Limnol. Oceanogr., 53(2), 2008, 605–613

Identifying four phytoplankton functional types from space: An ecological approach

Dionysios E. Raitsos
School of Earth, Ocean, and Environmental Sciences (SEOES), University of Plymouth, Drake Circus, Plymouth, PL4 8AA, United Kingdom; Sir Alister Hardy Foundation for Ocean Science (SAHFOS), The Laboratory, Citadel Hill, Plymouth PL1 2PB, United Kingdom; Hellenic Centre for Marine Research (HCMR), 46,7 Km Athens-Sounio, P.O. Box 712, 190 13 Anavissos, Attica, Greece

Samantha J. Lavender
School of Earth, Ocean, and Environmental Sciences (SEOES), University of Plymouth, Drake Circus, Plymouth, PL4 8AA, United Kingdom

Christos D. Maravelias and John Haralabous
Hellenic Centre for Marine Research (HCMR), 46,7 Km Athens-Sounio, P.O. Box 712, 190 13 Anavissos, Attica, Greece

Anthony J. Richardson
Sir Alister Hardy Foundation for Ocean Science (SAHFOS), The Laboratory, Citadel Hill, Plymouth PL1 2PB, United Kingdom; Department of Mathematics, University of Queensland, St. Lucia, Queensland, 4072, Australia; CSIRO Marine and Atmospheric Research, P.O. Box 120, Cleveland, Queensland 4163, Australia

Philip C. Reid
Sir Alister Hardy Foundation for Ocean Science (SAHFOS), The Laboratory, Citadel Hill, Plymouth PL1 2PB, United Kingdom

Abstract

Deriving maps of phytoplankton taxa based on remote sensing data using bio-optical properties of phytoplankton alone is challenging. A more holistic approach was developed using artificial neural networks, incorporating ecological and geographical knowledge together with ocean color, bio-optical characteristics, and remotely sensed physical parameters. Results show that the combined remote sensing approach could discriminate four major phytoplankton functional types (diatoms, dinoflagellates, coccolithophores, and silicoflagellates) with an accuracy of more than 70%. Models indicate that the most important information for phytoplankton functional type discrimination is spatio-temporal information and sea surface temperature. This approach can supply data for large-scale maps of predicted phytoplankton functional types, and an example is shown.

As the foundation of the aquatic food chain, phytoplankton are an integral part of the ecosystem, affecting trophic dynamics, nutrient cycling, habitat condition, and fisheries resources (Irigoien et al. 2002). Phytoplankton are responsible for >45% of the total primary production of plants on Earth (Falkowski et al. 2004) and uptake of the greenhouse gas carbon dioxide (CO_2), and they contribute to the biological pump.

Although the role of marine phytoplankton is significant, knowledge of spatio-temporal distribution and abundance of functional types is limited, especially in the open oceans. Research has been restricted in both time and space because information is often obtained from relatively expensive ship-based in situ measurements. Deriving maps of phytoplankton functional types (PFTs) from remotely sensed data is a new and potentially important technological application which offers high spatio-temporal coverage. Anderson (2005) reported that ecology is poorly understood due to a lack of in situ data as well as functional-type information related to the chemical and physical regime, which in turn has hindered the development of a convincing PFT prediction model. Empirical relationships between in situ pigment measurements and remotely sensed ocean color data were determined by Alvain et al. (2005) who generated global maps of haptophytes, prochlorococcus, synechococcus-like cyanobacteria, and diatoms. Development of several bio-optical methods for the identification of different PFTs have been used to map coccolithophore bloom distributions (Brown

Acknowledgments

We thank present and past staff of SAHFOS who have contributed to the maintenance of the CPR time series. Special thanks to Abigail McQuatters-Gollop and Yaswant Pradhan for useful discussions and comments on the manuscript. D. E. Raitsos is supported by a scholarship from the University of Plymouth.

This study was also supported by the U.K. Natural Environment Research Council through the Atlantic Meridional Transect consortium (NERC/O/S/2001/00680) and Center for Observation of Air-Sea Interactions and Fluxes (CASIX). This is contribution No. 159 of the AMT programme and No. 46 for CASIX.

and Podestá, 1997), trichodesmium (Subramaniam et al. 2002), and diatoms (Sathyendranath et al. 2004). Although some innovative studies have provided promising results for discriminating PFTs from space, they have acknowledged weaknesses. One of these weaknesses is the limited availability of data needed to develop the robust relationships necessary for building accurate models. In situ data are not only restricted by the method of collection, but also a significant proportion of the data (~85%) are not matched with concurrent satellite data due primarily to cloud coverage (Sathyendranath et al. 2004; Raitsos et al. 2005). In addition, research has tended to separate PFTs using bio-optical properties alone (spectral absorption and backscattering). More robust interpretations should be possible when additional information about the physical, chemical, and biological environments that different PFTs prefer is included.

This research discriminates between four common and important PFTs. Diatoms account for ~20% of global carbon fixation, ~25% of global primary production, and a large amount of the carbon exported to the deep ocean via sinking particles (Armbrust et al. 2004). Diatoms are also a key food source for copepods and other zooplankton, which are subsequently consumed by larger predators such as fish and marine mammals (Irigoien et al. 2002), thereby transferring energy to higher levels of the marine food web. The second type, photosynthetic dinoflagellates, is an important aquatic primary producer. However, they are less nutritious than diatoms and can result in food webs culminating in non-fodder gelatinous organisms instead of fish and are thus sometimes considered trophic dead ends (Verity and Smetacek, 1996). Certain dinoflagellate species cause red tides and may impact fisheries, aquaculture, and marine mammal and human health by introducing toxins into the food chain (Nixon 1995). The third type is coccolithophores, which are capable of forming spatially extensive blooms (Raitsos et al. 2006). Calcifiers such as coccolithophores contribute to some of the densest ballasts observed in sinking particles (Klaas and Archer 2002). They are major producers of dimethyl-sulphide (DMS), calcium carbonate, and organic carbon, all of which affect climate (Holligan et al. 1993). Growing at extensive scales, their role in and contribution to the oceanic and atmospheric environment (Tyrrell and Merico, 2004), as well as to the local heat budget and biogeochemical cycle is important at a global scale (Holligan et al. 1993). Finally, silicoflagellates represent a minor fraction of the total microplankton assemblage in the pelagic environment; they are a major component in coastal and estuarine waters (Jochem and Babenerd 1989). Silicoflagellates are also good indicators of water masses and have been used in reconstructions of the paleoenvironment (Onodera and Takahashi 2005). Nejstgaard et al. (2001) reported that certain bloom-forming silicoflagellate species may negatively affect copepod reproduction in the sea.

