



# A statistical approach to unveil phytoplankton adaptation to ocean fronts

Théo Garcia<sup>1</sup>, Laurina Oms<sup>2</sup>, Xavier Milhaud<sup>1</sup>, Andrea M. Doglioli<sup>2</sup>, Monique Messié<sup>3</sup>,  
Pierre Vandekerckhove<sup>4</sup>, Claire Lacour<sup>4</sup>, Gérald Grégori<sup>2</sup>, and Denys Pommeret<sup>1</sup>

<sup>1</sup>Aix Marseille Univ., CNRS, I2M, Marseille, France

<sup>2</sup>Université Aix Marseille, Université de Toulon, CNRS, IRD, MIO, Marseille, France

<sup>3</sup>Monterey Bay Aquarium Research Institute, Moss Landing, CA, USA

<sup>4</sup>Université Gustave Eiffel, LAMA (UMR 8050), 77420 Champs-sur-Marne, France

**Correspondence:** Théo Garcia (theo.garcia@univ-amu.fr)

Received: 24 July 2025 – Revised: 12 November 2025 – Accepted: 12 January 2026 – Published: 30 January 2026

**Abstract.** Fine-scale oceanic fronts are ubiquitous and ephemeral physical features that separate contrasting water masses, creating significant heterogeneity in the physical seascape and plankton distributions. Because phytoplankton community composition (PCC) is a key driver of marine ecosystem functioning, understanding the extent to which fine-scale fronts influence PCC is a critical challenge. However, studying PCC across and within fronts is particularly difficult due to data scarcity and high biophysical variability. We developed a tailored statistical model to characterize PCC within an oceanic front we studied in the Mediterranean Sea. We modeled the frontal community as a finite mixture model with three components: two communities of adjacent water masses and a potential front-adapted community. Each component was further considered as a discrete mixture of an unknown number of multivariate Gaussian sub-components. First, we used an Expectation–Maximization algorithm to estimate the Gaussian parameters and determine the optimal number of sub-components based on in situ datasets of the PCC within a frontal zone and its adjacent water masses. Second, a hierarchical Bayesian approach was applied to estimate the weight of all components within the frontal dataset. Our analysis suggests that within the front a new community component, distinct from those in adjacent water masses, accounts for 70 % of the frontal community, indicating that a specific phytoplankton community can emerge in fine-scale oceanic fronts. Despite the limited number of frontal observations, our Bayesian modelling approach provides statistical evidence of the front's influence on phytoplankton community composition, effectively overcoming data scarcity and high variability.

## 1 Introduction

The oceanic seascape resembles a dynamic mosaic of contrasting water bodies, separated by boundaries known as fronts (Acha et al., 2015). Fine-scale fronts (1–100 km, day–weeks) arise from the interaction of water masses with distinct origins and characteristics (such as temperature and salinity) and are ubiquitous in the ocean (McWilliams, 2021). These fronts influence the environment from the surface to deeper ocean layers, impacting biogeochemical processes by modulating material transport; both by acting as horizontal barriers and by generating vertical fluxes (Mahadevan and Archer, 2000). In particular, upward currents can transport

nutrients from deeper layers, supporting enhanced biodiversity and biomass (Lévy et al., 2015; Clayton et al., 2017).

Among the biological communities affected by fine-scale frontal dynamics, phytoplankton are especially affected due to their limited motility. Phytoplankton communities (i.e. specific assemblages of taxa) form the base of the trophic chain, produce oxygen by photosynthesis and play a key role in the biogeochemical cycling of carbon, nitrogen, and phosphorus, thereby regulating marine ecosystem functioning and contributing to global climate processes (Litchman et al., 2010; Eggers et al., 2014). Significant heterogeneity in phytoplankton communities is observed throughout the ocean,

and key questions remain regarding the factors that shape their composition and the underlying drivers of their remarkable diversity (Sournia et al., 1991; Bianchi and Morri, 2000; Coll et al., 2010). One plausible hypothesis is that fronts delineate distinct habitats, thereby maintaining diversity by structuring species distributions and interactions. Given their potential impact on biological processes across the trophic chain, fine-scale fronts are a critical area of study. Moreover, fine-scale effects on biogeochemical cycles in the context of global warming are of great concern (Yang et al., 2023; Lévy et al., 2024).

The study of frontal phytoplankton communities needs dedicated cruises (Lévy et al., 2024) with high-frequency sampling to have enough data to perform robust statistical analysis. A few *in situ* studies suggested that fronts are either i) areas where environmental conditions allow the development of an inherent phytoplankton community (Taylor et al., 2012; Mangolte et al., 2023; Clifton Gray et al., 2024), or ii) simple boundaries between two contrasting water masses and their associated phytoplankton communities (Clayton et al., 2014; Mousig et al., 2016; Marrec et al., 2018; Tzortzis et al., 2021). However, these suggestions are hindered by significant challenges of obtaining *in situ* measurements within fine-scale fronts as they are small, short-lived, and difficult to track, leading to a lack of observations (Lévy et al., 2012). In addition, phytoplankton organisms respond rapidly to their environment (Collins et al., 2014), with large variations in abundance and biomass, which in turn result in highly variable datasets (i.e. non-Gaussian, skewed or multimodal distributions).

Consequently, a first key step lies in applying statistical analyses to scarce variable observations. When priors (i.e. assumed distribution before incorporating any data or observations) are properly defined, Bayesian statistics are known for their ability to capture signals even with scarce and highly variable data, providing reliable statistical inference even with small sample sizes (McNeish, 2016). A second key step in studying the phytoplankton community composition (hereafter “PCC”) in different areas, such as fronts and their adjacent water masses, is to manage the complexity of multidimensional datasets characterized here by different phytoplankton types. Gaussian mixture modelling (hereafter “GMM”) is used to model multiple signals that are assumed to follow normal distributions (McLachlan and Peel, 2000). By modelling multiple Gaussian components, GMM can model complex (i.e., non-Gaussian) distributions (Birgé, 1983). Originally introduced by Pearson (1894) to model heterogeneous biological data, GMM has since been widely applied in oceanography, for example to analyse krill cohort dynamics (Shaw et al., 2021) and phytoplankton classification (Hyrkas et al., 2016). Phytoplankton communities consist of different groups (e.g., cyanobacteria, picophytoplankton constituted by cells between 0.2 and 2–3  $\mu\text{m}$  in size, nanophytoplankton constituted by cells between 2–3 and 20  $\mu\text{m}$  in size, etc.). Applying GMM to a multivariate dataset would

help quantify the ecological signals of different phytoplankton communities (i.e., bloom of a specific group, a response to a nutrient pulse), especially when dealing with complex, overlapping distribution of groups. This is useful when the assumptions of traditional statistical tests are not met in the case of a mixture (e.g. normality and homoscedasticity).

Our study focuses on the Mediterranean Sea due to its combination of moderately energetic physical processes and oligotrophic conditions, which resemble those in the global ocean (Bethoux et al., 1999). We build on the previous study by Tzortzis et al. (2021) conducted south of Balearic islands where a front separating Atlantic waters recently entering the Mediterranean from saltier surface waters of the western Mediterranean was observed. That study demonstrated that this front plays a significant role in the structuring of PCC by segregating different classes of phytoplankton sizes between the two adjacent water masses, resulting in two distinct communities. However, a potential front-adapted phytoplankton community could not be identified due to *in situ* sampling limitations leading to a small number of observations in the front.

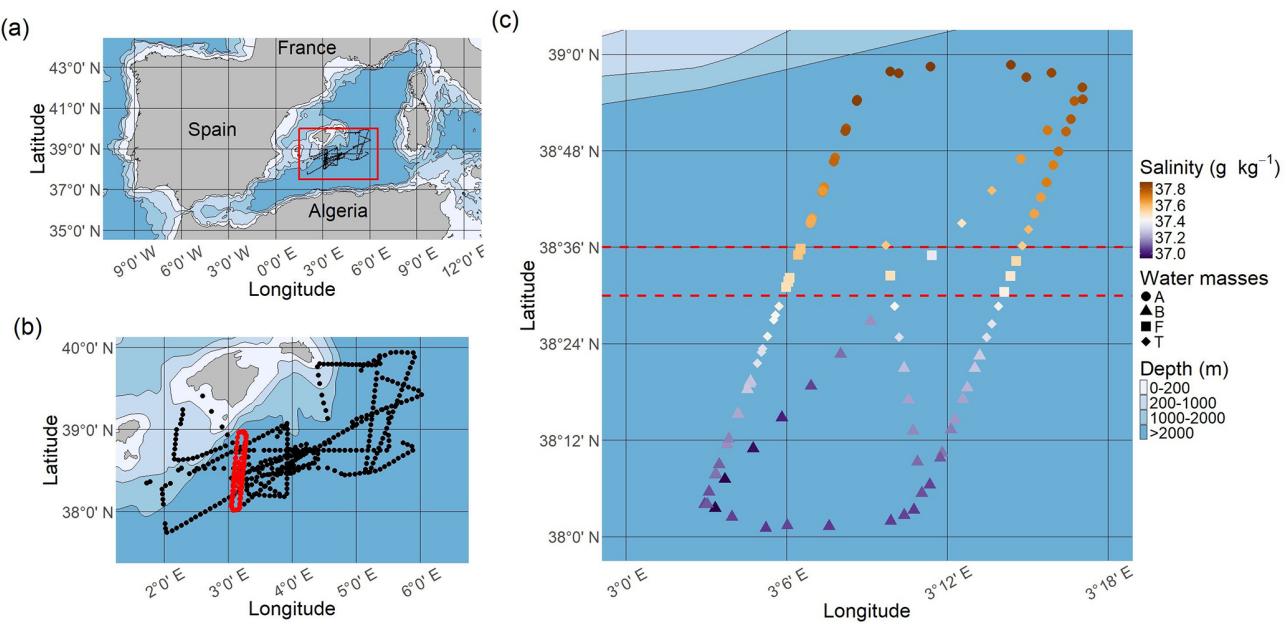
In this article we developed a statistical approach combining GMM and Bayesian methods that allowed us to estimate the presence of communities using the phytoplankton biomass data within the well-defined physical frontal region previously studied by Tzortzis et al. (2021). This provides a novel methodological framework for investigating the complex interactions between fine-scale physical and biological seascapes, while accounting for the challenges of obtaining data at such scales. We ask the following questions: *What is the structure of the community that might be formed at the front? Is there a frontal community as a mixture, where the expected community results from the combination of the adjacent water communities, or is there another community resulting from the intrinsic frontal characteristics?* Answering these questions would provide valuable insights into the role of fine-scale oceanic fronts in the distribution of marine biodiversity. This is particularly important in frontal areas where observations are rare.

This article is structured as follows: In Sect. 2, we describe the data, followed by the modelling approach in Sect. 3. In Sect. 4, we present the results, which are discussed in Sect. 5. Finally, we conclude the study in Sect. 6.

## 2 Study area and data collection

### 2.1 Cruise strategy and hydrology

During the PROTEVSMED-SWOT campaign (May 2018, south of the Balearic Islands, Dumas, 2018), we implemented a sampling strategy to cross a frontal zone separating two distinct water masses several times, with a North-South, “hippodrome” shaped route (hereinafter NS-Hippodrome, Fig. 1) (Tzortzis et al., 2021). High-resolution physical and biological surface measurements were collected using a CTD



**Figure 1.** Study area and sampling strategy. (a) Map of the north-western Mediterranean Sea; the red rectangle corresponds to the PROTEVSMED-SWOT study area. (b) Map of the sampling area; the black dots correspond to cytometric samples collected throughout the entire cruise (*outside dataset*) and the red dots to samples from the NS-Hippodrome. (c) Absolute salinity distribution ( $\text{g kg}^{-1}$ ) along the NS-Hippodrome. The shape of the dots depends on the water masses they belong to. The dashed red lines correspond to the latitudinal limits of the front. Note that to highlight the sampling points, no geographic projection is used in this panel. Maps were produced using Natural Earth open access data.

sensor mounted on a towed vehicle, a thermosalinograph (TSG) and an automated flow cytometer installed on the surface water intake of the TSG circuit. By employing an adaptive Lagrangian sampling strategy, we tracked physical and biological structures in both space and time, identifying a fine-scale frontal zone separating the two water masses *A* and *B*, each characterized by contrasting abundances of nine phytoplankton groups defined by flow cytometry (Tzortzis et al., 2021). To capture the phytoplankton diel cycle, both water masses were continuously sampled along the NS-Hippodrome from 11 to 13 May, 2018. This approach allowed us to capture the diel cycle in both *A* and *B* similarly, reducing biases from cell size and division. As a result, any differences in cell abundance between the *A* and *B* water masses are not related to diel cycle variations, making the observations independent and identically distributed (hereafter i.i.d.).

Based on extensive analysis of temperature and salinity in the water column and across the zone, Tzortzis et al. (2021) characterized the frontal area around  $38.32^\circ\text{N}$ . They stated that surface salinity was a good marker for the water masses in the visited area (Fig. 1c). The salinity gradient during the cruise (see Fig. A1) indicates that water mass *A* is characterized by a salinity  $\geq 37.6 \text{ g kg}^{-1}$ , and water mass *B* by a salinity  $\leq 37.3 \text{ g kg}^{-1}$ . The front *F* lies between these two isohalines. The frontal zone definition also followed a geographic criteria,  $38.36^\circ\text{N} \geq \text{Latitude} \geq 38.30^\circ\text{N}$ , cor-

responding to the measurement locations (Tzortzis et al., 2021). To focus on the frontal zone, measurement points that lay within the salinity range of the front but outside of its geographical boundaries were considered as part of a transitional zone (*T*) and were not used for the data analysis. In total, 30 samples were collected in *A*, 44 in *B*, 11 in *F* and 17 in *T*.

