] ou de tourbillons [77], les propriétés des patchs de plancton [337], ou encore les interactions entre les couches de zooplancton et de phytoplancton [153, 156, 355]. Enfin, des interactions (compétition, parasitisme. . .) entre organismes ont pu être observées directement dans les images *in situ* [152].

...et même faire la lumière sur le comportement du plancton.

De plus, certains de ces instruments, comme l'*In Situ* Ichthyoplankton Imaging Sytem (ISIIS) [85] ou l'Underwater Vision Profiler (UVP) [317, 318], peuvent capturer des images d'organismes planctoniques sans les perturber, pouvant ainsi révéler une position particulière, un comportement ou des interactions avec d'autres organismes [299]. Par exemple, Ohman et al. [293] ont observé le comportement alimentaire des copépodes. Chez les rhizaires (Eucaryotes unicellulaires), une orientation préférentielle [140] et un comportement de prédation potentielle [258] ont également été détectés. Au-delà du comportement, les traits individuels (taille, opacité, statut reproducteur. . .) peuvent également être mesurés à partir d'images *in situ* : la morphologie et l'activité des copépodes varient avec la fonte de la glace en baie de Baffin [409], la morphologie des particules de neige marine change pendant le bloom de printemps [401].

Une avalanche de données à traiter.

Bien que les outils d'imagerie *in situ* échantillonnent généralement de plus petits volumes que les filets [243] (à l'exception de l'ISIIS, > 100 L s^{-1} en général), leur nombre croissant et leur facilité d'utilisation génèrent un volume de données de plus en plus important, si bien que

le traitement de ces données devient un goulot d'étranglement [253]. Par exemple, une heure de déploiement d'ISIIS génère 100 milliards de pixels (~11 millions d'objets) [192]. Pour traiter efficacement une telle quantité de données, les écologistes doivent désormais se tourner vers des méthodes numériques automatisées.

A.1.4 Écologie numérique du plancton

L'écologie numérique est un champ de recherche qui consiste à appliquer des méthodes statistiques et numériques pour répondre à des questions écologiques [226]. Devant la quantité croissante de données collectées, ces approches sont de plus en plus pertinentes [315]. Les progrès computationnels les rendent plus efficaces et accessibles [323], sans pouvoir toutefois rattraper le taux d'acquisition des données [253]. De nombreuses méthodes rentrent dans le cadre de l'écologie numérique, dont certaines sont décrites ci-dessous.

L'écologie numérique : les statistiques au service de l'écologie.

A.1.4.1 *Data mining*

Le data mining – aussi appelé exploration de données – consiste à extraire des connaissances à partir de grandes quantités de données [54, 424]. Dans le cadre plus large du processus d'analyse des données, le data mining est typiquement précédé par l'extraction et le nettoyage des données, et suivi par la visualisation, l'interprétation du modèle et la confirmation ou réfutation de l'hypothèse de travail [424]. Ces approches ont été utilisées pour révéler les variations des traits morphologiques des copépodes [409] et des particules de neige marine [401].

Data mining : découvrir des connaissances dans une avalanche de données.

A.1.4.2 *Intelligence artificielle*

L'intelligence artificielle (*Artificial Intelligence*, AI) se définit comme l'intelligence – percevoir, synthétiser et inférer des informations – mise en oeuvre par une machine. Si la théorie fut développée dans les années 50 [264], seuls des problèmes triviaux purent être abordés en raison des limites technologiques, conduisant à un désintérêt du domaine [344]. Ce n'est que dans les années 90 que le champ de l'AI revint sur le devant de la scène, si bien qu'elle est aujourd'hui omniprésente dans notre vie de tous les jours.

AI : intelligence mise en oeuvre par une machine.

A.1.4.3 *Apprentissage machine*

Le ML trouve des motifs dans les données.

Au sein de l'AI, les algorithmes d'apprentissage automatique (*Machine Learning*, ML) sont capables d'identifier des motifs dans des données d'apprentissage et peuvent éventuellement effectuer des prédictions sur de nouvelles données. De nombreux types de modèles peuvent être utilisés dans le cadre du ML : de la simple régression linéaire à des modèles plus élaborés comme les *Support Vector Machines* (SVM) [81] ou les *Random Forests* (RF) [169] pour ne citer que les plus célèbres.

L'écologie du plancton bénéficie déjà de ML,...

Le ML a de nombreuses potentielles applications dans le domaine de l'écologie planctonique [192]. Les modèles de régression tels que les *Boosted Regression Trees* (BRT) ou les RF sont souvent utilisés pour modéliser la distribution d'espèces [116], la richesse spécifique [221] ou encore la biomasse planctonique en fonction des conditions environnementales [111].

...et le ML est très utile pour accélérer le traitement des données.

De plus, le ML peut considérablement faciliter le traitement des données, par exemple l'identification automatique d'images de plancton [192]. Si les SVM [185, 250, 372] ou les RF [149] peuvent être utilisés pour classifier automatiquement des images d'organismes planctoniques, ils ne peuvent apprendre directement sur les images brutes, mais utilisent à la place des propriétés (taille, niveaux de gris...) des images. Ces modèles ont l'avantage d'être faciles à utiliser et peu coûteux en termes de puissance de calcul, mais sont aujourd'hui surpassés par d'autres modèles plus récents et plus complexes.

A.1.4.4 *Apprentissage profond*

Le DL : du ML basé des réseaux neuronaux avec plusieurs couches,...

Le *Deep Learning* (DL) est une branche du ML basé sur des réseaux neuronaux artificiels à plusieurs couches : les perceptrons multicouches (*multilayer perceptron*, MLP) [309], le terme « profond » faisant référence aux couches cachées. L'architecture de ces réseaux est inspirée de celle du cerveau animal, dans lequel les neurones sont les unités de base, reliées entre elles par des connexions synaptiques.

... qui sont très polyvalents.

Les MLP étant très polyvalents, ils sont appliqués à diverses tâches en écologie planctonique, et plus particulièrement à la classification d'images [89, 119, 367, 416]. Bien que les MLP puissent travailler sur des images brutes au contraire des approches de ML classique, le nombre de connexions augmente de façon quadratique avec la taille de l'image, limitant la taille des images pouvant être traitées. Des architectures plus récentes permettent désormais de traiter ce type d'entrées.

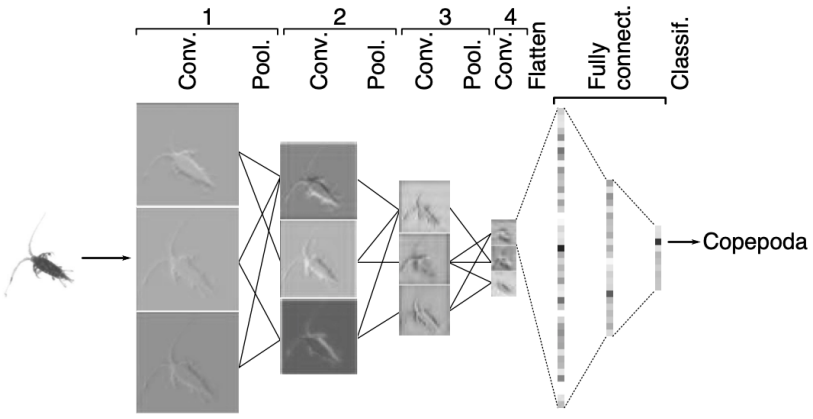


Figure A.3 : Représentation schématique d'en CNN dédié à la classification d'images de plancton. Conv = couche de convolution, Pool = couche de pooling , Fully connect. = couche complètement connectée, Classif. = couche de classification. Pour simplifier la visualisation, seules quelques connexions sont représentées. Crédits : JO Irisson.

Réseaux de neurones à convolution Les réseaux de neurones à convolution (*Convolutional Neural Network*, CNN) sont un type spécifique de réseau de neurones, dont l'architecture est inspirée du cortex visuel animal, et tire parti de l'autocorrélation spatiale au sein des images pour réduire le nombre de connexions. Un CNN est composé d'un *feature extractor* (extracteur de features) suivi de couches complètement connectées (i.e. un MLP), et se termine par une couche de sortie (Figure A.3). Ainsi, il n'est plus nécessaire d'extraire des features manuellement : cette étape est intégrée dans le modèle. Développés dans les années 90 [220], ces modèles sont devenus très populaires dans les années 2010 [213, 343] et sont maintenant l'approche de référence pour la classification d'images [222].

CNN : un problème partagé est un problème réduit de moitié.

Par conséquent, les CNN sont particulièrement utilisés en écologie du plancton, notamment pour automatiser le tri des images [59, 77, 93, 118, 124, 224, 247, 249, 327, 337, 354, 395]. Toutefois, les applications vont au-delà de la simple classification : les CNN permettent aussi de détecter ou de segmenter des objets dans des images (Figure A.4).

De nombreuses applications pour l'écologie du plancton.

Ainsi, ces approches numériques permettent aujourd'hui d'automatiser le traitement de grandes quantités de données, et ce grâce à des progrès notables dans les outils informatiques.

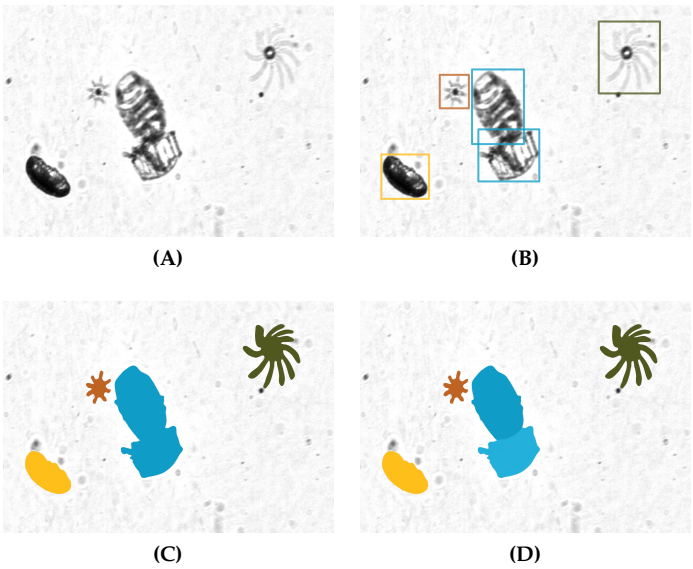


Figure A.4 : Différentes tâches de détection et de segmentation pouvant être effectuées avec des CNNs, sur des images de plancton collectées par l'ISIIS. **(A)** image brute, **(B)** détection d'objet, **(C)** segmentation sémantique, **(D)** segmentation par instance. Pour **(C)** et **(D)**, le fond a été laissé blanc intentionnellement bien qu'il constitue une classe en lui-même. Quatre classes de plancton sont représentées : Scyphozoa ephyra (Cnidaria) en jaune, Acantharia (Rhizaria) en marron, Doliolida en bleu et Rhopalonematidae (Cnidaria) en vert.

A.1.5 Les outils

Le succès de l'AI repose tout d'abord sur l'augmentation de la puissance de calcul disponible : elle double environ tous les deux ans tandis que le coût des ordinateurs est divisé par deux dans le même temps [278]. Le développement des processeurs graphiques (GPU) [73] a été un élément clé pour le succès des CNN, puisque ces derniers reposent sur un grand nombre de calculs unitaires simples qui peuvent facilement être exécutés en parallèle sur un GPU. De plus, la disponibilité de grands jeux de données annotés comprenant plusieurs millions d'images comme ImageNet [101] a été décisive pour entraîner et évaluer les performances des modèles de classification. Finalement, le troisième pilier est la disponibilité de bibliothèques open source clé en main telles que Tensorflow [1] ou Pytorch [308], permettant de concevoir et d'entraîner des modèles ML/DL.

Les 3 piliers de l'AI : puissance de calcul, grands jeux de données, librairies clé en main.

Ces outils sont en développement perpétuel. Pour donner un peu de contexte, l'un des outils au cœur de ce travail n'était pas encore publié lorsque le doctorat a commencé en septembre 2019. Lorsqu'il a été publié en février 2020, notre serveur de calcul, alors âgé de 3 ans, était dépassé pour son exécution. Les calculs ont été en partie réalisés sur un nouveau serveur de calcul, ainsi que sur le supercalculateur Jean-Zay et un serveur de calcul appartenant à la plateforme de bioinformatique ABiMS de Roscoff.

Ces outils sont en développement constant.

A.1.6 Objectifs de la thèse

A.1.6.1 *Questions écologiques*

Ce travail vise à améliorer nos connaissances sur la distribution du plancton et ses facteurs forçant, de l'échelle globale à la subméso-échelle. (i) Tout d'abord, la typologie globale des communautés de plancton est étudiée à l'échelle globale, en relation avec les facteurs environnementaux. (ii) Ensuite, la distribution des particules et du plancton est étudiée à travers un front de méso-échelle, pendant le bloom de printemps. (iii) Enfin, le lien entre la distribution d'organismes planctoniques mixotrophes et leur environnement est étudié à subméso-échelle. Pour cela, trois jeux de données d'imagerie *in situ*, collectés à trois échelles différentes, sont analysés (Table A.1).

Résoudre la distribution du plancton à différentes échelles.

Table A.1 : Principales caractéristiques des jeux de données analysés.

Dataset	Échelle		Instrument	Images	
	Temporelle	Spatiale		Total	Plancton
UVP5 global	10 ans	globale	UVP5	6.8×10^6	330,000
glider + UVP6	5 mois	méso-échelle	UVP6	1.1×10^6	30,000
VISUFRONT	10 jours	subméso-échelle	ISIIS	1.6×10^8	1.3×10^7

A.1.6.2 *Plan du travail*

Ce travail aborde...

Ce travail se divise en quatre parties : cette introduction, les développements méthodologiques, les résultats écologiques et la discussion des résultats par rapport aux connaissances existantes. La partie ii traite de l'implémentation d'un pipeline en deux étapes basé sur l'AI pour traiter la très grande quantité de données récoltées pendant la campagne VISUFRONT. Le chapitre 2 aborde la détection des organismes planctoniques dans les images ISIIS brutes et le chapitre 3 présente une comparaison de modèles de classification d'images de plancton, dont celui utilisé pour trier les organismes planctoniques détectés dans les images ISIIS. La partie iii couvre les résultats écologiques émanant de ce travail, de la plus grande à la plus petite échelle. Le chapitre 4 porte sur la répartition mondiale des types de communautés planctoniques. Le chapitre 5 décrit la dynamique printanière du plancton et des particules à travers un front de méso-échelle. Le chapitre 6 met en évidence la distribution et le comportement à très fine échelle des organismes planctoniques. Enfin, la partie iv est consacrée à la discussion des résultats, tant méthodologiques qu'écologiques.

... des développements méthodologiques...

... et des analyses écologiques.

A.2 L'intelligence artificielle au service du traitement des données ISIIS

Deux étapes de traitement : segmentation et classification.

Dans cette section, les deux étapes du traitement des données ISIIS sont présentées. La segmentation, c'est-à-dire la détection des organismes planctoniques dans les images brutes, est la première étape de ce traitement entièrement automatisé. La seconde étape consiste en l'identification taxonomique des organismes planctoniques précédemment détectés, en utilisant un modèle de classification. Les méthodes

appliquées au cours de ces deux étapes sont décrites dans les deux articles inclus dans ce chapitre.

A.2.1 Segmentation intelligente d'images de plancton

Comme expliqué dans l'introduction, les instruments d'imagerie *in situ* collectent de grandes quantités d'images, dont la grande majorité sont des images de particules de neige marine ou des artefacts d'imagerie. Parmi eux, l'*In Situ* Ichthyoplankton Imaging System (ISIIS) possède le plus grand taux d'échantillonnage (> 100 L s^{-1}) et génère donc de très grandes quantités de données. Pour extraire des quantités raisonnables d'informations écologiques à partir de ces images *in situ*, nous proposons de nous concentrer sur les organismes planctoniques dès le début du processus de traitement des données, c'est-à-dire à l'étape de la segmentation. Nous avons comparé trois méthodes de segmentation, en nous focalisant sur les cibles les plus petites, pour lesquelles le plancton représente moins de 1% des objets : (i) un seuillage naïf d'image, (ii) un détecteur d'objets basé sur les régions extrémales maximalement stables (maximally stable extremal regions, MSER), et (iii) un détecteur d'objets sensible au contenu, basé sur des réseaux de neurones à convolutions (CNN).

Trouver les organismes planctoniques dans un déluge de neige marine…

… grâce à des méthodes basées sur l'AI.

Ces méthodes ont été évaluées sur un sous-ensemble de données ISIIS collectées dans la mer Méditerranée, dont est extrait un ensemble de données de vérification de plus de 3 000 organismes manuellement détourés. La méthode naïve de seuillage a détecté 97,3% de ces organismes, mais a produit environ 340 000 segments, dont 99,1% n'étaient donc pas du plancton (rappel = 97,3%, précision = 0,9%). En combinant le seuillage avec un CNN, quelques organismes planctoniques supplémentaires ont été manqués (rappel = 91,8%) mais le nombre de segments a été divisé par 18 (la précision passant à 16,3%). Le détecteur MSER a produit quatre fois moins de segments que le seuillage (précision = 3,5%) mais a manqué plus d'organismes (rappel = 85,4%), en étant toutefois considérablement plus rapide. Étant donné que le seuillage naïf produit ~525 000 objets à partir d'une minute de déploiement ISIIS, les méthodes de segmentation intelligentes améliorent considérablement le traitement des données ISIIS et facilitent la future classification taxonomique objets segmentés, pour un coût limité en termes de rappel, en particulier pour la méthode CNN. Ces approches sont désormais standard en vision par ordinateur et pourraient être ap-

Les CNN ont donné le meilleur compromis pour une détection efficace du plancton.

plicables à d'autres dispositifs d'imagerie du plancton, dont la majorité partagent un problème de gestion et traitement d'une grande quantité de données.

A.2.2 Classification d'images de plancton

Trier de grandes quantités d'images de plancton est une tâche ardue.

Le traitement des images collectées par les instruments d'imagerie du plancton est souvent automatisé via des algorithmes d'apprentissage automatique. Cependant, la classification des images de plancton est une tâche informatique difficile en soi : les jeux de données sont fortement déséquilibrés ; les classes dominantes sont souvent sans intérêt biologique (artefacts, bulles) et/ou d'aspect très hétérogène (neige marine) ; et les images couvrent une large gamme de tailles. Malgré de nombreux rapports sur les performances des classifieurs automatiques d'images de plancton, il reste difficile de savoir comment les méthodes se comparent entre elles et pour quelles tâches on peut s'y fier. Ceci est principalement dû au fait que ces rapports s'appuient sur des jeux de données non publiés et souvent petits, qui ne sont pas nécessairement représentatifs d'échantillons biologiques réels en termes de taille, de nombre de classes et de proportions.

Comparé à une approche classique,...

... les CNN améliorent la détection des classes peu abondantes.

Nous présentons ici les performances d'une méthode de classification classique (Random Forest sur des propriétés extraites manuellement des images) et d'une méthode plus récente (un réseau de neurones à convolutions) sur de grands jeux de données ayant vocation à être publiés, provenant de six instruments d'imagerie du plancton largement utilisés. Nous montrons que l'utilisation d'un réseau de neurones à convolutions améliore les performances de classification, mais seulement de façon notable sur les classes peu abondantes (quelques centaines d'images). Enfin, nous montrons la différence entre les prédictions des deux classifieurs et une validation manuelle par un expert taxonomiste sur plusieurs ensembles de données du monde réel, afin de donner un aperçu des questions écologiques qui peuvent ou ne peuvent pas être étudiées à partir de classifications automatiques uniquement.

A.3 Distribution du plancton à différentes échelles

Ensuite, la distribution des organismes planctoniques a été étudiée à différentes échelles spatiales et temporelles. Chaque chapitre de cette section aborde la distribution du plancton à une échelle donnée, de

la plus grande – l'échelle globale – aux échelles plus fines – méso et subméso.

A.3.1 Typologie globale des communautés de plancton

Les océans sont généralement divisés verticalement en zones épipélagique (< 200 m), mésopélagique (200-500 m) et bathypélagique (> 500 m). Bien que plusieurs tentatives aient été faites pour partitionner les écosystèmes océaniques en grands biomes comme leurs homologues terrestres, cela reste difficile en raison du manque d'observations homogènes à l'échelle globale. Les biogéographies océaniques sont principalement basées sur des données biogéochimiques combinant des données optiques *in situ* (fluorescence, atténuation de la lumière), la télédétection et parfois les résultats de modèles biogéochimiques. La cohérence entre ces régionalisations, principalement biogéochimiques, et la distribution spatiale des organismes planctoniques reste cependant non résolue.

Les partitions verticales et horizontales de l'océan...

... reflètent-elles la distribution des organismes planctoniques ?

En utilisant l'imagerie *in situ*, nous avons étudié la distribution globale des organismes méso- et macroplanctoniques (> 600 µm de diamètre sphérique équivalent). Nous avons utilisé un jeu de données global de 2500 profils verticaux CTD équipés d'un *Underwater Vision Profiler* 5 (UVP5). Parmi les 6,8 millions d'objets imagés, 330 000 étaient des grands organismes zooplanctoniques ou des colonies de phytoplancton, tandis que le reste était principalement constitué de particules de neige marine. En appliquant des méthodes statistiques multivariées d'ordination et de régression, nous avons décrit les grands types de communautés planctoniques ainsi que leur lien avec les conditions environnementales dans les couches épipélagique et mésopélagique supérieure.

À partir d'un jeu de données global d'imagerie in situ...

Dans les deux couches, trois types de communautés planctoniques ont été décrites. Les communautés planctoniques épipélagiques étaient dominées par des *Trichodesmium* dans l'Atlantique intertropical, par des copépodes aux hautes latitudes et dans les zones d'upwelling, et par des rhizaires dans les zones oligotrophes. Dans la couche mésopélagique, les communautés planctoniques étaient dominées par des copépodes aux latitudes élevées et dans l'océan Atlantique, par les rhizaires dans le système d'upwelling péruvien, tandis que des communautés mixtes ont été trouvées ailleurs. La comparaison entre la distribution de ces communautés et un ensemble de régionalisa-

... nous caractérisons les grands types de communautés de plancton.

tions existantes de l'océan suggère que la structure des communautés planctoniques décrites ci-dessus est principalement déterminée par les conditions environnementales régionales plutôt que par les conditions à proximité immédiate du site d'échantillonnage.

A.3.2 Évolution temporelle de la distribution du plancton et des particules à travers un front à méso-échelle pendant le bloom de printemps

Mais qu'en est-il des plus petites échelles ?

Comme vu dans l'introduction, l'effet des dynamiques à méso-échelle sur la distribution des organismes planctoniques est relativement bien documenté. Cependant, l'interaction entre ces dynamiques spatiales et l'échelle temporelle, qui peut entraîner des augmentations soudaines de la biomasse planctonique, est moins connue et encore moins décrite à haute résolution.

Au travers d'un front à méso-échelle…

Nous avons étudié un front permanent de méso-échelle dans le nord-ouest de la mer Méditerranée. Ce front a été échantillonné de manière répétée entre janvier et juin 2021 en utilisant un planeur équipé d'un UVP6, un imageur *in situ* polyvalent. Nous nous sommes efforcés de décrire la distribution à méso-échelle du plancton et des particules pendant le bloom de printemps, afin d'évaluer si le front était un lieu de concentration accrue pour les organismes zooplanctoniques, et si cette structure contraint la distribution des particules. Pendant ces 5 mois, le planeur a effectué plus de 5 000 plongées et l'UVP6 a collecté 1,1 million d'images. Nous avons concentré notre analyse sur les transects peu profonds (300 m), avec une résolution horizontale de 900 m. Certaines images ont été triées manuellement, et d'autres prédites avec un algorithme d'apprentissage automatique. Au final, environ 13 000 images d'organismes planctoniques ont été retenues.

… échantillonné de façon répétée à l'aide d'imagerie in situ pendant le bloom de printemps…

Des méthodes statistiques d'ordination ont révélé des périodes contrastées pendant le bloom, au cours desquelles les changements dans l'abondance et la taille des particules pouvaient être expliqués par les changements dans la communauté planctonique. Le front a eu une forte influence sur la distribution des particules, alors que le signal n'était pas aussi clair pour le plancton, probablement en raison du nombre relativement faible d'organismes imagés. En outre, nous avons également détecté des structures à subméso-échelle telles que des événements de subduction et des tourbillons cohérents de subméso-échelle. Ce travail confirme la nécessité d'échantillonner à la fois le plancton

… nous montrons l'effet du front sur la distribution des particules et du plancton.

et les particules à fine échelle pour comprendre leur interaction, une tâche pour laquelle l'imagerie *in situ* est particulièrement adaptée.

A.3.3 Étude du comportement écologique complexe de mixotrophes géants grâce à l'imagerie *in situ*

L'imagerie in situ permet d'étudier des organismes peu étudiés...

... tels que les rhizaires.

Bien que les organismes planctoniques fassent l'objet de recherches scientifiques depuis des siècles, certains organismes sont passés à travers les mailles du filet, comme les rhizaires. En effet, ces eucaryotes unicellulaires sont particulièrement fragiles et souvent endommagés par les outils classiques d'échantillonnage. Bien que certains rhizaires soient connus comme mixotrophes hébergeant des symbiotes photosynthétiques, des lacunes persistent quant à leur écologie trophique. Les connaissances concernant leur cycle de reproduction sont encore plus rares. Toutefois, leur contribution substantielle à la biomasse planctonique a récemment été mise en évidence grâce à l'imagerie *in situ*. En effet, cette approche permet l'étude de ces organismes dans leur environnement non perturbé.

Leur distribution à fine-échelle a pu être reliée...

... à leur écologie trophique...

... ainsi qu'aux étapes de leur cycle de vie.

En exploitant les données d'imagerie *in situ* collectées à haute fréquence, nous avons étudié la distribution à fine échelle et la position *in situ* de ~230 000 organismes appartenant à trois groupes de rhizaires (Acantharia, Collodaria et Phaeodaria). Nous avons mis en évidence des différences dans la distribution verticale entre les sous-groupes, probablement causées par des stratégies de vie différentes et des différences de capacités dans le contrôle de la flottabilité. Nous avons également détecté une orientation préférentielle, non documentée auparavant, de certains organismes. Enfin, nous avons essayé de relier certaines de nos observations aux étapes présumées du cycle de vie méconnu des Collodaria, révélant potentiellement des variations du contrôle de la flottabilité des organismes afin d'atteindre l'environnement dans lequel se déroule l'étape suivante de leur cycle.

A.4 Discussion

A.4.1 L'imagerie *in situ* pour étudier la distribution du plancton à de nombreuses échelles

A.4.1.1 *Micro-échelle*

L'imagerie in situ permet d'aborder...

Les interactions entre les organismes planctoniques et leur environnement se produisent à micro-échelle, $\mathcal{O}$(1 mm) [19]. Cependant, peu d'études à micro-échelle ont été réalisées *in situ*, car peu d'outils sont adaptés, mais ils ne sont pas inexistants [132, 260]. Ces outils ont ainsi permis d'étudier la position *in situ* des organismes planctoniques [288, 396]. Plus généralement, certains instruments d'imagerie *in situ* peuvent être utilisés pour étudier les positions individuelles dans le plancton, que cela soit pour en déduire des interactions [152] ou bien une orientation préférentielle (Chapitre 6), à condition que l'image soit prise par le côté et non par le dessus. C'est ainsi que Gaskell, Ohman et Hull [140] ont pu mettre en évidence une orientation préférentielle chez les foraminifères.

... les interactions entre organismes ou leur orientation,...

A.4.1.2 *Fine-échelle*

... les agrégations au niveau de couches fines de plancton...

Les couches fines de plancton sont des structures de moins de 5 m d'épaisseur, pouvant s'étendre horizontalement sur plusieurs kilomètres [325], et sont composées de divers objets : phytoplancton, zooplancton, agrégats de neige marine... [5, 269]. Leur formation se fait via des mécanismes biologiques (croissance locale, déplacement actif) ou physiques (accumulation sur des gradients de densité) [112, 325]. Des organismes zooplanctoniques ou des poissons sont souvent associés à ces couches fines [31, 33], ce qui suggère un rôle écologique clé [112]. L'étude des couches fines de plancton a été envisagée au début de ce travail, mais de telles structures n'ont pas été détectées dans les données collectées pendant la campagne VISUFRONT. D'abord, l'algorithme de segmentation était peu efficace pour détecter les fibres de diatomées (Chapitre 2). Ensuite, étant donné les conditions d'oligotrophie en été dans la mer Ligure, une couche mince n'aurait pu se former en dehors du maximum profond de chlorophylle (*Deep Chlorophyll Maximum*, DCM).

Plus épais que les couches minces [112], les DCM sont également situés autour de discontinuités (e.g. pycnocline), à l'interface entre les

eaux de surface pauvres en nutriments et les eaux profondes limitées en lumière [175], c'est-à-dire un compromis entre la disponibilité en nutriments et en lumière où le phytoplancton peut prospérer. Si l'accumulation passive de cellules de phytoplancton sur les gradients de densité peut contribuer à la création de DCMs [241], les organismes zooplanctoniques peuvent aussi s'agréger sur des discontinuités [150, 168]. Dans notre étude sur la distribution à fine échelle des rhizaires à travers le Front Ligure (Chapitre 6), nous avons détecté une distribution préférentielle de plusieurs groupes de Rhizaria autour du DCM. Si de nombreux indices étaient en faveur d'un contrôle actif de la flottabilité pour les Collodaria solitaires occupant le DCM (source de nourriture et de symbiontes potentiels), les Aulacanthidae (Phaeodaria) quant à eux sont plus vraisemblablement soumis à une accumulation passive sur le gradient de densité. Par ailleurs, les Aulacanthidae étaient entraînés en profondeur par des mouvements d'eau descendants, ce qui est en faveur d'un faible contrôle de leur flottabilité.

... ou au niveau du DCM,...

A.4.1.3 *Subméso-échelle*

Comme expliqué dans l'introduction, la distribution du phytoplancton est fortement affectée par des structures à subméso-échelles, en particulier dans les zones frontales [230]. Les données ISIIS collectées pendant la campagne VISUFRONT nous ont permis d'étudier la distribution verticale des organismes planctoniques à l'échelle du mètre. Parmi ceux-ci, les Aulacanthidae (Phaeodaria) ont été les seuls dont la distribution était affectée par la recirculation à subméso-échelle (voir Figure C.1 en annexe). Si ce résultat n'est pas tellement surprenant pour les organismes planctoniques capables de se déplacer, il est beaucoup plus frappant pour les organismes non motiles (e.g. les rhizaires), et suggère une certaine capacité de contrôle de la flottabilité, comme nous en avons fait l'hypothèse pour les Collodaria (Chapitre 6). Les vitesses verticales des eaux descendantes mesurées à partir d'un courantomètre acoustique à effet doppler (ADCP) auraient été d'une grande aide pour confirmer ou infirmer nos hypothèses. Malheureusement, un dysfonctionnement du système pendant la campagne a rendu les données inutilisables pour une analyse à cette échelle.

... les déplacements induits par des mouvements de masses d'eau,...

A.4.1.4 *Méso-échelle*

. . . la distribution au travers de structures à méso-échelle,. . .

