JOURNAL OF PLANKTON RESEARCH | VOLUME 30 | NUMBER 8 | PAGES 857–872 | 2008

Modelling the influence of abiotic and biotic factors on plankton distribution in the Bay of Biscay, during three consecutive years (2004–06)

LUCIA ZARAUZ[1]*, XABIER IRIGOIEN[2] AND JOSE A. FERNANDES[2]

[1]MARINE RESEARCH DIVISION, AZTI FOUNDATION, TXATXARRAMENDI UGARTEA, Z/G 48395, SUKARRIETA (BIZKAIA), SPAIN AND [2]MARINE RESEARCH DIVISION, AZTI FOUNDATION, HERRERA KAIA PORTUALDEA, Z/G 20110, PASAIA (GUIPUZKOA), SPAIN

*CORRESPONDING AUTHOR: lzarauz@suk.azt.es

Received November 21, 2007; accepted in principle February 27, 2008; accepted for publication April 9, 2008; published online May 2, 2008

Corresponding editor: Roger Harris

Three surveys were carried out in the Bay of Biscay during the springs of 2004, 2005 and 2006. Hydrographic, nano–microplankton (diatoms, ciliates and unidentified particles) and mesozooplankton biomass data were collected at mesoscale spatial resolution. Generalized additive models (GAMs) based on a combination of hydrographic, geographic and biological terms were used to understand the factors affecting distribution. The final models accounted for 66% of the variability in the biomass distribution of unidentified particles, 60% for diatoms, 44% for mesozooplankton and 23% for ciliates. The contribution of hydro-geographical terms was greater than the information described by the biological variables. Geographical location (latitude and longitude) was the main explanatory factor for all of the plankton groups identified, revealing that the presence of mesoscale fronts related to geographical structures is more relevant than the hydrographic variables per se, to describe plankton distribution.

INTRODUCTION

Predictive habitat models are becoming a common tool in terrestrial ecology (Guisan and Zimmermann, 2000). Such models allow researchers to describe the present day distributions with high accuracy, as well as to predict relatively reliably (i.e. Hirzel *et al.*, 2006) the effect of changes in the environment (climate change, habitat modification, etc). In marine systems, habitat models are beginning to be used and the specific requirements for the modelling of different marine organisms (such as fish) recognized (Loukos *et al.*, 2003; Ward and Myers, 2006). However, plankton studies present specific difficulties to fully benefit from the use of predictive habitat models. It is well known that plankton distribution is "patchy" in nature. This spatial heterogeneity covers a wide range of time-space scales, and is shaped by a variety of physical and biological processes, acting both together and independently (Haury *et al.*, 1978). The main discussion on patchiness has been centred upon whether biological or physical agents are the main factors responsible for the observed spatial structure (Levin and Segel, 1976; Haury *et al.*, 1978; Abraham, 1998; Martin, 2003). To resolve the patches, mesoscale (0–100 km) sampling is needed. At this scale physical processes are particularly powerful and have great impacts on biological variability (Mann and Lazier, 1991). However, traditional plankton surveys have been limited by the laborious and time-consuming nature of sampling and analysis. Such a problem of sampling planktonic organisms at the relevant scales has hampered the understanding of control mechanisms in marine systems. Thus, our understanding of the functioning of marine ecosystems lags behind that of terrestrial ones (Duarte, 2007).

doi:10.1093/plankt/fbn049, available online at www.plankt.oxfordjournals.org

The field of phytoplankton research has benefited from the presence of photosynthetic pigments, which has permitted the use of colour and fluorescence to study the distribution with high spatial resolution (by means of satellite imaging methods, *in situ* fluorescence measurements, etc), and with some taxonomic differentiation methods (HPLC). However, to fully understand the ecosystem, information is needed on the distribution of prey and predators: relative abundances, size distribution, overlapping, etc. Until recently, obtaining high-resolution distribution of heterotrophs regularly at sea, especially microzooplankton, was not possible. However, within the last few decades, new image analysis systems have been developed for rapid and high-resolution plankton data acquisition. These systems are designed for the study of both autotrophic and heterotrophic plankton, over a broad range of sizes (Ashjian *et al.*, 2001; Davis *et al.*, 2004; Grosjean *et al.*, 2004; See *et al.*, 2005; Benfield *et al.*, 2007); and constitute a powerful alternative to the traditional manual treatment of plankton samples. At present, automatic sampling methods still lack taxonomic detail (Hu and Davis, 2006). Nevertheless, recent studies have demonstrated the power of machine learning and data mining used to classify field-collected plankton organisms of different taxa (Culverhouse *et al.*, 2003; Blaschko *et al.*, 2005; Hu and Davis, 2006), in achieving good accuracy levels in terms of abundance and biomass of major taxonomic groups (Culverhouse *et al.*, 2003, 2006). These systems, as of yet, do not have the resolving power to identify plankton to the level of species and life stage, but can provide important information on coarse taxonomic composition (Davis *et al.*, 2005).

Another problem, common to marine and terrestrial ecosystems, is the difficulty in modelling the complex non-linear relationships that exist between physical and biological processes. In this field, new statistical methods have been developed to model spatial and temporal dependence in ecological data. The generalized additive model (GAM) method (Hastie and Tibshirani, 1990) constitutes a very flexible modelling approach to describe the influence of different explanatory factors in a given variable. Through the use of spline functions, GAMs allows the data to accommodate the shape of the response curves, to almost any functional form. Moreover, GAMs are able to handle multicollinearity between variables (Yee and Mitchell, 1991), and to minimize the effects of extreme observations (Wood, 2006). Because of all this, GAMs have demonstrated a great ability to model complex non-linear relationships between variables, and so have been applied in many fields of ecological research, both for terrestrial (Guisan *et al.*, 2002) and marine systems (Augustin *et al.*, 1998; Beare *et al.*, 2000; Stratoudakis *et al.*, 2003; Planque *et al.*, 2007).

The combined use of the above-mentioned techniques provides approaches for the study of plankton distributions at the mesoscale level, with some taxonomic information and statistical tools to disentangle non-linear relations. This combination permits further understanding of the mechanisms that define the observed spatial structure. The aim of the present study was to investigate how hydrographical and biological factors affect biomass distribution of four broad planktonic groups (mesozooplankton, diatom chains, ciliates and unidentified particles), and to evaluate their potential to describe interannual variations in plankton distribution, for a period of three consecutive years.

METHODS

Sample collection and analysis

Three surveys were carried out in the Bay of Biscay during the springs of 2004, 2005 and 2006. Hydrographic, nano–microplankton and mesozooplankton biomass data were collected. Stations were distributed along transects perpendicular to the coast, covering the shelf and shelf-break (Fig. 1). The dates and the number of stations for each cruise are presented in Table I. Both hydrographic and plankton distribution data were represented in contour plots, performed using interpolation by krigging with Surfer software (Golden software Inc.).

Hydrographic data

The hydrographic characteristics of the surveyed stations were sampled using a CTD (model RBR XR420) fitted to a mesozooplankton net. The difference in seawater density between 100 m (or 5 m above the bottom) and the surface was used as an index of the water column stratification.

Nano- and microplankton

Nano–microplankton samples were collected at a depth of 3 m using 1.5 L NISKIN bottles. They were analysed onboard, by means of a FlowCAM (Sieracki *et al.*, 1998), to determine the biomass and size structure of the nano- and microplankton community. Fluorescence measurements were not included in the analysis; therefore, every particle (phytoplankton, zooplankton, detritus and inorganic) was counted and imaged. For each sample, a maximum of either 10 000 particles or a volume of 20 mL were analysed; except for the 2004 cruise, where either 2000 particles or 10 mL were processed.

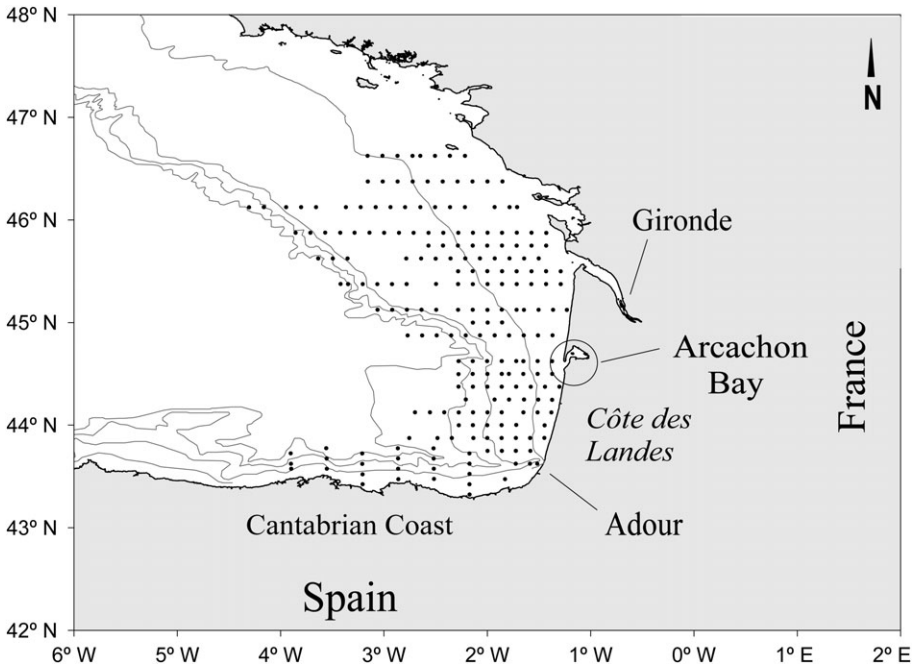

Fig. 1. General location of the study area. Main estuaries and geographical areas are indicated. Nano–microplankton sampling stations are plotted for the 2006 cruise. 100, 200, 1000 and 2000 m isobaths are shown.