## 2.2 Flow cytometry

Automated flow cytometry enables high-frequency seawater sampling and analysis to identify phytoplankton groups based on their optical scattering and fluorescence properties (Dubelaar et al., 1989; Thyssen et al., 2009, 2015). The Cytosense flow cytometer (Cytobuoy b.v., Netherlands) uses a sheath fluid of  $0.1 \mu\text{m}$  filtered seawater to align and guide individual particles (cells) through a  $488 \text{ nm}$  laser beam. As cells interact with the laser beam, multiple optical signals are simultaneously recorded for each particle (cell).

First, forward scatter (*FWS*) and sideward scatter (*SWS*) are measured, providing insights into particle size, shape, and granularity. Second, fluorescence signals from photosynthetic pigments are also detected using photomultiplier tubes: red fluorescence (*FLR*) from chlorophyll and orange (*FLO*) fluorescence from phycoerythrin. Sequential protocols are run sequentially every 30 min, to analyse samples by phytoplankton size class. The first protocol (*FLR6*) had a *FLR*

trigger threshold fixed at 6 mV and could analyze a volume of 1.5 cm<sup>3</sup>. It was dedicated to the analysis of the picophytoplankton (< 2 µm). The second protocol (FLR25) targeted nanophytoplankton and microphytoplankton (> 2 µm) with a FLR trigger level set at 25 mV and an analysed volume of 4 cm<sup>3</sup>.

Data acquisition was performed using CytoUSB software (Cytobuoy) and analyzed with CytoClus (Cytobuoy). The cytometer produces 2D cytograms, graphical representations that plot individual particles according to their optical signals, highlighting distinct populations based on scattering and fluorescence properties. Within these dot clouds, we manually identified clusters that serve as proxies for functional phytoplankton group (Peeters et al., 1989; Thyssen et al., 2008). CytoClus provides cell abundances (cells cm<sup>-3</sup>) and mean optical signal intensities for each phytoplankton group.

Nine phytoplankton groups were identified (Tzortzis et al., 2021): one cyanobacterial group, *Synechococcus* (Syne, 1,µm); four picoeukaryote groups (Pico1, Pico2, Pico3, PicoHFLR, 0.2–2 µm); two nanoplankton groups (SNano, RNano, 2–20 µm), cryptophytes (Crypto, 10–50 µm); and one microphytoplankton group (Micro, 20–200 µm). Abundances (number of cells) are converted into carbon biomass (mmolC m<sup>-3</sup>) using allometric relationships described in Tzortzis et al. (2023) and Oms et al. (2024). Importantly, there are huge size, biomass and abundance contrasts between the nine phytoplankton groups (Figs. A2 and A3).

### 3 Methods

#### 3.1 Model formulation

We denote by Com the random vector characterizing a community. It is composed of the biomass of the 9 phytoplankton groups described previously. We assume that the biomass distribution of the community in the front, denoted by Com<sup>F</sup>, is mathematically described as a finite discrete mixture of three random components corresponding, respectively, to the communities of water masses A (Com<sup>A</sup>) and B (Com<sup>B</sup>) and an unknown community (Com<sup>C</sup>), as follows:

$$\text{Com}^F = \mathbb{I}_{U=A} \text{Com}^A + \mathbb{I}_{U=B} \text{Com}^B + \mathbb{I}_{U=C} \text{Com}^C, \quad (1)$$

where, for any generic condition  $\mathcal{T}$ , the indicator function is defined as:

$$\mathbb{I}_{\mathcal{T}} = \begin{cases} 1 & \text{if condition } \mathcal{T} \text{ is satisfied,} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Here,  $U$  is an unobserved categorical random variable taking values in the set  $\{A, B, C\}$ , with the following probabilities:

$$P(U = A) = \lambda_A, P(U = B) = \lambda_B, P(U = C) = \lambda_C, \quad \text{with } \lambda_A + \lambda_B + \lambda_C = 1. \quad (3)$$

Each observation is thus assumed to originate from one of the three communities  $A$ ,  $B$ , or  $C$ , with respective weights  $\lambda_A$ ,  $\lambda_B$ , and  $\lambda_C$ . The case where  $\lambda_C$  is significantly different from 0 indicates the presence of the new community  $C$ , while  $\lambda_C \rightarrow 0$  corresponds to a mixture involving only communities  $A$  and  $B$ .

The collected observations (i.i.d., cf. Sect. 2.1) of Com<sup>F</sup> is a table of dimension  $n \times m$ , where  $n = 11$  biomass measurements and  $m = 9$  phytoplankton groups.

Multivariate normal distribution of  $A$ ,  $B$ ,  $F$  and  $T$  were assessed with the Henze–Zirkler's test (Korkmaz et al., 2014). Except for  $A$ , the empirical biomass distribution in  $B$ ,  $F$  and  $T$  revealed non-Gaussian shapes associated with high variability for each phytoplankton group (see Fig. A2), suggesting underlying mixture structures of Gaussian distributions. Given this, we proposed that Com<sup>A</sup>, Com<sup>B</sup> and Com<sup>C</sup>, are themselves issued from a mixture of, respectively,  $j$ ,  $k$  and  $l$  multivariate Gaussian components, which model potential sub-communities within  $A$ ,  $B$  and  $C$  as follows:

$$\text{Com}^A \sim \sum_{r=1}^j \alpha_{Ar} \cdot \mathcal{N}(\boldsymbol{\mu}_{Ar}, \boldsymbol{\Sigma}_{Ar}), \quad (4)$$

$$\text{Com}^B \sim \sum_{s=1}^k \alpha_{Bs} \cdot \mathcal{N}(\boldsymbol{\mu}_{Bs}, \boldsymbol{\Sigma}_{Bs}), \quad (5)$$

$$\text{Com}^C \sim \sum_{t=1}^l \alpha_{Ct} \cdot \mathcal{N}(\boldsymbol{\mu}_{Ct}, \boldsymbol{\Sigma}_{Ct}), \quad (6)$$

where  $\alpha$  denote the mixture weights (summing to 1 in each case), and  $\boldsymbol{\mu} \in \mathbb{R}^m$ ,  $\boldsymbol{\Sigma} \in \mathbb{R}^{m \times m}$  are the mean vectors and covariance matrices of the respective multivariate Gaussian components.

To estimate parameters  $\boldsymbol{\mu}$ ,  $\boldsymbol{\Sigma}$ ,  $\alpha$  and  $\lambda$ , and the number of components  $j$ ,  $k$  and  $l$  in Com<sup>A</sup>, Com<sup>B</sup> and Com<sup>C</sup>, respectively, we combined two estimation strategies: (i) an Expectation–Maximization (EM) algorithm ; and (ii) a two step Bayesian approach.

#### 3.2 Estimation of Gaussian parameters using the Expectation–Maximization algorithm

Communities  $A$  and  $B$  were, respectively estimated using in situ samples using large dataset from water mass  $A$  and water mass  $B$  (step 1 in Table 1). We considered varying numbers of components  $j$ ,  $k \in \{1, \dots, 10\}$ .

The potential community  $C$  cannot be directly observed in the front (due to the limited number of observations). Thus, we estimated the Gaussian parameters of likely communities to be in  $C$  from the larger dataset of the rest of the cruise, hereafter the *outside dataset*, which consists of 461 observations (black dots on Fig. 1b). The limitations of this approach are discussed in more detail in the Discussion section. We assumed that the community in the *outside dataset* denoted by Com<sup>C'</sup> is a mixture of sub-communities, large enough to

**Table 1.** Summary of data analysis workflow. Note that  $\text{Com}^{C'}$  refers to the community inferred from the *outside dataset*, while  $\text{Com}^C$  denotes the latent community in the frontal zone. By construction,  $\text{Com}^C \subset \text{Com}^{C'}$ .

Step	Purpose	Data	Parameters estimated	Method
1	Define communities $A$ and $B$ from water mass observations	Water mass $A$ ( $n = 30$ ) Water mass $B$ ( $n = 44$ )	$\mu_{A1, \dots, Aj}; \Sigma_{A1, \dots, Aj};$ $\alpha_{A1, \dots, Aj}; \mu_{B1, \dots, Bk};$ $\Sigma_{B1, \dots, Bk}; \alpha_{B1, \dots, Bk}$	EM
2	Find $l'$ candidate communities from the <i>outside dataset</i> ( $n = 461$ )	<i>Outside dataset</i> ( $n = 461$ )	$\mu_{C'1, \dots, C'l'}; \Sigma_{C'1, \dots, C'l'}$	EM
3	<i>Exploratory model:</i> Select the $l$ communities that characterize community $C$	Front data ( $n = 11$ )	$\lambda_A; \lambda_B; \lambda_C; \alpha_{C1, \dots, Cl'}$	Bayesian
4	<i>Final model:</i> Estimate the weight of communities $A$ , $B$ , and $C$ , with only the most important $l$ components in $C$	Front data ( $n = 11$ )	$\lambda_A; \lambda_B; \lambda_C; \alpha_{C1, \dots, Cl}$	Bayesian

represent the latent community  $\text{Com}^C$ . In other terms, we considered  $\text{Com}^C$  to be a subset of  $\text{Com}^{C'}$ . Here, we considered varying numbers of components  $l' \in \{1, \dots, 20\}$  to be able to propose several candidates for  $C$  (step 2 in Table 1).

For all combinations  $j, k \in \{1, \dots, 10\}$  and  $l' \in \{1, \dots, 20\}$ , the EM algorithm explored 14 models, each corresponding to a different structure of the covariance matrix  $\Sigma$ , ranging from diagonal to fully parameterized. Diagonal matrices imply no interaction between phytoplankton groups, while off-diagonal terms capture inter-group correlations.

Model selection was guided by the Integrated Completed Likelihood (ICL) criterion (McLachlan and Rathnayake, 2014), which penalizes model complexity and cluster overlap. First, optimal values of  $j$ ,  $k$  and  $l'$  were chosen by averaging ICL values across covariance structures. Then, the best covariance model was selected for parameter estimation. In addition to  $\mu$  and  $\Sigma$  values, the EM algorithm estimated the  $\alpha$  weights for  $\text{Com}^A$  and  $\text{Com}^B$ . Note that for  $\text{Com}^{C'}$ , only the  $\mu_{C'}$  and  $\Sigma_{C'}$  values were used. The weights of the candidates for  $\text{Com}^C$ ,  $\alpha_C$ , will be further estimated by the Bayesian model (see step 3 and 4 in Table 1). The R package *mclust* (Scrucca et al., 2023) was used to estimate the Gaussian parameters.

### 3.3 Estimation of components weights with Hierarchical Bayesian sampling based on scarce frontal dataset

Since very few observations were collected in the frontal region ( $n = 11$ ), a Bayesian approach was used to estimate the weights of components  $\lambda_A, \lambda_B, \lambda_C$ , see Eq. (3), and the sub components weights  $\alpha_{C1}, \dots, \alpha_{Cl}$ , see Eq. (6). Dirichlet distributions, which represent a distribution over probability distributions often used to model multivariate proportions, were used here to represent the component weights. These distributions were parameterized by a vector of positive real num-

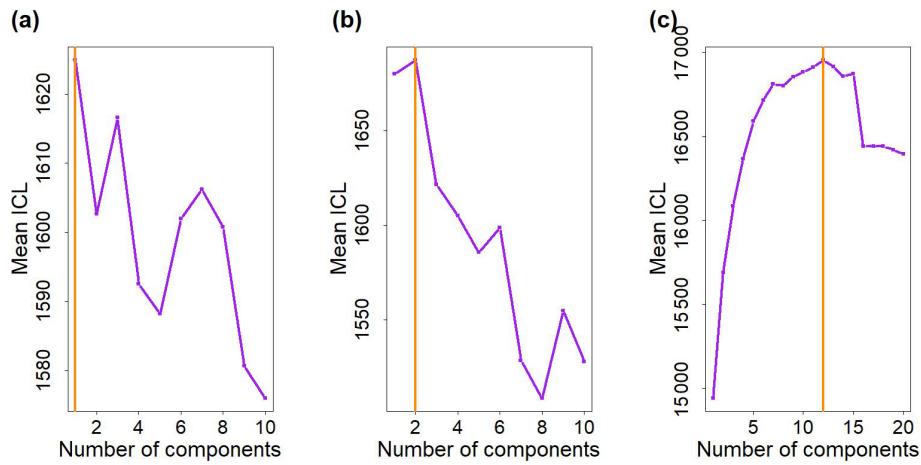
bers and we proposed a non-informative *prior*, as follow:

$$\lambda_A, \lambda_B, \lambda_C \sim \text{Dirichlet}(1, 1, 1) \quad (7)$$

$$\alpha_{C1}, \dots, \alpha_{Cl} \sim \text{Dirichlet}(1, \dots, 1), \quad (8)$$

assigning the same weight to all coefficients. The Hierarchical Bayesian Model is decomposed in two steps. In the *exploratory model* (Step 3, Table 1), the  $l'$  candidate components identified in  $\text{Com}^{C'}$  via EM were used to estimate their associated weights  $\alpha_{C1}, \dots, \alpha_{Cl'}$  in Eq. (6). This step allowed to select only the most significant components among them, i.e. the  $l$  components with the highest posterior  $\alpha_C$  values, to define  $\text{Com}^C$ . In the *final model* (Step 4, Table 1), we estimated the weights of the components  $\lambda_A, \lambda_B, \lambda_C$ , see Eq. (3), and the sub-components weights  $\alpha_{C1}, \dots, \alpha_{Cl}$ , see Eq. (6), with the selected  $l$  components in  $C$ .

