Reconstructing Global Chlorophyll-a Variations Using a Non-linear Statistical Approach

Elodie Martinez, Thomas Gorgues, Matthieu Lengaigne, Clément Fontana, Raphaëlle Sauzède, Christophe E. Menkès, Julia Uitz, Emanuele Di Lorenzo, Ronan Fablet

Reconstructing global Chlorophyll-a variations using a non-linear statistical approach

E. Martinez[1†,2*], T. Gorgues[1], M. Lengaigne[3], C. Fontana[2], R. Sauzède[4†,2], C. Menkes[5], J. Uitz[6], E. Di Lorenzo[7], R. Fablet[8]

[1] Laboratoire d'Océanographie Physique et Spatiale (LOPS), IUEM, IRD, Univ. Brest, Ifremer, CNRS, F-29280, Brest, France
[2] Ecosystèmes Insulaires Océaniens (EIO), IRD, Ifremer, UPF and ILM, Tahiti, French Polynesia
[3] Sorbonne Universités (UPMC, Univ Paris 06) -CNRS-IRD-MNHN, LOCEAN Laboratory, IPSL, Paris, France
[4] Sorbonne Universités, CNRS-INSU, Institut de la Mer de Villefranche,181 Chemin du Lazaret, 06230 Villefranche-Sur-Mer, France
[5]UMR ENTROPIE (IRD—Université de La Réunion—CNRS), Laboratoire d'Excellence Labex-CORAIL, Institut de 7 Recherche pour le Développement, BP A5, 98848 Nouméa Cedex, New Caledonia
[6] CNRS and Sorbonne Université, Laboratoire d'Océanographie de Villefranche, LOV, 06230 Villefranche-sur-mer, France
[7] Program in Ocean Science & Engineering, Georgia Institute of Technology, Atlanta, USA
[8] IMT-Atlantic, Lab-STICC, Brest, France

* Corresponding author: elodie martinez (elodie.martinez@ird.fr)

† current addresses

Key Points:

- A machine learning approach is applied to reconstruct the surface phytoplankton biomass at global scale over three decades.
- Chlorophyll variability derived from this statistical approach accurately reproduces satellite observations (possibly better than biogeochemical models).

- The sole use of surface predictors allows to accurately reproduce chlorophyll variability, in spite of its known sensitivity to three-dimensional processes.

Abstract

Monitoring the spatio-temporal variations of surface chlorophyll-a concentration (Chl, a proxy of phytoplankton biomass) greatly benefited from the availability of continuous and global ocean color satellite measurements from 1997 onwards. These two decades of satellite observations are however still too short to provide a comprehensive description of Chl variations at decadal to multi-decadal timescales. This paper investigates the ability of a machine learning approach (a non-linear statistical approach based on Support Vector Regression, hereafter SVR) to reconstruct global spatio-temporal Chl variations from selected surface oceanic and atmospheric physical parameters. With a limited training period (13 years), we first demonstrate that Chl variability issued from a 32-years global physical-biogeochemical simulation can generally be skillfully reproduced with an SVR using the model surface variables as input parameters. We then apply the SVR to reconstruct satellite Chl observations using the physical predictors from the above numerical model and show that the Chl reconstructed by this SVR more accurately reproduces some aspects of observed Chl variability and trends compared to the model simulation. The SVR is able to reproduce the main modes of interannual Chl variations depicted by satellite observations in most regions, including El Niño signature in the tropical Pacific and Indian Ocean. In stark contrast with the trends simulated by the biogeochemical model, it also accurately captures spatial patterns of Chl trends estimated by satellites data, with a Chl increase in most extratropical regions and a Chl decrease in the center of the subtropical gyres, although the amplitude of these trends are underestimated by half. Results from our SVR reconstruction over the entire period (1979-2010) also suggest that the Interdecadal Pacific Oscillation drives a significant part of decadal Chl variations in both the tropical Pacific and Indian Ocean. Overall, this study demonstrates that non-linear statistical reconstructions can be complementary tools to *in situ* and

satellite observations as well as conventional physical-biogeochemical numerical simulations to reconstruct and investigate Chl decadal variability.

1 INTRODUCTION

Phytoplankton—the microalgae that populate the upper lit layers of the ocean—fuels the oceanic food web and regulates oceanic and atmospheric carbon dioxide levels through photosynthetic carbon fixation. The launch of the "Coastal Zone Color Scanner" (CZCS) onboard the Nimbus-7 spacecraft in October 1978 (Hovis et al., 1980) provided the first synoptic view of near-surface chlorophyll-a concentration (Chl, a proxy of phytoplankton biomass). Although primarily focusing on coastal regions, CZCS also provided global pictures of Chl distribution and a new perspective on phytoplankton biomass seasonal variability (Campbell and Aarup, 1992; Longhurst et al., 1995; Yoo and Son, 1998; Banse and English, 2000).

After the failure of CZCS in 1986, ocean color observations were not available for more than a decade. The launch of the modern radiometric Sea-viewing Wide Field-of-View Sensor (SeaWiFS; McClain et al., 2004) in late 1997 followed later by other satellites allowed monitoring and understanding the spatio-temporal Chl variations at global scale over the past two decades. For instance, it revealed that El Niño events induce a Chl decrease in the central and eastern equatorial Pacific in response to reduced upwelled nutrients to the surface layers (e.g., Chavez et al., 1999; Wilson and Adamec, 2001; McClain et al., 2002; Radenac et al., 2012) but also a Chl signature outside the tropical Pacific through atmospheric teleconnections (Behrenfeld et al., 2001; Yoder and Kennelly, 2003; Dandonneau et al., 2004; Messie and Chavez, 2012). It also allowed identifying the Indian Ocean Dipole (IOD; Saji et al., 1999) as the main climate mode driving Chl interannual variations in the Indian Ocean (e.g. Murtugudde et al., 1999; Wiggert et al., 2009; Currie et al., 2013) and monitoring a Chl increase in the subpolar North Atlantic related to the positive phase of the North Atlantic Oscillation (NAO) (Martinez et al., 2016). Aside from the robust Chl decrease monitored in the mid-ocean gyres over the first decade of the XXIst century

(Polovina et al., 2008; Irwin and Oliver, 2009; Vantrepotte and Melin, 2009; Signorini and McClain, 2012), the reliability of the long-term trends derived from these satellite data are more questionable and led to conflicting results in the past literature (Behrenfeld et al., 2006; Vantrepotte and Mélin, 2011; Siegel et al., 2013; Gregg and Rousseaux, 2014). These discrepancies suggest that detection of robust global trend may require several decades of continuous observations (Beaulieu et al. 2013).

The production of longer, consistent ocean color time series can partly alleviate this issue. The combination of the global CZCS and SeaWiFS datasets provided an insight on the Chl response to natural decadal climate variations (Martinez et al., 2009; D'Ortenzio et al., 2012), such as the Pacific Decadal Oscillation (PDO; Mantua et al., 1997) and the Atlantic Multidecadal Oscillation (AMO; Enfield et al., 2001). However, blending these two archives or reconstructing them using compatible algorithms also led to contrasting results (Gregg and Conkright, 2002; Antoine et al., 2005).

The time span of the modern radiometric observations (~20 years), as well as the CZCS-SeaWiFS reprocessed time series, are still too short to investigate Chl decadal variations and longer-term trends. Longer, continuous and consistent records are required. *In situ* biogeochemical observatories can provide such long and continuous records, but their inhomogeneous spatial distribution and varying record length prevent a confident assessment of Chl long-term changes at the scale of a basin (Henson et al., 2016).

Coupled physical-biogeochemical ocean model simulations can provide additional, valuable information's in areas with limited observational coverage. These models resolve reasonably well the seasonal to interannual biogeochemical variability (Dutkiewicz et al., 2001; Wiggert et al., 2006; Aumont et al., 2015). They can however diverge in capturing Chl variations

at a timescale of a decade (Henson et al., 2009a,b; Patara et al., 2011), in particular phytoplankton regime shifts (Henson et al., 2009b). Different biological models are often coupled to different physical models, which renders the attribution of the different modelled responses to their physical or biological components difficult. The decadal or longer variability of the simulated primary producers should then be interpreted cautiously.

In this context, statistical methods reconstructing past Chl variations may be useful alternatives to overcome limitations associated with both observations and numerical models. While statistical reconstructions are now commonly used to extend physical variables back in time (e.g. Smith et al., 2012; Huang et al., 2017; Nidheesh et al., 2017), reconstructions of surface Chl are still in their infancy. Phytoplankton distribution is strongly controlled by physical processes, such as mixing and uplifting, fueling nutrients in the upper-lit layer (i.e., bottom up processes). Thus, relevant physical variables may allow to reconstruct Chl past variations. To our knowledge, a single study allowed the derivation of spatio-temporal surface Chl variations over the 1958-2008 period in the tropical Pacific (Schollaert et al., 2017). This reconstruction used a linear canonical correlation analysis on Sea Surface Temperature (SST) and Sea Surface Height (SSH) to improve the description of the Chl response to the diversity of observed El Niño events and decadal climate variations in the tropical Pacific.

The objective of the present study is to explore the potential of an alternative statistical technique to reconstruct Chl at global scale over a 32-year time-series (i.e., 1979-2010). The considered machine learning technique is based on a Support Vector Regression (SVR) which accounts for non-linearities between predictors and Chl. First, the SVR is trained over 1998-2010 on a self-consistent dataset of physical and Chl variables, all extracted from a forced ocean model simulation that includes a biogeochemical component (i.e., the NEMO-PISCES model). Then,

126 modelled physical variables are used to reconstruct Chl over 1979-2010. The feasibility and
127 robustness of the proposed reconstruction process is assessed through the comparison of modelled
128 *vs*. reconstructed Chl. In a second step, this framework is applied to satellite ocean color
129 observations.

130 2 DATA AND METHODS

131 2.1 The NEMO-PISCES simulation

132 In this study, we used the "Nucleus for European Modelling of the Ocean" (NEMO)
133 modelling framework (Madec, 2008). The NEMO configuration used displays a coarse resolution
134 with 31 vertical levels and a 2° horizontal grid with a refined 0.5° resolution in the equatorial band.
135 The model includes a biogeochemical component, the Pelagic Interaction Scheme for Carbon and
136 Ecosystem Studies (PISCES; Aumont et al., 2015). PISCES is a model of intermediate complexity
137 designed for global ocean applications (Aumont and Bopp, 2006), which uses 24 prognostic
138 variables and simulates biogeochemical cycles of oxygen, carbon and the main nutrients
139 controlling phytoplankton growth (nitrate, ammonium, phosphate, silicic acid and iron). It
140 simulates the lower trophic levels of marine ecosystems distinguishing four plankton functional
141 types based on size: two phytoplankton groups (small=nanophytoplankton and large=diatoms) and
142 two zooplankton groups (small=microzooplankton and large=mesozooplankton). Chl from
143 PISCES (hereafter referred to as Chl_{PISCES}) is defined as the sum of the simulated diatoms and
144 nanophytoplankton Chl content.

145 The NEMO-PISCES simulation is forced with atmospheric fields from the interannual
146 Drakkar Forcing Set 5 (DFS5.2, Dussin et al., 2014) for wind, air temperature and humidity,
147 precipitation, shortwave and longwave radiations. It is initialized with the World Ocean Atlas 2005

(WOA05) climatology for temperature, salinity, phosphate, nitrate and silicate (Garcia et al., 2006), while iron initial state is similar to the model climatology employed by Aumont and Bopp (2006). The model simulation was spun up using 3 repetitions of the 30 years' DFS5.2 forcing set, and finally ran over 1979-2010.

Although successfully used in a variety of biogeochemical studies (e.g. Bopp et al., 2005; Gehlen et al., 2006; Lengaigne et al., 2007; Schneider et al., 2008; Steinacher et al., 2010; Tagliabue et al., 2010; Séférian et al., 2013; Aumont et al., 2015; Keerthi et al., 2017; Parvathi et al., 2017 and references therein), the ability of the PISCES model to reproduce satellite surface Chl is briefly illustrated in section 3.1.

2.2 Chl derived from satellite radiometric observations

Satellite surface Chl for Case I waters is provided by the Ocean Colour – Climate Change Initiative (OC-CCI, hereafter referred to as $Chl_{OC\text{-}CCI}$) from the European Space Agency (http://www.esaoceancolour-cci.org/). This product combines multi-sensor, global, ocean-color products while attempting to reduce inter-sensor biases for climate research (Storm et al., 2013). OC-CCI extends the time series beyond that provided by single satellite sensors and perform better in terms of long-term consistency than other products from multi-mission initiatives (Belo Couto et al., 2016).

Only deep oceanic areas (depth>200 m) are considered to avoid coastal waters where specific non-case-1 waters products are required. The Chl Level-3 product is binned on a regular 1° grid with a monthly resolution over January 1998-December 2010. This time period does not extend beyond 2010 to be consistent with the NEMO-PISCES simulation. $Chl_{OC\text{-}CCI}$ is used to

169 evaluate the PISCES model performances in Section 3.1, and to train the statistical method in
170 Section 4.

171 2.3 Predictors and Chl variables

172 The variability of phytoplankton biomass is driven in many regions of the world ocean and
173 at many timescales by physical processes (e.g., Wilson and Adamec, 2002; Wilson and Coles,
174 2005; Kahru et al., 2010; Feng et al., 2015; Messie and Chavez, 2015). Our statistical architecture
175 relates to 12 predictors and one biological variable (Chl). A sample thus refers to 13 variables. The
176 12 predictors (7 physical variables from NEMO-DFS5.2, 2 temporal and 3 spatial parameters) are
177 detailed in **Table 1**, including their influence on Chl variations and the references supporting this
178 influence.

179 **Table 1:** physical predictors, the drivers of which they are proxy and associated references.

Proxy used as predictors	Relevance to Chl variations	References
SST	- Vertical mixing and upwelling - Impacts on phytoplankton metabolic rates	- Behrenfeld et al. (2006); Polovina et al. (2008); Martinez et al. (2009); Thomas et al. (2012) - Lewandowska et al. (2014)
Sea Level Anomaly	Thermocline/pycnocline depths	Wilson and Adamec (2001; 2002); Radenac et al. (2012)
Zonal and meridional wind components	Surface momentum flux forcing and vertical motions driven by Ekman pumping	Martinez et al. (2011); Thomas et al. (2012)
Zonal and meridional surface current components	Horizontal advective processes	Messie and Chavez (2012); Radenac et al. (2013)
Short-Wave radiations	Photosynthetically Active Radiation	Sakamoto et al. (2011)
Month (cos and sin)	Periodicity of the day of the year (day 1 is very similar to day 365 from a seasonal perspective)	Sauzède et al. (2015)
Longitude (cos and sin) **Latitude** (sin)	Periodicity (longitude 0°= longitude 360°)	Sauzède et al. (2015)

180

We purposely limited the predictors to surface variables because our objectives are 1) to reconstruct Chl from physical observations, which are mainly available through remotely sensed surface data (oceanic observations below the surface are indeed usually not accessible at large spatial-scales or interannual time-scales); 2) to build a statistical scheme that can complement more complex numerical models (here, NEMO-PISCES) which simulate complex three-dimensional processes and are costly to run.