A ×4 objective was used in the sample analysis. The instrument was calibrated using beads of a known size. Invalid recordings (i.e. bubbles, repeated images) were removed from the image database through visual recognition. The biovolume of each cell was calculated from its equivalent spherical diameter (ESD), and was converted into biomass using the equation given by Montagnes (Montagnes *et al.*, 1994) for marine diatoms and dinoflagellates.

Mesozooplankton

Mesozooplankton samples were obtained using vertical hauls of a 150-μm PAIROVET net (0.25 m net diameter). The nets were lowered to a maximum depth of either 100 m, or 5 m above the bottom, at the shallower stations.

Net samples were preserved immediately after collection with 4% borax-buffered formalin. The samples were stained for 24 h with 4 mL of 1% eosin, which stains the cell cytoplasm and the muscle protein. This stain creates enough contrast to be recognized by image analysis and reduces the counting of the detrital material. Net sub-samples were scanned at 24 bits colour, at a resolution of 600 dpi and using an HP Scanjet8200 series scanner.

The resulting jpg images were imported into the Zooimage software (GPL license, http://www.sciviews.org/zooimage) in order to count, measure and classify objects in the digital images. The volume of each zooplankter was calculated from the ESD of each particle >165 μm; it was converted to carbon and corrected for shrinkage caused by formalin (Alcaraz *et al.*, 2003). Artefacts (i.e. fibres) were identified automatically and removed from the data base.

For both nano–microplankton and mesozooplankton, the biomass of every individual was obtained. Very small and very large sizes were considered to be underrepresented because of: (a) the relative low volumes analysed by the FlowCAM and the ZooImage or (b) being undersampled by the collection gear (bottles and nets). Consequently, nano–microplankton biomass was estimated for the size range 7–50 μm ESD and mesozooplankton biomass was calculated for organisms ranging from 165 to 1300 μm ESD. It must be taken into account that nano–microplankton biomass and hydrographic parameters are surface measurements, while net tows integrate mesozooplankton biomass from the bottom (or 100 m) to the surface.

Nano- and microplankton classification

Nano–microplankton sized particles (7–50 μm ESD) were automatically classified into three broad groups such as diatom chains, ciliates and unidentified particles, using the WEKA toolkit (GNU license, http://www.cs.waikato.ac.nz/ml/weka; Witten and Frank, 2005). Two different algorithms were tested: Random Forest (RF; Breiman, 2001) and Tree Augmented Naive Bayes (TAN; Friedman *et al.*, 1997). These techniques

Table I: Date, number of stations sampled, mean values and range (in brackets) of plankton biomass data and hydrographic variables, for the cruises carried out in the years 2004, 2005 and 2006

		Stations (*n*)			Biomass (mg m^{-3})				CTD		
Cruise	Dates (May)	Nmi	Mzoo	CTD	diatoms	unidentified	ciliates	mesozoo	Temperature (°C)	Salinity (UPS)	Strat. (kg m^{-3})
2004	2–22	173	409	319	42 (0.01–217)	330 (23–1093)	23 (0.01–183)	11 (1–69)	13.7 (12.6–15.5)	34.1 (28.5–36)	1.1 (−0.1–5)
2005	8–27	200	418	416	19 (0.39–96)	127 (19–424)	24 (0.01–115)	10 (0.85–44)	14.9 (12.7–17)	34.3 (29.3–35.2)	1.3 (0.41–4.8)
2006	5–23	215	393	402	25 (0.01–457)	148 (13–514)	18 (1–110)	13 (0.01–75)	15.5 (12.6–18.3)	34.5 (32.5–35.4)	1.5 (0.45–0.94)

Nmi, nano–microplankton; Mzo, mesozooplankton.

search for decision trees (RF) and probabilistic relationships between the predictor variables (TAN), and are learnt using a training set where the items are labelled according to their corresponding class.

In this study, each object was described by five morphological parameters: cell area, cell ESD, maximum length, minimum length and number of particles per chain. Data used to establish the training set originated from stations sampled in the Bay of Biscay. These stations were selected on the basis of their high abundance and diversity of plankton. To obtain the training set, particles measured at these stations were classified by a human expert, by looking at the FlowCAM images and categorizing them. Due to the limited resolution of the images, only diatoms larger than 20 μm ESD and ciliates larger than 50 μm ESD could be identified as such; the rest were classified as unidentified particles.

A preliminary analysis showed that the main source of error was false negative-type errors for diatoms (diatoms classified as unidentified particles). To improve on this limitation, caused by the imbalanced nature of the problem, diatoms instances were duplicated in the training set (Witten and Frank, 2005). The final training set was composed of 317 ciliates; 442 diatoms and 15 869 unidentified particles. This disparity in the number of individuals reflected the different abundances present in the field.

The results were evaluated using 10-fold cross-validation. In 10-fold cross validation, the data are split into 10 randomly chosen and approximately equal partitions. Each part is held out in turn, and the learning scheme trained on the remaining nine-tenths; the error rate is then calculated on the holdout set. This learning procedure is performed a total of 10 times, omitting a different subset each time. Finally, the 10 error estimates are averaged to yield an overall error estimate (Witten and Frank, 2005).

A simple way to estimate the error of our classification is to calculate the percentage of particles that have been classified correctly (overall accuracy). However, the main objective of the present classification was to distinguish ciliates and diatoms, so total accuracy was not enough to evaluate the performance of the classifier. A confusion matrix, precision and recall for each class were calculated to evaluate the classification of the two different algorithms.

The confusion matrix is a table where the true counts (manual counts) for each group are presented in the rows, while counts by automatic identification are given in the columns. Good results correspond to large numbers down the main diagonal and small, ideally zero, off-diagonal elements.

Precision is defined as the number of correct results divided by the number of all results returned as positive:

$$\pi = \frac{\mathrm{tp}}{\mathrm{tp} + \mathrm{fp}}$$

The recall rate is defined as the number of correct results divided by the number of results that should have come back as positive:

$$\rho = \frac{\mathrm{tp}}{\mathrm{tp} + \mathrm{fn}}$$

where tp are true positive instances, fp are false positives and fn are false negatives.

It must be pointed out that the "unidentified particles" category, incorporates the fraction of nano–microplankton which could not be identified due to the limited resolution of the images obtained with the FlowCAM. These are particles sized between 7 and 50 μm ESD, comprising both autotrophic and heterotrophic organisms (e.g. nanoflagellates, dinoflagellates, centric or pennate diatoms . . .). As the FlowCAM does not discriminate between living and inert particles, this class may also include detrital and inorganic material. On the other hand, due to the low quality of the mesozooplankton images, it was not possible to discriminate zooplankters beyond the copepod level. As such, no further classification was carried out for this group. An example of the nano–microplankton and mesozooplankton images is given in Fig. 2.

Statistical analysis

The data collected during the three years of sampling were used to model the biomass of each plankton group as a function of hydrographic and biological variables, using GAMs (Hastie and Tibshirani, 1990). The analyses presented here were performed using the "mgcv" library in the R-statistical software. R is available from the Comprehensive R Archive Network, http://cran.r-project.org. A further description of GAM methods and their applications to habitat modelling can be found in the literature (Hastie and Tibshirani, 1990; Wood, 2006; Planque *et al.*, 2007; Zarauz *et al.*, 2007).

The transformed biomass values, $\log_{10}$ (plankton biomass + 1), were used to develop the GAMs. The hydrographic parameters used were surface salinity, surface temperature and stratification index. The geographical location (latitude and longitude) and the depth of the water column (in $\log_{10}$ scale) were included as a proxy of physical structures which have a persistent location over time, i.e. the low-salinity plume of the estuaries, or the internal waves at the shelf-break. Latitude and longitude were included in a two-dimensional spline to allow interactions between the two covariates. Biological explanatory variables were defined on the basis of the biomass of the different plankton categories (diatom chains, ciliates, unidentified particles, mesozooplankton), representing the various ecological relationships (i.e. predator-prey, competition) existing between plankton groups.

During the first stage, GAMs based on individual covariates were applied to identify the relationship between individual explanatory factors and the response variable. Non-linear dependence in the data was

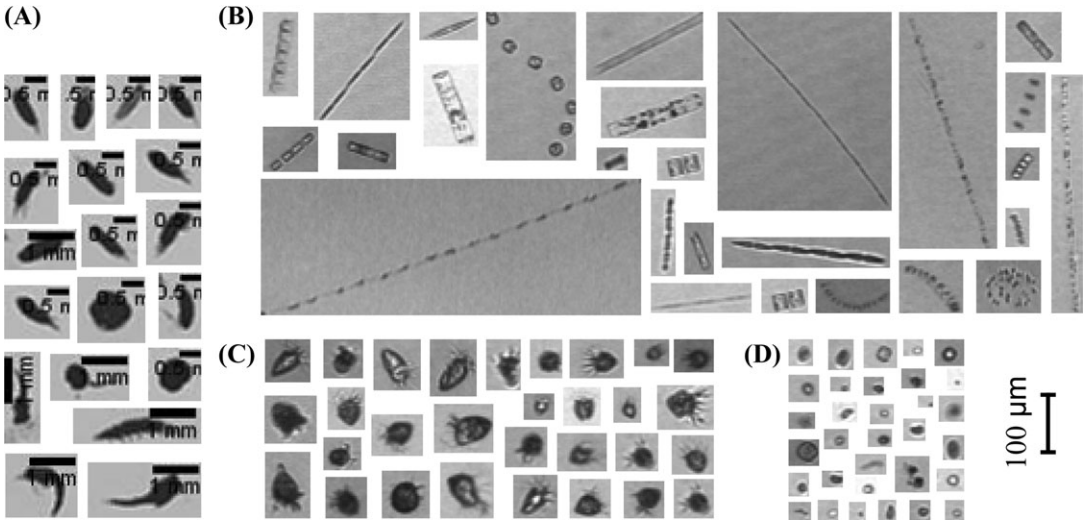

Fig. 2. Examples of the identified plankton groups: (**A**) mesozooplankton, (**B**) diatom chains, (**C**) ciliates and (**D**) unidentified particles. A scale bar is included in each mesozooplankton image. For diatom chains, ciliates and unidentified particles, the scale bar is placed outside the images.

