Aix*Marseille université
Socialement engagée

NNT/NL : 2020AIXM0001/001ED000

THÈSE DE DOCTORAT

Soutenue à Aix-Marseille Université
le 16 septembre 2022 par

Robin Fuchs

Méthodes neuronales et données mixtes : Vers une meilleure résolution spatio-temporelle des écosystèmes marins et du phytoplancton

Discipline
Mathématiques

Spécialité
Océanographie

École doctorale
ED 184 MATHEMATIQUES ET INFORMATIQUE

Laboratoires/Partenaires de recherche
Institut de Mathématiques de Marseille (I2M)
Institut Méditerranéen d'Océanologie (MIO)

Composition du jury

Christophe BIERNACKI University Lille 1 (France)	Rapporteur
Emilie POISSON-CAILLAULT Université du Littoral (France)	Rapporteure
Melika BAKLOUTI Institut Méditerranéen d'Océanologie (France)	Examinatrice
Pierre LATOUCHE Université de Paris (France)	Examinateur
Geoffrey MCLACHLAN Université du Queensland (Australie)	Examinateur
Mridul THOMAS Université de Genève (Suisse)	Examinateur
Denys POMMERET Institut de Mathématiques de Marseille (France)	Directeur de thèse
Melilotus THYSSEN Institut Méditerranéen d'Océanologie (France)	Codirectrice de thèse
Gérald GREGORI Institut Méditerranéen d'Océanologie (France)	Membre invité
Samuel SOUBEYRAND INRAE (France)	Membre invité

Affidavit

Je soussigné, Robin Fuchs, déclare par la présente que le travail présenté dans ce manuscrit est mon propre travail, réalisé sous la direction scientifique de Denys Pommeret et Melilotus Thyssen, dans le respect des principes d'honnêteté, d'intégrité et de responsabilité inhérents à la mission de recherche. Les travaux de recherche et la rédaction de ce manuscrit ont été réalisés dans le respect à la fois de la charte nationale de déontologie des métiers de la recherche et de la charte d'Aix-Marseille Université relative à la lutte contre le plagiat.

Ce travail n'a pas été précédemment soumis en France ou à l'étranger dans une version identique ou similaire à un organisme examinateur.

Fait à Marseille le 1er juin mai 2022

Liste de publications et participation aux conférences

Liste des publications réalisées dans le cadre du projet de thèse :

1. Fuchs R., Pommeret D., Viroli C., "Mixed Deep Gaussian Mixture Model : A clustering model for mixed datasets", Advances in Data Analysis and Classification, 2021 (publié)
2. Fuchs R., Pommeret D., Stocksieker S., "MIAMI : MIxed data Augmentation MIxture", 22nd International Conference on Computational Science and Its Applications, 2022 (publié)
3. Fuchs R. , Thyssen M., Creach V., Dugenne M., Izard L., Latimier M., Louchart A., Marrec P., Rijkeboer M., Grégori G., Pommeret D., "Automatic recognition of flow cytometric phytoplankton functional groups using Convolutional Neural Networks", Limnology and Oceanography : Methods, 2022 (publié)
4. Fuchs R., Baumas C.M.J., Garel M., Nerini D., Le Moigne F.A.C., Tamburini C., "A RUpture-Based detection method for the Active mesopeLagIc Zone (RUBALIZ) : a crucial step towards rigorous carbon budget assessments", Limnology and Oceanography : Methods, 2022 (soumis)
5. Fuchs R., Rossi V., Caille C., Bensoussan N., Pinazo C., Grosso O., Thyssen M., "Intermittent upwelling events trigger delayed, major, and reproducible pico-nanophytoplankton responses in coastal oligotrophic waters", Geophysical Research Letters, 2022 (soumis)
6. Barrillon S., Fuchs R., Petrenko A., Comby C., Bosse A., Yohia C., Fuda J.-L., Bhairy N., Berline L., Cyr F., Doglioli A., Grégori G., Tzortzis R., d'Ovidio F., and Thyssen M., "Intense storm in the north-western Mediterranean Sea strongly shaped local physics and generated significant phytoplankton reaction", Biogeosciences, 2022 (soumis)

Participation aux conférences et écoles d'été au cours de la période de thèse :

1. Fuchs R., Pommeret D., Viroli C., "MDGMM for mixed data clustering", Intervenant invité, King Abdullah University of Science and Technology, Arabie Saoudite, Février 2020

2. Fuchs R., Pommeret D., Viroli C. "Deep Gaussian Mixture Models for mixed type data", 51èmes Journées de Statistique de la Société Française de Statistique, Nice, France, Juin 2020
3. Fuchs R., Odic M., Rossi V., Bensoussan N., Caille C., Grosso O., Thyssen M., "Pico-nanophytoplankton dynamics in a coastal Mediterranean station : assessing hourly changes in community structure controlled by wind-driven events", ASLO 2021 Aquatic Sciences Meeting, Palma de Majorque, Espagne, Juin 2021
4. Thyssen M., Fuchs R., Creach V., Artigas L.F., Grégori G., Marrec P., Dugenne M., Rijkeboer M., Latimier M., Louchart A., "Standard vocabulary, consensual functional groups and automated classification for phytoplankton high throughput datasets using automated flow cytometry", ASLO 2021 Aquatic Sciences Meeting, Palma de Majorque, Espagne, Juin 2021
5. Fuchs R., Pommeret D., Stocksieker S., "MIAMI : MIxed data Augmentation MIxture", 52èmes Journées de Statistique de la Société Française de Statistique, Nice, France, Juin 2021
6. Fuchs R., Pommeret D., Viroli C., "Clustering mixed data with MDGMM", Intervenant invité, Université KU Leuven, Louvain, Belgique, Mars 2022
7. Fuchs R., Pommeret D., Stocksieker S., "MIAMI : MIxed data Augmentation MIxture", 22nd International Conference on Computational Science and Its Applications (ICCSA), Malaga, Espagne, Juillet 2022
8. Fuchs R., Pommeret D., Viroli C., "Mixed data : An integrated model to cluster and generate synthetic data", Intervenant invité, Conférence CMStatistics2022, King's College, Londres, Royaume-Uni, Décembre 2022

Résumé

Le phytoplancton constitue un des premiers maillons du réseau trophique et génère jusqu'à 50% de la production primaire mondiale. L'étude du phytoplancton et de son environnement physique nécessite des observations ayant une résolution inférieure à la journée et au kilomètre, ainsi que la prise en compte des types hétérogènes de données impliquées et des structures de dépendance spatio-temporelles des écosystèmes marins.

Cette thèse s'applique à développer des méthodes statistiques dans ce contexte en s'appuyant sur des technologies comme la cytométrie en flux automatisée. Les développements théoriques ont porté sur les modèles de mélanges gaussiens profonds (DGMM) introduits par Viroli et McLachlan (2019). Afin de mieux caractériser les niches écologiques du phytoplancton, nous avons étendu ces modèles aux données mixtes (présentant des variables continues et non continues) souvent présentes en océanographie. Une méthode de *clustering* a ainsi été proposée ainsi qu'un algorithme de génération de données mixtes synthétiques.

Concernant l'étude haute fréquence à proprement parler, des réseaux neuronaux convolutifs ont été introduits pour traiter les sorties de cytométrie en flux et étudier six groupes fonctionnels du phytoplancton en zone littorale et en océan ouvert. Des réactions différenciées et reproductibles de ces groupes ont été identifiées à la suite d'événements impulsionnels induits par le vent, soulignant l'importance du couplage entre la physique et la biologie. À cet égard, une méthode de détection de rupture a été proposée pour délimiter les zones épipélagique et mésopélagique, proposant ainsi une nouvelle base pour le calcul de budgets carbone mésopélagiques.

Mots clés : Modèles de mélange, Données synthétiques, Détection de rupture, Phytoplancton, Niches écologiques, Forçage physique

Abstract

Phytoplankton are one of the first links in the food web and generate up to 50% of the world's primary production. The study of phytoplankton and their physical environment requires observations with a resolution of less than a day and a kilometer, as well as the consideration of the heterogeneous types of data involved and the spatio-temporal dependency structures of marine ecosystems.

This thesis aims to develop statistical methods in this context by using technologies such as automated flow cytometry. Theoretical developments focused on Deep Gaussian Mixture Models (DGMM) introduced by Viroli and McLachlan (2019). To better characterize phytoplankton ecological niches, we extended these models to mixed data (exhibiting continuous and non-continuous variables) often found in oceanography. A clustering method has been proposed as well as an algorithm for generating synthetic mixed data.

Regarding the high-frequency study itself, convolutional neural networks have been introduced to process flow cytometry outputs and to study six functional groups of phytoplankton in the littoral zone and the open ocean. Differentiated and reproducible responses of these groups were identified following wind-induced pulse events, highlighting the importance of the coupling between physics and biology. In this regard, a change-point detection method has been proposed to delineate epipelagic and mesopelagic zones, providing a new basis for the calculation of mesopelagic carbon budgets.

Keywords: Mixture models, Data augmentation, Rupture detection, Phytoplankton, Ecological niches, Physical forcing

Remerciements

J'aimerais tout d'abord remercier mes encadrants, Denys Pommeret et Melilotus Thyssen pour leur engagement pendant ces trois années. Denys, je te remercie pour ta curiosité, ta volonté d'explorer des nouveaux horizons et méthodes, qui a donné naissance au MDGMM et MIAMI. La relation personnelle que nous avons développée a été importante pour moi pendant ces trois ans. J'ai pu compter sur ta bienveillance et ton écoute. Lotty merci d'avoir été un soutien indéfectible au quotidien et d'avoir toujours su te rendre disponible à n'importe quel moment. Ton encadrement, délaissant le cadre hiérarchique traditionnel, centré sur l'échange et la transmission, m'a permis de réellement progresser. J'ai eu le temps d'expérimenter, de me familiariser avec les thématiques océanographiques tout en étant très libre dans les méthodes employées. Ton intégrité, ton refus des messages scientifiques simplistes et vendeurs (pour lesquels les injonctions ne cessent de se multiplier), ainsi que ton humilité vis-à-vis de ce qui reste inconnu m'ont guidé et restent un réel exemple pour moi.

Je voudrais aussi remercier Gérald Grégori et Samuel Soubeyrand, qui constituent mes encadrants de l'ombre, mon "shadow cabinet" pour paraphraser la vie politique anglaise. Gérald, ton enthousiasme, ta vision des défis scientifiques futurs et nos discussions régulières m'ont enrichi au cours de ces trois années. Samuel, tes conseils méthodologiques sur le MDGMM et le présent manuscrit notamment ont réellement rendu ces travaux meilleurs. Un grand merci également aux membres de mon comité de suivi de thèse, Nicolas Chopin, Mathias Gauduchon et Pierre Pudlo.

Je remercie tous les gens que j'ai pu croiser au MIO qui m'ont beaucoup apporté au cours d'échanges plus ou moins formels. C'est particulièrement le cas de Caroline Lory, dont j'ai eu la chance de partager le bureau et avec laquelle j'ai fait ce tumultueux trajet de trois ans, partageant nos doutes et avancées. C'est aussi le cas de Chloé Baumas, pour sa bonne humeur et les projets que nous avons pu mener à bien ensemble, et de Marc Garrel avec qui je partage, entre autres, un réel amour de la programmation. Merci à Nolan Lezzoche pour son aide sur la mise en production du CNN et pour les sessions escalade (parfois à risque). Merci également au groupe natation du vendredi soir à Endoume. Ces sorties "in situ" ont réellement rythmé ma thèse et ont constitué chaque semaine une véritable respiration (sans mauvais jeu de mots). De manière générale, j'ai pu compter sur les conseils et commentaires très enrichissants de tous les coauteurs avec lesquels j'ai pu travailler.

Côté enseignement, j'adresse mes remerciements à Laurent Vigouroux pour l'autonomie qu'il m'a laissée en termes de pédagogie ainsi que pour ses retours sur

le ressenti des étudiants. Merci à Manuela Carenzi pour l'organisation parfaite des enseignements durant ces trois années, marquées par les enseignements présentiels/distanciels.

Je remercie également ma famille. Maman, merci de m'avoir donné la détermination, le goût du travail bien fait et de m'avoir forgé un mental en acier trempé. J'ai appris de toi que beaucoup d'objectifs deviennent réalisables à partir du moment où l'on se fixe un plan et que l'on s'y tient. Grâce à toi, j'ai toujours été libre de choisir mon chemin tout en restant attentif aux autres. Tu es la petite voix dans ma tête qui m'accompagne dans la plupart de mes décisions. Papa, tu m'as donné le goût de l'ambition, le besoin de me forger un avis sur les choses et d'être indépendant. Dino, je ne pouvais avoir plus de chances de tomber sur un frère comme toi, auquel je suis lié de tout mon être. Tu es empli de créativité, de passion et de facilités dans tout ce que tu entreprends. Je ne te remercie en revanche pas pour les défaites aux échecs que j'ai pu subir quasi-quotidiennement ces deux dernières années. Bibi, j'aime le regard singulier que tu as sur les choses, ton amour de la technique et le joueur incarné que tu es. Je suis très heureux de pouvoir aussi compter sur ma famille d'adoption, Christelle, Philippe, Maxence et Aurélie, pour leur soutien et leur joie de vivre. Vous avez toujours été là pour moi. Bien entendu, je pense fortement au reste de ma famille de Rognac et d'Alsace et à mon frère Sébastien.

J'ai la chance de pouvoir compter également sur vous mes amis. Simon dont j'aime la vivacité et l'enthousiasme permanent, Justine, élévée à Rock&Folk, dont j'aime la sensibilité et les convictions, Sarah pour son optimisme et sa capacité d'écoute infinie, Maxime pour son dévouement au produit (et bien entendu l'humour particulier qui va avec), Frankie pour la vérité de son amitié, ses préoccupations écologiques et bien sûr son goût inégalé pour le matériel sportif, Sacha pour lequel la nuit est toujours trop courte. Merci à Géo et Philippine pour leur amitié et ces sorties géniales à Marseille et dans les alentours. Merci à Matthieu Lagarde pour son ouverture d'esprit, sa simplicité et son intelligence hors du commun. Je suis heureux de pouvoir compter sur Adrien Campagne pour son amitié de longue date, son goût prononcé pour le rock et sa vision lucide des gens et de la société. J'ai aussi une très grosse pensée pour toi Anto, avec lequel j'aurais aimé avoir encore tant d'échanges et de discussions.

J'adresse le plus gros remerciement à Candice Roger. Depuis plus de onze ans, je nous vois évoluer ensemble. Tu as toujours été mon point d'ancrage sur terre, celle qui m'a empêché de rester enfermé dans le monde des idées. Les quêtes passionnantes, de la connaissance comme de toutes les choses pures, finissent par assécher l'âme et j'ai gardé pied grâce à toi. Je suis content que cette page se tourne pour en ouvrir d'autres avec toi.

Pour finir, je suis conscient de la chance qui m'a été offerte. Je suis conscient tout d'abord, outre le soutien affectif déjà mentionné, des ressources culturelles et financières familiales dont j'ai pu bénéficier. Je suis également conscient des ressources

publiques significatives qui m'ont été allouées, effet Matthieu oblige, au travers notamment de l'Ecole Normale Supérieure Paris-Saclay.
Cet investissement engage et ne donne pas le droit de se contenter de la facilité.

Entouré de tant d'amour et de conditions plus que favorables, il est impossible d'échouer.

Contents

List of Figures

List of Tables

Foreword

> Science is built up with facts, as a house is with stones. But a collection of facts is no more a science than a heap of stones is a house.
>
> *Henri Poincaré*

This Ph.D. thesis aimed to develop statistical methodologies to handle the complex datasets associated with the dynamic aspect of oceanic systems at different spatial and temporal levels. As a result, the contributions of the present work are twofold:

— Introducing mixed data clustering and data augmentation models to characterize marine environments and phytoplankton ecological niches;

— Designing proper statistical methods to create a sound study framework for high-frequency oceanographic phenomena.

These two points give its structure to the manuscript. The introduction (Chapter 1) presents the tackled oceanographic issues along with the general principles of the methods introduced to address them. The associated statistical details and complete literature review are given in the corresponding chapters to ease the reading.

Chapter 2 introduces the mixed data models, namely the MDGMM (Section 2.1) and MIAMI (Section 2.2), along with the RUBALIZ methodology (Section 2.3). The MDGMM was developed to characterize the ecological niches of the phytoplankton, i.e. identify the optimal environmental conditions for each phytoplankton group. MIAMI extended the MDGMM to simulate the effect of environmental changes in seawater temperature and nutrients on the phytoplankton. Finally, the RUBALIZ approach takes its roots in change-point methods and permitted the vertical separation between the epipelagic zone, hosting the phytoplankton populations, and the mesopelagic zone.

Chapter 3 lays emphasis on the high-frequency responses of phytoplankton functional groups to wind-induced events. First, the general approach is detailed and the systematic study of wind-induced events is motivated by the evidenced impact of a storm in the Ligurian Sea during the FUMSECK cruise (Section 3.1). Then, Section 3.2 introduces a convolutional neural network to track phytoplankton functional groups at high-frequency using automated flow cytometry. Finally, we provide a generalization of the phytoplankton responses observed during the FUMSECK cruise based on

twenty north-westerly events using the CNN and change-point detection methods (Section 3.3).

In general, this manuscript has been written with the concern to link the proposed statistical methods to the oceanographic questions they answer. The goal was to allow the readers of both fields of research to understand the general approach, avoiding as much as possible simplifications and omissions of the key concepts of both disciplines. Therefore in the sequel, an oceanographic context has been given to articles with a statistical focus in the dedicated chapters. Inversely, the statistical aspects that cannot be developed in articles with an oceanographic focus were either presented at the beginning of the chapters or in the appendix.

1. Introduction

I suppose it is tempting, if the only tool you have is a hammer, to treat everything as if it were a nail.

Abraham Maslow about the need for new statistical tools in oceanography

1. Identifying phytoplankton spatial distribution and ecological niches

Phytoplankton refer to a wide variety of unicellular autotrophic organisms, synthesizing their organic matter mainly from mineral elements, whose size varies from a few sub-micrometers to a few millimeters (see Figure 1.1). The phytoplankton perform photosynthesis using solar energy and dissolved CO_2. Their contribution to primary production, *i.e.* to the fixation of dissolved CO_2 by photosynthesis and per unit of time, is equivalent to all of the terrestrial primary production. A part of this so-fixed carbon then goes through the food web as phytoplankton cells are for instance grazed by zooplankton or lysed by viruses, and the remaining fraction sinks to the sediment in the deep ocean, a mechanism called biological carbon pump (BCP). The fraction of the primary production reaching the sediment is small but represents a significant total amount of carbon at the global ocean level. The biological carbon pump hence ensures a crucial role in slowing down the global warming process. The amount of carbon transferred to the sediment by the BCP is however subject to a high uncertainty depending on the estimation method used: 21 gigatons of carbon per year ($GtCyr^{-1}$) for the first estimates by Eppley et al. 1979, $12GtCyr^{-1}$ in the case of Laws et al. 2000 or more recently Henson et al. 2011 have estimated it to $5GtCyr^{-1}$. A part of this uncertainty derives directly from the extrapolation to the global ocean of relationships measured in a discrete number of locations (11 locations in the case of Laws et al. 2000, 306 in the case of Henson et al. 2011). This thus advocates for a clear understanding of phytoplankton adaptation to their environment and reliable estimation methods to resolve their local and high-frequency patterns.

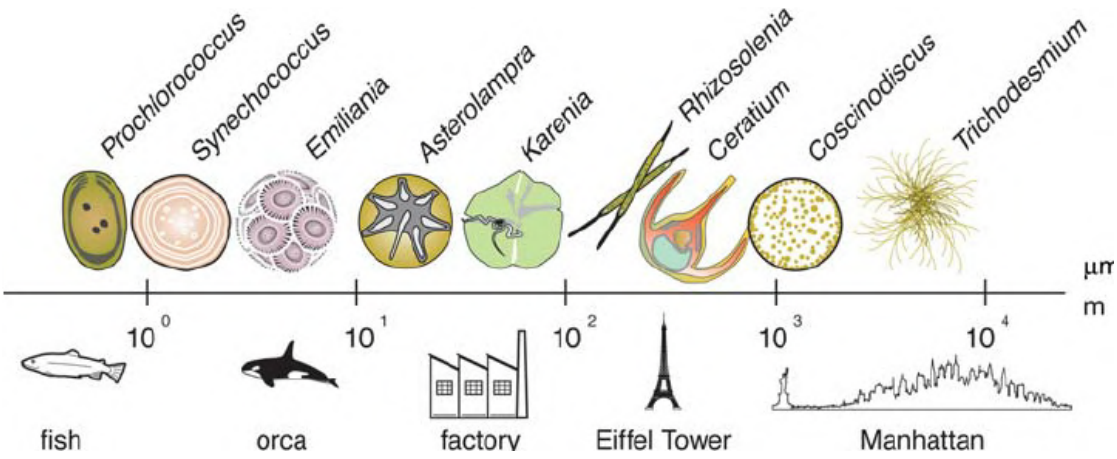


Figure 1.1. – Illustration of the phytoplankton size range in comparison to macroscopic objects (from Finkel et al. 2010).

1.1. Global oceanic circulation and biomes

Phytoplankton cells are mostly non-motile or lowly motile such as dinoflagellates or some coccolithophores (Ross et al. 2007) in a large spatial referential. Their global spatial dynamics are therefore strongly ruled by the oceanic general circulation under the control of light and temperature. The oceanic general circulation can be partitioned into a rapid surface circulation and a deeper slower circulation (also called thermohaline circulation) (Talley 2011). Surface oceanic circulation is mostly triggered by wind-driven processes coupled with the Coriolis force, i.e. the force created by the Earth rotation, as well as temperature and salinity diel variations. Deep ocean circulation is slower and mainly stems from differences in temperature and salinity of the water masses that generate density differentials and movements. These two circulations are not independent of each other and communicate in particular areas. For instance, in the Antarctic ocean, as the seawater gets colder and salter (due to the ice formation letting the remaining salt in the water), the surface seawater becomes denser and plunges deeper. Conversely, the Equator area sees deep waters coming back to the surface.

The oceanic general circulation along with other physical and atmospheric drivers such as the solar radiative flux creates connected but contrasted oceanic regions. There has been a long research tradition to characterize the continuous-in-nature oceanic environment into consistent regional provinces starting with the pioneering work of Somerville 1854. More recently, Longhurst 1995 have proposed a well-known partition of the oceans into four biomes based on the work of Yentsch et al. 1986 using satellite data. The four biomes have been delimited using the physical and biogeochemical characteristics of the water column such as the depth of the mixing

layer [1] or the nutricline [2] and are represented in Figure 1.2.

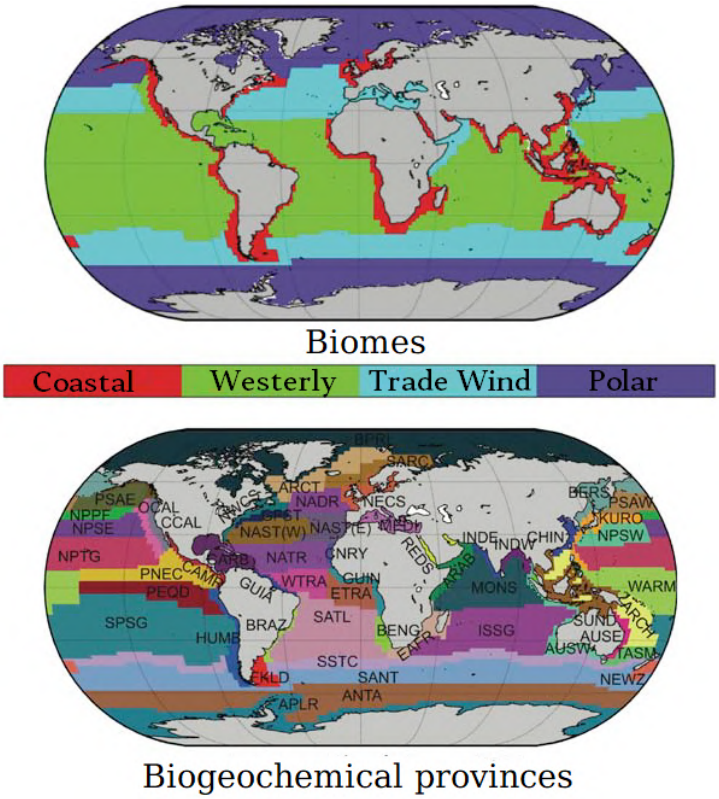


Figure 1.2. – Longhurst biomes and provinces from REYGONDEAU 2013.

The Polar zone, represented in purple in Figure 1.2, is very influenced by freshwaters and the fronts they are forming with saltwater. The Westerly zone is conspicuous for the high seasonal variability triggered by contrasted solar irradiance and wind patterns that fuel a significant phytoplankton bloom in spring. The Trade Wind zone on the contrary presents a low seasonal variability and the associated water column is very stratified (the vertical layers are well separated). The phytoplankton primary production occurring in this zone is very low. Finally, the coastal zone is characterized by shallow sea bottoms ($\leq 200m$) and is under the influence of regional physical processes such as upwellings during which surface coastal waters are replaced by offshore deep waters, colder and richer in nutrients. As evidenced in Figure 1.6, the data used in this work comes from the four Longhurst's biomes.

These four biomes are useful as they define a simple analytical framework but cannot account for the variability among each biome. As a result, they were later refined

1. Zone near the surface where the turbulences induced by the winds or differences in temperature from diel solar radiations have homogenized the temperature.

2. Zone of strong variation of nutrient concentrations often observed at the limit of the deep chlorophyll maximum, below the surface stratified layer.

by Longhurst 2010 into 56 provinces hosting consistent biological communities (Reygondeau et al. 2012).

Similarly to the fixed horizontal partition of the biomes and regions proposed by Longhurst 1995 and Longhurst 2010, an equivalent standard vertical partition exists. The water column is hence classically divided from surface to bottom between the epipelagic, the mesopelagic, the bathypelagic, abyssopelagic, and hadalpelagic zones (see Figure 1.3). The epipelagic zone, often assimilated to the euphotic zone, is traditionally located between 0 and 200m deep (Hedgpeth 1957). It is passed through by solar rays and hence contains photosynthetic organisms such as the phytoplankton. The mesopelagic layer is located traditionally between 200 and 1000m, hosts substantial fish resources, and plays a major role in the evoked biological carbon pump. Finally, the bathypelagic, abyssopelagic, and hadalpelagic zones are located between 1000m and 4000m, 4000 and 6000m and from 6000m to the seabed, respectively, and will not be studied as such in the sequel.

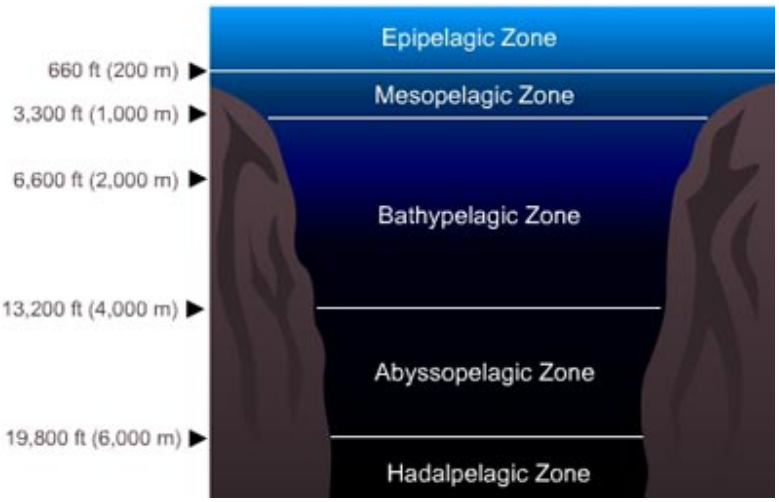


Figure 1.3. – Classical vertical oceanic layers boundaries (from the deep ocean facts website).

Yet, both of the presented standard vertical and horizontal partitions of the global ocean are static in nature and do not vary by location or per season. As a result, these static partitions are not able to resolve fine-scale dynamics, at the core of this study.

1.2. Revisiting classical oceanographic vertical boundaries with a dynamic perspective

This vertical partition of the ocean serves as a basis to perform carbon budgets. The particular organic carbon (POC) synthesized by photosynthetic primary producers

in the epipelagic zone sinks and is remineralized in the mesopelagic zone, mainly by heterotrophic organisms (Cho et al. 1988). The deeper the remineralization process occurs, the more sequestrated from the atmosphere the carbon will be (Kwon et al. 2009). The choice of the boundaries of the epipelagic and mesopelagic zones is therefore of prime importance for the construction of carbon budgets. In this respect, using the classical static boundaries leads to underestimating the impact of dynamic processes on local carbon budgets. Local boundary determination methods have thus to take into account the submesoscale features (~1-10km horizontal resolution) and dynamics of water masses to delimit proper boundaries.

The main historical criteria used to define water masses rely on temperature and salinity (Jacobsen 1927), and their resulting density. Temperature and salinity are often completed by the dioxygen, and the nutrients as in the Optimal Multi Parametric analysis (OMP) (Tomczak 1981; Tomczak et al. 1989). The OMP delimits the water masses assuming that separated water masses can mix, creating a front, according to a mixing process defined by a set of linear equations. This system linearity assumes the system to be of a closed nature, i.e. assumes no exchange of water or heat with the exterior of the system of interest. It also regards temperature and salinity as conservative variables (which undergo no modification within the system), but also O_2 and nutrients as conservative, which is a stronger hypothesis given the influence of the biology on these variables. Under these hypotheses, OMP determines the parameters of the mixture by ordinary least squares. OMP could be coupled with inverse box models (Wunsch 1996; Wunsch 2006) that directly model the geostrophic forces, wind forcing, and the resulting Ekman transport (wind-induced water transport) to integrate more regional insights into the analysis and get a good understanding of the water masses circulation and composition (Gasparin 2012).

Nevertheless, the OMP-related models consider the boundaries as the result of a dilution/water mixing process. As a result, they put physical drivers to the forefront and give less explanatory power to biogeochemical features. Moreover, OMP approaches strongly rely on the linearity assumptions and a mechanistic view of the processes. Alternatively, we propose statistical modeling giving a balanced weight to biochemical and physical drivers. The approach identifies the boundaries of the epipelagic and mesopelagic zones as ruptures in the signals of the variable characterizing the water masses. Based on the work by Truong et al. 2020, a kernelized mean-change model was applied to look for change points in the potential temperature, the salinity, the density, the fluorescence, and the dioxygen to determine the local boundaries of the epipelagic and mesopelagic layers. The approach was called RUBALIZ for "RUpture-Based detection method for the Active mesopeLagIc Zone". It is presented in section 2.3 preceded by a literature review of change-point detection methods.

1.3. Phytoplankton ecological niches

The classical horizontal notions of biomes and provinces, and standard vertical definitions of the epipelagic and mesopelagic zones encompass both physical, geochemical, and biological considerations but were not designed to resolve local patterns. Conversely, the evoked methods to determine local vertical boundaries have a finer spatial resolution. Yet, they consider the biological component as a whole and do not make differences between phytoplankton functional groups.

As a result, the concept of Hutchinsonian ecological niche (Hutchinson 1957) is maybe the most appropriate to describe both the physical and biological aspects of the studied local phenomena. More precisely, Hutchinson 1957 distinguished the fundamental niche from the realized niche. The fundamental niche covers all the conditions necessary for an organism or species to exist whereas the realized niche is the part of the fundamental niche that the organism/species can occupy due to the competition with other species. The competition between phytoplankton groups and the predation of the phytoplankton by the zooplankton was not studied as such here. We focused more on the impact of the water temperature, salinity, light, and nutrients on phytoplankton abundances (number of cells per unit of volume) and biomass. The determinants of the fundamental niche were hence explored considering the determinants of the realized niche as exogenous.

In our case, the datasets describing the ecological niches were tabular data of mixed nature, a widespread type of dataset in ecology. Tabular datasets are two-dimensional datasets presenting the observations (individuals, cell, sampling date) as rows and the variables for each observation as columns (temperature, salinity, cell size, etc.). Mixed datasets are datasets that contain continuous and non-continuous variables. Mixed variable types could be grouped into five main types:

— Categorical variables: That exhibit a finite and non-ordered number of modalities (e.g. the ocean name to which the sample belongs);
— Ordinal variables presenting a finite and ordered number of modalities (for example: the nutrient concentration defined as "high", "medium" or "low");
— Binary variables taking only two modalities (e.g: is it daytime or nighttime ?);
— Count variables having a finite [3] countable and ≥ 2 number of modalities (for instance: the number of samples collected per day);
— Continuous variables: Infinite and uncountable number of modalities (e.g. temperature, salinity, or dioxygen).

Given the heterogeneous nature of these datasets, the notion of similarity or distance between observations is particularly difficult to characterize for mixed data and requires the application of dedicated models. In this respect, the Mixed Deep Gaussian Mixture Model (MDGMM) was introduced. The MDGMM has a neural

3. Variables presenting an infinite but countable number of modalities were considered as continuous.

structure as made visible in Figure 1.4. It took advantage of the flexible parametric form of the Deep Gaussian Mixture Model (DGMM) (Viroli et al. 2019) and coupled it with Generalized Linear Latent Variable Models (GLLVM) (Moustaki et al. 2000; Moustaki 2003) to deal with mixed data. The formulation of these two models and of the MDGMM are given in section 2.1.2 along with a literature review of the main families of approaches handling mixed data.

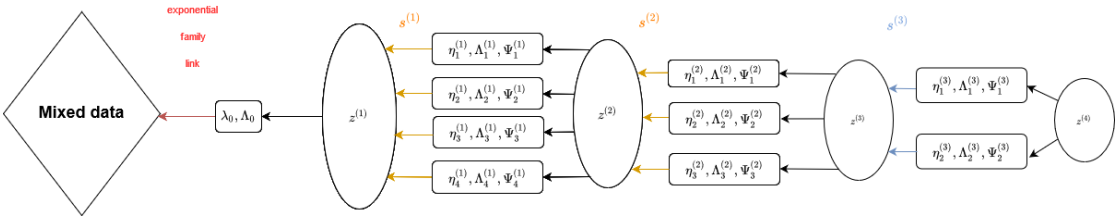


Figure 1.4. – Graphical model of a MDGMM.

2. Characterising high-frequency and submesoscale responses of phytoplankton functional groups

So far, a proper spatial and ecological framework has been set for the phytoplankton study. The focus is now turned on presenting the approaches resolving the local impact of sporadic events at high-temporal frequency, and more precisely the influence of short and intense wind events on the phytoplankton community. This class of events belongs to submesoscale phenomena which are of primary importance to the phytoplankton dynamics (Lévy et al. 2012). The infra-day frequency of the physical and chemical drivers (wind, drop in seawater temperature, nutrient pulses) and of the phytoplankton cell reactions (the cells can divide several times a day) necessitates dedicated hardware and treatment methods.

2.1. Flow Cytometry as high-frequency acquisition hardware

Flow cytometry (FC) is a high-throughput method that counts and characterizes the properties of single cells, such as the shape or the pigment content. This method has been introduced in the 1950s and 1960s, notably by Fulwyler 1965, and is used in disciplines that need to resolve the cell cycle such as cancerology, immunology, oceanography, and limnology. It was applied for the first time in marine research by Yentsch et al. 1983 and Olson et al. 1983. FC necessitates less manual treatment than microscopy or High-Performance Liquid Chromatography (HPLC). It is also more suited than satellite data in coastal areas and provides a finer phytoplankton community resolution, which has motivated the choice of this acquisition method.

The FC used here, a CytoSense manufactured by Cytobuoy (b.v.), is an automated flow cytometer, i.e. is equipped with automated and programmable acquisition protocols that simplify the data collection spanning long time periods. This FC resolves a large size range of phytoplankton groups with cell diameters between $0.5\mu m$ to more than $800\mu m$ in width and several millimeters in length. To do so, the CytoSense FC can take images of the biggest phytoplankton cells (not treated in this work) or collect a set of five curves per cell characterizing their shape and pigment content, called pulse shapes. These FC are hence called Automated pulse-shape recording Flow CytoMeters (AFCM). Traditionally, the cells collected by AFCM are manually gathered in groups sharing similar size and pigment characteristics called Phytoplankton Functional Groups (PFG), or cytometric Phytoplankton Functional Groups (cPFG) to distinguish them from proper functional groups as detailed for instance by Le Quere et al. 2005. This cell assignation process is called manual gating and is detailed in section 3.1.

2.2. A step further into Flow cytometry standardization

Nevertheless, the manual gating process lacks standardization concerning the group nomenclature used and the assignation process itself. AFCM is used in diversified environments by a significant community and the absence of standardization complicates a sound comparison of the results between studies. This inter-expert gating heterogeneity is however poorly documented in the literature with notable exceptions such as in Garcia et al. 2014. As such, a part of the work presented in section 3.2 is dedicated to the estimation of the magnitude of the manual gating bias.

This thesis hence aimed to contribute to the standardization of the functional groups recognition by AFCM. The possible improvements concern two aspects: providing a standardized cPFG nomenclature and standardizing the gating process.
The former was initiated by Thyssen et al. 2021, a group of more than 30 international experts in phytoplankton recognition by AFCM. The authors established that the groups observed worldwide can be gathered in 13 common cPFGs and that they present similar diffusion and fluorescence properties in contrasted world locations (http://vocab.nerc.ac.uk/collection/F02/current/).

We rely on the introduced nomenclature and focused on six of these groups ordered by growing average size and carbon content: Redpicopro, Orgpicopro, Redpicoeuk, Rednano, Orgnano, and Red/Orgmicro. These groups were often previously referred to as *Prochlorococcus*, *Synechococcus*, picoeukaryotes, nanoeukaryotes, cryptophytes, and microphytoplankton, respectively. Redpicopro are cyanobacteria belonging to the *Prochlorococcus* genus. They emit red auto-fluorescence when excited by a blue laser and no orange fluorescence as they do not contain phycoerythrin. On the contrary, Orgpicopro which are cyanobacteria belonging to the *Synechoccocus* genus, are rich in phycoerythrin and emit a strong orange fluorescence signal. Redpicoeuk are defined using a size criterion as they are phytoplankton cells smaller than $3\mu m$ in diameters. They are eukaryotes of polyphyletic origin and present a red fluorescence signal higher

than Redpicopro. Rednano are also eukaryotes characterized by a size between $3\mu m$ and $20\mu m$ that result in bigger diffusion signals than Redpicoeuk cells but also present a stronger red fluorescence due to their high blue-laser excited pigment content. Orgnano cells have a similar size range and red fluorescence as Rednano, but emit a higher orange fluorescence due to their significant phycocyanin and phycoerythrin content. They are constituted mainly of red algae, cyanobacteria, and cryptophytes. Finally, Redmicro and Orgmicro cells are bigger than $20\mu m$ and exhibit the strongest diffusion and fluorescence signals of the previously introduced cPFGs. They are mostly composed of dinoflagellates or diatoms and can be distinguished from Rednano and Orgnano by their higher FLR and FLO signals, respectively. In the presented works, Redmicro and Orgmicro were hardly distinguishable and gathered in a unique group: the Micro functional group.

Starting from this nomenclature, the second mentioned standardization aspect was addressed: proposing an automatic and reliable gating process. From a model standpoint, a convolutional neural network (CNN) as represented in Figure 1.5, initially designed to handle images, was introduced to deal with the pulse shape data. Using neural methods enabled not to manually design features from the pulse shapes, and to let the model identify the most group-discriminant pieces of information in the signal. From a data standpoint, pulse-shape data were treated by AFCM experts, and only consensual cells were kept to train the CNN to ensure a better learning process. More details concerning the data treatment, the models existing in the literature to deal with the pulse shape data, and the choice of the CNN architecture are given in section 3.2.

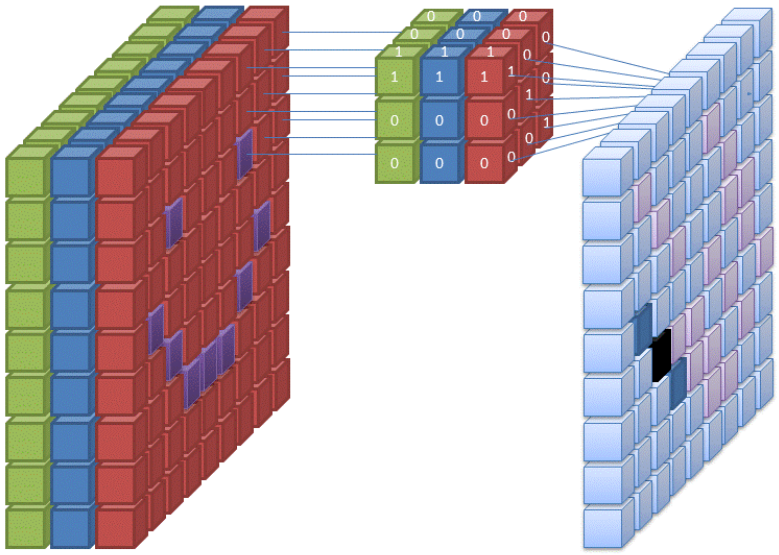


Figure 1.5. – Graphical representation of the convolutions performed by a CNN (under Wikimedia Commons licence).

2.3. Coupling physics and biology to resolve pulse events

Recently in oceanography, numerous studies have used FC to characterize the impact of infra-day frequency phenomena over the phytoplankton community such as Jacquet et al. 2002 and Sosik et al. 2003. Thyssen et al. 2008 and Dugenne et al. 2014 have used FC to study the impact of wind-induced events in the Mediterranean Sea and the Berre lagoon, respectively. Ribalet et al. 2015 have characterized the growth patterns occurring during the day and mortality during the night of *Prochlorococcus* (belonging to the Redpicopro group) in the subtropical Pacific Ocean. Hunter-Cevera et al. 2020 have shown in the North Atlantic Ocean that the phytoplankton spring bloom of *Synechococcus* (belonging to the Orgpicopro group) was due mainly to a change in their daily growth rate.

The works presented in Chapter 3 hence belong to this literature and analyze the impact of submesoscale wind-induced events on phytoplankton functional groups thanks to the introduced common nomenclature, CNN automatic recognition, and rupture detection methods. The unique features of each cPFG in terms of carbon content or nycthemeral cycle suggest differentiated responses between groups and motivate this study. The local zone of interest in this chapter is the Mediterranean sea. Indeed, the Mediterranean Sea can be regarded as a "hotspot" for climate change

(Group et al. 2011) due to the short residence time of the water masses, the variable meteorological conditions, and the strong anthropogenic pressure it undergoes.

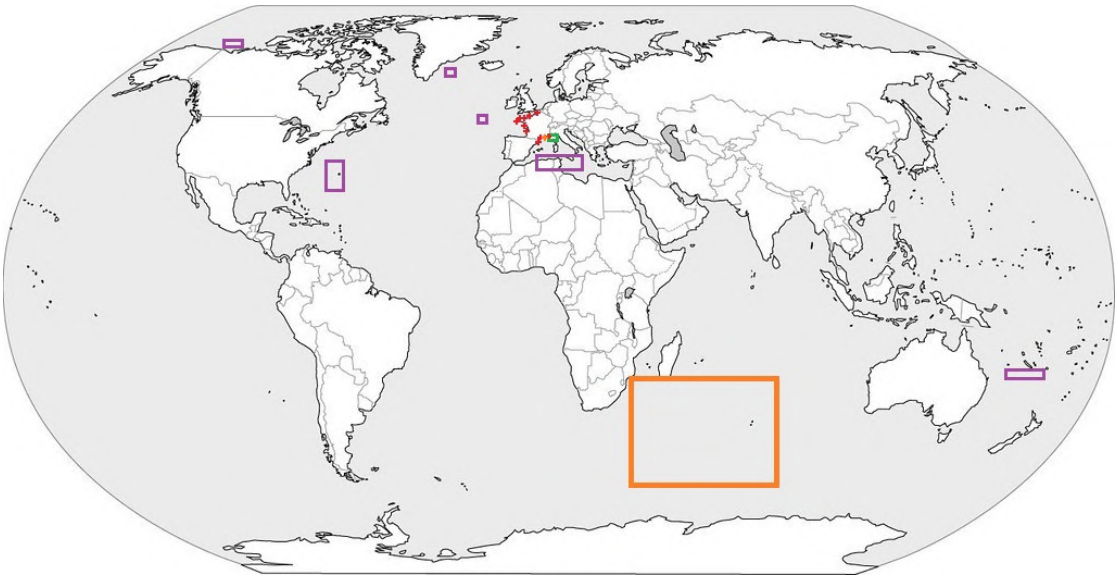


Figure 1.6. – Data sampling zones of the works presented here. Samples acquired during cruises are represented by rectangles and fixed-point data collections by crosses. In Chapter 2, ecological niches were determined on data denoted with red crosses (SOMLIT data), and vertical epipelagic boundaries on purple rectangle located data. In Chapter 3, evidence of the phytoplankton functional group response was highlighted using the data represented by a green rectangle (FUMSECK data). The convolutional neural network was trained with the orange data (SSL@MM station and GEOTRACES SWINGS cruise). Finally, generic reproducible pico-nanophytoplankton group responses to wind-induced upwelling events were determined at the SSL@MM station (orange cross).

2. Unraveling phytoplankton ecological niches and vertical spatial boundaries

Sommaire

> Nothing is lost, nothing is created,
> everything can be seen as a mixture
>
> *(Adapted from) Antoine Lavoisier concerning the MDGMM and water masses*

This chapter presents the Mixed Deep Gaussian Mixture Model (MDGMM) and its extension, the MIxed data Augmentation MIxture (MIAMI), which constitute the main theoretical contributions of this Ph.D. thesis. The MDGMM belongs to mixed data clustering models and was used to characterize phytoplankton ecological niches. The MIAMI model was applied to the same data to prospect the effect of environmental shifts.

1. Clustering ecological niches using Mixed Deep Gaussian Mixture Models

The goal of the clustering task is to create a meaningful partition that divides the observations into K groups named clusters. The observations gathered in the same group should present similar features contrary to observations belonging to different clusters. As evoked earlier, the notion of similarity in the mixed data case is made harder by the heterogeneous nature of the variables. To address this issue, Ahmad et al. 2019 distinguished between five families of clustering algorithms in the mixed data case: Partitional models, hierarchical models, model-based models, neural-networks models, and other models, shortly presented hereafter.

1.1. The MDGMM: A neural and model-based approach

The partitional algorithms have received more attention in the literature and rely on three main ingredients: defining a center for each cluster, a distance that encapsulates the evoked notion of similarity, and a cost function to minimize. The most reknown models of this family are derived from the k-means model, such as the k-modes or k-Prototypes modes (Huang 1997; Huang 1998). These algorithms have the advantage to exhibit a $o(n)$ complexity, with n the number of observations, which makes them particularly suited for big datasets. Yet, the initialization of the clusters and the choice of the total number of clusters constitute the main limitations of this family of models.

The second family of mixed data clustering models is the family of hierarchical models (e.g. Philip et al. 1983). These models are built upon two main principles: a similarity criterion that determines the pairwise level of similarity between two observations and a linkage criterion, that determines the distance between two clusters. The clustering process is performed either by creating numerous clusters and iteratively merging them or conversely by iteratively splitting the observations until the desired

number of clusters is reached. The resulting clustering is then easily interpretable using the suite of iterative splits or merges performed during the training, which could be visualized as a phylogenetic tree. The main limit of these models is their complexity in time ($o(n^3)$) and memory ($o(n^2)$), which make them designed to handle only small to medium-size datasets.

The third family of models is composed of model-based clustering methods. As their name suggests it, these methods are constructed over a properly defined statistical framework generally involving statistical distributions, latent spaces and mixture models (e.g. Browne et al. 2012; McParland et al. 2016). These models are often trained by likelihood maximization using algorithms based on the Expectation-Maximisation algorithm. These models remain interpretable but present relatively long training times in general.

The approaches relying on neural networks constitute the fourth class of methods in the Ahmad et al. 2019 typology. These models are structured as layers of neurons and their high-dimensional parametric space can capture complex patterns observed in the data. The major developments have concerned Self-Organizing Maps (SOM) (Kohonen 1990) and Adaptive Resonance Theory networks (Carpenter et al. 2010). The handling of categorical and ordinal data by these models is limited and they are thus often encoded as binary variables before performing the clustering.

Finally, the last family of models is composed of the models that did not fit in the first four families. It includes for instance spectral clustering (Ng et al. 2001; David et al. 2012), density-based clustering (Ester et al. 1996) or tree-ensemble methods (Lin et al. 2018).

The MDGMM belongs to the model-based and neural-network-based families. Indeed, it is built up upon the Generalized Linear Latent Variable Model (GLLVM) (Moustaki et al. 2000; Moustaki 2003) that belongs to model-based approaches, and the Deep Gaussian Mixture Model (DGMM) (Viroli et al. 2019) that also belongs to model-based approaches but can also be regarded as neural-based methods. As a result, it is able to capture complex patterns in the data (thanks to its neural structure) but also provides an interpretation of the clustering and latent representations learned (thanks to its model-based construction).

1.2. The MDGMM as a generalization model

To properly introduce the MDGMM, a brief definition of the DGMM and GLLVM is given before providing the MDGMM derivations and results in the associated paper.

Let $y = (y^C, y^D)$ denote the data of dimension $n \times p$, with n the number of observations and p the number of variables, y^C the continuous data and y^D the non-continuous data. We define i as the observation index such that $i \in [1, n]$ and j the

variable index such that $j \in [1, p]$.

Presentation of the DGMM

The DGMM was introduced by Viroli et al. 2019 to perform clustering on fully continuous data. The DGMM can be viewed as a generalization of Factor Models (FA), Mixtures of Factor Analyzers (MFA), and Gaussian Mixtures.
Factor models were first presented in Harman 1976 to compress the signal contained in the original data in a latent space of a much lower dimension (r). Referring to the already introduced notations, the continuous data y^C of dimension $n \times p_C$ are compressed into a Gaussian latent space of dimension r with $p_C >> r$.

$$y_i^C = \eta + \Lambda z_i + u_i,$$

with η a constant vector of size p_C, $z_i \sim \mathcal{N}(0, I_r)$, $u_i \sim \mathcal{N}(0, \Psi)$ $\forall i \in [1, n]$, and Λ the factor loading matrix of dimension $p_C \times r$. The loading matrix is then used to interpret the relationship existing between the data and their new representation.

This model can be extended assuming that different groups of observations in the data have different latent representations, the model then becomes a Mixture of Factor Analyzers (Ghahramani et al. 1996).

$$y_i^C = \eta_{k_1} + \Lambda_{k_1} z_i + u_{ik_1} \text{ with probability } \pi_{k_1},$$

for $k_1 \in [1, K_1]$ and with $z_i \sim \mathcal{N}(0, I_p)$ and $p > r_1$.

It is possible to extend once again the MFA model by assuming that z_i is no more drawn from a multivariate Gaussian but is itself a MFA. The corresponding model is a two hidden layers DGMM (Viroli et al. 2019):

$$\begin{cases} y_i^C = \eta_{k_1}^{(1)} + \Lambda_{k_1}^{(1)} z_i^{(1)} + u_{ik_1}^{(1)} \text{ with probability } \pi_{k_1}^{(1)} \\ z_i^{(1)} = \eta_{k_2}^{(2)} + \Lambda_{k_2}^{(2)} z_i^{(2)} + u_{ik2}^{(2)} \text{ with probability } \pi_{k_2}^{(2)}, \end{cases} \tag{2.1}$$

with $z_i^{(2)} \sim \mathcal{N}(0, I_{r_2})$, $k_0 \in \{1\}$, $k_1 \in [1, K_1]$ et $k_2 \in [1, K_2]$ and $p > r_1 > r_2$.

Deeper DGMMs can be defined by rewriting iteratively the last latent variable as a MFA. Doing so, one ends up with the following L-layers deep DGMM:

$$\begin{cases} y_i^C = \eta_{k_1}^{(1)} + \Lambda_{k_1}^{(1)} z_i^{(1)} + u_{ik_1}^{(1)} \text{ with probability } \pi_{k_1}^{(1)} \\ z_i^{(1)} = \eta_{k_2}^{(2)} + \Lambda_{k_2}^{(2)} z_i^{(2)} + u_{ik_2}^{(2)} \text{ with probability } \pi_{k_2}^{(2)} \\ \ldots \\ z_i^{(L-1)} = \eta_{k_L}^{(L)} + \Lambda_{k_L}^{(L)} z_i^{(L)} + u_{ik_L}^{(L)} \text{ with probability } \pi_{k_L}^{(L)} \\ z_i^{(L)} \sim \mathcal{N}(0, I_{r_L}), \end{cases} \tag{2.2}$$

for $k_1 \in [1, K_1]$,..., $k_L \in [1, K_L]$, and $p > r_1 > r_2 > ... > r_L$.

However, the DGMM can only deal with continuous data. To apply a DGMM to discrete datasets, one has first to find a continuous representation of mixed data. To do so, we have integrated the Generalized Linear Latent Variable Model (GLLVM) framework within the DGMM framework.

Presentation of the GLLVM

Generalized Linear Latent Variable Models were introduced by Moustaki et al. 2000 and Moustaki 2003 and can deal with mixed data. They assume that the data in the original variable space can be projected into a lower-dimensional latent space that is assumed to be Gaussian. The fundamental assumption of such models, called the "conditional independence assumption", states that the variables are mutually independent conditionally to the latent variable. In other words, the latent variable is assumed to account for all the dependence structure between the original variables. The original variables are linked to the latent space using link functions that depend on the variable type and belong to an exponential family.

More formally, $\forall j \in [1, p]$, the variables $y_j \in \mathbb{R}^n$ are mutually independent with respect to the latent variables still denoted by $z^{(1)}$ (as in Equations 2.1 and 2.2). The function linking each original variable to the latent variable has the following form:

$$f(y_j|z^{(1)}) = \exp\Big(\frac{y_j\theta_j - b_j(\theta_j)}{\phi_j} + c_j(y_j, \phi_j)\Big),$$

with θ_j, ϕ_j, c_j coefficients to estimate that indirectly depend on $z^{(1)}$. If y_j is a binary variable, $f(y_j|z^{(1)})$ could for example be specified to be a Bernoulli distribution. Similarly, if y_j is a categorical variable, one could specify $f(y_j|z^{(1)})$ to be a multinomial distribution. For a continuous variables, a Gaussian or a Gamma distribution could for instance be used.

To summarize, the MDGMM uses the GLLVM to plunge the mixed data into a continuous latent space modeled as a DGMM. The clustering process and the learning of the best parameters for the latent space are performed jointly.

Introducing the MDGMM

The following study introduces the Mixed Deep Gaussian Mixture Models. It dwells on the mathematical aspects of this class of models, presents the training process, a new initialization strategy, and an automatic architecture selection procedure. The performance of the MDGMM is compared to the other model families mentioned by Ahmad et al. 2019. Additional details are provided in Appendix A.

Mixed Deep Gaussian Mixture Model: A clustering model for mixed datasets

Robin Fuchs*
CNRS, Centrale Marseille, I2M, MIO, Aix-Marseille Univ.
and
Denys Pommeret
Univ Lyon, UCBL, ISFA LSAF EA2429
and
Cinzia Viroli
Department of Statistical Sciences, Univ. of Bologna.

March 11, 2021

Abstract

Clustering mixed data presents numerous challenges inherent to the very heterogeneous nature of the variables. A clustering algorithm should be able, despite of this heterogeneity, to extract discriminant pieces of information from the variables in order to design groups. In this work we introduce a multilayer architecture model-based clustering method called Mixed Deep Gaussian Mixture Model (MDGMM) that can be viewed as an automatic way to merge the clustering performed separately on continuous and non-continuous data. This architecture is flexible and can be adapted to mixed as well as to continuous or non-continuous data. In this sense we generalize Generalized Linear Latent Variable Models and Deep Gaussian Mixture Models. We also design a new initialisation strategy and a data driven method that selects the best specification of the model and the optimal number of clusters for a given dataset "on the fly". Besides, our model provides continuous low-dimensional representations of the data which can be a useful tool to visualize mixed datasets. Finally, we validate the performance of our approach comparing its results with state-of-the-art mixed data clustering models over several commonly used datasets.

Keywords: Binary and count data; Deep Gaussian Mixture Model; Generalized Linear Latent Variable Model; MCEM algorithm; Ordinal and categorical data; Two-heads architecture.

This preprint has not undergone any post-submission improvements or corrections.
The Version of Record of this article is published in Advances in Data Analysis and Classification, and is available online at https://doi.org/10.1007/s11634-021-00466-3

*robin.fuchs@univ-amu.fr

1 Introduction

Mixed data consist of variables of heterogeneous nature that can be divided into two categories: the continuous data generated by real-valued random variables, and the non-continuous data which are composed of categorical and ordinal data (non-ordered or ordered data taking a finite number of modalities), binary data (that take either the value 1 or the value 0), and count data (taking values in $\mathbb{N}$). By language abuse, these non-continuous variables will also be referred to as discrete variables in the following.

Due to their different natures, mixed variables do not share common scales and it is typically hard to measure the similarity between observations. There has been a significant and long interest in the statistical literature for mixed and continuous data clustering, which can be framed into four main categories, as described in Ahmad and Khan (2019): (i) partitional clustering minimizes the distance between observations and center groups by iterative optimization, as in K-modes or K-prototypes (Huang, 1997, 1998); (ii) hierarchical algorithms perform nested clusterings and merge them to create the final clustering (Philip and Ottaway, 1983; Chiu et al., 2001); (iii) model-based clustering (McLachlan and Peel, 2000; Fraley and Raftery, 2002; Melnykov et al., 2010), as their name suggests, rely on probability distributions; (iv) finally Neural Networks-based algorithms (Kohonen, 1990) design the clusters thanks to connected neurons that learn complex patterns contained in the data.

Within the framework of model-based clustering we propose a model for clustering mixed data, in which the different non-continuous variables are merged via a Generalized Linear Latent Variable Model (GLLVM) (Moustaki, 2003; Moustaki and Knott, 2000). GLLVMs assume that there exists a link function between the non-continuous observable space (composed of non-continuous variables) and a latent continuous data space, consisting of Gaussian latent variables. Recently, Cagnone and Viroli (2014) have extended this approach by considering latent variables that are no more Gaussian but follow some mixtures of Gaussians (Fraley and Raftery, 2002) so as the observations are naturally clustered into groups. Since the latent dimension is chosen to be strictly lower than the original dimension, the model also performs dimension reduction. By abuse of language, we will refer

to this extended version when mentioning GLLVMs in the sequel.

Our work generalizes this idea by considering a Deep Gaussian Mixture Model (DGMM) in the latent space (see Viroli and McLachlan, 2019). DGMMs can be seen as a series of nested Mixture of Factor Analyzers (MFA) (Ghahramani et al., 1996; McLachlan et al., 2003). As such, this approach performs clustering via subsequent dimensionally reduced latent spaces in a very flexible way.

To adapt the GLLVM to mixed data we propose a multilayer architecture inspired by the idea that composing simple functions enables to capture complex patterns, as in supervised neural networks. We design two versions of our model. In the first one, denoted by M1DGMM, continuous and non-continuous data goes through the GLLVM model which acts as an embedding layer. The signal is then propagated to the following layers. In the second version, called M2DGMM, discrete data are still handled by the GLLVM model but continuous data are embedded separately by a DGMM head. The two signals are then merged by a "common tail". This second architecture is analogous to multi-inputs Supervised Deep Learning architectures used for instance when data are composed of both images and text.

Our model implementation relies on automatic differentiation (Baydin et al., 2017) that helps keeping an acceptable running time even when the number of layers increases. Indeed, using auto-differentiation methods provided for instance by the autograd package (Maclaurin et al., 2015) cuts the computational running time. For instance, for the special case of GLLVM models, Niku et al. (2019) reported significant computational gains from using auto-differentiation methods.

To summarize, our work has three main aims: it first extends the GLLVM and DGMM frameworks to deal with mixed data. Secondly, a new initialisation method is proposed to provide a suitable starting point for the MDGMM and more generally for GLLVM-based models. This initialization step combines Multiple Correspondence Analysis (MCA) or Factor analysis of mixed data (FAMD) which generalizes it, GMM, MFA and the Partial Least Squares (PLS) algorithm. As mixed data are plunged into a multilayer continuous space we call this new initialisation Nested Spaces Embedding Procedure (NSEP). Thirdly,

a model selection procedure is designed to identify the architecture of the model that best fits a given dataset.

Since the models are quite complex we propose to develop the method within the article and to reduce some mathematical developments by reporting them in a Supplementary Materials.

The paper is organized as follows: Section 2 provides a detailed description of the proposed model. In Section 3 the EM algorithms used for estimation are developed. Section 4 deals with the identifiability constraints of the model. Section 5 presents the initialization procedure NSEP and some practical considerations are given that can serve as a user manual. The performance of the model is compared to other competitor models in Section 6. In conclusion, Section 7 analyses the contributions of this work and highlights directions for future research.

2 Model presentation

2.1 The MDGMM as a generalization of existing models

In the sequel we assume that we observe n random variables $y_1, \cdots, y_n$, such that $\forall i = 1, \cdots, n$, $y_i = (y_i^C, y_i^D)$, where y_i^C is a p_C-dimensional vector of continuous random variables and y_i^D is a p_D-dimensional vector of non-continuous random variables. From what precedes, each y_i is hence a vector of mixed variables of dimension $p = p_C + p_D$.

The architecture of the MDGMM is based on two models. First, Mixtures of Factor Analyzers generalized by the Deep Gaussian Mixture Models are applied on continuous variables, and second, a Generalized Linear Latent Variable Model coupled with a DGMM is applied on non-continuous variables. Mixtures of Factor Analyzers represent the most elementary building block of our model and can be formulated as follows:

$$y_i^C = \eta_k + \Lambda_k z_i + u_{ik}, \text{ with probability } \pi_k,$$

where $k \in [1, K]$ identifies the group, η_k is a constant vector of dimension p_C, $z_i \sim \mathrm{N}(0, I_r)$, $u_{ik} \sim \mathrm{N}(0, \Psi_k)$ and Λ_k is the factor loading matrix of dimension $p_C \times r$, r being the

dimension of the latent space. The underlying idea is to find a latent representation of the data of lower dimension r, with $r < p_C$. For each group k, the loading matrix is then used to interpret the relationship existing between the data and their new representation.

The DGMM approach consists in extending the MFA model by assuming that z_i is no more drawn from a multivariate Gaussian but is itself a MFA. By repeating L times this hypothesis we obtain a L-layers DGMM defined by:

$$\begin{cases} y_i^C = \eta_{k_1}^{(1)} + \Lambda_{k_1}^{(1)} z_i^{(1)} + u_{ik_1}^{(1)}, \text{ with probability } \pi_{i,k_1}^{(1)} \\ z_i^{(1)} = \eta_{k_2}^{(2)} + \Lambda_{k_2}^{(2)} z_i^{(2)} + u_{ik_2}^{(2)}, \text{ with probability } \pi_{i,k_2}^{(2)} \\ \dots \\ z_i^{(L-1)} = \eta_{k_L}^{(L)} + \Lambda_{k_L}^{(L)} z_i^{(L)} + u_{ik_L}^{(L)}, \text{ with probability } \pi_{i,k_L}^{(L)} \\ z_i^{(L)} \sim \mathcal{N}(0, I_{r_L}), \end{cases} \tag{1}$$

where, for $\ell = 1, \cdots, L$, $k_\ell \in [1, K_\ell]$, $u_{ik_\ell}^{(\ell)} \sim N(0, \Psi_{k_\ell}^{(\ell)})$, $z_i^{(L)} \sim \mathrm{N}(0, I_{r_L})$ and where the factor loading matrices $\Lambda_{k_\ell}^{(\ell)}$ have dimension $r_{\ell-1} \times r_\ell$, with the constraint $p > r_1 > r_2 > ... > r_L$. Identifiability constraints on the parameters $\Lambda_{k_\ell}^{(\ell)}$ and $\Psi_{k_\ell}^{(\ell)}$ will be discussed in Section 4.

The DGMM described in (1) can only handle continuous data. In order to apply a DGMM to discrete data we propose to integrate a Generalized Linear Latent Variable Model (GLLVM) framework within (1). This new integrated model will be called Discrete DGMM (DDGMM).

A GLLVM assumes that, $\forall j \in [1, p_D]$, the discrete random variables y_j^D are associated to one (or more) continuous latent variable through an exponential family link (see the illustrations given in Cagnone and Viroli (2014)), under the so-called *conditional independence assumption*, according to which variables are mutually independent conditionally to the latent variables.

Hence, one can combine the previously introduced DGMM architecture and the GLLVM to deal with mixed data. In order to do so, we propose two specifications of the MDGMM: a one head version (the M1DGMM) and a two heads version (the M2DGMM). In the M1DGMM, the continuous variables pass through the GLLVM layer by defining a link function between y^C and $z^{(1)}$ and one assumes that the *conditional independence assump-*

tion evoked earlier holds. On the contrary, by specifying a second head to deal with the continuous data specifically, one can relax this assumption: the continuous variables are not assumed to be mutually independent with respect to the latent variables. Instead, each continuous variable is only conditionally independent from the discrete variables but not from the other continuous variables. The two-heads architecture is also more flexible than the one-head specification as its "link function" between y^C and z^C is a mixture of mixture rather than a probability distribution belonging to an exponential family. This flexibility comes at the price of additional model complexity and computational costs which has to be evaluated in regard of the performances of each specification.

The intuition behind the M2DGMM is simple. The two heads extract features from the data and pass them to the common tail. The tail reconciles both information sources, designs common features and performs the clustering. As such, any layer on the tail could in principle be used as clustering layer. As detailed in Section 5.2, one could even use several tail layers to perform several clustering procedures (with different latent dimensions or numbers of clusters) in the same model run. The same remarks applies for the hidden layers of the M1DGMM.

To summarize the different setups that can be handled by DGMM-based models:

- Use the M1DGMM or the M2DGMM when data are mixed,
- Use the DDGMM when data are non-continuous,
- Use the DGMM when data are continuous.

2.2 Formal definition

Let y be the $n \times p$ matrix of the observed variables. We will denote by $i \in [1, n]$ the observation index and by $j \in [1, p]$ the variable index. We can decompose the data as $y = (y^C, y^D)$ where y^C is the $n \times p_C$ matrix of continuous variables and y^D is the $n \times p_D$ matrix of discrete variables. The global architecture of the M2DGMM is analogous to (1)

with an additional GLLVM step for the discrete head as follows:

$$
\begin{aligned}
&\text{Discrete head}: \begin{cases} y_i^D \to z_i^{(1)D} \text{ through GLLVM link via } (\lambda^{(0)}, \Lambda^{(0)}) \\ z_i^{(1)D} = \eta_{k_1}^{(1)D} + \Lambda_{k_1}^{(1)D} z_i^{(2)D} + u_{i,k_1}^{(1)D} \text{ with probability } \pi_{i,k_1}^{(1)D} \\ \ldots \\ z_i^{(L_D)D} = \eta_{k_{L_D}}^{(L_D)D} + \Lambda_{k_{L_D}}^{(L_D)D} z_i^{(L_D+1)} + u_{k_{L_D}}^{(L_D)D}, \text{ with probability } \pi_{i,k_{L_D}}^{(L_D)D} \end{cases} \\
&\text{Continuous head}: \begin{cases} y_i^C = \eta_{k_1}^{(1)C} + \Lambda_{k_1}^{(1)C} z_i^{(1)C} + u_{i,k_1}^{(1)C} \text{ with probability } \pi_{i,k_1}^{(1)C} \\ z_i^{(1)C} = \eta_{k_1}^{(1)C} + \Lambda_{k_1}^{(1)C} z_i^{(2)C} + u_{i,k_1}^{(1)C} \text{ with probability } \pi_{i,k_2}^{(2)C} \\ \ldots \\ z_i^{(L_C)C} = \eta_{k_{L_C}}^{(L_C)C} + \Lambda_{k_{L_C}}^{(L_C)C} z_i^{(L_C+1)} + u_{k_{L_C}}^{(L_C)}, \text{ with probability } \pi_{i,k_{L_C+1}}^{(L_C+1)C} \end{cases} \\
&\text{Common tail}: \begin{cases} z_i^{(L_0+1)} = \eta_{k_{L_0+1}}^{(L_0+1)} + \Lambda_{k_{L_0+1}}^{(L_0+1)} z_i^{(L_0+2)} + u_{k_{L_0+1}}^{(L_0+1)}, \text{ with probability } \pi_{i,k_{L_0+2}}^{(L_0+1)} \\ \ldots \\ z_i^{(L-1)} = \eta_{k_{L-1}}^{(L-1)} + \Lambda_{k_{L-1}}^{(L-1)} z_i^{(L)} + u_{k_{L-1}}^{(L-1)} \text{ with probability } \pi_{i,k_L}^{(L-1)} \\ z_i^{(L)} \sim \mathcal{N}(0, I_{r_L}). \end{cases}
\end{aligned}
\tag{2}
$$

The architecture of the M1DGMM is the same except that there is no "continuous head" and that the y_i^C goes through the GLLVM link. Figure 1 presents the graphical models associated with both specifications. In the M2DGMM case one can observe that $L_0 = \max(L_C, L_D)$, that is, the first layer of the common tail is the L_0+1-th layers of the model. For simplicity of notation, we assume in the sequel that

$$L_C = L_D = L_0,$$

but all the results are easily obtained in the general case. It is assumed that the random variables $(u_{k_\ell}^{(\ell)})_{k_\ell,\ell}$ are all independent. The two heads only differ from each other by the fact that for the discrete head, a continuous representation of the data has first to be determined before information is fed through the layers. The GLLVM layer is parametrized by (λ_0, Λ_0). $\lambda_0 = (\lambda_{0bin}, \lambda_{0count}, \lambda_{0ord}, \lambda_{0categ})$ contains the intercept coefficients for each discrete data

sub-type. Λ_0 is a matrix of size $p_D \times r_1$, with r_1 the dimension of the first Discrete DGMM layer.

The notation remains the same as in the previous subsection and only a superscript is added to specify for each variable the head or tail to which it belongs. For instance $z^C = (z^{(1)C}, ..., (z^{(L_C)C})$ is the set of latent variables of the continuous head. This subscript is omitted for the common head. The ℓ-th layer of the head h contains K_ℓ^h components which is the number of components of the associated mixture. L_D and L_C are the number of layers of the discrete and continuous head, respectively.

Each path from one layer to the next is the realization of a mixture. In this sense we introduce, $s^{(\ell)h} \in [1, K_\ell^h]$ the latent variable associated with the index of the component k_ℓ^h of the layer ℓ of the head h. More generally, the latent variable associated with a path going from the first layer to the last layer of one head h is denoted by $s^h = (s^{(1)h}, ..., s^{(L_0)h})$. Similarly, the random variable associated to a path going through all the common tail layers is denoted by $s^{(L_0+1:)} = (s^{(L_0+1)}, ..., s^{(L)})$. By extension, the variable associated with a full path going from the beginning of head h to the end of the common tail is $s^{(1h:L)} = (s^h, s^{L_0+1:})$. $s^{(1h:L)}$ belongs to Ω^h the set of all possible paths starting from one head of cardinal $S^h = \prod_{\ell=1}^{L} K_\ell^h$. The variable associated with a path going from layer ℓ of head h to layer L will be denoted $s^{(\ell h:L)}$. Finally, by a slight abuse of notation a full path going through the component k_ℓ^h of the ℓ-th layer of head h will be denoted: $s^{(1:k_\ell^h:L)}$ or more simply $s^{(:k_\ell^h:)}$.

In order to be as concise as possible, we group the parameters of the model by defining:

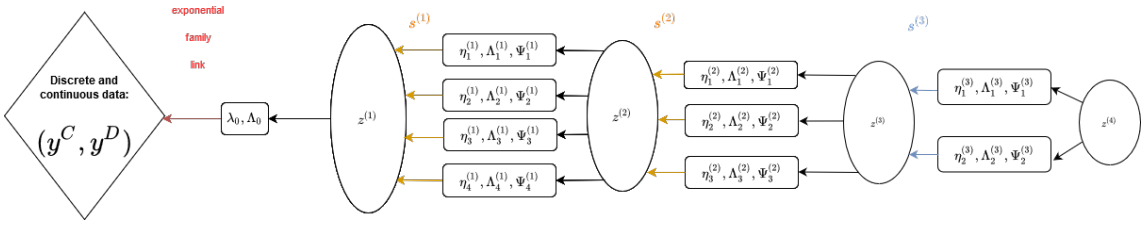
$$\Theta_D = (\Theta_{emb}, \Theta_{DGMM}) = \left((\lambda_0, \Lambda_0), (\eta_{k_\ell}^{(\ell)D}, \Lambda_{k_\ell}^{(\ell)D}, \Psi_{k_\ell}^{(\ell)D})_{k_\ell \in [1, K_\ell^D], \ell \in [1, L_0]}\right),$$

$$\Theta_C = (\eta_{k_\ell}^{(\ell)C}, \Lambda_{k_\ell}^{(\ell)C}, \Psi_{k_\ell}^{(\ell)C})_{k_\ell \in [1, K_\ell^C], \ell \in [1, L_0]}, \quad \Theta_{L_0+1:} = (\eta_{k_\ell}^{(\ell)}, \Lambda_{k_\ell}^{(\ell)}, \Psi_{k_\ell}^{(\ell)})_{k_\ell \in [1, K_\ell], \ell \in [L_0+1, L]},$$

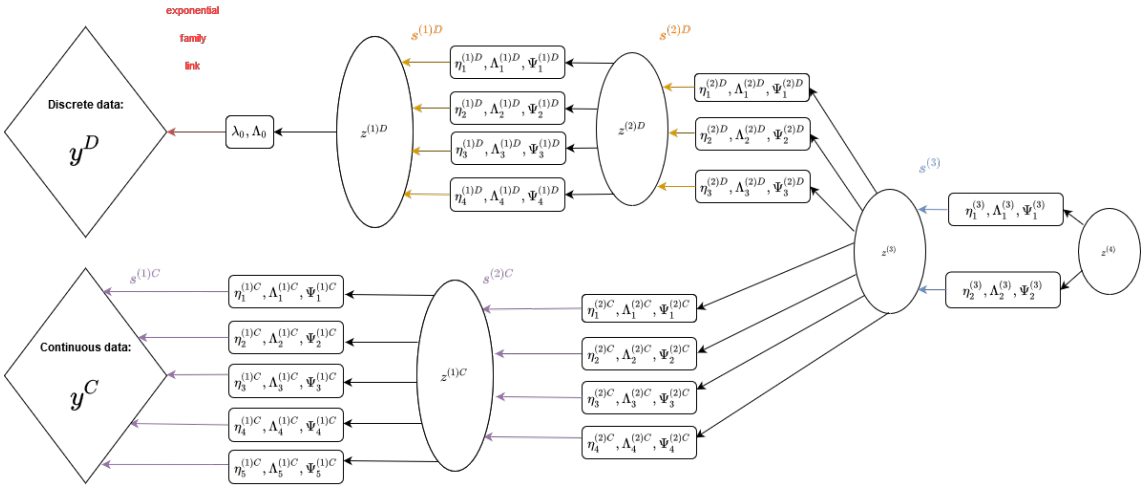
with *emb* standing for embedding.

As an illustration, Figure 1 gives an example graphical models for M1DGMM and M2DGMM.

In Figure 1, for the M2DGMM case we have $K_C = (5, 4)$, $K_D = (4, 3)$, $K = (2, 1)$, $L_C = L_D = L_0 = 2$, $S^C = 40$ and $S^D = 24$. The decreasing size of the $(z^{(\ell)})_\ell$ illustrates the decreasing dimensions of the latent variables.



(a) M1DGMM



(b) M2DGMM

Figure 1: Graphical model of: (a) M1DGMM, (b) M2DGMM

3 Model estimation

We deal only with the M2DGMM, the M1DGMM may be handled in much the same way. The complete density of the M2DGMM is given by:

$$
\begin{aligned}
&L(y^C, y^D, z^C, z^D, z^{(L_0+1:)}, s^C, s^D, s^{(L_0+1:)} | \Theta_C, \Theta_D, \Theta_{L_0+1:}) \\
&\quad = L(y^C | z^{(1)C}, s^C, s^{(L_0+1:)}, \Theta_C, \Theta_{L_0+1:}) L(z^C | z^{(L_0+1:)}, s^C, s^{(L_0+1:)}, \Theta_C, \Theta_{L_0+1:}) \\
&\quad \times L(y^D | z^{(1)D}, s^D, s^{(L_0+1:)}, \Theta_D, \Theta_{L_0+1:}) L(z^D | z^{(L_0+1:)}, s^D, s^{(L_0+1:)}, \Theta_D, \Theta_{L_0+1:}) \\
&\quad \times L(z^{(L_0+1:)} | s^C, s^D, s^{(L_0+1:)}, \Theta_C, \Theta_D, \Theta_{L_0+1:}) L(s^C, s^D, s^{(L_0+1:)} | \Theta_C, \Theta_D, \Theta_{L_0+1:}),
\end{aligned}
$$

which comes from the fact that we assume the two heads of the model to be conditionally independent with respect to the tail layers. Moreover, the layers of both heads and tail share

the Markov property derived from the graphical model: $(z^{(\ell)h} \perp\!\!\!\perp z^{(\ell+2)h}, ..., z^{(L)h}) \Big| z^{(l+1)h}$, $\forall h \in \{C, D, (L_0+1:)\}$.

The aim of the training is to maximize the expected log-likelihood, i.e. to maximize:

$$\mathbb{E}_{z^C, z^D, z^{(L_0+1:)}, s^C, s^D, s^{(L_0+1:)} | y^C, y^D, \hat{\Theta}_C, \hat{\Theta}_D, \hat{\Theta}_{L_0+1:}} [\log L(y^C, y^D, z^C, z^D, z^{(L_0+1:)}, s^C, s^D, s^{(L_0+1:)} | \Theta_C, \Theta_D, \Theta_{L_0+1:})]$$

that we derive in the Supplementary Materials.

The model is fitted using a Monte Carlo version of the EM algorithm (MCEM) introduced by Wei and Tanner (1990). Three types of layers have here to be trained: the GLLVM layer, the regular DGMM layers and the common tail layers that join the two heads.

3.1 Generalized Linear Latent Variable Model Embedding Layer

In this section we present the canonical framework of GLLVMs for discrete data based on Moustaki (2003) and Moustaki and Knott (2000).

By the conditional independence assumption between discrete variables, the likelihood can be written as:

$$f(y^D | \Theta_D, \Theta_{L_0+1:}) = \int_{z^{(1)D}} \prod_{j=1}^{p_D} f(y_j^D | z^{(1)D}, \Theta_D, \Theta_{L_0+1:}) f(z^{(1)D} | \Theta_D, \Theta_{L_0+1:}) dz^{(1)D}, \quad (3)$$

where y_j^D is the jth component of y^D. The density $f(y_j^D | z^{(1)D}, \Theta_D, \Theta_{L_0+1:})$ belongs to an exponential family and in our empirical study we chose a Bernoulli distribution for binary variables, a binomial distribution for count variables and an ordered multinomial distribution for ordinal data. The whole expressions of the densities can be found in Cagnone and Viroli (2014). In order to train the GLLVM layer, we maximize

$$\begin{aligned}
&\mathbb{E}_{z^{(1)D}, s^D, s^{(L_0+1:)} | y^D, \hat{\Theta}_D, \hat{\Theta}_{L_0+1:}} [\log L(y^D | z^{(1)D}, s^D, s^{L_0+1:}, \Theta_D, \Theta_{L_0+1:})] \\
&= \mathbb{E}_{z^{(1)D} | y^D, \hat{\Theta}_D, \hat{\Theta}_{L_0+1:}} [\log L(y^D | z^{(1)D}, \Theta_D, \Theta_{L_0+1:})] \\
&= \int f(z^{(1)D} | y^D, \hat{\Theta}_D, \hat{\Theta}_{L_0+1:}) \log L(y^D | z^{(1)D}, \Theta_D, \Theta_{L_0+1:}) dz^{(1)D},
\end{aligned}$$

the second equality being due to the fact that y^D is related to $(s^D, s^{(L_0+1:)})$ only through $z^{(1)D}$.

3.1.1 MC Step

Draw $M^{(1)}$ observations from $f(z^{(1)D}|s^D, s^{(L_0+1:)}, \hat{\Theta}_D, \hat{\Theta}_{L_0+1:})$.

3.1.2 E step

Hence the E step consists in determining $f(z^{(1)D}|y^D, \hat{\Theta}_D, \hat{\Theta}_{L_0+1:})$, which can be rewritten as:

$$f(z^{(1)D}|y^D, \hat{\Theta}_D, \hat{\Theta}_{L_0+1:}) = \sum_{s'} f(z^{(1)D}|y^D, s', \hat{\Theta}_D, \hat{\Theta}_{L_0+1:}) f(s^{(1D:L)} = s'|y^D, \hat{\Theta}_D, \hat{\Theta}_{L_0+1:})..$$
(4)

The detailed calculus is given in the Supplementary Materials.

3.1.3 M step

There are no closed-form solutions for the estimators of (λ_0, Λ_0) that maximize

$$\mathbb{E}_{z^{(1)D}|y^D, \hat{\Theta}_D, \hat{\Theta}_{L_0+1:}}[\log L(y^D|z^{(1)D}, \Theta_D, \hat{\Theta}_{L_0+1:})].$$

We then use optimisation methods (see Supplementary Materials).

3.2 Determining the parameters of the DGMM layers

In this section, we omit the subscript $h \in \{C, D\}$ on the z^h, y^h and s^h variables because the reasoning is the same for both cases. For $\ell \in [1, L_0]$, we aim to maximize

$$\mathbb{E}_{z^{(\ell)}, z^{(\ell+1)}, s|y, \hat{\Theta}}[\log L(z^{(\ell)}|z^{(\ell+1)}, s, \Theta)].$$

Here the conditional distribution under which the expectation is taken depends on variables located in 3 different layers.

3.2.1 MC Step

At each layer ℓ, $M^{(\ell)}$ pseudo-observations are drawn for each of the previously obtained $\prod_{j=1}^{\ell-1} M^{(j)}$ pseudo-observations. Hence, in order to draw from $f(z^{(\ell)}, z^{(\ell+1)}, s|y, \hat{\Theta})$ at layer ℓ:

- If $\ell = 1$, reuse the $M^{(1)}$ pseudo-observations drawn from $f(z^{(1)}|s, \hat{\Theta})$,
- otherwise reuse the $M^{(\ell-1)}$ pseudo-observations from $f(z^{(\ell-1)}|y, s, \hat{\Theta})$ and the $M^{(\ell)}$ pseudo-observations from $f(z^{(\ell)}|z^{(\ell-1)}, s, \hat{\Theta})$ computed for each pseudo-observation of the previous DGMM layer.
- Draw $M^{(\ell+1)}$ observations from $f(z^{(\ell+1)}|z^{(\ell)}, s, \hat{\Theta})$ for each previously sampled $z^{(\ell)}$.

3.2.2 E Step

The conditional expectation distribution has the following decomposition:

$$\begin{aligned} f(z^{(\ell)}, z^{(\ell+1)}, s|y, \hat{\Theta}) &= f(z^{(\ell)}, s|y, \hat{\Theta}) f(z^{(\ell+1)}|z^{(\ell)}, s, y, \hat{\Theta}) \\ &= f(z^{(\ell)}|y, s, \hat{\Theta}) f(s|y, \hat{\Theta}) f(z^{(\ell+1)}|z^{(\ell)}, s, \hat{\Theta}), \end{aligned} \tag{5}$$

and we develop this term in the Supplementary Materials.

3.2.3 M step

The estimators of the DGMM layer parameters $\forall \ell \in [1, L_0]$ are given by:

$$\begin{cases} \hat{\eta}_{k_\ell}^{(\ell)} = \dfrac{\sum_{i=1}^{n} \sum_{\tilde{s}_i^{(:k_\ell:)}} f(s_i^{(:k_\ell:)} = \tilde{s}_i^{(:k_\ell:)}|y, \hat{\Theta}) \left[E[z_i^{(\ell)}|s_i^{(:k_\ell:)} = \tilde{s}_i^{(:k_\ell:)}, y_i, \hat{\Theta}] - \Lambda_{k_\ell}^{(\ell)} E[z_i^{(\ell+1)}|\tilde{s}_i^{(:k_\ell:)}, y_i, \hat{\Theta}] \right]}{\sum_{i=1}^{n} \sum_{\tilde{s}_i^{(:k_\ell:)}} f(s_i^{(:k_\ell:)} = \tilde{s}_i^{(:k_\ell:)}|y_i, \hat{\Theta})} \\ \hat{\Lambda}_{k_\ell}^{(\ell)} = \dfrac{\sum_{i=1}^{n} \sum_{\tilde{s}_i^{(:k_\ell:)}} f(s_i^{(:k_\ell:)} = \tilde{s}_i^{(:k_\ell:)}|y_i, \hat{\Theta}) \left[E[(z_i^{(\ell)} - \hat{\eta}_{k_\ell}^{(\ell)}) z_i^{(\ell+1)T}|s_i^{(:k_\ell:)} = \tilde{s}_i^{(:k_\ell:)}, y_i, \hat{\Theta}] \right]}{\sum_{i=1}^{n} \sum_{\tilde{s}_i^{(:k_\ell:)}} f(s_i^{(:k_\ell:)} = \tilde{s}_i^{(:k_\ell:)}|y_i, \hat{\Theta})} E[z_i^{(\ell+1)} z_i^{(\ell+1)T}|\tilde{s}_i^{(:k_\ell:)}, y_i, \hat{\Theta}]^{-1} \\ \hat{\Psi}_{k_\ell}^{(\ell)} = \dfrac{\sum_{i=1}^{n} \sum_{\tilde{s}_i} f(s_i^{(:k_\ell:)} = \tilde{s}_i^{(:k_\ell:)}|y_i, \hat{\Theta}) E\left[\left(z_i^{(\ell)} - (\eta_{k_l}^{(\ell)} + \Lambda_{k_l}^{(\ell)} z_i^{(\ell+1)}) \right) \left(z_i^{(\ell)} - (\eta_{k_l}^{(\ell)} + \Lambda_{k_l}^{(\ell)} z_i^{(\ell+1)}) \right)^T |\tilde{s}_i^{(:k_\ell:)}, y_i, \hat{\Theta} \right]}{\sum_{i=1}^{n} \sum_{\tilde{s}_i^{(:k_\ell:)}} f(s_i^{(:k_\ell:)} = \tilde{s}_i^{(:k_\ell:)}|y_i, \hat{\Theta})}, \end{cases}$$

with $\tilde{s}_i^{(:k_\ell:)} = (\tilde{k}_1, ..., \tilde{k}_{\ell-1}, k_\ell, \tilde{k}_{\ell+1}, ..., \tilde{k}_L)$, a path going through the network and reaching the component k_ℓ. The details of the computation are given in the Supplementary Materials.

