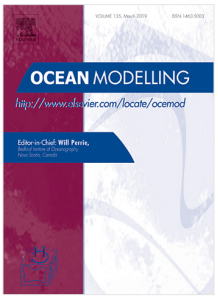
Journal Pre-proof



Assessment of a regional physical-biogeochemical stochastic ocean model. Part 2: Empirical consistency

Vassilios D. Vervatis, Pierre De Mey-Frémaux, Nadia Ayoub, John Karagiorgos, Stefano Ciavatta, Robert J.W. Brewin, Sarantis Sofianos

PII: S1463-5003(21)00020-2
DOI: https://doi.org/10.1016/j.ocemod.2021.101770
Reference: OCEMOD 101770

To appear in: *Ocean Modelling*

Received date : 18 May 2020
Revised date : 19 January 2021
Accepted date : 7 February 2021

Please cite this article as: V.D. Vervatis, P. De Mey-Frémaux, N. Ayoub et al., Assessment of a regional physical-biogeochemical stochastic ocean model. Part 2: Empirical consistency. *Ocean Modelling* (2021), doi: https://doi.org/10.1016/j.ocemod.2021.101770.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2021 Published by Elsevier Ltd.

Journal Pre-proof

Assessment of a regional physical-biogeochemical stochastic ocean model. Part 2: empirical consistency

Vassilios D. Vervatis (1), Pierre De Mey-Frémaux (2), Nadia Ayoub (2), John Karagiorgos (1), Stefano Ciavatta (3, 4), Robert J. W. Brewin (3, 5), and Sarantis Sofianos (1)

(1) University of Athens, Department of Physics, Athens, Greece. (2) LEGOS/CNRS, Toulouse, France. (3) Plymouth Marine Laboratory, Plymouth, UK. (4) National Centre for Earth Observation, Plymouth, UK. (5) University of Exeter, College of Life and Environmental Sciences, Penryn, Cornwall, UK.

Correspondence to: Vassilios D. Vervatis (vervatis@oc.phys.uoa.gr)

Abstract. In this Part 2 article of a two-part series, observations based on satellite missions were used to evaluate the empirical consistency of model ensembles generated via stochastic modelling of ocean physics and biogeochemistry. A high-resolution Bay of Biscay configuration was used as a case study to explore the model error subspace in both the open and coastal ocean. In Part 1 of this work, three experiments were carried out to generate model ensembles by perturbing only physics, only biogeochemistry, and both of them simultaneously. In Part 2 of this work, empirical consistency was checked, first by means of rank histograms projecting the data onto the model ensemble classes, and second, by pattern-selective consistency criteria in the space of "array modes" (eigenvectors of the representer matrix). Rank histograms showed large dependency on geographical region and on season for sea surface temperature (SST), sea-level anomaly (SLA), and phytoplankton functional types (PFT), shifting from consistent model-data configurations to large biases because of model ensemble underspread. Consistency for SST array modes was found to be verified at large, small and coastal scales soon after the ensemble spin-up. Array modes for the along-track sea-level showed useful consistent information at large scales and at the mesoscale; for the gridded SLA was verified only at large scale. Array modes showed that biogeochemical model uncertainties generated by stochastic physics, were effectively detected by PFT measurements at large scales, as well as at mesoscale and small-scale. By contrast, perturbing only biogeochemistry, with an identical physical forcing across the ensemble, limits the potential of PFT measurements at detecting and possibly correcting small-scale biogeochemical model errors. When an ensemble was found to be inconsistent with observations along a particular direction (here, an array mode), a plausible reason is that other error processes must have been active in the model, in addition to the ones at work across the ensemble.

Keywords: stochastic modelling, ensembles, phytoplankton functional types, prior error covariances, array modes, Bay of Biscay

1 Introduction

In the coastal parts of regional ocean models, many factors can complicate the assimilation of data. One of them is the characterization and specification of model errors, which are critical in any assimilation scheme, but extremely challenging in the coastal zone. Ocean model errors strongly depend on spatiotemporal scales, though any attempt at separation is confounded by strong nonlinearity in the dynamics that can couple variability at different frequencies and wavenumbers (*Auclair et al.*, 2003). Another factor complicating data assimilation is the specification of observational errors, which are made up of measurement (usually small) and representativity errors (usually large or unknown), and cross-correlations between observations (usually unknown) (*Oke and Sakov*, 2008).

Most studies point at the benefit of advanced assimilation methods with built-in error propagation (*Kourafalou et al.*, 2015a; 2015b), such as the Ensemble Kalman Filter (EnKF;

Evensen, 2003) and variants. As a first step, one must characterize the forecast uncertainties under various error regimes (situations where the model is in error), either in response to time-varying forcing errors (e.g. uncertainties in boundary conditions; *Ghantous et al.*, 2020) or due to internal sources (e.g. model parameterizations; *Brankart et al.*, 2015), and include realistic error dynamics through stochastic modelling. For instance, depending on the wind regime, or seasonal baroclinic instabilities of the slope current, shelf errors can be different in terms of spatiotemporal scales (*Auclair et al.*, 2003). Biogeochemical model errors can also be different stemming from unresolved scales and biodiversity (*Garnier et al.*, 2015). It should be noted that stochastic modelling, in itself, may not yield realistic error dynamics, for example, it could inflate the ensemble (*Anderson*, 2009).

In the companion article Part 1 (*Vervatis et al.*, 2021), we configured a high-resolution (1/36°) stochastic ocean model for the Bay of Biscay performing physical-biogeochemical ensemble simulations. We carried out quantitative assessment of model-data misfits and qualitative evaluation of multivariate incremental analysis. We found that the skill of the perturbation method to generate model errors was improved (in general) for physics compared with biogeochemical perturbations in source and sink terms (i.e. larger ensemble spread when physics is perturbed) and that the data assimilation performance to correct those model errors, was largely dependent on the chosen multivariate analysis. On the other hand, the biogeochemical model spread was found under-dispersive, leading to disjoint supports[1] of the model and data probability density functions (pdf), and performance was mainly defined by the assimilation of ocean colour data, possibly because of weak cross-covariances between ocean physics and biogeochemistry.

In light of these findings, in this companion article Part 2, we present two methods to evaluate the consistency with respect to observations (hereafter empirical consistency) of a stochastic coupled ocean model, which comes in the form of an ensemble of multivariate ocean states. We focus on observational products including physics and biogeochemistry, derived from satellite missions monitoring upper-ocean properties, such as the sea surface temperature (SST), the sea-level anomaly (SLA) and an ocean colour product of phytoplankton functional types (PFT; *Brewin et al.*, 2010; 2015; 2017). Several biogeochemical studies have focused on improving the model's predictive skill using remote sensing and in-situ observations, incorporating data assimilation and probabilistic attribution (*Candille et al.*, 2015; *Song et al.*, 2016; *Gharamti et al.*, 2017; *Mattern et al.*, 2018; *Ford*, 2019). Here, we focus our analysis on the use of ocean colour PFT observations following recent advances in data assimilation to improve marine ecosystem simulations (*Ciavatta et al.*, 2018; 2019).

In a probabilistic framework, the support of the joint pdf of observed and forecast values should be non-null in order to enable data assimilation. In other words, the prior model state and the data must be compatible with each other given their respective uncertainties in order for assimilation to be meaningful. Therefore, one important question we address in this study is the following: is the distribution of the forecast errors estimated from the ensemble (the prior distribution) compatible with the distribution of the data to be assimilated? Another important question is about how ensembles (e.g. the state of a stochastic model) can be validated – in effect, ensembles are not used solely for assimilation – other uses include: array design, probabilistic forecasting, a learning base for artificial intelligence applications, etc.

As a first step, we assess the empirical consistency of ensembles by means of rank histograms (also referred to as Talagrand diagrams; *Candille and Talagrand*, 2005) projecting the data

[1] The support of a probability density function *(pdf)* is the smallest closed set outside of which the *pdf* vanishes. For a *pdf* defined in $\mathbb{R}$, the *pdf* envelope (i.e. here the ensemble envelope) is the range between the minimum and maximum values of the support.

onto the model ensemble classes. Rank histograms are useful for diagnosing reliability and inferring systematic biases in an ensemble prediction system (*Hamill,* 2001). As a second step, we apply a *consistency diagnostic on innovations* inspired by the similarly-named diagnostic developed in *Andersson* (2003) and *Desroziers et al.* (2005) in the framework of the assessment of the well-posedness of data assimilation schemes. Our implementation is however specific in two ways: (1) we use that metric here outside of a data assimilation scheme, to check how consistent our ensembles are with respect to innovations; (2) we project the innovation consistency diagnostic on *Array Modes*, with the objective of permitting pattern-dependent consistency analysis. Our definition of array modes follows previous publications (*Le Hénaff et al.,* 2009; *Lamouroux et al.*, 2016; and *Charria et al.*, 2016). This is detailed below.

This study is organized as follows. Section 2 describes the stochastic approach to generate model ensembles, following the companion article Part 1. The specifications of the observational networks are presented in Section 3, including the description of the ocean colour PFT. A summary of the consistency analysis framework based on rank histograms and array modes is given in Section 4, which presents also a new criterion in "array space". The results and conclusions are discussed in Sections 5 and 6.

2 Stochastic modelling

We used the NEMO platform (Nucleus for European Modelling of the Ocean; http://www.nemo-ocean.eu/; *Madec,* 2012) and its biogeochemical component PISCESv2 (Pelagic Interactions Scheme for Carbon and Ecosystem Studies volume 2; *Aumont et al.,* 2015). The ocean model domain encompasses the Bay of Biscay and the western part of the English Channel (*Quattrocchi et al.,* 2014; *Vervatis et al.,* 2016). The physical model is coupled online (one-way with high coupling frequency for the conservation of tracers) with a biogeochemical model at $1/36°$ horizontal resolution.

Three seasonal-range ensembles were carried out using stochastic modelling of ocean physics and biogeochemistry as part of the project SCRUM (Stochastic Coastal/Regional Uncertainty Modelling; cf. companion article Part 1 by *Vervatis et al*. 2021). Perturbations were modelled using first-order auto-regressive processes - AR(1) - in the context of stochastic perturbed parameterized tendencies (SPPT; *Buizza et al.*, 1999) and stochastic perturbed parameters (SPP; *Ollinaho et al.*, 2017). The AR(1) processes were different for each perturbed tendency or parameter, and varied under the assumption of spatiotemporal correlated scales. A deterministic free run was performed from July 2011 to November 2011, to serve as a five-month model spin-up starting from ocean analyses for the whole state vector, and then extended from December 2011 to June 2012, to serve as reference for the main physical and biogeochemical processes in the region and deduce the statistical properties for the AR(1) stochastic parameterizations. The ensemble simulations were initialized from the five-month spin-up and carried out from December 2011 to June 2012. The 40-member ensembles were used to estimate the forecast error covariance matrix perturbing different sources of model errors.

In Table 1, we summarize the stochastic protocol for the three ensembles, hereafter referring to them as: EnsP - perturbing physics under the assumption of atmospheric forcing uncertainties and model uncertainties in physical parameterizations; EnsB - perturbing biogeochemistry under the assumption of model uncertainties in biogeochemical sources and sinks; and EnsPB - perturbing both physics and biogeochemistry. The biogeochemical model parameters were not perturbed. The initial conditions were also not perturbed. The EnsPB is statistically identical to EnsP for model errors in ocean physics, because there is no feedback to the physics from the biogeochemical model. By contrast, EnsPB is statistically different to EnsB and EnsP because uncertainties in the physical forcing are found to have a large impact

on biogeochemical model properties. Also, for the EnsB the physics and hence the circulation are the same for all members. For details of the physical-biogeochemical model and for model data ensemble-based misfits, the reader is referred to the companion article Part 1.

The model uncertainties in physical and biogeochemical properties in the upper ocean are described in Part 1. Here we illustrate the typical patterns that we obtained for the spread in SST, sea surface height (SSH) and chlorophyll abundance at specific dates that will also serve as a reference for the analyses in Section 5. Figure 1a shows, the Bay of Biscay SST from the deterministic simulation on May 31, 2012. Fig. 1b shows the EnsP model ensemble spread for SST on the same date. Figs. 1a-b show that the ocean model solution and its associated error structures for SST are statistically consistent. In more detail, error regimes are linked to physical processes controlling SST, for example, the extension of the Ushant thermal tidal front west of Brittany in the English Channel. At that time of the year, the stratification is strong and the freshwater front (due to the Loire and Gironde rivers runoff in the Armorican shelf) leads to a large spread locally exceeding $0.8\ °C$. Figs. 1c-d give an insight of the model circulation patterns on February 25, 2012, depicted by SSH and their error structures. After a two-month spin-up period, the maxima of SSH model spread are collocated with mesoscale eddies observed in the abyssal plain, and with (most likely) inertial barotropic waves observed over the shelves. The perturbation mechanism works throughout the whole period. Therefore, the SSH spread continues to grow in the abyssal plain due to mesoscale decorrelation of eddies across members (not shown).

Figure 2a shows the total chlorophyll abundance of the deterministic simulation at the onset of the spring bloom on March 28, 2012. In Figs. 2b-d, and for the same date, we show model uncertainties in total chlorophyll for the three ensembles EnsP, EnsB and EnsPB. We emphasize two important findings: first, the model errors in physics have a larger impact on chlorophyll model uncertainties compared with those generated by perturbing the biogeochemical model source and sink terms; second, the chlorophyll spread is increased when model physics and biogeochemistry are perturbed simultaneously. However, there are also exceptions of decreasing spread (e.g. occurring locally in the presence of coherent eddies). Interestingly, an uneven ensemble spread is also observed between nanophytoplankton and diatoms chlorophyll (Figs. 2e-f). Model errors for each phytoplankton type follow the different patterns of their chlorophyll concentrations, with large uncertainties observed in the abyssal plain for nanophytoplankton, and on the shelves and the English Channel for diatoms.

3 Observational networks

In this study, although we examine the same broad categories of observations as in the companion article Part 1, we have included more satellite products observing the upper-ocean physical and biogeochemical properties. Therefore, for the sake of clarity, we have included a full description of observations. We consider global and regional satellite products monitoring the Bay of Biscay, using high-level L3 and L4 merged data from multi-missions for both real-time and delayed modes. All datasets are provided at daily frequency. This level of data quality is often required in ocean forecasting systems for model validation and data assimilation. Table 2 summarizes the datasets used for the empirical consistency analysis and their specifications are given briefly in the paragraphs below.

3.1 SST gridded observations

We chose two high-resolution SST L4 products, namely the OSTIA SST global dataset and the regional Atlantic European north west shelf (NWS) SST. High-resolution SST data are necessary when it comes to validate eddy-resolving model ensembles. Both networks are able

to resolve mesoscale eddy variability and frontal activity near the coasts, ranging from a few kilometres over the shelves to scales of tens of kilometres in the open ocean.

The OSTIA SST is provided on a global regular grid at 0.05^o resolution (*Donlon et al*., 2012) and is by construction free of diurnal variability with a reference depth of 10 meters, referred also as foundation SST. The reconstructed fields consist of L4 daily gap-free maps. The product includes multi-sensor measurements from both infra-red and microwave instruments, as well as in-situ observations retrieved from ships of opportunity, drifters and buoys. The data were reprocessed using optimal interpolation and include a bias correction.

The NWS SST product is delivered at 0.04^o high horizontal resolution and consists of L4 daily gap-free regional maps. Calculations are derived from infra-red measurements using only night-time observations, and thus the NWS SST data are not affected by diurnal warming.