First, the SVR is trained on physical predictors from NEMO and DFS5.2 *vs*. Chl_{PISCES}. The reconstructed Chl time-series is referred to as $Chls_{vr\text{-}PISCES}$. Second, the SVR is trained using the same physical predictors but *vs*. satellite Chl observations ($Chl_{OC\text{-}CCI}$). The reconstructed Chl time-series is referred to as $Chls_{vr\text{-}CCI}$.

2.4 Climate indices

Climate indices are provided by the National Oceanic and Atmospheric Administration website (www.esrl.noaa.gov/psd): the AMO, the Multivariate El Niño Southern Oscillation (ENSO) Index (MEI) and the Interdecadal Pacific Oscillation (IPO).

2.5 Support Vector Regression

The statistical reconstruction technique is based on a SVR. This method belongs to kernel methods in Statistical Learning Theory and relates to the Support Vector Machine (SVM, Vapnik, 1998). SVM is a kernel-based supervised learning method (Vapnik, 2000) developed for classification purpose in the early 1990s and then extended for regression by Vapnik (1995). The basic idea behind SVR is to map the variables into a new non-linear space using the kernel function, so that the regression task becomes linear in this space. The learning step estimates the parameters of the regression model according to a linear quadratic optimization problem, which

can be solved efficiently. SVR also uses a robust error norm based on the principle of structural risk minimization, where both the error rates and the model complexity should be minimized simultaneously. Because SVR can efficiently capture complex non-linear relationships, it has been used in a variety of fields, and more specifically for oceanographic, meteorological and climate impact studies (Aguilar-Martinez and Hsieh, 2009; Descloux et al., 2012; Elbisy, 2015; Neetu et al., 2020), as well as in marine bio-optics (Kim et al., 2014; Hu et al., 2018; Tang et al., 2019).

Predictors and Chl are normalized by removing their respective average and dividing them by their standard deviations. Two SVR are trained over 1998-2010: one on Chl_{PISCES} and one on $Chl_{OC\text{-}CCI}$ (Step A in **Figure 1**). This time period has been chosen as 1998 is the first complete year of the satellite $Chl_{OC\text{-}CCI}$ time-series, and 2010 is the last year available of the modelled Chl_{PISCES}. The two resulting SVR schemes are applied on the NEMO-DFS5.2 physical predictors over 1979-2010. Finally, the annual means and standard deviations initially removed are applied to perform the back transformation and reconstruct either $Chls_{vr\text{-}PISCES}$ or $Chls_{vr\text{-}CCI}$ (Step B in **Figure 1**).

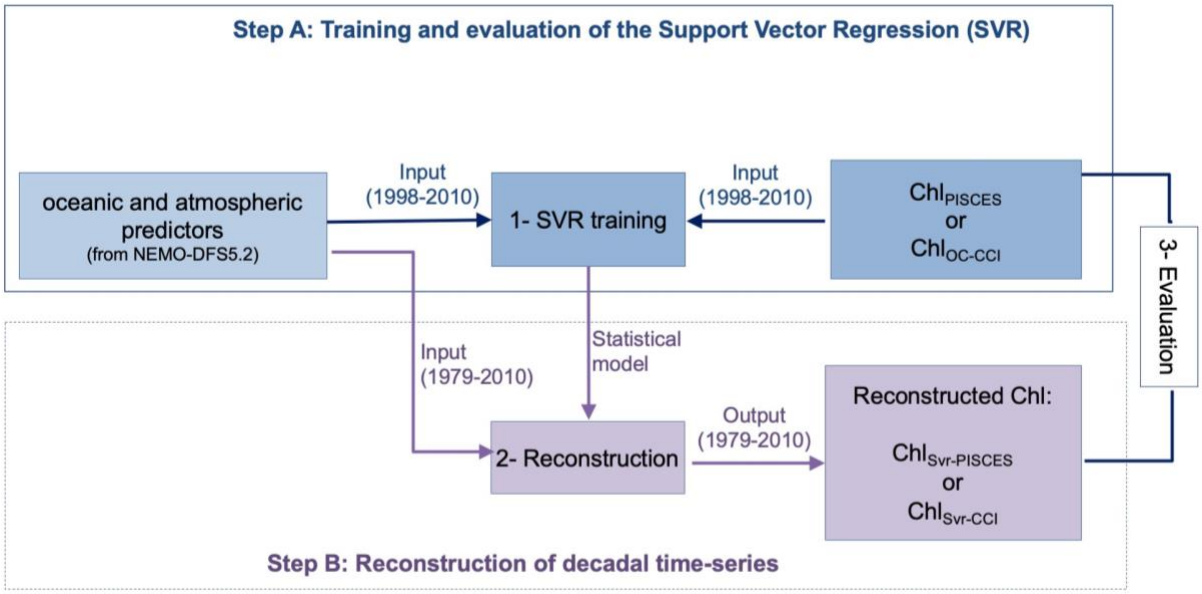


Figure 1 | Steps performed to train the SVR and reconstruct Chl time-series

Considering a Gaussian kernel, SVR only involves the selection of two hyperparameters: the penalty parameter C of the error term and the kernel coefficient gamma, driving the reduction of the cost function. C and gamma values are 1 and 0.1, respectively when the SVR is trained on Chl_{PISCES}, and 2 and 0.3 when trained on $Chl_{OC\text{-}CCI}$ (See details in the Supplementary material, and **Figure Supp 1A**). Sensitivity tests to an increasing portion of the sample total number (from 0.2 % to 9 % of the full dataset) used in the training process are performed (See Supplementary materials and **Figure Supp 1B**). The mean absolute error stabilizes for a sample number higher than 6 %, suggesting that the SVR skills don't improve much afterward. This observation combined with computational limitations lead us to present the 9% experiment hereafter.

2.6 Empirical Orthogonal Function analysis

The SVR skills to reconstruct Chl interannual to decadal variations are investigated performing Empirical Orthogonal Function analysis on Chl_{PISCES}, $Chl_{OC\text{-}CCI}$, $Chls_{vr\text{-}PISCES}$ and $Chls_{vr\text{-}CCI}$. First, Chl data are centered and reduced (i.e., the monthly climatology is removed and the induced anomalies are divided by their standard deviations) to avoid an overly dominant contribution of high values on the analysis (Emery and Thompson, 1997) over the periods of interest (i.e., 1998-2010 or 1979-2010). A 5-month running mean is applied to focus on the interannual/decadal signal. The analysis is separately performed for the Atlantic, Pacific and Indian Oceans north of 40°S until 60°N, and for the 40°S-60°S region hereafter referred to as the Austral Ocean. Indeed, the large area covered by the Pacific Ocean and its dominant modes in climate variability (i.e., ENSO/IPO), could regionally dampen other modes of variability. Basin-scale spatial maps are then gathered to a global one, refer to as EOF. The associated time-series refer to as the Principal Components (PCs).

3 Synthetic reconstruction from a physical-biogeochemical ocean model

This section assesses the reliability and robustness of the SVR approach using a complete and coherent dataset extracted from a global simulation performed with a coupled physical-biogeochemical ocean model. The SVR is first trained over 1998-2010 on Chl_{PISCES}, and $Chls_{vr\text{-}PISCES}$ is reconstructed over 1979-2010. Chl_{PISCES} and $Chls_{vr\text{-}PISCES}$ are then compared over 32 years to evaluate the consistency of the proposed data-driven reconstruction scheme.

3.1 Evaluation of Chl_{PISCES} at global scale

The ability of the NEMO-PISCES model to reproduce the satellite Chl over 1998-2010 is briefly presented here. Boreal winter and summer climatology from Chl_{PISCES} compare reasonably well with those of $Chl_{OC\text{-}CCI}$ (**Figures 2A *vs*. B, and C *vs*. D**). The model correctly represents the main spatial patterns with, for instance, higher Chl and a stronger seasonal cycle at high latitudes, despite an overestimated biomass in the Southern Ocean (Launois et al., 2015). The model also captures low Chl in the subtropical gyres, with some underestimation. This discrepancy may be explained by the lack of acclimation dynamics to oligotrophic conditions or by the assumption of constant stoichiometry either in phytoplankton or in organic matter in the model (Aumont et al., 2015; Ayata et al., 2013). The model underestimates Chl values in the equatorial Atlantic and Arabian Sea. In this latter region, mesoscale and submesoscale processes unresolved by the model have been shown to be of critical importance (Hood et al., 2003; Resplandy et al., 2011). Finally, the parameterization of nitrogen-fixing organisms not explicitly modelled in that PISCES version could explain the Chl_{PISCES} underestimation in the western Pacific in austral summer (Dutheil et al., 2018).

High Chl are accurately simulated in the eastern boundary upwelling systems. In two of the three main High Nutrient Low Chlorophyll (HNLC) regions, i.e. the equatorial Pacific and the eastern subarctic Pacific, the model successfully reproduces the moderate $Chl_{OC\text{-}CCI}$. However, the model overestimates $Chl_{OC\text{-}CCI}$ east of Japan because of an incorrect representation of the Kuroshio current trajectory. This common bias in coarse resolution models (i.e., Gnanadesikan et al., 2002; Dutkiewicz et al., 2005; Aumont and Bopp, 2006) is potentially related to too deep mixed layer simulated in winter inducing very strong spring blooms (Aumont et al., 2015). In the Southern Ocean, the third and largest main HNLC region, the model overestimates $Chl_{OC\text{-}CCI}$ values, especially during summer. However, the standard satellite algorithms that deduce Chl from reflectance tend to underestimate *in situ* observations by a factor of about 2 to 2.5, especially for intermediate concentrations (e.g., Dierssen and Smith, 2000; Kahru and Mitchell, 2010). It is to note that Chl in physical-biogeochemical coupled models is commonly overestimated in the Southern Ocean, and systematically underestimated in the oligotrophic gyres (Séférian et al., 2013).

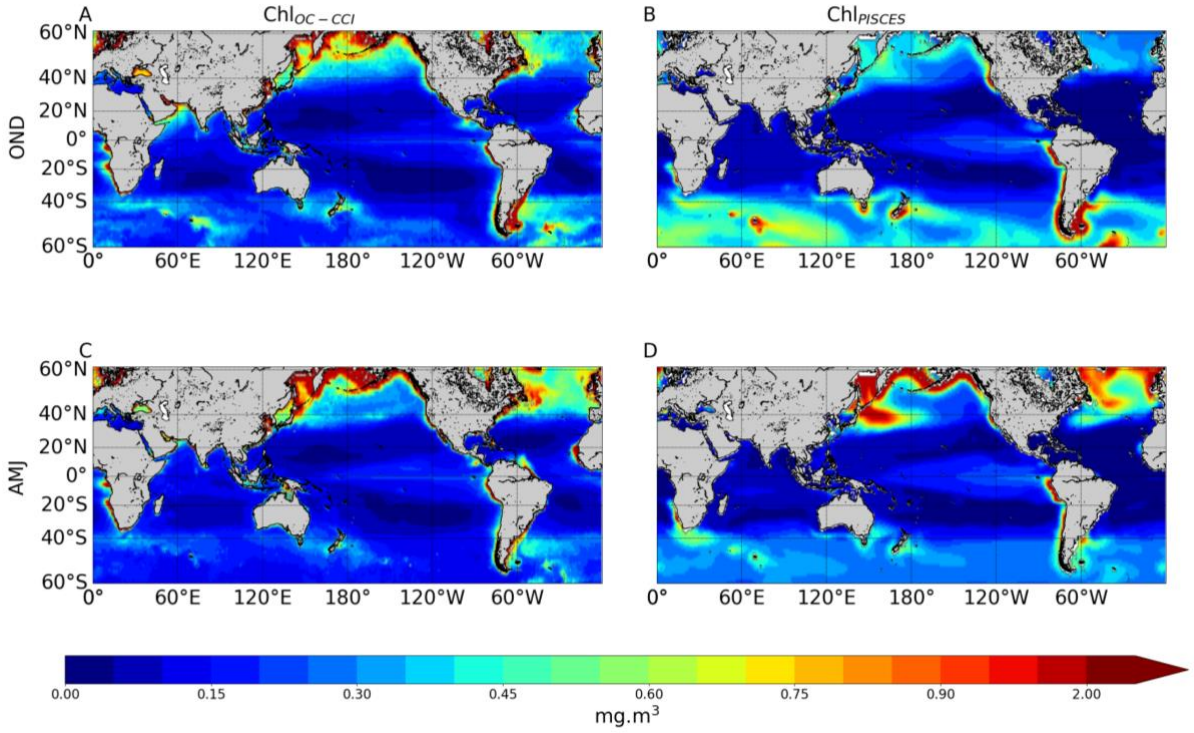


Figure 2 | Surface seasonal mean of Chl (mg.m^{-3}) over 1998-2010 derived from satellite (left panels) and the PISCES model (right panels), in October-November-December (**A-B**) and April-May-June (**C-D**).

The 1st mode of the EOF analysis performed on interannual Chl displays close percent of total variance for $Chl_{OC\text{-}CCI}$ and Chl_{PISCES} (16.6% *vs*. 21.1%, respectively). Their PCs in the Pacific Ocean are well correlated with the MEI (r = 0.71 and 0.89 with p=0.0015 and p<0.001, respectively; **Figure 3C**). PCs show the greatest positive values in January 1998 during the peak of the strong 1997/1998 El Niño event and the greatest negative values during the following La Niña beginning of 1999. The associated EOFs display a Chl horseshoe pattern (**Figure 3A and 3B**), reminiscent of the ENSO pattern on SST from NEMO (SST_{NEMO}; **Figure Supp 2;** Messie and Chavez, 2012). While the tropical Pacific experiences a Chl decrease during El Niño events, the North and South Pacific display a Chl increase, and inversely during La Niña. This typical ENSO pattern is also related to remote Chl anomalies outside the Pacific induced by atmospheric teleconnections, such as a Chl decrease in the tropical North Atlantic and in the South Indian Ocean

during El Niño. Although the Atlantic and Indian Ocean's PCs are not correlated with the MEI (0.14 and 0.05, respectively), their EOFs are similar to those obtained from analysis performed at global scale (*vs*. basin scale here) and which have been largely discussed in the past (e.g., Behrenfeld et al., 2001, 2006; Chavez et al., 2011; Yoder and Kennelly, 2003). Chl_{PISCES} reasonably well captures the first mode of $Chl_{OC\text{-}CCI}$ interannual variability over 1998-2010 in the Pacific and Atlantic Oceans, with 0.89 and 0.77 ($p<0.001$) correlations between their PCs, respectively, but not in the Indian Ocean, where the PCs correlation is far weaker (0.13) and insignificant (**Figure 3C-E**).