handled within the GAMs using spline functions. At a further step, GAMs of increasing complexity were built by adding hydro-geographical and biological covariates to the models. The selection of the explanatory factors to include in each GAM was performed applying a stepwise elimination, based on the method proposed by Wood and Augustin (Wood and Augustin, 2002). The following criteria were used to identify the significant terms: (i) the generalized cross validation (GCV; the lower, the better); (ii) the level of deviance explained (0–100%; the higher, the better); (iii) the confidence region for the smoothing (which was not to include zero through the range of the prediction). Additionally, parametric terms were identified regarding the estimated degrees of freedom of each term. When the degrees of freedom were close to their lower limit (i.e. close to 1 for univariate smooths), the covariate was defined by a linear coefficient and included in the GAM as a parametric term (Daskalov, 1999; Wood, 2008). In GAM formulae and figures, linear coefficients are omitted and smooth functions are referred to as $s(\,)$, following the notation used in R statistical software.

To avoid overfitting, a Jackknife (JK) procedure was applied to validate the final models using an independent data set (Lobo and Martin-Piera, 2002; Zarauz *et al.*, 2007). The *P*-value and the coefficient of determination (r^2) of the least squares linear regression between the GAM predicted and observed biomass data were used to validate the model.

RESULTS

Hydrography

The hydrography of the Bay of Biscay is influenced largely by the outflows of the main rivers in the area, the Gironde and Adour. In our study, the lowest values of surface salinity corresponded to the plume of the Gironde, which affected the French coastal and shelf waters. The influence of the Adour outflow, on southern French waters, was also observed throughout the three years. The Cantabrian coast was characterized by slightly lower salinity values, but no riverine output was distinguished (Fig. 3).

The three cruises were carried out at similar times of the year. However, significant differences in the surface temperature and stratification were observed between them (Fig. 3). In 2004, surface temperatures were the lowest, ranging from 13.7 to 15.6°C (Table I). Higher

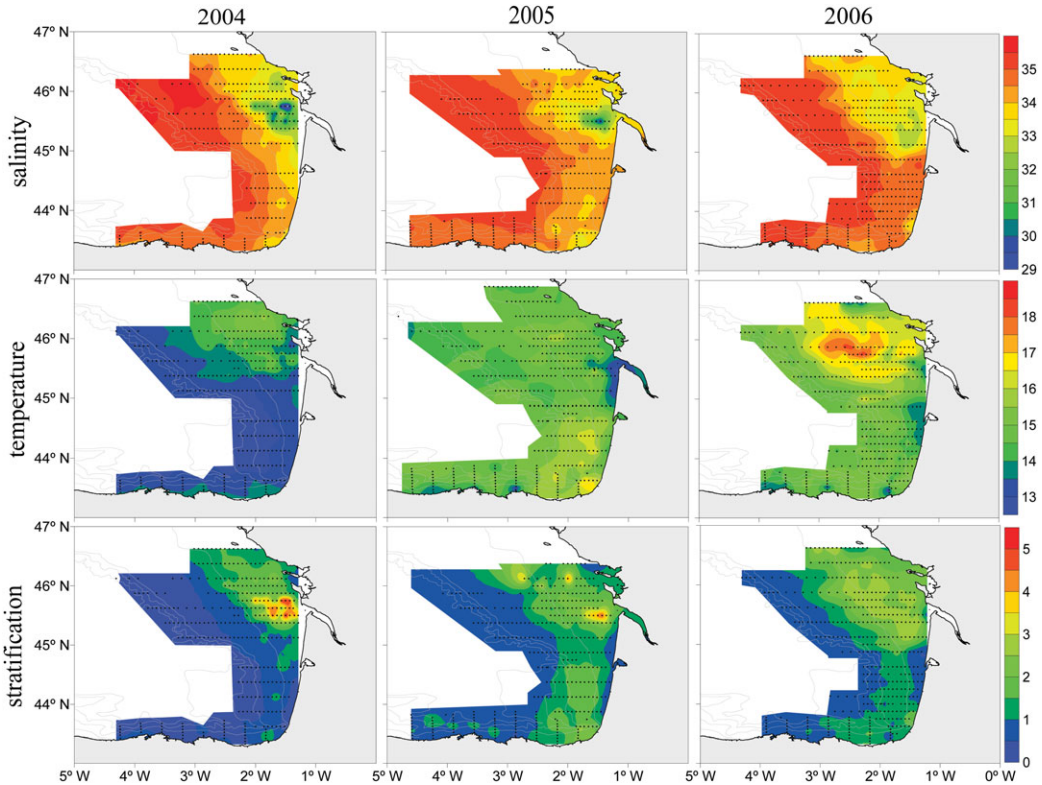

Fig. 3. Spatial distribution of sea surface salinity (PSU), surface temperature (°C) and stratification index (kg m^{-3}), for the cruises in 2004, 2005 and 2006. The sampling grid is plotted and bathymetry is similar to Fig. 1.

temperatures were found in the north of the bay, revealing a warming of the surface waters during the last days of the cruise. Stratification in this campaign was due to low salinity waters in the area under the influence of the Gironde and Adour outflows.

Surface temperatures in 2005 and 2006 were significantly higher, reaching maximum values of 18.3°C (Table I, Fig. 3). The coldest waters were found in the estuary of the Gironde, in the Bay of Arcachon and in some coastal areas of the Cantabrian shore. In these two years, the stratification index ranged from 0.4 to 4.8 kg m^{-3}, revealing the onset of summer thermal stratification. An area characterized by low temperatures and low stratification, probably related to an upwelling, can be observed along the French west coast, between the Gironde and the Adour estuaries.

Accuracy of nano–microplankton classification

Significance tests showed that overall accuracy levels achieved using RF were considerably higher than the accuracy levels achieved with the TAN algorithm (Table II). Moreover, RF has previously been described as an efficient algorithm to perform plankton classification (Grosjean *et al.*, 2004). As such, it was selected to perform the classification of nano–microplankton data collected in the present study.

The results of the 10-fold cross-validation showed an overall accuracy of 99.4% for the classification carried out using the RF algorithm (Table III). Confusion matrix showed that the main source of error was due to false negative-type errors for ciliates, i.e. ciliates classified as unidentified particles (Table IV). This type of error is evaluated in the recall scores, for each individual group. The recall score represents the percentage of instances of a given class which are correctly classified in that category, in relation to the total number of instances that correspond to that class. In the case of ciliates, this would mean dividing the number of correctly classified ciliates (297) by the total existing number of ciliates in our training set (317). This is slightly different to precision which evaluates the percentage of instances that really correspond to the class in relation to the total instances classified in a given class. In the experiment undertaken here, diatoms and unidentified particles achieved the highest recall scores (99.7 and 95.2%, respectively) and ciliates the lowest (93.7%).

Table II: Output of the significance test

Iteration	RF	TAN	
1	99.43	99.1	*
2	99.47	99.15	*
3	99.52	99.12	*
4	99.48	99.07	*
5	99.43	99.09	*
	(v/ /*)	(0/0/5)	

The percentage of correctly classified instances using RF and TAN algorithms are compared. Annotations (v/ /*) correspond to the number of iterations in which TAN algorithm is significantly better (v), similar () or worse (*) than RF.

Table III: Precision, recall and overall accuracy for the classifier based upon RF algorithm, computed using 10-fold cross validation

	Precision	Recall
diatoms	0.998	0.997
Unidentified	0.909	0.952
Ciliates	0.955	0.937
Overall accuracy	99.43%	

Table IV: A confusion matrix for the classifier based upon RF algorithm

Classified	Diatoms	Unidentified	Ciliates
diatoms	**421**	20	1
unidentified	40	**15 816**	13
ciliates	2	18	**297**

True counts (manual counts) for each group are presented in the rows, while counts by automatic identification are given in the columns. Good results correspond to large numbers in the main diagonal (bolded), and ideally zero off-diagonal elements.

Nano–microplankton and mesozooplankton biomass distribution

The biomass of diatom chains, unidentified particles and ciliates ranged, respectively, from 0.01 to 456.7; from 12.9 to 1093; and from 0.01 to 183.6 mg C m^{-3}. Mesozooplankton biomass varied from 0.01 to 68.8 mg C m^{-3} (Table I). The distributions of all the four plankton groups presented differences between the three cruises (Fig. 4).

In 2004, diatoms showed an uneven distribution covering some small areas near Arcachon Bay, in the Cantabrian coast and over the shelf break. On the 2005 cruise, the diatom biomass had diminished significantly; in 2006, they reached their highest levels, covering a small area associated with the Landes upwelling and Gironde river mouth.