3.2 SLA gridded and along-track observations

We selected two products based on different levels of processing. The L3 along-track sea-level product is based on several altimetric missions, specifically on Envisat and Cryosat-2 satellites. The L4 gridded product was generated by reprocessing multi-mission altimetry data over many years.

For the along-track data, we used the filtered sub-sampled L3 product with 14 km distance between successive points along the altimetry track. The effective resolution of the data is coarser than the aforementioned distance (*Pujol et al*., 2016). The inverse barometer response and the tides have been removed from the data (and the model, cf. companion article Part 1).

The L4 product is a merge of multi-mission satellites spanning the last two decades. The result is a homogenous and consistent gridded dataset at a resolution of 0.25^o. The scales resolved by the L4 gridded product depend on various factors, such as the quality of the L3 input data, the sampling of the altimeter constellation and the optimal interpolation, which limits the capability of the data to resolve part of the mesoscale (*Chelton et al*., 2011).

3.3 Phytoplankton functional types ocean colour observations

We selected the regionally-tuned PFT satellite model of *Brewin et al*. (2017). The approach is a modification of the three-component model of *Brewin et al*. (2010; 2015) and produces estimates of the chlorophyll concentration of four phytoplankton groups (picophytoplankton, nanophytoplankton, diatoms and dinoflagellates), as a function of the total chlorophyll concentration, accounting for the influence of SST on model parameters. The model was tuned and validated using a large dataset collected in the North Atlantic, inclusive of our region of interest. The approach was run using ocean colour data from the ESA ocean colour climate change initiative (OC-CCI) and OSTIA SST data. Products were produced on a regular grid at 0.05^o resolution. The data has been used successfully in PFT data assimilation exercises in the North Atlantic (*Ciavatta et al*., 2018; *Skakala et al*., 2018). The products were developed in the framework of the Copernicus Marine Environment Monitoring Service (CMEMS) "Towards Operational Size-class Chlorophyll Assimilation (TOSCA)" Service Evolution project.

The mismatch between biogeochemical model variables and ocean colour products has been highlighted recently in an IOCCG Report No. 19 (*IOCCG*, 2020). The biogeochemical model PISCES distinguishes two classes of chlorophyll, namely the nanophytoplankton and diatoms. These two chlorophyll classes were given this name, as the traits of these modelled functional types are closest to these two abundant groups in the natural world. However, these two classes are designed to be representative of the entire phytoplankton community, and therefore are far broader than the names imply. Each of the two model compartments encompasses a large range

of cell sizes, approximately representing smaller and larger phytoplankton types in the biogeochemical model. In order to compare the observations with the model ensembles, we decided to follow a size class-based categorization (*Sieburth et al.*, 1978) and group the four ocean colour PFT into two phytoplankton types and keep the total chlorophyll concentration in the data unchanged (i.e. the sum of all four types). For this, we combined the concentrations of ocean colour pico- and nano- PFT into one small class comparable with the PISCES nanophytoplankton, and the diatoms and dinoflagellates ocean colour PFT into one large class comparable with the PISCES diatoms. This categorization took into consideration many processes and biogeochemical cycles in the model, for example, not just the silicate cycle but also the carbon and nitrogen cycles, and the size-selective feeding by microzooplankton and mesozooplankton. However, it was unavoidable to combine one PFT with characteristic silicate limitation with one without (i.e. diatoms and dinoflagellates; in supplementary material we provide also an analysis without this combination). The choice to include all four PFT was supported by the fact that their sum ensures the total biomass (chlorophyll) and therefore, model and satellite total chlorophyll data can be compared like-for-like. In addition, the total chlorophyll concentration (sum of the four PFT) in the model is of the same magnitude as the total chlorophyll of other ocean colour products used for the validation of our ensembles, as shown in the companion article Part 1 and in the supplementary material.

3.4 Observational errors

An estimate of the observational error is given by the measurement and representativity errors (usually unknown) (*Desroziers et al.*, 2005; *Oke and Sakov*, 2008; *Janjić et al.*, 2018). The measurement errors are usually small and refer to the instrument's sensor accuracy and data processing. The representativity errors are usually large and can account for different model and observation sampling schemes, or for the physical-biogeochemical signal that is contained in the observations but is not represented in the model. Some of the data used here were retrieved from archives and included a spatial distribution of their errors. While this information is important for a dynamically heterogeneous system such as the Bay of Biscay, we chose not to use it because it would be hard to interpret pattern-dependent consistency results. Instead, we used a representative constant error for each observational network provided by the CMEMS infrastructure and the data providers.

Assuming that errors are uncorrelated and that the innovation variance is therefore close to the sum of model and data error variances, we considered an error standard deviation for SST equal to $0.5\ ^{o}C$, for SLA equal to $0.05\ m$, and for PFT equal to $0.3\ mg/m^3$ (Table 2). The satellite chlorophyll *a* data are provided with a scaled (%) observational error in comparison to the signal and therefore, the error has spatial distribution. We chose this constant PFT error over an error that scales with the satellite chlorophyll *a* signal for the reason explained in the paragraph above. The static PFT error is representative for the region, though moderately larger for total chlorophyll and nanophytoplankton in the open ocean, and underestimated for diatoms over the shelves, if compared with the PFT errors estimated by *Brewin et al*. (2017).

In the following Section 4, and for each of the consistency analysis methods, we discuss the perturbation of observations to generate data distributions and the anamorphosis functions applied to transform those distributions in the proper space (*Simon and Bertino*, 2009).

4 Consistency analysis framework

In this study, we implemented two empirical consistency analysis methods to evaluate model ensembles with respect to observations. The first approach was based on rank histograms and the second on array modes. The calculations were performed with a toolbox (nicknamed "scrumcat": SCRUM consistency analysis toolbox) built upon the Sequoia Data Assimilation

Platform (SDAP; De Mey-Frémaux, 2020: https://sourceforge.net/projects/sequoia-dap/). The implementation details are given below.

4.1 Rank histograms

The rank histogram (Talagrand diagram; *Candille and Talagrand*, 2005) is a verification metric testing the empirical consistency between model samples and observations. An ensemble of m members defines $m+1$ ranks including the rank of one observation sorted within the model ensemble. The steps to calculate rank histograms are (1) project ensemble samples in data space using an observation operator, (2) sorting the ensemble samples by value, (3) ranking the observed value within the sorted ensemble at each observation location, (4) tally over many observations in space and time. We note that for the rank histograms the observations are not perturbed and the model ensemble distributions are not checked for gaussianity.

The rank histogram is useful for determining the reliability of ensemble forecasts (*Hamill,* 2001) and is often used as a tool to infer systematic biases of an ensemble prediction system, as well as differing spreads between model and data (e.g. small spread translates into most data falling into the outer classes of the rank histogram). The rank histogram flatness is a measure of the reliability of the ensemble prediction system and is estimated according to the definition of statistical consistency by *Anderson* (1996) and *Candille and Talagrand* (2005):

$$\Delta = \sum_{i=1}^{m+1}\left(S_i - \frac{p}{m+1}\right)^2,\ \Delta_o = \frac{p*m}{m+1},\ \delta = \frac{\Delta}{\Delta_o} \sim 1, \tag{1}$$

where p is the number of observations, m the number of ensemble members, S_i the outcomes in each rank (i.e. number of observations of rank i), Δ the deviation of the histogram from flatness, Δ_o the expectation for a reliable system and δ the ratio measuring the reliability for a scalar variable of the ensemble prediction system.

4.2 Ensemble consistency in array space

Rank histograms provide a way to check ensemble empirical consistency in a global, distribution-like manner. In a complementary manner, we wish to complete the global analysis with pattern-selective consistency analysis, using patterns which would be (A) hierarchized, (B) representative of error covariances (as estimated from the ensemble) and (C) observable by the data array. To that end, we chose to assess consistency in the space of *Array Modes*, hereafter nicknamed "array space".

We use whenever possible the unified notations of data assimilation as in *Ide et al.* (1997). We define *array modes* $\boldsymbol{\mu}$ as the eigenmodes of the *scaled representer matrix* $\boldsymbol{\chi}$ defined as:

$$\boldsymbol{\chi} \equiv \boldsymbol{R}^{-\frac{1}{2}}\boldsymbol{H}\boldsymbol{P}^{f}\boldsymbol{H}^{T}\boldsymbol{R}^{-\frac{1}{2}} = \boldsymbol{\mu}\boldsymbol{\sigma}\boldsymbol{\mu}^{T} \tag{2}$$

where $\boldsymbol{P}^f$ is the *forecast* (*prior*) *error covariance matrix*, here approximated as an ensemble covariance, $\boldsymbol{R}$ is the *observational error covariance matrix*, and $\boldsymbol{H}$ is the "classic" linear *observation operator*. Because an observing array will not be concomitant, all operators are four-dimensional in that they span time in addition to space.

These array modes have been discussed in several articles in the literature. First, *Le Hénaff et al.* (2009) presented the theoretical background for the representer matrix spectra methodology, aiming at assessing the performance of observational networks at detecting model errors. In their study, prior model errors were generated via stochastic modelling of the wind forcing and various altimetry and in-situ array deployment strategies were tested. Two other studies followed, based on array modes, by *Lamouroux et al.* (2016) and *Charria et al*. (2016), where authors designed optimal observation network experiments based on array modes, for the future implementation of efficient integrated ocean observing systems monitoring the coastal

environment. Array modes are the eigenmodes of the representer matrix as approximated as an ensemble covariance (e.g. from Section 2 ensembles). It follows that array modes meet the pattern definition criteria (A-C) above.

The matrix $\boldsymbol{\chi}$ in Eq. (2) expresses how the array "sees" the model uncertainties; the diagonal matrix $\boldsymbol{\sigma} = diag\{\sigma_k\}$ offers the same information as $\boldsymbol{\chi}$ in array space, but in a diagonal form: it is the *array mode spectrum*. Intuitively, the higher the eigenvalues in $\boldsymbol{\sigma}$, the better the array can detect (and help correct) errors of a prior estimate. *Le Hénaff et al.* (2009) show that a useful choice of a discriminating spectral value is 1. In order to clarify this choice, let us consider the *innovation vector* $\boldsymbol{d}$ and its second-order statistics:

$$\boldsymbol{d} \equiv \boldsymbol{y^o} - \boldsymbol{H x^f} \tag{3}$$

$$\langle \boldsymbol{d d^T} \rangle = \boldsymbol{R} + \boldsymbol{H P^f H^T} \tag{4}$$

where $\boldsymbol{y^o}$ is the *observation vector* and $\boldsymbol{x^f}$ is the *forecast* (*prior) state vector*. Equation (4) is the same as Eq. (1) in *Desroziers et al*. (2005). An intuitive criterion of array performance is as follows:

1) If observational errors $\boldsymbol{R}$ dominate in Eq. (4), then most of the model-data discrepancies in $\boldsymbol{d}$ are attributable to observational error, and observations are not being very useful at detecting model uncertainties.

2) If prior state errors $\boldsymbol{H P^f H^T}$ (the *representer matrix*) dominate, then most of the discrepancies are attributable to model uncertainties, and observations can be expected to be useful at identifying and correcting them.

Let us examine this criterion in array space. Using the orthogonality of array modes, it is easy to show that the innovation covariance in Eq. (4) projected into array space, becomes:

$$\boldsymbol{E^i} \equiv \boldsymbol{\mu^T R^{-\frac{1}{2}}} < \boldsymbol{d d^T} > \boldsymbol{R^{-\frac{1}{2}} \mu} = \boldsymbol{I} + \boldsymbol{\sigma}, \tag{5}$$

hence the choice of 1 as a discriminating spectral value. It follows:

Criterion ArM1: Array performance is measured by the number of ranks k *for which:*

$$1 \leq \sigma_k$$

Note that (1) this first criterion is only based on the space-time sampling scheme and on the ensemble covariance; it does not require the actual values of observations; (2) it is not yet expressed as an empirical consistency criterion for our ensemble. Now, when we have the values of observations, Eq. (4) can be used to derive an empirical ensemble consistency criterion. Innovation Eq. (3) can be formed, and $\langle \boldsymbol{d d^T} \rangle$ can be calculated empirically and compared to the right-hand side of Eq. (4). In effect, the innovation spread is the result of prior uncertainties of both the model and observations; therefore, that spread should be statistically consistent with the sum of prior model uncertainty estimates and observational uncertainty estimates.

As above, this criterion is more advantageously examined in array space, i.e. in the form of Eq. (5). Since $\boldsymbol{I} + \boldsymbol{\sigma}$ is diagonal, we focus on the diagonal of $\boldsymbol{E^i} = diag\{e_k^i\}$, i.e. the innovation variance in array space. Our second criterion writes:

Criterion ArMCA1: Empirical consistency is measured by the number of ranks k for which:

$$1-\tau \leq \frac{1+\sigma_k}{e_k^i} \leq 1+\tau$$

with τ the user-selected tolerance (e.g. $\tau = 0.1$ for 10% tolerance).

In this study, we use the ArM1 and ArMCA1 criteria as implemented in the ArM tools library *(De Mey-Frémaux, pers. comm.,* 2020*),* distributed as open source within the SDAP assimilation platform (*De Mey-Frémaux, 2020:* https://sourceforge.net/projects/sequoia-dap/). We check consistency in array space, also qualifying consistency results with their associated multivariate patterns in observation space. In short, criterion ArM1 is an "array performance" criterion based on how the observational array "observes" model uncertainties; criterion ArMCA1 tests ensemble empirical consistency in terms of variances; other criteria are present in the ArM tools but are not used here.

Our approach can be seen as an extension of the *Desroziers et al.* (2005) consistency diagnostic on innovations (their Eq. (1)), which we project on the space of array modes.

We remove ensemble averages prior to analysis. In contrast to rank histogram analysis, observations are perturbed, using a Gaussian random number to generate data distributions with proper error estimates for each network (Table 2). Although the methodology allows for nondiagonal *observational error covariance matrix*, we consider it to be diagonal in this study, meaning that observational errors are considered statistically independent from each other. In case of non-Gaussian distributions, an anamorphosis transformation is applied so as to calculate array modes in the transformed space. For instance, in this study, model and data chlorophyll samples are log-transformed prior to computing array modes.

5 Results

5.1 Rank histograms consistency analysis

In Figs. 3-5, we show Hovmöller plots of rank histograms as a means to analyse space-time ensemble consistency. Under-dispersive rank histograms are characterized by a U-shape diagram, since observations fall mostly in the outer ranks of the ensemble. In the context of a Hovmöller plot of rank histograms, under-dispersion is illustrated by contrasting colours between outer and central ranks (Fig. 3). In the same line of thinking, bias is depicted by having large values only on one side of the histogram and flatness (i.e. $\delta \sim 1$) is verified for evenly distributed pdf values at $\frac{1}{41} \sim 0.024$ across all ranks. Rank histograms for SST verify that EnsP is generally under-dispersive and biased, though results vary depending on seasons and on the use of different networks (Fig. 3). On the other hand, there are cases where observations are evenly distributed within the ensemble spread despite persistent under-dispersion. The rank histograms are in close agreement with the consistency analysis based on the OSTIA SST innovation samples, presented in the companion article Part 1.

