Prediction of the annual cycle of phytoplankton production in the North East Atlantic

E.D. Clarke[1], S.N. Wood[2], M.R. Heath[1], D.C. Speirs[3], W.S.C. Gurney[3] & S.J. Holmes[1]

[1] FRS Marine Laborartory, PO Box 101, 375 Victoria Road, Aberdeen, AB11 9DB, UK

[2] Department of Statistics, 15 University Gardens, University of Glasgow, Glasgow, G12 8QQ, UK

[3] Department of Statistics and Modelling Science, University of Strathclyde, Livingstone Tower, 26 Richmond Street, Glasgow, G1 1XH, UK

Abstract

SeaWiFS satellite measurements of ocean colour, available for the whole globe on a weekly basis, are routinely converted into estimates of chlorophyll a concentration using calibration algorithms. These calibration algorithms algorithms are validated using bottle measurements collected in clear, open ocean waters, where chlorophyll a concentrations are low. The North East Atlantic has large areas of coastal and polar waters, where these calibration algorithms do not work well. Thus, although the satellite measurements describe the broad-scale patterns in chlorophyll a concentration, quantitative agreement between satellite estimates and bottle measurements is poor. The actual relationship between bottle measurements and satellite estimates is complex and varies with water characteristics, which in turn vary with depth and time of year. We therefore model bottle measurements as a function of SeaWiFS estimates, depth and time of year, using three dimensional thin plate regression splines. The resulting model provides plausible predictions which capture the main features of both the satellite and bottle data.

1 Introduction

As in the terrestrial environment, the seasonal cycle of vegetation is the most conspicuous feature of the biota in the ocean. Even though the ocean vegetation is predominantly made up of microscopic unicellular algae (phytoplankton), these algae are clearly visible even from space by the colour that their photosynthetic pigments confer on the surface waters. The seasonality of phytoplankton abundance varies most strongly with latitude, and is governed by the balance between diffusion, grazing by planktonic herbivores and availability of nutrients and light to support photosynthesis. At high latitudes, appreciable concentrations of algae occur during a brief bloom in the spring or early summer, which follows the receding edge of the polar sea-ice (Engelsen *et al.*, 2002). In temperate latitudes the bloom is more protracted but centred around the spring and autumn, whilst in equatorial waters the seasonal variations are slight (Longhurst, 1998). The carbon fixed by the blooms of algae supports the pelagic food web in the upper ocean, and ultimately the fisheries and marine top-predators (Parsons *et al.*, 1984). Surplus algal production and waste from the

pelagic food web rains down to the seabed which may act as an important sink for carbon (Mann and Lazier, 1991). Changes in climate certainly affect ocean phytoplankton (Reid *et al.*, 1998), with implications for food web productivity, but phytoplankton may also be important regulators of global carbon dioxide and hence future climate through their role in the carbon sink. Hence, there is considerable interest in mapping and monitoring changes in ocean phytoplankton at local, regional, ocean basin and global scales.

Most measurements of phytoplankton abundance are made not in terms of cell numbers or carbon biomass, which is difficult to discriminate from non-living or non-algal carbon-rich particles in the water, but in terms of phytosynthetic pigment content — in particular chlorophyll *a* which is endemic across all taxonomic groups of algae. For direct measurements, fine particulate material in water bottle samples, including algal cells, is collected on glass fibre filters and chlorophyll extracted into an organic solvent which is then analysed chromatographically or spectrophotometrically (Strickland and Parsons, 1972). Such analysis of water samples collected by ships or buoys provides accurate data but there is little prospect of collecting sufficient data in space and time to fully characterise the dynamics of algal abundance even on a regional scale, far less on an ocean or global scale. An alternative is to utilise the fluorescent properties of chlorophyll to provide a less accurate in situ estimate of concentration without solvent extraction (Yentsch and Phinney, 1985). This allows for continuous monitoring by towed or moored in situ sensors or in pumped flow-lines. However, this still does not materially improve the scope for matching the sampling to the scale of the problem.

The spectral properties of light reflected from the sea surface can also be used to estimate the concentration of algal pigments in the water (Sathyendranath *et al.*, 1994). The advantage of exploiting reflectance is that data can be collected remotely by airborne or orbiting satellite-borne sensors, providing sampling at scales and resolutions relevant to the problem, subject to cloud cover. The Coastal Zone Colour Scanner (CZCS, 1978–1986), was the first satellite borne ocean reflectance sensor. More recently, the Sea-viewing Wide Field-of-view Sensor (SeaWiFS), launched by NASA in 1997 continues to provide global high resolution data. The algorithms used to derive chlorophyll concentration from reflectance data are calibrated by careful comparison of simultaneous overhead and ground-truth measurements, the primary data source being MOBY, a fixed optical mooring off the west coast of Lanai (Hawaii) (Hooker and McClain, 2000). The radiance measurements received by the satellite are corrected for back-scattering caused by air molecules and other particles in the atmosphere, and for reflections from glint and foam (atmospheric correction), to obtain water-leaving radiance, from which chlorophyll *a* concentration can be estimated using bio-optical algorithms. Both these calibrations are based on clear open ocean (case I) waters (Hooker and McClain, 2000; O'Reilly *et al.*, 2000) and so do not take into account near infrared reflectance from suspended sediment in turbid coastal (case II) waters or the effect of coloured organic matter dissolved in the water (gelbstoff).

These particles can have a substantial effect on ocean colour in coastal and polar waters, and the algorithms do not perform well in these areas (Dierssen and Smith, 2000; Burenkov *et al.*, 2001; Sathyendranath *et al.*, 2001). There is no explicit methodology for correcting reflectance-derived chlorophyll measurements for such effects. Alternatives have been to apply region-specific algorithms to the reflectance data based on the relative optical properties of sediment, chlorophyll and gelbstoff (*e.g.* Burenkov *et al.*, 2001; Westbrook *et al.*, 2001; Chen *et al.*, 2002), or employ a statistical method to 'blend' together satellite-derived and bottle-derived data so that the two are consistent over time and space (Gregg *et al.*, 2001; Gregg *et al.*, 2002). Few alternatives to the 'blending' methodology have so far been explored. In this paper we describe the novel application of thin plate regression splines for this purpose. We model sparsely distributed chlorophyll data from water sampling conducted over a number of years in the northeast Atlantic (NEA), with multi-annual composite SeaWiFS chlorophyll data as one of three covariates.