To discriminate between the PFTs, a Probabilistic Neural Network (PNN) utilized ecological (phytoplankton) and geographical knowledge along with ocean color bio-optical characteristics and remotely sensed physical parameters. Physical and optical variables include chlorophyll *a* (Chl *a*), solar radiation, sea surface temperature (SST), wind stress and normalized water-leaving radiances (nLw); they were supplemented by spatio-temporal information including longitude, latitude, and season. Phytoplankton information derived from the Continuous Plankton Recorder (CPR), which is an upper-layer plankton monitoring program in the North Atlantic Ocean and North Sea operating since 1931 (Reid et al. 2003), was used to train the PNN.

Methods and data analysis

Methodological approach—All datasets were processed for the northern North Atlantic (46°N–66°N, 52°W–4°W) between September 1997 and December 2003 (Fig. 1). Within this region concurrent match-ups between SeaWiFS and in situ CPR samples (phytoplankton biomass) were compared (*see* Raitsos et al. 2005 for methodological details). A data matrix was then produced with concurrent satellite remote sensing and CPR measurements for the same spatial and temporal coverage (~300 weekly composite images of the North Atlantic). In this way, no interpolation or averaging of datasets for the area of study was necessary, therefore allowing the inclusion of local variations and extreme events in high spatio-temporal resolution comparisons in the analysis. Thus, after screening the satellite dataset for CPR match-ups, 3,732 (of the available 14,001) samples could be used for comparison (Fig. 1).

Phytoplankton functional types—Measurements of phytoplankton abundance (cell counts) were derived from the CPR survey, the largest and longest running plankton survey in the world. Samples were collected by a high-speed plankton recorder (~15–20 km h^{-1}) that is towed behind 'ships of opportunity' in the surface layer of the ocean (~6–10 m deep); one sample represents ~18 km of tow (Richardson et al. 2006). Plankton were filtered onto a constantly moving band of silk. CPR analysis involves the taxonomic identification of species and cell counts for each sample. In this study, the total number of species per sample for each of the four functional types (diatoms, dinoflagellates, coccolithophores, and silicoflagellates) was used. Each type is comprised of many species (*see* list in Richardson et al. 2006 for details). The dominant phytoplankton type for each sample was estimated using the Z factor standardized method

$$Z_i = \frac{n_i - \bar{x}_i}{s_i}$$

Where n_i is the cell count for phytoplankton type i in a sample, $\bar{x}_i$ is the overall mean of all cell counts for each type i, and s_i is the standard deviation of all samples for type i. The largest Z_i for each sample was used as the dominant species. This standardized method was used to derive the dominant type because the number of cells between each of the four types was substantially different. For instance, diatoms form more concentrated blooms than silicoflagellates (mean cell counts 140,000 and 34,000 respectively). Whenever CPR samples indicated that there

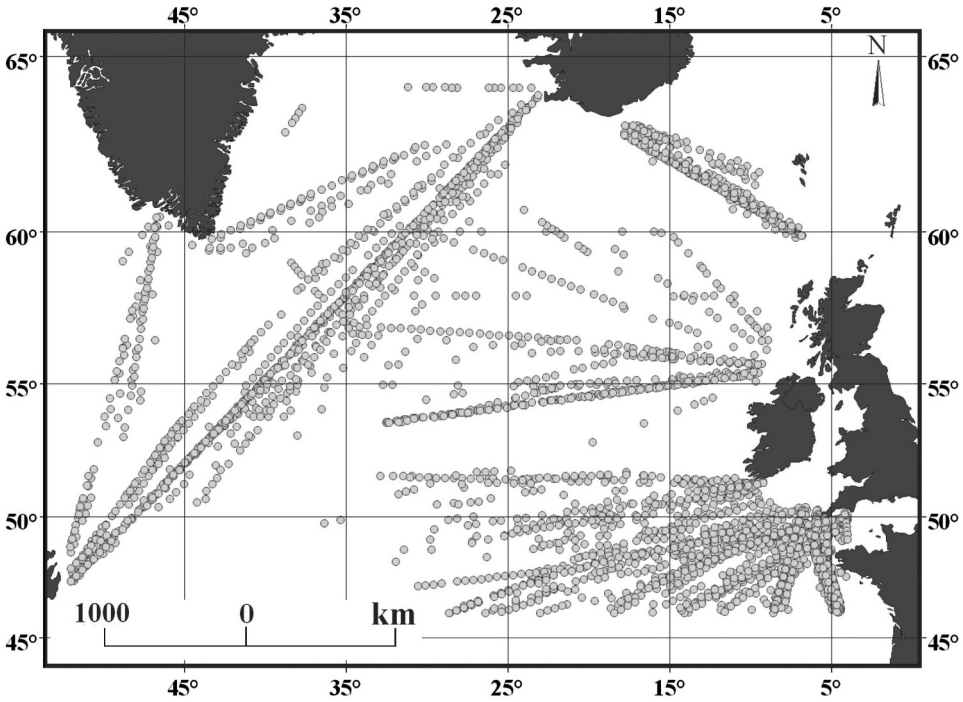

Fig. 1. CPR and satellite match-ups between 1997 and 2003 in the North Atlantic Ocean.

were no phytoplankton cells, this category was named "no-dominance." Although phytoplankton ≥ 10 μm were caught (qualitatively), most of the biomass of pico- and ultraplankton were not included in the patterns (Richardson et al. 2006). Thus, this research focused on the larger autotrophic component and samples classified as having no-dominance may in fact have been dominated by cells <10 μm in size that could not be considered in this study.