In order to illustrate dependency based on geographical region, we focus on three distinct areas in the Bay of Biscay, namely the abyssal plain, the Armorican shelf and the English Channel (Fig. 3). There is evidence of local error regimes where EnsP is either warm or cold biased with respect to observations. From the latter we can link SST observational biases and underestimated model errors potential to local dynamics. For instance, rank histogram flatness is occasionally verified in the abyssal plain, whereas in the Armorican shelf and the English Channel the ensemble is shifting from being warm biased in winter to cold biased in spring with respect to observations. The results are associated with the spring shoaling of the thermocline in the open ocean, as well as to coastal processes controlled by frontal activity of

freshwater river discharges over the Armorican shelf, and to tidal mixing in the English Channel. There is also a marked difference between the two SST networks when used to validate consistency against EnsP. The OSTIA SST fits better within the ensemble classes during winter, whilst the consistency for the NWS SST is improved during spring. Overall, the EnsP is notably under-dispersive and on average warm biased with respect to observations, whilst cold biased in the English Channel.

Figure 4a shows a Hovmöller plot of rank histograms and Fig. 4b a map of ranks as a means to analyse space-time ensemble consistency for sea-level. The SLA model equivalent is calculated using a mean dynamic topography as in the companion article Part 1. The gridded SLA product appears to be under-dispersive having a bias against the ensemble shifted between seasons (Fig. 4a). This bias is most likely attributed to the seasonal steric cycle contributing to sea-level variability being weak in the observations. As a consequence, rank histogram flatness is not verified throughout the series.

Similar results are presented in the map of rank histograms for the along-track SLA product (Fig. 4b). As in the companion article Part 1, we focus our analysis in late-winter for a few consecutive days after having a model ensemble statistical spin-up for EnsP (Fig. 1d) and also observing noticeable sea-level variability in the satellite observations. Strong biases are apparent, both in the deeper areas of the domain and in the shelves. These biases also coincide with the location of inconsistent tidal residual signals (both in the model and data) in the English Channel. Contrasting colours in the Celtic Sea and near the shelf break hint at the presence of high-frequency error processes currently unaccounted for in the model ensemble. Under-dispersion and missing errors in the high-frequency band (e.g. open boundary conditions in the shelves, which do not deal properly with high-frequency processes), is also confirmed by the fact that there are only a few central ranks depicted in the map.

In Fig. 5, we present Hovmöller plots of rank histograms as a means to analyse ensemble consistency with respect to different classes of chlorophyll. For this, we use the PFT dataset against the model ensemble EnsPB which has the largest spread among the three ensembles (Figs. 2b-d). Rank histograms for EnsPB verify that all ensembles are under-dispersive, with EnsB being the least dispersive and EnsPB the most dispersive. The chlorophyll rank histograms for EnsP (not shown) are comparable to those of EnsPB (Fig. 5), but with lower levels of consistency. Rank histograms indicate a rather long statistical biogeochemical spin-up period, on the order of 3 months (verified by rank histogram flatness for evenly distributed pdf values), and persistent model underestimation of chlorophyll abundance. The spin-up period corresponds only to the winter season characterised by low primary production. After the spin-up period, the consistency is in general improved for total chlorophyll, as a result of the onset of the nanophytoplankton spring bloom. On the other hand, diatoms appear to degenerate the consistency of total chlorophyll. During successive spring blooms, inconsistencies are seen in the peak, but better agreement in the onset and relaxation of the blooms. This is explained by the fact that ocean colour peak values are of an order of magnitude larger compared with model outputs, whilst PFT data fall within the ensemble spread during the onset and relaxation phases of a bloom event. Overall, the most consistent configurations are the nanophytoplankton in early-spring (i.e. primary bloom mainly in the abyssal plain) and the diatoms in late-spring (i.e. secondary bloom mainly coastal and over the shelves).

5.2 Array-space empirical consistency analysis

In Fig. 6a, we show in a Hovmöller plot the variations in time of OSTIA SST array mode spectra vs. modal rank (i.e. the rank of the array mode), including an EnsP consistency check against OSTIA SST data. We used 39 array modes, corresponding to the 39 “degrees of freedom” characterizing a 40-member de-biased ensemble. The ensemble spin-up period

appears as a period of low eigenvalues near the beginning of the time series, lasting from a few days (in the head of the spectra) to a few weeks (in the tails). The array performance at "detecting" model uncertainties, as per criterion ArM1, appears satisfactory from spin-up time to the end of the series, with values above 1 across the spectra. Eigenvalues get larger in the last two months, likely reflecting both the onset of a seasonal thermocline and upper-ocean processes within the error subspace generated by our stochastic protocol (e.g. in response to wind uncertainties). The loss of empirical consistency of EnsP with respect to OSTIA SST as per criterion ArMCA1 appears as white "pixels". Consistency appears to be almost always verified along the dominant array modes, mostly associated with large-scale patterns (as illustrated in Fig.7). The number of ensemble inconsistency cases increases as one moves to the tails of the spectra, mostly associated with smaller-scale and coastal patterns. Overall, pattern consistency is fairly good with OSTIA SST.

Figure 6b shows the impact of subsampling the OSTIA data, mimicking the impact of using a lower-resolution SST product for ensemble validation. However, although the general pattern is similar, the statistical significance of the consistency analysis degrades significantly; only ~50% of the array modes pass the ArM1 criterion (many more than with the product without subsampling) and appear inconsistent in the ArMCA1 criterion. It is important to note that this degradation is also felt for the dominant array modes, mostly associated with large-scale patterns, except possibly near the end of the time series. Such a (random and systematic) subsampling leads to a significant loss of information that could be used to constrain the high-resolution model ensemble, if EnsP was to be used as an estimate of the model errors. Also, high-resolution SST products appear necessary when it comes to validating eddy-resolving ensembles, such as $1/36°$ here.

Is this degradation dependent upon the data product? The NWS SST observations are of higher resolution compared with the OSTIA SST and therefore are able to "detect" a broader range of model errors, including in the tail modes (not shown). A subsampling rate of one-sixth brings up a resolution similar to the one-fifth subsampling of OSTIA dataset (~25 km). The consistency analysis results were found to be comparable for both networks (Figs. 6b, c).

Let us now turn to data-space patterns of array modes, i.e. the eigenvectors over which our empirical ensemble consistency analysis of EnsP is carried out. Figure 7 shows examples of such patterns for both one-fifth subsampled OSTIA (7a-c) and one-sixth NWS (7d), along with the result of the ArMCA1 criterion for each mode. We focus on May 31, 2012 which is characterized by a slightly higher energy in the spectra compared with neighbouring periods (Figs. 6b-c). Unlike rank histograms depicted in Fig. 3, the OSTIA and NWS array modes are very similar (as illustrated in Figs.7a, d for mode 1), because the subsampled observational schemes are similar, but consistency analysis results do not have to be. Not surprisingly for a regularly-spaced dataset, the first array mode patterns (Fig.7a, d) resembles the structure of the SST spread (Fig. 1b).

For the datasets at hand, the dominant patterns appear to be mostly typical of large-scale and mesoscale processes (Figs.7a, b); seeing zero amplitudes in low-order modes for the English Channel indicates that the array will not be able to correct the SST statistically dominant model errors. The spatial scales appear to decrease as one moves to the tail of the spectrum, and the coastal features become more pronounced (as illustrated in Fig.7c); simultaneously, the ensemble consistency is verified less frequently.

When an ensemble is found to be inconsistent with observations along a particular direction (here, an array mode), a plausible reason is that other error processes must be active in the model in addition to the ones which are at work across the ensemble. This can be verified by checking for model underdispersion. However, as is the case in the examples above,

insufficient resolution of the observations can also lead to statistical inconsistency, impacting mostly small scales and coastal scales.

In Fig. 8a, we show a Hovmöller plot of variations in time of the gridded SLA array mode spectra, including an EnsP consistency check against gridded SLA data. Again, we used 39 array modes. The dataset is relatively coarse at ~25 km resolution and therefore is not subsampled. As before, spectral values smaller than one (light grey in Fig.8a) are associated with poor array performance at "detecting" model uncertainties as per the ArM1 criterion, and white areas depict inconsistent array modes as per ArMCA1 criterion. On average, on the order of 20 array modes sit above the observational noise floor. However, only a fraction of those modes, on the order of 10 on average, show pattern consistency as per ArMCA1 criterion. Overall, consistency of EnsP with respect to the gridded SLA product is not as good as it was with respect to the SST product.

Let us turn to the patterns of eigenvectors in an effort to explain the weak consistency. In Figs. 8b-e, we show examples of consistent and inconsistent patterns of gridded SLA array modes on April 30, 2012. Consistency is verified for the first two array modes shown (ranks 1 and 5), featuring large-scale sea-level gradients over the whole domain including the shallow Celtic Sea and the Eastern Biscay shelf (Fig. 8b-c). Both higher array modes (ranks 10 and 20) shown in Figs. 8d-e are inconsistent – they seem to be dominated by mesoscale and submesoscale sea-level signals (or at least, their 0.25° representation) as well as regional-scale sea-level signals over the shelves and in the English Channel. Small-scale inconsistent higher array modes in the abyssal plain can be explained by the mesoscale low-frequency decorrelation between members, generating phase differences between mesoscale features. From these results we can infer that assimilating the 0.25° gridded product in EnsP would probably lead to unsatisfactory results at the mesoscale and in coastal areas, since error vicinities in the model ensemble and in observations seem at least partly inconsistent with each other in that scale range. Of course, one can introduce representativity errors in the observational error covariance matrix to make things consistent again, but small scales are expected to be inadequately controlled by assimilation.

We now check the consistency of EnsP against the along-track sea-level data. We focus on the same period and on the same tracks as in the rank histograms analysis with the same dataset, starting on February 25, 2012, for 10 consecutive days. The analysis is performed considering together all tracks that would be assimilated in a 10-day assimilation cycle. Therefore, each array mode, characteristic of the 10-day multi-track network, spans both space and time. Figure 9a shows the array mode spectrum over the 10-day period. As per criterion ArM1, almost all eigenvalues are larger than the observational noise threshold (except a few tail modes) and as per criterion ArMCA1 most of them are consistent. This is significantly improved compared to the consistency of EnsP with respect to the gridded sea-level product.

In order to get a bit further into which patterns and underlying processes appear to be (in)consistent, we examine the corresponding eigenvectors. Figures 9b-c show the first two ranks of array modes for the multi-track network; Figs. 9d-e are the same for the higher array modes with ranks 19 and 20 respectively. Both Figs. 9b and 9c show large-scale sea-level gradients, encompassing open-ocean and shelf regions. The patterns in Fig. 9b are generally of a bipolar nature with near zero amplitudes along the continental shelf break for most of the tracks in the region, hinting at shelf/open-ocean exchange processes at work in the model's error subspace. In Figs. 9d-e, we show two examples of higher array modes, with ranks 19 and 20 being consistent and inconsistent, respectively. Higher array modes show a wide range of spatial scales, including large-scale over the shelves and tracks characteristic of mesoscale signals in the open-ocean, with gradients more pronounced than in the gridded product. In most

cases, the higher array modes are consistent and because it is above observational noise (as shown in Fig. 9a) their consistency is meaningful. We therefore have evidence that along-track SLA data could, if assimilated using EnsP for covariances, bring useful consistent information at large scales and at the mesoscale.

Central to this work is the consistency of biogeochemical model ensembles against our PFT product as derived from satellite ocean colour measurements. In Fig. 10, we show Hovmöller plots of variations in time of univariate PFT array mode spectra, including EnsP, EnsB and EnsPB consistency analysis against PFT data, using again 39 array modes. Calculations were performed in log array space (i.e. involving log transformation from data space prior to array mode calculations), without subsampling data. The array mode spectra in Fig.10 exhibit strong variations in time, with several peaks corresponding to differential blooms across ensembles for the various chlorophyll classes during the onset and relaxation of those events. We note however that this is a gridded product with gaps, so part of those variations are also explained by the spatial data coverage variations with time (Fig. 10a; superimposed black line). This does not question our statistical approach, since poor data coverage will logically lead to poorer array performance at detecting ensemble variance; but that dependency must be kept in mind when interpreting the results in terms of processes.

With the ensemble strategy we adopted, one can examine the array performance and the consistency individually for physical (EnsP) and biogeochemical (EnsB) uncertainties, and for both (EnsPB) together, all in the same PFT data space; in that manner, we can see the individual and combined effects of both components (Fig. 10). As already noted in Part 1 article, it appears that physical perturbations (EnsP) have greater impact on biogeochemical model errors than biogeochemical sources and sinks perturbations. The higher EnsP ensemble variance for PFT variables explains also the shorter statistical spin-up period on the order of 1 month, whereas EnsB shows a spin-up period on the order of 3 months, shown as overly inconsistent (i.e. the first three months in Figs. 10a-c vs. 10d-f). It also appears that visually "adding" the nanophytoplankton and diatoms spectra does not come close to describing the total chlorophyll spectra, in particular for EnsP and EnsPB (Figs. 10a-c and 10g-i), whilst being partially verified for EnsB (Figs. 10d-f). The spectra derived from diatoms in EnsP and EnsPB appears not to contribute to the total chlorophyll spectra, the latter having eigenvalues similar to those of nanophytoplankton.

We see three possible classes of explanations to this result: (1) it is likely that the constant observational error of $0.3\ mg/m^3$ over the whole domain and for all classes, is unrealistic, particularly for diatoms exhibiting higher local uncertainties in the shelves and coastal regions; (2) it is possible that our decomposition of total chlorophyll into two size classes (or four binned into two as discussed in Section 3.3) is not entirely relevant process-wise (i.e. combining one PFT with characteristic silicate limitation with one without), effectively leading to a statistical ill-posed issue; (3) total chlorophyll as resulting from a variety of processes (especially when physical perturbations are applied) will have higher statistical complexity (higher number of "degrees of freedom") than either nanophytoplankton or diatoms taken independently, leading to "purple" spectra with more tail modes not represented on Figs. 10a and 10g, while Figs. 10c and 10i show "redder" diatoms spectra.

Figure 11 shows examples of low and high gridded PFT array modes and consistency status of EnsP and EnsB vs. the PFT dataset (EnsPB eigenvectors are similar to EnsP; not shown). Let us focus on March 28, 2012, where for this specific date the PFT are gap-free over the Bay of Biscay facilitating our analysis at all scales. As for rank histogram analyses, we choose to investigate the period during early spring, because consistency can be easily explained by the onset of the chlorophyll abundance primary bloom (Fig. 2). In effect, for a positive oriented

variable such as chlorophyll, bloom-related model spread increases when abundance increases during a bloom event, therefore chances for pattern consistency are higher. We note that this is a very different result compared to rank histograms which show degraded consistency during blooms. This is because rank histograms are sensitive to model-data biases amplified during blooms (in contrast to array modes free from biases). The latter is also supported by the fact that the biogeochemical model ensembles are under-dispersive, due to model underestimation of chlorophyll abundance.

The PFT first array modes using EnsP and EnsB show consistent patterns and are compatible in structure with the model's second-order statistical moments respectively (Figs. 11a, f vs. Figs. 2b, c). Another remark is that the first array modes using EnsP and EnsB are different (Figs. 11a-c vs. Figs. 11d-f). When physics is perturbed (i.e. EnsP) the total chlorophyll eigenvectors appear as large-scale open ocean patterns controlled by nanophytoplankton primary production (Figs. 11a-b). When biogeochemical source and sink terms are perturbed (i.e. EnsB) we observe mesoscale patterns for all classes, with nanophytoplankton again defining the eigenvectors of total chlorophyll (Figs. 11d-e). For both ensembles, diatom measurements appear to detect model errors mostly on the shelves, but their contribution for the first array mode for total chlorophyll is limited (Figs. 11c, f).