2 Water sampling (bottle) data

Point measurements of chlorophyll *a* concentration (mg m^{-3}) in the NEA (56°N–72°N, 30°W–20°E) from analyses of aqueous acetone or methanol extracted pigments in water bottle samples identified by their date, latitude, longitude and depth of collection, were gathered from a variety of sources. Anonymous data collected between 1960 and 1999 were obtained from the ICES Oceanographic Data Centre in Copenhagen and the British Oceanographic Data Centre. Other data were obtained from databases of the EU-ICOS, EU-TASC, and EU-ESOP2 projects, and from institutional records at FRS Aberdeen.

The entire data set available to us contained about 65,000 bottle measurements dating back to 1960, sampled at different depths at about 18,500 space-time locations (stations). We were interested in the average concentration in the top 5 m, so stations with no samples in the top 5 m were removed, leaving approximately 17,000 stations. Since environmental conditions have changed over the past 40 years, we removed observations collected before 1986, leaving 13,000 stations. Sampling has increased since the late 1980's and so these data retain most of the spatial coverage present in the full data set. We used the trapezium rule to estimate the concentration in the top 5 m, linearly interpolating between samples (Fig. 1). If there was only one sample in the top 30 m, and it was between [0,5] m, then this was taken as the average concentration in the top 5 m. If there were samples at 0 m and 5 m, we linearly interpolated between these and any samples in between, then integrated under the resulting polygon. If there was no sample at 0 m, we extrapolated from the nearest two samples to estimate the value at 0 m before integrating. If the extrapolation resulted in a negative concentration, this was set to zero. If there was no sample at 5 m, but one was available at a deeper depth (up to 30 m), then we interpolated between this sample and the first sample shallower than 5 m before integrating. If no sample was taken below 5 m, then we extrapolated from the nearest two samples, as for

zero depths.

The water bottle data set included data from several time series studies, in which sampling was carried out repeatedly at the same location thereby forming a detailed record of the seasonal cycle. The data from two of these locations, Stonehaven and Ocean Weatherstation (OWS) Mike, are referred to later in more detail. The sampling site at Stonehaven lies approximately 6 km off the northeast coast of Scotland in the North Sea (56°58′N, 02°06′W), and samples have been collected weekly since January 1997 (Heath *et al.*, 1999). The vessel MV Polarfront was on station at OWS Mike in the Norwegian Sea (66°N, 2°E) throughout 1997, and daily sampling was carried out by personnel involved in the EU-TASC project for most of the year (Heath *et al.*, 2000).

3 SeaWiFS Data

Output from the 2002 NASA reprocessing of the SeaWiFS data archive were compiled by staff at Plymouth Marine Laboratory into a multi-annual (1997-2002) average data set for the northeast Atlantic area. Valid SeaWiFS data were averaged over approximately 5′ longitude × 5′ latitude cells and successive 8-day intervals during the year.

These raw SeaWiFS data were stored as integer values for each pixel, v, on [0,255], which are converted to SeaWiFS predictions of chlorophyll concentration, z, in mg m^{-3}, by the formula: $z = 10^{0.015v-2}$. The raw data had been preprocessed on the v scale to fill in missing values and form climatological averages. However, this still left many missing values in the North in Winter, due to low sun angles. We therefore interpolated to fill in the missing values using a nearest neighbour method. We first checked for valid pixels in the four neighbouring pixels. Any valid pixels were averaged. If none were found, the next nearest neighbours surrounding these pixels were searched, and so on, in increasing circles up to 7 pixels from the original missing value. If no valid pixels were found, the data for the previous 8-day period were searched in the same way, and if still none were found then the 8-day period preceding that was searched. If valid pixels were found, then the data for the 8-day period following the missing pixel were searched, and the period after that if necessary. If valid pixels were found both before and after the period of the missing pixel, then a weighted average of the valid pixels was calculated. Otherwise the pixel remained missing. Averages were rounded to integer values. Pixels interpolated in this way were not used to calculate interpolated values for other pixels. The interpolated data were then converted to chlorophyll concentrations. Monthly averages of the SeaWiFS data (Fig. 2) show the broad scale patterns in chlorophyll production in the NEA.

4 Model Formulation

Whilst accurate mapping of phytoplankton is of interest in itself, here we were also concerned with the role

of phytoplankton as the main food source for the boreal copepod, *Calanus finmarchicus*, which in turn is eaten by many harvestable fish resources. Gurney *et al.* (2001) describe a spatially explicit physiologically structured model of *Calanus finmarchicus* in the NEA over an average climatological year. However this model does not include two key stages in the life cycle of *Calanus finmarchicus* which are resource dependent: the onset of egg production (Hind *et al.*, 2000, Richardson *et al.*, 1999) and diapause (Heath and Jónasdóttir, 1999). Incorporating these responses into the model requires accurate predictions of the annual cycle of chlorophyll *a* concentration over the entire NEA. Since the spring bloom affects the onset of egg production and the end of the autumn bloom triggers diapause in *Calanus finmarchicus*, corect estimation of the timing of these blooms was particularly important. *Calanus finmarchicus* feed in the surface waters, and surface chlorophyll concentration has been found to correlate well with the total chlorophyll *a* standing stock (Shiomoto *et al.*, 2002; Engelsen *et al.*, 2002) so we used the average chlorophyll *a* concentration in the top 5 m of the water column as our response variable.