Satellite data—SeaWiFS (Sea-viewing Wide Field-of-view Sensor): Reprocessed data (version 5.1) produced by the Ocean Biology Processing Group was acquired from the NASA Oceancolor website (http://oceancolor.gsfc.nasa.gov/). Data were level three, 8-d composite products (9 km^2 $\times$ 9 km^2 resolution) of near-surface Chl *a* (mg m^{-3}), normalized water leaving radiance (nLw) at 555 nm (mW cm^{-2} μm^{-1} sr^{-1}) and Photosynthetically Active Radiation (PAR) (E m^{-2} d^{-1}). Chl *a* concentration was estimated using the Ocean Chlorophyll 4–version 4 (OC4-v4) algorithm (O'Reilly et al. 1998), which performs well in the open waters that dominate the study area (Fig. 1). Phytoplankton optical properties (light absorption and backscattering) have been found to vary among different phytoplankton types, thus the nLw product was used as a proxy for backscattering in the discrimination procedure (Alvain et al. 2005). The PAR product is the incoming solar radiation or insolation that can be simply defined as the light intensity received at the surface of the Earth (http://oceancolor.gsfc.nasa.gov/DOCS/). Light intensity is clearly fundamental to photosynthesis (Nanninga and Tyrrell 1996).

Advanced Very High Resolution Radiometer (AVHRR): The nighttime AVHRR Pathfinder 5 weekly means of sea surface temperature (SST) at 4 km^2 $\times$ 4 km^2 resolution were obtained from the NASA PO.DAAC website (http://poet.jpl.nasa.gov/). Nighttime SST products were used so that the solar radiation bias (the diurnal fluctuation in SST) that can occur from surface heating during daytime could be avoided (Raitsos et al. 2006). Generally, temperature has major direct (e.g., metabolic) and indirect (e.g., through stratification) effects on phytoplankton (Edwards and Richardson, 2004).

European Remote Sensing Satellites (ERS–2) and NASA-QuikSCAT (QS): Weekly composites of mean wind stress data ($0.5^\circ \times 0.5^\circ$ spatial resolution) were obtained from CERSAT, IFREMER (http://www.ifremer.fr/cersat/en/index.htm). Data were available from ERS–2 and QS, and preliminary intersensor comparisons between the derived wind speeds have indicated that the sensors are compatible (http://www.ifremer.fr/cersat/en/research/validation/qscat_vs_topex_ ers.htm). Because wind stress is responsible for vertical mixing of the water column, it may have an indirect effect on phytoplankton through nutrient availability. Wind stress is a function of wind speed, the non-dimensional drag coefficient, and the boundary layer air density (Pickard and Pond 1978). The spatial variation of wind stress over the ocean causes surface divergence of horizontal flow that in turn gives rise to vertical mass flux through Ekman pumping (Pond and Pickard 1983).

Potential data biases—Weekly mean satellite data (for Chl *a*, the 8-d standard NASA product) was used to establish CPR match-ups. When daily satellite data were used, ~85% of CPR data were unusable, but with weekly mean composites the loss was reduced to 73%. Results from a daily comparison indicated that they were not

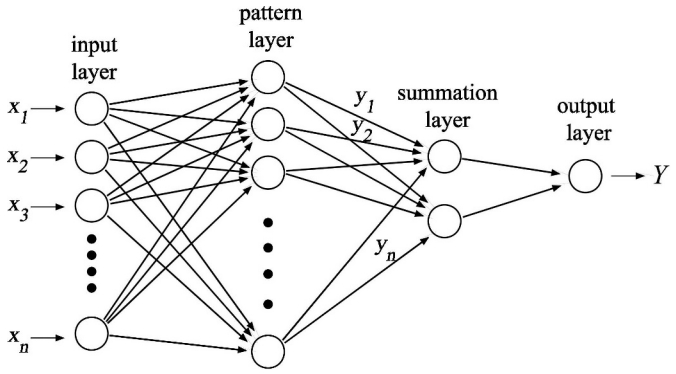

Fig. 2. A schematic representation of a probabilistic neural network structure.

statistically different from the weekly comparison (data not shown). However, the weekly relationship was more robust (based on more samples), and the 95% confidence limits were reduced considerably, indicating an improvement of the relationship. In summary, the weekly products capture the dynamic information that is present in daily images.

Consistency and comparability of the methodology used in the CPR survey has been studied in some depth and can be found elsewhere (Batten et al. 2003; Richardson et al. 2006). Although the CPR provides unique spatial and temporal coverage of North Atlantic waters, samples are not equally distributed in space or time. In addition, data are screened for match-up comparisons, which may lead to a potential bias if a particular month or week is consistently cloudy each year. Therefore, the data and hence the derived relationships might not be representative of the entire study area at all times. However, this issue was relatively minor in the study because the samples were distributed relatively uniformly seasonally, interannually, and spatially.

The study area included both Case I (open ocean) and Case II (coastal) waters. In optically-complex Case II waters, Chl *a* cannot readily be distinguished from particulate matter and/or yellow substances (colored dissolved organic matter) and so global chlorophyll algorithms (such as OC4-v4) are less reliable (IOCCG 2000). Because the majority of the study area was comprised of Case I water, this bias influenced only a small proportion of the data points (Fig. 1).