For the higher modes we observe several combinations of consistent and inconsistent modes across phytoplankton types and total chlorophyll. In general, when both phytoplankton types are found to be consistent (inconsistent), then total chlorophyll is also likely to be found consistent (inconsistent). Mixed consistency results between chlorophyll classes are also observed (less frequently), but the non-mixed examples are the most informative to investigate (Figs. 11g-l). When physics is perturbed the PFT array mode 25 is able to detect model errors for both chlorophyll classes at small-scale (Figs. 11g-i). Surprisingly, for the higher array modes using EnsP the diatoms appear to contribute to the spatial variability of total chlorophyll at small-scale, especially in the shelves (Figs. 11g-i). By contrast, when biogeochemical source and sink terms are perturbed the model-based PFT array mode 25 appears not to be consistent with PFT data at the mesoscale (Figs. 11j-l). The fact that EnsB generates chiefly eigenvectors with mesoscale patterns, implies that having identical ocean physics across all members limits the potential of the PFT measurements to detect biogeochemical model errors at small-scale (as shown for example for EnsP in Figs. 11g-i).

6 Discussion and conclusions

This study is Part 2 of a two-part series following a companion article Part 1 aimed at generating ocean model ensembles. Part 2 article focuses on: (1) sophisticated ensemble model-data comparison methods, one of them published here for the first time based on a new criterion in the space of array modes, and (2) applying those methods for the first time with phytoplankton functional type data derived from ocean colour.

The empirical consistency analysis focused on satellite observations (SST, SLA and ocean colour), in concert with model ensembles of ocean physics and biogeochemistry. We used a high-resolution configuration for the Bay of Biscay, as a means to investigate the model error subspace generated by our stochastic protocol in a companion article Part 1, both in the open and coastal ocean. In order to better understand the couplings between physics and biogeochemistry we examined three model ensemble experiments: perturbing only physics, perturbing only biogeochemical source and sink terms, and perturbing both simultaneously. Below, we synthesize results from the two consistency methods in an attempt to assess the reliability of model ensembles with respect to observations.

Rank histograms for physical properties showed a large dependency based on geographical region and on season for both SST and SLA networks. In most cases, rank histograms showed large biases between model and data, and were limited because of model ensemble underspread and because of weak variability in the observations. In addition, there were high-frequency errors in the observations that were not present in the model ensemble. Rank histograms for the ocean colour PFT showed persistent model underestimation and underspread for chlorophyll abundance, with some improvement for nanophytoplankton after a spin-up period on the order of 3 months during the winter low primary productivity. Rank histograms have been used successfully as a reliability tool identifying on several occasions consistent model-data configurations and attributing this result to physical and biogeochemical processes, such as the spring shoaling of the thermocline, the frontal activity in the shelves explained by river plume migration, the tidal mixing in the English Channel (*Karagiorgos et al*., 2020), and to a lesser extent the chlorophyll abundance during spring, mainly on the onset and relaxation of bloom events.

Array modes were performed using innovation samples in array space and therefore, model-data distributions were free from biases hindering empirical consistency of error patterns (as in some cases for the rank histograms). A typical result using array modes in the context of stochastic modelling is that insufficient resolution of the verifying observations can lead to statistical inconsistency, impacting mostly small and coastal scales. High-resolution datasets (and model ensembles) can improve array modes consistency at small-scale, as long as the following assumptions are valid: (a) data errors are uncorrelated, (b) data errors are small in comparison to model-data misfit and (c) the model ensembles are not under-dispersive. When ensembles are found to be inconsistent with observations along an array mode, a plausible reason is that other error processes must be active in the model, in addition to the ones at work across the ensemble. For instance, model errors in physical processes currently unaccounted for are attributable to residual tidal errors (due to local tidal fronts and occasional Kelvin waves propagating along the coasts), to the non-isostatic response of atmospheric pressure, and to high-frequency errors in open boundaries over the shelves (*Vervatis et al.*, 2021). In addition, complex processes in the biogeochemical model are simplified by making use of only a few model parameters. Biogeochemical parameterizations controlling the growth rate and grazing of phytoplankton classes in the model are based on approximations and therefore, stochastic modelling of these parameters may improve model performance (*Garnier et al.*, 2015).

Array modes for both SST networks (i.e. OSTIA and NWS) showed that despite their differences in production and resolution, if subsampled, can provide comparable information to detect model errors, from shortly after the ensemble spin-up to the end of the series (in contrast to rank histograms). Their consistency was verified for the large-scale and open ocean, but also for the small-scale and coastal SST patterns. Array modes pattern consistency between the model ensemble perturbing physics and the sea-level datasets was in general verified at large-scale. The gridded SLA product showed unsatisfactory results at the mesoscale and in coastal areas, whereas the along-track sea-level showed that, if assimilated, can bring potentially useful information at the mesoscale.

The most important findings for the array modes using PFT against ocean biogeochemical ensembles can be summarised as follows. A large ensemble variance can lead to an improved PFT array performance, considering biogeochemical model uncertainties stemming mainly from stochastic physics. Consistency results can be further enhanced if we consider a two-way coupling and a feedback from the bio-optical model (due to changes in chlorophyll, leading to increased model spread) to the physical model for the solar radiation penetration in the water column. We performed a few simulations (not shown) but we decided not to activate the two-way coupling, because the onset and relaxation of the spring bloom may occur earlier and more

rapidly in the model compared with observations in the region (*Gutknecht et al.*, 2019). Additive spectra of nanophytoplankton and diatoms cannot describe total chlorophyll spectra. One plausible explanation is that EnsP and EnsB state vectors (i.e. their anomalies from the ensemble mean) are not statistically independent (perhaps as an artefact of the limited ensemble size). This is proven also by the fact that ensemble variance for EnsPB sometimes can be unexpectedly small, as if physics and biogeochemistry processes compensated, cf. companion article Part 1. Consequently, there is no reason for array mode eigenspectra to be additive on a rank-by-rank basis nor hierarchised in the same way (in contrast to other spectra, e.g. Fourier and Wavelet Transform). Another point is that there is no reason our ensembles should be orthogonal, because they are not built this way. The array mode spectra give an idea of the number of "degrees of freedom" of the ensembles in data space, and of the very high temporal variability of that number. For instance, the bloom periods are characterized by larger numbers of "degrees of freedom", apparently peaking at 39 (40 minus mean). Other plausible explanations may be the unrealistic observational error set globally at $0.3\ mg/m^3$ for all PFT and also, the total chlorophyll decomposition of four PFT binned into two not being entirely relevant process-wise.

In general, the data PFT errors are not negligible and are related to the size-class (*Brewin et al.*, 2017; *Laiolo et al.*, 2021). Errors are lower for the small size classes and higher for the large size classes (with dinoflagellates having the higher errors of all PFT), and increased with increasing chlorophyll. This is owing to the nature of how the satellite PFT error is computed. Chlorophyll *a* uncertainties (e.g. root mean square error and bias) for each PFT are assigned based on optical water type membership at a given satellite pixel (*Brewin et al.*, 2017). It is expected that observational errors are cross-correlated, both across functional types and spatially for each PFT. In fact, each PFT is computed as a function of the same total chlorophyll product, and the same procedure is adopted to compute the errors of the different PFT (*Brewin et al.*, 2017). The total chlorophyll product itself is expected to have spatially correlated errors. This is due to both atmospheric effects (e.g. cloud cover) and in-water sources of errors, in particular in the coastal zone, due to the coloured dissolved organic matter (CDOM) inputs that degrade ocean colour near the coast (e.g. terrestrial CDOM in river plumes). This is also confirmed by the PFT spatial error gradients and the gradients of the concentrations themselves, being stronger for diatoms than nanophytoplankton in the coastal zone (*Brewin et al.*, 2017; *Ciavatta et al.*, 2018). Cross-correlations are expected to be lower in the open ocean and deeper areas. The representation of such uncertainty and correlated errors has been so far neglected in ocean colour data assimilation (*IOCCG*, 2020). In our application, if we used such PFT uncertainties and data correlations, the array modes consistency would have been degraded, pertaining to the fact that ocean colour datasets are provided in high-resolution (and therefore, spatial data correlations may not be negligible).

Pattern-selective consistency analysis showed that low-rank eigenvectors appear as large-scale and mesoscale, mainly controlled by nanophytoplankton in the open ocean; diatom measurements appear to detect model errors mostly on the shelves. For the higher modes, biogeochemical model errors appear to be detected at small-scale only when physics is perturbed. By contrast, when only biogeochemical source and sink terms are perturbed, the model-based high-rank modes appear not to be consistent with PFT data at the mesoscale, limiting also the potential to detect model errors at small-scale due to identical physical processes across all members.

We recommend using methods adapting and estimating the $\boldsymbol{R}$ and $\boldsymbol{Q}$ *observational* and *model* error covariance matrices, respectively. The ArM methodology can provide pattern- and scale-dependent ensemble consistency checks, facilitating the qualification of ensembles to provide useful error covariance estimates in regional systems for ulterior coastal downscaling.

The next step of this study will be dedicated to the investigation of ensemble consistency based on array modes: (1) performing multivariate analysis between physics and biogeochemistry (e.g. SST vs. total chlorophyll), and between phytoplankton size classes, using information for the spatial distribution of observational errors; and (2) taking into account correlated observational errors. To that end, an additional criterion will be introduced to check the diagonality of the innovation covariance matrix in the space of array modes.

Acknowledgments. This work was carried out as part of the Copernicus Marine Environment Monitoring Service (CMEMS) "Stochastic Coastal/Regional Uncertainty Modelling (SCRUM)" Service Evolution project. CMEMS is implemented by Mercator Ocean International in the framework of a delegation agreement with the European Union. Part of this research was also made possible through the IKY Scholarships Programme and co-financed by the European Union (European Social Fund-ESF) and Greek national funds through the action entitled "Reinforcement of Postdoctoral Researchers", in the framework of the Operational Programme "Human Resources Development Programme, Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) 2014-2020. The contribution of P. De Mey-Frémaux and N. Ayoub is supported by Centre National de la Recherche Scientifique (CNRS). We acknowledge the use of the ECMWF's computing and archive facilities in this research. This work was also supported by computational time granted from the Greek Research & Technology Network (GRNET) in the National HPC facility – ARIS – under project ID PA002007. We thank five anonymous reviewers for their constructive comments.

References

Anderson, J., 1996. A method for producing and evaluating probabilistic forecasts from Ensemble model integrations, J. Climate, 9, 1518–1530.

Andersson, E., 2003. Modelling of innovation statistics. Proceedings of the Workshop on recent developments in data assimilation for atmosphere and oceans. ECMWF, Reading, UK, 153-164.

Anderson, J.L., 2009. Spatially and temporally varying adaptive covariance inflation for ensemble filters. Tellus A, 61: 72-83. doi:10.1111/j.1600-0870.2008.00361.x.

Auclair, F., P. Marsaleix, and P. de Mey-Frémaux, 2003. Space-time structure and dynamics of the forecast error in a coastal circulation model of the Gulf of Lions. Dynamics of Atmospheres and Oceans, Elsevier, 36 (4), pp.309-346. 10.1016/S0377-0265(02)00068-4.

Aumont, O., Ethé, C., Tagliabue, A., Bopp, L., and Gehlen, M., 2015. PISCES-v2: an ocean biogeochemical model for carbon and ecosystem studies. Geosci. Model Dev. 8, 2465-2513.

Brewin, R. J. W., Ciavatta S., Sathyendranath S., Jackson T., Tilstone G., Curran K., Airs R. L., Cummings D., Brotas V., Organelli E., Dall' Olmo G., and Raitsos D. E., 2017. Uncertainty in Ocean-Colour Estimates of Chlorophyll for Phytoplankton Groups. Front. Mar. Sci. 4:104. doi: 10.3389/fmars.2017.00104.

Brewin, R. J. W., S. Sathyendranath, T. Hirata, S. J. Lavender, R. Barciela, and N. J. Hardman-Mountford, 2010. A three-component model of phytoplankton size class for the Atlantic Ocean, Ecol. Modell., 221, 1472–1483, doi: 10.1016/j.ecolmodel.2010.02.014.

Brewin, R. J. W., Sathyendranath, S., Jackson, T., Barlow, R., Brotas, V., Airs, R., et al., 2015. Influence of light in the mixed layer on the parameters of a three-component model of phytoplankton size structure. Remote Sens. Environ. 168, 437–450. doi: 10.1016/j.rse.2015.07.004.

Candille, G., and Talagrand, O., 2005. Evaluation of probabilistic prediction systems for a scalar variable. Q. J. R. Meteorol. Soc., 131, 2131-2150.

Candille, G., Brankart, J.-M., and Brasseur, P., 2015. Assessment of an ensemble system that assimilates Jason-1/Envisat altimeter data in a probabilistic model of the North Atlantic ocean circulation. Ocean Science 11, 425–438.

Charria, G., Lamouroux, J., and De Mey, P., 2016. Optimizing observational networks combining gliders, moored buoys and FerryBox in the Bay of Biscay and English Channel. Journal of Marine Systems, 162, 112-125, http://dx.doi.org/10.1016/j.jmarsys. 2016.04.003.

Chelton, D. B., Schlax, M. G., Samelson, R. M., 2011. Global observations of nonlinear mesoscale eddies, Prog. Oceanogr., 91, 167–216, doi:10.1016/j.pocean.2011.01.002.

Ciavatta, S., Brewin, R. J. W., Skákala, J., Polimene, L., de Mora, L., Artioli, Y., and Allen, J. I., 2018. Assimilation of Ocean-Colour Plankton Functional Types to Improve Marine Ecosystem Simulations. Journal of Geophysical Research: Oceans, 123(2), 834–854, https://doi.org/10.1002/2017JC013490.

Ciavatta, S., Kay, S., Brewin, R. J. W., Cox, R., Di Cicco, A., Nencioli, F., and Tsapakis, M., 2019. Ecoregions in the Mediterranean Sea Through the Reanalysis of Phytoplankton Functional Types and Carbon Fluxes. Journal of Geophysical Research: Oceans, 124(10), 6737–6759, https://doi.org/10.1029/2019JC015128.

Desroziers, G., Berre, L., Chapnik, B., Poli, P., 2005. Diagnosis of observation, background and analysis-error statistics in observation space. Q. J. R. Meteorol. Soc. 131, 3385–3396. https://doi.org/10.1256/qj.05.108.

Donlon, C.J., Martin, M., Stark, J., Roberts-Jones, J., Fiedler, E., and Wimmer, W., 2012. The Operational Sea Surface Temperature and Sea Ice Analysis (OSTIA) system. Remote Sensing of the Environment, doi: 10.1016/j.rse.2010.10.017 2011.

Evensen, G., 2003. The Ensemble Kalman Filter: theoretical formulation and practical implementation. Ocean Dyn. 53, 343-367.

Ford, D., 2019. Assessing the role and consistency of satellite observation products in global physical-biogeochemical ocean reanalysis. Ocean Sci. Discuss. 1–25. https://doi.org/10.5194/os-2019-118.

Garnier, F., Brankart, J.M., Brasseur, P., and Cosme, E., 2016. Stochastic parameterizations of biogeochemical uncertainties in a 1/4° NEMO/PISCES model for probabilistic comparisons with ocean color data. Journal of Marine Systems, 155, 59–72.