As explained above, the relationship between satellite and bottle data is complex and thus a simple calibration between the two was not enough, especially at in polar regions and on the continental shelf. Furthermore, the satellite data estimate the climatogical average of chlorophyll concentration at the surface over a small region in time and space, whereas the bottle data estimate average chlorophyll concentration in the top 5 m in a precise location on a specific date. It is therefore not surprising that the satellite data did not correlate well with the bottle data available for the NEA (Fig. 3). Comparisons of SeaWiFS data with bottle data for the timeseries at Stonehaven and OWS Mike show discrepancies between the two which depend on location and time of year (Fig. 4). The SeaWiFS data for Stonehaven, which is on the shelf, was too high in the spring and too low in the summer, whereas at OWS Mike, in the ocean, the SeaWiFS data was accurate in early spring but too high in the autumn. We required an alternative method of predicting chlorophyll *a* concentrations, which would capture the broad scale patterns in the satellite data but also correlate well with the bottle data.

If we modelled the bottle data as a function of SeaWiFS then we could use SeaWiFS to predict chlorophyll concentrations in areas where bottle data have not been collected. We would expect bottle data to be smoothly related to SeaWiFS, but this relationship is affected by weather and water characteristics, which in turn vary with ocean depth and time of year. Since we had good coverage of depth and time of year, we used these as proxies for the things that are really controlling the relationship but had not been measured.

The *Calanus* model uses log(chlorophyll) as the index of food abundance, so we modelled the log of the bottle chlorophyll concentrations. We therefore also took logs of SeaWiFS, which reduced leverage caused by a few high SeaWiFS values. Depth was square-root transformed to reduce leverage of points with high depth values.

Both variables were then scaled to have a similar range to time of year (1–365). Thus we modelled observed log(bottle chlorophyll concentration) as a smooth function $g\,(s,\, h,\, t)$ where $s = 10^{k_s}\text{scaled}\{\log(\text{SeaWiFS})\}$, $h = 10^{k_h}\text{scaled}(\sqrt{\text{depth}})$, t =time of year, given scaling parameters k_s and k_h. In order to to minimise the prediction error in terms of log(chlorophyll), we used penalised least squares to fit the model, using the smoothing parameter λ to trade off fidelity to the data with wiggliness of the fitted function. That is, we minimised:

$$S(g) = \sum_{i=1}^{n} \{y_i - g(s_i, h_i, t_i)\}^2 + \lambda J_{md}(g) \tag{1}$$

where y_i is log(chlorophyll) for the ith observation, λ, k_s and k_h are treated as smoothing parameters to be estimated by GCV and wiggliness is measured by the thin plate spline penalty functional, $J_{md}(g)$. λ and g were estimated using the package `mgcv`, version 0.8.0 (Wood, 2003) in the `R` environment (Ihaka and Gentleman, 1996). We used a grid-search to find smoothing parameters k_s and k_h, thus fitting an anisotropic function using the isotropic measure $J_{md}(g)$. Further details are given in the Appendix.

5 Practical details

Satellite values were assigned to the bottle data by averaging valid SeaWiFS concentrations within a $15'$ by $15'$ box surrounding the bottle value, and linearly interpolating between the values for the two 8-day periods surrounding the bottle value. Some bottle values were considered to be on land by the satellite data, and these were removed from the data set.

The survey coverage was very uneven in space and time. For example about 6,700 of the stations were in the Skagerrak, whilst only 540 were above 65°N. A perfect model would not be affected by this uneven coverage, however, we did not have a perfect model, even the covariates we were using were only proxies for those that we really believed were driving the process. Furthermore, the uneven coverage in this data set was rather extreme, and the model for the open ocean, the major part of the NEA, and the area in which we were most interested, was likely to be unduly influenced by the data on the shelf. We therefore selected a random sub-sample of the data with a more even spatial coverage, hereafter referred to as the fitting data.

To achieve a more even spatial coverage, the sample was obtained using inclusion probabilities inversely proportional to the density of points in a neighbourhood surrounding each observation, the constant of proportionality depending on location. The neighbourhood of a particular observation was defined by a three-dimensional bin, centred on the observation, with sides approximately 50 nm by 50 nm by 15 days. Thus the density of observations in the neighbourhood of observation i, ρ_i was given by:

$$\rho_i = \sum_{j \neq i} L_{ij}$$

where

$$L_{ij} = \begin{cases} 1 & p_i - \delta p \le p_j \le p_i + \delta p,\ q_i - \delta q \le q_j \le q_i + \delta q,\ t_i - \delta t \le t_j \le t_i + \delta t \\ 0 & \text{otherwise} \end{cases}$$

where p represents longitude, q latitude, t time of year, and δp, δq and δt represent half the length of the corresponding sides of the box defining the neighbourhood.

To increase the number of observations selected from the open ocean relative to the shelf, we doubled the inclusion probabilities of observations in the ocean. The shelf is often defined as the area with depths less than 400m. We retained that definition, but further defined the southern shelf to include the North Sea and the shelf around Scotland and Ireland. This region was approximated by an ellipse of sides 10° longitude, 7° latitude, centered on 56°N, 10°E (Fig. 5). To reduce the number of observations in the Skagerrak, we divided the inclusion probabilities in this region by 5. Finally, inclusion probabilities were set at a maximum of 0.9. Thus the inclusion probability of observation i, P_i, can be expressed as follows:

$$P_i = \begin{cases} \min\left(\frac{1}{\rho_i}, 0.9\right) & i \in \text{southern shelf} \\ \min\left(\frac{1}{5\rho_i}, 0.9\right) & i \in \text{Skagerrak} \\ \min\left(\frac{2}{\rho_i}, 0.9\right) & i \in \text{otherwise} \end{cases}$$

Having randomly selected the fitting data set, a further validation data set was selected from the remaining observations. Setting a maximum inclusion probability of 0.9 improved the chance of having data representative of all locations in space and time in the validation data set as well as in the fitting data set. This procedure resulted in a fitting data set of 1,540 observations and a validation data set of 668 observations, both of which were relatively well-spaced in space and time (Fig. 5). Bottle values were set at a minimum of 0.003 mg m^{-3} before taking logs. This affected 18 observations out of the 1,540 (17 of which were zero).