Comme décrit dans le chapitre 5, l'agrégation d'organismes planctoniques est habituelle au niveau des fronts [304]. Tandis que la distribution du plancton peut être étudiée à l'aide d'un échantillonnage au filet au travers de fronts relativement larges (e.g., [48, 49]), l'imagerie *in situ* en donner une vue plus détaillée [154, 248]. En effectuant un échantillonnage répété au travers du front Ligure pendant le bloom de printemps à l'aide d'un planeur sous-marin équipé d'un UVP6, nous n'avons pu clairement mettre en évidence ni une accumulation de zooplancton ou de chlorophylle au niveau du front, ni une distribution des organismes d'un côté ou de l'autre du front, qui semblait bien agir comme une barrière contraignant la distribution des particules, comme précédemment rapporté [148, 384]. Cependant, le taux d'échantillonnage de l'UVP6-LP était probablement trop faible pour imager suffisamment d'organismes planctoniques et détecter un tel effet.

Mais ce même front avait déjà été échantillonné à une résolution beaucoup plus élevée avec l'ISIIS (> 100 L s^{-1}) lors de la campagne VISUFRONT, bien que ces données ne montrent qu'un instantané estival de la distribution du plancton. En ce qui concerne la distribution à méso-échelle du plancton par rapport au front, certains organismes (Appendicularia, Doliolida, Hydrozoa, Pyrocystis et Siphonophorae) étaient contraints du côté côtier du front (Figure C.1), conformément aux résultats précédents [124]. Cependant, ces données ne mettent pas en évidence d'accumulation de plancton au niveau du front en été.

A.4.1.5 *Échelle globale*

. . . et enfin la distribution globale à condition d'agréger des jeux de données.

Enfin, la distribution globale du plancton peut être étudiée à partir des données d'imagerie *in situ* en agrégeant des jeux de données cohérents collectés en divers endroits. Une telle approche nécessite des instruments standardisés, inter-calibrés et disponibles dans le commerce tels que l'UVP [317, 318]. Par exemple, Kiko et al. [205] a récemment publié un jeu de données mondial (8 805 profils UVP5) sur la distribution de la taille des particules. Un jeu de données similaire pour les organismes planctoniques imagés par l'UVP5 est sur le point d'être publié. C'est ce jeu de données qui a été utilisé dans le travail présenté dans le chapitre 4 et qui a également permis d'estimer la biomasse globale du plancton [111], une étude à laquelle j'ai contribué (voir l'annexe B). J'ai personnellement participé à l'effort de tri pour construire ce jeu

de données, en validant ~250 000 images de plancton, à l'aide d'un guide taxonomique établi spécifiquement pour les images UVP5. Une version antérieure, plus petite, de ce jeu de données a été utilisée pour mettre en lumière la contribution inattendue des rhizaires à la biomasse planctonique [42].

Pas aussi standardisé que l'UVP, l'ISIIS a néanmoins été produit en plusieurs exemplaires qui ont été déployés dans de nombreux écosystèmes à travers le monde, permettant la construction d'un jeu de données suffisamment cohérent pour mener des études comparatives, par exemple sur les dolioles [159]. Pour cette étude (Annexe B), j'ai fourni des données pour la mer Méditerranée (n $\approx$ 80 000 images de dolioles). Ces données ont été traitées par les pipelines de segmentation et de classification présentés dans les chapitres 2 et 3.

A.4.2 L'écologie à l'ère du big data

A.4.2.1 *Échantillonner plus pour faire des découvertes*

Un échantillonnage intensif permet de trouver des objets rares,...

Grâce au taux d'échantillonnage de l'ISIIS, nous avons collecté des millions d'images de plancton, ce qui est l'occasion d'échantillonner des objets rares tels que des nouvelles espèces ou des organismes aux caractéristiques particulières. Cependant, leur détection peut s'apparenter à la recherche d'une aiguille dans une botte de foin. Dans notre cas, les collodaires solitaires porteurs de vacuoles (décrits au chapitre 6) auraient pu passer entre les mailles du filet si un œil averti (celui de Tristan Biard) ne les avait pas repérés. Ils représentaient moins de 3% des collodaires solitaires, eux-mêmes représentant une fraction minuscule des organismes imagés.

... de façon fortuite,...

... via une inspection méticuleuse des données,...

En l'absence d'une détection opportuniste, des observations inattendues peuvent être faites grâce à un examen méticuleux des images. Par exemple, Greer et al. [157] ont étudié le mimétisme batésien chez les larves de poissons à partir des images ISIIS, une étude qui a nécessité l'examen de plus de 1 million d'images et demandé des centaines d'heures, pour quelques centaines de larves de poissons détectées. La science participative a été envisagée pour faciliter ces observations, mais cette approche n'a pas été jugée entièrement concluante [336] Enfin, les approches automatisées telles que le *vector embedding* ont la capacité de détecter efficacement des objets non détectés auparavant [254].

... ou grâce à des méthodes automatisées.

A.4.2.2 *Vers une science guidée par les données*

L'abondance des données offre un nouveau paradigme scientifique...

Avec le développement d'outils d'observation plus efficaces, polyvalents et rentables, le taux de collecte de données augmente de façon spectaculaire [208] et de plus en plus de données d'observation deviennent disponibles pour les écologistes [167]. Nous sommes ainsi entrés dans l'ère de l'écologie pilotée par les données (*data-driven*), basée sur l'analyse exploratoire de grands jeux de données pour en extraire des connaissances, une approche considérée comme le quatrième paradigme scientifique [179, 208]. Cette approche requiert toutefois des méthodes très efficaces pour traiter les grandes quantités de données collectées [18]. Dans ce contexte, les approches en temps réel telles que la segmentation T-MSER des images de plancton présentée dans le chapitre 2 sont particulièrement appréciées. Les méthodes d'écologie numérique, présentées dans l'introduction, ainsi que les progrès dans la puissance de calcul contribuent à la capacité de traiter de grandes quantités de données.

... auquel l'écologie marine n'échappe pas,...

... bien que des progrès restent à faire.

Bien entendu, le domaine de l'écologie marine n'échappe pas à ces changements : de plus en plus de capteurs sont embarqués sur des plateformes autonomes [72]. Divers instruments d'imagerie du plancton existent [243], de telle sorte que les bases de données connaissent une croissance exponentielle [192]. À l'heure actuelle, 201 millions d'objets sont contenus dans la base de données EcoTaxa [316]. Cependant, ces outils d'imagerie génèrent de grandes quantités de données au prix d'une certaine résolution taxonomique (les organismes ne peuvent pas être manipulés), et certains objets restent non identifiés. Malgré l'abondance des données, nous manquons toujours de données annotées et nettoyées pour entraîner et tester les modèles [192], bien que les données d'entraînement peuvent être simulées [253]. Le traitement des données étant devenu une étape critique, elle doit être réalisée aussi efficacement que possible.

A.4.3 Considérations méthodologiques

A.4.3.1 *Trier efficacement des images de plancton*

Commencer par travailler sur un sous-ensemble représentatif...

Ce paragraphe est une réflexion sur la façon dont nous avons traité les données collectées par l'UVP6, dans le but de mettre en évidence ce qui aurait pu être fait plus efficacement. Pendant les 5 mois passés en mer, l'UVP6 a collecté 1 123 123 images qui ont dû être triées. Bien que ce

chiffre soit très faible comparé aux données ISIIS, le tri d'un tel nombre d'images nécessite tout de même une stratégie bien pensée en amont. En effet, notre plan initial consistait à trier certaines de ces images puis nous fier aux prédictions d'un modèle pour les autres, sans toutefois avoir réfléchi à quelles images valider manuellement, ni à la manière de garantir la qualité des prédictions, ce qui a entraîné une certaine perte de temps. Si cela était à refaire, ce paragraphe explique comment je m'y prendrais.

Il faut tout d'abord choisir sa stratégie : validation de toutes les images ou utilisation des prédictions sans vérification. Ensuite, quelle que soit la stratégie, il faut générer un sous-ensemble *représentatif* des données (e.g. $1/n$ profil régulièrement répartis sur l'ensemble des données), qui sera validé entièrement et utilisé pour entraîner un modèle de classification. Sa taille résulte d'un compromis entre l'effort de validation requis et le besoin d'un nombre suffisant d'exemples (~100 objets par classe d'intérêt pour la validation complète, ~500-1000 objets par classe d'intérêt pour les prédictions). Grâce l'échantillonnage stratifié, sa composition devrait être proche de celle de l'ensemble des données. Ce dernier point est très important car il permet d'éviter le *dataset shift* qui se produit quand la distribution des nouvelles données est différente de celle des données d'entraînement [280].

. . . permet de généraliser au reste des images,. . .

Pour la stratégie basée sur les prédictions, il convient d'être en mesure d'estimer les performances du modèle de classification. Le sous-ensemble entièrement validé doit ainsi être divisé en deux parties : un jeu d'entraînement avec ~70% des données, un jeu de validation (15% des données) pour ajuster le modèle et un jeu de test (15% des données) pour évaluer le modèle sur un jeu de données indépendant, non vu par le modèle à l'entraînement, afin d'éviter d'en surestimer les capacités [108, 192].

. . . qu'elles soient toutes inspectées ou non.

A.4.3.2 *Tirer le meilleur de l'imagerie in situ*

Parmi les nombreuses choses que j'ai pu explorer au cours de ce doctorat, le résultat que je trouve le plus frappant est la distribution différentielle des stades de vie des Collodaria, et en particulier le contrôle de la flottabilité que cela implique. Des organismes unicellulaires, remontant des profondeurs à l'aide d'alvéoles qui semblent fonctionner comme des montgolfières ; c'est *fascinant*. Cependant, comme expliqué dans le chapitre 6, nous ne sommes pas en mesure de démontrer ce processus, nous racontons simplement une histoire cohérente basée

De nombreuses questions restent sans réponse,. . .

. . . par exemple sur le cycle de vie des collodaires.

sur nos observations. À mes yeux, affiner ces observations pour confirmer (ou infirmer !) nos hypothèses serait un sujet de recherche très intéressant. En outre, d'autres questions se posent : où les collodaires solitaires acquièrent-ils leurs symbiotes lorsqu'il n'y a pas de DCM ? Comment leur concentration varie-t-elle selon les saisons ? La reproduction sexuelle est-elle saisonnière ? Les formes coloniales prévalent-elles lorsque de nouveaux symbiotes ne peuvent être acquis ?

L'imagerie in situ. . .

. . . combinée à l'intelligence artificielle. . .

. . . semble particulièrement appropriée pour aborder ces questions.

Comme nous l'avons vu, l'ISIIS est un excellent instrument pour imager de grands volumes d'eau et ainsi détecter des organismes rares, sans les perturber. Il semble donc être l'instrument parfait pour ce type d'étude, car les collodaires solitaires avec vacuoles étaient particulièrement rares. Néanmoins, un algorithme de détection efficace devrait être développé pour cibler spécifiquement ces organismes, qu'il s'agisse de formes solitaires ou coloniales, afin de parcourir efficacement les grandes quantités de données générées par ISIIS. De plus, il serait particulièrement intéressant de pouvoir résoudre la composition de la communauté phytoplanctonique à proximité immédiate des collodaires. Ceci pourrait être fait avec des capteurs de fluorescence détectant d'autres pigments que la chlorophylle a ou avec d'autres instruments d'imagerie ciblant des tailles plus petites, et pourrait apporter des informations concernant les symbiontes potentiels ou la source de nourriture des collodaires solitaires. Un tel échantillonnage devrait être effectué à différentes saisons afin de décrire les variations en concentration des différentes phases du cycle de vie de Collodaria. Enfin, la cerise sur le gâteau serait d'échantillonner physiquement une cellule porteuse de vacuoles et de pouvoir rechercher des symbiotes à l'intérieur. Bien que cela semble extrêmement difficile, cela contribuerait à démontrer que, comme nous le supposons, ces cellules sont nouvellement formées et encore exemptes de symbiotes.

B

Collaborative works

B.1 Global Distribution of Zooplankton Biomass Estimated by *In Situ* Imaging and Machine Learning

Laetitia Drago, **Thelma Panaïotis**, Jean-Olivier Irisson, Marcel Babin, Tristan Biard, François Carlotti, Laurent Coppola, Lionel Guidi, Helena Hauss, Lee Karp-Boss, Fabien Lombard, Andrew M. P. McDonnell, Marc Picheral, Andreas Rogge, Anya M. Waite, Lars Stemmann and Rainer Kiko

Frontiers in Marine Science 9
DOI: 10.3389/fmars.2022.894372

Abstract Zooplankton plays a major role in ocean food webs and biogeochemical cycles, and provides major ecosystem services as a main driver of the biological carbon pump and in sustaining fish communities. Zooplankton is also sensitive to its environment and reacts to its changes. To better understand the importance of zooplankton, and to inform prognostic models that try to represent them, spatially-resolved biomass estimates of key plankton taxa are desirable. In this study we predict, for the first time, the global biomass distribution of 19 zooplankton taxa (1-50 mm Equivalent Spherical Diameter) using observations with the Underwater Vision Profiler 5, a quantitative *in situ* imaging instrument. After classification of 466,872 organisms from more than 3,549 profiles (0-500 m) obtained between 2008 and 2019 throughout the globe, we estimated their individual biovolumes and converted them to biomass using taxa-specific conversion factors. We then associated these biomass estimates with climatologies of environmental variables (temperature, salinity, oxygen, etc.), to build habitat models using boosted regression trees. The results reveal maximal zooplankton biomass values around 60° N and 55° S as well as minimal values around the oceanic gyres. An increased zooplankton biomass is also predicted for the equator. Global integrated biomass (0-500 m) was

estimated at 0.403 PgC. It was largely dominated by Copepoda (35.7%, mostly in polar regions), followed by Eumalacostraca (26.6%) Rhizaria (16.4%, mostly in the intertropical convergence zone). The machine learning approach used here is sensitive to the size of the training set and generates reliable predictions for abundant groups such as Copepoda (R_2 ~20-66%) but not for rare ones (Ctenophora, Cnidaria, $R_2 < 5\%$). Still, this study offers a first protocol to estimate global, spatially resolved zooplankton biomass and community composition from *in situ* imaging observations of individual organisms. The underlying dataset covers a period of 10 years while approaches that rely on net samples utilized datasets gathered since the 1960s. Increased use of digital imaging approaches should enable us to obtain zooplankton biomass distribution estimates at basin to global scales in shorter time frames in the future.

B.2 *In situ* imaging across ecosystems to resolve the fine-scale oceanographic drivers of a globally significant planktonic grazer

Adam Greer, Moritz S Schmid, Patrick Duffy, Kelly Robinson, Mark Genung, Jessica Luo, **Thelma Panaïotis**, Christian Briseño-Avena, Marc Frischer, Su Sponaugle and Robert K Cowen

Limnology and Oceanography
DOI: 10.1002/lno.12259i

Abstract Doliolids are common gelatinous grazers in marine ecosystems around the world and likely influence carbon cycling due to their large population sizes and high growth and excretion rates. Aggregations or blooms of these organisms occur frequently, but they are difficult to measure or predict because doliolids are fragile, under sampled with conventional plankton nets, and can aggregate on fine spatial scales (1-10 m). Moreover, ecological studies typically target particular regions that do not encompass the range of possible habitats favoring doliolid proliferation. To address these limitations, we combined in situ imaging data from six coastal ecosystems, including the Oregon shelf, northern California, southern California Bight, northern Gulf of Mexico, Straits of Florida, and Mediterranean Sea, to resolve and compare doliolid habitat associations during warm months when en-

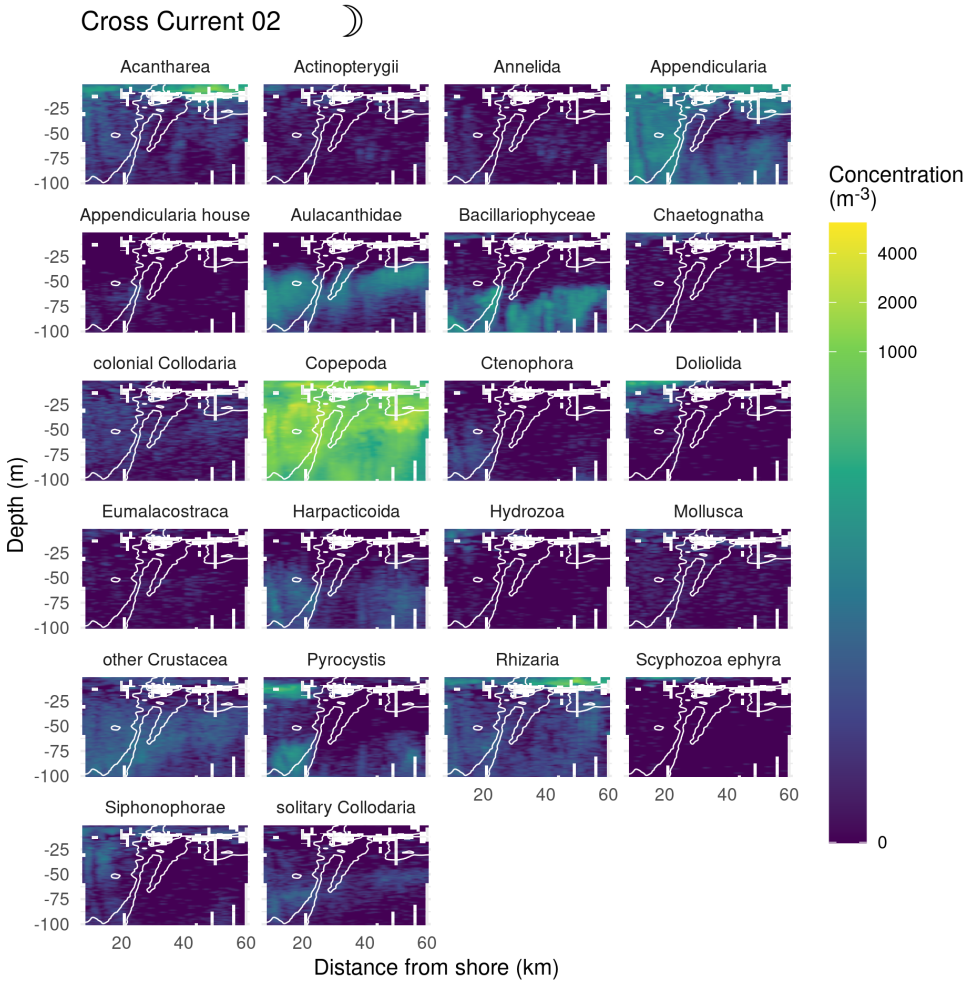
vironmental gradients are strong and doliolid blooms are frequently documented. Higher ocean temperature was the strongest predictor of elevated doliolid abundances across ecosystems, with additional variance explained by chlorophyll-a fluorescence and oxygen. The relative abundance of the nurse stage tended to increase when total doliolid abundance was low, but this pattern did not hold in upwelling ecosystems, indicating that nurses occupy less favorable habitats in established populations with wider shelf habitats. The doliolids tended to be most aggregated in oligotrophic systems (Mediterranean and southern California), suggesting that microhabitats within the water column favor proliferation on fine spatial scales. Similar comparative approaches can resolve the realized niche of fast-reproducing marine animals, thus improving predictions of population changes in response to oceanographic conditions.

C

Distribution maps of planktonic organisms imaged during the VISUFRONT campaign

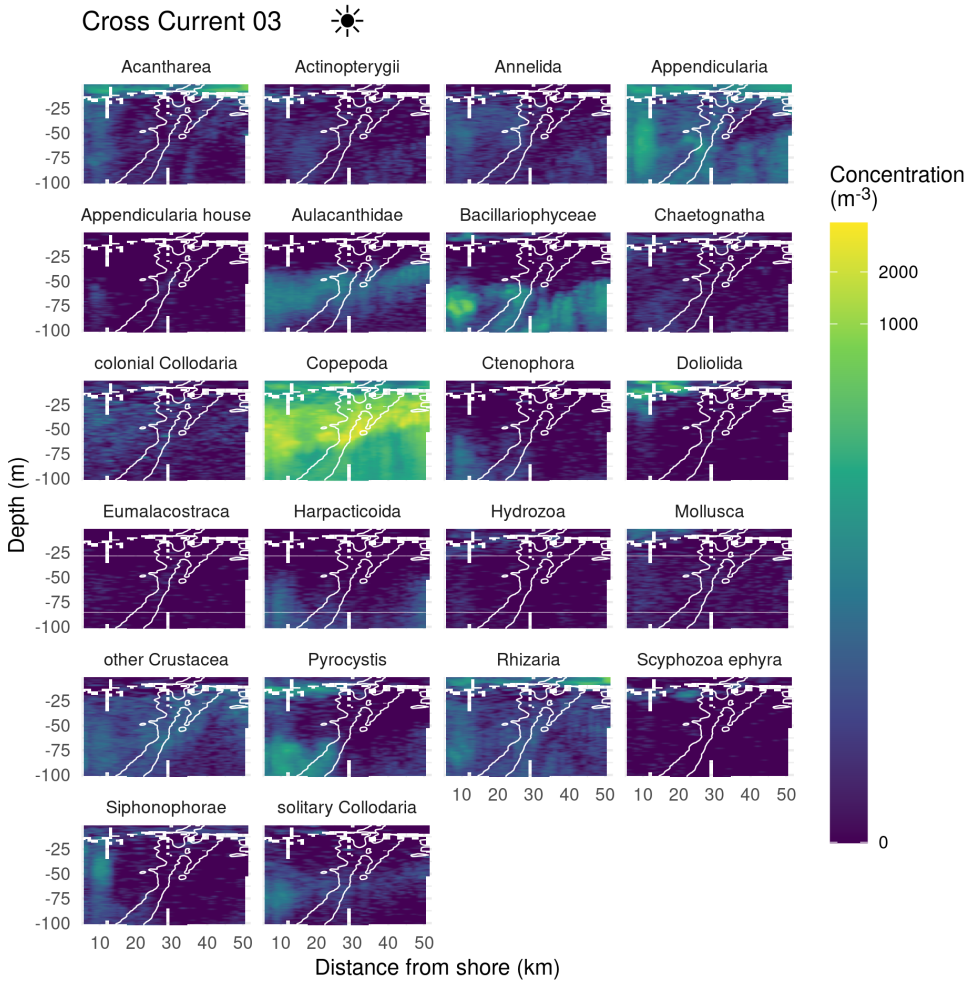
C.1 Cross front transects

Seven cross front transects were performed between 23/07/2013 09:27:00 and 28/07/2013 18:51:00. As the name suggests, these transects were performed perpendicularly to the front, starting either inshore or offshore, for a duration of 7-8 h each. Unfortunately, the first transect had to be interrupted and was thus excluded from our analyses. Distribution maps for 22 taxonomic groups on the 6 retained cross front transects are presented in Figure C.1.



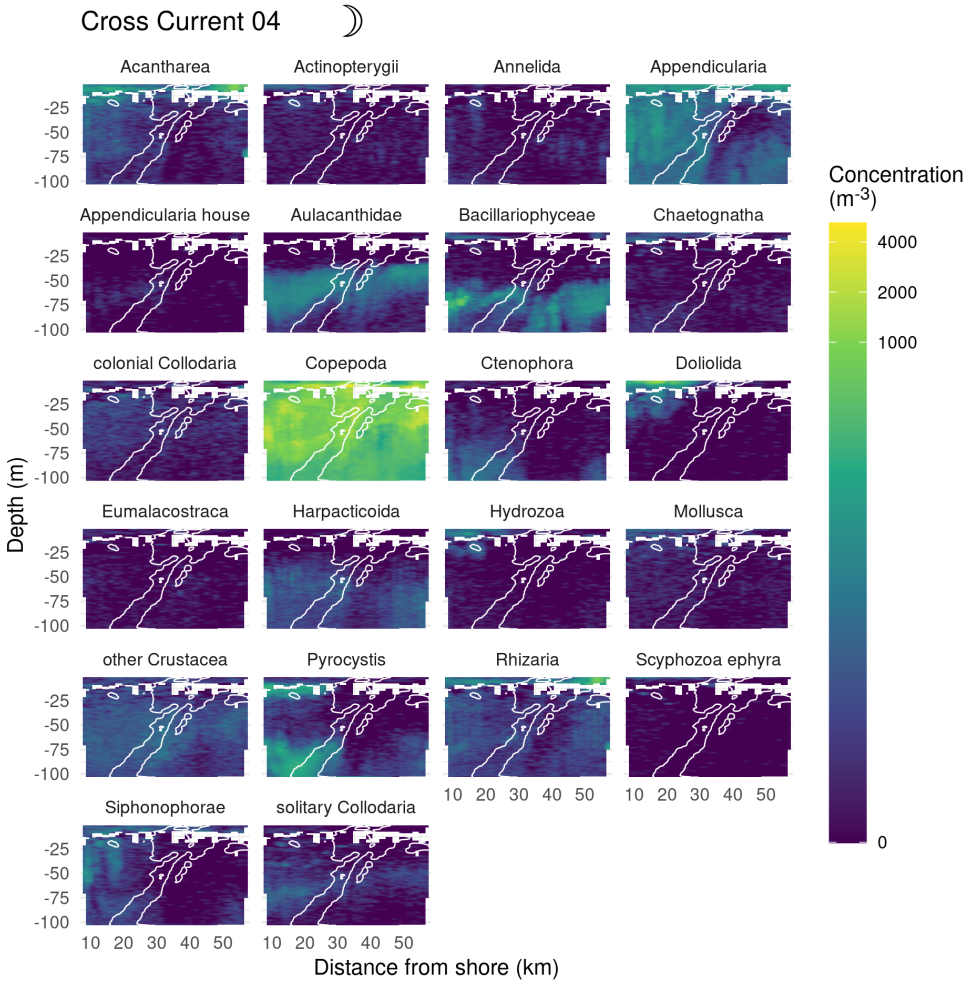
(A)

Figure C.1: *(Figure continues)*



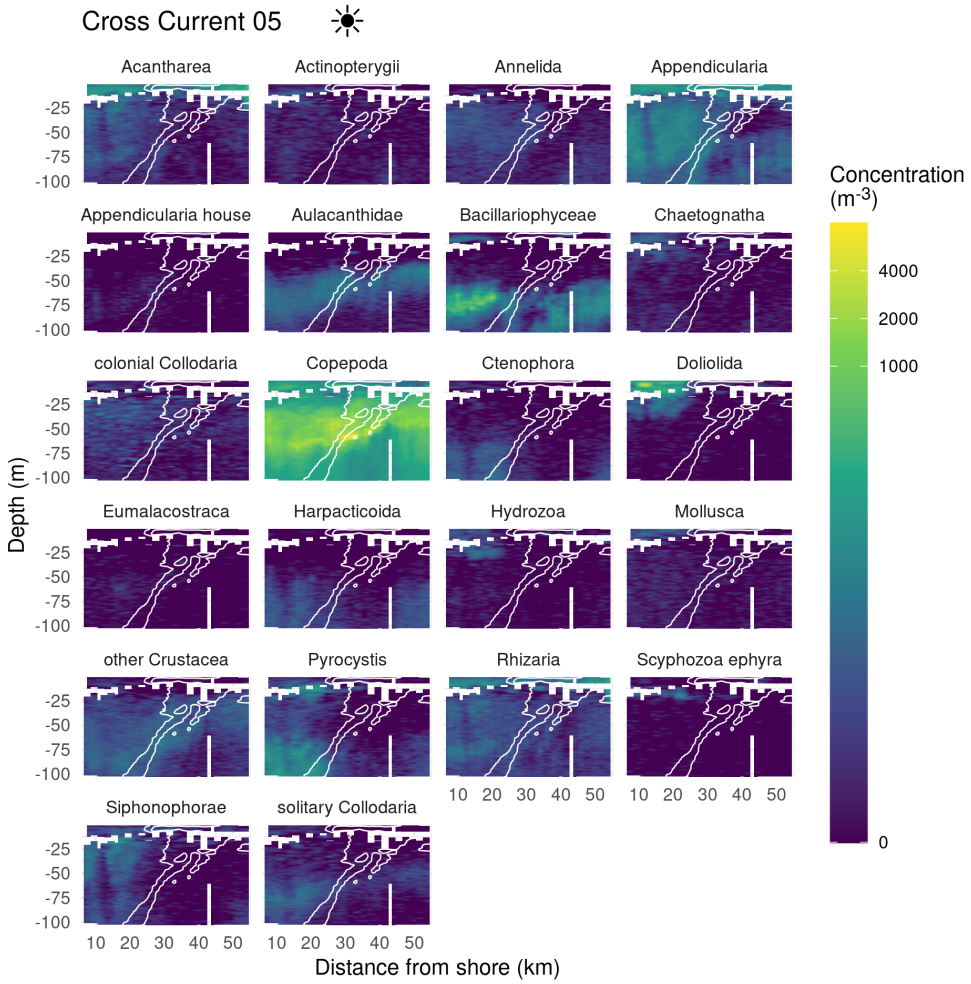
(B)

Figure C.1: *(Figure continues)*



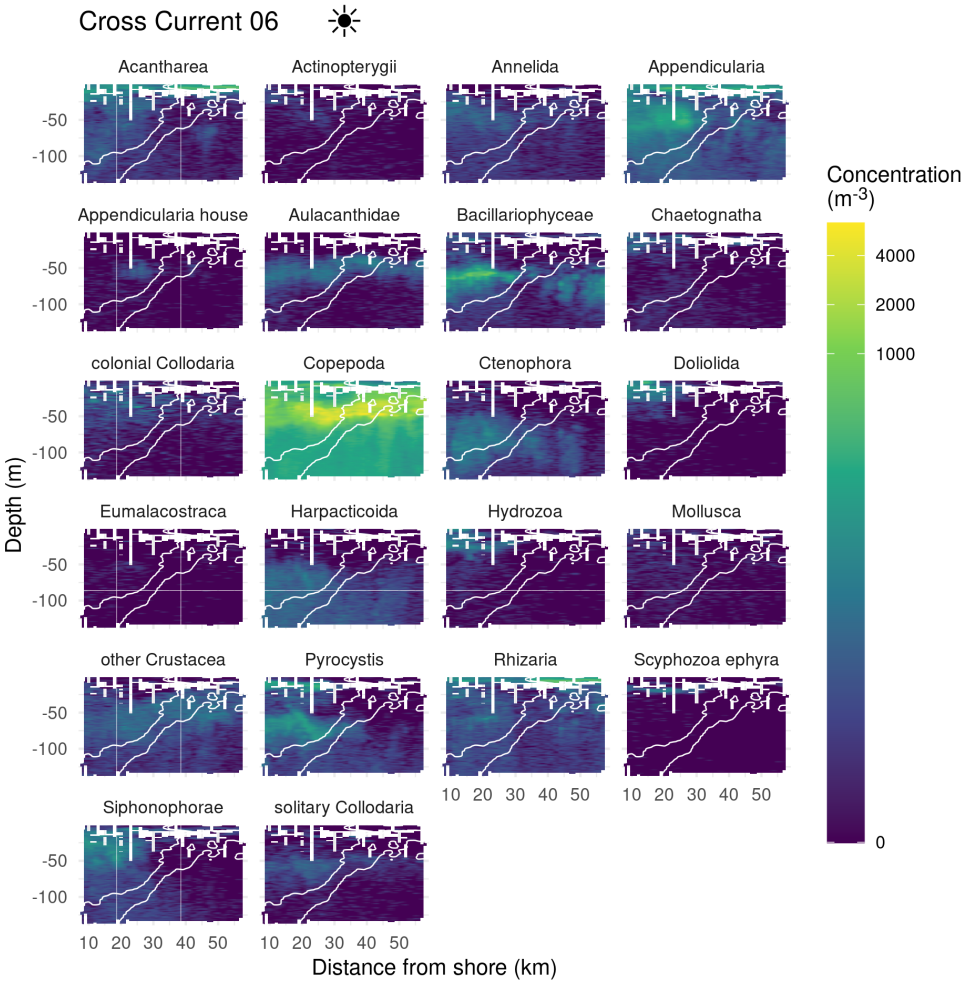
(C)