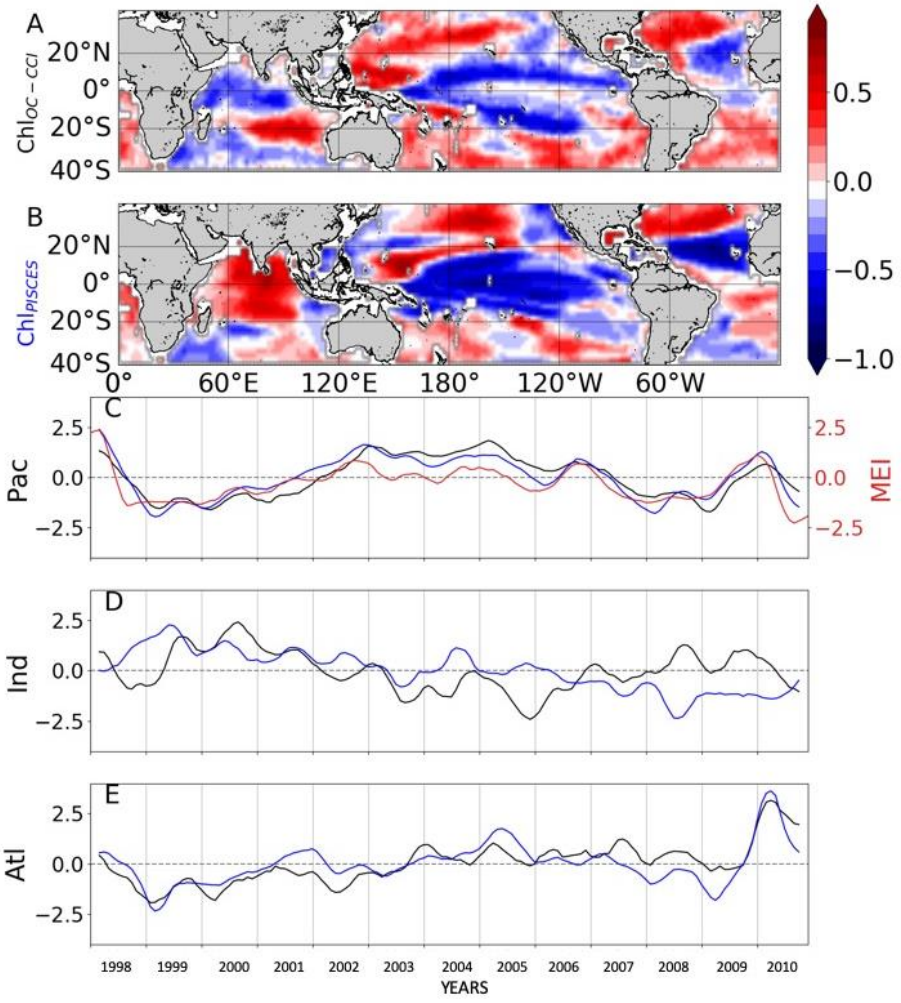


Figure 3 | 1st mode of basin-scale EOFs of interannual **A**) $Chl_{OC\text{-}CCI}$ and **B**) Chl_{PISCES}, and their corresponding PCs over 1998-2010 in the **C**) Pacific, **D**) Indian and **E**) Atlantic Oceans. $Chl_{OC\text{-}CCI}$ and Chl_{PISCES} PCs are represented by the black and blue lines, respectively. The MEI index is reported in red (right y-axis) on C.

3.2 Evaluation of the SVR method trained on synthetic data only

3.2.1 Statistical performances

A first evaluation of the SVR applied on the synthetic dataset (i.e., both physical and biogeochemical model outputs) is provided for the dedicated subset (i.e., 20% of 9% of the total data set) over the 1998-2010 training time period. Chl_{PISCES} and $Chls_{vr\text{-}PISCES}$ datasets display a determination coefficient of 0.95 and a root mean square error (RMSE) of 0.22 (see **Figure Supp 1C**), indicating at first glance a very good ability of the SVR to reconstruct Chl_{PISCES}. The SVR reconstruction is very accurate when comparing the full modelled and reconstructed Chl for (i) the 1998-2010 training time period, (ii) the 1979-1997 fully independent dataset and (iii) the 1979-2010 whole dataset, both at global and basin scales (**Table 2** and **Figure 4).** For each oceanic basin, determination coefficients between both datasets exceed 0.85, except in the Austral Ocean where they get down to 0.71. RMSE are lower than 0.14 and associated with a slope ranging from 0.84 in the Austral to 0.97 in the Atlantic (**Figure 4**). In addition, the quality of the reconstructed $Chls_{vr\text{-}PISCES}$ over the 1979-1997 independent time period is only marginally degraded compared to the 1998-2010 training period or the 1979-2010 full period.

Table 2 | Statistical performances between Chl_{PISCES} *vs.* $Chl_{Svr\text{-}PISCES}$ normalized monthly anomalies for the global ocean and the 4 oceanic basins over the 1998-2010, 1979-1997 and the whole 1979-2010 time period. The determination coefficient, NRMSE and the slope of the Chl_{PISCES} *vs.* $Chl_{Svr\text{-}PISCES}$ regression line are indicated.

	1998-2010				*1979-1997*				*1979-2010*			
	R^2	RMSE	slope	number of bins	R^2	RMSE	slope	number of bins	R^2	RMSE	slope	number of bins
Global 60°S-60°N	0.96	0.1	0.95	4 487 457	0.94	0.12	0.95	6 562 272	0.94	0.11	0.95	11 049 729
Pacific	0.92	0.1	0.93	1 945 173	0.88	0.12	0.9	2 843 501	0.9	0.11	0.91	4 788 674
Indian	0.86	0.11	0.85	612 612	0.85	0.11	0.89	895 356	0.85	0.11	0.87	1 507 968
Austral	0.81	0.06	0.8	1 026 876	0.71	0.08	0.75	1 506 260	0.76	0.07	0.77	2 533 136
Atlantic	0.96	0.12	0.96	902 796	0.94	0.14	0.94	1 317 155	0.94	0.13	0.95	2 219 951

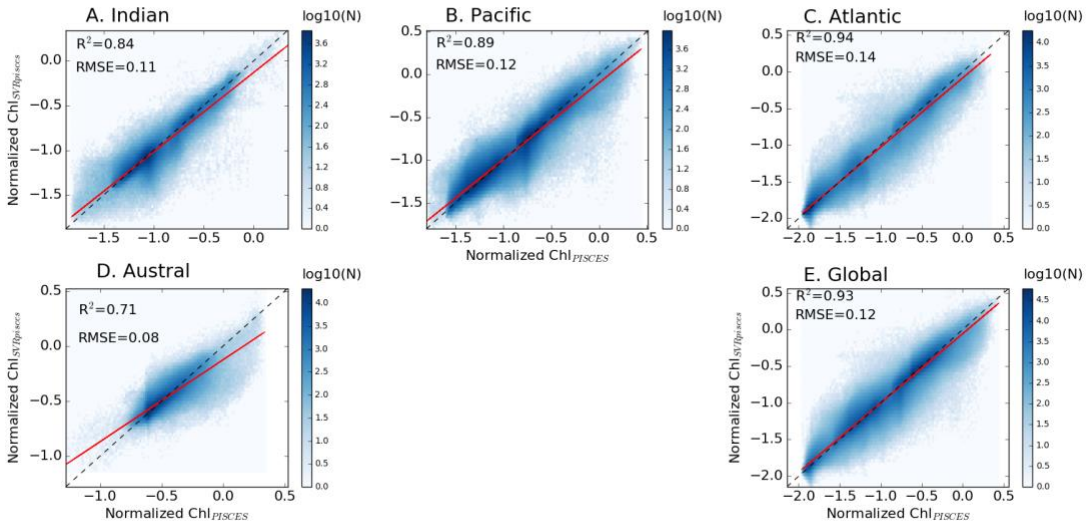


Figure 4 | Scatter plots of Chl_{PISCES} *vs.* $Chl_{Svr\text{-}PISCES}$ normalized monthly anomalies over 1979-1997, **A-D**) for each basin and **E**) at global scale between 60°S and 60°N. The Chl_{PISCES} *vs.* $Chl_{Svr\text{-}PISCES}$ and the 1:1 regression lines are plotted as the red and black lines, respectively. The figure is color-coded according to the density of observations.

3.2.2 Evaluation of the reconstructed Chl spatio-temporal variability

The Normalized Root-Mean-Square-Error (NRMSE, i.e. RMSE normalized by the average of Chl used to train the SVR) between Chl_{PISCES} and $Chl_{Svr\text{-}PISCES}$ filtered with a 5-month running

mean (to discard the high frequency signal) shows an error ranging between 10 and 20% over 1998-2010 (**Figure 5A**). Their correlation exceeds 0.7 ($p<0.001$) over most of the global ocean (**Figure 5B**). At mid-latitudes they are generally larger than 0.8, and they range between 0.6 and 0.9 in the equatorial Pacific. This accurate reconstruction demonstrates that a strong relationship exists between physical processes and Chl at global scale. However, the reconstructed Chl field can be regionally less accurate. For instance, the edges of the oligotrophic gyres (delimited by the 0.1 $mg.m^{-3}$ contour in **Figure 5A**) exhibit the highest NRMSE and lowest correlations. Large NRMSE are also evident in the Gulf Stream region while the western tropical Atlantic exhibits lower correlations than 0.5.

Those discrepancies could be due first to the zooplankton grazing pressure (top-down control) which is often overestimated in PISCES simulations. It results in an underestimated nanophytoplankton biomass in the oligotrophic gyres, emphasized along their edges (Laufkötter et al., 2015). Because the top-down control is not accounted for by the SVR, Chl variability induced by the overgrazing in these areas might not be captured. Second, in the equatorial Pacific Ocean, a minimum iron threshold value has been imposed (0.01 $nmol.L^{-1}$) in the biogeochemical model. Without that threshold Chl is too low on both sides of the equator, resulting in a strong accumulation of macronutrients and a spurious poleward migration of the subtropical gyre boundaries (Aumont et al., 2015). While the existence of such a threshold suggests that a minor but regionally important source of iron is missing in PISCES, it also suggests the inability of the SVR in reproducing ecosystem dynamics related to such artificial input of micro-nutrient. Finally, atmospheric input of iron through desert dust deposition is known to be stronger in the Atlantic than in the Pacific Ocean (Jickells et al., 2005). Such signal cannot be accounted for by the SVR

with the given predictors, which might (with meso – and sub-mesoscale activities) explain the higher NRMSE in the north western Atlantic than in the north-western Pacific.

As expected, areas of high NRMSE and low correlations between Chl_{PISCES} and $Chl_{Svr\text{-}PISCES}$ identified over 1998-2010 (**Figure 5, left column**) extend and strengthen over 1979-1997 (**Figure 5, right column**). Indeed, the correlations significantly decrease in the tropical Pacific while they slightly decrease in mid-latitudes between the two periods. Correlations remain high and NRMSE low in the North-West Pacific, North and South-West Atlantic, and South Indian Oceans as well as over a large part of the Southern Ocean providing confidence for analyses extended beyond the training period of the SVR.

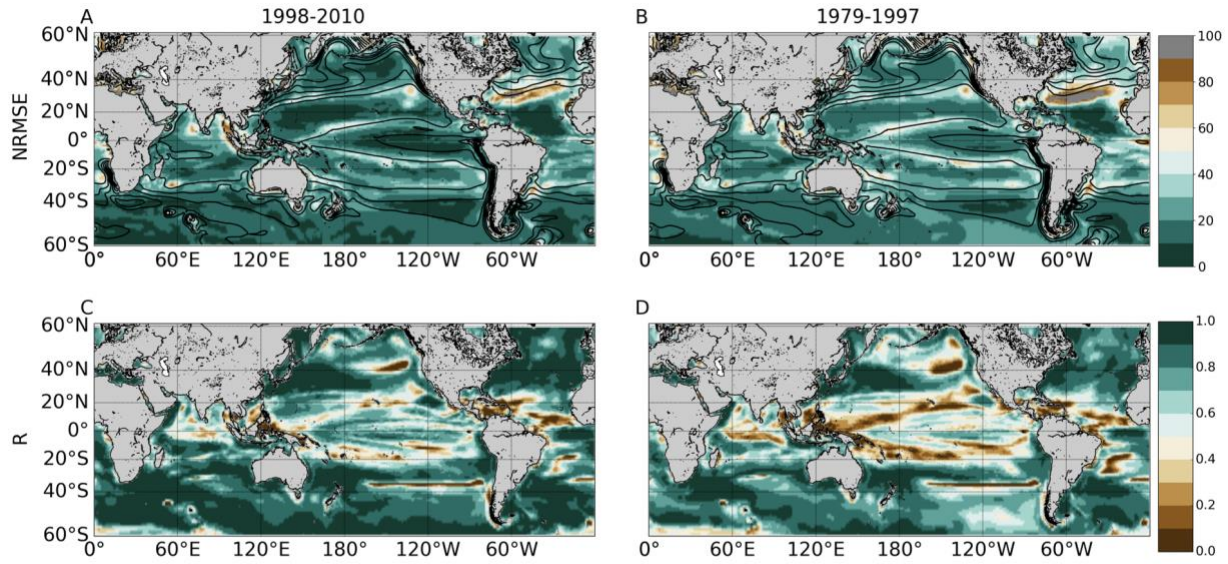


Figure 5 | A-B) NRMSE (in %) and **C-D)** correlation between Chl_{PISCES} *vs*. $Chl_{Svr\text{-}PISCES}$ after applying a 5 month-running mean on both time-series. These 2 diagnostics are calculated over 1998-2010 (left column) and 1979-1997 (right column). Contours on the upper panels show their respective 1998-2010 Chl time average (every 0.1 $mg.m^{-3}$).

The analysis is now extended to the 1979-2010 time-period to investigate the skills of the SVR in reproducing phytoplankton interannual/decadal cycles. The 1st EOFs of Chl_{PISCES} *vs*.

373 $Chls_{vr\text{-}PISCES}$ have the same sign of variability over 72% of the global ocean (**Figures 6A-B**). Both
374 EOFs are similar in the Pacific and Atlantic Oceans and their PCs are highly correlated over 1979-
375 2010 (**Table 3; Figures 6C,6E**). In the Pacific, these EOFs strongly resemble the typical horseshoe
376 pattern of IPO with SST anomalies of opposite polarities in the tropical and extra-tropical Pacific
377 regions (**Figure Supp 3**). Correlations between Chl_{PISCES} and $Chls_{vr\text{-}PISCES}$ 1st PCs and the IPO
378 index are high (0.94 and 0.95 with $p<0.001$, respectively; blue and black *vs*. red lines in **Figure**
379 **6C**). It highlights that the 1st mode of Chl variability in the Pacific is strongly driven by the IPO.
380 In the Atlantic, both PCs are strongly correlated with the AMO (-0.8 for $Chls_{vr\text{-}PISCES}$ and -0.85 for
381 Chl_{PISCES} with $p<0.001$; **Figure 6E**). The AMO shifts from a cold to a warm phase in the mid-
382 1990's (**Figure Supp3**), and is associated with a decrease in Chl (**Figures 6A-B**).