3.3 Training of the common tail layers

In this section we aim to maximise $\forall \ell \in [L_0 + 1, L]$, the following expression:

$$\mathbb{E}_{z^{(\ell)}, z^{(\ell+1)}, s^C, s^D, s^{(L_0+1:)}|y^C, y^D, \hat{\Theta}_C, \hat{\Theta}_D, \hat{\Theta}_{L_0+1:}} [\log L(z^{(\ell)}|z^{(\ell+1)}, s^C, s^D, s^{(L_0+1:)}, \Theta_C, \Theta_D, \Theta_{L_0+1:})].$$

3.3.1 MC Step

The MC step remains the same as for regular DGMM layers except that the conditioning concerns both types of data (y^C and y^D) and not only discrete or continuous data as in the heads layers.

3.3.2 E Step

The distribution of the conditional expectation is $f(z^{(\ell)}, z^{(\ell+1)}, s^C, s^D, s^{(L_0+1:)}|y^C, y^D, \hat{\Theta}_C, \hat{\Theta}_D, \hat{\Theta}_{L_0+1:})$ that we can express as previously. We detail the calculus in the Supplementary Materials.

3.3.3 M Step

The estimators of the junction layers keep the same form as the regular DGMM layers except once again that the two types of data and paths exist in the conditional distribution of the expectation.

3.4 Determining the path probabilities

In this section, we determine the path probabilities by optimizing the parameters of the following expression derived from the expected log-likelihood:

$$\mathbb{E}_{s^C, s^D, s^{(L_0+1:)}|y^C, y^D, \hat{\Theta}_C, \hat{\Theta}_D, \hat{\Theta}_{L_0+1:}}[\log L(s^C, s^D, s^{(L_0+1:)}|\Theta_C, \Theta_D, \Theta_{L_0+1:})],$$

with respect to π_s^h, $\forall h \in \{C, D\}$ and $\pi_s^{(L_0+1:)}$.

3.4.1 E step

By mutual independence of s^C, s^D and $s^{L_0+1:}$, estimating the distribution of the expectation boils down to computing three densities: $f(s^{(\ell)D} = k_\ell|y^D, \hat{\Theta}_D, \hat{\Theta}_{L_0+1:})$, $f(s^{(\ell)C} = k_\ell|y^C, \hat{\Theta}_C, \hat{\Theta}_{L_0+1:})$, and $f(s^{(\ell)} = k_\ell|y^C, y^D, \hat{\Theta}_C, \hat{\Theta}_D, \hat{\Theta}_{L_0+1:})$ (details are given in the Supplementary Materials).

3.4.2 M step

Estimators for each head h and for the common tail are given respectively by (see the Supplementary Materials):

$$\hat{\pi}_{k_\ell}^{(\ell)h} = \frac{\sum_{i=1}^{n} f(s_i^{(\ell)h} = k_\ell | y_i^h, \hat{\Theta}_h, \hat{\Theta}_{L_0+1:})}{n} \quad \text{and} \quad \hat{\pi}_{k_\ell}^{(\ell)} = \frac{\sum_{i=1}^{n} f(s_i^{(\ell)} = k_\ell | y_i^C, y_i^D, \hat{\Theta}_C, \hat{\Theta}_D, \hat{\Theta}_{L_0+1:})}{n}.$$

4 Identifiability

In this section, we combine both GLLVM and DGMM identifiability constraints proposed in Cagnone and Viroli (2014) and Viroli and McLachlan (2019), respectively, to make our model identifiable.

4.1 GLLVM identifiability constraints

Both the GLLVM model and the Factor Analysis model assume that the latent variables are centered and of unit variance. This can be obtained by rescaling iteratively all the latent layers parameters from the last common layer to the first head layers as follows:

$$\begin{cases} \eta_{k_\ell}^{(\ell)hnew} = (A^{(\ell)h})^{-1T} \left[\eta_{k_\ell}^{(\ell)h} - \sum_{k'_\ell} \pi_{k'_\ell}^{(\ell)h} \eta_{k'_\ell}^{(\ell)h} \right] \\ \Lambda_{k_\ell}^{(\ell)hnew} = (A^{(\ell)h})^{-1T} \Lambda_{k_\ell}^{(\ell)h} \\ \Psi_{k_\ell}^{(\ell)new} = (A^{(\ell)h})^{-1T} \Psi_{k_\ell}^{(\ell)h} (A^{(\ell)h})^{-1}. \end{cases}$$

where $A^{(\ell)h} = Var(z^{(\ell)h})\ \forall \ell \in [1, L], h \in \{C, D, L_0 + 1 :\}$ and the subscript "new" denotes the rescaled version of the parameters. The details are given in the Supplementary Materials. In the same way, the coefficients of $\Lambda^{(0)}$ of the discrete head are rescaled as follows: $\Lambda^{(0)new} = \Lambda^{(0)} A^{-1T}$.

In GLLVM models, the number of coefficients of the $\Lambda^{(0)}$ matrix for binary and count data leads to a too high number of degrees of freedom. Thus, to ensure the identifiability of the model, one has to reduce the number of free coefficients. As in Cagnone and Viroli (2014) the upper triangular coefficients of $\Lambda^{(0)}$ are constrained to be zero for binary and

count data. This constraint is explicitly taken into account during the optimisation phase, as the optimisation program is looking for solutions for $\Lambda^{(0)}$ that are upper triangular.

4.2 DGMM identifiability constraints

We assume first that the latent dimension is decreasing through the layers of each head and tail *i.e.* $p_h > r_1^h > ... > r_L$. Secondly, we make the assumption that $\Lambda_{k_\ell}^{(\ell)hT}\Psi_{k_\ell}^{(\ell)-1h}\Lambda_{k_\ell}^{(\ell)h}$ is diagonal with elements in decreasing order $\forall \ell \in [1, L]$. Fruehwirth-Schnatter and Lopes (2018) obtained sufficient conditions for MFA identifiability, including the so-called *Anderson-Rubin* (AR) condition, which requires that $r_\ell \leq \frac{r_{\ell-1}-1}{2}$. Enforcing this condition would prevent from defining a MDGMM for all datasets that present less than 7 variables of each type which is far too restrictive. Then, we implement a transformation to ensure the diagonality of $\Lambda_{k_\ell}^{(\ell)hT}\Psi_{k_\ell}^{(\ell)-1h}\Lambda_{k_\ell}^{(\ell)h}$ as follows: once all parameters have been estimated by the MCEM algorithm, the following transformation is applied over $\Lambda_{k_\ell}^{(\ell)h}$:

- Compute $B = \Lambda_{k_\ell}^{(\ell)hT}\Psi_{k_\ell}^{(\ell)-1h}\Lambda_{k_\ell}^{(\ell)h}$.
- Decompose B according to the eigendecomposition $B = PDP^{-1}$, with D the matrix of the eigenvalues and P the matrix of eigenvectors.
- Define $\Lambda_{k_\ell}^{(\ell)hnew} = \Lambda_{k_\ell}^{(\ell)h}P$.

5 Practical considerations

5.1 Initialisation procedure

EM-based algorithms are known to be very sensitive to their initialisation values as shown for instance by Biernacki et al. (2003) for Gaussian Mixture models. In our case, using purely random initialization as in Cagnone and Viroli (2014) made the model diverge most of the time when the latent space was of high dimension. This can be explained by the fact that the clustering is performed in a projected continuous space of which one has no prior knowledge about. Initialising at random the latent variables $(\eta_{k_\ell}^{(\ell)h}, \Lambda_{k_\ell}^{(\ell)h}, \Psi_{k_\ell}^{(\ell)h}, s^{(\ell)h}, z^{(\ell)h})_{k_\ell,\ell,h}$ and the exponential family links parameters $(\lambda^{(0)}, \Lambda^{(0)})$

seems not to be a good practice. This problem gets even worse as the number of DGMM layers grows. To stabilize our algorithm we propose the NSEP approach which combines MCA, GMM, FA and PLS algorithm in the M2DGMM case.

- For discrete head initialisation, the idea used here is to perform a Multiple Correspondence Analysis (MCA) (Nenadic and Greenacre, 2005) to determine a continuous low dimensional representation of the discrete data and use it as a first approximation of the latent variables $z^{(1)D}$. The MCA considers all variables as categorical, thus the more the dataset actually contains this type of variables the better the initialisation should in theory be. Once this is done, a Gaussian Mixture Model is fitted in order to determine groups in the continuous space and to estimate $(\pi_{k_\ell}^{(\ell)})$. For each group a Factor Analysis Model (FA) is fitted to determine the parameters of the model $(\eta_{k_\ell}^{(\ell)}, \Lambda_{k_\ell}^{(\ell)}, \Psi_{k_\ell}^{(\ell)})$ and the latent variable of the following layer $z^{(\ell+1)}$. Concerning the GLLVM parameters, logistic regressions of y_j^D over $z^{(1)D}$ are fitted for each original variable of the discrete head: an ordered logistic regression for ordinal variables, an unordered logistic regression for binary, count and categorical variables.

- For the continuous head and the common tail, the same described GMM coupled with FA procedure can be applied to determine the coefficients of the layer. The difficulty concerns the initialisation of the first tail layer with latent variable $z^{(L_0+1)}$. Indeed, $z^{(L_0+1)}$ has to be the same for both discrete and continuous last layers. As Factor Models are unsupervised models, one cannot enforce such a constraint on the latent variable generated from each head. To overcome this difficulty, $z^{(L_0+1)}$ has been determined by applying a PCA over the stacked variables $(z^{(L_0)C}, z^{(L_0)D})$. Then the DGMM coefficients $(\eta_{k_{L_0}}^{(L_0)h}, \Lambda_{k_{L_0}}^{(L_0)h}, \Psi_{k_{L_0}}^{(L_0)}h)$ of each head have been separately determined using Partial Least Square (Wold et al., 2001) of each head last latent variable over $z^{(L_0+1)}$.

The same ideas are used to initialize the M1DGMM. As the data going through the unique head of the M1DGMM are mixed, Factor analysis of mixed data (Pagès, 2014) is employed instead of MCA as it can handle mixed data.

5.2 Model and number of clusters selection

The selection of the best MDGMM architecture is performed using the pruning methodology which is widely used in the field of supervised neural networks (Blalock et al., 2020) but also for tree-based methods (Patil et al., 2010). The idea is to determine the simplest architecture that could describe the data. In order to do so, one starts with a complex architecture, and deletes the coefficients that do not carry enough information. Deleting those coefficients at some point during the training process is known as "pre-pruning" and performing those deletions after full convergence is known as "post-pruning". In our case, we use a pre-pruning strategy to estimate the best number of components k_ℓ, the best number of factors r_ℓ and the best number of layers for the heads and tails. The reason not to use post-pruning instead of pre-pruning is that very complex architectures tend to show long running times and a higher propensity not to converge to good maxima in our simulations.

Classical approaches to model specification based on information criteria, such as AIC (Akaike, 1998) or BIC (Schwarz et al., 1978), need the estimation of all the possible specifications of the model. In contrast, our approach needs only one model run to determine the best architecture which is far more computationally efficient.

In the following, we give a summary of our pruning strategy (extensive details are provided in the Supplementary Materials). The idea is to determine the best number of components on each layer k_ℓ^h by deleting the components associated with very low probabilities $\pi_{k_\ell}^{(\ell)h}$ as they are the least likely to explain the data.

The choice of the latent dimensions of each layer r_ℓ^h is performed by looking at the dimensions that carry the most important pieces of information about the previous layer. The goal is to ensure the circulation of relevant information through the layers without transmitting noise information. This selection is conducted differently for the GLLVM layer compared to the regular DGMM layers. For the GLLVM layer, we perform logistic regressions of y^C over $z^{(1)C}$ and delete the dimensions that were associated with non-significant coefficients in a vast majority of paths. Concerning the regular DGMM layers, information carried by the current layer given the previous layer has been modeled using

a Principal Component Analysis. We compute the contribution of each original dimension to the first principal component analysis and keep only the dimensions that present a high correlation with this first principal component, so that to drop information of secondary importance carried out through the layers.

Finally, the choice of the total number of layers is guided by the selected r_ℓ. For instance, if a dimension of two is selected for a head layer (or a dimension of one for a tail layer), then according to the identifiability constraint $p_h > r_1^h > ... > r_\ell^h > ... > r_L$, the following head (or tail) layers are deleted.

Given that this procedure also selects the number of components on the tail layers, it can also be used to automatically find the optimal number of clusters in the data. The user specifies a high number of components on the clustering layer and let the automatic selection operate. The optimal number of clusters is then the number of components remaining on the clustering layer at the end of the run. This feature of the algorithm is referred to as the "autoclus mode" of the MDGMM in the following and in the code.

Alternatively, in case of doubt about the number of clusters in the data, the MDGMM could be used in "multi-clustering" mode. For example, if the number of clusters in the data is assumed to be two or three, one can define a MDGMM with three components on the first tail layer and two on the second tail layer. The first layer will output a three groups clustering and the second layer a two groups clustering. The two partitions obtained can then be compared to chose the best one. This can be done with the silhouette coefficient (Rousseeuw, 1987) as implemented in our code. In the "multi-clustering" mode, the same described model selection occurs. The only exception is that the number of components of the tail layers remains frozen (as it corresponds to the tested number of clusters in the data).

For all clustering modes of the MDGMM, the architecture selection procedure is performed at the end of some iterations chosen by the user before launching the algorithm. Note that once the optimal specification has been determined, it is better to refit the model using the determined specification rather than keeping the former output. Indeed, changing the architecture "on the fly" seems to disturb the quality of the final clustering.

Finally, in EM-based algorithms, the iteration which presents the best likelihood (the last one in general) is returned as the final output of the model. The likelihood of the model informs about how good the model is at explaining the data. However, it does not give direct information about the clustering performance of the model itself. Therefore, in the MDGMM we retain the iteration presenting the best silhouette coefficient (Rousseeuw, 1987) among all iterations. To summarize: the likelihood criterion was used as a stopping criterion to determine the total number of iterations of the algorithm and the best silhouette score was used to select the iteration returned by the model.

6 Real Applications

In this section we illustrate the proposed models on real datasets. First, we will present the continuous low dimensional representations of the data generated by the Discrete DGMM (DDGMM) and the M2DGMM. Then, the performance will be properly evaluated by comparing them to state-of-the-art mixed data clustering algorithms, the one-head version of the MDGMM (M1DGMM) provided with a Gaussian link function, the NSEP and the GLLVM. As some of the clustering models can deal with discrete data only (GLLVM, DDGMM) and other with mixed data (M1DGMM, MDGMM) we consider both types of data sets. The code of the introduced models is available on Github under the name MDGMM_suite. The associated DOI is 10.5281/zenodo.4382321.

6.1 Data description

For the discrete data specification, we present results obtained on three datasets: the Breast cancer, the Mushrooms and the Tic Tac Toe datasets.

- The Breast cancer dataset is a dataset of 286 observations and 9 discrete variables. Most of the variables are ordinal.
- The Tic Tac Toe dataset is composed of 9 variables corresponding to each cell of a 3×3 grid of tic-tac-toe. The dataset presents the grids content at the end of 958 games. Each cell can be filled with one of the player symbol (x or o), or left blanked

(b) if the play has ended before all cells were filled in. Hence all the variables are categorical in contrast with the Breast cancer data.
The goal is here to retrieve which game has led to victory of player 1 or of player 2 (no even games are considered here).

- Finally, the Mushrooms dataset is a two-class dataset with 22 attributes and 5644 observations once the missing data have been removed. The majority of the variables are categorical ones.

For mixed datasets, we have used the Australian credit, the Heart (Statlog) and the Pima Indians diabetes Datasets.

- The Heart (Stalog) dataset is composed of 270 observations, five continuous variables, three categorical variables, three binary variables and two ordinal variables.
- The Pima Indians Diabetes dataset presents several physiological variables (e.g. the blood pressure, the insulin rate, the age) of 768 Indian individuals. 267 individuals suffer from diabetes and the goal of classification tasks over this dataset is to distinguish the sound people from the sick ones. This dataset counts two discrete variables considered here respectively as binomial and ordinal and seven continuous variables.
- Finally, the Australian credit (Stalog) dataset is a binary classification dataset concerning credit cards. It is composed of 690 observations, 8 discrete categorical variables and 6 continuous variables. It is a small dataset with a high dimension.

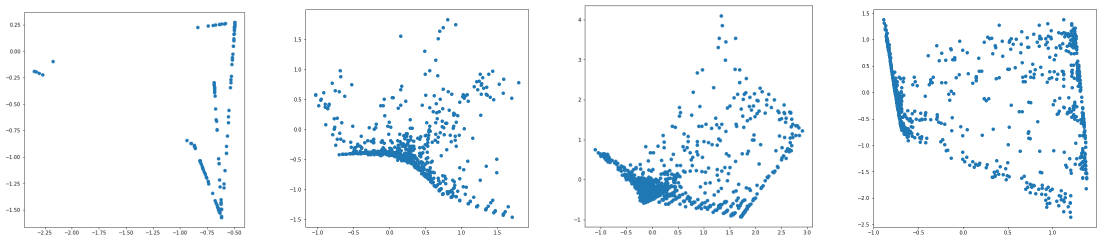
In the analysis, all the continuous variables have been centered and reduced to ensure the numeric stability of the algorithms. All the datasets are available in the UCI repository (Dua and Graff, 2017).

6.2 Clustering vizualisation

According to their multi-layer structures, the DDGMM and the M2DGMM perform several dimension reductions of information while the signal goes through their layers. As such, they provide low dimensional continuous representations of complex data than can be

discrete, mixed or potentially highly dimensional. These representations are useful to understand how observations are clustered through the training process. They could also be reused to train other algorithms in the same spirit as for supervised Neural Networks (Jogin et al., 2018).

Figure 2 shows the evolution of the latent representation during the training of the clustering layer of a DDGMM for the tic tac toe dataset. Four illustrative iterations have been chosen to highlight the training process. The clustering layer has a dimension of $r_\ell = 2$ and tries to distinguish $k_\ell = 2$ groups in the data. At the beginning of the training at $t1$, it is rather difficult to differentiate two clusters in the data. However, through the next iterations, one can clearly distinguish that two sets of points are pushed away from each other by the model. Moreover in $t3$ the frontier between the two clusters can be drawn as a straight line in a two dimensional space. In $t4$ at the end of the training, the model seems to have found a simpler frontier to separate the groups as only a vertical line, i.e. a separation in a one dimensional space is needed. This highlight the information sorting process occurring through the layers in order to keep only the simplest and the more discriminating parts of the signal.

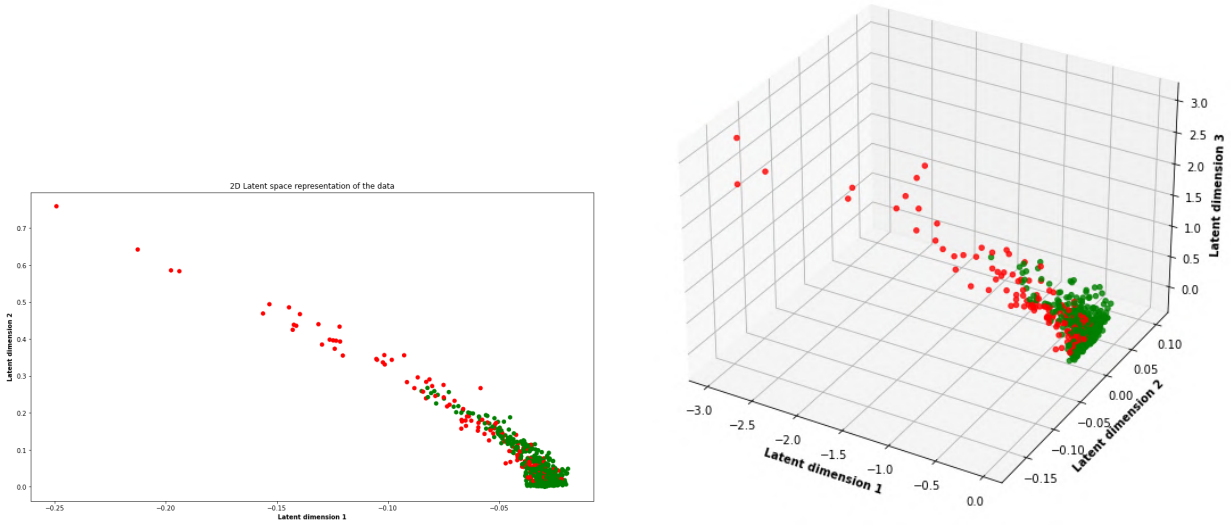


(a) Training at t1 (b) Training at t2 (c) Training at t3 (d) Training at t4

Figure 2: Continuous representation of the Tic Tac Toe dataset through the training of a DDGMM

The next two figures illustrate graphical properties of the M2DGMM. Figure 3 presents two continuous representations of the Pima Diabetes data. These are obtained during the training of a M2DGMM with two hidden tail layers of respectively $r_{L_0+1} = 3$ and $r_{L_0+1} = 2$ during the same iteration. Two clusters are looked for in each case ($K_{L_0+1} = K_{L_0+2} = 2$)

and are associated with green and red colors on the figure.



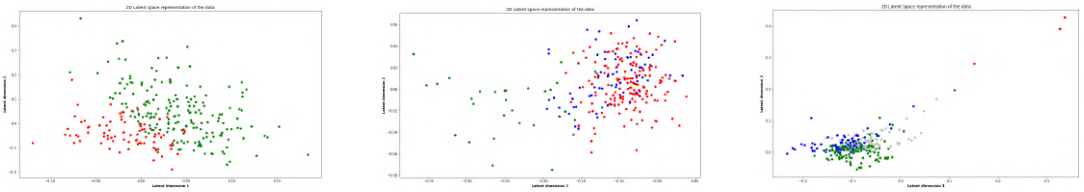
(a) 2D representation

(b) 3D representation

Figure 3: Continuous representations of the Pima Diabetes dataset provided by a M2DGMM

On both layers the clusters are quite well separated. The signal carried seems coherent between the two layers with a very similar structure. For the same computational cost, *i.e.* one run of the model, several latent representations of the data in different dimensions can therefore be obtained.

Finally, the graphical representations produced by the M2DGMM are useful tools to identify the right number of clusters in the data. Three M2DGMM have been run by setting $r_{L_0+1} = 2$ and with respectively $K_{L_0+1} = 2, K_{L_0+1} = 3$ and $K_{L_0+1} = 4$. The associated latent variables are presented in Figure 4 with a different color for each identified cluster.



(a) 2 clusters (b) 3 clusters (c) 4 clusters

Figure 4: Continuous representations of the Heart dataset at the end of the training of three M2DGMMs with different numbers of clusters specified

The representations with three and four clusters present points that are intertwined, with no clear distinctions between clusters. On the contrary, when the number of clusters searched in the data is two this separation appears distinctly. Hence, this representation advocates for a two groups distinction in the data as it is suggested by the supervised labels of the dataset (absence or presence of heart disease). The four clusters representation also shows that the three points associated with the red cluster might be outliers potentially important to study.
As evoked in Subsection 5.2, this visual diagnostic can be completed by using the "autoclus mode" of the M2DGMM where the model automatically determines the best number of clusters in the data.

6.3 Performance comparison

In order to benchmark the performance of the proposed strategy, we consider alternative algorithms coming from each family of approaches identified by Ahmad and Khan (2019), namely k-modes, k-Prototypes, Hierarchical Clustering, Self-Organising Maps (SOM), and DBSCAN (Ester et al., 1996).

For each dataset, we have set the number of unsupervised clusters to the "ground truth" classification number. In order to present a fair report, several specifications of the benchmark models have been run. For each specification, the models have been launched 30 times. The reported results correspond to the best specification of each benchmark model with respect to each metric on average over the 30 runs. The set of specifications evaluated

Metrics	Silhouette	Micro	Macro
Algorithms	Breast Cancer		
GLLVM (random init)	0.215 (0.093)	0.673 (0.080)	0.570 (0.113)
GLLVM (with NSEP)	0.305 (0.023)	0.728 (0.025)	0.671 (0.018)
NSEP	0.303 (0.000)	0.722 (0.000)	0.664 (0.000)
DDGMM	0.268 (0.043)	0.696 (0.074)	0.648 (0.048)
k-Modes	0.174 (0.000)	0.592 (0.000)	0.534 (0.000)
k-Prototypes	0.293 (0.024)	0.729 (0.014)	0.666 (0.011)
Hierarchical	0.303 (0.000)	0.755 (0.000)	0.855 (0.000)
SOM	0.091 (0.088)	0.668 (0.060)	0.593 (0.011)
DBSCAN	0.264 (0.000)	0.726 (0.000)	0.860 (0.000)
	Tic Tac Toe dataset		
GLLVM (random init)	0.094 (0.031)	0.591 (0.052)	0.536 (0.100)
GLLVM (with NSEP)	0.110 (0.005)	0.550 (0.029)	0.545 (0.028)
NSEP	0.137 (0.000)	0.602 (0.021)	0.597 (0.019)
DDGMM	0.118 (0.016)	0.559 (0.028)	0.533 (0.036)
k-Modes	0.104 (0.002)	0.611 (0.000)	0.586 (0.000)
k-Prototypes	∅(∅)	∅(∅)	∅(∅)
Hierarchical	0.078 (0.000)	0.654 (0.000)	0.827 (0.000)
SOM	0.082 (0.010)	0.650 (0.000)	0.560 (0.000)
DBSCAN	∅(∅)	0.653 (0.000)	0.327 (0.000)
	Mushrooms dataset		
GLLVM (random init)	0.266 (0.103)	0.685 (0.107)	0.613 (0.255)
GLLVM (with NSEP)	0.351 (0.107)	0.803 (0.102)	0.854 (0.135)
NSEP	0.354 (0.064)	0.811 (0.101)	0.861 (0.074)
DDGMM	0.317 (0.078)	0.760 (0.131)	0.809 (0.116)
k-Modes	0.395 (0.000)	0.852 (0.000)	0.898 (0.000)
k-Prototypes	0.328 (0.081)	0.742 (0.136)	0.818 (0.086)
Hierarchical	0.395 (0.000)	0.854 (0.000)	0.904 (0.000)
SOM	0.155 (0.015)	0.710 (0.000)	0.814 (0.001)
DBSCAN	0.294 (0.000)	0.624 (0.000)	0.811 (0.000)

Table 1: Average results and standard errors over 30 runs of the best specification for each model over three discrete datasets

Algorithms	Silhouette	Micro	Macro
Metrics	Heart		
NSEP	0.165 (0.049)	0.738 (0.068)	0.739 (0.070)
M1DGMM	0.253 (0.003)	0.820 (0.012)	0.820 (0.012)
M2DGMM	0.146 (0.011)	0.710 (0.015)	0.712 (0.014)
k-Modes	0.247 (0.000)	0.811 (0.000)	0.813 (0.000)
k-Prototypes	0.044 (0.000)	0.593 (0.000)	0.585 (0.000)
Hierarchical	0.263 (0.000)	0.811 (0.000)	0.809 (0.000)
SOM	0.257 (0.000)	0.795 (0.000)	0.793 (0.000)
DBSCAN	0.177 (0.000)	0.556 (0.000)	0.724 (0.000)
	Pima		
NSEP	0.189 (0.013)	0.666 (0.056)	0.651 (0.051)
M1DGMM	0.227 (0.020)	0.633 (0.029)	0.607 (0.029)
M2DGMM	0.195 (0.079)	0.647 (0.019)	0.586 (0.068)
k-Modes	0.049 (0.033)	0.581 (0.000)	0.482 (0.000)
k-Prototypes	∅ (∅)	∅ (∅)	∅ (∅)
Hierarchical	0.391 (0.000)	0.656 (0.000)	0.826 (0.000)
SOM	0.232 (0.000)	0.644 (0.000)	0.610 (0.003)
DBSCAN	0.391 (0.000)	0.654 (0.000)	0.826 (0.000)
	Australian Credit		
NSEP	0.165 (0.034)	0.754 (0.098)	0.753 (0.110)
M1DGMM	0.170 (0.032)	0.707 (0.112)	0.806 (0.036)
M2DGMM	0.224 (0.080)	0.575 (0.040)	0.680 (0.104)
k-Modes	0.222 (0.007)	0.785 (0.008)	0.784 (0.007)
k-Prototypes	0.163 (0.000)	0.562 (0.000)	0.780 (0.000)
Hierarchical	0.399 (0.000)	0.849 (0.000)	0.847 (0.000)
SOM	0.127 (0.096)	0.649 (0.001)	0.676 (0.002)
DBSCAN	0.201 (0.000)	0.570 (0.000)	0.740 (0.000)

Table 2: Average results and standard errors over 30 runs of the best specification for each model over three mixed datasets

for each benchmark model is given in the Supplementary Materials. Concerning our models, the architectures were automatically selected and then fitted 30 times on each dataset. Here we use one unsupervised metric and two supervised metrics to assess the clustering quality: the silhouette coefficient, the micro precision and the macro precision. The silhouette coefficient measures how close on average a point is from the points of the same group with respect to the points of the other groups. The Euclidian distance cannot be used here due to the mixed feature space and hence the Gower distance (Gower, 1971) is used instead. The silhouette coefficient ranges between 1 (perfect clustering) and -1 (meaningless clustering). The micro precision corresponds to the overall accuracy, *i.e.* the proportion of correctly classified instances. The macro precision computes the proportion of correctly classified instances per class and then returns a non-weighted mean of those proportions. These two quantities tend to differ when the data are not balanced. The formal expressions of the metrics are given in the Supplementary Materials. Note that we cannot use AIC or BIC criteria here since their values are not available for all methods.

Tables 1-2 present the best average results obtained by the algorithms and the associated standard error over the 30 runs in parenthesis. The best algorithm for a given dataset and metric is associated with a green cell and the worst with a red cell. An empty set symbol means that the metric was not defined for this algorithm on that dataset. For the special case of the k-prototypes algorithm, the empty set symbol means that the dataset contained only one type of discrete data which is a situation that the algorithm is not designed for.

6.3.1 Results on discrete data

The new initialisation (NSEP) enables the GLLVM to achieve better performances on the Mushrooms dataset and on the Breast dataset where the GLLVM attains the best silhouette score. It also stabilizes the GLLVM as the standard errors obtained are divided by at least a factor two for all metrics of the Breast Cancer and of the Tic Tac Toe datasets.
The NSEP in itself gives good results for all metrics and is often among the best two performing models. Finally, over the Tic Tac Toe dataset the DDGMM performs slightly better than the GLLVM, but less on the two other datasets.

Hence, compared to the other methods, the models introduced in this work represent solid baseline models. On the contrary, some alternative methods appear to fit some datasets well and poorly other ones. This is the case for instance of DBSCAN which performs well on the Breast cancer dataset, but much less on the Mushrooms and the Tic Tac Toe datasets (the algorithm could find only one group in the Tic Tac Toe data which explains that the silhouette score is not defined). Another example is k-Modes which obtains substantial results on the Mushrooms dataset but under-average results for the two other datasets. Finally, among all methods, the hierarchical clustering is the algorithm that performs best on a majority of metrics and datasets.

6.3.2 Results on mixed data

As clear from results in Table 2, the NSEP seems again to be a good starting point for both algorithms and certainly also explains the fact that the M1DGMM reaches the best micro and macro scores on the Heart dataset.
The M1DGMM achieves better average results than the M2DGMM except for the silhouette score on the Australian Credit dataset and the micro precision on the Pima dataset. The two specifications tend to often present opposite patterns in terms of standard errors: when the M2DGMM results are stable the M1DGMM results tend to be more volatile and vice versa. Hence, depending on whether one wants to minimize the bias or the variance of the estimation, the two specifications seem complementary and could be used in turn to conduct clustering on a large diversity of datasets.
As in the discrete data results, the models introduced and especially the M1DGMM, give satisfactory performance on all datasets on average, whereas other models such as SOM, DBSCAN or k-modes perform well on some datasets only. Similarly, the hierarchical clustering method seems to provide the best results on a large set of metrics and datasets.

7 Conclusion

This work aimed to provide a reliable and flexible model for clustering mixed data by borrowing ingredients from the GLLVM and the DGMM recent approaches. Several sub-

models have been introduced and could be used on their own:

- a new initialisation procedure called NSEP for GLLVM-based models,
- a Discrete DGMM (DDGMM) for discrete data,
- a one-head (M1DGMM) and a two-heads (M2DGMM) DGMM for mixed data.

This suite of models handles the usual clustering issues concerning architecture selection and the choice of the number of clusters in the data in an automated manner.
From the experiments carried out on real data, the MDGMM performances are in line with the other state-of-the-art models. It can be regarded as a baseline model over a general class of data. Its use of nested Mixtures of Factor Analyzers enables it to capture a very wide range of distributions and patterns.
Despite of its complexity, the MDGMM remains interpretable. From a practical viewpoint, the structure of the latent space can be observed through the model training with the help of the graphical utilities presented in section 6.2. Thus, they allow the user to perform visual diagnostics of the clustering process. From a theoretical standpoint, the parameters of the model remain interpretable as the link between parameters and clustering results is proper thanks to the identifiability of the model. The set of identifiability constraints presented here could seem quite restrictive. However, it forces the model to stay in a quite well delimited parameter space and to avoid for instance a too significant explosion of the norm of the parameters values. The implementation of these constraints can nevertheless be improved by considering a Bayesian re-writing of our model on Variational principles. Indeed, it should make identification requirements easier to meet, as one can keep only the posterior draws that meet the identifiability requirements. Niku et al. (2019) have rewritten the GLLVM model in a variational fashion and exhibit high running time and accuracy gains. Following their path, one could adapt the MDGMM to the variational framework.
Finally considering the training process, the choice of an EM-based algorithm was motivated by its extensive use in the Gaussian Mixture Model literature. The EM-related algorithms are however very sensitive to the initialisation, which was in our case particularly tricky given the size of the parameter space. Combining Multiple Correspondence

Analysis with Gaussian Mixture Models, Factor Analysis and Partial Least Squares into NSEP has however enabled us to significantly stabilize the estimation process. Yet, new initialisation and training processes could be designed to help the model to better rationalize latent structures in the data within its very highly dimensional space.

Acknowledgments

Thanks to the LIA LYSM (agreement between AMU, CNRS, ECM and INdAM) for having funded a mission to Bologna. Thanks also to Nicolas Chopin and Samuel Soubeyrand for their helpful advices.

References

Ahmad, A. and S. S. Khan (2019). Survey of state-of-the-art mixed data clustering algorithms. *IEEE Access 7*, 31883–31902.

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pp. 199–213. Springer.

Baydin, A. G., B. A. Pearlmutter, A. A. Radul, and J. M. Siskind (2017). Automatic differentiation in machine learning: a survey. *The Journal of Machine Learning Research 18*(1), 5595–5637.

Biernacki, C., G. Celeux, and G. Govaert (2003). Choosing starting values for the em algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Computational Statistics & Data Analysis 41*(3-4), 561–575.

Blalock, D., J. J. G. Ortiz, J. Frankle, and J. Guttag (2020). What is the state of neural network pruning? arXiv preprint arXiv:2003.03033.

Cagnone, S. and C. Viroli (2014). A factor mixture model for analyzing heterogeneity and cognitive structure of dementia. *AStA Advances in Statistical Analysis 98*(1), 1–20.

Chiu, T., D. Fang, J. Chen, Y. Wang, and C. Jeris (2001). A robust and scalable clustering algorithm for mixed type attributes in large database environment. In *Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 263–268.

Conn, A. R., N. I. Gould, and P. L. Toint (2000). *Trust region methods*, Volume 1. Siam.

Dua, D. and C. Graff (2017). UCI machine learning repository.

Ester, M., H.-P. Kriegel, J. Sander, X. Xu, et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, Volume 96, pp. 226–231.

Fletcher, R. (2013). *Practical methods of optimization*. John Wiley & Sons.

Fraley, C. and A. E. Raftery (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association 97*(458), 611–631.

Fruehwirth-Schnatter, S. and H. F. Lopes (2018). Sparse bayesian factor analysis when the number of factors is unknown. arXiv preprint arXiv:1804.04231.

Ghahramani, Z., G. E. Hinton, et al. (1996). The em algorithm for mixtures of factor analyzers. Technical report, Technical Report CRG-TR-96-1, University of Toronto.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics 27*(4), 857–871.

Huang, Z. (1997). Clustering large data sets with mixed numeric and categorical values. In *Proceedings of the 1st pacific-asia conference on knowledge discovery and data mining,(PAKDD)*, pp. 21–34. Singapore.

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery 2*(3), 283–304.

Jogin, M., M. Madhulika, G. Divya, R. Meghana, S. Apoorva, et al. (2018). Feature extraction using convolution neural networks (cnn) and deep learning. In *2018 3rd IEEE*

International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), pp. 2319–2323. IEEE.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE 78*(9), 1464–1480.

Maclaurin, D., D. Duvenaud, and R. P. Adams (2015). Autograd: Effortless gradients in numpy. In *ICML 2015 AutoML Workshop*, Volume 238, pp. 5.

McLachlan, G. J. and D. Peel (2000). *Finite mixture models*, Volume 299 of *Probability and Statistics – Applied Probability and Statistics Section*. New York: Wiley.

McLachlan, G. J., D. Peel, and R. W. Bean (2003). Modelling High-Dimensional Data by Mixtures of Factor Analyzers. *Computational Statistics and Data Analysis 41*(3-4), 379–388.

Melnykov, V., R. Maitra, et al. (2010). Finite mixture models and model-based clustering. *Statistics Surveys 4*, 80–116.

Moustaki, I. (2003). A general class of latent variable models for ordinal manifest variables with covariate effects on the manifest and latent variables. *British Journal of Mathematical and Statistical Psychology 56*(2), 337–357.

Moustaki, I. and M. Knott (2000). Generalized latent trait models. *Psychometrika 65*(3), 391–411.

Nenadic, O. and M. Greenacre (2005). Computation of multiple correspondence analysis, with code in r. Technical report, Universitat Pompeu Fabra.

Niku, J., W. Brooks, R. Herliansyah, F. K. Hui, S. Taskinen, and D. I. Warton (2019). Efficient estimation of generalized linear latent variable models. *PloS one 14*(5), 481–497.

Pagès, J. (2014). *Multiple factor analysis by example using R.* CRC Press.

Patil, D. D., V. Wadhai, and J. Gokhale (2010). Evaluation of decision tree pruning algorithms for complexity and classification accuracy. *International Journal of Computer Applications 11*(2), 23–30.

Philip, G. and B. Ottaway (1983). Mixed data cluster analysis: an illustration using cypriot hooked-tang weapons. *Archaeometry 25*(2), 119–133.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics 20*, 53–65.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics 6*(2), 461–464.

Viroli, C. and G. J. McLachlan (2019). Deep gaussian mixture models. *Statistics and Computing 29*(1), 43–51.

Virtanen, P., R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods 17*, 261–272.

Wei, G. C. and M. A. Tanner (1990). A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *Journal of the American statistical Association 85*(411), 699–704.

Wold, S., M. Sjöström, and L. Eriksson (2001). Pls-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems 58*(2), 109–130.

1.3. Application to the determination of phytoplankton ecological niches

Data presentation

The previous paper has introduced the MDGMM models along with additional visualization tools useful to interpret the clustering output. The MDGMM is here applied to characterize the link existing between important environmental variables and piconanophytoplankton cells. To do so, we rely on the SNO SOMLIT data (Service National d'Observation - Service d'Observation en Milieu Littoral) https://www.somlit.fr/. The SNO SOMLIT is a French National marine monitoring program which has started in 1996 and relies on a network of eleven marine stations based in the Mediterranean sea, the Gironde River, the Atlantic Ocean, and the Channel (see Figure 2.1). It collects surface hydrobiological variables ("hydro" series) such as the temperature, salinity, pH, dissolved oxygen, nutrient concentrations (phosphate, nitrite, ammonium, phosphate, dissolved silica), suspended matter, chlorophyll-a, particulate organic carbon (POC), and nitrogen (PON). Cytometric phytoplankton functional group data (in the "piconano" series) are also acquired at the surface or in the epipelagic zone, and deal with five of the presented cPFGs: Orgpicopro, Redpicopro, Redpicoeuk, Rednano, and Orgnano. Finally, temperature, fluorescence, salinity, and PAR data ("CTD" series) are collected over the water column using Conductivity-Temperature-Depth sensors (CTD). The data are acquired on a bimensual basis.

The MDGMM clustering was performed on the five cPFGs tracked by the piconano series, and on the temperature, salinity, and nutrients: ammonium (NH_{4^+}), nitrates (NO_{3^-}), nitrites (NO_{2^-}), phosphates ($PO_{4^{3-}}$), of the hydro series. The corresponding dataset contained observations collected from 2009 to 2021. To take the spatial and dependence structure of the data into account, three additional variables were included. First, a variable presented the month during which the sample was added to encompass the seasonality of the data ("MONTH" variable, 12 modalities). Besides, the ocean/sea/river of origin was included ("ZONE" variable, 4 modalities). Finally, the depth at which the observation was performed was made part of the dataset ("DEPTH" variable, 3 modalities) to include the vertical spatial dependency (all depth levels were not available for each station). The final dataset is hence composed of eleven continuous variables, one ordinal variable, and two categorical variables, and counts 2700 observations. The M1DGMM minimal architecture was used to obtain the most stable results.

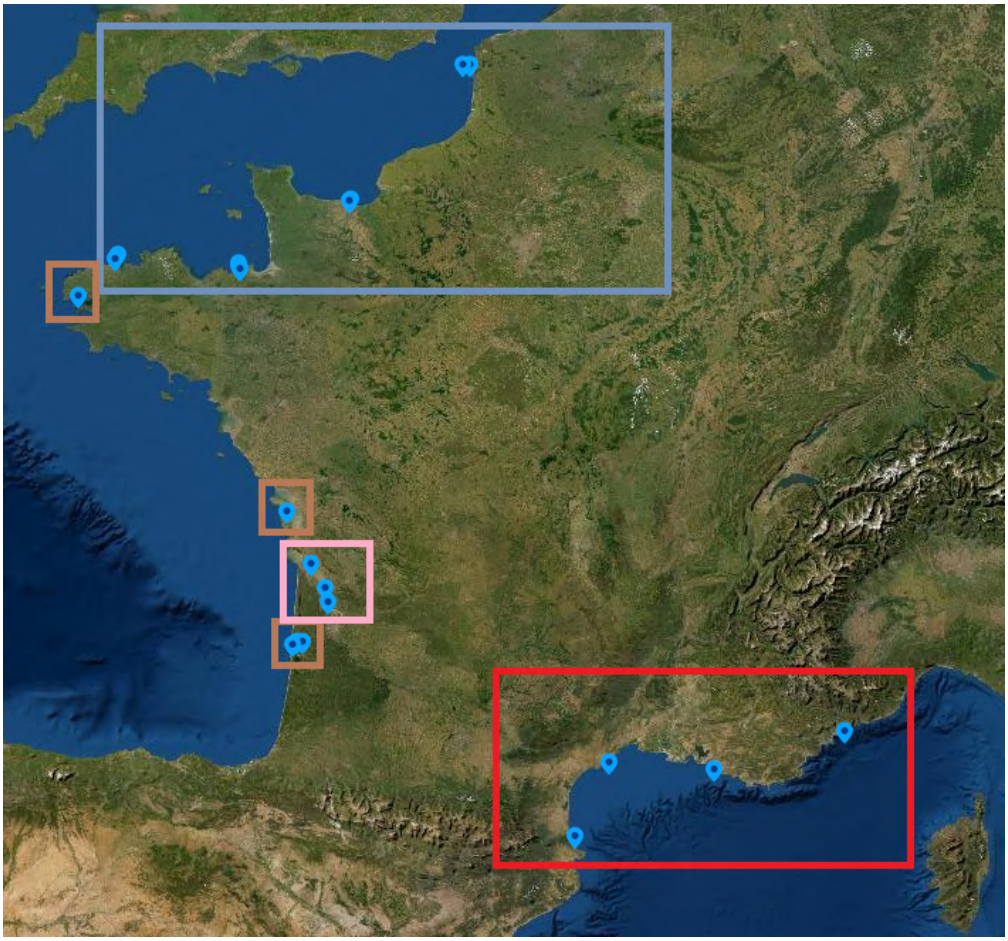


Figure 2.1. – Maps of the eleven SOMLIT stations and the associated zones: The Mediterranean Sea stations are denoted by a red rectangle, the Atlantic stations are in brown, the Gironde River stations in pink and the Channel-related stations in blue (based on the Leaflet map library).

Results

The MDGMM captures the correlations existing in the dataset thanks to its latent space. One can trace back the construction of the latent space by computing the associations between the original variables and the two newly created latent dimensions, as represented in Figure 2.2 (see Appendix A for more details on associations matrices). The variables that contribute the most to the latent space were the "ZONE", "MONTH", "DEPTH" variables (along with the "Redpicoeuk" variable). Hence, the spatio-temporal dependence was the most powerful structuring signal identified by the MDGMM. More precisely, the contribution of "ZONE" and "DEPTH" was higher than the contribution of "MONTH": the inter-location variability was more discriminant than inter-seasonal variability.

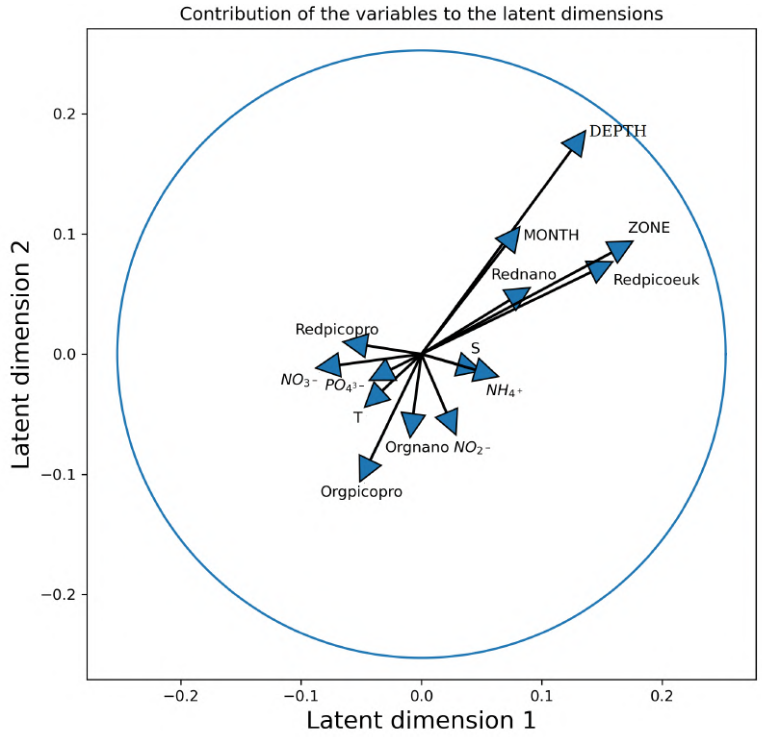


Figure 2.2. – Contributions of the original dataset variables to the MDGMM latent dimensions. The biggest the arrow, the most contributing the original variable is. Two arrows sharing the same sign and direction carry similar pieces of information concerning the latent space. The association between a continuous variable and each latent dimension lies in [-1,1], while it lies in [0,1] for the association of a non-continuous variable with the latent dimensions. Thus, the sign of the arrow is directly interpretable for continuous variables but not for the non-continuous variables ("ZONE", "MONTH", and "DEPTH"): only the norm and direction have a direct interpretation.

Concerning the environment characterization, the most saline environments were the richest in ammonium and nitrite but poorest in nitrate and phosphate. Seawater temperature and salinity were not expressed on the same latent dimensions, as the corresponding arrows were nearly orthogonal. It underlined the diversity of temperature/salinity configurations. Finally, the zones which were rich in Redpicoeuk and Rednano were generally poorer in Orgpicopro. Conversely, the abundances in Redpicopro and Orgnano were relatively independent of each other in the latent space.

The temporal and spatial dependence structured the latent space and hence had a significant influence on the data clustering process as presented in Figure 2.3. The model found two distinct clusters with several intermediate points. It separated mainly the Mediterranean data from the three other data sources (Figure 2.3 a and b). Moreover, the ecological assemblages found in the Gironde River were more similar

to the Channel assemblages than to the Atlantic Ocean and the Mediterranean Sea assemblages. The shallowest samples were located mainly in the center bottom of the latent space, whereas intermediate-depth samples were represented mainly in the top left area, and the deepest samples were in the right part of the latent space (Figure 2.3 c). Finally, data acquired in the late months of the year were plotted at the bottom of the latent space, whereas early sampling months could be found at the top of the latent space (Figure 2.3 d).

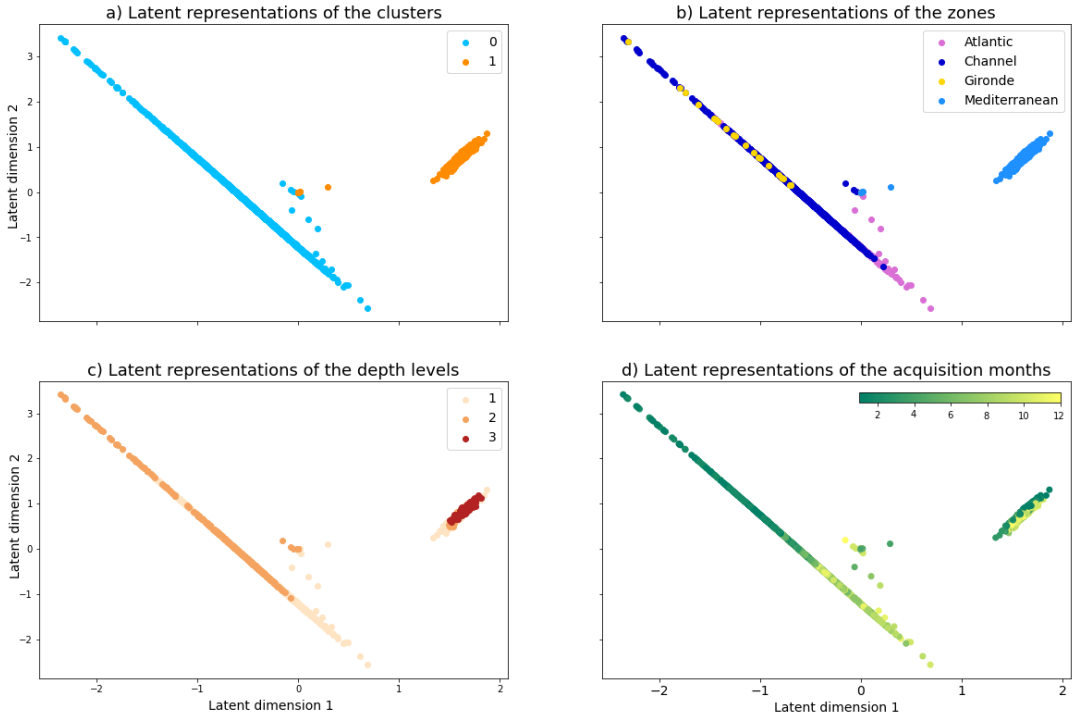


Figure 2.3. – Latent representation of the SOMLIT data. a) Latent representation colored by MDGMM cluster number (the model identifies two clusters here, numbered 0 and 1). b) Latent representation of the data colored by the zone of belonging ("ZONE" variable). c) Latent representation of the observations colored by sampling depth ("DEPTH" variable). d) Latent representation of the data colored by sampling month ("MONTH" variable), 1 corresponds to January and 12 to December.

The fundamental Hutchinson ecological niches (Hutchinson 1957) refer to the set of conditions necessary for some organisms or species to exist. Here we have made a slight abuse of this concept as we have extended it to the functional group level. The ecological conditions that generated the highest abundances for each cPFG reflect the optimal conditions for a cPFG to thrive and hence give information about the ecological niche of a given cPFG. On the contrary, the conditions leading to the lowest abundances could be identified as detrimental to a given cPFG development. Figures 2.4 and 2.5 represent the top and low 5% abundances observed in the SOMLIT data for the Orgpicopro and Redpicoeuk. These two groups have been chosen as examples, but

similar plots could be generated for the other groups. The representations of the best and worst conditions for these two cPFGs were located in well-delimited and dense regions of the latent space. This confirmed the proper identification of the ecological niches for these two groups by the MDGMM. The ecological niches of Redpicoeuk and Orgpicopro were the opposite in the Mediterranean sea, as the best conditions for Orgpicopro corresponded to the worst conditions for the Redpicoeuk (Figures 2.4 a and 2.5 a). This was not the case in the Channel and in the Atlantic where the two groups shared similar niches.

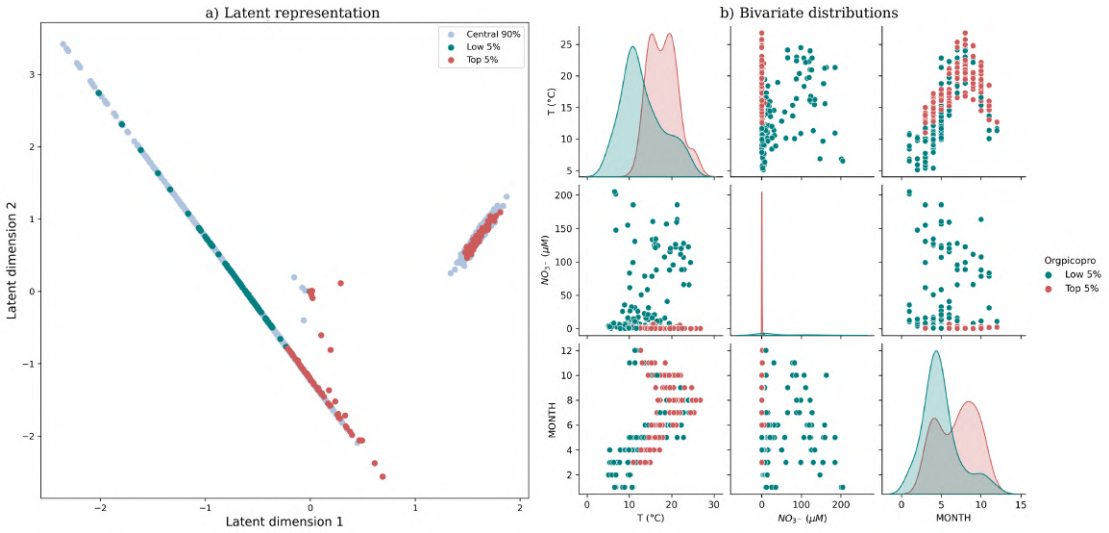


Figure 2.4. – Orgpicopro distribution representations. a) Representation in the latent space of the lowest 5% abundances, central 90% abundances and top 5% abundances. b) Bivariate distribution of the temperature, nitrate concentration and month broken down between the lowest 5% and top 5% Orgpicopro abundances. The diagonal plots correspond to the marginal distributions of each "environmental" variable for the top 5% (red distribution) and lowest 5% (blue distribution) Orgpicopro abundances.

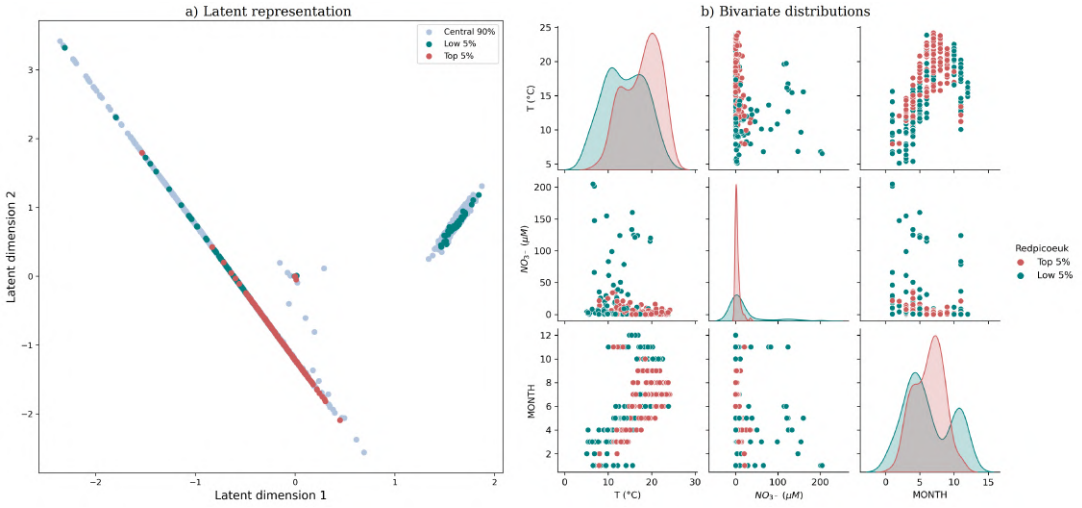


Figure 2.5. – Redpicoeuk distribution representations. a) Representation in the latent space of the lowest 5% abundances, central 90% abundances and top 5% abundances. b) Bivariate distribution of the temperature, nitrate concentration and month broken down between the lowest 5% and top 5% Redpicoeuk abundances. The diagonal plots correspond to the marginal distributions of each "environmental" variable for the top 5% (red distribution) and lowest 5% (blue distribution) Redpicoeuk abundances.

In all the considered zones, the Atlantic Ocean, the Mediterranean sea, the Gironde River, and the Channel, the Redpicoeuk and Orgpicopro were most abundant in warm and poor in nitrate waters. Yet, the Redpicoeuk ecological niche was located in warmer and richer in nitrate waters than the Orgpicopro ecological niche (Figures 2.4 b and 2.5 b). The optimal abundance months for Redpicoeuk span from March to October whereas Orgpicopro are more numerous from April to December.

To summarize, spatial dependence was one of the most discriminant pieces of information between the SOMLIT observations. Running the MDGMM without spatial and temporal dependence resulted in a less structured latent representation due to the introduced omitted variable bias (result not shown). The Mediterranean Sea was the most differentiated zone whereas the Atlantic Ocean, the Gironde River, and the Channel could be regarded as offering closest environment-phytoplankton assemblages. The spatial dependence was identified as stronger than the temporal dependence, which is certainly due to partial coverage of this dependence source by the data. The temporal dependence could be conceptually broken up into three non-exclusive categories: the inter-annual, the inter-seasonal, and the intra-seasonal dependence patterns. The annual dependence structure seemed not to be properly captured by the model as no trend from one year to another was captured by the MDGMM, even when the corresponding variable was explicitly included in the clustering process (results not shown). This might be explained by the moderate number of sampling years (12 years). The inter-seasonal patterns were correctly accounted

for in the latent space (Figure 2.3 d). It highlights the strong pico-nanophytoplankton seasonal dynamics, ruled out by nutrient concentrations and temperature, but also by light and nitrogen-fixing organisms not included here (Otero-Ferrer et al. 2018; Fowler et al. 2020; Farnelid et al. 2021). Finally, the intra-seasonal dependence could not be properly addressed due to the data sampling frequency of 15 days. The next chapter is dedicated to this question.

Concerning the ecological niches themselves, the most oligotrophic environments in terms of nitrates and phosphate (the Mediterranean Sea and the Atlantic Ocean) favored small cyanobacteria (Redpicopro and Orgpicopro), whereas less oligotrophic oceanic environments (e.g. in the Channel) favored the biggest phytoplankton functional groups. This is consistent with the works of Glibert et al. 2016 and Otero-Ferrer et al. 2018 that found that picophytoplankton (Redpicoeuk) were favored by higher nitrate concentrations compared to picocyanobacteria such as Orgpicopro. The dominance of Redpicoeuk in the late summer is not in contradiction with the predominance of picoeukaryotes in autumn as observed by Pulina et al. 2017 in a Sardinian lagoon. Yet, the bimodal Orgpicopro temporal niche in late spring and autumn contradicts the finding of these authors. These patterns are however consistent with the spring and autumn Orgpicopro blooms evidenced in the Marseille Bay in Section 3 of Chapter 3. This difference could thus simply reflects differentiated bloom patterns between the different sea, ocean, and river, that the model tried to account for.

2. Prospecting environmental changes with MIxed data Augmentation MIxture

The MDGMM provides an interpretable information summary and clustering process through its use of latent space. The low performance difference between the M1DGMM and M2DGMM in Section 1.2 highlighted the fact that the conditional independence assumption is not a so strong hypothesis. In other words, the latent variable $z^{(1)}$ takes into account the majority of the dependence structure between the original variables. Hence, this latent structure is highly informative and could be used to generate synthetic observations that present the same dependence structure as the original data.

In this section, the MIxed data Augmentation MIxture (MIAMI) is introduced and applied to American Census data to reproduce missing multivariate modalities (more results are available in Appendix B). In a second time, prospective analyses are performed on the SOMLIT data presented in the previous section.

2.1. MIAMI: presentation

MIAMI: MIxed data Augmentation MIxture *

Robin Fuchs[1], Denys Pommeret[1,2], and Samuel Stocksieker[2]

[1] Aix Marseille Univ, CNRS, Centrale Marseille, I2M, Marseille, France
robin.fuchs@univ-amu.fr; samuel.stocksieker@univ-amu.fr
[2] Lyon 1 Univ, ISFA, Lab. SAF EA2429, F-69366, Lyon, France.
denys.pommeret@univ-amu.fr

Abstract. Performing data augmentation for mixed datasets remains an open challenge. We propose an adaptation of the Mixed Deep Gaussian Mixture Models (MDGMM) to generate such complex data. The MDGMM explicitly handles the different data types and learns a continuous latent representation of the data that captures their dependence structure and can be exploited to conduct data augmentation. We test the ability of our method to simulate crossings of variables that were rarely observed or unobserved during training. The performances are compared with recent competitors relying on Generative Adversarial Networks, Random Forest, Classification And Regression Trees, or Bayesian networks on the UCI Adult dataset.

Keywords: Mixed data · Data Augmentation · Mixture Models · Unbalanced data

1 Introduction

Data augmentation is a powerful methodology to deal with unbalanced data, with data containing missing values, as well as to produce synthetic and anonymous datasets. Most data augmentation approaches are designed for a single data type: either continuous or non-continuous, with a particular focus on the continuous case. In the continuous data framework, the main methods are k-nearest neighbors (kNN), perturbation methods adding random noises to the data [13, 16], methods based on the dependence structure obtained by modeling joint distribution or copulas [25], Gaussian Mixture Models [24], Generative Adversarial Networks (GAN) [14, 22], and Variational Autoencoder (VAE) [17]. For non-continuous data, methods often rely on kNN [7] using adapted metrics, Classification And Regression Trees (CART) or Random Forest [21].

Methods dealing with each data type separately aim at capturing the dependence structure of the observations and using it to generate data. However, when the data are mixed, performing data augmentation can be challenging since the approaches have to simultaneously model categorical, binary, ordinal,

* Granted by the Research Chair DIALog under the aegis of the Risk Foundation, an initiative by CNP Assurances

discrete, and continuous data. These mixed distributions often contain multimodal marginal densities, non-standard continuous distributions, and unbalanced modalities for the binary, categorical and ordinal variables.

There exist several recent works in the literature dedicated to the problem of mixed data. Some of them are adaptations of the previously cited methods, such as kNN with a specific distance [3], probabilistic models based on conditional copulas as synthetic data generators [10], or conditional GANs [4, 27]. This generalization of GANs was introduced to overcome the fact that traditional GANs had difficulties reproducing complex distributions such as multimodal distributions (e.g. mixture distributions) or modeling entire distributions and to keep the full dependence structure of the data [15]. The Bayesian framework also constitutes a powerful family of methods to deal with mixed data. Dirichlet process mixtures can be used as latent spaces to generate data [9, 20]. The Bayesian framework can also be combined with Gaussian copulas to generate fully-synthetic mixed data [5]. Yet, one of the main difficulties of Bayesian models remains in the choice of the priors to reflect the underlying model and the complexity of the dependence structure.

In this work, we introduce an approach based on the Mixed Deep Gaussian Mixture Model (MDGMM) [6]. The MDGMM learns a continuous representation of the dataset and can be inverted to generate pseudo-observations. The proposed methodology keeps the dependence structure of mixed datasets in a flexible way considering the flexible parametric distribution of the latent space which relies on Deep Gaussian Mixture Models [26]. Furthermore, all mixed data types are handled explicitly, especially the ordinal data type that is often assimilated to categorical or continuous data by competitor methods. The MDGMM is hence used as a data generator and coupled with an acceptation-rejection procedure to select observations presenting the desired characteristics. Our main objective is here to reconstruct unobserved regions of the mixed multivariate support of the data. We call this complete procedure “MIAMI", standing for “MIxed data Augmentation MIxture".

2 MDGMM brief presentation

The MDGMM is an unsupervised multi-layer model designed for mixed data clustering introduced by [6]. The mixed data are mapped into a continuous latent space using a Generalized Linear Latent Variable Model (GLLVM) [2, 18, 19]. The latent space is a Deep Gaussian Mixture Model [26] which enables the latent space to capture a broad range of possible distributions. More formally, denoting $Y = (Y_1, \cdots, Y_n)$ the n observations of dimension p, we have for $i \in [1, n]$:

$$\begin{cases} Y_i \to z_i^{(1)} \text{ through GLLVM link via } (\lambda^{(0)}, \Lambda^{(0)}) \\ z_i^{(1)} = \eta_{k_1}^{(1)} + \Lambda_{k_1}^{(1)} z_i^{(2)} + u_{i,k_1}^{(1)} \text{ with probability } \pi_{k_1}^{(1)} \\ ... \\ z_i^{(L-1)} = \eta_{k_{L-1}}^{(L-1)} + \Lambda_{k_{L-1}}^{(L-1)} z_i^{(L)} + u_{k_{L-1}}^{(L-1)} \text{ with} \\ \text{probability } \pi_{k_L}^{(L-1)} \\ z_i^{(L)} \sim \mathcal{N}(0, I_{r_L}), \end{cases} \tag{1}$$

where "GLLVM link" refers to the link functions relating the original mixed variable space to the continuous latent space. These link functions $f(Y_i|z_i^{(1)}, \Theta)$ are part of exponential families and the parameters Θ are learned during training (more details are given in [6]). To illustrate the possible link functions, if the jth component of the ith observation, Y_{ij}, is a count variable, one can choose a a Binomial distribution:

$$f(Y_{ij}|z^{(1)}, \Theta) = \binom{n_j}{Y_{ij}} h(z^{(1)})^{Y_{ij}} (1 - h(z^{(1)}))^{n_j - Y_{ij}}, \tag{2}$$

with n_j the upper bound of the count variable support. Other examples of link functions are given in [2].

In the simulations, the ordinal variables are linked to the latent space using ordered multinomial distributions, the categorical variables using unordered multinomial distributions, the count variables using Binomial distributions, the binary variables with Bernoulli distributions, and the continuous variables with Gaussian distributions.

The graphical model of the MDGMM described in (1) is presented in Figure 1. This architecture was introduced as the M1DGMM in [6]. There exists a second architecture, M2DGMM, which merges the embeddings learned separately on continuous and non-continuous variables. We have chosen a simple one-layer deep M1DGMM architecture with a two-dimensional latent space and $K_1 = 4$ components, which has proven to be the more stable architecture [6] and will ensure to obtain more reproducible results.

3 Data augmentation procedure

The latent representation of the data is first determined by training the MDGMM on the data Y. The parameters learned are hereafter denoted by a tilde. The model is then inverted to generate pseudo-observations and only the observations with the desired characteristics are kept.

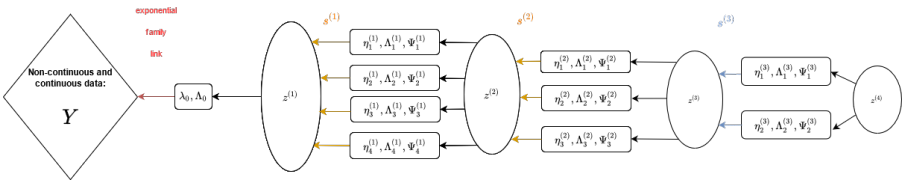


Fig. 1. Graphical model of a M1DGMM

Model training is described in the original MDGMM and enables drawing $z^{(1)}$ samples from a $DGMM(\tilde{\Theta})$ distribution. The model is then inverted using Bayes rule:

$$f(Y|\tilde{\Theta}) = \frac{f(\tilde{z}^{(1)}|\tilde{\Theta})f(Y|\tilde{z}^{(1)},\tilde{\Theta})}{f(\tilde{z}^{(1)}|Y,\tilde{\Theta})} \tag{3}$$

$$\propto f(\tilde{z}^{(1)}|\tilde{\Theta}) \prod_{j=1}^{p} f(Y_j|\tilde{z}^{(1)},\tilde{\Theta}). \tag{4}$$

The passage from (3) to (4) comes from the fact that, by construction, there is mutual independence between the original variables given the latent variable. This means that the latent representation captures all the dependence structure existing in the original dataset, which is a suitable feature to perform data augmentation in the mixed data case.

Let C be the set of wanted characteristics, being for example a region of the original variable space, missing crossings of variables, or unbalanced modalities for non-continuous variables. One can simulate the N^* pseudo-observations presenting the characteristics C using the procedure described in Algorithm 1.

Algorithm 1 MIAMI

Input: Y, N^*, C, m_O, m_1.
Initialize $s = 0$.
repeat
  Generate m_0 draws of $z^{(1)}$ from $DGMM(\tilde{\Theta})$.
  Use these draws to sample m_1 pseudo-observations from $f(Y^*|\tilde{z}^{(1)},\tilde{\Theta})$ using (4).
  for $i = 1$ **to** m_1 **do**
    if Y_i^* satisfies condition C **then**
      Add Y_i^* to Y^*.
      $s = s + 1$.
    end if
  end for
until $s \geq N^*$
Output: N^* draws of Y^*.

In this algorithm, "Y^* satisfies condition C" means that the pseudo-observations (Y^*) present the wanted characteristics (C). In this sense, Algorithm 1 can be viewed as an oversampling method creating pseudo-observations with the desired features. The N^* pseudo-observations can be simulated by changing the number of copies of the latent variables $z^{(1)}$, m_0, or the number of pseudo-observations Y^*, m_1, to draw from each $z^{(1)}$. More $z^{(1)}$ draws ensure better coverage of the latent space while more Y^* draws per $z^{(1)}$ gives more information about the link existing between each latent point and the original variable space.

4 Numerical illustration

4.1 Competitors

We propose to compare our approach with four recent competitors:

- *CTGAN* [1] is part of the SDV project and relies on a GAN-based Deep Learning data synthesizer to deal with continuous as well as categorical data.
- *Synthpop* proposed in the synthpop package in R [21]. Operating in a non-parametric framework, Synthpop generates the synthetic dataset sequentially by using either a CART procedure or a Random Forest (RF) approach. It is suitable for continuous as well as categorical, ordinal, and binary data. We consider both approaches, namely SynthPop-CART and SynthPop-RF, as competitors.
- *DataSynthesizer* [23] in the DataSynthesizer package in Python captures the underlying correlation structure between the different attributes through a Bayesian network and then draws samples from this model. It is suitable for continuous as well as categorical or binary data.

It is worth pointing out that we had also tested extensions of the SMOTE algorithm [3]: the SMOTE-NC algorithm with the HEOM distance and the so-called Adasyn algorithm [8]. However, these methods obtained much worse results than the other competitors on this dataset. Thus, their results are not shown here.

4.2 Evaluation metrics

The model performances are here evaluated graphically and by using properly defined metrics. First, the dependence structure between couples of variables can be graphically assessed using Associations Matrices (AM) which are a generalization of correlations matrices for mixed data. In AM, the standard correlations are used to compare pairs of ordered variables (continuous, ordinal, and count variables), correlation ratios are used to compare an ordered variable with a non-ordered variable, and the Cramer's V to compare two non-ordered variables.

Secondly, the performances are also measured by three metrics:

- The association distance which is the Mean Relative Absolute Errors (MRAE) between the test and the generated datasets obtained by summing the absolute relative differences between the values of their association matrices.
- The MAE (Mean Absolute Error) between proportions for binary and categorical variables.
- The Kullback-Liebler divergence between the multivariate continuous distributions of the test and generated datasets.

The association distance hence summarizes how well each method captures the dependence structure, the MAE the quality of the marginal distributions reconstruction for categorical and binary variables, and the Kullback-Liebler a pseudo-distance between the multivariate continuous distributions.
The presented results are obtained over ten runs for each competitor and the formulas of the MAE and Kullback-Liebler divergence are given in Appendix.

4.3 Dataset

We test our approach on the Adult Census Income data. This dataset contains weighted census data extracted from the 1994 and 1995 current population surveys conducted by the U.S. Census Bureau [11]. The dataset contains n=32.561 observations and is composed of three continuous variables, six categorical variables, two binary variables, and three ordinal variables as follows A detailed list of the variables is given in Appendix according to the UCI documentation.

4.4 Experimental designs

In this work, the ability of the competitor models to generate observations presenting a given combination of two and three variables is tested. This combined modality is either weakly present (10 observations, called "Unbalanced design" hereafter) or completely missing (hereafter "Absent design") in the training set. We have then four designs:

- Absent for a bivariate modality (“Bivariate Absent"),
- Absent for a trivariate modality (“Trivariate Absent"),
- Unbalanced design for a bivariate modality (“Bivariate Unbalanced"),
- Unbalanced design for a trivariate modality (“Trivariate Unbalanced").

The bivariate modality is in our case, women of more than 60 years old (age>60 & sex==“Female") and the trivariate modality is widowed women of more than 60 years old (age>60 & sex==“Female" & Marital.status==“Widowed").

The stability of the methods is evaluated using a 10-fold approach: for each experimental design, ten training sets of 1000 observations are drawn from the original dataset. The test sets are composed of the observations presenting the desired modality and that are not included in the train set. The number of pseudo-observations presenting the desired crossing of variables N^* to draw is 200 for each competitor.

4.5 Results analysis

We restrict our attention here to the most representative results.

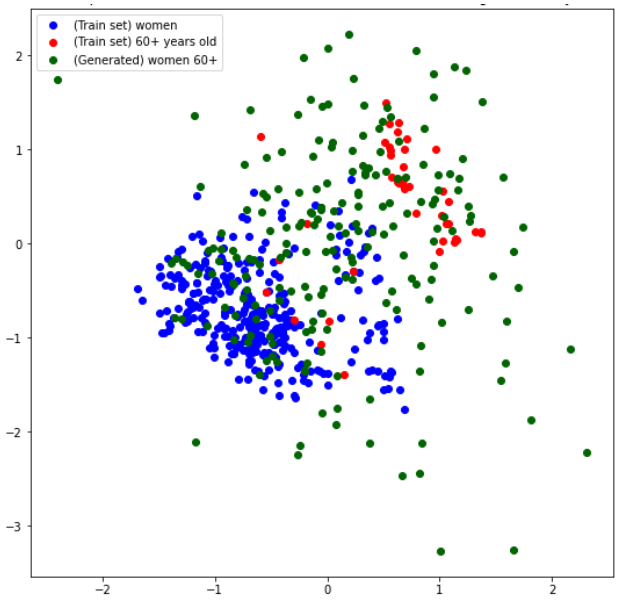


Fig. 2. Latent representation of women (in blue) and 60+ years old individuals (in red) coming from the train dataset, and women of 60+ years old generated by MIAMI (in green) in the Absent bivariate design.

As shown in Figure 2, the characteristics of the original dataset are well mapped into the latent space ($z^{(1)}$), which is the first layer of the MDGMM as illustrated in Figure 1. Women and individuals of more than 60 years old are represented in two different and coherent zones. The generated individuals which present both characteristics are mainly generated near these two zones denoting that this global area well encodes the modality crossing.

Concerning the reproduction of the dependence structure through the association distances (Figure 3), CTGAN generally obtains the best performance followed by MIAMI, DataSynthesizer, and SynthPop-CART. MIAMI is especially competitive on the Absent bivariate and Unbalanced trivariate designs. The SynthPop-RF approach outperforms most methods on trivariate designs but fails to capture the dependence structure of the bivariate designs.

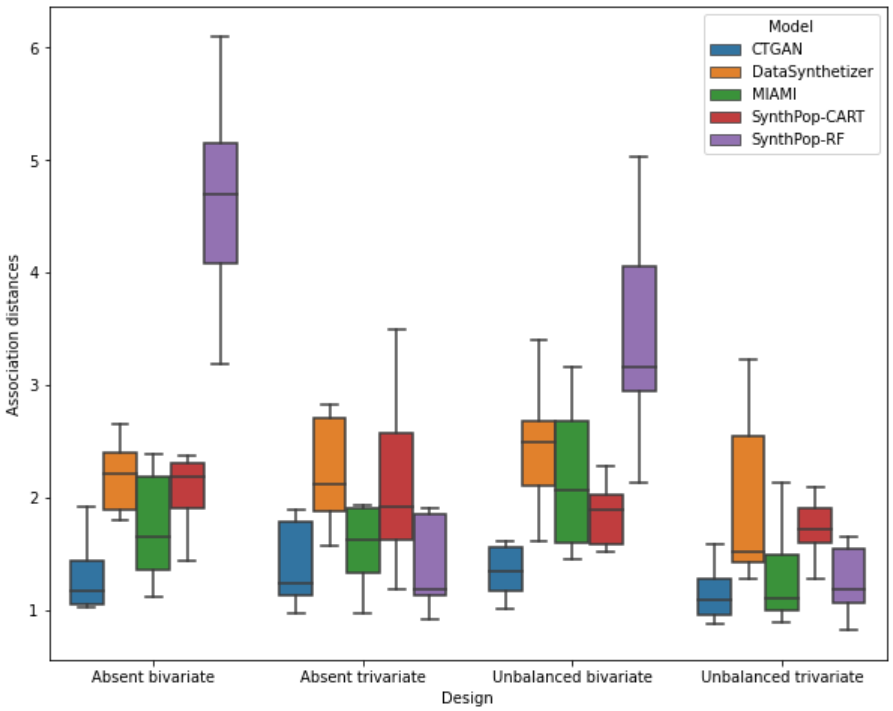
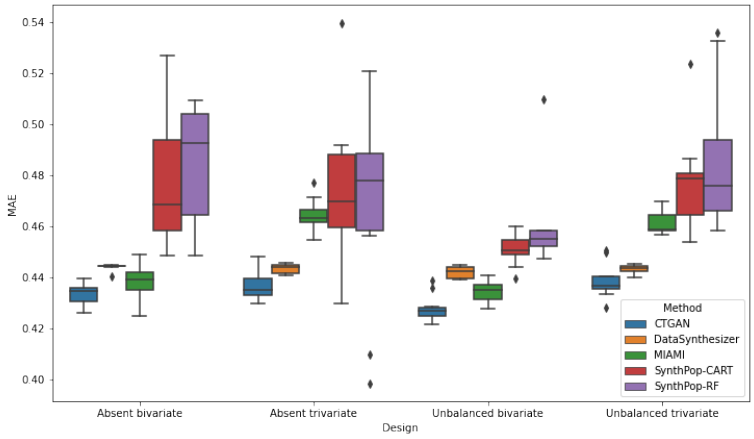


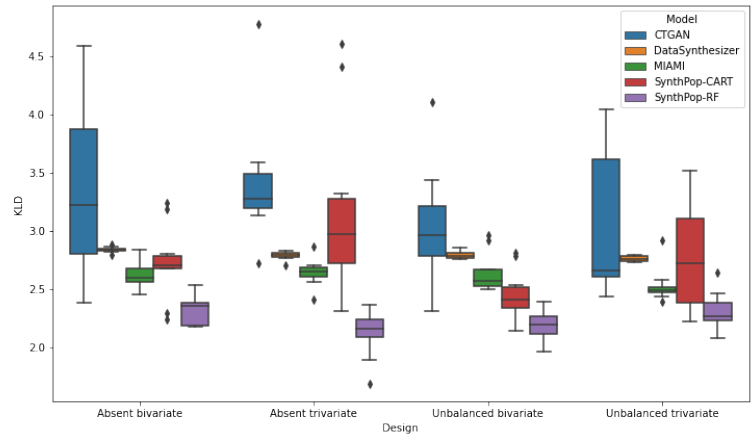
Fig. 3. Association distance for the four designs. The lower the distance, the best the dependence structure is reproduced.

CTGAN and MIAMI still obtain the best results for the reconstruction of categorical and binary variables but DataSynthesizer also performs well (Figure 4(a)). The SynthPop methods present a much higher variance than the other competitors for three designs out of four, while MIAMI and CTGAN obtain comparable variances and DataSynthesizer a much lower variance.

Besides, the continuous variables are best accounted for by SynthPop-RF, MIAMI, and SynthPop-CART (Figure 4(b)). The distributions generated by CTGAN are furthest from the test ones and show more variability, especially in comparison with DataSynthesizer and MIAMI. This pattern can also be observed in Figure 5 representing the bivariate distribution of the age and flnwgt variables on the test dataset and for the observations generated by MIAMI and CTGAN. The continuous distributions reconstructed by MIAMI are more concentrated and the density maximum is closest to the one of the test dataset when compared to CTGAN.



(a) MAE



(b) Multivariate Kullback-Liebler

Fig. 4. MAE (a) and Kullback-Liebler divergence (b) between the test and the generated datasets for all designs

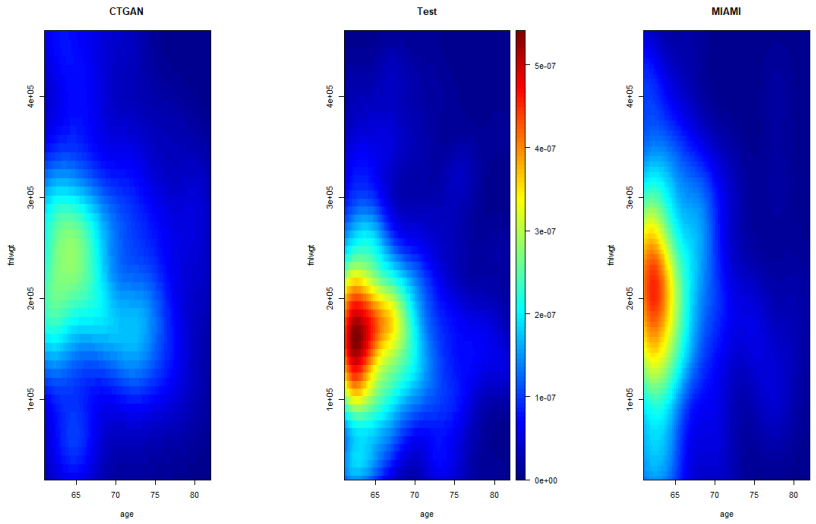


Fig. 5. Density estimation of the bivariate distribution (Fnlweight, Age) for CTGAN, on the test set, and for MIAMI for the Unbalanced bivariate design.

Figure 6 gives an illustration of the generation of the seven modalities of the variable “Marital Status" for the Bivariate Unbalanced and Absent designs. It can be seen that only MIAMI and DataSynthesizer manage to generate all the possible modalities in the first design. MIAMI is closest to the observations while DataSynthesizer creates unobserved modalities. CTGAN is concentrated on only one modality. In the Absent design, none of the methods manage to cover all the modalities while DataSynthesizer once again proposes an unobserved modality.

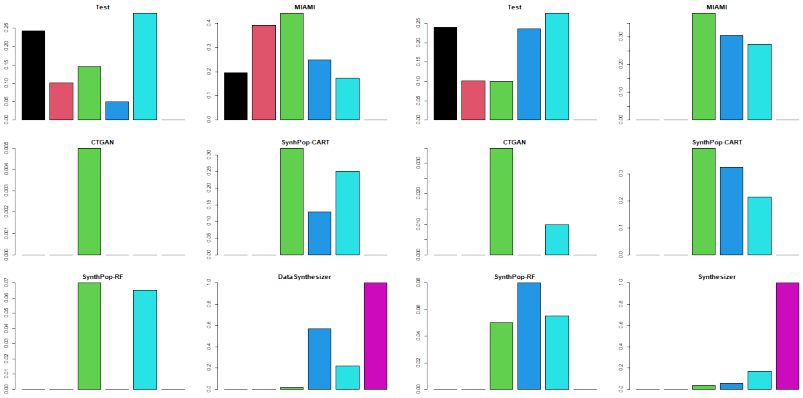


Fig. 6. Modality proportions for the Marital-status under the Bivariate Unbalanced design (left) and the Bivariate Absent design (right)

Finally, compared to the competitor methods, MIAMI gives associations close to the true associations existing in the test set but with a slightly lower intensity than in the test set (Figure 7). The associations between “Education.num" and "Occupation" or the one between the marital status and the "Relationship" variable are captured. These two associations are also well reproduced by SynthPop-CART, which tends however to create nonexistent associations between most variables. CTGAN has more difficulty in reproducing the original patterns. For concision purposes, the results of SynthPop-RF and DataSynthesizer are not presented. Indeed, SynthPop-RF exaggerates the associations existing between variables even more than SynthPop-CART, and DataSynthesizer fails to reproduce the main patterns of the test association matrix.

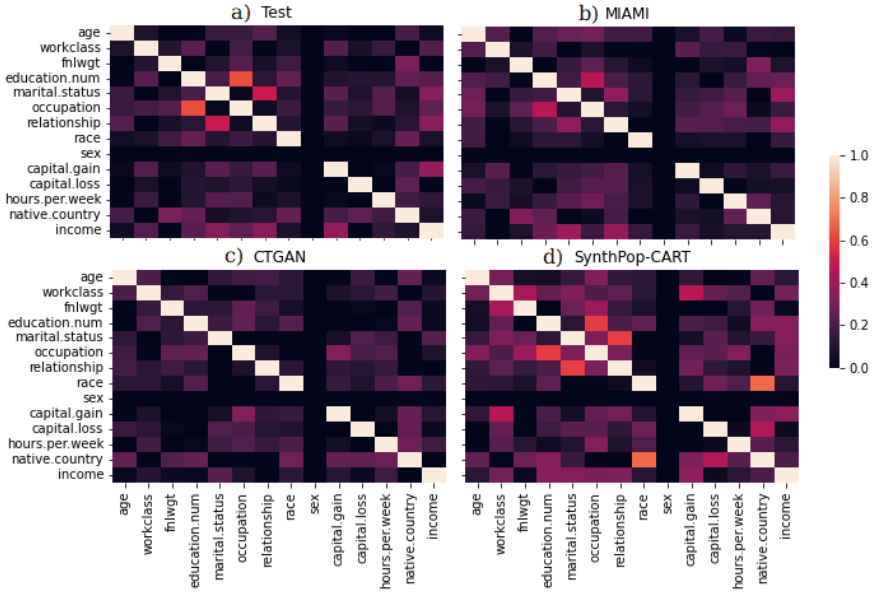


Fig. 7. Association matrices for women of more than 60 years old existing in the test set of the Unbalanced bivariate design (a), generated by MIAMI (b), by CTGAN (c) and by SynthPop-CART (d) .

The code to reproduce the results is available at `https://github.com/RobeeF/M1DGMM`.

5 Discussion and perspective

MIAMI is an algorithm dedicated to mixed data which oversamples desired areas of the sampling space while preserving the multivariate dependence structure of the data. Based on our numerical study, we conclude that MIAMI seems to reconstruct well the joint dependence of the mixed data, the univariate non-continuous distributions, and the multivariate continuous distributions. Its major competitor seems to be CTGAN in terms of dependence structure and non-continuous variable, and SynthPop-CART on the multivariate continuous distributions reconstruction.

The flexibility of its latent space enables MIAMI to properly reconstruct areas of high density for continuous variables and to generate a wide range of modalities for non-continuous variables while the other methods often reproduce only the most represented ones (Figure 6 and Figures 8-9 in Appendix). In this work, the latent space takes the form of a simple one hidden layer architecture. More

complex architectures and hyperparameters could be investigated in future works.