Figure C.1: *(Figure continues)*



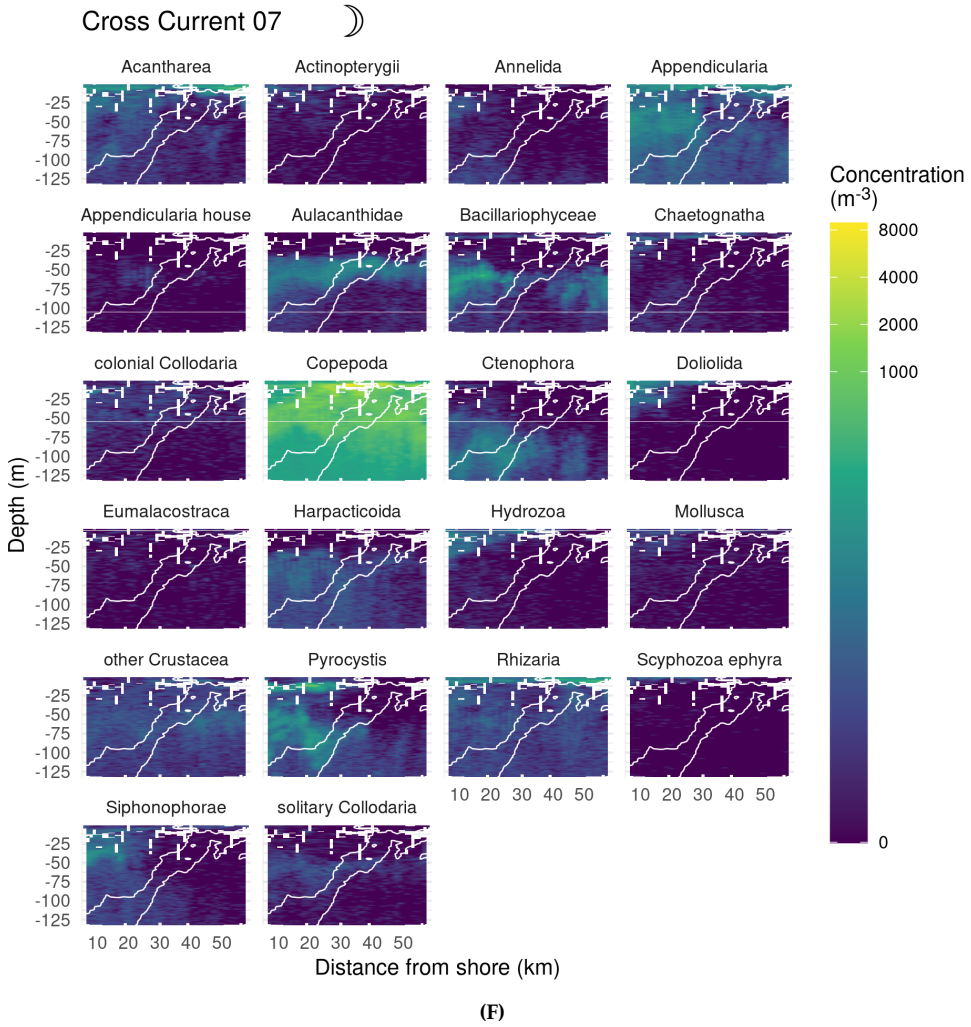
(D)

Figure C.1: *(Figure continues)*



(E)

Figure C.1: *(Figure continues)*

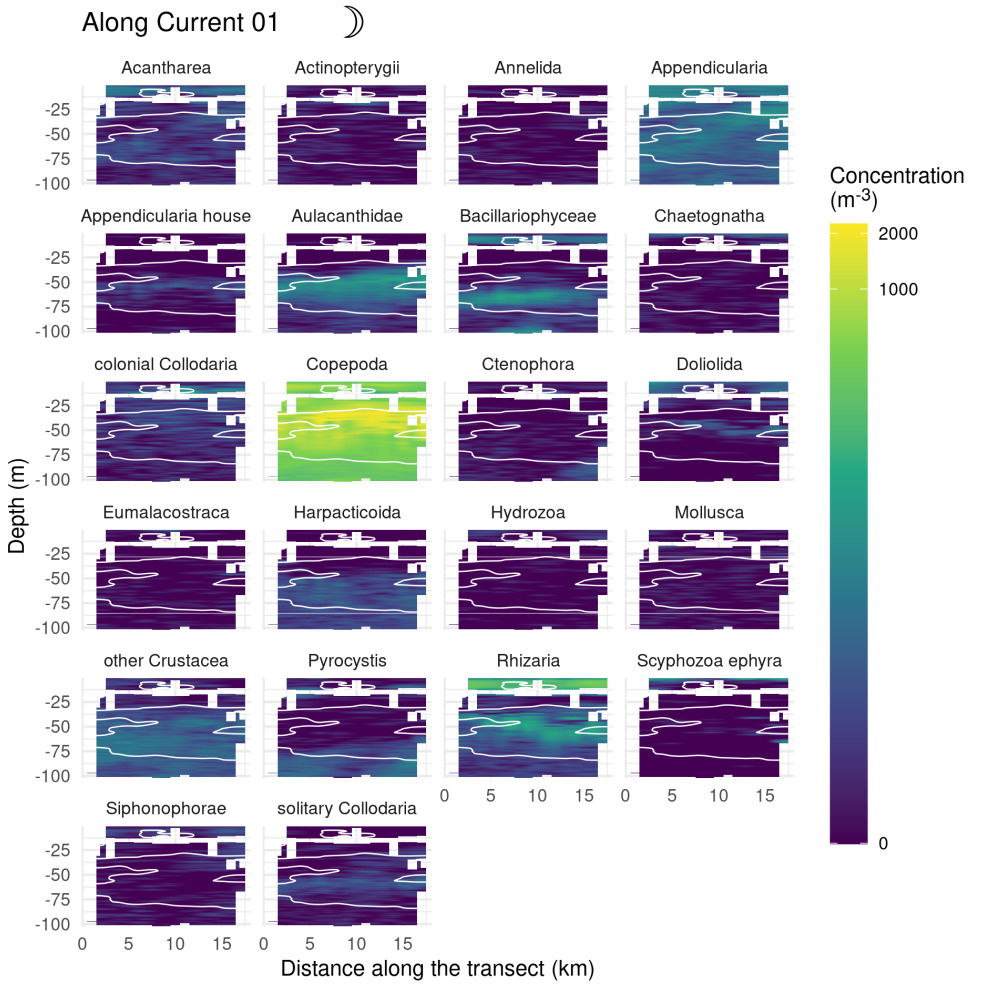


(F)

Figure C.1: Distribution maps of 22 taxonomic groups for cross front transects. **(A)** Cross front 2 (nighttime), **(B)** Cross front 3 (daytime), **(C)** Cross front 4 (nighttime), **(D)** Cross front 5 (daytime), **(E)** Cross front 6 (daytime), **(F)** Cross front 7 (nighttime). White lines represent the 38.2 and 38.3 isohalines delineating the Ligurian front. Note that the colour scale is log-transformed.

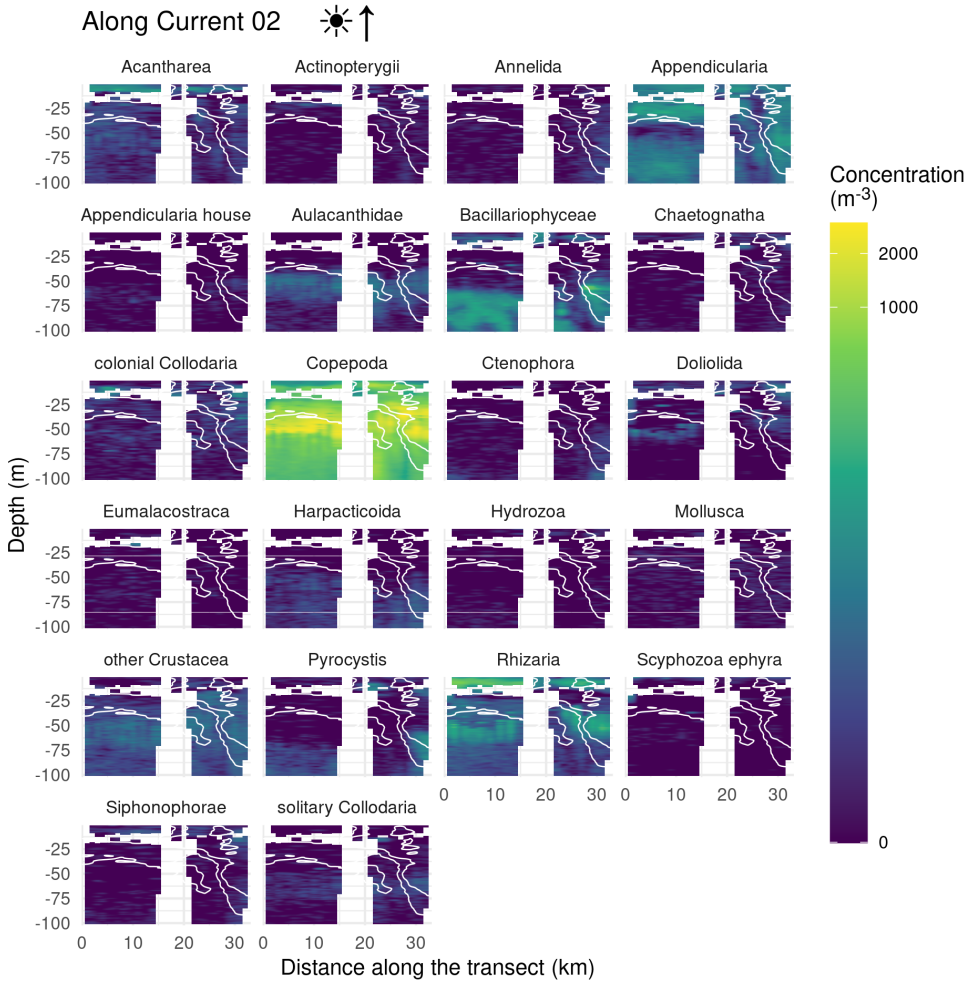
C.2 Along front transects

Besides cross front transects, 7 along front transects were performed parallel to the front. Some were conducted at dawn or dusk to visualise plankton diel vertical migration. Distributions are shown in Figure C.2.



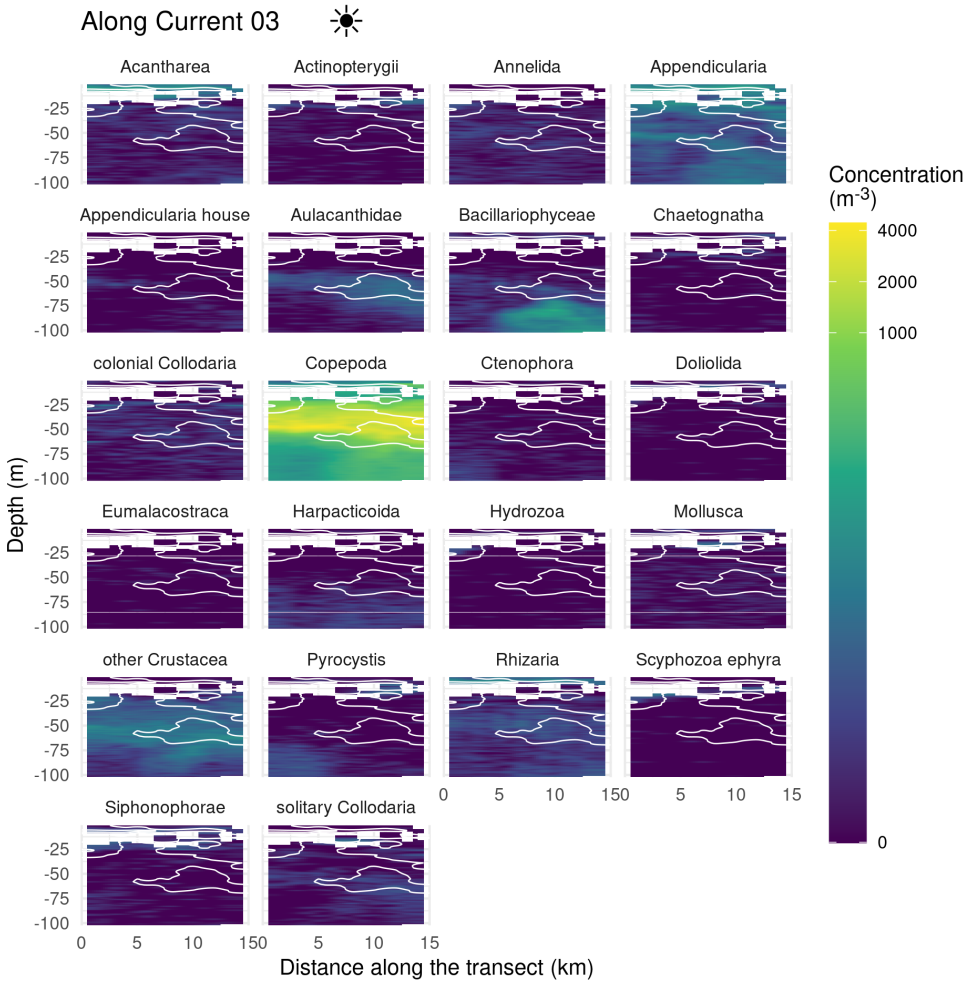
(A)

Figure C.2: *(Figure continues)*



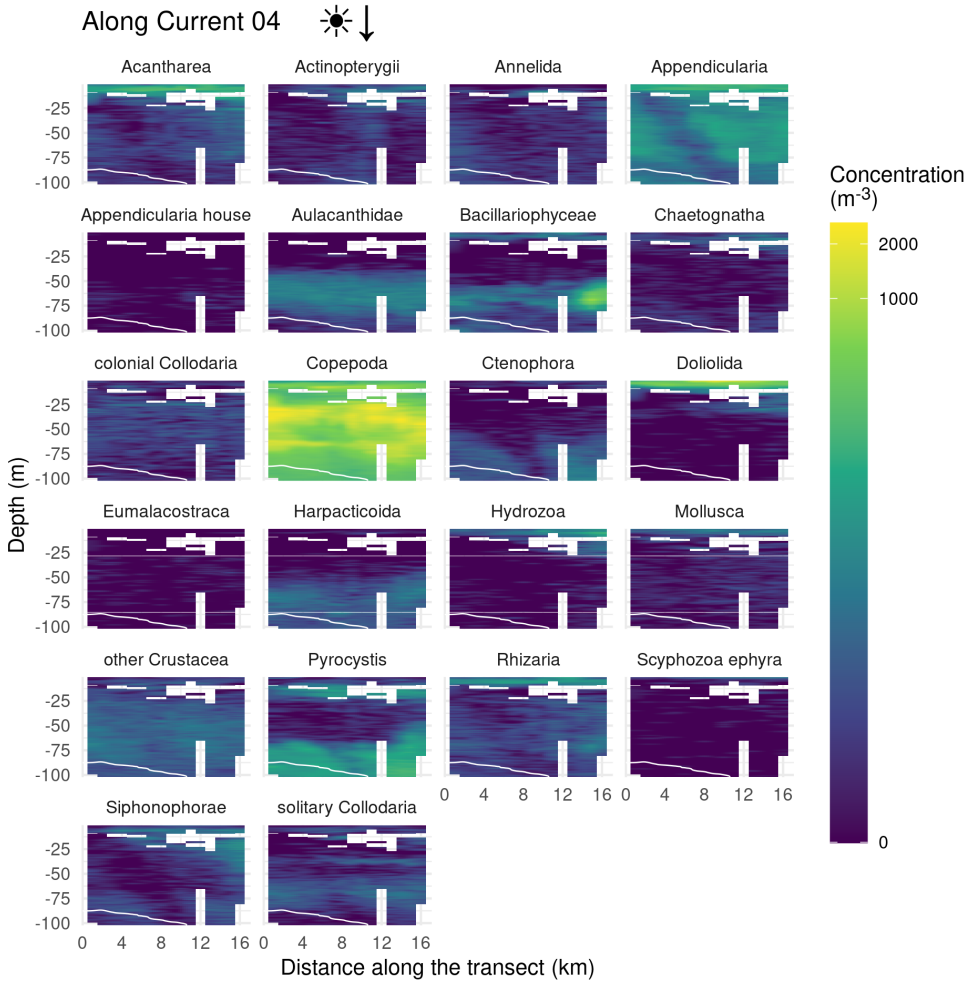
(B)

Figure C.2: *(Figure continues)*



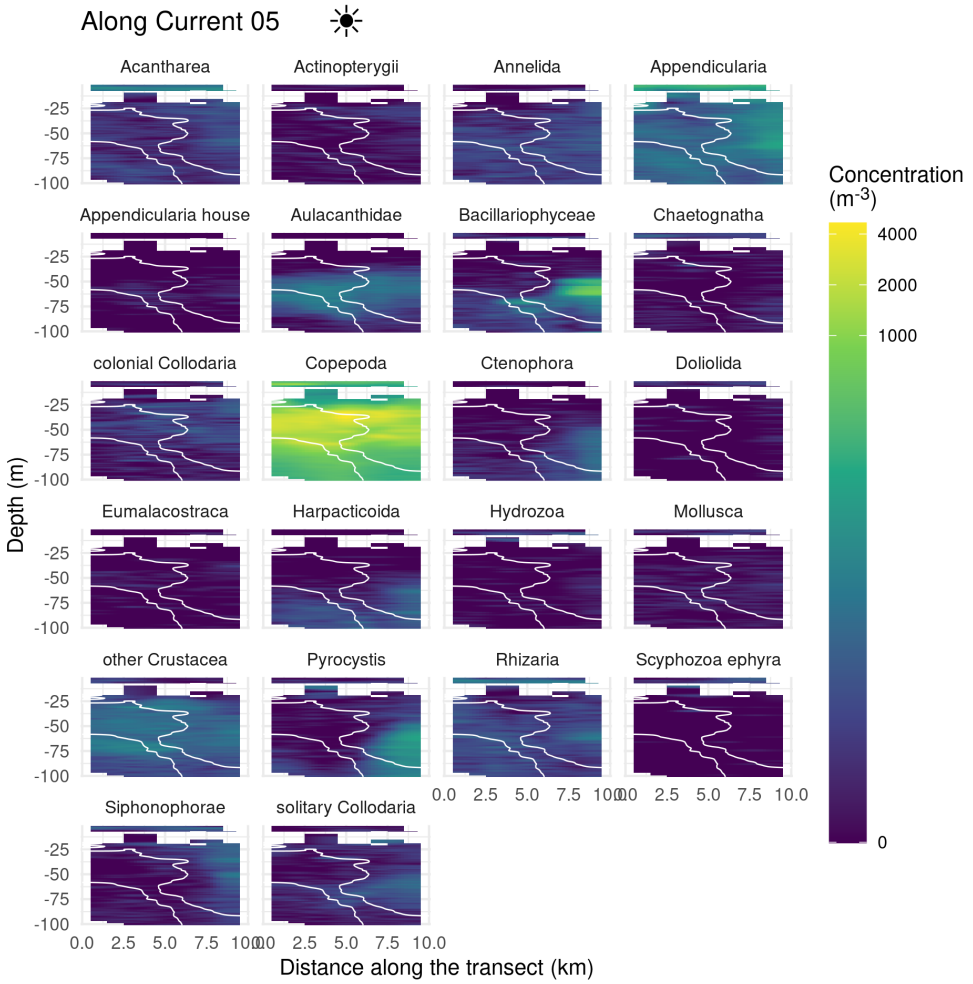
(C)

Figure C.2: *(Figure continues)*



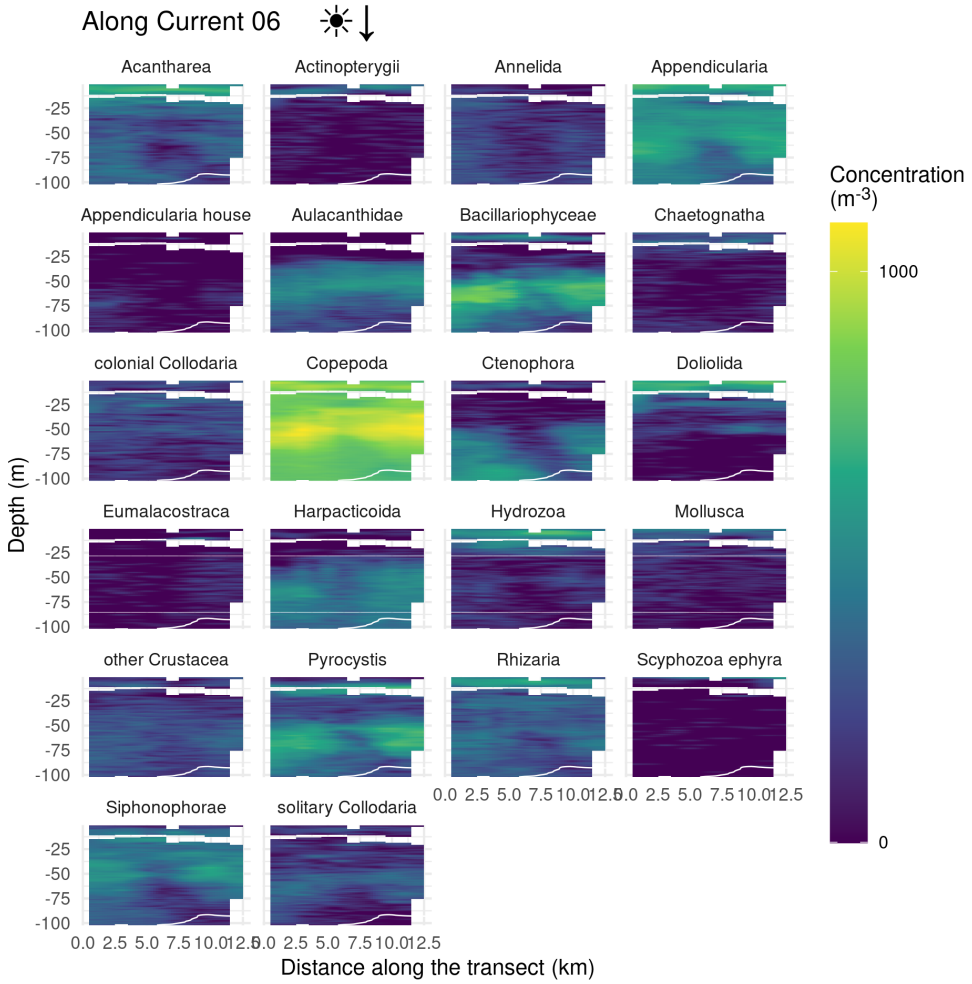
(D)

Figure C.2: *(Figure continues)*



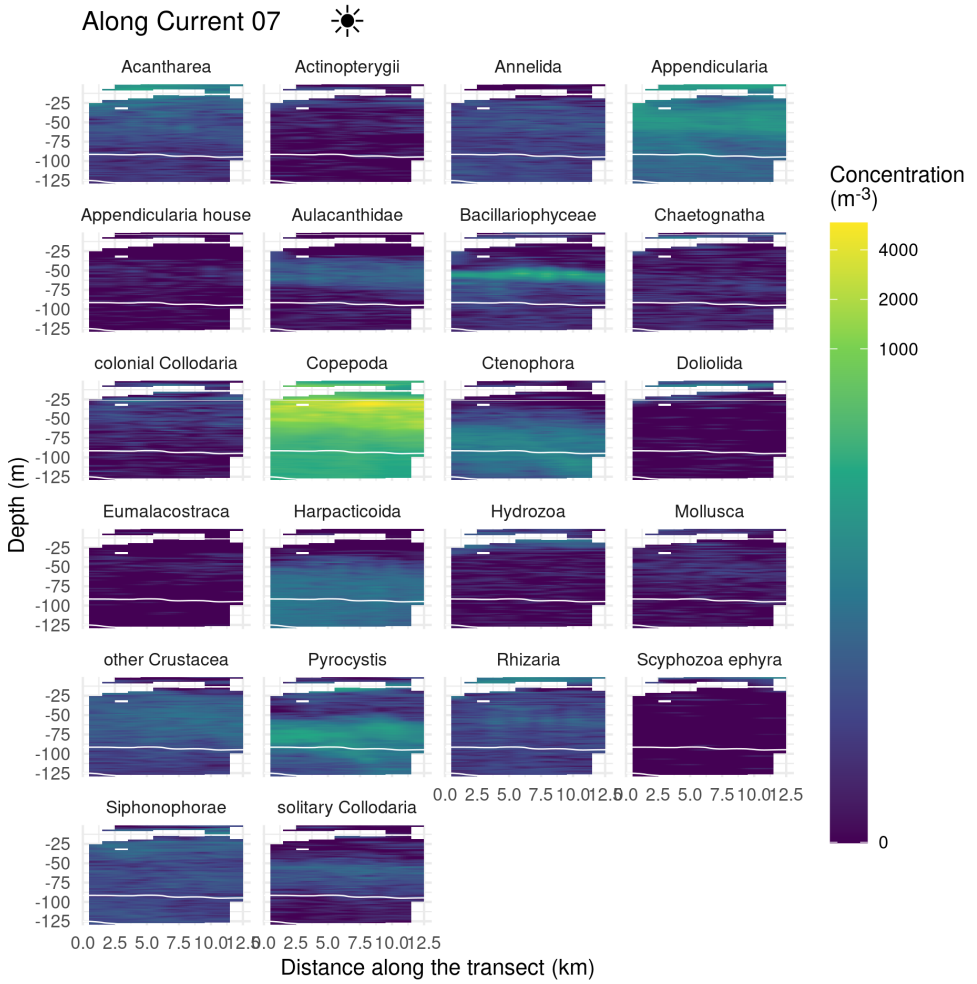
(E)

Figure C.2: *(Figure continues)*



(F)

Figure C.2: *(Figure continues)*

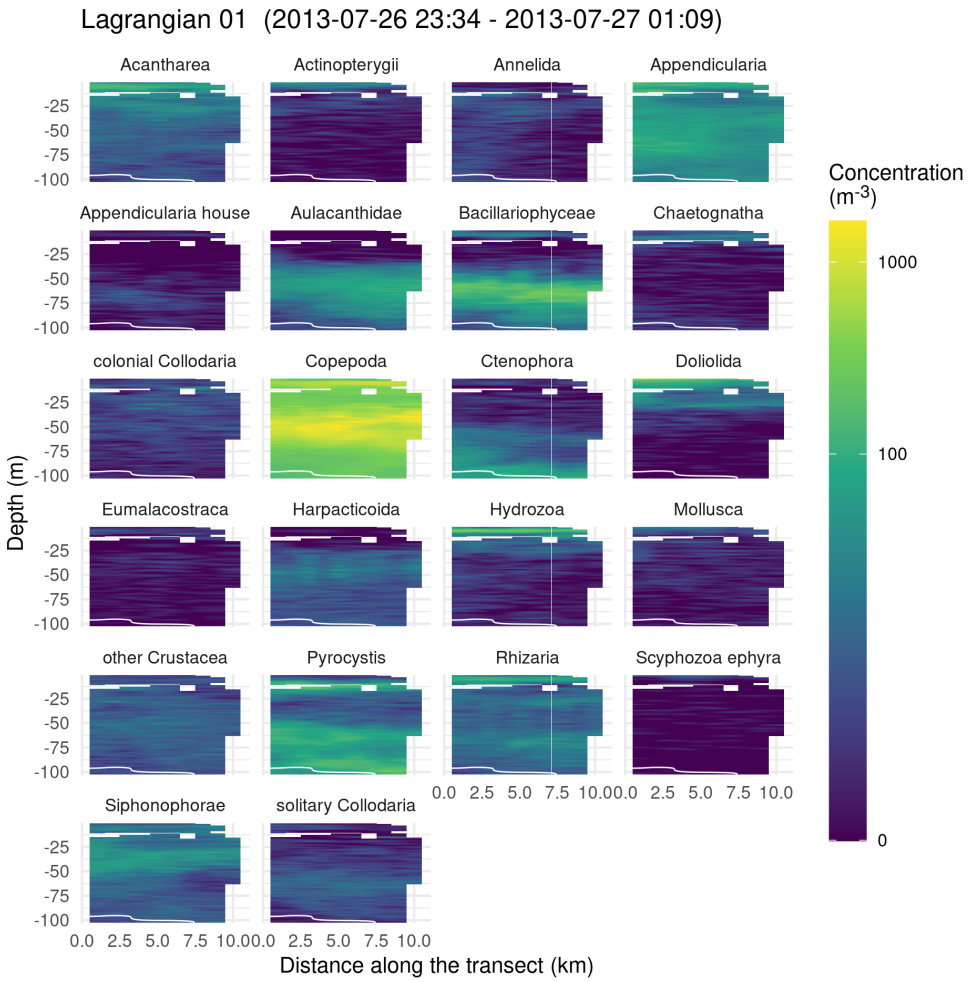


(G)

Figure C.2: Distribution maps of 22 taxonomic groups for along front transects. **(A)** Along front 1 (nighttime), **(B)** Along front 2 (dawn), **(C)** Along front 3 (daytime), **(D)** Along front 4 (dusk), **(E)** Along front 5 (daytime), **(F)** Along front 6 (dusk), **(G)** Along front 7 (daytime). White lines represent the 38.2 and 38.3 isohalines delineating the Ligurian front. Note that the colour scale is log-transformed.