383 The 1st two modes explain a similar percent variance for Chl_{PISCES} and $Chls_{vr\text{-}PISCES}$ in the
384 four oceanic basins, with the exception of the 1st mode in the Atlantic Ocean (see **Table 3**). In this
385 basin $Chls_{vr\text{-}PISCES}$ percent variance is underestimated by a factor 2 compared to Chl_{PISCES}, while
386 their 1st EOFs and PCs are well correlated. One explanation might be that the AMO is the climate
387 cycle with the longest period (80 years) when compared to the IPO. Thus, it might be the most
388 difficult signal to reproduce as the SVR is trained over a relatively "short" 12 years' time-period.

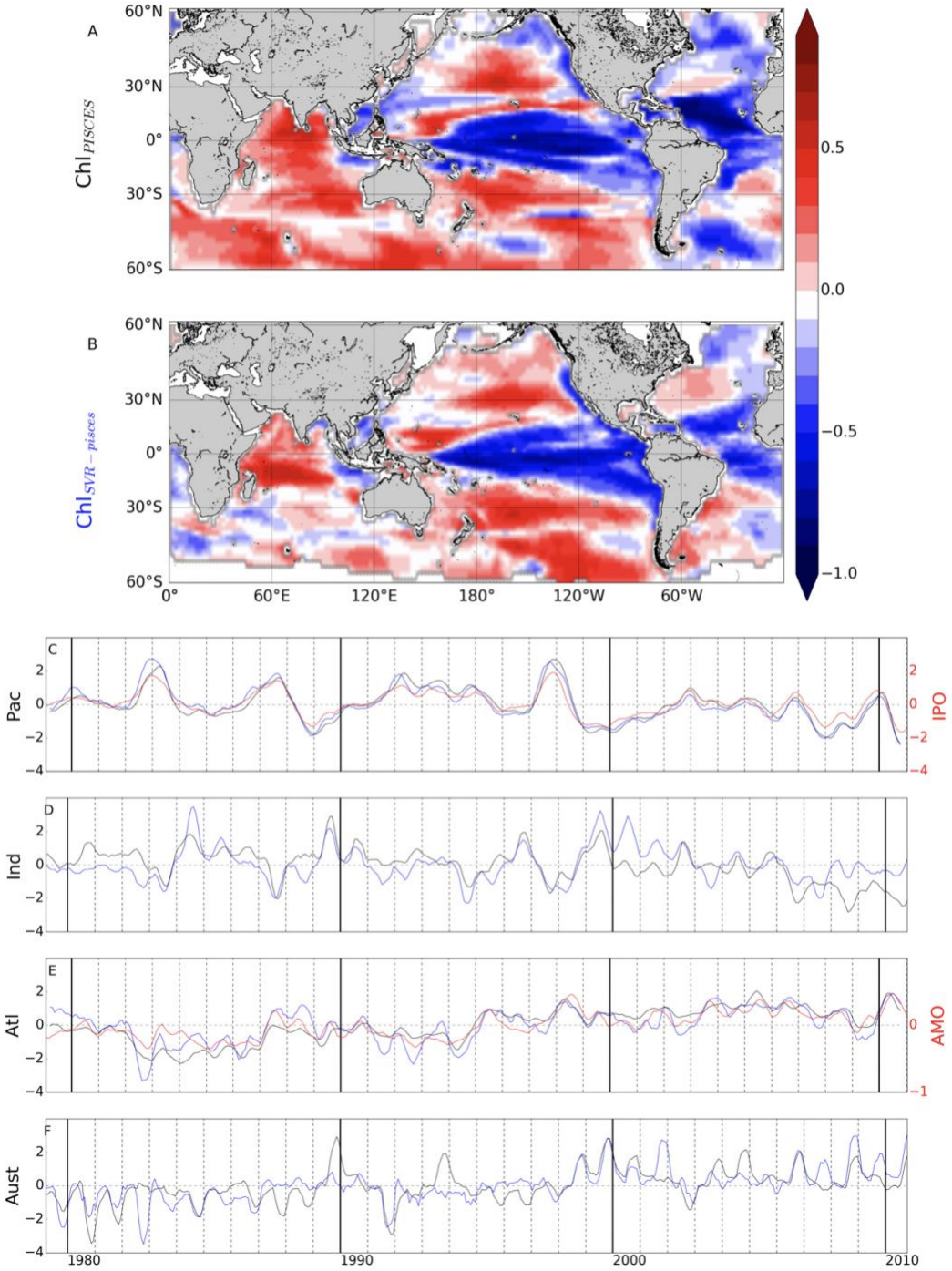
389 The agreement between Chl_{PISCES} and $Chls_{vr\text{-}PISCES}$ 1st EOFs is not as good in the Austral
390 and Indian Oceans when compared to the Atlantic and Pacific Oceans (**Table 3** and **Figures 6A-**
391 **B**). In the Indian Ocean, the Chl_{PISCES} EOF exhibits a maximum positive variability along the
392 western Arabian Sea, while it is located north-east of Madagascar for $Chls_{vr\text{-}PISCES}$. In the Austral
393 Ocean, Chl_{PISCES} and $Chls_{vr\text{-}PISCES}$ EOFs roughly follow a zonal distribution.

394 A strong correspondence between SST and Chl has been previously reported over a large
395 part of the global ocean (Behrenfeld et al., 2006; Martinez et al., 2009; Siegel et al., 2013),

demonstrating the close interrelationship between ocean biology and climate variations. Consequently, it is not surprising to observe strong correlations between Chl_{PISCES} or $Chls_{vr\text{-}PISCES}$ and climatic indexes mostly built on SST anomalies (**Figure Supp 3**).

Table 3 | Percent variance explained by the first two modes of the Empirical Orthogonal Function analysis performed on Chl_{PISCES} and $Chls_{vr\text{-}PISCES}$ for each oceanic basin over 1979-2010. The correlation (r) between the $Chls_{vr\text{-}PISCES}$ and Chl_{PISCES} PCs is also reported with a significant level of p<0.001 (*) and p<0.002 (**).

	1st mode			***2nd mode***		
	Chl_{PISCES}	$Chls_{vr\text{-}PISCES}$	r	Chl_{PISCES}	$Chls_{vr\text{-}PISCES}$	r
Pacific	19.7	23.5	0.95*	7.8	6.7	0.6*
Indian	13.1	14.1	0.58*	10.5	9.4	0.78*
Austral	13.8	12.1	0.62*	11	8.4	0.47**
Atlantic	23.2	13.9	0.81*	9.6	10.9	0.73*



403

404 **Figure 6 |** 1st mode of basin-scale EOFs of interannual **A**) Chl_{PISCES} and **B**) $Chl_{Svr\text{-}PISCES}$, and their corresponding PCs
405 over 1979-2010 in the **C**) Pacific, **D**) Indian **E**) Atlantic and **F**) Austral Oceans (black and blue lines, respectively). Climate indices
406 are reported in red (right y-axis).

407

The 2nd mode of variability of Chl_{PISCES} is also well reproduced by the SVR. The percent variances are close (**Table 3**) as well as their spatio-temporal variability in the four oceanic basins (**Figure Supp 4**). The high correlations between the first two modes of Chl_{PISCES} *vs*. $Chls_{vr\text{-}PISCES}$ highlight the SVR ability to relatively well reproduce the Chl_{PISCES} low-frequency variability.

4 Application to satellite radiometric observations

4.1 SVR statistical performances and sensitivity tests

In this section, the SVR uses the same physical predictors from NEMO-DFS5.2 as in Section 3, but it is trained on satellite radiometric observations (e.g., $Chl_{OC\text{-}CCI}$). The same procedure is followed (See **Figure Supp 5A-B**). A first validation is performed for 20% of 9% of the full data set and over the 1998-2010 training period showing a high determination coefficient of 0.87 and RMSE of 0.37 between $Chl_{OC\text{-}CCI}$ and $Chls_{vr\text{-}CCI}$ (**Figure Supp 5C**).

As expected, the regression lines of $Chl_{OC\text{-}CCI}$ *vs*. $Chls_{vr\text{-}CCI}$ for each oceanic basin and at global scale are farther away from the 1:1 line than for the synthetic study over the training period, but still remain close (higher slope than 0.8, except in the Austral Ocean; **Figure 7**). The SVR trained on NEMO-DFS5.2 predictors *vs*. satellite Chl is expected to be less efficient than the SVR trained on the coherent NEMO-DFS5.2-PISCES physical-biogeochemical dataset. Some of the biological interactions/processes (such as the diversity of the prey-predator relationships, the complexity of photoacclimation phenomena) are not yet optimally formulated by model equations inducing that Chl derived from numerical modelling is oversimplified compared to the complexity of the real ocean. Not to mention that satellite Chl may itself be partially affected by other components that are not Chl, such as colored dissolved organic matter (CDOM; Morel and Gentili, 2009) and suspended particulate matter (SPM). Phytoplankton can also adjust their intracellular

Chl according to light and nutrient availability (e.g., Laws and Bannister, 1980; Behrenfeld et al., 2015). The induced Chl changes are no longer ascribed to changes in biomass. All these signatures on satellite Chl could explain $Chl_{Svr\text{-}CCI}$ underestimation. Nevertheless, correlations between $Chl_{Svr\text{-}CCI}$ and $Chl_{OC\text{-}CCI}$ remain high over the training time period (0.92, 0.94 and 0.93 for the Indian, Pacific and Atlantic Oceans, respectively, **Figure 7**).

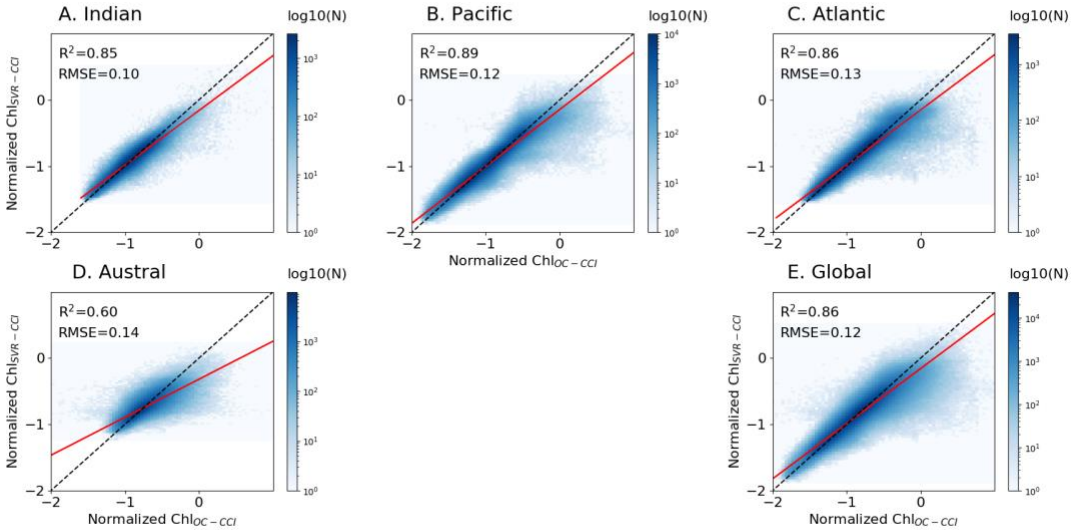


Figure 7 | Same as **Figure 4** for $Chl_{OC\text{-}CCI}$ *vs*. $Chl_{Svr\text{-}CCI}$ normalized monthly anomalies over 1998-2010.

The NRMSE between $Chl_{OC\text{-}CCI}$ *vs*. $Chl_{Svr\text{-}CCI}$ is lower than 20% over most of the global ocean (**Figure 8A**). Correlations higher than 0.9 ($p<0.001$) are evident over large subtropical areas in the Atlantic, Indian and Pacific Oceans as well as in the Equatorial Pacific (**Figure 8B**). Interestingly, the SVR generally does a better job at reconstructing the satellite Chl than the modeled one (**Figure 5A-B** *vs*. **Figure 8**). NRMSE are higher at high latitudes and along the oligotrophic area boundaries, although to a less extent than for Chl_{PISCES}. Because $Chl_{OC\text{-}CCI}$ can only be retrieved under clear sky conditions, gaps in satellite observations (especially during

wintertime) likely alters the SVR learning and could explain such a degradation of $Chls_{vr\text{-}CCI}$ as moving towards high latitudes.

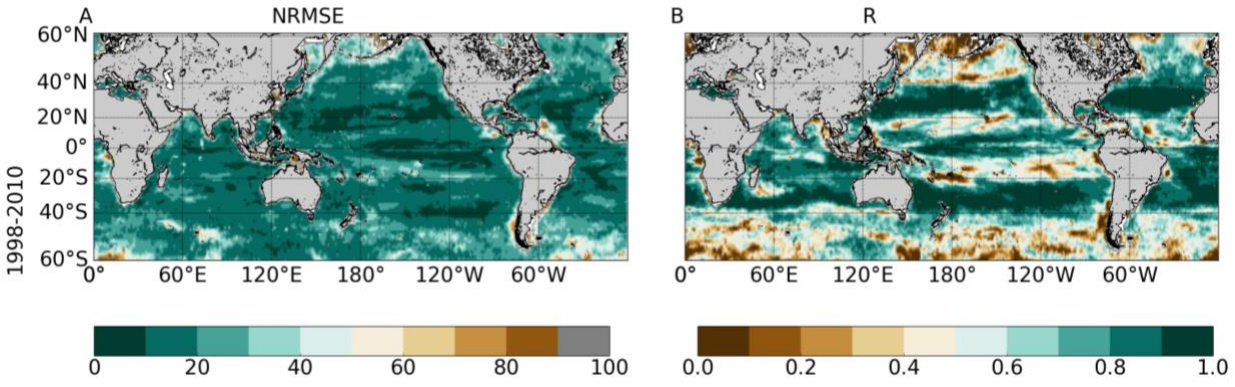


Figure 8 | A) NRMSE (in %) and **B**) correlation between $Chl_{OC\text{-}CCI}$ *vs*. $Chls_{vr\text{-}CCI}$ over 1998-2010 after applying a 5 month-running mean on both time-series. Contours on the NRMSE show the 1998-2010 $Chl_{OC\text{-}CCI}$ time average (every 0.1 $mg.m^{-3}$). Correlations <0.73 and 0.6 are significant with a p-value<0.001 and 0.01, respectively.

4.2 Reconstruction of satellite Chl interannual to decadal variability and trends

The SVR ability to replicate $Chl_{OC\text{-}CCI}$ interannual variability is now investigated over 1998-2010 (**Figure 9)**. In the Pacific Ocean, $Chl_{OC\text{-}CCI}$ and $Chls_{vr\text{-}CCI}$ 1st EOFs are close (**Figure 9A** *vs*. **9B**), their PCs are highly correlated (r=0.89, p<0.001; **Figure 9C**), and their percent variance are similar (**Table 4**). As presented in Section 3.1, this mode of Chl variability can be attributed to ENSO, given their EOFs pattern as well as their PCs highly correlated with the MEI ($r_{OC\text{-}CCI/MEI}$=0.71 and $r_{Svr\text{-}CCI/MEI}$=0.91, with p=0.0015 and p<0.001, respectively). Interestingly, $Chls_{vr\text{-}CCI}$ EOFs are closer to $Chl_{OC\text{-}CCI}$ than Chl_{PISCES} in several areas such as in the north-western Pacific, the south-western Atlantic and the Indian Ocean from Madagascar to the western coast of Australia (**Figures 9A-B** *vs*. **Figure 3B**). Consistently, correlations between $Chl_{OC\text{-}CCI}$ and $Chls_{vr\text{-}CCI}$ PCs in the three basins and for the 1st two modes are higher than between $Chl_{OC\text{-}CCI}$ and Chl_{PISCES} (**Table 4**).