Furthermore, MIAMI here generated pseudo-observations using the entire latent space, and only the pseudo-observations presenting the desired characteristics were kept. However, as shown in Figure 2, some regions of the latent space are more likely than others to generate these pseudo-observations of interest. Hence, the sampling of the latent space could be adapted through the procedure to increase the generation/acceptation ratio. One could for instance exploit the link functions of the MDGMM or rely on Bayesian optimization methods. In the latter solution, the distribution of the latent variable would be taken as a prior and the task will be to estimate the posterior areas presenting high-acceptation rates.

References

1. Buuren, S.V., Brand, J.P., Groothuis-Oudshoorn, C.G., Rubin, D.B.: Fully conditional specification in multivariate imputation. Journal of Statistical Computation and Simulation **76**(12), 1049–1064 (2006)
2. Cagnone, S., Viroli, C.: A factor mixture model for analyzing heterogeneity and cognitive structure of dementia. AStA Advances in Statistical Analysis **98**(1), 1–20 (2014)
3. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: Synthetic minority over-sampling technique. J. Artif. Int. Res. **16**(1), 321–357 (jun 2002)
4. Engelmann, J., Lessmann, S.: Conditional wasserstein gan-based oversampling of tabular data for imbalanced learning. Expert Syst. Appl. **174**, 114582 (2021)
5. Feldman, J., Kowal, D.: A bayesian framework for generation of fully synthetic mixed datasets (2021)
6. Fuchs, R., Pommeret, D., Viroli, C.: Mixed deep gaussian mixture model: a clustering model for mixed datasets. Advances in Data Analysis and Classification pp. 1–23 (2021)
7. Guu, K., Lee, K., Tung, Z., Pasupat, P., Chang, M.: Retrieval augmented language model pre-training. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 3929–3938. PMLR (13–18 Jul 2020)
8. He, H., Bai, Y., Garcia, E.A., Li, S.: Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In: in: IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), IJCNN 2008. pp. 1322–1328 (2008)
9. Hu, J., Reiter, J.P., Wang, Q., et al.: Dirichlet process mixture models for modeling and generating synthetic versions of nested categorical data. Bayesian Analysis **13**(1), 183–200 (2018)
10. Kamthe, S., Assefa, S., Deisenroth, M.: Copula flows for synthetic data generation (2021)
11. Kohavi, R.: Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. p. 202–207. KDD'96, AAAI Press (1996)
12. Kullback, S., Leibler, R.A.: On information and sufficiency. Ann. Math. Statist. **22**(1), 79–86 (1951)

13. Lee, S.S.: Noisy replication in skewed binary classification. Computational Statistics and Data Analysis **34**(2), 165–191 (2000)
14. Liu, Y., Liu, Y., Liu, Y., Zhou, Y., Liu, X., Dong, F., Wang, C., Wang, Z.: Wasserstein gan-based small-sample augmentation for new-generation artificial intelligence: A case study of cancer-staging data in biology. Engineering (2019)
15. Lucic, M., Kurach, K., Michalski, M., Bousquet, O., Gelly, S.: Are gans created equal? a large-scale study. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. p. 698–707. NIPS'18, Curran Associates Inc., Red Hook, NY, USA (2018)
16. Menardi, G., Torelli, N.: Training and assessing classification rules with imbalanced data. Data Mining and Knowledge Discovery **28**, 92–122 (2012)
17. Moreno-Barea, F.J., Jerez, J.M., Franco, L.: Improving classification accuracy using data augmentation on small data sets. Expert Syst. Appl. **161**, 113696 (2020)
18. Moustaki, I.: A general class of latent variable models for ordinal manifest variables with covariate effects on the manifest and latent variables. British Journal of Mathematical and Statistical Psychology **56**(2), 337–357 (2003)
19. Moustaki, I., Knott, M.: Generalized latent trait models. Psychometrika **65**(3), 391–411 (2000)
20. Murray, J.S., Reiter, J.P.: Multiple Imputation of Missing Categorical and Continuous Values via Bayesian Mixture Models With Local Dependence. Journal of the American Statistical Association **111**(516), 1466–1479 (October 2016)
21. Nowok, B., Raab, G.M., Dibben, C.: synthpop: Bespoke creation of synthetic data in R. Journal of Statistical Software **74**(11), 1–26 (2016). `https://doi.org/10.18637/jss.v074.i11`
22. Park, N., Mohammadi, M., Gorde, K., Jajodia, S., Park, H., Kim, Y.: Data synthesis based on generative adversarial networks. Proc. VLDB Endow. **11**(10), 1071–1083 (jun 2018)
23. Ping, H., Stoyanovich, J., Howe, B.: Datasynthesizer: Privacy-preserving synthetic datasets. In: Proceedings of the 29th International Conference on Scientific and Statistical Database Management. SSDBM '17, Association for Computing Machinery, New York, NY, USA (2017)
24. Richardson, E., Weiss, Y.: On gans and gmms. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 31. Curran Associates, Inc. (2018)
25. Sun, Y., Cuesta-Infante, A., Veeramachaneni, K.: Learning vine copula models for synthetic data generation. In: AAAI (2019)
26. Viroli, C., McLachlan, G.J.: Deep gaussian mixture models. Statistics and Computing **29**(1), 43–51 (2019)
27. Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K.: Modeling tabular data using conditional gan. In: NeurIPS (2019)

2.2. Assessing environmental change effects on phytoplankton distribution

The previously learned latent space by the MDGMM is here used to generate synthetic observations with MIAMI and assess the potential shifts in phytoplankton communities triggered by structural environmental changes.
The effects of two changes were evaluated separately: an increase in seawater temperature and the phosphate (PO_4^{3-}) concentration in the Mediterranean sea. The Mediterranean Sea was chosen as it will be the main zone of interest of Chapter 3.

The first scenario is an increase of the seawater temperature by 2°C in winter. Such an increase could be experienced due to the global warming process by the end of the century as evidenced by Sakallı 2017 and Pastor et al. 2020 (even if most of this warming is expected to occur in early summer). The current 90% interval of fluctuation for the winter seawater temperature in the SOMLIT Mediterranean data is [11°C, 15°C]. Hence, synthetic data presenting temperatures between 13°C and 17°C in winter were generated to simulate the effect of a stand-alone rise in the temperature.

The second scenario deals with the effect of a pulse of nutrients in summer. In summer, the water column is stratified, and the surface is relatively poor in nutrients. As such, the effect of a nutrient pulse is assumed to be the strongest during this period. More precisely, an increase of 10% of the phosphate concentration was simulated from the actual 90% fluctuation interval of [0.01, 0.13] μM. This type of nutrient pulse is notably observed during coastal upwelling events studied in Chapter 3.

The first scenario of increasing seawater temperature in winter is represented in Figure 2.6. The increase in the seawater temperature significantly modified the mean abundances of all groups (Bonferroni-corrected Student-Welch at a 1% level) except the Redpicopro. This was especially the case for the Orgpicopro and the Orgnano which both experienced abundance rises by 52%, and for the Redpicoeuk abundance that grew by 39%. In general, the distributions were flatter for the simulated data than for the current data.

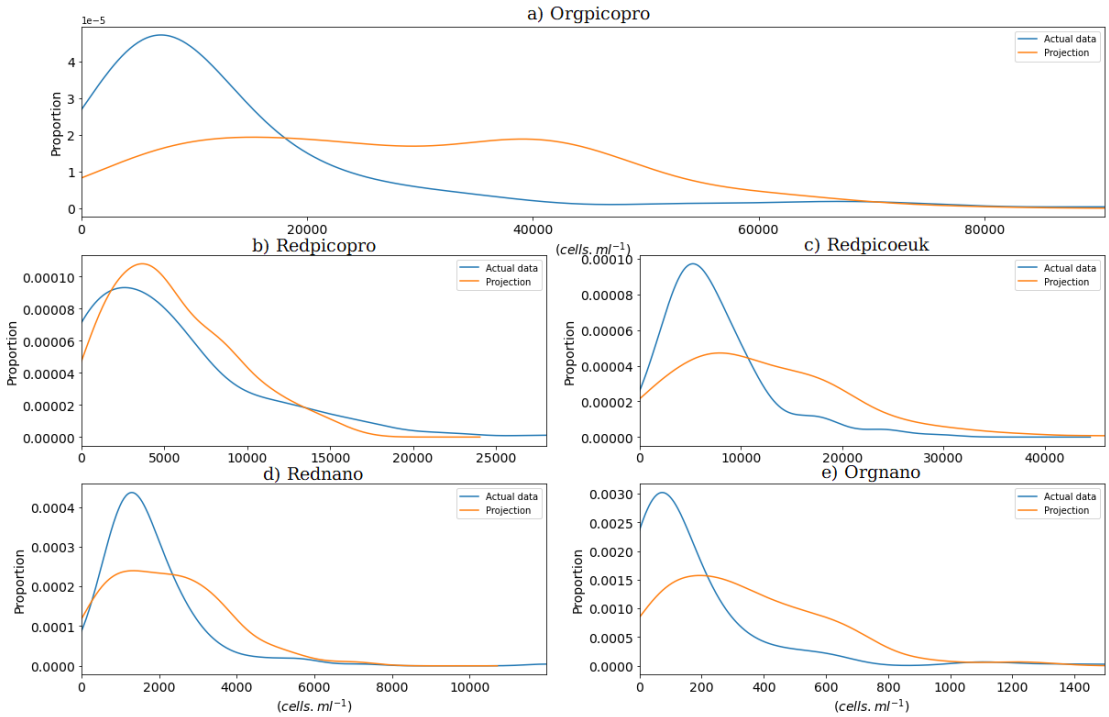


Figure 2.6. – Distribution of the functional group abundances in the actual SOMLIT data and for a simulated increase in water temperature by 2°C in winter (n = 180 in both cases). The distribution of the data is shown for the Orgpicopro (a), Redpicopro (b), Redpicoeuk (c), Rednano (d), and Orgnano (e). The mean of each cPFG actual and simulated distributions are significantly different (Bonferroni-corrected Student-Welch test, p<0.01).

The increase in phosphate also affected the cPFG abundances (Bonferroni-corrected Student-Welch at a 1% level for all cPFGs) as evidenced in Figure 2.7. The mean Orgpicopro abundance was curbed by 18% whereas the abundance of the Redpicoeuk and Orgnano rose up to +86% and +89%, respectively. Yet, the distributions of these two groups became flatter.

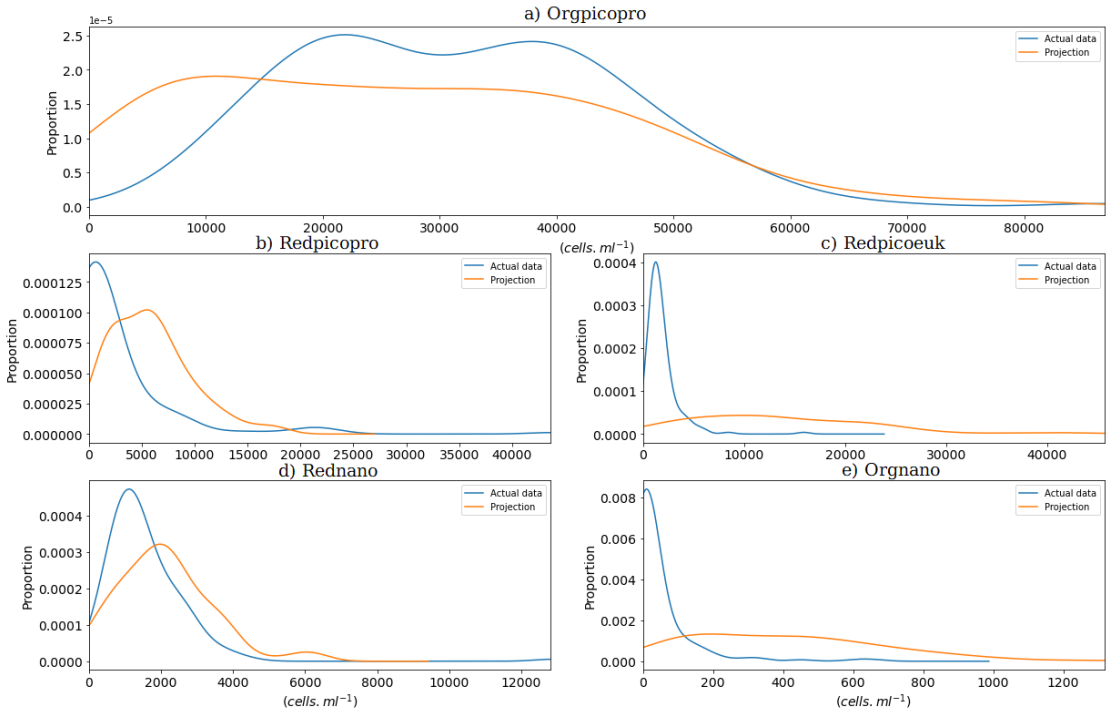


Figure 2.7. – Distribution of the functional group abundances in the actual SOMLIT data and for a simulated increase in phosphate concentration increase by 10% in summer ($n = 318$ in both cases). The distribution of the data is shown for the Orgpicopro (a), Redpicopro (b), Redpicoeuk (c), Rednano (d), and Orgnano (e). The mean of each cPFG actual and simulated distributions are significantly different (Student-Welch test p<0.01).

To summarize, the temperature increase had a positive effect on most functional groups. This could for instance be explained by the fact that increasing temperatures in low-temperature regimes favor high division rates. The non-significative impact on Redpicopro could underline the model uncertainties concerning the Redpicopro ecological niche. It might also highlight that the current temperature patterns in winter are currently near-optimal for Redpicopro and this cPFG could not take advantage of even warmer waters.
Besides, the simulated phosphate pulse had an overall stronger impact than the temperature rise: the cPFG populations may be more limited in nutrients during summer when the water column is stratified than by temperature in winter. The phosphate pulse fostered all groups except the Orgpicopro. This is consistent with high Orgpicopro abundances in very oligotrophic waters as demonstrated in Figure 2.4, and by Glibert et al. 2016 and Otero-Ferrer et al. 2018. Concerning the Redpicoeuk and Orgnano, the flat simulated distributions might reflect the multiplicity of the possible ecological niches for these groups or the model uncertainties for these two cPFGs.

Finally, it is worth noting that the estimation of the phosphate pulse impact is more subject to caution than the effect of a temperature rise. Indeed, contrary to the tem-

perature, phytoplankton also have a reverse influence on the nutrient levels: high nutrient rates foster phytoplankton abundances, but high phytoplankton abundances curb the nutrient concentrations. The estimation process hence suffers from potential endogeneity. The low temporal resolution of the SOMLIT data does not allow to disentangle the intertwined links of the phytoplankton-nutrient relationship and evidences the need to use high-resolution data to fully resolve this issue. Hence, the presented results were given more as a detailed demonstration of the possibilities offered by the MDGMM and MIAMI rather than a proper study of the underlying oceanographic processes. These possibilities are numerous and other scenarios could be tested. Simultaneous variations of different nutrients could be for example simulated to study the impact of the co-limitations between nitrogen-related nutrients ("NH_{4^+}", "NO_{3^-}", "NO_{2^-}") and phosphorus-related nutrients ("$PO_{4^{3-}}$") on phytoplankton abundances. Similarly, conversions of abundances into biomass could be conducted to assess the extent in terms of carbon budgets of such structural changes, especially for the Rednano which generally represents the highest contribution to the global biomass in the Mediterranean sea.

3. Delimiting the epipelagic zone from the mesopelagic zone

When all things beautiful and bright
sink in the night

Peter Gabriel about the mesopelagic zone.

As shown in Section 1, the vertical spatial signal is crucial to differentiate between ecological niches. In the SOMLIT data, the depth levels were determined locally for each station to provide, when possible, a suited coverage of the epipelagic zone hosting phytoplankton photosynthesis. Yet, in general, the maximal depth of the local epipelagic zone is not known in advance as during cruises. As a result, we have introduced a method called RUBALIZ to overcome this issue.

The epipelagic and mesopelagic zones are, as mentioned earlier, generally defined using fixed depths (0-200m and 200-1000m, respectively). To provide more local definitions of these zones, these boundaries were identified as change points in the vertical profiles of five characteristic variables of the water column.

3.1. Change point methods: A short literature review

Identifying the changes in a signal is an issue treated by a dedicated field in statistics called rupture detection or change-point detection methods. They are mainly two

types of contexts: "online contexts" where the goal is to detect ruptures in an incoming stream of data, in opposition to "offline" contexts for which the dataset is fixed, all data points have already been collected, and the analysis is performed *a posteriori*. The goal of the approach is here to determine the ruptures in the full signal acquired on the entire water column: our approach is hence rooted in the offline detection context. In this respect, the water column is partitioned using five characteristic variables: the potential temperature [1], the salinity, the water density, the fluorescence, and the dioxygen. Except for the fluorescence, which gives information about the photosynthetic biological content of the water column, the other quantities are hence close to OMP models (Tomczak 1981; Tomczak et al. 1989). The change points looked for in the work presented in Section 3.2 correspond to the local boundaries of the classically defined epipelagic and mesopelagic zones. The number of rupture points is hence known in advance and equal to two. More detailed vertical partitions could of course be determined by setting a higher number of change points.

The rupture detection method implemented is based on the work by Truong et al. 2020 and can be considered as a particular optimization problem that aims to split the signals into a given number of more homogeneous sub-signals. The "homogeneity" of the sub-signals is captured by a cost function that is minimized using an optimization method. This class of methods operating in a statistical frequentist framework can be applied to a large class of problems: multivariate signals or rupture detection in a regression setup for instance. Bayesian change-point models (Rabiner 1989) such as the Hidden-Markov Models (HMM) (Chen et al. 2012) were not studied here but could constitute a solid alternative.

Classical cost functions lie in two main families: parametric and non-parametric frameworks. In the parametric framework, assumptions about the data distribution are made and shape the solution contrary to non-parametric methods. The main cost functions used in the parametric framework are based on standard likelihood-based methods (as in Page 1955 or Lavielle 1999), piecewise linear models that identify structural changes in the link (assumed linear) between several variables (Qu et al. 2007), and their extensions using Mahalanobis-type metrics (Lajugie et al. 2014). The most common cost functions in the non-parametric case are the non-parametric maximum likelihood that relies on empirical cumulative distribution functions (Zou et al. 2014), rank-based detection methods using the observation ranks rather than observation values (Lung-Yut-Fong et al. 2015), or kernel methods that perform the rupture detection after projecting the data in a reproducing kernel Hilbert space (rkhs) (Harchaoui et al. 2007).

To minimize these cost functions, two types of optimization procedures can be used: exact and approximated optimization procedures. For a known number of change

1. "Potential temperature is a conserved quantity for adiabatic (energy conserving) motions and is equal to the temperature of a water parcel restored adiabatically to a reference surface pressure" (adapted from North et al. 2014).

points, exact optimization relies on dynamic programming: the rupture points are determined recursively, one after the other (Bai et al. 2003). This procedure can be extended for instance to return a list of the most likely partitions rather than a single partition as in Guédon 2013. Yet the number of observations or change points can become too large for exact methods to work in a reasonable running time. In this case, approximate procedures could be used. The window-sliding approach (Lung-Yut-Fong et al. 2012) is a popular method. It consists of a twofold window sliding over the signal. When the signal highly differs between both parts of the window, then it is likely that the window is located on a rupture point. Window-sliding methods are fast but require the selection of sensitive hyper-parameters such as the window size. Alternatively, divisive (or top-down) and agglomerative (or bottom-up) approaches are interesting options (Duda et al. 1973). While top-down methods work by iteratively splitting the signal into sub-signals, bottom-up approaches merge sub-signals until they obtain the desired number of rupture points. Divisive methods are easy to implement but can have trouble detecting change points close to each other, while agglomerative approaches are of linear complexity in the number of samples but can be unstable in the early iterations when the sub-signals to merge are of small sizes.

3.2. The RUBALIZ method and results

Using this rupture detection framework, we have proposed an alternative to the traditional boundaries of the epipelagic zone (0m-200m deep) and the mesopelagic zone (200m-1000m) on a local basis (The epipelagic zone was assimilated to the euphotic zone in the study). To do so, the five mentioned variables were collected using Conductivity-Temperature-Depth sensors in thirteen stations belonging to seven cruises (represented in purple in Figure 1.6). Supplementary Material of the paper is given in Appendix C.

A RUpture-Based detection method for the Active mesopeLagIc Zone (RUBALIZ): a crucial step towards rigorous carbon budget assessments

Robin Fuchs[1,2*], Chloé M.J. Baumas[1*], Marc Garel[1], David Nerini[1], Frédéric A.C. Le Moigne[3], Christian Tamburini[1]

[1]*Mediterranean Institute of Oceanography (MIO), UM110 (CNRS/INSU, Aix-Marseille Université, IRD), Marseille, France.*

[2]*Aix Marseille Univ, CNRS, Centrale Marseille, I2M, Marseille, France*

[3] *LEMAR, CNRS, Brest, France*

[*]*Both authors contributed equally*

Corresponding author: Chloé M.J. Baumas[1*], email address: chloe.baumas@mio.osupytheas.fr

Keywords: Mesopelagic zone boundaries, Biological carbon pump, Detection rupture method, Hydrological profiles, Carbon budget

Abstract:

Determining mesopelagic organic carbon budgets is essential to characterize the ocean's role as a carbon dioxide sink. This is because the biological processes observed in the mesopelagic zone are crucial for understanding the biological carbon pump. Yet, field assessments of carbon budgets are often unbalanced with the carbon demand exceeding its supply. This underlines either methodological issues in the budget calculations or incomplete knowledge of the mesopelagic carbon cycling with potentially missing sources. Carbon budgets are built by partitioning the ocean into vertical depth zones. Vertical boundaries are conventionally defined between 200 and 1000m depth or using various thresholds. Such

approaches lack consistent methodology preventing robust comparison of mesopelagic carbon budget from region to region. Here, using a statistical rupture detection method applied to CTD-cast variables (fluorescence, O_2 concentration, potential temperature, salinity, and density), we aim to provide independent estimates of mesopelagic boundaries. We demonstrate that the so-determined upper boundary is highly correlated with the knee points of the POC fluxes estimated by a power law and that over 90% of the POC flux attenuation occurs within our method boundaries. The identified zone therefore corresponds to the most active part of the conventional mesopelagic zone and we name it the "active mesopelagic zone". We find that the depths of the mesopelagic zone depend on the region considered. Our results demonstrate that the mesopelagic carbon budget discrepancy can vary up to four folds depending on the boundaries chosen and hence provide novel grounds to reassess existing and future mesopelagic carbon budgets.

Introduction:

In the euphotic zone of the ocean, phytoplankton convert carbon dioxide (CO_2) into biogenic carbon (C). A fraction of this biogenic C escapes the euphotic zone and crosses the mesopelagic zone of the ocean. The vertical export processes and the fate of Organic Carbon (OC) in the mesopelagic zone have been considered of major interest for the past decade and have received increased attention from the international community in recent years (Buesseler and Boyd 2009; Robinson et al. 2010; Siegel et al. 2016; Martin et al. 2020). The mesopelagic zone harbors substantial fish resources and above all, plays a key role in biogeochemical cycles, in particular in the downward pumping of biogenic carbon in the ocean. Mesopelagic organisms intercept about 90% of Particulate Organic Carbon (POC) before sinking deeper, and then respire CO_2 back into the water (Arístegui et al. 2005, 2009; Robinson et al. 2010; Costello and Breyer 2017). The mesopelagic zone is therefore a key component of the efficiency of the Biological Carbon Pump

(BCP), a crucial ecosystemic service being defined as the sum of all biological processes transporting C into the deep ocean (Eppley and Peterson 1979; Siegel et al. 2016; Le Moigne 2019).

Despite its paramount role in the BCP and thus in climate regulation, the mesopelagic zone, its global composition, and its ecology remain poorly known (Buesseler and Boyd 2009; Burd et al. 2010; Martin et al. 2020). The conventional sampling methods do not allow to gather representative data due to the vast size of the ocean, vertical heterogeneity, short temporal-scale research ship activities, hydrostatic pressure, and the avoidance tactics of metazoan (Robinson et al. 2010). In this respect, the lack of consensus concerning the boundaries of the mesopelagic zone is a stumbling block since the scientific community has failed to reconcile the mesopelagic C budget. Indeed, in most cases, measurements and estimates have shown a biological carbon demand often greater than the amount of POC exported (Reinthaler et al. 2006; Steinberg et al. 2008; Burd et al. 2010; Collins et al. 2015). In other words, the measured POC flux cannot support the measured metabolic C demand of prokaryotes and zooplankton altogether in the mesopelagic zone. In order to assess mesopelagic C budgets, C demand needs to be integrated over the whole mesopelagic zone, which by definition requires knowing its boundaries. Analyzing the work by Giering et al. (2014), it is worth noting that the choice of these boundaries significantly impacts the budget estimate, leading the balance towards a deficit, surplus, or a balanced C budget (see Extended Data Figure 5 in Giering et al. (2014)). In addition, the mesopelagic zone encompasses strong gradients in environmental conditions suggesting that the mesopelagic zone should not be considered as a homogeneous block towards the ocean. For these reasons, the mesopelagic zone boundaries need to be wisely and consistently defined before trying to provide interpretations about the mesopelagic C budget.

Similarly to Longhurst (2007), some studies have shown that the mesopelagic zone could be horizontally divided into 13 to 33 ecoregions by clustering physical or/and biological data (Proud

et al. 2017; Reygondeau et al. 2018). However, concerning the vertical boundaries of the mesopelagic zone, fewer comprehensive data-based approaches have been proposed. In practice, this partition of the water column is often performed using fixed boundaries or thresholds. The mesopelagic zone is conventionally defined between 200m and 1000m (Hedgpeth 1957). However, evidence begins to show that these boundaries can vary among oceanic biogeochemical provinces (Reygondeau et al. 2018), preventing accurate comparison between locations and studies. Besides, Buesseler et al. (2020) demonstrate that a fixed depths approach is not suitable for BCP efficiency assessment. Alternatively, criteria based on light and photosynthesis are often used (Lee et al. 2007). The upper boundary of the mesopelagic zone is then located where light is not sufficient for photosynthesis (between 0.1% and 1% of the surface Photosynthetically Active Radiation (PAR) value). Yet, PAR-based approaches can only be implemented using Conductivity Temperature Depth (CTD) profiles acquired during the day and greatly depend on water turbidity (being therefore dependent on POC fluxes). Besides, they do not take into account the whole PAR profile but only two values: the surface value and the value at the limit depth of the euphotic zone. Other methods such as Deep Scattering Layer (DSL) based on horizons where biomass-rich communities of zooplankton and fish stop during their daily migration have been proposed by (Proud et al. 2017). These depths are readily detectable by echosounders but such methods require different measurements along the day and night as the depths of echos change, additional instruments, treatment skills, and time (e.g. Proud et al. (2015)). The mesopelagic upper boundary can also be fixed below the Mixed Layer Depth (MLD) (Giering et al. 2014; Belcher et al. 2016; Reygondeau et al. 2018). Two techniques exist to determine this depth, i.e. depth where the temperature was 0.5°C lower than surface temperature (Monterey and Levitus 1997) or the depth at which a change from the surface density of 0.125 kg m^{-3} has occurred (Levitus 1982). The main disadvantage of this method is the dependence of the result on the season and the chosen technique which provide significantly different results (Lukas and Lindstrom 1991). Instead of the euphotic zone, Owens et al. (2015) use the primary production

zone (PPZ), considered as the zone between 0 m and the depth at which fluorescence reaches 10% of its highest value (Marra et al. 2014; Owens et al. 2015). Finally, these methods only provide an upper boundary for the mesopelagic zone but no lower boundary. Reygondeau et al. (2018) proposed to use the depth where the vertical POC fluxes gradient is sufficiently close to zero as a lower boundary. However, this approach determines the integration boundaries of the biogeochemical data using the biogeochemical data themselves. It hence presents an endogeneity problem for our purpose and could not be compared with the presented results.

Variables such as temperature, salinity, dissolved O_2 concentration, density, and fluorimetry data are well known and widely measured using sensors from CTD profiles, or casts, throughout the whole water column. Aside from their worldwide availability, these variables, among others, are considered to be significant ecological drivers and proxy measures of community structure or abundance (Sutton et al. 2017). These reasons make CTD profiles good candidates for moving towards a consistent and robust determination of the vertical boundaries of the mesopelagic zone.

In this study, we propose to use automatic rupture detection methods (Truong et al. 2020) applied to the CTD profiles to identify both the upper and lower boundaries of the mesopelagic zone. We name our approach RUBALIZ: a RUpture-Based detection method for the Active mesopeLagIc Zone. RUBALIZ boundaries are independent of biogeochemical data contrary to PAR-based threshold methods which rely on particle concentration (and can operate only during daytime) or to the DSL method that depends on migration activities. It can therefore be easily used for any cruise without taking care of the daytime or of the region.

In order to characterize the importance of the boundary determination over the mesopelagic C budget, we present the associated integrated Prokaryotic Heterotrophic Production (PHP) and POC flux based on the boundaries estimated by the different methods. To highlight the readiness of the proposed method, RUBALIZ has been applied to seven cruises that occurred in the North

Atlantic, the Mediterranean, the South Pacific, and the Arctic areas with contrasted stations in distinct oceanic biogeochemical provinces.

Material & procedures:

In this study, data from seven cruises and thirteen stations from distinct oceanic biogeochemical provinces were gathered (Table 1). These data include the potential temperature, salinity, dissolved O_2 concentration, density, and fluorimetry from CTD profiles, as well as PHP and POC fluxes. The CTD profiles were processed using the SeaDataProcess software. Only the downward CTD profiles have been used for all stations considered in this study.

Table 1: References and sources of the data

Cruise	Station	Region	Dates	CTD data	POC fluxes	PHP
D341	PAP	North-Atlantic	Jul-Aug 2009	BODC	(Giering et al. 2014)	(Giering et al. 2014)
DY032	PAP	North-Atlantic	Jun-Jul 2015	BODC	(Belcher et al. 2016)	Baumas et al. 2021
KN207-01*	QL-1	North-Atlantic	Apr-May 2012	BCO-DMO	(Collins et al. 2015)	(Collins et al. 2015)
KN207-01*	QL-2	North-Atlantic	Apr-May 2012	BCO-DMO	(Collins et al. 2015)	(Collins et al. 2015)
KN207-03	PS-1	North-Atlantic	Jul 2012	BCO-DMO	(Collins et al. 2015)	(Collins et al. 2015)
KN207-03	PS-3&4	North-Atlantic	Jul 2012	BCO-DMO	(Collins et al. 2015)	(Collins et al. 2015)
MALINA	430	Arctic	Jul-Aug 2009	SEANOE	(Forest et al. 2013; Miquel et al. 2015)	(Ortega-Retuerta et al. 2012)
MALINA	540	Arctic	Jul-Aug 2009	SEANOE	(Forest et al. 2013; Miquel et al. 2015)	(Ortega-Retuerta et al. 2012)
MALINA	620	Arctic	Jul-Aug 2009	SEANOE	(Forest et al. 2013; Miquel et al. 2015)	(Ortega-Retuerta et al. 2012)
PEACETIME	FAST	Mediterranean Sea	Jun 2017	(Guieu et al. 2020)	(Guieu et al. 2020)	(Marañón et al. 2021)
PEACETIME	ION	Mediterranean Sea	Jun 2017	(Guieu et al. 2020)	(Guieu et al. 2020)	(Marañón et al. 2021)
PEACETIME	TYRR	Mediterranean Sea	Jun 2017	(Guieu et al. 2020)	(Guieu et al. 2020)	(Marañón et al. 2021)
TONGA	Station 8	South-Pacific	Dec 2019	(Guieu and Bonnet 2019)	(Bressac et al. in prep.)	(Van Wambeke unpublished)

*The potential temperature and density were not recorded during the KN207-01 cruise. Hence the rupture detection was performed only using the salinity, dissolved O_2 concentration, and fluorimetry profiles for the KN207-01 cruise.

Carbon fluxes

Sinking POC flux

The POC fluxes were measured using drifting sediment trap data, except for TONGA and DY032 for which they are derived from RESPIRE measurements (Boyd et al. 2015) and a MSC (Riley et al. 2012), respectively. They were communicated or already published elsewhere for this purpose (cruises and references indicated in Table 1). POC fluxes throughout the mesopelagic zone were calculated assuming that the data followed a power law as in (Martin et al. 1987) for each of the thirteen stations (Table 1). The knee points of the power-law curves were estimated using the Unit Invariant Knee (UIK) method (Christopoulos 2016).

POC fluxes were estimated at the depths determined by the RUBALIZ method (see below). In the special case of the PEACETIME cruise, measured data were available from 200 to 1000m. PEACETIME POC fluxes appeared to be constant throughout this zone, indicating that the major attenuation of interest likely occurred in shallower water. In order to obtain a proper integrable profile, the POC flux at 100m-depth was estimated using the method from Henson et al. (2011). This algorithm links export efficiency (*e-eff*) to sea surface temperature (SST):

$e\text{-}eff = 0.23 \times e^{(-0.08 \times SST)}$. The *e-eff* is then multiplied by the primary production (PP) to estimate the exported F_{POC}. The POC flux was not available for the PS-1 station of the KN207-03 cruise.

Prokaryotic Heterotrophic Production (PHP)

PHP was measured by incorporation of ^{3}H-Leucine as described in (Kirchman et al. 1985) and following two different protocols according to the different studies: i) filtration on 0.2µm 25mm nitrocellulose filter and ii) microcentrifugation technique (see references in Table 1 for details). In brief, for both protocols, a volume of seawater samples was collected with a Niskin bottle and was incubated in the dark with 20nM (saturating concentration) of ^{3}H-Leucine between 2-8h according to the depth at in situ temperature. Then samples were processed according to the protocols of the authors and counted with a scintillation counter as described by different authors in Table 1. For depths deeper or equal to 1000m, samples were incubated with 10nM (saturating

concentration) of ^{3}H-Leucine for 10h at in situ temperature. At the end of the experiment, incubations were stopped by adding formaldehyde, filtered on 0.2µm 25mm nitrocellulose, and counted with a scintillation counter. To calculate the PHP, we used the empirical conversion factor of 1.55ng C pmol^{-1} of incorporated ^{3}H-Leu, assuming an isotopic dilution equal to 1, according to Simon and Azam (1989).

Prokaryotic Respiration (PR)

PR was estimated from measured PHP and a Prokaryotic Growth Efficiency (PGE) according to the equation from del Giorgio and Cole (1998). We use a fixed PGE of 7%, defined as the median of 32 values, measured or estimated from literature data that were computed using a conversion factor of 1.55ng C pmol^{-1} Leu between 50 and 1000 m (Arístegui et al. 2005; Reinthaler et al. 2006; Baltar et al. 2010; Collins et al. 2015). The choice of the PGE, as well as the conversion factor value, are known to strongly impact the C budget (Burd et al. 2010; Giering and Evans 2022). However, asserting their impact is out of the scope of the present work and will be conducted in a dedicated study.

Prokaryotic Carbon Demand (PCD) and C budget discrepancy (Δ_{POC})

The Prokaryotic Carbon Demand (PCD) was computed as the sum of PHP and PR. The C budget discrepancy, $\Delta_{POC,}$ was calculated using the boundaries determined by the RUBALIZ method and the following formula:

$$\mathbf{\Delta_{POC}} = POC_{input} - PCD,$$

with POC_{input} being the POC flux available at the benchmark methods or RUBALIZ upper boundary and PCD = PHP + PR.

In order to characterize the impact of integration boundaries over the C budget discrepancy, the discrepancy obtained for each method in a given station was compared to the average discrepancy obtained by the methods in this station using z-scores:

$$\text{z-score} = (\Delta_{POC} - \mu) / \sigma,$$

with Δ_{POC} the discrepancy of the method of interest, and μ and σ the mean and standard error of the discrepancies obtained by all the methods. The Z-scores are given in standard deviation (SD) to the discrepancy mean. The higher the z-score, the higher the method discrepancy is compared to the other methods in this station and conversely.

Mesopelagic boundaries detection and PHP integration

Integration boundaries: the RUBALIZ method

As explained in the introduction, RUBALIZ relies on routinely collected variables: potential temperature, salinity, dissolved O_2 concentration, density, and fluorimetry data to determine the boundaries of the mesopelagic zone. Density is determined by the salinity and potential temperature. However, the functional form relating these three quantities is highly complex (Roquet et al. 2015) and cannot be retrieved by the rupture detection method. Taking into account the density signal therefore provides additional information and has an influence on the observed rupture (see Figure S2). All CTD profiles sources are indicated in Table 1.

The CTDs signals of the five variables were resampled using linear interpolation in order to have a value at each meter depth between the minimum and maximum depths considered, $\underline{z}$ and $\bar{z}$ respectively. For each station, all CTD profiles were set to the same length and pulled together as a matrix *y*. *y* has ($\bar{z}$ - $\underline{z}$) rows and a number of columns equal to five times the number of CTDs (each profile is made of five curves: the potential temperature, salinity, dissolved O_2 concentration, density, and fluorimetry). The rupture detection is performed over *y* and looks for common rupture points over all CTD and flux signals. In order for all CTD variables to be within the same magnitude, *y* was centered and reduced before performing the rupture detection. The number of CTD profiles available for each station is given in Table S1 in supplementary information.

The rupture detection method was based on a kernelized mean change (Harchaoui and Cappe 2007; Truong et al. 2020). This was motivated by the fact that ruptures in the signal seemed more related to mean changes rather than changes in other statistical moments such as the variance. Besides, the kernelized mean change cost function did not make parametric assumptions about the shape of the statistical distribution of the data.

More formally, y was plunged into a reproducing Hilbert space H (rkhs) associated with a kernel function $k(.,.)$: $R^d \, x \, R^d \rightarrow R$ such that $k(y_z, y_{z'}) = exp(-\gamma||y_z - y_{z'}||)$, with γ a positive bandwidth parameter. The mapping function between the original space and the rkhs is denoted by ϕ: $R^d \rightarrow H$ and is such that:

$$< \phi(y_z) \mid \phi(y_{z'})_H > = k(y_z, y_{z'}) \text{ and } ||\phi(y_z)||_H^2 = k(y_z, y_z), \quad (1)$$

for all embedded samples $(\phi(y_z), \phi(y_{z'})) \in R^d x \, R^d$ and ||.|| the euclidean norm.

Intuitively, the algorithm tries to split the full embedded signal $\{\phi(y_z)\}_{z \in [\underline{z}, \bar{z}]}$ into sub-signals $\{\phi(y_z)\}_{z \in [a,b], \underline{z} \leq a \leq b \leq \bar{z},}$ such that each subpart of the signal is the closest to its mean and the farthest from the mean of the other subparts of the signal. In practice, this is captured by the following cost function c_{kernel} to minimize:

$$c_{kernel}(y_{a..b}) := \sum_{z=a}^{b} || \phi(y_z) - \bar{\mu}_{a..b} ||_H^2, H$$

with $y_{a..b}$ the subsignal between depths a and b, $\bar{\mu}_{a..b}$ the mean of the embedded subsignal $\{\phi(y_z)\}_{z \in [a,b], \underline{z} \leq a \leq b \leq \bar{z},}$, and $||\,.\,||_H^2$ as defined in (1).

This cost function was minimized using a binary search method (Olshen et al. 2004), which determined an approximate minimum of the cost function using a sequence of two-fold partitions of the signal.

The main hyperparameters to set in the method are γ the bandwidth parameter, $\underline{z}$ and $\bar{z}$. γ was set to the inverse of the median of the pairwise squared euclidean distances between all the samples of the full signal, following the heuristic given by (Truong et al. 2020). As no consensus on the absolute upper and lower boundaries of the mesopelagic zone exists, several values of $\underline{z}$ and $\bar{z}$ could be specified to run the method. In order to assess the sensitivity of the approach to the choice of these two hyperparameters, we have estimated the boundaries for 10 equally spaced $\underline{z}$ and $\bar{z}$ values. The boundaries determined by the method correspond to the mean boundary values found for these 10 $\underline{z}$ and $\bar{z}$ values. The associated standard errors give an indication about the sensitivity of the results to the choice of $\underline{z}$ and $\bar{z}$. Thus, we have set $\underline{z}$ to 0m and let $\bar{z}$ vary between 280 and 320m to determine the upper boundary of the mesopelagic zone. To determine the lower boundary of the mesopelagic zone, the algorithm was run between the identified upper boundary and $\bar{z}$ varying between 1000 and 1300m. The identified upper and lower boundaries are referred to as z_{upper} and z_{lower}, respectively.

A summary of the full rupture detection pipeline is given in Figure S1.

PHP integration

The relationship between daily PHP flux and depth Z is commonly considered as a power law function of the form:

$$PHP = kZ^m \quad (2)$$

where k and m are parameters. When taking a log transformation so that setting X = ln(Z) and Y = ln(PHP), model (2) can be re-expressed as

$$Y = b + aX,$$

where b = ln(k) and a = m. Estimation of parameters k and m is then achieved by linear regression using an observed sample ($x_i = \ln(z_i)$, $y_i = \ln(php_i)$), i = 1,...,n of size n. However, the observation

of the scatterplot of observations (x_i, y_i) (Fig 1, a) rather suggests that variables X and Y are connected through a piecewise linear model such that an estimate $\hat{y}$ is expressed as

$$\hat{y}(x) = \begin{cases} a_1 x + b_1 \; if \;\; x \leq \;\; x_t \\ a_2 x + b_2 \; if \;\; x > \;\; x_t \end{cases}$$

under some continuity constraint of the form $\hat{y}_1(x_t) = \hat{y}_2(x_t)$.

Slope parameters a_1, a_2, intercept parameters b_1, b_2, and threshold parameter x_t are estimated when minimizing the sum of squares of the errors between data and model such that

$$SSE(a_1, a_2, b_1, b_2, x_t) = \sum_{i=1}^{n} [y_i - \hat{y}(x_i)]^2$$

For a fixed value of x_t, the vector of parameters $\alpha = (b_1, a_1, a_2)'$ and the parameter b_2 are solution of the linear system given with

$$\begin{cases} \widehat{\alpha} \; = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}. \\ \widehat{b_2} = (\hat{a}_1 - \widehat{a_2}) \times x_t + \; \widehat{b_1}, \end{cases} \qquad (3)$$

where $\mathbf{y} = (y_{(1)},...,y_{(n)})'$ is the vector of the observations y_i when the observations x_i have been sorted in ascending order. Matrix $\mathbf{X}$ is the $n \times 3$ design matrix such that

$$X = \begin{bmatrix} 1 & x_{(1)} & 0 \\ \vdots & \vdots & \vdots \\ 1 & x_{(n_1)} & 0 \\ 1 & x_t & x_{(n_1+1)} - x_t \\ \vdots & \vdots & \vdots \\ 1 & x_t & x_{(n)} - x_t \end{bmatrix}$$

where $x_{(i)}$, i = 1,...,n is the sequence of the observations x_i sorted in ascending order, $1 \leq n_1 \leq n$ is the position of the last observation x_{n_1} such that $x_{n_1} \leq x_t$. Only the search for the optimal value of x_t is achieved numerically (using any well-suited 1D root-finding algorithm) such that the solution is given by:

$$\hat{x}_t = \underset{x_{(1)} \leq x_t \leq x_{(n)}}{\operatorname{argmin}} \; SSE$$

As shown in equation (3), the system owns 5 parameters but only 4 degrees of freedom since the value of intercept b_2 is constrained by the continuity between both lines. The piecewise linear model is equivalent to a regression spline of degree 1 with one free knot (the threshold x_t).

Once an optimum threshold value $\hat{x}_t$ has been found, the theory of linear regression provides tools for drawing confidence intervals for parameters a_1, a_2, b_1, and b_2. Under some normality assumption of the residuals, the parameter α follows approximately a multivariate normal distribution with estimated mean $\hat{\mu} = (\hat{b}_1, \hat{a}_1, \hat{a}_2)$' and estimated covariance matrix $\hat{\Sigma} = \hat{\sigma}^2(X'X)^{-1}$ where $\hat{\sigma}^2 = \frac{1}{n}SSE(\hat{a}_1, \hat{a}_2, \hat{b}_1, \hat{b}_2, \hat{x}_t)$ is the estimated variance of the residuals. The value of $\hat{b}_2$ is deduced from equation (3). A 95% confidence interval can then be computed for the piecewise linear model and plotted into the original system of coordinates (Figure 1 b). The confidence interval lengths in the linear regression case were three times bigger than in the piecewise regression case (not shown).

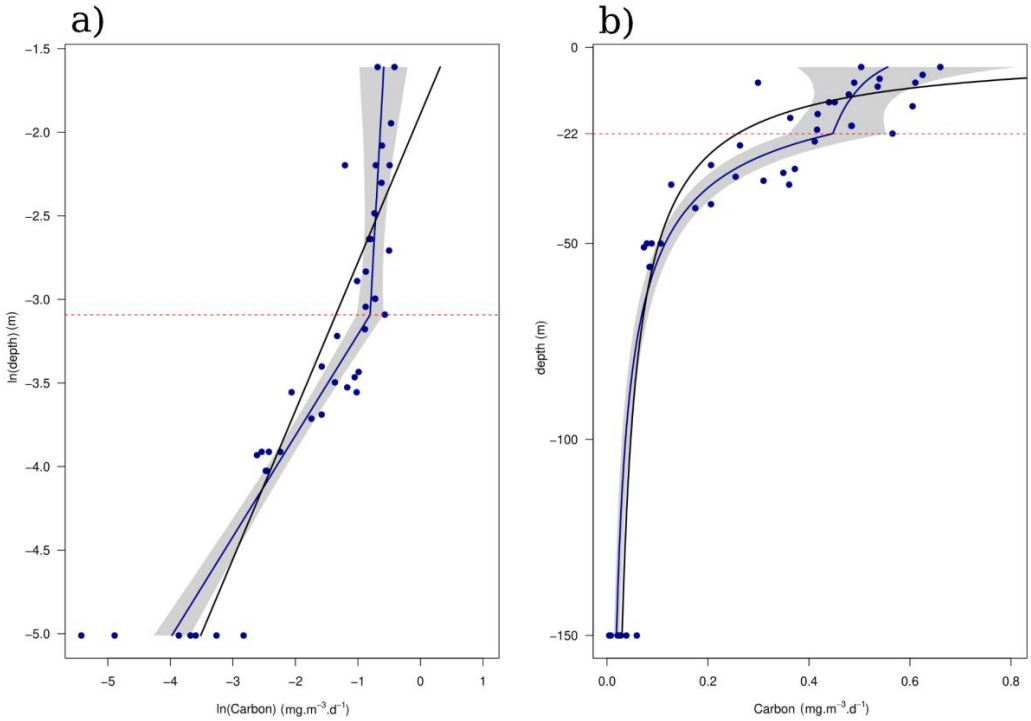


Figure 1: Example of simple linear model and piecewise linear model fits on KN-207 03 PS3&4 data. The blue points are the observations, the black curve represents the simple linear fit, and the blue curve the piecewise fit. The red dashed line is the estimated threshold depth where the piecewise model changes and gray areas are 95% confidence intervals. The model fits are shown on the log-data (a) and on the original data (b).

Once every piecewise linear model has been fitted, computation of integrated PHP fluxes along depth is achieved using the explicit formulation for the integral

$$Is = \frac{exp(\hat{b}_1)}{\hat{a}_1 + 1}\left(z_t^{\hat{a}_1+1} - z_{upper}^{\hat{a}_1+1}\right) + \frac{exp(\hat{b}_2)}{\hat{a}_2 + 1}\left(z_{lower}^{\hat{a}_2+1} - z_t^{\hat{a}_2+1}\right)$$

where $\hat{a}_1, \hat{a}_2, \hat{b}_1, \hat{b}_2$ and $\hat{z}_t = exp(\hat{x}_t)$ are estimated parameters from piecewise regression of data (z_i, php_i), i = 1,...,n sampled on cruise s. As these parameters are associated with a 95% confidence interval, it is also possible to appreciate the uncertainty of the estimations of integrated carbon fluxes.

Benchmarks methods

Finally, the benchmark approaches, namely the approaches based on the PAR values (Ez0.1 and Ez1), on the Mixed Layer Depth or PPZ, enabled to determine the beginning of the mesopelagic

zone but not its end. Hence, in order to revise C budget estimations, we have set the end of the mesopelagic zone for the benchmark approaches to 1000 meters deep in agreement with the literature conventional value.

General approach summarized

The general approach of the paper is summarized in Figure 2 and the rupture detection itself in Figure S1. The code and data to reproduce the results are available at https://github.com/RobeeF/rubaliz_paper and the DOI associated specifically with the RUBALIZ package is: DOI:10.5281/zenodo.6425452

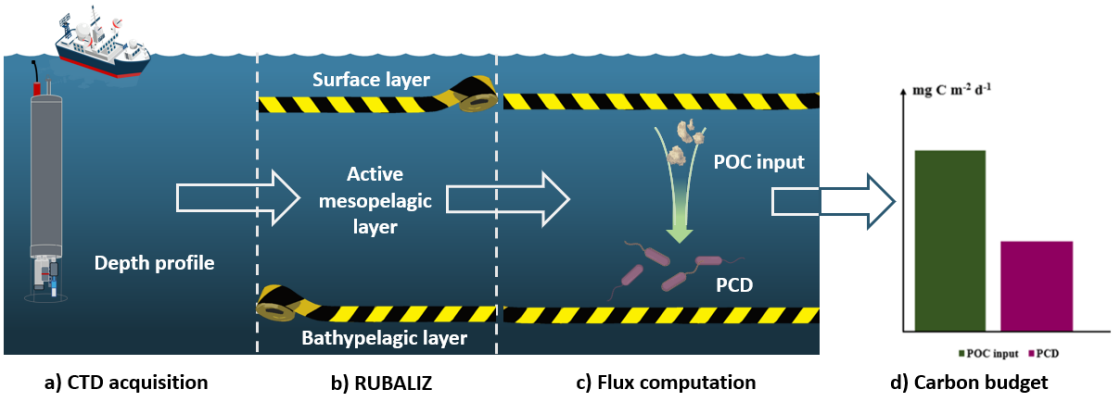


Figure 2: Presentation of the RUBALIZ rupture detection pipeline. (a) Several potential temperature, salinity, dissolved O_2 concentration, density, and fluorimetry depth profiles are acquired. (b) RUBALIZ takes these five profiles and identifies the upper and lower boundaries of the active mesopelagic zone. (c) These bounds are used to compute the gravitational POC flux input to the active mesopelagic zone and to integrate the PCD profiles and to provide C budgets (d).

Results:

In this section, we first compare the RUBALIZ approach to existing methods. Then, we show that the zone identified by RUBALIZ matches the biogeochemically active part of the mesopelagic zone, which is the zone of interest in C budget assessments. Finally, the estimated boundaries are used to integrate biogeochemical data and we compute the related C budgets.

Assessment of the Approach

RUBALIZ boundaries

The method determined the upper and lower boundaries of the mesopelagic zone for the thirteen locations considered. An illustration of the boundaries found for DY032 PAP, MALINA 430, PEACETIME ION, and KN207-01 QL-2 along with the associated variables of a CTD-cast is shown in Figure 3. The profiles were not smoothed before performing the rupture detection and presented small amplitude fluctuations that did not influence the boundaries found. The upper boundaries identified by RUBALIZ were located right below the fluorescence peaks and below the significant O_2 variations, i.e. at 126, 76, 117, and 189m deep for DY032 PAP, MALINA 430, PEACETIME ION, and KN207-01 QL-2, respectively. This result is confirmed by the sensitivity analysis reported in Figure S2, which showed that the main variables driving the upper boundary estimation are O_2 and fluorescence. The upper boundaries of the other stations presented comparable values (e.g. 109m for D341 PAP and 149m for Tonga Station 8).

The lower boundaries were estimated between 540m (KN207-03 PS-1) and 781m (KN207-01 QL-2) (see Table S1). As presented in Figure 2, these boundaries reflected brutal changes in most variables (PAP), which were located below an inflection point in some profiles (e.g. O_2 at KN207-01 QL2, or potential temperature at Station 430), or at a slope rupture (e.g. O_2 signal at PEACETIME ION). The sensitivity analysis (Figure S2) highlighted the prime importance of the O_2 signal in the lower boundary determination, followed by the salinity and potential temperature. The upper boundaries were more precisely estimated than the lower boundaries (Table S1 and Figure S4).

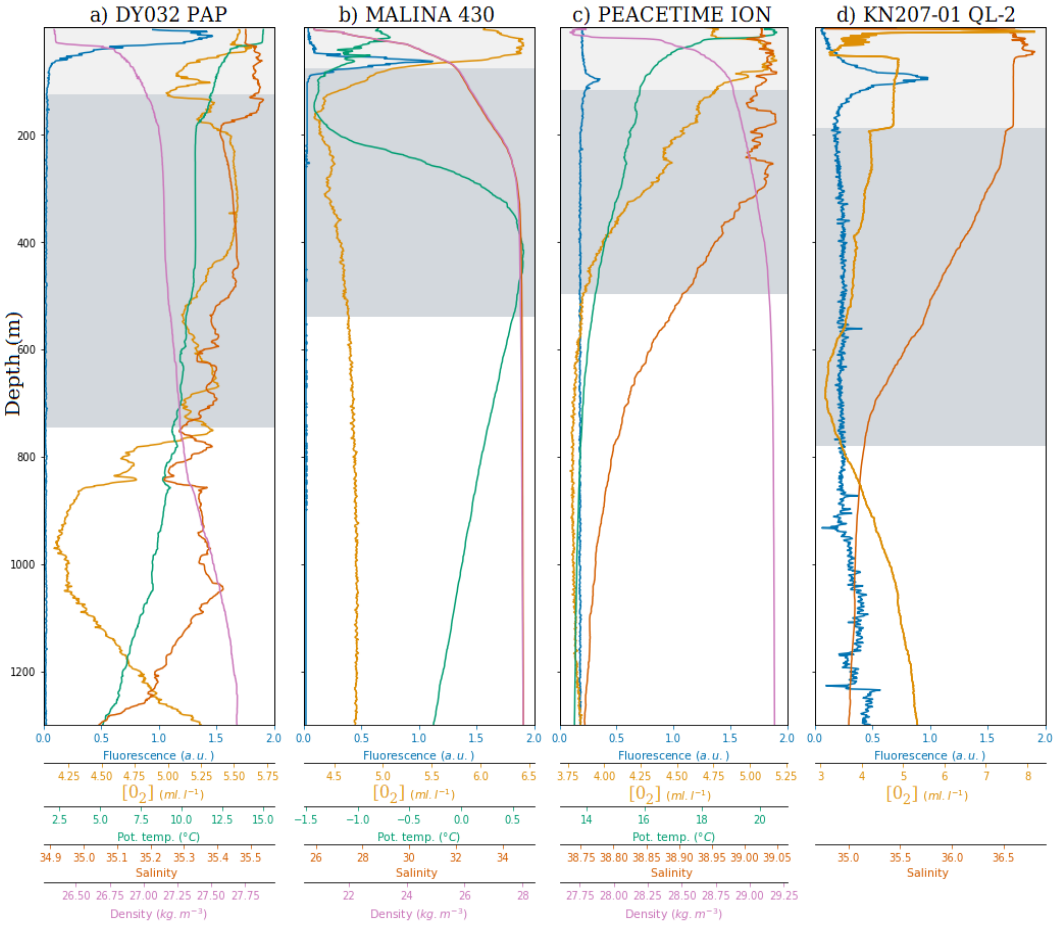


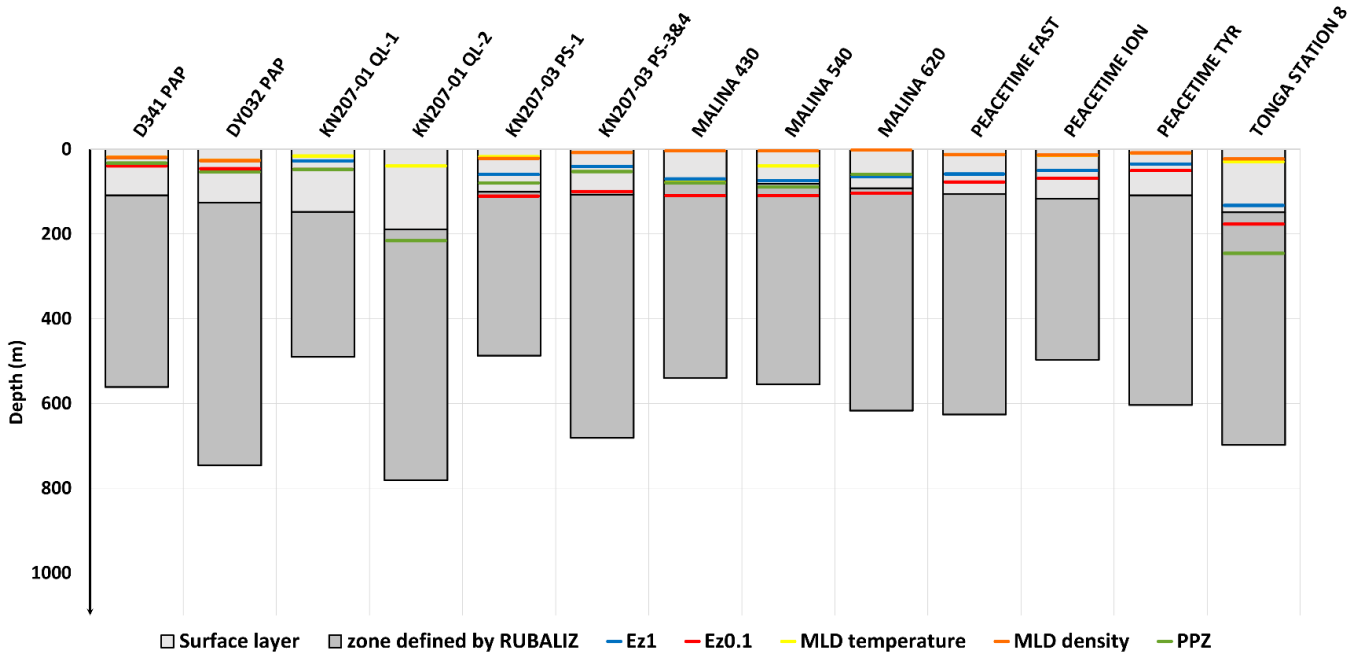
Figure 3: Illustration of the boundaries of the mesopelagic zone along with one CTD signal for a) PAP DY032, b) MALINA 430, c) PEACETIME ION and d) KN207-01 QL-2.

Comparison with benchmark methods

Figure 3 depicts the mesopelagic vertical boundaries established by the RUBALIZ approach with regards to existing approaches, namely 1% and 0.1% PAR (Ez1 and Ez0.1), MLD computed on temperature or density, the PPZ, and the usual fixed 200-1000m boundaries. Regardless of the method, the upper boundary was always set shallower than the standard 200m value, except for the PPZs of stations QL-2 (KN207-01) and Station 8 (TONGA) with 216 and 246m, respectively. The RUBALIZ upper boundary was generally deeper than the upper boundary of the other methods (PAP for both cruises, QL-1, and the three PEACETIME stations) or equivalent to Ez0.1

368 and PPZ (KN207-03 stations, MALINA stations). In all cases, the shallowest depth appeared to
369 be determined by MLD temperature or MLD density and the deepest by RUBALIZ, PPZ, or
370 Ez0.1 (Figure 3). The upper boundaries often present the same general depth ordering, from the
371 shallowest to the deepest: MLD density, MLD temperature, EZ1, PPZ, EZ0.1, and RUBALIZ.

372



374 *Figure 4: Comparison of the upper boundaries found by each method and presentation of the lower boundary found by*
375 *RUBALIZ. The missing bars are due to inoperant methods at a given station (e.g. unavailable data, the variable threshold*
376 *used by the method did not exist)*

377

378

379 Concerning the lower boundary, RUBALIZ shallowest results corresponded to the PS-1 station
380 (487m deep), the deepest to the QL-2 station (781m deep), and a mean depth of 606m for all the
381 thirteen stations. Therefore, these lower boundaries were always shallower than the 1000m
382 classically used to define the end of the mesopelagic zone.

383

RUBALIZ targets the active mesopelagic zone

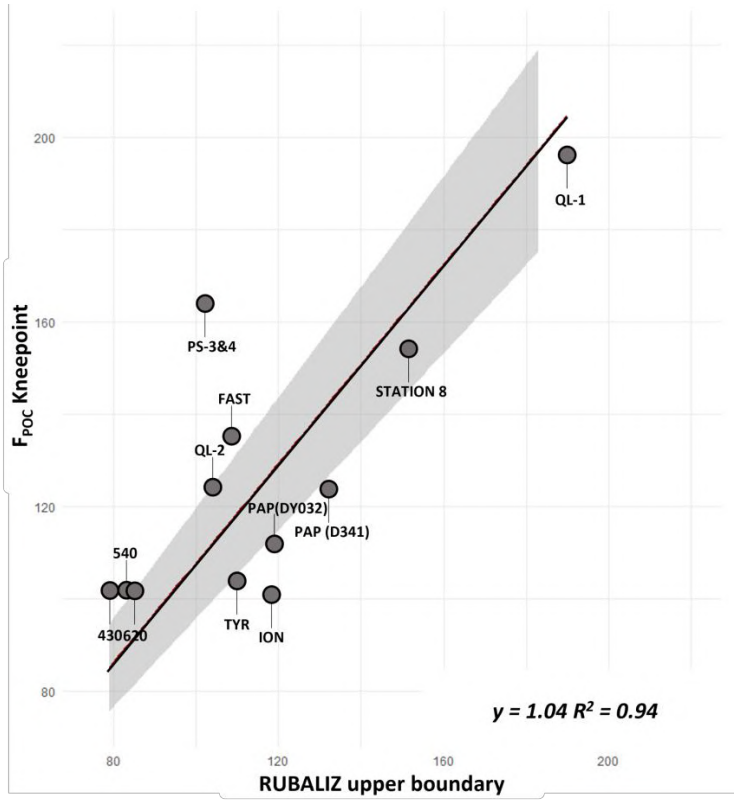


Figure 5: Linear relationship between the boundaries detected by RUBALIZ and biogeochemistry data (symbolized by the knee points of the respective POC fluxes estimated by a power law). The intercept coefficient was not significant and the p-value of the slope coefficient was $6.63x10^{-08}$. The shaded area corresponds to the 95% confidence interval

The vertical boundaries were determined by RUBALIZ using the five physical CTD profiles independently from the biological fluxes. However, we highlight an important result: the identified upper and lower boundaries are closely linked to the major attenuation of the POC flux. Firstly, the onset of the POC flux attenuation begins at the RUBALIZ upper boundary as indicated in Figure 5. Indeed, Figure 5 shows a 1:1 line between the POC curve knee points and the RIBALIZ upper boundary ($R^2 = 0.94$) with an average spread of ± 26m. This indicates that the onset of the POC flux attenuation begins at the RUBALIZ upper boundary. Secondly, below the RUBALIZ lower boundary the POC flux attenuation is limited. Since POC fluxes are represented

by a power law, the difference between the maximum and minimum values at the related depths quantifies the attenuation of the POC flux. We calculated the attenuation within the RUBALIZ boundary and compared it to the attenuation between the lower RUBALIZ boundary and the 1000m depth. We found that over 90% of the POC flux is attenuated within the RUBALIZ boundaries.

Regardless of seasons and locations, the area bounded by the upper and lower boundaries of the RUBALIZ always appears to be located near the maximum attenuation and in the vicinity of the depth where the POC flux attenuation strongly slows down. The boundaries determined on the physical conditions are hence consistent with the patterns observed on the biological fluxes. As a result, we propose to call the "active mesopelagic zone" the zone determined by RUBALIZ and this denomination will be used in the sequel.

C budget assessment

Integrating biogeochemical rates data

The active mesopelagic zone boundaries presented above for each station were used to integrate PHP fluxes and construct C budgets. The cruises presenting the highest integrated PHP were TONGA (54.08mg C m^{-2} d^{-1}), PEACETIME (26.86mg C m^{-2} d^{-1} on average) and DY032 (24.94mg C m^{-2} d^{-1}). The different stations of a given cruise presented analogous PHP except for PEACETIME FAST (~2 times higher than the two other stations) and MALINA station 620 (~10 times higher than the two other stations). The R^2, which informs about how well the estimated relationship described the data, was higher than 0.62 for all stations, except for stations QL-2, with a mean of 0.85 (see Table2). The best estimations were performed for TONGA and PEACETIME ($R^2 \geq 0.92$). The largest confidence intervals with respect to the estimated PHP

were due either to a low R^2 (QL-2) or to a limited number of points (MALINA), but the confidence interval size remained inferior or equal to the estimated PHP for all stations.

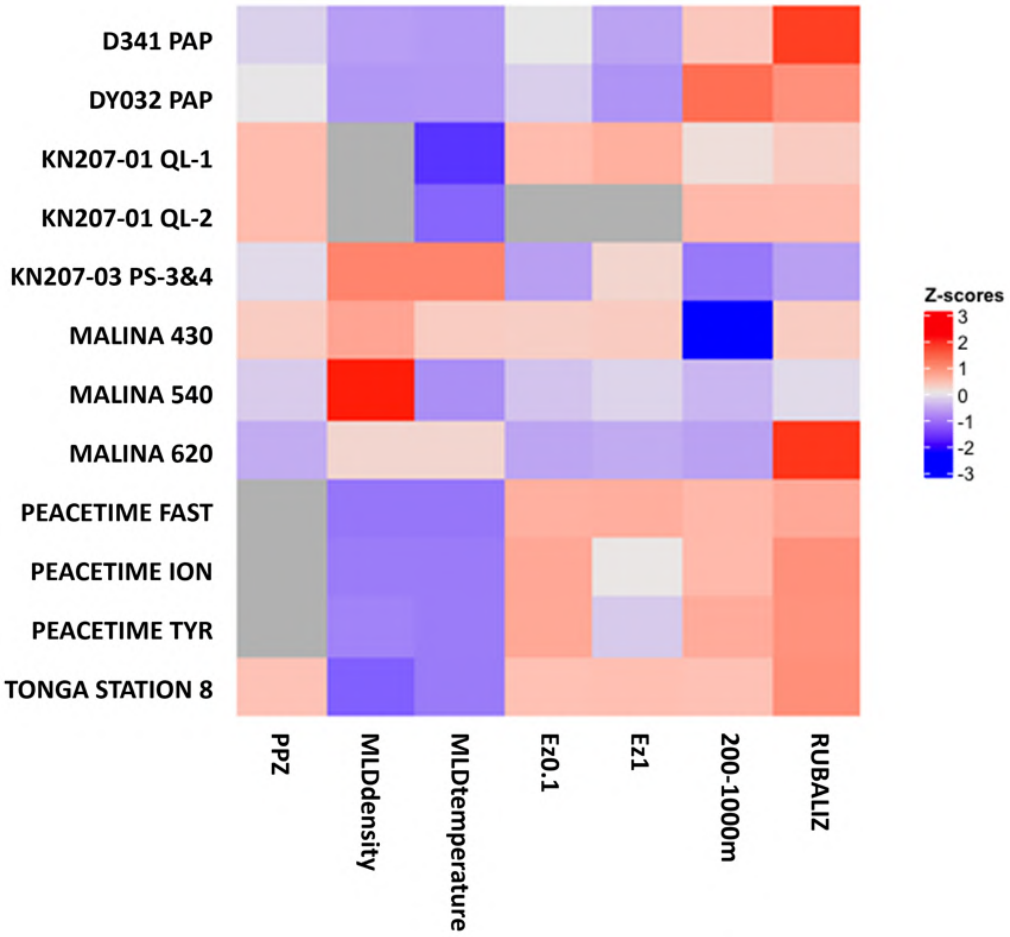
Table 2: Estimated integrated PHP fluxes using the detected boundaries of RUBALIZ

cruise	station	Active mesopelagic zone boundaries	PHP estimated (mg C m^{-2} d^{-1})	PHP Confidence Interval (mg C m^{-2} d^{-1})	R^2	Number of points
D341	PAP	(109 ; 561)	12.68	(9.01 ; 18.2)	0.82	16
DY032	PAP	(126 ; 746)	24.94	(21.58 ; 28.91)	0.89	82
KN207-01	QL-1	(148 ; 490)	11.41	(9.15 ; 14.39)	0.70	39
KN207-01	QL-2	(189 ; 781)	6.85	(4.89 ; 11.42)	0.42	24
KN207-03	PS-1	(101 ; 487)	13.98	(12.32 ; 16.05)	0.80	28
KN207-03	PS-3&4	(107 ; 681)	13.36	(11.42 ; 15.93)	0.88	42
MALINA	430	(76 ; 540)	3.54	(2.38 ; 5.96)	0.62	6
MALINA	540	(81 ; 555)	3.22	(2.45 ; 4.54)	0.93	6
MALINA	620	(92 ; 617)	31.20	(22.64 ; 44.8)	0.98	6
PEACETIME	FAST	(106 ; 626)	38.54	(31.86 ; 46.98)	0.92	54
PEACETIME	ION	(117 ; 497)	18.46	(16.46 ; 20.78)	0.96	31
PEACETIME	TYR	(109 ; 604)	23.59	(19.7 ; 28.26)	0.95	25
TONGA	STATION 8	(149 ; 698)	54.08	(45.72 ; 64.37)	0.96	14

Assessing active mesopelagic zone C budget

The C budget discrepancy, Δ_{POC}, *i.e.* the difference between gravitational sinking POC_{input} and PCD, was negative except in the KN207-03 PS-3&4 station for PPZ, Ez1, MLD temperature, and density methods (Figure S3). This implies that POC gravitational input is not sufficient to satisfy the PCD in most cases regardless of the method. Δ_{POC}, the estimated discrepancy, depends significantly on the boundary determination methods used and cruise with a mean of -323.31 ± 279.69, a maximum of -1074.34, and a minimum of 91.04mg C m^{-2} d^{-1} for respectively TONGA station 8 and KN207-03 PS-3&4 both from MLD density. Figure 6 presents the z-score per station associated with all the benchmark methods and RUBALIZ. For all negative carbon discrepancies, the higher the z-score is, the less negative the discrepancy is. Thus, using RUBALIZ boundaries

436 reduces the C budget discrepancy compared to the other methods. In the special case of KN207-
437 03 PS-3&4, RUBALIZ found a slightly more negative discrepancy compared to the other
438 methods. Compared to the usual 200-1000m boundaries, the other benchmark methods estimated
439 more pronounced negative discrepancies, especially the MLD-based methods (Figure S3).
440



441
442 *Figure 6: Z-scores per station of the C carbon budget discrepancy for all methods. The z-score is the number of standard*
443 *deviations separating a raw score from the mean. The gray cells correspond to stations for which a given method could not*
444 *determine an upper boundary. The POC flux of KN207-03 PS-1 was not available and the associated z-score was not represented*
445 *here.*
446
447
448
449
450

Discussion:

A robust methodology for boundary determinations

The mesopelagic zone is the theater of the highest attenuation rate of POC flux, a key point for assessing carbon sequestration across the ocean (Robinson et al. 2010). However, to date, no easy-to-use and universal method exist to define the boundaries of the mesopelagic zone in a meaningful and consistent way.

The customary definition of the mesopelagic zone between 200 and 1000m depth (since Hedgpeth (1957)) is practical from a theoretical point of view but not relevant to compare studies of different biogeochemical provinces. Indeed, recent research (e.g. Reygondeau et al. 2018) has challenged this view and demonstrated the variability in time and space of these vertical boundaries.

Here, we propose to revisit the mesopelagic boundary determination by taking into account the vertical variability of five variables well known to characterize the water column: fluorescence, potential temperature, salinity, density, and O_2 concentration (Sprintall and Cronin 2001; Lavigne et al. 2015). The complete vertical profiles of these five variables were used all together contrary to existing methods that define a threshold operating on a single variable. As demonstrated in our sensitivity analysis (Figure S2), all five variables participated in the determination of the boundaries, whereas the benchmark approaches were based on a single variable. Furthermore, using the whole profiles and a non-parametric mean-change kernel rather than a single threshold, makes RUBALIZ less sensitive to outlier points frequent in in situ data, and robust to missing profiles as during the KN207-01 cruise. The general trends shared by different CTD casts at a given station were captured without being influenced by cast-specific background noise.

The choice of the depth intervals on which the upper and lower boundaries were determined ($\underline{z}$ and $\bar{z}$) constituted one of the main limits of our approach. However, the low estimation variances,

especially for the upper boundary, underlined that this choice was not the main source of variability in our estimates and that the lower boundary was more difficult to estimate. Other rupture criteria than the kernelized mean change, such as Gaussian process change point models or least absolute deviation methods, were implemented (results not shown) but focused more on local features of the profiles rather than on the changes in trends and inflection points.

Analysis of the ruptures found

The upper boundary delimited by RUBALIZ was located deeper than the ones provided by the benchmark models but shallower than the 200m boundary. Conversely, the RUBALIZ lower boundaries were all located above 1000m, certainly denoting that most of the mesopelagic remineralization occurs before 1000m (Robinson et al. 2010). We have calculated that over 90% of the POC flux is attenuated within the RUBALIZ boundaries. The RUBALIZ boundaries were determined using exogenous CTD physical data, which enabled to separate the zone boundary determination problem from the integration of biological fluxes problem. Yet, the so-determined upper boundaries matched POC flux attenuation knee points (Figure 5), and the lowest boundaries delimit the end of the maximum POC fluxes attenuation zone, indicating that the five physical CTD variables used shared common information with the biological POC flux and motivated the denomination of "active mesopelagic zone". The link between the active mesopelagic zone and biogeochemical processes could be explained by the influence of environmental variables on how prokaryotes degrade POC. Indeed, prokaryotes diversity (DeLong et al. 2006; Ghiglione et al. 2008; Severin et al. 2016; Garel et al. 2019; Sebastián et al. 2021), growth efficiency (del Giorgio and Cole 1998; Nagata et al. 2010) or even gene expression (Bergauer et al. 2018) are known to be dynamic according to physical variables. Most of these processes are still poorly understood and the associated data are scarce (Burd et al. 2010; Robinson et al. 2010; Baumas et al. 2021; Giering and Evans 2022). However, the close link between environmental physical variables and prokaryotic activities could, together, strongly drive how POC flux is attenuated.

RUBALIZ as a ready-to-use shipboard tool

Given the link existing between POC flux and RUBALIZ boundaries, the present method could be a useful tool to adapt the sampling strategy during seagoing cruises. From an operational perspective, given the cost, manpower, specific equipment, low sea state, and post-analysis efforts required to use sediment traps and get POC fluxes from them (McDonnell et al. 2015), RUBALIZ could help to optimize sediment trap position and deployments at sea. From our results, we recommend using at least one CTD cast down to 1300m to fully resolve the active mesopelagic zone. Indeed, as shown in Figure S4 in the PEACETIME FAST case, RUBALIZ provided reliable estimates from the first acquired CTD cast, the spread with the final estimation being less than 1m and 35m for the upper and lower boundaries, respectively. Hence, the physical and biological sampling strategy can be designed at the very beginning of a station occupation, when only a few CTDs are available. Similarly, after strong weather events, hydrography could be significantly modified (Lavigne et al. 2015) and RUBALIZ could be used to rapidly adapt the sampling strategy.

C budget and perspectives

PHP data are usually integrated by trapezoidal rule (e.g. Reinthaler et al. (2006); Gazeau et al. (2021)) or using a power law (e.g. Giering et al. (2014)). Here, we showed that using a piecewise model with a single node on the log-data provided a better fit to the data, by increasing the R^2 of the fit and decreasing the confidence interval of the PHP fluxes estimated (see Figure S5). These PHP fluxes were integrated using the boundaries identified by the benchmark methods and RUBALIZ at each station to compute C budgets. Our results emphasized that regardless of the method used to determine the integration boundaries, the C budget discrepancies were systematically negative (except for KN207-03 PS-3&4). This problem implying that the estimated POC gravitational input is not sufficient to satisfy the estimated PCD has been an issue for several decades (Burd et al. 2010). However, we show that RUBALIZ significantly reduced

this discrepancy (Figure 6), and thus provided a solid basis for comparison between mesopelagic C budgets from different regions and seasons.

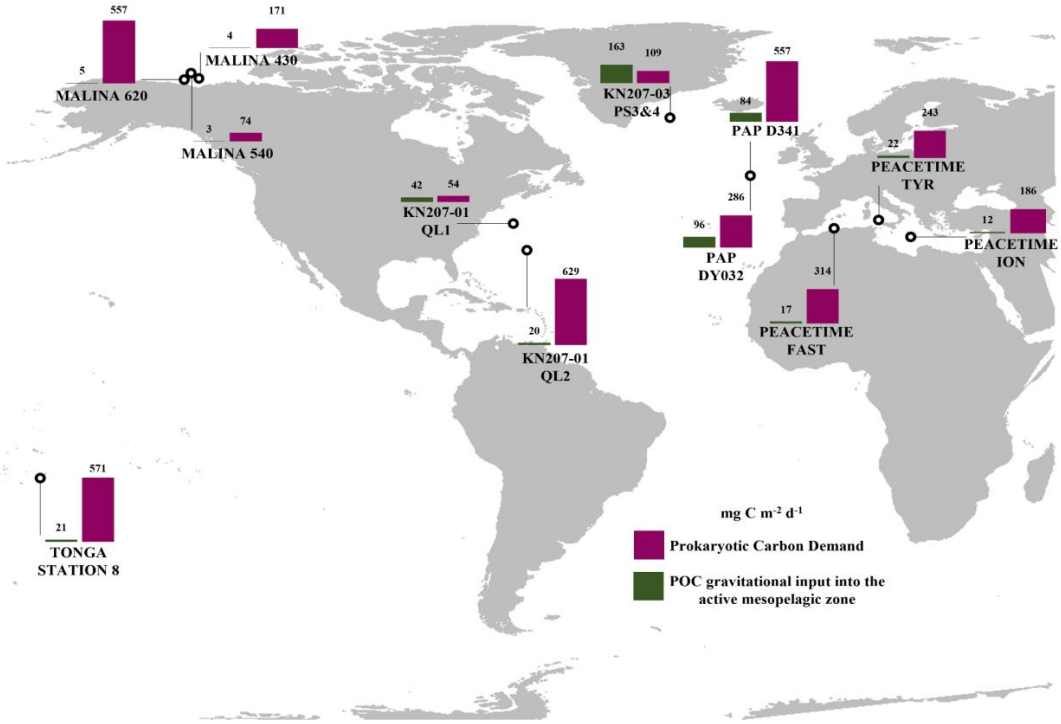


Figure 7: RUBALIZ derived C budget of the active mesopelagic zone. Pink bars represent the Prokaryotic Carbon Demand (PCD) and green bars the POC_{input}. The POC flux of KN207-03 PS-1 was not available and the associated C budget was not represented here.

Figure 7 presents a first attempt to consistently compare thirteen C budgets to one another. By comparing these various C budgets from contrasting regions and seasons all together, we conclude that this discrepancy remains a widespread feature of the ocean and that additional research is still needed to resolve this issue. RUBALIZ is only a first step and a better estimation of C fluxes in the mesopelagic zone still requires further investigation about i) the validity of PGE used to estimate PR (Burd et al. 2010) and of the CF Leu/C used to convert leucine incorporation into PHP (Giering and Evans 2022), ii) the role of attached to sinking particles prokaryotes which are not included here as we only use free-living PHP data from Niskin (Baumas et al. 2021), iii) the high-pressure effect as is it now proved that pressure can have an

important effect on prokaryotic activities and diversity, especially at depths below 200 m (Garel et al. 2019; Tamburini et al. 2021), iv) the additional C sources such as from particles injection pump or nychthemeral migrations of zooplankton or micronekton (Steinberg and Landry 2017; Aumont et al. 2018; Boyd et al. 2019), and v) assessing the contribution of chemolithoautotrophs as a new source of organic C in the dark ocean (Herndl and Reinthaler 2013). The research field aiming to decipher the C cycle in the mesopelagic zone would benefit from a worldwide effort in mapping the RUBALIZ active mesopelagic zone across regions and seasons. In that sense, autonomous and semi-autonomous platforms such as Argo floats data covering the global ocean could be used to understand how the active mesopelagic zone is varying and as a second step how to model it to predict how POC sequestration may evolve in the future. Finally, RUBALIZ, in addition to a precise sampling strategy directly on board, provides a first step towards a world mapping in Longhurst et al. style of the active mesopelagic zone. Such large-scale mapping could in turn be linked to particle flux composition, prokaryotic diversity and activities, zooplankton ecology, and POC degradation processes in order to set a new regionalization of the BCP efficiency in response to changing ocean dynamics.

Acknowledgments:

We wish to express our gratitude to D. Lefèvre, F. D'ortenzio, G. Reygondeau and F. Van Wambeke for stimulating and informative discussions. We warmly thank J.R. Collins, E.O. Retuerta, P. Massicotte, A. Martin, M. Bressac, F. Van Wambeke, F. Gazeau and T. Blasco, N. Bhairy for sharing data. We thank the SAM platform from the Mediterranean institute of oceanography as well as the PNIO-DT INSU for their technical expertise and facilities. This work was supported by the French National program LEFE (Les Enveloppes Fluides et l'Environnement), through the PARTY project (remineralisation des PARticules marines et

Transfert vers les abYsses) awarded to FACLM. This manuscript is a contribution of the ANR APERO (project number ANR-21-CE01-0027).

References and citations:

Arístegui, J., C. M. Duarte, J. M. Gasol, and L. Alonso-Sáez. 2005. Active mesopelagic prokaryotes support high respiration in the subtropical northeast Atlantic Ocean. Geophys. Res. Lett. **32**: 1–4. doi:10.1029/2004GL021863

Arístegui, J., J. M. Gasol, C. M. Duarte, and G. J. Herndld. 2009. Microbial oceanography of the dark ocean's pelagic realm. Limnol. Oceanogr. **54**: 1501–1529. doi:10.4319/lo.2009.54.5.1501

Aumont, O., O. Maury, S. Lefort, and L. Bopp. 2018. Evaluating the Potential Impacts of the Diurnal Vertical Migration by Marine Organisms on Marine Biogeochemistry. Global Biogeochem. Cycles **32**: 1622–1643. doi:10.1029/2018GB005886

Baltar, F., J. Arístegui, E. Sintes, J. M. Gasol, T. Reinthaler, and G. J. Herndl. 2010. Significance of non-sinking particulate organic carbon and dark CO 2 fixation to heterotrophic carbon demand in the mesopelagic northeast Atlantic. Geophys. Res. Lett. **37**: n/a-n/a. doi:10.1029/2010GL043105

Baumas, C. M. J., F. A. C. Le Moigne, M. Garel, and others. 2021. Mesopelagic microbial carbon production correlates with diversity across different marine particle fractions. ISME J. **15**: 1695–1708. doi:10.1038/s41396-020-00880-z

Belcher, A., M. Iversen, S. Giering, V. Riou, S. A. Henson, L. Berline, L. Guilloux, and R. Sanders. 2016. Depth-resolved particle-associated microbial respiration in the northeast Atlantic. Biogeosciences **13**: 4927–4943. doi:10.5194/bg-13-4927-2016

Bergauer, K., A. Fernandez-Guerra, J. A. L. Garcia, R. R. Sprenger, R. Stepanauskas, M. G. Pachiadaki, O. N. Jensen, and G. J. Herndl. 2018. Organic matter processing by microbial communities throughout the Atlantic water column as revealed by metaproteomics. Proc. Natl. Acad. Sci. U. S. A. **115**: E400–E408. doi:10.1073/pnas.1708779115

Boyd, P. W., H. Claustre, M. Levy, D. A. Siegel, and T. Weber. 2019. Multi-faceted particle pumps drive carbon sequestration in the ocean. Nature. doi:10.1038/s41586-019-1098-2

602 Boyd, P. W., A. McDonnell, J. Valdez, D. LeFevre, and M. P. Gall. 2015. RESPIRE: An in
603 situ particle interceptor to conduct particle remineralization and microbial dynamics studies
604 in the oceans' Twilight Zone. Limnol. Oceanogr. Methods **13**: 494–508.
605 doi:10.1002/lom3.10043
606 Bressac, M., E. C. Laurenceau-Cornec, A. Santoro, F. Kennedy, and P. W. Boyd.
607 Deconstructing drivers of Carbon flux attenuation in the oceans' biological pump, in
608 prep.
609 Buesseler, K. O., and P. W. Boyd. 2009. Shedding light on processes that control particle
610 export and flux attenuation in the twilight zone of the open ocean. Limnol. Oceanogr.
611 **54**: 1210–1232. doi:10.4319/lo.2009.54.4.1210
612 Buesseler, K. O., P. W. Boyd, E. E. Black, and D. A. Siegel. 2020. Metrics that matter for
613 assessing the ocean biological carbon pump. Proc. Natl. Acad. Sci. **117**: 9679–9687.
614 doi:10.1073/pnas.1918114117
615 Burd, A. B., D. A. Hansell, D. K. Steinberg, and others. 2010. Assessing the apparent
616 imbalance between geochemical and biochemical indicators of meso- and bathypelagic
617 biological activity: What the @$#! is wrong with present calculations of carbon
618 budgets? Deep. Res. Part II Top. Stud. Oceanogr. 1557–1571.
619 doi:10.1016/j.dsr2.2010.02.022
620 Christopoulos, D. 2016. Introducing Unit Invariant Knee (UIK) As an Objective Choice for
621 Elbow Point in Multivariate Data Analysis Techniques. SSRN Electron. J. 1–7.
622 doi:10.2139/ssrn.3043076
623 Collins, J. R., B. R. Edwards, K. Thamatrakoln, J. E. Ossolinski, G. R. DiTullio, K. D. Bidle,
624 S. C. Doney, and B. A. S. Van Mooy. 2015. The multiple fates of sinking particles in
625 the North Atlantic Ocean. Global Biogeochem. Cycles **29**: 1471–1494.
626 doi:10.1002/2014GB005037
627 Costello, M. J., and S. Breyer. 2017. Ocean Depths: The Mesopelagic and Implications for
628 Global Warming. Curr. Biol. **27**: R36–R38. doi:10.1016/j.cub.2016.11.042
629 DeLong, E. F., C. M. Preston, T. Mincer, and others. 2006. Community Genomics Among
630 Stratified Microbial Assemblages in the Ocean's Interior. Science (80-.). **311**: 496–
631 503. doi:10.1126/science.1120250
632 Eppley, R. W., and B. J. Peterson. 1979. Particulate organic matter flux and planktonic new
633 production in the deep ocean. Nature **282**: 677–680. doi:10.1038/282677a0
634 Forest, A., M. Babin, L. Stemmann, and others. 2013. Ecosystem function and particle flux
635 dynamics across the Mackenzie Shelf (Beaufort Sea, Arctic Ocean): an integrative

636 analysis of spatial variability and biophysical forcings. Biogeosciences **10**: 2833–2866.
637 doi:10.5194/bg-10-2833-2013
638 Garel, M., P. Bonin, S. Martini, S. Guasco, M. Roumagnac, N. Bhairy, F. Armougom, and
639 C. Tamburini. 2019. Pressure-Retaining Sampler and High-Pressure Systems to Study
640 Deep-Sea Microbes Under In Situ Conditions. Front. Microbiol. **10**: 453.
641 doi:10.3389/FMICB.2019.00453
642 Gazeau, F., F. Van Wambeke, E. Marañón, and others. 2021. Impact of dust addition on the
643 metabolism of Mediterranean plankton communities and carbon export under present
644 and future conditions of pH and temperature. Biogeosciences **18**: 5423–5446.
645 doi:10.5194/bg-18-5423-2021
646 Ghiglione, J. F., C. Palacios, J. C. Marty, and others. 2008. Role of environmental factors
647 for the vertical distribution (0–1000 m) of marine bacterial communities in the NW
648 Mediterranean Sea. Biogeosciences **5**: 1751–1764. doi:10.5194/bg-5-1751-2008
649 Giering, S. L. C., and C. Evans. 2022. Overestimation of prokaryotic production by leucine
650 incorporation—and how to avoid it. Limnol. Oceanogr. 1–13. doi:10.1002/lno.12032
651 Giering, S. L. C., R. Sanders, R. S. Lampitt, and others. 2014. Reconciliation of the carbon
652 budget in the ocean's twilight zone. Nature **507**: 480–483. doi:10.1038/nature13123
653 del Giorgio, P. A., and J. J. Cole. 1998. BACTERIAL GROWTH EFFICIENCY IN
654 NATURAL AQUATIC SYSTEMS. Annu. Rev. Ecol. Syst. **29**: 503–541.
655 doi:10.1146/annurev.ecolsys.29.1.503
656 Guieu, C., and S. Bonnet. 2019. TONGA cruise, RV
657 L'Atalante.doi:https://doi.org/10.17600/18000884
658 Guieu, C., K. Desboeufs, S. Albani, and others. 2020. BIOGEOCHEMICAL dataset
659 collected during the PEACETIME cruise.doi:https://doi.org/10.17882/75747
660 Harchaoui, Z., and O. Cappe. 2007. Retrospective Mutiple Change-Point Estimation with
661 Kernels. *2007 IEEE/SP 14th Workshop on Statistical Signal Processing*. IEEE. 768–
662 772.
663 Hedgpeth, J. W. 1957. Classification of marine environments, *In* Treatise on Marine Ecology
664 and Paleoecology. Geological Society of America.
665 Henson, S. A., R. Sanders, E. Madsen, P. J. Morris, F. Le Moigne, and G. D. Quartly. 2011.
666 A reduced estimate of the strength of the ocean's biological carbon pump. Geophys.
667 Res. Lett. **38**: n/a-n/a. doi:10.1029/2011GL046735
668 Herndl, G., and T. Reinthaler. 2013. Microbial control of the dark end of the biological
669 pump. Nat. Geosci. **6**: 718–724. doi:10.1038/ngeo1921

Kirchman, D., E. K'nees, and R. Hodson. 1985. Leucine incorporation and its potential as a measure of protein synthesis by bacteria in natural aquatic systems. Appl. Environ. Microbiol. **49**: 599–607. doi:10.1128/aem.49.3.599-607.1985

Lavigne, H., F. D'Ortenzio, M. Ribera D'Alcalà, H. Claustre, R. Sauzède, and M. Gacic. 2015. On the vertical distribution of the chlorophyll a concentration in the Mediterranean Sea: A basin-scale and seasonal approach. Biogeosciences **12**: 5021–5039. doi:10.5194/bg-12-5021-2015

Lee, Z., A. Weidemann, J. Kindle, R. Arnone, K. L. Carder, and C. Davis. 2007. Euphotic zone depth: Its derivation and implication to ocean-color remote sensing. J. Geophys. Res. **112**: C03009. doi:10.1029/2006JC003802

Levitus, S. 1982. Climatological Atlas of the World Ocean. 173.