Time of year, seabed depth and SeaWiFS were chosen as covariates because they all had good coverage over most of their range. Taking transformations of SeaWiFS and depth further improved the coverage and thus avoided problems caused by high leverage of a few large values. Since the logged bottle data were being modelled, we took the log of SeaWiFS too, and after experimentation, we took the square-root of depth. SeaWiFS values were set at a minimum of $\exp(-2.5)$ before being logged, which removed the gap between the zeros and the next smallest value on the transformed scale. 270 observations had zero SeaWiFS values, and these were mainly in the winter months. Time of year was measured using julian day and therefore had a range 1–365, so after transformation, $\sqrt{\text{depth}}$ and log(SeaWiFS) were scaled so that they had similar ranges to time of year. These scaled covariates were then each multiplied by 10^{k_s} and 10^{k_h} respectively, where k_s and k_d are scaling parameters which control the relative degree of smoothing in each dimension.

The maximum degrees of freedom allowed in the model was set to 300, approximately 20% of the sample size. The grid of scaling parameters was evenly spaced over the range -3 to 1. This allowed a wide range of relative smoothing, although *a priori* we would expect that SeaWiFS and depth to be smoothed more than time of year, *i.e.* have negative scaling parameters. Preliminary results showed that the GCV score was relatively smooth over the grid and so we chose an interval of 0.4 between the nodes on the grid. Multiplying a covariate by 10^{-3} makes its values very small compared to the other covariates. This means it has little effect in the wiggly part of the model but since its effect in the polynomial part could be large, it might still be an important covariate.

The data have poor spatial coverage in the winter months, and so the model predictions could have produced unrealistic values at the ends of the year. Having used the data described above to find the scaling parameters, we then added 22 'structural zeros' evenly spaced over the model arena, with zero bottle and satellite values, at julian day -10 and 375. The model was then refitted to the combined data set using the chosen scaling parameters. This new model was used to make predictions.

6 Results

The best model had $R^2 = 66\%$, measured on the scale of the logged chlorophyll values. Diagnostic plots for this model are shown in Fig. 6. Scaling parameters were -1.4 for SeaWiFS and -1.0 for depth, indicating that most of the 188 effective degrees of freedom in the model were used by time of year. If we had just used logged SeaWiFS values to predict logged bottle values in the fitting data set, we would have obtained an R^2 of 37%. Comparing model predictions with observed values for the validation data set resulted in $R^2 = 65\%$, which is almost as good as the R^2 for the data we fitted the model to, a remarkably good result. Model predictions for the time series at Stonehaven and OWS Mike (Fig. 7) wiggled a lot over time, partly because the model responds to SeaWiFS values, which also vary a lot over time. However, the R^2 we obtained for the validation data set indicates that the model was not overfitted. Comparison of Figs 4 and 7 show that, for these time series, the model predicted the onset of the spring bloom on the shelf and the magnitude of the autumn bloom in the ocean better than the satellite data. Comparisons of the monthly midpoint predictions (Figs 2 and 8) show that the model retained the broad scale patterns in the SeaWiFS plots, but altered the magnitudes where necessary. For example the satellite data predicted chlorophyll beginning to appear along the coasts in the North Sea in February (Fig. 2), but the SeaWiFS predictions were too high at Stonehaven (Fig. 4). The model also predicted chlorophyll appearing around the coasts, but at levels lower than the satellite (Fig. 8), resulting in a better fit to the data at Stonehaven (Fig. 7). Similarly, the satellite data predicted high concentrations along the coasts and in shallower waters in September (Fig. 2), but these predictions were too high at Stonehaven and OWS Mike (Fig. 4). The model predictions retained the same

spatial patterns as the satellite data, but again with lower levels (Fig. 8), resulting in an improved fit at the time series locations (Fig. 7). Furthermore, the model provided sensible predictions from November to January, when valid SeaWiFS data were not available.

To test whether we really needed SeaWiFS data to provide adequate predictions of the phytoplankton bloom, we fitted another model to the data, the same as that above, except that log(SeaWiFS) was replaced by latitude as a covariate. Latitude is a natural choice of alternative covariate since the phytoplankton bloom can be described in terms of latitude, time of year and depth, as in the introduction to this paper. Transformations of latitude were considered but found to be unnecessary, $\sqrt{\text{depth}}$ was again considered the best depth transformation. The model selected had a similar GCV score to the model using SeaWiFS, with an $R^2 = 64\%$ and 208 effective degrees of freedom. The scaling parameters were -0.4 for latitude and -0.8 for $\sqrt{\text{depth}}$. Predictions for the validation data set had $R^2 = 61\%$, and predictions for the two time series at Stonehaven and OWS Mike were similar to those in Fig. 7. Thus the model with latitude instead of SeaWiFS appeared to perform almost as well as the model with SeaWiFS as a covariate. However, monthly predictions over the NEA (Fig. 9) did not reflect the broad scale behaviour of the phytoplankton bloom. They clearly had artefacts related to latitude, due to the paucity of data at high latitudes. In particular, predictions for June and August had high bands of chlorophyll at around 60-65°N, which were not present in the satellite predictions, whereas predictions for September were too low along the Norwegian coast. Although we can fit a good model to the available data, we do not have the spatial coverage to be able to predict well in areas without data unless we use SeaWiFS satellite data in the model.

7 Discussion

Satellite estimates of chlorophyll *a* concentration are available for the whole world and are routinely used to quantify primary production. We have shown that the SeaWiFS data do not correlate well with field measurements of chlorophyll concentration for the North East Atlantic. This is because the conversion algorithms are applied on a global scale and do not take different water characteristics into account. Furthermore, the satellites measure the reflection off the surface of the ocean whereas interest usually lies in the chlorophyll concentration in the water column, which we approximate with that in the top 5 m. Many attempts have been made to improve the algorithms converting satellite radiance measurements into chlorophyll concentration (*e.g.* Land and Haigh, 1996; Hu *et al.*, 2000; Ruddick *et al.*, 2000; Burenkov *et al.*, 2001; Westbrook *et al.*, 2001; Chen *et al.*, 2002), but these rely on assumptions of the state of the sea when the measurements were taken. Although satellite estimates are frequently compared to data collected in situ (Kahru and Mitchell, 1999; Moore *et al.*, 1999; Dierssen and Smith, 2000; Burenkov *et al.*, 2001; Sathyendranath *et al.*, 2001), the two estimates are rarely combined. The only method we are aware of is Gregg *et al.* (2001),

in which satellite and in situ measurements are 'blended', using a method previously applied to sea surface temperature (Reynolds, 1988), in which the in situ (bottle) estimates are unadajusted in the final product. Here, however, we also removed the sampling and inter-annual variability of the bottle estimates to obtain a climatological average.