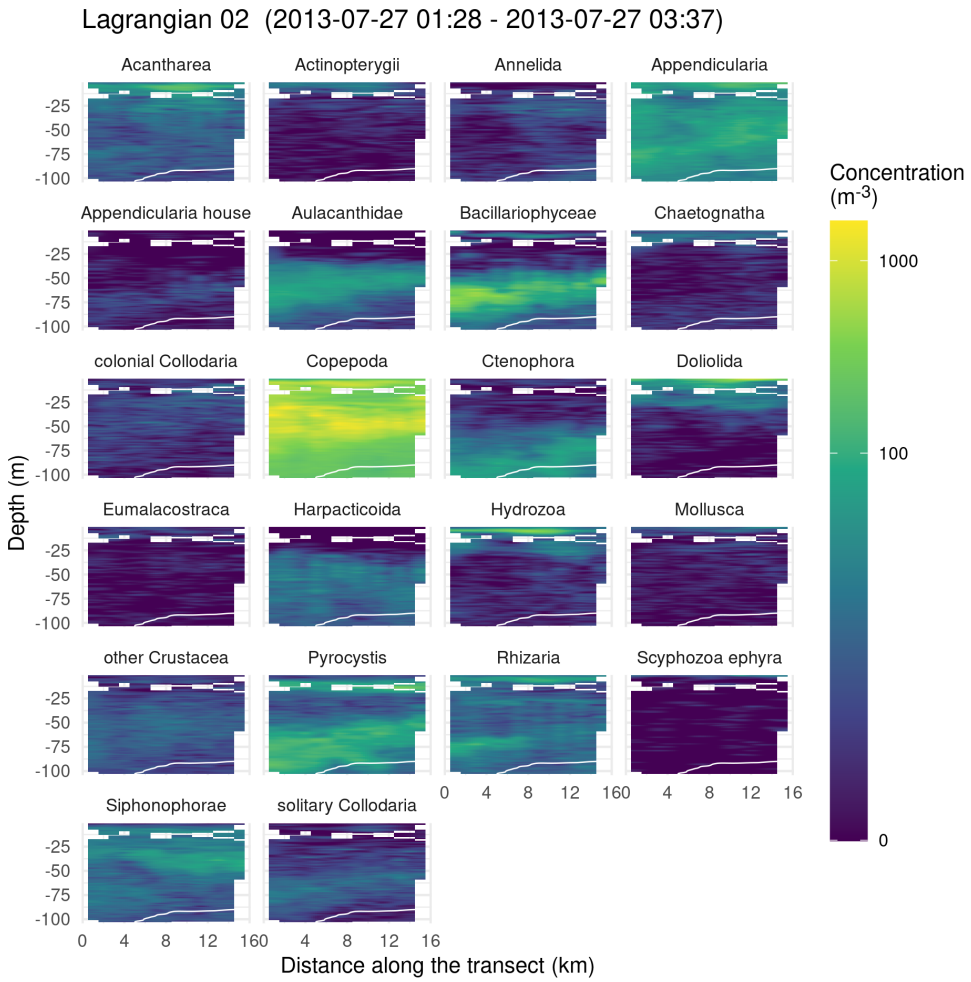
C.3 Lagrangian transects

Finally, 14 Lagrangian transects were conducted to follow a water mass for 24 h and inspect potential changes in the plankton community during this period. Unfortunately, both external drives (original and back-up) containing the data collected during the 3rd transect had malfunctions, so that data could not be retrieved. In addition, during the 10th transect, the connection with ISIIS was lost, leading to a premature end just after a few minutes of deployment. This transect is thus absent from the data. Distribution for the 12 retained transects are presented in Figure C.3.



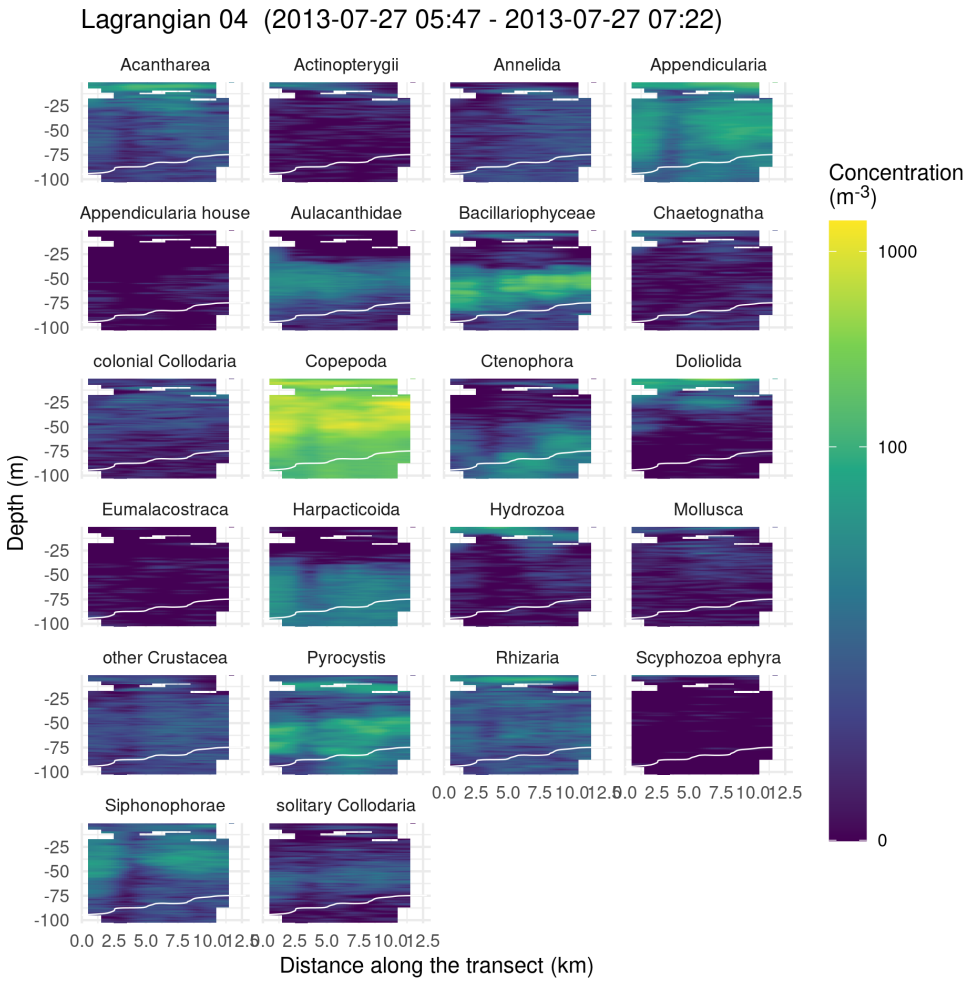
(A)

Figure C.3: *(Figure continues)*



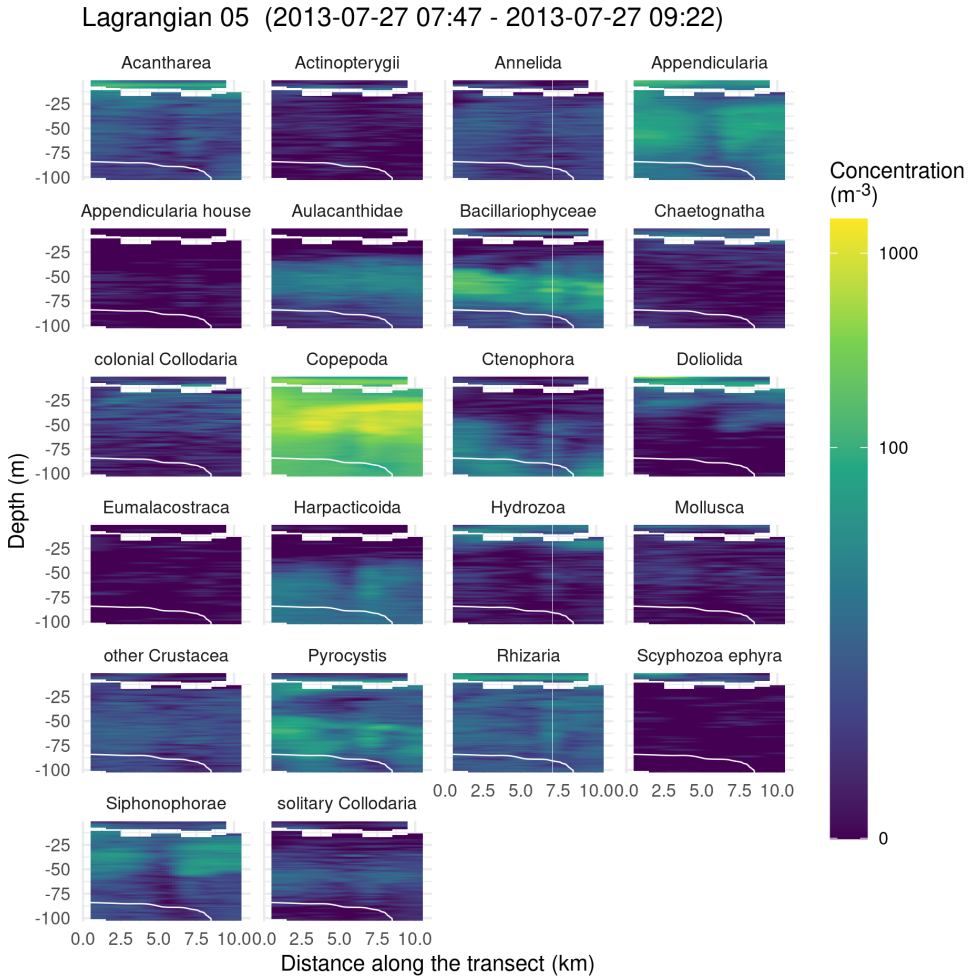
(B)

Figure C.3: *(Figure continues)*



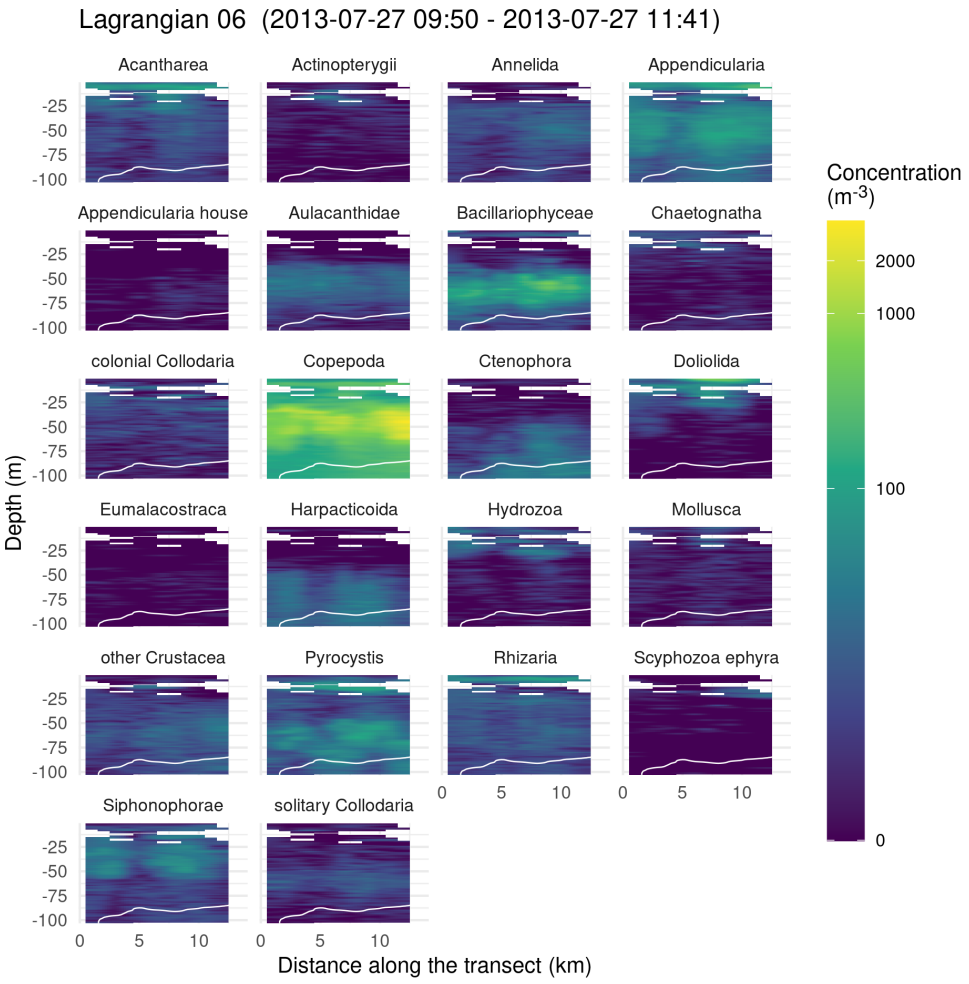
(C)

Figure C.3: *(Figure continues)*



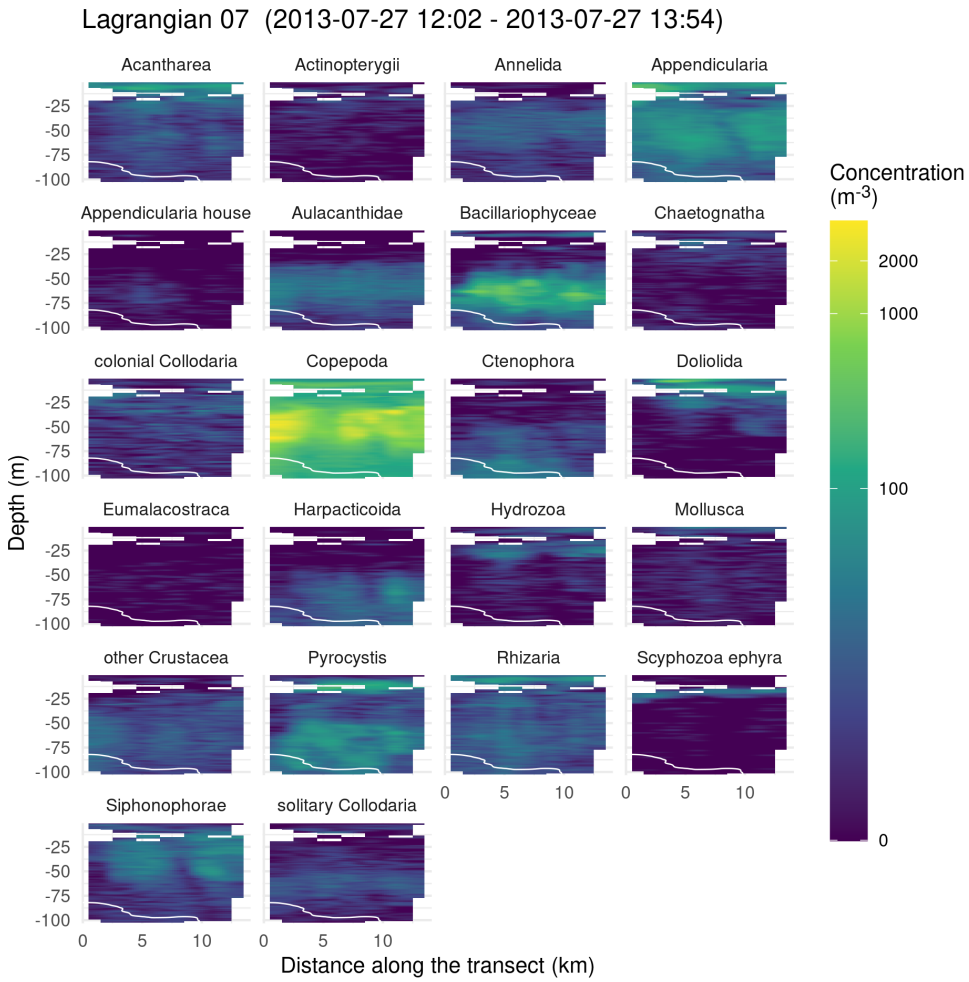
(D)

Figure C.3: *(Figure continues)*



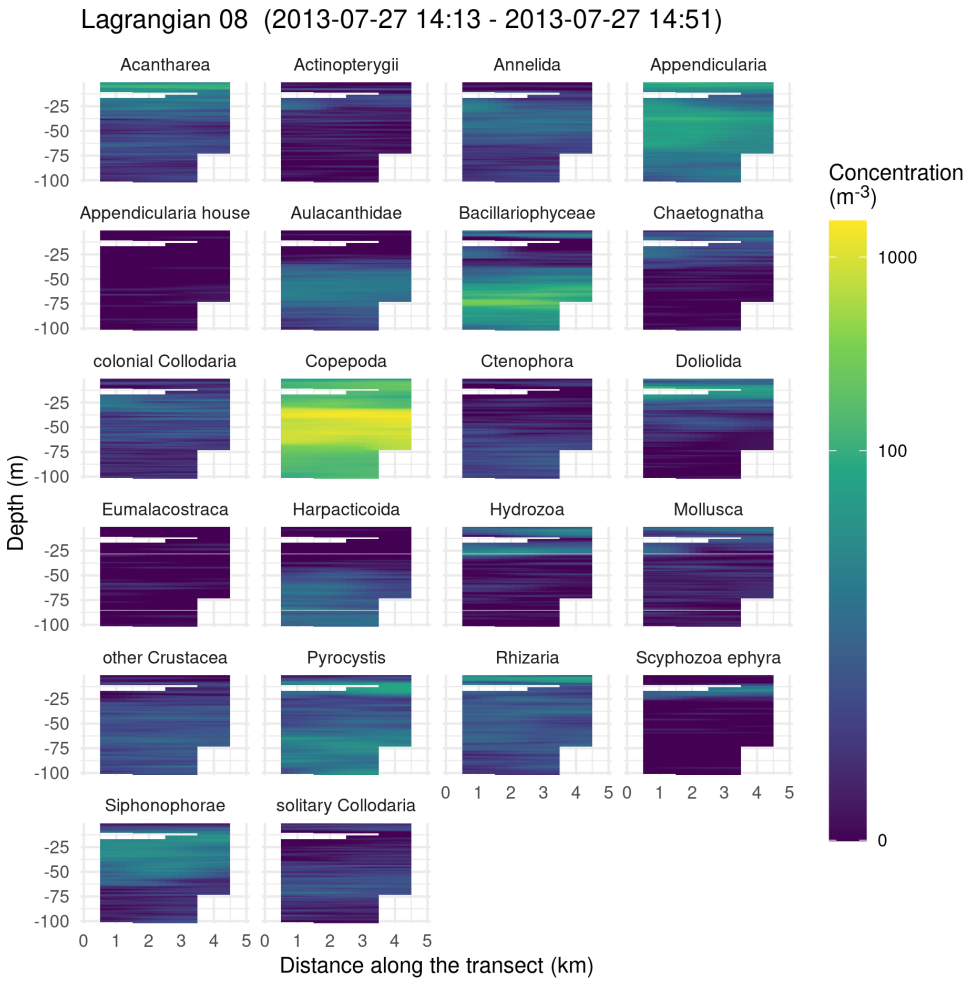
(E)

Figure C.3: *(Figure continues)*



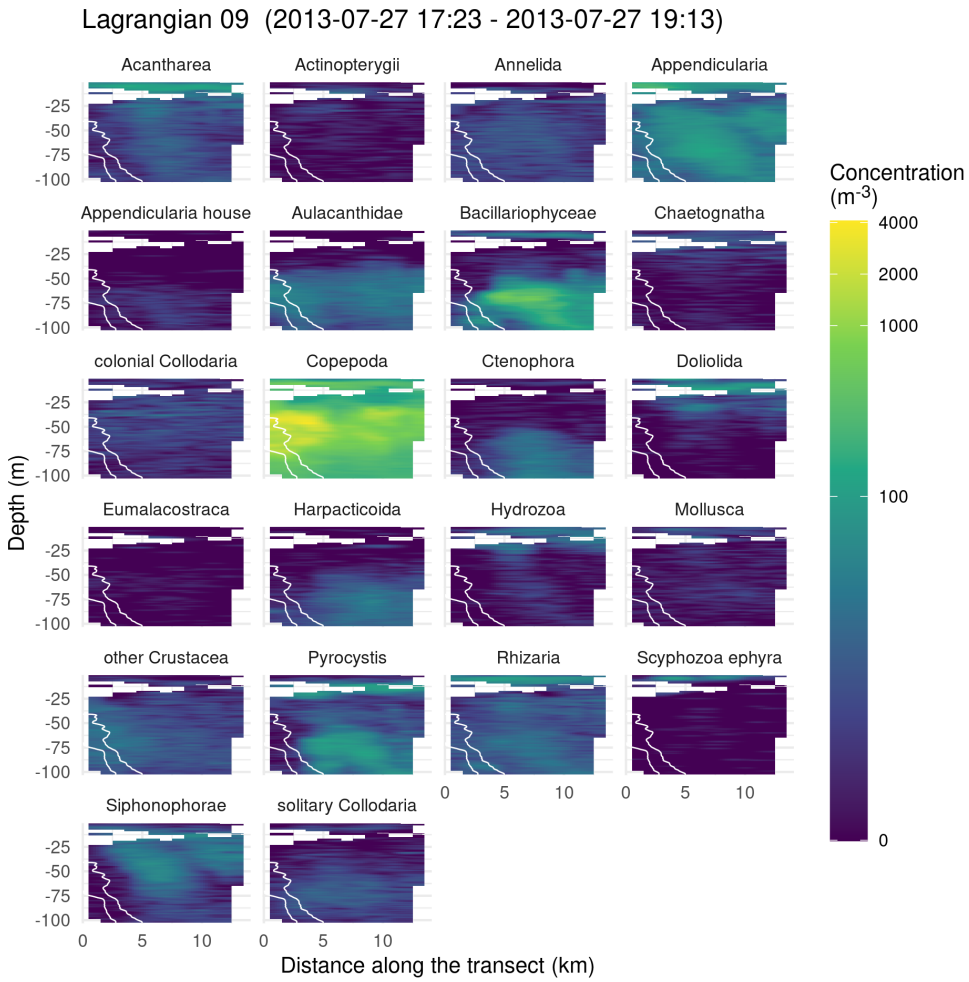
(F)

Figure C.3: *(Figure continues)*



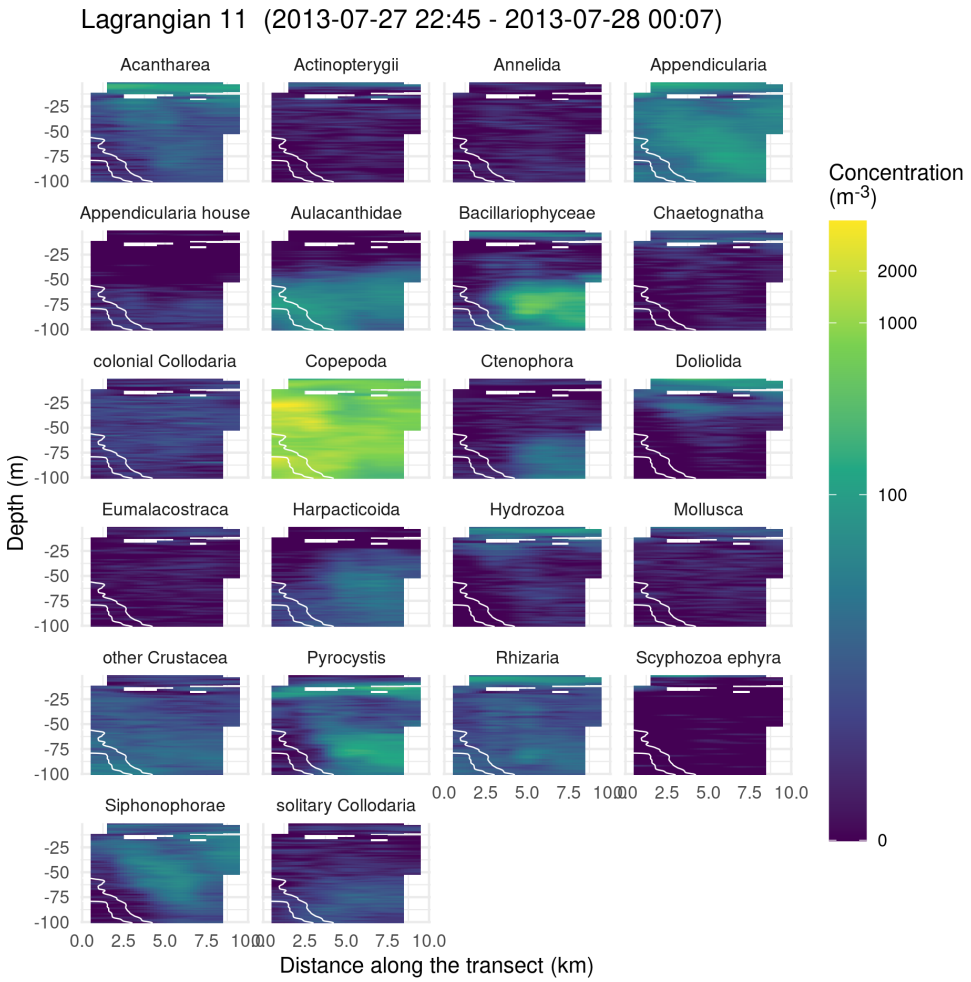
(G)

Figure C.3: *(Figure continues)*



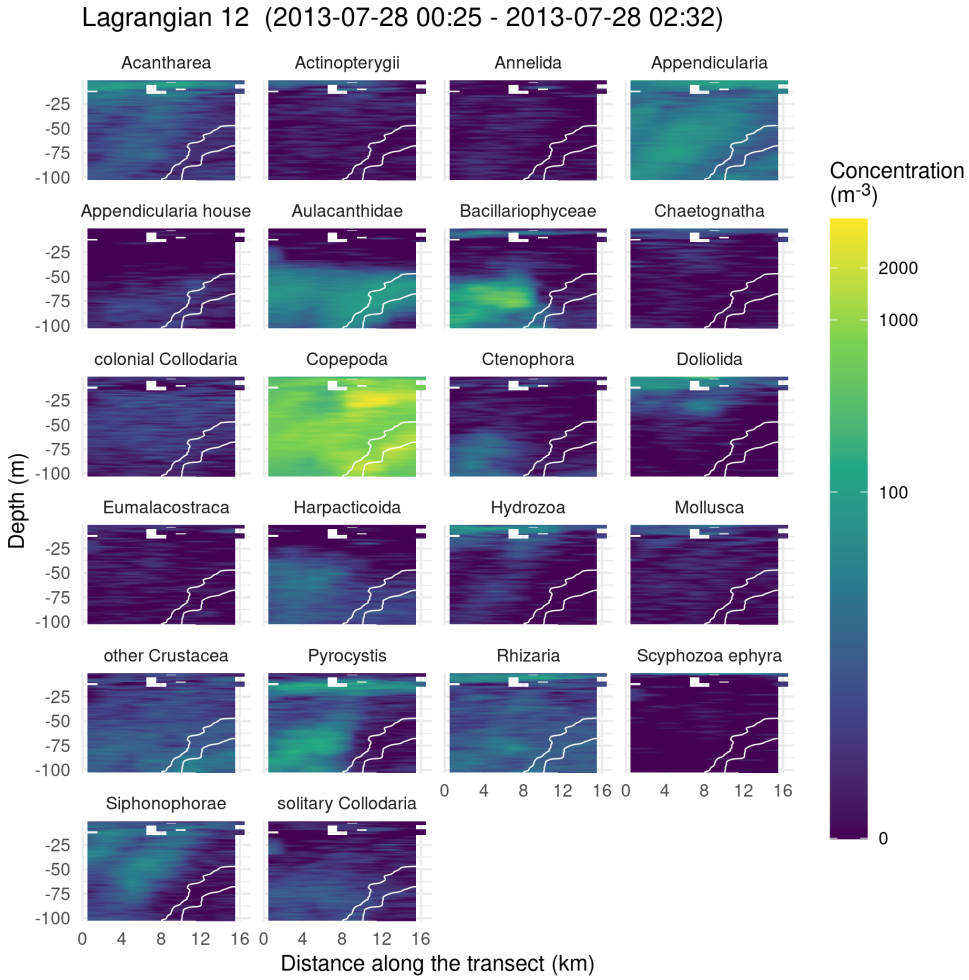
(H)

Figure C.3: *(Figure continues)*



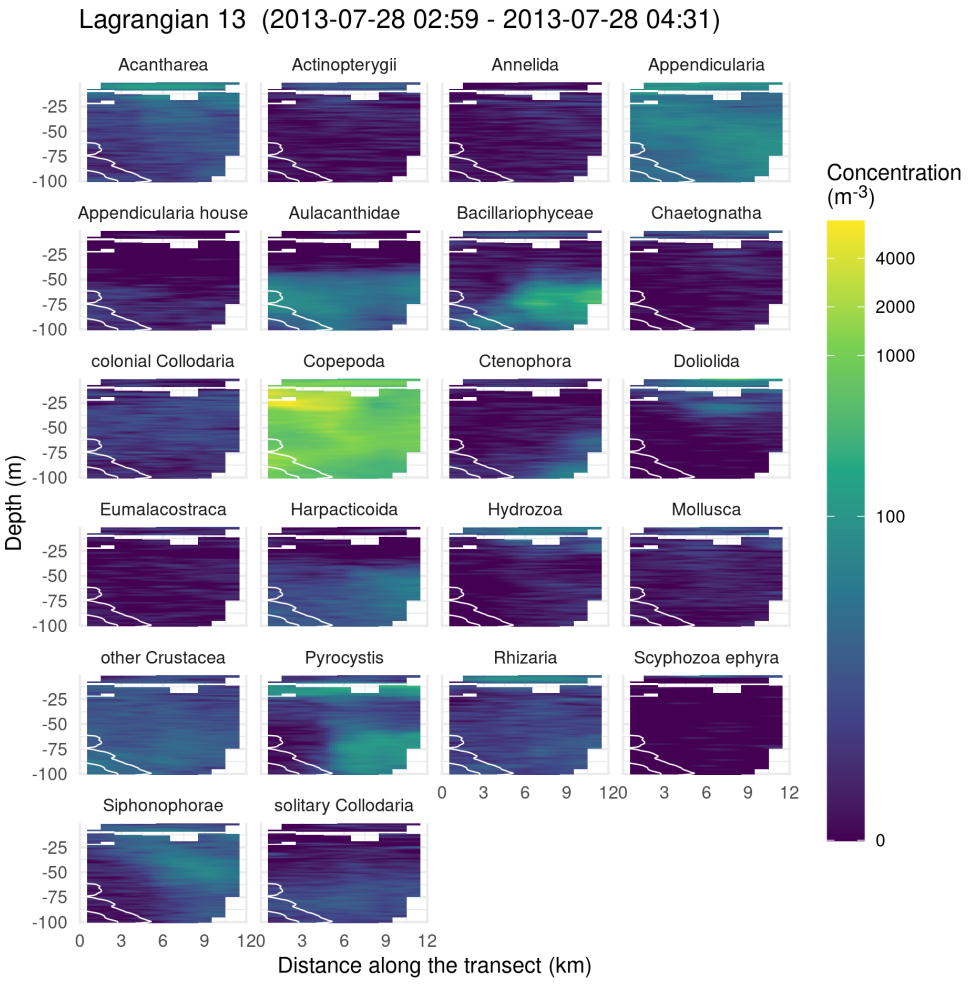
(I)

Figure C.3: *(Figure continues)*



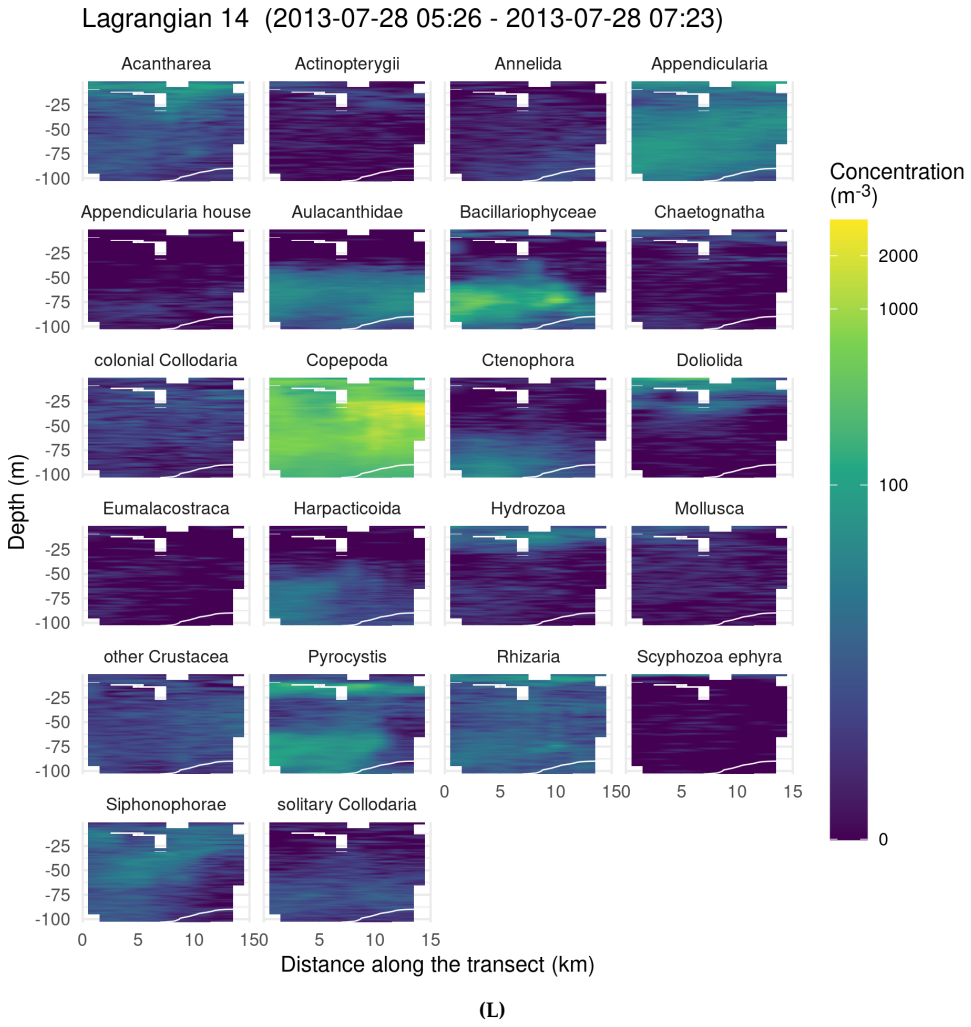
(J)

Figure C.3: *(Figure continues)*



(K)

Figure C.3: *(Figure continues)*



(L)

Figure C.3: Distribution maps of 22 taxonomic groups for Lagrangian transects. **(A)** Lagrangian 1, **(B)** Lagrangian 2, **(C)** Lagrangian 4, **(D)** Lagrangian 5, **(E)** Lagrangian 6, **(F)** Lagrangian 7, **(G)** Lagrangian 8, **(H)** Lagrangian 9, **(I)** Lagrangian 11, **(J)** Lagrangian 12, **(K)** Lagrangian 13, **(L)** Lagrangian 14. White lines represent the 38.2 and 38.3 isohalines delineating the Ligurian front. Note that the colour scale is log-transformed.