Unidentified particles showed maximum values of biomass during the 2004 cruise. They extended over the entire French shelf, with several areas of higher biomass associated with the Gironde river plume, the southeast

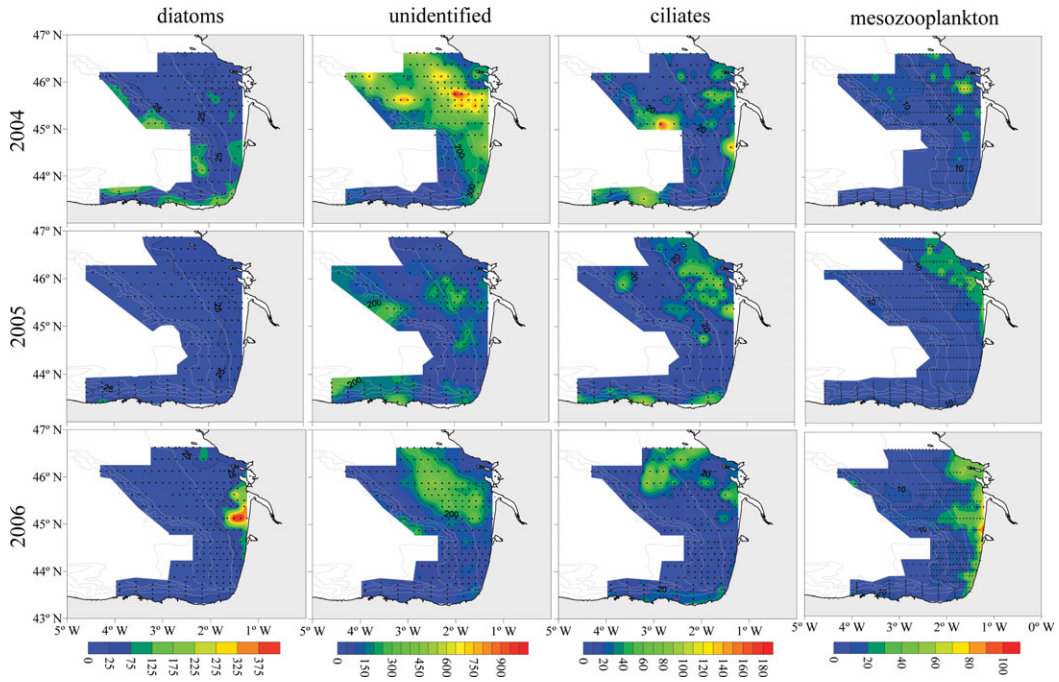

Fig. 4. Spatial distribution of diatom chains, unidentified particles, ciliates and mesozooplankton, for the cruises in 2004, 2005 and 2006. The colour spectra represent biomass (mg C m^{-3}; note the differing scales). The sampling grid is plotted and bathymetry is similar to Fig. 1.

part of the shelf (*Côte des Landes*) and the shelf-break. In 2005, biomass was significantly lower, and concentrated in small areas in front of the Gironde, at the shelf-break and along the Cantabrian coast. Finally, during the 2006 cruise, the biomass of unidentified particles was concentrated in the plume waters of the Gironde and extended towards the north French shelf.

During the 2004 cruise, ciliates also showed a patchy distribution, similar to that of the diatom chains. In 2005 and 2006, a higher biomass of ciliates was found in front of the Gironde estuary, on the northern French shelf and along the Cantabrian coast.

Mesozooplankton was distributed near the French coast during all the cruises, with maximum levels located near the Gironde river mouth and along the northern French coast. In 2006, the biomass of mesozooplankton was also high along the southern French coast, from the Gironde to the Adour estuaries. In 2004 and 2006, an area of slightly increased biomass extended towards the outer shelf and the shelf-break.

GAMs based on single explanatory variables

The geographical location of each sample (expressed as a two-dimensional spline of latitude and longitude) was shown to be the most relevant covariate for all of the plankton categories, achieving the highest levels of deviance explained and the lowest GCV, as well as the highest JK-r^2 (Table V). Using this single explanatory factor, 47.5% of the deviance in unidentified particles distribution was described; 38% for diatoms, 32.2% for mesozooplankton and 10.5% for ciliates. For the diatom chains, the depth of the water column (expressed on a $\log_{10}$ scale), together with the temperature, was the following factors in order of relevance, explaining 19.5 and 19.8% of their distribution. The depth of the water column also explained the considerable levels of deviance in the mesozooplankton biomass distribution (25.7%). For unidentified nano–microplankton, salinity and stratification were also significant covariates, explaining 27.9 and 25.3% of distribution of biomass. The ciliates were the group that was most poorly described by physical factors.

Biological variables explained less of the variability in plankton biomass distribution than physical–geographical parameters. The biomass of ciliates explained 11.4% of the deviance in the distribution of diatoms and 10.7% of the distribution of unidentified particles. For the two zooplankton categories (ciliates and mesozooplankton), diatoms were the only significant variable, explaining 6.5 and 8.2% of the deviance in their biomass distribution, respectively.

GAMs based upon combined explanatory variables

GAMs based upon multiple explanatory variables were established. The results of the GAM models are shown

Table V: Single variable-based GAMs, for diatom chains, unidentified particles, ciliates and mesozooplankton biomass

	Diatom chains				Unidentified particles				Ciliates				Mesozooplankton			
Models	% Dev	GCV	JK-r ^{2}	JK-p	% Dev	GCV	JK-r ^{2}	JK-p	% Dev	GCV	JK-r ^{2}	JK-p	% Dev	GCV	JK-r ^{2}	JK-p
Hydro-geographic																
∼*s* (S)	10.7	0.291	0.10	<0.001	27.9	0.095	0.26	<0.001	5.68	0.176	0.05	<0.001	12.4	0.072	0.09	<0.001
∼*s* (T)	19.8	0.265	0.18	<0.001	20.2	0.105	0.19	<0.001	3.41	0.181	0.02	<0.001	4.2	0.077	0.02	<0.001
∼*s* (Str)	15.8	0.280	0.13	<0.001	25.3	0.099	0.23	<0.001	7.67	0.175	0.03	<0.001	4.56	0.076	0.03	<0.001
∼*s* (D)	19.5	0.267	0.17	<0.001	21	0.105	0.18	<0.001	6.64	0.177	0.04	<0.001	25.7	0.060	0.23	<0.001
∼*s*(latitude and longitude)	38	0.219	0.32	<0.001	47.5	0.074	0.41	<0.001	10.5	0.175	0.05	<0.001	32.2	0.058	0.26	<0.001
Biologic																
∼*s* (Di)		–				n.s.			6.52	0.175	0.05	<0.001	8.22	0.075	0.05	<0.001
∼*s* (Unid)		n.s.				–				n.s.				n.s.		
∼*s* (Cil)	11.4	0.291	0.09	<0.001	10.7	0.117	0.09	<0.001		–				n.s.		
∼*s*(Mzoo)	7.6	0.304	0.05	<0.001		n.s.				n.s.				–		

For each model, the percentage of deviance explained (Dev. %), the generalized cross validation (GCV) score, and the parameters of the Jackknife procedure, r^2 (JK-r^2) and *P*-value (JK-*p*) are given. The smoothing functions are referred to as *s*(). Explanatory variables used in each model appear in brackets: S, surface salinity; T, surface temperature; Str, stratification index; D, water depth; (Lat, Long), latitude and longitude; Di, diatom biomass; Unid, unidentified particles biomass; Cil, ciliate biomass; Mzoo, mesozooplankton biomass. n.s. correspond to non-significant terms ($P > 0.001$).

as plots of the best-fitting smooths for the conditional effect of the covariates on the parameter of interest, i.e. plankton biomass (Figs 5–8). The *y*-axis reflects the influence of each covariate on plankton biomass, given that the other variables are included in the model (Maravelias *et al.*, 2000). The influence of the two-dimensional spline, *s* (latitude and longitude), is represented in a surface plot. Approximate confidence curves have been obtained for each univariate function (Figs 5–8).

Diatoms

Using the stepwise elimination method, five variables were considered significant enough to model the distribution of diatom chain biomass in the Bay of Biscay. This model achieved a 60.2% of deviance explained with a JK-r^2 of 0.52 (Table VI). Variables were ranked on the basis of their contribution to improve the fit of the model: geographical location (latitude and longitude), surface temperature, stratification index, the

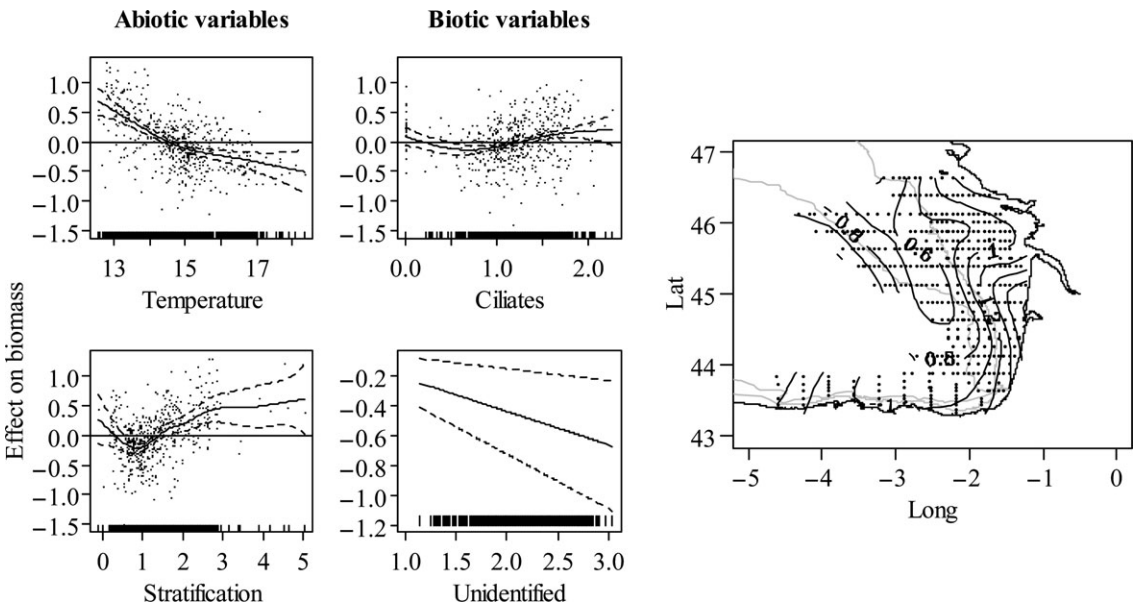

Fig. 5. Output of the generalized additive models (GAMs) are given in italics in Table VI, for the diatom chains. The partial effects of each individual covariate (surface temperature, stratification, biomass of ciliates, biomass of unidentified particles) are plotted as smoothed fits. Broken lines correspond to 2 standard errors, above and below the estimate of the smooth being plotted. Short vertical lines located on the *x*-axes of each plot indicate the values at which observations were made. The partial effect of the two-dimensional geographic term (latitude and longitude) is represented in a surface plot.