Data analysis using probabilistic neural networks—In this study PNNs, a type of Artificial Neural Network (ANN), were used to discriminate between four PFTs based on environmental, optical, and spatio-temporal variables. The PNN is in essence a combination of neural networks and Bayesian statistics (Specht 1988; Specht 1990). Bayes theory takes into account the relative likelihood of events and uses a priori information to improve prediction. The network paradigm uses Parzen estimators and spheres of influence that were developed to construct probability density functions required by Bayes theory.

The most common choice of kernel is the basic Gaussian kernel, which involves only the Gaussian function and one sphere of influence parameter σ. A schematic representation of a typical PNN structure is given in Fig. 2. As Bayesian approximators, the basic Gaussian kernel PNN built to map x_k as function of x_1; x_2;,… , x_{k-1} will have $k - 1$ neurons in the input layer, one neuron in the first hidden layer (pattern layer) for each case in the training set, $k - 1$ neurons in the second hidden layer (summation layer), and one neuron in the output layer (decision layer).

The current PNN was implemented on a random subset of available data (training set) and then applied to the remaining data (validation or testing set). The training set consisted of a random 80% of the available data (2,986 cases) with the remaining random 20% (746 cases) used as the testing set. This holdout partitioning technique (Kohavi 1995) was repeated five times to test the validity of the model. The testing set was used for calibration, which prevented overtraining the networks, making them generalise well on new data. Calibration adjusted the weight of each neuron by computing the distance metric between a given classification and the network results for all outputs over all patterns. The genetic adaptive algorithm (Specht 1991) was applied during this process to test a range of smoothing factors. Individual smoothing factors were used as a sensitivity analysis tool, as the larger the factor for a given input, the more important that input was to the model, at least, as far as the test set was concerned.

The following performance criteria were evaluated: (1) sensitivity, the percentage of true presences correctly identified; (2) specificity, the percentage of true absences correctly identified; and (3) accuracy, the total fraction of the sample correctly identified. When applied to training sets, the accuracy provides a measure of the recognition performance, whereas when applied to testing sets it gives a measure of prediction performance.

An important property of ANNs is that they are adaptive, i.e., they can learn from new data. This ability does not depend upon the prior knowledge of rules. ANNs have the ability to extract essential process information from data. As new training data become available, the network can be updated to represent the process more accurately. Moreover, with only a few exceptions, neural networks are essentially nonlinear, and they are capable of learning complex interactions among the input variables in a system even when those interactions are difficult to find and describe. Consequently, neural networks can provide solutions for problems that do not have an algorithmic solution or for which an algorithmic solution is too complex to be found. A further important advantage of neural networks is that they are capable of generalization, i.e., they can correctly process information that only broadly resembles the original training data. They are also fault tolerant by being capable of properly handling noisy or incomplete data. Additionally, ANNs work well with various types of data because there are no conditions put on the predicted variables, i.e., they can be true/false, continuous values, and so forth.

Using a PNN in addition to its nonlinear and multimodal properties has several advantages. First, a PNN network structure is dictated by the dimensionality of the samples as opposed to some other ANNs such as multilayered perceptrons whose network structure is determined either by a trial-and-error or a rule-of-thumb

Table 1. Percentage values of sensitivity, specificity and classification accuracy of the five types, applying the PNN five times. Each time a random sample of 80% of cases was used as the training set and the remaining 20% as the testing set. Results are given analytically for each sample in training and testing. Mean values of the five samples are given in bold.

		Set					
		Training			Testing		
Type	Sample	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
No-dominance	1	86.7	96.2	93	75.6	83.1	80.4
	2	86.6	96.3	93	76	82.9	80.4
	3	87	96.2	93.1	76.3	82.6	80.4
	4	87.4	96	93.1	77.1	81.6	80
	5	86.8	96.1	93	77.1	82.4	80.6
	$\bar{x}$	**86.9**	**96.2**	**93**	**76.4**	**82.5**	**80.4**
	SD	0.3	0.1	0.1	0.7	0.6	0.2
Diatom	1	83.8	96.3	93.3	74	83.1	81
	2	83.9	96.2	93.3	75.1	83.7	81.6
	3	84.1	96.5	93.5	75.1	83.3	81.4
	4	84.1	96.5	93.6	73.5	82.8	80.6
	5	83.2	96.1	93.1	75.1	83	81.1
	$\bar{x}$	**83.8**	**96.3**	**93.4**	**74.6**	**83.2**	**81.1**
	SD	0.4	0.2	0.2	0.8	0.3	0.4
Dinoflagellate	1	90.1	93.7	92.9	73.6	85.4	82.8
	2	90	93.9	93.1	72.3	85	82.3
	3	89.7	93.5	92.7	71.1	85.5	82.4
	4	89.8	93.6	92.8	72.3	85.7	82.8
	5	88.9	93.8	92.7	70.4	85	81.9
	$\bar{x}$	**89.7**	**93.7**	**92.8**	**72**	**85.3**	**82.5**
	SD	0.5	0.2	0.2	1.2	0.3	0.4
Coccolithophore	1	83.5	98.1	96.2	72.8	92.5	90.4
	2	83.2	98.2	96.3	71.6	92.6	90.4
	3	85.1	98	96.4	75.3	92.9	91
	4	82.4	98	96.1	74.1	92.3	90.4
	5	82.7	98.1	96.2	70.4	92.6	90.2
	$\bar{x}$	**83.4**	**98.1**	**96.2**	**72.8**	**92.6**	**90.5**
	SD	1	0.1	0.1	2	0.2	0.3
Silicoflagellate	1	85	98.1	96.9	70.2	95.1	92.9
	2	84.3	98.1	96.8	67.2	94.9	92.4
	3	86.1	98.2	97.1	74.6	95	93.2
	4	83.9	98.1	96.8	62.7	95.3	92.4
	5	84.7	98	96.8	65.7	95.1	92.5
	$\bar{x}$	**84.8**	**98.1**	**96.9**	**68.1**	**95.1**	**92.7**
	SD	0.8	0.1	0.1	4.6	0.2	0.4

approach. Second, the training process requires only one pass of the training set in order to compute the optimal smoothing factor. Last, it works well on small training sets, and when the sample size increases, it provides a very close approximation to the real density function. Nevertheless, its major drawback is the need to store all training cases in the pattern layer for future classifications.