We modelled bottle data as a smooth function of SeaWiFS, depth and time of year, using anisotropic thin plate regression splines. The model predictions were a vast improvement on the satellite data, R^2 increasing from 37% to 66%. A similar R^2 was obtained for a validation data set, indicating that the model is not overfitting, despite the wiggliness of the predictions over time. The model did not predict the magnitude of the peak of the spring bloom very well. This is because we fitted the model on the log-scale, so the difference between the model and the data was not as important as it appears in Fig. 7. The model has the advantage that we now have sensible predictions for the winter months, when SeaWiFS data are not available. Where data were available for comparison, the model tended to reduce predictions round the coasts, where SeaWiFS predictions were too high, particularly in the Skagerrak, for which the SeaWiFS data predicts very high concentrations from March to October, probably because the use of standard conversion algorithms is inappropriate for this region and time of year. The model also increased predictions in the open water, particularly in the spring bloom, when SeaWiFS data were too low. A comparison with a similar model using latitude instead of SeaWiFS showed that although superficially similar fits to the data could be obtained, the predictions were not sensible in areas where little data had been collected. Thus we conclude that satellite data are necessary to successfully predict chlorophyll *a* concentrations in the absence of bottle data.

The statistical innovation in this paper is the use of thin plate regression splines for anisotropic smoothing. This was achieved by scaling the covariates relative to each other, choosing these scaling parameters objectively by GCV. Although the models use different basis functions and so are no longer nested, it appears that this approximation works well. The use of a simple grid-search for scaling parameter estimation is computationally expensive but simple to program as the individual models can be fitted using the `mgcv` package in `R`. This is therefore a simple objective method of fitting anisotropic smooths.

Acknowledgements

This work was supported by the NERC Special Topic in Marine Productivity (GR02/2749).

We would like to thank the organisations who made the bottle data available to us: Harry Dooley and Elsa Green at the ICES Oceanographic Data Centre in Copenhagen, and Polly Hadziabdic at the British Oceanographic Data Centre for extracting archived data; Kjell Bakkeplass of the Institute of Marine Research Bergen, Ole-Petter Pedersen of the University of Tromso, Astthor Gislason of the Marine Research Institute Reykjavic, and Sigrun Jónasdóttir of the Danish Institute for Marine Research for helping to access data

collected during the EU-TASC project.

We would also like to thank the SeaWiFS Project (Code 970.2) and the Distributed Active Archive Center (Code 902) at the Goddard Space Flight Center, Greenbelt, MD 20771, for the production and distribution of these data, respectively, activities that are sponsored by NASA's Mission to Planet Earth Program. Thanks also to Steve Groom, Tim Smyth, Gareth Mottram and Luke Tudor in the Remote Sensing Group at Plymouth Marine Laboratory, who compiled the SeaWiFS composite data.

References

Burenkov, V. I., Vedernikov, V. I., Ershova, S. V., Kopelevich, O. V. and Sheberstov, S. V. (2001) Application of the ocean colour data gathered by the SeaWiFS satellite scanner for estimating the bio-optical characteristics of waters in the Barents Sea. *Oceanology*, **41**, 461–468.

Chen, C., Jonsson, L. and Larson, M. (2002) Parameters to characterize biological conditions in marine and coastal waters retrieved from SeaWiFS data. *MTS Journal*, **36**, 14–22.

Dierssen, H. M. and Smith, R. C. (2000) Bio-optical properties and remote sensing ocean colour algorithm for Antarctic penisula waters. *J. Geophys. Res. – Oceans*, **105**, 26301–26312.

Engelsen, O., Hegseth, E. N., Hop, H., Hansen, E. and Falk-Petersen, S. (2002) Spatial variability of chlorophyll *a* in the Marginal Ice Zone of the Barents Sea, with relations to sea ice and oceanographic conditions. *J. Mar. Syst.*, **15**, 79–97.

Gregg, W. W. and Congkright, M. E. (2001) Global seasonal climatologies of ocean chlorophyll: Blending in situ and satellite data for the Coastal Zone Color Scanner era. *J. Geophys. Res.*, **106**, 2499–2515.

Gregg, W. W. and Congkright, M. E. (2002) Decadal changes in global ocean chlorophyll. *Geophys. Res. Lett.*, **29**, 20–26.

Green, P. J. and Silverman, B. W. (1994) *Nonparametric Regression and Generalized Linear Models.* London: Chapman and Hall.

Gu, C. and Wahba, G. (1991) Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method. *SIAM J. Sci. Stat. Comp.*, **12**(2), 383–398.

Gurney, W. S. C., Speirs, D. C., Wood, S. N., Clarke, E. D. and Heath, M. R. (2001) Simulating spatially and physiologically structured populations. *Journal of Animal Ecology*, **70**, 881–894.

Heath, M. R., Adams, R. D., Brown, F., Dunn, J., Frase, S., Hay, S. J., Kelly, M. C., Macdonald, E. M.,

Robertson, S. and Wilson, C. (1999) Plankton monitoring off the east coast of Scotland in 1997 and 1998. *Fisheries Research Services Report 13/99.*

Heath, M. R., Astthorsson, O. S., Dunn, J., Ellertsen, B., Gislason, A., Gaard, E., Gurney, W. S. C., Hind, A. T., Irigoien, X., Melle, W., Niehoff, B., Olsen, K., Skreslet, S. and Tande, K. S. (2000) Comparative analysis of *Calanus finmarchicus* demography at locations around the northeast Atlantic. *ICES Journal of Marine Science*, **57**(6), 1562–1580.