Table 4 | Percent variance explained by the first two modes of the Empirical Orthogonal Function analysis performed on $Chl_{OC\text{-}CCI}$, $Chl_{Svr\text{-}CCI}$ and Chl_{PISCES} for each oceanic basin over 1998-2010. The correlation (r) between $Chl_{OC\text{-}CCI}$ *vs.* $Chl_{Svr\text{-}CCI}$ and between $Chl_{OC\text{-}CCI}$ *vs.* Chl_{PISCES} is also reported with a significant level of P<0.001 (*) and P<0.02 (**).

	1st mode					**2nd mode**				
	% of variance			r $Chl_{OC\text{-}CCI}$ *vs.*		% of variance			r $Chl_{OC\text{-}CCI}$ *vs.*	
	$Chl_{OC\text{-}CCI}$	$Chl_{Svr\text{-}CCI}$	Chl_{PISCES}	$Chl_{Svr\text{-}CCI}$	Chl_{PISCES}	$Chl_{OC\text{-}CCI}$	$Chl_{Svr\text{-}CCI}$	Chl_{PISCES}	$Chl_{Svr\text{-}CCI}$	Chl_{PISCES}
Pacific	16.6	23.7	21.1	0.89*	0.89*	10.7	12.5	13.6	0.81*	0.52**
Indian	16.9	16.6	17.3	0.57**	0.13	11.8	12.2	15.1	0.48	0.36
Atlantic	14	17.9	19.4	0.85*	0.77*	10.7	9.1	12.5	0.82*	0.59**

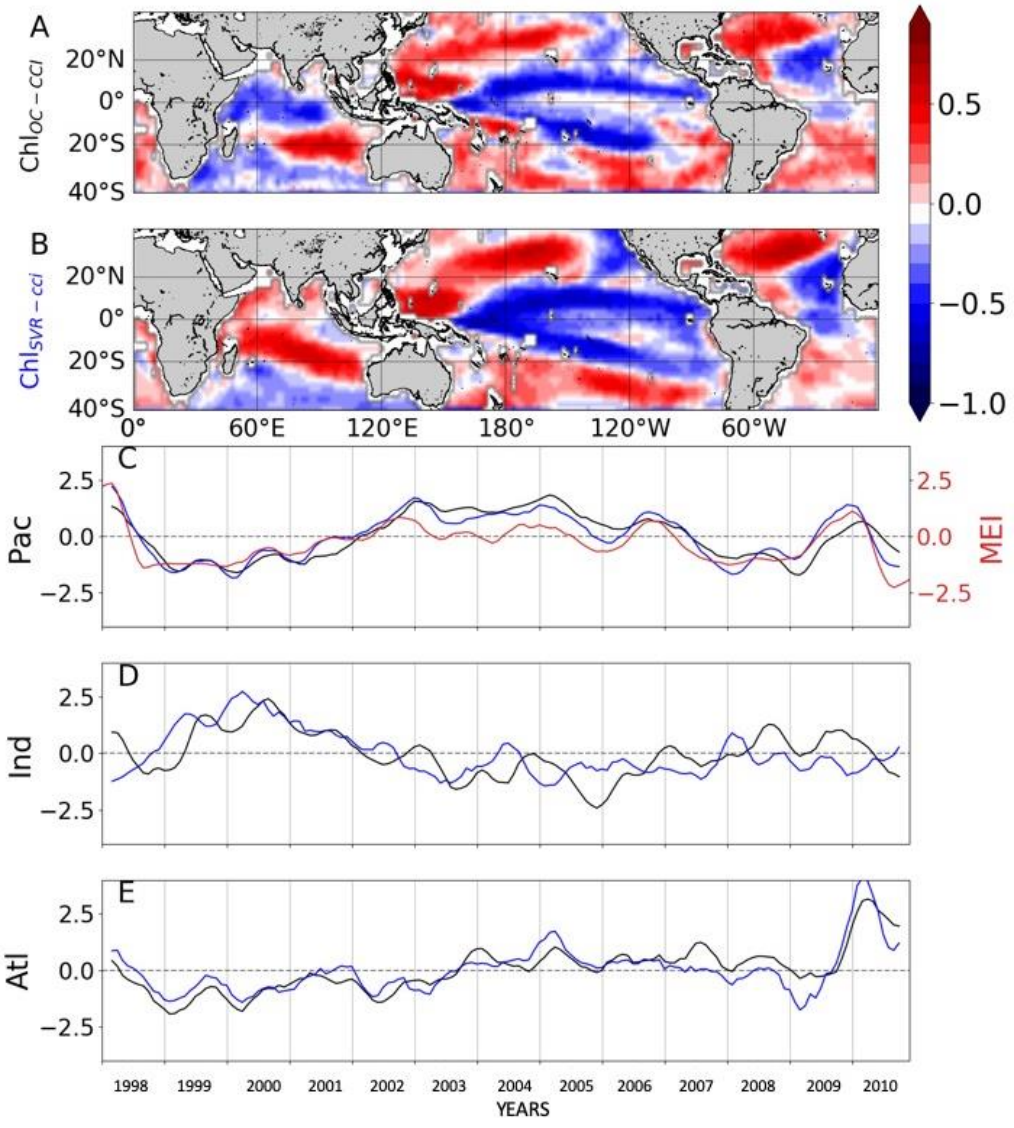


Figure 9: 1st mode of basin-scale EOFs of interannual **A**) $Chl_{OC\text{-}CCI}$ and **B**) $Chl_{Svr\text{-}CCI}$ and their associated PCs over 1998-2010 in the **C**) Pacific, **D**) Indian and **E**) Atlantic Oceans as the black and blue lines, respectively (left y-axis). The climate indices are reported in red on the right y-axis.

$Chl_{OC\text{-}CCI}$ linear trends over 1998-2010 exhibit large areas of increase or decrease (red and blue areas in **Figure 10A**, respectively). Productive regions at high latitudes and along the equatorial and upwelling areas generally exhibit positive $Chl_{OC\text{-}CCI}$ trends, albeit many underlying regional nuances. Contrastingly, trends are generally negative in the center of the gyres. These regional trends are consistent with those extracted from the first 13 years of the SeaWiFS record and discussed by Siegel et al. (2013) (see their Figures 5b and 8b). The negative trends in the oligotrophic gyres were also reported by Signorini et al. (2015) who attributed this behavior to MLD shallowing trends. Surface water density variability induced by changes in temperature and salinity, combined with wind stirring, are effective drivers of vertical mixing, which in turn control the renewal of nutrients from the rich-deep layers toward the euphotic zone. Thus, shallower MLD would decrease nutrient uplift and phytoplankton growth in the oligotrophic areas.

$Chls_{vr\text{-}CCI}$ trends agree qualitatively well with those of $Chl_{OC\text{-}CCI}$ at global scale (**Figures 10B** *vs*. **10A,** respectively). Indeed, decline of $Chls_{vr\text{-}CCI}$ can be observed in the center of the gyres, while outside $Chls_{vr\text{-}CCI}$ generally increases in a similar way to $Chl_{OC\text{-}CCI}$. $Chls_{vr\text{-}CCI}$ accurately captures the largest $Chl_{OC\text{-}CCI}$ increase observed in the Southern Ocean along the Antarctic Circumpolar Current. While Gregg and Casey (2004) reported a substantial negative bias in the SeaWiFS data for this region when compared to *in situ* observations, which could hamper the reliability of satellite trends discussed in this area (e.g. Siegel et al., 2013), the SVR remains able to reproduce the positive observed trend. Despite qualitative spatial agreements, it is noteworthy

that the SVR underestimates by half the magnitude of the satellite trend (see scales in **Figure 10A** *vs*. **10B**).

Interestingly, trends in Chl_{PISCES} generally differs from $Chl_{OC\text{-}CCI}$ (**Figure 10C**). This is striking for the North Pacific and Atlantic high latitudes, but also in the equatorial Atlantic and Arabian Sea with opposite trends when compared with $Chl_{OC\text{-}CCI}$ and $Chls_{vr\text{-}CCI}$, and in a more mitigated manner in the Austral Ocean.

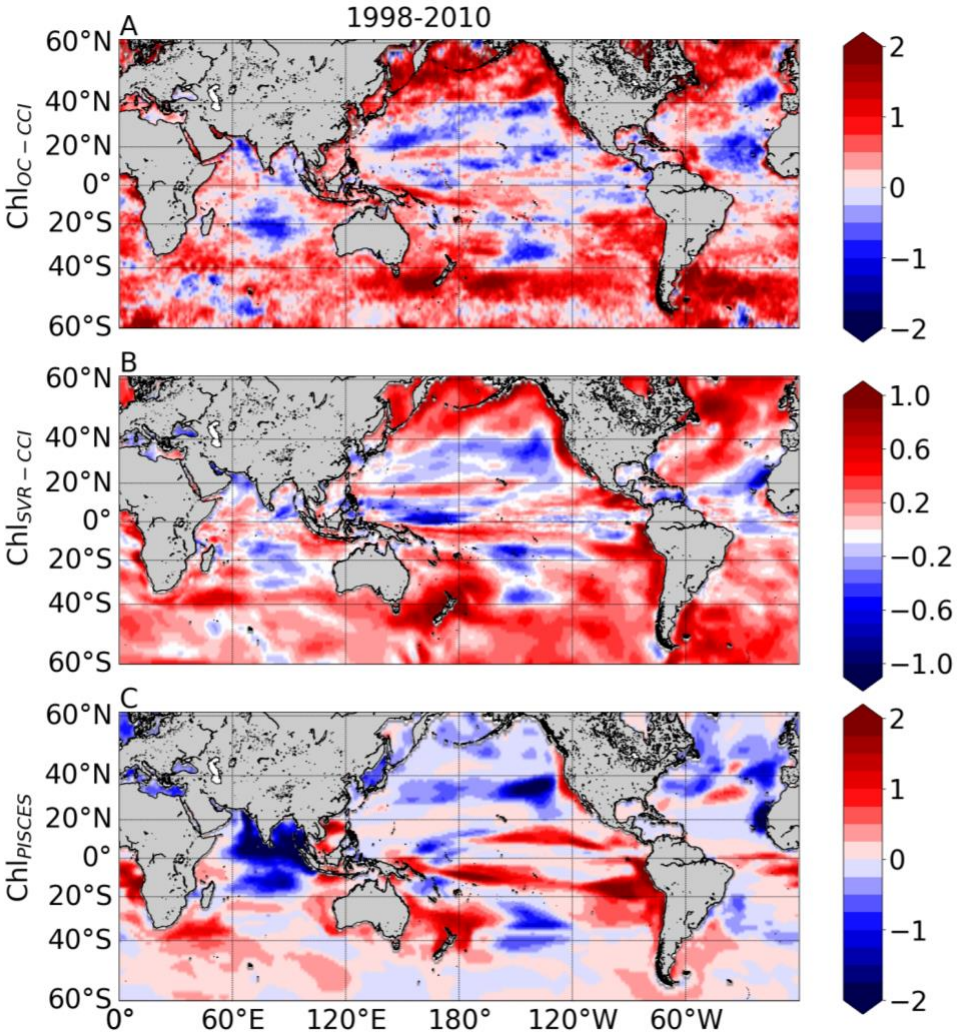


Figure 10 | Linear trends (in % year $^{-1}$) calculated over 1998-2010 from the monthly **A**) ln($Chl_{OC\text{-}CCI}$), **B**) ln($Chls_{vr\text{-}CCI}$), **C**) ln(Chl_{PISCES}). Note that the scale is divided by 2 for ln($Chls_{vr\text{-}CCI}$).

$Chls_{vr\text{-}CCI}$ is also compared with the only historical consistent dataset built by Antoine et al. (2005) who reanalyzed ocean color time series from CZCS (1979–1983) and SeaWiFS (1998–2002). A 22% global mean increase of Chl between the two era was reported. It was mainly due to large increases in the intertropical areas and to a lesser extent in higher latitudes, while oligotrophic gyres displayed declining concentrations (**Figure 11A**). SST from the SODA reanalysis was used as a proxy of ocean stratification and opposite Chl and SST changes over 60% of the ocean between 50°S and 50°N was reported (light blue and yellow in **Figure 11B,** adapted from Martinez et al., 2009). This inverse relationship was used to hypothesized that multidecadal changes in global phytoplankton abundances were related to basin-scale oscillations of the ocean dynamics. Briefly, SST changes were related to a regime shift of the PDO (although the use of the basin-scale IPO would have been more appropriate) from a warm to a cold phase in the Pacific and Indian Oceans leading to an increase of Chl, and inversely in the Atlantic Ocean with a regime shift from a cold to a warm phase of the AMO leading to a Chl decrease.

Observed Chl changes over the last decades are accurately reproduced by $Chls_{vr\text{-}CCI}$, including a Chl increase in the equatorial Pacific and the southern tropical Indian Oceans, as well as a Chl decline in both the Atlantic and Pacific oligotrophic gyres (**Figure 11C**). However, the magnitude of the SVR reconstructed Chl is underestimated (note that the Chl ratio is multiplied by 2 in **Figure 11C** to allow the comparison with **Figure 11A**). On average, the inverse relationship between $Chls_{vr\text{-}CCI}$ and SST_{NEMO} (**Figure Supp 4**) occurs over 69.4% of the global ocean between 50°S and 50°N in a similar way to that reported by Martinez et al. (2009), especially in the Pacific Ocean (see **Figure 11D** *vs*. **Figure 11B**). In the Indian Ocean, although Chl mainly increases in both studies, it is here associated with a SST decrease. Interestingly, this inverse Chl-SST relationship in the Indian Ocean (yellow area in **Figure 11D**) was reported in Behrenfeld et al.

(2006) over the SeaWiFS era, suggesting that the SST dataset used in Martinez et al. (2009) may have decadal discrepancies for this region.