Longhurst, A. R. 2007. Ecological geography of the sea. 542.

Lukas, R., and E. Lindstrom. 1991. The mixed layer of the western equatorial Pacific Ocean. J. Geophys. Res. **96**: 3343. doi:10.1029/90JC01951

Marañón, E., F. Van Wambeke, J. Uitz, and others. 2021. Deep maxima of phytoplankton biomass, primary production and bacterial production in the Mediterranean Sea. Biogeosciences **18**: 1749–1767. doi:10.5194/bg-18-1749-2021

Marra, J. F., V. P. Lance, R. D. Vaillancourt, and B. R. Hargreaves. 2014. Resolving the ocean's euphotic zone. Deep. Res. Part I Oceanogr. Res. Pap. **83**: 45–50. doi:10.1016/j.dsr.2013.09.005

Martin, A., P. Boyd, K. Buesseler, and others. 2020. The oceans' twilight zone must be studied now, before it is too late. Nature. doi:10.1038/d41586-020-00915-7

Martin, J. H., G. A. Knauer, D. M. Karl, and W. W. Broenkow. 1987. VERTEX: carbon cycling in the northeast Pacific. Deep Sea Res. Part A. Oceanogr. Res. Pap. **34**: 267–285. doi:10.1016/0198-0149(87)90086-0

McDonnell, A. M. P., P. J. Lam, C. H. Lamborg, and others. 2015. The oceanographic toolbox for the collection of sinking and suspended marine particles. Prog. Oceanogr. **133**: 17–31. doi:10.1016/j.pocean.2015.01.007

Miquel, J.-C., B. Gasser, J. Martín, C. Marec, M. Babin, L. Fortier, and A. Forest. 2015. Downward particle flux and carbon export in the Beaufort Sea, Arctic Ocean; the role of zooplankton. Biogeosciences **12**: 5103–5117. doi:10.5194/bg-12-5103-2015

Le Moigne, F. A. C. 2019. Pathways of organic carbon downward transport by the oceanic biological carbon pump. Front. Mar. Sci. **6**: 1–8. doi:10.3389/fmars.2019.00634

Monterey, G., and S. Levitus. 1997. Seasonal variability of mixed layer depth for the world

ocean, p. 96. *In* NOAA ATLAS, NESDlS ,14, Washington, D.C.

Nagata, T., C. Tamburini, J. Arístegui, and others. 2010. Emerging concepts on microbial processes in the bathypelagic ocean – ecology, biogeochemistry, and genomics. Deep Sea Res. Part II Top. Stud. Oceanogr. **57**: 1519–1536. doi:10.1016/j.dsr2.2010.02.019

Olshen, A. B., E. S. Venkatraman, R. Lucito, and M. Wigler. 2004. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics **5**: 557–572. doi:10.1093/biostatistics/kxh008

Ortega-Retuerta, E., W. H. Jeffrey, M. Babin, and others. 2012. Carbon fluxes in the Canadian Arctic: Patterns and drivers of bacterial abundance, production and respiration on the Beaufort Sea margin. Biogeosciences **9**: 3679–3692. doi:10.5194/BG-9-3679-2012

Owens, S. A., S. Pike, and K. O. Buesseler. 2015. Thorium-234 as a tracer of particle dynamics and upper ocean export in the Atlantic Ocean. Deep Sea Res. Part II Top. Stud. Oceanogr. **116**: 42–59. doi:10.1016/j.dsr2.2014.11.010

Proud, R., M. J. Cox, and A. S. Brierley. 2017. Biogeography of the Global Ocean's Mesopelagic Zone. Curr. Biol. **27**: 113–119. doi:10.1016/j.cub.2016.11.003

Proud, R., M. J. Cox, S. Wotherspoon, and A. S. Brierley. 2015. A method for identifying Sound Scattering Layers and extracting key characteristics A. Tatem [ed.]. Methods Ecol. Evol. **6**: 1190–1198. doi:10.1111/2041-210X.12396

Reinthaler, T., H. van Aken, C. Veth, J. Arístegui, C. Robinson, P. J. L. B. Williams, P. Lebaron, and G. J. Herndl. 2006. Prokaryotic respiration and production in the meso- and bathypelagic realm of the eastern and western North Atlantic basin. Limnol. Oceanogr. **51**: 1262–1273. doi:10.4319/lo.2006.51.3.1262

Reygondeau, G., L. Guidi, G. Beaugrand, and others. 2018. Global biogeochemical provinces of the mesopelagic zone. J. Biogeogr. **45**: 500–514. doi:10.1111/jbi.13149

Riley, J. S., R. Sanders, C. Marsay, F. A. C. Le Moigne, E. P. Achterberg, and A. J. Poulton. 2012. The relative contribution of fast and slow sinking particles to ocean carbon export. Global Biogeochem. Cycles **26**: 1–10. doi:10.1029/2011GB004085

Robinson, C., D. K. Steinberg, T. R. Anderson, and others. 2010. Mesopelagic zone ecology and biogeochemistry – a synthesis. Deep Sea Res. Part II Top. Stud. Oceanogr. **57**: 1504–1518. doi:10.1016/j.dsr2.2010.02.018

Roquet, F., G. Madec, T. J. McDougall, and P. M. Barker. 2015. Accurate polynomial expressions for the density and specific volume of seawater using the TEOS-10 standard. Ocean Model. **90**: 29–43. doi:10.1016/j.ocemod.2015.04.002

738 Sebastián, M., E. Ortega-Retuerta, L. Gómez-Consarnau, M. Zamanillo, M. Álvarez, J.
739 Arístegui, and J. M. Gasol. 2021. Environmental gradients and physical barriers drive
740 the basin-wide spatial structuring of Mediterranean Sea and adjacent eastern Atlantic
741 Ocean prokaryotic communities. Limnol. Oceanogr. 1–19. doi:10.1002/lno.11944
742 Severin, T., C. Sauret, M. Boutrif, and others. 2016. Impact of an intense water column
743 mixing (0-1500 m) on prokaryotic diversity and activities during an open-ocean
744 convection event in the NW Mediterranean Sea. Environ. Microbiol. **00**.
745 doi:10.1111/1462-2920.13324
746 Siegel, D. A., K. O. Buesseler, M. J. Behrenfeld, and others. 2016. Prediction of the Export
747 and Fate of Global Ocean Net Primary Production: The EXPORTS Science Plan. Front.
748 Mar. Sci. **3**: 22. doi:10.3389/fmars.2016.00022
749 Simon, M., and F. Azam. 1989. Protein content and protein synthesis rates of planktonic
750 marine bacteria. Mar. Ecol. Prog. Ser. doi:10.3354/meps051201
751 Sprintall, J., and M. F. Cronin. 2001. Upper Ocean Vertical Structure, p. 3120–3128. *In*
752 Encyclopedia of Ocean Sciences. Elsevier.
753 Steinberg, D. K., and M. R. Landry. 2017. Zooplankton and the Ocean Carbon Cycle.
754 http://dx.doi.org/10.1146/annurev-marine-010814-015924 **9**: 413–444.
755 doi:10.1146/ANNUREV-MARINE-010814-015924
756 Steinberg, D. K., B. A. S. Van Mooy, K. O. Buesseler, P. W. Boyd, T. Kobari, and D. M.
757 Karl. 2008. Bacterial vs. zooplankton control of sinking particle flux in the ocean's
758 twilight zone. Limnol. Oceanogr. **53**: 1327–1338. doi:10.4319/lo.2008.53.4.1327
759 Sutton, T. T., M. R. Clark, D. C. Dunn, and others. 2017. A global biogeographic
760 classification of the mesopelagic zone. Deep Sea Res. Part I Oceanogr. Res. Pap. **126**:
761 85–102. doi:10.1016/j.dsr.2017.05.006
762 Tamburini, C., M. Garel, A. Barani, and others. 2021. Increasing Hydrostatic Pressure
763 Impacts the Prokaryotic Diversity during Emiliania huxleyi Aggregates Degradation.
764 Water **13**: 2616. doi:10.3390/w13192616
765 Truong, C., L. Oudre, and N. Vayatis. 2020. Selective review of offline change point
766 detection methods. Signal Processing **167**: 107299. doi:10.1016/j.sigpro.2019.107299
767 Van Wambeke, F. Mediterranean Institute of Oceanography; Université Aix-Marseille:
768 Marseille, France; unpublished results from TONGA cruise.
769

Chapter conclusion

Data in oceanography are conspicuous for their significant spatial and temporal dependence structure. More precisely, the MDGMM analysis has shown on the SOMLIT data that the spatial dependence was the most powerful structuring variable followed by the temporal dependence. This was confirmed by the environmental scenarios simulated thanks to MIAMI, which underlined that ecological niches were very sea-dependent. Similarly, the change points determined on fluorescence and physical profiles by RUBALIZ matched the POC flux change points, highlighting the common oceanic forces at stake in shaping ecological niches. Thus, at this point of the study, two points can be made and structure the next chapter. First, there is a need for physics-biology joint approaches. Second, the sharp contrast between zones and seasons pushes to develop proper statistical methods to understand phytoplankton fine-scale and high-frequency dynamics.

3. High-frequency phytoplankton response to pulse events

Sommaire

> But storms won't last, they clear the air,
> For something new.
> The sun came out and brought you through.
>
> *Avended Sevenfold about the FUMSECK storm*

Phytoplankton cells experience substantial infra-day variations, notably through the nycthemeral cycle: The cells grow during the day relying on photosynthesis and generally divide at the end of the day when solar energy is scarcer. These infra-day variations give the phytoplankton a strong adaptability and response capacity to their continuously changing direct environment.

This adaption ability to local intense and brief environmental perturbations of phytoplankton functional groups is under study in this chapter. First, the response potential of phytoplankton groups to a single storm is highlighted during the FUMSECK cruise. Then, a more general characterization of the impact of several wind-induced events is proposed. It relies on a dedicated convolutional neural network to automate the Flow Cytometry data treatment of several thousand acquisitions, and on rupture detection methods to characterize the magnitude and duration of the phytoplankton responses.

1. General approach and phytoplankton response first characterization

1.1. A physics and biology joint approach centered around flow cytometry

As evoked earlier, resolving phytoplankton behavior at high temporal frequency necessitates dedicated instruments to measure both physical and biological variables. For physical quantities, thermo-salinometers, acoustic Doppler current profilers (ADCP), and gliders were for instance deployed and enabled infra-hour samples. Concerning phytoplankton measurements, a Cytosense Automated Flow Cytometer made it possible to follow the phytoplankton functional groups with an acquisition frequency of 4MHz (corresponding to a collection capacity of up to 10,000 cells per second). Cytosense FC collects four signals for each cell, two diffusion signals (ForWard Scatter, FWS, and SideWard Scatter, SWS), and two fluorescence signals (Red FLuorescence, FLR, and Orange/Yellow FLuorescence, FLO or FLY). From the FWS signal, a fifth signal is computed: the Curvature describing how much a cell presents a curved shape. These signals can be summarized by simple descriptors (mean, variance, area under the curve, etc.) before storage. In this case, the data are stored in "listmode" format (Dubelaar et al. 1999). Conversely, the full signal of the curves can be stored as such and the associated storage format is then referred to as the "pulse shapes" format.

The manual gating, *i.e.* the association of the cells to a functional group, is performed using the data in listmode format. A series of 2D plots is generated, each plot presenting one descriptor as a function of another one (e.g. the mean of the FLR signal vs. the area under the curve of the FWS signal) for each cell. Popular choices for these 2D plots are Total FWS vs. Total FLR, Total FLR vs. Total FLO, or Total FLR vs Total SWS ("Total" standing for the area under the curve). On these 2D plots, cells presenting similar descriptors are hence concentrated around several density centers associated with each cPFG. The experts then draw borders, or gates, around these dense zones on each 2D plot and associate them with existing cPFGs. An example of a manual gating procedure is given in Appendix D.

Given the phytoplankton size range, a two-threshold acquisition protocol is often designed to deal with the smallest and the biggest cells separately using two FLR thresholds as in Marrec et al. 2018. Using such thresholds enables the collection of phytoplankton cells without collecting too many noise particles (decomposing cells, non-organic particles, electronic artifacts). The noise particles can represent the majority of the particles and could saturate the processing capacity of the FC. Thus, a low FLR threshold is used to capture the smallest cells without saturating the sampling capacities and a higher threshold is used to count only the most fluorescent (and less abundant) cells. This two-threshold protocol and the FC manual gating strategy were used during the FUMSECK cruise.

1.2. High response of phytoplankton functional groups during a storm: a case in point

The FUMSECK cruise (DOI 10.17600/18001155) occurred in the Ligurian Sea (North-Western Mediterranean Sea) from April 30, 2019, to May 07, 2019 (see Figure 1.6). The cruise was marked by a violent storm that pushes deep seawater to the surface and trigger a sharp biological reaction. Using Flow Cytometry associated with a glider, satellite data, and an atmospheric model, this study shows the response potential of phytoplankton groups to a particularly intense event. As a result, this study can be seen as a motivation to better characterize the phytoplankton response to such events.

Intense storm in the north-western Mediterranean Sea strongly shaped local physics and generated significant phytoplankton reaction

Barrillon Stéphanie[1], Fuchs Robin[1,2], Petrenko Anne[1], Comby Caroline[1], Bosse Anthony[1], Yohia Christophe[3], Fuda Jean-Luc[1], Bhairy Nagib[1], Berline Léo[1], Cyr Frédéric[4], Doglioli Andrea[1], Grégori Gérald[1], Tzortzis Roxane[1], d'Ovidio Francesco[5], and Thyssen Melilotus[1]

[1]Aix Marseille Univ., Université de Toulon, CNRS, IRD, MIO UM 110 , 13288, Marseille, France
[2]Aix Marseille Univ, CNRS, Centrale Marseille, I2M, Marseille, France
[3]Aix Marseille Univ., Université de Toulon, CNRS, IRD, OSU Pytheas UAR 3470 , 13288, Marseille, France
[4]Fisheries and Oceans Canada, Northwest Atlantic Fisheries Centre, St. John's, Canada
[5]LOCEAN, UMR CNRS / Université P. et M. Curie / IRD / MNHM, F-75005 Paris, France

Correspondence: Barrillon Stéphanie (stephanie.barrillon@mio.osupytheas.fr)

Abstract.

The study of extreme weather events and their impact on ocean physics and biogeochemistry is challenged by the difficulty of collecting data in situ. Yet, recent research pointed out the major influence of such physical forcing events on microbiological organisms. In May 2019, an intense storm occurred in the Ligurian Sea (north-western Mediterranean Sea) and was captured during the FUMSECK cruise. In situ sensors (onboard ADCP, thermosaligraph, fluorometer, and flow cytometer; tracked Moving Vessel Profiler; and a glider) along with a 3D atmospherical model were used to characterise the fine-scale dynamics occurring in the impacted oceanic zone. The most affected area was marked by a lower water temperature (1.1°C less), and in average 2.5 times more surface chla and 7.4 times more nitrate concentrations, exhibiting strong gradients with respect to the surrounding waters. Our results show that this storm physical forcing led to a deepening of the mixed layer depth and a dilution of the deep chlorophyll maximum. As a result, the phytoplankton biomass of most groups identified by automated flow cytometry increased up to a factor of two. Conversely, the phytoplankton carbon-chlorophyll ratio of most groups dropped down by a factor of 2, evidencing significant changes in the phytoplankton cell compositions. This observational evidence of an immediate reaction of phytoplankton community to a physical forcing due to a storm highlights the need for high-resolution coupled physics-biology measurements.

1 Introduction

Meteorological impulse wind-events such as storms, and their effects on oceanic physics and biogeochemistry, are poorly explored with in situ data. Such events generate mixing and stirring of the surface layer and, depending on the strength and duration of the events, can trigger transitional peaks in primary production, mainly explained by nitracline shoaling and grazers dilution (Lomas et al., 2009; Menkes et al., 2016), or by diluting the deep chlorophyll maximum. In oligotrophic ocean conditions, Han et al. (2012) and Babin et al. (2004) observed from satellite ocean colour the sudden and consequent increase of chlorophyll, lasting several weeks, after summer Hurricane-storms. Babin et al. (2004) explored the resulting increase in surface chlorophyll-a (chla) and generated values were close to the spring bloom values. His study also suggested that a Hurricane can trigger primary productivity equal to meso-scale eddies, but conclusions could not reach further processes understanding as in situ observations were lacking. Only few studies combined high-resolution physical description of the phenomena coupled to phytoplankton functional groups resolution. Some coastal studies, such as Fuchs et al. (2022), have evidenced pico-nanophytoplankton abundance and biomass reactions within two to four days following wind-induced events in a coastal station located in the north-western Mediterranean Sea in stratified conditions. Again, the authors showed that extreme events can generate daily biomass increases of the same order of magnitude as those observed during the spring bloom. Similarly, Anglès et al. (2015) have studied the reaction of nano-microphytoplankton to tropical cyclones generating wind-physical forcing and substantial rains in the Western Gulf of Mexico. They highlighted abundance reaction delays consistent with Fuchs et al. (2022) and strong abundance peaks following the storms.

The classical spring bloom as observed in temperate oceans is triggered by the intermittent shoaling of the mixed layer when passing from the winter convection to the spring stratification (Behrenfeld, 2010), which ends when no more nutrients are available in the euphotic layer or when grazers overpass the absorption capacity. This is particularly the case in the north-western (NW) Mediterranean Sea characterized by spring blooms of different intensities that can be detected from satellite images (d'Ortenzio and Ribera d'Alcalà, 2009; Mayot et al., 2016). The area is affected by strong Northerly's winds, and their intensity in winter define the bloom intensity (Conan et al., 2018). In summer stratified conditions, meteorological impulse wind-events could induce additional submesoscale vertical mixing. Observing the effect of these events on phytoplankton dynamics and distribution, in particular under stratified oligotrophic conditions where they may trigger patches of high production, is challenging and requires the deployment of dedicated automated and high frequency sampling tools. Being able to monitor phytoplankton distribution at a functional level, by integrating small and rapid scale dynamics into larger space and time scales, such as basin and annual scales, would precise the role of phytoplankton in biogeochemical processes.

The FUMSECK (Facilities for Updating the Mediterranean Submesoscale - Ecosystem Coupling Knowledge, https://doi.org/10.17600/18001155, PI S. Barrillon (Barrillon et al., 2020)) cruise was conducted in spring 2019 in the Ligurian Sea (NW Mediterranean Sea). FUMSECK aimed at combining physical and biological oceanography for the study of fine scales dynamics, which imply structures such as eddies, filaments or fronts over a horizontal spatial range of 1 to 100 km, a vertical one

of 0.1 to 1 km, and a temporal range of days to a few weeks (Giordani et al., 2006; Ferrari and Wunsch, 2009; McWilliams, 2019). During this one-week cruise, we deployed towed instruments and an underwater glider (Testor et al., 2019) to measure physical properties at high resolution. These measurements have been paired with ship-board measurements of phytoplankton functional groups from an automated pulse shape recording flow cytometer, based on cell sizes and pigment contents (Dugenne et al., 2014; Thyssen et al., 2014; Bonato et al., 2015; Louchart et al., 2020). During this cruise, a particularly intense episode of wind hit the south of France and the Ligurian Sea. Right after the storm for which we had to take shelter, the ship came back to the wind-exposed zone to collect data. Meanwhile, the glider collected data in the storm-exposed zone.

After the description of the material and methods, the results section show the general hydrodynamics and biogeochemical conditions in the Ligurian Sea, before focusing on the data collected just after the storm showing significant differences in both physical and biological characteristics, with respect to the general ones. The discussion explores the observed biological reaction to the storm.

2 Material and methods

The FUMSECK cruise took place from 30 April 2019 to 7 May 2019, in the Ligurian Sea (NW Mediterranean Sea), onboard the RV *Téthys II*. Figure 1 shows the cruise trajectory together with the positions of the 7 stations, and the glider trajectory. Several in situ instruments for measuring physics and biogeochemistry were deployed and are described in this section within the first two parts : transect measurements, and glider. The satellite data exploited to guide the cruise and obtain a synoptic view of the region are described in the third part, followed by the meteorological model. The last part deals with the comparison of the fluorescence and chla concentrations from the different measurements.

2.1 Transect measurements

The vessel-mounted Acoustic Doppler Current Profiler (VM-ADCP, RDI Ocean Surveyor 75 kHz) ran continuously during the cruise. The vertical range in depth is $[18\,\mathrm{m};\ 562\,\mathrm{m}]$ with a $8\,\mathrm{m}$ resolution. Current data are averaged and stored every 2 minutes, corresponding to a horizontal resolution of $0.5\,\mathrm{km}$ for a vessel speed of 8 knots. Resulting horizontal currents have been treated by the Cascade 7.2 package (Le Bot et al., 2011).

Surface-water flow-through system pumped seawater at a $2\,\mathrm{m}$ depth with a flow rate of about $60\,\mathrm{L\,min^{-1}}$. A thermosalinograph (TSG, SeaBird SBE 21) acquired sea surface temperature (SST) and salinity (SSS) data every minute. A fluorometer (Turner Designs, 10-AU-005-CE) recorded simultaneously sea surface red fluorescence $> 680\,\mathrm{nm}$ after excitation in the blue (Rfluo_tsg (a.u.), a.u. standing for arbitrary units) as a proxy of chla content.

A Moving Vessel Profiler (MVP200) was deployed with the Multi Sensor Free Fall Fish (MSFF) set of instruments, including a μCTD (AML S/n 7373 PDC-B0204), a fluorometer (WETLabs ECOFL S/n FLRTD-1581), a Laser Optical Plankton

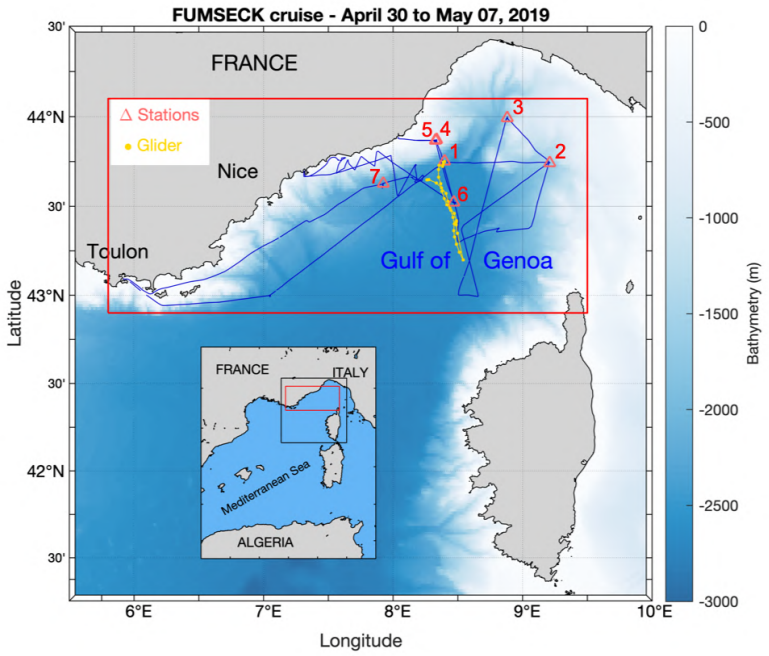


Figure 1. FUMSECK cruise (blue line), superimposed with the bathymetry. The geographical domain is represented in red, the stations with orange triangles, and the glider trajectory in yellow.

Counter (LOPC, particle size range: $100\,\mu\mathrm{m} - 1920\,\mu\mathrm{m}$), and an ODDI attitude sensor (rotation measurement on the 3 axes). Temperature and salinity profiles were treated with the LatexTool Package (Doglioli and Rousselet, 2013). In total, 507 profiles have been performed along $680.4\,\mathrm{km}$ of route (58h25min of effective measurements), separated in 7 transects (MVP 1 to 7) with a mean duration of 8h20min each and a general vessel speed of 8 knots.

Along the cruise, 26 samples for Phosphate (PO_4^{-3}), Nitrate (NO_3), Nitrite (NO_2) and Silicate ($SiOH_4$) concentrations were collected from the flow-through system in $20\,\mathrm{mL}$ high-density polyethylene bottles poisoned with HgCl2 to a final concentration of $20\,\mathrm{mg\,L^{-1}}$ and stored at $4°\mathrm{C}$ before being analysed in the laboratory a few weeks later. Nutrient concentrations were determined using a Seal AA3 auto-analyser following the method of Aminot and Kérouel (2007) with analytical precision of $0.01\,\mu\mathrm{mol\,L^{-1}}$ and quantification limits of 0.02, 0.05 and 0.30 $\mu\mathrm{mol\,L^{-1}}$ for PO_4^{-3}, NO_3 (and NO_2) and $SiOH_4$, respectively.

Similarly, chla concentration (chla_insitu, $\mathrm{ng\,mL^{-1}}$) was extracted from a total of 20 samples filtered from $500 \pm 20\,\mathrm{mL}$ of seawater through $25\,\mathrm{mm}$ glass-fiber pyrolysed filters (Whatman® GF/F) and immediately frozen at $-20\,°\mathrm{C}$. Filters were placed in glass tubes containing $5\,\mathrm{mL}$ of pure methanol and allowed to extract for 30 min as described by Aminot and Kérouel (2007). Fluorescence of the extract was determined by using a Turner Fluorometer AU10 equipped with the Welschmeyer

kit to avoid chlorophyll-b interference (Welschmeyer, 1994). The fluorometer was zeroed with methanol turbidity blank. The detection limit was $0.01\,\mathrm{ng\,mL^{-1}}$. Calibration was performed using a pure chla standard (Sigma Aldrich®, ref: C5753, pure spinach chlorophyll).

Phytoplankton abundances and functional groups were resolved using an automated pulse shape recording flow cytometer, a Cytosense (AFCM, cytobuoy b.v.; NL) plugged on the flow-through system, which automatically analyzed samples for phytoplankton counts in the size range of $0.6 - 800\,\mu\mathrm{m}$ in width. The cells contained in a volume of water were first surrounded by an isotonic sheath fluid, aligned in a laminar flow and went through a $488\,\mathrm{nm}$ laser beam laser beam thanks to a weight calibrated sample peristaltic pump. Doing so, a set of optical curves, called pulse shapes, was generated for each cell. The pulse shapes of side-ward scatter (SWS, $488\,\mathrm{nm}$) and fluorescence emissions were separated by a set of optical filters (orange fluorescence (FLO, $552\breve{}652\,\mathrm{nm}$) and red fluorescence (FLR, $> 652\,\mathrm{nm}$) and collected on photomultiplier tubes. The pulse shapes of forward scatter (FWS) are collected on left and right angle photodiodes and used to validate the laser alignment. A total of 409 samples were acquired at a 20-minutes frequency, corresponding to a mean resolution of 3.9 km during transects. The samples were stabilised in a $300.\mathrm{mL}$ sub-sampling chamber before acquisition, the instrument and the acquisition protocol are described in Marrec et al. (2018).

The identification of the phytoplankton groups relied on the standard vocabulary description in "Flow cytometry cluster names for marine waters definition": http://vocab.nerc.ac.uk/collection/F02/current/. Two protocols were successively run, one triggering on FLR $6\,\mathrm{mV}$ for 5 min targeting Orgpicopro and a second one triggering on FLR25 for 10 min targeting the Redpicoeuk, Rednano, Orgnano, and Redmicro phytoplankton groups, as presented on Fig. 2. Phytoplankton groups were manually classified using the CytoClus® software by generating several two-dimensional cytograms plotting descriptors of the four pulse shape such as the area under the curve of the pulse shape signal (FWS_cyto, SWS_cyto, Ofluo_cyto, Rfluo_cyto). Groups abundances and cell properties were processed by the software.

The size of the different phytoplankton cells was estimated based on the relationship between silica beads (1.0, 2.01, 3.13, 5.02, 7.27 $\mu\mathrm{m}$ non-functionalised silica microspheres, Bangs Laboratories, Inc.) real size and FWS_cyto signal and converted into equivalent spherical diameter (ESD) and biovolume (BV, $\mu\mathrm{m}^3$). A power law relationship ($\log(\mathrm{BV}) = 0.912 \times \log(\mathrm{FWS\_cyto}) - 5.540$, $\mathrm{r}^2 = 0.89$, $\mathrm{n} = 7$) allowed the conversion of the FWS signal into cell size. The stability of the optical unit and the flow rates were checked using Beckman Coulter Flowcheck™ fluorospheres ($2\,\mu\mathrm{m}$) before, during and after installation. Phytoplankton biomass per group were computed in $\mathrm{pgC\,mL^{-1}}$ from the power law $\mathrm{aBV^b}$, to get a mean carbon cellular quota (C, $\mathrm{pgC\,cell^{-1}}$), with a and b conversion factors reported by Menden-Deuer and Lessard (2000) and Verity et al. (1992).

2.2 Glider

An automous Alseamar's SeaExplorer glider was deployed during the whole cruise in order to perform complementary measurements on the dynamics and biogeochemistry around the area of the cruise. It performed saw-tooth cycles with a pitch angle

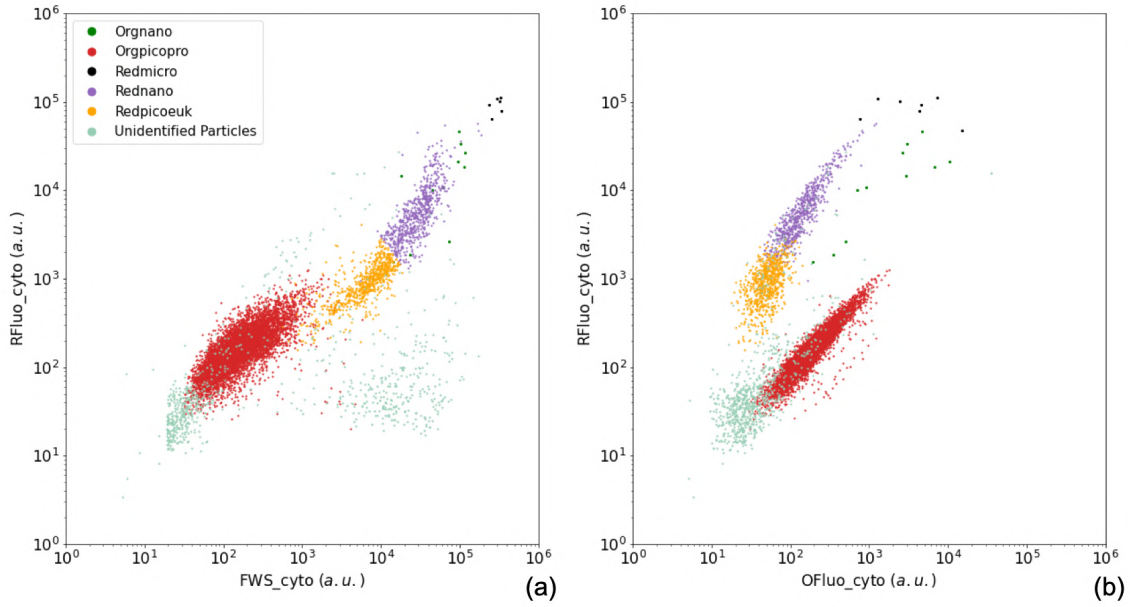


Figure 2. Manual identification of the main phytoplankton functional groups. Two dimension cytograms representing: (a) the area under the curve of red fluorescence (RFluo_cyto, (a.u.)) versus forward scatter (FWS_cyto (a.u.)) of each particle, depicting the main cytometric functional groups identified, namely Orgnano (green dots), Orgpicopro (red dots), Redmicro (black dots), Rednano (purple dots), Redpicoeuk (orange dots) and the Unidentified particles group (green dots). (b) the area under the curve of red fluorescence (RFluo_cyto (a.u.)) versus orange fluorescence (Ofluo_cyto (a.u.)) of each particle, evidencing the Orgpicopro and the Orgnano groups.

of about 20–25° from the surface to $600\,\mathrm{m}$ depth in about 2 h, resulting in distance between consecutive vertical profiles of about $1\,\mathrm{km}$. The glider was equipped with a pumped Seabird CTD probe (Glider Payload CTD), and a Wetlab ECO-puck with chla fluorescence channel sampling at 0.25 Hz, corresponding to a vertical resolution of $0.5 - 0.8\,\mathrm{m}$.

The raw counts from the ECO-puck were converted to chla fluorescence using manufacturer's calibration coefficients and was then corrected near the surface during day-light time from non-photochemical quenching following (Xing et al., 2012). To do so, the mixed layer depth was evaluated using a $0.1\,^{\circ}\mathrm{C}$ criterion on the conservative temperature profiles relative to a reference depth of $10\,\mathrm{m}$ (Houpert et al., 2015). The relative differences of fluorescence are used as a quantitative proxy of the evolution in the distribution of the chla concentration. The glider fluorescence data haven't been calibrated against reference measurements, but agree well with the surface measurements of the ship's adjusted chla concentrations.

2.3 Satellite data

The FUMSECK cruise benefited before, during and after the cruise from the automatic SPASSO software (https://spasso.mio.osupytheas.fr, last access 7 April 2022), which performs real-time processing of CMEMS satellite products (Petrenko et al., 2017; d'Ovidio et al., 2015; Nencioli et al., 2011). The onshore team interpreted the results and sent their daily recommendations on the routes to be taken and the choice of stations to target specific oceanic fine-scale processes like fronts or eddies (Petrenko et al., 2017; Doglioli et al., 2013). Near-real-time products of SSH (Sea Surface Height) and associated geostrophic currents, SST (Sea Surface Temperature), and Sea Surface chla concentration, together with Lagrangian calculations such as FSLE (Finite-Size Lyapunov Exponents) have been used daily from the 2 April 2019 to the 3 July 2019, and all the results are available online on https://spasso.mio.osupytheas.fr/FUMSECK/. The details of the satellite products can be found in Barrillon et al. (2020). A total of 11 daily bulletins (from 23 April to 7 May) have been released and are available online on https://spasso.mio.osupytheas.fr/FUMSECK/Bulletin_web/.

2.4 Meteorological model

The WRF (Weather Research and Forecasting) model, a non-hydrostatic model developed by NCAR (Skamarock et al., 2019), was run with the core ARW (Advanced Research Weather). The horizontal resolution is 2km, the vertical grid is defined with 34 vertical levels. The ARAKAWA-C grid was used one-way with 350 points in X direction and 280 points in Y direction. ARW was forced every six hours by the ECMWF (European Centre for Medium-Range Weather Forecasts) coupling model.

The surface net heat flux and winds were extracted from the model at hourly outputs to characterise the storm event. The net heat flux from the atmosphere to the land/sea surface was computed as : $Q_{net} = Q_{sw} + Q_{lw} + Q_{sens} + Q_{lat}$ with respectively Q_{sw}, Q_{lw} the shortwave and longwave radiations, Q_{sens} the sensible and Q_{lat} the latent heat flux. All fluxes are here downward positive.

2.5 Fluorescences and chlorophyll-a

Different sources to estimate chla concentration were used during the cruise and compared. Absolute chla concentration from Chl_insitu was used as the reference to convert red fluorescence from AFCM and TSG fluorometer into chla concentration based on the significant correlations between them (Fig. 3a).

Fluorescence from the TSG (RFluo_tsg) was converted into units of chla concentration (Chl_tsg, $\mathrm{ng\,mL^{-1}}$) using the significant correlation with Chl_insitu, $\mathrm{Chl\_tsg} = 0.85 \times \mathrm{Rfluo\_tsg} - 0.19$, $\mathrm{r^2} = 0.79$, $\mathrm{n} = 20$. AFCM chla concentration (Chl_cyto) was estimated from the Rfluo_cyto. Values were normalised with 2 μm Polyscience beads, and multiplied by the abundance of each group to get the total normalised Rfluo_cyto per unit of volume (nRFluo_cyto ($\mathrm{a.u\,mL^{-1}}$)). nRFluo_cyto was then compared to the Chl_insitu (Fig. 3a and b). A set of samples from a minicosm experiment (PIANO, unpublished data), acquired with the same chla extraction protocol and the same Cytosense instrument was added to the observations. These samples pre-

sented higher chla concentration values, strengthening the relationship. The linear relation between nRfluo_cyto and Chl_insitu was used to estimate chla concentration for each AFCM phytoplankton group (Chl_cyto, $\mathrm{ng\,mL^{-1}}$) following the linear regression $\mathrm{Chl\_cyto} = 0.11 \times \mathrm{nRFluo\_Cyto}$, $\mathrm{r^2} = 0.97$, $\mathrm{n} = 41$ (Fig. 3b). The origin of the linear regression was not significantly different from 0.

Sea surface chla concentration estimates from three different satellite ocean color algorithms (Chl_ACRI, Chl_MEDOCL3, Chl_MEDOCL4, the product details can be found in Barrillon et al. (2020)) were compared to the other sets of chla concentration estimates for sea surface chla validation (Fig. 3a). Comparisons were done on the period 6:00-18:00 UTC in order to minimise the effect of night extrapolated points. The glider sampling did not follow the ship's route, but a comparison of the 0–5m signal when the ship to glider distance was smaller than 15km showed non-significant difference with the ship's adjusted surface chla concentrations ($0.04 \pm 0.13\,\mathrm{ng\,mL^{-1}}$).

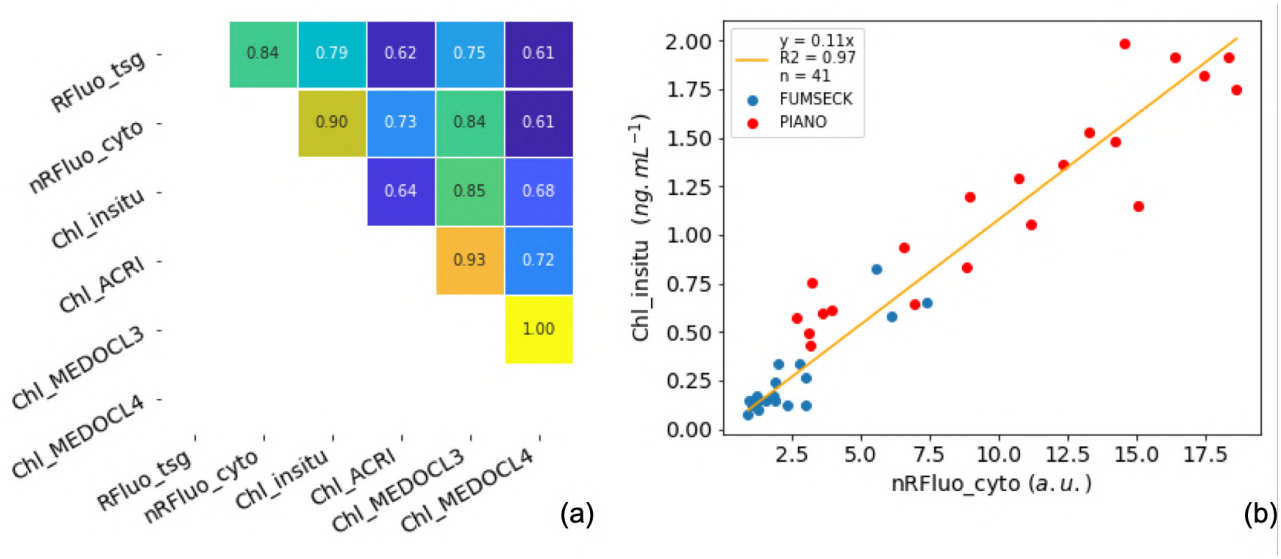


Figure 3. (a) Correlation plot between different sources of fluorescence and chla concentration estimation per unit of volume: Fluorescence from the flow-through fluorometer (Rfluo_tsg (a.u.), $\mathrm{n} = 8543$), sum of all phytoplankton cells normalised red fluorescence from the CytoSense (nRFluo_cyto ($\mathrm{a.u\,mL^{-1}}$), $\mathrm{n} = 403$), chla from in situ discrete sampling (Chl_insitu ($\mathrm{ng\,mL^{-1}}$), $\mathrm{n} = 20$), from the ACRI ocean color product for the 6:00-18:00 UTC day time (Chl_ACRI ($\mathrm{ng\,mL^{-1}}$), $\mathrm{n} = 4555$), from the MEDOCL3 product for the 6:00-18:00 UTC day time (Chl_MEDOCL3 ($\mathrm{ng\,mL^{-1}}$), $\mathrm{n} = 4555$), and from the MEDOCL4 product for the 6:00-18:00 UTC day time (Chl_MEDOCL4 ($\mathrm{ng\,mL^{-1}}$), $\mathrm{n} = 4555$). All the presented correlations were significant at a 0.01 level using a Pearson test. (b) Linear regression between the chla concentration from in situ discrete sampling (Chl_insitu ($\mathrm{ng\,mL^{-1}}$), $\mathrm{n} = 41$) and the sum of all phytoplankton cells normalised red fluorescence from the CytoSense (nRFluo_cyto ($\mathrm{a.u\,mL^{-1}}$)). Two data sets are shown using the same instrument (PIANO and FUMSECK). The intercept coefficient of the regression was not significant at at 10% level (t-test).

3 Results

3.1 Overall circulation

The Ligurian Sea is characterised by a cyclonic general circulation pattern with a geostrophic flow along the coastal line (Esposito and Manzella, 1982). The Northern Current (Millot, 1999), hereafter called NC, is a boundary current in the Northern part of the western Mediterranean circulation.

The general oceanic circulation during the FUMSECK cruise is schematised in Fig. 4. In Fig. 4a the horizontal current velocities averaged over $25-150\,\mathrm{m}$ are shown for the stations, superimposed with the mean chla concentration measured by satellite (Chl_MEDOCL4) from the 1 to the 6 May 2019. The horizontal current velocities are obtained with the vessel-mounted ADCP, averaged during the 20 min preceding the arrival at each station. The boundaries of the different hydrodynamic zones were drawn based on Chl_MEDOCL4 concentration isolines. The region of the NC (hatched in purple, $< 0.12\,\mathrm{ng.mL^{-1}}$) corresponds to the lowest Chl_MEDOCL4 concentration. The southeastern part of the cyclonic recirculation (hatched in orange, $> 0.15\,\mathrm{ng.mL^{-1}}$) shows the highest Chl_MEDOCL4 concentrations. These two zones are separated by a region, hereafter referred to as the intermediate zone (hatched in green, $0.1-0.15\,\mathrm{ng.mL^{-1}}$). The vessel-mounted ADCP horizontal currents at $26.5\,\mathrm{m}$-depth along the cruise (Fig. 4b) show the high-intensity of the NC ($0.43\,\mathrm{m\,s^{-1}}$ mean velocity in the core of the NC) with respect to the cyclonic recirculation zone ($0.18\,\mathrm{m\,s^{-1}}$ mean velocity).

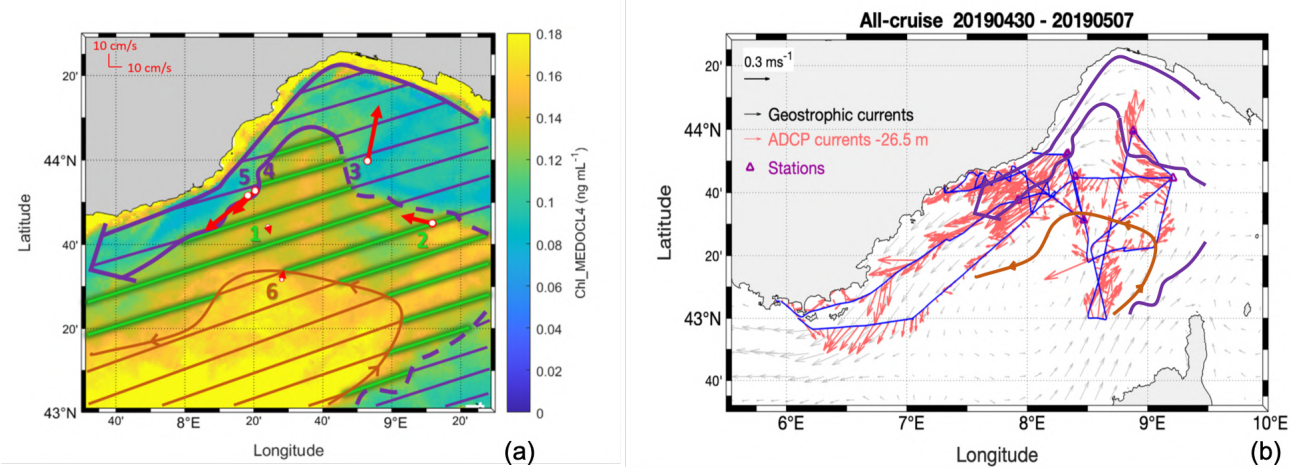


Figure 4. (a) Satellite chla averaged concentration (Chl_MEDOCL4, $\mathrm{ng\,mL^{-1}}$) from 1 to 6 May 2019, used to set the drawn hatched boundaries between the hydrodynamic zones, superimposed with horizontal velocities (VM-ADCP, red vectors) at the stations, averaged over $25-150\,\mathrm{m}$. (b) ADCP horizontal currents at $26.5\,\mathrm{m}$-depth superimposed on surface geostrophic currents from satellite altimetry.

3.2 Storm

During this cruise, an episode of particularly intense winds hit the south of France and the Ligurian Sea. In particular the Ligurian Sea was exposed to two main winds : NW (Mistral wind) with intensities between 93 and 130 $\mathrm{km\,h^{-1}}$, and N (Tramontana wind) with intensities between 74 and 93 $\mathrm{km\,h^{-1}}$. In this zone, this episode began during the night between the 4 and 5 May 2019 reached its maximum intensity on the 5 May around 5:00 am, and finished on 05 in the evening.

Although the conjunction of these two winds is a classical situation in the Ligurian Sea, this event was particularly intense. The analysis of coastal data in the South of France by Meteo France shows winds of intensity $> 100\,\mathrm{km\,h^{-1}}$ occur on average 6 times per year, but only once every 4 years for winds $> 150\,\mathrm{km\,h^{-1}}$.

The ship came back in the storm zone during the night between 5 and 6 May. The model shows that at the storm maximum on the 5 May around 5 am, the ship-sampled zone (marked with squares Fig. 5) was affected by a wind intensity peak of 26 $\mathrm{m\,s^{-1}}$ (108 $\mathrm{km\,h^{-1}}$) associated to an intense negative net heat flux of -400 $\mathrm{W\,m^{-2}}$. This sampled zone was in the core of a corridor area (8° E 42.5-44.5° N) with strong wind intensities and high negative heat fluxes (Fig. 5). The glider was on-site during the storm, on its northward return transect.

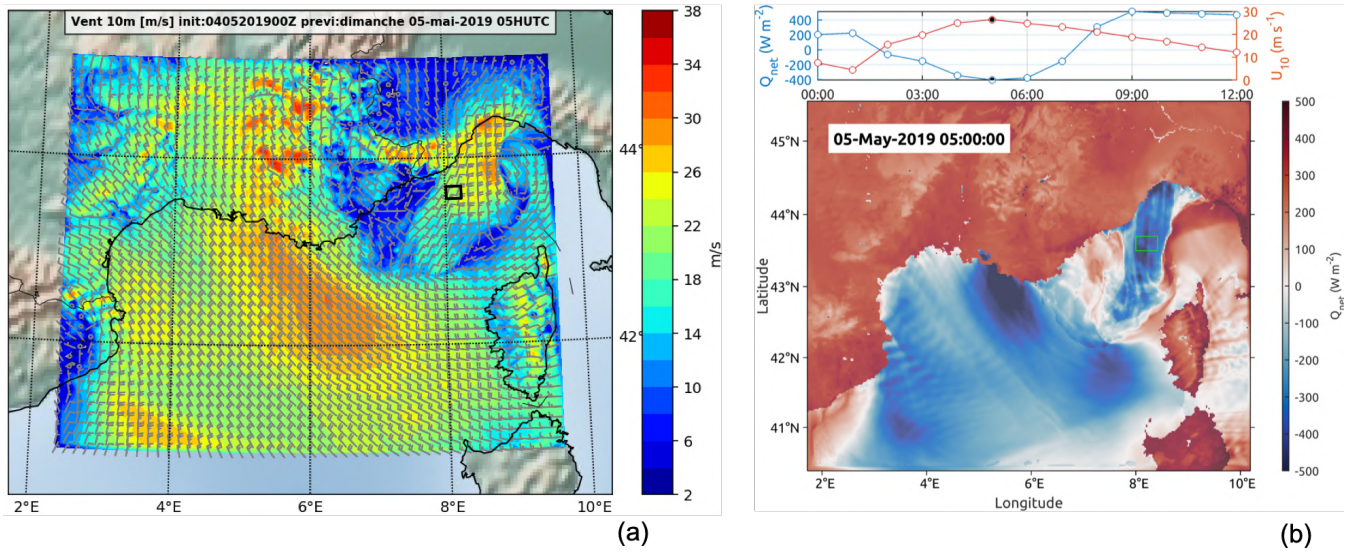


Figure 5. Results of the wind situation on the 5 May (WRF model WRF-ARW v4.2.1). (a) Wind intensity at 10 m on the 5 May, 05:00. (b) Temporal distribution of wind intensity and heat flux on 5 May between 0:00 and 12:00 (top) in the green squared area marked on the bottom plot representing the heat flux on the 5 May, 5:00.

3.3 Surface hydrodynamics and hydrology

The general properties of the surface waters include surface conservative temperature, absolute salinity, chla concentration (Chl_tsg and Chl_insitu), and in situ Nitrate (NO_3) concentration (Fig. 6, 7). The conservative temperature is globally warmer near the coast and in the NC (mean value of 15.7°C in the NC), and cooler in the intermediate and recirculation zone (mean value of 15.4°C in the recirculation zone). The absolute salinity is lower near the coast and in the NC (mean value of $38.12\mathrm{g.kg}^{-1}$ in the NC), and higher in the intermediate and recirculation zone (mean value of $38.38\mathrm{g.kg}^{-1}$ in the recirculation zone). The TSG chla (Chl_tsg) concentration mean value is 0.29 $\mathrm{ng\,mL}^{-1}$ over the whole cruise, with a lower mean value in the NC (0.21 $\mathrm{ng\,mL}^{-1}$) than in the recirculation zone (0.33 $\mathrm{ng\,mL}^{-1}$).

When the ship came back offshore less than 24 h after the maximum storm intensity, we observed a patch of low-temperature ($< 14.8°\mathrm{C}$) and high-salinity ($> 38.28\,\mathrm{g\,kg}^{-1}$) water, with a sharp horizontal gradient separating it from surrounding waters (Fig. 6). This patch was associated with an increase in mean chla : Chl_insitu rised up to $0.65\,\mathrm{ng\,mL}^{-1}$ while the mean value for the whole cruise was $0.25\,\mathrm{ng\,mL}^{-1}$, similarly, chla_tsg maximal value inside the patch was of $1.11\,\mathrm{ng\,mL}^{-1}$. The nutrients also showed an increase, in particular the NO_3 concentration which was up to $1.25\,\mu\mathrm{M}$, for a mean value of $0.15\,\mu\mathrm{M}$ for the whole cruise (Fig. 7b). This particular zone of interest is highlighted in cyan in Fig. 6 and 7 and corresponds to longitudes between $8°$ E and $8°15'$ E and latitudes between $43°33'$ N and $43°42'$ N.

The TS diagram is classically used to describe water masses. In Fig. 8, we have classified the water masses using the absolute salinity $\mathrm{S_A}$ and the conservative temperature Θ, from black for deeper, denser waters ($\mathrm{S_A} \geq 38.61\,\mathrm{g\,kg}^{-1}$) to lighter orange/yellow tones for the shallower ones ($\mathrm{S_A} < 38.61\,\mathrm{g\,kg}^{-1}$). Hence surface waters include mostly yellow waters ($\mathrm{S_A} \leq 38.46\,\mathrm{g\,kg}^{-1}$ for $\Theta \leq 13.8°\mathrm{C}$, and $\mathrm{S_A} \leq 38.38\,\mathrm{g\,kg}^{-1}$ for $\Theta > 13.8°\mathrm{C}$) and orange waters ($38.38\,\mathrm{g\,kg}^{-1} < \mathrm{S_A} \leq 38.62\,\mathrm{g\,kg}^{-1}$ & $\Theta > 13.8°\mathrm{C}$). As can be seen in Fig. 8b, the yellow waters are present at the surface in the NC area and in the intermediate zone and will be thereafter named NC waters; while the orange waters are localised at the surface offshore in the recirculation zone of the basin-scale cyclonic circulation and will be thereafter called recirculation waters.

The cold surface water patch was encountered, after the storm, by the ship in the geographical cyan area in Fig. 6, and in addition by the glider, during the storm in its ascending route (Fig. 12a). The characteristics of this cold surface water patch ($38.31\,\mathrm{g\,kg}^{-1} \leq \mathrm{S_A} \leq 38.45\,\mathrm{g\,kg}^{-1}$ for $14\,°\mathrm{C} \leq \Theta \leq 14.5\,°\mathrm{C}$ and $38.28\,\mathrm{g\,kg}^{-1} \leq \mathrm{S_A} \leq 38.38\,\mathrm{g\,kg}^{-1}$ for $14.5°\mathrm{C} \leq \Theta \leq 14.78°\mathrm{C}$) are superimposed in cyan on the TS diagram. They correspond to either NC or recirculation waters, with a density around $28.37 - 28.70\,\mathrm{kg\,m}^{-3}$, and are present around $30 - 40\,\mathrm{m}$ depth before the storm, as can be seen in Fig. 9a. Between $43°\ 31'$ N and $43°\ 39'$ N, these waters have been detected between $50\,\mathrm{m}$ and the surface, by both the MVP during its 7^{th} transect (after the storm) and the glider at the end of his ascending route (during the storm), as can be seen in Fig. 9b. These waters, thereafter called newly-mixed waters, are present up to the surface in a very localised spot in space and time (Fig. 8c), and are represented

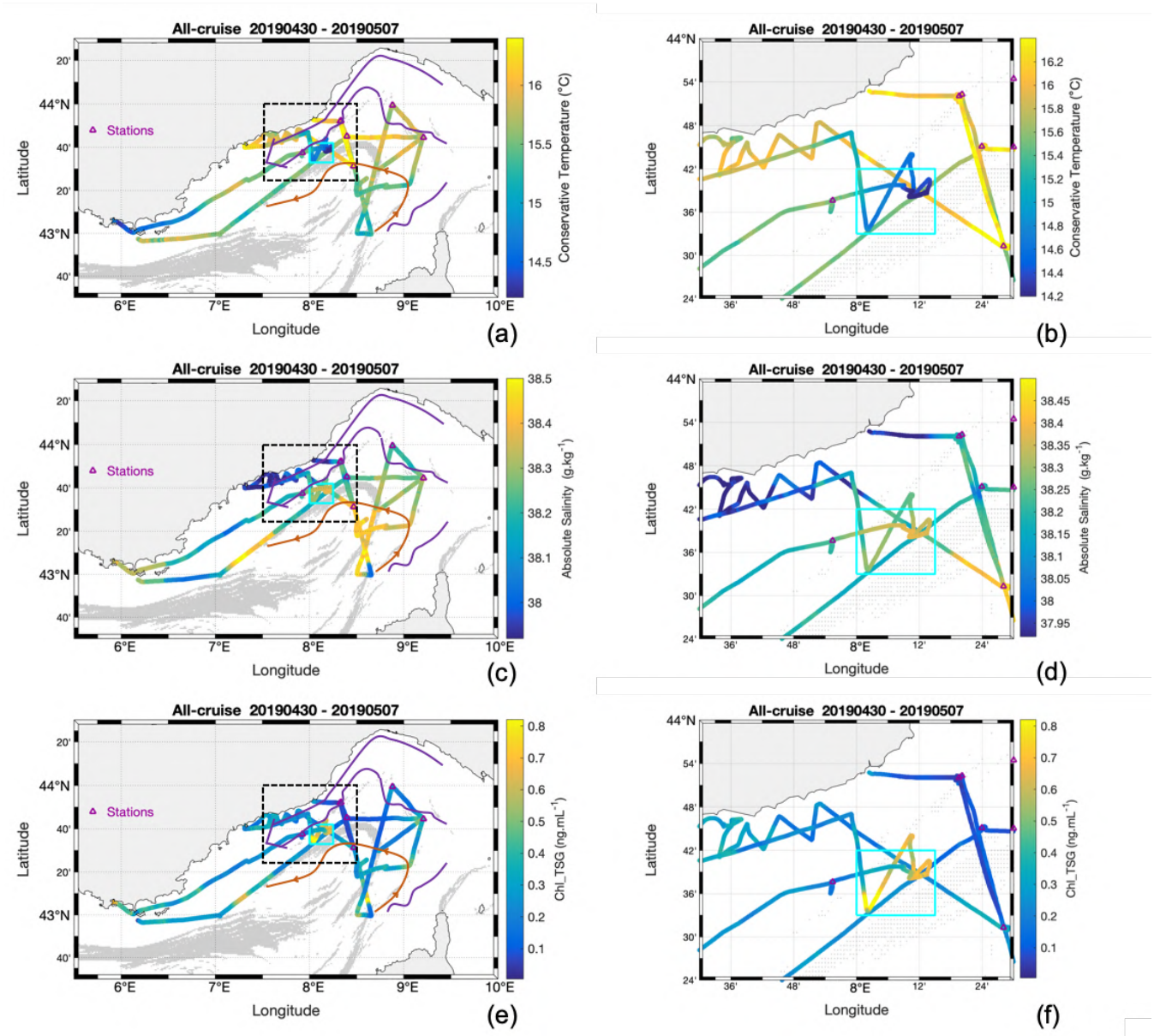


Figure 6. Water surface characteristics from TSG along the cruise, superimposed with FSLE calculated from altimetry. (a) (b) Sea surface conservative temperature. (c) (d) Absolute salinity. (e) (f) Chla_tsg concentration Stations are indicated by purple triangles. Left panels show the whole geographic region of the cruise and right panels illustrate the zoom of the indicated region (black dotted square), identifying the particular TSG region of interest (cyan square) sampled one day after the storm.

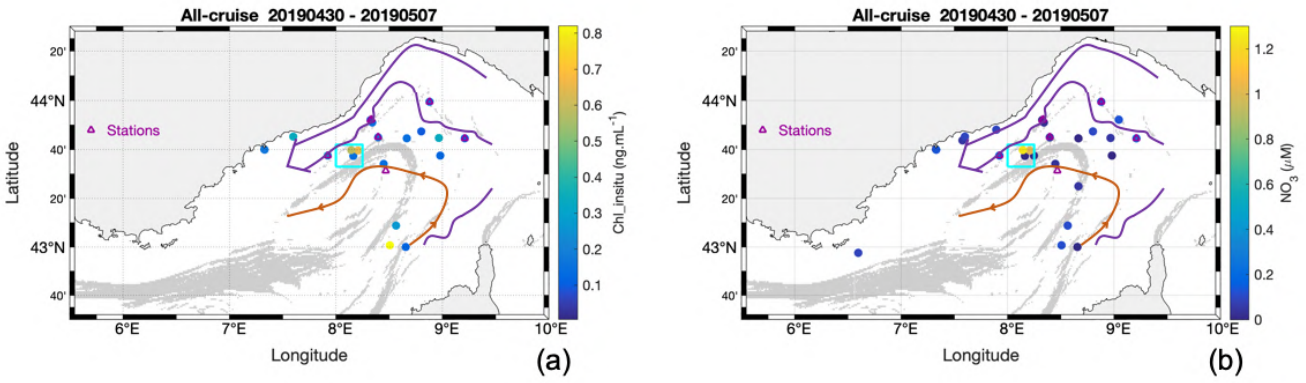


Figure 7. Water surface characteristics from discrete in situ sampling along the cruise. (a) Chl_insitu concentration. (b) NO_3 concentration. Stations are indicated by purple triangles. The cyan square identifies the particular TSG region of interest sampled one day after the storm.

in cyan through the paper.

250 The vessel crossed these surface newly-mixed waters on the 6 May between 2:32 am and 2:53 am , 3:03 am and 4:03 am , and 5:32 am and 11:36 am , with the vessel moving in and out of these waters. The glider encountered the surface newly-mixed waters on its way North around 10 am on the 5 May. It was at this time at about 85 km from the ship and stayed in these waters until its recovery on the morning of the 6 May.

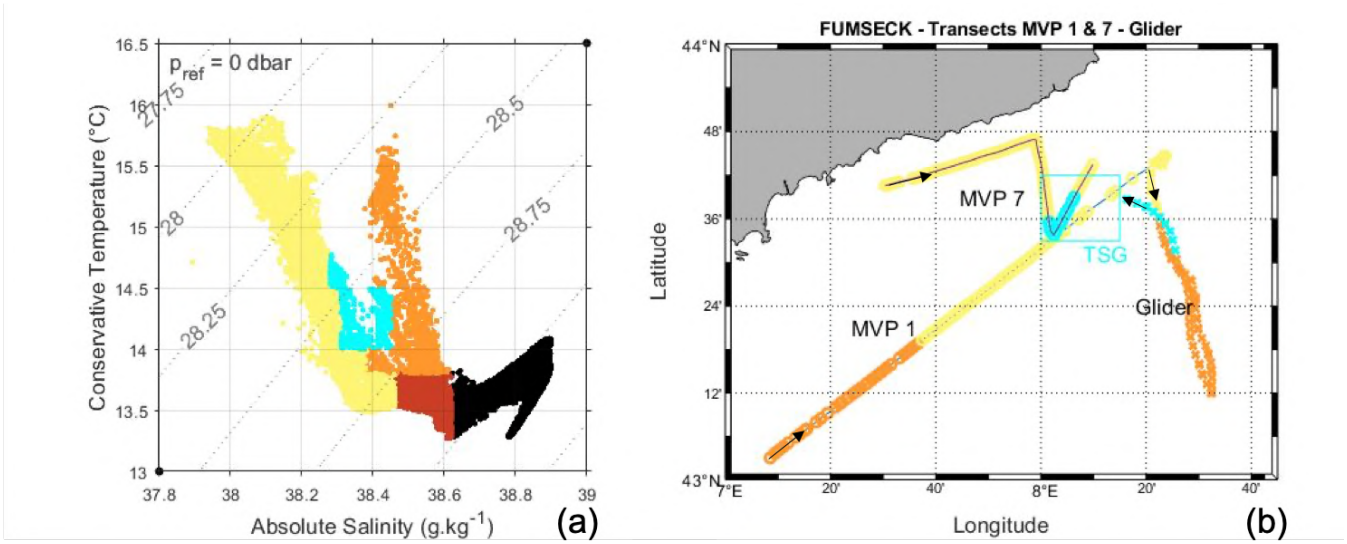


Figure 8. Water masses types measured by the MVP (MVP 1 from 30 April 21:29 to 1 May 7:50, and MVP 7 from 5 May 19:22 to 6 May 5:06) and the glider (descending from 1 May 8:50 to 04 May 0:29, and ascending from 4 May 0:29 to 6 May 3:42). (a) TS diagram from MVP 7 and ascending glider data. (b) Map with the surface colored waters measured by MVP and glider, and TSG zone of interest. The colors are as follows: recirculation waters in orange, NC waters in yellow, and newly-mixed waters in cyan.

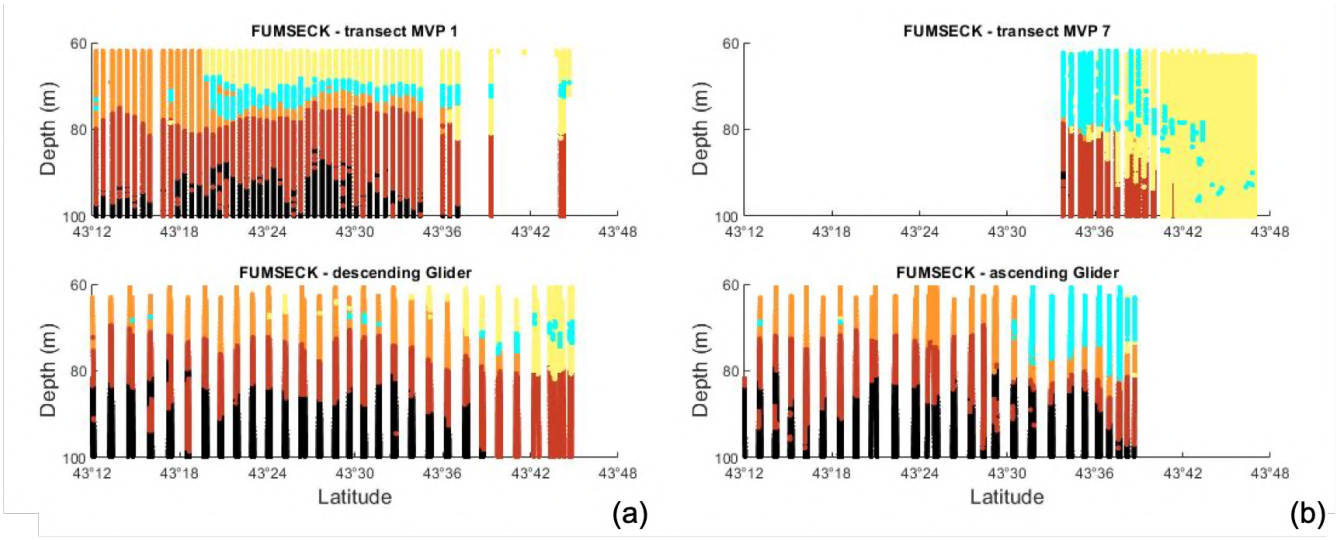


Figure 9. (a) Vertical transects versus longitude with associated colored waters, top panel for the MVP 1, bottom panel for the descending glider. (c) Vertical transects versus longitude with associated colored waters, top panel for the MVP 7, bottom panel for the ascending glider.

3.4 Chlorophyll-a and global biomass

Chl_insitu varies between 0.07 and 0.82 $\mathrm{ng\,mL^{-1}}$ with a mean $\pm$ sd of 0.25 $\pm$ 0.21 $\mathrm{ng\,mL^{-1}}$, with 20 samples collected all along the cruise (Fig. 7a, Fig. 10a). The standard deviations are representative of the variability, not the measurement errors. Chl_cyto values follow a similar trend with minimal and maximal values of 0.03 and 0.94 respectively, and a mean $\pm$ sd of 0.26 $\pm$ 0.16 $\mathrm{ng\,mL^{-1}}$ (Fig. 10a, 7b). Average spatial resolution is of 2.9 km, with a 20 min sampling strategy and 403 points. Chl_tsg varies between undetectable values and 1.11 $\mathrm{ng\,mL^{-1}}$, with a mean of 0.29 $\pm$ 0.16 $\mathrm{ng\,mL^{-1}}$ and a mean spatial resolution of 0.16 km with a total of 8453 points (Fig. 6d, Fig. 10b).

Ocean color chla match-ups with Chl_cyto are selected during day time (6:00-18:00, Fig. 10b) and are significantly higher for Chl_ACRI than for Chl_MEDOCL4 (0.27 $\pm$ 0.07 and 0.15 $\pm$ 0.05 $\mathrm{ng\,mL^{-1}}$, $p < 0.001$, block-bootstrap test (appenndix A). Maximal values of Chl_ACRI and Chl_MEDOCL4 (0.48 and 0.51 $\mathrm{ng\,mL^{-1}}$, respectively) are below the maximal values of Chl_cyto and Chl_tsg.

Total biomass of phytoplankton ranges between 13.75 and 77.94 $\mathrm{ngC\,mL^{-1}}$ with a mean of 33.05$\pm$ 11.23 $\mathrm{ngC\,mL^{-1}}$ and follows Chl_cyto trends with a correlation of 0.52 (n=403) when considering the entire data set, and of 0.72 (n=382) when removing the data from the newly-mixed waters. For the newly-mixed waters, correlation is of 0.78 (n=21).

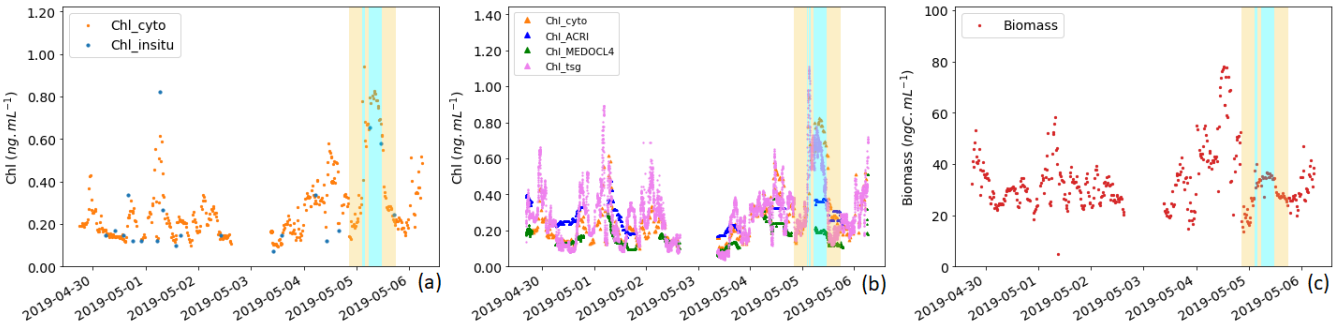


Figure 10. (a) Comparison between the chla concentration measured in situ (Chl_insitu, $\mathrm{ng\,mL^{-1}}$) and the chla concentration estimation obtained from AFCM (Chl_cyto, $\mathrm{ng\,mL^{-1}}$). (b) Comparison between the chla concentration estimated by the AFCM (Chl_cyto), the ACRI (Chl_ACRI), the MEDOCL4 (Chl_MEDOCL4) approach and the fluorometer (Chl_tsg). (c) Total phytoplankton biomass variation through the cruise ($\mathrm{ngC\,mL^{-1}}$). The period corresponding to the surface crossing of the newly-mixed waters (in cyan) and the surrounding NC ones (6 hours before and after the newly-mixed ones, in yellow) are indicated.

3.5 Phytoplankton groups and reaction

The most abundant group belongs to the Orgpicopro, followed by the Rednano, Redpicoeuk, Orgnano and Redmicro (see Table 1). Inversely, the Rednano biomass was the highest, followed by the Orgpicopro biomass. Redpicoeuk biomass was the

lowest. Chlorophyll per group per unit of volume regarding the overall study area was also the highest for the Rednano followed by the Orgpicopro.The biomass/Chl_cyto ratio was above 127 for all phytoplankton groups when considering the entire study area.

For all phytoplankton groups except for Orgpicopro, abundances and biomass per unit of volume are twice higher in newly-mixed waters (cyan in Fig. 8a) compared to NC surrounding waters (yellow in Fig. 8a) as shown in Tab. 1 and 11. All groups have higher chla values in the newly-mixed waters (Tab. 1). Conversely, Rednano and Redpicoeuk estimated average sizes are higher with a concomitant higher biomass per cell in the NC surrounding waters than in the newly-mixed ones (Tab. 1). The biomass/Chl_cyto ratios is lower in NC waters and even more in newly-mixed waters compared to the overall area (see Fig. 14) for all groups, despite lower carbon content per cell. In short, the newly-mixed waters evidence higher abundances and higher chla concentration and biomass per unit of volume but smaller sizes and biomass per cell (mainly for the Redpicoeuk and Rednano).

Observable	*Waters*	*Orgpicopro* **mean ± SD**	*Redpicoeuk* ***mean ± SD***	*Rednano* ***mean ± SD***	*Orgnano* ***mean ± SD***	*Redmicro* ***mean ± SD***
Abundance (cell mL^{-1})	Overall	51556 ± 21827	1017 ± 473	3686 ± 887	211 ± 192	3 ± 2
	NC surrounding	63239 ± 29087	**1175 ± 397**	**2746 ± 546**	**160 ± 84**	**4 ± 2**
	Newly-mixed	61162 ±4898	**2334 ± 392**	**4597 ± 333**	**325 ± 34**	**6 ± 1**
Size (ESD, μm)	Overall	0.98 ± 1.02	2.18 ± 1.76	3.30 ± 2.45	5.22 ± 4.50	11.38 ± 10.72
	NC surrounding	0.96 ± 0.97	**2.14 ± 1.69**	**3.18 ± 2.35**	4.91 ± 4.40	9.98 ± 8.46
	Newly-mixed	0.94 ± 1.02	**1.92 ± 1.61**	**3.02 ±2.29**	4.83 ± 4.45	9.64 ± 8.16
Biovolume (μm^3)	Overall	0.50 ± 0.56	5.53 ± 2.95	19.16 ± 7.91	75.24 ± 47.90	1051.26 ± 1586.96
	NC surrounding	0.47 ± 0.50	**5.24 ± 2.58**	**16.97 ± 6.86**	62.6 ± 45.00	585.94 ± 566.35
	Newly-mixed	0.45 ± 0.57	**3.72 ± 2.18**	**14.54 ± 6.32**	59.33 ± 46.46	470.97 ± 286.68
Biomass/cell (pgC $cell^{-1}$)	Overall	0.14 ± 0.16	1.13 ± 0.66	5.52 ± 2.57	18.00 ± 12.20	165.31 ± 216.90
	NC surrounding	0.14 ± 0.14	**1.08 ± 0.59**	**4.98 ± 2.28**	15.36 ± 11.56	103.26 ± 90.60
	Newly-mixed	0.13 ± 0.16	**0.80 ± 0.81**	**4.36 ± 0.12**	14.68 ± 11.82	87.70 ± 58.49
Chl_cyto (ng mL^{-1})	Overall	0.061 ± 0.051	0.006 ± 0.006	0.184 ± 0.104	0.014 ± 0.015	0.003 ± 0.002
	NC surrounding	**0.100 ± 0.600**	**0.009 ± 0.005**	**0.173 ± 0.083**	**0.012 ± 0.001**	**0.004 ± 0.002**
	Newly-mixed	**0.160 ± 0.014**	**0.025 ± 0.005**	**0.526 ± 0.079**	**0.032 ± 0.000**	**0.006 ± 0.001**
Biomass (ngC mL^{-1})	Overall	7.47 ± 3.97	1.12 ± 0.43	20.30 ± 5.58	3.77 ± 3.75	4.81 ± 4.82
	NC surrounding	7.72 ± 2.71	**1.22 ± 0.29**	**13.58 ± 2.18**	**2.38 ± 1.09**	**3.77 ± 1.94**
	Newly-mixed	7.90 ± 3.61	**1.86 ± 0.24**	**20.05 ± 1.39**	**4.77 ± 4.81**	**5.34 ± 1.26**
Biomass/Chl_cyto	Overall	158.10 ± 56.3	268.4 ± 99.2	127.4 ± 43.2	292.1 ± 71.6	206.4 ± 202.5
	NC surrounding	**86.0 ± 26.3**	**158.0 ± 40.9**	**87.5 ± 21.5**	**218.5 ± 33.2**	115.6 ± 94.2
	Newly-mixed	**48.6 ± 3.8**	**76.8 ± 9.5**	**38.6 ± 4.3**	**148.4 ± 9.9**	93.6 ± 142.0

Table 1. Mean ± standard deviation values for abundance, size (equivalent spherical diameter ESD), biovolume per cell, biomass per cell, chla per unit of volume (Chl_cyto), biomass per unit of volume and the ratio biomass over Chl_cyto for the overall sampling waters (n=400), the NC surrounding waters (n=20) and the newly-mixed waters (n=43) (Fig. 8) and for the five AFCM phytoplankton groups identified. The surrounding NC waters corresponds to the NC waters 6 hours before and after the newly-mixed ones. A Moving Blocks Bootstrap test between NC surrounding and newly-mixed waters reveal significant differences, bold values are significantly different at a Bonferroni-corrected 5%.

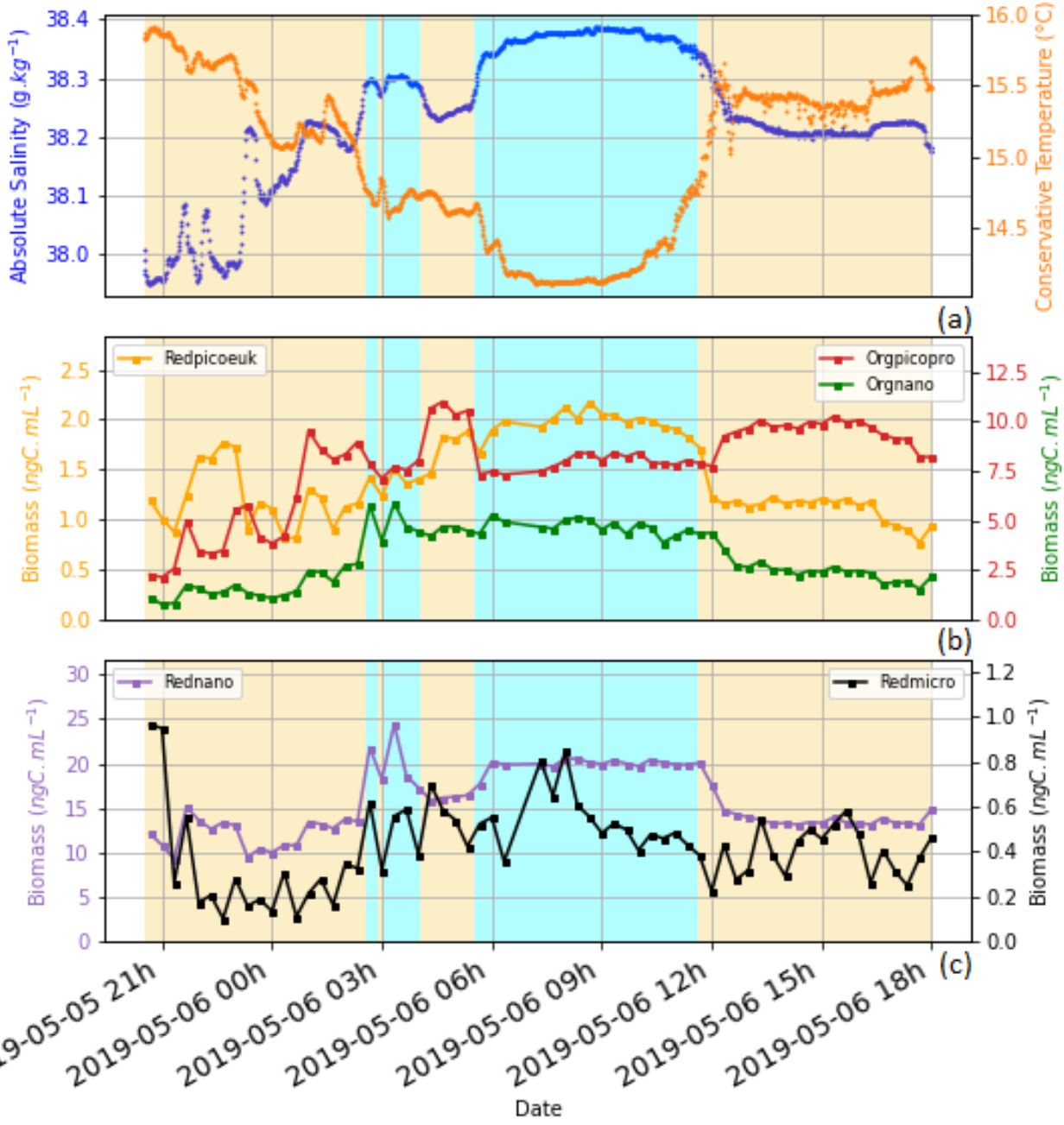


Figure 11. Illustration of the newly-mixed waters (in cyan spans) and their direct surroundings (NC waters, in yellow spans), in terms of temperature, salinity and biomass per phytoplankton group ($\mathrm{ngC\,mL^{-1}}$). (a) Variation of the Absolute Salinity (blue dots) and Conservative temperature (orange dots). (b) Variation of the biomass for Redpicoeuk (Orange line), Orgpicopro (Red line) and Orgnano (Green line). (c) Variation of the biomass for Rednano (violet line) and the Redmicro (black line).

3.6 Subsurface fluorescence signal observed by the glider

Referring to the surface water masses of section 3.3, the glider entered the newly-mixed surface waters on its northward return transect on the 5 May, leaving behind recirculation waters (Fig. 8c). Down to approx. 60 m depth, the surface temperature and salinity steeply decrease (see Fig. 12), moving from recirculation waters to newly-mixed waters. The fluorescence near the surface increases rapidly by a factor four (Fig. 13b) as the mixed layer depth recorded by the glider deepens from 15 to 50 m (Fig. 13a). However, the integrated fluorescence content in the upper 100 m did not show any significant variation (Fig. 13b). This indicates that the increase in chla concentration observed near the surface (Fig. 6) is likely due to the dilution by vertical mixing of the phytoplankton cells within the mixed layer.

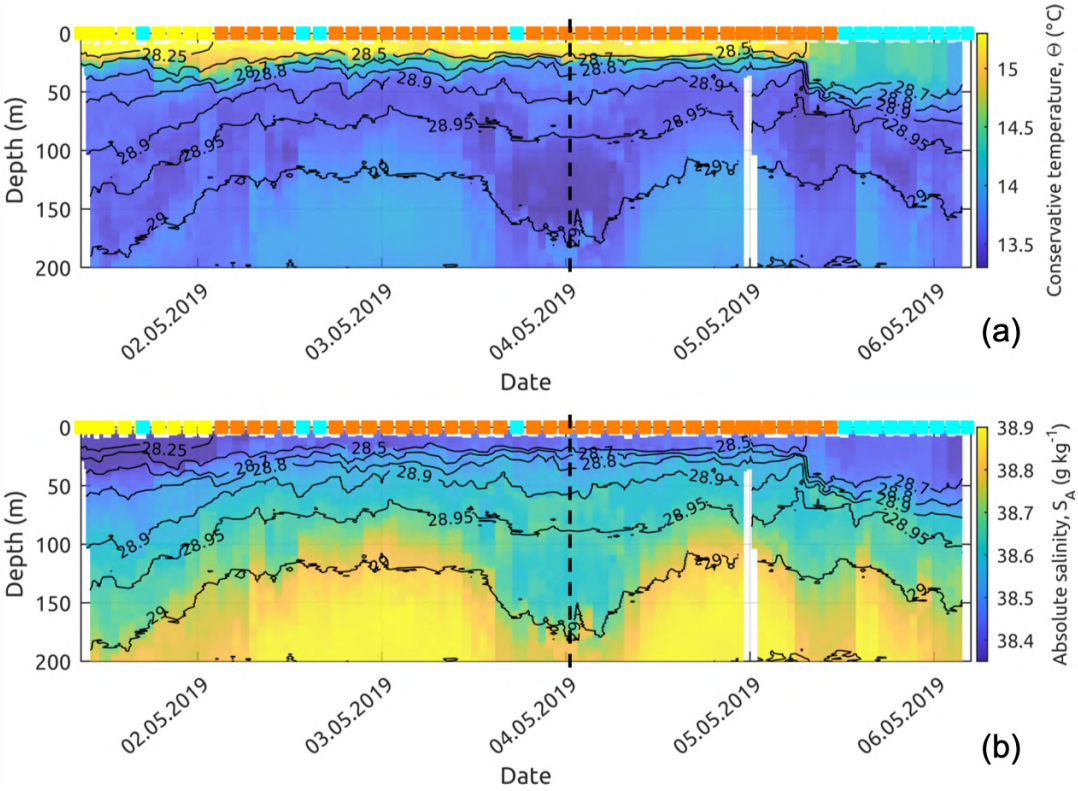


Figure 12. Glider profiles of (a) conservative temperature and (b) absolute salinity. The colored squares correspond to the dominant water mass according to Fig. 8 observed at 10m depth by the glider. The dashed vertical line represents the time separating the descending and ascending transects.