A sensitivity analysis was performed to test the robustness of the Bayesian inference. In particular, the sensitivity analysis aimed at assessing the robustness of the model according to the number of observations in the front, and the robustness of the model to false positive detection (i.e. detecting a new community when no communities are present in the data). For this, numerical sampling of “known” frontal community were done in two cases. First, we considered the case where front observations are only a mixture of  $\text{Com}^A$  and  $\text{Com}^B$  (i.e. simulations do not include the component  $\text{Com}^C$ ,  $\lambda_C = 0$ ). Five scenarios were assessed : (1)  $\lambda_A = \lambda_B = 0.5$ ; (2)  $\lambda_A = 0.4$  and  $\lambda_B = 0.6$ ; (3)  $\lambda_A = 0.6$  and  $\lambda_B = 0.4$ ; (4)  $\lambda_A = 0.7$  and  $\lambda_B = 0.3$ ; (5)  $\lambda_A = 0.3$  and  $\lambda_B = 0.7$ . Second, we considered the case where a new community exists, i.e. simulations include the component  $\text{Com}^C$ ,  $\lambda_C \neq 0$ . Here the  $\lambda$  values used are the same as observed in the in situ dataset (see the results Sect. 4.2), and with proportion  $\lambda_A = 0.45$ ,  $\lambda_B = 0.2$  and  $\lambda_C = 0.35$ . In all cases, 5, 10, 20, 30 and 50 frontal observations were simulated 10 times. Then the Bayesian model was computed to estimate the  $\lambda$  values of



**Figure 2.** Values of the mean Integrated Complete-data Likelihood (ICL) in function of the number of multivariate Gaussian components. (a) Water mass  $A$  (to model  $\text{Com}^A$ ). (b) Water mass  $B$  (to model  $\text{Com}^B$ ). (c) *Outside dataset* (to propose likely parameters to model  $\text{Com}^C$ ). The vertical orange lines correspond to the number of multivariate Gaussian components that reach the highest ICL values.

the known communities from the simulated datasets. In addition a comparison test for equal distribution between transitional waters  $T$  and front  $F$  was performed (Székely and Rizzo, 2004).

The posterior probability distribution sampling was computed with STAN's (Carpenter et al., 2017) Hamiltonian Monte Carlo (HMC) algorithm. For all the models (i.e. *exploratory* and *final* models and sensitivity analysis) we performed four chains of 11 000 iterations to study the convergence. The first 10 000 draws of each chain were discarded (i.e. burn-in) to avoid initial sample bias. Thus, the last 1000 iterations of the four chains were used for analysis of the posterior probability distribution. Convergence was assessed with  $\hat{R}$  statistic and effective sample size. The models were computed in R (R Core Team, 2021) by means of the rstan package as interface with STAN (Stan Development Team, 2020).

## 4 Results

### 4.1 Selection of the number of components to describe phytoplankton communities

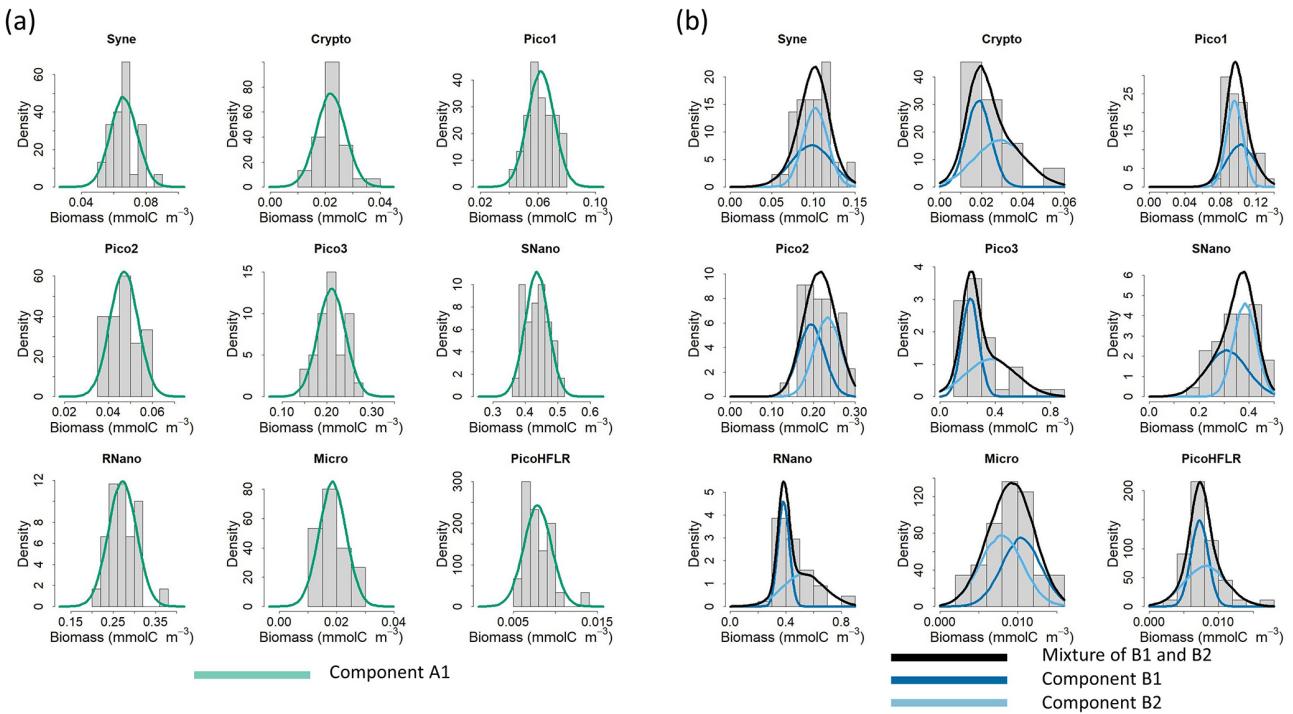
According to the mean ICL criterion, one component,  $A1$ , is enough to model  $\text{Com}^A$  (Fig. 2a), while two components (hereafter  $B1$  and  $B2$ ) are necessary to model  $\text{Com}^B$  (Fig. 2b). The weights of  $B1$  and  $B2$  are, respectively  $\alpha_{B1} = 0.47$  and  $\alpha_{B2} = 0.53$ . In the *outside dataset*, 12 candidates components were selected (Fig. 2c).

Figure 3 shows that the estimated parameters of the multivariate Gaussian fitted well to the observed phytoplankton biomass in water mass  $A$  (Fig. 3a) and water mass  $B$  (Fig. 3b). The mixture of two components in  $\text{Com}^B$  allows to model complex biomass distributions, for e.g. skewed dis-

tribution for Crypto, Pico1, Pico3, SNano, RNano or Pico-HFLR, compared with  $\text{Com}^A$ .

### 4.2 Modelling of the frontal phytoplankton community

The *exploratory model* was performed with 12 candidates components, denoted  $C'$  as they were estimated from the *outside dataset* community, to describe  $\text{Com}^C$  (see step 2 in Table 1). Using the proposed candidates, the *exploratory model* (whose trajectories and values of  $\hat{R}$  and effective sample size are consistent with those of a converged chain, see Fig. A4 and Table A1) was used to estimate the weights  $\alpha_C$  and  $\lambda_A, \lambda_B$  and  $\lambda_C$  in the mixture (Fig. 4a and b). Among the three communities,  $\text{Com}^C$  has a higher weight ( $\lambda$ ) in the mixture (0.787, quantile 2.5 % = 0.123, quantile 97.5 % = 0.902), followed by  $\text{Com}^A$  (0.203, quantile 2.5 % = 0.051, quantile 97.5 % = 0.459) and  $\text{Com}^B$  (0.065, quantile 2.5 % = 0.002, quantile 97.5 % = 0.127) (Fig. 4a and Table A1). Among the 12 candidates components in  $\text{Com}^C$  (Fig. 4b), components  $C'8$  and  $C'10$  present the highest weight,  $\alpha$ , approx. 0.2, and to a lesser extent  $C'5$  and  $C'6$  with weights reaching 0.1. The weights of the other 8 components are below 0.05 (see Table A1 for quantiles values of the posterior distribution for each components). For the *final model*, performed with the most significant components, only  $C'8$  and  $C'10$  were used in  $\text{Com}^C$  as they display the highest weight. We considered  $C'8$  and  $C'10$ , the most significant components in the *exploratory model*, as the two components of the community  $C$ ,  $\text{Com}^C$ , which we call hereafter  $C1$  and  $C2$ , respectively. The number of components in  $\text{Com}^C$  was chosen to be the most parsimonious possible. For the sake of simplicity, the components  $C'5$  and  $C'6$  were not used in the *final model* as they did not show strong differences relative to using only  $C'8$  and  $C'10$  in  $\text{Com}^C$ .



**Figure 3.** Fit of the multivariate Gaussian component with the observed biomass for each phytoplankton group in (a) water mass A and (b) water mass B. Histograms show the observed phytoplankton biomass distributions, and lines correspond to the density curves of the multivariate Gaussian components estimated by EM. Note that in (b) two components are necessary to model  $\text{Com}^B$ ; the components  $B1$  and  $B2$  are weighted by their  $\alpha$  values and the black line corresponds to the finite mixture (i.e. here the sum) of these two components.

In the *final model* (that converged, see Fig. A5 for the trajectories, and Table A2 for convergence metrics), the estimations of  $\lambda_A$ ,  $\lambda_B$  and  $\lambda_C$  did not vary drastically – 0.203 (quantile 2.5 % = 0.057, quantile 97.5 % = 0.467) for  $\text{Com}^A$ , 0.06 (quantile 2.5 % = 0.003, quantile 97.5 % = 0.281) for  $\text{Com}^B$  and 0.714 (quantile 2.5 % = 0.439, quantile 97.5 % = 0.901) for  $\text{Com}^C$ , Fig. 4c – relative to those observed before in the model with 12 components. In this second model, the weights of  $C1$  ( $\alpha_{C1} = 0.55$ , quantile 2.5 % = 0.258, quantile 97.5 % = 0.823) and  $C2$  ( $\alpha_{C2} = 0.45$ , quantile 2.5 % = 0.177, quantile 97.5 % = 0.742) were almost equivalent in the mixture (Fig. 4d and Table A2).

Figure 5 shows how the *final model* fits the observed data in the front. As expected when looking at the weight  $\lambda_B$  (0.06, see in Fig. 4c), components  $B1$  (dark blue curve) and  $B2$  (light blue curve) contribute little to the global mixture (black lines), which is mostly driven by components  $A1$  (light green curve),  $C1$  (dark orange curve), and  $C2$  (orange curve). Overall, the mixture of these five components captures the phytoplankton groups biomass distribution well. In some cases, the biomass distribution is bimodal (for Syne, Pico2, RNano) or skewed (for Pico1, Micro). For Pico3 and RNano a difference remained between the estimated density and the observed biomass in the front (Fig. 5). For Pico3, the  $\mu$  values for  $C1$  and  $C2$  are the lowest (see Table 2). How-

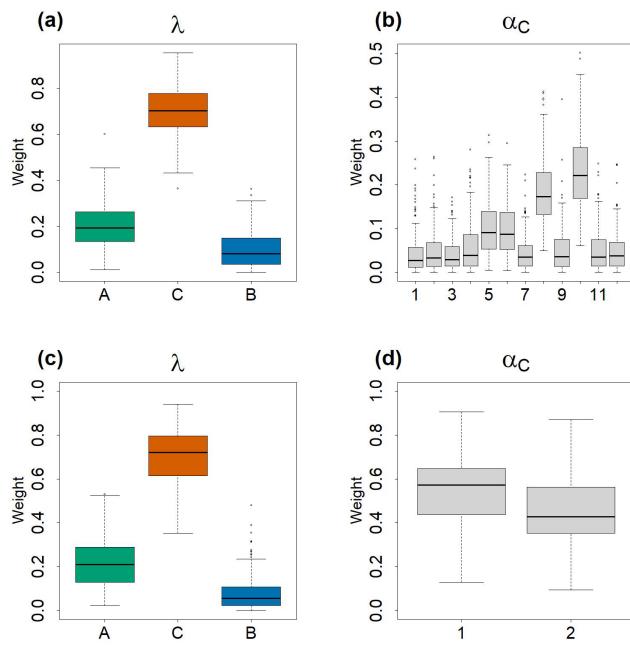
**Table 2.** Rounded  $\mu$  ( $\text{mmolC m}^{-3}$ ) values estimated by Expectation–Maximization algorithm for the components in  $\text{Com}^A$  ( $A1$ ),  $\text{Com}^B$  ( $B1$  and  $B2$ ) and  $\text{Com}^C$  ( $C1$  and  $C2$ ). The rows in bold correspond to the three phytoplankton groups presenting the highest  $\mu$  values (i.e. biomass) within a component. For a better comparison between components for the same phytoplankton groups see Fig. 6.