Results

Table 1 illustrates the discrimination results of the PNN for the training, testing, and overall datasets. Training outcomes (percentages) can be considered to be an indication of the quality of the data used to develop relationships for the final discrimination model. Sensitivity analysis (true presences) in the training stage performed well, because each phytoplankton functional type was classified with a precision of >83%, and dinoflagellates had the highest percentage (>89.7%). Regarding true absences (specificity), the PNN model performance was >93%, with coccolithophores and silicoflagellates having the highest percentages. The classification accuracy had a recognition performance of >92%, with coccolithophores and silicoflagellates again showing the highest accuracy (Table 1). Using the training relationships derived from 80% of the dataset, the ability of the PNN model to discriminate types within the remaining (random) 20% of the samples was examined (testing). The sensitivity results showed that true presences were correctly identified with a precision of >68%; the highest mean discrimination performance occcurred in the no-dominance (76.4%), diatom (74.6%) and coccolithophore (72.8%) types; whereas, dinoflagellate and silicoflagellate types performed slightly lower (72.0% and 68.1%, respectively). From Table 1 also shows that the specificity performed considerably better; every type was >82%, with silicoflagellate and coccolithophore types (95.1% and 92.6%, respectively) performing better, followed by the dinoflagellate (85.3%), diatom (83.2%) and

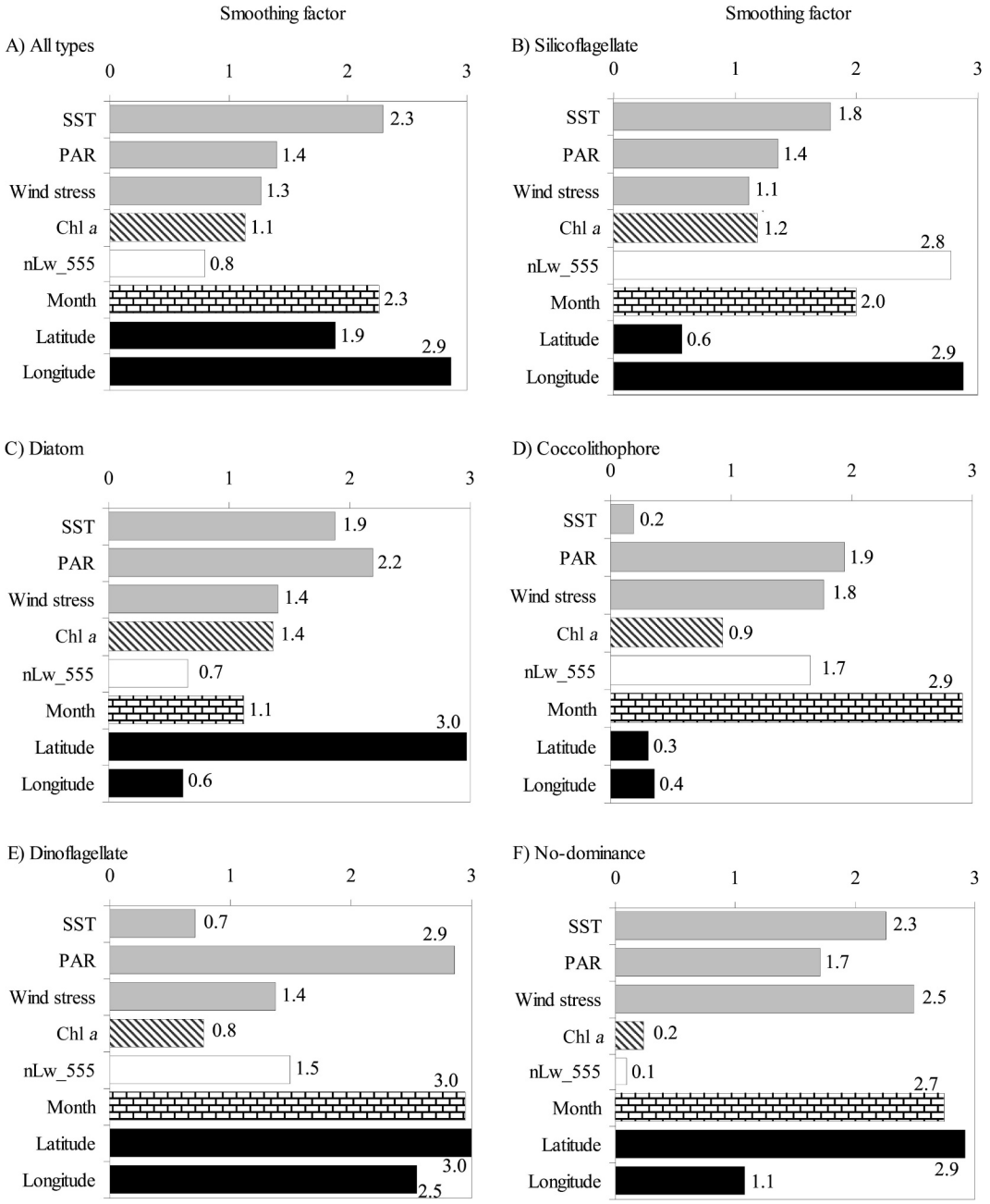

Fig. 3. The relative impact (smoothing factor) of the eight variables indicating the importance of each variable in distinguishing the types from each other as separated functional types (All types) and discriminating one type from the others (i.e., diatoms from the remaining types). Note that the different coloration or shading separates the physical, biological, optical, temporal, and spatial variables (respectively). (A) All types, (B) Silicoflagellates, (C) Diatoms, (D) Coccolithphores, (E) Dinoflagellates, and (F) No-dominance.

no-dominance (82.5%) types. Although silicoflagellates and dinoflagellates had the lowest sensitivity performance in testing, the opposite pattern was observed for specificity.