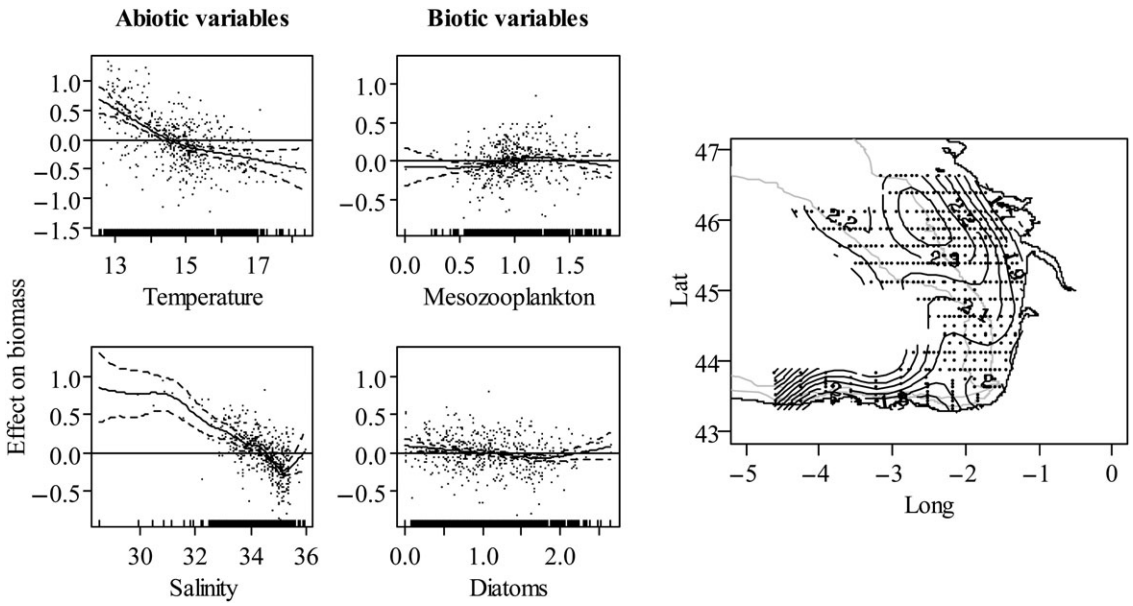

Fig. 6. Output of the GAMs are given in italics in Table VI, for the unidentified particles. The partial effects of each individual covariate (surface temperature, surface salinity, biomass of mesozooplankton, biomass of diatom chains) are plotted as smoothed fits. The partial effect of the two-dimensional geographic term (latitude and longitude) is represented in a surface plot. Plot annotation is similar to Fig. 5.

biomass of ciliates and the biomass of unidentified particles.

The biomass of diatom chains was higher near the French and the Cantabrian coast, with a second maximum at the shelf-break (Fig. 5). Moreover, the highest levels of biomass were related to the lowest values of temperature and the highest levels of stratification. Although biological variables had a weaker influence on diatom biomass than physical factors, a direct relationship with ciliates and an inverse relationship with unidentified plankton can be observed.

Unidentified particles

The GAM established for unidentified particles, explained the highest percentage of the deviance of biomass distribution (65.8%), with a JK-r^2 of 0.58 (Table VI). The variables selected to build the model were, in order of relevance: geographical location (latitude and longitude), surface temperature, surface salinity, mesozooplankton biomass and diatom biomass.

The surface plot showed that the highest biomass of unidentified particles was found over the central French shelf, at the same latitude as the Gironde, and extending towards the north (Fig. 6). A second biomass maximum was found on the west Cantabrian coast and over the shelf-break. The biomass of unidentified particles was at the lowest of temperatures (<15°C). Lower salinities are also related to areas of higher biomass, although a slight increase in biomass was also observed for higher salinity waters. Once again, the relationship between the biomass of unidentified particles and the biological variables was weak. However, an increase in the biomass of ciliates can be observed, with increasing mesozooplankton and decreasing diatom biomass.

Ciliates

The final GAM for ciliates explained 22.7% of the deviance of the biomass distribution (Table VI). The selected variables were: geographical location (latitude and longitude), surface temperature, biomass of unidentified particles, depth, biomass of diatoms and biomass of mesozooplankton.

Areas of enhanced ciliate biomass were located in front of the bay of Arcachon and over the northern French shelf (Fig. 7). Two more areas of higher biomass were observed for the west Cantabrian coast and over the shelf break. Higher biomass was associated with shallower water depths and higher temperatures. Biomass of ciliates peaked at intermediate values of unidentified plankton. It also showed a direct relationship with the biomass of diatom chains, and an inverse relationship with mesozooplankton.

Mesozooplankton

Five covariates were selected to establish the final GAM for mesozooplankton: geographical location (latitude

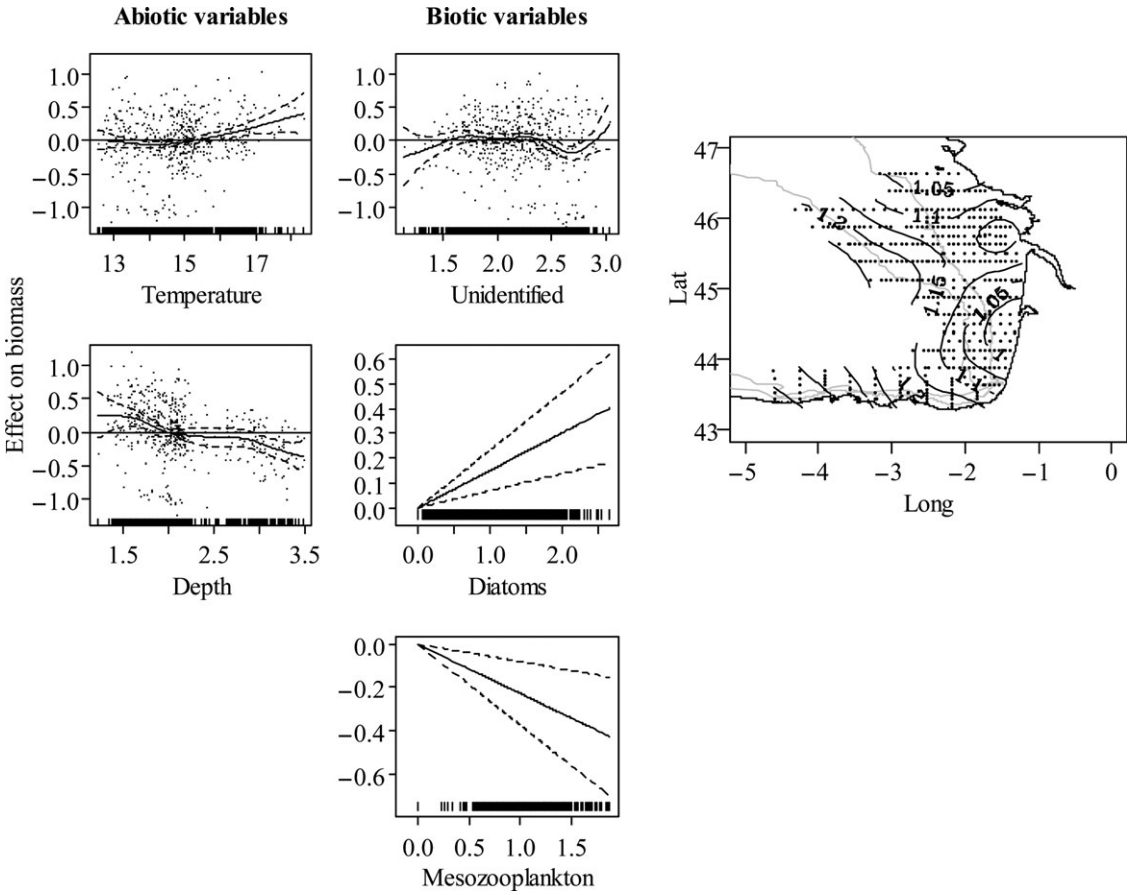

Fig. 7. Output of the GAMs are given in italics in Table VI, for the ciliates. The partial effects of each individual covariate (surface temperature, depth, biomass of unidentified particles, biomass of diatom chains, biomass of mesozooplankton) are plotted as smoothed fits. The partial effect of the two-dimensional geographic term (latitude and longitude) is represented in a surface plot. Plot annotation is similar to Fig. 5.

and longitude), surface salinity, surface temperature, depth and biomass of ciliates. Added to the same model, these variables explained 43.7% of the distribution of mesozooplankton and achieved a JK-r^2 of 0.31 (Table VI).