	$\mu_{A1}$	$\mu_{B1}$	$\mu_{B2}$	$\mu_{C1}$	$\mu_{C2}$
Syne	0.07	0.1	0.1	0.06	0.12
Crypto	0.02	0.02	0.03	0.03	0.02
Pico1	0.06	0.1	0.1	0.07	0.1
Pico2	0.05	0.19	0.23	0.08	0.13
<b>Pico3</b>	<b>0.21</b>	<b>0.22</b>	<b>0.37</b>	<b>0.2</b>	<b>0.13</b>
SNano	<b>0.44</b>	<b>0.31</b>	<b>0.38</b>	<b>0.46</b>	<b>0.32</b>
RNano	<b>0.27</b>	<b>0.38</b>	<b>0.53</b>	<b>0.5</b>	<b>0.51</b>
Micro	0.02	0.01	0.01	0.02	0.01
PicoHFLR	0.01	0.01	0.01	0.01	0.01

ever due to high variance in  $\Sigma$  matrices the mode around 0.14–0.16 is not captured by the model.

#### 4.3 Characteristics of the phytoplankton communities

The parameters of the multivariate Gaussian estimated by the EM algorithm are referenced in Table 2 for  $\mu$  parameters and



**Figure 4.** Boxplots of the estimated values of  $\alpha_C$  and  $\lambda$  values by the Bayesian models. In total 100 values were used to construct the boxplots. Each 40 values among the 4000 iterations of the posterior distributions, generated by the HMC chains, were taken in order to avoid autocorrelation within the chains. (a) and (b) represent, respectively  $\lambda_A, \lambda_B$  and  $\lambda_C$  and  $\alpha_{C1,\dots,C12}$  of *exploratory model*. (c) and (d) represent, respectively  $\lambda_A, \lambda_B$  and  $\lambda_C$  and  $\alpha_{C1,C2}$  of the *final model*.

in Tables A3–A7 for  $\Sigma$  covariance matrices. The  $\mu$  values and the variances in the diagonal of the  $\Sigma$  matrices of SNano, RNano and Pico3 are the highest. These results highlight the dominance and a large variability in biomass of these phytoplankton groups during the cruise.

In  $\text{Com}^B$ , the covariance matrices  $\Sigma_{B1}$  and  $\Sigma_{B2}$  are diagonal. This suggests that the addition of interactions between phytoplankton groups would not have improved the modeling of  $\text{Com}^B$ . By contrast, for  $\text{Com}^A$  and  $\text{Com}^C$ , the covariance matrices are not diagonal which allow to model positive or negative interactions between each phytoplankton group. The covariances matrices  $\Sigma_{C1}$  and  $\Sigma_{C2}$  present similar patterns and highlight mostly the positive interactions of SNano and RNano with most of the phytoplankton groups, except for Pico3. In these two communities Pico3 and Pico2 have a negative interaction.  $\Sigma_{A1}$  presents similar pattern than in  $\Sigma_{C1}$  and  $\Sigma_{C2}$ , but the main differences are negative interactions of Syne and Crypto with SNano, RNano, Micro and PicoHFLR, and strong interactions between Pico3 and Crypto (positive) and PicoHFLR (negative).

Overall, the relative biomass (i.e. calculated from  $\mu$  and  $\alpha$  values) of the phytoplankton groups of  $\text{Com}^C$  is intermediate between  $\text{Com}^A$  and  $\text{Com}^B$ . However, RNano and Pico3 in  $\text{Com}^C$  show a relative biomass that is higher and lower, re-

spectively than in  $\text{Com}^A$  and  $\text{Com}^B$  (Fig. 6a). This pattern is clearly observed when looking at the relative biomass at the sub-component scale (Fig. 6b). Where the relative biomass in  $C1$  and  $C2$  for RNano and Pico3 are, respectively higher and lower than in the other three components. Nevertheless, Fig. 6b, shows that two sub-components of the same community show different patterns for the same phytoplankton group. This is the case of  $C1$  and  $C2$  for Syne, which reach their lowest and highest relative biomass, respectively for these components.

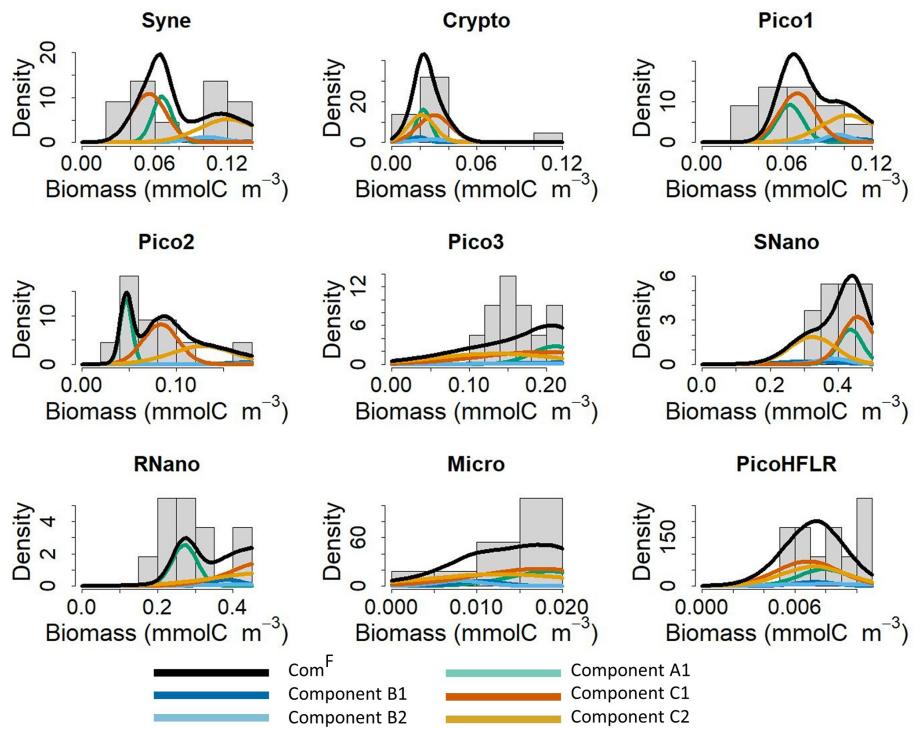
#### 4.4 Sensitivity analysis of the Bayesian inference

Figure 7 shows the results of the sensitivity analysis performed on simulated data. The posterior distribution obtained from the Bayesian inference showed satisfactory mean estimation of the unknown parameters, leading to close estimates compared to the true values independently of the number of simulated observations. Note that while the average values of the posterior distributions of estimated parameters are reliable even for the lowest number of observations, increasing the number of simulated data lead to a decrease in the standard deviation of the posterior distribution (see in Fig. A6). In the case that the simulated data is only coming from a mixture of components  $\text{Com}^A$  and  $\text{Com}^B$  ( $\lambda_C = 0$ ), the sensitivity test shows that the estimated values of  $\lambda_C$  are very close to 0, meaning that this component is not important in the mixture (comparing to the cases where  $\lambda \neq 0$ ) (Fig. 7). In addition, the model can detect slight changes in the proportions of components even in the case that the simulated data is coming from a mixture of the three components  $\text{Com}^A$ ,  $\text{Com}^B$  and  $\text{Com}^C$ . Overall, the sensitivity analysis highlighted the robustness of the approach, even with fewer observations than the actual number of observations in the in situ data. Finally, the comparison test for equal distribution (Székely and Rizzo, 2004) between transitional waters and front rejected the  $H_0$  hypothesis ( $H_0: T \stackrel{d}{=} F$ ,  $p$  value  $< 0.05$ ), which suggests that frontal and transitional water communities are different.

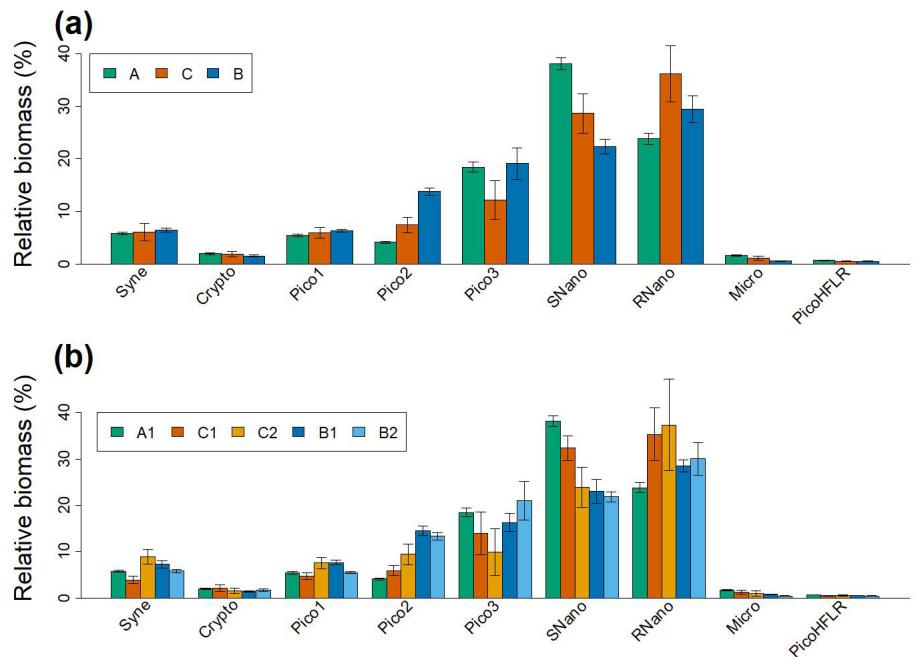
## 5 Discussion

### 5.1 A new approach to identify the frontal community

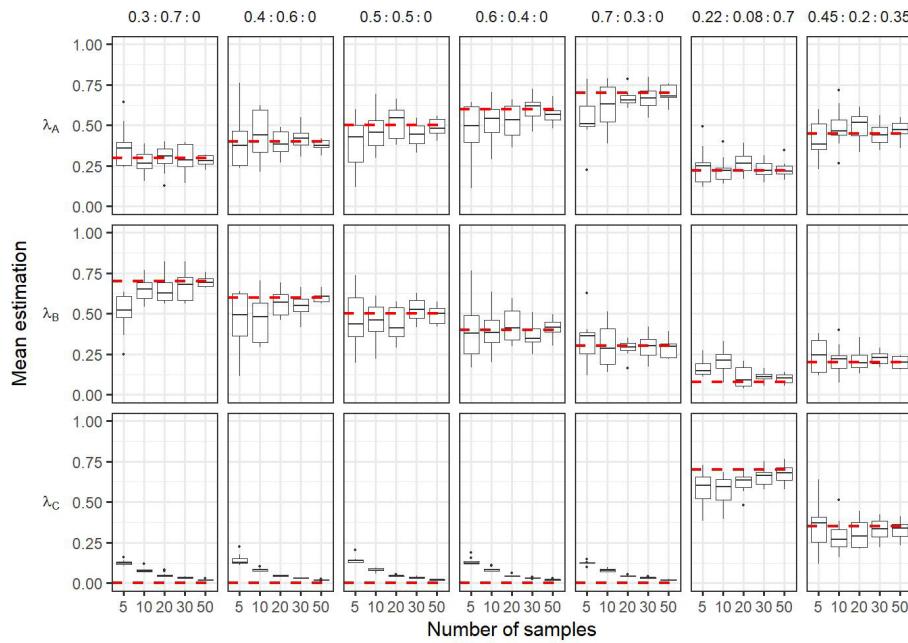
We developed a statistical approach to address key challenges in detecting and confirming fine-scale frontal-adapted phytoplankton communities, despite the limited and highly variable data from an oceanographic campaign. We represented the phytoplankton biomass distribution across and within a frontal region – reflecting the phytoplankton community composition – using a multivariate Gaussian mixture of distinct sub-communities. The critical objective was to determine which community and sub-community has the highest weight within the front. Here, we combined two approaches, Expectation–Maximization (EM) algorithm and



**Figure 5.** Fit of the multivariate Gaussian component with the observed biomass for each phytoplankton group in the front  $F$ . The histograms show the observed phytoplankton biomass distributions in the front, and the lines correspond to the density curves of the multivariate Gaussian components estimated by EM and by the Bayesian model. Note that five components are necessary to model community  $\text{Com}^F$ ; all components are weighted by their  $\lambda$  and  $\alpha$  values. The black line correspond to the estimated  $\text{Com}^F$  from the finite mixture (i.e. here the sum) of these five components.



**Figure 6.** (a) Relative biomass (%) of the nine phytoplankton groups for the modeled communities  $\text{Com}^A$ ,  $\text{Com}^B$  and  $\text{Com}^C$ . Biomass was calculated from  $\mu$  values and weighted by  $\alpha$  values of sub-components (for  $\text{Com}^B$  and  $\text{Com}^C$ ). (b) Relative biomass (%) of the nine phytoplankton groups for the five sub-components. In (a) and (b) errors bars correspond to the 95 % confidence interval of the  $\mu$  values.