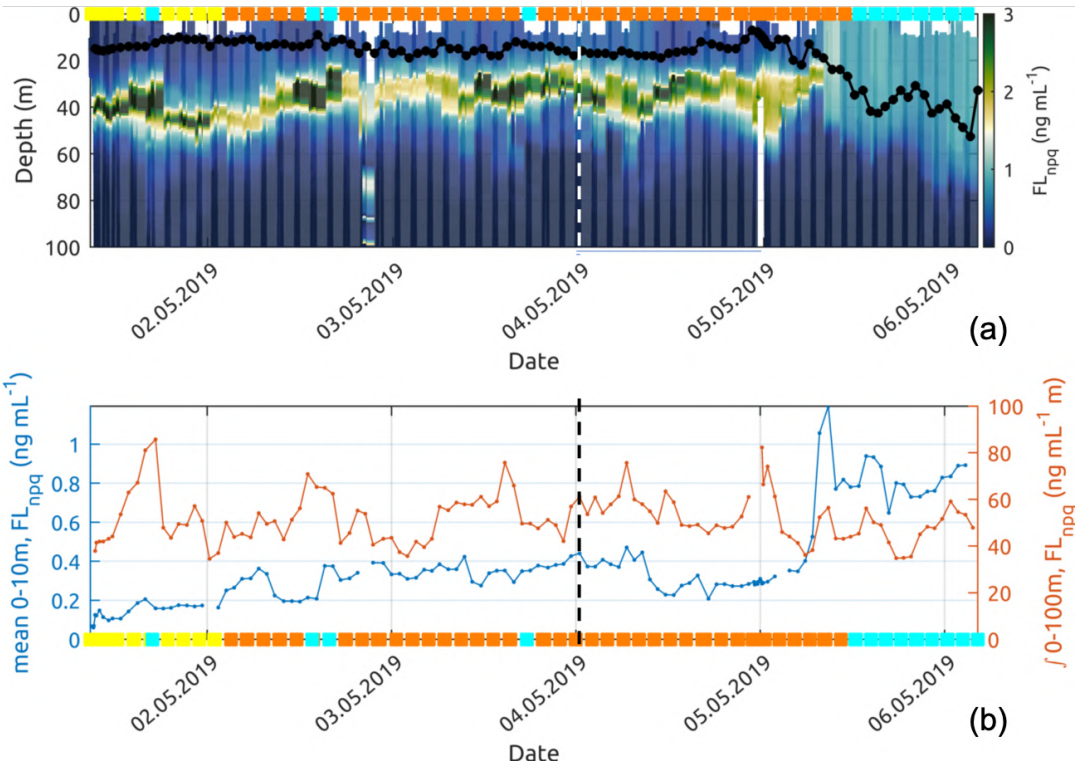


Figure 13. . (a) Fluorescence observed by the glider and corrected from non—photochemical quenching following (Xing et al., 2012). The black line with dots shows the mixed layer depth (MLD) at each glider profile. (b) Near-surface (0-10 m average) and integrated over 0-100 m chla fluorescence concentration along the glider track. The colored squares correspond to the dominant water mass according to Fig. 8 observed at 10m depth by the glider. The dashed vertical line represents the time separating the descending and ascending transects.

4 Discussion

During the FUMSECK cruise, occurring in May, an intense storm dominated by Northerly winds impacted the water column which is at that time of the year considered as stratified with surface nutrient availability nearly undetectable (Pasqueron De Fommervault et al., 2015). The physical and biogeochemical data, collected thanks to the deployment of high-resolution sensors, show a clear localised event after the storm, with a steep change of temperature and salinity, an increase of chla corroborated by a significant phytoplankton increase in terms of biomass and abundances. Above this area, a wind intensity peak of 30 $\mathrm{m\,s^{-1}}$ associated to an intense negative net heat flux of -400 $\mathrm{W\,m^{-2}}$ was observed, around 5 am.

In general, abundances of phytoplankton groups were much higher than the ones observed at the same location during the OSCAHR cruise in November 2015 (Marrec et al., 2018) for the Rednano and the Orgpicopro, but similar for the Redpicoeuk. The size of Rednano and Redpicoeuk were smaller in average than the ones observed during the OSCAHR cruise, but larger for the Orgpicopro. diagram (Fig. 8). The conversion into chla from the total red fluorescence displayed Rednano as the main contributor during the entire study, with a similar picture for its biomass. All groups exhibited higher Chla_cyto in the newly-mixed waters while cells were almost smaller. Similarly, fluorescence per group were much higher in the cold core of the OSCAHR eddy than in the surrounding warm water, but the difference was not higher than 1.5, compared to our study, where Chl_cyto for Rednano and for Redpicoeuk were nearly 3 times higher in the cold newly-mixed waters. Such increase in chlorophyll after the deepening of the mixed layer depth during post bloom periods and linked to wind events is not obvious as demonstrated by Andersen and Prieur (2000).

The carbon/chla ratio calculated in this paper aims at contributing to the estimated ratio from field studies with much higher precision thanks to the clear separation between phytoplankton and bulk particulate organic carbon given by AFCM. The conversion into biomass of carbon can be discussed by the possible shift in size estimates from single cell scatter, affecting directly biomass conversion, but also biomass conversion factors from the literature. Nevertheless, the high variability in the ratio values per phytoplankton group indicates they are not having similar metabolisms, Redpicoeuk having much higher ratio (268) that Rednano (127). The higher ratio are similar to the ones observed in coastal areas, and the lowest are similar to the one observed in open surface waters, as observed in the study of (Calvo-Díaz et al., 2008) were values for picoeukaryotes varied from 0.07 to 282. Generally, the carbon/chla ratios presented in our study are high compared to the traditional value of 50, and are much higher than values found in high nutrient environments with lower light conditions (Jakobsen and Markager, 2016). The carbon/chla ratio integrating all groups varies from approx. 90 to 250 in surface conditions but drops down to 50 in the newly-mixed waters (Fig. 14). The general high values could evidence the high light and low nutrient conditions of the post bloom oligotrophic period sampled in the Ligurian Sea. The remarkable drop in the ratio observed in the cold water patch is a clear signature of a sudden change in phytoplankton cell physiology and translates the unadapted configuration of the cells to high light conditions (Jakobsen and Markager, 2016).

While surface observations only suggested an increase in chla concentrations (Fig. 6 and 7), the integrated chla values from the glider fluorometer clearly evidenced that this surface increase is due to a dilution of the deep chlorophyll maximum in the mixed layer during the storm (Fig. 13). The deepening of the mixed layer depth and the dilution of phytoplankton cells previously concentrated in the well-know limited layer of deep chlorophyll maximum by vertical mixing is a punctual event with potential consequences on the carbon fluxes in this oligotrophic area. Indeed, the increase in nutrients in the water column due to the uplift of the nitracline, followed by a spreading of the phytoplankton in the upper layer, and a possible dilution of the grazers, could lead to an increase in integrated primary production by enhancing division rate (Behrenfeld, 2010), followed by an accumulation of biomass. Although the increase in biomass in the newly mixed water column is hypothetical and was not observed because we were not on site for a longer period, this expected small scale post-bloom situation leads to a different process than a classical spring bloom setup as it originates from a DCM dilution.

Because the cruise was ending, the results presented here only captured the short-term physical and phytoplankton reaction to the storm, seriously limiting the interpretations in terms of long lasting responses. For future work, the objective will be to study the medium to long term reaction, after the so-called reaction period, and for each observed phytoplankton group. Indeed, such events are critical, as they may affect the primary production annual budgets. These results highlight the need of concomitant observations of physics and biology with high spatio-temporal resolution in order to understand the effect of physical forcing events, such as storms, on marine ecosystems.

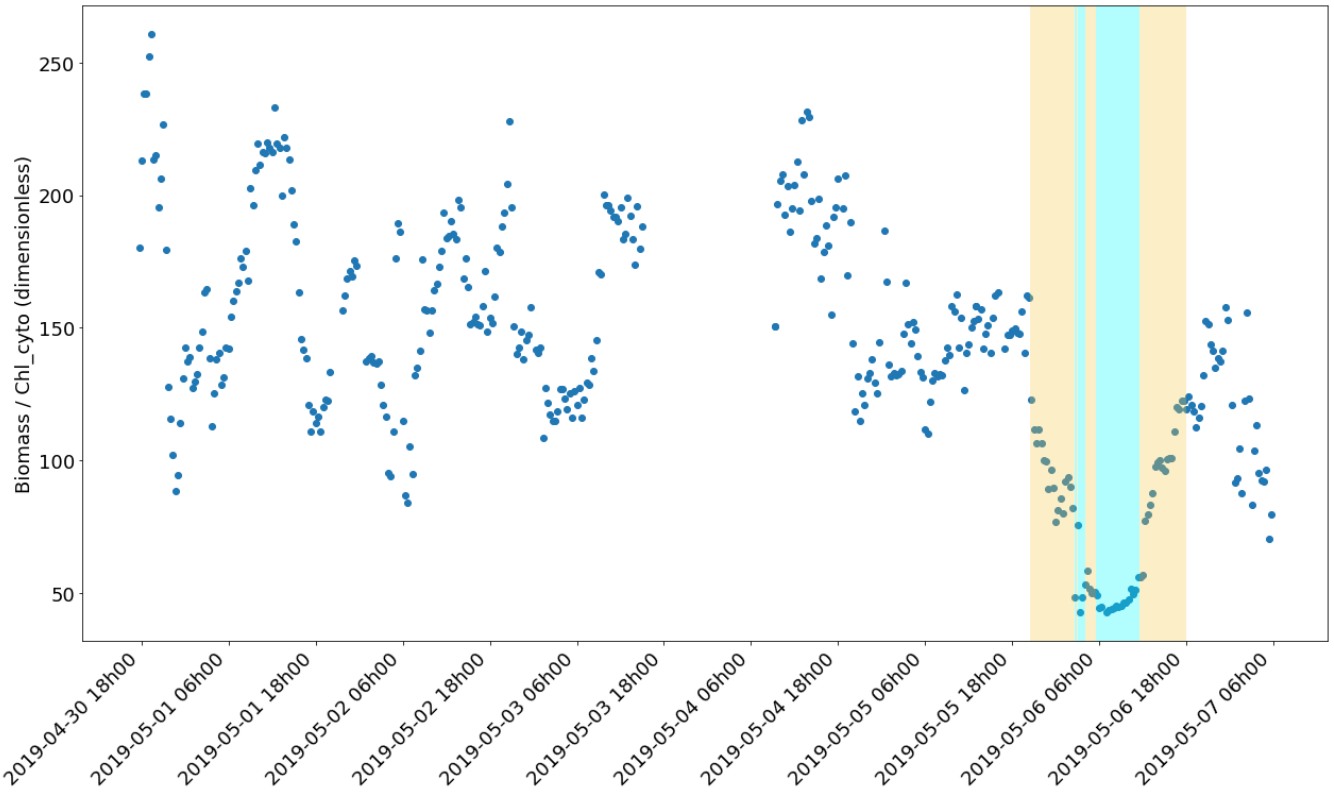


Figure 14. Evolution of the Biomass ($\mathrm{ngC\,mL^{-1}}$) / Chl_cyto ($\mathrm{ng\,mL^{-1}}$) ratio through the cruise. The yellow and cyan color spans correspond to the water masses of Fig. 11.

5 Conclusion

During the FUMSECK cruise, the deployment of high-frequency and high-resolution instruments has enabled to observe the link existing between the fine-scale physical structures and phytoplankton size-class distribution in the Ligurian Sea. The studied area was under typical post-bloom physical and biological characteristics of the NW Mediterranean Sea with surface stratified conditions, a deep chlorophyll maximum and close to undetectable surface concentration of chlorophyll, where cells < 4-5 μm dominated biomass.

A storm of high intensity occurred during the cruise period and effects on the water column and the phytoplankton were specifically studied thanks to the concordance between a glider, a vertical moving profiler, a surface thermo-salinometer, an automated flow cytometer, strengthened by satellite data and discrete samples collected for nutrients and chla concentration. The zone of interest affected by the storm was characterized by surface waters coming up from depths down to $60\,\mathrm{m}$, with a clear dilution of the deep chlorophyll maximum, leading to abrupt changes in the phytoplankton abundances in surface waters. Furthermore, the study of phytoplankton at the single cell level showed clear physiological changes as a signature of sudden ecosystem changes, as evidenced by a drop in carbon/chla ratio but an increase in abundances and biomass. The storm, although identified as a rare event in this area, should be considered as an important feature to study within the fine-scale physical biological coupling, especially in oligotrophic conditions, where nutrients increases in the stratified surface waters can trigger pulsed production and affect global biogeochemical budgets. Moreover, such violent event occurrences may rise in the future in the context of the global change.

These results pave the way for future oceanic cruises, and in particular for the BioSWOT-Med cruise in 2023. This cruise is planned as part of the "Adopt a Cross Over" initiative organising simultaneous oceanographic cruises around the world during the fast sampling phase of the new satellite SWOT (Surface Water and Ocean Topography) (d'Ovidio et al., 2019), that will allow the precise observation of fine-scale ocean dynamics. The aim is to study the fine-scale features and their influence on biology, with methodology supporting offshore, multi-instrumental, multi-technique, multi-scale, and multi-disciplinary observations.

Code availability. TEXT

Data availability. TEXT

Data are available here:
https://dataset.osupytheas.fr/geonetwork/srv/eng/catalog.search#/metadata/5bda8ab8-79e7-4dec-9bcb-25a3196e2f9a

Code and data availability. TEXT

Sample availability. TEXT

Video supplement. TEXT

Appendix A: Testing the mean differences of the phytoplankton groups in different water types

The significance of the differences in means of each phytoplankton group between water types was tested using two-tailed tests based on the Moving Blocks Bootstrap principle (Liu et al., 1992). Using a bootstrap-based test avoids relying on the gaussian distribution assumption, which was in our case violated in nearly all samples. Instead, the stationarity of the samples originating from each water mass was assumed. Sampling the observations by block of adjacent observations enables to preserve the serial auto-correlation existing in the sample. The size of the blocks is in practice left to the practitioner and values in [1,4] were tested and did not influence the results. The number of bootstrap samples used to perform the tests was 3000 draws. The level of the tests was 5% with a Bonferroni correction (Dunn, 1961) in order to account for multiple hypotheses testing.

Author contributions. TEXT

Jean-Luc Fuda (JLF) prepared the instruments prior to the cruise, and deployed them onboard, together with Stéphanie Barrillon (SB), Andrea Doglioli (SD), Gérald Grégori (GG), Melilotus Thyssen (MT), and Roxane Tzortzis (RT). Anne Petrenko operated SPASSO and analysed the water masses, Caroline Comby analysed the current data. GG and MT prepared and operated the flow cytometer, MT and Robin Fuchs analysed and interpreted the flow cytometry data. Nagib Bhairy prepared and piloted the glider, Frédéric Cyr performed the first treatment of the glider data and Anthony Bosse (AB) analysed the glider data. Christophe Yohia performed the model data, and analysed the model data together with AB. AD and Léo Berline analysed the MVP data. Francesco d'Ovidio and AD initiated the project. SB designed the experiment, lead the research and prepared the manuscript with contributions from all co-authors. All authors participated to the manuscript.

Competing interests. TEXT

The authors declare that they have no conflict of interest.

Disclaimer. TEXT

Acknowledgements. We thank the captain and the crew of the RV *Téthys II* for the cruise and their help for the deployement of instruments. All of this research was supported by CNES (BioSWOT project) and by the French National program LEFE (Les Enveloppes Fluides et l'Environnement), FUMSECK-vv project. The flow cytometer was funded by the CHROME (PI M. Thyssen, funded by the Excellence Initiative of Aix-Marseille University – A*MIDEX, a French "Investissements d'Avenir" program), and the FEDER fundings (PRECYM flow cytometry platform). SPASSO is operated with the support of the SIP (Service Informatique de Pythéas) and in particular C. Yohia, J. Lecubin. D. Zevaco and C. Blanpain (Institut Pythéas, Marseille, France)

References

Aminot, A. and Kérouel, R.: Dosage automatique des nutriments dans les eaux marines: méthodes en flux continu, Editions Quae, 2007.

Andersen, V. and Prieur, L.: One-month study in the open NW Mediterranean Sea (DYNAPROC experiment, May 1995): overview of the hydrobiogeochemical structures and effects of wind events, DEEP-SEA RESEARCH PART I-OCEANOGRAPHIC RESEARCH PAPERS, 47, 397–422, https://doi.org/10.1016/S0967-0637(99)00096-5, 2000.

Anglès, S., Jordi, A., and Campbell, L.: Responses of the coastal phytoplankton community to tropical cyclones revealed by high-frequency imaging flow cytometry, Limnology and Oceanography, 60, 1562–1576, 2015.

Babin, S., Carton, J. A., Dickey, T. D., and Wiggert, J. D.: Satellite evidence of hurricane-induced phytoplankton blooms in an oceanic desert, Journal of Geophysical Research, 109, 2004.

Barrillon, S., Bataille, H., Bhairy, N., Comby, C., Coulon, T., Doglioli, A., d'Ovidio, F., Fuda, J.-L., Grégori, G., Petrenko, A., et al.: FUMSECK cruise report, https://archimer.ifremer.fr/doc/00636/74854/, 2020.

Behrenfeld, M. J.: Abandoning Sverdrup's critical depth hypothesis on phytoplankton blooms, Ecology, 91, 977–989, 2010.

Bonato, S., Christaki, U., Lefebvre, A., Lizon, F., Thyssen, M., and Artigas, L. F.: High spatial variability of phytoplankton assessed by flow cytometry, in a dynamic productive coastal area, in spring: The eastern English Channel, Estuar. Coast. Shelf S., 154, 214–223, 2015.

Calvo-Díaz, A., Morán, X. A. G., and Suárez, L. A.: Seasonality of picophytoplankton chlorophyll a and biomass in the central Cantabrian Sea, southern Bay of Biscay, J. Mar. Syst., 72, 271–281, 2008.

Conan, P., Testor, P., Estournel, C., D'Ortenzio, F., Pujo-Pay, M., and Durrieu de Madron, X.: Preface to the Special Section: Dense Water Formations in the Northwestern Mediterranean: From the Physical Forcings to the Biogeochemical Consequences, Journal of Geophysical Research: Oceans, 123, 6983–6995, https://doi.org/https://doi-org.insu.bib.cnrs.fr/10.1029/2018JC014301, 2018.

Doglioli, A. and Rousselet, L.: Users guide for latextools, 2013.

Doglioli, A., Nencioli, F., Petrenko, A., Fuda, J.-L., Rougier, G., and Grima, N.: A software package and hardware tools for in situ experiments in a Lagrangian reference frame, J. Atmos. Ocean. Tech., pp. 1945–1950, https://doi.org/10.1175/JTECH-D-12-00183.1, 2013.

d'Ortenzio, F. and Ribera d'Alcalà, M.: On the trophic regimes of the Mediterranean Sea: a satellite analysis, Biogeosciences, 6, 139–148, 2009.

d'Ovidio, F., Penna, A. D., Trull, T. W., Nencioli, F., Pujol, I., Rio, M. H., Park, Y.-H., Cotté, C., Zhou, M., and Blain, S.: The biogeochemical structuring role of horizontal stirring: Lagrangian perspectives on iron delivery downstream of the Kerguelen plateau, Biogeosciences Discuss., pp. 779–814, 2015.

d'Ovidio, F., Pascual, A., Wang, J., Doglioli, A., Jing, Z., Moreau, S., Gregori, G., Swart, S., Speich, S., Cyr, F., Légresy, B., Chao, Y., Fu, L., and Morrow, R.: Frontiers in fine scale in-situ studies: opportunities during the SWOT fast sampling phase, Front. Mar. Sci., p. 168, https://doi.org/10.3389/fmars.2019.00168, 2019.

Dugenne, M., Thyssen, M., Nerini, D., Mante, C., Poggiale, J.-C., Garcia, N., Garcia, F., and Grégori, G. J.: Consequence of a sudden wind event on the dynamics of a coastal phytoplankton community: an insight into specific population growth rates using a single cell high frequency approach, Front. Microbiol., 5, 485, 2014.

Dunn, O. J.: Multiple comparisons among means, J. Am. Stat. Assoc., 56, 52–64, 1961.

Esposito, A. and Manzella, G.: Current circulation in the Ligurian Sea, in: Elsev. Oceanogr. Serie, vol. 34, pp. 187–203, Elsevier, 1982.

Ferrari, R. and Wunsch, C.: Ocean circulation kinetic energy: Reservoirs, sources, and sinks, Annu. Rev. Fluid Mech., 41, 253–282, https://doi.org/10.1146/annurev.fluid.40.111406.102139, 2009.

Fuchs, R., Rossi, V., Chloé, C., Nathaniel, B., Pinazo, C., Grosso, O., and Thyssen, M.: Intermittent upwelling triggers delayed, yet major and reproducible, pico-nanophytoplankton responses in oligotrophic waters, (in prep.), 2022.

Giordani, H., Prieur, L., and Caniaux, G.: Advanced insights into sources of vertical velocity in the ocean, Ocean Dynam., 56, 513–524, https://doi.org/10.1007/s10236-005-0050-1, 2006.

Han, G., Ma, Z., and Chen, N.: Hurricane Igor impacts on the stratification and phytoplankton bloom over the Grand Banks, Journal of Marine Systems, 100, 19–25, 2012.

Houpert, L., Testor, P., De Madron, X. D., Somot, S., D'ortenzio, F., Estournel, C., and Lavigne, H.: Seasonal cycle of the mixed layer, the seasonal thermocline and the upper-ocean heat storage rate in the Mediterranean Sea derived from observations, Prog. Oceanogr., 132, 333–352, 2015.

Jakobsen, H. H. and Markager, S.: Carbon-to-chlorophyll ratio for phytoplankton in temperate coastal waters: Seasonal patterns and relationship to nutrients, Limnol. Oceanogr., 61, 1853–1868, https://doi.org/https://doi.org/10.1002/lno.10338, 2016.

Le Bot, P., Kermabon, C., Lherminier, P., and Gaillard, F.: CASCADE V6. 1: Logiciel de validation et de visualisation des mesures ADCP de coque, 2011.

Liu, R. Y., Singh, K., et al.: Moving blocks jackknife and bootstrap capture weak dependence, Exploring the limits of bootstrap, 225, 248, 1992.

Lomas, M., Roberts, N., Lipschultz, F., Krause, J., Nelson, D., and Bates, N.: Biogeochemical responses to late-winter storms in the Sargasso Sea. IV. Rapid succession of major phytoplankton groups, Deep Sea Research Part I: Oceanographic Research Papers, 56, 892–908, 2009.

Louchart, A., Lizon, F., Lefebvre, A., Didry, M., Schmitt, F. G., and Artigas, L. F.: Phytoplankton distribution from Western to Central English Channel, revealed by automated flow cytometry during the summer-fall transition, Cont. Shelf Res., 195, 104 056, 2020.

Marrec, P., Grégori, G., Doglioli, A. M., Dugenne, M., Della Penna, A., Bhairy, N., Cariou, T., Hélias Nunige, S., Lahbib, S., Rougier, G., Wagener, T., and Thyssen, M.: Coupling physics and biogeochemistry thanks to high-resolution observations of the phytoplankton community structure in the northwestern Mediterranean Sea, Biogeosciences, 15, 1579–1606, https://doi.org/10.5194/bg-15-1579-2018, 2018.

Mayot, N., d'Ortenzio, F., Ribera d'Alcalà, M., Lavigne, H., and Claustre, H.: Interannual variability of the Mediterranean trophic regimes from ocean color satellites, Biogeosciences, 13, 1901–1917, 2016.

McWilliams, J. C.: A survey of submesoscale currents, Geoscience Letters, 6, 1–15, 2019.

Menden-Deuer, S. and Lessard, E. J.: Carbon to volume relationships for dinoflagellates, diatoms, and other protist plankton, Limnol. Oceanogr., 45, 569–579, 2000.

Menkes, C. E., Lengaigne, M., Lévy, M., Éthé, C., Bopp, L., Aumont, O., Vincent, E., Vialard, J., and Jullien, S.: Global impact of tropical cyclones on primary production, Global Biogeochemical Cycles, 30, 767–786, 2016.

Millot, C.: Circulation in the western Mediterranean Sea, Journal of Marine Systems, 20, 423–442, 1999.

Nencioli, F., d'Ovidio, F., Doglioli, A., and Petrenko, A.: Surface coastal circulation patterns by in-situ detection of Lagrangian Coherent Structures, Geophys. Res. Lett., https://doi.org/10.1029/2011GL048815, 2011.

Pasqueron De Fommervault, O., Migon, C., D'Ortenzio, F., Ribera D 'alcala, M., and Coppola, L.: Temporal variability of nutrient concentrations in the northwestern Mediterranean sea (DYFAMED time-series station), Deep Sea Research Part I: Oceanographic Research Papers, 100, 1–12, https://doi.org/10.1016/j.dsr.2015.02.006, 2015.

Petrenko, A., Doglioli, A., Nencioli, F., Kersalé, M., Hu, Z., and d'Ovidio, F.: A review of the LATEX project: mesoscale to submesoscale processes in a coastal environment, Ocean Dynam., https://doi.org/10.1007/s10236-017-1040-9, 2017.

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Liu, Z., Berner, J., Wang, W., Powers, J. G., Duda, M. G., Barker, D. M., et al.: A description of the advanced research WRF model version 4, National Center for Atmospheric Research: Boulder, CO, USA, p. 145, 2019.

Testor, P., de Young, B., Rudnick, D. L., Glenn, S., Hayes, D., Lee, C. M., Pattiaratchi, C., Hill, K., Heslop, E., Turpin, V., et al.: OceanGliders: a component of the integrated GOOS, Frontiers in Marine Science, 6, 422, 2019.

Thyssen, M., Grégori, G. J., Grisoni, J.-M., Pedrotti, M. L., Mousseau, L., Artigas, L. F., Marro, S., Garcia, N., Passafiume, O., and Denis, M. J.: 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, Front. Microbiol., 5, 387, https://doi.org/10.3389/fmicb.2014.00387, 2014.

Verity, P. G., Robertson, C. Y., Tronzo, C. R., Andrews, M. G., Nelson, J. R., and Sieracki, M. E.: Relationships between cell volume and the carbon and nitrogen content of marine photosynthetic nanoplankton, Limnol. Oceanogr., 37, 1434–1446, 1992.

Welschmeyer, N. A.: Fluorometric analysis of chlorophyll a in the presence of chlorophyll b and pheopigments, Limnol. Oceanogr., 39, 1985–1992, 1994.

Xing, X., Claustre, H., Blain, S., d'Ortenzio, F., Antoine, D., Ras, J., and Guinet, C.: Quenching correction for in vivo chlorophyll fluorescence acquired by autonomous platforms: A case study with instrumented elephant seals in the Kerguelen region (Southern Ocean), Limnol. Oceanogr.-Meth., 10, 483–495, 2012.

2. Automating the flow cytometry gating process with convolutional neural networks

The results presented during the FUMSECK cruise highlighted the potentially intense response of phytoplankton functional groups to wind-induced events. During the FUMSECK cruise, the assignation of the cells was performed manually, which is the most commonly used procedure to process FC data. Yet, this procedure is time-consuming and error-prone. As a result, a method based on convolutional neural networks was introduced to automate the manual gating process and provide a full characterization of the effect of several wind events on the cPFGs.

2.1. Designing convolutional networks to deal with Flow Cytometry pulse shapes

The manual gating process and most supervised learning models (see Section 2.2 for a review of these models) use the listmode data format which summarizes the pulse shapes using simple descriptors (e.g., mean, variance, area under the curve). Conversely, the Convolutional Neural Network (CNN) proposed here, does not rely on these simple descriptors and deals directly with the pulse shapes to predict six phytoplankton classes: Redpicopro, Orgpicopro, Redpicoeuk, Rednano, Orgnano, and Micro cells, and two noise classes: particles $\leq 1\mu m$ or particles $\geq 1\mu m$. The CNN automatically determines the best features to extract from the signal to perform the classification.

CNNs are supervised neural networks contrary to the unsupervised neural networks presented in Section 1.1 such as the SOM or ART models. The general architecture of a supervised neural network is given in Figure 3.1. It represents a Feed-Forward Neural Network (FFNN), also called Multi-Layer Perceptron (MLP) here designed to perform classification.

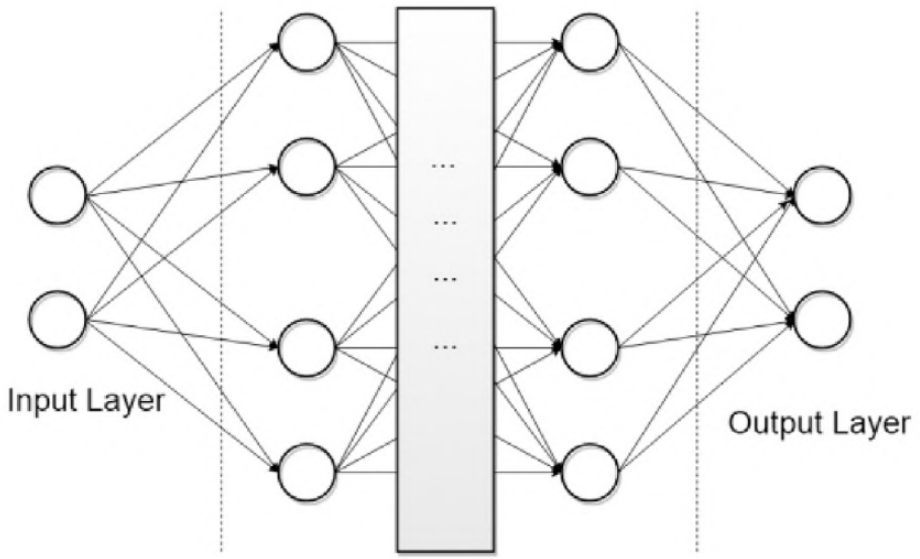


Figure 3.1. – Graphical representation of a Feed Forward Neural Network (FFNN) (under Wikimedia Commons licence).

Supervised neural networks for classification: a general presentation

In a neural network, the data are given as input in the so-called "Input Layer", go through "hidden layers" that capture the patterns contained in the data and the prediction itself is performed in the output layer. When the number of hidden layers is high the associated learning process is referred to as Deep Learning. In the case of a network designed for classification, the number of neurons of the output layer is equal to the number of classes K to predict for the variable of interest y ($K = 2$ in Figure 3.1, $K = 8$ in the phytoplankton case). Alternatively, a network designed for a regression task has only one output neuron. The sequel will deal only with classification networks but most of the introduced notions can be applied to regression tasks (and to unsupervised neural networks).

At each hidden layer, the data outputted by the previous hidden layer, X^l for layer l undergo the following transformation:

$$X^{l+1} = f(WX^l + b), \tag{3.1}$$

with $l \in [1, L]$, $l = 0$ the input layer index, $y = X^L$ by construction, and W and b a matrix and a vector of weights, respectively, and f an function called activation function. The hidden layers described in 3.1 are called "dense layers" as every neuron of a layer is linked with every neuron of the next layer. W and b are the trainable weights of the network that capture the information contained in the data, and the activation function acts as a filter keeping only the interesting pieces of information. Popular choices of activation functions include Rectified Linear Unit (Relu), hyperbolic tangent

(tanh), sigmoid, and softmax functions, with the following expressions:

$$\begin{aligned} Relu(x) &= \max(0,x), \text{ with image in } [0,+\infty[\\ tanh(x) &= \frac{e^x - e^{-x}}{e^x + e^{-x}}, \text{ with image in }]-1,1[\\ sigmoid(x) &= \frac{1}{1+e^{-x}}, \text{ with image in }]0,1[\\ softmax(x) &= \frac{e^{x_k}}{\sum_{k'}^{K} e^{x'_k}}, \text{ with image in }]0,1[^K, \end{aligned} \tag{3.2}$$

The first two activation functions are mainly used in the hidden layers. Conversely, the sigmoid and the softmax are mostly used in the output layer in the classification framework to compute the probability that an observation belongs to each class. The quality of the prediction can be asserted by defining a distance between the actual label to predict and the network prediction. This distance is computed and aggregated for all the observations and takes the form of a cost function to minimize. The most usual loss function for the classification task is the categorical cross-entropy or negative log-likelihood (negLL). Its expression is given by :

$$negLL = -\sum_{k=1}^{K}\sum_{i=1}^{n}(y_{i,k} * log(\hat{p}_{i,k})), \tag{3.3}$$

with $i \in [1,n]$ the observation index, $k \in [1,K]$ the class index, $y_{i,k}$ equals 1 if observation i is in class k, and $\hat{p}_{i,k}$ is the probability that observation i belongs to class k according to the model. The loss given in 3.3 gives the same weights to all errors whatever the actual class of the observations. Conversely, different weights could be assigned to some observations or classes using a weighted version of the categorical cross-entropy.

Remark 1 *Note that the similarity between equation 3.1 and the expressions in (2.2) in the DGMM case. The DGMM case uses the identity function as activation function f. Yet, the MDGMM specifies an error term assumed null in equation 3.1. Besides, the 3.1 relationship relates the next layer to the previous layer in the FFNN case, whereas it relates the previous layer to the next layer in the DGMM case.*

From equations 3.1, 3.2 and 3.3, the loss function can be written as a function of the trainable weights and the activation functions. This function is derivable nearly everywhere on its support (depending on the activation function used). Hence, the loss is derived with respect to the weights, and the error gradients are propagated from the output layer to the input layer to update the weights: the training is thus said to occur by "back-propagation" of the error. The errors are generally computed and summed on a set of observations, called a "batch" of observations, to iteratively update the weights rather than updating the weights using the whole dataset or on a single observation basis. Training the model with batches of observations acts as

a regularization force, i.e. pushes the model to learn general and reusable features, without focusing on unnecessary and dataset-specific patterns, a general pitfall of machine learning models called overfitting. Contrary to most machine learning methods, several passes over the dataset are performed, i.e. the observations are given more than a unique time to neural networks. A complete pass over all the observation batches of a dataset is called an epoch. The number of epochs typically ranges from more than a dozen to several thousand as networks are not able to extract all the pieces of information contained in the dataset in one unique pass to converge towards optimal weights.

The strategy to update the weights with respect to the errors is ruled by the optimizer and the update pace by the learning rate of the optimizer. Popular choices of optimizers are based on Stochastic Gradient Descent (SGD) such as RMSProp (Rezende et al. 2014) or Adam (Kingma et al. 2014). As the batch size, the learning rate of the optimizer can be seen as a regularization force. Other regularization processes exist such as the dropout rate and the batch normalization methods. The principle of the dropout rate (Srivastava et al. 2014) is to leave a share of the neurons untrained during the training phase. The untrained neurons are randomly selected at each epoch, which slows down the learning but forces the learning to be more general and robust to the noise. Alternatively, a method called Batch normalization (BN) has been developed by Ioffe et al. 2015 and is now often preferred to dropout. During training, the weights of the hidden layers might be perturbed by differences in statistical distributions existing between batches of observations or by the way the weights of the previous layers have been initialized. To mitigate this effect, called "internal covariate shift", BN proposes to normalize the mean and variance of the signal going out from a layer before passing it to the next layer. Doing so, BN is supposed to reduce overfitting and notably permit the use of a higher learning rate.

To summarize, the choice of the network architecture (number of neurons and layers), the activation functions, the loss, the optimizer, and the regularization methods are the main quantities, or hyper-parameters, to tune in a supervised neural network. Until now, only dense neural networks, made of dense layers, have been presented. Dense networks are particularly suited for tabular data, but less suited for signals presenting more dimensions such as images or pulse shape data in our case. Convolutional neural networks generalize the dense networks and are more adapted to this end.

Convolutional neural networks

The idea of Convolutional Neural Networks (CNN) is to reduce the number of parameters to train in the network by replacing the first dense layers with convolutional layers. In computer vision tasks, the goal is to extract information from the image pixels. The number of pixels in a black and white image is equal to the width times the height of the image. This number is multiplied by three for a colored image, the red,

green, and blue pixels being stored in three different channels. Hence, the number of pixels can reach more than 10^6 pixels per image, and connecting each of these pixels to a neuron in a dense network would lead to a far too high number of parameters. Furthermore, two consecutive pixels in an image are likely to contain very similar pieces of information and this information redundancy unnecessarily increases the computational cost. Convolutional layers, instead of connecting every pixel to a neuron, compute convolutions using a sliding window on the image, which significantly reduces the number of parameters to learn. A representation of this process is given in Figure 1.5.

These sliding windows can be viewed as filters that capture and recognize shape patterns in the images (e.g. straight and curved lines, object borders, differences in contrast in different areas of the pictures, etc.) as presented in Figure 3.2. While the signal goes through the network, the filters of the deepest convolutional layers focus on more precise patterns than the shallowest layers. Convolutional layers are hence regarded as feature extractors: they design the most relevant filters with respect to the task to perform (ex: classification, regression). The features learned are generally then passed to dense layers that perform the classification/regression itself. Some networks only contain convolutional layers and no dense layers and are thus called "fully convolutional neural networks". The feature extraction process of convolutional layers can also be regularized using pooling layers that summarize parts of the signals by computing local means or maxima, reducing the number of parameters and signal complexity. The corresponding pooling layers are called average-pooling max-pooling layers, respectively.

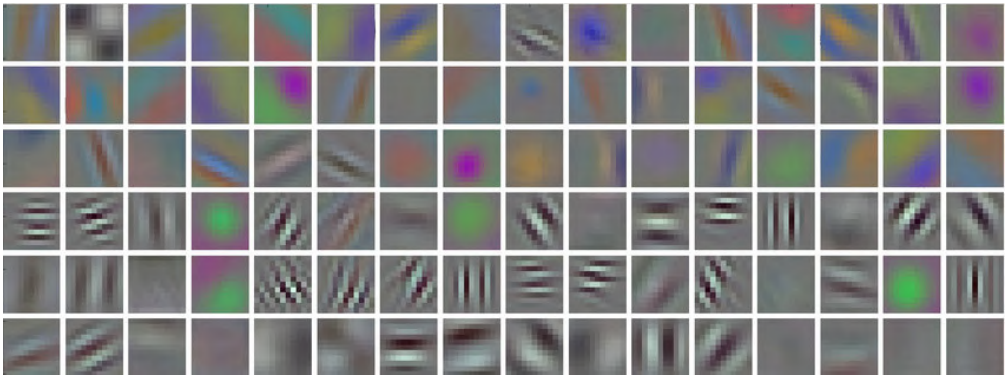


Figure 3.2. – Example of filters learnt by the first convolutional layer in A. Brachmann, and C. Redies (2016).

Convolutional networks were first introduced by LeCun et al. 1989 under the LeNet name and popularized by Krizhevsky et al. 2012 with the Alexnet network. LeNet had two convolutional layers separated by average pooling layers, and ended with three dense layers. The activation function used was a sigmoid. LeNet was tested for the classification of ten handwritten digits from the well-known MNIST dataset (Deng

2012), stored as grey images (one channel) of 28 × 28 pixels. The Alexnet architecture was inspired by LeNet but treated images in color (composed of three channels). Compared to the LeNet architecture, an additional block of three stacked convolutional layers ended by a pooling layer has been added, which makes it much deeper. The pooling layers were max-pooling layers and no more average-pooling layers and the sigmoid activation function from LeNet was replaced by the Relu activation function. In a nutshell, AlexNet has set the real basis of modern CNN architectures by standardizing the usage of the Relu activations and Max-pooling layers but necessitated much more computational power delivered by GPU computing. Most of the recent architectures are based on these principles, such as the VGG architectures that deepened the AlexNet architecture, and used thirteen convolutional and three dense layers in the VGG-16 architecture.

2.2. Creating a fully automated recognition procedure

Inspired by the VGG architecture, we propose to use a CNN to deal with the pulse shapes issued by FC. The five pulse shapes per cell were interpolated to a fixed length and stacked together as a matrix. Doing so, we used Relu activation functions, average-pooling layers, dropout, and a refinement of the Adam optimizer to perform the phytoplankton functional group classification. The training data came from the SSL@MM station and SWINGS cruise (see Figure 1.6). The data were obtained by making FC experts classify several acquisitions and by keeping only consensual particles. This expert classification made it possible to assess the inter-expert gating heterogeneity, which as mentioned earlier, was rarely available in the literature. The following article addresses these points and more information could be found in Appendix D.

Automatic recognition of flow cytometric phytoplankton functional groups using Convolutional Neural Networks

Robin Fuchs[a,b], Melilotus Thyssen[b,1], Véronique Creach[c], Mathilde Dugenne[d], Lloyd Izard[e], Marie Latimier[f], Arnaud Louchart[g,h], Pierre Marrec[i], Machteld Rijkeboer[j], Gérald Grégori[b], Denys Pommeret[a,k,l,m]

[a] Aix Marseille Univ, CNRS, Centrale Marseille, I2M, Marseille, France; [b] Aix Marseille Univ, Université de Toulon, CNRS, IRD, MIO, Marseille, France; [c] Cefas, Pakefield Road, NR33 0HT Lowestoft, Suffolk, UK; [d] Department of Oceanography, University of Hawai'i at Mānoa, Honolulu, Hawai'i, USA; [e] Sorbonne Université, CNRS, IRD, MNHN, Laboratoire d'Océanographie et du Climat : Expérimentations et Approches Numériques (LOCEAN-IPSL), Paris, France; [f] IFREMER, DYNECO PELAGOS, F-29280 Plouzane, France; [g] Department of Integrative Marine Ecology, Stazione Zoologica Anton Dohrn, Villa Comunale, 80121, Napoli, Italy; [h] IFREMER, Laboratoire Environnement & Ressources, F-62321 Boulogne sur mer, France; [i] Graduate School of Oceanography, University of Rhode Island, Narragansett, Rhode Island, USA; [j] Laboratory for Hydrobiological Analysis, Rijkswaterstaat (RWS), Zuiderwagenplein 2, 8224 AD Lelystad, The Netherlands, [k] Université Claude Bernard Lyon 1, 43 boulevard du 11 Novembre 1918 69622 Villeurbanne cedex, France; [l] ISFA, 50 Avenue Tony Garnier, 69007 Lyon, France; [m] Laboratoire de Sciences Actuarielle et Financière (SAF) EA2429, Lyon France.

[1] Corresponding author: melilotus.thyssen@mio.osupytheas.fr

Abstract

The variability of phytoplankton distribution has been unraveled by high-frequency measurements. Such a resolution can be approached by automated pulse-shape recording flow cytometry (AFCM) operating at hourly sampling resolution. AFCM records morphological and physiological traits as single-cell optical pulse shapes that can be used to classify cells into Phytoplankton Functional Groups (PFG). However, the associated manual post-processing of the data coupled with the increasing size and number of datasets is time-consuming and error-prone. Machine learning models are increasingly used to run automatic classification. Yet, most of the existing methods either present a long training process, need to manually design features from the raw optical pulse shapes, or are dedicated to images only. In this study, we present a Convolutional Neural Network (CNN) to classify several PFGs using AFCM pulse shapes. The uncertainties of manual classification were first estimated by comparing experts' recognition of six PFGs. Consensual particles from the manual PFG classification were used to train and validate the CNN. The CNN obtained competitive performances compared to other models used in the literature and remained robust across several sampling areas, and instrumental hardware and settings. Finally, we assessed the ability of this classifier to predict phytoplankton counts at a Mediterranean coastal station and from a cruise in the South-West Indian Ocean, providing a comparison with the manual classification over three-month periods and a two-hour frequency. These promising results strengthen the near real-time observation of PFGs, especially required with the increasing use of AFCM in monitoring research programs.

Keywords— phytoplankton | pulse-shape recording flow cytometry | automatic classification | deep learning | high frequency

Introduction

Phytoplankton cells are major actors in marine environments and in biogeochemical cycles. The amount of seawater dissolved CO_2 absorbed by phytoplankton cells per unit of time, called autotrophic carbon fixation, is estimated to be equivalent to all of the primary terrestrial production. This is the case even if they represent less than 1% of the total autotrophic biomass (Field et al. 1998), suggesting a rapid growth capacity and high turnover rates (Fowler et al. 2020). Currently, models estimating primary production in the ocean present a wide uncertainty range (Carr et al. 2006; Saba et al. 2011; Buitenhuis et al. 2012), mainly due to the coarse resolution of the datasets collected (Lévy et al. 2012). Indeed, the heterogeneous distributions of phytoplankton combined with a high structural and functional diversity highlight the need for infra kilometer spatial resolution and infra hour temporal resolution (Kavanaugh et al. 2016).

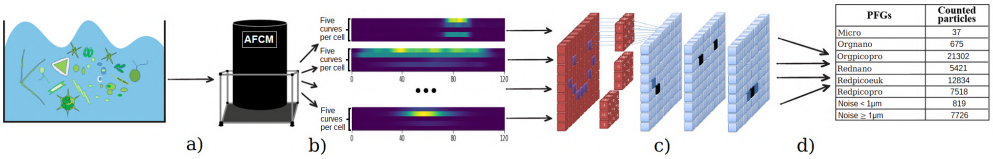
Phytoplankton functional diversity, biomass, and distribution are listed as Essential Ocean Variables (EOV) (Miloslavich et al. 2018), but datasets with resolutions inferior to 10 km are scarce. Automated pulse-shape recording flow cytometry (AFCM) such as the CytoSense instrument (Cytobuoy, b.v., (Dubelaar et al. 1999; Dubelaar and Gerritzen 2000)) enables vast automated data acquisition with hourly sampling strategies of several phytoplankton groups at a single-cell level resolution. AFCM is now involved in numerous oceanographic field studies and benefits from the growing scientific interest for automated single-cell approaches (Boss et al. 2020) in monitoring programs.

The CytoSense AFCMs generate a set of pulse shapes or flow cytometric curves (FCCs) which represent the optical pro-

files of scatter and fluorescences emitted by each particle (detritus, cell, or colony) when crossing a laser beam. Scatter signals collected at small and large angles (forward scatter (FWS) and sideward scatter (SWS) respectively) are related to the particle size and structure (granularity), while red (FLR) and yellow-orange fluorescence (FLY or FLO) signals are reflecting pigment contents of the photosynthetic cells (such as chlorophyll a or phycoerythrin). From the difference between left-angled and right-angled FWS pulses, a fifth signal named Curvature is extracted. Instruments can process up to 10 000 particles per second thanks to a frequency acquisition of 4 MHz, with sampled volume up to 5 mL routinely.

Groups recognition and identification are based on seminal papers (Olson et al. 1985; Chisholm et al. 1988; Green et al. 1996; Jacquet et al. 2002; Metfies et al. 2010; Ribeiro et al. 2016; Hamilton et al. 2017; van den Engh et al. 2017; Marrec et al. 2018) describing the most common groups observed by flow cytometry in natural seawater. In addition to these groups of pico-nanophytoplankton, AFCM resolves microphytoplankton size classes with a coarse taxonomic level identification (typically up to the genus) using recent integration of image-in-flow devices (Dugenne et al. 2014). A dedicated vocabulary, relying on these papers, has been recently suggested by a wide group of flow cytometry experts (`http://vocab.nerc.ac.uk/collection/F02/current/`). These size and pigment-related groups belong to several phytoplankton functional groups (PFG), since they fit the initial definition of sets of species sharing similar ecological and biogeochemical functionalities (Le Quere et al. 2005), and will hereafter be identified as cytometric PFG (cPFG).

Raw data recorded by AFCM has to



PFGs	Counted particles
Micro	37
Orgnano	675
Orgpicopro	21302
Rednano	5421
Redpicoeuk	12834
Redpicopro	7518
Noise < 1µm	819
Noise ≥ 1µm	7726

Figure 1: Explanatory scheme of the predictive pipeline. (a) Particles are sampled from seawater by AFCM. (b) The five flow cytometric curves (FCCs = SWS, FWS, FLR, FLO, Curvature) generated for each particle as they cross a laser beam are interpolated to a fixed length and stacked together into matrices. (c) The CNN predicts the class of each particle using Convolutional layers (red) and Dense layers (blue). (d) The number of particles per group (phytoplankton or background noise) is computed and returned.

be manually processed. This processing, called manual gating of cPFG, is performed on 2D projections of reduced statistics of the FCCs (such as pulse maximum height, area under the curve, pulse width). The long periods of assiduity required, coupled with experts' diversity of practices and the significant differences in cPFG abundances can be substantial sources of errors. Furthermore, the spread of the AFCM technology generates datasets too numerous to be manually processed, constraining the collection of valuable high-frequency cPFGs datasets. In order to facilitate the work of an increasing number of AFCM users and decrease the uncertainties linked to manual gating, the classification of cPFGs has to be semi- or fully automated. The automation can be achieved using supervised machine learning methods that assign a label to an observation based on its characteristics, a task named classification.

In the case of phytoplankton, automatic classification generally relies on image processing and computer vision. One can for example cite the count of coccoliths using shallow Neural Networks (Beaufort and Dollfus 2004) or more re-

cent works based on Residual Neural Networks and transfer learning (Yosinski et al. 2014) in order to classify images from diverse laboratory cultures and in situ monitoring (Dunker 2019; González et al. 2019). However, cameras resolution is relatively low for the identification of pico-nanophytoplankton size classes, which show limited morphological diversity. As such, using the FCCs offers an alternative since it deals also with these small particles that can represent up to 90% of the total phytoplankton biomass (Li et al. 1983; Detmer and Bathmann 1997; Ribeiro et al. 2016). A second main advantage in working on the automatic classification of optical profiles is the shorter training process due to the absence of transfer learning (Pan and Yang 2009) required to fine-tune heavy Neural Networks like Residual Networks (He et al. 2016) for image recognition.

Automatic recognition of cPFGs from the FCCs has received less attention than image-based identification and can be gathered in two main types of approaches. The first family of approaches applies machine learning methods on a set of reduced statistics derived from the FCCs. Boddy et al. (1994) started to use neural methods to classify cells at the species level. Wacquet et al. (2013) developed original statistical methods and implemented them along with existing statistical methods in the R package RclusTool. Thomas et al. (2018) and Schmidt et al. (2020) used Random Forests to respectively discriminate between phytoplankton cells of different populations and between phytoplankton and non-phytoplankton particles. Abdelaal et al. (2019) used Linear Discriminant Analysis (LDA) and present performances outperforming Deep Learning approaches.

The second family of approaches, to which this study belongs, relies on the

entire FCC signal to perform classification. For example, Malkassian et al. (2011) plunged the FCCs into a Fourier basis and calculated distances to discriminate between populations. Del Barrio et al. (2019) created curves templates to classify AFCM non-marine cells using Wasserstein distance and optimal transport. Finally, Caillault et al. (2009) relied on the Elastic Matching coupled with standard classifiers. While these two families of approaches attempt to classify cPFGs in an objective and reproducible manner, they all present unique advantages and trade-offs. A comparison of all these approaches has yet to be reported.

In this article, we provide a comparison of expert manual classifications of cPFGs detected by AFCM. We used the consensual particles to develop, for the first time, a CNN trained on pulse shapes recorded by AFCM as described in Figure 1. We compared the performance of our CNN, along with other automatic approaches, and tested its robustness across two instruments and multiple study areas. Finally, the CNN was used to generate predictions spanning three months in a coastal station of the Mediterranean Sea and two months in the South-West Indian Ocean, both at a two-hour sampling frequency. The robustness and extremely fast process of the applied CNN open the way to near real-time cPFG analysis.

Material and procedures

Data origin and collection

Two datasets collected using different approaches were used in this study. The first one, referred to as SSLAMM data, was acquired in different Mediterranean areas using the same flow cytometer and settings: at a coastal marine Mediterranean station (the SSLAMM, SeaWater Sensing Laboratory At MIO Marseille,

France), between September 2019 and December 2019 and in an open Mediterranean sea area, during the FUMSECK cruise (DOI 10.17600/18001155) in the Gulf of Genoa from April 30, 2019, to May 05, 2019. The second dataset, named hereafter SWINGS data, originated from the South-West Indian Ocean and the Southern Ocean and was collected onboard the R/V Marion Dufresne II, from January 11 to March 8, 2021, in the frame of the MAP-IO project (Marion Dufresne Atmospheric Program - Indian Ocean, University of la Reunion) during the GEOTRACES SWINGS cruise (South-West Indian Geotraces Section, DOI 10.13155/83989, SWINGS data). Two distinct CytoSense flow cytometers (Cytobuoy b.v.), hereafter identified as SSLAMM-AFCM, and MAP-IO-AFCM were deployed. A map indicating the location of the different sampling areas is given in Figure 1 in Supplemental Information.

For both datasets, seawater was continuously pumped in situ and the flow cytometers ran automated acquisitions scheduled every two hours. The SSLAMM coastal seawater was gently pumped with a VerderFlex40 peristaltic pump at 10 meters away from the coast at a depth of 3 meters, and was delivered unaltered into the laboratory where analyses were conducted. The FUMSECK data were collected onboard the R/V le Tethys II from the underway clean seawater supply pumped at 2 m depth. Onboard the Marion Dufresne II, the seawater was collected from the underway clean seawater supply pumped at 7 m depth, using a centrifugal pump.

The two automated CytoSense flow cytometers (Cytobuoy b.v.) were operated similarly in the three conditions. They pumped samples from a dedicated external chamber of 300 ml. The volume an-

alyzed for each sample was estimated using a calibrated peristaltic pump. Before entering the flow cell, the sample was surrounded by a 0.1 μm filtered seawater sheath fluid and the generated laminar flow aligned each particle before crossing a 488 nm laser beam (Coherent, 120 mW). Both instruments recorded the optical pulse shapes emitted resulting in forward scatter (FWS), sideward scatter (SWS), and two fluorescences. The SSLAMM-AFCM collected wavebands of > 652 nm (red fluorescence, FLR) and between $552 - 652$ nm (orange fluorescence, FLO). The MAP-IO-AFCM collected wavebands between $668 - 726nm$ (FLR) and $516 - 650nm$ (yellow fluorescence, FLY). Particles were recorded in the size range $< 1 - 800$ μm in width and up to a few mm in length for chain-forming cells.

Laser scattering at frontal angles (FWS) was collected by two distinct photodiodes to check for the sample core alignment. The difference between left and right photodiodes signatures generated the Curvature curve. SWS, FLR, and FLY were collected with photomultiplier tubes. To follow the stability of the flow cytometers, 2.0 μm fluorescing polystyrene beads (Polyscience ®) were regularly analyzed. Silica beads (1.01 μm, 2.56 μm, 3.13 μm, 5.02 μm, 7.27 μm in diameter, Bangs Laboratory®) were also used to calibrate FWS into particle size.

Because of the current memory and computational limitations, optimally sampling the entire size range of the phytoplankton community in natural marine waters requires some compromises. To collect small cells, the AFCM settings were set on high sensitivity: the red fluorescence trigger threshold was set on 6 mV (FLR6) for SSLAMM-AFCM and on 5 mV (FLR5) for MAP-IO-AFCM.

As a result, the sample was filled by a majority of small and/or dimly fluorescent particles and electrical background noise, hereafter simply called noise. Since the smallest phytoplankton cells are the most abundant in natural samples, they were counted in volumes between 0.5 ml and 1 ml.

To collect the largest but less concentrated cells, a second protocol was applied with a red fluorescence trigger threshold (high trigger level) set up to 25 mV (FLR25) for SSLAMM-AFCM, and to 20 mV (FLR20) for MAP-IO-AFCM and a volume analyzed reaching 5 ml. With this setting, the small particles and background noise generating acquisition limitations were not recorded. Except for their use of two different thresholds, the two protocols (FLR5/FLR6 and FLR20/FLR25) used the same AFCM settings (same sample pump speed, similar filter mesh sizes, same optical chamber, similar sampling frequency, similar gains).

Flow cytometry groups nomenclature

A set of six phytoplankton functional groups determined by their optical properties were selected in this study. They were identified and labeled using the flow cytometry consensual nomenclature (`http://vocab.nerc.ac.uk/collection/F02/current/`): Redpicopro, Orgpicopro, Redpicoeuk, Rednano, Orgnano, Redmicro, Orgmicro. A correspondence table between this new nomenclature and previous denominations observed in the literature is given in Supplemental Information in Table 1. There were not enough Redmicro and Orgmicro cells in situ to distinguish between these two groups and they will be gathered together in the sequel under the name “Micro” cells. The HSnano,

Redredpico, Redrednano, and Orgpico groups defined in the nomenclature were not abundant enough to be resolved or not found in our case.

In addition to these six phytoplankton functional groups, the datasets contained non-phytoplankton particles thereafter called noise particles or events. Noise events were heterogeneous and have been subdivided into $< 1\ \mu m$ and $\geq 1\ \mu m$ groups using silica beads as a size reference (Figure 2 in Supplemental Information). $\geq 1\ \mu m$ noise mainly contained large detrital particles or predators such as ciliates or flagellates cells that have ingested some phytoplankton cells. Conversely, $< 1\ \mu m$ noise often contained optical noise from the sensors, non-fluorescing heterotrophic prokaryotes, or decaying cells.

The total number of Orgpicopro and Redpicopro cells was obtained from the FLR5/FLR6 files and the total number of Orgnano, Redpicoeuk, Rednano, and Micro cells was obtained from the corresponding FLR20/FLR25 files.

Manual gating methodology and heterogeneity estimation

The raw data collected by the AFCM are composed of a series of five curves exhibiting variable heights, areas, and lengths. Experts use a dedicated software, CytoClus4©, and single values for each curve, typically the area under the curve or the maximal value of the curve, to perform their gating. With the summary statistics, experts obtain a point of dimension five for each observation and the dataset can be represented by a series of 2D projections. For example, experts commonly project the Total FLR (the area under the FLR curve) against the Total FLO or FLY (the area under the FLO or FLY curve) to separate Orgpicopro and Orgnano from red only fluorescing parti-

cles. Total FLR vs Total FWS are commonly used to separate Redpicoeuk, Rednano and Micro size classes, while Total FLR vs Total SWS (or Maximal height of SWS) can help in gating the Redpicopro group. The manual gating procedure is illustrated in Figure 3 in Supplemental Information.

The heterogeneity amongst six AFCM manual classifications was assessed on multiple SSLAMM and SWINGS acquisitions (6 and 20 respectively), spanning multiple seasons, study areas, and times of the day. The list of the cPFGs was given, along with two acquisitions of 2.0 μm polystyrene (Polyscience ®) and 3.13 μm silica beads (Bangs Laboratory ®).

The heterogeneity was measured by computing the Adjusted Rand Indices (ARIs) Steinley (2004) on the experts' overall classification and the coefficients of variation (CVs) of each cPFG count. The ARIs indicate the similarity between two experts' overall classifications. The closest the ARI is to 1, the more similar the classifications between two experts are. The ARIs have been computed for all pairs of experts and all files.

Additionally, the coefficient of variation of each cPFG is computed as the standard error divided by the mean of the expert counts for that cPFG. The closest it is to zero, the more the experts agreed on the count of the given cPFG. As a result, the ARIs assessed the overall agreement between experts' classifications whereas the CVs summarized the similarities of manual classifications at the cPFG level.

Beyond the initial training samples, one of the experts has manually gated three months of data from the SSLAMM station (from mid-September 2019 to mid-December 2019) and the entire dataset from the MAP-IO-SWINGS cruise. The classification obtained from

the CNN was then compared with the manual gating.

Data processing for automatic classification

Only the consensual particles, defined as particles for which 2/3 of the experts assigned the same label were kept to train and evaluate statistical models.

Due to the acquisition limitations of the two cytometers and because they present dim fluorescence in surface waters, the Redpicopro are hard to distinguish from $< 1\ \mu m$ noise events and a curve shape criterion was used to distinguish between them. Indeed, Redpicopro cells are likely to be spherical cells, and their SWS signals are expected to look like bell curves, whereas $< 1\ \mu m$ noise events can present a significant variety of shapes. Therefore among the consensual Redpicopro cells, only the bell-curved SWS cells were kept to train and validate the models.

The consensual particles were split into three sets: the training set, the validation set, and the test set. The training set is used by the models to learn how to distinguish between cPFGs, the validation to compare several specifications of a given model, and the test set to compare the best specifications of different models. In order to reach a substantial total dataset size and to reduce the imbalance between groups that affect the training process, the over-represented groups were undersampled in the training set.

Yet, as Figure 2 highlights it, the density of points is not uniform in 2D cytograms. Pure random particles sampling tends to let some of the low-density areas of 2D cytograms nearly empty, preventing machine learning models to learn which class to predict for particles in these areas. Hence, additional particles were sampled to fill low-density areas in

the limit of 5% of the dataset size. The impact of these zones on the confidence of the CNN cPFG predictions can for instance be seen on Figure 4 in Supplemental Information.

Before undersampling, the number of particles of the most represented group in the training set was 130 times higher than the less represented one. After undersampling, it was only 8 times higher at most for the two datasets.

Conversely, the validation set was undersampled in a stratified manner, i.e. non-rebalanced. Finally, the test set was constituted of three genuine files to give the best representation possible of in situ conditions at different seasons and times of the day. The total size of the training, validation, and test sets were 33 791, 50 682, and 134 313 particles for the SSLAMM data, and 57 241, 365 863, and 224 426 particles for the SWINGS data. Tables 2 and 3 in Supplemental Information describe the number of particles of each group in the training, validation, and test sets.

The length of each AFCM curve is closely linked to the size of the particle (the bigger the particle the longer the sequence). The size distribution of the FCCs suggested that 75% of our observations were recorded with 120 or fewer values.

In order to train the CNN, which needs a fixed data format for all observations, the curves have been all interpolated to the fixed length of 120 values using quadratic interpolation (see Figure 5 in Supplemental Information for an illustration). The choice of the third quartile was motivated by the fact that, intuitively, less information is destroyed when small curves are interpolated to be bigger than the reverse. Besides, as the curves were not truncated and the profile shapes were preserved, the choice of this length is not

expected to be of prime importance regarding the performance of the model.

Convolutional Neural Network specification

The core of the predictive pipeline is a Convolutional Neural Network initially designed for image recognition. The general idea of such a network is to learn a series of filters that detect some patterns in images and help to discriminate between the classes. More formally, these filters are tables of coefficients iteratively used to compute convolutional operations on the data going through the layers. Compared to Dense layers, the Convolutional ones rely on the assumption that regions in the images convey useful information and that close pixels often carry redundant information. As a result, the total number of parameters of the model is reduced and the training of the model is kept tractable. The Convolutional layers automatically extract features from the signal, which are then used by Dense layers at the end of the network to perform the classification itself.

As both images and AFCM data can be represented as tables of coefficients, the same Convolutional Neural Networks can be used to treat both data types with minor adjustments. The CNN architecture is presented in Supplemental Information (see Figure 6). The architecture was inspired by the VGG architecture (Simonyan and Zisserman 2014). Other architectures such as the Inception Architecture (Szegedy et al. 2015) have been implemented but brought no additional performance (result not shown). The number of observations was not sufficient to implement deeper architectures such as Residual Networks (He et al. 2016).
In our network, features are first extracted by three blocks of convolutional layers separated by "local" average pool-

ing layers to reduce the redundant parts of the signal and to automatically design features useful for the classification. These convolutional features are then pooled together using a global average pooling layer so that they can be treated by two dense layers. At the end of the dense layers, a softmax activation function computes the probabilities that an observation belongs to each class and the loss of the model is evaluated.

The loss measures the gap existing between the class probabilities outputted by the model and the actual class of the observation. This gap represents an error, back-propagated to update the parameters of the network accordingly. The negative-likelihood also called the categorical cross-entropy is the most widely used loss for single-label multivariate classification (each observation belongs to one class only) and is the one used here. More refined versions of the categorical cross-entropy such as the weighted version of the categorical cross-entropy, the Focal Loss (FL) (Lin et al. 2017), or the Focal Class-Balanced loss (FCBL) (Cui et al. 2019) have been implemented but brought no additional performances.

Beyond the choice of the loss specification, another important choice is the one of the optimizer which deals with how the network parameters are updated with respect to the loss. Ranger (Liu et al. 2019), a generalization of the widely used Adam optimizer (Kingma and Ba 2014), was here used. Ranger comes from the combination of two recent publications: RectifiedAdam (or Radam) (Liu et al. 2019) and Lookahead (Zhang et al. 2019).
In order for the optimization process not to remain stuck in very local minima, it is a common practice to slowly update the parameters of the models at the beginning of the training, when promising parameter regions are not yet identified.

This adaptation rate of the parameters with respect to the loss is called the learning rate of the model and is hence often chosen to be small in the early stages of the training process (Popel and Bojar 2018). Radam adapts the learning rate to avoid the learning rate variance to grow too substantially, which is often detrimental to the learning process according to the authors. On the other hand, Lookahead enables the network to get a better understanding of the loss topology. To do so, two sets of weights are used by Lookahead: a faster set of weights that is frequently updated to "explore" the loss surface and a slower set of weights (less frequently updated) to ensure the stability of the learning process. The faster set of weights is updated using not all the data but only a set of several observations batches to get a raw idea of the promising regions to explore. In the Ranger case, these fast weights are updated thanks to the Radam optimizer.

Comparison with other classification algorithms

The CNN has been benchmarked against other supervised models to compare the performance of individual machine learning algorithms. The benchmark models have been published in the literature: the k-Nearest Neighbors (kNN) and the Linear Discriminant Analysis (LDA) (Abdelaal et al. 2019). Tree-based methods such as Random Forest were represented by the Light Gradient Boosting Machine (LGBM) (Ke et al. 2017) which is more recent and takes advantage of gradient-boosting methods.

The data from the manual classifications comparison experiment were used for models evaluation. Once interpolated to a fixed length, the CNN was trained over the five FCCs per particle, while the benchmark models (which cannot deal

with the raw curves) were trained on the hand-designed features computed from these FCCs (commonly referred to as "Listmode features"). The choice of the features created from the signal highly influences the performances of the models and has to be considered when presenting the results. We rely on the thirteen features per curve created by default by the CytoClus4© software. The feature list is given in Supplemental Information (see section 1).

Most parts of statistical models are ruled by a set of hyper-parameters chosen by the user (e.g. number of neurons and layers, number of neighbors, learning rate, batch size). The number of possible combinations is far too high for all the combinations to be tested and then to select the best models specifications.
One popular approach relies on Bayesian Hyperoptimisation algorithms (Bergstra et al. 2013), implemented in our case in the Python libraries Hyperopt and Hyperas (Hyperopt for Keras). The idea of Hyperoptimisation methods is to consider hyperparameters as statistical random variables with a prior and to identify posterior regions that present a low loss value. Hence, some draws are taken from the prior distributions, the model is evaluated and low loss regions are identified and focused on. It avoids spending substantial computational efforts on non-promising regions of the hyper-parameters space as it is often the case using standard line search. The hyperparameters spaces used are given in section 2 in Supplemental Information.

The performances of the CNN and of benchmark models were evaluated using the standard per-class precision and recall metrics. The precision is the proportion of particles actually belonging to class k among all those identified as belonging to class k by the algorithm. The

recall is the proportion of particles effectively belonging to class k among all the particles of class k existing in the dataset. The closer both precision and recall are to 100%, the closer the classification of a model is to the "true" labels.

The Python code used to produce the results of this work is freely available as a Github repository named phyto_curves_reco with the following DOI: 10.5281/zenodo.5681642.

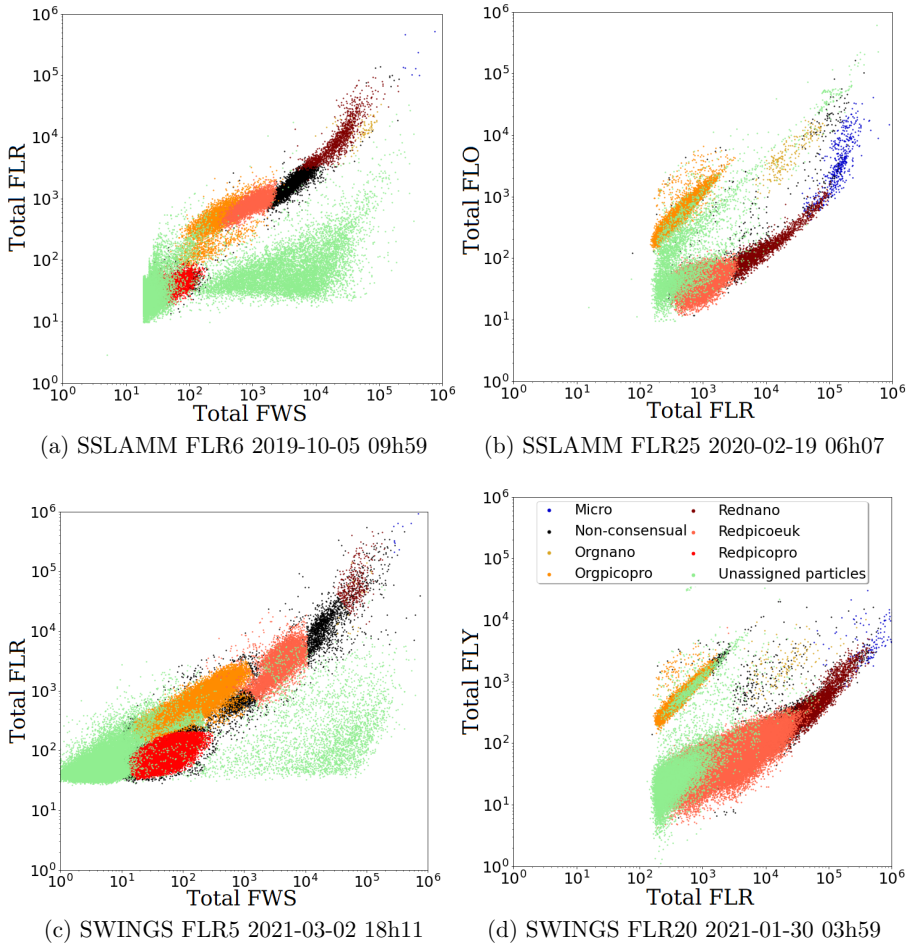
Results

Manual gating uncertainty estimation

The main groups observed by AFCM are represented on Figure 2. It presents descriptive 2D cytograms associated with two files for each data source. The non-consensual particles - on which less than 2/3 of the experts agreed - were located mainly at the frontiers between groups. The less consensual demarcation lines were between Rednano and Redpicoeuk and between Redpicopro and the background noise events.

The uncertainties of manual classification for individual cPFGs are reported in Supplemental Information (Figures 7 and 8). The patterns observed in terms of ARIs and CVs were similar between SSLAMM and SWINGS data. For both data sources, 75% of the pairwise ARIs were higher than 0.78. However, these high ARIs were driven by several over-represented cPFGs which were also well identified.

This was the case of Orgpicopro cells that obtained CVs between 0.01 and 0.14 for the SSLAMM data and between 0.02 and 0.50 for the SWINGS data and the case of Redpicoeuk (SSLAMM $CV \in [0.05, 0.44]$ and SWINGS $CV \in [0.03, 0.28]$). Conversely, Micro cells (SSLAMM $CV \in [0.27, 1.55]$ and SWINGS



(a) SSLAMM FLR6 2019-10-05 09h59

(b) SSLAMM FLR25 2020-02-19 06h07

(c) SWINGS FLR5 2021-03-02 18h11

(d) SWINGS FLR20 2021-01-30 03h59

Figure 2: 2D cytograms showing the particles contained in two files from the SSLAMM data (a and b) and two files from the SWINGS data (c and d). Cytograms (a) and (c) present the Total Red Fluorescence (a.u., Total FLR) as a function of the Total Forward Scatter (a.u., Total FWS) and cytograms (b) and (d) show the Total Orange/Yellow Fluorescence (a.u., Total FLO, Total FLY) as a function of the Total Red Fluorescence (a.u., Total FLR). Total refers to the area under the curve of the optical variable. Each dot represents a particle. A particle is considered as consensual if 2/3 of the experts have voted for the same cPFG for this particle. Non-consensual particles are represented in black.

$CV \in [0.12, 1.26]$), Orgnano (SSLAMM $CV \in [0.50, 0.85]$ and SWINGS $CV \in [0.21, 1.75]$), Rednano (SSLAMM $CV \in [0.25, 0.92]$ and SWINGS $CV \in [0.10, 1.34]$), and Redpicopro (SSLAM $CV \in [0.13, 2.45]$ and SWINGS $CV \in [0.56, 1.07]$) were far less identified (Figure 8 in Supplemental Information).

Model benchmark on the test set

Figures 3 and 4 report the precision and the recall obtained by the four models for each cPFG and noise classes.

Based on the specific precision and recall values, the CNN and the LGBM obtained the best performances on the quasi-totality of cPFGs. The kNN presented the worst performances for both datasets. The LDA results are mixed as it distinguished noise events from phytoplankton particles classified but got the worst precision on three cPFGs on the SWINGS data.

The best manually identified cPFGs were also the best classified by machine learning models i.e., Orgpicopro and Redpicoeuk cells. Similarly, the Redpicopro and Orgnano cells were weakly manually identified and less well gated by machine learning models. Finally, Micro and Rednano cells that experienced poor manual identifiability presented good precision and recall values for near all methods.

The generalization capacity of the models was tested by training them on one data source (SSLAMM or SWINGS) and by making predictions on the other data source. Results are given in Figures 9 and 10 in Supplemental Information.

When the models were trained on the SWINGS data and used to predict SSLAMM data, the CNN obtained the best performances, with precisions higher than 90% for five out of the eight classes and

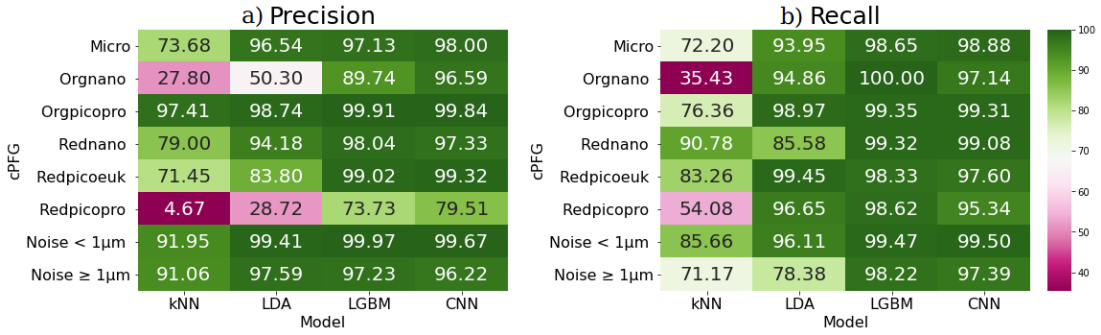


Figure 3: Precision (a) and recall (b) (%) of the benchmarked models on SSLAMM data

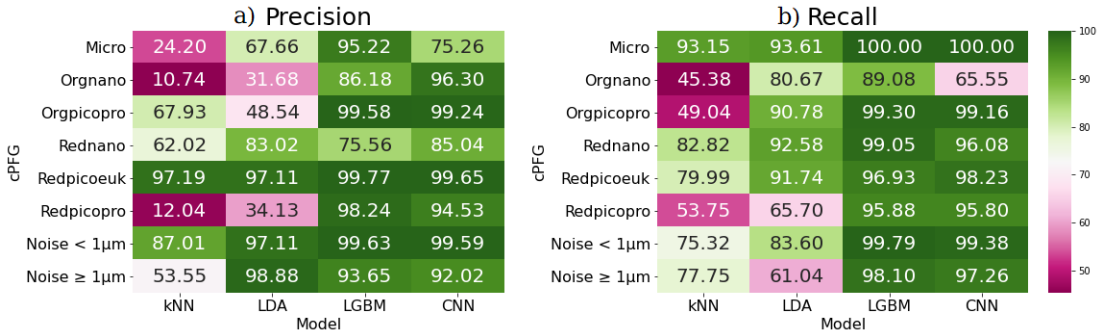


Figure 4: Precision (a) and recall (b) (%) of the benchmarked models on SWINGS data

kNN the worst performances. Concerning the cPFGs, noise events and Orgpicopro were the best classified, and Redpicopro and Micro cells were the less well gated.

When trained on the SSLAMM data and used to predict SWINGS data, the LGBM obtained the best performances and LDA the worst. Redpicopro cells and noise events $\geq 1\mu m$ were the worst identified by the models. Rednano cells obtained precisions lower than 34% but recall values higher than 87%. The opposite

pattern was observed for the Redpicoeuk class, denoting that a significant number of manually identified Redpicoeuk cells were predicted as Rednano cells by the models.

The running time of the models is given in Supplemental Information (Table 4).

Automatic classification on the full datasets

Figure 5 presents the regression between the automatically and manually counted cPFGs particles from the SSLAMM files and the SWINGS files.

The R^2 and the slope coefficients on Figure 5 are close to 1.0 for the majority of the cPFGs of both data sources: The counts resulting from the manual and CNN gatings are in adequation. The main exceptions are the Micro and Rednano cells from the SSLAMM data and the Redpicopro cells from the SWINGS data. In the SSLAMM data, Micro cells were scarce (less than 300 cells per file) which made the identification of this population difficult. The CNN counted twice as many Micro cells as the manual expert, but the counts seemed to be proportional ($R^2 = 0.84$). Concerning the Rednano cells, the R^2 of 0.61 is partly explained by a different Redpicoeuk/Rednano frontier between the CNN and the expert. This is confirmed by the 0.84 slope coefficients of the SSLAMM Redpicoeuk cells: the largest manually gated Redpicoeuk cells were regarded as Rednano cells by the CNN. The automatic Redpicopro count from SWINGS data presented a strong correlation with the manual count ($R^2 = 0.91$). However, the CNN was more conservative and considered some of the manually gated Redpicopro cells as $noise < 1\mu m$ cells. Finally, the R^2 for the noise particles was equal to 1.0 for both data sources (data not shown). The CNN and the manual expert hence dis-

criminated similarly between phytoplankton and non-phytoplankton particles (the counts only differed by 2.5%).

The CNN average prediction time for each file of the series was of 66 seconds (seven seconds for the prediction itself and more than a minute for the pre-processing steps). We ran the pipeline on two machines in parallel and the total prediction time was of 15 CPU usage hours for the 1639 files of the SSLAMM time series and 10 hours for the 1184 files of the SWINGS time series.

Discussion

The use of automated sensors is often mandatory to get resolutive datasets, common in the field of physical oceanography, but still limited in marine microbial ecology. Microbial populations in marine environments are influenced by physics, chemistry, and biological interactions that shape their distribution. Yet, they also have internal clocks and specific physiological-morphological characteristics that affect their fitness and require sensors integrating biodiversity and dynamic processes (Dutkiewicz et al. 2020). Flow cytometry measurements of phytoplankton cell abundances and single-cell morphological traits have already provided numerous insights into their interaction with environmental factors (Ribalet et al. 2015; Hyun et al. 2020), such as physical conditions (Partensky et al. 1999; Marrec et al. 2018; Louchart et al. 2020) and trophic network interactions (Christaki et al. 2011). The collected morphological traits have also enabled hourly growth rates and primary production assessments per phytoplankton group (Sosik et al. 2003; Hunter-Cevera et al. 2014; Dugenne et al. 2014).

Although AFCM is a powerful tool for the study of phytoplankton functional

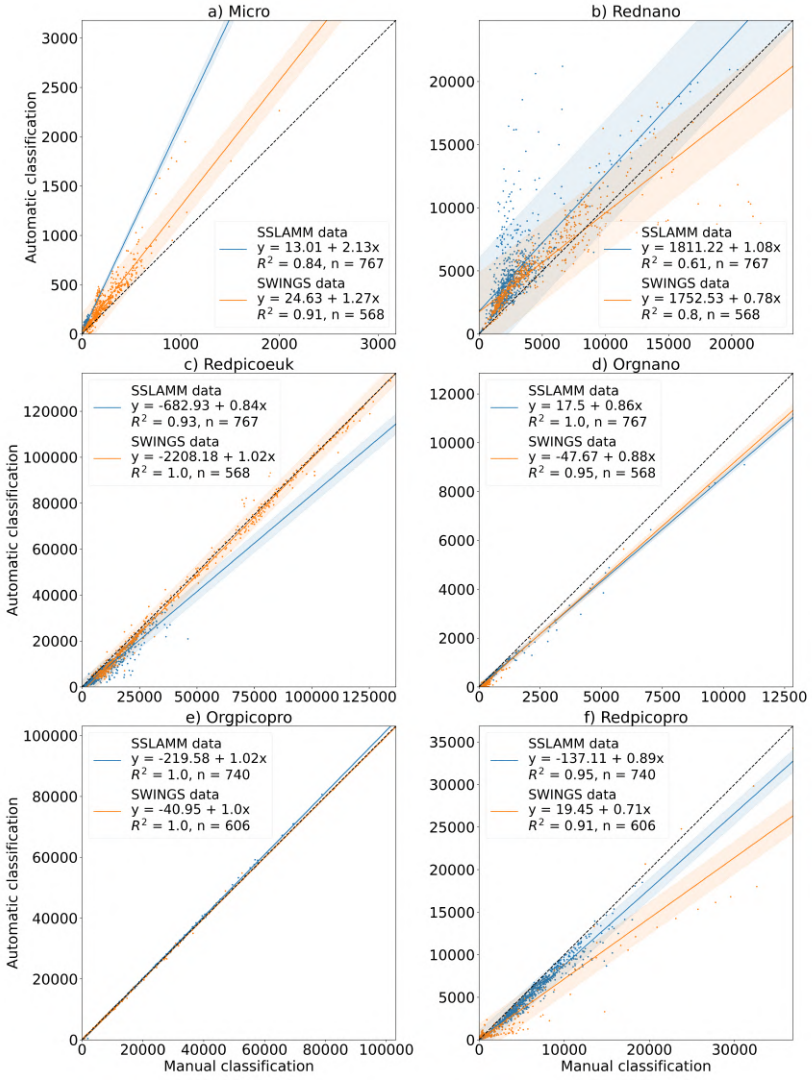


Figure 5: Automatic classification count (number of particles) as a function of the manual gating count (number of particles) for each cPFG: the Micro (a), the Rednano (b), the Redpicoeuk (c), the Orgnano (d), the Orgpicopro (e), the Redpicopro (f). Blue dots are for SSLAMM data, Orange dots are for SWINGS data. For each cPFG a linear regression has been fitted and the corresponding regression coefficients and R^2 are reported. The resulting 95% confidence intervals are illustrated as light orange and blue bands. The black dashed line indicates a 1:1 ratio between the manual and automatic classifications. 24

groups and benefits from recent technological advances, AFCM data post-processing is often performed manually. Yet, this post-processing (also named manual gating) is prone to subjectivity, and assessments of the heterogeneity between experts classifications are rarely performed in flow cytometric studies. Garcia et al. (2014) evidenced up to 20% variability between two experts on two groups of bacterioplankton. In the present study, a consensus between six experts from different laboratories was evaluated on six cPFGs and noise events. The overall classification methodology was shared by the experts as confirmed by the high pairwise Adjusted Rand Indices. On the contrary, the uncertainties existing in the exact manual gates frontiers coupled with the under-representation of several cPFGs led to significant differences in cPFG counts.

The most abundant cPFGs, Orgpicopro and Redpicoeuk, were identified by all experts with small error margins. This can be attributed to the high number of cells, combined with the very characteristic orange fluorescence of Orgpicopro particles. On the contrary, there was a lack of consensus concerning the boundaries between Redpicoeuk and Rednano, with counts variations of more than 100% between experts for Rednano cells. The origin of this discrepancy came from the non-consensual criteria used to differentiate these groups using 2D projections. Some experts used the 3.13 μm silica beads provided to them for the experiment, while other experts used a threshold between the 2 and 3.13 μm beads. The choice of a criterion to distinguish Redpicoeuk from Rednano is an issue already reported in Buitenhuis et al. (2012). In addition, the observation of Redpicopro cells by AFCM has been enabled only recently thanks to advances

in filtration of the sheath fluid or more powerful lasers Marrec et al. (2018). Yet, these particles still remain close to the flow cytometer detection limits and Red-picopro cells were hardly distinguished from the noise $< 1\mu m$ by the experts. Finally, the differences in cPFG relative abundances made the manual classification of rare cPFGs equivocal and entailed divergences in Micro, Rednano and Orgnano counts.

As such, the intercomparison highlighted the necessity of consensual rules and criteria to distinguish groups and the need for peer-reviewed data to obtain reliable cPFG observations for automation purposes. Such multi-reviewed datasets are increasing in popularity in the machine learning community, the best example being the ImageNet repository (Fei-Fei 2010).

Despite the heterogeneity in manual gating, a robust and reliable dataset has been built by keeping the particles that were consensual between experts. Using the consensual observations, three statistical models were trained and their performances compared with the ones of the Convolutional Neural Network presented here.

On the SSLAMM and SWINGS test sets, the CNN model proposed in this study achieved precision and recall values competitive with the ones of the LGBM and higher than the ones of the kNN and the LDA. It exhibited performances higher than 90% in a vast majority of cases. When compared to a manual expert gating the CNN has evidenced its reliability to track the cPFG abundance in near real-time in two very different contexts. The small discrepancies between manual and automatic classifications can be considered marginal when compared to the length and the high temporal and functional diversity resolution

of the predicted time series. Furthermore, the CNN exhibited significant generalization properties when trained on the SWINGS data and used for prediction on the SSLAMM data. When trained on the SSLAMM data to predict SWINGS data, the generalization power of the CNN was still solid but lower. This may be due to the lower diversity and number of observations of SSLAMM data, where pico-nanophytoplankton cells dominated all over the year, compared to the SWINGS data collected in very contrasted areas of the South-West Indian and Southern oceans, the latter being considered as dominated by nano-microphytoplankton cells (Rembauville et al. 2017).

More generally, the training sets used in this study are of moderate sizes ($\sim 10^4$ observations compared to $\sim 10^6$ observations generally encountered in CNN image classification as in Simonyan and Zisserman (2014)). Yet, deep learning methods seem to take a bigger advantage of dataset sizes than traditional machine learning methods (Ng 2017), at least when the dataset size grows from a moderate to substantial size (several millions of observations) (Sun et al. 2017; Neyshabur et al. 2017; Hestness et al. 2017). Thus, the current increase in AFCM dataset sizes and dataset number should give an additional edge to the CNN over the LGBM which currently present comparable performances.

In summary, this preliminary and highly promising work applies a CNN on interpolated raw pulse shapes acquired on an hourly basis by pulse-shape recording flow cytometry. It opens the way to the integration of cPFGs into forecasting biogeochemical models, depending on near real-time data inputs. High-frequency sampling of phytoplankton and determination of the communities structure and abundances will permit a better

integration of pulsed events and responses capacities of some functional groups in these models. It will also enable to adjust near real-time spatial sampling strategies where influences of physical structures such as fronts and eddies directly affect the distribution of phytoplankton groups (d'Ovidio et al. 2019).