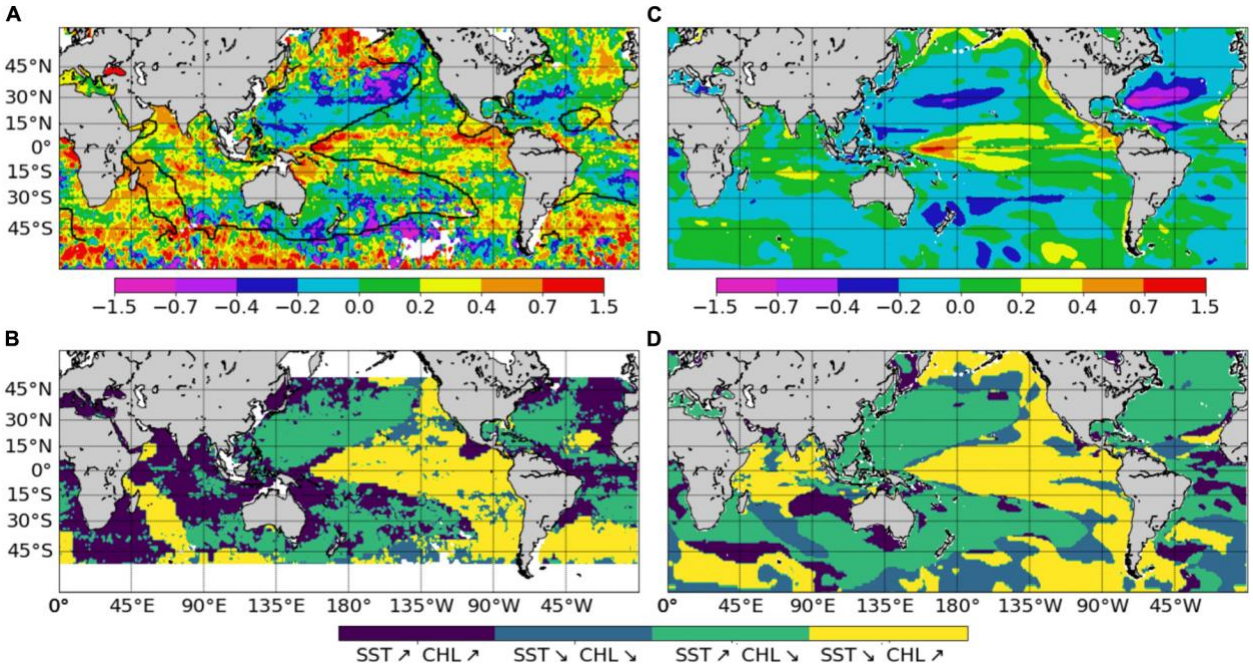


Figure 11 | Chl change from the CZCS (1979–1983) to the SeaWiFS (1998–2002) era, expressed as the logarithm of the ratio of the average values over the two time periods **A**) from satellite Chl adapted from Antoine et al. (2005), **C**) from $Chl_{Svr\text{-}CCI}$. Note that this ratio is multiplied by 2 to fit the same color bar as in **A**. Maps of areas with concomitant parallel or opposite changes of Chl and SST **B**) from Chl satellite and SST from the SODA reanalysis adapted from Martinez et al. (2009) and **D**) from $Chl_{Svr\text{-}CCI}$ and SST_{NEMO}. The respective SST zero differences are shown on the maps as a thick black curve.

In their study, Martinez et al. (2009) analyzed two 5-year time periods apart each other by 15 years. They suggested that averaging observations separately over the two time-periods may have dampen the effect of interannual variability and reveal the decadal one. Most of the changes observed between the time periods covered by the two satellites are here confirmed based on the reconstructed $Chl_{Svr\text{-}CCI}$. However, the continuous 30-year time series of $Chl_{Svr\text{-}CCI}$ provides new insights on the observed regime shifts (**Figure 12**). In the Pacific Ocean, the 1st EOF of $Chl_{Svr\text{-}CCI}$ (**Figure 12A**) is close to the Chl spatial patterns obtained from the CZCS to SeaWiFS era (**Figure 11C)** and the PC remains highly correlated with the IPO over 1979-2010 ($r=0.94$ with $p<0.001$, **Figure 12B**). The Chl increase in the Indian Ocean, north-east of Madagascar toward the west

543 coast of Australia, between the 1980's and the 2000's also appears on the $Chls_{vr\text{-}CCI}$ EOF. These
544 temporal changes might also be related to the IPO variability (correlation between the IPO index
545 and $Chls_{vr\text{-}CCI}$ PC =0.6, $p<0.001$; **Figure 12C**).

546 In the Atlantic Ocean, CZCS-SeaWiFS Chl and $Chls_{vr\text{-}CCI}$ 1st EOFs also share some
547 similarities, including a decrease of Chl in the subtropical gyres and an increase in the
548 equatorial/tropical regions. The associated PC (**Figure 12E**), exhibits a shift between 1979-1983
549 and 1998-2002 consistently with Figure 2C of Martinez et al. (2009). In this latter study, this
550 change is attributed to a regime shift of the AMO. However, the AMO index is not correlated with
551 the 1st $Chls_{vr\text{-}CCI}$ PC ($r=0.03$, $p=0.43$) but rather with the 2nd mode ($r = 0.43$ with $p=0.003$), likely
552 explaining the spatial discrepancies in **Figures 11A** *vs*. **11C**. Although the detailed analysis of Chl
553 decadal variability is beyond the scope of the present study, these initial findings underscore the
554 importance of continuous time series at regional/global scales to combine spatial and temporal
555 information's and properly investigate Chl long-term variability.

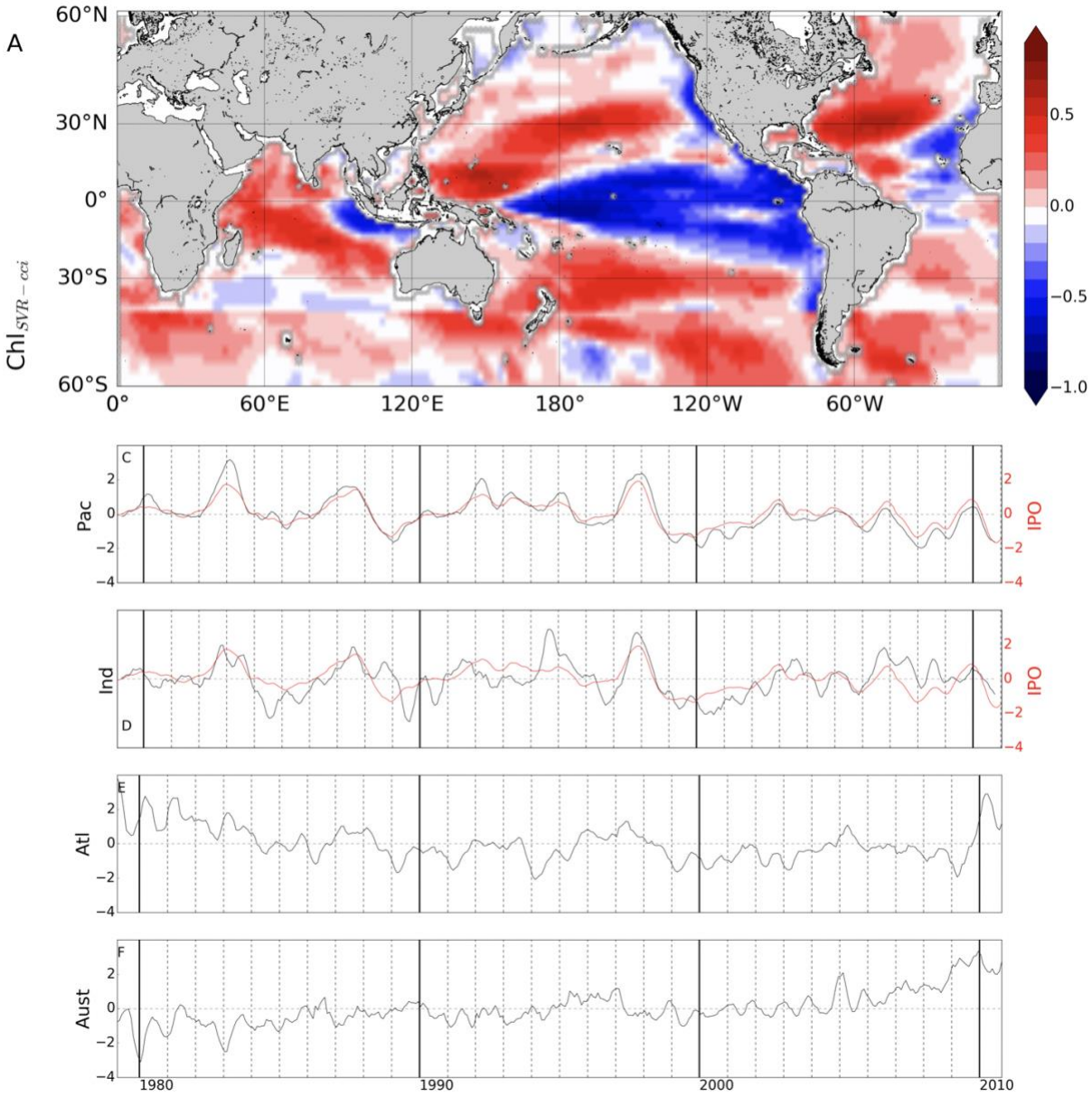


Figure 12 | A) 1st mode of basin-scale EOFs of interannual $Chl_{Svr\text{-}CCI}$ over 1979-2010 and their corresponding PCs in the **B**) Pacific (23.2% of the total variance), **C**) Indian (15.2% of the total variance), **D**) Atlantic (13.5% of the total variance) and **E**) Austral Oceans (11.4% of the total variance). IPO is reported in red (right y-axis).

5 Summary and conclusions

In this paper, we assess the efficiency of a machine learning statistical approach based on support vector regression (SVR) to reconstruct surface Chl from oceanic and atmospheric variables. We first apply this strategy on self-consistent global dataset gathering physical predictors and Chl data simulated by a coupled physical-biogeochemical model simulation. Our results indicate that this nonlinear method accurately hindcasts interannual-to-decadal variations of the phytoplankton biomass simulated at global scale by the model, except at the boundaries of the subtropical gyres where the strong top-down control of zooplankton grazing in the numerical model is not accounted for by the SVR. Likewise, this statistical approach cannot yet reproduce Chl variability induced by nutrient inputs that are not directly related to our selected physical predictors, such as atmospheric iron deposit.

The SVR was then trained on satellite Chl observations. It accurately reproduces observed interannual Chl variations in most regions, including El Niño signature in the tropical Pacific and Indian Ocean as well as the main modes of Atlantic Chl variability. Despite an amplitude underestimation by half, it also accurately captures spatial patterns of Chl trends over the satellite period, with a Chl increase in most extratropical regions and a Chl decrease in the center of the subtropical gyres, as well as their changes between the CZCS and SeaWiFS era, while the model simulation generally fails to capture these long-term trends. Interestingly, while Chl_{PISCES} magnitude is closer to $Chl_{OC\text{-}CCI}$ than $Chl_{Svr\text{-}CCI}$, interannual variability and spatial trends of Chl_{PISCES} are farther than $Chl_{Svr\text{-}CCI}$ to $Chl_{OC\text{-}CCI}$. Equations representing the processes that govern the evolution of biogeochemical variables in a biogeochemical model are obviously less complex than the ones at play in the real ocean. We thus anticipated the modeled Chl to be easier to reconstruct than the satellite one. Additional complications were also expected through the

584 reconstruction of satellite Chl from the model oceanic and atmospheric predictors, which may be
585 less realistic than physical parameters derived from satellite measurements. As a consequence, the
586 SVR is indeed slightly less efficient at reproducing the major satellite Chl patterns compared to
587 the model ones but is surprisingly more efficient at capturing observed Chl temporal variations.
588 This results in a normalized root mean square error generally weaker when reconstructing satellite
589 data compared to the model one, although the predictors used are identical.

590 Machine learning techniques are powerful tools to statistically model non-linear processes.
591 They require a significant amount of data to be trained and are well-suited to analyze remote
592 sensing data. While several attempts have been made over the last decade to retrieve oceanic Chl
593 content (Kwiatkowska and Fargion, 2003; Zhan et al., 2003; Camps-Valls et al., 2009; Jouini et
594 al., 2013; Blix and Eltoft, 2018), the present work is one of the first attempt to use such machine
595 learning techniques to reconstruct past time series of phytoplankton biomass at global scale. To
596 our knowledge only Schollaert et al. (2017) tried to reconstruct the Chl multi-decadal variability
597 in the tropical Pacific using a canonical correlation analysis built only from SST and SSH. Our
598 SVR approach leads to higher correlations between reconstructed and satellite Chl in the tropical
599 Pacific, highlighting the strength of such non-linear machine-learning methods with multiple
600 predictors. These results emphasize deep learning approaches as promising tools to reconstruct
601 multidecadal Chl time series in the global ocean, based on the knowledge of physical conditions.
602 The successful use of surface variables only in reproducing Chl variability which is influenced by
603 3D-processes is here clearly noteworthy, and investigation of variable importance in the Chl
604 reconstruction will deserve some future insights.

605 An obvious short-term perspective of the current study is to train a wider range of such
606 statistical models with physical predictors derived from surface satellite observations but also

607 observations within the water column which could be derived from Argo data (i.e., mixed layer
608 and thermocline depth). Including complementary variables such as satellite particulate
609 backscattering coefficient (as a proxy of the Particulate Organic Carbon) in the
610 training/reconstruction process should also be considered. It would allow to investigate the extent
611 to which the Chl variability reflects changes in phytoplankton biomass *vs*. cellular changes in
612 response to light (e.g., Siegel et al., 2005; Westberry et al., 2008; Behrenfeld et al., 2015). The use
613 of longitude and latitude as predictors may limit the ability to capture long-term trends in the
614 evolution of the biogeochemical province boundaries, such as the expansion of the oligotrophic
615 areas (Irwin and Oliver, 2009; Polovina et al., 2008; Staten et al. 2018). Thus, exploring deep
616 learning schemes which may not explicitly depend on longitude and latitude, especially
617 convolutional representations (Lecun et al., 2015), are particularly appealing. Further efforts need
618 also to be dedicated to alleviate the issue of the underestimation of the long-term Chl trends. For
619 instance, it would be noteworthy to investigate secular trends such as the 30% Chl decrease
620 reported at global scale over the last century by Boyce et al. (2010), which remains largely debated
621 (Mackas, 2011; McQuatters-Gollop et al., 2011; Rykaczewski and Dunne, 2011).

622 Whatever the methodology used (i.e. numerical models, satellite or *in situ* observations),
623 they all have both advantages and drawbacks. *In situ* observations are considered as ground truth
624 (with some errors/uncertainties depending for instance on the field measurement protocols), but
625 are heterogeneous in time and space. Satellite Chl data provide a spatio-temporal synoptic view
626 but they have their own measurements issues and uncertainties (e.g., sensor radiometric and
627 spectral properties, atmospheric corrections, water constituents and their optical properties) and
628 are limited to 20 years in their record length. Biogeochemical models are useful tools to (i)
629 interpolate or extrapolate in time and space biogeochemical tracers such as Chl and to (ii)

investigate complex three-dimensional processes responsible for their variations. However, those models suffer from biases and are farther from *in situ* data than satellite observations. They are also not straightforward to run and require large computing resources. Thus, machine learning statistical schemes could be seen as a complementary tool to the "interpolate/extrapolate" use of biogeochemical models in providing a long-term synoptic surface view built from observations (being aware of the uncertainties associated with the variables used in the training schemes). Such methods, applied on observations only, will then provide an independent tool that may either question or enforce conclusions drawn from model simulations. Comparison between both methods and observations will also help to improve biogeochemical models with acute quantification of model biases and identification of the most meaningful predictors that may point to missing processes in biogeochemical models. As a conclusion, machine learning is a versatile tool that, associated with biogeochemical models and observations, may greatly enhance our view of global biogeochemistry.