Bibliography

[1] M. Abadi et al. “TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems”. 2016.

[2] J. Alheit and M. Niquen. “Regime Shifts in the Humboldt Current Ecosystem”. In: *Progress in Oceanography*. Regime Shifts in the Ocean. Reconciling Observations and Theory 60.2 (2004), pp. 201–222. DOI: 10.1016/j.pocean.2004.02.006.

[3] A. L. Alldredge, G. Gorsky, M. Youngbluth, and D. Deibel. “The Contribution of Discarded Appendicularian Houses to the Flux of Particulate Organic Carbon from Oceanic Surface Waters”. In: *Response of marine ecosystems to global change: Ecological impact of appendicularians* (2005), pp. 309–326.

[4] A. L. Alldredge and L. P. Madin. “Pelagic Tunicates: Unique Herbivores in the Marine Plankton”. In: *BioScience* 32.8 (1982), pp. 655–663. DOI: 10.2307/1308815.

[5] A. L. Alldredge, T. J. Cowles, S. MacIntyre, J. E. B. Rines, P. L. Donaghay, C. F. Greenlaw, D. V. Holliday, M. M. Dekshenieks, J. M. Sullivan, and J. R. V. Zaneveld. “Occurrence and Mechanisms of Formation of a Dramatic Thin Layer of Marine Snow in a Shallow Pacific Fjord”. In: *Marine Ecology Progress Series* 233 (2002), pp. 1–12. DOI: 10.3354/meps233001.

[6] A. L. Alldredge, T. C. Granata, C. C. Gotschalk, and T. D. Dickey. “The Physical Strength of Marine Snow and Its Implications for Particle Disaggregation in the Ocean”. In: *Limnology and Oceanography* 35.7 (1990), pp. 1415–1428. DOI: 10.4319/lo.1990.35.7.1415.

[7] A. L. Alldredge and M. W. Silver. “Characteristics, Dynamics and Significance of Marine Snow”. In: *Progress in Oceanography* 20.1 (1988), pp. 41–82. DOI: 10.1016/0079-6611(88)90053-5.

[8] V. Allken, S. Rosen, N. O. Handegard, and K. Malde. “A Deep Learning-Based Method to Identify and Count Pelagic and Mesopelagic Fishes from Trawl Camera Images”. In: *ICES Jour-*

nal of Marine Science 78.10 (2021), pp. 3780–3792. DOI: 10.1093/icesjms/fsab227.

[9] J. M. Alston and J. A. Rick. "A Beginner's Guide to Conducting Reproducible Research". In: *The Bulletin of the Ecological Society of America* 102.2 (2021), e01801. DOI: 10.1002/bes2.1801.

[10] M. Álvarez-Noriega, S. C. Burgess, J. E. Byers, J. M. Pringle, J. P. Wares, and D. J. Marshall. "Global Biogeography of Marine Dispersal Potential". In: *Nature Ecology & Evolution* 4.9 (9 2020), pp. 1196–1203. DOI: 10.1038/s41559-020-1238-y.

[11] O. R. Anderson and S. M. Gupta. "Evidence of Binary Division in Mature Central Capsules of a Collosphaerid Colonial Radiolarian: Implications for Shell Ontogenetic Patterns in Modern and Fossil Species". In: *Palaeontologia Electronica* 1 (1998), pp. 1–13.

[12] O. R. Anderson. *Radiolaria*. Springer. New York, NY: Springer Science & Business Media, 1983. ISBN: 1-4612-5536-8.

[13] S. Anglès, A. Jordi, and L. Campbell. "Responses of the Coastal Phytoplankton Community to Tropical Cyclones Revealed by High-Frequency Imaging Flow Cytometry". In: *Limnology and Oceanography* 60.5 (2015), pp. 1562–1576. DOI: 10.1002/lno.10117.

[14] H. Auel and H. M. Verheye. "Hypoxia Tolerance in the Copepod Calanoides Carinatus and the Effect of an Intermediate Oxygen Minimum Layer on Copepod Vertical Distribution in the Northern Benguela Current Upwelling System and the Angola–Benguela Front". In: *Journal of Experimental Marine Biology and Ecology* 352.1 (2007), pp. 234–243. DOI: 10.1016/j.jembe.2007.07.020.

[15] P. Ayón, M. I. Criales-Hernandez, R. Schwamborn, and H.-J. Hirche. "Zooplankton Research off Peru: A Review". In: *Progress in Oceanography*. The Northern Humboldt Current System: Ocean Dynamics, Ecosystem Processes, and Fisheries 79.2 (2008), pp. 238–255. DOI: 10.1016/j.pocean.2008.10.020.

[16] N. Baker, H. Lu, G. Erlikhman, and P. J. Kellman. "Deep Convolutional Networks Do Not Classify Based on Global Object Shape". In: *PLOS Computational Biology* 14.12 (2018), e1006613. DOI: 10.1371/journal.pcbi.1006613.

[17] N. Baker, H. Lu, G. Erlikhman, and P. J. Kellman. "Local Features and Global Shape Information in Object Classification by Deep Convolutional Neural Networks". In: *Vision Research* 172 (2020), pp. 46–61. DOI: 10.1016/j.visres.2020.04.003.

[18] R. G. Baraniuk. "More Is Less: Signal Processing and the Data Deluge". In: *Science* 331.6018 (2011), pp. 717–719. DOI: 10.1126/science.1197448.

[19] G. Basterretxea, J. S. Font-Muñoz, and I. Tuval. "Phytoplankton Orientation in a Turbulent Ocean: A Microscale Perspective". In: *Frontiers in Marine Science* 7 (2020).

[20] G. Beaugrand, C. Luczak, and M. Edwards. "Rapid Biogeographical Plankton Shifts in the North Atlantic Ocean". In: *Global Change Biology* 15.7 (2009), pp. 1790–1803. DOI: 10.1111/j.1365-2486.2009.01848.x.

[21] G. Beaugrand, M. Edwards, and L. Legendre. "Marine Biodiversity, Ecosystem Functioning, and Carbon Cycles". In: *Proceedings of the National Academy of Sciences* 107.22 (2010), pp. 10120–10124. DOI: 10.1073/PNAS.0913855107.

[22] G. Beaugrand, P. C. Reid, F. Ibañez, J. A. Lindley, and M. Edwards. "Reorganization of North Atlantic Marine Copepod Biodiversity and Climate". In: *Science* 296.5573 (2002), pp. 1692–1694. DOI: 10.1126/science.1071329.

[23] G. Beaugrand, I. Rombouts, and R. R. Kirby. "Towards an Understanding of the Pattern of Biodiversity in the Oceans". In: *Global Ecology and Biogeography* 22.4 (2013), pp. 440–449. DOI: 10.1111/geb.12009.

[24] M. J. Behrenfeld. "Climate-Mediated Dance of the Plankton". In: *Nature Climate Change* 4.10 (2014), pp. 880–887. DOI: 10.1038/nclimate2349.

[25] M. J. Behrenfeld and E. S. Boss. "Resurrecting the Ecological Underpinnings of Ocean Plankton Blooms". In: *Annual Review of Marine Science* 6.1 (2014), pp. 167–194. DOI: 10.1146/annurev-marine-052913-021325.

[26] I. M. Belkin, P. C. Cornillon, and K. Sherman. "Fronts in Large Marine Ecosystems". In: *Prog. Oceanogr.* 81.1-4 (2009), pp. 223–236. DOI: 10.1016/J.POCEAN.2009.04.015.

[27] G. Bell, T. Hey, and A. Szalay. "Beyond the Data Deluge". In: *Science* 323.5919 (2009), pp. 1297–1298. DOI: 10.1126/science.1170411.

[28] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick. "Inside-Outside Net: Detecting Objects in Context With Skip Pooling and Recurrent Neural Networks". In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016, pp. 2874–2883.

[29] M. Benavides et al. "Sinking Trichodesmium Fixes Nitrogen in the Dark Ocean". In: *The ISME Journal* (2022), pp. 1–8. DOI: 10.1038/s41396-022-01289-6.

[30] M. Benfield et al. "RAPID: Research on Automated Plankton Identification". In: *Oceanography* 20.2 (2007), pp. 172–187. DOI: 10.5670/oceanog.2007.63.

[31] K. J. Benoit-Bird, M. A. Moline, C. M. Waluk, and I. C. Robbins. "Integrated Measurements of Acoustical and Optical Thin Layers I: Vertical Scales of Association". In: *Continental Shelf Research*. The Ecology and Oceanography of Thin Plankton Layers 30.1 (2010), pp. 17–28. DOI: 10.1016/j.csr.2009.08.001.

[32] K. J. Benoit-Bird et al. "Prey Patch Patterns Predict Habitat Use by Top Marine Predators with Diverse Foraging Strategies". In: *PLOS ONE* 8.1 (2013), e53348. DOI: 10.1371/journal.pone.0053348.

[33] K. Benoit-Bird. "Dynamic 3-Dimensional Structure of Thin Zooplankton Layers Is Impacted by Foraging Fish". In: *Mar. Ecol. Prog. Ser.* 396 (2009), pp. 61–76. DOI: 10.3354/meps08316.

[34] C. Berney et al. "UniEuk: Time to Speak a Common Language in Protistology!" In: *Journal of Eukaryotic Microbiology* 64.3 (2017), pp. 407–411. DOI: 10.1111/jeu.12414.

[35] R. E. Bernstein, P. R. Betzer, R. A. Feely, R. H. Byrne, M. F. Lamb, and A. F. Michaels. "Acantharian Fluxes and Strontium to Chlorinity Ratios in the North Pacific Ocean". In: *Science* 237.4821 (1987), pp. 1490–1494. DOI: 10.1126/science.237.4821.1490.

[36] R. E. Bernstein, P. R. Betzer, and K. Takahashi. "Radiolarians from the Western North Pacific Ocean: A Latitudinal Study of Their Distributions and Fluxes". In: *Deep Sea Research Part A. Oceanographic Research Papers* 37.11 (1990), pp. 1677–1696. DOI: 10.1016/0198-0149(90)90071-3.

[37] H. Bi, Z. Guo, M. C. Benfield, C. Fan, M. Ford, S. Shahrestani, and J. M. Sieracki. "A Semi-Automated Image Analysis Procedure for In Situ Plankton Imaging Systems". In: *PLOS ONE* 10.5 (2015), e0127121. DOI: 10.1371/journal.pone.0127121.

[38] T. Biard. "Diversity and Ecology of Radiolaria in Modern Oceans". In: *Environmental Microbiology* 24.5 (2022), pp. 2179–2200. DOI: 10.1111/1462-2920.16004.

[39] T. Biard, J. W. Krause, M. R. Stukel, and M. D. Ohman. "The Significance of Giant Phaeodarians (Rhizaria) to Biogenic Silica Export in the California Current Ecosystem". In: *Global Biogeochemical Cycles* 32.6 (2018), pp. 987–1004. DOI: 10.1029/2018GB005877.

[40] T. Biard and M. D. Ohman. "Vertical Niche Definition of Test-Bearing Protists (Rhizaria) into the Twilight Zone Revealed by in Situ Imaging". In: *Limnology and Oceanography* 65.11 (2020), pp. 2583–2602. DOI: 10.1002/lno.11472.

[41] T. Biard, L. Pillet, J. Decelle, C. Poirier, N. Suzuki, and F. Not. "Towards an Integrative Morpho-molecular Classification of the Collodaria (Polycystinea, Radiolaria)". In: *Protist* 166.3 (2015), pp. 374–388. DOI: 10.1016/j.protis.2015.05.002.

[42] T. Biard, L. Stemmann, M. Picheral, N. Mayot, P. Vandromme, H. Hauss, G. Gorsky, L. Guidi, R. Kiko, and F. Not. "In Situ Imaging Reveals the Biomass of Giant Protists in the Global Ocean". In: *Nature* 532.7600 (2016), pp. 504–507. DOI: 10.1038/nature17652.

[43] M. B. Blaschko et al. "Automatic In Situ Identification of Plankton". In: *2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05) - Volume 1*. 2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05) - Volume 1. Vol. 1. 2005, pp. 79–86. DOI: 10.1109/ACVMOT.2005.29.

[44] S. Bojinski, M. Verstraete, T. C. Peterson, C. Richter, A. Simmons, and M. Zemp. "The Concept of Essential Climate Variables in Support of Climate Research, Applications, and Policy". In:

Bulletin of the American Meteorological Society 95.9 (2014), pp. 1431–1443. DOI: 10.1175/BAMS-D-13-00047.1.

[45] D. Borcard, F. Gillet, and P. Legendre. *Numerical Ecology with R*. Springer, 2018. ISBN: 3-319-71404-X.

[46] A. Bosse et al. "Scales and Dynamics of Submesoscale Coherent Vortices Formed by Deep Convection in the Northwestern Mediterranean Sea". In: *Journal of Geophysical Research: Oceans* 121.10 (2016), pp. 7716–7742. DOI: 10.1002/2016JC012144@10.1002/(ISSN)2169-9291.DENSEWATER01.

[47] A. Bosse et al. "A Submesoscale Coherent Vortex in the Ligurian Sea: From Dynamical Barriers to Biological Implications". In: *Journal of Geophysical Research: Oceans* 122.8 (2017), pp. 6196–6217. DOI: 10.1002/2016JC012634.

[48] J. Boucher, F. Ibanez, and L. Prieur. "Daily and Seasonal Variations in the Spatial Distribution of Zooplankton Populations in Relation to the Physical Structure in the Ligurian Sea Front". In: *Journal of Marine Research* 45.1 (1987), pp. 133–173. DOI: 10.1357/002224087788400891.

[49] J. Boucher. "Localization of Zooplankton Populations in the Ligurian Marine Front: Role of Ontogenic Migration". In: *Deep Sea Res. Part A. Oceanogr. Res. Pap.* 31.5 (1984), pp. 469–484. DOI: 10.1016/0198-0149(84)90097-9.

[50] C.-F. Boudouresque. "Taxonomy and Phylogeny of Unicellular Eukaryotes". In: *Environmental Microbiology: Fundamentals and Applications: Microbial Ecology*. Ed. by J.-C. Bertrand, P. Caumette, P. Lebaron, R. Matheron, P. Normand, and T. Sime-Ngando. Dordrecht: Springer Netherlands, 2015, pp. 191–257. ISBN: 978-94-017-9118-2. DOI: 10.1007/978-94-017-9118-2_7.

[51] P. W. Boyd and P. P. Newton. "Does Planktonic Community Structure Determine Downward Particulate Organic Carbon Flux in Different Oceanic Provinces?" In: *Deep Sea Research Part I: Oceanographic Research Papers* 46.1 (1999), pp. 63–91. DOI: 10.1016/S0967-0637(98)00066-1.

[52] P. Boyd and P. Newton. "Evidence of the Potential Influence of Planktonic Community Structure on the Interannual Variability of Particulate Organic Carbon Flux". In: *Deep Sea Research Part I: Oceanographic Research Papers* 42.5 (1995), pp. 619–639. DOI: 10.1016/0967-0637(95)00017-Z.

[53] T. P. Boyer et al. “World Ocean Atlas 2018. NOAA National Centers for Environmental Information. Dataset.” In: (2018).

[54] M. Bramer. *Principles of Data Mining*. Undergraduate Topics in Computer Science. London: Springer, 2016. ISBN: 978-1-4471-7306-9 978-1-4471-7307-6. DOI: 10.1007/978-1-4471-7307-6.

[55] M. C. Brandão et al. “Macroscale Patterns of Oceanic Zooplankton Composition and Size Structure”. In: *Scientific Reports* 11.1 (1 2021), p. 15714. DOI: 10.1038/s41598-021-94615-5.

[56] K. Brandt. “Beiträge Zur Kenntnis Der Colliden.” In: *Archiv für Protistenkunde* 1 (1902), pp. 59–88.

[57] L. Breiman. “Random Forests”. In: *Machine Learning* 45.1 (2001), pp. 5–32. DOI: 10.1023/A:1010933404324.

[58] A. Bricaud and A. Morel. “Light Attenuation and Scattering by Phytoplanktonic Cells: A Theoretical Modeling”. In: *Applied Optics* 25.4 (1986), pp. 571–580. DOI: 10.1364/AO.25.000571.

[59] C. Briseño-Avena, M. S. Schmid, K. Swieca, S. Sponaugle, R. D. Brodeur, and R. K. Cowen. “Three-Dimensional Cross-Shelf Zooplankton Distributions off the Central Oregon Coast during Anomalous Oceanographic Conditions”. In: *Progress in Oceanography* 188 (2020), p. 102436. DOI: 10.1016/j.pocean.2020.102436.

[60] J. H. Brown and M. V. Lomolino. *Biogeography*. Sunderland, Massachussetts: Sinauer Associates, Inc., 1998.

[61] J. Browne and E. J. Browne. *The Secular Ark: Studies in the History of Biogeography*. JSTOR, 1983. ISBN: 0-300-02460-6.

[62] K. W. Bruland and M. W. Silver. “Sinking Rates of Fecal Pellets from Gelatinous Zooplankton (Salps, Pteropods, Doliolids)”. In: *Marine Biology* 63.3 (1981), pp. 295–300. DOI: 10.1007/BF00395999.

[63] P. Brun, M. R. Payne, and T. Kiørboe. “Trait Biogeography of Marine Copepods – an Analysis across Scales”. In: *Ecology Letters* 19.12 (2016), pp. 1403–1413. DOI: 10.1111/ele.12688.

[64] A. B. Burd and G. A. Jackson. “Particle Aggregation”. In: *Annual Review of Marine Science* 1.1 (2009), pp. 65–90. DOI: 10.1146/annurev.marine.010908.163904.

[65] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos. "A Unified Multi-scale Deep Convolutional Neural Network for Fast Object Detection". 2016.

[66] D. G. Capone, J. P. Zehr, H. W. Paerl, B. Bergman, and E. J. Carpenter. "Trichodesmium, a Globally Significant Marine Cyanobacterium". In: *Science* 276.5316 (1997), pp. 1221–1229. DOI: 10.1126/science.276.5316.1221.

[67] D. A. Caron. "Mixotrophy Stirs up Our Understanding of Marine Food Webs". In: *Proceedings of the National Academy of Sciences* 113.11 (2016), pp. 2806–2808. DOI: 10.1073/pnas.1600718113.

[68] D. A. Caron. "The Rise of Rhizaria". In: *Nature* 532.7600 (7600 2016), pp. 444–445. DOI: 10.1038/nature17892.

[69] D. A. Caron, P. D. Countway, A. C. Jones, D. Y. Kim, and A. Schnetzer. "Marine Protistan Diversity". In: *Annual review of marine science* 4.1 (2012), pp. 467–493.

[70] D. A. Caron et al. "Probing the Evolution, Ecology and Physiology of Marine Protists Using Transcriptomics". In: *Nature Reviews Microbiology* 15.1 (1 2017), pp. 6–20. DOI: 10.1038/nrmicro.2016.160.

[71] E. L. Cavan, M. Trimmer, F. Shelley, and R. Sanders. "Remineralization of Particulate Organic Carbon in an Ocean Oxygen Minimum Zone". In: *Nature Communications* 8.1 (1 2017), p. 14847. DOI: 10.1038/ncomms14847.

[72] F. Chai, K. S. Johnson, H. Claustre, X. Xing, Y. Wang, E. Boss, S. Riser, K. Fennel, O. Schofield, and A. Sutton. "Monitoring Ocean Biogeochemistry with Autonomous Platforms". In: *Nature Reviews Earth & Environment* 1.6 (6 2020), pp. 315–326. DOI: 10.1038/s43017-020-0053-y.

[73] K. Chellapilla, S. Puri, and P. Simard. "High Performance Convolutional Neural Networks for Document Processing". In: Tenth International Workshop on Frontiers in Handwriting Recognition. Suvisoft, 2006.

[74] B. Chen, H. Liu, W. Xiao, L. Wang, and B. Huang. "A Machine-Learning Approach to Modeling Picophytoplankton Abundances in the South China Sea". In: *Progress in Oceanography* 189 (2020), p. 102456. DOI: 10.1016/j.pocean.2020.102456.

[75] T. Chen, J. Li, W. Ju, and J. Sun. "Object Detection and Abundance Analysis for Fountain-Flow Imaging of Marine Plankton". In: *OCEANS 2021: San Diego–Porto*. IEEE, 2021, pp. 1–9. ISBN: 0-692-93559-2.

[76] K. Cheng, X. Cheng, Y. Wang, H. Bi, and M. C. Benfield. "Enhanced Convolutional Neural Network for Plankton Identification and Enumeration". In: *PLOS ONE* 14.7 (10 juil. 2019), e0219570. DOI: 10.1371/journal.pone.0219570.

[77] S. Christiansen et al. "Particulate Matter Flux Interception in Oceanic Mesoscale Eddies by the Polychaete Poeobius Sp." In: *Limnology and Oceanography* 63.5 (2018), pp. 2093–2109. DOI: 10.1002/lno.10926.

[78] H. Claustre, L. Legendre, P. W. Boyd, and M. Levy. "The Oceans' Biological Carbon Pumps: Framework for a Research Observational Community Approach". In: *Frontiers in Marine Science* 8 (2021).

[79] F. Colas et al. "The ZooCAM, a New in-Flow Imaging System for Fast Onboard Counting, Sizing and Classification of Fish Eggs and Metazooplankton". In: *Progress in Oceanography*. Multidisciplinary Integrated Surveys 166 (2018), pp. 54–65. DOI: 10.1016/j.pocean.2017.10.014.

[80] N. Cooper. "A Guide to Reproducible Code in Ecology and Evolution". In: (2017).

[81] C. Cortes and V. Vapnik. "Support-Vector Networks". In: *Machine Learning* 20.3 (1995), pp. 273–297. DOI: 10.1007/BF00994018.

[82] M. J. Costello and S. Breyer. "Ocean Depths: The Mesopelagic and Implications for Global Warming". In: *Current Biology* 27.1 (2017), R36–R38. DOI: 10.1016/j.cub.2016.11.042.

[83] M. J. Costello, P. Tsai, P. S. Wong, A. K. L. Cheung, Z. Basher, and C. Chaudhary. "Marine Biogeographic Realms and Species Endemicity". In: *Nature Communications* 8.1 (1 2017), p. 1057. DOI: 10.1038/s41467-017-01121-2.

[84] M. J. Costello, Z. Basher, R. Sayre, S. Breyer, and D. J. Wright. "Stratifying Ocean Sampling Globally and with Depth to Account for Environmental Variability". In: *Scientific Reports* 8.1 (2018), pp. 1–9. DOI: 10.1038/s41598-018-29419-1.

[85] R. K. Cowen and C. M. Guigand. "In Situ Ichthyoplankton Imaging System (ISIIS): System Design and Preliminary Results". In: *Limnology and Oceanography: Methods* 6.2 (2008), pp. 126–132. DOI: 10.4319/lom.2008.6.126.

[86] R. K. Cowen, S. Sponaugle, K. Robinson, J. Luo, O. S. University, and H. M. S. Center. *PlanktonSet 1.0: Plankton Imagery Data Collected from F.G. Walton Smith in Straits of Florida from 2014-06-03 to 2014-06-06 and Used in the 2015 National Data Science Bowl (NCEI Accession 0127422)*. 2015.

[87] J. Cui, B. Wei, C. Wang, Z. Yu, H. Zheng, B. Zheng, and H. Yang. "Texture and Shape Information Fusion of Convolutional Neural Network for Plankton Image Classification". In: *2018 OCEANS - MTS/IEEE Kobe Techno-Oceans (OTO)*. 2018 OCEANS - MTS/IEEE Kobe Techno-Oceans (OTO). 2018, pp. 1–5. DOI: 10.1109/OCEANSKOBE.2018.8559156.

[88] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie. "Class-Balanced Loss Based on Effective Number of Samples". In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019, pp. 9268–9277.

[89] P. Culverhouse et al. "Automatic Classification of Field-Collected Dinoflagellates by Artificial Neural Network". In: *Marine Ecology Progress Series* 139 (1996), pp. 281–287. DOI: 10.3354/meps139281.

[90] P. F. Culverhouse, R. Williams, B. Reguera, V. Herry, and S. González-Gil. "Do Experts Make Mistakes? A Comparison of Human and Machine Indentification of Dinoflagellates". In: *Marine Ecology Progress Series* 247 (2003), pp. 17–25. DOI: 10.3354/meps247017.

[91] F. D'Ortenzio and M. Ribera d'Alcalà. "On the Trophic Regimes of the Mediterranean Sea: A Satellite Analysis". In: *Biogeosciences* 6.2 (2009), pp. 139–148. DOI: 10.5194/bg-6-139-2009.

[92] F. d'Ovidio, S. D. Monte, A. D. Penna, C. Cotté, and C. Guinet. "Ecological Implications of Eddy Retention in the Open Ocean: A Lagrangian Approach". In: *Journal of Physics A: Mathematical and Theoretical* 46.25 (2013), p. 254023. DOI: 10.1088/1751-8113/46/25/254023.

[93] J. Dai, R. Wang, H. Zheng, G. Ji, and X. Qiao. "ZooplanktoNet: Deep Convolutional Network for Zooplankton Classification". In: *OCEANS 2016 - Shanghai*. OCEANS 2016 - Shanghai. 2016, pp. 1–6. DOI: 10.1109/OCEANSAP.2016.7485680.

[94] J. Dai, K. He, and J. Sun. "Instance-Aware Semantic Segmentation via Multi-Task Network Cascades". In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016, pp. 3150–3158.

[95] C. S. Davis, F. T. Thwaites, S. M. Gallager, and Q. Hu. "A Three-Axis Fast-Tow Digital Video Plankton Recorder for Rapid Surveys of Plankton Taxa and Hydrography". In: *Limnology and Oceanography: Methods* 3.2 (2005), pp. 59–74. DOI: 10.4319/lom.2005.3.59.

[96] C. de Boyer Montégut, G. Madec, A. S. Fischer, A. Lazar, and D. Iudicone. "Mixed Layer Depth over the Global Ocean: An Examination of Profile Data and a Profile-Based Climatology". In: *Journal of Geophysical Research* 109 (2004), p. C12003. DOI: 10.1029/2004JC002378.

[97] J. Decelle, S. Colin, and R. A. Foster. "Photosymbiosis in Marine Planktonic Protists". In: *Marine Protists: Diversity and Dynamics*. Ed. by S. Ohtsuka, T. Suzaki, T. Horiguchi, N. Suzuki, and F. Not. Tokyo: Springer Japan, 2015, pp. 465–500. ISBN: 978-4-431-55130-0. DOI: 10.1007/978-4-431-55130-0_19.

[98] J. Decelle, P. Martin, K. Paborstava, D. W. Pond, G. Tarling, F. Mahé, C. de Vargas, R. Lampitt, and F. Not. "Diversity, Ecology and Biogeochemistry of Cyst-Forming Acantharia (Radiolaria) in the Oceans". In: *PLOS ONE* 8.1 (2013), e53598. DOI: 10.1371/journal.pone.0053598.

[99] J. Decelle and F. Not. "Acantharia". In: *eLS*. John Wiley & Sons, Ltd, 2015, pp. 1–10. ISBN: 978-0-470-01590-2. DOI: 10.1002/9780470015902.a0002102.pub2.

[100] J. Decelle, I. Probert, L. Bittner, Y. Desdevises, S. Colin, C. de Vargas, M. Galí, R. Simó, and F. Not. "An Original Mode of Symbiosis in Open Ocean Plankton". In: *Proceedings of the National Academy of Sciences* 109.44 (2012), pp. 18000–18005.

[101] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. "ImageNet: A Large-Scale Hierarchical Image Database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009 IEEE Conference on Computer Vision and Pattern Recognition. 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[102] L. Deng. "The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]". In: *IEEE Signal Processing Magazine* 29.6 (2012), pp. 141–142. DOI: 10.1109/MSP.2012.2211477.

[103] K. L. Denman and T. M. Powell. "Effects of Physical Processes on Planktonic Ecosystems in the Coastal Ocean". In: *Oceanography and Marine Biology* 22 (1984), pp. 125–168.

[104] M. R. Dennett, D. A. Caron, A. F. Michaels, S. M. Gallager, and C. S. Davis. "Video Plankton Recorder Reveals High Abundances of Colonial Radiolaria in Surface Waters of the Central North Pacific". In: *Journal of Plankton Research* 24.8 (2002), pp. 797–805. DOI: 10.1093/plankt/24.8.797.