Acknowledgments

We thank Cytobuoy b.v. for their assistance to design special CytoClus4© features mandatory to conduct this work. We thank Olivier Grosso and Michel Durand for technical assistance at the SeaWater Sensing Laboratory At MIO Marseille (SSLAMM), and the support of the MIO Service Atmosphere Mer (Deny Malengros and Fabrice Garcia) and UMS OSU Pytheas divers, Laurent VanBostal, Christian Marshal, and Dorian Guillemain for installing and maintaining the pumping inlet. Supports for the SSLAMM were provided by Aix Marseille Université, MIO, and OSU PYTHEAS. We thank Stéphanie Barrillon and the participants of the FUMSECK cruise, and the captain and crew of the R/V Tethys II. MAP-IO is a scientific program led by the LACy/La Réunion University and was funded by the European Union through the ERDF program,

the University of Reunion, the SGAR-Réunion, the région Réunion, the CNRS, the TAAF, the IFREMER and the Flotte Océanographique Française. The authors thank the technical team of the LACy engaged in the data acquisition and the maintenance of the instruments of the MAP-IO program. We are also very thankful to Maiwenn Hascoët, Gaëtan Viardot, and Chloé Caille for their participation in the manual gating process and in the data post-processing. Funding of R.F. PhD thesis was provided by the Ministry of Higher Education, Research and Innovation. The project leading to this publication has received funding from the ERDF under project 1166-39417. The project leading to this publication has received funding from Excellence Initiative of Aix-Marseille University - A*MIDEX, a French "Investissements d'Avenir" program.

References

Abdelaal, T., V. van Unen, T. Höllt, F. Koning, M. J. Reinders, and A. Mahfouz 2019. Predicting cell populations in single cell mass cytometry data. *Cytometry Part A 95*(7), 769–781.

Beaufort, L. and D. Dollfus 2004. Automatic recognition of coccoliths by dynamical neural networks. *Marine Micropaleontology 51*(1-2), 57–73.

Bergstra, J., D. Yamins, and D. D. Cox 2013. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. *Proceedings of the 30th International Conference on Machine Learning 28(1)*, 115–123.

Boddy, L., C. Morris, M. Wilkins, G. Tarran, and P. Burkill 1994. Neural network analysis of flow cytometric data for 40 marine phytoplankton species.

Cytometry: The Journal of the International Society for Analytical Cytology 15(4), 283–293.

Boss, E., A. M. Waite, J. Uitz, and others 2020. Recommendations for plankton measurements on the go-ship program with relevance to other sea-going expeditions. *SCOR Working Group GOSHIP Report 154*, 1–70.

Buitenhuis, E. T., W. K. Li, D. Vaulot, and others 2012. Picophytoplankton biomass distribution in the global ocean. *Earth System Science Data 4*(1), 37–46.

Caillault, É., P.-A. Hébert, and G. Wacquet 2009. Dissimilarity-based classification of multidimensional signals by conjoint elastic matching: application to phytoplanktonic species recognition. In *International Conference on Engineering Applications of Neural Networks*, pp. 153–164. Springer.

Carr, M.-E., M. A. Friedrichs, M. Schmeltz, and others 2006. A comparison of global estimates of marine primary production from ocean color. *Deep Sea Research Part II: Topical Studies in Oceanography 53*(5-7), 741–770.

Chisholm, S. W., R. J. Olson, E. R. Zettler, R. Goericke, J. B. Waterbury, and N. A. Welschmeyer 1988. A novel free-living prochlorophyte abundant in the oceanic euphotic zone. *Nature 334*(6180), 340–343.

Christaki, U., C. Courties, R. Massana, P. Catala, P. Lebaron, J. M. Gasol, and M. V. Zubkov 2011. Optimized routine flow cytometric enumeration of heterotrophic flagellates using sybr green i. *Limnology and Oceanography: Methods 9*(8), 329–339.

Cui, Y., M. Jia, T.-Y. Lin, Y. Song, and S. Belongie 2019. Class-balanced

loss based on effective number of samples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9268–9277.

Del Barrio, E., H. Inouzhe, J.-M. Loubes, C. Matrán, and A. Mayo-Íscar 2019. optimalflow: Optimal-transport approach to flow cytometry gating and population matching. arXiv preprint arXiv:1907.08006.

Detmer, A. and U. Bathmann 1997. Distribution patterns of autotrophic pico- and nanoplankton and their relative contribution to algal biomass during spring in the atlantic sector of the southern ocean. *Deep Sea Research Part II: Topical Studies in Oceanography 44*(1-2), 299–320.

Dubelaar, G. and P. Gerritzen 2000. Cytobuoy: a step forward towards using flow cytometry in operational oceanography. *Scientia Marina 64*(2), 255–265.

Dubelaar, G. B., P. L. Gerritzen, A. E. Beeker, R. R. Jonker, and K. Tangen 1999. Design and first results of cytobuoy: A wireless flow cytometer for in situ analysis of marine and fresh waters. *Cytometry: The Journal of the International Society for Analytical Cytology 37*(4), 247–254.

Dugenne, M., M. Thyssen, D. Nerini, C. Mante, J.-C. Poggiale, N. Garcia, F. Garcia, and G. J. Grégori 2014. Consequence of a sudden wind event on the dynamics of a coastal phytoplankton community: an insight into specific population growth rates using a single cell high frequency approach. *Frontiers in microbiology 5*, 485.

Dunker, S. 2019. Hidden secrets behind dots: Improved phytoplankton taxonomic resolution using high-

throughput imaging flow cytometry. *Cytometry Part A 95*(8), 854–868.

Dutkiewicz, S., P. Cermeno, O. Jahn, M. J. Follows, A. E. Hickman, D. A. Taniguchi, and B. A. Ward 2020. Dimensions of marine phytoplankton diversity. *Biogeosciences 17*(3), 609–634.

d'Ovidio, F., A. Pascual, J. Wang, and others 2019. Frontiers in fine-scale in situ studies: Opportunities during the swot fast sampling phase. *Frontiers in Marine Science 6*, 168.

Fei-Fei, L. 2010. Imagenet: crowdsourcing, benchmarking & other cool things. In *CMU VASC Seminar*, Volume 16, pp. 18–25.

Field, C. B., M. J. Behrenfeld, J. T. Randerson, and P. Falkowski 1998. Primary production of the biosphere: integrating terrestrial and oceanic components. *science 281*(5374), 237–240.

Fowler, B. L., M. G. Neubert, K. R. Hunter-Cevera, R. J. Olson, A. Shalapyonok, A. R. Solow, and H. M. Sosik 2020. Dynamics and functional diversity of the smallest phytoplankton on the northeast us shelf. *Proceedings of the National Academy of Sciences 117*(22), 12215–12221.

Garcia, F. C., A. Lopez-Urrutia, and X. A. G. Moran 2014. Automated clustering of heterotrophic bacterioplankton in flow cytometry data. *Aquatic Microbial Ecology 72*(2), 175–185.

González, P., A. Castaño, E. E. Peacock, J. Díez, J. J. Del Coz, and H. M. Sosik 2019. Automatic plankton quantification using deep features. *Journal of Plankton Research 41*(4), 449–463.

Green, J., P. Course, and G. Tarran 1996. The life-cycle of emiliania huxleyi: A brief review and a study of relative ploidy levels analysed by flow cytome-

try. *Journal of marine systems 9*(1-2), 33–44.

Hamilton, M., G. M. Hennon, R. Morales, and others 2017. Dynamics of teleaulax-like cryptophytes during the decline of a red water bloom in the columbia river estuary. *Journal of Plankton Research 39*(4), 589–599.

He, K., X. Zhang, S. Ren, and J. Sun 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Hestness, J., S. Narang, N. Ardalani, G. Diamos, H. Jun, H. Kianinejad, M. Patwary, M. Ali, Y. Yang, and Y. Zhou 2017. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*.

Hunter-Cevera, K. R., M. G. Neubert, A. R. Solow, R. J. Olson, A. Shalapyonok, and H. M. Sosik 2014. Diel size distributions reveal seasonal growth dynamics of a coastal phytoplankter. *Proceedings of the National Academy of Sciences 111*(27), 9852–9857.

Hyun, S., M. R. Cape, F. Ribalet, and J. Bien 2020. Modeling cell populations measured by flow cytometry with covariates using sparse mixture of regressions. arXiv preprint arXiv:2008.11251.

Jacquet, S., M. Heldal, D. Iglesias-Rodriguez, A. Larsen, W. Wilson, and G. Bratbak 2002. Flow cytometric analysis of an emiliana huxleyi bloom terminated by viral infection. *Aquatic Microbial Ecology 27*(2), 111–124.

Kavanaugh, M. T., M. J. Oliver, F. P. Chavez, R. M. Letelier, F. E. Muller-Karger, and S. C. Doney 2016. Seascapes as a new vernacular for pelagic ocean monitoring, management and conservation. *ICES Journal of Marine Science 73*(7), 1839–1850.

Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu 2017. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems*, pp. 3146–3154.

Kingma, D. P. and J. Ba 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Le Quere, C., S. P. Harrison, I. Colin Prentice, and others 2005. Ecosystem dynamics based on plankton functional types for global ocean biogeochemistry models. *Global Change Biology 11*(11), 2016–2040.

Lévy, M., R. Ferrari, P. J. Franks, A. P. Martin, and P. Rivière 2012. Bringing physics to life at the submesoscale. *Geophysical Research Letters 39*(14).

Li, W., D. S. Rao, W. Harrison, J. Smith, J. Cullen, B. Irwin, and T. Platt 1983. Autotrophic picoplankton in the tropical ocean. *Science 219*(4582), 292–295.

Lin, T.-Y., P. Goyal, R. Girshick, K. He, and P. Dollár 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988.

Liu, L., H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han 2019. On the variance of the adaptive learning rate and beyond. arXiv preprint arXiv:1908.03265.

Louchart, A., F. Lizon, A. Lefebvre, M. Didry, F. G. Schmitt, and L. F. Artigas 2020. Phytoplankton distribution from western to central english channel, revealed by automated flow cytometry during the summer-fall transition. *Continental Shelf Research 195*, 104056.

Malkassian, A., D. Nerini, M. A. van Dijk, M. Thyssen, C. Mante, and

G. Gregori 2011. Functional analysis and classification of phytoplankton based on data from an automated flow cytometer. *Cytometry part A 79*(4), 263–275.

Marrec, P., G. Grégori, A. M. Doglioli, and others 2018. Coupling physics and biogeochemistry thanks to high-resolution observations of the phytoplankton community structure in the northwestern mediterranean sea. HAL preprint. HAL Id: hal-01735426.

Metfies, K., C. Gescher, S. Frickenhaus, and others 2010. Contribution of the class cryptophyceae to phytoplankton structure in the german bight 1. *Journal of Phycology 46*(6), 1152–1160.

Miloslavich, P., N. J. Bax, S. E. Simmons, and others 2018. Essential ocean variables for global sustained observations of biodiversity and ecosystem changes. *Global change biology 24*(6), 2416–2433.

Neyshabur, B., S. Bhojanapalli, D. McAllester, and N. Srebro 2017. Exploring generalization in deep learning. *arXiv preprint arXiv:1706.08947*.

Ng, A. 2017. Machine learning yearning. *URL: https://www.deeplearning.ai 139*.

Olson, R., D. Vaulot, and S. Chisholm 1985. Marine phytoplankton distributions measured using shipboard flow cytometry. *Deep Sea Research Part A. Oceanographic Research Papers 32*(10), 1273–1280.

Pan, S. J. and Q. Yang 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering 22*(10), 1345–1359.

Partensky, F., J. Blanchot, and D. Vaulot 1999. Differential distribution and

ecology of prochlorococcus and synechococcus in oceanic waters: a review. *Bulletin-Institut Oceanographique Monaco Special Number 19*, 457–476.

Popel, M. and O. Bojar 2018. Training tips for the transformer model. *The Prague Bulletin of Mathematical Linguistics 110*(1), 43–70.

Rembauville, M., N. Briggs, M. Ardyna, J. Uitz, P. Catala, C. Penkerc'h, A. Poteau, H. Claustre, and S. Blain 2017. Plankton assemblage estimated with bgc-argo floats in the southern ocean: Implications for seasonal successions and particle export. *Journal of Geophysical Research: Oceans 122*(10), 8278–8292.

Ribalet, F., J. Swalwell, S. Clayton, and others 2015. Light-driven synchrony of prochlorococcus growth and mortality in the subtropical pacific gyre. *Proceedings of the National Academy of Sciences 112*(26), 8008–8012.

Ribeiro, C. G., A. L. dos Santos, D. Marie, V. H. Pellizari, F. P. Brandini, and D. Vaulot 2016. Pico and nanoplankton abundance and carbon stocks along the brazilian bight. *PeerJ 4*, e2587.

Saba, V. S., M. A. Friedrichs, D. Antoine, and others 2011. An evaluation of ocean color model estimates of marine primary productivity in coastal and pelagic regions across the globe. *Biogeosciences 8*(2), 489–503.

Schmidt, K. C., S. L. Jackrel, D. J. Smith, G. J. Dick, and V. J. Denef 2020. Genotype and host microbiome alter competitive interactions between microcystis aeruginosa and chlorella sorokiniana. *Harmful Algae 99*, 101939.

Simonyan, K. and A. Zisserman 2014. Very deep convolutional networks for

large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sosik, H. M., R. J. Olson, M. G. Neubert, A. Shalapyonok, and A. R. Solow 2003. Growth rates of coastal phytoplankton from time-series measurements with a submersible flow cytometer. *Limnology and Oceanography 48*(5), 1756–1765.

Steinley, D. 2004. Properties of the hubert-arable adjusted rand index. *Psychological methods 9*(3), 386.

Sun, C., A. Shrivastava, S. Singh, and A. Gupta 2017. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pp. 843–852.

Szegedy, C., W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9.

Thomas, M. K., S. Fontana, M. Reyes, and F. Pomati 2018. Quantifying cell densities and biovolumes of phytoplankton communities and functional groups using scanning flow cytometry, machine learning and unsupervised clustering. *PloS one 13*(5), e0196225.

van den Engh, G. J., J. K. Doggett, A. W. Thompson, M. A. Doblin, C. N. Gimpel, and D. M. Karl 2017. Dynamics of prochlorococcus and synechococcus at station aloha revealed through flow cytometry and high-resolution vertical sampling. *Frontiers in Marine Science 4*, 359.

Wacquet, G., É. P. Caillault, D. Hamad, and P.-A. Hébert 2013. Constrained spectral embedding for k-way data clustering. *Pattern Recognition Letters 34*(9), 1009–1017.

Yosinski, J., J. Clune, Y. Bengio, and H. Lipson 2014. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pp. 3320–3328.

Zhang, M., J. Lucas, J. Ba, and G. E. Hinton 2019. Lookahead optimizer: k steps forward, 1 step back. In *Advances in Neural Information Processing Systems*, pp. 9593–9604.

3. Evidencing reproducible and differentiated phytoplankton patterns with automatic recognition and change points

The FUMSECK study has evidenced a clear impact of wind-induced events on the phytoplankton functional groups in the Ligurian Sea. Yet, using manual treatment processes over a single event, one was not able to characterize reproducible response patterns of the cPFGs. After introducing the change-point and CNN methodologies, we were able to estimate this reaction at the SSL@MM station for 20 wind-driven events spanning more than two years. The associated study is presented below and its supplementary material is given in Appendix E. The data treated in this work represented more than 14 000 FC acquisitions and more than 300 Go of data automatically treated.

manuscript submitted to *Geophysical Research Letters*

Intermittent upwelling events trigger delayed, major, and reproducible pico-nanophytoplankton responses in coastal oligotrophic waters

R. Fuchs[1,2*], **V. Rossi**[2†], **C. Caille**[3‡], **N. Bensoussan**[2§], **C. Pinazo**[2¶], **O. Grosso**[2‖], **M. Thyssen**[2**]

[1]Aix Marseille Univ, CNRS, Centrale Marseille, I2M, Marseille, France
[2]Aix Marseille Univ, Université de Toulon, CNRS, IRD, MIO, Marseille, France
[3]Sorbonne Université, CNRS, LOMIC, Banyuls-sur-Mer, France

Key Points:

- Phytoplankton abundance (biomass, resp.) reactions start less than 2 days (4 days, resp.) after the upwelling onset and last 2 to 5 days.
- Except for *Synechococcus*, all group biomasses increase by 50-173% (up to +400%) during each event, then sharply decrease back to normal.
- Biomass peaks and daily rates of increase induced by the most extreme upwellings are of the same magnitude as the spring bloom ones.

*robin.fuchs@univ-amu.fr
†vincent.rossi@mio.osupytheas.fr
‡caillec@obs-banyuls.fr
§nathaniel.bensoussan@mio.osupytheas.fr
¶christel.pinazo@mio.osupytheas.fr
‖olivier.grosso@mio.osupytheas.fr
**melilotus.thyssen@mio.osupytheas.fr

Corresponding author: Melilotus Thyssen, melilotus.thyssen@mio.osupytheas.fr

Abstract

Pico-nanophytoplankton organisms are dominant in oligotrophic areas of the ocean thanks to competitive skills in nutrient-depleted waters. Their small cell size and highly adaptive growth rates make their contribution to the oceanic carbon cycle difficult to estimate. Despite the recent recognition of rapid and marked environmental shifts impact on microbial communities, the response capacities of pico-nanophytoplankton remain poorly studied. Here we address this knowledge gap in a coastal Mediterranean station influenced by intermittent wind gusts causing sporadic upwelling events. Within a few days after the wind rises, upwellings result in short-lived nutrient pulses and seawater temperature drops of up to 10°C lasting six days on average. Using a CytoSense flow cytometer continuously operating at a two-hour frequency from September 2019 to November 2021, we monitored the abundances and biomass of five phytoplankton functional groups over two complete annual cycles. Using unsupervised signal rupture-detection methods, our investigations focus on events forced by north-westerlies when the water column is stratified in late spring, summer, and early fall, corresponding to oligotrophic conditions. We show that despite their short durations, these events repeatedly trigger delayed increases in both abundances and biomasses for most pico-nanophytoplankton groups that can overpass spring bloom values. These positive biological reactions last two to five days and are immediately followed by an overall drop evidencing a clear physical driver of the biomass peaks. Not considering these submesoscale events, which are currently not reproduced by climate models, and the fast and salient biological responses they trigger may significantly bias carbon budgets.

Plain Language Summary

Short-lived north-westerlies in the Mediterranean sea replace surface coastal waters with colder and richer in nutrients deeper waters from offshore. This phenomenon, called a sporadic upwelling event, lasts only a few days after the wind stops and induces brutal environmental shifts. During summer, upwellings generate drops in surface water temperature of up to 10°C and are expected to have a significant impact on phytoplankton cells. Small phytoplankton cells are conspicuous for their fast response to environmental changes thanks to their high division rates (up to several times a day). As a result, the biological response to wind-induced upwellings has to be studied using high-frequency measurements. Using four attributes for each of the five studied phytoplankton groups, we show that the number of cells of most groups rose strongly in less than two days after the temperature drop according to remarkable repeatable patterns. Similarly, their carbon content increased after less than four days. The reactions themselves lasted up to five days before going back near to the initial level. The described phytoplankton reactions to local upwelling events can be as important as the ones observed during the spring bloom, often regarded as the most important seasonal event for phytoplankton communities.

1 Introduction

Coastal zones play a significant role in the global carbon cycle as they sustain, despite large uncertainties, up to 30% of the global oceanic primary production (Gattuso et al., 1998). Previous research suggested the importance of taking into account the diversity and variability of near-shore ecosystems, which remain poorly known and under the influences of complex physical forcing (Borges et al., 2005; Bauer et al., 2013; Wimart-Rousseau et al., 2020) that strongly shapes phytoplankton communities (Morel & André, 1991; Antoine et al., 1995; Bosc et al., 2004; Armbrecht et al., 2014). Furthermore, there is evidence of the fast response capacities of phytoplankton after environmental changes, notably considering the prominence of meso and submesoscale processes in the ocean (Lévy et al., 2012). This is especially true for the pico-nanophytoplankton cells that present

adaptive growth rates enhancing their competitive strategies (Lomas et al., 2009). The pico-nanophytoplankton size class is composed of polyphyletic unicellular photosynthetic microorganisms that dominate primary production in oligotrophic basins (Li, 1995; Grob et al., 2007) and are dominant in less oligotrophic conditions outside of the main spring and autumn bloom periods (Bolaños et al., 2020). They contribute substantially to the export of organic carbon into the deep layers mainly by aggregation or via grazing and subsequent sinking of organic materials (Richardson & Jackson, 2007; Lomas & Moran, 2011).

To assess the typical speed and frequency of community shifts that inform the capacity of pico-nanophytoplankton adaptation to abrupt changes in their environment, long-term and high-frequency sampling strategies allowing the separation of phytoplankton cells into functionally meaningful size classes are required. Martin-Platero et al. (2018) relied on a time series composed of daily samples for 93 days to show that physical forcing strongly shapes phytoplankton communities and that the observed patterns were highly dependent on the sampling frequency. Similarly, Martiny et al. (2016) have demonstrated positive significant correlations of cyanobacteria, pico and nanoeukaryotes abundances with temperature as well as nutrients using weekly samples over three years. Hunter-Cevera et al. (2020) used a 16-year long time series at an hourly frequency to highlight the seasonal cycles of *Synechococcus* abundances and proposed an explanation for *Synechococcus* blooms relying on growth rates variations. Wilkerson et al. (2006) demonstrated that wind-induced upwelling events followed by relaxation periods trigger optimal growth conditions for phytoplankton cells, depleting the upwelled nutrients and fostering a community of large phytoplanktonic cells (e.g. large diatoms), in line with Rossi et al. (2013). In more oligotrophic coastal areas, the responses of phytoplanktonic communities to short-lived enrichment events are more puzzling (Armbrecht et al., 2014) and suggest the prominence of small-sized phytoplanktonic cells. Thyssen et al. (2008) and Dugenne et al. (2014) have indeed shown important responses of pico-nanophytoplankton groups after strong north-westerlies events in the Bay of Marseille. Apart from atmospheric or riverine inputs and other classes of submesoscale frontal dynamics, sporadic wind-driven upwelling events are one major source of nutrients in the surface layers of various oligotrophic coastal areas (Millot, 1979; Bakun & Agostini, 2001; Palma & Matano, 2009; Rossi et al., 2014). While their hydrographic impacts, temperature cooling and nutrient enrichment of surface waters, are relatively well documented, little information exists on how they influence phytoplankton communities. The Bay of Marseille constitutes a natural laboratory to study the biological impacts of such events since they are common during stratified summer periods (Odic et al., 2022).

To our knowledge, all previous studies did not focus on wind events exclusively (Martiny et al., 2016; Hunter-Cevera et al., 2020), had low statistical power (Thyssen et al., 2008; Dugenne et al., 2014; Martin-Platero et al., 2018), had an insufficient temporal resolution (daily frequency for Wilkerson et al. (2006), weekly frequency in Martiny et al. (2016)) or did not fully resolve the pico-nanophytoplankton size class (Wilkerson et al., 2006; García-Reyes et al., 2014; Hunter-Cevera et al., 2020). In this study, we analyzed twenty short-lived wind-driven events occurring when the water column was stratified (late spring, summer, and early fall) allowing the detection of clear upwelling signatures in comparison to unstratified periods. The causal effect of the physical forcing was identified using a bi-hourly time series capturing the dynamics of five phytoplankton functional groups as resolved by Automated Flow Cytometry (Dubelaar & Gerritzen, 2000; Olson et al., 2003) over two complete years. The area of interest is the French Bay of Marseille, which is considered oligotrophic in stratified periods during which it is generally affected by the regional offshore bloom occurring in winter-early spring and fall seasons (d'Ortenzio & Ribera d'Alcalà, 2009). It is dominated by pico-nanophytoplankton size classes and its hydrology is strongly influenced by North-westerlies winds generating regularly short-lived upwelling events (Bensoussan et al., 2010; Pairaud et al., 2011; Fraysse et al., 2013; Lajaunie-Salla et al., 2021; Odic et al., 2022).

2 Materials and Methods

The temperature, nutrients, and phytoplankton data were collected from September 19, 2019 to November 31, 2021, at the Sea Water Sensing Laboratory @ MIO Marseille (SSL@MM), a coastal marine station in the North-West Mediterranean Sea (43°17' N, 5°22' E). Seawater was continuously pumped at 10 meters from the coastline at a depth of 3 meters and delivered into the laboratory using a VerderFlex 40 peristaltic pump. The seawater was coarsely pre-filtered by a PVC strainer (3 mm) and routed by polypropylene pipes that are cleaned monthly.

The temperature data were acquired every hour using an STPS sensor from the NKE-manufacturer presenting a temperature accuracy of 0.05°C. Nutrient samples were collected every four days on average and stored at -20°C until they were analyzed in a laboratory using a Technicon Autoanalyser® (SEAL Analytical) as in Tréguer and Le Corre (1975).

2.1 Phytoplankton Acquisition by Automated Pulse-shape Recording Flow Cytometry

Phytoplankton data were sampled every two hours using an Automated pulse-shape recording Flow Cytometer (Dubelaar et al., 1999; Dubelaar & Gerritzen, 2000) with the same protocol as in Marrec et al. (2018). We relied on the nomenclature proposed by Thyssen et al. (2021) (`http://vocab.nerc.ac.uk/collection/F02/current/`) and resolved five phytoplankton functional groups (PFGs): Redpicopro, Orgpicopro, Redpicoeuk, Rednano, and Orgnano, which were previously often referred to as *Prochlorococcus*, *Synechococcus*, picoeukaryotes, nanoeukaryotes, and cryptophytes, respectively. Microphytoplankton cells were collected but were not representative enough to be reported here: 75% of the samples presented less than 13 particles per milliliter. Each cell was assigned to a PFG by a Convolutional Neural Network (CNN) introduced in Fuchs et al. (2022).

2.2 Phytoplankton Biovolume, Biomass, and Growth Rate Estimations

Biovolume and biomass were estimated through empirical relationships (see Figure S1, sections 1.2 and 1.3 in Supplemental Information) following Marrec et al. (2018). The functional groups growth rate was estimated from the cell biovolumes using a size-structured population model introduced by Sosik et al. (2003) and adapted by Ribalet et al. (2015).

2.3 Wind-driven Upwelling Signatures

The occurrence and strength of each upwelling event were assessed based on the positive values of the Wind-driven Upwelling/Downwelling Index (WUDI) developed and extensively validated by Odic et al. (2022). The drop in temperature generated during an upwelling-favorable wind was evaluated as the difference between the measured water temperature and its low-pass filtered time series using a cut-off frequency of 15 days as in Rossi et al. (2014) and Odic et al. (2022). These temperature drops, or anomalies, were used to delimit three physical phases (Figure 2): (i) a pre-anomaly phase when the water temperature is stable and high, (ii) an anomaly phase when the temperature drops, stays cool for a few hours/days to then warm-up slowly, and (iii) a post-anomaly phase when the temperature has returned to a warmer and more stable state. These anomalies are particularly significant during the summer when the water column is stratified. A period was considered stratified when the filtered temperature was higher than the annual average temperature and conversely for unstratified periods as in Odic et al. (2022). Among the 54 events recorded over two years, only 20 events occurred during stratified periods and had temperature and flow cytometry data available. Besides, all successive events marked with negative seawater temperature anomalies separated by less than one

day were not considered in order to have for each event a minimal relaxation time. In other words, we retain here only the significant wind-driven events happening in stratified periods that are surrounded by relatively calm periods, denoted "Stratified period Wind-induced Upwelling Event", SWUE.

The spring blooms occurring in unstratified periods were used to benchmark the biomass (and abundance) increases generated by SWUEs as the spring blooms are expected to be the most productive periods (Fraysse et al., 2013). The bloom dates were determined using the threshold method (Sapiano et al., 2012; Brody et al., 2013) and the median biomass and abundance per PFG during the bloom were used as the reference benchmark level. The biomass increase imputable to the blooms was computed using the median biomass during the week preceding the bloom as a reference value.

2.4 Rupture Detection and Response Characterization

The biological response of each PFG to the SWUE was evaluated in terms of both abundances and biomasses using a statistically-based rupture detection method presented in Truong et al. (2020). This mathematically well-founded method looked for ruptures in causal time series. It is here employed to detect potential changes in the link existing between the temperature signal and each PFG abundance or biomass. The link was here assumed to be linear (Bai & Perron, 2003) and rupture detections were performed on biomasses and abundances separately. This methodology encompasses the idea that PFGs respond to a change in their environment, and delimited the start and end of the reactions for each PFG. The response of each PFG is hence composed of three phases: a pre-reaction, a reaction, and a post-reaction phase (called the relaxation phase).

Based on the identified ruptures, four key variables per PFG were used to characterize the duration and magnitude of the biological responses as presented in Figure 2 a). The reaction delay is the time taken by a PFG to react after the rise of physical forcing, i.e. between the start of the water cooling and the beginning of the PFG automatically identified reaction. The reaction duration measures the length of the reaction phase. The reaction and relaxation magnitudes are computed as the difference in medians during the pre-reaction and reaction phases and during the reaction and relaxation phases, respectively. To capture only PFGs causal responses to sporadic upwelling events, only the PFG responses for which the reactions occurred after the beginning of the anomaly phase were considered, which was the case for most events and PFGs. The number of SWUEs taken into account for each PFG is given in Figure 3.

More material and method details are given in Supplemental Information (section 1 and Figure S2).

3 Results

3.1 Seawater Temperature and Nutrients as Markers of Sporadic Upwelling Events

The annual mean temperature over the three years was 17.8°C in 2019, 17.1°C in 2020, and 17.3°C in 2021. The associated stratified periods started on May, 8 in 2020, and May, 25 in 2021 (not available in 2019), and ended on November, 13 in 2019, October, 27 in 2020, and October, 31 in 2021. The number of significant and distinct SWUEs during the stratified periods was two in 2019, ten in 2020, and eight in 2021. The median duration anomaly phase of the SWUEs was of six days and the subsequent drops in water temperature (difference between both maximal and minimal values recorded during each SWUE) varied from 0.7°C to 9.9°C, with a median value of 4.7°C (see also Odic et al. (2022)).

Nutrient concentrations and N/P ratio were higher during unstratified periods as compared to stratified periods, except for phosphate concentration (Figure S3 in Supplemental Information; Kruskal-Wallis test, p-value $\leq$ 1.0E-7 for nitrites, nitrates, and N/P ratio, p-value $\leq$ 0.05 for ammonium). In stratified periods, the nitrite concentration and N/P ratios were higher and nitrate concentration lower during SWUEs than outside the SWUEs. The concentrations of phosphate and ammonium were however comparable during and outside the SWUEs. The N/P ratio was 25.15 in the unstratified period, 17.33 during SWUEs, and 13.05 in the stratified period outside of the SWUEs. Yet, only the nitrite concentrations recorded during and outside SWUEs under stratified conditions were significantly different (Kruskal-Wallis test, p-value = 0.034). The concentrations are given in Table S1 in Supplemental Information.

3.2 Wind-induced Upwelling Events Trigger Peaks of Biomass and Abundances

All SWUEs triggered noticeable peaks of biomass for most PFGs (Figure 1 and Figure S4 in Supplemental Information). The pico-nanophytoplankton biomass was dominated in both stratification regimes by Rednano cells, followed by Orgnano, Orgpicopro, Redpicoeuk, and Redpicopro cells (Table S2 in Supplemental Information). Orgnano exceeded their median bloom biomass during one-third of the SWUEs. Similarly, more than half of the Orgpicopro and Rednano peaks went over their median bloom values. Finally, Redpicoeuk and Redpicopro biomass peak values were higher than their median bloom values in 4/5 SWUEs and all SWUEs, respectively.

In terms of abundance, the SWUEs generated peaks for most PFGs (Figure S5 in Supplemental Material). Over the whole series, the most abundant PFGs were the Orgpicopro, followed by the Redpicopro, Redpicoeuk, Rednano, and Orgnano cells (Table S3 in Supplemental Information). Near the half of Orgnano and Orgpicopro SWUE abundance peaks exceeded their median bloom abundances. Besides, more than 4/5 of SWUEs saw Rednano, Redpicoeuk and Redpicopro abundances go higher than their respective median abundances during the spring bloom.

Figure 1. Time series of (a) Wind-driven Upwelling/Downwelling Index (WUDI, $m^3.s^{-1}m^{-1}$) and temperature (C°) as well as (b, c) phytoplankton biomasses ($\mu gC.mL^{-1}$) monitored at the SSL@MM coastal station. The blue rectangles correspond to the 20 studied SWUEs. The event shown in Figure 2 is bounded by a dark blue box. The horizontal dashed colored lines correspond to the median biomasses observed during the spring bloom (except for 2019, not available) for each PFG (according to the color code).

3.3 Characterization of the Phytoplankton Response: A Single Event Illustration

The typical effect of wind-induced upwellings on temperature and pico-nanophytoplankton biomass is illustrated in Figure 2, showing differentiated responses among the PFGs. This event was fueled by three periods of intense wind forcings, or intensification periods, that generated an abrupt drop in temperature (-7.6°C) followed by the maintenance of cold waters for six days. As shown in Figure S6 in Supplemental Information, during these three sub-events, the N/P ratio rose after each wind intensification with a short delay, especially after the third one that multiplied the nitrates, nitrites, and phosphates concentration by a factor of 19, 5, and 5, respectively.

The biomass reactions of the Redpicopro, Orgpicopro, and Orgnano groups to this SWUE were quasi-instantaneous while they appeared after a short delay for the Redpicoeuk and Rednano cells (∼3 days). The biomass reaction magnitude was +42.7% for the Rednano, +123.7% for the Orgnano, +178.7% for the Redpicoeuk, +377.3% for the Redpicopro, and -82.1% for the Orgpicopro. Biomass levels decreased in the relaxation phase for all PFGs except the Orgnano.
The estimated hourly growth rates (Figure S7 in Supplemental Information) varied inversely with respect to the biomass (Figure 2) and the abundance (data not shown): when the PFG was high in biomass, its growth rate was estimated to be low and conversely.

3.4 Detailed Characterization of the Phytoplankton Response

The PFG abundances showed reaction delays ranging between 24h and 36h in median (Figure 3a). The reaction duration of the PFGs lasted between three and four days in median, with a lower Inter-Quartile Range (IQR)/median ratio than the reaction delay (Figure 3e). Concerning the reaction magnitude, the Orgnano and Orgpicopro abundances decreased while the other PFGs generally saw their abundances rising (Figure 3c). The Redpicopro and Redpicoeuk presented the largest increases in abundance. Their large IQRs were explained by some intense positive reactions for the majority of the SWUEs while only five presented moderately negative reactions for both groups. The abundance levels in the relaxation period decreased for all PFGs with median variations ranging from -28.96% to -52.85% (Figure 3g).

In terms of biomass, the Orgpicopro reacted in less than one day, the Orgnano and Redpicopro in less than two days, and Rednano and Redpicoeuk median reaction delay was three days (Figure 3b). The majority of reaction durations lasted between two and five days (Figure 3f). The signs of the reactions remained the same as for the abundance, except for the Orgnano that experienced a positive biomass reaction (Figure 3d). In the relaxation periods, the biomass levels decreased for all PFGs (-27.58% to -61.90% in median). However, positive relaxation magnitudes were observed in five SWUEs both for Orgpicopro and Rednano, explaining higher variance than for other PFGs (Figure 3h).

The estimated growth rates of the PFGs tended to slow down during the reaction phase and then increase during the relaxation phase (Figure S8 in Supplemental Information), except for the Orgpicopro. This pattern was however significant for Redpicoeuk cells only (Kruskal-Wallis test, p-value ≤ 0.01).

4 Discussion

The Bay of Marseille located in the NW Mediterranean upwelling system is a natural laboratory to explore the impact of wind-driven coastal processes on oligotrophic communities because of the unique intensities and short duration of upwelling events (Odic et al., 2022). During the stratified periods, the SWUEs had a clear signature on the seawater surface temperature. The expected signature on nutrient enrichment was less sig-

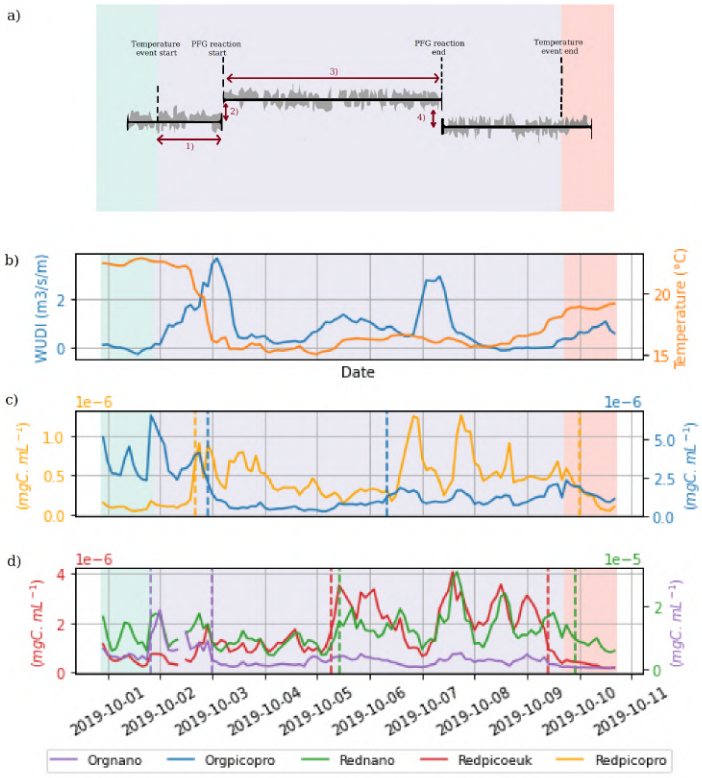


Figure 2. Illustrative view of a typical SWUE (highlighted by a dark blue box in Figure 1). a) Characterisation of the biological response to an SWUE. The grey-shaded time series represents a schematic PFG time series and the background shading corresponds to the temperature anomaly phases defining the physical event: pre-anomaly (green), anomaly (violet), and post-anomaly phase (red). The characterization is performed using four attributes: (1) the reaction delay, (2) the reaction magnitude, (3) the reaction duration, (4) and the relaxation magnitude. b) Variation of the WUDI ($m^3.s^{-1}m^{-1}$, blue line) and the temperature (°C, orange line) , c) Biomass ($mgC.mL^{-1}$) of Redpicopro and Orgpicopro d) Biomass ($mgC.mL^{-1}$) of Redpicoeuk, Rednano, and Orgnano. The vertical dashed lines represent the ruptures automatically detected by the statistical method for each PFG, according to the color code.

nificant, probably due to the littoral conditions, the delay needed for upwelled nutrients to reach the surface sampling point, but also largely to the low and irregular nutrient sampling rates (see Figure S3 in Supplemental Information).

As mentioned in García-Reyes et al. (2014), Rossi et al. (2014), and Armbrecht et al. (2014), the physically-driven temperature drops and nutrient enrichments are key indicators to characterize the impact of SWUEs over the phytoplankton community. Using a statistical rupture detection method, the causal effects of the environmental shifts over the pico-nanophytoplankton functional groups were assessed, capturing more than simple correlations and evidencing differentiated response patterns.

The phytoplankton functional groups reacted to the SWUEs in one to five days, a delay consistent with several studies evidencing phytoplankton biomass peaks two to

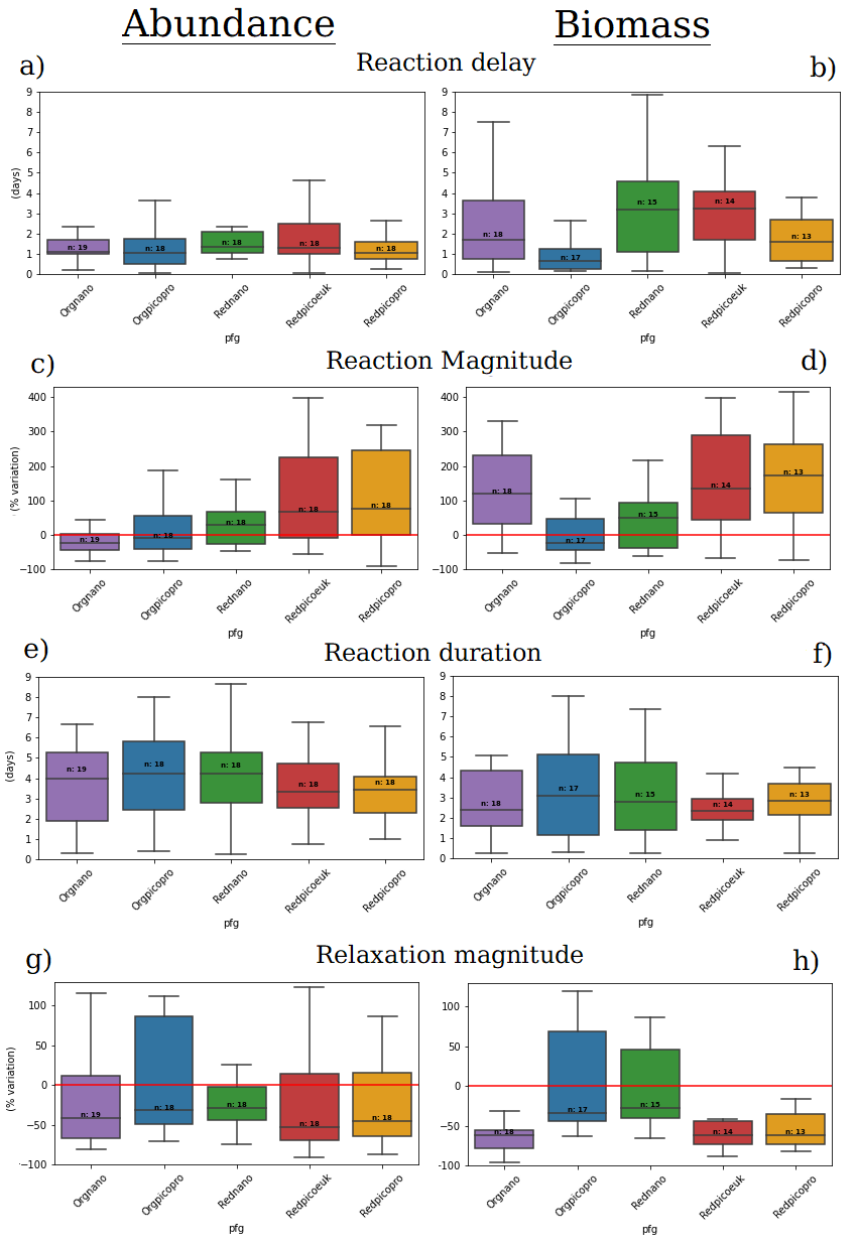


Figure 3. Boxplots of the reaction delay (a and b), the reaction magnitude (c and d), the reaction duration (e and f) and the relaxation magnitude (g and h) in terms of abundance and biomass, respectively, for five different PFGs. The horizontal red lines represent a variation of 0%. n denotes the number of SWUE for each PFG on which the boxplot has been constructed.

five days after nutrient enrichment (Edwards et al., 2005; Hauss et al., 2012; Teixeira et al., 2018). The reaction durations lasted between two and five days and were positive for all PFG abundances except for the Orgnano and Orgpicopro cells and for all PFG

biomasses except for the Orgpicopro cells. The comparison with previous studies is complicated by the different phytoplankton nomenclatures used. For instance, the increase in cyanobacteria abundance shown by Martin-Platero et al. (2018) is difficult to match with either Orgpicopro decreases or Redpicopro increases in abundance. Yet, the joint Redpicopro abundance positive reaction and increase in N/P ratio during the event is consistent with Martiny et al. (2016). Similarly, the co-occurrence of strong biological and N/P variability is in accordance with (Martz et al., 2014). The negative sign of Orgnano reaction could be compared to the curbing abundance of cluster C5 identified in Dugenne et al. (2014) after a wind event. Similarly, Thyssen et al. (2008) have shown that two groups that presented similar red fluorescence/yellow fluorescence profiles as the Orgpicopro and Orgnano groups reacted differently than the other functional groups to the SWUEs.

After the reaction, the PFGs presented mostly negative relaxation patterns except for Orgpicopro and Orgnano during some SWUEs. As presented in Figure S9 in Supplemental Material, there seems to exist an inverse relationship between these two phases for most PFG abundances and biomasses: the more positive the reaction was, the more negative the relaxation will be for a given PFG. This can be interpreted as environmental forces pushing back to the steady state. These forces remain however to be identified and could be of various nature: nutrient depletion (Wilkerson et al., 2006), competition between functional groups (Martin-Platero et al., 2018), viral lysis or predation (Sun et al., 2007; Coello-Camba et al., 2020). Following Hunter-Cevera et al. (2014), the effect of these forces can be estimated using the model loss, i.e. the difference between the observed PFG population growth rates and their estimations by the size-structured model. The authors showed that the more correlated the loss is to the growth rate, the more likely these losses are caused by biological factors. As made visible in Figure S10 in Supplemental Information, only the Rednano and Orgnano losses were significantly but weakly correlated ($r \leq 0.31$) with their growth rates in the relaxation phase. These low or non-significant correlations between growth rates and PFG losses seem to indicate that physical forces, such as water masses mixing or water column re-stratification, as well as biogeochemical hindrances (e.g. nutrient depletion or co-limitation) are dominant during this phase as compared with grazing and viral lysis.

The PFG responses have been characterized thanks to a fine temporal and functional-level resolution. As evoked in Martin-Platero et al. (2018), the chosen taxonomic level (taxa, genera, etc.) along with the temporal frequency have a strong impact on the response patterns observed. In their studies, Martin-Platero et al. (2018) have used Operational taxonomic units (OTUs) based on rRNA sequences similarity, while Martiny et al. (2016) relied on functional groups close to the ones of this study obtained using diagnostic pigments. We used automated pulse-shape recording flow cytometry to obtain an infra-day resolution over a long period and a resolution up to the cytometric functional group. Each functional group contains several ecotypes which could affect the estimated growth rates (Hunter-Cevera et al., 2014) and add uncertainty to the size-structured model. The effect of complete PFG population replacements that could occur during extremely strong SWUEs may additionally impact the presented estimations. This is also the case of the independence between predator behaviors and the phytoplankton cell sizes assumed by the model that could not be tested here. As a result, the estimated growth rates were principally used to give context to the underlying phenomena and to emphasize the fast and remarkable impacts of SWUE on phytoplankton dynamics. Future research could hence use the introduced high-frequency methodology to derive the proper impact of SWUE on phytoplankton primary production.

Similarly, while the temporal aspects of such tight biophysical coupled mechanisms are well-resolved by our sampling strategy and numerical approach, the present study did not offer a comprehensive view of the spatial variability at stake. When coupling physics with biology, the observed biological response of the PFGs could dramatically vary de-

pending on whether the water masses originated for example from areas near the Deep Chlorophyll Maximum, the nitracline, or the phosphacline. The phytoplankton biomass spatial dynamics, approached by chlorophyll-a concentration, have been extensively tracked by satellite (Wu et al., 2008; d'Ortenzio & Ribera d'Alcalà, 2009; Mayot et al., 2016; Lehahn et al., 2017; El Hourany et al., 2019). However, the satellites typically have issues resolving coastal areas and submesoscale patterns, focus on surface waters, have lower temporal resolutions (e.g. daily for sea surface temperature, weekly for clear chlorophyll-a maps) and hence could not properly resolve the phytoplankton nycthemeral cycles.

In this respect, multi-year high-frequency in situ measurements, such as the ones performed at the SSL@MM coastal laboratory, could bring crucial missing pieces of information. It could for instance be complementary to the work of Alvain et al. (2008) that matched chlorophyll-a anomalies resolved by satellite with phytoplankton community structures collected in situ. Other methods such as autonomous vehicle fleets (Jaffe et al., 2017), coastal radars (HFRs) (Cianelli et al., 2017), or 3D models coupling physics and biogeochemistry (Fraysse et al., 2013) could be used jointly with the SSL@MM data to gain further insights about spatial dynamics and help guide future modeling efforts.

In summary, the SWUEs have generated significant abundance and biomass responses from the pico-nanophytoplankton community. From our data, the biggest daily biomass increase due to a single wind-induced upwelling represented 97.6% of the daily biomass increase imputable to the spring bloom. The consistent time scales and magnitudes of biological responses reported here for sporadic wind-induced events using an innovative sampling strategy and an advanced statistical methodology could provide new insights on how to observe, and perhaps model, the impact of other submesoscale events on phytoplankton communities.

Acknowledgments
The authors are grateful to Cytobuoy b.v. for the personalized software developments performed on the CytoClus4© software features. At the SSL@MM station, the data could not have been collected without the support of Michel Durand, the MIO Service Atmosphere Mer (Deny Malengros and Fabrice Garcia), and UMS OSU Pytheas (Christian Marshal and Dorian Guillemain) that maintain the pumping inlet. Additional support for the SSL@MM was provided by Aix Marseille Université, MIO, and OSU PYTHEAS. The authors are also very thankful to Ivane Pairaud and the team exploiting the MESURHO buoy for the PAR data, and to the MIO-PACEM platform for the nutrients analysis. Funding for R.F.'s Ph.D. thesis was provided by the Ministry of Higher Education, Research, and Innovation. The project leading to this publication has received funding from the ERDF under project 1166-39417. The project leading to this publication has received funding from the Excellence Initiative of Aix-Marseille University - A*MIDEX, a French "Investissements d'Avenir" program.

Open Research

The code and data to reproduce the presented results of the paper are available at `https://github.com/RobeeF/PhytoUpwellingPaper`

References

Alvain, S., Moulin, C., Dandonneau, Y., & Loisel, H. (2008). Seasonal distribution and succession of dominant phytoplankton groups in the global ocean: A satellite view. *Global Biogeochemical Cycles*, *22*(3).

Antoine, D., Morel, A., & André, J.-M. (1995). Algal pigment distribution and primary production in the eastern mediterranean as derived from coastal zone color scanner observations. *Journal of Geophysical Research: Oceans*, *100*(C8),

16193–16209.

Armbrecht, L. H., Roughan, M., Rossi, V., Schaeffer, A., Davies, P. L., Waite, A. M., & Armand, L. K. (2014). Phytoplankton composition under contrasting oceanographic conditions: Upwelling and downwelling (eastern australia). *Continental Shelf Research*, *75*, 54–67.

Bai, J., & Perron, P. (2003). Critical values for multiple structural change tests. *The Econometrics Journal*, *6*(1), 72–78.

Bakun, A., & Agostini, V. N. (2001). Seasonal patterns of wind-induced upwelling/downwelling in the mediterranean sea. *Scientia Marina*, *65*(3), 243–257.

Bauer, J. E., Cai, W.-J., Raymond, P. A., Bianchi, T. S., Hopkinson, C. S., & Regnier, P. A. (2013). The changing carbon cycle of the coastal ocean. *Nature*, *504*(7478), 61–70.

Bensoussan, N., Romano, J.-C., Harmelin, J.-G., & Garrabou, J. (2010). High resolution characterization of northwest mediterranean coastal waters thermal regimes: to better understand responses of benthic communities to climate change. *Estuarine, Coastal and Shelf Science*, *87*(3), 431–441.

Bolaños, L. M., Karp-Boss, L., Choi, C. J., Worden, A. Z., Graff, J. R., Haëntjens, N., ... others (2020). Small phytoplankton dominate western north atlantic biomass. *The ISME journal*, *14*(7), 1663–1674.

Borges, A. V., Delille, B., & Frankignoulle, M. (2005). Budgeting sinks and sources of co2 in the coastal ocean: Diversity of ecosystems counts. *Geophysical research letters*, *32*(14).

Bosc, E., Bricaud, A., & Antoine, D. (2004). Seasonal and interannual variability in algal biomass and primary production in the mediterranean sea, as derived from 4 years of seawifs observations. *Global Biogeochemical Cycles*, *18*(1).

Brody, S. R., Lozier, M. S., & Dunne, J. P. (2013). A comparison of methods to determine phytoplankton bloom initiation. *Journal of Geophysical Research: Oceans*, *118*(5), 2345–2357.

Cianelli, D., D'Alelio, D., Uttieri, M., Sarno, D., Zingone, A., Zambianchi, E., & d'Alcalà, M. R. (2017). Disentangling physical and biological drivers of phytoplankton dynamics in a coastal system. *Scientific reports*, *7*(1), 1–15.

Coello-Camba, A., Diaz-Rua, R., Duarte, C. M., Irigoien, X., Pearman, J. K., Alam, I. S., & Agusti, S. (2020). Picocyanobacteria community and cyanophage infection responses to nutrient enrichment in a mesocosms experiment in oligotrophic waters. *Frontiers in Microbiology*, *11*, 1153. Retrieved from `https://www.frontiersin.org/article/10.3389/fmicb.2020.01153` doi: 10.3389/fmicb.2020.01153

d'Ortenzio, F., & Ribera d'Alcalà, M. (2009). On the trophic regimes of the mediterranean sea: a satellite analysis. *Biogeosciences*, *6*(2), 139–148.

Dubelaar, G. B., & Gerritzen, P. L. (2000). Cytobuoy: a step forward towards using flow cytometry in operational oceanography. *Scientia Marina*, *64*(2), 255–265.

Dubelaar, G. B., Gerritzen, P. L., Beeker, A. E., Jonker, R. R., & Tangen, K. (1999). Design and first results of cytobuoy: A wireless flow cytometer for in situ analysis of marine and fresh waters. *Cytometry: The Journal of the International Society for Analytical Cytology*, *37*(4), 247–254.

Dugenne, M., Thyssen, M., Nerini, D., Mante, C., Poggiale, J.-C., Garcia, N., ... Grégori, G. J. (2014). Consequence of a sudden wind event on the dynamics of a coastal phytoplankton community: an insight into specific population growth rates using a single cell high frequency approach. *Frontiers in microbiology*, *5*, 485.

Edwards, V., Icely, J., Newton, A., & Webster, R. (2005). The yield of chlorophyll from nitrogen: a comparison between the shallow ria formosa lagoon and the deep oceanic conditions at sagres along the southern coast of portugal. *Estuarine, Coastal and Shelf Science*, *62*(3), 391–403.

El Hourany, R., Abboud-abi Saab, M., Faour, G., Mejia, C., Crépon, M., & Thiria, S. (2019). Phytoplankton diversity in the mediterranean sea from satellite data using self-organizing maps. *Journal of Geophysical Research: Oceans*, *124*(8), 5827–5843.

Fraysse, M., Pinazo, C., Faure, V. M., Fuchs, R., Lazzari, P., Raimbault, P., & Pairaud, I. (2013). Development of a 3d coupled physical-biogeochemical model for the marseille coastal area (nw mediterranean sea): what complexity is required in the coastal zone? *PloS one*, *8*(12), e80012.

Fuchs, R., Thyssen, M., Creach, V., Dugenne, M., Izard, L., Latimier, M., ... Pommeret, D. (2022). Automatic recognition of flow cytometric phytoplankton functional groups using convolutional neural networks. *Limnology and Oceanography: Methods (submitted)*.

García-Reyes, M., Largier, J. L., & Sydeman, W. J. (2014). Synoptic-scale upwelling indices and predictions of phyto-and zooplankton populations. *Progress in Oceanography*, *120*, 177–188.

Gattuso, J., Frankignoulle, M., & Wollast, R. (1998). Carbon and carbonate metabolism in coastal aquatic ecosystems. *Annual Review of Ecology, Evolution, and Systematics*, *29*, 405-434.

Grob, C., Ulloa, O., Claustre, H., Huot, Y., Alarcon, G., & Marie, D. (2007). Contribution of picoplankton to the total particulate organic carbon concentration in the eastern south pacific. *Biogeosciences*, *4*(5), 837–852.

Hauss, H., Franz, J. M., & Sommer, U. (2012). Changes in n: P stoichiometry influence taxonomic composition and nutritional quality of phytoplankton in the peruvian upwelling. *Journal of sea Research*, *73*, 74–85.

Hunter-Cevera, K. R., Neubert, M. G., Olson, R. J., Shalapyonok, A., Solow, A. R., & Sosik, H. M. (2020). Seasons of syn. *Limnology and oceanography*, *65*(5), 1085–1102.

Hunter-Cevera, K. R., Neubert, M. G., Solow, A. R., Olson, R. J., Shalapyonok, A., & Sosik, H. M. (2014). Diel size distributions reveal seasonal growth dynamics of a coastal phytoplankter. *Proceedings of the National Academy of Sciences*, *111*(27), 9852–9857.

Jaffe, J. S., Franks, P. J., Roberts, P. L., Mirza, D., Schurgers, C., Kastner, R., & Boch, A. (2017). A swarm of autonomous miniature underwater robot drifters for exploring submesoscale ocean dynamics. *Nature communications*, *8*(1), 1–8.

Lajaunie-Salla, K., Diaz, F., Wimart-Rousseau, C., Wagener, T., Lefèvre, D., Yohia, C., ... Pinazo, C. (2021). Implementation and assessment of a carbonate system model (eco3m-carbox v1. 1) in a highly dynamic mediterranean coastal site (bay of marseille, france). *Geoscientific Model Development*, *14*(1), 295–321.

Lehahn, Y., Koren, I., Sharoni, S., d'Ovidio, F., Vardi, A., & Boss, E. (2017). Dispersion/dilution enhances phytoplankton blooms in low-nutrient waters. *Nature Communications*, *8*(1), 1–8.

Lévy, M., Ferrari, R., Franks, P. J., Martin, A. P., & Rivière, P. (2012). Bringing physics to life at the submesoscale. *Geophysical Research Letters*, *39*(14).

Li, W. (1995). Composition of ultraphytoplankton in the central north atlantic. *Marine Ecology Progress Series*, *122*, 1–8.

Lomas, M. W., & Moran, S. B. (2011). Evidence for aggregation and export of cyanobacteria and nano-eukaryotes from the sargasso sea euphotic zone. *Biogeosciences*, *8*(1), 203–216.

Lomas, M. W., Roberts, N., Lipschultz, F., Krause, J., Nelson, D., & Bates, N. (2009). Biogeochemical responses to late-winter storms in the sargasso sea. iv. rapid succession of major phytoplankton groups. *Deep Sea Research Part I: Oceanographic Research Papers*, *56*(6), 892–908.

Marrec, P., Grégori, G., Doglioli, A. M., Dugenne, M., Della Penna, A., Bhairy,

N., ... Thyssen, M. (2018). Coupling physics and biogeochemistry thanks to high-resolution observations of the phytoplankton community structure in the northwestern mediterranean sea. *Biogeosciences*, *15*(5), 1579–1606. Retrieved from `https://bg.copernicus.org/articles/15/1579/2018/` doi: 10.5194/bg-15-1579-2018

Martin-Platero, A. M., Cleary, B., Kauffman, K., Preheim, S. P., McGillicuddy, D. J., Alm, E. J., & Polz, M. F. (2018). High resolution time series reveals cohesive but short-lived communities in coastal plankton. *Nature communications*, *9*(1), 1–11.

Martiny, A. C., Talarmin, A., Mouginot, C., Lee, J. A., Huang, J. S., Gellene, A. G., & Caron, D. A. (2016). Biogeochemical interactions control a temporal succession in the elemental composition of marine communities. *Limnology and Oceanography*, *61*(2), 531–542.

Martz, T., Send, U., Ohman, M. D., Takeshita, Y., Bresnahan, P., Kim, H.-J., & Nam, S. (2014). Dynamic variability of biogeochemical ratios in the southern california current system. *Geophysical Research Letters*, *41*(7), 2496–2501.

Mayot, N., d'Ortenzio, F., Ribera d'Alcalà, M., Lavigne, H., & Claustre, H. (2016). Interannual variability of the mediterranean trophic regimes from ocean color satellites. *Biogeosciences*, *13*(6), 1901–1917.

Millot, C. (1979). Wind induced upwellings in the gulf of lions. *Oceanologica Acta*, *2*(3), 261–274.

Morel, A., & André, J.-M. (1991). Pigment distribution and primary production in the western mediterranean as derived and modeled from coastal zone color scanner observations. *Journal of Geophysical Research: Oceans*, *96*(C7), 12685–12698.

Odic, R., Bensoussan, N., Pinazo, C., Taupier-Letage, I., & Rossi, V. (2022). Sporadic wind-driven upwelling/downwelling and associated cooling/warming along the north-west mediterranean coastlines. *(in prep.)*.

Olson, R. J., Shalapyonok, A., & Sosik, H. M. (2003). An automated submersible flow cytometer for analyzing pico-and nanophytoplankton: Flowcytobot. *Deep Sea Research Part I: Oceanographic Research Papers*, *50*(2), 301–315.

Pairaud, I., Gatti, J., Bensoussan, N., Verney, R., & Garreau, P. (2011). Hydrology and circulation in a coastal area off marseille: Validation of a nested 3d model with observations. *Journal of marine systems*, *88*(1), 20–33.

Palma, E. D., & Matano, R. P. (2009). Disentangling the upwelling mechanisms of the south brazil bight. *Continental Shelf Research*, *29*(11-12), 1525–1534.

Ribalet, F., Swalwell, J., Clayton, S., Jiménez, V., Sudek, S., Lin, Y., ... Armbrust, E. V. (2015). Light-driven synchrony of prochlorococcus growth and mortality in the subtropical pacific gyre. *Proceedings of the National Academy of Sciences*, *112*(26), 8008–8012.

Richardson, T. L., & Jackson, G. A. (2007). Small phytoplankton and carbon export from the surface ocean. *Science*, *315*(5813), 838–840.

Rossi, V., Garçon, V., Tassel, J., Romagnan, J.-B., Stemmann, L., Jourdin, F., ... Morel, Y. (2013). Cross-shelf variability in the iberian peninsula upwelling system: Impact of a mesoscale filament. *Continental Shelf Research*, *59*, 97–114.

Rossi, V., Schaeffer, A., Wood, J., Galibert, G., Morris, B., Sudre, J., ... Waite, A. M. (2014). Seasonality of sporadic physical processes driving temperature and nutrient high-frequency variability in the coastal ocean off southeast australia. *Journal of Geophysical Research: Oceans*, *119*(1), 445–460.

Sapiano, M., Brown, C., Schollaert Uz, S., & Vargas, M. (2012). Establishing a global climatology of marine phytoplankton phenological characteristics. *Journal of Geophysical Research: Oceans*, *117*(C8).

Sosik, H. M., Olson, R. J., Neubert, M. G., Shalapyonok, A., & Solow, A. R. (2003). Growth rates of coastal phytoplankton from time-series measurements with a submersible flow cytometer. *Limnology and Oceanography*, *48*(5), 1756–1765.

572 Sun, J., Feng, Y., Zhang, Y., & Hutchins, D. (2007, 09). Fast microzooplankton
573 grazing on fast-growing, low-biomass phytoplankton: A case study in spring in
574 chesapeake bay, delaware inland bays and delaware bay. *Hydrobiologia*, *589*,
575 127-139. doi: 10.1007/s10750-007-0730-6
576 Teixeira, I., Arbones, B., Froján, M., Nieto-Cid, M., Álvarez-Salgado, X. A., Cas-
577 tro, C. G., ... Figueiras, F. (2018). Response of phytoplankton to enhanced
578 atmospheric and riverine nutrient inputs in a coastal upwelling embayment.
579 *Estuarine, Coastal and Shelf Science*, *210*, 132–141.
580 Thyssen, M., Fuchs, R., Créach, V., Artigas, L. F., Grégori, G., Marrec, P., ... oth-
581 ers (2021). Standard vocabulary, consensual functional groups and automated
582 classification for phytoplankton high throughput datasets using automated flow
583 cytometry. In *Aslo 2021.*
584 Thyssen, M., Mathieu, D., Garcia, N., & Denis, M. (2008). Short-term variation of
585 phytoplankton assemblages in mediterranean coastal waters recorded with an
586 automated submerged flow cytometer. *Journal of Plankton Research*, *30*(9),
587 1027–1040.
588 Tréguer, P., & Le Corre, P. (1975). Manuel d'analyse des sels nutritifs dans l'eau de
589 mer (utilisation de l'autoanalyseur ii technicon), 110, lab. d'océanogr. *Chim.,*
590 *Univ. de Bretagne Occident., Brest, France*.
591 Truong, C., Oudre, L., & Vayatis, N. (2020). Selective review of offline change point
592 detection methods. *Signal Processing*, *167*, 107299.
593 Wilkerson, F. P., Lassiter, A. M., Dugdale, R. C., Marchi, A., & Hogue, V. E.
594 (2006). The phytoplankton bloom response to wind events and upwelled
595 nutrients during the coop west study. *Deep Sea Research Part II: Topical*
596 *Studies in Oceanography*, *53*(25-26), 3023–3048.
597 Wimart-Rousseau, C., Lajaunie-Salla, K., Marrec, P., Wagener, T., Raimbault, P.,
598 Lagadec, V., ... others (2020). Temporal variability of the carbonate system
599 and air-sea co2 exchanges in a mediterranean human-impacted coastal site.
600 *Estuarine, Coastal and Shelf Science*, *236*, 106641.
601 Wu, Y., Platt, T., Tang, C. C., Sathyendranath, S., Devred, E., & Gu, S. (2008). A
602 summer phytoplankton bloom triggered by high wind events in the labrador
603 sea, july 2006. *Geophysical Research Letters*, *35*(10).

Chapter conclusion

Chapter 2 has introduced a proper framework to characterize the general link of phytoplankton functional groups with their direct marine environment. This link has then been studied at a submesoscale resolution and high-temporal frequency during wind-induced events in the current chapter. The FUMSECK study has evidenced the possibly high influence of these events in stratified periods, as they result in short and intense nutrient enrichments in a previously strongly nutrient-limited environment. The introduced CNN and change-point methodologies have confirmed the pattern observed in a single event: the impact of coastal wind-induced events has a significant and cPFG-differentiated impact on the phytoplankton biomass (and abundance) in oligotrophic waters. The daily impact of the most extreme events was comparable with the one observed during the spring bloom.

4. Conclusion and perspectives

The crucial role played by the phytoplankton in biogeochemical processes along with their morphological diversity and response capacities have created a need for dedicated hardware and statistical methodologies. The MDGMM and MIAMI have given insights into the phytoplankton ecological niches and possible reactions to long-term environmental changes, while RUBALIZ will provide adapted epipelagic boundaries for future FC sampling vertical strategies. To complete the investigation, high-temporal frequency and local responses of phytoplankton cells have been studied in this properly defined framework, using FC, supervised neural methods, and change-point detection models. In this final chapter, the main features and goals of the introduced approaches are first summarized, notably in Table 4.1. Then, perspectives to overcome the current limitations of the approaches or to improve their accuracy are presented.

Model/Method	Dataset type	Dataset size	Data dimension	Oceanographic context	Code and data
MDGMM/ MIAMI	Tabular mixed	Moderate	Moderate	Ecological niches determination / Environmental shifts prospective	Code and data
RUBALIZ	Depth-dependent	Moderate	High	Determination of local epipelagic and mesopelagic vertical boundaries	Code and data
CNN	Functional multivariate	High	Low	Automatic recognition of phytoplankton functional groups	Code and data
Change-points	Time series	Low	Low	Characterizing phytoplankton response to wind-induced events	Code and data

Table 4.1. – Summary of the models introduced per type of data, data characteristics, and oceanographic question. The dataset size was evaluated by considering datasets of less than 1 000 observations as small, from 1 000 to 50 000 as moderate, and superior to 50 000 as big datasets. Similarly, datasets with dimensions inferior to 10 were regarded as low-dimensional, between 10 and 100 as datasets of moderate dimension and higher to 100 as high-dimension datasets.

1. Characterization of the ecological niches and vertical zone boundaries by the MDGMM and RUBALIZ methodologies

1.1. MDGMM and MIAMI

The Mixed Deep Gaussian Models (MDGMM) have proven to be solid benchmark models. Unlike other neural-based methods, the MDGMM also preserved good results interpretability thanks notably to visualization tools. Doing so, the prominence of the spatio-temporal dependence in the SOMLIT data was put in evidence with highly contrasted conditions between the Atlantic Ocean, the Mediterranean Sea, the Channel, and the Gironde River. The Mediterranean Sea environmental variables and cPFG assemblages seemed to be the more contrasted and differed the most compared to the other zones. MDGMM and MIAMI gave a fine characterization of the phytoplankton functional groups ecological niches. All other things equal, rising temperature in winter significantly fostered the abundances of all cPFGs except Redpicopro. Similarly, simulated phosphate pulse fueled even more the abundances of all groups except for Orgpicopro which demonstrated substantial adaptation skills in oligotrophic environments.

These results were derived using the minimal MDGMM and MIAMI architectures. Using deep MDGMM architectures would have given more explanatory power but also significantly greater variability in the results as expected from the work by Selosse et al. 2020 on the DGMM and confirmed by the experiments conducted in the MDGMM study (Chapter 2 Section 1.2). These experiments highlighted that bigger MDGMM architectures led to less stable partitions from one run to another and were not sufficiently compensated by better clustering performances. This is especially the case for the M2DGMM architecture which presented lower performances than the M1DGMM despite its more flexible assumptions concerning the continuous data dependence structure. This could be explained by the MCEM training in a high-dimensional latent space, by the choice of the GLLVM as an embedding layer, or by the choice of the likelihood as the objective function to maximize.

The conditional independence assumption and the exponential link functions of the GLLVM highly structure the signal going through the GLLVM layer. Yet, potential mismatches between the link function distributions and the data could prevent the clustering process to be properly conducted. Moreover, the signal reaching the DGMM layers could be too structured/simple for additional DGMM layers to be useful. Hence, other embedding layers could be tested as embedding layers based on Genetic Programming (GP) algorithms (Poli et al. 2008). Doing so, the training would not consist in learning the coefficients of a user-chosen link function but rather to learn the best link function relating the original and latent space. As most of the latent distributions depend on the chosen link distributions, the training by MCEM would

also have to be adapted. The simplest way would be to automatically differentiate the GP link function as it is already done with GLLVM link functions, which requires making compatible GP packages with the automatic differentiation packages such as JAX (Bradbury et al. 2018) or Autograd (Maclaurin et al. 2015). Another way would be to change the MCEM training for a gradient descent-based approach, which is the main family of approaches used in the supervised neural network case to maximize the criterion of interest, here the negative log-likelihood of the model. Indeed, the MDGMM due to its model-based structure hence pursues two potentially conflicting goals: to provide a good representation of the mixed original data (*i.e.* a dimension reduction task) and to separate the observations into homogeneous and distinct groups (*i.e.* a clustering task). In our implementation, the total number of iterations was ruled by the likelihood criterion accounting for the dimension reduction fitness. Among these iterations, the iteration presenting the highest silhouette score, reflecting the quality of the clustering process, was selected. Other possibilities exist to address the trade-off between these two goals, such as maximizing a weighted sum of the likelihood and silhouette score criteria.

Combining a new embedding layer with a new criterion to maximize may ensure better stability of deep MDGMM architectures and better use of the flexibility provided by the model. MIAMI, which extends the MDGMM to generate synthetic data, could also benefit from such changes. In parallel, possible improvements of the MIAMI model are numerous. For instance, the acceptance rate of the synthetic data generating process could be improved using Bayesian optimization (Frazier 2018) to target the latent areas corresponding to the desired synthetic data characteristics rather than randomly sampling the latent space.

1.2. Determination of the epipelagic and active mesopelagic layer boundaries

The determination of surface phytoplankton ecological niches is of primary importance and gives a proper physical and biological framework to study pico-nanophytoplankton cells behavior. In the same spirit, the RUBALIZ method gives insights into the limit of the epipelagic zone in which the phytoplankton cells operate and separate the epipelagic zone from the most active part of the mesopelagic zone.

The determination of the RUBALIZ boundaries relied on a frequentist change-point detection method. The depth range to look for the upper and lower boundary respectively was manually specified based on values of the literature before launching the algorithm. Alternatively, a Bayesian framework could be more suited to integrate prior information about where each boundary is likely to be located. This could be done by giving a high prior probability to the likely depth ranges and null or nearly-null probability mass otherwise. The estimation variance would then not be evaluated on a grid of depth ranges but derived from the variance of the posterior distribution.

As demonstrated in the paper, the mesopelagic boundaries identified on physical fluxes by RUBALIZ were highly consistent with Particular Organic Carbon (POC) inputs. It highlights that these boundaries captured both the physical and biological dimensions of the water column, and as a result could help to build more founded carbon budgets. Yet, by themselves, these new boundaries did not fully resolve the now-standard negative carbon budget discrepancy (Burd et al. 2010), *i.e.* the fact that the estimated carbon demand exceeds the estimated carbon supply in the mesopelagic zone.

Other variables and estimation parameters also have a crucial influence on this carbon budget discrepancy. This is the case of the Leucine-to-Carbon Conversion Factors (CF) and the prokaryotic growth efficiency (PGE). These two sets of parameters are often set to fixed values (1.55 or 0.44 kgC mol^{-1} and 0.08, respectively) based on literature medians for all world locations, depths, and seasons (Giering et al. 2014). In a work currently in preparation, we rely on model inversion to provide local values of such parameters using RUBALIZ boundaries as the mesopelagic zone limits. In this study, the mesopelagic trophic network is modeled according to Anderson et al. 2010. The model takes as input the CFs and PGEs associated with both attached-to-particles prokaryotes and free-living prokaryotes. The model inversion is conducted by matching model outputs with their *in situ* measurements. Four *in situ* measurements are used in this respect: the heterotrophic production of non-sinking prokaryotes, the heterotrophic production of sinking prokaryotes, the respiration of sinking prokaryotes, and the respiration of zooplankton. As a result, the CFs and PGEs leading to the closest model and *in situ* output fluxes are identified as the most likely. The so-determined values of CFs and PGEs are hence the most consistent with the current knowledge of the mesopelagic trophic network as modeled by Anderson et al. 2010.

2. High temporal frequency resolution of phytoplankton responses

The low monitoring cost of automated flow cytometry makes it a good candidate for long-term high-frequency phytoplankton group tracking. Yet, the lack of consensus concerning the way to assign the cells to functional groups currently strongly limits this ability as highlighted in Section 3.2.2.

2.1. Automating the flow cytometry manual gating process

After the introduction of a common nomenclature by Thyssen et al. 2021, the automation of the classification process by a CNN constituted a second step towards the full automation of FC. This automation has taken advantage of the whole pulse-shape signal issued by FC rather than simple descriptors (e.g. the mean or the maximum of each pulse shape curve) as most supervised learning models in the literature did. Yet, the CNN focused on the pulse shapes and did not use the images taken by FC. Given

the low physiological variability of the smallest PFGs and the current resolution of the FC camera, FC images are more valuable for nano-microphytoplankton than for picophytoplankton. This is all the more the case that the microphytoplankton cells were not well resolved in the presented works using the pulse shapes. Indeed, the highest FLR-threshold-acquisition protocol (e.g. FLR25 at the SSL@MM or FLR30 during the SWINGS cruise, with a sampling volume of approx. 5 mL) did not always sample enough microphytoplankton cells to be representative. Besides, the fluorescence and diffusion signals of the biggest cells tend to saturate the FC photomultipliers and lead to distorted pulse shapes. Thus, the resulting low representativity of the microphytoplankton group has for example lead to its non-consideration in Section 3.3.

Several solutions can be designed to better resolve the whole phytoplankton size range. First, one can use a third pulse shape acquisition protocol based for instance on a size-dependent threshold (e.g. a FWS or SWS threshold) rather than on a fluorescence-dependent threshold (a FLR threshold). Secondly, one could include a second head to the CNN to deal with both images and pulse-shape signals, and thus cancel out the limitation of both types of signals at each end of the size distribution (image limitations for the smallest cells and pulse shape limitations for the biggest cells). Third, the pico-nanophytoplankton could be resolved by the CNN introduced here and the nano-microphytoplankton by a network specialized in images such as the one developed by the CEREGE in the RAPP project (Reconnaissance Automatique du Plancton et Pollens, ECCOREV) and integrated to the ParticleTrieur software (Marchant et al. 2020).

Going back to the pulse shape handling, several improvements are possible concerning individual predictions and the predictions of successive FC acquisitions. First, concerning individual predictions, the CNN architecture could be improved by adding attention mechanisms to better capture the dependence existing between each value of the interpolated pulse shapes, as it is often done in Natural Language Processing tasks (Vaswani et al. 2017). Besides, the CNN predictions were performed for each cell separately, not at the functional group level. Building bridges between these two levels during the training process could be a future axis of research. This could be done by adding community-dependent penalties to the CNN loss. Conversely, other methods than neural methods might be better suited for this task. This is for instance the case of reinforcement learning methods (Sutton et al. 2018) which could learn to reproduce the gates drawn by manual experts and associate them with functional groups.

Secondly, the dependence between successive FC acquisitions was not addressed, and the prediction on each acquisition was performed independently from the previous and following one. However, when looking at successive acquisitions, the same populations can be tracked through the day and most of the time the associated 2D scatter plots only slightly move through the day during the cell nycthemeral cycles. This procedure was therefore not the most efficient as it boils down to performing a

slightly different version of the same task multiple times. The main reason for this methodology was that it is the simplest from a modeling point of view, but was also due to the fact that the classes to predict were strongly unbalanced. The training is hence performed over a strongly re-balanced training set compiling multiple acquisitions to collect enough data for the unbalanced classes, rather than on raw successive acquisitions. Training the CNN on re-balanced successive acquisitions and making its loss dependent on the positions of each cPFG in the previously predicted acquisitions might be a way to proceed. External models could also be used to post-treat the results generated by the CNN and track time-evolving communities.

Beyond these model improvements, the current theoretical shift from model-centric approaches to data-centric approaches as stated by Strickland 2022, tends to shift the priority away from identifying the best model to focusing more on the data quality and pre-processing.
The data quality was here enhanced by asking six FC experts to gate acquisitions and only keeping the cells that were similarly labeled by a 2/3 majority of experts. This approach was inspired by the one used for the ImageNet repository (Deng et al. 2009) which gathers images annotated and voted by the community. The so-collected data were stored on the ERDDAP repository (Simons et al. 2012) and made accessible to the FC community. The number of observations in both the SSL@MM and GEOTRACES SWINGS datasets was 50 000 observations. This medium size dataset might not reveal the full potential of neural methods and also limit the depth of the implemented network. The most obvious way to overcome this issue would be to replicate the multi-expert labeling approach in other oceanic zones to tend towards a better representation of the global ocean. The CNN has demonstrated a high ability to generalize between very different oceanic zones. Training the CNN on a global ocean representative dataset could thus enable the deployment of a unique CNN for the majority of oceanic areas. Alternatively, for a given dataset size, simple data augmentation procedures could be enforced. The easiest procedure is maybe to add a centered and small-variance Gaussian noise to the pulse-shape values (Lee 2000). Alternatively, as each cell goes through the laser in a random orientation, the pulse shapes could simply be reverted to mimic the fact that a cell went through the laser after a 180° rotation.

The pulse shapes are for the moment interpolated to a fixed length of 120 values. This 120-value length was based on the observed third quartile length of the data with the intuition that interpolating to a lower length destroys more signal than interpolating to a higher length. Manual experiments have been conducted to investigate the role of the fixed length over the performance and showed little impact on the performance. They were however performed by letting the other hyper-parameters fixed. A joint change of all hyperparameters might cause the fixed interpolation length to have an impact on the performance but also on the computational burden of the model (reducing the burden if a lower fixed length was selected). Optimizing directly the interpolation length as in Talebi et al. 2021 with the other hyper-parameters could

also enhance the CNN performance.

Finally, the pulse-shape trained CNN could be used for other tasks and in other contexts. The CNN presented here was designed for cPFG classification purposes but could perform biovolume and biomass estimations (in a regression framework). In the FUMSECK and SSL@MM studies, the biovolume was first estimated by an empirical relationship between the Total FWS (the area under the curve of the Forward Scatter pulse shape) and used to derive the biomass using relationships coming from the literature (Verity et al. 1992; Menden-Deuer et al. 2000). The errors performed at each stage of this two-stage procedure were hence cumulative. To overcome this issue, a CNN could be trained to directly predict the biomass (or biovolume) of a cell from its pulse-shape signal. Indeed, the FWS and SWS signals convey strong information about the cell shape and the two fluorescence signals reflect the fluorescent pigment content and location in the cell. The main challenge of this approach would actually be to generate reliable ground-truth biomasses for each cell (*i.e.* the variable to predict). This could be done by passing a sample in a flow cytometer, collecting the pulse shape of the cells contained in the sample, and estimating their biomass (or biovolume) distribution using filters of different sizes. Alternatively, the cells could be first sorted and separated by functional group using the CNN to estimate cPFG-biomass distributions.

The CNN could also be used in other operational contexts. It was here used in an offline/a posteriori manner to retrospectively determine the cPFG abundances. Yet, it could be used in an online fashion, *i.e.* as a prospective tool during cruises to adapt the sampling strategy. It could for example be implemented as a complement to the SPASSO system (`https://spasso.mio.osupytheas.fr/`) that tracks phytoplankton spatial distribution and circulation using satellite chlorophyll-a estimations. The CNN *in situ* measurements would allow to go beyond this phytoplankton "bulk" estimation by chlorophyll-a and to provide a near-real-time adaptive sampling at the cPFG level.

To ease the future embedding of the CNN in scientific programs or workflows, the trained models and developed utilities were embedded in a predictive workflow taking the form of a Docker container built with the help of a software engineer. This workflow takes as input the new acquisitions performed and store the associated predictions. These predictions can then be visualized by the user to ensure their quality (see Figure 4.1). If satisfactory, the predictions are sent to a database along with the necessary metadata (date, GPS coordinates, hardware identifier, etc.). This workflow can therefore be viewed as a cross-platform stand-alone software and may contribute to the usage of convolutional neural networks by a wider oceanographic community.

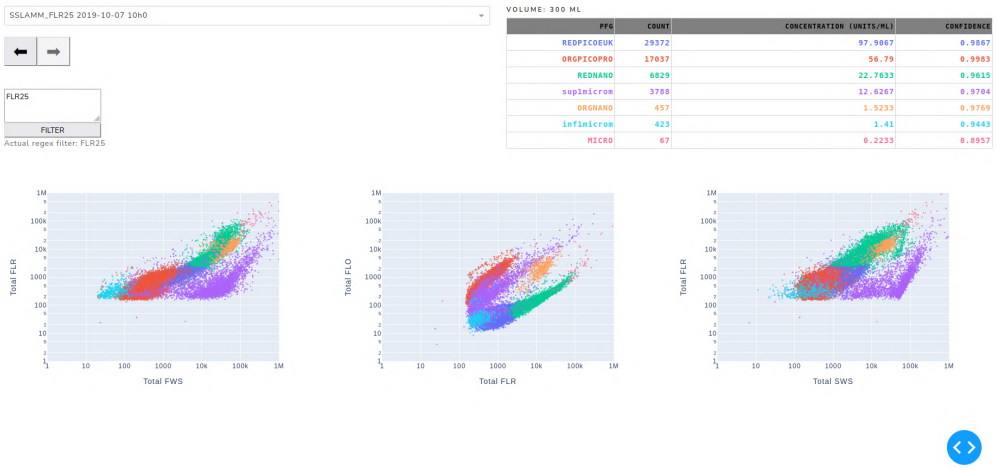


Figure 4.1. – Window example of the CNN prediction workflow.

2.2. Resolving the effect of sporadic wind-induced events on phytoplankton functional groups

The extensive use of automated flow cytometry made it possible to track phytoplankton functional group responses to wind-related events, a storm in the FUMSECK case, and sporadic coastal upwelling events at the SSL@MM station. The two studies evidenced fast biological responses to physical forcing events. While the FUMSECK study focused on a particular event, the study at the SSL@MM station evidenced reproducible patterns notably helpful for the oceanographic modeling community. These two studies confirmed fast and major changes in phytoplankton communities and highlighted the necessity to account for such events in global carbon budgets. Besides, the FUMSECK and SSL@MM studies showed that the changes in phytoplankton communities were strongly explained by water masses replacements in the first place followed by group-specific adaptations of the phytoplankton cells to their new environment. The hydrodynamics at stake was best described in the FUMSECK study as it coupled *in situ* ship-located and glider-collected data with satellite data and outputs from an atmospherical model. The approach confirmed the consistency between the measurements coming from different instruments and offered a detailed view of the currentology and water mixing patterns.

Yet, starting from the SSL@MM and FUMSECK studies, several bottom-up and top-down phytoplankton regulation forces have to be better resolved in future research. This is the case of nutrients, listed by the GOOS expert panels as Essential Oceanic Variables (Miloslavich et al. 2018), and zooplankton and virus populations.
The data collection of the main phytoplankton nutrients (nitrates, nitrites, phosphates, ammonium, and silicate [1]) is performed manually in most studies including

1. The silicate is mainly used by diatoms which are not well resolved here as mentioned earlier. Hence silicate concentration estimation is not evoked in this paragraph

ours. Samples are traditionally collected and stabilized *in situ* before being analyzed in a laboratory. This procedure significantly limits the current nutrient data frequency. *In situ* nutrient sensors have been in development for several decades but still exhibit numerous limitations such as insufficient detection limits, difficult-to-assess reliability, sensibility to biofouling, or the necessity of data post-treatment (Daniel et al. 2020). Yet, they could be deployed in addition to manual measurements to obtain high-frequency series or at least information on long-term trends. More precisely for the SSL@MM station, nitrate could be resolved by Ultraviolet Optical Sensors as provided for example by the ISUS instrument (Sakamoto et al. 2017), and phosphate by Electrochemical Sensors (Jońca et al. 2013). These two methods indeed have a better measurement frequency than methods based on wet chemistry (Daniel et al. 2020), and could be used during the whole annual cycle (or just during the wind-induced events to provide a better contextualization).

Furthermore, the zooplankton grazing and viral lysis were absent from our analysis. The zooplankton compartment is however essential to well-describe the total plankton succession scheme, especially in coastal areas (Anderson 1998; Hereu et al. 2006). Following heterotrophic predators at high-frequency could be conducted by the Cytopro flow cytometer (Silovic et al. 2017), which is currently in the final stages of development. The Cytopro adds an automating staining module to the Cytosense FC and marks the heterotrophes with a SYBR Green dye (manufactured by Invitrogen™) before incubating them for nearly 30 minutes. The resulting sampling frequency is hence inferior to the hour.
Conversely, other automatic or semi-automatic approaches based on images or acoustic methods could be implemented. Semi-automatic zooplankton image-based recognition systems have for instance been introduced in Romagnan et al. 2016. The zooplankton cells were collected using a net of predefined mesh size, and a Random Forest model (Breiman 2001) was trained to recognize the zooplankton groups from images issued by the zooscan software (Gorsky et al. 2010). Yet, notably due to its manual zooplankton collection process, hourly temporal resolution seems for the moment out-of-reach for this method. Conversely, approaches relying on acoustic-based systems such as the Acoustic Water Column Profilers (AWCP), provide a high-frequency temporal resolution (infra-minute) but with a low zooplankton taxonomic resolution as in Borstad et al. 2010. Hence, in our case, acoustic-based methods to track the global zooplankton compartment may be better suited than image-based methods.

Concerning the resolution of oceanic viruses, Breitbart 2012 have identified the collection of comprehensive virus datasets as a key challenge for the field. Compared to other existing methods such as microfluidic digital PCR (Tadmor et al. 2011), epifluorescence microscopy-based methods (Allers et al. 2013) or viral genome identification (Mizuno et al. 2013; Roux et al. 2014), Viral Tagging (VT) seems the more suited for high-frequency studies (Brum et al. 2015). VT stains the DNA of wild viruses with SYBR Gold and then incubates these viruses with cultivated host organisms. A Flow cytometer sorts the infected cells from non-infected cells and metagenomics methods

could give further insights into the identity of the viruses infecting the different hosts. Doing so, one could in principle determine the proportion of phytoplankton infected cells and which virus strain infected each cPFG. VT has been deployed successfully for *Synechococcus*-infecting viruses (Deng et al. 2013; Deng et al. 2014). Additional research is necessary to obtain a fully-automatized method for most cPFGs-related viruses, but viral tagging seems to be a promising path for the joint high-frequency study of cPFG and viruses in the future.