Ghantous, M., Ayoub, N., De Mey-Frémaux, P., Vervatis, V., and Marsaleix, P., 2020. Ensemble downscaling of a regional ocean model. Ocean Modell., 145, https://doi.org/10.1016/j.ocemod.2019.101511.

Gharamti, M. E., J. Tjiputra, I. Bethke, A. Samuelsen, I. Skjelvan, M. Bentsen, and L. Bertino, 2017. Ensemble data assimilation for ocean biogeochemical state and parameter estimation at different sites, Ocean Modelling, 112, 65-89, https://doi.org/10.1016/j.ocemod.2017.02.006.

Gutknecht, E., Reffray, G., Mignot, A., Dabrowski, T., and Sotillo, M. G., 2019. Modelling the marine ecosystem of Iberia–Biscay–Ireland (IBI) European waters for CMEMS operational applications, Ocean Sci., 15, 1489–1516, https://doi.org/10.5194/os-15-1489-2019.

Hamill, T.M., 2001. Interpretation of Rank Histograms for Verifying Ensemble Forecasts. Mon. Wea. Rev., 129, 550-560, https://doi.org/10.1175/1520-0493(2001)129<0550: IORHFV>2.0.CO;2.

Ide, K., Courtier P., Ghil M., and Lorenc A., 1997. Unified notations for data assimilation: operational, sequential and variational. J. Met. Soc. Japan, 75(1B):181-189.

IOCCG, 2020. Synergy between Ocean Colour and Biogeochemical/Ecosystem Models. Dutkiewicz, S. (ed.), IOCCG Report Series, No. 19, International Ocean Colour Coordinating Group, Dartmouth, Canada.

Janjić, T., Bormann N., Bocquet M., Carton J.A., Cohn S.E., Dance S.L., Losa S.N., Nichols N.K., Potthast R., Waller J.A., Weston P, 2018. On the representation error in data assimilation, Q J R Meteorol Soc.,144,1257–1278.

Karagiorgos, J., V. Vervatis, and S. Sofianos, 2020. The Impact of Tides on the Bay of Biscay Dynamics. J. Mar. Sci. Eng., 8, 617.

Kourafalou V.H., P. De Mey, M. Le Hénaff, G. Charria, C.A. Edwards, R. He, M. Herzfeld, A. Pasqual, E.V. Stanev, J. Tintoré, N. Usui, A.J. Van Der Westhuysen, J. Wilkin and Zhu, X., 2015a. Coastal Ocean Forecasting: system integration and validation. Journal of Operational Oceanography, http://dx.doi.org/10.1080/1755876X.2015.1022336.

Kourafalou, V.H., P. De Mey, J. Staneva, N. Ayoub, A. Barth, Y. Chao, M. Cirano, J. Fiechter, M. Herzfeld, A. Kurapov, A.M. Moore, P. Oddo, J. Pullen, A.J. van der Westhuysen and Weisberg, R.H., 2015b. Coastal Ocean Forecasting: science foundation and user benefits. Journal of Operational Oceanography, http://dx.doi.org/10.1080/ 1755876X.2015.1022348.

Lamouroux, J., G. Charria, P. De Mey, S. Raynaud, C. Heyraud, P. Craneguy, F. Dumas and Le Hénaff, M., 2016. Objective assessment of the contribution of the RECOPESCA network to the monitoring of 3D coastal ocean variables in the Bay of Biscay and the English Channel. Ocean Dynamics, 66(4), 567-588, http://dx.doi.org/10.1007/s10236-016-0938-y.

Le Hénaff, M., P. De Mey and Marsaleix, P., 2009. Assessment of observational networks with the Representer Matrix Spectra method – Application to a 3-D coastal model of the Bay of Biscay. Ocean Dynamics, 59, 3-20, DOI 10.1007/s10236-008-0144-7.

Leonardo, L., R. Matear, M. Soja-Woźniak, D. J. Suggett, D. J. Hughes, M. E. Baird, M. A. Doblin, 2021. Modelling the impact of phytoplankton cell size and abundance on inherent optical properties (IOPs) and a remotely sensed chlorophyll-a product, Journal of Marine Systems, 213, 103460, https://doi.org/10.1016/j.jmarsys.2020.103460.

Madec, G., 2012. Nemo ocean engine, Tech. rep., NEMO team.

Mattern, J.P., Edwards, C.A., and Moore, A.M., 2018. Improving Variational Data Assimilation through Background and Observation Error Adjustments. Mon. Wea. Rev., 146, 485–501, https://doi.org/10.1175/MWR-D-17-0263.1.

Oke, P. R., and P. Sakov, 2008. Representation Error of Oceanic Observations for Data Assimilation. J. Atmos. Oceanic Technol., 25, 1004–1017, https://doi.org/10.1175/2007JTECHO558.1.

Pujol, M.-I., Faugère, Y., Taburet, G., Dupuy, S., Pelloquin, C., Ablain, M., and Picot, N., 2016. DUACS DT2014: the new multi-mission altimeter data set reprocessed over 20 years, Ocean Sci., 12, 1067-1090, doi:10.5194/os-12-1067-2016.

Quattrocchi, G., P. De Mey, N. Ayoub, V. Vervatis, C.-E. Testut, G. Reffray, J. Chanut and Y. Drillet, 2014. Characterisation of errors of a regional model of the Bay of Biscay in response

to wind uncertainties: a first step toward a data assimilation system suitable for coastal sea domains. Journal of Operational Oceanography, Volume 7, Number 2, August 2014, pp. 25-34(10).

Sakov, P., G. Evensen and L. Bertino, 2010. Asynchronous data assimilation with the EnKF. Tellus A. 62(1): 24-29, http://dx.doi.org/10.1111/j.1600-0870.2009.00417.x.

Sieburth, John McN. Smetacek, Victor Lenz, Jürgen, 1978. Pelagic ecosystem structure: Heterotrophic compartments of the plankton and their relationship to plankton size fractions, Limnology and Oceanography, 23, doi: 10.4319/lo.1978.23.6.1256.

Simon, E., and Bertino, L., 2009. Application of the Gaussian anamorphosis to assimilation in a 3-D coupled physical-ecosystem model of the North Atlantic with the EnKF: a twin experiment. Ocean Sci. 5, 495-510.

Skakala, J., Ford, D., Brewin, R. J. W., Mc Ewan, R., Kay, S., Taylor, B., de Mora, L., and Ciavatta, S., 2018. The assimilation of phytoplankton functional types for operational forecasting in the North-West European Shelf. Journal of Geophysical research - Oceans, 123 (8), 5230-5247, doi: 10.1029/2018JC014153.

Song, H., Edwards, C.A., Moore, A.M., Fiechter, J., 2016. Data assimilation in a coupled physical-biogeochemical model of the California current system using an incremental lognormal 4-dimensional variational approach: Part 3—Assimilation in a realistic context using satellite and in situ observations. Ocean Model. 106, 159–172. https://doi.org/10.1016/j.ocemod.2016.06.005.

Vervatis, V. D., C.E. Testut, P. De Mey, N. Ayoub, J. Chanut, and G. Quattrocchi, 2016. Data assimilative twin-experiment in a high-resolution Bay of Biscay configuration: 4D EnOI based on stochastic modelling of the wind forcing. Ocean Modelling, 100, 1-19, http://dx.doi.org/10.1016/j.ocemod.2016.01.003.

Vervatis, V. D., De Mey-Frémaux, P., Ayoub, N., Karagiorgos, J., Ghantous, M., Kailas, M., Testut, C.-E., and Sofianos, S., 2021. Assessment of a regional physical-biogeochemical stochastic ocean model. Part 1: ensemble generation, Ocean modell., under review.

Journal Pre-proof

Table 1. Stochastic model ensembles (cf. companion article Part 1 by *Vervatis et al*., 2021).

Ensemble	SPPT/SPP-AR(1)		
EnsP	Physics	Atm. forcing	U_{air}, T_{air}, SLP
		Parameters	c_d, c_e, c_h, c_b
EnsB	Biogeochemistry	sources-minus-sinks of 24 prognostic variables $SMS(C)$	
EnsPB	Physics & Biogeochemistry	EnsP & EnsB	

abbreviations: SPPT - stochastic perturbed parameterized tendencies; SPP - stochastic perturbed parameters; AR(1) - first-order autoregressive processes; U_{air} - wind velocities; T_{air} - air temperature; SLP - sea level pressure; c_d, c_e, c_h - wind drag and turbulent coefficients; c_b - bottom drag; $SMS(C)$ - sources minus sinks of biogeochemical tracers C.

Table 2. Observational networks.

CMEMS Product Identifiers (http://marine.copernicus.eu/)*		Error
OSTIA SST L4 gridded 0.05^o	SST_GLO_SST_L4_NRT_OBS_010_001	$0.5\ ^oC$
NWS SST L4 gridded 0.04^o	SST_ATL_SST_L4_REP_OBS_010_026	
SLA L3 along-track 14 km	SEALEVEL_GLO_PHY_L3_REP_OBS_008_062	$0.05\ m$
SLA L4 gridded 0.25^o	SEALEVEL_GLO_PHY_L4_REP_OBS_008_047	
Chlorophyll *a* gridded 0.05^o	Phytoplankton Functional Types - PFT (*Brewin et al*., 2017)	$0.3\ mg/m^3$

*All datasets are provided at daily frequency.

Journal Pre-proof

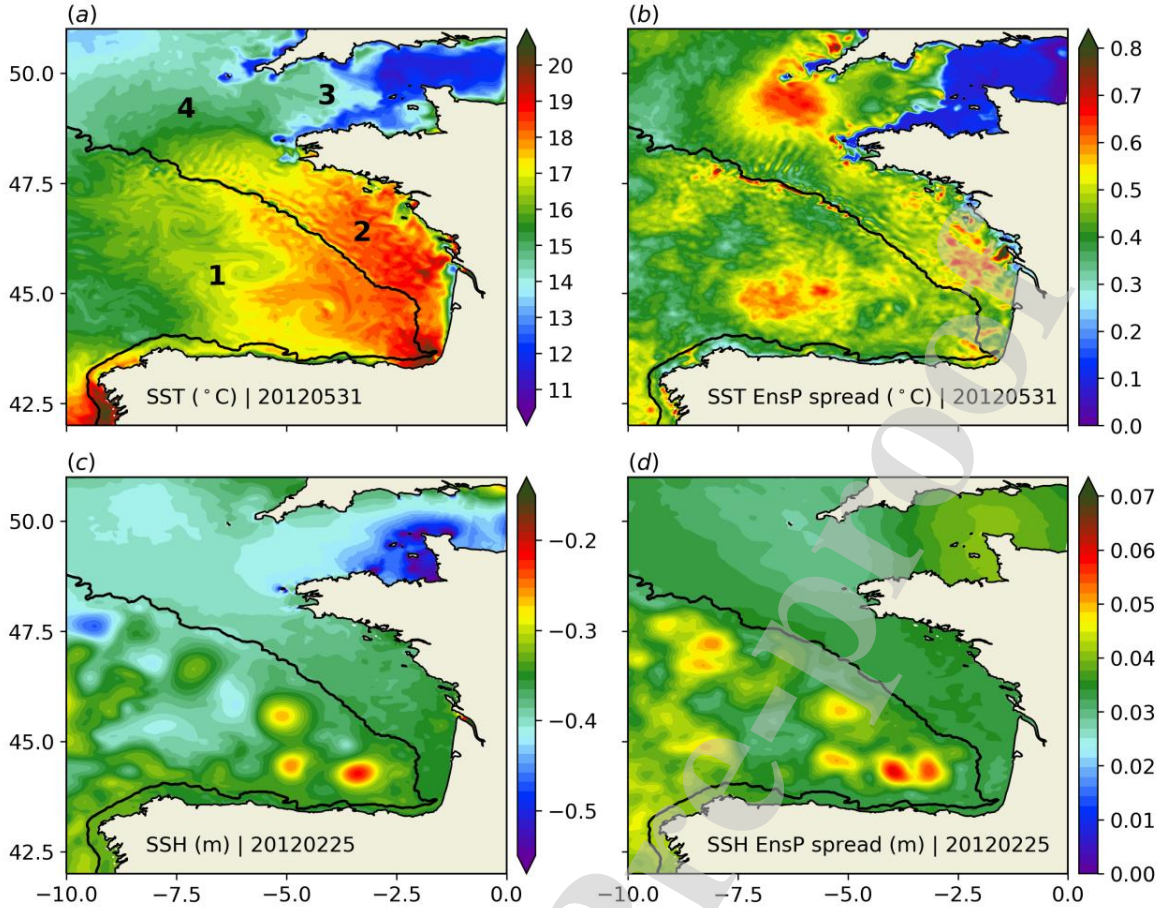


Figure 1 (a-b) Deterministic run SST and model ensemble EnsP spread ($°C$) on May 31, 2012. (c-d) Same for SSH (m) on February 25, 2012. Fig. 1a depicts the characteristic areas in the Bay of Biscay discussed in the text: 1-abyssal plain, 2-Armorican shelf, 3-English Channel, 4-Celtic shelf. Black line denotes the 200 m isobaths.

Journal Pre-proof

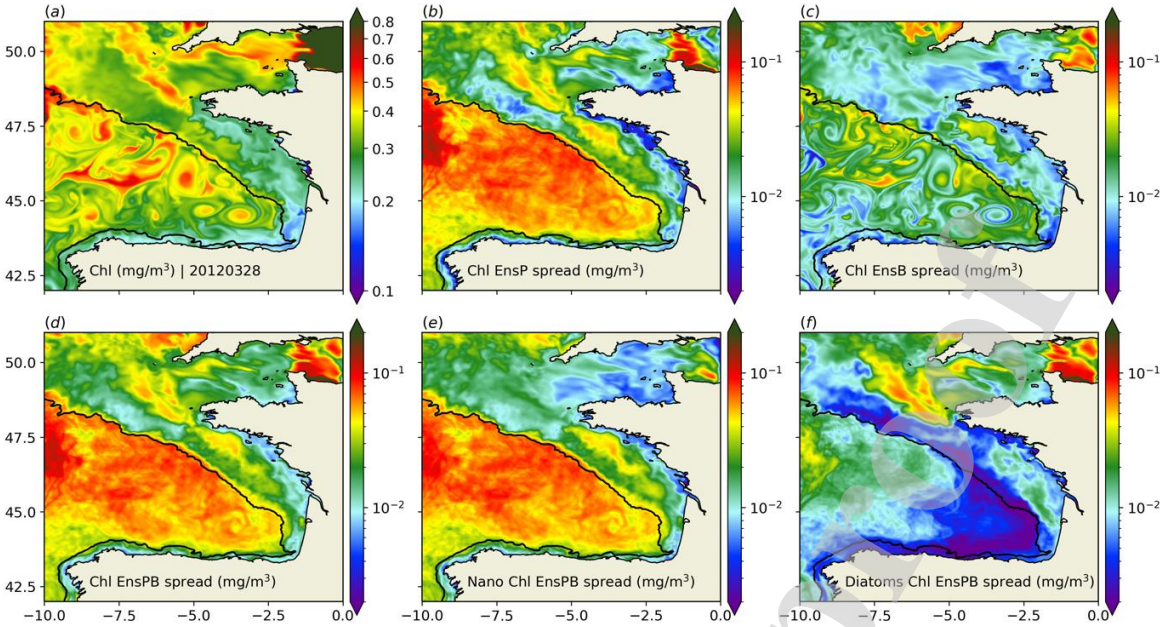


Figure 2 (a-d) Total chlorophyll deterministic run and model ensembles spreads EnsP, EnsB, EnsPB (mg/m^3) on March 28, 2012. (e-f) Same with (d) model ensemble spread EnsPB for nanophytoplankton and diatoms chlorophyll (mg/m^3) respectively.