Acknowledgments, Samples, and Data

This work was supported by CNES under contract n°160515/00 within the framework of the PhytoDev project. We thank the two reviewers who helped to improve this manuscript.

Authors contribution

EM led the project, analyzed results and wrote the 1st draft of the manuscript. TG provided physical model outputs. TG, ML provided support in the analysis and the writing of the manuscript. CF processed the machine learning approach with support from R. Sauzede. R. Fablet provided feedbacks on the statistical approach. All authors contributed to the development of the manuscript and provided feedbacks throughout its many stages of preparation.

654

655 **References**

656 Aguilar-Martinez, S., and Hsieh, W. W. (2009). Forecasts of tropical pacific sea surface
657 temperatures by neural networks and support vector regression. *Int. J. Oceanog*., 13,
658 doi:10.1155/2009/167239.

659 Antoine, D., Morel, A. , Gordon, H. R., Banzon, V. F., and Evans, R. H. (2005). Bridging ocean
660 color observations of the 1980s and 2000s in search of long-term trends. *J. Geophys. Res.*
661 *Oceans* 110, C06009.

662 Aumont, O., and Bopp, L. (2006). Globalizing results from ocean in situ iron fertilization
663 studies. *Global Biogeochemical Cycles*, 20, GB2017, doi:10.1029/2005GB002591, 2006.

664 Aumont, O., Ethé, C., Tagliabue, A., Bopp, L., and Gehlen, M. (2015). PISCES-v2: an ocean
665 biogeochemical model for carbon and ecosystem studies. *Geosci. Model Dev*., 8, 2465–2513,
666 2015, doi:10.5194/gmd-8-2465-2015.

667 Ayata, S. D., Lévy, M., Aumont, O., Sciandra, A., Sainte-Marie, J., Tagliabue, A., and Bernard,
668 O. (2013). Phytoplankton growth formulation in marine ecosystem models: should we take
669 into account photoacclimation and variable stoichiometry in oligotrophic areas? *J. Marine*
670 *Syst.*, 125, 29–40, doi:10.1016/j.jmarsys.2012.12.010, 2013. 1433.

671 Banse, K., and English, D.C. (2000). Geographical differences in seasonality of CZCS-derived
672 phytoplankton pigment in the Arabian Sea for 1978-1986. *Deep-Sea Res. II*, 47: 1623-1677.

673 Beaulieu, C., Henson, S. A., Sarmiento, J. L., Dunne, J. P., Doney, S. C., Rykaczewski, R. R., and
674 Bopp, L. (2013). Factors challenging our ability to detect long-term trends in ocean
675 chlorophyll. *Biogeosciences*, 10(4), 2711-2724.

Behrenfeld, M. J., Randerson, J. T., McClain, C. R., Feldman, G. C., Los, S. O., Tucker, C. J., et al. (2001). Biospheric primary production during an ENSO transition. *Science*, 291(5513), 2594-2597.

Behrenfeld, M. J., O'Malley, R. T., Siegel, D. A., McClain, C. R., Sarmiento, J. L. et al. (2006). Climate driven trends in contemporary ocean productivity. *Nature*, 444(7120), 752–755.

Behrenfeld, M. J., O'Malley, R. T., Boss, E. S., Westberry, T. K., Graff, J. R., Halsey, K. H., et al. (2015). Revaluating ocean warming impacts on global phytoplankton. *Nat. Clim. Chang*. 6, 323–330. doi:10.1038/nclimate2838.

Belo Couto, A., Brotas, V., Mélin, F., Groom, S., and Sathyendranath, S. (2016). Inter-comparison of OC-CCI chlorophyll-a estimates with precursor data sets. *International Journal of Remote Sensing*, 37(18), 4337-4355.

Blix, K., and Eltoft, T. (2018). Machine Learning Automatic Model Selection Algorithm for Oceanic Chlorophyll-a Content Retrieval. *Remote Sensing*, 10(5), 775.

Bopp, L., Aumont, O., Cadule, P., Alvain, S., and Gehlen, M. (2005). Response of diatoms distribution to global warming and potential implications: A global model study. *Geophys. Res. Lett*., 32(19).

Boyce, D. G., Lewis, M. R., and Worm, B. (2010). Global phytoplankton decline over the past century. *Nature*, 466(7306), 591-596.

Campbell, J. W., and Aarup, T. (1992). New production in the North Atlantic derived from seasonal patterns of surface chlorophyll. *Deep-Sea Res. Prt A.*, 39(10), 1669-1694.

Camps-Valls, G., Muñoz-Marí, J. L., Gómez-Chova, K. R., and Calpe-Maravilla, J. (2009). Biophysical Parameter Estimation With a Semisupervised Support Vector Machine. *IEEE Geoscience and Remote Sensing Letters*, 6 (2), 248 – 252.

699 Chavez, F. P., Strutton, P. G., Friederich, G. E., Feely, R. A., Feldman, G. C., Foley, D. G., and
700 McPhaden, M. J. (1999). Biological and chemical response of the equatorial Pacific Ocean to
701 the 1997-98 El Niño. *Science*, 286(5447), 2126-2131.
702 Chavez, F. P., Messié, M., and Pennington, J. T. (2011). Marine primary production in relation to
703 climate variability and change. *Annu. Rev. Mar. Sci*., 3, 227–260,
704 doi:10.1146/annurev.marine.010908.163917.
705 Currie, J. C., Lengaigne, M., Vialard, J., Kaplan, D., Aumont, O., Naqvi, S. W. A., and Maury, O.
706 (2013). Indian Ocean dipole and El Nino/southern oscillation impacts on regional chlorophyll
707 anomalies in the Indian Ocean. *Biogeosciences*, 10(10), 6677-6698.
708 Dandonneau, Y., Deschamps, P. Y., Nicolas, J. M., Loisel, H., Blanchot, J., Montel, Y., et al.
709 (2004). Seasonal and interannual variability of ocean color and composition of phytoplankton
710 communities in the North Atlantic, equatorial Pacific and South Pacific. *Deep-Sea Res. Prt II*,
711 51(1-3), 303-318.
712 Descloux, E., Mangeas, M., Menkes, C. E., Lengaigne, M., Leroy, A., Tehei, T., et al. (2012).
713 Climate-Based Models for Understanding and Forecasting Dengue Epidemics. *PLoS*
714 *Neglected Tropical Diseases*, 6(2), e1470. doi:10.1371/journal.pntd.0001470.
715 Dierssen, H. M., and Smith, R. C. (2000). Bio-optical properties and remote sensing ocean color
716 algorithms for Antarctic Peninsula waters. *J. Geophys. Res. Oceans*, 105, 26301–26312,
717 doi:10.1029/1999JC000296, 2000. 1434, 1497.
718 D'Ortenzio, F., Antoine, D., Martinez, E., and Ribera d'Alcalà, M. (2012). Phenological changes
719 of oceanic phytoplankton in the 1980s and 2000s as revealed by ocean-color remote-sensing
720 observations. *Global Biogeochem. Cycles*, Vol. 26, GB4003.

721 Dussin, R., Barnier, B., Brodeau, L., and Molines, J. M. (2014). The making of Drakkar forcing
722 set DFS5, DRAKKAR/MyOcean Rep.
723 Dutheil, C., Aumont, O., Gorguès, T., Lorrain, A., Bonnet, S., Rodier, M., et al. (2018). Modelling
724 the processes driving Trichodesmium sp. spatial distribution and biogeochemical impact in
725 the tropical Pacific Ocean. *Biogeosciences Discussions*, 1–34. https://doi.org/10.5194/bg-
726 2017-559.
727 Dutkiewicz, S., Follows, M., Marshall, J., and Gregg, W.W (2001). Interannual variability of
728 phytoplankton abundances in the North Atlantic. *Deep Sea Res. II*, Volume 48, Pages 2323-
729 2344.
730 Dutkiewicz, S., Follows, M. J., and Parekh, P. (2005). Interactions of the iron and phosphorus
731 cycles: a three-dimensional model study. *Global Biogeochem. Cycles*, 19, GB1021,
732 doi:10.1029/2004GB002342, 2005. 1412, 1430, 1433.
733 Elbisy, M. S. (2015). Sea Wave Parameters Prediction by Support Vector Machine Using a Genetic
734 Algorithm. *J. Coastal Res*., 31, 4, 892 –899.
735 Emery, W., and Thomson, R. (1997). Data Analysis in Physical Oceanography , 634 pp.,
736 Pergamon, New York.
737 Enfield, D. B., Mestas Nunez, A. M., and Trimble, P. J. (2001). The Atlantic multidecadal
738 oscillation and its relation to rainfall and river flows in thecontinental U.S. *Geophys. Res.*
739 *Lett*., 28, 2077–2080.
740 Feng, J., Durant, J. M., Stige, L. C., Hessen, D. O., Hjermann, D. Ø., Zhu, L., et al. (2015).
741 Contrasting correlation patterns between environmental factors and chlorophyll levels in the
742 global ocean. *Global Biogeochem. Cycles*, 29, doi:10.1002/2015GB005216.

743 Garcia, H. E., Locarnini, R. A., Boyer, T. P., and Antonov, J. I. (2006). World Ocean Atlas 2005,
744 Volume 4: Nutrients (phosphate, nitrate, silicate). S. Levitus, Ed. NOAA Atlas NESDIS 64,
745 U.S. Government Printing Office, Washington, D.C., 396 pp.
746 Gehlen, M., Bopp, L., Emprin, N., Aumont, O., Heinze, C., and Ragueneau, O. (2006).
747 Reconciling surface ocean productivity, export fluxes and sediment composition in a global
748 biogeochemical ocean model. *Biogeosciences, European Geosciences Union*, 2006, 3 (4),
749 pp.521-537. hal-00330316
750 Gnanadesikan, A., Slater, R. J., Gruber, N., and Sarmiento, J. L. (2002). Oceanic vertical exchange
751 and new production: a comparison between models and observations. *Deep-Sea Res. Pt. II*,
752 49, 363–401, 2002. 1433.
753 Gregg, W. W., and Rousseaux, C. S. (2014). Decadal trends in global pelagic ocean chlorophyll:
754 A new assessment integrating multiple satellites, in situ data, and models. *J. Geophys. Res.*
755 *Oceans*, 119(9), 5921-5933.
756 Gregg, W. W., and Casey, N. W. (2004). Global and regional evaluation of the SeaWiFS
757 chlorophyll data set. *Remote Sensing of Environment*, 93(4), 463-479.
758 Gregg, W. W., and Conkright, M. E. (2002). Decadal changes in global ocean chlorophyll.
759 *Geophys. Res. Lett.*, 29(15), 20-1.
760 Henson, S. A., Dunne, J. P., and Sarmiento, J. L. (2009a). Decadal variability in North Atlantic
761 phytoplankton blooms. *J. Geophys. Res. Oceans*, 114, C04013.
762 Henson, S. A., Raitsos, D., Dunne, J. P., and McQuatters-Gollop, A. (2009b). Decadal variability
763 in biogeochemical models: Comparison with a 50-year ocean colour dataset. *Geophys. Res.*
764 *Lett.*,, 36, L21061.

765 Henson, S. A., Beaulieu, C. and Lampitt, R. (2016). Observing climate change trends in ocean
766 biogeochemistry: when and where. *Glob Change Biol*, 22: 1561-1571.
767 doi:10.1111/gcb.13152.
768 Hood, R. R., Kohler, K. E., McCreary, J. P., and Smith, S. L. (2003). A four- dimensional
769 validation of a coupled physical-biological model of the Arabian Sea. *Deep-Sea Res. Pt. II*,
770 50, 2917–2945.
771 Hovis, W. A., Clark, D. K., Anderson, F., Austin, R. W., Wilson, W. H., Baker, E. T., et al. (1980).
772 Nimbus-7 Coastal Zone Color Scanner: system description and initial
773 imagery. *Science*, 210(4465), 60-63.
774 Hu, S., Liu, H., Zhao, W., Shi, T., Hu, Z., Li, Q., and Wu, G. (2018). Comparison of machine
775 learning techniques in inferring phytoplankton size classes. *Remote Sensing*, 10(3), 191.
776 Huang, B., Thorne, P. W., Banzon, V. F., Boyer, T., Chepurin, G., Lawrimore, et al. (2017).
777 Extended reconstructed sea surface temperature, version 5 (ERSSTv5): upgrades, validations,
778 and intercomparisons. *Journal of Climate*, *30*(20), 8179-8205.
779 Irwin, A. J., and Oliver, M. J. (2009). Are ocean deserts getting larger? *Geophys. Res. Lett.*, 36,
780 L18609, doi:10.1029/2009GL039883.
781 Jickells, T. D., An, Z. S., Andersen, K. K., Baker, A. R., Berga- metti, G., Brooks, N., et al. (2005).
782 Global iron connections between desert dust, ocean biogeochemistry, and climate. *Nature*,
783 308, 67–71, doi:10.1126/science.1105959.
784 Jouini, M., Lévy, M., Crépon, M., and Thiria, S. (2013). Reconstruction of satellite chlorophyll
785 images under heavy cloud coverage using a neural classification method. *Remote Sens.*
786 *Environ.*, 131, 232-246.

Kahru, M. and Mitchell, B. G. (2010). Blending of ocean colour algorithms applied to the Southern Ocean. *Remote Sensing Letters*, 1, 119–124. 1434.

Kahru, M., Gille, S. T., Murtugudde, R., Strutton, P. G., Manzano-Sarabia, M., Wang, H., and Mitchell, B. G. (2010). Global correlations between winds and ocean chlorophyll. *J. Geophys. Res. Oceans* 115, C12040, doi:10.1029/2010JC006500.

Keerthi, M.G., Lengaigne, M., Levy, M., Vialard, J., and de Boyer Montegut, C. (2017). Physical control of interannual variations of the winter chlorophyll bloom in the northern Arabian Sea. *Biogeosciences*, *14*(15), 3615, doi:10.5194/bg-14-3615-2017.

Kim, Y. H., Im, J., Ha, H. K., Choi, J. K., and Ha, S. (2014). Machine learning approaches to coastal water quality monitoring using GOCI satellite data. *GIScience & Remote Sensing*, 51(2), 158-174.

Kwiatkowska, E.J., and Fargion, G.S. (2003). Application of Machine-Learning Techniques Toward the Creation of a Consistent and Calibrated Global Chlorophyll Concentration Baseline Dataset Using Remotely Sensed Ocean Color Data. *IEEE Transactions on Geoscience and Remote Sensing*, 41 (12), 2844 – 2860.

Laufkötter, C., Vogt, M., Gruber, N., Aita-Noguchi, M., Aumont, O., Bopp, L., et al. (2015). Drivers and uncertainties of future global marine primary production in marine ecosystem models. *Biogeosciences*, 12, 6955-6984, https://doi.org/10.5194/bg-12-6955-2015.