Finally, the results presented in the FUMSECK and SSL@MM studies took place in the open Ligurian Sea and in a coastal station of the Northwestern Mediterranean Sea. As mentioned earlier, the Mediterranean Sea is a well-suited laboratory to study the impact of sudden and intense wind-induced forcing due to its "hotspot" for climate change (Group et al. 2011) status. As a result, we expect the main detailed patterns evidenced here to give insightful perspectives about phytoplankton response capacities to sporadic events in oligotrophic waters.

> Who has the means to save us from ourselves?
> To pull us from the vicious cycles feeding back again.
>
> *Trivium discussing our climatic future*

References

[AK19] Amir Ahmad and Shehroz S Khan. "Survey of state-of-the-art mixed data clustering algorithms". In: *IEEE Access* 7 (2019), pp. 31883–31902 (cit. on pp. 29, 30, 32).

[All+13] Elke Allers, Cristina Moraru, Melissa B Duhaime, Erica Beneze, Natalie Solonenko, Jimena Barrero-Canosa, Rudolf Amann, and Matthew B Sullivan. "Single-cell and population level viral infection dynamics revealed by phage FISH, a method to visualize intracellular and free viruses". In: *Environmental microbiology* 15.8 (2013), pp. 2306–2318 (cit. on p. 226).

[AT10] Thomas R Anderson and Kam W Tang. "Carbon cycling and POC turnover in the mesopelagic zone of the ocean: Insights from a simple model". In: *Deep Sea Research Part II: Topical Studies in Oceanography* 57.16 (2010), pp. 1581–1592 (cit. on p. 221).

[And98] V Anderson. "Salp and pyrosomid blooms and their importance in biogeochemical cycles". In: *The biology of pelagic tunicates* (1998), pp. 125–137 (cit. on p. 226).

[BP03] Jushan Bai and Pierre Perron. "Computation and analysis of multiple structural change models". In: *Journal of applied econometrics* 18.1 (2003), pp. 1–22 (cit. on p. 90).

[Bor+10] Gary Borstad, Leslie Brown, Mei Sato, David Lemon, Randy Kerr, and Peter Willis. "Long zooplankton time series with high temporal and spatial resolution". In: *OCEANS 2010 MTS/IEEE SEATTLE*. IEEE. 2010, pp. 1–9 (cit. on p. 226).

[Bra+18] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. *JAX: composable transformations of Python+NumPy programs*. Version 0.2.5. 2018. URL: http://github.com/google/jax (cit. on p. 220).

[Bre01] Leo Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32 (cit. on p. 226).

[Bre12] Mya Breitbart. "Marine viruses: truth or dare". In: *Annual review of marine science* 4 (2012), pp. 425–448 (cit. on p. 226).

[BM12] Ryan P Browne and Paul D McNicholas. “Model-based clustering, classification, and discriminant analysis of data with mixed type”. In: *Journal of Statistical Planning and Inference* 142.11 (2012), pp. 2976–2984 (cit. on p. 30).

[BS15] Jennifer R Brum and Matthew B Sullivan. “Rising to the challenge: accelerated pace of discovery transforms marine virology”. In: *Nature Reviews Microbiology* 13.3 (2015), pp. 147–159 (cit. on p. 226).

[Bur+10] Adrian B Burd, Dennis A Hansell, Deborah K Steinberg, Thomas R Anderson, Javier Arístegui, Federico Baltar, Steven R Beaupre, Ken O Buesseler, Frank DeHairs, George A Jackson, et al. “Assessing the apparent imbalance between geochemical and biochemical indicators of meso- and bathypelagic biological activity: What the ... is wrong with present calculations of carbon budgets?” In: *Deep Sea Research Part II: Topical Studies in Oceanography* 57.16 (2010), pp. 1557–1571 (cit. on p. 221).

[CG10] Gail A Carpenter and Stephen Grossberg. *Adaptive resonance theory*. 2010 (cit. on p. 30).

[CG12] Jie Chen and Arjun K Gupta. “Parametric statistical change point analysis: with applications to genetics, medicine, and finance”. In: (2012) (cit. on p. 89).

[CA88] BC Cho and Farooq Azam. “Major role of bacteria in biogeochemical fluxes in the ocean’s interior”. In: *Nature* 332.6163 (1988), pp. 441–443 (cit. on p. 21).

[Dan+20] Anne Daniel, Agathe Laës-Huon, Carole Barus, Alexander D Beaton, Daniel Blandfort, Nathalie Guigues, Marc Knockaert, Dominique Munaron, Ian Salter, E Malcolm S Woodward, et al. “Toward a harmonization for using in situ nutrient sensors in the marine environment”. In: *Frontiers in Marine Science* 6 (2020), p. 773 (cit. on p. 226).

[DA12] Gil David and Amir Averbuch. “SpectralCAT: categorical spectral clustering of numerical and nominal data”. In: *Pattern Recognition* 45.1 (2012), pp. 416–433 (cit. on p. 30).

[Den+09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. “Imagenet: A large-scale hierarchical image database”. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255 (cit. on p. 223).

[Den12] Li Deng. “The mnist database of handwritten digit images for machine learning research”. In: *IEEE Signal Processing Magazine* 29.6 (2012), pp. 141–142 (cit. on p. 161).

[Den+13] Li Deng, Ann Gregory, Suzan Yilmaz, Bonnie T Poulos, Philip Hugenholtz, and Matthew B Sullivan. “Contrasting life strategies of viruses that infect photo-and heterotrophic bacteria, as revealed by viral tagging”. In: *MBio* 4.1 (2013), e00516–12 (cit. on p. 227).

[Den+14] Li Deng, J Cesar Ignacio-Espinoza, Ann C Gregory, Bonnie T Poulos, Joshua S Weitz, Philip Hugenholtz, and Matthew B Sullivan. "Viral tagging reveals discrete populations in Synechococcus viral genome sequence space". In: *Nature* 513.7517 (2014), pp. 242–245 (cit. on p. 227).

[Dub+99] George BJ Dubelaar, Peter L Gerritzen, Arnout ER Beeker, Richard R Jonker, and Karl Tangen. "Design and first results of CytoBuoy: A wireless flow cytometer for in situ analysis of marine and fresh waters". In: *Cytometry: The Journal of the International Society for Analytical Cytology* 37.4 (1999), pp. 247–254 (cit. on p. 126).

[DHS73] Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification and scene analysis*. Vol. 3. Wiley New York, 1973 (cit. on p. 90).

[Dug+14] Mathilde Dugenne, Melilotus Thyssen, David Nerini, Claude Mante, Jean-Christophe Poggiale, Nicole Garcia, Fabrice Garcia, and Gérald J Grégori. "Consequence of a sudden wind event on the dynamics of a coastal phytoplankton community: an insight into specific population growth rates using a single cell high frequency approach". In: *Frontiers in microbiology* 5 (2014), p. 485 (cit. on p. 26).

[EP79] Richard W Eppley and Bruce J Peterson. "Particulate organic matter flux and planktonic new production in the deep ocean". In: *Nature* 282.5740 (1979), pp. 677–680 (cit. on p. 17).

[Est+96] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. "A density-based algorithm for discovering clusters in large spatial databases with noise." In: *Kdd*. Vol. 96. 1996, pp. 226–231 (cit. on p. 30).

[FL+21] Hanna Farnelid, Catherine Legrand, et al. "Seasonality of Coastal Picophytoplankton Growth, Nutrient Limitation, and Biomass Contribution." In: *Frontiers in microbiology* 12 (2021), pp. 786590–786590 (cit. on p. 70).

[Fin+10] Zoe V Finkel, John Beardall, Kevin J Flynn, Antonietta Quigg, T Alwyn V Rees, and John A Raven. "Phytoplankton in a changing world: cell size and elemental stoichiometry". In: *Journal of plankton research* 32.1 (2010), pp. 119–137 (cit. on p. 18).

[Fow+20] Bethany L Fowler, Michael G Neubert, Kristen R Hunter-Cevera, Robert J Olson, Alexi Shalapyonok, Andrew R Solow, and Heidi M Sosik. "Dynamics and functional diversity of the smallest phytoplankton on the Northeast US Shelf". In: *Proceedings of the National Academy of Sciences* 117.22 (2020), pp. 12215–12221 (cit. on p. 70).

[Fra18] Peter I Frazier. "Bayesian optimization". In: *Recent advances in optimization and modeling of contemporary problems*. Informs, 2018, pp. 255–278 (cit. on p. 220).

[Ful65] Mack J Fulwyler. "Electronic separation of biological cells by volume". In: *Science* 150.3698 (1965), pp. 910–911 (cit. on p. 23).

[GLM14] Francisca C Garcia, Angel Lopez-Urrutia, and Xose Anxelu G Moran. "Automated clustering of heterotrophic bacterioplankton in flow cytometry data". In: *Aquatic Microbial Ecology* 72.2 (2014), pp. 175–185 (cit. on p. 24).

[Gas12] Florent Gasparin. "Caractéristiques des masses d'eau, transport de masse et variabilité de la circulation océanique en mer de corail (Pacifique sud-ouest)". PhD thesis. Toulouse 3, 2012 (cit. on p. 21).

[GH+96] Zoubin Ghahramani, Geoffrey E Hinton, et al. *The EM algorithm for mixtures of factor analyzers*. Tech. rep. Technical Report CRG-TR-96-1, University of Toronto, 1996 (cit. on p. 31).

[Gie+14] Sarah LC Giering, Richard Sanders, Richard S Lampitt, Thomas R Anderson, Christian Tamburini, Mehdi Boutrif, Mikhail V Zubkov, Chris M Marsay, Stephanie A Henson, Kevin Saw, et al. "Reconciliation of the carbon budget in the ocean's twilight zone". In: *Nature* 507.7493 (2014), pp. 480–483 (cit. on p. 221).

[Gli+16] Patricia M Glibert, Frances P Wilkerson, Richard C Dugdale, John A Raven, Christopher L Dupont, Peter R Leavitt, Alexander E Parker, JoAnn M Burkholder, and Todd M Kana. "Pluses and minuses of ammonium and nitrate uptake and assimilation by phytoplankton and implications for productivity and community composition, with emphasis on nitrogen-enriched conditions". In: *Limnology and Oceanography* 61.1 (2016), pp. 165–197 (cit. on pp. 70, 87).

[Gor+10] Gaby Gorsky, Mark D Ohman, Marc Picheral, Stéphane Gasparini, Lars Stemmann, Jean-Baptiste Romagnan, Alison Cawood, Stéphane Pesant, Carmen García-Comas, and Franck Prejger. "Digital zooplankton image analysis using the ZooScan integrated system". In: *Journal of plankton research* 32.3 (2010), pp. 285–303 (cit. on p. 226).

[Gro+11] The MerMex Group, X Durrieu de Madron, C Guieu, R Sempéré, P Conan, D Cossa, F D'Ortenzio, C Estournel, F Gazeau, C Rabouille, et al. "Marine ecosystems' responses to climatic and anthropogenic forcings in the Mediterranean". In: *Progress in Oceanography* 91.2 (2011), pp. 97–166 (cit. on pp. 27, 227).

[Gué13] Yann Guédon. "Exploring the latent segmentation space for the assessment of multiple change-point models". In: *Computational Statistics* 28.6 (2013), pp. 2641–2678 (cit. on p. 90).

[HC07] Zaid Harchaoui and Olivier Cappé. "Retrospective mutiple change-point estimation with kernels". In: *2007 IEEE/SP 14th Workshop on Statistical Signal Processing*. IEEE. 2007, pp. 768–772 (cit. on p. 89).

[Har76] Harry H Harman. *Modern factor analysis*. University of Chicago press, 1976 (cit. on p. 31).

[Hed57] Joel W Hedgpeth. "Classification of marine environments". In: *Treatise on marine ecology and paleoecology* 50 (1957), pp. 17–28 (cit. on p. 20).

[Hen+11] Stephanie A Henson, Richard Sanders, Esben Madsen, Paul J Morris, Frédéric Le Moigne, and Graham D Quartly. "A reduced estimate of the strength of the ocean's biological carbon pump". In: *Geophysical Research Letters* 38.4 (2011) (cit. on p. 17).

[Her+06] Clara M Hereu, Bertha E Lavaniegos, Gilberto Gaxiola-Castro, and Mark D Ohman. "Composition and potential grazing impact of salp assemblages off Baja California during the 1997–1999 El Niño and La Niña". In: *Marine Ecology Progress Series* 318 (2006), pp. 123–140 (cit. on p. 226).

[Hua97] Zhexue Huang. "Clustering large data sets with mixed numeric and categorical values". In: *Proceedings of the 1st pacific-asia conference on knowledge discovery and data mining,(PAKDD)*. Singapore. 1997, pp. 21–34 (cit. on p. 29).

[Hua98] Zhexue Huang. "Extensions to the k-means algorithm for clustering large data sets with categorical values". In: *Data mining and knowledge discovery* 2.3 (1998), pp. 283–304 (cit. on p. 29).

[Hun+20] Kristen R Hunter-Cevera, Michael G Neubert, Robert J Olson, Alexi Shalapyonok, Andrew R Solow, and Heidi M Sosik. "Seasons of Syn". In: *Limnology and oceanography* 65.5 (2020), pp. 1085–1102 (cit. on p. 26).

[Hut57] G Evelyn Hutchinson. "Concluding remarks". In: *Cold Spring Harbor symposia on quantitative biology*. Vol. 22. Cold Spring Harbor Laboratory Press. 1957, pp. 415–427 (cit. on pp. 22, 67).

[IS15] Sergey Ioffe and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift". In: *International conference on machine learning*. PMLR. 2015, pp. 448–456 (cit. on p. 160).

[Jac27] JP Jacobsen. "Eine graphische Methode zur Bestimmungd des Vermischungskoeffizienten im Meer". In: *Gerl. Beitr. z. Geophysik* 16 (1927) (cit. on p. 21).

[Jac+02] Stéphan Jacquet, Mikal Heldal, Debora Iglesias-Rodriguez, Aud Larsen, William Wilson, and Gunnar Bratbak. "Flow cytometric analysis of an Emiliana huxleyi bloom terminated by viral infection". In: *Aquatic Microbial Ecology* 27.2 (2002), pp. 111–124 (cit. on p. 26).

[JCG13] Justyna Jońca, M Comtat, and V Garçon. "In situ phosphate monitoring in seawater: today and tomorrow". In: *Smart Sensors for Real-Time Water Quality Monitoring*. Springer, 2013, pp. 25–44 (cit. on p. 226).

[KB14] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014) (cit. on p. 160).

[Koh90] Teuvo Kohonen. "The self-organizing map". In: *Proceedings of the IEEE* 78.9 (1990), pp. 1464–1480 (cit. on p. 30).

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25 (2012) (cit. on p. 161).

[KPS09] Eun Young Kwon, François Primeau, and Jorge L Sarmiento. "The impact of remineralization depth on the air–sea carbon balance". In: *Nature Geoscience* 2.9 (2009), pp. 630–635 (cit. on p. 21).

[LBA14] Rémi Lajugie, Francis Bach, and Sylvain Arlot. "Large-margin metric learning for constrained partitioning problems". In: *International Conference on Machine Learning*. PMLR. 2014, pp. 297–305 (cit. on p. 89).

[Lav99] Marc Lavielle. "Detection of multiple changes in a sequence of dependent variables". In: *Stochastic Processes and their applications* 83.1 (1999), pp. 79–102 (cit. on p. 89).

[Law+00] Edward A Laws, Paul G Falkowski, Walker O Smith Jr, Hugh Ducklow, and James J McCarthy. "Temperature effects on export production in the open ocean". In: *Global biogeochemical cycles* 14.4 (2000), pp. 1231–1246 (cit. on p. 17).

[LHC+05] Corinne Le Quere, Sandy P Harrison, I Colin Prentice, et al. "Ecosystem dynamics based on plankton functional types for global ocean biogeochemistry models". In: *Global Change Biology* 11.11 (2005), pp. 2016–2040 (cit. on p. 24).

[LeC+89] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. "Backpropagation applied to handwritten zip code recognition". In: *Neural computation* 1.4 (1989), pp. 541–551 (cit. on p. 161).

[Lee00] Sauchi Stephen Lee. "Noisy replication in skewed binary classification". English. In: *Computational Statistics and Data Analysis* 34.2 (2000), pp. 165–191 (cit. on p. 223).

[Lév+12] Marina Lévy, Raffaele Ferrari, Peter JS Franks, Adrian P Martin, and Pascal Rivière. "Bringing physics to life at the submesoscale". In: *Geophysical Research Letters* 39.14 (2012) (cit. on p. 23).

[LAR18] Sangdi Lin, Bahareh Azarnoush, and George C Runger. "Crafter: a tree-ensemble clustering algorithm for static datasets with mixed attributes and high dimensionality". In: *IEEE Transactions on Knowledge and Data Engineering* 30.9 (2018), pp. 1686–1696 (cit. on p. 30).

[Lon95] Alan R Longhurst. "Seasonal cycles of pelagic production and consumption". In: *Progress in oceanography* 36.2 (1995), pp. 77–167 (cit. on pp. 18, 20).

[Lon10] Alan R Longhurst. *Ecological geography of the sea*. Elsevier, 2010 (cit. on p. 20).

[LLC12] Alexandre Lung-Yut-Fong, Céline Lévy-Leduc, and Olivier Cappé. "Distributed detection/localization of change-points in high-dimensional network traffic data". In: *Statistics and Computing* 22.2 (2012), pp. 485–496 (cit. on p. 90).

[LLC15] Alexandre Lung-Yut-Fong, Céline Lévy-Leduc, and Olivier Cappé. "Homogeneity and change-point detection tests for multivariate data using rank statistics". In: *Journal de la Société Française de Statistique* 156.4 (2015), pp. 133–162 (cit. on p. 89).

[MDA15] Dougal Maclaurin, David Duvenaud, and Ryan P Adams. "Autograd: Effortless gradients in numpy". In: *ICML 2015 AutoML Workshop*. Vol. 238. 2015, p. 5 (cit. on p. 220).

[Mar+20] R Marchant, M Tetard, A Pratiwi, M Adebayo, and T de Garidel-Thoron. "Automated analysis of foraminifera fossil records by image classification using a convolutional neural network". In: *Journal of Micropalaeontology* 39.2 (2020), pp. 183–202. DOI: 10.5194/jm-39-183-2020. URL: https://jm.copernicus.org/articles/39/183/2020/ (cit. on p. 222).

[Mar+18] Pierre Marrec, Gŕald Grégori, Andrea. M. Doglioli, Mathilde Dugenne, Alice Della Penna, Nagib Bhairy, Thierry Cariou, Sandra Hélias Nunige, Soumaya Lahbib, Gilles Rougier, Thomas Wagener, and Melilotus Thyssen. "Coupling physics and biogeochemistry thanks to high-resolution observations of the phytoplankton community structure in the northwestern Mediterranean Sea". In: *Biogeosciences* 15.5 (2018), pp. 1579–1606. DOI: 10.5194/bg-15-1579-2018. URL: https://bg.copernicus.org/articles/15/1579/2018/ (cit. on p. 127).

[MG16] Damien McParland and Isobel Claire Gormley. "Model based clustering for mixed data: clustMD". In: *Advances in Data Analysis and Classification* 10.2 (2016), pp. 155–169 (cit. on p. 30).

[ML00] Susanne Menden-Deuer and Evelyn J Lessard. "Carbon to volume relationships for dinoflagellates, diatoms, and other protist plankton". In: *Limnology and oceanography* 45.3 (2000), pp. 569–579 (cit. on p. 224).

[MBS+18] Patricia Miloslavich, Nicholas J Bax, Samantha E Simmons, et al. "Essential ocean variables for global sustained observations of biodiversity and ecosystem changes". In: *Global change biology* 24.6 (2018), pp. 2416–2433 (cit. on p. 225).

[Miz+13] Carolina Megumi Mizuno, Francisco Rodriguez-Valera, Nikole E Kimes, and Rohit Ghai. "Expanding the marine virosphere using metagenomics". In: *PLoS genetics* 9.12 (2013), e1003987 (cit. on p. 226).

[Mou03] Irini Moustaki. "A general class of latent variable models for ordinal manifest variables with covariate effects on the manifest and latent variables". In: *British Journal of Mathematical and Statistical Psychology* 56.2 (2003), pp. 337–357 (cit. on pp. 23, 30, 32).

[MK00] Irini Moustaki and Martin Knott. "Generalized latent trait models". In: *Psychometrika* 65.3 (2000), pp. 391–411 (cit. on pp. 23, 30, 32).

[NJW01] Andrew Ng, Michael Jordan, and Yair Weiss. "On spectral clustering: Analysis and an algorithm". In: *Advances in neural information processing systems* 14 (2001) (cit. on p. 30).

[NPZ14] Gerald R North, John A Pyle, and Fuqing Zhang. *Encyclopedia of atmospheric sciences*. Vol. 1. Elsevier, 2014 (cit. on p. 89).

[Ols+83] Robert J Olson, Sheila L Frankel, Sallie W Chisholm, and Howard M Shapiro. "An inexpensive flow cytometer for the analysis of fluorescence signals in phytoplankton: chlorophyll and DNA distributions". In: *Journal of Experimental Marine Biology and Ecology* 68.2 (1983), pp. 129–144 (cit. on p. 23).

[Ote+18] Jose Luis Otero-Ferrer, Pedro Cermeño, Antonio Bode, Bieito Fernández-Castro, Josep M Gasol, Xosé Anxelu G Morán, Emilio Marañon, Victor Moreira-Coello, Marta M Varela, Marina Villamaña, et al. "Factors controlling the community structure of picoplankton in contrasting marine environments". In: *Biogeosciences* 15.20 (2018), pp. 6199–6220 (cit. on pp. 70, 87).

[Pag55] ES Page. "A test for a change in a parameter occurring at an unknown point". In: *Biometrika* 42.3/4 (1955), pp. 523–527 (cit. on p. 89).

[PVK20] Francisco Pastor, Jose Antonio Valiente, and Samiro Khodayar. "A warming Mediterranean: 38 years of increasing sea surface temperature". In: *Remote sensing* 12.17 (2020), p. 2687 (cit. on p. 85).

[PO83] G Philip and BS Ottaway. "Mixed data cluster analysis: an illustration using Cypriot hooked-tang weapons". In: *Archaeometry* 25.2 (1983), pp. 119–133 (cit. on p. 29).

[Pol+08] Ricardo Poli, William B. Langdon, Nicholas F. McPhee, and John R. Koza. "A field guide to genetic programming". In: *Published via http://lulu. com and freely http://www. gp-field-guide. org. uk (with contributions by JR Koza). GPBiB* (2008) (cit. on p. 219).

[Pul+17] Silvia Pulina, Cecilia Teodora Satta, Bachisio Mario Padedda, Anna Maria Bazzoni, Nicola Sechi, and Antonella Lugliè. "Picophytoplankton seasonal dynamics and interactions with environmental variables in three Mediterranean coastal lagoons". In: *Estuaries and Coasts* 40.2 (2017), pp. 469–478 (cit. on p. 70).

[QP07] Zhongjun Qu and Pierre Perron. "Estimating and testing structural changes in multivariate regressions". In: *Econometrica* 75.2 (2007), pp. 459–502 (cit. on p. 89).

[Rab89] Lawrence R Rabiner. "A tutorial on hidden Markov models and selected applications in speech recognition". In: *Proceedings of the IEEE* 77.2 (1989), pp. 257–286 (cit. on p. 89).

[REY13] Gabriel REYGONDEAU. “L’océan mondial peut-il être divisé en grandes unités écologiques? L’approche biogéographique de Longhurst”. In: *Fiches de synthèse de l’Institut océanographique - Fondation Albert Ier, Prince de Monaco* (2013) (cit. on p. 19).

[Rey+12] Gabriel Reygondeau, Olivier Maury, Gregory Beaugrand, Jean Marc Fromentin, Alain Fonteneau, and Philippe Cury. “Biogeography of tuna and billfish communities”. In: *Journal of Biogeography* 39.1 (2012), pp. 114–129 (cit. on p. 20).

[RMW14] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. “Stochastic backpropagation and approximate inference in deep generative models”. In: *International conference on machine learning*. PMLR. 2014, pp. 1278–1286 (cit. on p. 160).

[Rib+15] Francois Ribalet, Jarred Swalwell, Sophie Clayton, Valeria Jiménez, Sebastian Sudek, Yajuan Lin, Zackary I Johnson, Alexandra Z Worden, and E Virginia Armbrust. “Light-driven synchrony of Prochlorococcus growth and mortality in the subtropical Pacific gyre”. In: *Proceedings of the National Academy of Sciences* 112.26 (2015), pp. 8008–8012 (cit. on p. 26).

[Rom+16] Jean Baptiste Romagnan, Lama Aldamman, Stéphane Gasparini, Paul Nival, Anaıs Aubert, Jean Louis Jamet, and Lars Stemmann. “High frequency mesozooplankton monitoring: Can imaging systems and automated sample analysis help us describe and interpret changes in zooplankton community composition and size structure—An example from a coastal site”. In: *Journal of Marine Systems* 162 (2016), pp. 18–28 (cit. on p. 226).

[RS07] Oliver N Ross and Jonathan Sharples. “Phytoplankton motility and the competition for nutrients in the thermocline”. In: *Marine Ecology Progress Series* 347 (2007), pp. 21–38 (cit. on p. 18).

[Rou+14] Simon Roux, Alyse K Hawley, Monica Torres Beltran, Melanie Scofield, Patrick Schwientek, Ramunas Stepanauskas, Tanja Woyke, Steven J Hallam, and Matthew B Sullivan. “Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by single-cell-and metagenomics”. In: *elife* 3 (2014), e03125 (cit. on p. 226).

[Sak17] Abdulla Sakallı. “Sea surface temperature change in the Mediterranean Sea under climate change: a linear model for simulation of the sea surface temperature up to 2100”. In: (2017) (cit. on p. 85).

[Sak+17] Carole M Sakamoto, Kenneth S Johnson, Luke J Coletti, Tanya L Maurer, Gene Massion, J Timothy Pennington, Joshua N Plant, Hans W Jannasch, and Francisco P Chavez. “Hourly in situ nitrate on a coastal mooring: a 15-year record and insights into new production”. In: *Oceanography* 30.4 (2017), pp. 114–127 (cit. on p. 226).

[Sel+20] Margot Selosse, Claire Gormley, Julien Jacques, and Christophe Biernacki. "A bumpy journey: exploring deep Gaussian mixture models". In: *"I Can't Believe It's Not Better!"NeurIPS 2020 workshop*. 2020 (cit. on p. 219).

[Sil+17] Tina Silovic, Gérald Grégori, Mathilde Dugenne, Melilotus Thyssen, François Calendreau, Thibaut Cossart, Harrie Kools, George Dubelaar, and Michel Denis. "A new automated flow cytometer for high frequency in situ characterisation of heterotrophic microorganisms and their dynamics in aquatic ecosystems". In: (2017) (cit. on p. 226).

[SM12] RA Simons and R Mendelssohn. "ERDDAP-a brokering data server for gridded and tabular datasets". In: *AGU Fall Meeting Abstracts*. Vol. 2012. 2012, IN21B–1473 (cit. on p. 223).

[Som54] Mary Somerville. *Physical geography*. Blanchard and Lea, 1854 (cit. on p. 18).

[Sos+03] Heidi M Sosik, Robert J Olson, Michael G Neubert, Alexi Shalapyonok, and Andrew R Solow. "Growth rates of coastal phytoplankton from time-series measurements with a submersible flow cytometer". In: *Limnology and Oceanography* 48.5 (2003), pp. 1756–1765 (cit. on p. 26).

[Sri+14] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. "Dropout: a simple way to prevent neural networks from overfitting". In: *The journal of machine learning research* 15.1 (2014), pp. 1929–1958 (cit. on p. 160).

[Str22] Eliza Strickland. "Andrew Ng, AI Minimalist: The Machine-Learning Pioneer Says Small is the New Big". In: *IEEE Spectrum* 59.4 (2022), pp. 22–50 (cit. on p. 223).

[SB18] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018 (cit. on p. 222).

[Tad+11] Arbel D Tadmor, Elizabeth A Ottesen, Jared R Leadbetter, and Rob Phillips. "Probing individual environmental bacteria for viruses by using microfluidic digital PCR". In: *Science* 333.6038 (2011), pp. 58–62 (cit. on p. 226).

[TM21] Hossein Talebi and Peyman Milanfar. "Learning to resize images for computer vision tasks". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 497–506 (cit. on p. 223).

[Tal11] Lynne D Talley. *Descriptive physical oceanography: an introduction*. Academic press, 2011 (cit. on p. 18).

[Thy+21] Melilotus Thyssen, Robin Fuchs, Véronique Créach, Luis Felipe Artigas, Gérald Grégori, Pierre Marrec, Mathilde Dugenne, Machteld Rijkeboer, Marie Latimier, Louchart Arnaud, et al. "Standard vocabulary, consensual functional groups and automated classification for phytoplankton high throughput datasets using automated flow cytometry". In: *ASLO 2021*. 2021 (cit. on pp. 24, 221).

[Thy+08] Melilotus Thyssen, Delphine Mathieu, Nicole Garcia, and Michel Denis. "Short-term variation of phytoplankton assemblages in Mediterranean coastal waters recorded with an automated submerged flow cytometer". In: *Journal of Plankton Research* 30.9 (2008), pp. 1027–1040 (cit. on p. 26).

[Tom81] Matthias Jr Tomczak. "A multi-parameter extension of temperature/salinity diagram techniques for the analysis of non-isopycnal mixing". In: *Progress in Oceanography* 10.3 (1981), pp. 147–171 (cit. on pp. 21, 89).

[TL89] Matthias Jr Tomczak and Daniel GB Large. "Optimum multiparameter analysis of mixing in the thermocline of the eastern Indian Ocean". In: *Journal of Geophysical Research: Oceans* 94.C11 (1989), pp. 16141–16149 (cit. on pp. 21, 89).

[TOV20] Charles Truong, Laurent Oudre, and Nicolas Vayatis. "Selective review of offline change point detection methods". In: *Signal Processing* 167 (2020), p. 107299 (cit. on pp. 21, 89).

[Vas+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017) (cit. on p. 222).

[Ver+92] Peter G Verity, Charles Y Robertson, Craig R Tronzo, Melinda G Andrews, James R Nelson, and Michael E Sieracki. "Relationships between cell volume and the carbon and nitrogen content of marine photosynthetic nanoplankton". In: *Limnology and Oceanography* 37.7 (1992), pp. 1434–1446 (cit. on p. 224).

[VM19] Cinzia Viroli and Geoffrey J McLachlan. "Deep gaussian mixture models". In: *Statistics and Computing* 29.1 (2019), pp. 43–51 (cit. on pp. 23, 30, 31).

[Wun96] Carl Wunsch. *The ocean circulation inverse problem*. Cambridge University Press, 1996 (cit. on p. 21).

[Wun06] Carl Wunsch. *Discrete inverse and state estimation problems: with geophysical fluid applications*. Cambridge University Press, 2006 (cit. on p. 21).

[YG86] CHARLES S Yentsch and JEAN C Garside. "Patterns of phytoplankton abundance and biogeography". In: *UNESCO Technical Papers in Marine Science* 49 (1986), pp. 278–284 (cit. on p. 18).

[Yen+83] CHARLES S Yentsch, PK Horan, K Muirhead, Q Dortch, EM Haugen, L Legendre, LS Murphy, D Phinney, SA Pomponi, RW Spinrad, et al. "Flow cytometry and sorting: a powerful technique with potential applications in aquatic sciences". In: *Limnol. Oceanogr* 28 (1983), pp. 1275–1280 (cit. on p. 23).

[Zou+14] Changliang Zou, Guosheng Yin, Long Feng, and Zhaojun Wang. "Nonparametric maximum likelihood approach to multiple change-point problems". In: *The Annals of Statistics* 42.3 (2014), pp. 970–1002 (cit. on p. 89).

Appendix

A. MDGMM: Supplementary Material

8 Supplementary Material

This preprint has not undergone any post-submission improvements or corrections.
The Version of Record of this article is published in Advances in Data Analysis and Classification, and is available online at https://doi.org/10.1007/s11634-021-00466-3

8.1 Expression of the expected Log-Likelilhood

The expected log-likelihood can be expressed as:

$$
\begin{aligned}
&\mathbb{E}_{z^C,z^D,z^{(L_0+1:)},s^C,s^D,s^{(L_0+1:)}|y^C,y^D,\hat{\Theta}_C,\hat{\Theta}_D,\hat{\Theta}_{L_0+1:}}[\log L(y^C,y^D,z^C,z^D,z^{(L_0+1:)},s^C,s^D,s^{(L_0+1:)}|\Theta_C,\Theta_D,\Theta_{L_0+1:})]\\
&=\mathbb{E}_{z^{(1)D},s^D,s^{(L_0+1:)D}|y^D,\hat{\Theta}_D,\hat{\Theta}_{L_0+1:}}[\log L(y^D|z^{(1)D},s^D,s^{(L_0+1:)},\Theta_D,\Theta_{L_0+1:})]\\
&+\mathbb{E}_{z^{(1)C},s^C,s^{(L_0+1:)C}|y^C,\hat{\Theta}_C,\hat{\Theta}_{L_0+1:}}[\log L(y^C|z^{(1)C},s^C,s^{(L_0+1:)},\Theta_C,\Theta_{L_0+1:})]\\
&+\sum_{h\in\{C,D\}}\sum_{l=1}^{L_0}\mathbb{E}_{z^{(l)h},z^{(l+1)h},s^h,s^{(L_0+1:)}|y^h,\hat{\Theta}_h,\hat{\Theta}_{L_0+1:}}[\log L(z^{(l)h}|z^{(l+1)h},s^h,s^{(L_0+1:)},\Theta_h,\Theta_{(L_0+1:)})]\\
&+\sum_{l=L_0+1}^{L-1}\mathbb{E}_{z^{(l)},z^{(l+1)},s^C,s^D,s^{(L_0+1:)}|y^C,y^D,\hat{\Theta}_C,\hat{\Theta}_D,\hat{\Theta}_{L_0+1:}}[\log L(z^{(l)}|z^{(l+1)},s^C,s^D,s^{(L_0+1:)},\Theta_C,\Theta_D,\Theta_{L_0+1:}]\\
&+\mathbb{E}_{z^{(L)}|y^C,y^D,\hat{\Theta}_C,\hat{\Theta}_D,\hat{\Theta}_{L_0+1:}}[\log L(z^{(L)}|\Theta_C,\Theta_D,\Theta_{L_0+1:}))]\\
&+\mathbb{E}_{s^C,s^D,s^{(L_0+1:)}|y^C,y^D,\hat{\Theta}_C,\hat{\Theta}_D,\hat{\Theta}_{L_0+1:}}[\log L(s^C,s^D,s^{(L_0+1:)}|\Theta_C,\Theta_D,\Theta_{L_0+1:})],
\end{aligned}
\tag{6}
$$

with a slight abuse of notation in the double sum as we have set $z^{(L_0+1)} = z^{(L_0+1)C} = z^{(L_0+1)D}$. $\hat{\Theta}_h$ are the provisional estimate of Θ_h through the iterations of the algorithm.

8.2 GLLVM embedding layer mathematical derivations

8.2.1 E step for the GLLVM embedding layer

We consider the conditional density

$$
f(z^{(1)D}|y^D,\hat{\Theta}_D,\hat{\Theta}_{L_0+1:}) = \sum_{s'} f(z^{(1)D}|y^D,s',\hat{\Theta}_D,\hat{\Theta}_{L_0+1:}) f(s^{(1D:L)}=s'|y^D,\hat{\Theta}_D,\hat{\Theta}_{L_0+1:}).
\tag{7}
$$

The Bayes rule for the first term gives :

$$
f(z^{(1)D}|y^D,s',\hat{\Theta}_D,\hat{\Theta}_{L_0+1:}) = \frac{f(z^{(1)D}|s',\hat{\Theta}_D,\hat{\Theta}_{L_0+1:}) f(y^D|z^{(1)D},\hat{\Theta}_D,\hat{\Theta}_{L_0+1:})}{f(y^D|s',\hat{\Theta}_D,\hat{\Theta}_{L_0+1:})},
\tag{8}
$$

and we have

$$
(z^{(1)D}|s',\hat{\Theta}_D,\hat{\Theta}_{L_0+1:}) \sim N(\mu_{s'}^{(1D:L)},\Sigma_{s'}^{(1D:L)}),
$$

where the mean and covariance parameters $(\mu_{s'}^{(1D:L)}, \Sigma_{s'}^{(1D:L)})$ are detailed in Section 8.3.1. Moreover, $f(y^D|z^{(1)D}, \hat{\Theta}_D, \hat{\Theta}_{L_0+1:})$ belongs to an exponential family. Finally, $f(y^D|s', \hat{\Theta}_D, \hat{\Theta}_{L_0+1:})$ has to be numerically approximated. This is here performed by Monte Carlo estimation by simulating $M^{(1)}$ copies of $z^{(1)D}$ as follows

$$\begin{aligned} f(y^D|s', \hat{\Theta}_D, \hat{\Theta}_{L_0+1:}) &= \int_{z^{(1)D}} f(y^D|z^{(1)D}, \hat{\Theta}_D, \hat{\Theta}_{L_0+1:}) f(z^{(1)D}|s', \hat{\Theta}_D, \hat{\Theta}_{L_0+1:}) dz^{(1)D} \\ &\approx \sum_{m=1}^{M^{(1)}} f(y^D|z_m^{(1)D}, \hat{\Theta}_D, \hat{\Theta}_{L_0+1:}, \hat{\Theta}) f(z_m^{(1)D}|s', \hat{\Theta}_D, \hat{\Theta}_{L_0+1:}). \end{aligned}$$

The second term of (7) can be written as a posterior density:

$$f(s^{(1D:L)} = s'|y^D, \hat{\Theta}_D, \hat{\Theta}_{L_0+1:}) = \frac{f(s^{(1D:L)} = s'|\hat{\Theta}_D, \hat{\Theta}_{L_0+1:}) f(y^D|s', \hat{\Theta}_D, \hat{\Theta}_{L_0+1:})}{\sum_{s''} f(s^{(1D:L)} = s''|\hat{\Theta}_D, \hat{\Theta}_{L_0+1:}) f(y^D|s^{(1D:L)} = s'', \hat{\Theta}_D, \hat{\Theta}_{L_0+1:})}, \tag{9}$$

and we have $(s^{(1D:L)}|\hat{\Theta}_D, \hat{\Theta}_{L_0+1:}) \sim M(\pi_s^{(1D:L)})$ a multinomial distribution with parameters $\pi_s^{(1D:L)}$ which is the probability of a full path through the network starting from the discrete head. The density $f(y^D|s', \hat{\Theta}_D, \hat{\Theta}_{L_0+1:})$ is once again approximated by Monte Carlo.

8.2.2 M step for the GLLVM embedding layer

To maximize $\mathbb{E}_{z^{(1)D}|y^D, \hat{\Theta}_D, \hat{\Theta}_{L_0+1:}}[\log L(y^D|z^{(1)D}, \Theta_D, \hat{\Theta}_{L_0+1:})]$, we use optimisation methods. All methods belong to the Python scipy.optimize package (Virtanen et al., 2020). For binary, count and categorical variables, the optimisation program is unconstrained and the BFGS (Fletcher, 2013) algorithm is used. Concerning ordinal variables, the optimisation program is constrained as the intercept coefficients have to be ordered. The method used is a trust-region algorithm (Conn et al., 2000). All the gradients are computed by automatic differentiation using the autograd package (Maclaurin et al., 2015), which significantly speeds up the optimization process compared to hand-coded gradients.

8.3 DGMM layers mathematical derivations

8.3.1 E step for the DGMM layers

Recall that we have:

$$\begin{aligned} f(z^{(\ell)}, z^{(\ell+1)}, s|y, \hat{\Theta}) &= f(z^{(\ell)}, s|y, \hat{\Theta}) f(z^{(\ell+1)}|z^{(\ell)}, s, y, \hat{\Theta}) \\ &= f(z^{(\ell)}|y, s, \hat{\Theta}) f(s|y, \hat{\Theta}) f(z^{(\ell+1)}|z^{(\ell)}, s, \hat{\Theta}). \end{aligned} \tag{10}$$

The first term can be rewritten and approximated as follows:

$$\begin{aligned} f(z^{(\ell)}|y, s, \hat{\Theta}) &= \int_{z^{(\ell-1)}} f(z^{(\ell)}|z^{(\ell-1)}, s, \hat{\Theta}) f(z^{(\ell-1)}|y, s, \hat{\Theta}) dz^{(\ell-1)} \\ &\approx \sum_{m=1}^{M^{(\ell-1)}} f(z^{(\ell)}|z_m^{(\ell-1)}, s, \hat{\Theta}) f(z_m^{(\ell-1)}|y, s, \hat{\Theta}). \end{aligned} \tag{11}$$

This expression is hence calculable in a recurrent manner $\forall \ell \in [2, L_0]$, starting with $f(z^{(1)}|y, s', \hat{\Theta})$ given by (8). The second term of (10) can be expressed as in (9), and the last term is given by the Bayes rule:

$$f(z^{(\ell+1)}|z^{(\ell)}, s, \hat{\Theta}) = \frac{f(z^{(\ell)}|z^{(\ell+1)}, s, \hat{\Theta}) f(z^{(\ell+1)}|s, \hat{\Theta})}{f(z^{(\ell)}|s, \hat{\Theta})}. \tag{12}$$

Clearly, the denominator does not depend on $z^{(\ell+1)}$ and is hence considered as a normalisation constant. Besides, we have that $f(z^{(\ell)}|z^{(\ell+1)}, s, \hat{\Theta}) = N(\eta_{k_\ell}^{(\ell)} + \Lambda_{k_\ell}^{(\ell)} z^{(\ell+1)}, \Psi_{k_\ell}^{(\ell)})$. Finally, by construction of the DGMM, we have

$$f(z^{(\ell+1)}|s, \hat{\Theta}) = f(z^{(\ell+1)}|s^{(l+1:L)}, \hat{\Theta}) = N(\mu_{s^{(:k_{\ell+1}:)}}^{(\ell+1)}, \Sigma_{s^{(:k_{\ell+1}:)}}^{(\ell+1)}). \tag{13}$$

It follows that (12) is also a Gaussian distribution of parameters $(\rho_{k_{\ell+1}}^{(\ell+1)}, \xi_{k_{\ell+1}}^{(\ell+1)})$.

The formulas of the Gaussian parameters are obtain as follows: the DGMM can be written at each layer as a regular Gaussian Mixture with a number of components equal to the number of paths starting from that layer. The Gaussian mean and covariance matrix of each path starting from the k_ℓ component of layer ℓ can be computed in the following way:

$$\mu_{\tilde{s}^{(k_\ell:)}}^{(\ell)} = \eta_{k_\ell}^{(\ell+1)} + \sum_{j=\ell+1}^{L} \Big(\prod_{m=\ell}^{j-1} \Lambda_{k'_m}^{(m)} \Big) \eta_{k'_j}^{(j)},$$

and

$$\Sigma^{(\ell)}_{\tilde{s}^{(k_\ell:)}} = \Psi^{(\ell)}_{k_\ell} + \sum_{j=\ell+1}^{L} \Big(\prod_{m=\ell}^{j-1} \Lambda^{(m)}_{k'_m} \Big) \big(\Psi^{(j)}_{k'_j} + \Lambda^{(j)}_{k'_\ell} \Lambda^{(j)T}_{k'_j} \big) \Big(\prod_{m=\ell}^{j-1} \Lambda^{(m)}_{k'_m} \Big)^T.$$

In addition, we have that the random variable $(z^{(\ell+1)}|z^{(\ell)}, \tilde{s}, \hat{\Theta})$ also follows a multivariate Gaussian distribution with mean and covariance parameters $(\rho^{(\ell+1)}_{k_{\ell+1}}, \xi^{(\ell+1)}_{k_{\ell+1}})$:

$$\rho^{(\ell+1)}_{k_{\ell+1}} = \xi^{(\ell+1)}_{k_{\ell+1}} \Big(\Lambda^{(\ell+1)T}_{k_{\ell+1}} (\Psi^{(\ell+1)}_{k_{\ell+1}})^{-1} (z^{(\ell)} - \eta^{(\ell+1)}_{k_{\ell+1}}) + \Sigma^{(\ell+1)}_{\tilde{s}^{(:k_{\ell+1}:)}} \mu^{(\ell+1)}_{\tilde{s}^{(:k_{\ell+1}:)}} \Big),$$

and

$$\xi^{(\ell+1)}_{k_{\ell+1}} = \Big(\Sigma^{(\ell+1)}_{\tilde{s}^{(:k_{\ell+1}:)}} + \Lambda^{(\ell+1)T}_{k_{\ell+1}} (\Psi^{(\ell+1)}_{k_{\ell+1}})^{-1} \Lambda^{(\ell+1)}_{k_{\ell+1}} \Big)^{-1}.$$

8.3.2 M Step for the DGMM layers

We now turn on to the log-likelihood expression and give the estimators of the ℓ-th DGMM layer parameters $\forall \ell \in [1, L_h]$, $\forall h \in \{C, D\}$. In this section the h superscripts are omitted for simplicity of notation.

$$\log L(z^{(\ell)}_i | z^{(\ell+1)}_i, s_i, \Theta) =$$
$$-\frac{1}{2} \left[\log(2\pi) + \log\det(\Psi^{(\ell)}_{k_\ell}) + \Big(z^{(\ell)}_i - (\eta^{(\ell)}_{k_\ell} + \Lambda^{(\ell)}_{k_\ell} z^{(\ell+1)}_i) \Big)^T \Psi^{(\ell)-1}_{k_\ell} \Big(z^{(\ell)}_i - (\eta^{(\ell)}_{k_\ell} + \Lambda^{(\ell)}_{k_\ell} z^{(\ell+1)}_i) \Big) \right].$$

The derivatives of this quantity with respect to $\eta^{(\ell)}_{k_\ell}, \Lambda^{(\ell)}_{k_\ell}, \Psi^{(\ell)}_{k_\ell}$ are given by

$$\begin{cases} \frac{\partial \log L(z^{(\ell)}_i | z^{(\ell+1)}_i, s_i, \Theta)}{\partial \eta^{(\ell)}_{k_\ell}} = \Psi^{(\ell)-1}_{k_\ell} \Big(z^{(\ell)}_i - (\eta^{(\ell)}_{k_\ell} + \Lambda^{(\ell)}_{k_\ell} z^{(\ell+1)}_i) \Big) \\ \frac{\partial \log L(z^{(\ell)}_i | z^{(\ell+1)}_i, s_i, \Theta)}{\partial \Lambda^{(\ell)}_{k_\ell}} = \Psi^{(\ell)-1}_{k_\ell} \Big(z^{(\ell)}_i - (\eta^{(\ell)}_{k_\ell} + \Lambda^{(\ell)}_{k_\ell} z^{(\ell+1)}_i) \Big) z^{(\ell+1)T}_i \\ \frac{\partial \log L(z^{(\ell)}_i | z^{(\ell+1)}_i, s_i, \Theta)}{\partial \Psi^{(\ell)}_{k_\ell}} = -\frac{1}{2} \Psi^{(\ell)-1}_{k_\ell} \left[I_{r_1} - \Big(z^{(\ell)}_i - (\eta^{(\ell)}_{k_\ell} + \Lambda^{(\ell)}_{k_\ell} z^{(\ell+1)}_i) \Big) \Big(z^{(\ell)}_i - (\eta^{(\ell)}_{k_\ell} + \Lambda^{(\ell)}_{k_\ell} z^{(\ell+1)}_i) \Big)^T \Psi^{(\ell)-1}_{k_\ell} \right]. \end{cases}$$

Taking the expectation of the derivative with respect to $\eta^{(\ell)}_{k_\ell}$ and equalizing it to zero, it

follows that:

$$\mathbb{E}_{z^{(\ell)},z^{(\ell+1)},s|y,\hat{\Theta}}\left[\frac{\partial \log L(z^{(\ell)}|z^{(\ell+1)},s,\Theta)}{\partial \eta_{k_\ell}^{(\ell)}}\right]=0$$

$$\iff \Psi_{k_\ell}^{(\ell)-1}\sum_{i=1}^{n}\mathbb{E}_{z_i^{(\ell)},z_i^{(\ell+1)},s_i|y_i,\hat{\Theta}}\left[z_i^{(\ell)}-(\eta_{k_\ell}^{(\ell)}+\Lambda_{k_\ell}^{(\ell)}z_i^{(\ell+1)})\right]=0$$

$$\iff \sum_{i=1}^{n}\mathbb{E}_{z_i^{(\ell)},z_i^{(\ell+1)},s_i|y_i,\hat{\Theta}}\left[z_i^{(\ell)}-(\eta_{k_\ell}^{(\ell)}+\Lambda_{k_\ell}^{(\ell)}z_i^{(\ell+1)})\right]=0,\ \text{since } \Psi_{k_\ell}^{(\ell)} \text{ is positive semi-definite.}$$

$$\iff \sum_{i=1}^{n}\sum_{\tilde{s}_i^{(:k_\ell:)}} f(s_i^{(:k_\ell:)}=\tilde{s}_i^{(:k_\ell:)}|y_i,\hat{\Theta})\left[\mathbb{E}_{z_i^{(\ell)}|\tilde{s}_i^{(:k_\ell:)},y_i,\hat{\Theta}}[z_i^{(\ell)}]-\eta_{k_\ell}^{(\ell)}-\Lambda_{k_\ell}^{(\ell)}\mathbb{E}_{z_i^{(\ell+1)}|\tilde{s}_i^{(:k_\ell:)},y_i,\hat{\Theta}}[z_i^{(\ell+1)}]\right]=0.$$

Therefore, the estimator of $\eta_{k_\ell}^{(\ell)}$ is given by

$$\hat{\eta}_{k_\ell}^{(\ell)}=\frac{\sum_{i=1}^{n}\sum_{\tilde{s}_i^{(:k_\ell:)}} f(s_i^{(:k_\ell:)}=\tilde{s}_i^{(:k_\ell:)}|y,\hat{\Theta})\left[E[z_i^{(\ell)}|s_i^{(:k_\ell:)}=\tilde{s}_i^{(:k_\ell:)},y_i,\hat{\Theta}]-\Lambda_{k_\ell}^{(\ell)}E[z_i^{(\ell+1)}|\tilde{s}_i^{(:k_\ell:)},y_i,\hat{\Theta}]\right]}{\sum_{i=1}^{n}\sum_{\tilde{s}_i^{(:k_\ell:)}} f(s_i^{(:k_\ell:)}=\tilde{s}_i^{(:k_\ell:)}|y_i,\hat{\Theta})},$$

with

$$\begin{aligned}E[z_i^{(\ell+1)}|s_i^{(:k_\ell:)}=\tilde{s}_i^{(:k_\ell:)},y_i,\hat{\Theta}]&=\int_{z_i^{(\ell)}} f(z_i^{(\ell)}|\tilde{s}_i^{(:k_\ell:)},y_i,\hat{\Theta})\int_{z_i^{(\ell+1)}} f(z_i^{(\ell+1)}|z_i^{(\ell)},\tilde{s}_i^{(:k_\ell:)},\hat{\Theta})z_i^{(\ell+1)}dz_i^{(\ell+1)}dz_i^{(\ell)}\\&\approx\sum_{m_\ell=1}^{M^{(\ell)}} f(z_{i,m_\ell}^{(\ell)}|\tilde{s}_i^{(:k_\ell:)},y_i,\hat{\Theta})\sum_{m_{\ell+1}=1}^{M^{(\ell+1)}} z_{i,m_{\ell+1}}^{(\ell+1)},\end{aligned}$$

where $z_{i,m_{\ell+1}}^{(\ell+1)}$ has been drawn from $f(z_{i,m_{\ell+1}}^{(\ell+1)}|z_{i,m_\ell}^{(\ell)},s)$. Using the same reasoning for $\Lambda_{k_\ell}^{(\ell)}$ we

obtain

$$\begin{aligned}&\mathbb{E}_{z^{(\ell)},z^{(\ell+1)},s|y,\hat{\Theta}}\left[\frac{\partial \log L(z^{(\ell)}|z^{(\ell+1)},s,\Theta)}{\partial \Lambda_{k_\ell}^{(\ell)}}\right]=0\\&\iff \Psi_{k_\ell}^{(\ell)-1}\sum_{i=1}^{n}\left[\mathbb{E}_{z_i^{(\ell)},z_i^{(\ell+1)},s_i|y_i,\hat{\Theta}}[(z_i^{(\ell)}-(\eta_{k_\ell}^{(\ell)}+\Lambda_{k_\ell}^{(\ell)}z_i^{(\ell+1)}))z_i^{(\ell+1)T}]\right]=0\\&\iff \sum_{i=1}^{n}\sum_{\tilde{s}_i^{(:k_\ell:)}} f(s_i^{(:k_\ell:)}=\tilde{s}_i^{(:k_\ell:)}|y_i,\hat{\Theta})\Big[\mathbb{E}_{z_i^{(\ell)},z_i^{(\ell+1)}|\tilde{s}_i^{(:k_\ell:)},y_i,\hat{\Theta}}\\&\qquad\qquad[(z_i^{(\ell)}-\eta_{k_l}^{(\ell)})z_i^{(\ell+1)T}]-\Lambda_{k_\ell}^{(\ell)}\mathbb{E}_{z_i^{(\ell+1)}|\tilde{s}_i^{(:k_\ell:)},y_i,\hat{\Theta}}[z_i^{(\ell+1)}z_i^{(l+\ell)T}]\Big]\\&\quad=0.\end{aligned}$$

Hence the estimator of $\Lambda_{k_\ell}^{(\ell)}$ is given by

$$\hat{\Lambda}_{k_\ell}^{(\ell)} = \frac{\sum_{i=1}^n \sum_{\tilde{s}_i^{(:k_\ell:)}} f(s_i^{(:k_\ell:)} = \tilde{s}_i^{(:k_\ell:)} | y_i, \hat{\Theta}) \Big[E[(z_i^{(\ell)} - \hat{\eta}_{k_\ell}^{(\ell)}) z_i^{(\ell+1)T} | \tilde{s}_i^{(:k_\ell:)}, y_i, \hat{\Theta}] \Big]}{\sum_{i=1}^n \sum_{\tilde{s}_i^{(:k_\ell:)}} f(s_i^{(:k_\ell:)} = \tilde{s}_i^{(:k_\ell:)} | y_i, \hat{\Theta})} E[z_i^{(\ell+1)} z_i^{(\ell+1)T} | \tilde{s}_i^{(:k_\ell:)}, y_i, \hat{\Theta}]^{-1},$$

with

$$\begin{aligned} E[(z_i^{(\ell)} - \hat{\eta}_{k_\ell}^{(\ell)}) z_i^{(\ell+1)T} | \tilde{s}_i^{(:k_\ell:)}, y_i, \hat{\Theta}] &= \int_{z_i^{(\ell)}} f(z_i^{(\ell)} | \tilde{s}_i^{(:\ell:)}, y_i, \hat{\Theta}) \int_{z_i^{(\ell+1)}} f(z_i^{(\ell+1)} | z_i^{(\ell)}, \tilde{s}_i^{(:k_\ell:)}, \hat{\Theta}) [(z_i^{(\ell)} - \hat{\eta}_{k_\ell}^{(\ell)}) z_i^{(\ell+1)T}] dz_i^{(\ell+1)} dz_i^{(} \\ &\approx \sum_{m_\ell=1}^{M^{(\ell)}} f(z_{i,m_\ell}^{(\ell)} | \tilde{s}_i^{(:k_\ell:)}, y_i, \hat{\Theta}) \sum_{m_{\ell+1}=1}^{M^{(\ell+1)}} [(z_{i,m_\ell}^{(\ell)} - \hat{\eta}_{k_l}^{(\ell)}) z_{i,m_{\ell+1}}^{(\ell+1)T}], \end{aligned}$$

where $z_{i,m_\ell}^{(\ell)}$ has been drawn from $f(z_{i,m_\ell}^{(\ell)} | s, \hat{\Theta})$ and $z_{i,m_{\ell+1}}^{(\ell+1)}$ from $f(z_{i,m_{\ell+1}}^{(\ell+1)} | z_{i,m_\ell}^{(\ell)}, s, \hat{\Theta})$.
Finally, we write

$$\begin{aligned} &\mathbb{E}_{z^{(\ell)}, z^{(\ell+1)}, s | y, \hat{\Theta}} \left[\frac{\partial \log L(z^{(\ell)} | z^{(\ell+1)}, s, \Theta)}{\partial \Psi_{k_\ell}^{(\ell)}} \right] = 0 \\ &\iff -\frac{1}{2} \Psi_{k_\ell}^{(\ell)-1} \sum_{i=1}^n \mathbb{E}_{z_i^{(\ell)}, z_i^{(\ell+1)}, s_i |, y_i, \hat{\Theta}} \Big[I_{r_1} - \Big(z_i^{(\ell)} - (\eta_{k_\ell}^{(\ell)} + \Lambda_{k_\ell}^{(\ell)} z_i^{(\ell+1)}) \Big) \Big(z_i^{(\ell)} - (\eta_{k_\ell}^{(\ell)} + \Lambda_{k_\ell}^{(\ell)} z_i^{(\ell+1)}) \Big)^T \Psi_{k_\ell}^{(\ell)-1} \Big] = 0 \\ &\iff \sum_{i=1}^n \mathbb{E}_{z_i^{(\ell)}, z_i^{(\ell+1)}, s_i | y_i, \hat{\Theta}} \Big[I_{r_1} - e^{(\ell)} e^{(\ell)T} \Psi_{k_\ell}^{(\ell)-1} \Big] = 0 \\ &\iff \sum_{i=1}^n \sum_{\tilde{s}_i^{(:k_\ell:)}} f(s_i^{(:k_\ell:)} = \tilde{s}_i^{(:k_\ell:)} | y_i, \hat{\Theta}) I_{r_1} = \sum_{i=1}^n \sum_{\tilde{s}_i^{(:k_\ell:)}} f(s_i^{(:k_\ell:)} = \tilde{s}_i^{(:k_\ell:)} | y_i, \hat{\Theta}) \mathbb{E}_{z_i^{(\ell)}, z_i^{(\ell+1)} | \tilde{s}_i^{(:k_\ell:)}, y_i, \hat{\Theta}} \Big[e^{(\ell)} e^{(\ell)T} \Big] \Psi_{k_\ell}^{(\ell)-1} = 0 \\ &\iff \sum_{i=1}^n \sum_{\tilde{s}_i^{(:k_\ell:)}} f(s_i^{(:k_\ell:)} = \tilde{s}_i^{(:k_\ell:)} | y_i, \hat{\Theta}) \Psi_{k_\ell}^{(\ell)} = \sum_{i=1}^n \sum_{\tilde{s}_i^{(:k_\ell:)}} f(s_i^{(:k_\ell:)} = \tilde{s}_i^{(:k_\ell:)} | y_i, \hat{\Theta}) \mathbb{E}_{z_i^{(\ell)}, z_i^{(\ell+1)} | \tilde{s}_i^{(:k_\ell:)}, y_i, \hat{\Theta}} \Big[e^{(\ell)} e^{(\ell)T} \Big], \end{aligned}$$

with $e^{(\ell)} = \Big(z_i^{(\ell)} - (\eta_{k_\ell}^{(\ell)} + \Lambda_{k_\ell}^{(\ell)} z_i^{(\ell+1)}) \Big)$. Hence the estimator of $\Psi_{k_\ell}^{(\ell)}$ has the form

$$\hat{\Psi}_{k_\ell}^{(\ell)} = \frac{\sum_{i=1}^n \sum_{\tilde{s}_i^{(:k_\ell:)}} f(s_i^{(:k_\ell:)} = \tilde{s}_i^{(:k_\ell:)} | y_i, \hat{\Theta}) E\Big[\Big(z_i^{(\ell)} - (\eta_{k_\ell}^{(\ell)} + \Lambda_{k_\ell}^{(\ell)} z_i^{(\ell+1)}) \Big) \Big(z_i^{(\ell)} - (\eta_{k_\ell}^{(\ell)} + \Lambda_{k_\ell}^{(\ell)} z_i^{(\ell+1)}) \Big)^T | \tilde{s}_i^{(:k_\ell:)}, y_i, \hat{\Theta} \Big]}{\sum_{i=1}^n \sum_{\tilde{s}_i^{(:k_\ell:)}} f(s_i^{(:k_\ell:)} = \tilde{s}_i^{(:k_\ell:)} | y_i, \hat{\Theta})}.$$

8.4 Common tail layers mathematical derivations (E step)

The conditional expectation $f(z^{(\ell)}, z^{(\ell+1)}, s^C, s^D, s^{(L_0+1:)}|y^C, y^D, \hat{\Theta}_C, \hat{\Theta}_D, \hat{\Theta}_{L_0+1:})$ can be rewritten as:

$$\begin{aligned} f(z^{(\ell)}, z^{(\ell+1)}, s^C, s^D, & s^{(L_0+1:)}|y^C, y^D, \hat{\Theta}_C, \hat{\Theta}_D, \hat{\Theta}_{L_0+1:}) \\ &= f(z^{(\ell)}|s^C, s^D, s^{(L_0+1:)}, y^C, y^D, \hat{\Theta}_C, \hat{\Theta}_D, \hat{\Theta}_{L_0+1:}) f(z^{(\ell+1)}|z^{(\ell)}, s^{(L_0+1:)}, \hat{\Theta}_{L_0+1:}) \\ &\quad \times f(s^C, s^D, s^{(L_0+1:)}|y^C, y^D, \hat{\Theta}_C, \hat{\Theta}_D, \hat{\Theta}_{L_0+1:}). \end{aligned} \tag{14}$$

$\forall \ell \in [L_0+1, L]$, the first term of (14) can be proportionally expressed as:

$$\begin{aligned} f(z^{(\ell)}|s^C, s^D, s^{(L_0+1:)}, y^C, y^D, \hat{\Theta}_C, \hat{\Theta}_D, \hat{\Theta}_{L_0+1:}) \quad \propto \quad & f(y^C|z^{(\ell)}, s^C, s^{(L_0+1:)}, \hat{\Theta}_C, \hat{\Theta}_{L_0+1:}) \\ & \times f(z^{(\ell)}|s^D, s^{(L_0+1:)}, y^D, \hat{\Theta}_D, \hat{\Theta}_{L_0+1:}). \end{aligned}$$

One can compute $f(y^C|z^{(\ell)}, s^C, s^{(L_0+1:)}, \hat{\Theta}_C, \hat{\Theta}_{L_0+1:})$ using Bayes rule and $f(z^{(\ell)}|s^D, s^{(L_0+1:)}, y^D, \hat{\Theta}_D, \hat{\Theta}_{L_0+1:})$ is known from (11). Finally the second term of (14) can be computed as in (12). By mutual independence of s^C, s^D, and $s^{(L_0+1:)}$, the third term reduces to the product of three densities which are given in Section 8.5.1.

8.5 Path probabilities mathematical derivations

8.5.1 E step for determining the path probabilities

We consider the three following densities: $f(s^{(\ell)D} = k_\ell|y^D, \hat{\Theta}_D, \hat{\Theta}_{L_0+1:})$, $f(s^{(\ell)C} = k_\ell|y^C, \hat{\Theta}_C, \hat{\Theta}_{L_0+1:})$, and $f(s^{(\ell)} = k_\ell|y^C, y^D, \hat{\Theta}_C, \hat{\Theta}_D, \hat{\Theta}_{L_0+1:})$. The first density can be computed from (9) as

$$f(s^{(\ell)D} = k_\ell|y^D, \hat{\Theta}_D, \hat{\Theta}_{L_0+1:}) = \sum_{\tilde{s} \in \Omega^{(:k_\ell:)D}} f(s^{(1D:L)} = \tilde{s}|y^D, \hat{\Theta}_D, \hat{\Theta}_{L_0+1:}),$$

where $\Omega^{(:k_\ell:)D}$ is the set of the full paths going through the component k_ℓ of layer ℓ of head D. The second density can be computed similarly using the fact that $f(y^C|s^C, s^{(L_0+1:)}, \hat{\Theta}_C, \hat{\Theta}_{L_0+1:})$ is Gaussian with parameters $(\mu^{(1C:L)}, \Sigma^{(1C:L)})$. Concerning the last density, we have to compute $p(s^{(L_0+1:)}|y^C, y^D, \hat{\Theta}_D, \hat{\Theta}_C, \hat{\Theta}_{(L_0+1:)})$. We are still making the two following conditional assumptions:

$$(y^C \perp\!\!\!\perp y^D)|z^{(L_0+1)} \quad \text{and} \quad (z^D \perp\!\!\!\perp z^C)|z^{(L_0+1)},$$

with $z^{(L_0+1)}$ the first common tail layer. We then have:

$$
\begin{aligned}
&p(s^{(L_0+1:)}|y^C, y^D, \hat{\Theta}_D, \hat{\Theta}_C, \hat{\Theta}_{(L_0+1:)}) \\
&= \frac{p(s^{(L_0+1:)}, y^C, y^D|\hat{\Theta}_D, \hat{\Theta}_C, \hat{\Theta}_{(L_0+1:)})}{p(y^C, y^D)} \\
&\propto p(s^{(L_0+1:)}, y^C, y^D|\hat{\Theta}_D, \hat{\Theta}_C, \hat{\Theta}_{(L_0+1:)}) \\
&= \sum_{s^C}\sum_{s^D}\int_{z^{(L_0+1)}} p(s^{(L_0+1:)}, y^C, y^D, z^{(L_0+1)}, s^C, s^D|\hat{\Theta}_D, \hat{\Theta}_C, \hat{\Theta}_{(L_0+1:)})dz^{(L_0+1)} \\
&= \sum_{s^C}\sum_{s^D}\int_{z^{(L_0+1)}} p(y^C|s^C, s^{(L_0+1:)}, z^{(L_0+1)}, \hat{\Theta}_C, \hat{\Theta}_{(L_0+1:)})p(y^D|s^D, s^{(L_0+1:)}, z^{(L_0+1)}, \hat{\Theta}_D, \hat{\Theta}_{(L_0+1:)}) \\
&\times p(z^{(L_0+1)}|s^{(L_0+1:)}, \hat{\Theta}_{(L_0+1:)})\prod_{\ell=1}^{L_0} p(s^{(\ell)C}|\hat{\Theta}_C, \hat{\Theta}_{(L_0+1:)})p(s^{(\ell)D}|\hat{\Theta}_D, \hat{\Theta}_{(L_0+1:)})\prod_{\ell=L_0+1}^{L} p(s^{(\ell)}|\hat{\Theta}_{(L_0+1:)})dz^{(L_0+1)},
\end{aligned}
$$

by independence of the $(s^{(\ell)h})_{\ell,h}$. The first two terms are computed as for (14), the third term is a Gaussian given in (13) and each density of the products are multinomial densities whom coefficients have already been estimated.

8.5.2 M step for determining the path probabilities

Using again the conditional independence, we maximise

$$
\begin{aligned}
&E_{s^C,s^D,s^{(L_0+1)}|y^C,y^D,\hat{\Theta}_C,\hat{\Theta}_D,\hat{\Theta}_{L_0+1:}}[\log L(s^C, s^D, s^{(L_0+1:)}|\Theta_C, \Theta_D, \Theta_{L_0+1:})] \\
&= \sum_{\ell=1}^{L_0}\mathbb{E}_{s^{(\ell)C}|y^C,\hat{\Theta}_C,\hat{\Theta}_{L_0+1:}}[\log L(s^{(l)C}|\Theta_C, \Theta_{L_0+1:})] \\
&+ \sum_{\ell=1}^{L_0}\mathbb{E}_{s^{(\ell)D}|y^D,\hat{\Theta}_D,\hat{\Theta}_{L_0+1:}}[\log L(s^{(l)D}|\Theta_D, \Theta_{L_0+1:})] \\
&+ \sum_{\ell=L_0+1}^{L}\mathbb{E}_{s^{(\ell)}|y^C,y^D,\hat{\Theta}_C,\hat{\Theta}_D,\hat{\Theta}_{L_0+1:}}[\log L(s^{(\ell)}|\Theta_C, \Theta_D, \Theta_{L_0+1:})],
\end{aligned}
$$

with respect to $\pi_{k_\ell}^{(\ell)h}$, $\forall h \in \{C, D\}, \ell \in [1, L_0], k_\ell \in [1, K_\ell]$ and with respect to $\pi_{k_\ell}^{(\ell)}$, $\forall \ell \in [L_0, L], k_\ell \in [1, K_\ell]$.

Each of heads and tail estimators can be computed in the same way. Let k_ℓ be the index of a component of layer ℓ for which we want to derive an estimator and $\tilde{k}_\ell$ another component index. The associated probabilities are respectively π_{k_ℓ} and $\pi_{\tilde{k}_\ell}$. We omit the head subscript h for better readability.

We have

$$
\begin{aligned}
&E_{s^{(\ell)}|y,\hat{\Theta}}[\log L(s^{(\ell)}|\Theta)] \\
&= \sum_{i=1}^{n} \sum_{k'_\ell=1}^{K_\ell} f(s^{(\ell)} = k'_\ell|y,\hat{\Theta}) \log L(s^{(\ell)} = k'_\ell|\Theta) \\
&= \sum_{i=1}^{n} \sum_{\substack{k'_\ell=1 \\ k'_\ell \neq \tilde{k}_\ell}}^{K_\ell} f(s^{(\ell)} = k'_\ell|y,\hat{\Theta}) \log L(s^{(\ell)} = k'_\ell|\Theta) + \sum_{i=1}^{n} f(s^{(\ell)} = \tilde{k}_\ell|y,\hat{\Theta}) \log L(s^{(\ell)} = \tilde{k}_\ell|\Theta) \\
&= \sum_{i=1}^{n} \sum_{\substack{k'_\ell=1 \\ k'_\ell \neq \tilde{k}_\ell}}^{K_\ell} f(s^{(\ell)} = k'_\ell|y,\hat{\Theta}) \log \pi^{(\ell)}_{k'_\ell} + \sum_{i=1}^{n} f(s^{(\ell)} = \tilde{k}_\ell|y,\hat{\Theta}) \log(1 - \sum_{\substack{k'_\ell=1 \\ k'_\ell \neq \tilde{k}_\ell}} \pi^{(\ell)}_{k'_\ell}).
\end{aligned}
$$

Taking the derivative with respect to $\pi^{(\ell)}_{k_\ell}$ and equalizing to zero yields

$$
\begin{aligned}
\frac{\partial \mathbb{E}_{s^{(\ell)}|y,\hat{\Theta}}[\log L(s^{(\ell)}|\Theta)]}{\partial \pi^{(\ell)}_{k_\ell}} = 0 \Leftrightarrow \frac{\sum_{i=1}^{n} f(s^{(\ell)} = k_\ell|y,\hat{\Theta})}{\pi^{(\ell)}_{k_\ell}} &= \frac{\sum_{i=1}^{n} f(s^{(\ell)} = \tilde{k}_\ell|y,\hat{\Theta})}{\pi^{(\ell)}_{\tilde{k}_\ell}} \\
\Leftrightarrow \pi^{(\ell)}_{\tilde{k}_\ell} &= \frac{\sum_{i=1}^{n} f(s^{(\ell)} = \tilde{k}_\ell|y,\hat{\Theta})}{\sum_{i=1}^{n} f(s^{(\ell)} = k_\ell|y,\hat{\Theta})} \pi^{(\ell)}_{k_\ell}.
\end{aligned}
$$

Finally, summing over $\tilde{k}_\ell$ we get

$$
\pi^{(\ell)}_{\tilde{k}_\ell} = \frac{\sum_{i=1}^{n} f(s^{(\ell)} = \tilde{k}_\ell|y,\hat{\Theta})}{\sum_{i=1}^{n} f(s^{(\ell)} = k_\ell|y,\hat{\Theta})} \pi^{(\ell)}_{k_\ell} \Leftrightarrow 1 = \frac{n}{\sum_{i=1}^{n} f(s^{(\ell)} = k_\ell|y,\hat{\Theta})} \pi^{(\ell)}_{k_\ell} \Leftrightarrow \hat{\pi}^{(\ell)}_{k_\ell} = \frac{\sum_{i=1}^{n} f(s^{(\ell)} = k_\ell|y,\hat{\Theta})}{n}.
$$

As a result, the probability estimator for each head h is:

$$
\hat{\pi}^{(\ell)h}_{k_\ell} = \frac{\sum_{i=1}^{n} f(s^{(\ell)h} = k_\ell|y^h,\hat{\Theta}_h,\hat{\Theta}_{L_0+1:})}{n}.
$$

For the common tail the estimator is of the form, $\forall \ell \in [L_0+1, L]$:

$$
\hat{\pi}^{(\ell)}_{k_\ell} = \frac{\sum_{i=1}^{n} f(s^{(\ell)} = k_\ell|y^C,y^D,\hat{\Theta}_C,\hat{\Theta}_D,\hat{\Theta}_{L_0+1:})}{n}.
$$

8.6 Latent variables identifiabiliy rescaling

The GLLVM and Factor Analysis models assume that the latent variable is centered and of unit variance, *i.e.* that $z^{(1)C}$ and $z^{(1)D}$ are centered-reduced in our setup.

We iteratively center and reduce each layer latent variable $z^{(\ell)}$ starting from the last layer of the

common tail to the first layers of each head h in order for all $(z^{(\ell)h})_{h,\ell}$ to be centered-reduced. As the latent variable of the last layer of DGMM family models is a centered-reduced Gaussian, by induction, $z^{(1)C}$ and $z^{(1)D}$ are centered and reduced.

Assuming that the latent variable of the next layer is centered-reduced, the mean and variance of the latent variable of the current layer $l \in [1, L]$ of head or tail $h \in \{C, D, L_0 + 1 :\}$ is:

$$\begin{cases} E(z^{(\ell)h}) = \sum_{k'_\ell} \pi_{k'_\ell}^{(\ell)h} \eta_{k'_\ell}^{(\ell)h} \\ Var(z^{(\ell)h}) = \sum_{k'_\ell} \pi_{k'_\ell}^{(\ell)h} (\Lambda_{k'_\ell}^{(\ell)h} \Lambda_{k'_\ell}^{(\ell)hT} + \Psi_{k'_\ell}^{(\ell)h} + \eta_{k'_\ell}^{(\ell)h} \eta_{k'_\ell}^{(\ell)hT}) - (\sum_{k'_\ell} \pi_{k'_\ell}^{(\ell)h} \eta_{k'_\ell}^{(\ell)h})(\sum_{k'_\ell} \pi_{k'_\ell}^{(\ell)h} \eta_{k'_\ell}^{(\ell)h})^T. \end{cases}$$

Let $A^{(\ell)h}$ be the Cholesky decomposition of $Var(z^{(\ell)h})$ $\forall k_\ell \in [1, K_\ell]$, then we rescale the layer parameters in the following way:

$$\begin{cases} \eta_{k_\ell}^{(\ell)hnew} = A^{(l)h-1T} \left[\eta_{k_\ell}^{(\ell)h} - \sum_{k'_\ell} \pi_{k'_\ell}^{(\ell)h} \eta_{k'_\ell}^{(\ell)h} \right] \\ \Lambda_{k_\ell}^{(\ell)hnew} = A^{(l)h-1T} \Lambda_{k_\ell}^{(\ell)h} \\ \Psi_{k_\ell}^{(\ell)new} = A^{(l)h-1T} \Psi_{k_\ell}^{(\ell)h} A^{(l)h-1}, \end{cases}$$

with the subscript "new" denoting the rescaled version of the parameters.

8.7 Monte Carlo scheme

The number of Monte Carlo copies $M^{(\ell)h}$ to draw at each layer has to be chosen before running the MCEM. Wei and Tanner (1990) advise to let M grow through the iterations starting with a very low M. Doing so, one does not get stuck into very local optima at the beginning of the algorithm and ends up in a precisely estimated expectation state. The growth scheme of M_t^ℓ through the iterations t implemented here is:

$$M_t^\ell = \left\lfloor \frac{40}{\log(n)} \times t \times \sqrt{r_\ell} \right\rfloor.$$

M_t^ℓ grows linearly with the number of iterations t to follow Wei and Tanner (1990) advice. In order to explore the latent space at each layer, M^ℓ also grows with the dimension of the layer. The square root rate just ensures that the running time remains affordable, whereas if there were no additional computational costs we would certainly have let M^ℓ grow much more with r_ℓ. Finally, we make the hypothesis that the more observations in the dataset the stronger the signal is and hence the fewer draws of latent variables are needed to train the model.

Remark 8.1 *In this Monte Carlo version contrary to the regular EM algorithm, the likelihood does not increase necessary through the iterations. In classical EM-based models, the training is often stopped once the likelihood increases by less than a given threshold between two iterations. The stopping process had then to be adapted to account for temporary losses in likelihood. Hence, we have defined a patience parameter which is the number of iterations without log-likelihood increases to wait before stopping the algorithm. Typically, we set this parameter to 1 iteration in the simulations.*

8.8 Model selection details

This section gives additional details about the way model selection is performed on the fly.

A component of the ℓth layer is considered useless if its probability is inferior to $\frac{1}{4k_\ell}$, where k_ℓ denotes the number of components of the layer. For instance, if a layer is formerly composed of four components, the components associated with a probability inferior to 0.0625 are removed from the architecture.

For the GLLVM layer, logistic and linear regressions were fitted to determine which of the dimensions had a significant effect over each y_j^D for each path $\tilde{s}$. We have fitted a logistic LASSO for each binary, count and categorical variable and an ordinal logistic regression for each ordinal variable. In the M1DGMM case, we have fitted a linear LASSO for each continuous variable. The variables associated with coefficients identified as being zero (or not significant at a 10% level) for at least 25% of the paths were removed.

The same voting idea was used for the regular DGMM layers to determine the useless dimensions. As our algorithm generate draws of $(z^{(\ell+1)}|z^{(\ell)}, s)$ of dimension $r_{\ell+1}$, it is possible to perform a PCA on this variable for each path and each of the $M^{(\ell)}$ points simulated for $z^{(\ell)}$. Doing so, one can compute the average contribution of each dimension of r_{l+1} to the first principal component and set a threshold under which a dimension is deleted. We have set this threshold to 0.2 for our simulations. The intuition behind this is that the first component of the PCA conveys the majority of the pieces of information that $z^{(\ell+1)}$ has on $z^{(\ell)}$. If a dimension shares no common information with this first component, hence it is not useful to keep it.

The dimension of the junction layer (the first DGMM layer on the common tail) is chosen according to this procedure too. The two heads decide which dimensions of the junction layer is important and each dimension important for at least one head is kept. This is rather conser-

vative but avoids that contradictory information coming from the two heads disrupt the global architecture.

The number of layers on the heads and tails is fully determined by the selection of the layers dimensions in order to keep the model identifiable. If the dimension of an intermediate tail layer ℓ is selected to be one then $r_\ell > r_{\ell+1} > ... > r_L$ does not hold anymore. Thus, the following tail layers are deleted.
Similarly, if an head layer has a selected dimension of two, then the following head layers are deleted. Indeed, the tail has to have minimal dimensions of two and one on its last layers. This is not compatible with previous head layers of dimension inferior or equal to two.

In the case of head layers deletion, we restart the algorithm (initialisation and proper model run) with the new architecture. Otherwise it would be necessary to re-determine all the path and DGMM coefficients values to bridge the gap between the previous head layer and the junction layer. There were no easy way to do such thing and restarting the algorithm seemed the best to do. Note that in our simulations defining several heads layers did not give good results. Intuitively, it could too much dilute information before passing it to the common tail, resulting in poor performance. We advise to keep only one or two head layers before running the MDGMM. Doing so, this restarting procedure would not be often performed in practice.

8.9 Metrics

A true positive (TP) prediction of the model is an observation that has been assigned to the same class as the "ground truth" label. On the contrary, a False Positive (FP) means that the class predicted by the model and the label do not match. k denotes the class index and K the cardinal of the set of all possible classes. n_k is the number of points in the class k and $y_{i,k}$ an observation of class k.

The formulas of the two precision metrics are :

$$\text{Micro precision} = \frac{\sum_{k=1}^{K}\sum_{i=1}^{n} TP_{i,k}}{\sum_{k=1}^{K}\sum_{i=1}^{n} TP_{i,k} + FP_{i,k}},$$

$$\text{Macro precision} = \frac{1}{K}\sum_{k=1}^{K}\frac{\sum_{i=1}^{n} TP_{i,k}}{\sum_{i=1}^{n} TP_{i,k} + FP_{i,k}}.$$

The formula of the silhouette coefficient is:

$$\text{Silhouette coefficient} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_k}\sum_{i=1}^{n}\frac{d\_inter(i,k) - d\_intra(i,k)}{\max\left(d\_intra(i,k), d\_inter(i,k)\right)},$$

with $d\_intra(i,k) = \frac{1}{n_k - 1}\sum_{i' \neq i} d(y_{i,k}, y_{i',k})$ and $d\_inter(i,k) = \min_{k' \neq k}\frac{1}{n_k}\sum_{i' \neq i}\sum_{k'=1}^{K} d(y_{i,k}, y_{i',k'})$
With d a distance, the Gower distance (Gower, 1971) in our case.

8.10 Benchmark models specifications

A standard Grid Search has been performed to find the best specification of the hyperparameters of the benchmark models. The best value for each metric is reported independently from the other metrics. As such for a given model, the best silhouette score, micro and macro precisions can actually be achieved by three different specifications. The silhouette metric seemed to us the most appropriate since it is unsupervised, but we did not want to favor any metric against the others. Besides, all of the benchmark models are not built upon a likelihood principle which prevents from performing model selection using a common criterion such as the Bayesian Information Criterion (BIC) or the Aikake Information Criterion (AIC). Therefore, this performance report aims at illustrating the clustering power of different algorithms compared to the ones introduced in this work rather than presenting the metrics associated with the best selected specification of each benchmark model.

The following hyperparameters search spaces were used :

K-modes (from the kmodes package)

- Initialisation ∈ {'Huang', 'Cao', 'random'}.

K-prototypes (from the kmodes package)

- Initialisation ∈ {'Huang', 'Cao', 'random'}.

Agglomerative clustering (from the scikit-learn package)

- linkages ∈ {'complete', 'average', 'single'}.

This model was trained using the Gower Distance Matrix computed on the data.

Self-Organizing Map (from the SOMPY package)

- sigma ∈ [0.001, 0.751, 1.501, 2.250, 3.000]
- lr ∈ [0.001, 0.056, 0.111, 0.167, 0.223, 0.278, 0.333, 0.389, 0.444, 0.500].

DBSCAN (from the scikit-learn package)

- leaf_size ∈ {10, 20, 30, 40, 50}
- eps ∈ {0.01, 1.258, 2.505, 3.753, 5.000}
- min_samples ∈ {1, 2, 3, 4}
- Data used: 'scaled data', 'Gower Distance'.

DBSCAN was trained on two versions of the dataset: on the data themselves and using the Gower Distance Matrix computed on the data. Each time the best performing specification was taken.

GLMLVM

- $r \in [1, 5]$
- $k = 2$.

NESP DDGMM (MCA + GMM + FA)

- $r \in [1, 13]$
- $k = 2$.

DDGMM

The starting architecture over which automatic architecture selection was performed was:

- $r = \{5, 4, 3\}$
- $k = \{4, 2\}$
- Number of maximum iterations = 30.

NESP M2DGMM (MCA + GMM + FA + PLS)

The architectures considered had at most 2 layers on each head and 3 layers on the tail.

- r: All the minimal identifiable architectures.
- k: Random draws for each $k_\ell \in \{2, 3, 4\}$
- Number of maximum iterations = 30.

M1DGMM

The starting architecture over which automatic architecture selection was performed was:

- $r = \{5, 4, 3\}$
- $k = \{4, 2\}$
- Number of maximum iterations = 30.

M2DGMM

The starting architecture over which automatic architecture selection was performed was:

- $r_c = \{p_c\}, r_d = \{5\}, r_t = \{4, 3\}$
- $k_c = \{1\}, k_d = \{3\}, k_{L_0+1:} = \{2, 1\}$
- Number of maximum iterations = 30.

$k_c = \{1\}$ and $r_c = \{p_c\}$ are imposed by construction as the first layer of the continuous head are the data themselves.

B. MIAMI: Supplementary Material

Appendix

A Datasets details

The variables of the Adult dataset are according to the UCI documentation:

- Income: binary (>50K, <=50K).
- Age: continuous.
- Workclass: categorical (Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked).
- Fnlwgt: continuous.
- Education-num: ordinal.
- Marital-status: categorical (Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse).
- Occupation: categorical (Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces).
- Relationship: categorical (Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried).
- Race: categorical (White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black).
- Sex: binary (Female, Male).
- Capital-gain: ordinal.
- Capital-loss: ordinal.
- Hours-per-week: continuous.
- Native-country: categorical (United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands).

This preprint has not undergone any post-submission improvements or corrections.
This article was accepted in "22nd International Conference on Computational Science and Its Applications - ICCSA 2022", and will be soon available online (the DOI number will be given then).

B Evaluation metrics details

Our overall criterion between a test dataset and a generated dataset is the association distance obtained as follows:

$$DA = \frac{1}{P} \sum_{1 \leq i < j \leq p}^{p} |M_{ij}(test) - M_{ij}(gen)| / M_{ij}(test),$$

where p denotes the number of variables (14 in Adult dataset), $P = (p^2 - p)/2$, and $M_{ij}(test)$ (resp. $M_{ij}(gen)$ is the (i, j)th entry of the test (resp. generated) Association Matrix.

To measure the similarity between the dependence structures of the vectors formed by the three continuous variables (Age, Fnlwgt, and Hours) we used the multivariate Kullback Leibler divergence [12].

For qualitative data, we chose the mean absolute errors (MAE) between proportions. More precisely, for a kth intersection of modalities we consider

$$MAE(k) = |p_k(test) - p_k(gen)|,$$

where $p_k(test)$ (resp. $p_k(gen)$) stands for kth test (resp. generated) proportion. The final MAE is the mean of all the $MAE(k)$ over all the possibilities.

C Additional results: Unobserved marginal density reconstruction

When it comes to reconstructing univariate densities, MIAMI generates well-identified unimodal densities contrary to DataSynthesizer which generates flat densities, or SynthPop-CART which generates multi-modal densities (Figure 8). More precisely, Figures 8 and 9 represent the estimations of the observed density versus the generated one for Age in the case of the Bivariate and Trivariate Unbalanced designs. We observe that MIAMI can recover the right distribution, yet not observed in the training set. It means that MIAMI captures well the dependence structure of such a partially unobserved variable. CTGAN and SynthPop-RF also seem to work well for both designs. DataSynthesizer shows a larger variance. This illustration shows that we cannot clearly decide between the methods by looking only at the marginal distributions. Only a criterion like the association distance can take into account a more complex multivariate dependence structure.