[105] C. de Vargas et al. "Eukaryotic Plankton Diversity in the Sunlit Ocean". In: *Science* 348.6237 (2015), p. 1261605. DOI: 10.1126/SCIENCE.1261605.

[106] T. D. Dickey. "The Emergence of Concurrent High-Resolution Physical and Bio-Optical Measurements in the Upper Ocean and Their Applications". In: *Reviews of Geophysics* 29.3 (1991), pp. 383–413. DOI: 10.1029/91RG00578.

[107] S. Dieleman, J. De Fauw, and K. Kavukcuoglu. "Exploiting Cyclic Symmetry in Convolutional Neural Networks". 2016.

[108] T. Dietterich. "Overfitting and Undercomputing in Machine Learning". In: *ACM computing surveys (CSUR)* 27.3 (1995), pp. 326–327.

[109] E. M. Ditria, S. Lopez-Marcano, M. Sievers, E. L. Jinks, C. J. Brown, and R. M. Connolly. "Automating the Analysis of Fish Abundance Using Object Detection: Optimizing Animal Ecology With Deep Learning". In: *Frontiers in Marine Science* 7 (2020).

[110] J. Dolan and V. Raybaud. "Zooplankton I. Micro- and Mesozooplankton". In: *The Mediterranean Sea in the Era of Global Change 2*. John Wiley & Sons, Ltd, 2020, pp. 67–107. ISBN: 978-1-119-70478-2. DOI: 10.1002/9781119704782.ch3.

[111] L. Drago et al. “Global Distribution of Zooplankton Biomass Estimated by In Situ Imaging and Machine Learning”. In: *Frontiers in Marine Science* 9 (2022). DOI: 10.3389/fmars.2022.894372.

[112] W. M. Durham and R. Stocker. “Thin Phytoplankton Layers: Characteristics, Mechanisms, and Consequences”. In: *Annual Review of Marine Science* 4.1 (2012), pp. 177–207. DOI: 10.1146/annurev-marine-120710-100957.

[113] C. Eggert, S. Brehm, A. Winschel, D. Zecha, and R. Lienhart. “A Closer Look: Small Object Detection in Faster R-CNN”. In: *2017 IEEE International Conference on Multimedia and Expo (ICME)*. 2017 IEEE International Conference on Multimedia and Expo (ICME). 2017, pp. 421–426. DOI: 10.1109/ICME.2017.8019550.

[114] C. Eggert, A. Winschel, D. Zecha, and R. Lienhart. “Saliency-Guided Selective Magnification for Company Logo Detection”. In: *2016 23rd International Conference on Pattern Recognition (ICPR)*. 2016 23rd International Conference on Pattern Recognition (ICPR). 2016, pp. 651–656. DOI: 10.1109/ICPR.2016.7899708.

[115] A. Elineau et al. *ZooScanNet: Plankton Images Captured with the ZooScan*. SEANOE, 2018. DOI: 10.17882/55741.

[116] J. Elith and J. R. Leathwick. “Species Distribution Models: Ecological Explanation and Prediction Across Space and Time”. In: *Annual Review of Ecology, Evolution, and Systematics* 40.1 (2009), pp. 677–697. DOI: 10.1146/annurev.ecolsys.110308.120159.

[117] J. Ellen, H. Li, and M. D. Ohman. “Quantifying California Current Plankton Samples with Efficient Machine Learning Techniques”. In: *OCEANS 2015 - MTS/IEEE Washington*. OCEANS 2015 - MTS/IEEE Washington. 2015, pp. 1–9. DOI: 10.23919/OCEANS.2015.7404607.

[118] J. S. Ellen, C. A. Graff, and M. D. Ohman. “Improving Plankton Image Classification Using Context Metadata”. In: *Limnology and Oceanography: Methods* 17.8 (2019), pp. 439–461. DOI: 10.1002/lom3.10324.

[119] K. V. Embleton, C. E. Gibson, and S. I. Heaney. “Automated Counting of Phytoplankton by Pattern Recognition: A Comparison with a Manual Counting Method”. In: *Journal of Plankton Research* 25.6 (2003), pp. 669–681. DOI: 10.1093/plankt/25.6.669.

[120] S. Emerson, P. Quay, D. Karl, C. Winn, L. Tupas, and M. Landry. "Experimental Determination of the Organic Carbon Flux from Open-Ocean Surface Waters". In: *Nature* 389.6654 (1997), pp. 951–954. DOI: 10.1038/40111.

[121] A. Engel, R. Kiko, and M. Dengler. "Organic Matter Supply and Utilization in Oxygen Minimum Zones". In: *Annual Review of Marine Science* 14.1 (2022), pp. 355–378.

[122] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. "The Pascal Visual Object Classes (VOC) Challenge". In: *International Journal of Computer Vision* 88.2 (2010), pp. 303–338. DOI: 10.1007/s11263-009-0275-4.

[123] R. Faillettaz. "Estimation Des Capacités Comportementales Des Larves de Poissons et Leurs Implications Pour La Phase Larvaire : Un Cas d'étude d'espèces Démersales de Méditerranée Nord-Occidentale". PhD thesis. UPMC, Ecole doctorale Sciences de l'Environnement d'Île-de-France (ED129), Paris. Encadrants: P Koubbi, JO Irisson., 2015.

[124] R. Faillettaz, M. Picheral, J. Y. Luo, C. Guigand, R. K. Cowen, and J.-O. Irisson. "Imperfect Automatic Image Classification Successfully Describes Plankton Distribution Patterns". In: *Methods in Oceanography* 15–16 (2016), pp. 60–77. DOI: 10.1016/J.MIO.2016.04.003.

[125] P. Falkowski. "Ocean Science: The Power of Plankton". In: *Nature* 483.7387 (2012), S17–S20. DOI: 10.1038/483S17a.

[126] E. Faure, F. Not, A.-S. Benoiston, K. Labadie, L. Bittner, and S.-D. Ayata. "Mixotrophic Protists Display Contrasted Biogeographies in the Global Ocean". In: *The ISME Journal* 13.4 (4 2019), pp. 1072–1083. DOI: 10.1038/s41396-018-0340-5.

[127] J. Febvre. "The Myoneme of the Acantharia (Protozoa): A New Model of Cellular Motility". In: *Biosystems* 14.3 (1981), pp. 327–336. DOI: 10.1016/0303-2647(81)90039-3.

[128] C. Febvre-Chevalier and J. Febvre. "Buoyancy and Swimming in Marine Planktonic Protists". In: *Mechanics and Physiology of Animal Swimming* (1994), pp. 13–26.

[129] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim. "Do We Need Hundreds of Classifiers to Solve Real World Classification Problems?" In: *The journal of machine learning research* 15.1 (2014), pp. 3133–3181.

[130] R. Ferrari. "A Frontal Challenge for Climate Models". In: *Science* 332.6027 (2011), pp. 316–317. DOI: 10.1126/science.1203632.

[131] C. B. Field, M. J. Behrenfeld, J. T. Randerson, and P. Falkowski. "Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components". In: *Science* 281.5374 (1998), pp. 237–240. DOI: 10.1126/science.281.5374.237.

[132] J. S. Font-Muñoz, R. Jeanneret, J. Arrieta, S. Anglès, A. Jordi, I. Tuval, and G. Basterretxea. "Collective Sinking Promotes Selective Cell Pairing in Planktonic Pennate Diatoms". In: *Proceedings of the National Academy of Sciences* 116.32 (2019), pp. 15997–16002. DOI: 10.1073/pnas.1904837116.

[133] A. Forest, L. Stemmann, M. Picheral, L. Burdorf, D. Robert, L. Fortier, and M. Babin. "Size Distribution of Particles and Zooplankton across the Shelf-Basin System in Southeast Beaufort Sea: Combined Results from an Underwater Vision Profiler and Vertical Net Tows". In: *Biogeosciences* 9.4 (2012), pp. 1301–1320. DOI: 10.5194/bg-9-1301-2012.

[134] A. Forest, M. Sampei, R. Makabe, H. Sasaki, D. G. Barber, Y. Gratton, P. Wassmann, and L. Fortier. "The Annual Cycle of Particulate Organic Carbon Export in Franklin Bay (Canadian Arctic): Environmental Control and Food Web Implications". In: *Journal of Geophysical Research: Oceans* 113.C3 (2008). DOI: 10.1029/2007JC004262.

[135] P. J. S. Franks and J. S. Jaffe. "Microscale Distributions of Phytoplankton: Initial Results from a Two-Dimensional Imaging Fluorometer, OSST". In: *Marine Ecology Progress Series* 220 (2001), pp. 59–72. DOI: 10.3354/meps220059.

[136] M. Frederiksen, M. Edwards, A. J. Richardson, N. C. Halliday, and S. Wanless. "From Plankton to Top Predators: Bottom-up Control of a Marine Food Web across Four Trophic Levels". In: *Journal of Animal Ecology* 75.6 (2006), pp. 1259–1268. DOI: 10.1111/j.1365-2656.2006.01148.x.

[137] J. A. Fuhrman, J. A. Steele, I. Hewson, M. S. Schwalbach, M. V. Brown, J. L. Green, and J. H. Brown. "A Latitudinal Diversity Gradient in Planktonic Marine Bacteria". In: *Proceedings of the National Academy of Sciences* 105.22 (2008), pp. 7774–7778. DOI: 10.1073/pnas.0803070105.

[138] C. Garrett and W. Munk. "Space-Time Scales of Internal Waves: A Progress Report". In: *Journal of Geophysical Research (1896-1977)* 80.3 (1975), pp. 291–297. DOI: 10.1029/JC080i003p00291.

[139] J. C. Garwood, R. C. Musgrave, and A. J. Lucas. "Life in Internal Waves". In: *Oceanography* 33.3 (2020), pp. 38–49.

[140] D. E. Gaskell, M. D. Ohman, and P. M. Hull. "Zooglider-Based Measurements of Planktonic Foraminifera in the California Current System". In: *Journal of Foraminiferal Research* 49.4 (2019), pp. 390–404. DOI: 10.2113/gsjfr.49.4.390.

[141] B. Gasser, G. Payet, J. Sardou, and P. Nival. "Community Structure of Mesopelagic Copepods (>500 µm) in the Ligurian Sea (Western Mediterranean)". In: *Journal of Marine Systems* 15.1 (1998), pp. 511–522. DOI: 10.1016/S0924-7963(97)00094-8.

[142] S. L. C. Giering, B. Hosking, N. Briggs, and M. H. Iversen. "The Interpretation of Particle Size, Shape, and Carbon Flux of Marine Particle Images Is Strongly Affected by the Choice of Particle Detection Algorithm". In: *Frontiers in Marine Science* 7 (2020), p. 564. DOI: 10.3389/fmars.2020.00564.

[143] R. W. Gilmer and G. R. Harbison. "Morphology and Field Behavior of Pteropod Molluscs: Feeding Methods in the Families Cavoliniidae, Limacinidae and Peraclididae (Gastropoda: Thecosomata)". In: *Marine Biology* 91.1 (1986), pp. 47–57. DOI: 10.1007/BF00397570.

[144] A. Goffart, J.-H. Hecq, and L. Prieur. "Contrôle du phytoplancton du bassin ligure par le front liguro-provençal (secteur corse)". In: *Oceanologica Acta* 18 (1995).

[145] P. González, E. Álvarez, J. Díez, Á. López-Urrutia, and J. J. del Coz. "Validation Methods for Plankton Image Classification Systems". In: *Limnology and Oceanography: Methods* 15.3 (2017), pp. 221–237. DOI: 10.1002/lom3.10151.

[146] G. Gorsky, L. Prieur, I. Taupier-Letage, L. Stemmann, and M. Picheral. "Large Particulate Matter in the Western Mediterranean: I. LPM Distribution Related to Mesoscale Hydrodynamics". In: *Journal of Marine Systems*. MATER: MAss Transfer and Ecosystem Response 33–34 (2002), pp. 289–311. DOI: 10.1016/S0924-7963(02)00063-5.

[147] G. Gorsky, N. L. da Silva, S. Dallot, P. Laval, J. C. Braconnot, and L. Prieur. "Midwater Tunicates: Are They Related to the Permanent Front of the Ligurian Sea (NW Mediterranean)?" In: *Marine Ecology Progress Series* 74.2/3 (1991), pp. 195–204.

[148] G. Gorsky, M. Picheral, and L. Stemmann. "Use of the Underwater Video Profiler for the Study of Aggregate Dynamics in the North Mediterranean". In: *Estuarine, Coastal and Shelf Science*. Visualization in Marine Science 50.1 (2000), pp. 121–128. DOI: 10.1006/ecss.1999.0539.

[149] G. Gorsky, M. D. Ohman, M. Picheral, S. Gasparini, L. Stemmann, J.-B. Romagnan, A. Cawood, S. Pesant, C. Garcia-Comas, and F. Prejger. "Digital Zooplankton Image Analysis Using the ZooScan Integrated System". In: *Journal of Plankton Research* 32.3 (2010), pp. 285–303. DOI: 10.1093/plankt/fbp124.

[150] W. M. Graham, F. Pagès, and W. M. Hamner. "A Physical Context for Gelatinous Zooplankton Aggregations: A Review". In: *Hydrobiologia* 451.1 (2001), pp. 199–212. DOI: 10.1023/A:1011876004427.

[151] A. T. Greer, L. M. Chiaverano, J. Y. Luo, R. K. Cowen, and W. M. Graham. "Ecology and Behaviour of Holoplanktonic Scyphomedusae and Their Interactions with Larval and Juvenile Fishes in the Northern Gulf of Mexico". In: *ICES Journal of Marine Science* 75.2 (2018), pp. 751–763. DOI: 10.1093/icesjms/fsx168.

[152] A. T. Greer, L. M. Chiaverano, L. M. Treible, C. Briseño-Avena, and F. J. Hernandez. "From Spatial Pattern to Ecological Process through Imaging Zooplankton Interactions". In: *ICES Journal of Marine Science* 78.8 (2021), pp. 2664–2674. DOI: 10.1093/icesjms/fsab149.

[153] A. T. Greer, A. D. Boyette, V. J. Cruz, M. K. Cambazoglu, B. Dzwonkowski, L. M. Chiaverano, S. L. Dykstra, C. Briseño-Avena, R. K. Cowen, and J. D. Wiggert. "Contrasting Fine-Scale Distributional Patterns of Zooplankton Driven by the Formation of a Diatom-Dominated Thin Layer". In: *Limnology and Oceanography* 65.9 (2020), pp. 2236–2258. DOI: 10.1002/lno.11450.

[154] A. T. Greer, R. K. Cowen, C. M. Guigand, and J. A. Hare. "Fine-Scale Planktonic Habitat Partitioning at a Shelf-Slope Front Revealed by a High-Resolution Imaging System". In: *Journal of Marine Systems* 142 (2015), pp. 111–125. DOI: 10.1016/j.jmarsys.2014.10.008.

[155] A. T. Greer, R. K. Cowen, C. M. Guigand, J. A. Hare, and D. Tang. "The Role of Internal Waves in Larval Fish Interactions with Potential Predators and Prey". In: *Progress in Oceanography* 127 (2014), pp. 47–61. DOI: 10.1016/j.pocean.2014.05.010.

[156] A. T. Greer, R. K. Cowen, C. M. Guigand, M. A. McManus, J. C. Sevadjian, and A. H. V. Timmerman. "Relationships between Phytoplankton Thin Layers and the Fine-Scale Vertical Distributions of Two Trophic Levels of Zooplankton". In: *Journal of Plankton Research* 35.5 (2013), pp. 939–956. DOI: 10.1093/plankt/fbt056.

[157] A. T. Greer, C. B. Woodson, C. M. Guigand, and R. K. Cowen. "Larval Fishes Utilize Batesian Mimicry as a Survival Strategy in the Plankton". In: *Marine Ecology Progress Series* 551 (2016), pp. 1–12. DOI: 10.3354/meps11751.

[158] A. T. Greer et al. "High-Resolution Sampling of a Broad Marine Life Size Spectrum Reveals Differing Size- and Composition-Based Associations With Physical Oceanographic Structure". In: *Frontiers in Marine Science* 7 (2020), p. 1125. DOI: 10.3389/fmars.2020.542701.

[159] A. T. Greer et al. "In Situ Imaging across Ecosystems to Resolve the Fine-Scale Oceanographic Drivers of a Globally Significant Planktonic Grazer". In: *Limnology and Oceanography* 68.1 (2023), pp. 192–207. DOI: 10.1002/lno.12259.

[160] P. Grosjean, M. Picheral, C. Warembourg, and G. Gorsky. "Enumeration, Measurement, and Identification of Net Zooplankton Samples Using the ZOOSCAN Digital Imaging System". In:

ICES Journal of Marine Science 61.4 (2004), pp. 518–525. DOI: 10.1016/j.icesjms.2004.03.012.

[161] L. Guidi, G. A. Jackson, L. Stemmann, J. C. Miquel, M. Picheral, and G. Gorsky. "Relationship between Particle Size Distribution and Flux in the Mesopelagic Zone". In: *Deep Sea Research Part I: Oceanographic Research Papers* 55.10 (2008), pp. 1364–1374. DOI: 10.1016/J.DSR.2008.05.014.

[162] L. Guidi, L. Legendre, G. Reygondeau, J. Uitz, L. Stemmann, and S. A. Henson. "A New Look at Ocean Carbon Remineralization for Estimating Deepwater Sequestration". In: *Global Biogeochemical Cycles* 29.7 (2015), pp. 1044–1059. DOI: 10.1002/2014GB005063.

[163] L. Guidi, L. Stemmann, G. A. Jackson, F. Ibanez, H. Claustre, L. Legendre, M. Picheral, and G. Gorskya. "Effects of Phytoplankton Community on Production, Size, and Export of Large Aggregates: A World-Ocean Analysis". In: *Limnology and Oceanography* 54.6 (2009), pp. 1951–1963. DOI: 10.4319/lo.2009.54.6.1951.

[164] L. Guidi et al. "Does Eddy-Eddy Interaction Control Surface Phytoplankton Distribution and Carbon Export in the North Pacific Subtropical Gyre?" In: *Journal of Geophysical Research: Biogeosciences* 117.G2 (2012). DOI: 10.1029/2012JG001984.

[165] L. Guidi et al. "Plankton Networks Driving Carbon Export in the Oligotrophic Ocean". In: *Nature* 532.7600 (7600 2016), pp. 465–470. DOI: 10.1038/nature16942.

[166] E. Haeckel. "Report on the Radiolaria Collected by HMS Challenger during the Years 1873-1876". In: *Report of the Voyage of HMS Challenger, Zoology* 18 (1887), pp. i–clxxxviii, 1–1803.

[167] S. Hampton, C. Strasser, J. Tewksbury, W. Gram, A. Budden, A. Batcheller, C. Duke, and J. Porter. "Big Data and the Future of Ecology". In: *Frontiers in Ecology and the Environment* 11.3 (2013), pp. 156–162.

[168] W. Harder. "Reactions of Plankton Organisms to Water Stratification". In: *Limnology and Oceanography* 13.1 (1968), pp. 156–168. DOI: 10.4319/lo.1968.13.1.0156.

[169] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media, 2009. ISBN: 0-387-84858-4.

[170] L. R. Haury, J. A. McGowan, and P. H. Wiebe. "Patterns and Processes in the Time-Space Scales of Plankton Distributions". In: *Spatial Pattern in Plankton Communities*. Ed. by J. H. Steele. NATO Conference Series. Boston, MA: Springer US, 1978, pp. 277–327. ISBN: 978-1-4899-2195-6. DOI: 10.1007/978-1-4899-2195-6_12.

[171] S. P. Hawser, J. M. O'Neil, M. R. Roman, and G. A. Codd. "Toxicity of Blooms of the Cyanobacterium Trichodesmium to Zooplankton". In: *Journal of Applied Phycology* 4.1 (1992), pp. 79–86. DOI: 10.1007/BF00003963.

[172] G. C. Hays, A. J. Richardson, and C. Robinson. "Climate Change and Marine Plankton". In: *Trends in Ecology & Evolution* 20.6 (2005), pp. 337–344. DOI: 10.1016/J.TREE.2005.03.004.

[173] K. He, G. Gkioxari, P. Dollar, and R. Girshick. "Mask R-CNN". In: Proceedings of the IEEE International Conference on Computer Vision. 2017, pp. 2961–2969.

[174] A. K. Heinrich. "The Life Histories of Plankton Animals and Seasonal Cycles of Plankton Communities in the Oceans". In: *J. Cons. Int. Explor. Mer* 27.1 (1962), pp. 15–24.

[175] A. Herbland and B. Voituriez. "Hydrological Structure Analysis for Estimating the Primary Production in the Tropical Atlantic Ocean". In: *J. mar. Res* 37.1 (1979), pp. 87–101.

[176] A. W. Herman. "Vertical Copepod Aggregations and Interactions with Chlorophyll and Production on the Peru Shelf". In: *Continental Shelf Research* 3.2 (1984), pp. 131–146. DOI: 10.1016/0278-4343(84)90003-7.

[177] A. W. Herman and M. R. Mitchell. "Counting and Identifying Copepod Species with an in Situ Electronic Zooplankton Counter". In: *Deep Sea Research Part A. Oceanographic Research Papers* 28.7 (1981), pp. 739–755. DOI: 10.1016/0198-0149(81)90133-3.

[178] S. Hernández-León et al. "Large Deep-Sea Zooplankton Biomass Mirrors Primary Production in the Global Ocean". In: *Nature Communications* 11.1 (1 2020), p. 6048. DOI: 10.1038/s41467-020-19875-7.

[179] T. Hey, S. Tansley, and K. M. Tolle. *Jim Gray on eScience: A Transformed Scientific Method*. 2009.

[180] U. Hofmann Elizondo, D. Righetti, F. Benedetti, and M. Vogt. "Biome Partitioning of the Global Ocean Based on Phytoplankton Biogeography". In: *Progress in Oceanography* 194 (2021), p. 102530. DOI: 10.1016/j.pocean.2021.102530.

[181] A. C. Hollande and M. Cachon-Enjumet. "Contribution à l'étude Biologique Des Spaerocollides:(Radiolaires Collodaires et Radiolaires Polycyttaires) et de Leurs Parasites. Thalassicollidae, Physematidae, Thalassophysidae". In: . Annales des sciences naturelles, 1953.

[182] C. R. Horne, A. G. Hirst, D. Atkinson, A. Neves, and T. Kiørboe. "A Global Synthesis of Seasonal Temperature–Size Responses in Copepods". In: *Global Ecology and Biogeography* 25.8 (2016), pp. 988–999. DOI: 10.1111/geb.12460.

[183] C. Hörstmann, P. L. Buttigieg, U. John, E. J. Raes, D. Wolf-Gladrow, A. Bracher, and A. M. Waite. "Microbial Diversity through an Oceanographic Lens: Refining the Concept of Ocean Provinces through Trophic-Level Analysis and Productivity-Specific Length Scales". In: *Environmental Microbiology* 24.1 (2022), pp. 404–419. DOI: 10.1111/1462-2920.15832.

[184] H. J. T. Hoving, P. Neitzel, H. Hauss, S. Christiansen, R. Kiko, B. H. Robison, P. Silva, and A. Körtzinger. "In Situ Observations Show Vertical Community Structure of Pelagic Fauna in the Eastern Tropical North Atlantic off Cape Verde". In: *Scientific Reports* 10.1 (1 2020), p. 21798. DOI: 10.1038/s41598-020-78255-9.

[185] Q. Hu and C. Davis. "Automatic Plankton Image Recognition with Co-Occurrence Matrices and Support Vector Machine". In: *Marine Ecology Progress Series* 295 (2005), pp. 21–31. DOI: 10.3354/meps295021.

[186] G. E. Hutchinson. "The Paradox of the Plankton". In: *The American Naturalist* 95.882 (1961), pp. 137–145.

[187] T. H. Huxley. "Upon Thalassicolla, a New Zoophyte. Zoological Notes and Observations Made on Board HMS Rattlesnake". In: *Ann Mag Nat Hist: Series* 2.8 (1851), pp. 433–442.

[188] F. M. Ibarbalz et al. "Global Trends in Marine Plankton Diversity across Kingdoms of Life". In: *Cell* 179.5 (2019), 1084–1097.e21. DOI: 10.1016/j.cell.2019.10.008.

[189] ICES. *Working Group on Machine Learning in Marine Science (WGMLEARN; Outputs from 2021 Meeting)*. report. ICES Scientific Reports, 2022. DOI: 10.17895/ices.pub.10060.

[190] T. Ikeda. "Metabolic Rates of Epipelagic Marine Zooplankton as a Function of Body Mass and Temperature". In: *Marine Biology* 85.1 (1985), pp. 1–11. DOI: 10.1007/BF00396409.

[191] T. Ikenoue, K. Kimoto, Y. Okazaki, M. Sato, M. C. Honda, K. Takahashi, N. Harada, and T. Fujiki. "Phaeodaria: An Important Carrier of Particulate Organic Carbon in the Mesopelagic Twilight Zone of the North Pacific Ocean". In: *Global Biogeochemical Cycles* 33.8 (2019), pp. 1146–1160. DOI: 10.1029/2019GB006258.

[192] J.-O. Irisson, S.-D. Ayata, D. J. Lindsay, L. Karp-Boss, and L. Stemmann. "Machine Learning for the Study of Plankton and Marine Snow from Images". In: *Annual Review of Marine Science* 14.1 (2022), pp. 277–301. DOI: 10.1146/annurev-marine-041921-013023.

[193] A. G. Ivakhnenko. "Polynomial Theory of Complex Systems". In: *IEEE Transactions on Systems, Man, and Cybernetics* SMC-1.4 (1971), pp. 364–378. DOI: 10.1109/TSMC.1971.4308320.

[194] N. Iyer. *Machine Vision Assisted in Situ Ichthyoplankton Imaging System*. Purdue University, 2012. ISBN: 1-369-11309-9.

[195] G. B. Jeffery. "The Motion of Ellipsoidal Particles Immersed in a Viscous Fluid". In: *Proceedings of the Royal Society of London. Series A, Containing papers of a mathematical and physical character* 102.715 (1922), pp. 161–179.

[196] D. Jon Furbish and A. J. Arnold. "Hydrodynamic Strategies in the Morphological Evolution of Spinose Planktonic Foraminifera". In: *GSA Bulletin* 109.8 (1997), pp. 1055–1072. DOI: 10.1130/0016-7606(1997)109<1055:HSITME>2.3.CO;2.

[197] J. Kämpf and P. Chapman. *Upwelling Systems of the World*. Cham: Springer International Publishing, 2016. ISBN: 978-3-319-42522-1 978-3-319-42524-5. DOI: 10.1007/978-3-319-42524-5.

[198] L. Karp-Boss, E. Boss, and P. A. Jumars. "Motion of Dinoflagellates in a Simple Shear Flow". In: *Limnology and Oceanography* 45.7 (2000), pp. 1594–1602. DOI: 10.4319/lo.2000.45.7.1594.

[199] K. Katija, P. L. D. Roberts, J. Daniels, A. Lapides, K. Barnard, M. Risi, B. Y. Ranaan, B. G. Woodward, and J. Takahashi. "Visual Tracking of Deepwater Animals Using Machine Learning-Controlled Robotic Underwater Vehicles". In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2021, pp. 860–869.

[200] J. D. Kelleher, B. Mac Namee, and A. D'arcy. *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*. MIT press, 2020. ISBN: 0-262-36110-8.

[201] S. R. Kerr and L. M. Dickie. *The Biomass Spectrum: A Predator-Prey Theory of Aquatic Production*. Columbia University Press, 2001. ISBN: 0-231-50734-8.

[202] T. Kerr, J. R. Clark, E. S. Fileman, C. E. Widdicombe, and N. Pugeault. "Collaborative Deep Learning Models to Handle Class Imbalance in FlowCam Plankton Imagery". In: *IEEE Access* 8 (2020), pp. 170013–170032. DOI: 10.1109/ACCESS.2020.3022242.

[203] R. Kiko, P. Brandt, S. Christiansen, J. Faustmann, I. Kriest, E. Rodrigues, F. Schütte, and H. Hauss. "Zooplankton-Mediated Fluxes in the Eastern Tropical North Atlantic". In: *Frontiers in Marine Science* 7 (2020).

[204] R. Kiko and H. Hauss. "On the Estimation of Zooplankton-Mediated Active Fluxes in Oxygen Minimum Zone Regions". In: *Frontiers in Marine Science* 6 (2019).

[205] R. Kiko et al. "A Global Marine Particle Size Distribution Dataset Obtained with the Underwater Vision Profiler 5". In: *Earth System Science Data Discussions* (2022), pp. 1–37. DOI: 10.5194/essd-2022-51.

[206] K. Kimoto. "Planktic Foraminifera". In: *Marine Protists*. Springer, 2015, pp. 129–178.

[207] T. Kiørboe. "Formation and Fate of Marine Snow: Small-Scale Processes with Large- Scale Implications". In: *Scientia Marina* 65.S2 (S2 2001), pp. 57–71. DOI: 10.3989/scimar.2001.65s257.

[208] R. Kitchin. "Big Data, New Epistemologies and Paradigm Shifts". In: *Big Data & Society* 1.1 (2014), p. 2053951714528481. DOI: 10.1177/2053951714528481.

[209] P. Klein, G. Lapeyre, L. Siegelman, B. Qiu, L.-L. Fu, H. Torres, Z. Su, D. Menemenlis, and S. Le Gentil. "Ocean-Scale Interactions From Space". In: *Earth and Space Science* 6.5 (2019), pp. 795–817. DOI: 10.1029/2018EA000492.

[210] S. A. Kling and D. Boltovskoy. "Radiolaria Phaeodaria". In: *South Atlantic Zooplankton* 1 (1999), pp. 231–264.

[211] M. Koski and F. Lombard. "Functional Responses of Aggregate-Colonizing Copepods". In: *Limnology and Oceanography* n/a.n/a (). DOI: 10.1002/lno.12187.

[212] B. Krawczyk. "Learning from Imbalanced Data: Open Challenges and Future Directions". In: *Progress in Artificial Intelligence* 5.4 (2016), pp. 221–232. DOI: 10.1007/s13748-016-0094-0.

[213] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems 25*. Ed. by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger. Curran Associates, Inc., 2012, pp. 1097–1105.

[214] S. P. Kyathanahally, T. Hardeman, E. Merz, T. Bulas, M. Reyes, P. Isles, F. Pomati, and M. Baity-Jesi. "Deep Learning Classification of Lake Zooplankton". In: *Frontiers in Microbiology* 12 (2021).

[215] C. M. Lalli and T. R. Parsons. "Phytoplankton and Primary Production". In: *Biological Oceanography: An Introduction*. Elsevier, 1997, pp. 39–73.

[216] W. Lampert. "The Adaptive Significance of Diel Vertical Migration of Zooplankton". In: *Functional Ecology* 3.1 (1989), pp. 21–27. DOI: 10.2307/2389671.

[217] R. S. Lampitt, I. Salter, and D. Johns. "Radiolaria: Major Exporters of Organic Carbon to the Deep Ocean". In: *Global Biogeochemical Cycles* 23.1 (2009), n/a–n/a. DOI: 10.1029/2008GB003221.