Journal Pre-proof

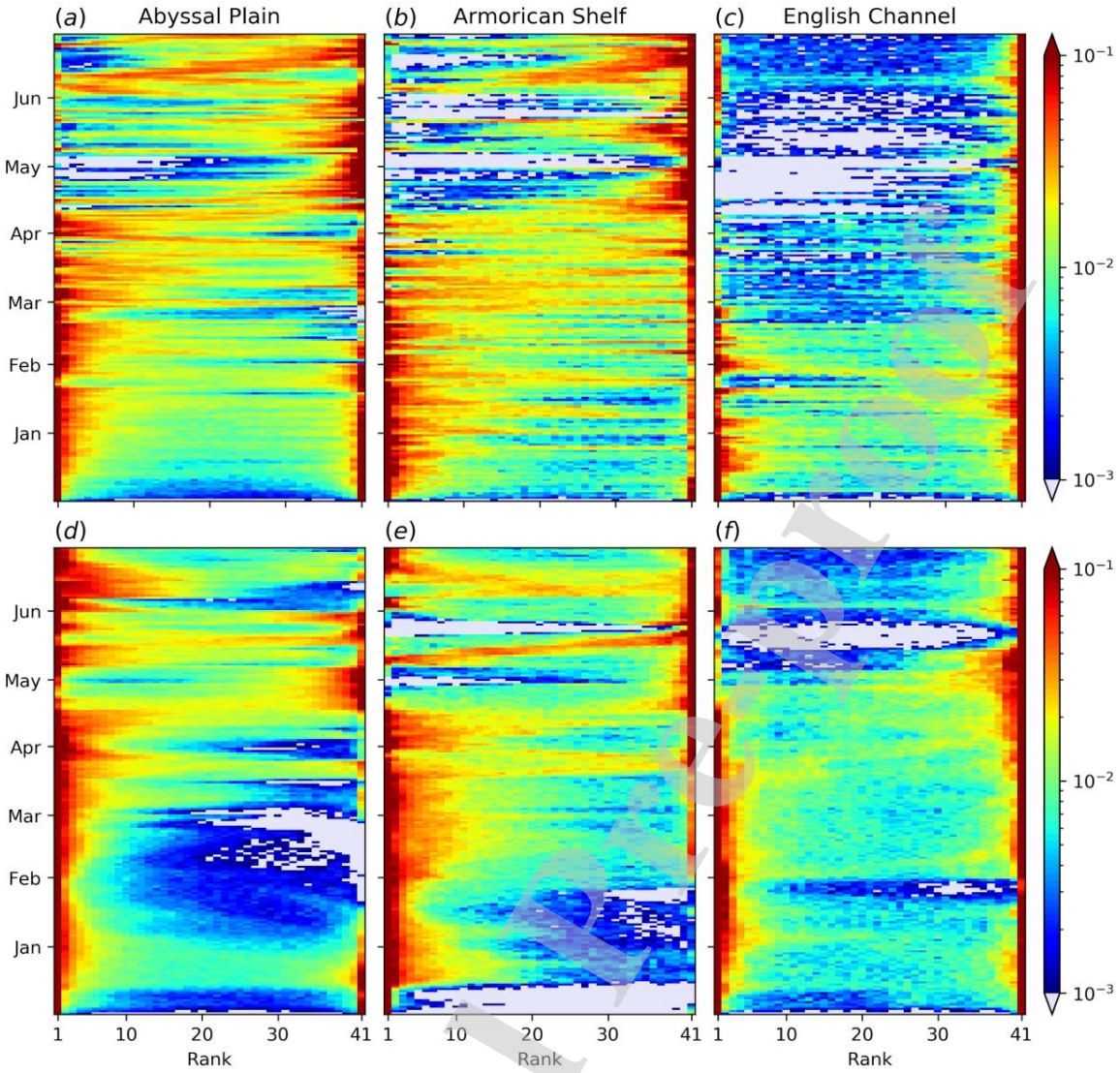


Figure 3 Hovmöller plot of rank histograms (x-axis: rank, y-axis: time, colorbar: pdf) between ensemble EnsP and SST L4 datasets: (a-c) OSTIA, (d-f) NWS. Regional consistency is checked for the (left to right): abyssal plain, Armorican shelf and English Channel. Rank histogram flatness is verified for pdf values at $\frac{1}{41}$ ~0.024 (green-yellow in colorbar).

Journal Pre-proof

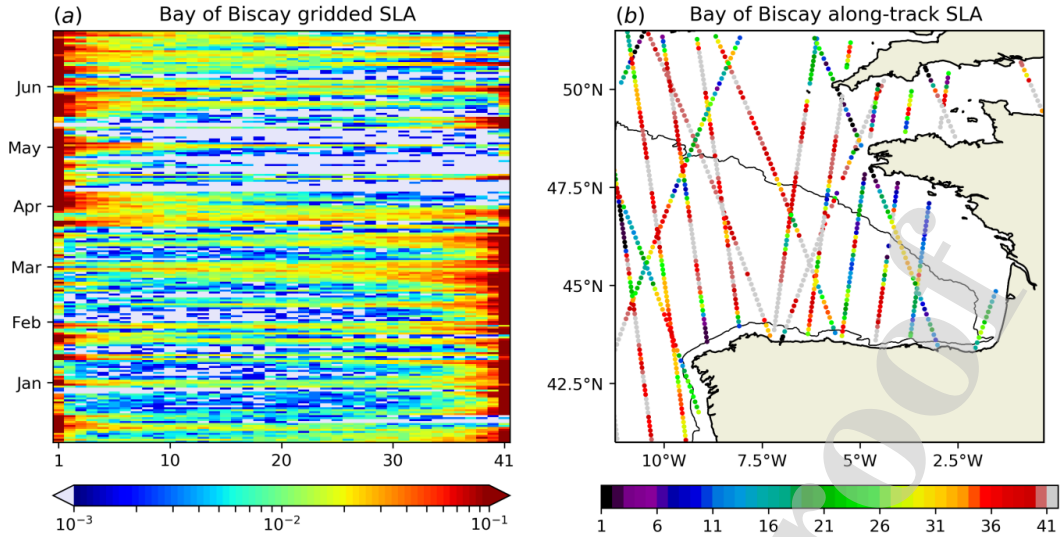


Figure 4 (a) Hovmöller plot of rank histograms (x-axis: rank, y-axis: time, colorbar: pdf) between ensemble EnsP and SLA L4 gridded dataset for the Bay of Biscay, and (b) map of ranks for along-track SLA L3 dataset with respect to the ensemble EnsP (colorbar: rank), starting on February 25, 2012, and for 10 consecutive days, including ascending and descending tracks in the domain.

Journal Pre-proof

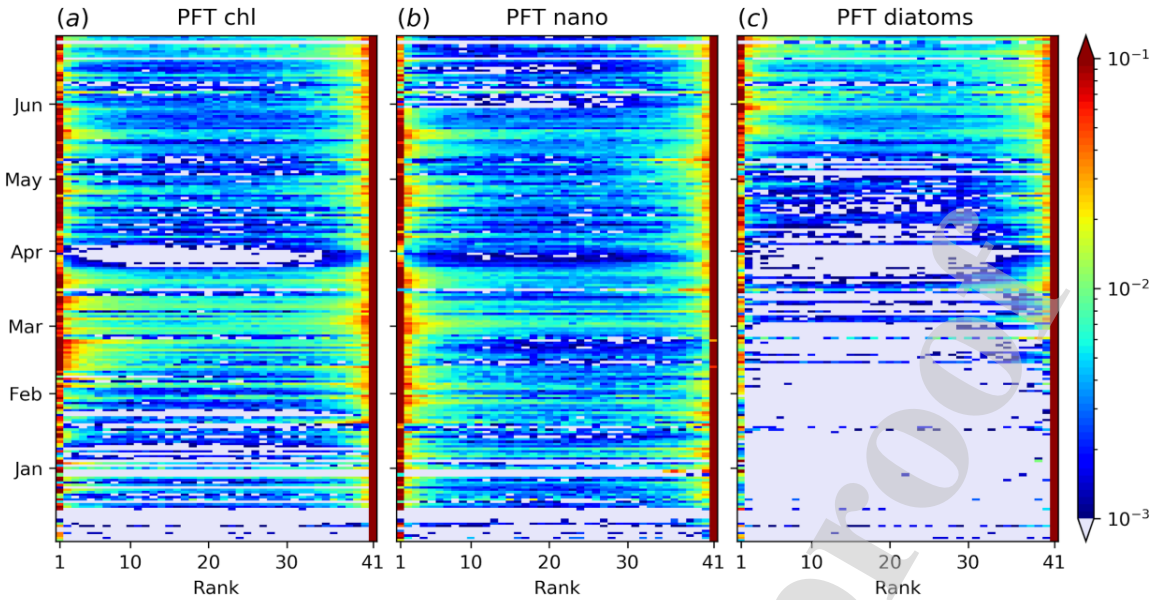


Figure 5 Hovmöller plot of rank histograms (x-axis: rank, y-axis: time, colorbar: pdf) between ensemble EnsPB and PFT dataset: (a) total chlorophyll, (b) nanophytoplankton and (c) diatoms chlorophyll.

Journal Pre-proof

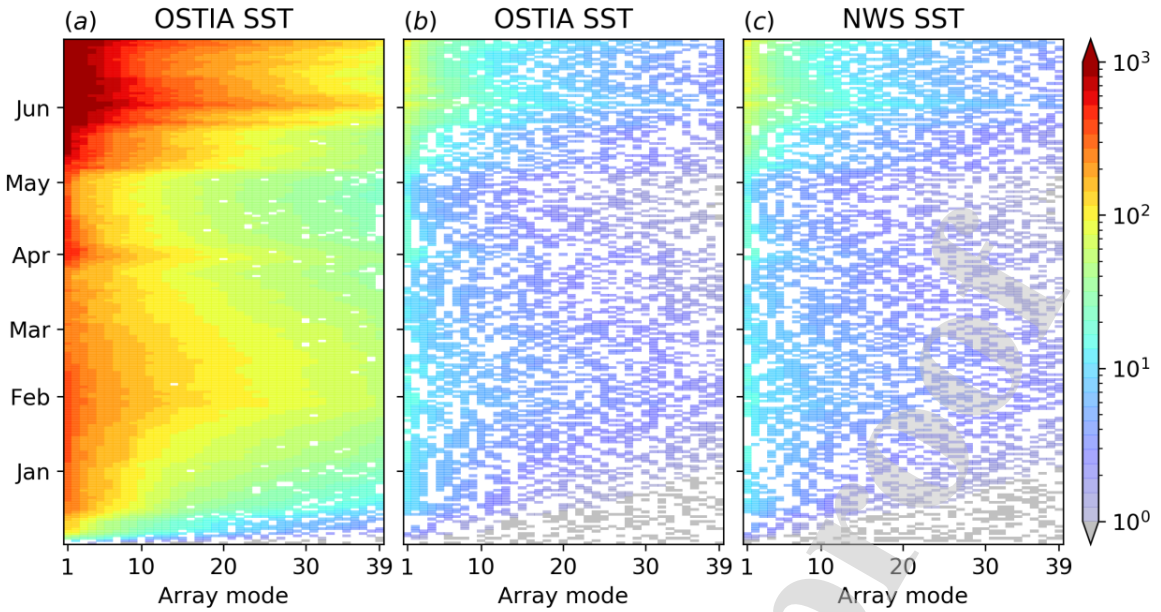


Figure 6 Hovmöller plot of variations in time of SST array mode spectra vs. modal rank (i.e. the rank of the array mode), including an EnsP consistency check against (a-b) OSTIA and (c) NWS SST data. Colorbar: array mode spectra as of criterion ArM1; eigenvalues smaller than one (the observational noise floor in array space; grey pixels) denote error modes marginally detectable by the array. White pixels depict inconsistent array modes as per criterion ArMCA1 (the higher modes are also mostly inconsistent as per criterion ArM1). (a) No data subsampling, (b) one-fifth subsampling rate (i.e. one data point every fifth OSTIA point retained), and (c) one-sixth subsampling rate.

Journal Pre-proof

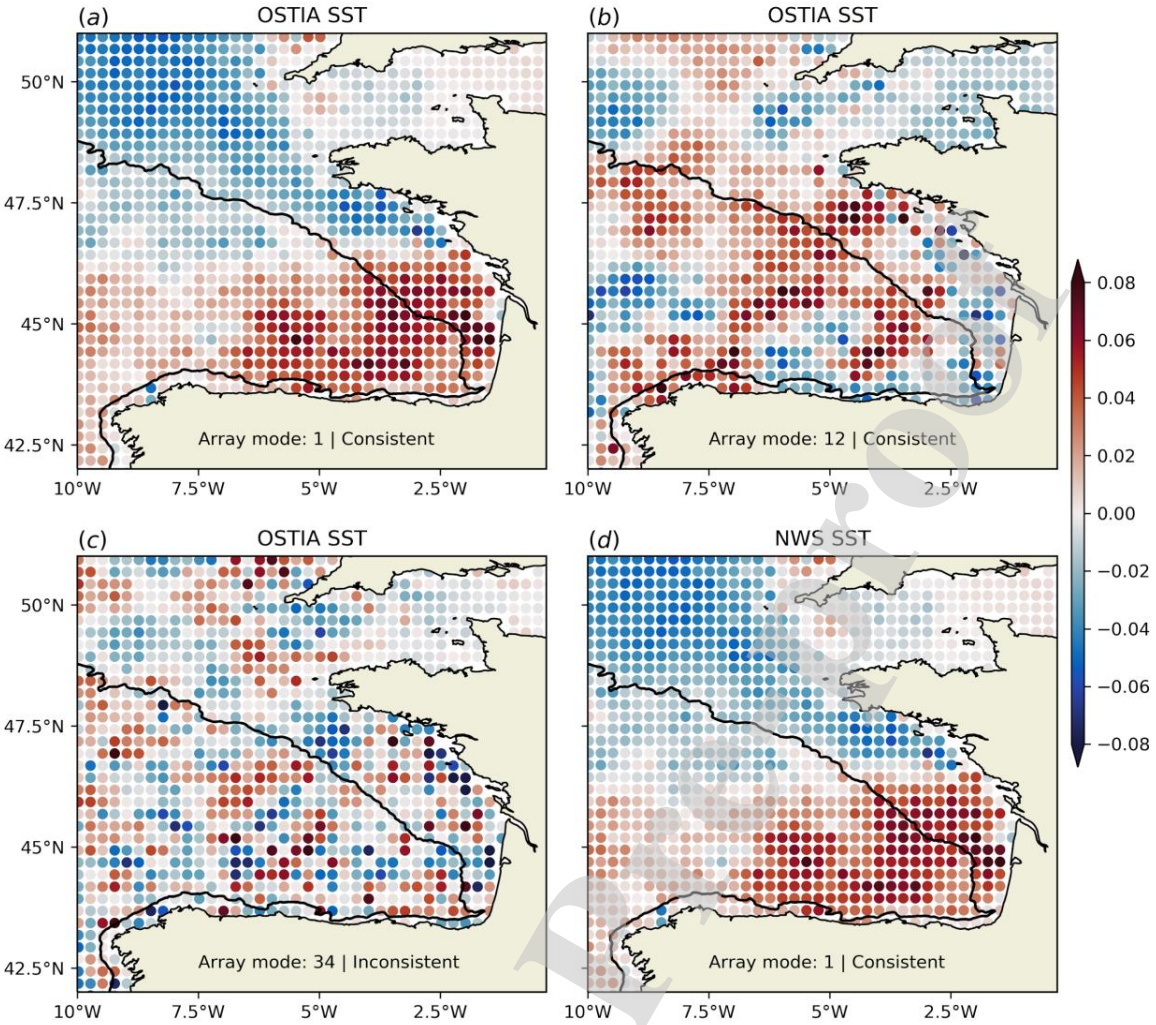


Figure 7 (a-c) Array modes 1, 12 and 34 for EnsP as "seen" from OSTIA SST on May 31, 2012, using one-fifth subsampling rate, and (below as text) corresponding results of the ArMCA1 consistency criterion using OSTIA SST observations; the 1st and 12th array modes are consistent; array mode 34 is inconsistent. (d) Same as (a) for NWS SST and one-sixth subsampling rate. Colorbar: array mode amplitude (no units) on (subsampled) data grid.