Mesozooplankton biomass was higher near the French coast, with two areas of enhanced biomass associated with the Gironde and the Adour estuaries. A third area of higher biomass was described for the shelf-break (Fig. 8). The biomass of mesozooplankton showed a direct relationship with salinity and temperature, and highest levels were found at the shallowest depths. Finally, higher biomasses of mesozooplankton were associated with low ciliate biomass.

The final models were applied to the original data set. The plot made with the estimated plankton biomass for each cruise agreed individually with the observed biomass distribution. For all of the plankton categories, the models highlighted the main high biomass areas described above, although the size of the patches was not always accurate (Fig. 9).

DISCUSSION

To our knowledge, this is the first time that mesoscale-resolution distribution of ciliate biomass has been described in repeated large-scale surveys. In the present study, an overall description of the distribution of plankton has been achieved, covering a broad range of sizes, from nano- to mesozooplankton. In the nano–microplankton size range, three broad groups of ecological significance were distinguished (diatom chains, ciliates and unidentified particles) with levels of accuracy higher than 95%. Both physical and plankton sampling were undertaken at a relevant spatial resolution. The average distance between stations was <10 nautical miles, which has been considered enough to detect the main mesoscale structures present in the Bay of Biscay, and to match physical and biological processes (Gil *et al.*, 2002; Albaina and Irigoien, 2004). The results obtained indicate that imaging technology, combined with automatic recognition techniques, has the potential to enable surveys of micro and

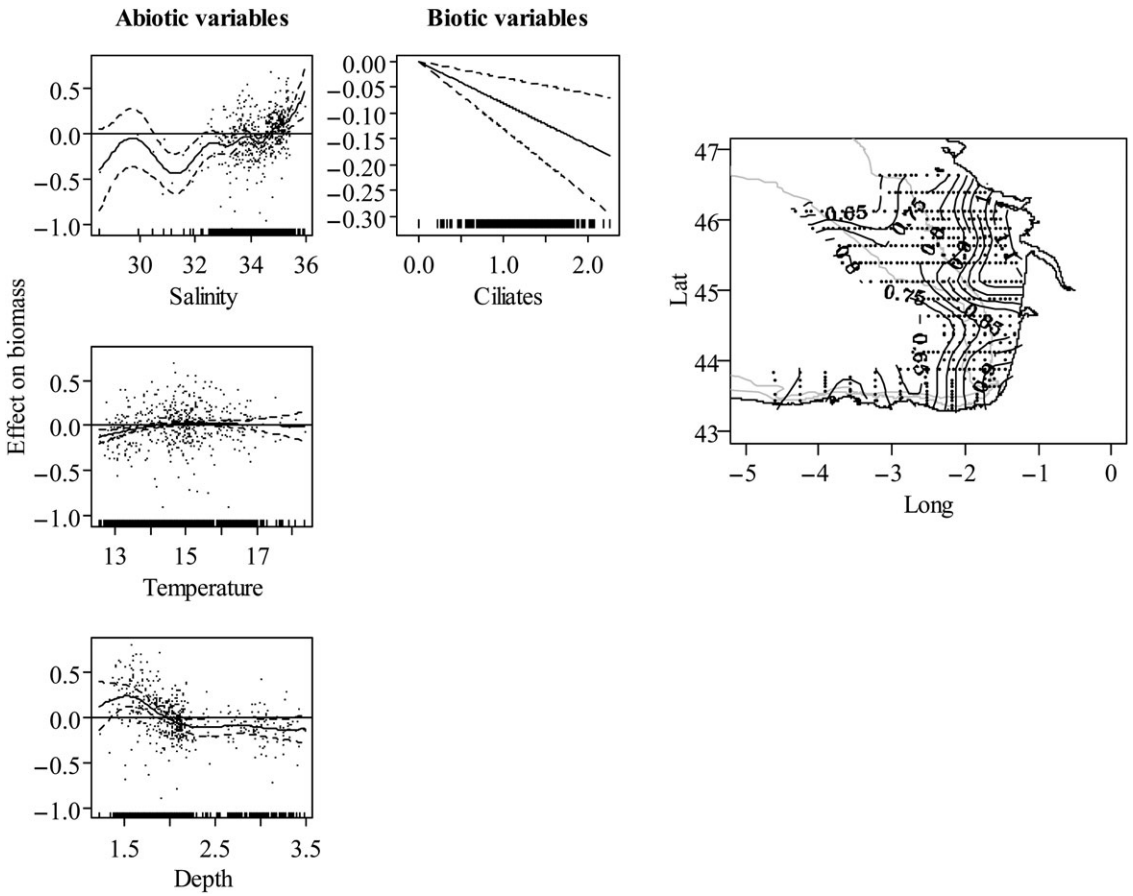

Fig. 8. Output of the GAMs are given in italics in Table VI, for the mesozooplankton. The partial effects of each individual covariate (surface salinity, surface temperature, depth, biomass of ciliates) are plotted as smoothed fits. The partial effect of the two-dimensional geographic term (latitude and longitude) is represented in a surface plot. Plot annotation is similar to Fig. 5.

mesozooplankton to be conducted on an equal level of spatial resolution compared with those of physics and phytoplankton.

GAM method has proved to be a useful tool to use in describing the interactions between plankton biomass and physical and biological variables, in order to produce models with a good descriptive capacity (Table VI). GAMs based on multiple covariates were able to describe a significant percentage of the deviance existing in plankton biomass distribution, and were validated by low GCV and high JK-r^2 scores (Table VI). The variability in the distribution of unidentified particles biomass was explained at 66%, and that of diatom chains at 60%. Mesozooplankton and ciliate distributions were described by GAMs at 44 and 23%, respectively. All four models used included a combination of hydrographic, geographic and biological terms; they all account for a considerable proportion of the variability in the biomass distributions.

Furthermore, being able to sample different trophic levels at the same spatial resolution as the physical variables is essential for a better understanding of the factors controlling plankton distribution. It permits the evaluation of the relative importance of physical conditions versus food, as well as bottom-up versus top-down mechanisms. In the present study, the contribution of the biological variables was less than the information described in hydro-geographical terms, for all of the plankton categories (Table V). However, the type of relationship between hydro-geographical variables and biomass differed for the plankton groups considered. Interestingly, for all of the identified plankton groups, geographical location (latitude and longitude) was the main explanatory factor. This reveals that, at this scale, the distribution is not determined by the physical parameters *per se*, but rather by the hydrographical structures. In the Bay of Biscay, the most relevant physical fronts are persistent structures, related to geographical features, such as the plume of the Gironde and the Adour rivers; the shelf-break (Pingree and Mardell, 1981; Pingree *et al.*, 1986; New, 1988) and occasional upwelling events located along the *Côte des*

Table VI: Multiple, variable-based GAMs for diatom chains, unidentified particles, ciliates and mesozooplankton biomass

GAMs based upon multiple covariates	% Dev	GCV	JK-r^2	JK-p
Diatom chains				
Response: log(Di+1)				
Terms				
~*s* (latitude and longitude)	38	0.2195	0.316	<0.001
~*s* (latitude and longitude)+*s* (T)	52.5	0.1700	0.469	<0.001
~*s* (latitude and longitude)+*s* (T)+*s* (Str)	56.7	0.1587	0.497	<0.001
~*s* (latitude and longitude)+*s* (T)+*s* (Str)+*s* (Cil)	59	0.1522	0.515	<0.001
~s (latitude and longitude)+s (T)+s (Str)+s (Cil)+Unid	*60.2*	*0.1492*	*0.523*	*<0.001*
Unidentified particles				
Response: log(Unid +1)				
Terms				
~*s* (latitude and longitude)	47.5	0.0744	0.414	<0.001
~*s* (latitude and longitude)+*s* (T)	57.3	0.0625	0.509	<0.001
~*s* (latitude and longitude)+*s* (T)+*s* (S)	63.4	0.0550	0.571	<0.001
~*s* (latitude and longitude)+*s* (T)+*s* (S)+*s* (Mzoo)	64.7	0.0537	0.580	<0.001
~s (latitude and longitude)+s (T)+s (S)+s (Mzoo)+s (Di)	*65.8*	*0.0529*	*0.584*	*<0.001*
Ciliates				
Response: log(Cil+1)				
Terms				
s (latitude and longitude)	10.5	0.1751	0.048	<0.001
~*s* (latitude and longitude)+*s* (T)	16.2	0.1691	0.080	<0.001
~*s* (latitude and longitude)+*s* (T)+(Di)	18.7	0.1650	0.104	<0.001
~*s* (latitude and longitude)+*s* (T)+(Di)+*s* (Unid)	20	0.1625	0.111	<0.001
~*s* (latitude and longitude)+*s* (T)+(Di)+*s* (Unid)+*s* (D)	20	0.1599	0.128	<0.001
~s (latitude and longitude)+s (T)+(Di)+s (Unid)+s (D)+Mzoo	*22.7*	*0.1577*	*0.124*	*<0.001*
Mesozooplankton				
Response: log(Mzoo +1)				
Terms				
~*s* (latitude and longitude)	32.2	0.0579	0.257	<0.001
~*s* (latitude and longitude)+*s* (S)	37.4	0.0553	0.273	<0.001
~*s* (latitude and longitude)+*s* (S)+*s* (T)	40.8	0.0536	0.294	<0.001
~*s* (latitude and longitude)+*s* (S)+s(T)+*s* (D)	42.6	0.0527	0.304	<0.001
~s (latitude and longitude)+s (S)+s(T)+s (D)+Cil	*43.7*	*0.0518*	*0.309*	*<0.001*

For each model, the percentage of deviance explained (Dev. %), the generalized cross validation (GCV) score, and the parameters of the Jackknife procedure, r^2 (JK-r^2) and *P*-value (JK-*p*), are given. In the formulation of GAM models, linear coefficients are omitted and smooth functions are referred to as *s*(). Explanatory variables used in each model appear in brackets. Abbreviations are similar to those listed in Table V. Models analysed in the Discussion are given in italics.