The accuracy for the total fraction of the functional types correctly identified indicated the final prediction ability of the PNN model and had an accuracy of >80%. Performance results occurred in the following order: silicoflagellates (92.7%), coccolithophores (90.5%), dinoflagellates (82.5%), diatoms (81.1%), and no-dominance (80.4%).

The contribution of each predictor used for discrimination varied within and among PFTs. Thus, the analysis was made up of two approaches: (1) distinguishing the types from each other as separate functional types (Fig. 3: all types plot), and (2) discriminating one type from the others (i.e., diatoms from the remaining types; Fig. 3).

The relative impact (smoothing factor) of the eight predictors is shown in Fig. 3, with the first plot (all types) indicating the importance of each variable in distinguishing the functional types. Spatio-temporal information such as

longitude, month, and latitude made an important contribution to the discrimination of the PFTs (2.9, 2.3, and 1.9, respectively). Regarding the effect of the physical regime, SST was found to be the key factor (2.3), followed by light intensity (1.4) and wind stress (1.3). Chl *a* and nLw_555 did contribute to the overall discrimination, but they were of lesser importance compared with other parameters (1.1 and 0.8, respectively).

In terms of diatoms, latitude along with PAR and SST (3.0, 2.2, and 1.9, respectively) were the key factors for discriminating this functional type from other types (Fig. 3). In addition, wind stress and Chl *a* played an important discriminatory role (1.4 and 1.4, respectively), whereas longitude (0.6) had the lowest impact. Dinoflagellates were distinguished mostly based on spatiotemporal information, i.e., latitude (3.0), month (3.0), and longitude (2.5). In terms of physical variables, PAR (2.9) was the key discriminator, followed by nLw_555 and wind stress. SST and Chl *a* contributed less to the dinoflagellate discrimination (Fig. 3). Discrimination of coccolithophores appeared to be regulated by the seasonal cycle (month, 2.9), with PAR (1.9), wind stress (1.8) and nLw_555 (1.7) playing key roles, while SST (0.2) appeared not to make a significant contribution. Longitude (2.9) and nLw_555 (2.8) had the highest impact for discriminating silicoflagellates, followed by month, SST, and PAR. Although Chl *a* and wind stress contributed to the final discrimination, their impacts appeared to be less significant (Fig. 3). Finally, for the no-dominance type almost all parameters played key roles, but contributions of Chl *a* and nLw_555 (0.2 and 0.1 respectively) were smaller.

Discussion

The discrimination of PFTs from remotely sensed data is usually based on bio-optical properties and does not incorporate spatio-temporal or environmental knowledge (Sathyendranath et al. 2004; Alvain et al. 2005). In this study, current research was extended by incorporating geographical, temporal, biological, physical, and bio-optical information. Results demonstrated that neural networks are able to discriminate and identify four major functional types (diatoms, dinoflagellates, coccolithophores, and silicoflagellates) with an accuracy of $>70\%$. This is the first step toward an ecological approach which will ultimately be able to predict PFTs without the geographical (longitude and latitude) information.

Because geographical information was considered during the training of the ANN and appeared to be a very important variable, the forecasted phytoplankton maps should be focused only on the northeast Atlantic Ocean. However, once additional CPR dataset become available, an attempt to exclude the geographical information could be made. Although this would decrease the accuracy of the model, the results could become more globally applicable. In addition, nutrients are key regulators of phytoplankton abundance (Redfield et al. 1963), and the mixed layer depth is responsible for the stratification and supply of nutrients. They were not used because remote sensing cannot offer these variables, and in situ data at a fine spatial and temporal scale, like those used in this study, do not exist. However, wind stress that is a measure of vertical mixing was included.

An important property of successful presence/absence ecological models, when applied to independent datasets, is their ability to predict presence accurately. Hence, sensitivity is considered to be of primary importance compared to specificity (and overall accuracy) because the latter can suffer from the prevalence effect of the types, (i.e., frequency of occurrence; Maravelias et al. 2003). In this study, true absence outnumbered true presence. Therefore, prevalence effect was reflected in every section of the analysis (training and testing) because the PNN model predicted true absence better than true presence. However, in the testing section, the ability of the model to successfully discriminate and identify the true cases was $\sim 70\%$.

Several studies have dealt with the identification of key processes controlling the growth of PFTs (e.g., Platt et al. 2005), but this study investigated the importance of parameters regarding species discrimination. Although identification of key processes and species discrimination are similar, they are not the same. For example, a parameter such as SST might be vital for a particular type such as coccolithophores (Cokacar et al. 2004; Raitsos et al. 2006). However, this study found that light intensity, wind stress, and seasonal information were more important than SST when separating coccolithophores from other types. Although all predictors played a role, the variables responsible for the highest percentages of discrimination accuracy were the spatio-temporal variables as well as the physical ones such as SST, PAR (light intensity) and wind stress (vertical mixing). The fact that bio-optical information was of lesser importance does not mean that the model would have performed adequately without this information. Platt et al. (2005) argued that growth as well as community and size structure of phytoplankton assemblages are controlled by physical factors. Physical parameters sometimes reflect the habitat of the epipelagic zone because they have significant direct and indirect impacts on phytoplankton. Therefore, one factor alone was not sufficient to identify/separate the PFTs. Alvain et al. (2005) reported that Chl *a* alone will never be able to discriminate PFTs; however, using this factor in combination with other variables may make possible the ultimate goal of deriving maps of these types using remote sensing data. Based on the study results, future research aimed at the discrimination/identification of functional types from remotely sensed data should include fundamental information on the physical environment.