Journal Pre-proof

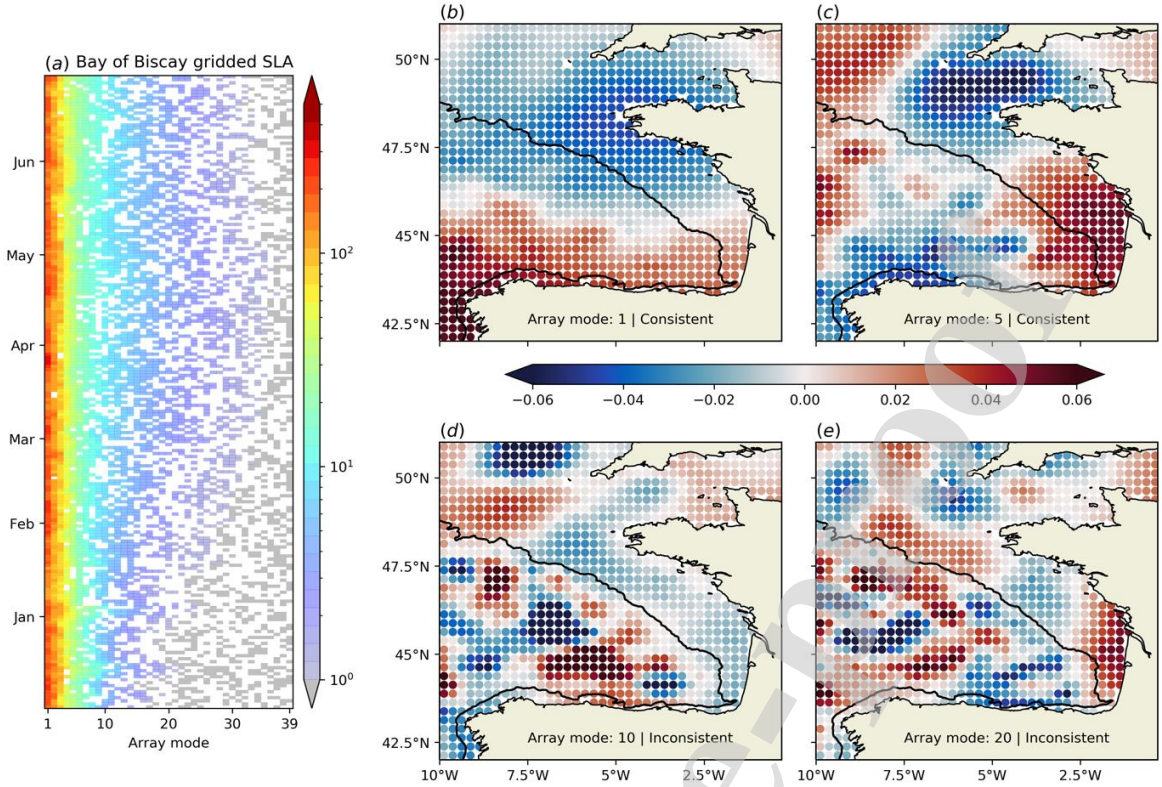


Figure 8 (a) Hovmöller plot of variations in time of gridded SLA array mode spectra vs. modal rank, including an EnsP consistency check against gridded SLA data. (b-e) Examples of consistent and inconsistent gridded SLA array modes on April 30, 2012 for EnsP. Colorbars and units as in Figs. 6 and 7.

Journal Pre-proof

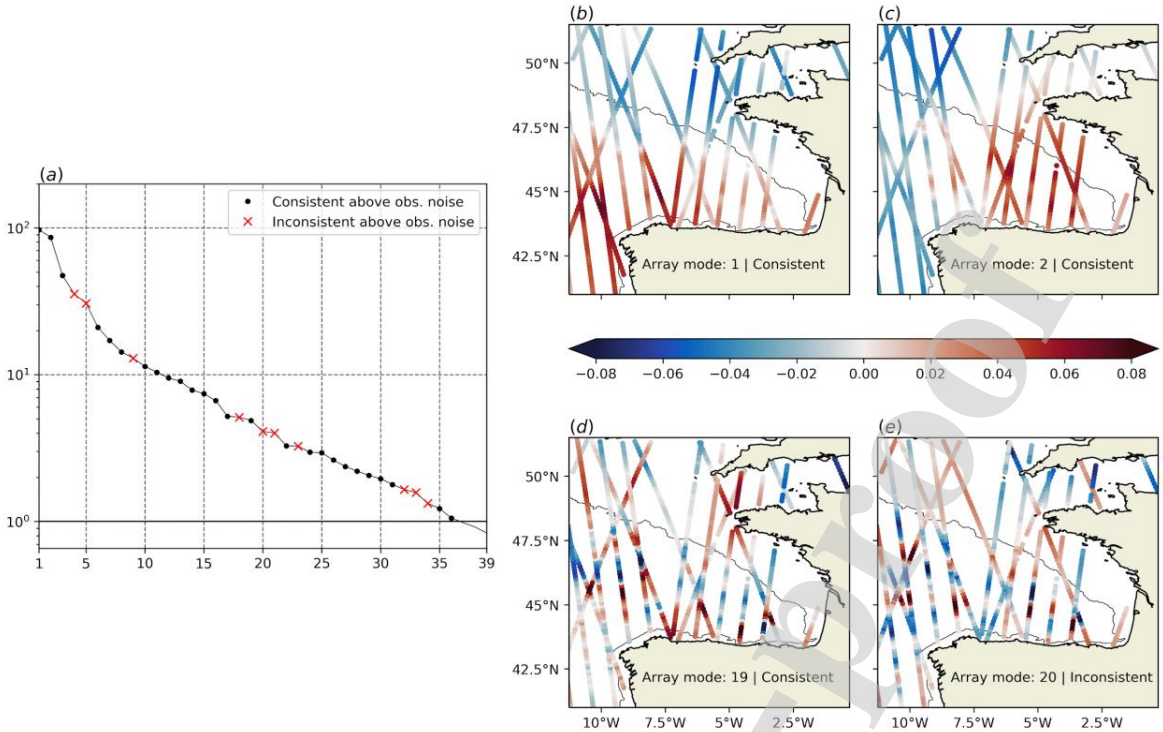


Figure 9 (a) Along-track SLA L3 array mode spectrum for the same period as in Fig. 4b, considering all tracks that would be assimilated in a 10-day assimilation cycle. Black and red markers for eigenvalues above the observational noise floor (= 1 in array space) denote consistent and inconsistent modes respectively. (b-c) Multi-track network consistent array modes of ranks 1 and 2 respectively (no units; all times "flattened", i.e. 2D representation of array modes including time space); (d-e) same as (b-c) for a higher consistent array mode of rank 19 and for an inconsistent array mode of rank 20 respectively. Colorbar and units as in Figs. 6 and 7.

Journal Pre-proof

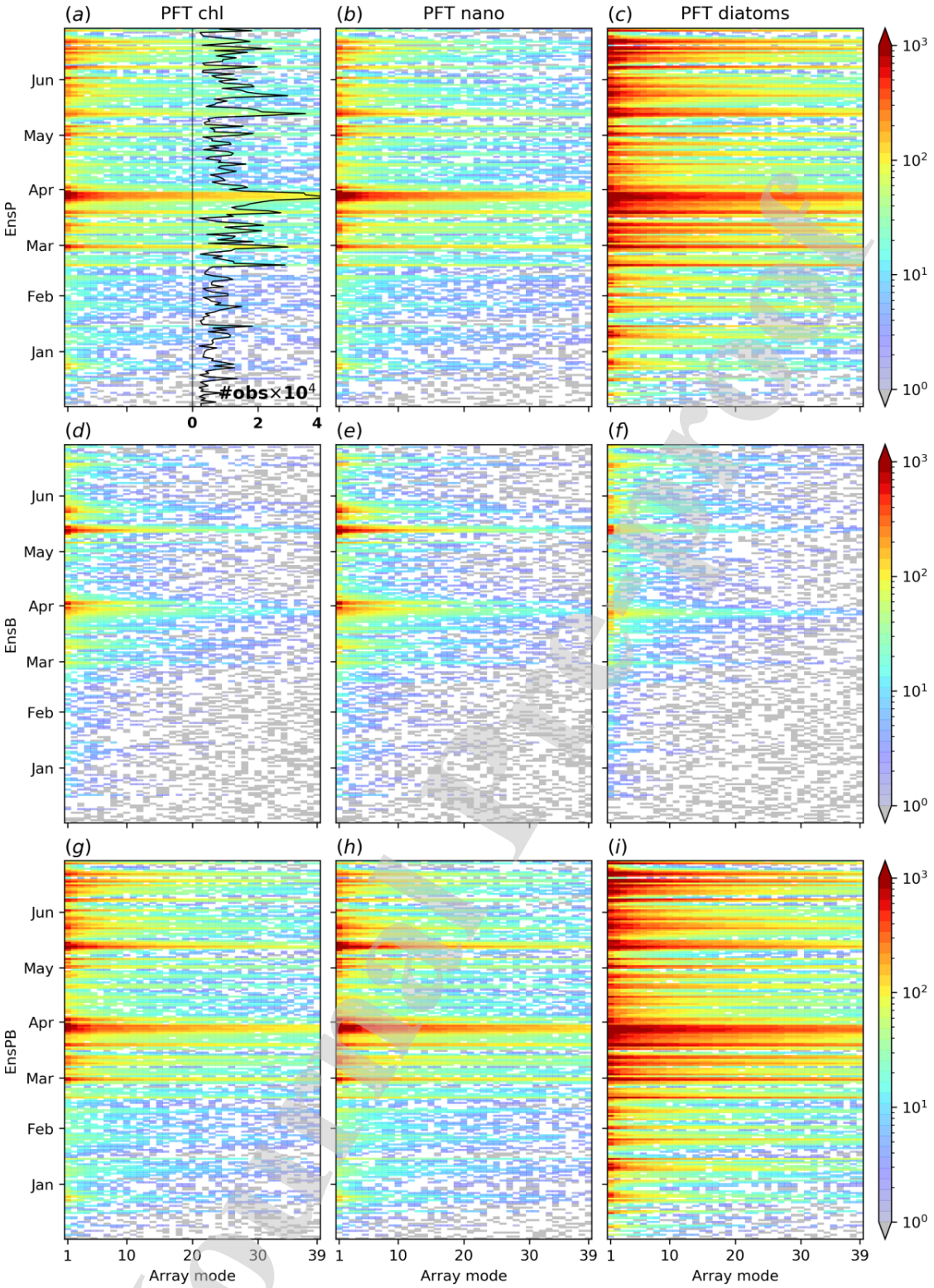


Figure 10 (a-c) Hovmöller plots of variations in time of PFT array mode spectra vs. modal rank, including EnsP consistency checks against PFT data for total chlorophyll, nanophytoplankton, and diatoms respectively as shown; (d-f) same as (a-c) for EnsB; (g-i) same as (a-c) for EnsPB. Chlorophyll distributions have been log transformed. Colorbars and units as in Fig. 6. Fig. 10a superimposed black line: number of data with largest values corresponding to full data coverage of the domain i.e. #obs $\sim(4 \cdot 10^4)$.

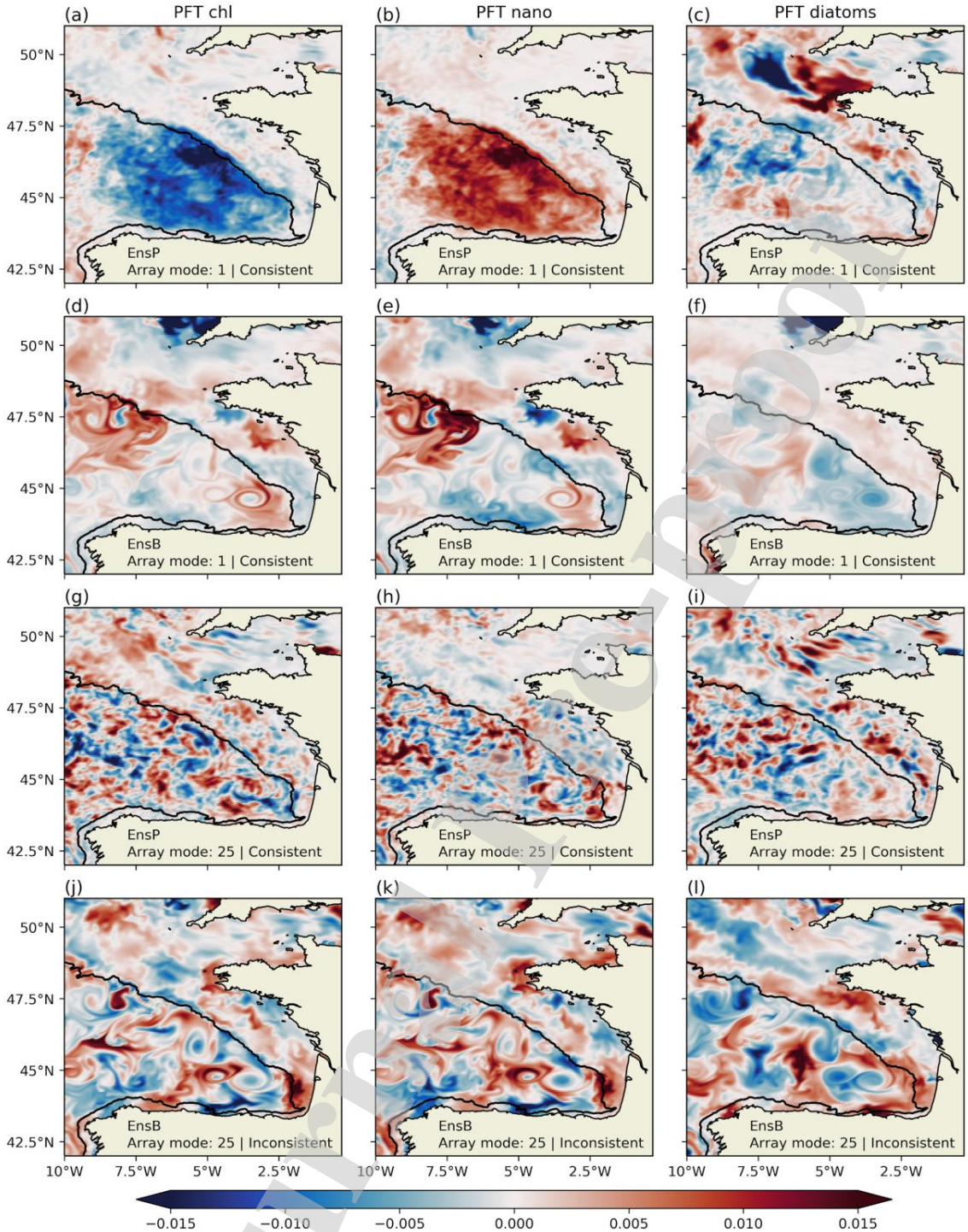


Figure 11 Examples of array modes for EnsP and EnsB on March 28, 2012, as "seen" from PFT data points, and (below as text) corresponding results of the ArMCA1 consistency criterion using PFT data: (a-c) array modes at rank 1 using EnsP (all consistent); (d-f) array modes at rank 1 using EnsB (all consistent, but different); (g-i) array modes at rank 25 using EnsP (all consistent); (j-l) array modes at rank 25 using EnsB (all inconsistent). PFT gridded data (gappy in general) have no gaps on that particular date in the Bay of Biscay. (left column) Total chlorophyll; (center) nanophytoplankton; (right column) diatoms. No data subsampling. Array modes calculated in log array space. Colorbar and units as in Fig. 7.