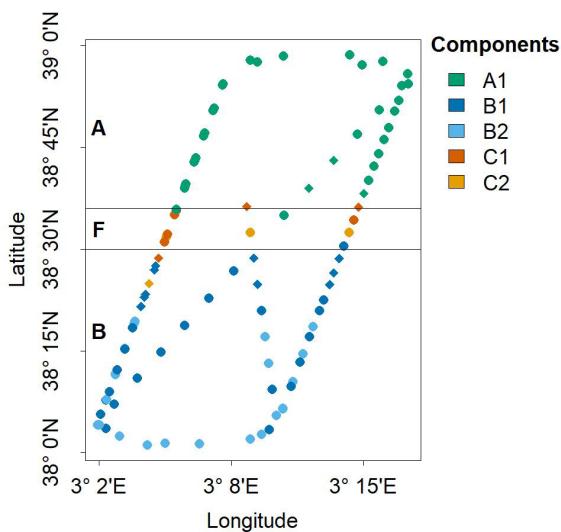
**Figure 7.** Boxplots of the mean of the posterior distributions during the sensitivity analysis. Each column correspond to the  $\lambda_A : \lambda_B : \lambda_C$  estimation of a same condition: In the case frontal community is composed only by a mixture of adjacent water masses (i.e.  $\lambda_C = 0$ ) with varying proportion of simulated  $\lambda_A : \lambda_B : \lambda_C$  (i.e. 0.3 : 0.7 : 0; 0.4 : 0.6 : 0; 0.5 : 0.5 : 0; 0.6 : 0.4 : 0; 0.7 : 0.3 : 0). The two last columns correspond to simulations where frontal community includes a new community  $\text{Com}^C$  ( $\lambda_C \neq 0$ ), here the proportion used for simulations are the same as observed in the *in situ dataset* (i.e.  $\lambda_A = 0.22$ ,  $\lambda_B = 0.08$ ,  $\lambda_C = 0.7$ ) and 0.45 : 0.2 : 0.35. The dashed red horizontal lines correspond to the true  $\lambda$  values used to simulate the data in the mixtures. In each conditions the number of observations in the simulated data in the front varies from 5 to 50. Each boxplot is based on the 10 values of the mean calculated on the 10 simulated datasets for a same hypothesis and a same number of observations.

Bayesian modelling, to characterize the nature of frontal phytoplankton communities from sparse *in situ* data. The EM algorithm allowed us to estimate parameters ( $\mu$  and  $\Sigma$ ) of Gaussian distributions and identify the number of communities and sub-communities (from relatively large datasets), while the Bayesian approach, known to be robust even with few observations, enabled us to determine their relative weights ( $\lambda$  and  $\alpha$ ) within the frontal community. Sensitivity analysis (Fig. 7) confirmed that the Bayesian inference was robust even for fewer observation (here down to 5) than the actual *in situ* frontal dataset (i.e. 11 observations).

The parameters  $\mu$  (average biomass) and  $\Sigma$  (variance and covariance) provided a realistic overview of the phytoplankton community composition (PCC), highlighting the global dominance of two nanophytoplankton groups (SNano, RNano) and the largest picoeukaryote (Pico3), as well as the interactions between these groups that shape specific PCC (Tables 2, A3, A6, and A7). In the Mediterranean Sea, *Synechococcus* species (Syne) are the most dominant group of phytoplankton in abundance (Moutin et al., 2002). However, certain physical forcings, such as frontal structuring, may alter their presence by locally modifying environmental conditions (e.g., nutrient inputs), which can favor larger cells (Siokou-Frangou et al., 2010). In frontal zones, different

types of interactions between plankton organisms, such as shading or shared predation, can lead to distinct community structures (Mangolte et al., 2022). Notably, the differences observed between the covariance matrices (i.e.  $\Sigma_{A1}$ ,  $\Sigma_{C1}$ , and  $\Sigma_{C2}$ ) suggest that interactions between phytoplankton are different within distinct communities (Tables A3, A6, and A7). Estimated values of parameters  $\lambda$  corresponding to the weight of communities within the front and its adjacent water masses provided information to answer our questions: “*What is the structure of the community that might be formed at the front? Is the frontal community a mixture, where the expected community results from the combination of the adjacent water communities, or is there another community resulting from intrinsic frontal characteristics?*”. In particular,  $\lambda_C$  (0.714 in the *final model*, Fig. 4c) represents the proportion of the frontal community attributed to  $\text{Com}^C$ . Since  $\lambda_C > 0$ , our results suggest that the phytoplankton frontal community is not a mixture of adjacent communities, but instead is a specific frontal-adapted community. More precisely,  $\lambda_C$  indicates that  $\text{Com}^C$  represents more than 70 % of the frontal community  $\text{Com}^F$  (Fig. 4c).

Figure 8 shows the spatial projection of each sample point, with shapes and colors representing their community and sub-community classification, identified as the domi-



**Figure 8.** Spatial distribution of dominant components within the NS-Hippodrome. Dots are colored according to the multivariate Gaussian component, i.e. sub-communities ( $A_1$ ,  $B_1$ ,  $B_2$ ,  $C_1$  and  $C_2$ ) that reach the highest density (in nine dimension) for the sample phytoplankton groups observed biomass. The horizontal black lines correspond to the frontal area (latitude between  $38.5$  and  $38.6^{\circ}\text{N}$ , as in Fig. 1c, Component  $A_1$  ( $A$  community) was mostly dominant in the North of the Hippodrome. Components  $B_1$  and  $B_2$  ( $B$  sub-communities) were mostly dominant in the south of the Hippodrome. Components  $C_1$  and  $C_2$  (unknown  $C$  sub-communities) are mostly dominant in the front. Note that diamond shaped points correspond to samples transitional waters,  $T$ , that were not taken into account during the characterisation of the phytoplankton nine communities.

nant component and sub-component of the Gaussian mixtures (highest  $\lambda$  and  $\alpha$ ). We reach two key conclusions. First, our approach successfully reconstructed the initial pattern observed by Tzortzis et al. (2021), characterized by a distribution of two communities on either side of the frontal region (here identified as  $\text{Com}^A$  and  $\text{Com}^B$ ). A notable refinement was the identification of two sub-communities within  $\text{Com}^B$  ( $B_1$  and  $B_2$ ), which could be attributed to significant fine-scale meandering activity in the southern part of the front (i.e., within the Algerian Basin) (Millot, 1999). We hypothesize that such dynamics could lead to a closer cohabitation of different sub-communities. Second, our approach appeared to successfully detect the presence of the unknown  $\text{Com}^C$  located within the frontal community ( $\text{Com}^F$ ).

## 5.2 Phytoplankton communities across frontal areas

Our findings suggest that the frontal region acted as a selective environment, structuring phytoplankton communities by promoting certain phytoplankton groups while disadvantaging others (Fig. 6). According to our results, the frontal zone during PROTEVSMED-SWOT represented a narrow habitat for communities  $C_1$  and  $C_2$ .

Mangolte et al. (2022) described the impact of frontal responses of plankton groups using the terms “winners” and “losers”. In the Californian Current Ecosystem, larger phytoplankton (e.g., microphytoplankton, diatoms) were classified as “winners” (increased abundance within fronts) and smaller picophytoplankton as “losers” (decreased abundance within fronts). Taking into account the whole  $\text{Com}^C$ , our result showed that RNano was clearly a “winner” and Pico3 was a “loser” within the front (Fig. 6). However, at a smaller scale, we showed that even within a same community, phytoplankton assemblage were different. For example, this was the case of *Synechococcus* (Syne) that, respectively showed the lowest and highest  $\mu$  values with  $C_1$  and  $C_2$  (Fig. 6b). This suggests that Syne can simultaneously be both “winner” and “loser”, depending on local conditions. This pattern may result from differences in the origins of  $C_1$  and  $C_2$  communities, driven by advection or stirring of distinct water masses, or from biological interactions that either favored or hindered Syne (Lévy et al., 2018; Hernández-Hernández et al., 2020, 2021). Mangolte et al. (2023) highlighted that different plankton communities can be observed at a smaller scale (1–5 km) than the width of the front scale (10–30 km).

## 5.3 Limitations

The strong assumption that the potential front-adapted community existed within the *outside dataset* implies two limitations. On the one hand, the *outside dataset* is not an exhaustive dataset of the region. Actually, phytoplankton communities of the southern water masses may not be efficiently represented in the *outside dataset*, since the water masses off the Algerian coast (south of the sampling area) were not sampled. In addition, the inclusion of the stations close to the Balearic coasts might have led to an over-representation of coastal phytoplankton communities (different than the one observed in the open sea). But excluding coastal stations and selecting only data near the NS-Hippodrome transect (e.g., between  $38$ – $39^{\circ}\text{N}$  and  $3$ – $5^{\circ}\text{E}$ ) did not drastically affect our results. On the other hand, frontal conditions could be unique in both space and time and might have not been sampled elsewhere than in the NS-Hippodrome transect. Actually, the communities identified in  $\text{Com}^C$ ,  $C_1$  and  $C_2$ , were mostly observed in stations in the same range of temperature and salinity that were close to the studied front, to the east, and were sampled a few days before the NS-Hippodrome transect sampling (Fig. A7). Hydrodynamic circulation across the frontal area was eastward (Tzortzis et al., 2021). This suggests that the communities observed at these sites in the *outside dataset* may have been advected from the front.

As Fig. 5 shows, our approach may not precisely capture the biomass distribution of certain phytoplankton groups (e.g. Pico3 and RNano). This is certainly because no biomass distributions that better fit the frontal data were observed in the “outside” dataset for these two groups. A more flexible option would be to estimate all parameters using a full

Bayesian approach (i.e. the number of Gaussian components,  $\mu$ ,  $\alpha$ ,  $\Sigma$  and  $\lambda$ ). However, as the number of parameters to be estimated far exceeded the actual number of observations at the front (number of observations = 11; number of parameters = 458), we opted to “fix” certain parameters (i.e.  $\mu$  and  $\Sigma$ ) using existing data (adjacent water masses and outside dataset). Nevertheless, the sensitivity analysis demonstrated the robustness of our approach, showing that the new components  $C1$  and  $C2$  in  $Com^C$  helped to model the community in the front more accurately, revealing the existence of a new frontal community.

## 6 Conclusions and perspectives

The re-analysis of the phytoplankton dataset from PROTEVSMED-SWOT using a novel statistical methodology allowed us to reveal a biological signal that remained undetected with classical statistical approaches due to the critical lack of data. This method effectively addresses one of the main challenges in in situ biological oceanography: the difficulty of collecting comprehensive datasets that integrate biological, physical, and biogeochemical measurements while maintaining high temporal and spatial resolution. Notably, without incorporating explicit spatial information or environmental variables into our analysis, our approach successfully captured the structuring effect of the front and detected the presence of a frontal-adapted phytoplankton community.

Importantly, our method reshaped our understanding of this moderately energetic front, previously considered merely a hydrodynamic barrier between two communities (Tzortzis et al., 2021). Instead, our results suggest that this front acted as a unique ecological environment where a distinct community seemed to have emerged. This study can be seen as a first attempt to assess this hypothesis, but due to the dataset scarcity, our results need further application on other in situ datasets to be generalizable. Thus, given the broad applicability of our methodology to plankton datasets, we plan to use it to further investigate whether fronts generally function as simple boundaries or as areas fostering the development of frontal-adapted communities. In addition, recent work has shown that frontal conditions appear to favor the presence of non-dominant phytoplankton groups relative to dominant ones (Oms et al., submitted). Such a “refuge effect” will be evaluated in further research that will analyse satellite-based data sets (e.g., ocean color and altimetry) to provide a global perspective on phytoplankton distribution in frontal regions. Additionally, the future research will include analyses of other in situ larger plankton datasets, such as those from BioSWOT-Med (Doglioli et al., 2024), which provide a more comprehensive environmental context. Including the complete dataset from BioSWOT-Med, integrating nutrient concentrations and fluxes, as well as zooplankton concentrations and grazing rates, will help disentangle

the key processes driving the observed phytoplankton community composition.