[218] M. Latasa, A. M. Cabello, X. A. G. Morán, R. Massana, and R. Scharek. "Distribution of Phytoplankton Groups within the Deep Chlorophyll Maximum". In: *Limnology and Oceanography* 62.2 (2017), pp. 665–685. DOI: 10.1002/lno.10452.

[219] P. Laval, J.-C. Braconnot, C. Carré, J. Goy, P. Morand, and C. Mills. "Small-Scale Distribution of Macroplankton and Micronekton in the Ligurian Sea (Mediterranean Sea) as Observed from the Manned Submersible Cyana". In: *Journal of Plankton Research* 11.4 (1989), pp. 665–685. DOI: 10.1093/plankt/11.4.665.

[220] Y. Le Cun, L. Jackel, B. Boser, J. Denker, H. Graf, I. Guyon, D. Henderson, R. Howard, and W. Hubbard. "Handwritten Digit Recognition: Applications of Neural Network Chips and Automatic Learning". In: *IEEE Communications Magazine* 27.11 (1989), pp. 41–46. DOI: 10.1109/35.41400.

[221] J. R. Leathwick, J. Elith, M. P. Francis, T. Hastie, and P. Taylor. "Variation in Demersal Fish Species Richness in the Oceans Surrounding New Zealand: An Analysis Using Boosted Regression Trees". In: *Marine Ecology Progress Series* 321 (2006), pp. 267–281. DOI: 10.3354/meps321267.

[222] Y. LeCun, Y. Bengio, and G. Hinton. "Deep Learning". In: *Nature* 521.7553 (7553 2015), pp. 436–444. DOI: 10.1038/nature14539.

[223] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel. "Handwritten Digit Recognition with a Back-Propagation Network". In: *Advances in neural information processing systems* 2 (1990), pp. 396–404.

[224] H. Lee, M. Park, and J. Kim. "Plankton Classification on Imbalanced Large Scale Database via Convolutional Neural Networks with Transfer Learning". In: *2016 IEEE International Conference on Image Processing (ICIP)*. 2016 IEEE International Conference on Image Processing (ICIP). 2016, pp. 3713–3717. DOI: 10.1109/ICIP.2016.7533053.

[225] L. Legendre and S. Demers. "Towards Dynamic Biological Oceanography and Limnology". In: *Canadian Journal of Fisheries and Aquatic Sciences* 41.1 (1984), pp. 2–19. DOI: 10.1139/f84-001.

[226] P. Legendre and L. Legendre. *Numerical Ecology*. Elsevier, 2012. 990 pp. ISBN: 0-444-53869-0.

[227] S. G. Leles et al. "Oceanic Protists with Different Forms of Acquired Phototrophy Display Contrasting Biogeographies and Abundance". In: *Proceedings of the Royal Society B: Biological Sciences* 284.1860 (2017), p. 20170664. DOI: 10.1098/rspb.2017.0664.

[228] M. Levy, P. Klein, and A.-M. Treguier. "Impact of Sub-Mesoscale Physics on Production and Subduction of Phytoplankton in an Oligotrophic Regime". In: *Journal of Marine Research* 59.4 (2001), pp. 535–565.

[229] M. Lévy, R. Ferrari, P. J. S. Franks, A. P. Martin, and P. Rivière. "Bringing Physics to Life at the Submesoscale". In: *Geophysical Research Letters* 39.14 (2012). DOI: 10.1029/2012GL052756.

[230] M. Lévy, P. J. S. Franks, and K. S. Smith. "The Role of Submesoscale Currents in Structuring Marine Ecosystems". In: *Nature Communications* 9.1 (1 2018), p. 4758. DOI: 10.1038/s41467-018-07059-3.

[231] C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata. "Detecting Outliers: Do Not Use Standard Deviation around the Mean, Use Absolute Deviation around the Median". In: *Journal of Experimental Social Psychology* 49.4 (2013), pp. 764–766. DOI: 10.1016/j.jesp.2013.03.013.

[232] Q. Li, X. Sun, J. Dong, S. Song, T. Zhang, D. Liu, H. Zhang, and S. Han. "Developing a Microscopic Image Dataset in Support of Intelligent Phytoplankton Detection Using Deep Learning". In: *ICES Journal of Marine Science* 77.4 (2020), pp. 1427–1439.

[233] W. K. W. Li, E. J. H. Head, and W. Glen Harrison. "Macroecological Limits of Heterotrophic Bacterial Abundance in the Ocean". In: *Deep Sea Research Part I: Oceanographic Research Papers* 51.11 (2004), pp. 1529–1540. DOI: 10.1016/j.dsr.2004.06.012.

[234] P. Licandro and P. Icardi. "Basin Scale Distribution of Zooplankton in the Ligurian Sea (North-Western Mediterranean) in Late Autumn". In: *Hydrobiologia* 617.1 (2009), pp. 17–40. DOI: 10.1007/s10750-008-9523-9.

[235] P. Licandro, F. Ibañez, and M. Etienne. "Long-Term Fluctuations (1974-99) of the Salps Thalia Democratica and Salpa Fusiformis in the Northwestern Mediterranean Sea: Relationships with Hydroclimatic Variability". In: *Limnology and Oceanography* 51.4 (2006), pp. 1832–1848. DOI: 10.4319/lo.2006.51.4.1832.

[236] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. "Feature Pyramid Networks for Object Detection". 2017.

[237] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. "Focal Loss for Dense Object Detection". 2018.

[238] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár. *Microsoft COCO: Common Objects in Context*. 2015. DOI: 10.48550/arXiv.1405.0312.

[239] H. Liu, J. Sticklus, K. Köser, H.-J. T. Hoving, H. Song, Y. Chen, J. Greinert, and T. Schoening. "TuLUMIS - a Tunable LED-based Underwater Multispectral Imaging System". In: *Optics Express* 26.6 (2018), pp. 7811–7828. DOI: 10.1364/OE.26.007811.

[240] N. Llopis Monferrer, D. Boltovskoy, P. Tréguer, M. M. Sandin, F. Not, and A. Leynaert. "Estimating Biogenic Silica Production of Rhizaria in the Global Ocean". In: *Global Biogeochemical Cycles* 34.3 (2020), e2019GB006286. DOI: 10.1029/2019GB006286.

[241] M. E. Lofton, T. H. Leach, B. E. Beisner, and C. C. Carey. "Relative Importance of Top-down vs. Bottom-up Control of Lake Phytoplankton Vertical Distributions Varies among Fluorescence-Based Spectral Groups". In: *Limnology and Oceanography* 65.10 (2020), pp. 2485–2501. DOI: 10.1002/lno.11465.

[242] F. Lombard and T. Kiørboe. "Marine Snow Originating from Appendicularian Houses: Age-dependent Settling Characteristics". In: *Deep Sea Research Part I: Oceanographic Research Papers* 57.10 (2010), pp. 1304–1313. DOI: 10.1016/j.dsr.2010.06.008.

[243] F. Lombard et al. "Globally Consistent Quantitative Observations of Planktonic Ecosystems". In: *Frontiers in Marine Science* 6 (2019). DOI: 10.3389/fmars.2019.00196.

[244] A. R. Longhurst. "Seasonal Cycles of Pelagic Production and Consumption". In: *Progress in Oceanography* 36.2 (1995), pp. 77–167. DOI: 10.1016/0079-6611(95)00015-1.

[245] A. R. Longhurst. *Ecological Geography of the Sea*. Academic Press, 2010. 542 pp. ISBN: 0-08-046557-9.

[246] A. R. Longhurst and W. Glen Harrison. "The Biological Pump: Profiles of Plankton Production and Consumption in the Upper Ocean". In: *Progress in Oceanography* 22.1 (1989), pp. 47–123. DOI: 10.1016/0079-6611(89)90010-4.

[247] A. Lumini and L. Nanni. "Deep Learning and Transfer Learning Features for Plankton Classification". In: *Ecological Informatics* 51 (2019), pp. 33–43. DOI: 10.1016/j.ecoinf.2019.02.007.

[248] J. Y. Luo, B. Grassian, D. Tang, J.-O. Irisson, A. T. Greer, C. M. Guigand, S. McClatchie, and R. K. Cowen. "Environmental Drivers of the Fine-Scale Distribution of a Gelatinous Zooplankton Community across a Mesoscale Front". In: *Marine Ecology Progress Series* 510 (2014), pp. 129–149. DOI: 10.3354/meps10908.

[249] J. Y. Luo, J.-O. Irisson, B. Graham, C. Guigand, A. Sarafraz, C. Mader, and R. K. Cowen. "Automated Plankton Image Analysis Using Convolutional Neural Networks". In: *Limnology and Oceanography: Methods* 16.12 (2018), pp. 814–827. DOI: 10.1002/lom3.10285.

[250] T. Luo, K. Kramer, S. Samson, A. Remsen, D. Goldgof, L. Hall, and T. Hopkins. "Active Learning to Recognize Multiple Types of Plankton". In: *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.* Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004. Cambridge, UK: IEEE, 2004, 478–481 Vol.3. ISBN: 978-0-7695-2128-2. DOI: 10.1109/ICPR.2004.1334570.

[251] D. L. Mackas, K. L. Denman, and M. R. Abbott. "Plankton Patchiness: Biology in the Physical Vernacular". In: *Bulletin of Marine Science* 37.2 (1985), pp. 652–674.

[252] A. Mahadevan. "The Impact of Submesoscale Physics on Primary Productivity of Plankton". In: *Annual Review of Marine Science* 8.1 (2016), pp. 161–184. DOI: 10.1146/annurev-marine-010814-015912.

[253] K. Malde, N. O. Handegard, L. Eikvil, and A.-B. Salberg. "Machine Intelligence and the Data-Driven Future of Marine Science". In: *ICES Journal of Marine Science* 77.4 (2020), pp. 1274–1285. DOI: 10.1093/icesjms/fsz057.

[254] K. Malde and H. Kim. "Beyond Image Classification: Zooplankton Identification with Deep Vector Space Embeddings". 2019.

[255] E. Malkiel, O. Alquaddoomi, and J. Katz. "Measurements of Plankton Distribution in the Ocean Using Submersible Holography". In: *Measurement Science and Technology* 10.12 (1999), p. 1142. DOI: 10.1088/0957-0233/10/12/305.

[256] K. H. Mann and J. R. Lazier. *Dynamics of Marine Ecosystems: Biological-Physical Interactions in the Oceans*. John Wiley & Sons, 2013. ISBN: 1-118-68791-4.

[257] D. Marrable, K. Barker, S. Tippaya, M. Wyatt, S. Bainbridge, M. Stowar, and J. Larke. "Accelerating Species Recognition and Labelling of Fish From Underwater Video With Machine-Assisted Deep Learning". In: *Frontiers in Marine Science* 9 (2022).

[258] M. Mars Brisbin, O. D. Brunner, M. M. Grossmann, and S. Mitarai. "Paired High-Throughput, in Situ Imaging and High-Throughput Sequencing Illuminate Acantharian Abundance and Vertical Distribution". In: *Limnology and Oceanography* 65.12 (2020), pp. 2953–2965. DOI: 10.1002/lno.11567.

[259] D. S. Marszalek. "The Role of Heavy Skeletons in Vertical Movements of Non-motile Zooplankton". In: *Marine Behaviour and Physiology* 8.4 (1982), pp. 295–303. DOI: 10.1080/10236248209387026.

[260] G. Massey and C. Friedrichs. "Laser In-situ Scattering and Transmissometer (LISST) Observations in Support of the Sensor Insertion System Duck, NC October, 1997". In: *Reports* (1998). DOI: 10.21220/v55m6d.

[261] J. Matas, O. Chum, M. Urban, and T. Pajdla. "Robust Wide-Baseline Stereo from Maximally Stable Extremal Regions". In: *Image and Vision Computing*. British Machine Vision Computing 2002 22.10 (2004), pp. 761–767. DOI: 10.1016/j.imavis.2004.02.006.

[262] N. Mayot, F. D'Ortenzio, V. Taillandier, L. Prieur, O. P. de Fommervault, H. Claustre, A. Bosse, P. Testor, and P. Conan. "Physical and Biogeochemical Controls of the Phytoplankton Blooms in North Western Mediterranean Sea: A Multiplatform Approach Over a Complete Annual Cycle (2012–2013 DEWEX Experiment)". In: *Journal of Geophysical Research: Oceans* 122.12 (2017), pp. 9999–10019. DOI: 10.1002/2016JC012052.

[263] N. Mayot, P. Nival, and M. Levy. "Primary Production in the Ligurian Sea". In: *The Mediterranean Sea in the Era of Global Change 1*. John Wiley & Sons, Ltd, 2020, pp. 139–164. ISBN: 978-1-119-70696-0. DOI: 10.1002/9781119706960.ch6.

[264] J. McCarthy, M. L. Minsky, N. Rochester, and C. E. Shannon. "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955". In: *AI Magazine* 27.4 (4 2006), pp. 12–12. DOI: 10.1609/aimag.v27i4.1904.

[265] S. McClatchie, R. Cowen, K. Nieto, A. Greer, J. Y. Luo, C. Guigand, D. Demer, D. Griffith, and D. Rudnick. "Resolution of Fine Biological Structure Including Small Narcomedusae across a Front in the Southern California Bight". In: *Journal of Geophysical Research: Oceans* 117.C4 (2012). DOI: 10.1029/2011JC007565.

[266] M. McFarland, A. R. Nayak, N. Stockley, M. Twardowski, and J. Sullivan. "Enhanced Light Absorption by Horizontally Oriented Diatom Colonies". In: *Frontiers in Marine Science* 7 (2020).

[267] D. J. McGillicuddy. "Mechanisms of Physical-Biological-Biogeochemical Interaction at the Oceanic Mesoscale". In: *Annual Review of Marine Science* 8.1 (2016), pp. 125–159. DOI: 10.1146/annurev-marine-010814-015606.

[268] D. J. McGillicuddy et al. "Eddy/Wind Interactions Stimulate Extraordinary Mid-Ocean Plankton Blooms". In: *Science* 316.5827 (2007), pp. 1021–1026. DOI: 10.1126/science.1136256.

[269] M. A. McManus, O. M. Cheriton, P. T. Drake, D. V. Holliday, C. D. Storlazzi, P. L. Donaghay, and C. F. Greenlaw. "Effects of Physical Processes on Structure and Transport of Thin Zooplankton Layers in the Coastal Ocean". In: *Marine Ecology Progress Series* 301 (2005), pp. 199–215. DOI: 10.3354/meps301199.

[270] F. Ménard, S. Dallot, G. Thomas, and J. C. Braconnot. "Temporal Fluctuations of Two Mediterranean Salp Populations from 1967 to 1990. Analysis of the Influence of Environmental Variables Using a Markov Chain Model". In: *Marine Ecology Progress Series* 104.1/2 (1994), pp. 139–152.

[271] A. F. Michaels. "Vertical Distribution and Abundance of Acantharia and Their Symbionts". In: *Marine Biology* 97.4 (1988), pp. 559–569. DOI: 10.1007/BF00391052.

[272] C. E. Mills. "Density Is Altered in Hypromedusae and Ctenophores in Response to Changes in Salinity". In: *The Biological Bulletin* 166.206-215 (1984). DOI: 10.2307/1541442.

[273] P. Miloslavich et al. “Essential Ocean Variables for Global Sustained Observations of Biodiversity and Ecosystem Changes”. In: *Global Change Biology* 24.6 (2018), pp. 2416–2433. DOI: 10.1111/gcb.14108.

[274] A. Mitra et al. “Defining Planktonic Protist Functional Groups on Mechanisms for Energy and Nutrient Acquisition: Incorporation of Diverse Mixotrophic Strategies”. In: *Protist* 167.2 (2016), pp. 106–120. DOI: 10.1016/j.protis.2016.01.003.

[275] J. C. Molinero, F. Ibanez, S. Souissi, E. Bosc, and P. Nival. “Surface Patterns of Zooplankton Spatial Variability Detected by High Frequency Sampling in the NW Mediterranean. Role of Density Fronts”. In: *Journal of Marine Systems*. Physical-Biological Interactions in the Upper Ocean 69.3 (2008), pp. 271–282. DOI: 10.1016/j.jmarsys.2005.11.023.

[276] K. O. Möller, M. S. John, A. Temming, J. Floeter, A. F. Sell, J.-P. Herrmann, and C. Möllmann. “Marine Snow, Zooplankton and Thin Layers: Indications of a Trophic Link from Small-Scale Sampling with the Video Plankton Recorder”. In: *Marine Ecology Progress Series* 468 (2012), pp. 57–69. DOI: 10.3354/meps09984.

[277] M. Moniruzzaman, S. M. S. Islam, M. Bennamoun, and P. Lavery. “Deep Learning on Underwater Marine Object Detection: A Survey”. In: *Advanced Concepts for Intelligent Vision Systems*. Ed. by J. Blanc-Talon, R. Penne, W. Philips, D. Popescu, and P. Scheunders. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2017, pp. 150–160. ISBN: 978-3-319-70353-4. DOI: 10.1007/978-3-319-70353-4_13.

[278] G. E. Moore. “Cramming More Components onto Integrated Circuits, Reprinted from Electronics, Volume 38, Number 8, April 19, 1965, Pp.114 Ff.” In: *IEEE Solid-State Circuits Society Newsletter* 11.3 (2006), pp. 33–35. DOI: 10.1109/N-SSC.2006.4785860.

[279] A. Morel and S. Maritorena. “Bio-Optical Properties of Oceanic Waters: A Reappraisal”. In: *Journal of Geophysical Research* 106.C4 (2001), pp. 7163–7180. DOI: 10.1029/2000JC000319.

[280] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera. “A Unifying View on Dataset Shift in Classification”. In: *Pattern Recognition* 45.1 (2012), pp. 521–530. DOI: 10.1016/j.patcog.2011.06.019.

[281] J. N. Moum, J. D. Nash, and J. M. Klymak. "Small-Scale Processes in the Coastal Ocean". In: *Oceanography* 21.4 (2008), pp. 22–33.

[282] M. Mucko, S. Bosak, R. Casotti, C. Balestra, and Z. Ljubešić. "Winter Picoplankton Diversity in an Oligotrophic Marginal Sea". In: *Marine Genomics* 42 (2018), pp. 14–24. DOI: 10.1016/j.margen.2018.09.002.

[283] J. X. Müller. *Über Die Thalassicollen, Polycystinen Und Acanthometren Des Mittelmeeres*. Druckerei der Königlichen Akademie der Wissenschaften, 1858.

[284] F. E. Muller-Karger et al. "Advancing Marine Biological Observations and Data Requirements of the Complementary Essential Ocean Variables (EOVs) and Essential Biodiversity Variables (EBVs) Frameworks". In: *Frontiers in Marine Science* 5 (2018).

[285] Y. Nakamura, R. Somiya, N. Suzuki, M. Hidaka-Umetsu, A. Yamaguchi, and D. J. Lindsay. "Optics-Based Surveys of Large Unicellular Zooplankton: A Case Study on Radiolarians and Phaeodarians". In: *Plankton and Benthos Research* 12.2 (2017), pp. 95–103. DOI: 10.3800/pbr.12.95.

[286] Y. Nakamura and N. Suzuki. "Phaeodaria: Diverse Marine Cercozoans of World-Wide Distribution". In: *Marine Protists: Diversity and Dynamics*. Ed. by S. Ohtsuka, T. Suzaki, T. Horiguchi, N. Suzuki, and F. Not. Tokyo: Springer Japan, 2015, pp. 223–249. ISBN: 978-4-431-55130-0. DOI: 10.1007/978-4-431-55130-0_9.

[287] A. R. Nayak, E. Malkiel, M. N. McFarland, M. S. Twardowski, and J. M. Sullivan. "A Review of Holography in the Aquatic Sciences: In Situ Characterization of Particles, Plankton, and Small Scale Biophysical Interactions". In: *Frontiers in Marine Science* 7 (2021), p. 1256. DOI: 10.3389/fmars.2020.572147.

[288] A. R. Nayak, M. N. McFarland, J. M. Sullivan, and M. S. Twardowski. "Evidence for Ubiquitous Preferential Particle Orientation in Representative Oceanic Shear Flows". In: *Limnology and Oceanography* 63.1 (2018), pp. 122–143. DOI: 10.1002/lno.10618.

[289] A. Niculescu-Mizil and R. Caruana. "Predicting Good Probabilities with Supervised Learning". In: *Proceedings of the 22nd International Conference on Machine Learning*. ICML '05. New York, NY, USA: Association for Computing Machinery, 2005, pp. 625–632. ISBN: 978-1-59593-180-1. DOI: 10.1145/1102351.1102430.

[290] K. Niewiadomska, H. Claustre, L. Prieur, and F. d'Ortenzio. "Submesoscale Physical-Biogeochemical Coupling across the Ligurian Current (Northwestern Mediterranean) Using a Bio-Optical Glider". In: *Limnology and Oceanography* 53 (5part2 2008), pp. 2210–2225. DOI: 10.4319/lo.2008.53.5_part_2.2210.

[291] Y. Nishibe, K. Takahashi, T. Ichikawa, K. Hidaka, H. Kurogi, K. Segawa, and H. Saito. "Degradation of Discarded Appendicularian Houses by Oncaeid Copepods". In: *Limnology and Oceanography* 60.3 (2015), pp. 967–976. DOI: 10.1002/lno.10061.

[292] P. Nival, F. Lombard, J. Cuzin, J. Goy, and L. Stemmann. "Zooplankton II. Macroplankton and Long-Term Series". In: *The Mediterranean Sea in the Era of Global Change 2*. John Wiley & Sons, Ltd, 2020, pp. 109–146. ISBN: 978-1-119-70478-2. DOI: 10.1002/9781119704782.ch4.

[293] M. D. Ohman, R. E. Davis, J. T. Sherman, K. R. Grindley, B. M. Whitmore, C. F. Nickels, and J. S. Ellen. "Zooglider: An Autonomous Vehicle for Optical and Acoustic Sensing of Zooplankton". In: *Limnology and Oceanography: Methods* 17.1 (2019), pp. 69–86. DOI: 10.1002/lom3.10301.

[294] M. D. Ohman, J. R. Powell, M. Picheral, and D. W. Jensen. "Mesozooplankton and Particulate Matter Responses to a Deep-Water Frontal System in the Southern California Current System". In: *Journal of Plankton Research* 34.9 (2012), pp. 815–827. DOI: 10.1093/plankt/fbs028.

[295] S. Ohtsuka, T. Suzaki, T. Horiguchi, N. Suzuki, and F. Not. "Marine Protists". In: *Diversity and Dynamics. Springer, Tokyo (Japan)* (2015).

[296] J. Oksanen et al. *Vegan: Community Ecology Package*. 2018.

[297] D. B. Olson, G. L. Hitchcock, A. J. Mariano, C. J. Ashjian, G. Peng, R. W. Nero, and G. P. Podestá. "Life on the Edge: Marine Life and Fronts". In: *Oceanography* 7.2 (1994), pp. 52–60.

[298] R. J. Olson and H. M. Sosik. "A Submersible Imaging-in-Flow Instrument to Analyze Nano-and Microplankton: Imaging FlowCytobot". In: *Limnology and Oceanography: Methods* 5.6 (2007), pp. 195–203. DOI: 10.4319/lom.2007.5.195.

[299] E. C. Orenstein et al. "Machine Learning Techniques to Characterize Functional Traits of Plankton from Image Data". 2021.

[300] E. C. Orenstein and O. Beijbom. "Transfer Learning and Deep Feature Extraction for Planktonic Image Data Sets". In: *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). 2017, pp. 1082–1088. DOI: 10.1109/WACV.2017.125.

[301] E. C. Orenstein, O. Beijbom, E. E. Peacock, and H. M. Sosik. "WHOI-Plankton- A Large Scale Fine Grained Visual Recognition Benchmark Dataset for Plankton Classification". 2015.

[302] E. C. Orenstein, D. Ratelle, C. Briseño-Avena, M. L. Carter, P. J. S. Franks, J. S. Jaffe, and P. L. D. Roberts. "The Scripps Plankton Camera System: A Framework and Platform for in Situ Microscopy". In: *Limnology and Oceanography: Methods* 18.11 (2020), pp. 681–695. DOI: 10.1002/lom3.10394.

[303] N. Otsu. "A Threshold Selection Method from Gray-Level Histograms". In: *IEEE transactions on systems, man, and cybernetics* 9.1 (1979), pp. 62–66.

[304] R. W. Owen. "Fronts and Eddies in the Sea: Mechanisms, Interactions and Biological Effects". In: *Analysis of marine ecosystems* (1981), pp. 197–233.

[305] G.-A. Paffenhöfer. "On the Ecology of Marine Cyclopoid Copepods (Crustacea, Copepoda)". In: *Journal of Plankton Research* 15.1 (1993), pp. 37–55. DOI: 10.1093/plankt/15.1.37.

[306] **T. Panaïotis**, L. Caray–Counil, B. Woodward, M. S. Schmid, D. Daprano, S. T. Tsai, C. M. Sullivan, R. K. Cowen, and J.-O. Irisson. "Content-Aware Segmentation of Objects Spanning a Large Size Range: Application to Plankton Images". In: *Frontiers in Marine Science* 9 (2022). DOI: 10.3389/fmars.2022.870005.

[307] D. Parikh, C. L. Zitnick, and T. Chen. "Exploring Tiny Images: The Roles of Appearance and Contextual Information for Machine and Human Object Recognition". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.10 (2012), pp. 1978–1991. DOI: 10.1109/TPAMI.2011.276.

[308] A. Paszke et al. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. 2019. DOI: 10.48550/arXiv.1912.01703.

[309] J. Patterson and A. Gibson. *Deep Learning: A Practitioner's Approach*. " O'Reilly Media, Inc.", 2017. ISBN: 1-4919-1423-8.

[310] F. Pedregosa et al. “Scikit-Learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12.85 (2011), pp. 2825–2830.

[311] M. L. Pedrotti and L. Fenaux. “Dispersal of Echinoderm Larvae in a Geographical Area Marked by Upwelling (Ligurian Sea, NW Mediterranean)”. In: *Marine Ecology Progress Series* 86.3 (1992), pp. 217–227.

[312] H. M. Pereira et al. “Essential Biodiversity Variables”. In: *Science* 339.6117 (2013), pp. 277–278. DOI: 10.1126/science.1229931.

[313] M. C. Pernice, C. R. Giner, R. Logares, J. Perera-Bel, S. G. Acinas, C. M. Duarte, J. M. Gasol, and R. Massana. “Large Variability of Bathypelagic Microbial Eukaryotic Communities across the World’s Oceans”. In: *The ISME Journal* 10.4 (4 2016), pp. 945–958. DOI: 10.1038/ismej.2015.170.

[314] F. Péron and C. A. Lesueur. “Tableau Des Caractères Génériques et Spécifiques de Toutes Les Espèces de Méduses Connues Jusqu’à Ce Jour”. In: *Annales Du Muséum d’Histoire Naturelle*. Vol. 14. 1810, pp. 325–366.

[315] S. Petrovskii and N. Petrovskaya. “Computational Ecology as an Emerging Science”. In: *Interface Focus* 2.2 (2012), pp. 241–254. DOI: 10.1098/rsfs.2011.0083.

[316] M. Picheral, S. Colin, and J.-O. Irisson. *EcoTaxa, a Tool for the Taxonomic Classification of Images*. 2017. URL: https://ecotaxa.obs-vlfr.fr/ (visited on 11/13/2020).

[317] M. Picheral, L. Guidi, L. Stemmann, D. M. Karl, G. Iddaoud, and G. Gorsky. “The Underwater Vision Profiler 5: An Advanced Instrument for High Spatial Resolution Studies of Particle Size Spectra and Zooplankton”. In: *Limnology and Oceanography: Methods* 8.9 (2010), pp. 462–473. DOI: 10.4319/lom.2010.8.462.

[318] M. Picheral et al. “The Underwater Vision Profiler 6: An Imaging Sensor of Particle Size Spectra and Plankton, for Autonomous and Cabled Platforms”. In: *Limnology and Oceanography: Methods* n/a.n/a (2021). DOI: 10.1002/lom3.10475.

[319] S. Pinca and S. Dallot. “Meso- and Macrozooplankton Composition Patterns Related to Hydrodynamic Structures in the Ligurian Sea (Trophos-2 Experiment, April-June 1986)”. In: *Ma-*

rine Ecology Progress Series 126 (1995), pp. 49–65. DOI: 10.3354/meps126049.

[320] M. H. Pinkerton, A. N. H. Smith, B. Raymond, G. W. Hosie, B. Sharp, J. R. Leathwick, and J. M. Bradford-Grieve. "Spatial and Seasonal Distribution of Adult Oithona Similis in the Southern Ocean: Predictions Using Boosted Regression Trees". In: *Deep Sea Research Part I: Oceanographic Research Papers* 57.4 (2010), pp. 469–485. DOI: 10.1016/j.dsr.2009.12.010.

[321] M. H. Pinkerton, M. Décima, J. A. Kitchener, K. T. Takahashi, K. V. Robinson, R. Stewart, and G. W. Hosie. "Zooplankton in the Southern Ocean from the Continuous Plankton Recorder: Distributions and Long-Term Change". In: *Deep Sea Research Part I: Oceanographic Research Papers* 162 (2020), p. 103303. DOI: 10.1016/j.dsr.2020.103303.

[322] L. Piterbarg, V. Taillandier, and A. Griffa. "Investigating Frontal Variability from Repeated Glider Transects in the Ligurian Current (North West Mediterranean Sea)". In: *Journal of Marine Systems* 129 (2014), pp. 381–395. DOI: 10.1016/j.jmarsys.2013.08.003.

[323] T. Poisot, R. Labrie, E. Larson, and A. Rahlin. *Data-Based, Synthesis-Driven: Setting the Agenda for Computational Ecology*. 2018. DOI: 10.1101/150128.

[324] S. Polet, C. Berney, J. Fahrni, and J. a. n. Pawlowski. "Small-Subunit Ribosomal RNA Gene Sequences of Phaeodarea Challenge the Monophyly of Haeckel's Radiolaria". In: *Protist* 155.1 (2004), pp. 53–63. DOI: 10.1078/1434461000164.

[325] J. C. Prairie, K. R. Sutherland, K. J. Nickols, and A. M. Kaltenberg. "Biophysical Interactions in the Plankton: A Cross-Scale Review". In: *Limnology and Oceanography: Methods* 2 (2012), pp. 121–145. DOI: 10.1215/21573689-1964713@10.1002/(ISSN)1541-5856.ECODAS-VI.

[326] L. Prieur, F. D'ortenzio, V. Taillandier, and P. Testor. "Physical Oceanography of the Ligurian Sea". In: *The Mediterranean Sea in the Era of Global Change 1*. John Wiley & Sons, Ltd, 2020, pp. 49–78. ISBN: 978-1-119-70696-0. DOI: 10.1002/9781119706960.ch3.

[327] O. Py, H. Hong, and S. Zhongzhi. "Plankton Classification with Deep Convolutional Neural Networks". In: *2016 IEEE Information Technology, Networking, Electronic and Automation Control Conference*. 2016 IEEE Information Technology, Networking, Electronic and Automation Control Conference. 2016, pp. 132–136. DOI: 10.1109/ITNEC.2016.7560334.

[328] S. Rahmstorf. "Ocean Circulation and Climate during the Past 120,000 Years". In: *Nature* 419.6903 (6903 2002), pp. 207–214. DOI: 10.1038/nature01090.