The proposed approach has considerable potential for mapping spatial and temporal (seasonal cycle) trends in PFTs using remote sensing data alone, which is particularly important for areas with minimal in situ data. The next step is to use the relationships (between the phytoplankton types and variables) obtained from the data analysis in training the current statistical model (PNN) to forecast the spatio-temporal distribution of PFTs. Fig. 4 illustrates an example of diatom presence in the study area derived using this method. Because a weekly image (second week of May 1999) was used, the number of match-ups was not sufficient for it

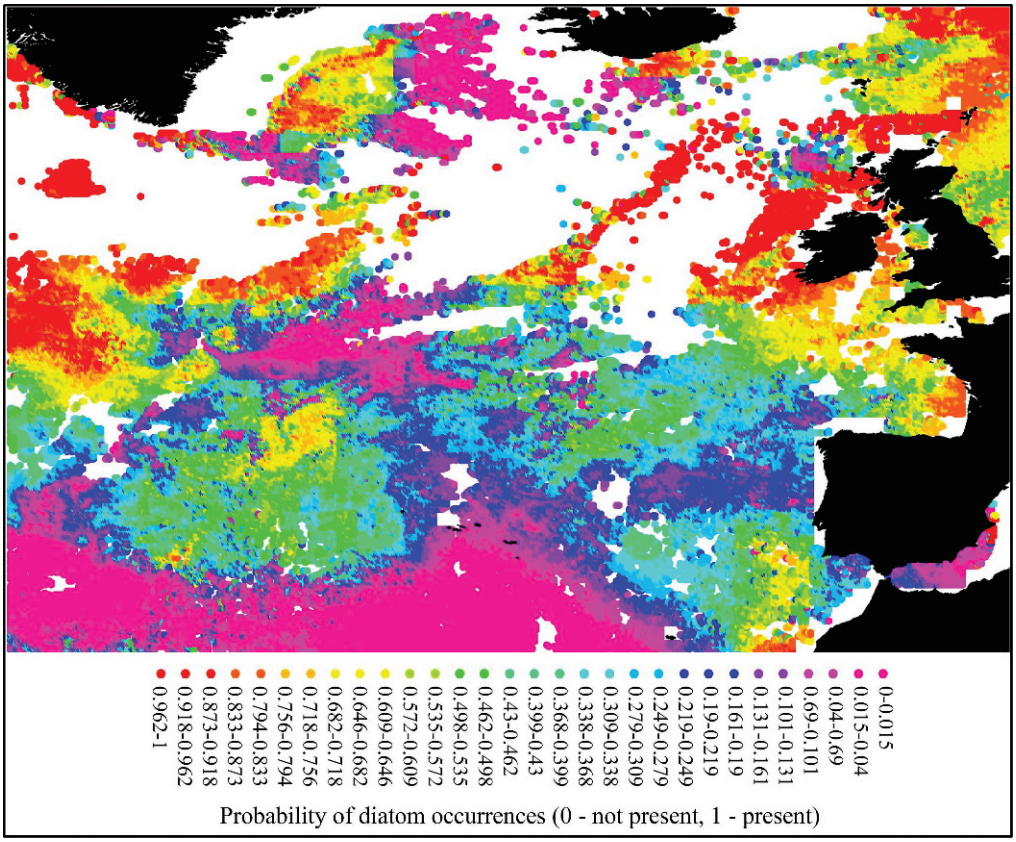

Fig. 4. Prediction of diatoms presence for the second week of May 1999. May is an important month for diatom blooming in the study area, and this week was relatively cloud free. The scale of the probability of diatoms occurrence is from 0 to 1 (not present to present).

to be directly compared with the CPR data. However, monthly composites of the PFTs can be produced and validated by comparison with the CPR maps and the output of papers such as Alvain et al. (2005). Predicted distributions of PFTs in global biogeochemical models have been less than convincing (Anderson 2005), probably because there is no information for validation at a global scale. Therefore, comparisons with these models are also important.

Satellites are not a substitute for ship-based sampling not only because in situ verification is always needed to improve/confirm the results, but also because satellites only receive information about the physical/biological regime within the top of the water column. However, Uitz et al. (2006) inferred phytoplankton biomass, its vertical distribution, and the community composition (microplankton, nanoplankton and picoplankton) from near-surface, satellite-derived Chl *a* concentrations, and coupled physical–biological models are moving toward the assimilation of biological data. Therefore, the future lies in the combined utilization of in situ data, remote sensing, and modelling.

References

Alvain, S., C. Moulin, Y. Dandonneau, and F. M. Breon. 2005. Remote sensing of phytoplankton groups in case 1 waters from global SeaWiFS imagery. Deep-Sea Res. Part I **52:** 1989–2004.

Armbrust, E. V., and others. 2004. The genome of the diatom Thalassiosira pseudonana: Ecology, evolution, and metabolism. Science **306:** 79–86.

Anderson, T. R. 2005. Plankton functional type modelling: running before we can walk? J. Plankton Res. **27**(11)**:** 1073–1081.

Batten, S. D., A. W. Walne, M. Edwards, and S. B. Groom. 2003. Phytoplankton biomass from continuous plankton recorder data: An assessment of the phytoplankton colour index. J. Plankton Res. **25:** 697–702.

Brown, C. W., and G. P. Podestá. 1997. Remote sensing of coccolithophore blooms in the western South Atlantic Ocean. Remote Sens. Environ. **60:** 83–91.

Cokacar, T., T. Oguz, and N. Kubilay. 2004. Satellite-detected early summer coccolithophore blooms and their interannual variability in the Black Sea. Deep-Sea Res. Part I **51**(8)**:** 1017–1031.