## Appendix A

**Table A1.** Summary of the statistics (mean, standard deviations, quantiles 2.5, 25, 50, 75 and 97.5 % of the posterior distributions) and convergence metrics (effective sample size ESS, and  $\hat{R}$ ) of the estimated parameters of the explanatory model.

	mean	sd	2.5 %	25 %	50 %	75 %	97.5 %	ESS	$\hat{R}$
$\alpha_{C1}$	0.048	0.045	0.001	0.015	0.035	0.067	0.167	5513.46	1
$\alpha_{C2}$	0.048	0.046	0.001	0.015	0.035	0.067	0.172	5391.567	1
$\alpha_{C3}$	0.049	0.047	0.001	0.014	0.034	0.069	0.177	6298.038	1
$\alpha_{C4}$	0.048	0.046	0.002	0.014	0.034	0.069	0.171	5107.524	0.999
$\alpha_{C5}$	0.097	0.063	0.013	0.049	0.085	0.131	0.255	4676.972	1
$\alpha_{C6}$	0.099	0.062	0.014	0.052	0.086	0.133	0.256	5474.442	1
$\alpha_{C7}$	0.048	0.046	0.001	0.014	0.035	0.069	0.167	5548.687	1
$\alpha_{C8}$	0.192	0.085	0.055	0.13	0.183	0.244	0.385	5195.142	0.999
$\alpha_{C9}$	0.047	0.045	0.001	0.014	0.034	0.065	0.167	4808.668	0.999
$\alpha_{C10}$	0.226	0.09	0.075	0.159	0.218	0.288	0.419	5193.024	1
$\alpha_{C11}$	0.048	0.046	0.001	0.014	0.034	0.069	0.17	4375.181	1
$\alpha_{C12}$	0.049	0.047	0.001	0.015	0.035	0.069	0.176	4054.995	1.001
$\lambda_A$	0.216	0.107	0.051	0.136	0.203	0.28	0.459	6066.953	0.999
$\lambda_B$	0.088	0.077	0.002	0.029	0.065	0.127	0.28	4854.213	0.999
$\lambda_C$	0.697	0.123	0.428	0.614	0.709	0.787	0.902	5406.117	0.999

**Table A2.** Summary of the statistics (mean, standard deviations, quantiles 2.5, 25, 50, 75 and 97.5 % of the posterior distributions) and convergence metrics (effective sample size ESS, and  $\hat{R}$ ) of the estimated parameters of the final model.

	mean	sd	2.5 %	25 %	50 %	75 %	97.5 %	ESS	$\hat{R}$
$\alpha_{C1}$	0.55	0.149	0.258	0.443	0.554	0.659	0.823	3200.431	1
$\alpha_{C2}$	0.45	0.149	0.177	0.341	0.446	0.557	0.742	3200.431	1
$\lambda_A$	0.217	0.106	0.057	0.138	0.203	0.28	0.467	3620.614	1
$\lambda_B$	0.083	0.075	0.003	0.025	0.06	0.119	0.281	3660.203	1.001
$\lambda_C$	0.701	0.121	0.439	0.622	0.714	0.791	0.901	3630.034	1.001

**Table A3.**  $\Sigma$  covariance matrix estimated by Expectation–Maximization algorithm for the component A1 in Com<sup>A</sup>.

	Syne	Crypto	Pico1	Pico2	Pico3	SNano	RNano	Micro	PicoHFLR
Syne	6.98e-05	1.69e-05	3.25e-05	1.73e-05	7.51e-05	-1.24e-04	-7.26e-05	-1.45e-05	-5.83e-06
Crypto	1.69e-05	2.84e-05	2.88e-06	7.46e-06	6.88e-05	-3.45e-05	-1.72e-05	-1.14e-05	-1.58e-06
Pico1	3.25e-05	2.88e-06	8.57e-05	2.89e-05	5.71e-05	7.38e-05	9.78e-05	9.50e-06	-4.26e-06
Pico2	1.73e-05	7.46e-06	2.89e-05	4.02e-05	1.32e-05	-1.83e-05	1.27e-05	-3.10e-06	1.43e-06
Pico3	7.51e-05	6.88e-05	5.71e-05	1.32e-05	9.37e-04	1.67e-04	2.66e-04	-1.42e-05	-2.37e-05
SNano	-1.24e-04	-3.45e-05	7.38e-05	-1.83e-05	1.67e-04	1.30e-03	9.54e-04	9.76e-05	-3.88e-06
RNano	-7.26e-05	-1.72e-05	9.78e-05	1.27e-05	2.66e-04	9.54e-04	1.12e-03	8.06e-05	-1.45e-05
Micro	-1.45e-05	-1.14e-05	9.50e-06	-3.10e-06	-1.42e-05	9.76e-05	8.06e-05	2.26e-05	1.06e-06
PicoHFLR	-5.83e-06	-1.58e-06	-4.26e-06	1.43e-06	-2.37e-05	-3.88e-06	-1.45e-05	1.06e-06	2.71e-06

**Table A4.**  $\Sigma$  covariance matrix estimated by Expectation–Maximization algorithm for the component  $B1$  in Com $^B$ .

	Syne	Crypto	Pico1	Pico2	Pico3	SNano	RNano	Micro	PicoHFLR
Syne	5.99e-04	0.00e+00							
Crypto	0.00e+00	3.49e-05	0.00e+00						
Pico1	0.00e+00	0.00e+00	2.72e-04	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00
Pico2	0.00e+00	0.00e+00	0.00e+00	9.98e-04	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00
Pico3	0.00e+00	0.00e+00	0.00e+00	0.00e+00	3.79e-03	0.00e+00	0.00e+00	0.00e+00	0.00e+00
SNano	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	6.72e-03	0.00e+00	0.00e+00	0.00e+00
RNano	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	1.64e-03	0.00e+00	0.00e+00
Micro	0.00e+00	6.30e-06	0.00e+00						
PicoHFLR	0.00e+00	1.57e-06							

**Table A5.**  $\Sigma$  covariance matrix estimated by Expectation–Maximization algorithm for the component  $B2$  in Com $^B$ .

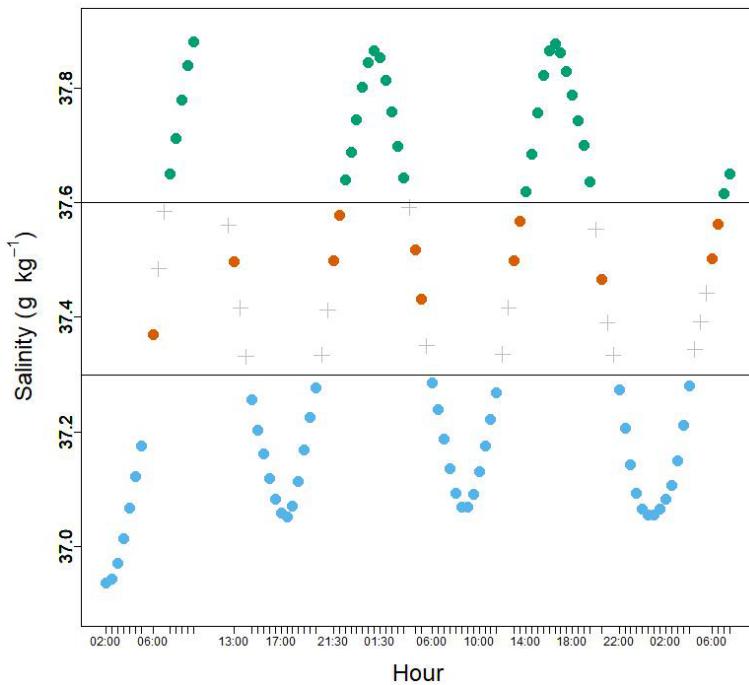
	Syne	Crypto	Pico1	Pico2	Pico3	SNano	RNano	Micro	PicoHFLR
Syne	2.17e-04	0.00e+00							
Crypto	0.00e+00	1.57e-04	0.00e+00						
Pico1	0.00e+00	0.00e+00	8.36e-05	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00
Pico2	0.00e+00	0.00e+00	0.00e+00	1.09e-03	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00
Pico3	0.00e+00	0.00e+00	0.00e+00	0.00e+00	3.31e-02	0.00e+00	0.00e+00	0.00e+00	0.00e+00
SNano	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	2.14e-03	0.00e+00	0.00e+00	0.00e+00
RNano	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	2.26e-02	0.00e+00	0.00e+00
Micro	0.00e+00	7.57e-06	0.00e+00						
PicoHFLR	0.00e+00	9.23e-06							

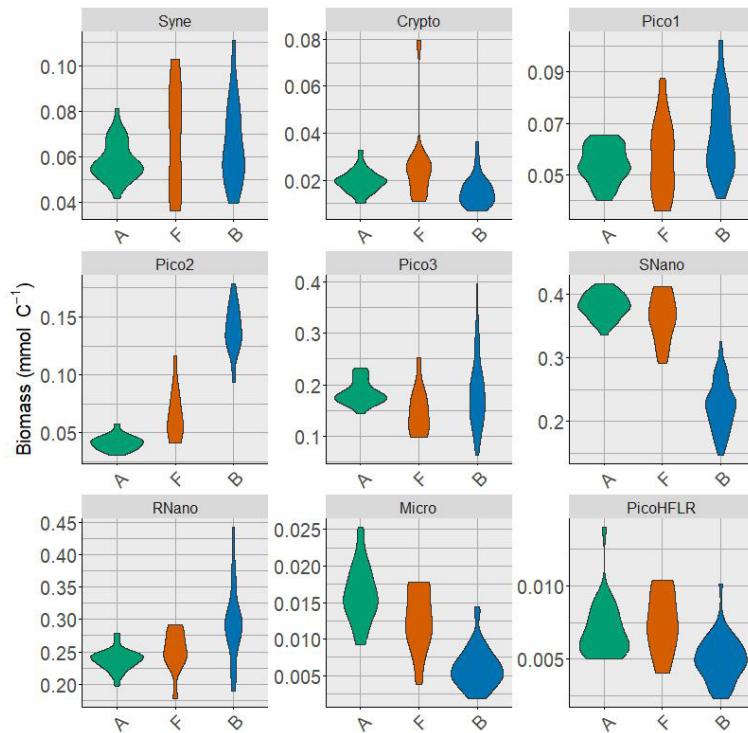
**Table A6.**  $\Sigma$  covariance matrix estimated by Expectation–Maximization algorithm for the component  $C1$  in Com $^C$ .

	Syne	Crypto	Pico1	Pico2	Pico3	SNano	RNano	Micro	PicoHFLR
Syne	2.04e-04	2.29e-05	2.12e-05	-1.30e-05	1.18e-04	1.18e-05	5.82e-04	1.84e-05	1.36e-06
Crypto	2.29e-05	1.37e-04	1.30e-05	-6.25e-06	8.33e-05	1.17e-04	3.35e-04	1.74e-05	4.82e-06
Pico1	2.12e-05	1.30e-05	1.65e-04	7.66e-07	5.03e-05	4.20e-05	3.02e-04	1.46e-05	2.57e-06
Pico2	-1.30e-05	-6.25e-06	7.66e-07	3.53e-04	-6.69e-04	1.61e-04	2.70e-04	-2.38e-05	-7.82e-07
Pico3	1.18e-04	8.33e-05	5.03e-05	-6.69e-04	6.77e-03	-2.26e-04	1.42e-04	5.24e-05	-1.43e-08
SNano	1.18e-05	1.17e-04	4.20e-05	1.61e-04	-2.26e-04	2.31e-03	2.47e-03	1.85e-04	2.02e-05
RNano	5.82e-04	3.35e-04	3.02e-04	2.70e-04	1.42e-04	2.47e-03	1.01e-02	3.72e-04	4.06e-05
Micro	1.84e-05	1.74e-05	1.46e-05	-2.38e-05	5.24e-05	1.85e-04	3.72e-04	5.66e-05	3.64e-06
PicoHFLR	1.36e-06	4.82e-06	2.57e-06	-7.82e-07	-1.43e-08	2.02e-05	4.06e-05	3.64e-06	4.18e-06

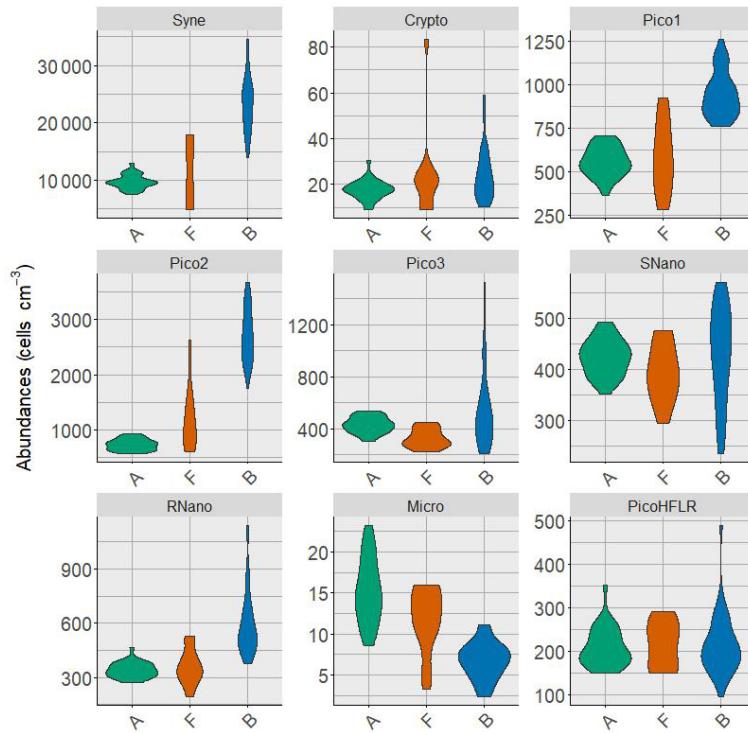
**Table A7.**  $\Sigma$  covariance matrix estimated by Expectation–Maximization algorithm for the component  $C2$  in Com $^C$ .