[329] S. Rahmstorf. "Thermohaline Circulation: The Current Climate". In: *Nature* 421.6924 (6924 2003), pp. 699–699. DOI: 10.1038/421699a.

[330] A. Remsen, T. L. Hopkins, and S. Samson. "What You See Is Not What You Catch: A Comparison of Concurrently Collected Net, Optical Plankton Counter, and Shadowed Image Particle Profiling Evaluation Recorder Data from the Northeast Gulf of Mexico". In: *Deep Sea Research Part I: Oceanographic Research Papers* 51.1 (2004), pp. 129–151. DOI: 10.1016/J.DSR.2003.09.008.

[331] G. Reygondeau, L. Guidi, G. Beaugrand, S. A. Henson, P. Koubbi, B. R. MacKenzie, T. T. Sutton, M. Fioroni, and O. Maury. "Global Biogeochemical Provinces of the Mesopelagic Zone". In: *Journal of Biogeography* 45.2 (2018), pp. 500–514. DOI: 10.1111/jbi.13149.

[332] A. Rice, P. Šmarda, M. Novosolov, M. Drori, L. Glick, N. Sabath, S. Meiri, J. Belmaker, and I. Mayrose. "The Global Biogeography of Polyploid Plants". In: *Nature Ecology & Evolution* 3.2 (2 2019), pp. 265–273. DOI: 10.1038/s41559-018-0787-9.

[333] D. J. Richter et al. "Genomic Evidence for Global Ocean Plankton Biogeography Shaped by Large-Scale Current Systems". In: *bioRxiv* (2020), p. 867739. DOI: 10.1101/867739.

[334] A. R. Robinson. "Overview and Summary of Eddy Science". In: *Eddies in Marine Science*. Ed. by A. R. Robinson. Topics in Atmospheric and Oceanographic Sciences. Berlin, Heidelberg: Springer, 1983, pp. 3–15. ISBN: 978-3-642-69003-7. DOI: 10.1007/978-3-642-69003-7_1.

[335] C. Robinson et al. "Mesopelagic Zone Ecology and Biogeochemistry – a Synthesis". In: *Deep Sea Research Part II: Topical Studies in Oceanography*. Ecological and Biogeochemical Interactions in the Dark Ocean 57.16 (2010), pp. 1504–1518. DOI: 10.1016/j.dsr2.2010.02.018.

[336] K. L. Robinson, J. Y. Luo, S. Sponaugle, C. Guigand, and R. K. Cowen. "A Tale of Two Crowds: Public Engagement in Plankton Classification". In: *Frontiers in Marine Science* 4 (2017).

[337] K. L. Robinson, S. Sponaugle, J. Y. Luo, M. R. Gleiber, and R. K. Cowen. "Big or Small, Patchy All: Resolution of Marine Plankton Patch Structure at Micro- to Submesoscales for 36 Taxa". In: *Science Advances* 7.47 (2021), eabk2904. DOI: 10.1126/sciadv.abk2904.

[338] F. C. M. Rodrigues, N. S. Hirata, A. A. Abello, T. Leandro, D. La Cruz, R. M. Lopes, and R. Hirata Jr. "Evaluation of Transfer Learning Scenarios in Plankton Image Classification." In: *VISIGRAPP (5: VISAPP)*. 2018, pp. 359–366.

[339] J.-B. Romagnan et al. "Comprehensive Model of Annual Plankton Succession Based on the Whole-Plankton Time Series Approach". In: *PLOS ONE* 10.3 (2015). Ed. by B. R. MacKenzie, e0119219. DOI: 10.1371/journal.pone.0119219.

[340] I. Rombouts, G. Beaugrand, F. Ibaňez, S. Gasparini, S. Chiba, and L. Legendre. "Global Latitudinal Variations in Marine Copepod Diversity and Environmental Factors". In: *Proceedings of the Royal Society B: Biological Sciences* 276.1670 (2009), pp. 3053–3062. DOI: 10.1098/rspb.2009.0742.

[341] F. Rosenblatt. "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain". In: *Psychological Review* 65 (1958), pp. 386–408. DOI: 10.1037/h0042519.

[342] D. L. Rudnick, R. E. Davis, C. C. Eriksen, D. M. Fratantoni, and M. J. Perry. "Underwater Gliders for Ocean Research". In: *Marine Technology Society Journal* 38.2 (2004), pp. 73–84. DOI: 10.4031/002533204787522703.

[343] O. Russakovsky et al. "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision* 115.3 (2015), pp. 211–252. DOI: 10.1007/s11263-015-0816-y.

[344] S. J. Russell. *Artificial Intelligence a Modern Approach*. Pearson Education, Inc., 2010. ISBN: 0-13-604259-7.

[345] S. Rutherford, S. D'Hondt, and W. Prell. "Environmental Controls on the Geographic Distribution of Zooplankton Diversity". In: *Nature* 400.6746 (1999), pp. 749–753. DOI: 10.1038/23449.

[346] V. Sandel, R. Kiko, P. Brandt, M. Dengler, L. Stemmann, P. Vandromme, U. Sommer, and H. Hauss. "Nitrogen Fuelling of the Pelagic Food Web of the Tropical Atlantic". In: *PLOS ONE* 10.6 (2015), e0131258. DOI: 10.1371/journal.pone.0131258.

[347] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. "MobileNetV2: Inverted Residuals and Linear Bottlenecks". 2019.

[348] H. Santander Bueno. "The Zooplankton in an Upwelling Area off Peru". In: *Coastal Upwelling*. American Geophysical Union (AGU), 1981, pp. 411–416. ISBN: 978-1-118-66532-9. DOI: 10.1029/CO001p0411.

[349] C. Sardet. *Plankton: Wonders of the Drifting World*. University of Chicago Press, 2015. ISBN: 0-226-26534-X.

[350] R. Sato, Y. Tanaka, and T. Ishimaru. "Species-Specific House Productivity of Appendicularians". In: *Marine Ecology Progress Series* 259 (2003), pp. 163–172. DOI: 10.3354/meps259163.

[351] R. Sauzède, H. Claustre, C. Jamet, J. Uitz, J. Ras, A. Mignot, and F. D'Ortenzio. "Retrieving the Vertical Distribution of Chlorophyll a Concentration and Phytoplankton Community Composition from in Situ Fluorescence Profiles: A Method Based on a Neural Network with Potential for Global-Scale Applications". In: *Journal of Geophysical Research: Oceans* 120.1 (2015), pp. 451–470. DOI: 10.1002/2014JC010355.

[352] K. L. Scales, P. I. Miller, C. B. Embling, S. N. Ingram, E. Pirotta, and S. C. Votier. "Mesoscale Fronts as Foraging Habitats: Composite Front Mapping Reveals Oceanographic Drivers of Habitat Use for a Pelagic Seabird". In: *Journal of The Royal Society Interface* 11.100 (2014), p. 20140679. DOI: 10.1098/rsif.2014.0679.

[353] M. S. Schmid, D. Daprano, K. M. Jacobson, C. Sullivan, C. Briseño-Avena, J. Y. Luo, and R. K. Cowen. *A Convolutional Neural Network Based High-Throughput Image Classification Pipeline - Code and Documentation to Process Plankton Underwater Imagery*

Using Local HPC Infrastructure and NSF's XSEDE. Zenodo, 2021. DOI: 10.5281/zenodo.4641158.

[354] M. S. Schmid, R. K. Cowen, K. Robinson, J. Y. Luo, C. Briseño-Avena, and S. Sponaugle. "Prey and Predator Overlap at the Edge of a Mesoscale Eddy: Fine-Scale, in-Situ Distributions to Inform Our Understanding of Oceanographic Processes". In: *Scientific Reports* 10.1 (2020), pp. 1–16. DOI: 10.1038/s41598-020-57879-x.

[355] M. S. Schmid and L. Fortier. "The Intriguing Co-Distribution of the Copepods Calanus Hyperboreus and Calanus Glacialis in the Subsurface Chlorophyll Maximum of Arctic Seas". In: *Elementa: Science of the Anthropocene* 7 (2019). Ed. by J. W. Deming and J. E. Keister, p. 50. DOI: 10.1525/elementa.388.

[356] S. Schmidt, J. A. Raven, C. Paungfoo-Lonhienne, S. Schmidt, J. A. Raven, and C. Paungfoo-Lonhienne. "The Mixotrophic Nature of Photosynthetic Plants". In: *Functional Plant Biology* 40.5 (2013), pp. 425–438. DOI: 10.1071/FP13061.

[357] S.-M. Schröder, R. Kiko, J.-O. Irisson, and R. Koch. "Low-Shot Learning of Plankton Categories". In: *Pattern Recognition*. Ed. by T. Brox, A. Bruhn, and M. Fritz. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019, pp. 391–404. ISBN: 978-3-030-12939-2. DOI: 10.1007/978-3-030-12939-2_27.

[358] S.-M. Schröder, R. Kiko, and R. Koch. "MorphoCluster: Efficient Annotation of Plankton Images by Clustering". In: *Sensors* 20.11 (11 2020), p. 3060. DOI: 10.3390/s20113060.

[359] J. Schulz, K. Barz, P. Ayon, A. Lüdtke, O. Zielinski, D. Mengedoht, and H.-J. Hirche. "Imaging of Plankton Specimens with the Lightframe On-Sight Keyspecies Investigation (LOKI) System". In: *Journal of the European Optical Society - Rapid publications* 5.0 (0 2010). DOI: 10.2971/jeos.2010.10017s.

[360] K. G. Sellner. "Trophodynamics of Marine Cyanobacteria Blooms". In: *Marine Pelagic Cyanobacteria: Trichodesmium and Other Diazotrophs*. Dordrecht: Springer Netherlands, 1992, pp. 75–94. DOI: 10.1007/978-94-015-7977-3_6.

[361] E. Ser-Giacomi, L. Zinger, S. Malviya, C. De Vargas, E. Karsenti, C. Bowler, and S. De Monte. "Ubiquitous Abundance Distribution of Non-Dominant Plankton across the Global Ocean". In: *Nature Ecology & Evolution* 2.8 (8 2018), pp. 1243–1249. DOI: 10.1038/s41559-018-0587-2.

[362] R. W. Sheldon and T. R. Parsons. "A Continuous Size Spectrum for Particulate Matter in the Sea". In: *Journal of the Fisheries Research Board of Canada* 24.5 (1967), pp. 909–915. DOI: 10.1139/f67-081.

[363] R. W. Sheldon, A. Prakash, and W. H. Sutcliffe. "The Size Distribution of Particles in the Ocean". In: *Limnology and Oceanography* 17.3 (1972), pp. 327–340. DOI: 10.4319/lo.1972.17.3.0327.

[364] K. Sherman. "The Large Marine Ecosystem Approach for Assessment and Management of Ocean Coastal Waters". In: *Sustaining Large Marine Ecosystems: The Human Dimension*. Ed. by T. Hennessey and J. Sutinen. Elsevier, 2005, pp. 3–16. DOI: 10.1016/S1570-0461(05)80025-4.

[365] C. K. Sieracki, M. E. Sieracki, and C. S. Yentsch. "An Imaging-in-Flow System for Automated Analysis of Marine Microplankton". In: *Marine Ecology Progress Series* 168 (1998), pp. 285–296. DOI: 10.3354/meps168285.

[366] S. R. Signorini, B. A. Franz, and C. R. McClain. "Chlorophyll Variability in the Oligotrophic Gyres: Mechanisms, Seasonality and Trends". In: *Frontiers in Marine Science* 2 (2015).

[367] R. Simpson, R. Williams, R. Ellis, and P. F. Culverhouse. "Biological Pattern Recognition by Neural Networks". In: *Marine Ecology Progress Series* 79.3 (1992), pp. 303–308.

[368] L. N. Smith. *A Disciplined Approach to Neural Network Hyper-Parameters: Part 1 – Learning Rate, Batch Size, Momentum, and Weight Decay*. 2018. DOI: 10.48550/arXiv.1803.09820.

[369] G. Sommeria-Klein, R. Watteaux, F. M. Ibarbalz, J. J. Pierella Karlusich, D. Iudicone, C. Bowler, and H. Morlon. "Global Drivers of Eukaryotic Plankton Biogeography in the Sunlit Ocean". In: *Science* 374.6567 (2021), pp. 594–599. DOI: 10.1126/science.abb3717.

[370] V. Sonnet, L. Guidi, C. B. Mouw, G. Puggioni, and S.-D. Ayata. "Length, Width, Shape Regularity, and Chain Structure: Time Series Analysis of Phytoplankton Morphology from Imagery". In: *Limnology and Oceanography* 67.8 (2022), pp. 1850–1864. DOI: 10.1002/lno.12171.

[371] H. M. Sosik, E. E. Peacock, and E. F. Brownlee. *WHOI-Plankton. Annotated Plankton Images - Data Set for Developing and Evaluating Classification Methods.* 2015.

[372] H. M. Sosik and R. J. Olson. "Automated Taxonomic Classification of Phytoplankton Sampled with Imaging-in-Flow Cytometry". In: *Limnology and Oceanography: Methods* 5.6 (2007), pp. 204–216. DOI: 10.4319/lom.2007.5.204.

[373] H. M. Sosik, E. E. Peacock, and E. F. Brownlee. "Annotated Plankton Images Data Set for Developing and Evaluating Classification Methods". In: *URL http://darchive. mblwhoilibrary. org/handle/1912/7341* (2015).

[374] Y. D. Soviadan, F. Benedetti, M. C. Brandão, S.-D. Ayata, J.-O. Irisson, J. L. Jamet, R. Kiko, F. Lombard, K. Gnandi, and L. Stemmann. "Patterns of Mesozooplankton Community Composition and Vertical Fluxes in the Global Ocean". In: *Progress in Oceanography* 200 (2022), p. 102717. DOI: 10.1016/j.pocean.2021.102717.

[375] M. D. Spalding, V. N. Agostini, J. Rice, and S. M. Grant. "Pelagic Provinces of the World: A Biogeographic Classification of the World's Surface Pelagic Waters". In: *Ocean & Coastal Management* 60 (2012), pp. 19–30. DOI: 10.1016/j.ocecoaman.2011.12.016.

[376] M. D. Spalding et al. "Marine Ecoregions of the World: A Bioregionalization of Coastal and Shelf Areas". In: *BioScience* 57.7 (2007), pp. 573–583. DOI: 10.1641/B570707.

[377] D. Speijer, J. Lukeš, and M. Eliáš. "Sex Is a Ubiquitous, Ancient, and Inherent Attribute of Eukaryotic Life". In: *Proceedings of the National Academy of Sciences* 112.29 (2015), pp. 8827–8834. DOI: 10.1073/pnas.1501725112.

[378] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: *The journal of machine learning research* 15.1 (2014), pp. 1929–1958.

[379] D. K. Steinberg and M. R. Landry. "Zooplankton and the Ocean Carbon Cycle". In: *Annual review of marine science* 9 (2017), pp. 413–444.

[380] L. Stemmann and E. Boss. "Plankton and Particle Size and Packaging: From Determining Optical Properties to Driving the Biological Pump". In: *Annual Review of Marine Science* 4 (2012), pp. 263–290.

[381] L. Stemmann, G. Gorsky, J.-C. Marty, M. Picheral, and J.-C. Miquel. "Four-Year Study of Large-Particle Vertical Distribution (0–1000m) in the NW Mediterranean in Relation to Hydrology, Phytoplankton, and Vertical Flux". In: *Deep Sea Research Part II: Topical Studies in Oceanography*. Studies at the DYFAMED (France JGOFS) Time-Series Station, N.W. M Editerranean Sea 49.11 (2002), pp. 2143–2162. DOI: 10.1016/S0967-0645(02)00032-2.

[382] L. Stemmann, A. Hosia, M. J. Youngbluth, H. Søiland, M. Picheral, and G. Gorsky. "Vertical Distribution (0–1000 m) of Macrozooplankton, Estimated Using the Underwater Video Profiler, in Different Hydrographic Regimes along the Northern Portion of the Mid-Atlantic Ridge". In: *Deep Sea Research Part II: Topical Studies in Oceanography* 55.1-2 (2008), pp. 94–105. DOI: 10.1016/J.DSR2.2007.09.019.

[383] L. Stemmann, M. Picheral, and G. Gorsky. "Diel Variation in the Vertical Distribution of Particulate Matter (>0.15mm) in the NW Mediterranean Sea Investigated with the Underwater Video Profiler". In: *Deep Sea Research Part I: Oceanographic Research Papers* 47.3 (2000), pp. 505–531. DOI: 10.1016/S0967-0637(99)00100-4.

[384] L. Stemmann, L. Prieur, L. Legendre, I. Taupier-Letage, M. Picheral, L. Guidi, and G. Gorsky. "Effects of Frontal Processes on Marine Aggregate Dynamics and Fluxes: An Interannual Study in a Permanent Geostrophic Front (NW Mediterranean)". In: *Journal of Marine Systems* 70.1 (2008), pp. 1–20. DOI: 10.1016/j.jmarsys.2007.02.014.

[385] L. Stemmann, M. J. Youngbluth, K. Robert, A. Hosia, M. Picheral, H. Paterson, F. Ibaňez, L. Guidi, F. Lombard, and G. Gorsky. "Global Zoogeography of Fragile Macrozooplankton in the Upper 100-1000 m Inferred from the Underwater Video Profiler".

In: *ICES Journal of Marine Science* 65.3 (2008), pp. 433–442. DOI: 10.1093/icesjms/fsn010.

[386] D. K. Stoecker, P. J. Hansen, D. A. Caron, and A. Mitra. "Mixotrophy in the Marine Plankton". In: *Annual Review of Marine Science* 9.1 (2017), pp. 311–335. DOI: 10.1146/annurev-marine-010816-060617.

[387] D. K. Stoecker, M. D. Johnson, C. de Vargas, and F. Not. "Acquired Phototrophy in Aquatic Protists". In: *Aquatic Microbial Ecology* 57.3 (2009), pp. 279–310. DOI: 10.3354/ame01340.

[388] M. R. Stukel, T. Biard, J. Krause, and M. D. Ohman. "Large Phaeodaria in the Twilight Zone: Their Role in the Carbon Cycle". In: *Limnology and Oceanography* 63.6 (2018), pp. 2579–2594. DOI: 10.1002/lno.10961.

[389] Z. Su, J. Wang, P. Klein, A. F. Thompson, and D. Menemenlis. "Ocean Submesoscales as a Key Component of the Global Heat Budget". In: *Nature Communications* 9.1 (1 2018), p. 775. DOI: 10.1038/s41467-018-02983-w.

[390] S. Sunagawa et al. "Structure and Function of the Global Ocean Microbiome". In: *Science (New York, N.Y.)* 348.6237 (2015), p. 1261359. DOI: 10.1126/science.1261359.

[391] S. Sunagawa et al. "Tara Oceans: Towards Global Ocean Ecosystems Biology". In: *Nature Reviews Microbiology* 18.8 (8 2020), pp. 428–445. DOI: 10.1038/s41579-020-0364-5.

[392] T. T. Sutton et al. "A Global Biogeographic Classification of the Mesopelagic Zone". In: *Deep Sea Research Part I: Oceanographic Research Papers* 126 (2017), pp. 85–102. DOI: 10.1016/J.DSR.2017.05.006.

[393] N. Suzuki and F. Not. "Biology and Ecology of Radiolaria". In: *Marine Protists: Diversity and Dynamics*. Ed. by S. Ohtsuka, T. Suzaki, T. Horiguchi, N. Suzuki, and F. Not. Tokyo: Springer Japan, 2015, pp. 179–222. ISBN: 978-4-431-55130-0. DOI: 10.1007/978-4-431-55130-0_8.

[394] N. Swanberg, P. Bennett, J. L. Lindsey, and O. R. Anderson. "The Biology of a Coelodendrid: A Mesopelagic Phaeodarian Radiolarian". In: *Deep Sea Research Part A. Oceanographic Research Papers* 33.1 (1986), pp. 15–25. DOI: 10.1016/0198-0149(86)90105-6.

[395] K. Swieca, S. Sponaugle, C. Briseño-Avena, M. S. Schmid, R. D. Brodeur, and R. K. Cowen. "Changing with the Tides: Fine-Scale Larval Fish Prey Availability and Predation Pressure near a Tidally Modulated River Plume". In: *Marine Ecology Progress Series* 650 (2020), pp. 217–238. DOI: 10.3354/meps13367.

[396] S. Talapatra, J. Hong, M. McFarland, A. R. Nayak, C. Zhang, J. Katz, J. Sullivan, M. Twardowski, J. Rines, and P. Donaghay. "Characterization of Biophysical Interactions in the Water Column Using in Situ Digital Holography". In: *Marine Ecology Progress Series* 473 (2013), pp. 29–51. DOI: 10.3354/meps10049.

[397] H. Tappan and A. R. Loeblich. "Evolution of the Oceanic Plankton". In: *Earth-Science Reviews* 9.3 (1973), pp. 207–240. DOI: 10.1016/0012-8252(73)90092-5.

[398] M. Thyssen, G. J. Grégori, J.-M. Grisoni, M. L. Pedrotti, L. Mousseau, L. F. Artigas, S. Marro, N. Garcia, O. Passafiume, and M. J. Denis. "Onset of the Spring Bloom in the Northwestern Mediterranean Sea: Influence of Environmental Pulse Events on the in Situ Hourly-Scale Dynamics of the Phytoplankton Community Structure". In: *Frontiers in Microbiology* 5 (2014).

[399] D. P. Tittensor, C. Mora, W. Jetz, H. K. Lotze, D. Ricard, E. V. Berghe, and B. Worm. "Global Patterns and Predictors of Marine Biodiversity across Taxa". In: *Nature* 466.7310 (2010), pp. 1098–1101. DOI: 10.1038/nature09329.

[400] E. Trudnowska, L. Stemmann, K. Błachowiak-Samołyk, and S. Kwasniewski. "Taxonomic and Size Structures of Zooplankton Communities in the Fjords along the Atlantic Water Passage to the Arctic". In: *Journal of Marine Systems* 204 (2020), p. 103306. DOI: 10.1016/j.jmarsys.2020.103306.

[401] E. Trudnowska, L. Lacour, M. Ardyna, A. Rogge, J. O. Irisson, A. M. Waite, M. Babin, and L. Stemmann. "Marine Snow Morphology Illuminates the Evolution of Phytoplankton Blooms and Determines Their Subsequent Vertical Export". In: *Nature Communications* 12.1 (1 2021), p. 2816. DOI: 10.1038/s41467-021-22994-4.

[402] G. Tsechpenakis, C. Guigand, and R. K. Cowen. "Image Analysis Techniques to Accompany a New In Situ Ichthyoplankton Imaging System". In: *OCEANS 2007 - Europe*. OCEANS 2007 - Europe. 2007, pp. 1–6. DOI: 10.1109/OCEANSE.2007.4302271.

[403] A. M. Turing. "Computing Machinery and Intelligence". In: *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*. Ed. by R. Epstein, G. Roberts, and G. Beber. Dordrecht: Springer Netherlands, 2009, pp. 23–65. ISBN: 978-1-4020-6710-5. DOI: 10.1007/978-1-4020-6710-5_3.

[404] J. T. Turner. "Zooplankton Fecal Pellets, Marine Snow, Phytodetritus and the Ocean's Biological Pump". In: *Progress in Oceanography* 130 (2015), pp. 205–248. DOI: 10.1016/J.POCEAN.2014.08.005.

[405] K. Uchida, M. Tanaka, and M. Okutomi. "Coupled Convolution Layer for Convolutional Neural Network". In: *Neural Networks* 105 (2018), pp. 197–205. DOI: 10.1016/j.neunet.2018.05.002.

[406] C. Uwizeye et al. "Cytoklepty in the Plankton: A Host Strategy to Optimize the Bioenergetic Machinery of Endosymbiotic Algae". In: *Proceedings of the National Academy of Sciences* 118.27 (2021), e2025252118. DOI: 10.1073/pnas.2025252118.

[407] G. Van Horn and P. Perona. "The Devil Is in the Tails: Fine-grained Classification in the Wild". 2017.

[408] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. "Attention Is All You Need". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017.

[409] L. Vilgrain, F. Maps, M. Picheral, M. Babin, C. Aubry, J.-O. Irisson, and S.-D. Ayata. "Trait-Based Approach Using in Situ Copepod Images Reveals Contrasting Ecological Patterns across an Arctic Ice Melt Zone". In: *Limnology and Oceanography* 66.4 (2021), pp. 1155–1167. DOI: 10.1002/lno.11672.

[410] A. Walsby. "The Properties and Buoyancy-Providing Role of Gas Vacuoles in Trichodesmium Ehrenberg". In: *British Phycological Journal* 13.2 (1978), pp. 103–116. DOI: 10.1080/00071617800650121.

[411] S. van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, and T. Yu. "Scikit-Image: Image Processing in Python". In: *PeerJ* 2 (2014), e453. DOI: 10.7717/peerj.453.

[412] Z. Wang, S. Chang, Y. Yang, D. Liu, and T. S. Huang. "Studying Very Low Resolution Recognition Using Deep Networks". In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016, pp. 4792–4800.

[413] D. M. Ware and R. E. Thomson. "Bottom-Up Ecosystem Trophic Dynamics Determine Fish Production in the Northeast Pacific". In: *Science* 308.5726 (2005), pp. 1280–1284. DOI: 10.1126/SCIENCE.1109049.

[414] L. Watling, J. Guinotte, M. R. Clark, and C. R. Smith. "A Proposed Biogeography of the Deep Ocean Floor". In: *Progress in Oceanography* 111 (2013), pp. 91–112. DOI: 10.1016/j.pocean.2012.11.003.

[415] T. K. Westberry and D. A. Siegel. "Spatial and Temporal Distribution of Trichodesmium Blooms in the World's Oceans". In: *Global Biogeochemical Cycles* 20.4 (2006). DOI: 10.1029/2005GB002673.

[416] M. F. Wilkins, L. Boddy, C. W. Morris, and R. Jonker. "A Comparison of Some Neural and Non-Neural Methods for Identification of Phytoplankton from Flow Cytomery Data". In: *Bioinformatics* 12.1 (1996), pp. 9–18. DOI: 10.1093/bioinformatics/12.1.9.

[417] M. Winder and J. E. Cloern. "The Annual Cycles of Phytoplankton Biomass". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 365.1555 (2010), pp. 3215–3226. DOI: 10.1098/rstb.2010.0125.

[418] R. S. Woodd-Walker, P. Ward, and A. Clarke. "Large-Scale Patterns in Diversity and Community Structure of Surface Water Copepods from the Atlantic Ocean". In: *Marine Ecology Progress Series* 236 (2002), pp. 189–203. DOI: 10.3354/meps236189.

[419] A. Z. Worden, M. J. Follows, S. J. Giovannoni, S. Wilken, A. E. Zimmerman, and P. J. Keeling. "Rethinking the Marine Carbon Cycle: Factoring in the Multifarious Lifestyles of Microbes". In: *Science* 347.6223 (2015), p. 1257594. DOI: 10.1126/science.1257594.

[420] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick. *Detectron2*. 2019.

[421] D. Yellowlees, T. A. V. Rees, and W. Leggat. "Metabolic Interactions between Algal Symbionts and Invertebrate Hosts". In: *Plant, Cell & Environment* 31.5 (2008), pp. 679–694. DOI: 10.1111/j.1365-3040.2008.01802.x.

[422] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. "How Transferable Are Features in Deep Neural Networks?" 2014.

[423] T. Yuasa and O. Takahashi. "Ultrastructural Morphology of the Reproductive Swarmers of Sphaerozoum Punctatum (Huxley) from the East China Sea". In: *European Journal of Protistology* 50.2 (2014), pp. 194–204. DOI: 10.1016/j.ejop.2013.12.001.

[424] M. J. Zaki, W. Meira Jr, and W. Meira. *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press, 2014. ISBN: 0-521-76633-8.

[425] D. Zawada. "Image Processing of Underwater Multispectral Imagery". In: *IEEE Journal of Oceanic Engineering* 28.4 (2003), pp. 583–594. DOI: 10.1109/JOE.2003.819157.

[426] T. Zebin, P. J. Scully, N. Peek, A. J. Casson, and K. B. Ozanyan. "Design and Implementation of a Convolutional Neural Network on an Edge Computing Smartphone for Human Activity Recognition". In: *IEEE Access* 7 (2019), pp. 133509–133520. DOI: 10.1109/ACCESS.2019.2941836.

[427] L. A. A. Zettler, O. R. Anderson, and D. A. Caron. "Towards a Molecular Phylogeny of Colonial Spumellarian Radiolaria". In: *Marine Micropaleontology* 36.2 (1999), pp. 67–79. DOI: 10.1016/S0377-8398(98)00028-0.

[428] Q. Zhao, Z. Basher, and M. J. Costello. "Mapping near Surface Global Marine Ecosystems through Cluster Analysis of Environmental Data". In: *Ecological Research* 35.2 (2020), pp. 327–342. DOI: 10.1111/1440-1703.12060.

[429] H. Zheng, R. Wang, Z. Yu, N. Wang, Z. Gu, and B. Zheng. "Automatic Plankton Image Classification Combining Multiple View Features via Multiple Kernel Learning". In: *BMC Bioinformatics* 18.16 (2017), p. 570. DOI: 10.1186/s12859-017-1954-8.

The End

This document was typeset with LaTeX using a modified version of the `classicthesis` theme developped by André Miede.

RÉSUMÉ

En tant que base des réseaux trophiques océaniques et élément clé de la pompe à carbone biologique, les organismes planctoniques jouent un rôle majeur dans les océans. Cependant, leur distribution à petite échelle, régie par les interactions biotiques entre organismes et les interactions avec les propriétés physico-chimiques des masses d'eau de leur environnement immédiat, est mal décrite *in situ*, en raison du manque d'outils d'observation adaptés. De nouveaux instruments d'imagerie *in situ* à haute résolution, combinés à des algorithmes d'apprentissage automatique pour traiter la grande quantité de données collectées, nous permettent aujourd'hui d'aborder ces échelles.

La première partie de ce travail se concentre sur le développement méthodologique de deux pipelines automatisés basés sur l'intelligence artificielle. Ces pipelines ont permis de détecter efficacement les organismes planctoniques au sein des images brutes, et de les classer en catégories taxonomiques ou morphologiques. Dans une deuxième partie, des outils d'écologie numérique ont été appliqués pour étudier la distribution du plancton à différentes échelles, en utilisant trois jeux de données d'imagerie *in situ*. Tout d'abord, nous avons mis en évidence un lien entre les communautés planctoniques et les conditions environnementales à l'échelle globale. Ensuite, nous avons décrit la distribution du plancton et des particules à travers un front de méso-échelle, et mis en évidence des périodes contrastées pendant le bloom de printemps.
Enfin, grâce aux données d'imagerie *in situ* à haute fréquence, nous avons étudié la distribution à fine échelle et la position préférentielle d'organismes appartement au groupe des Rhizaria, des protistes fragiles et peu étudiés, dont certains sont mixotrophes.

Dans l'ensemble, ce travail démontre l'efficacité de l'imagerie *in situ* combinée à des approches d'intelligence artificielle pour comprendre les interactions biophysiques dans le plancton et les conséquences sur sa distribution à petite échelle.