Heath, M. R. and Jónasdóttir, S. (1999) Distribution and abundance of overwintering *Calanus finmarchicus* in the Faroe-Shetland Channel. *Fisheries Oceanography*, **8**, 40–60.

Hind, A. T., Gurney, W. S. C., Heath, M. R. and Bryant, A. (2000) Overwintering strategies in *Calanus finmarchicus*. *Mar. Ecol. Prog. Ser.*, **193**, 95–107.

Hooker, S. B. and McClain, C. R. (2000) The calibration and validation of SeaWiFS data. *Progress in Oceanography*, **45**, 427–465.

Hu, C. M., Carder, K. L. and Muller-Karger, P. E. (2000) Atmospheric correction of SeaWiFS imagery over turbid coastal waters: A practical method. *Remote Sens. Environ.*, **74**, 195–206.

Ihaka, R. and Gentleman, R. (1996) R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, **5**(3), 299–314.

Kahru, M. and Mitchell, B. G. (1999) Empirical chlorophyll algorithm and preliminary SeaWiFS validation for the Californian Current. *Int. J. Remote Sens.*, **20**, 3423–3429.

Land, P. E. and Haigh, J. D. (1996) Atmospheric correction over case 2 waters with an iterative fitting algorithm. *Applied Optics*, **35**, 5443–5451.

Longhurst, A. R. (1998) *Ecological geography of the Sea.* San Diego: Academic Press.

Mann, K. H. and Lazier, J. R. N. (1991) *Dynamics of Marine Ecosystems: Biological-physical interactions in the Oceans.* Boston: Blackwell Science.

Moore, J. K., Abbott, M. R., Richman, J. G., Smith, W. O., Cowles, T. J., Coale, K. H., Gardner, W. D. and Barber, R. T. (1999) SeaWiFS satellite ocean color data from the Southern Ocean. *Geophys. Res. Lett.*, **26**, 1465–1468.

O'Reilly, J. E., Maritorena, S., Mitchell, B. G., Siegel, D. A., Carder, K. L., Garver, S. A., Kahru, M. and McClain, C. (1998) Ocean colour chlorophyll algorithms for SeaWiFS. *J. Geophys. Res. – Oceans*, **103**,

24937–24953.

Parsons, T. R., Takahashi, M. and Hargrave, B. (1984) *Biological oceanographic processes*. Oxford: Pergamon Press.

Reid, P. C., Edwards, M. E., Hunt, H. and Warner, A. R. (1998) Phytoplankton changes in the North Atlantic. *Nature*, **391**, 546.

Reynolds, R. W. (1988) A real-time global sea surface temperature analysis. *J. Clim.*, **1**, 75–86.

Richardson, K., Jónasdóttir, S. H., Hay, S. J., Christoffersen, A. (1999) *Calanus finmarchicus* egg production and food availability in the Faroe-Shetland Channel and northern North Sea: October-March. *Fish. Oceanogr.*, **8**(Suppl. 1), 153–162.

Ruddick, K. G., Ovidio, F. and Rijkeboer, M. (2000) Atmospheric correction of SeaWiFS imagery for turbid coastal and inland waters. *Appl. Optics*, **39**, 897–912.

Sathyendranath, S., Gota, G., Stuart, V., Maass, H., and Platt, T. (2001) Remote sensing of phytoplankton pigments: a comparison of empirical and theoretical approaches. *Int. J. Remote Sens.*, **22**, 249–273.

Sathyendranath, S., Hoge, F. E., Platt, T. and Swift, R. N. (1994). Detection of phytoplankton pigments from ocean colour — improved algorithms. *Applied Optics*, **33**, 1081–1089.

Shiomoto, A., Saitoh, S., Imai, K., Toratani, M., Ishida, Y. and Sasaoka, K. (2002) Interannual variation in phytoplankton biomass in the Bering Sea basin in the 1990s. *Prog. Oceanogr.*, **55**, 147–163.

Strickland, J. D. H. and Parsons, T. R. (1972) A practical handbook of seawater analysis. *Fish. Res. Bd. Canada Bull*, **167**, 1–311.

Westbrook, A. G., Pinkerton, M. H., Aiken, J., Pilgrim, D. A. (2001) Simulated performance of remote sensing ocean colour algorithms during the 1996 PRIME cruise. *Deep-Sea Res. Part II – Top. Stud. Oceanogr.*, **48**, 845–858.

Wood, S. N. (2000) Modelling and smoothing parameter estimation with multiple quadratic penalties. *JRSSB*, **62**(2), 413–428.

Wood, S. N. (2003) Thin plate regression splines. *JRSSB*, **65**(1), 95–114.

Yentsch, C. S. and Phinney, D. A. (1985) Spectral fluorescence: an ataxonomic tool for studying the structure of phytoplankton populations. *J. Plankton Res.*, **7**, 617–632.

Appendix: Estimation

The general form of the model is:

$$y_i = f(\mathbf{x}_i) + \varepsilon_i$$

where y_i is the response for the ith observation, f is a smooth function of d covariates, whose values for the ith observation are contained in the vector $\mathbf{x}_i$, and the ε_is are independent random errors with zero mean and equal variance. Here $\mathbf{x}_i = (s_i, h_i, t_i)^T$.