	Syne	Crypto	Pico1	Pico2	Pico3	SNano	RNano	Micro	PicoHFLR
Syne	6.07e-04	1.10e-04	1.26e-04	-3.41e-05	9.93e-05	1.18e-04	1.28e-03	5.43e-05	5.69e-06
Crypto	1.10e-04	8.94e-05	5.94e-05	-3.95e-05	7.81e-05	2.73e-04	7.65e-04	4.42e-05	5.78e-06
Pico1	1.26e-04	5.94e-05	3.65e-04	1.20e-05	4.66e-05	1.24e-04	6.73e-04	3.67e-05	6.57e-06
Pico2	-3.41e-05	-3.95e-05	1.20e-05	1.16e-03	-5.00e-04	2.72e-04	6.15e-04	-1.20e-04	-7.26e-06
Pico3	9.93e-05	7.81e-05	4.66e-05	-5.00e-04	6.05e-03	-5.93e-05	3.37e-04	4.75e-05	5.02e-07
SNano	1.18e-04	2.73e-04	1.24e-04	2.72e-04	-5.93e-05	4.56e-03	5.80e-03	3.96e-04	4.30e-05
RNano	1.28e-03	7.65e-04	6.73e-04	6.15e-04	3.37e-04	5.80e-03	2.29e-02	8.55e-04	9.31e-05
Micro	5.43e-05	4.42e-05	3.67e-05	-1.20e-04	4.75e-05	3.96e-04	8.55e-04	7.71e-05	6.62e-06
PicoHFLR	5.69e-06	5.78e-06	6.57e-06	-7.26e-06	5.02e-07	4.30e-05	9.31e-05	6.62e-06	4.41e-06

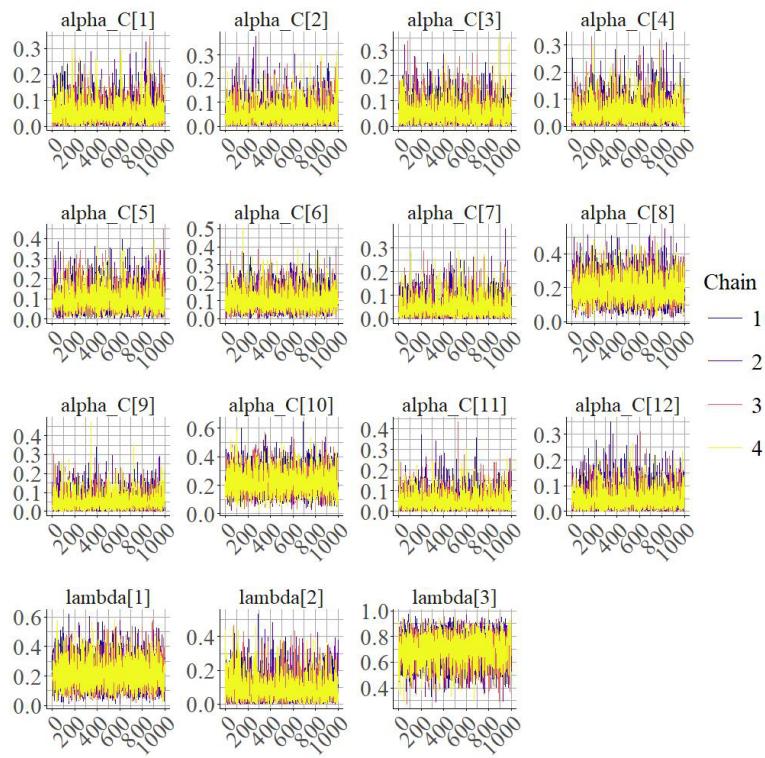
**Figure A1.** Variation of the salinity measurements during the cruise between 11 and 13 May. The horizontal lines correspond to the isohalines that were chosen to characterize the frontal area. Green dots correspond to the water mass  $A$ , blue dots to the water mass  $B$ . Within the frontal area, latitudinal limits were chosen according to Tzortzis et al. (2021). In this zone, orange dots correspond to the front, and grey crosses to the transitional waters,  $T$ , that are not taken for the data analyses.



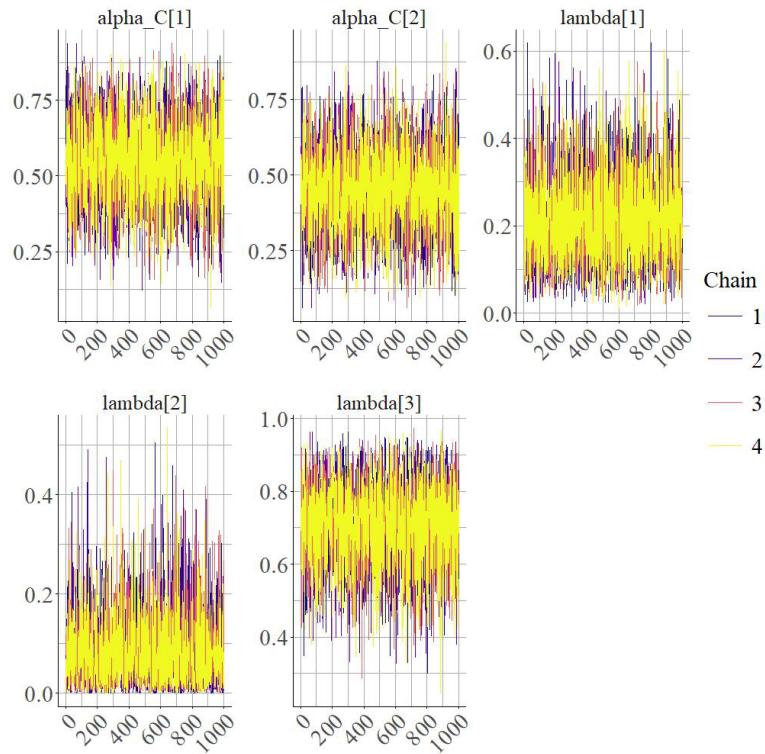
**Figure A2.** Violin plot of the phytoplankton groups biomass in the three water masses A, F and B. Biomasses are expressed in  $\text{mmolC m}^{-3}$ .



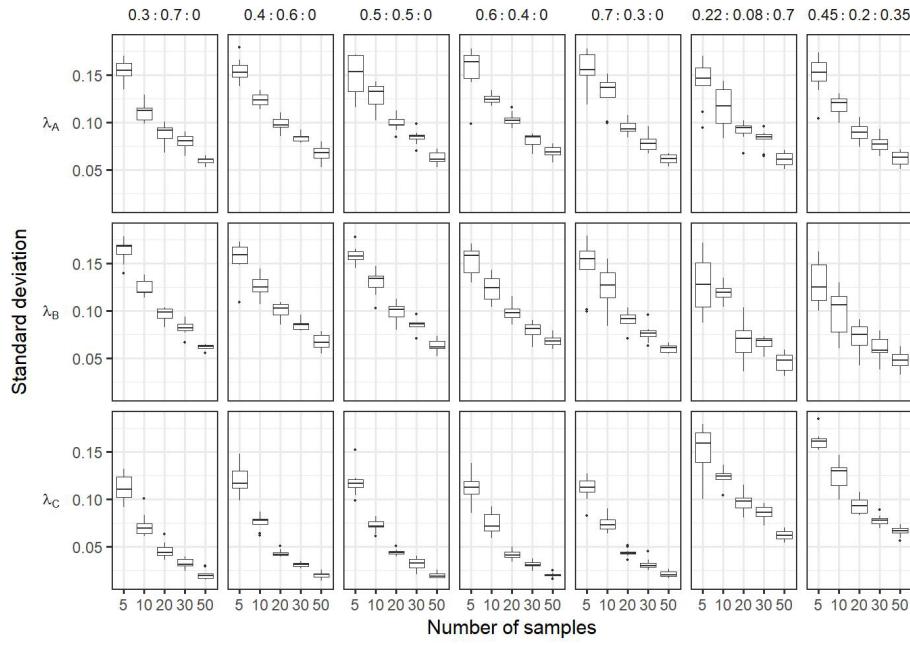
**Figure A3.** Violin plot of the phytoplankton groups abundances in the three water masses A, F and B. Abundances are expressed in  $\text{cells cm}^{-3}$ .



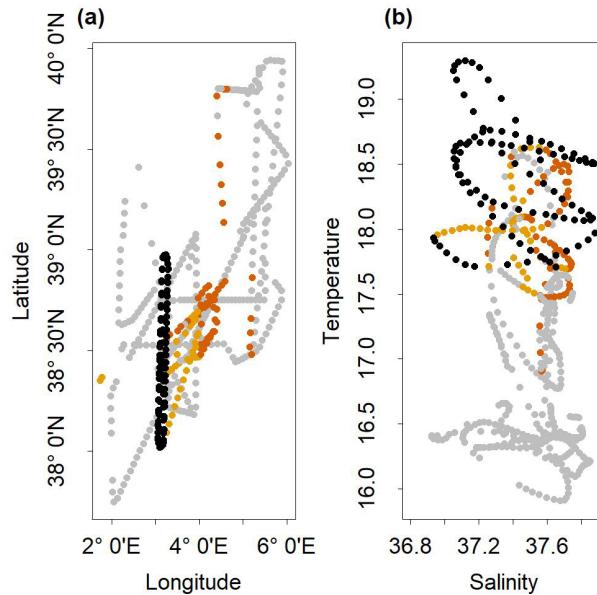
**Figure A4.** Trace of the posteriors distributions of the parameters estimated by the first Bayesian model, *exploratory model*.



**Figure A5.** Trace of the posteriors distributions of the parameters estimated by the second Bayesian model, *final model*.



**Figure A6.** Boxplots of the standard deviation of the posterior distributions of the sensitivity analysis. Each column correspond to the  $\lambda_A : \lambda_B : \lambda_C$  estimation of a same condition: In the case frontal community is composed only by a mixture of adjacent water masses (i.e.  $\lambda_C = 0$ ) with varying proportion of simulated  $\lambda_A : \lambda_B : \lambda_C$  (i.e. 0.3 : 0.7 : 0; 0.4 : 0.6 : 0; 0.5 : 0.5 : 0; 0.6 : 0.4 : 0; 0.7 : 0.3 : 0). The two last columns correspond to simulations where frontal community includes a new community  $\text{Com}^C$  ( $\lambda_C \neq 0$ ), here the proportion used for simulations are the same as observed in the *in situ* dataset (i.e.  $\lambda_A = 0.22$ ,  $\lambda_B = 0.08$ ,  $\lambda_C = 0.7$ ) and 0.45 : 0.2 : 0.35. The dashed red horizontal lines correspond to the true  $\lambda$  values used to simulate the data in the mixtures. In each conditions the number of observations in the simulated data in the front varies from 5 to 50. Each boxplot is based on the 10 values of the mean calculated on the 10 simulated datasets for a same hypothesis and a same number of observations.



**Figure A7.** (a) Spatial Distribution of stations belonging to cluster C1 (in dark orange), cluster C2 (in orange) and the NS-Hippodrome transect stations (in black). The grey dots are the others stations of the cruise. (b) Temperature/Salinity diagram of the stations of the cruise. The dots in dark orange correspond to cluster C1, in orange to cluster C2, in black to the NS-Hippodrome transect stations, and in grey are the others stations of the cruise.

**Code and data availability.** Code and data are available at: [https://github.com/theogarcia/Phytoplankton\\_in\\_front.git](https://github.com/theogarcia/Phytoplankton_in_front.git) (last access: 27 January 2026).

**Author contributions.** Conception and design of the study: TG, LO, XM, AD, MM, GG, DP. Formal analysis: TG. Writing – original draft preparation: TG and LO. Writing – review and editing: all authors.

**Competing interests.** The contact author has declared that none of the authors has any competing interests.

**Disclaimer.** Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. The authors bear the ultimate responsibility for providing appropriate place names. Views expressed in the text are those of the authors and do not necessarily reflect the views of the publisher.

**Acknowledgements.** Franck Dumas, PI of the cruise, the SHOM and the crew of the RV *Beautemps-Beaupré* are acknowledged for shipboard operations. The authors thank Melilotus Thyssen for providing the CytoBuoy flow cytometer and Roxane Tzortzis and Lloyd Izard for the cytometry data analysis. The authors acknowledge François Ribalet and the two anonymous for their valuable comments.

**Financial support.** This work was supported by the CNES under the BIOSWOT-AdAC project and the MIO Axes Transverses (AT-COUPAGE). This work is part of the rODEo project which is funded by the Institut des Mathématiques pour la Planète Terre which supports collaborations between mathematicians and life and earth scientists.

**Review statement.** This paper was edited by Chris Forest and reviewed by François Ribalet and two anonymous referees.

## References

- Acha, E. M., Piola, A., Iribarne, O., and Mianzan, H.: Ecological processes at marine fronts: oases in the ocean, Springer, <https://doi.org/10.1007/978-3-319-15479-4>, 2015.
- Bethoux, J. P., Gentili, B., Morin, P., Nicolas, E., Pierre, C., and Ruiz-Pino, D.: The Mediterranean Sea: a miniature ocean for climatic and environmental studies and a key for the climatic functioning of the North Atlantic, *Progress in Oceanography*, 44, 131–146, 1999.
- Bianchi, C. N. and Morri, C.: Marine biodiversity of the Mediterranean Sea: situation, problems and prospects for future research, *Marine Pollution Bulletin*, 40, 367–376, 2000.