Journal Pre-proof

Highlights

- Model-data biases and model underspread revealed in rank histograms
- Consistent SST array modes at large scales and at small-scale
- Along-track array modes showed useful consistent information at the mesoscale
- Consistent PFT array modes at small-scale perturbing physics
- Additional error processes active in the model for inconsistent configurations

Journal Pre-proof

Journal Pre-proof

Authors' contribution statement using CRediT

1) Vassilios D. Vervatis: conceptualization, methodology, software, visualization, formal analysis, investigation, writing – original draft, writing – review and editing.
2) Pierre De Mey-Frémaux: conceptualization, methodology, software, formal analysis, writing – original draft, writing – review and editing.
3) Nadia Ayoub: conceptualization, formal analysis, writing – original draft, writing – review and editing.
4) John Karagiorgos: software, visualization, formal analysis.
5) Stefano Ciavatta: data curation, writing – review and editing.
6) Robert J. W. Brewin: data curation, writing – review and editing.
7) Sarantis Sofianos: conceptualization, resources.

The corresponding author,

Vassilios Vervatis

Journal Pre-proof

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Journal Pre-proof

Supplementary material

Fig. S1 presents the temporal variability of the PFT data in the Bay of Biscay, during the period December, 2011 to June, 2012. We also show the ocean colour L4 total chlorophyll product used in the Part 1 article (*Vervatis et al.*, 2021) and we verify that concentrations are of the same order to those of the PFT total chlorophyll. According to this, we categorised the four satellite PFT (pico, nano, diatoms and dino) into the two broad size groups in PISCES (nano or diatoms), in a manner most representative, ensuring that the total biomass (chlorophyll) from the model and satellite data can be compared like-for-like.

PFT diatoms and nanophytoplankton contribute together approximately more than 80% in total chlorophyll, whereas picoplankton contributes at about 10% and dinoflagellates less than 10% (Fig. S1). PFT diatom chlorophyll concentration is an order of magnitude larger compared with dinoflagellates and nanophytoplankton is about three times larger than picoplankton chlorophyll concentration (Fig. S1). In Fig. S2, we show the spatial distribution of the four satellite PFT and the total chlorophyll during the peak of the spring bloom on March 28, 2012. We confirm that the satellite micro class (i.e. diatoms and dino) is driven primarily by diatoms, far more abundant in the satellite data, with the two functional types being highly correlated in spatial. Fig. S3 presents scatter plots of combined vs. non-combined PFT chlorophyll, verifying the close relationship between functional types in a size class-based approach.

We also present results from one-on-one comparisons between model and PFT data, as opposed to the size class-based categorization merging different functional types. Figure S4 shows Hovmöller plot of rank histograms between EnsPB and PFT, in the same way as Figs. 5b-c, with one main difference: in Figs. S4a-b we do not combine the nano functional type with pico, nor we combine diatoms with dino. Rank histogram results for the nano class are degraded when pico and nano PFT data are not combined together in late-winter and early-spring when a primary bloom occurs (Fig. 5b vs. Fig. S4a). The latter may suggest that PISCES nano can be representative of a broader phytoplankton community, accounting also for smaller size classes. Rank histogram results are almost identical throughout the whole period for the micro class, regardless of whether dino and diatoms are combined together or not (Fig. 5c vs. Fig. S4b). Model-data one-on-one array mode consistency results (not shown) are in practice indistinguishable by visual inspection with the results presented in Figs. 10 and 11, confirming the validity of the size class-based approach.

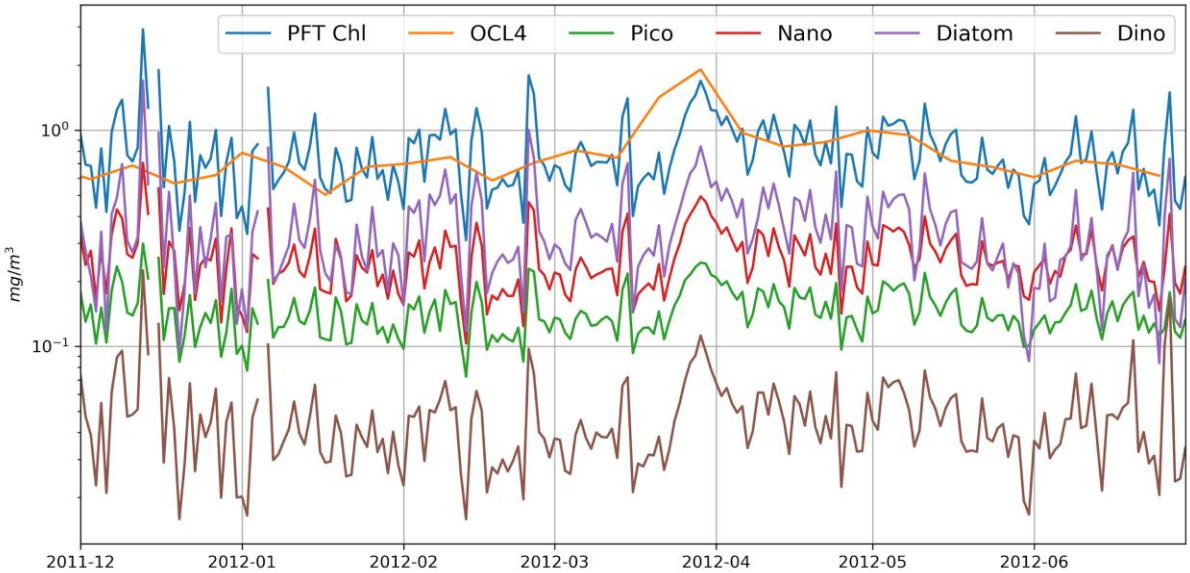


Figure S1 Ocean colour L4 (8-day frequency) and PFT (daily) chlorophyll concentration (mg/m^3) in the Bay of Biscay from December, 2011 to June, 2012.

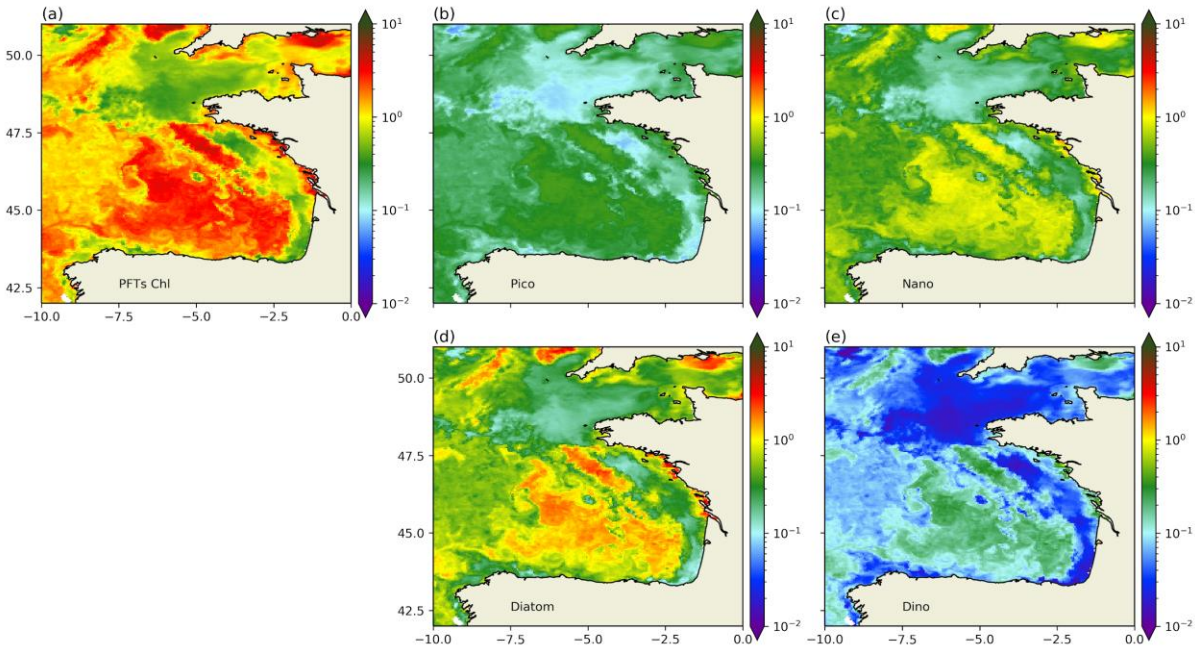


Figure S2 Spatial distribution of (a) PFT total chlorophyll and (b) pico, (c) nano, (d) diatoms, (e) dino concentrations (mg/m^3) on March 28, 2012.

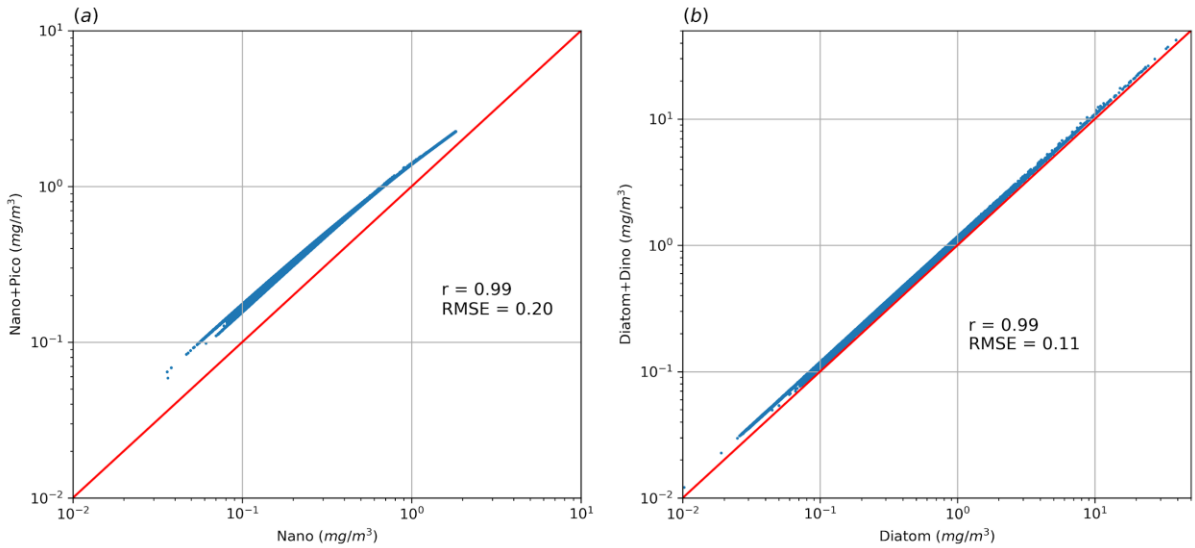


Figure S3 Scatter plots of chlorophyll concentrations in mg/m^3 on March 28, 2012: (a) PFT (nano and pico) vs. only nano, (b) PFT (diatoms and dino) vs. only diatoms. r is the correlation coefficient, RMSE the root mean square error (mg/m^3) and with red the 1:1 line.

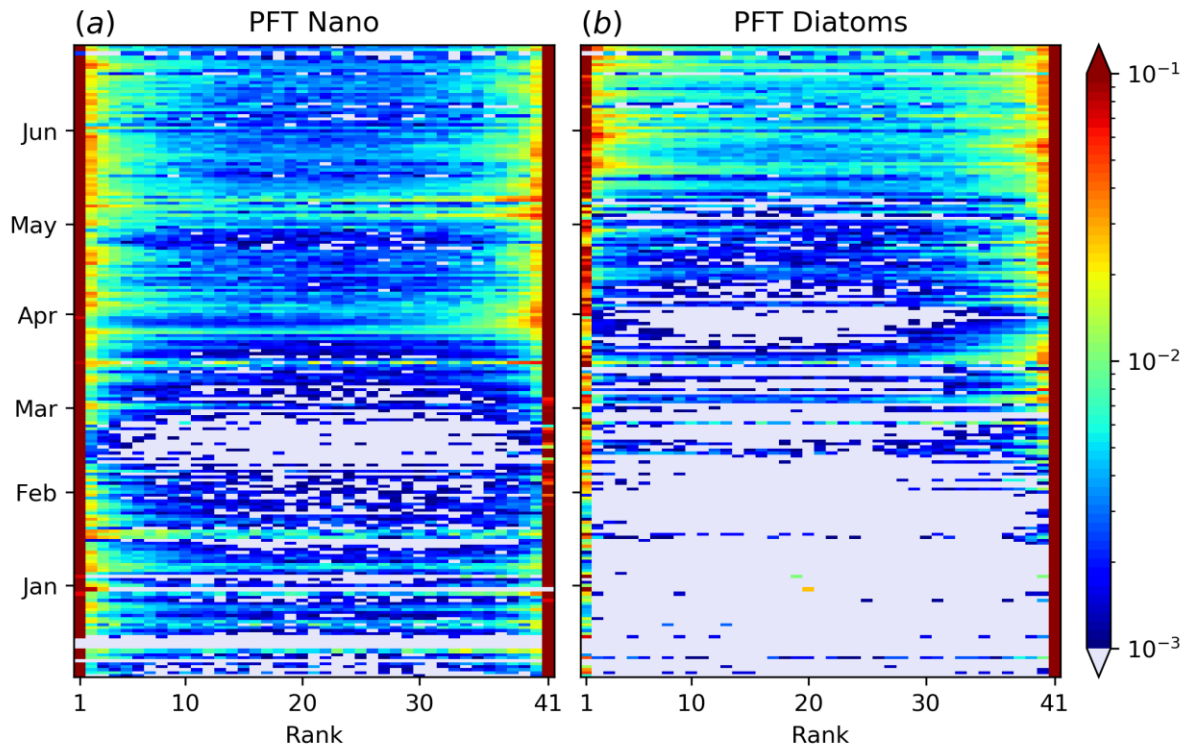


Figure S4 Hovmöller plot of rank histograms (same as in Figs. 5b-c) between EnsPB and PFT (a) nano not combined with pico, and (b) diatoms not combined with dino.