Edwards, M., and A. J. Richardson. 2004. Impact of climate change on marine pelagic phenology and trophic mismatch. Nature. **430:** 881–884.

Falkowski, P. G., M. E. Katz, A. H. Knoll, A. Quigg, J. A. Raven, O. Schofield, and F. J. R. Taylor. 2004. The evolution of modern eukaryotic phytoplankton. Science **305:** 354–360.

Holligan, P. M., and others. 1993. A biogeochemical study of the coccolithophore Emiliania huxleyi in the north Atlantic. Global Biogeochem. Cycles **7:** 879–900.

International Ocean-Color Coordinating Group (IOCCG). 2000. Remote sensing of ocean color in coastal, and other optically-complex waters. *In* Sathyendranath, S. [ed.], Reports of the International Ocean-Color Coordinating Group, No. 3. Dartmouth, Canada. 140 p.

Irigoien, X., and others. 2002. Copepod hatching success in marine ecosystems with high diatom concentrations. Nature **419:** 387–389.

Jochem, F., and B. Babenerd. 1989. Dictyocha speculum—a new type of phytoplankton bloom in the Western Baltic. Mar. Biol. **103:** 373–379.

Klaas, C., and D. E. Archer. 2002. Association of sinking organic matter with various types of mineral ballast in the deep sea: Implications for the rain ratio. Global Biogeochem. Cycles **16:** 1116 p.

Kohavi, R. 1995. The power of decision tables, p. 174–189. *In* Proceedings of the 8th European Conference on Machine Learning. Springer-Verlag. London.

Maravelias, C. D., J. Haralabous, and C. Papaconstantinou. 2003. Predicting demersal fish species distributions in the Mediterranean Sea using artificial neural networks. Mar. Ecol. Prog. Ser. **255:** 249–258.

Nanninga, H. J., and T. Tyrrell. 1996. Importance of light for the formation of algal blooms by Emiliania huxleyi. Mar. Ecol. Prog. Ser. **136:** 195–203.

Nejstgaard, J. C., B. H. Hygum, L. J. Naustvoll, and U. Bamstedt. 2001. Zooplankton growth, diet, and reproductive success compared in simultaneous diatom- and flagellate microzooplankton-dominated plankton blooms. Mar. Ecol. Prog. Ser. **221:** 77–91.

Nixon, S. 1995. Coastal marine eutrophication: A definition, social causes and future concerns. Ophelia **41:** 199–219.

Onodera, J., and K. Takahashi. 2005. Silicoflagellate fluxes and environmental variations in the northwestern North Pacific during December 1997–May 2000. Deep-Sea Res. Part I **52:** 371–388.

O'Reilly, J. E., and others. 1998. Ocean color chlorophyll algorithms for SeaWiFS. J. Geophys. Res. **103:** 24937–24953.

Pickard, G. L., and S. Pond. 1978. Introductory dynamic oceanography, 2nd ed. Pergamon.

Platt, T., H. Bouman, E. Devred, C. Fuentes-Yaco, and S. Sathyendranath. 2005. Physical forcing and phytoplankton distributions. Sci. Mar. **69:** 55–73.

Pond, S., and G. L. Pickard. 1983. Introductory dynamical oceanography. Oxford.

Raitsos, D. E., S. J. Lavender, Y. Pradhan, T. Tyrrell, P. C. Reid, and M. Edwards. 2006. Coccolithophore bloom size variation in response to the regional environment of the subarctic North Atlantic. Limnol. Oceanogr. **51:** 2122–2130.

———, P. C. Reid, S. J. Lavender, M. Edwards, and A. J. Richardson. 2005. Extending the SeaWiFS chlorophyll data set back 50 years in the northeast Atlantic. Geophys. Res. Lett. **32:** L06603.

Redfield, A. C., B. H. Ketchum, and F. A. Richards. 1963. The influence of organisms on the composition of sea water. p. 26–77. *In* M. N. Hill [ed.], The sea. Wiley.

Reid, P. C., J. B. L. Matthews, and M. A. Smith. 2003. Achievements of the Continuous Plankton Recorder survey and a vision for its future. Progr. Oceanogr. **58:** 115–358.

Richardson, A. J., and others. 2006. Using continuous plankton recorder data. Progr. Oceanogr. **68:** 27–74.

Sathyendranath, S., L. Watts, E. Devred, T. Platt, C. Caverhill, and H. Maass. 2004. Discrimination of diatoms from other phytoplankton using ocean-color data. Mar. Ecol. Prog. Ser. **272:** 59–68.

Specht, D. F. 1988. Probabilistic neural networks for classification, mapping, or associative memory (IEEE). Neural Networks **1:** 525–532.

———. 1990. Probabilistic neural networks (IEEE). Neural Networks **3:** 109–118.

———. 1991. A general regression neural network (IEEE). Neural Networks **2:** 568–576.

Subramaniam, A., C. W. Brown, R. R. Hodd, E. J. Carpenter, and D. G. Capone. 2002. Trichodesmium blooms in SeaWiFS imagery. Deep-Sea Res. Part II **49:** 107–121.

Tyrrell, T., and A. Merico. 2004. Emiliania huxleyi: Bloom observations and the conditions that induce them, p. 75–97. *In* H. R. Thiertein and J. R. Young [eds.], Coccolithophores: From molecular processes to global impact. Springer-Verlag.

Uitz, J., H. Claustre, A. Morel, and S. B. Hooker. 2006. Vertical distribution of phytoplankton communities in open ocean: An assessment based on surface chlorophyll. J. Geophys. Res. **111:** CO8005, doi:10.1029/2005JC003207.

Verity, P. G., and V. Smetacek. 1996. Organism life cycles, predation, and the structure of marine pelagic systems. Mar. Ecol. Prog. Ser. **130:** 277–293.

Received June 2006
Accepted 3 August 2007
Amended 11 October 2007