- Birgé, L.: Approximation et estimation dans les modèles à mélanges, *Annales de l'I. H. P. Probabilités et Statistiques*, 19, 295–320, 1983.
- Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., and Riddell, A.: Stan: A probabilistic programming language, *Journal of Statistical Software*, 76, 1–32, 2017.
- Clayton, S., Nagai, T., and Follows, M. J.: Fine scale phytoplankton community structure across the Kuroshio Front, *Journal of Plankton Research*, 36, 1017–1030, <https://doi.org/10.1093/plankt/fbu020>, 2014.
- Clayton, S., Lin, Y.-C., Follows, M. J., and Worden, A. Z.: Co-existence of distinct *Ostreococcus* ecotypes at an oceanic front, *Limnology and Oceanography*, 62, 75–88, <https://doi.org/10.1002/lno.10373>, 2017.
- Clifton Gray, P., Saveliev, I., Cassar, N., Lévy, M., Boss, E., Lehahn, Y., Bourdin, G., Thompson, K. A., Windle, A., Gronniger, J., Floge, S., Hunt, D. E., Silsbe, G., Johnson, Z. I., and Johnston, D. W.: Evidence for kilometer-scale biophysical features at the Gulf Stream front, *Journal of Geophysical Research: Oceans*, 129, e2023JC020526, 2024.
- Coll, M., Piroddi, C., Steenbeek, J., Kaschner, K., Ben Rais Lasram, F., Aguzzi, J., Ballesteros, E., Nike Bianchi, C., Corbera, J., Dailianis, T., Danovaro, R., Estrada, M., Froglio, C., Galil, B. S., Gasol, J. M., Gertwagen, R., Gil, J., Guilhaumon, F., Kesner-Reyes, K., Kitsos, M.-S., Koukouras, A., Lampadariou, N., Laxamana, E., López-Fé de la Cuadra, C. M., Lotze, H. K. Martin, D., Moullot, D., Oro, D., Raicevich, S., Rius-Barile, J., Saiz-Salinas, J. I., San Vicente, C., Somot, S., Templado, J., Turon, X., Vafidis, D., Villanueva, R., and Voultsiadou, E.: The biodiversity of the Mediterranean Sea: estimates, patterns, and threats, *PLOS one*, 5, e11842, <https://doi.org/10.1371/journal.pone.0011842>, 2010.
- Collins, S., Rost, B., and Ryneison, T. A.: Evolutionary potential of marine phytoplankton under ocean acidification, *Evolutionary Applications*, 7, 140–155, 2014.
- Doglioli, A. M., Grégori, G., d'Ovidio, F., Bosse, A., Pulido, E., Carlotti, F., Lescot, M., Barani, A., Barrillon, S., Berline, L., Berta, M., Bouruet-Aubertot, P., Chirurgien, L., Comby, C., Cornet, V., Cotté, C., Della Penna, A., Didry, M., Duhamel, S., Fuda, J.-L., Gastauer, S., Guilloux, L., Lefèvre, D., Le Merle, E., Martin, A., Mc Cann, D., Menna, M., Nunige, S., Oms, L., Paccaroni, M., Petrenko, A., Rolland, R., Rousselet, L., and Waggoner, E. M.: BioSWOT Med. Biological applications of the satellite Surface Water and Ocean Topography in the Mediterranean, Université Aix-Marseille, <https://doi.org/10.13155/100060>, 2024.
- Dubelaar, G. B., Groenewegen, A. C., Stokdijk, W., Van Den Engh, G., and Visser, J. W.: Optical plankton analyser: A flow cytometer for plankton analysis, II: Specifications, *Cytometry: The Journal of the International Society for Analytical Cytology*, 10, 529–539, 1989.
- Dumas, F.: PROTEVSMED\_SWOT\_2018\_LEG1 cruise, RV *Beautemps-Beaupré*, SHOM, [https://doi.org/10.17183/protevsmed\\_swot\\_2018\\_leg1](https://doi.org/10.17183/protevsmed_swot_2018_leg1), 2018.
- Eggers, S. L., Lewandowska, A. M., Barcelos e Ramos, J., Blanco-Ameijeiras, S., Gallo, F., and Matthiessen, B.: Community composition has greater impact on the functioning of marine phytoplankton communities than ocean acidification, *Global Change Biology*, 20, 713–723, 2014.

- Hernández-Hernández, N., Arístegui, J., Montero, M. F., Velasco-Senovilla, E., Baltar, F., Marrero-Díaz, Á., Martínez-Marrero, A., and Rodríguez-Santana, Á.: Drivers of plankton distribution across mesoscale eddies at submesoscale range, *Frontiers in Marine Science*, 7, 667, <https://doi.org/10.3389/fmars.2020.00667>, 2020.
- Hernández-Hernández, N., Santana-Falcón, Y., Estrada-Allis, S., and Arístegui, J.: Short-term spatiotemporal variability in picoplankton induced by a submesoscale front south of gran Canaria (Canary Islands), *Frontiers in Marine Science*, 8, 592703, <https://doi.org/10.3389/fmars.2021.592703>, 2021.
- Hyrkas, J., Clayton, S., Ribalet, F., Halperin, D., Virginia Armbrust, E., and Howe, B.: Scalable clustering algorithms for continuous environmental flow cytometry, *Bioinformatics*, 32, 417–423, 2016.
- Korkmaz, S., Göksülek, D., and Zararsiz, G.: MVN: An R package for assessing multivariate normality, *R journal*, 6, <https://doi.org/10.32614/RJ-2014-031>, 2014.
- Lévy, M., Ferrari, R., Franks, P. J., Martin, A. P., and Rivière, P.: Bringing physics to life at the submesoscale, *Geophysical Research Letters*, 39, <https://doi.org/10.1029/2012gl052756>, 2012.
- Lévy, M., Jahn, O., Dutkiewicz, S., Follows, M. J., and d’Ovidio, F.: The dynamical landscape of marine phytoplankton diversity, *Journal of The Royal Society Interface*, 12, 20150481, <https://doi.org/10.1098/rsif.2015.0481>, 2015.
- Lévy, M., Franks, P. J., and Smith, K. S.: The role of submesoscale currents in structuring marine ecosystems, *Nature Communications*, 9, 4758, <https://doi.org/10.1038/s41467-018-07059-3>, 2018.
- Lévy, M., Couespel, D., Haëck, C., Keerthi, M. G., Mangolte, I., and Prend, C. J.: The impact of fine-scale currents on biogeochemical cycles in a changing ocean, *Annual Review of Marine Science*, 16, 191–215, 2024.
- Litchman, E., de Tezanos Pinto, P., Klausmeier, C. A., Thomas, M. K., and Yoshiyama, K.: Linking traits to species diversity and community structure in phytoplankton, in: *Fifty years after the “Homage to Santa Rosalia”: Old and new paradigms on biodiversity in aquatic ecosystems*, Springer, 15–28, <https://doi.org/10.1007/s10750-010-0341-5>, 2010.
- Mahadevan, A. and Archer, D.: Modeling the impact of fronts and mesoscale circulation on the nutrient supply and biogeochemistry of the upper ocean, *Journal of Geophysical Research: Oceans*, 105, 1209–1225, <https://doi.org/10.1029/1999JC900216>, 2000.
- Mangolte, I., Lévy, M., Dutkiewicz, S., Clayton, S., and Jahn, O.: Plankton community response to fronts: winners and losers, *Journal of Plankton Research*, 44, 241–258, <https://doi.org/10.1093/plankt/fbac010>, 2022.
- Mangolte, I., Lévy, M., Haëck, C., and Ohman, M. D.: Sub-frontal niches of plankton communities driven by transport and trophic interactions at ocean fronts, *EGUphere* [preprint], <https://doi.org/10.5194/egusphere-2023-471>, 2023.
- 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.
- McLachlan, G. J. and Peel, D.: *Finite mixture models*, John Wiley & Sons, <https://doi.org/10.1002/0471721182>, 2000.
- McLachlan, G. J. and Rathnayake, S.: On the number of components in a Gaussian mixture model, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 4, 341–355, 2014.
- McNeish, D.: On using Bayesian methods to address small sample problems, *Structural Equation Modeling: A Multidisciplinary Journal*, 23, 750–773, 2016.
- McWilliams, J. C.: Oceanic frontogenesis, *Annual Review of Marine Science*, 13, 227–253, <https://doi.org/10.1146/annurev-marine-032320-120725>, 2021.
- Millot, C.: Circulation in the western Mediterranean Sea, *Journal of Marine Systems*, 20, 423–442, 1999.
- Mouling, E. A., Richardson, K., Bendtsen, J., Cetinić, I., and Perry, M. J.: Evidence of small-scale spatial structuring of phytoplankton alpha-and beta-diversity in the open ocean, *Journal of Ecology*, 104, 1682–1695, 2016.
- Moutin, T., Thingstad, T. F., Van Wambeke, F., Marie, D., Slawyk, G., Raimbault, P., and Claustre, H.: Does competition for nanomolar phosphate supply explain the predominance of the cyanobacterium *Synechococcus*?, *Limnology and Oceanography*, 47, 1562–1567, 2002.
- Oms, L., Messié, M., Poggiale, J.-C., Grégori, G., and Doglioli, A.: Fine-scale phytoplankton community transitions in the oligotrophic ocean: A Mediterranean Sea case study, *Journal of Marine Systems*, 246, 104021, 2024.
- Oms, L., Doglioli, A., Messié, M., d’Ovidio, F., Rousselet, L., Capet, X., Lévy, M., Berta, M., Petrenko, A., Bellacicco, M., Barillon, S., and Grégori, G.: “Living on the edge” Fine-scale observations reveal distinct frontal phytoplankton communities, *Nature Communications*, <https://doi.org/10.1038/s412120/v1>, submitted.
- Pearson, K.: Contributions to the mathematical theory of evolution, *Philosophical Transactions of the Royal Society of London A*, 185, 71–110, 1894.
- Peeters, J., Dubelaar, G., Ringelberg, J., and Visser, J.: Optical plankton analyser: A flow cytometer for plankton analysis, I: Design considerations, *Cytometry: The Journal of the International Society for Analytical Cytology*, 10, 522–528, 1989.
- R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, <https://www.R-project.org/> (last access: 19 January 2026), 2021.
- Scrucca, L., Fraley, C., Murphy, T. B., and Raftery, A. E.: Model-Based Clustering, Classification, and Density Estimation Using mclust in R, Chapman and Hall/CRC, <https://doi.org/10.1201/9781003277965>, 2023.
- Shaw, C. T., Bi, H., Feinberg, L. R., and Peterson, W. T.: Cohort analysis of *Euphausia pacifica* from the Northeast Pacific population using a Gaussian mixture model, *Progress in Oceanography*, 191, 102495, 2021.
- Siokou-Frangou, I., Christaki, U., Mazzocchi, M. G., Montresor, M., Ribera d’Alcalá, M., Vaqué, D., and Zingone, A.: Plankton in the open Mediterranean Sea: a review, *Biogeosciences*, 7, 1543–1586, <https://doi.org/10.5194/bg-7-1543-2010>, 2010.
- Sournia, A., Chrdiennot-Dinet, M.-J., and Ricard, M.: Marine phytoplankton: how many species in the world ocean?, *Journal of Plankton Research*, 13, 1093–1099, 1991.

- Stan Development Team: RStan: the R interface to Stan, R package version 2.32.7, <https://mc-stan.org/> (last access: 19 January 2026), 2020.
- Székely, G. J. and Rizzo, M. L.: Testing for equal distributions in high dimension, *InterStat*, 5, 1249–1272, 2004.
- Taylor, A. G., Goericke, R., Landry, M. R., Selph, K. E., Wick, D. A., and Roadman, M. J.: Sharp gradients in phytoplankton community structure across a frontal zone in the California Current Ecosystem, *Journal of Plankton Research*, 34, 778–789, <https://doi.org/10.1093/plankt/fbs036>, 2012.
- Thyssen, M., Mathieu, D., Garcia, N., and Denis, M.: Short-term variation of phytoplankton assemblages in Mediterranean coastal waters recorded with an automated submerged flow cytometer, *Journal of Plankton Research*, 30, 1027–1040, 2008.
- Thyssen, M., Garcia, N., and Denis, M.: Sub meso scale phytoplankton distribution in the North East Atlantic surface waters determined with an automated flow cytometer, *Biogeosciences*, 6, 569–583, <https://doi.org/10.5194/bg-6-569-2009>, 2009.
- Thyssen, M., Alvain, S., Lefèvre, A., Dessailly, D., Rijkeboer, M., Guiselin, N., Creach, V., and Artigas, L.-F.: High-resolution analysis of a North Sea phytoplankton community structure based on in situ flow cytometry observations and potential implication for remote sensing, *Biogeosciences*, 12, 4051–4066, <https://doi.org/10.5194/bg-12-4051-2015>, 2015.
- Tzortzis, R., Doglioli, A. M., Barrillon, S., Petrenko, A. A., d’Ovidio, F., Izard, L., Thyssen, M., Pascual, A., Barceló-Llull, B., Cyr, F., Tedetti, M., Bhairy, N., Garreau, P., Dumas, F., and Gregori, G.: Impact of moderately energetic fine-scale dynamics on the phytoplankton community structure in the western Mediterranean Sea, *Biogeosciences*, 18, 6455–6477, <https://doi.org/10.5194/bg-18-6455-2021>, 2021.
- Tzortzis, R., Doglioli, A. M., Messié, M., Barrillon, S., Petrenko, A. A., Izard, L., Zhao, Y., d’Ovidio, F., Dumas, F., and Gregori, G.: The contrasted phytoplankton dynamics across a frontal system in the southwestern Mediterranean Sea, *Biogeosciences*, 20, 3491–3508, <https://doi.org/10.5194/bg-20-3491-2023>, 2023.
- Yang, K., Meyer, A., Strutton, P. G., and Fischer, A. M.: Global trends of fronts and chlorophyll in a warming ocean, *Communications Earth & Environment*, 4, 489, <https://doi.org/10.1038/s43247-023-01160-2>, 2023.