Launois, T., Belviso, S., Bopp, L., Fichot, C. G., and Peylin, P. (2015). A new model for the global biogeochemical cycle of carbonyl sulfide–Part 1: Assessment of direct marine emissions with an oceanic general circulation and biogeochemistry model. *Atmospheric Chemistry and Physics*, 15(5), 2295-2312.

809 Laws, E. A., and Bannister, T. T. (1980). Nutrient-and light-limited growth of Thalassiosira
810 fluviatilis in continuous culture, with implications for phytoplankton growth in the ocean.
811 *Limnology and Oceanography*, 25(3), 457-473.
812 LeCun, Y., Bengio, Y. and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444,
813 https://doi.org/10.1038/nature14539.
814 Lengaigne, M., Menkes, C., Aumont, O., Gorgues, T., Bopp, L., André, J.-M., and Madec, G.
815 (2007). Influence of the oceanic biology on the tropical Pacific climate in a Coupled General
816 Circulation Model. *Climate Dynamics*, 28: 503-516, doi:10.1007/s00382-006-0200-2.
817 Lewandowska, A. M., Hillebrand, H., Lengfellner, K., and Sommer, U. (2014). Temperature
818 effects on phytoplankton diversity—The zooplankton link. *J. Sea Res*., 85, 359-364.
819 Longhurst, A., Sathyendranath, S., Platt, T., and Caverhill, C. (1995). An estimate of global
820 primary production in the ocean from satellite radiometer data. *J. Plankton Res*., 17(6), 1245-
821 1271.
822 Mackas, D. L. (2011). Does blending of chlorophyll data bias temporal trend? *Nature*, 472(7342),
823 E4-E5.
824 Madec, G. (2008). NEMO reference manual, ocean dynamics component: NEMO-OPA.
825 Preliminary version. Note du Pole de modélisation, Institut Pierre-Simon Laplace (IPSL),
826 France, (27), 1288-161.
827 Mantua, N.J., Hare, S.R., Zhang, Y., Wallace, J.M., and Francis, R.C. (1997). A Pacific
828 Interdecadal Climate Oscillation with Impacts on Salmon Production. *B. Amer. Meteo. Soc.*,
829 78, 1069.
830 Martinez, E., Antoine, D., D'Ortenzio, F., and Gentili, B. (2009). Climate-driven basin-scale
831 decadal oscillations of oceanic phytoplankton. *Science*, 36, 1253-1256.

832 Martinez, E., Antoine, D., D'Ortenzio, F., and de Boyer Montégut, C. (2011). Phytoplankton
833 spring and fall blooms in the North Atlantic in the 1980s and 2000s. *J. Geophys. Res. Oceans*,
834 116, C11029.
835 Martinez, E., Raitsos, D., and Antoine, D. (2016). Warmer, deeper and greener mixed layers in the
836 north Atlantic subpolar gyre over the last 50 years. *Global Change Biology*, 22, 604–612, doi:
837 10.1111/gcb.13100.
838 McClain, C. R., Christian, J. R., Signorini, S. R., Lewis, M. R., Asanuma, I., Turk, D., and Dupouy-
839 Douchement, C. (2002). Satellite ocean-color observations of the tropical Pacific
840 Ocean. *Deep-Sea Res. Pt. II*: Topical Studies in Oceanography, 49(13-14), 2533-2560.
841 McClain, C. R., Feldman, G., and Hooker, S. (2004). An overview of the SeaWiFS project and
842 strategies for producing a climate research quality global ocean bio-optical time series, *Deep*
843 *Sea Res. Part II*, 51, 5–42, doi:10.1016/j.dsr2.2003.11.001.
844 McQuatters-Gollop, A., Reid, P. C., Edwards, M., Burkill, P. H., Castellani, C., Batten, S., et al.
845 (2011). Is there a decline in marine phytoplankton? *Nature*, 472(7342), E6-E7.
846 Messié, M., and Chavez, F. P. (2012). A global analysis of ENSO synchrony: The oceans'
847 biological response to physical forcing. *J. Geophys. Res. Oceans*, 117, C09001,
848 doi:10.1029/2012JC007938.
849 Messié, M., and Chavez, F. P. (2015). Seasonal regulation of primary production in eastern
850 boundary upwelling systems. *Progress in Oceanography*, 134, 1-18.
851 Morel, A., and Gentili, B. (2009). The dissolved yellow substance and the shades of blue in the
852 Mediterranean Sea. *Biogeosciences*, 6(11).

853 Murtugudde, R. G., Signorini, S. R., Christian, J. R., Busalacchi, A. J., McClain, C. R., and Picaut,
854 J. (1999). Ocean color variability of the tropical Indo-Pacific basin observed by SeaWiFS
855 during 1997– 1998. *J. Geophys. Res. Oceans*, 104, 18351–18366.
856 Neetu S., Lengaigne, M., M. Mangeas, J. Vialard, J. Leloup, C. Menkes, et al. (2020). Quantifying
857 the benefits of non-linear methods for global statistical hindcasts of tropical cyclones intensity,
858 *Monthly Weather Review*, doi:10.1175/WAF-D-19-0163.1
859 Nidheesh, A. G., Lengaigne, M., Vialard, J., Izumo, T., Unnikrishnan, *A.* S., Meyssignac, B., et al.
860 (2017). Robustness of observation-based decadal sea level variability in the Indo-Pacific
861 Ocean. *Geophys. Res. Lett*, *44*(14), 7391-7400.
862 Parvathi, V., Suresh, I., Lengaigne, M., Ethé, C., Vialard, J., Levy, M., et al. (2017). Positive Indian
863 Ocean Dipole events prevent anoxia along the west coast of India. *Biogeosciences*, 14: 1541-
864 1559, doi:10.5194/bg-14-1541-2017.
865 Patara, L., Visbeck, M., Masina, S., Krahmann, G., and Vichi, M. (2011). Marine biogeochemical
866 responses to the North Atlantic Oscillation in a coupled climate model, *J. Geophys. Res.*
867 *Oceans*, 116, C07023.
868 Polovina, J. J., Howell, E. A., and Abecassis, M. (2008). Ocean's least productive waters are
869 expanding. *Geophys. Res. Lett*., 35, L03618, doi:10.1029/2007GL031745.
870 Radenac, M. H., Léger, F., Singh, A., and Delcroix, T. (2012). Sea surface chlorophyll signature
871 in the tropical Pacific during eastern and central Pacific ENSO events. *J. Geophys. Res.*
872 *Oceans*, 117(C4).
873 Radenac, M. H., Messié, M., Léger, F., and Bosc, C. (2013). A very oligotrophic zone observed
874 from space in the equatorial Pacific warm pool. *Remote sensing of environment*, 134, 224-
875 233.

876 Resplandy, L., Lévy, M., Madec, G., Pous, S., Aumont, O., and Kumar, D. (2011). Contribution
877 of mesoscale processes to nutrient budgets in the Arabian Sea. *J. Geophys. Res. Oceans*,
878 116(C11).
879 Rykaczewski, R. R., and Dunne, J. P. (2011). A measured look at ocean chlorophyll trends. *Nature*,
880 472(7342), E5-E6.
881 Saji, N. H., Goswami, B. N., Vinayachandran, P. N., and Yamagata, T. (1999). A dipole mode in
882 the tropical Indian Ocean. *Nature*, 401(6751), 360.
883 Sakamoto, T., Gitelson, A. A., Wardlow, B. D., Verma, S. B., and Suyker, A. E. (2011). Estimating
884 daily gross primary production of maize based only on MODIS WDRVI and shortwave
885 radiation data. *Remote Sensing of Environment*, 115(12), 3091-3101.
886 Sauzède, R., Claustre, H., Jamet, C., Uitz, J., Ras, J., Mignot, A., and D'Ortenzio, F. (2015).
887 Retrieving the vertical distribution of chlorophyll a concentration and phytoplankton
888 community composition from in situ fluorescence profiles: A method based on a neural
889 network with potential for global-scale applications. *J. Geophys. Res. Oceans*, 120, 451‑470,
890 doi:10.1002/ 2014JC010355.
891 Schneider, B., Bopp, L., Gehlen, M., Segschneider, J., Frölicher, T. L., Cadule, P., et al. (2008).
892 Climate-induced interannual variability of marine primary and export production in three
893 global coupled climate carbon cycle models. *Biogeosciences, European Geosciences Union*,
894 2008, 5 (2), pp.597-614. hal-00330705
895 Schollaert Uz, S., Busalacchi, A. J., Smith, T. M., Evans, M. N., Brown, C. W., Hackert, E. C., et
896 al. (2017). Interannual and Decadal Variability in Tropical Pacific Chlorophyll from a
897 Statistical Reconstruction: 1958–2008. *J. Clim.* 30, 7293–7315. doi:10.1175/JCLI-D-16-
898 0202.1.

899 Séférian, R., Bopp, L., Gehlen, M., Orr, J. C., Ethé, C., Cadule, P., et al. (2013). Skill assessment
900 of three earth system models with common marine biogeochemistry. *Clim. Dynam.*, 40, 2549–
901 2573, doi:10.1007/s00382-012-1362-8. 1379.
902 Siegel, D. A., Behrenfeld, M. J., Maritorena, S., McClain, C. R., Antoine, D., Bailey, S. W., at al.
903 (2013). Regional to global assessments of phytoplankton dynamics from the SeaWiFS
904 mission. *Remote Sensing of Environment*, 135, 77-91.
905 Siegel, D. A., Maritorena, S., Nelson, N. B., and Behrenfeld, M. J. (2005). Independence and
906 interdependencies among global ocean color properties: Reassessing the bio-optical
907 assumption. *J. Geophys. Res. Oceans*, 110(C7).
908 Signorini, S. R., Franz, B. A., and McClain, C. R. (2015). Chlorophyll variability in the
909 oligotrophic gyres: mechanisms, seasonality and trends. *Frontiers in Marine Science*, 2, 1.
910 Signorini, S. R., and McClain, C. R. (2012). Subtropical gyre variability as seen from satellites.
911 *Remote Sensing Letters*, 3(6), 471-479.
912 Smith, T. M., Arkin, P. A., Ren, L., and Shen, S. S. (2012). Improved reconstruction of global
913 precipitation since 1900. *J. Atmosph. and Ocean. Tech.*, 29(10), 1505-1517.
914 Staten, P. W., Lu, J., Grise, K. M., Davis, S. M., and Birner, T. (2018). Re-examining tropical
915 expansion. *Nature Climate Change*, 8(9), 768-775.
916 Steinacher, M., Joos, F., Frölicher, T. L., Bopp, L., Cadule, P., Cocco, V., et al. (2010). Projected
917 21st century decrease in marine productivity: a multi-model analysis. *Biogeosciences*, 7(3),
918 979-1005.
919 Storm, T., Boettcher, M., Grant, M., Zühlke, M., Fomferra, N., Jackson, T., and Sathyendranath,
920 S. (2013). Product User Guide, Ocean Colour Climate Change Initiative. 51. www.esa-
921 oceancolour-cci.org/?q=webfm_send/317.

922 Tagliabue, A., Bopp, L., Dutay, J. C., Bowie, A. R., Chever, F., Jean-Baptiste, P., et al. (2010).
923 Hydrothermal contribution to the oceanic dissolved iron inventory. *Nature Geoscience*, 3(4),
924 252-256.
925 Tang, W., Li, Z., and Cassar, N. (2019). Machine Learning Estimates of Global Marine Nitrogen
926 Fixation. *J. Geophys. Res.: Biogeosciences*, 124(3), 717-730.
927 Thompson, D., and Wallace, J. M. (2000). Annular modes in the extratropical circulation. Part I:
928 Month-to-month variability. *J. Climate*,13, 1000–1016.
929 Thomas, A. C., Strub, P. T., Weatherbee, R. A., and James, C. (2012). Satellite views of Pacific
930 chlorophyll variability: Comparisons to physical variability, local versus nonlocal influences
931 and links to climate indices. *Deep-Sea Res. Prt II*, 77, 99-116.
932 Vantrepotte, V., and Mélin, F. (2009). Temporal variability of 10-year global SeaWiFS time-series
933 of phytoplankton chlorophyll a concentration. *ICES J. Marine Science*, 66(7), 1547-1556.
934 Vapnik, V. (2000). Statistics for engineering and information science. The Nature of Statistical
935 Learning Theory.
936 Vapnik, V. (1998). The support vector method of function estimation. *In Nonlinear Modeling* (pp.
937 55-85). Springer, Boston, MA.
938 Vapnik, V. N. (1995). The Nature of Statistical Learning Theory. New York: Springer2 Verlag.
939 Westberry, T., Behrenfeld, M. J., Siegel, D. A., and Boss, E. (2008). Carbon-based primary
940 productivity modeling with vertically resolved photoacclimation. *Global Biogeochem. Cycles*,
941 22(2).
942 Wiggert, J. D., Murtugudde, R. G., and Christian, J. R. (2006). Annual ecosystem variability in
943 the tropical Indian Ocean: Results of a coupled bio-physical ocean general circulation model.
944 *Deep Sea Res. Part II*, Volume 53, Issues 5–7, Pages 644-676.

945 Wiggert, J. D., Vialard, J., and Behrenfeld, M. J. (2009). Basin wide modification of dynamical
946 and biogeochemical processes by the positive phase of the Indian Ocean Dipole during the
947 SeaWiFS era. In: Indian Ocean Biogeochemical Processes and Ecological Variability, vol.
948 185, edited by: J. D. Wiggert, R. R. Hood, S. Wajih, A. Naqvi, K. H. Brink, and S. L. Smith,
949 p. 350, 2009.
950 Wilson, C., and Adamec, D. (2001). Correlations between surface chlorophyll and sea surface
951 height in the tropical Pacific during the 1997-1999 El Nino-Southern event. *J. Geophys. Res. Oceans*, 106(C12),31175-31188.
953 Wilson, C., and Adamec, D. (2002). A global view of bio-physical coupling from SeaWiFS and
954 TOPEX satellite data, 1997 – 2001. *Geophys. Res. Lett*., 29(8), 1257,
955 doi:10.1029/2001GL014063.
956 Wilson, C., and Coles, V. J. (2005). Global climatological relationships between satellite
957 biological and physical observations and upper ocean properties. *J. Geophys. Res. Oceans,*
958 110, C10001, doi:10.1029/2004JC002724.
959 Yoder, J. A., and Kennelly, M. A. (2003). Seasonal and ENSO variability in global ocean
960 phytoplankton chlorophyll derived from 4 years of SeaWiFS measurements. *Global*
961 *Biogeochem. Cycles,* 17(4), 1112, doi:10.1029/2002GB001942.
962 Yoo, S. and Son, S.-H. (1998). Seasonal patterns of phytoplankton biomass change in the
963 Northwestern Pacific Marginal Seas observed by CZCS. *Ocean Res*., 20: 29-42.
964 Zhan, H., Shi, P., and Chen, C. (2003). Retrieval of Oceanic Chlorophyll Concentration Using
965 Support Vector Machines. *IEEE Transactions on Geoscience and Remote Sensing*, 41 (12),
966 2947–2951.