In principle we could use thin plate splines to estimate f by finding the function g which minimises

$$S(g) = \sum_{i=1}^{n} \{y_i - g(\mathbf{x}_i)\}^2 + \lambda J_{md}(g) \tag{2}$$

where

$$J_{md}(g) = \int \dots \int_{\Re^d} \sum_{v_1+\dots+v_d=m} \frac{m!}{v_1! \dots v_d!} \left(\frac{\partial^m g}{\partial x_1^{v_1} \dots \partial x_d^{v_d}} \right)^2 dx_1 \dots dx_d$$

and $2m > d$. $J_{md}(g)$ is isotropic and invariant under translations and rotations of the coordinate system. A good introduction to thin plate splines can be found in Green and Silverman (1994), and we briefly review them here. A function $g(\mathbf{x})$ is a thin plate spline on the data set $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ if it is of the form:

$$g(\mathbf{x}) = \sum_{i=1}^{n} \delta_i \eta_{md}(\|\mathbf{x} - \mathbf{x}_i\|) + \sum_{j=1}^{M} \alpha_j \phi_j(\mathbf{x}).$$

where

$$\eta_{md}(r) = \begin{cases} \frac{(-1)^{m+1+d/2}}{2^{2m-1}\pi^{d/2}(m-1)!(m-d/2)!} r^{2m-d} \log(r) & d \text{ even} \\ \frac{\Gamma(d/2-m)}{2^{2m}\pi^{d/2}(m-1)!} r^{2m-d} & d \text{ odd} \end{cases},$$

$M = \binom{m+d-1}{d}$ and the ϕ_j are M linearly independent polynomials spanning the M-dimensional space of polynomials in $\Re^d$ of total degree less than m. Thus g is linear in its parameters and is made up of a wiggly part $\sum_{i=1}^{n} \delta_i \eta_{md}(\|\mathbf{x} - \mathbf{x}_i\|)$ and a polynomial part $\sum_{j=1}^{M} \alpha_j \phi_j(\mathbf{x})$. A natural thin plate spline has the constraint that $\mathbf{T}^T \boldsymbol{\delta} = 0$, where $\boldsymbol{\delta} = (\delta_1, ..., \delta_n)^T$ and the $n \times M$ matrix $\mathbf{T}$ is defined by $T_{ij} = \phi_j(\mathbf{x}_i)$. This constraint ensures that $J_{md}(g)$ is finite. Since the ϕ_js have maximum degree $m-1$, $J_{md}(g)$ only contains terms resulting from differentiating the wiggly part of g, and for a natural thin plate spline,

$$J_{md}(g) = \boldsymbol{\delta}^T \mathbf{E} \boldsymbol{\delta} \tag{3}$$

where the $n \times n$ matrix $\mathbf{E}$ is given by $E_{ij} = \eta_{md}(\|\mathbf{x}_i - \mathbf{x}_j\|)$.

For a visibly smooth function without discontinuities at the knots, we require $2m > d + 1$. Thus for a three dimensional thin plate spline, $d = 3$, $\mathbf{x} = (x_1, x_2, x_3)^T$ and a natural choice for m is $m = 3$. Then

$$\eta_{md}(\|\mathbf{x} - \mathbf{x}_i\|) = \frac{1}{96\pi} \|\mathbf{x} - \mathbf{x}_i\|^3$$

and

$$\sum_{j=1}^{M} \alpha_j \phi_j(\mathbf{x}) = \alpha_1 + \alpha_2 x_1 + \alpha_3 x_2 + \alpha_4 x_3 + \alpha_5 x_1^2 + \alpha_6 x_2^2 + \alpha_7 x_3^2 + \alpha_8 x_1 x_2 + \alpha_9 x_2 x_3 + \alpha_{10} x_3 x_1.$$

To fit the model, we need to evaluate g at each observation. The vector

$$\mathbf{g} = (g(\mathbf{x}_1),\ g(\mathbf{x}_2),\ \ldots,\ g(\mathbf{x}_n))^T$$

is given by:

$$\mathbf{g} = \mathbf{E}\boldsymbol{\delta} + \mathbf{T}\boldsymbol{\alpha} \tag{4}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_M)^T$. Now substituting in (2) using (3) and (4), we can write:

$$S(g) = \|\mathbf{y} - \mathbf{E}\boldsymbol{\delta} - \mathbf{T}\boldsymbol{\alpha}\|^2 + \lambda \boldsymbol{\delta}^T \mathbf{E} \boldsymbol{\delta}, \tag{5}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^T$. Thus, to estimate f, we find the vectors of parameters $\boldsymbol{\delta}$ and $\boldsymbol{\alpha}$ that minimise (5), subject to $\mathbf{T}^T\boldsymbol{\delta} = 0$. The constraint gives us n parameters in total, $n - M$ in the wiggly part and M in the polynomial part.

Fitting this model is computationally intensive since we have a parameter for every data point, and the number of calculations is $O(n^3)$. However, by penalising the sum of squares with a wiggliness penalty we are *a priori* expecting some of these n parameters to be redundant. It would be convenient if we could find a wiggly basis, which uses fewer parameters, k say, and gives similar results to the standard basis with n parameters. One approach is to find a basis of rank k which simultaneously minimises the maximum change in fitted values and the maximum change in the penalty term (Wood, 2003). Set $\mathbf{E} = \mathbf{U}\mathbf{D}\mathbf{U}^T$, where $\mathbf{D}$ is a diagonal matrix of eigenvalues of $\mathbf{E}$ arranged in decreasing order, and $\mathbf{U}$ is a matrix containing the eigenvectors of $\mathbf{E}$ in the corresponding order. Then the best basis of rank k subject to the criteria above is given by $\mathbf{U}_k$ which contains the first k columns of $\mathbf{U}$ (Wood, 2003). $\mathbf{U}_k$ forms a k-dimensional orthonormal basis for the $\boldsymbol{\delta}$ parameter space, so that $\boldsymbol{\delta} = \mathbf{U}_k\boldsymbol{\delta}_k$, where $\boldsymbol{\delta}_k$ is a k-vector. A convenient way to include the constraints is to set $\boldsymbol{\delta} = \mathbf{U}_k\mathbf{Z}_k\tilde{\boldsymbol{\delta}}$, where $\mathbf{Z}_k$ is an orthogonal column basis such that $\mathbf{T}^T\mathbf{U}_k\mathbf{Z}_k = 0$. This ensures that $\mathbf{T}^T\boldsymbol{\delta} = 0$, and $S(g)$ becomes:

$$S(g) = \|\mathbf{y} - \mathbf{U}_k\mathbf{D}_k\mathbf{Z}_k\tilde{\boldsymbol{\delta}} - \mathbf{T}\boldsymbol{\alpha}\|^2 + \lambda\tilde{\boldsymbol{\delta}}^T\mathbf{Z}_k^T\mathbf{D}_k\mathbf{Z}_k\tilde{\boldsymbol{\delta}},$$

where $\mathbf{D}_k$ is the diagonal matrix containing the biggest k eigenvalues of $\mathbf{E}$ arranged in decreasing order. We now minimise $S(g)$ with respect to the $k + M$ parameters $\tilde{\boldsymbol{\delta}}$ and $\boldsymbol{\alpha}$ to find what we term a thin plate regression spline (Wood, 2003).