Landes (Castaing and Lagardere, 1983; Jegou and Lazure, 1995). These fronts are environments of enhanced plankton productivity (Le Fevre, 1986) and were easily reproduced by the models (Fig. 9). To describe plankton distribution at these locations, the presence of geographically defined gradients (fronts) was more relevant than the absolute values of each variable. This is important because plankton distribution is usually modelled on the basis of nutrients, light, temperature and salinity; and our results show that the simple inclusion of a geographical term could allow definition of areas of enhanced biomass associated with hydrographic fronts.

On the other hand, the relevance of temperature and salinity, as highlighted by the GAMs (Tables V and VI), is also related to the dynamics of the fronts. Fronts are regions of greater than average horizontal gradients of water properties such as temperature, salinity, density, turbidity or ocean colour (Joyce, 1983). In the study area, they are defined by the colder, low-salinity waters of the river outflows, together with the deep cold waters upwelled near the coast.

For both the diatoms and non-identified particles, geographical location and physical characteristics were the main explanatory factors. The results suggest that during this study period, phytoplankton, in particular diatoms, was mainly bottom-up controlled with little impact from ciliates and mesozooplankton. This pattern is not surprising, as complex food webs tend to dampen trophic cascades (Finke and Denno, 2004, 2005). On the other hand, ciliates and mesozooplankton show a lower dependence on hydrographic variables. In addition to the physical and chemical properties of the environment, spatial heterogeneity of zooplankton is largely determined by interactions between individuals as well as to the reactions of the organisms to their

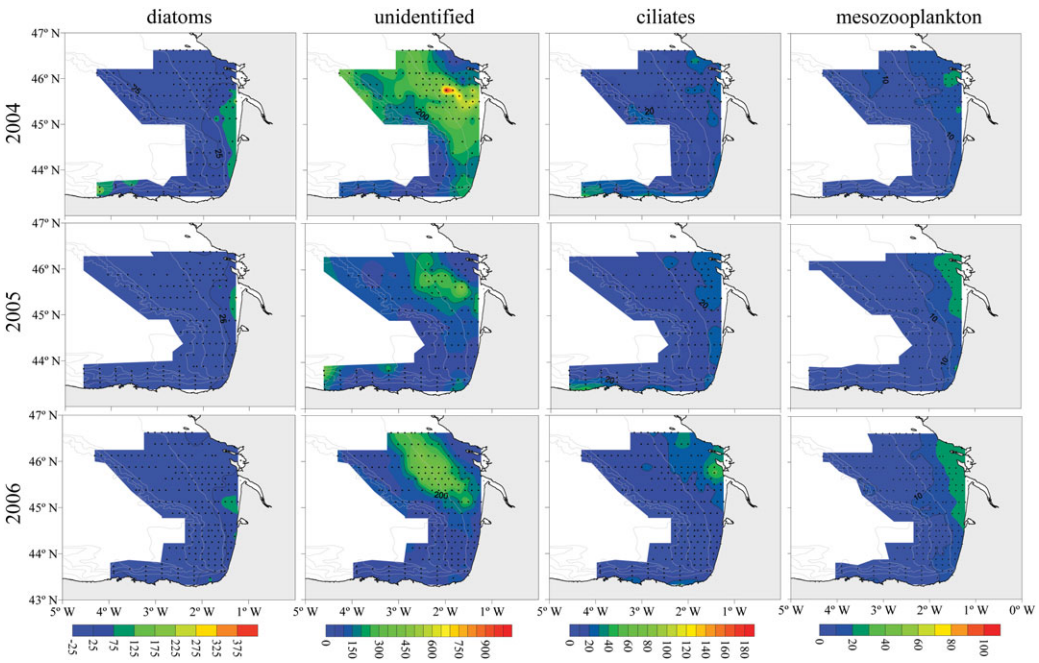

Fig. 9. Modelled distribution of diatom chains, unidentified particles, ciliates and mesozooplankton, for the cruises in 2004, 2005 and 2006. The colour spectra represent biomass (mg C m^{-3}; note the differing scales). Data are derived from the GAM outputs given in italics in Table VI, together with the original data sets. The sampling grid is plotted and bathymetry is similar to Fig. 1.

biological environment, including responses to patches of potential food organisms or to predators (Mauchline, 1998; Folt and Burns, 1999). Both behavioural (social and reproductive) and coactive processes (involving competition, predation and parasitism) need to be investigated in order to understand the development and maintenance of the distribution patterns observed in the ocean (Haury *et al.*, 1978).

Within this context, it should be noted that the influence of strictly hydrographic variables (surface salinity, surface temperature and stratification index), over the biomass of ciliates and mesozooplankton, was not significantly higher than the influence of biological variables (Table V). Moreover, the biological variables which explained the plankton distribution to a greater degree were the biomass of ciliates (explaining the distribution of diatoms and unidentified particles) and the biomass of diatoms (explaining the distribution of ciliates and mesozooplankton). All this highlights the influence of trophic predator–prey relationships, and indicates that biological information as a relevant variable to include in ecological models.

A noteworthy limitation to this study is that the categories into which plankton have been classified are very general; as such, they provide only a broad view of the ecological relationships existing between them. This makes it difficult to relate explanatory variables to the biomass distribution of plankton. Additionally, as only simple shape descriptors were used in the classification, difficulties were found to distinguish certain groups of similar size and shape, such as ciliates, large flagellates and circular diatoms. It may be possible to improve identification performance by using other contour representations, moment- and texture-based features (Blaschko *et al.*, 2005; Hu and Davis, 2006), as well as features specific to particular groups of organisms (Blaschko *et al.*, 2005). Better taxonomic grouping will provide an improved understanding of the functioning of the ecosystem and probably, increased predictive habitat models.

Additionally, the biological variables used to describe the distribution of plankton are just a snapshot of the existing biomass of a given group. Variables providing a deeper insight into biological processes, such as size or behaviour, could provide relevant information to improve our understanding of the spatial distribution of plankton. Additional information on nutrient concentration will also help to describe the spatial patterns of phytoplankton biomass.

Finally, it must be taken into account that statistical models have limited predictive ability outside of out of the range of measurements. However, they are not more limited than semi-mechanistic models, which are usually internally parameterized or optimized with empirical relations (Flynn, 2003; Mitra *et al.*, 2007).The present study covers three consecutive years, which is

not sufficient to describe interannual variability in plankton communities. Such variability is large for plankton biomass (Karentz and Smayda, 1998; Yallop, 2001), and is often more significant than any seasonal variability (Mercado *et al.*, 2005). If large-scale information could be obtained with sufficient resolution, and over long time periods, the envelop of the statistical model would increase, and this, in turn, would improve the usefulness of the predictive habitat models to levels similar to those attained in terrestrial ecology (Guisan and Zimmermann, 2000).

ACKNOWLEDGEMENTS

We are grateful to the captains and crews of the research vessel R/V Vizconde de Eza, and to the on-board scientists and analysts, for their support during the sampling and *in situ* FlowCAM analysis. Very special thanks are due to I. Gomez and N. Serrano, who analysed mesozooplankton samples in the lab, and to A. Urtizberea, for her help and advice with statistics. We also thank Professor M. Collins (SOES, University of Southampton and AZTI Tecnalia) for critical comments on the manuscript, and the three anonymous reviewers for their useful comments. This is contribution no. 403 from AZTI-Tecnalia (Marine Research).

FUNDING

This research was supported by several projects funded by the Spanish Ministry of Education and Research, and the Department of Agriculture and Fisheries of the Basque Country Government (BIOMAN, EIPZI). L.Z.'s work was supported by a Doctoral Fellowship of the Education, Universities and Research Department of the Basque Country Government. J.A.F. was supported by a doctoral fellowship from the Fundación Centros Tecnológicos Iñaki Goenaga.

REFERENCES

Abraham, E. R. (1998) The generation of plankton patchiness by turbulent stirring. *Nature*, **391**, 577–580.

Albaina, A. and Irigoien, X. (2004) Relationships between frontal structures and zooplankton communities along a cross shelf transect in the Bay of Biscay (1995 to 2003). *Mar. Ecol. Prog Ser.*, **284**, 65–75.

Alcaraz, M., Saiz, E., Calbet, A. *et al.* (2003) Estimating zooplankton biomass through image analysis. *Mar. Biol.*, **143**, 307–315.

Ashjian, C. J., Davis, C. S., Gallager, S. M. *et al.* (2001) Distribution of plankton, particles, and hydrographic features across Georges Bank described using the Video Plankton Recorder. *Deep-Sea Res. II*, **48**, 245–282.

Augustin, N. H., Borchers, D. L., Clarke, E. D. *et al.* (1998) Spatiotemporal modelling for the annual egg production method of stock assessment using generalized additive models. *Can. J. Fish. Aquat. Sci.*, **55**, 2608–2621.

Beare, D. J., Gislason, A., Astthorsson, O. S. *et al.* (2000) Assessing long-term changes in early summer zooplankton communities around Iceland. *ICES. J. Mar. Sci.*, **57**, 1545–1561.

Benfield, M. C., Grosjean, P., Culverhouse, P. *et al.* (2007) Research on automated plankton identification. *Oceanography*, **20**, 12–26.