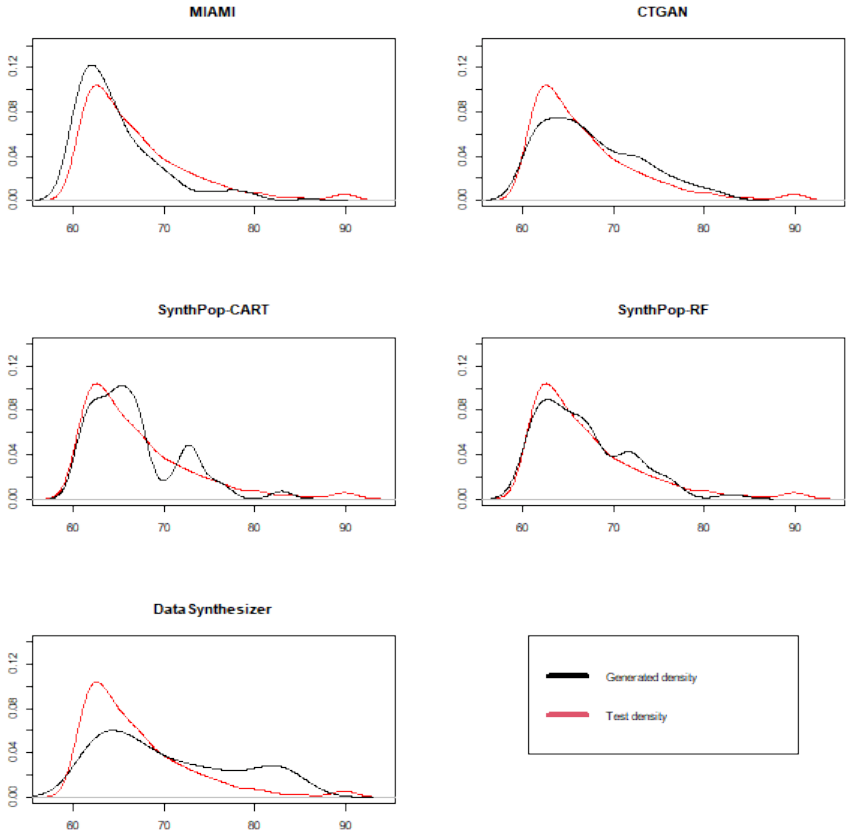


Fig. 8. Density estimations for Age based on the test dataset (red) or based on the Bivariate Unbalanced design (black)

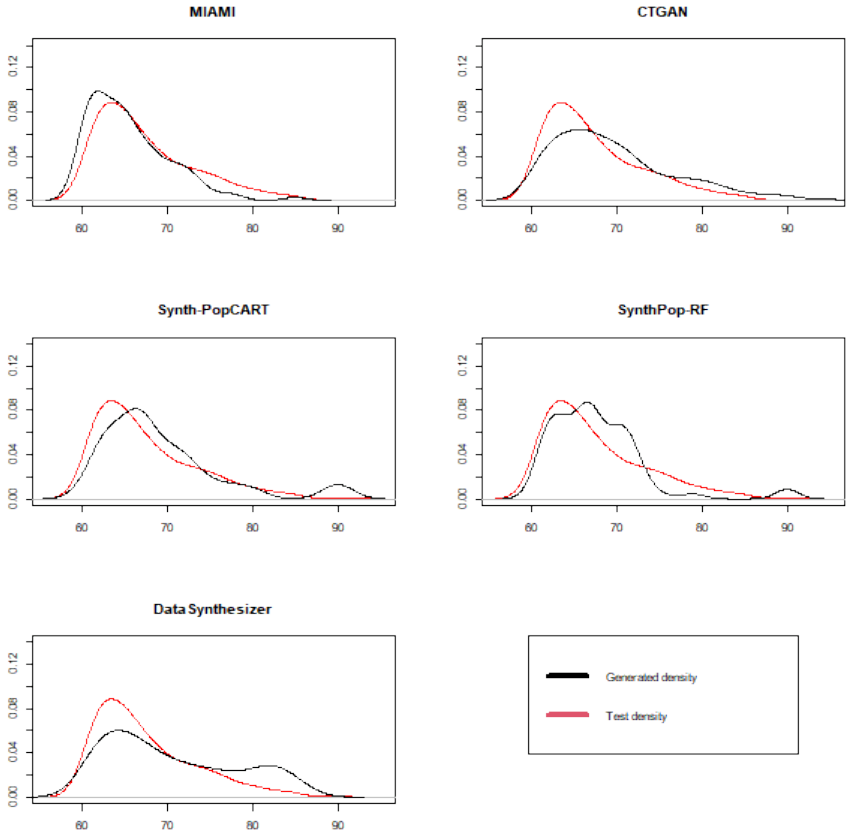


Fig. 9. Density estimations for Age based on the test dataset (red) or based on the Trivariate Unbalanced design (black)

C. RUBALIZ: Supplementary Material

Supplementary information for "A RUpture-Based detection method for the Active mesopeLagIc Zone (RUBALIZ): a crucial step towards rigorous carbon budget assessments" by Fuchs, Baumas et al.

Figure S1:

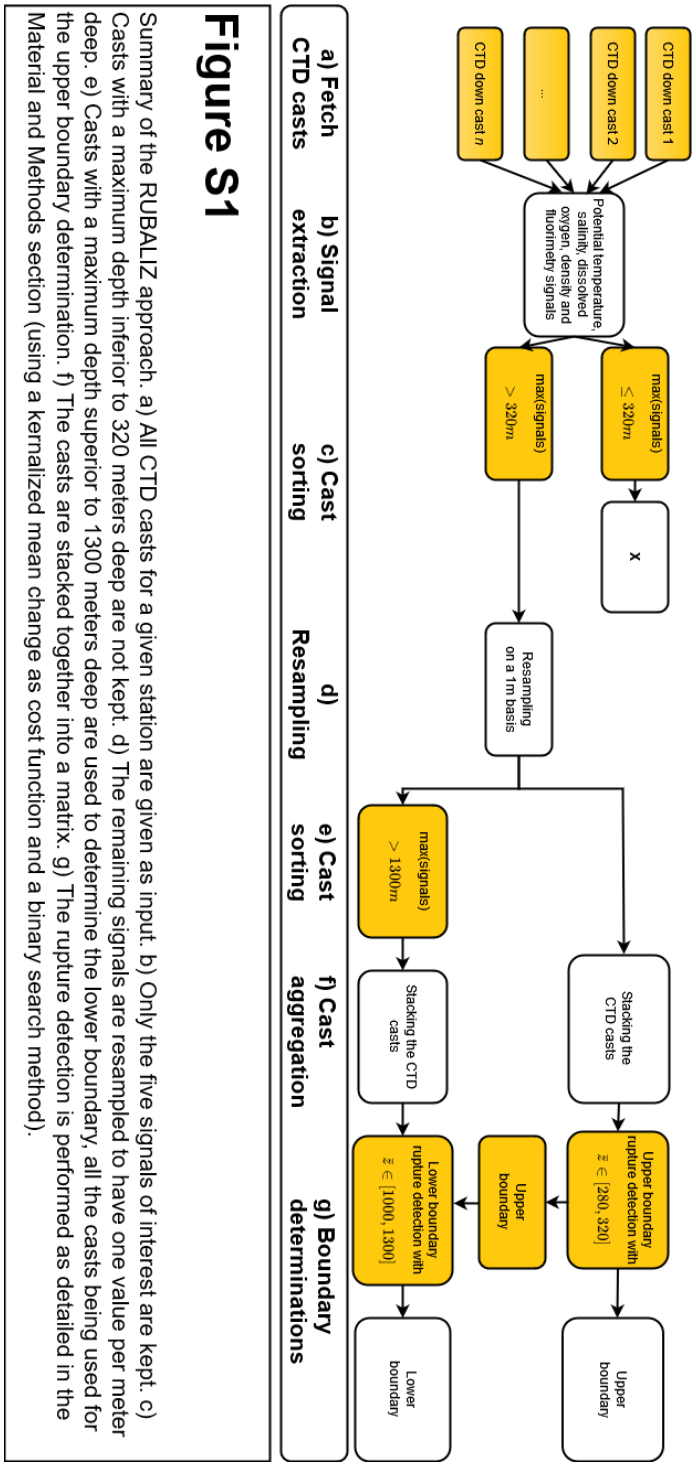


Figure S1

Summary of the RUBALIZ approach. a) All CTD casts for a given station are given as input. b) Only the five signals of interest are kept. c) Casts with a maximum depth inferior to 320 meters deep are not kept. d) The remaining signals are resampled to have one value per meter deep. e) Casts with a maximum depth superior to 1300 meters deep are used to determine the lower boundary, all the casts being used for the upper boundary determination. f) The casts are stacked together into a matrix. g) The rupture detection is performed as detailed in the Material and Methods section (using a kernalized mean change as cost function and a binary search method).

Table S1: Depths of the active mesopelagic zone boundaries determined by RUBALIZ

cruise	station	Upper boundary	std	Upper boundary CTD number	Lower boundary	std	Lower boundary CTD number
D341	PAP	109	1	16	561	5	1
DY032	PAP	126	34	16	746	26	3
KN207-01	QL-1	148	0	2	490	21	2
KN207-01	QL-2	189	0	1	781	198	1
KN207-03	PS-1	101	0	1	487	5	1
KN207-03	PS-3&4	107	1	1	681	4	1
MALINA	430	76	0	1	540	40	1
MALINA	540	81	1	1	555	75	1
MALINA	620	92	31	1	617	81	1
PEACETIME	FAST	106	0	21	626	27	8
PEACETIME	ION	117	2	11	497	25	6
PEACETIME	TYR	109	2	10	604	44	6
TONGA	STATION 8	149	2	5	698	135	3

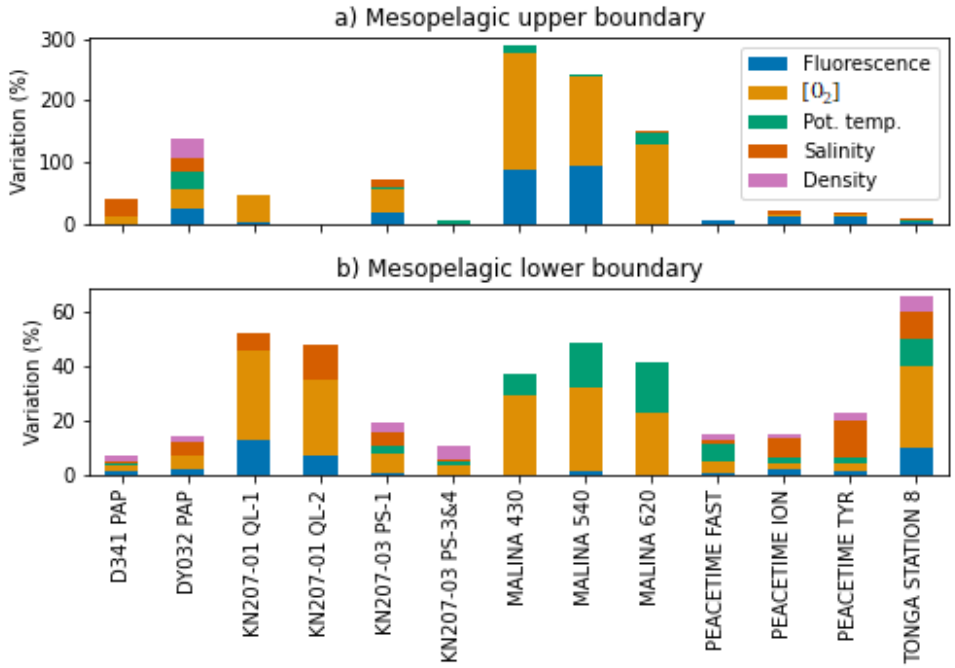


Figure S2: Variation of the boundary estimates due to the withdrawal of one variable from the CTD signal for the upper boundary (a) and lower boundary (b).

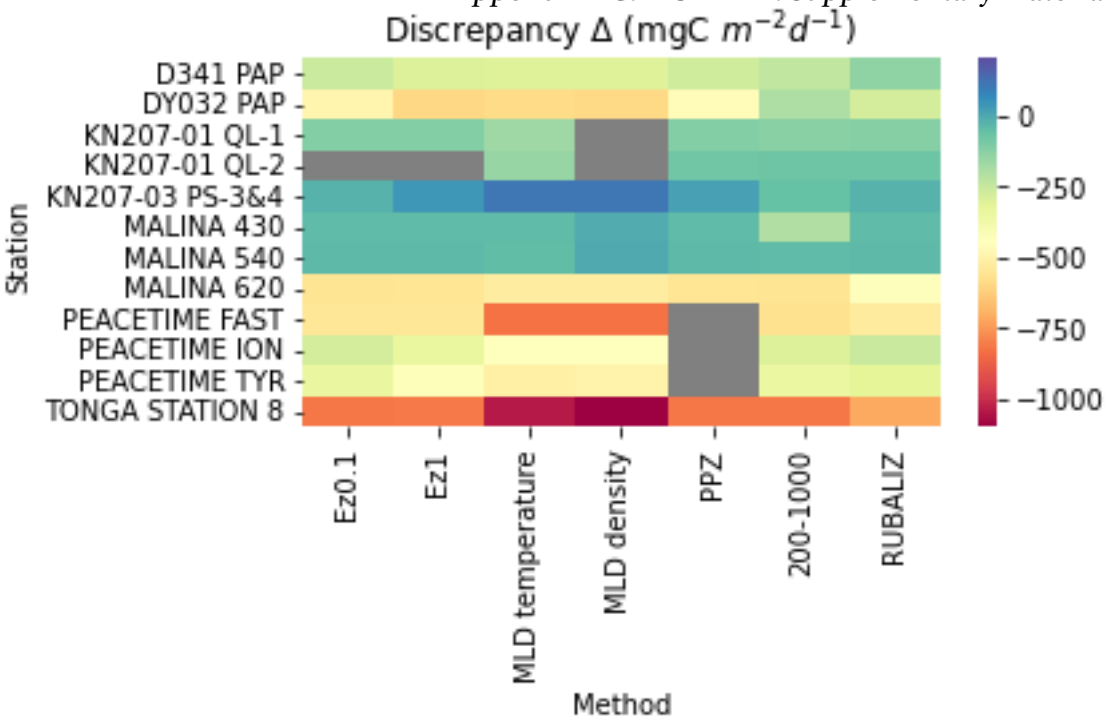


Figure S3: Discrepancy derived from assessment of C budget calculated from all different approaches including RUBALIZ. The gray cells correspond to stations for which a given method could not determine an upper boundary.

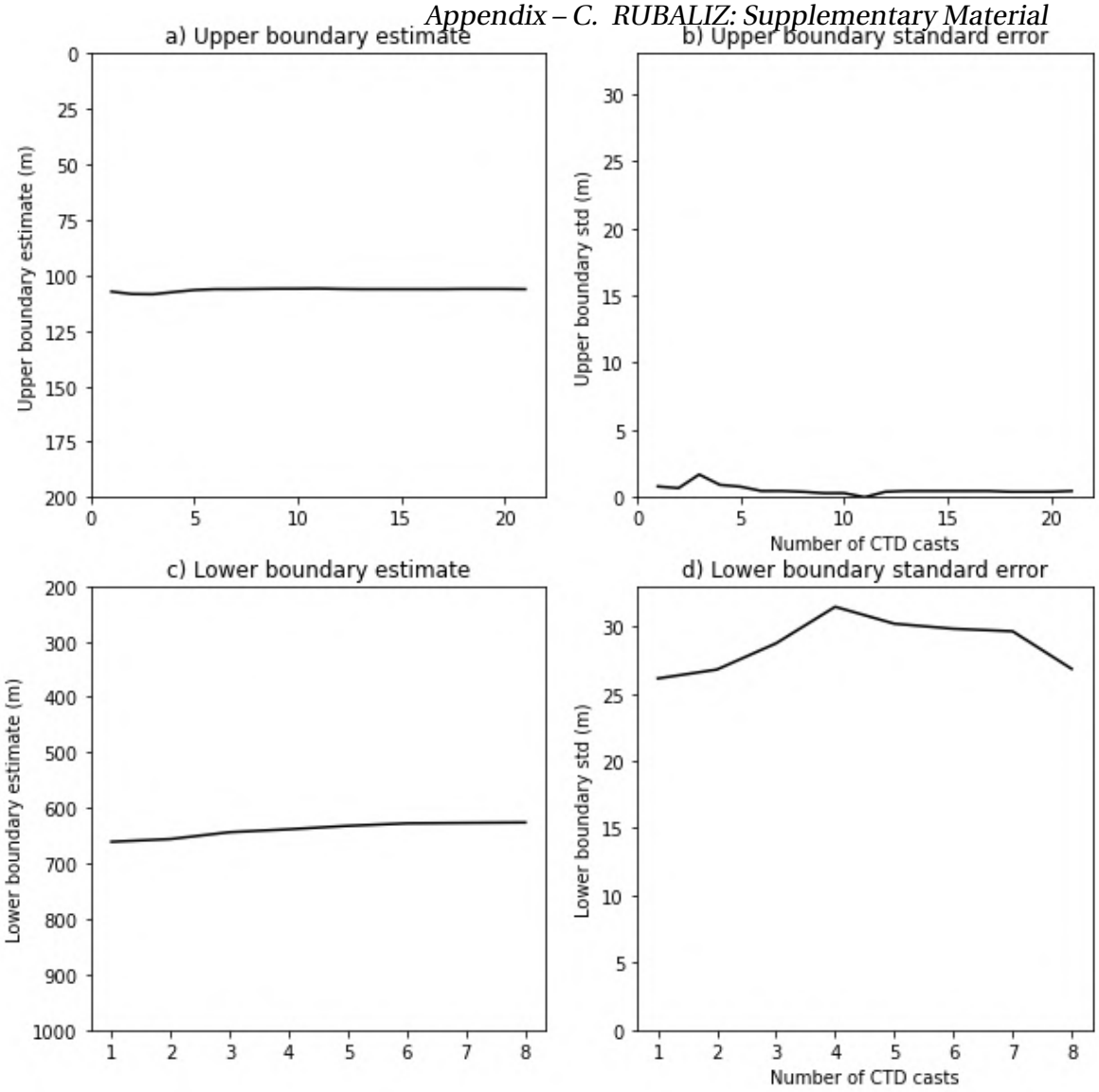


Figure S4: Example of the evolution of the boundaries estimate and associated standard errors (meters deep) when the number of CTDs available grows at the PEACETIME FAST station.

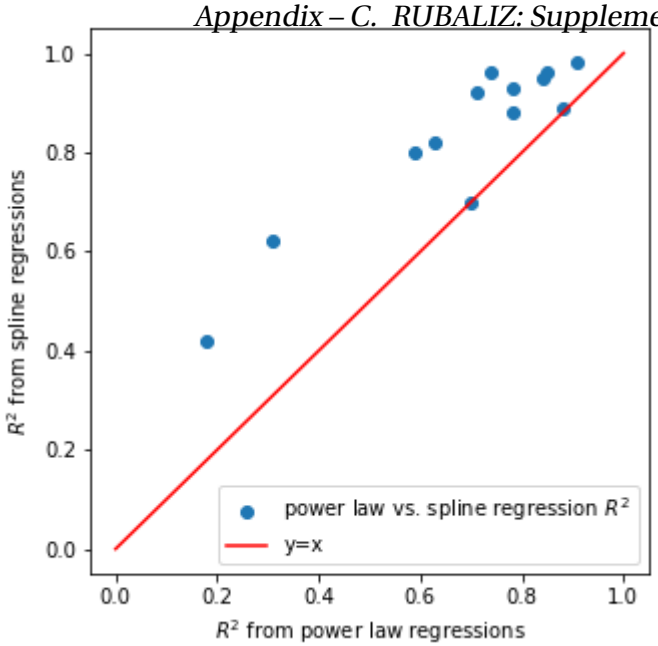


Figure S5: R^2 coefficients obtained using linear splines with one node on log-depth and log-PHP as a function of power law regression R2 coefficients for each station. The red line represents a situation where both methods have the same quality of fit, points above the line correspond to stations for which the spline regressions gave a better fit and conversely for points under the line.

D. CNN: Supplementary Material

Automatic recognition of flow cytometric phytoplankton functional groups using Convolutional Neural Networks Supplemental Information

Robin Fuchs, Melilotus Thyssen, Véronique Creach, Mathilde Dugenne, Lloyd Izard, Marie Latimier, Arnaud Louchart, Pierre Marrec, Machteld Rijkeboer, Gérald Grégori, Denys Pommeret

S1 Listmode features

For each optical curve the CytoClus4© software can output the following features:

- Asymmetry: an asymmetry coefficient of the curve.
- Average: the average value of the curve.
- Center of gravity: the center of gravity of the curve.
- Total: The area under the curve
- Fill factor: a coefficient that measures how slender the curve is.
- Inertia: indicates whether the high values of the curve are concentrated on the center or on the tails of the curve.
- Length: calculated length of the curve based on the signal length at half maximum.
- Maximum: the maximal value of the curve.
- Minimum: the minimal value of the curve.

- Number of cells: an estimate of the number of bumps of the curve (which is barely the number of cells in the particle).
- Time Of Flight (or impulsion length): the length of the curve.
- First: The first value of the curve
- Last: The last value of the curve

For several observations, some of the features were not available. These observations have then been removed from the Listmode files before model training.

S2 Hyperparameters space of the benchmark models

In order to use Bayesian hyper-optimization methods, one has to define a search space for the hyper-parameters. For each model, 30 combinations of parameters belonging to that space have been tested to keep an acceptable average running time for all models. We have used the Python implementations of all models presented in this work. The k-NN and LDA were implemented in the package scikit-learn, the LGBM in the lightgbm package, and the CNN in the Tensorflow and Keras packages. The search spaces specified for each model are given below and the argument names correspond to the denominations in the corresponding packages. Unless specified, uniform distributions (continuous or discrete) were used for all the parameters.

k-NN:

- n_neighbors = $[1, 50]$,
- weights = {'uniform','distance'},
- algorithm = {'ball_tree', 'kd_tree', 'brute'},
- p = $\{1, 2\}$.

LDA:

- solver = {'lsqr', 'eigen', 'svd'},
- shrinkage = $[0, 1]$,

- n_components = $[1, n\_classes - 1]$,
- tol = $[10^{-5}, 10^{-2}]$,
- priors: *in situ* cPFG relative abundances.

LGBM:

- learning_rate = $[10^{-3}, 10^{-2}]$,
- n_estimators = $[10, 1200]$,
- num_leaves = $\{6, 8, 12, 16\}$,
- boosting_type = {'gbdt', 'dart'},
- objective = 'binary',
- max_bin = $\{255, 510\}$,
- colsample_bytree = $\{0.64, 0.65, 0.66\}$,
- subsample = $\{0.7, 0.75\}$,
- reg_alpha = $\{1, 1.2\}$,
- reg_lambda = $\{1, 1.2, 1.4\}$,
- is_unbalance = {True, False}.
- class_weight = $\frac{1}{nb\_samples}$,

CNN

- early_stopping patience = 10,
- epochs = 120,
- loss = 'categorical_crossentropy',
- class_weight = $\frac{1}{nb\_samples}$,
- batch_size = $\{128, 256\}$,

- learning_rate = $[10^{-3}, 10^{-2}]$,
- optimizer = 'ranger'.

For the ranger optimizer the following sub-parameters have been explored:

- sync_period = $\{2, 6, 10\}$,
- $slow\_step\_size \sim \mathcal{N}(0.5, 0.1)$ with $\mathcal{N}(.)$ the Normal distribution.

S3 Additional Figures and Tables

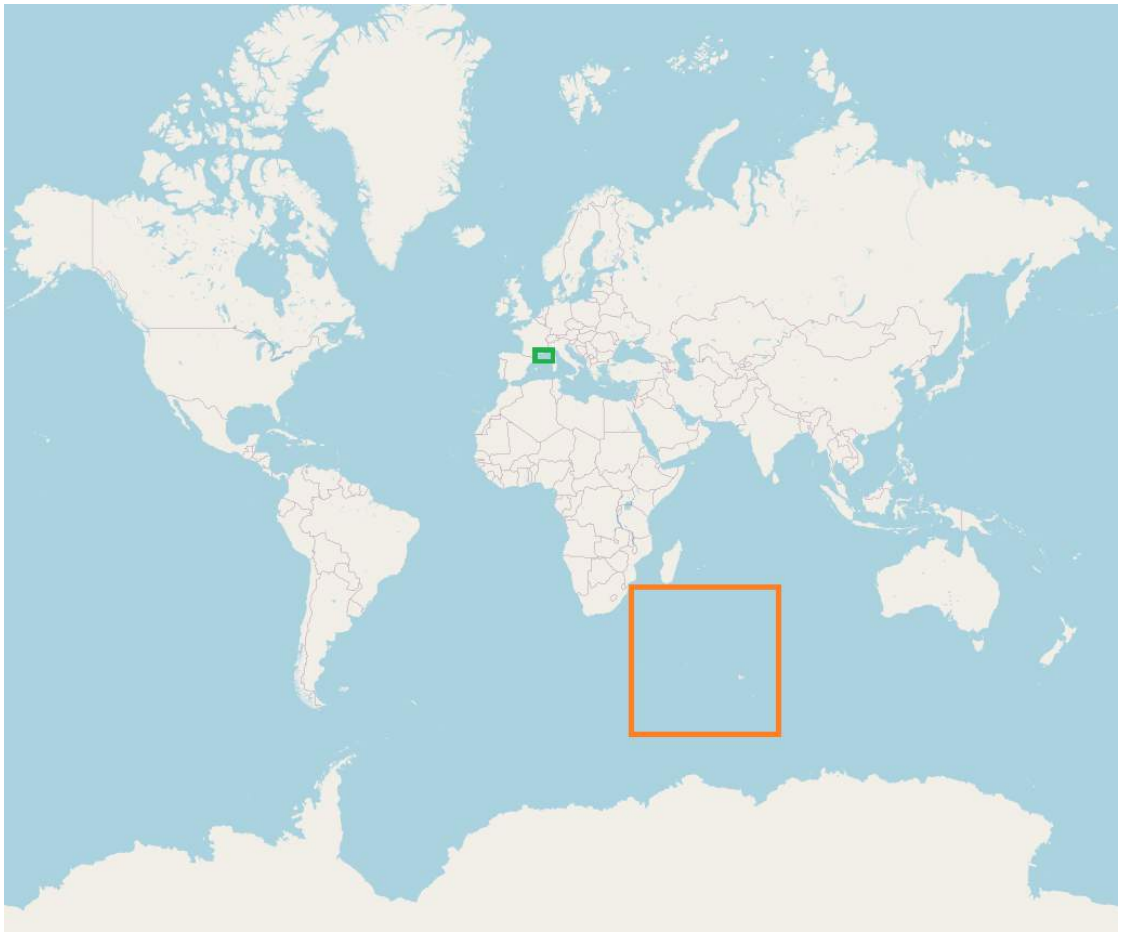


Figure S1: Locations of the samples collected. The green zone includes the SSLAMM station and the FUMSECK cruise areas while the orange zone refers to the SWINGS cruise area. Source: OpenStreetMap.

Interoperable nomenclature	**Expert suggested nomenclature**
Micro	Microphytoplankton
Orgnano	Cryptophytes-like
Orgpicopro	*Synechococcus*
Rednano	Nanoeukaryotes
Redpicoeuk	Picoeukaryotes
Redpicopro	*Prochlorococcus*

Table S1: Correspondence table between the SeaDataCloud Flow Cytometry Standardised Cluster Names, identified as the interoperable nomenclature and published by the Natural Environmental Research council, and the correspondence with an expert denomination.

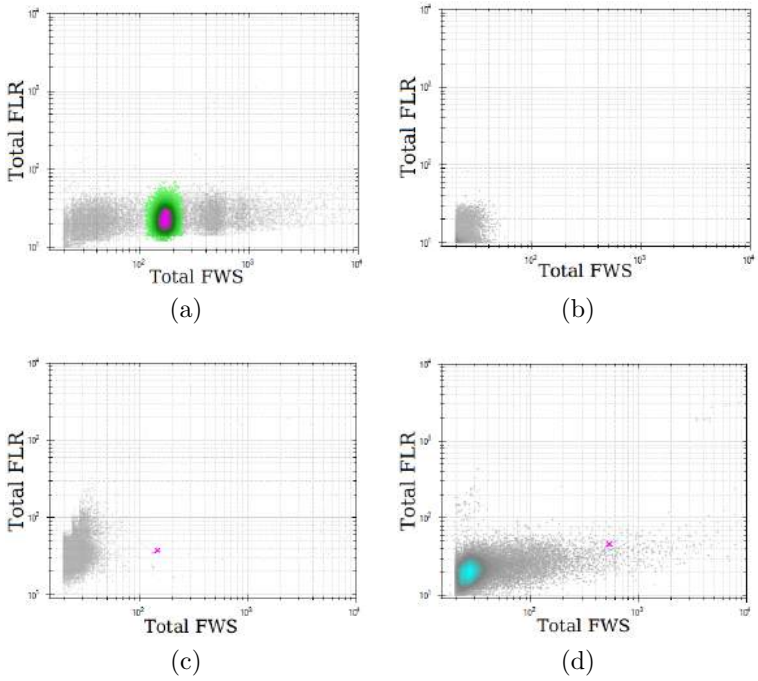
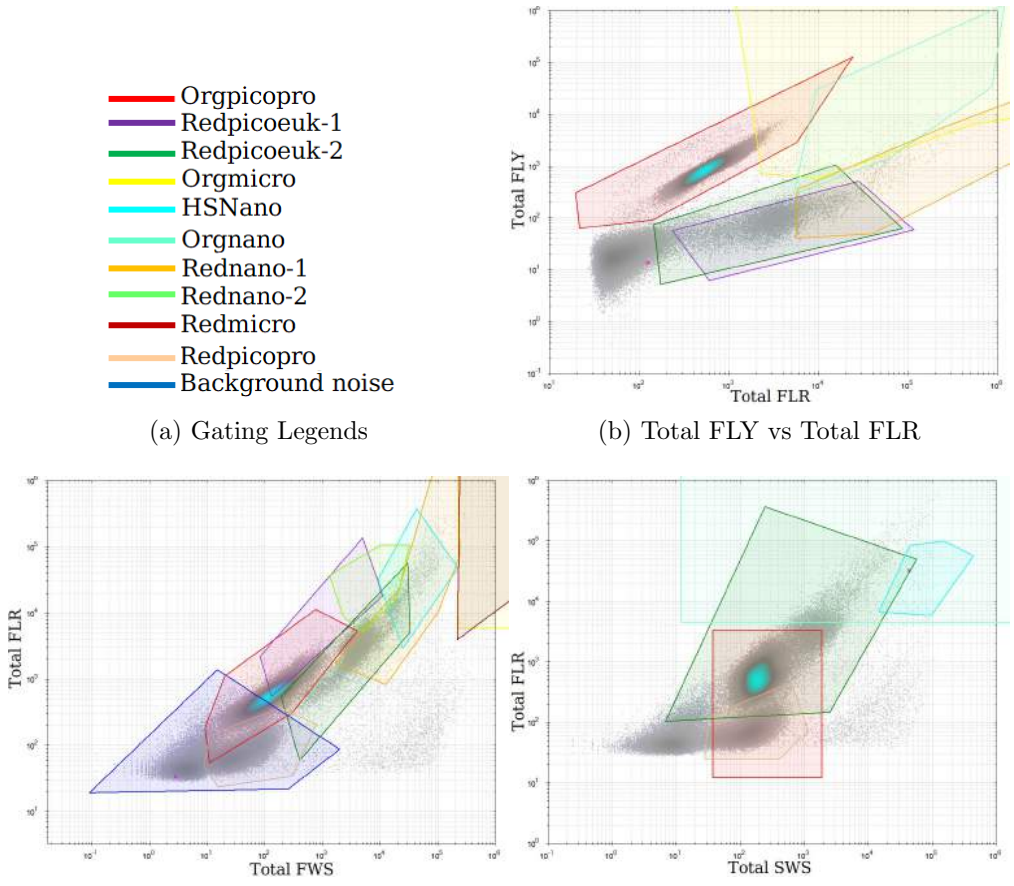


Figure S2: Illustration of negative control samples. 2D cytograms of total red fluorescence (a.u., Total FLR) vs. total forward scatter (a.u., Total FWS). (a) Cytogram of a solution of 1 µm silica beads diluted in ultra-pure water. The grey dots are noise particles. Green to pink dots correspond to the $1\mu m$ silica beads following a density gradient. The acquisition threshold was SWS9. (b) Acquisition of sheath liquid using similar acquisition settings as the *in situ* sampling protocol. The acquisition threshold was FLR6. (c) Acquisition of filtered seawater using a double polycarbonate $0.2\mu m$ syringe filter using similar acquisition settings as (b). (d) Acquisition of filtered seawater using a double polycarbonate $0.2\mu m$ syringe filter and similar acquisition settings as (a).



(a) Gating Legends

(b) Total FLY vs Total FLR

(c) Total FLR vs Total FWS

(d) Total FLR vs Total SWS

Figure S3: Illustration of the manual gating performed by an expert on a FLR 5 mV acquisition collected on 03/02/2021 at 6 pm during the SWINGS cruise. Gates were here created for each group listed in (a) on three 2D projections: (b) The total yellow fluorescence (a.u., Total FLY) as a function of the red fluorescence (a.u., Total FLR), (c) The total red fluorescence (a.u., Total FLR) as a function of the total forward scatter (a.u., Total FWS) and (d) The total red fluorescence (a.u., Total FLR) as a function of the total sideward scatter (a.u., Total SWS). Only the particles assigned by an expert to the same group in each 2D projection were assigned to this group. The other particles were regarded as unassigned particles along with the particles that did not belong to any gate. As each functional group was often composed of different sub-populations, they were gated separately and then gathered as a single functional group. In this study, Redpicoeuk-1, Redpicoeuk-2 were merged into the Redpicoeuk PFG, Orgmicro and Redmicro into the Micro PFG, the background noise and the unassigned particles were split according to their total FWS between $noise < 1\mu m$ and $noise \geq 1\mu m$, and the Rednano PFG was composed of the Rednano-1, Rednano-2, and HSnano cells.

8

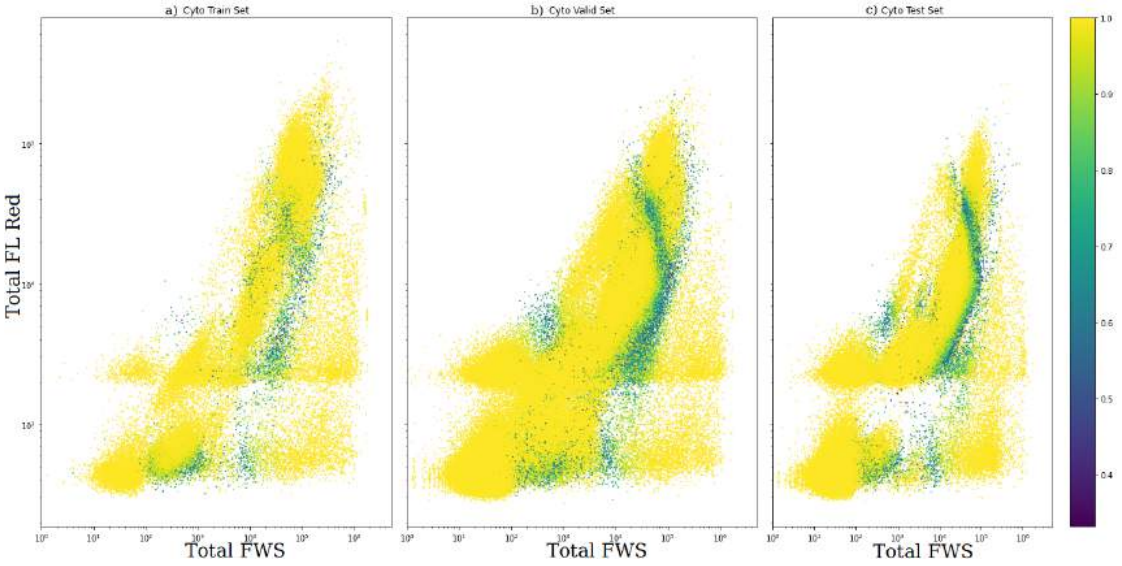


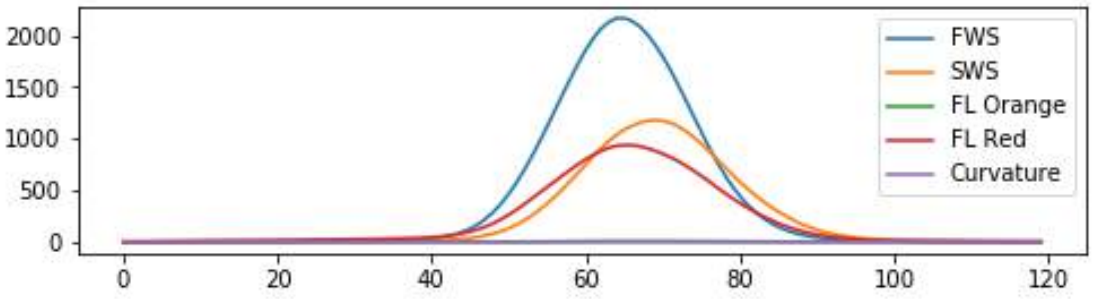
Figure S4: Total Forward scatter (a.u., Total FWS) vs. total red fluorescence (a.u., Total FL Red) cytograms representing the predictive confidence level of the CNN on the SWINGS train (a), validation (b), and test sets (c). The confidence level of the predicted class is the highest for the yellow dots and the lowest for the blue dots.

name	train	valid	test
Micro	779	134	446
Orgnano	1012	161	175
Orgpicopro	5500	17834	40598
Rednano	5500	2339	2060
Redpicoeuk	5000	8211	6868
Redpicopro	5000	1066	1372
Noise $< 1\mu m$	5500	13885	71097
Noise $\geq 1\mu m$	5500	7052	11697

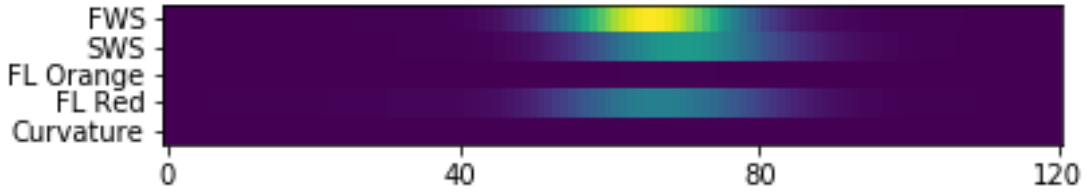
Table S2: Number of particles of each cPFG and noise classes in the train, validation, and test sets for SSLAMM data

name	train	valid	test
Micro	4947	774	219
Orgnano	5000	1264	119
Orgpicopro	7000	42538	15778
Rednano	8000	13391	6631
Redpicoeuk	8000	99096	83262
Redpicopro	8000	3841	7626
Noise $< 1\mu m$	8794	195994	103381
Noise $\geq 1\mu m$	7500	8965	7410

Table S3: Number of particles of each cPFG and noise classes in the train, validation, and test sets for SWINGS data



(a) Curves representation



(b) Matrix representation

Figure S5: Each cell passes in front of the AFCM laser beam and generates five flow cytometric curves (FCCs): Forward scatter (FWS), Sideward scatter (SWS), orange fluorescence (FL Orange), red fluorescence (FL Red) and Curvature of cell-size related length. These five FCCs are then interpolated quadratically to a fixed size of 120 values (a) using the Python scipy.interpolate.interp1D function and vertically stacked together as matrices containing the values of the five curves (b).

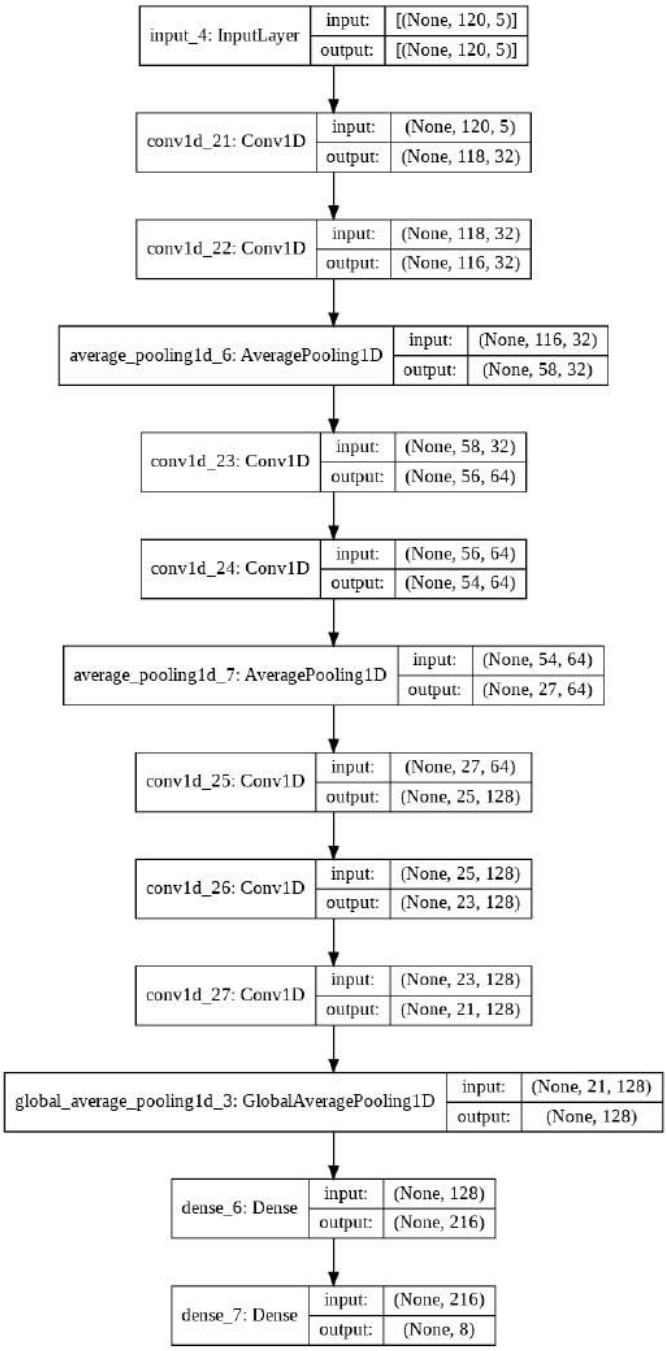
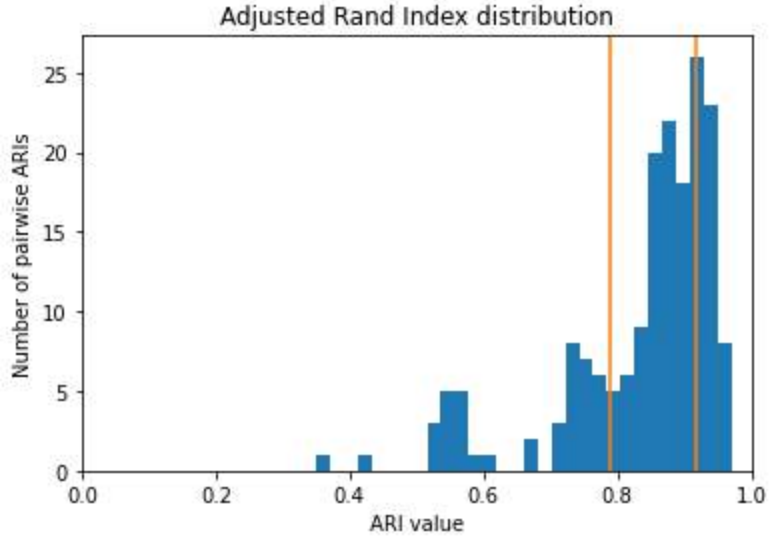
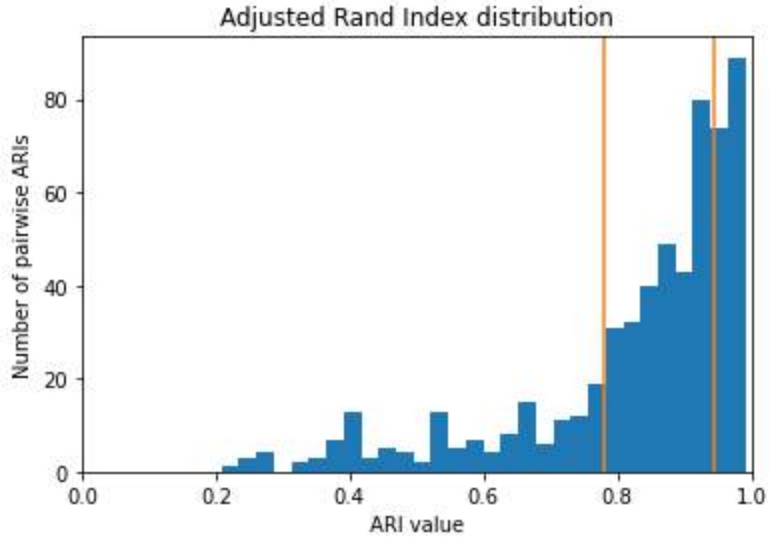


Figure S6: General architecture of the CNN used. By convention the first coordinate denoted by "None" corresponds to the batch size of the data.

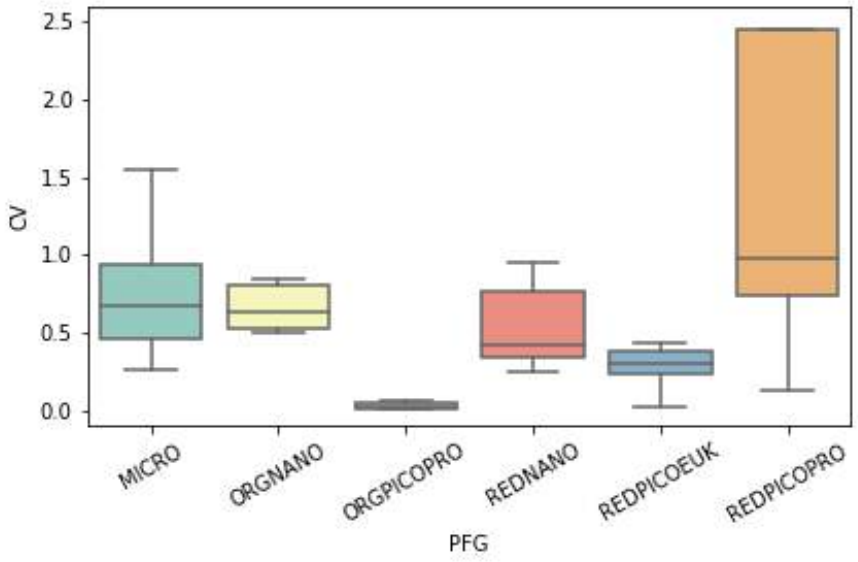


(a) SSLAMM

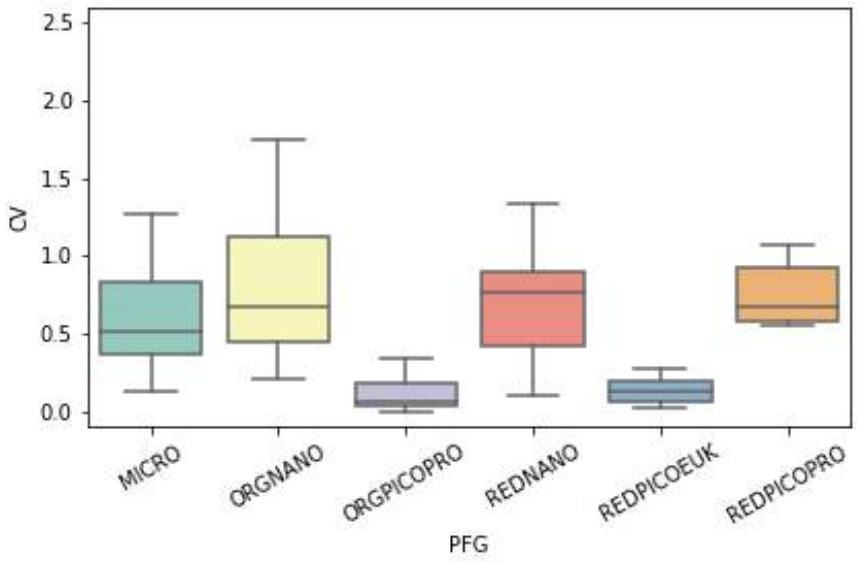


(b) SWINGS

Figure S7: Adjusted Rand Index on the SSLAMM manual gating data (a) and SWINGS manual gating data (b) for all files and all pairs of experts. The orange solid vertical lines represent the first and third quartiles of the distributions.



(a) SSLAMM



(b) SWINGS

Figure S8: Boxplots of the coefficients of variation (CV) of the manual experts counts per cPFG on the SSLAMM data, based on 6 acquisitions (a) and SWINGS data, based on 20 acquisitions (b). The low and high whiskers stand for the lowest and highest CVs per cPFG excluding outliers (defined as values greater than 1.5 times the inter-quartile range). The three bars of the box itself (bottom, central and top bars), represent the first quartile, the median and the third quartile of the cPFG CVs distribution, respectively.

Figure S9: Precision (a) and recall (b) (%) of the benchmarked models trained on the SWINGS data and used on SSLAMM data

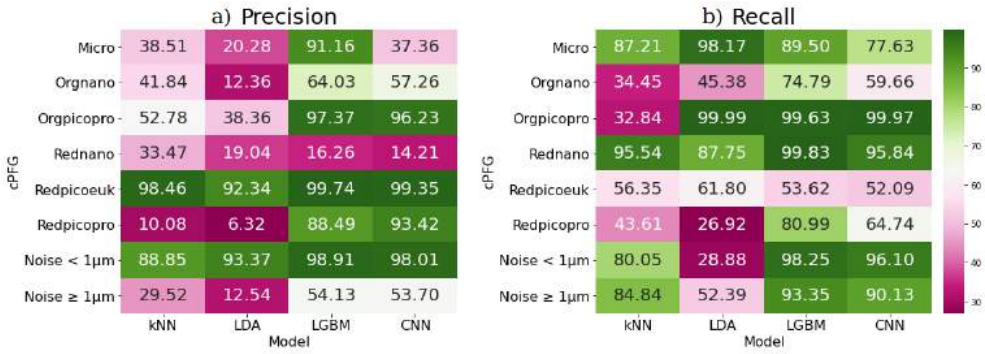


Figure S10: Precision (a) and recall (b) (%) of the benchmarked models trained on the SSLAMM data and used on SWINGS data

	kNN (CPU training)	LDA (CPU training)	LGBM (CPU training)	CNN (GPU training)	CNN (CPU training)
Training time	0.76	0.74	140.3	222.6	2281.32

Table S4: Running time in seconds of the specifications used to benchmark the models on SWINGS data. k-NN, LDA and LGBM scikit-learn implementations did not enable GPU training. However, they were trained using parallel CPU training (using all CPU cores).

S4 Datasets

Main data and models configurations necessary to reproduce the results of the study are available here: `https://erddap.osupytheas.fr/erddap/files/Automatic_recognition_CNN_material/`

E. GRL: Supplementary Material

Supporting Information for "Intermittent upwelling events trigger delayed, major, and reproducible pico-nanophytoplankton responses in coastal oligotrophic waters"

R. Fuchs[1,2] *, V. Rossi[2] †, C. Caille[3] ‡, N. Bensoussan[2] §, C. Pinazo[2] ¶, O. Grosso[2] ‖, M. Thyssen[2] ††

[1]Aix Marseille Univ, CNRS, Centrale Marseille, I2M, Marseille, France

[2]Aix Marseille Univ, Université de Toulon, CNRS, IRD, MIO, Marseille, France

[3]Sorbonne Université, CNRS, LOMIC, Banyuls-sur-Mer, France

Contents of this file

1. Material and Methods details

---

*robin.fuchs@univ-amu.fr

†vincent.rossi@mio.osupytheas.fr

‡caillec@obs-banyuls.fr

§nathaniel.bensoussan@mio.osupytheas.fr

¶christel.pinazo@mio.osupytheas.fr

‖olivier.grosso@mio.osupytheas.fr

†† melilotus.thyssen@mio.osupytheas.fr

June 1, 2022, 12:38pm

2. Figures S1 to S10

3. Tables S1 to S3

1. Materials and Methods Details

1.1. Stratified Period, Bloom Period and Salinity Data

1.1.1. Stratified periods characterization

The stratified period and the temperature anomalies were computed using a Butterworth digital and analog filter design (function butter of the Python "scipy.signal" subpackage). The bandwidth parameter was set to 60 days for the stratified periods determination and 15 days for the temperature anomaly. Events associated with temperature anomalies lasting less than eight hours were not considered.

1.1.2. Spring Bloom Periods Characterization

The dates of the spring bloom were determined using the threshold method (Sapiano et al., 2012; Brody et al., 2013) on the low-pass filtered biomass with a 5% threshold. The dates of the blooms in 2020 were from April 2 to April 30, 2020. There were two spring blooms in 2021, from March 25 to April 7 and from April 21 to May 12 (See Figures S4 and S5).

1.1.3. Salinity Data

The salinity data were acquired every hour using an STPS sensor from the NKE-manufacturer. Yet, salinity measurements from the STPS sensor were found not reliable and hence not used here.

1.2. Estimations of Phytoplankton Biovolume, Biomass and Growth Rates

1.2.1. Phytoplankton functional groups acquisition protocol summary

Phytoplankton organisms present significant differences in typical sizes and abundances (Finkel et al., 2010) so that two AFCM acquisition procedures are used to overcome this issue (as for example in Marrec et al. (2018)). Redpicopro and Orgpicopro pulse shape signals were acquired by setting a low red fluorescence threshold (6 mV) and by analyzing a volume of $850\mu L$ on average whereas the Redpicoeuk, Rednano, and Orgnano pulse shape signals were acquired using a high red fluorescence threshold (25 mV) and by analyzing volumes of $4000\mu L$ on average. The volume analyzed was quantified using a weight-calibrated peristaltic pump.

1.2.2. Biovolume estimation:

The biovolume of each phytoplankton cell was estimated using the relationship between AFCM Total forward scatter (the area under the FWS pulse shape) and the biovolume of Silica Beads and cell images taken by the AFCM (Figure S1). The Silica Beads were manufactured with a known size and the cell biovolumes from images were estimated according to Sun and Liu (2003). Even if the relationship existing between these two quantities is monotonic, its shape seemed not to be constant over all the possible Total FWS values. This pattern is due to the optical properties of the phytoplankton cell sizes relatively to the laser size. Indeed, for the cells exhibiting a Total FWS inferior to 2×10^2 a.u. the relationship seemed concave whereas it was convex for cells with Total FWS superior to 5×10^2 a.u. as made visible in Figure S1.

1.2.3. Biomass estimation:

The biomass of each cell was computed from the estimated biovolume (BV) using the following relationships:

• $Biomass = 0.260 \times BV^{0.860}$ for Redpicopro, Orgpicopro and Redpicoeuk cells according to Menden-Deuer and Lessard (2000).

• $Biomass = 0.433 \times BV^{0.863}$ for Rednano and Orgnano cells according to Verity et al. (1992).

1.3. Size-structured matrix population model

The size-structured model version introduced in Ribalet et al. (2015) was used. The corresponding code is available at `https://github.com/fribalet/ssPopModel` (version 1.1.0). By definition, the model is structured in size and the user has to define the number of classes along with a lower and upper bound of possible size for each PFG. In this study, the distribution of each PFG was discretized in 31 classes. The lower and upper bounds of a PFG size class were determined as the 1 over 1000 quantile and 999 over 1000 quantile of the PFG biovolume distribution during each SWUE, respectively. It prevented integrating outliers and avoided excluding a significant number of observations. The PFG data were linearly interpolated from a two-hour frequency to a one-hour frequency. The lightning data used by the model came from the MESURHO buoy (Cadiou et al., 2010) moored at the Rhone river mouth which is located about 40 kilometers away from the SSL@MM station. It provided the Photosynthetically Available Radiation data (PAR, $\mu E.m^{-2}.s^{-1}$) on a two hours basis. The PAR data were linearly interpolated on a 10 minutes basis. The PAR data were not available in 2021 due to a technical issue on the buoy and the growth rates were only calculated for 2019 and 2020.

1.4. PFG response identification

The rupture detection was conducted thanks to the Python "rupture" package: `https://github.com/deepcharles/ruptures`. A linear cost function with intercept was used to model the link between the water temperature and each PFG abundance or biomass signal. No observation subsampling was performed and a binary segmentation research method was used to minimize the cost function. As the goal was to identify the beginning and end of each PFG reaction, the number of rupture points was known and equal to two.

1.5. Computation of the additional biomass imputable to the Spring Bloom

The additional biomass generated between the start and the end of the bloom was computed by taking the median value over the preceding week before the bloom as a reference value. It was assumed that the biomass would have remained at this level during the whole period if the bloom did not occur. As a result, the daily additional biomass imputable to the bloom was computed as the difference between the actual total integrated biomass and the integrated reference level divided by the bloom duration in days.

2. Wind-driven Upwelling/Downwelling Index

The Wind-driven Upwelling/Downwelling Index is an hourly index that uses the sea surface wind speed and direction to estimate the Ekman transport perpendicular to the coastline (Bakun, 1973). A positive index value implies that surface waters are transported offshore (due to upwelling-favorable winds); conversely, a negative index value indicates that surface waters flow onshore (denoting wind favorable to downwelling events). An

upwelling event is a series of consecutive positive WUDI values. As in Odic, Bensoussan, Pinazo, Taupier-Letage, and Rossi (2022), events with average indices higher than $0.432 m^3.s^{-1}m^{-1}$ were considered as significant upwelling events. These events are associated with substantial changes in surface water temperature (more than 3°C on average, see Odic et al. (2022)), suggesting also measurable responses of both biogeochemistry (nutrients) and biology (phytoplankton). Events are considered distinct if they are separated from each other by at least one day (Millot, 1979).

References

Bakun, A. (1973). Coastal upwelling indices, west coast of north america, 1946-71. *NOAA technical report*.

Brody, S. R., Lozier, M. S., & Dunne, J. P. (2013). A comparison of methods to determine phytoplankton bloom initiation. *Journal of Geophysical Research: Oceans*, *118*(5), 2345–2357.

Cadiou, J.-F., Repecaud, M., Arnaud, M., Rabouille, C., Raimbaud, P., Radakovitch, O., ... Gaufrès, P. (2010). Mesurho: a high frequency oceanographic buoy at the rhone river mouth. In *39th ciesm congress-venice, italy, 10-14 may 2010.*

Finkel, Z. V., Beardall, J., Flynn, K. J., Quigg, A., Rees, T. A. V., & Raven, J. A. (2010). Phytoplankton in a changing world: cell size and elemental stoichiometry. *Journal of plankton research*, *32*(1), 119–137.

Marrec, P., Grégori, G., Doglioli, A. M., Dugenne, M., Della Penna, A., Bhairy, N., ... Thyssen, M. (2018). Coupling physics and biogeochemistry thanks to high-resolution observations of the phytoplankton community structure in the northwestern mediterranean sea. *Biogeosciences*, *15*(5), 1579–1606. Retrieved from `https://bg.copernicus.org/articles/15/1579/2018/` doi: 10.5194/bg-15-1579-2018

Menden-Deuer, S., & Lessard, E. J. (2000). Carbon to volume relationships for dinoflagellates, diatoms, and other protist plankton. *Limnology and oceanography*, *45*(3), 569–579.

Millot, C. (1979). Wind induced upwellings in the gulf of lions. *Oceanologica Acta*, *2*(3), 261–274.

Odic, R., Bensoussan, N., Pinazo, C., Taupier-Letage, I., & Rossi, V. (2022). Sporadic wind-driven upwelling/downwelling and associated cooling/warming along the north-west mediterranean coastlines. *(in prep.)*.

Ribalet, F., Swalwell, J., Clayton, S., Jiménez, V., Sudek, S., Lin, Y., ... Armbrust, E. V. (2015). Light-driven synchrony of prochlorococcus growth and mortality in the subtropical pacific gyre. *Proceedings of the National Academy of Sciences*, *112*(26), 8008–8012.

Sapiano, M., Brown, C., Schollaert Uz, S., & Vargas, M. (2012). Establishing a global climatology of marine phytoplankton phenological characteristics. *Journal of Geophysical Research: Oceans*, *117*(C8).

Sun, J., & Liu, D. (2003). Geometric models for calculating cell biovolume and surface area for phytoplankton. *Journal of plankton research*, *25*(11), 1331–1346.

Verity, P. G., Robertson, C. Y., Tronzo, C. R., Andrews, M. G., Nelson, J. R., & Sieracki, M. E. (1992). Relationships between cell volume and the carbon and nitrogen content of marine photosynthetic nanoplankton. *Limnology and Oceanography*, *37*(7), 1434–1446.

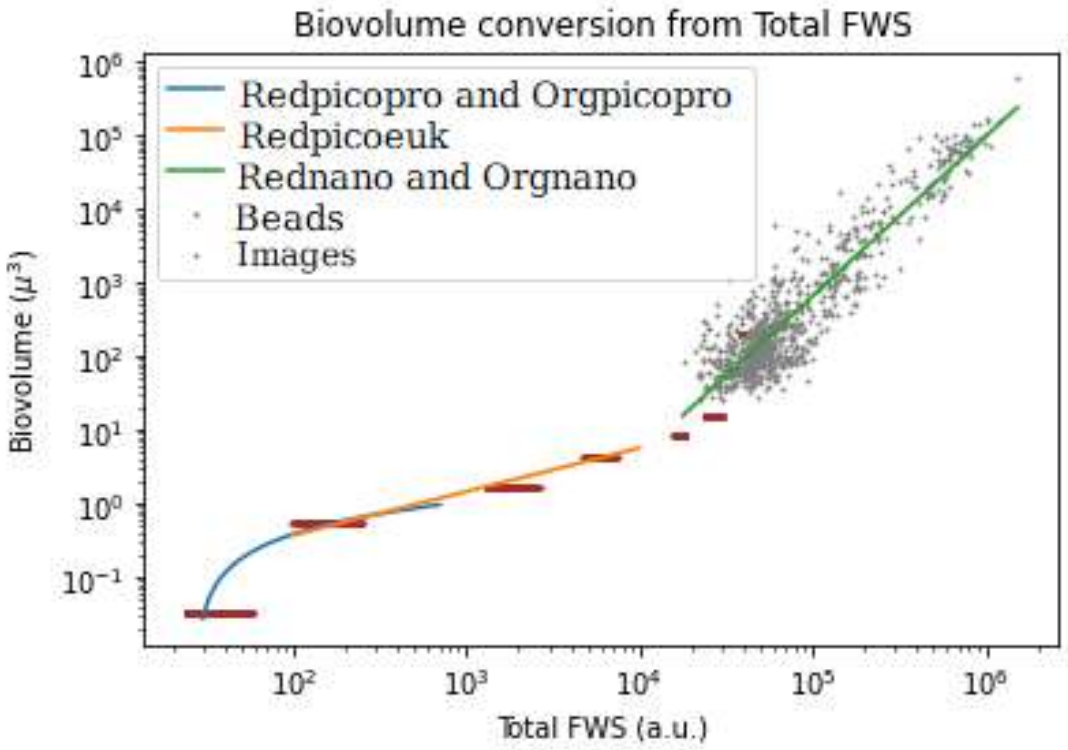


Figure S1. Summary of the empirical relationships used to convert the Total FWS signal of each cell into biovolume

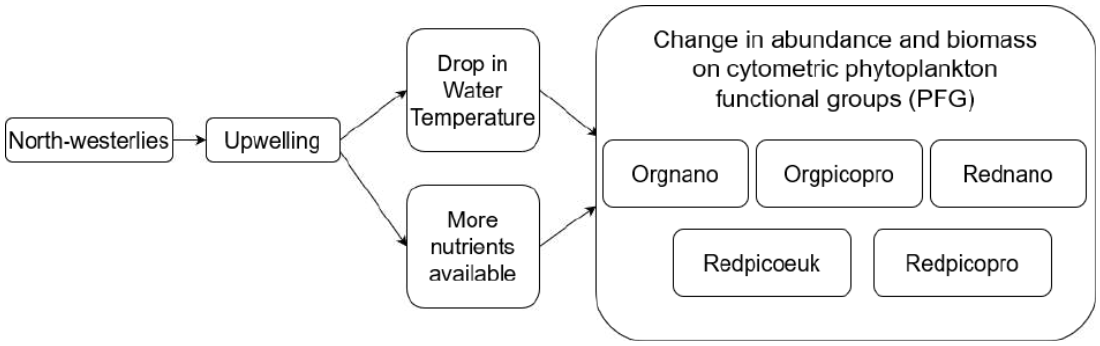


Figure S2. Summary of the causal relationships identified in this study. Reading of the underlying hydrodynamics: north-westerlies trigger offshore horizontal surface advection and upward vertical advection. Warm and nutrient-depleted surface water along with the associated phytoplankton (PFG) is exported offshore and replaced by deeper cold, nutrient-rich water, and the PFGs associated with these deeper water masses.

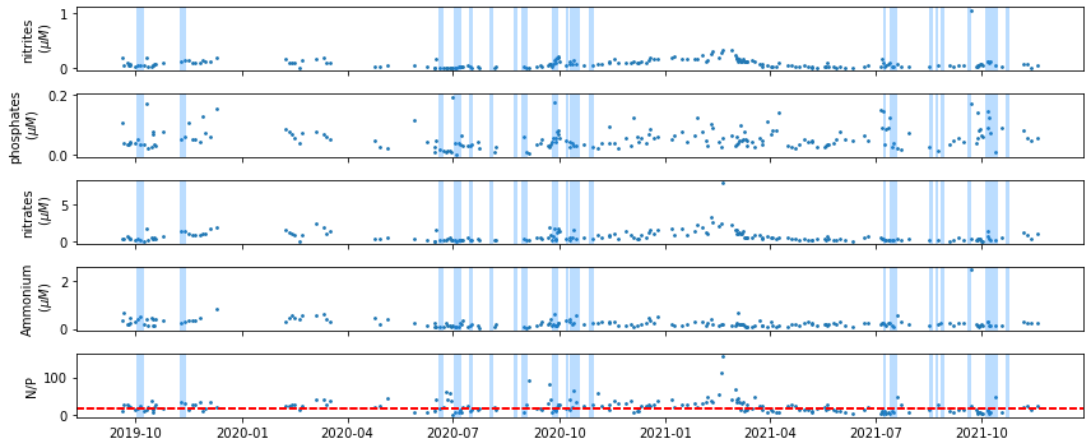


Figure S3. Nutrients over the two years of data. The colored rectangles correspond to the SWUEs considered in the study. The red dash line represent the N/P Redfield ratio (=16)

	Unstratified	Stratified Upwelling	Stratified Non SWUE
nitrites	0.10 (0.10)	0.05 (0.09)	0.03 (0.03)
phosphates	0.05 (0.03)	0.04 (0.04)	0.04 (0.04)
nitrates	0.90 (0.77)	0.26 (0.40)	0.36 (0.27)
Ammonium	0.22 (0.16)	0.20 (0.16)	0.19 (0.12)
N/P	25.15 (16.06)	17.33 (15.72)	13.06 (14.48)

Table S1. Medians and inter-quartile ranges (in parentheses) of the nutrients concentration (μM) for the nitrites, phosphates, nitrates, ammonium and N/P ratio during the unstratified periods, the SWUEs and unstratified period excluding SWUE.

	Unstratified	Stratified (SWUE reaction phase)	Stratified (Non SWUE)
Orgnano	4.03e-06 (6.74e-06)	2.55e-06 (2.87e-06)	3.66e-06 (5.23e-06)
Orgpicopro	1.55e-06 (2.38e-06)	2.16e-06 (1.55e-06)	3.12e-06 (2.49e-06)
Rednano	8.85e-06 (9.34e-06)	9.78e-06 (8.16e-06)	1.49e-05 (1.83e-05)
Redpicoeuk	1.64e-06 (2.11e-06)	9.37e-07 (1.03e-06)	6.52e-07 (6.88e-07)
Redpicopro	1.40e-07 (1.93e-07)	1.97e-07 (2.68e-07)	1.28e-07 (1.33e-07)

Table S2. Medians and inter-quartile ranges (in parentheses) of each PFG biomass (mgC.mL^{-1}) during the unstratified periods, the reaction of the PFG during SWUE, and in stratified periods outside of SWUEs.

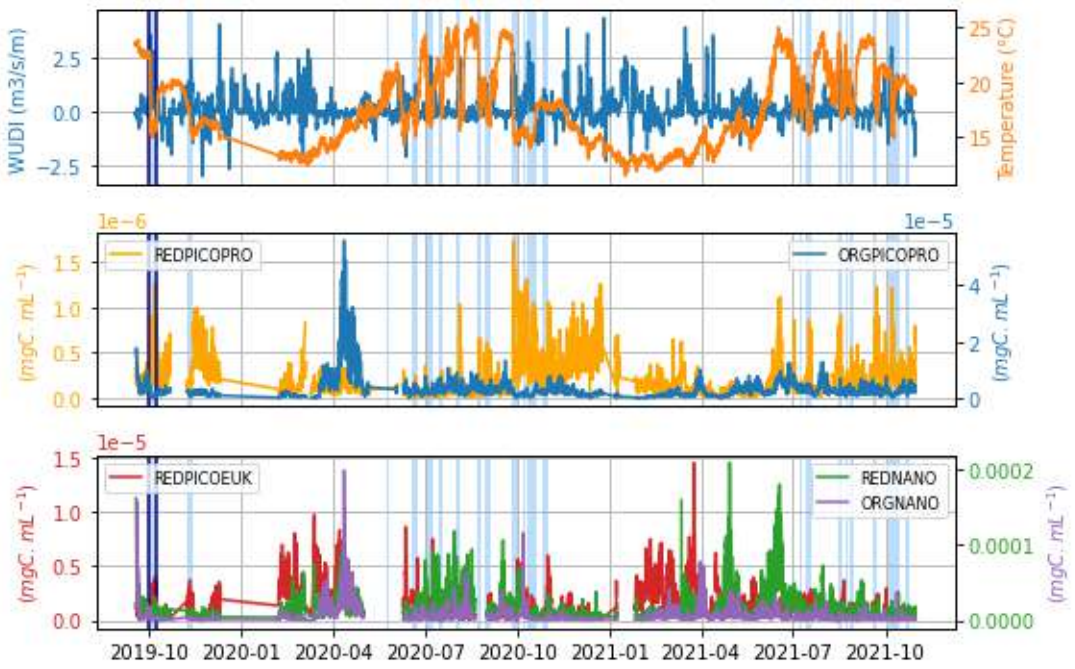


Figure S4. WUDI ($m^3.s^{-1}m^{-1}$) and temperature (C°) series (a), and phytoplankton biomass ($mgC.mL^{-1}$), at the SSL@MM station. The blue rectangles correspond to the studied SWUEs in the main text. The SWUE shown in Figure 2 in the main text is bounded by a dark blue box.

June 1, 2022, 12:38pm

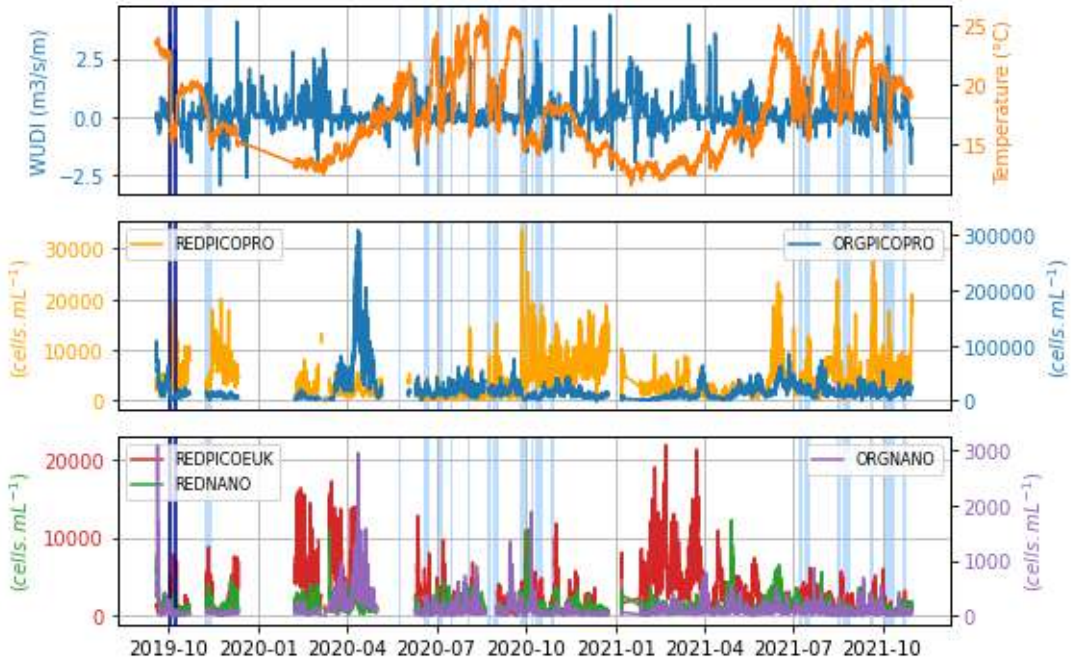


Figure S5. WUDI ($m^3.s^{-1}m^{-1}$) and temperature (C°) series (a), and phytoplankton abundances (cells.mL^{-1}), at the SSL@MM station. The blue rectangles correspond to the studied SWUEs in the main text. The SWUE shown in Figure 2 in the main text is bounded by a dark blue box.

June 1, 2022, 12:38pm

	Unstratified	Stratified (SWUE reaction phase)	Stratified (Non-SWUE)
Orgnano	69.21 (97.68)	58.19 (62.24)	77.73 (90.02)
Orgpicopro	8706.83 (14998.82)	13161.05 (9739.31)	18633.22 (15789.77)
Rednano	881.50 (853.54)	908.64 (634.03)	1052.81 (1013.43)
Redpicoeuk	2775.45 (4229.14)	1612.28 (1866.72)	997.38 (1019.65)
Redpicopro	2734.51 (3167.50)	4267.94 (5349.91)	2988.55 (3747.08)

Table S3. Medians and inter-quartile ranges (in parentheses) of each PFG abundance (cells.mL^{-1}) during the unstratified periods, the reaction of the PFG during the SWUEs, and in stratified periods outside of SWUEs.

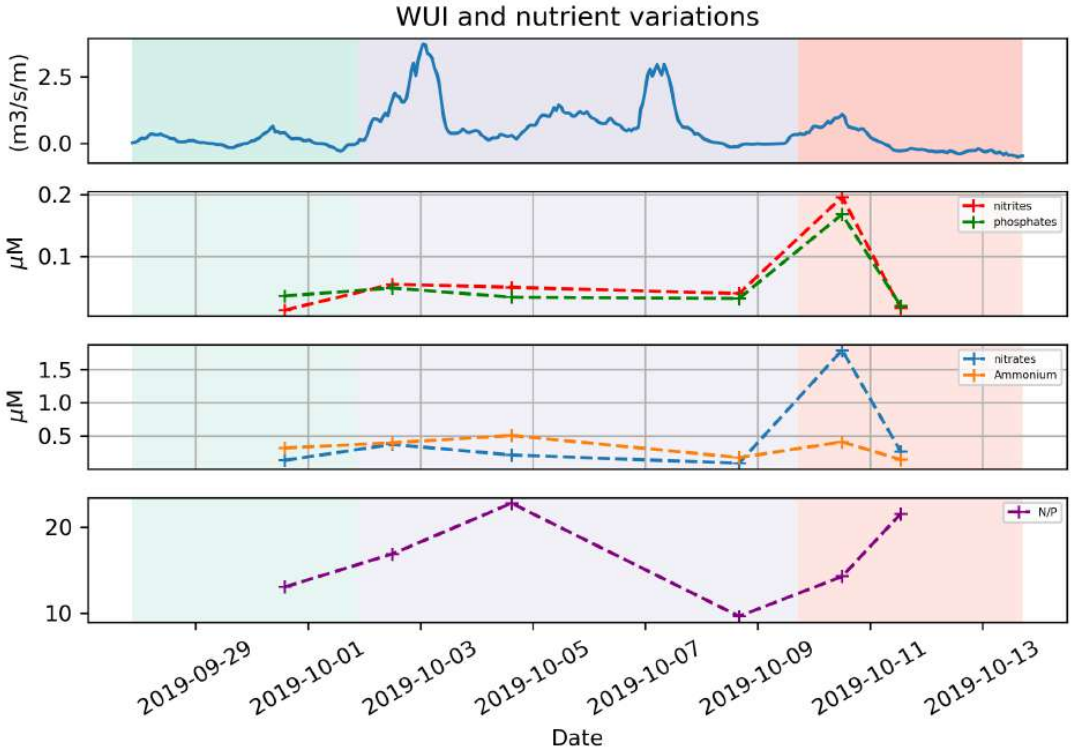


Figure S6. Nutrients and N/P ratio during the SWUE shown in Figure 2 in the main text.

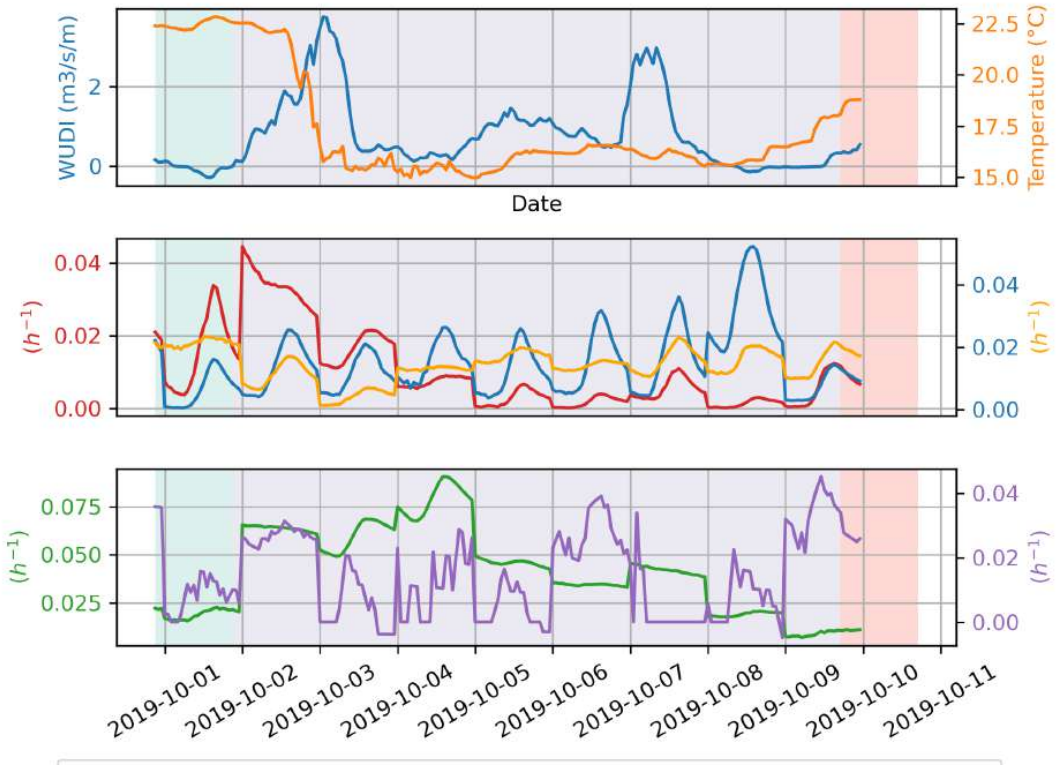


Figure S7. Hourly growth rates during the SWUE shown in Figure 2 in the main text.

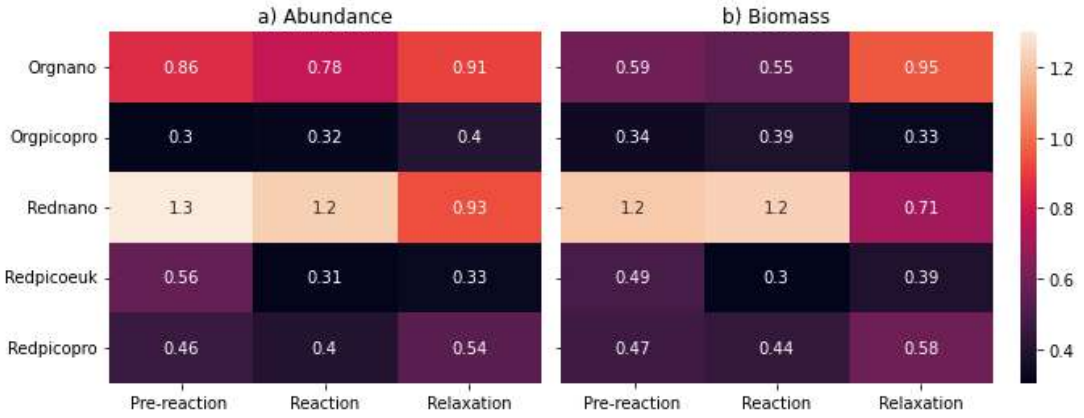


Figure S8. Estimated PFG daily growth rates during the three biological phases as defined by the abundance rupture points (a) or biomass rupture points (b). Only the Redpicoeuk growth rates significantly differed between the phases (for both abundance and biomass rupture points) and the Rednano using the biomass rupture points (Kruskal-Wallis test, p-value ≤ 0.05)

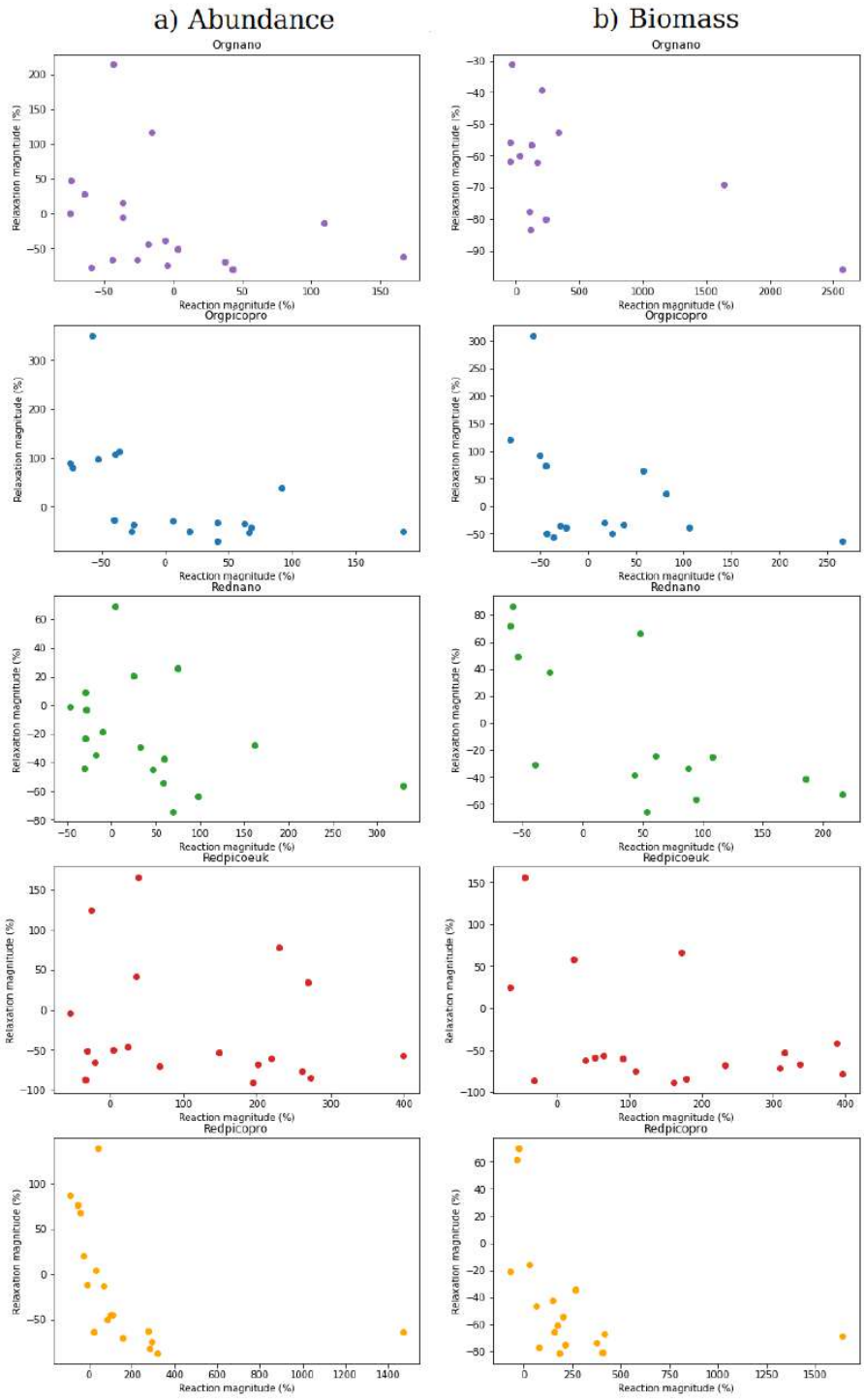


Figure S9. Inverse relationship existing between relaxation and reaction phases for all PFGs in both abundance (a) and biomass (b) illustrating a catch-up phenomenon.

June 1, 2022, 12:38pm

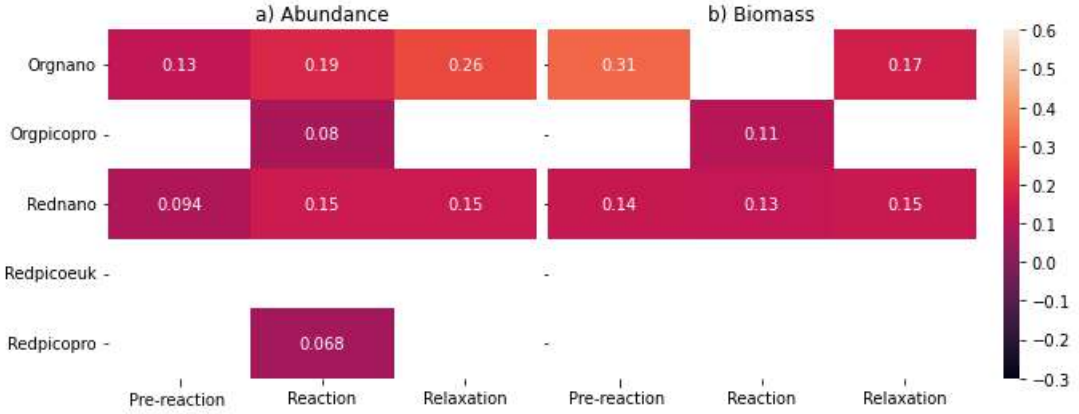


Figure S10. Spearman correlations between estimated growth and loss rates using the abundance (a) and biomass (b) rupture points for all PFG before their reaction, during their reaction and during their relaxation phase. Only correlations significant at 5% are displayed. The number of observations on which these correlations are computed is given in Figure 3 in the main text.