Blaschko, M., Holness, G., Mattar, M. *et al.* (2005) Automatic in situ identification of plankton. In Proceedings of IEEE Workshop on Applications of Computer Vision.

Breiman, L. (2001) Random forests. *Machine Learning*, **45**, 5–32.

Castaing, P. and Lagardere, F. (1983) Seasonal water temperature and salinity variations off the Aquitaine continental shelf. *Bull. Inst. Geol. Bass. Aquitaine*, **33**, 61–69.

Culverhouse, P., Williams, R., Reguera, B. *et al.* (2003) Do experts make mistakes? A comparison of human and machine indentification of dinoflagellates. *Mar. Ecol. Prog. Ser.*, **247**, 17–25.

Culverhouse, P. F., Williams, R., Benfield, M. *et al.* (2006) Automatic image analysis of plankton: future perspectives. *Mar. Ecol. Prog. Ser.*, **312**, 297–309.

Daskalov, G. (1999) Relating fish recruitment to stock biomass and physical environment in the Black Sea using generalized additive models. *Fish. Res.*, **41**, 1–23.

Davis, C., Hu, Q., Gallager, S. *et al.* (2004) Real-time observation of taxa-specific plankton distributions: an optical sampling method. *Mar. Ecol. Prog Ser.*, **284**, 77–96.

Davis, C. S., Thwaites, F. T., Gallager, S. M. *et al.* (2005) A three axis fast-tow digital Video Plankton Recorder for rapid surveys of plankton taxa and hydrography. *Limnol. Oceanogr. Methods*, **3**, 59–74.

Duarte, C. M. (2007) Marine ecology warms up to theory. *Trends Ecol. Evol.*, **22**, 331–333.

Finke, D. L. and Denno, R. F. (2004) Predator diversity dampens trophic cascades. *Nature*, **429**, 407–410.

Finke, D. L. and Denno, R. F. (2005) Predator diversity and the functioning of ecosystems: the role of intraguild predation in dampening trophic cascades. *Ecol. Lett.*, **8**, 1299–1306.

Flynn, K. J. (2003) Modelling multi-nutrient interactions in phytoplankton; balancing simplicity and realism. *Prog. Oceanogr.*, **56**, 249–279.

Folt, C. L. and Burns, C. W. (1999) Biological drivers of zooplankton patchiness. *Trends Ecol. Evol.*, **14**, 300–305.

Friedman, N., Geiger, D. and Goldszmidt, M. (1997) Bayesian network classifiers. *Machine Learning*, **29**, 131–163.

Gil, J., Valdés, L., Moral, M. *et al.* (2002) Mesoscale variability in a high-resolution grid in the Cantabrian Sea (southern Bay of Biscay), May 1995. *Deep-Sea Res. II*, **49**, 1591–1607.

Grosjean, P., Picheral, M., Warembourg, C. *et al.* (2004) Enumeration, measurement, and identification of net zooplankton samples using the ZOOSCAN digital imaging system. *ICES. J. Mar. Sci.*, **61**, 518–525.

Guisan, A. and Zimmermann, N. E. (2000) Predictive habitat distribution models in ecology. *Ecol. Model.*, **135**, 147–186.

Guisan, A., Edwards, T. C. and Hastie, T. (2002) Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecol. Model.*, **157**, 89–100.

Hastie, T. J. and Tibshirani, R. J. (1990) *Generalized Additive Models*. Chapman and Hall, London.

Haury, L. R., McGowan, J. A. and Wiebe, P. H. (1978) Patterns and processes in the time-space scales of plankton distributions. In Steele, J. H. (eds), *Spatial Pattern in Plankton Communities*. Plenum Press, New York, US, pp. 277–327.

Hirzel, A. H., Le Lay, G., Helfer, V. *et al*. (2006) Evaluating the ability of habitat suitability models to predict species presences. *Ecol. Model.*, **199**, 142–152.

Hu, Q. and Davis, C. (2006) Accurate automatic quantification of taxa-specific plankton abundance using dual classification with correction. *Mar. Ecol. Prog. Ser.*, **306**, 51–61.

Jegou, A. M. and Lazure, P. (1995) Quelques aspects de la circulation sur le plateau atlantique. In Cendrero, O. and Olaso, I. (eds), *Actas del IV Coloquio Internacional sobre Oceanografia del Golfo de Vizcaya*. Instituto Español de Oceanografía, Santander, pp. 99–106.

Joyce, T. (1983) Varieties of ocean fronts. In Stern, M. and Mellor, G. (eds), *Baroclinic Instability and Ocean Fronts. Technical Report*. Vol. 83–41. Woods Hole Oceanographic Institution, Woods Hole, WA.

Karentz, D. and Smayda, T. (1998) Temporal patterns and variations in phytoplankton community organization and abundance in Narragansett Bay during 1959–1980. *J. Plankton Res.*, **20**, 145.

Le Fevre, J. (1986) Aspects of the biology of frontal systems. *Adv. Mar. Biol.*, **23**, 163–299.

Levin, S. and Segel, L. (1976) Hypothesis for origin of planktonic patchiness. *Nature*, **259**, 659.

Lobo, J. M. and Martin-Piera, F. (2002) Searching for a predictive model for species richness of Iberian dung beetle based on spatial and environmental variables. *Conserv. Biol.*, **16**, 158–173.

Loukos, H., Monfray, P., Bopp, L. *et al*. (2003) Potential changes in skipjack tuna (*Katsuwonus pelamis*) habitat from a global warming scenario: modelling approach and preliminary results. *Fish. Oceanogr.*, **12**, 474–482.

Mann, K. H. and Lazier, J. R. N. (1991) *Dynamics of Marine Ecosystems*. Blackwell Science, Boston, MA.

Maravelias, C. D., Reid, D. G. and Swartzman, G. (2000) Modelling spatio-temporal effects of environment on Atlantic herring, *Clupea harengus*. *Environ. Biol. Fish.*, **58**, 157–172.

Martin, A. (2003) Phytoplankton patchiness: the role of lateral stirring and mixing. *Prog. Oceanogr.*, **57**, 125–174.

Mauchline, J. (1998) The biology of calanoid copepods. *Adv. Mar. Biol.*, **33**, 1–710.

Mercado, J. M., Ramirez, T., Cortes, D. *et al*. (2005) Seasonal and inter-annual variability of the phytoplankton communities in an upwelling area of the Alborán Sea (SW Mediterranean Sea). *Sci. Mar.*, **69**, 451–465.

Mitra, A., Flynn, K. and Fasham, M. (2007) Accounting for grazing dynamics in nitrogen–phytoplankton–zooplankton models. *Limnol. Oceanogr.*, **52**, 649–661.

Montagnes, D. J. S., Berges, J. A., Harrison, P. J. *et al*. (1994) Estimating carbon, nitrogen, protein and chlorophyll a from volume in marine phytoplankton. *Limnol. Oceanogr.*, **39**, 1044–1060.

New, A. L. (1988) Internal tidal mixing in the Bay of Biscay. *Deep-Sea Res. I*, **35**, 691–709.

Pingree, R. D. and Mardell, G. T. (1981) Slope turbulence, internal waves and phytoplankton growth at the Celtic Sea shelf-break. *Philos. Trans. R. Soc. Lond. A*, **302**, 663–682.

Pingree, R. D., Mardell, G. T. and New, A. L. (1986) Propagation of internal tides from the upper slopes of the Bay of Biscay. *Nature*, **321**, 154–158.

Planque, B., Bellier, E. and Lazure, P. (2007) Modelling potential spawning habitat of sardine (*Sardina pilchardus*) and anchovy (*Engraulis encrasicolus*) in the Bay of Biscay. *Fish. Oceanogr.*, **16**, 16–30.

See, J. H., Campbell, L., Richardson, T. L. *et al*. (2005) Combining new technologies for determination of phytoplankton community structure in the northern Gulf of Mexico. *J. Phycol.*, **41**, 305–310.

Sieracki, C. K., Sieracki, M. E. and Yentsch, C. S. (1998) An imaging-in-flow system for automated analysis of marine microplankton. *Mar. Ecol. Prog. Ser.*, **168**, 285–296.

Stratoudakis, Y., Bernal, M., Borchers, D. L. *et al*. (2003) Changes in the distribution of sardine eggs and larvae off Portugal, 1985–2000. *Fish. Oceanogr.*, **12**, 49–60.

Ward, P. J. and Myers, R. A. (2006) Do habitat models accurately predict the depth distribution of pelagic fishes? *Fish. Oceanogr.*, **15**, 60–66.

Witten, I. and Frank, E. (2005) *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco.

Wood, S. N. (2006) *Generalized Additive Models: an Introduction with R*. Chapman & Hall/CRC, FL.

Wood, S. N. (2008) Fast stable direct fitting and smoothness selection for generalized additive models. *J. R.I Statist. Soc. B*, **70**, 495–518.

Wood, S. N. and Augustin, N. H. (2002) GAMs with integrated model selection using penalized regression splines and applications to environmental modeling. *Ecol. Model.*, **157**, 157–177.

Yallop, M. L. (2001) Distribution patterns and biomass estimates of diatoms and autotrophic dinoflagellates in the NE Atlantic during June and July 1996. *Deep-Sea Res. II*, **48**, 825–844.

Yee, T. W. and Mitchell, N. D. (1991) Generalized additive models in plant ecology. *J. Veg. Sci.*, **2**, 587–602.

Zarauz, L., Irigoien, X., Urtizberea, A. *et al*. (2007) Mapping plankton distribution in the Bay of Biscay during three consecutive spring surveys. *Mar. Ecol. Prog. Ser.*, **345**, 27–39.