The smoothing parameter λ, which governs the wiggliness of the model, is selected by minimising the generalized cross-validation (GCV) score, using a method developed by Wood (2000) based on the method of Gu and Wahba (1991). GCV is a measure of prediction error and can be calculated with the following formula:

$$V(\lambda) = \frac{\|\mathbf{y} - \mathbf{A}\mathbf{y}\|^2}{[tr(\mathbf{I} - \mathbf{A})]^2}$$

where $\mathbf{A}$ is the hat matrix (see Green and Silverman (1994)). $tr(\mathbf{A})$ is the effective degrees of freedom in the model, and so we are simply dividing the residual sum of squares by the square of the residual degrees of freedom.

Thin plate regression splines are isotropic smoothers and the relative amount of smoothing of the covariates is related to their numerical ranges, regardless of the units they are measured in. The covariates therefore need to be scaled so that their relative ranges reflect the relative amount of smoothing they require in each direction. Thus if we have d covariates, we need to estimate $d - 1$ scaling parameters (k_s and k_h in this case). The estimation of the overall smoothing parameter, λ, is performed using GCV. We therefore also used GCV to estimate the scaling parameters. We used a simple grid search as follows. We chose a set of scaling parameters and scaled the covariates accordingly. The model was then fitted using these scaled covariates and the GCV score recorded. This was performed for each set of scaling parameters on the grid, and the set of scaling parameters which resulted in the smallest GCV score is chosen. This is a simple but computationally intensive method of finding the scaling parameters.

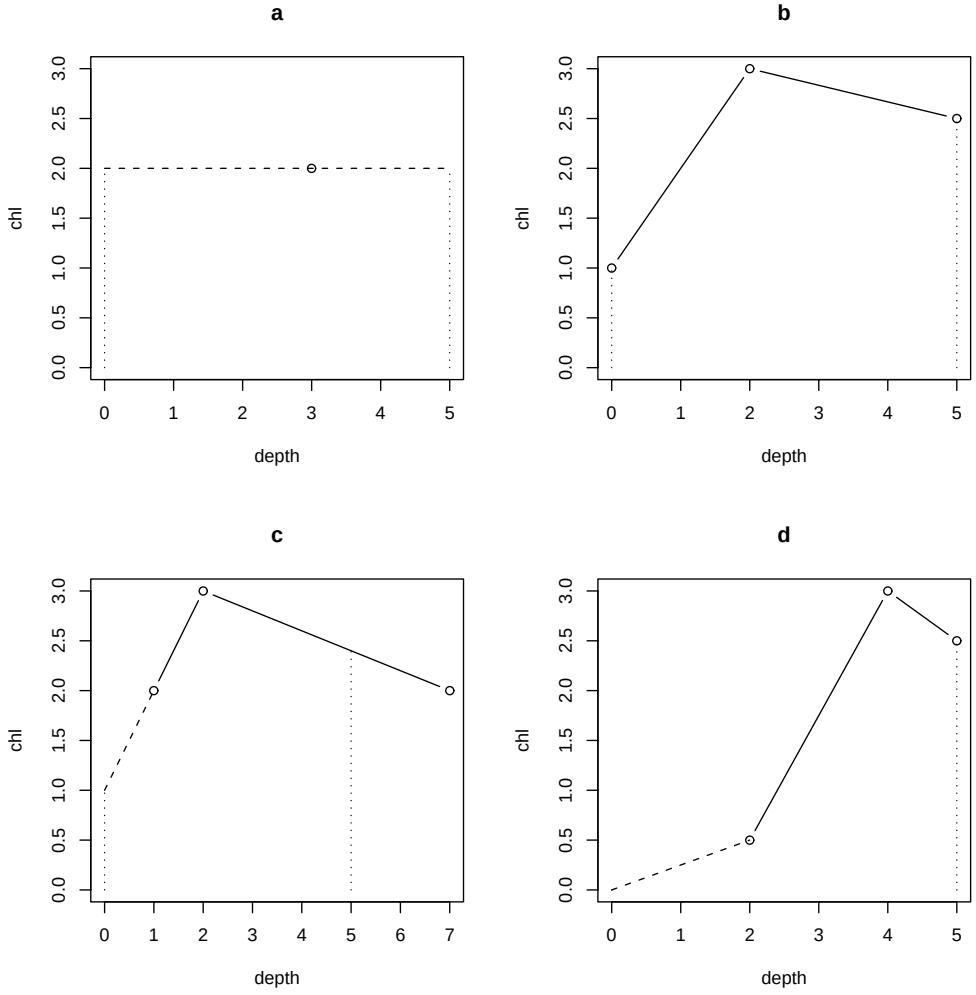

Figure 1: Diagrams showing the depth integration scheme used to obtain average chlorophyll concentration in the top 5 m of the water column from the bottle sample data. (a) Where there is a single measurement within at a particular location which is within the range [0,5] m, we use this as the average over the range [0,5] m. (b) Where there are measurements at zero and 5 m we interpolate between them and any measurements in between, then integrate beneath the resulting polygon. (c) Where there is no measurement at zero, we extrapolate back to obtain an estimate for zero. Where there is no measurement at 5 m but at least one measurement within the range [0,5) m and a measurement beyond 5 m, we interpolate to obtain an estimate at 5 m. (If there is no measurement beyond 5 m, we extrapolate.) (d) If any extrapolation results in a negative value, we set the estimate to zero.

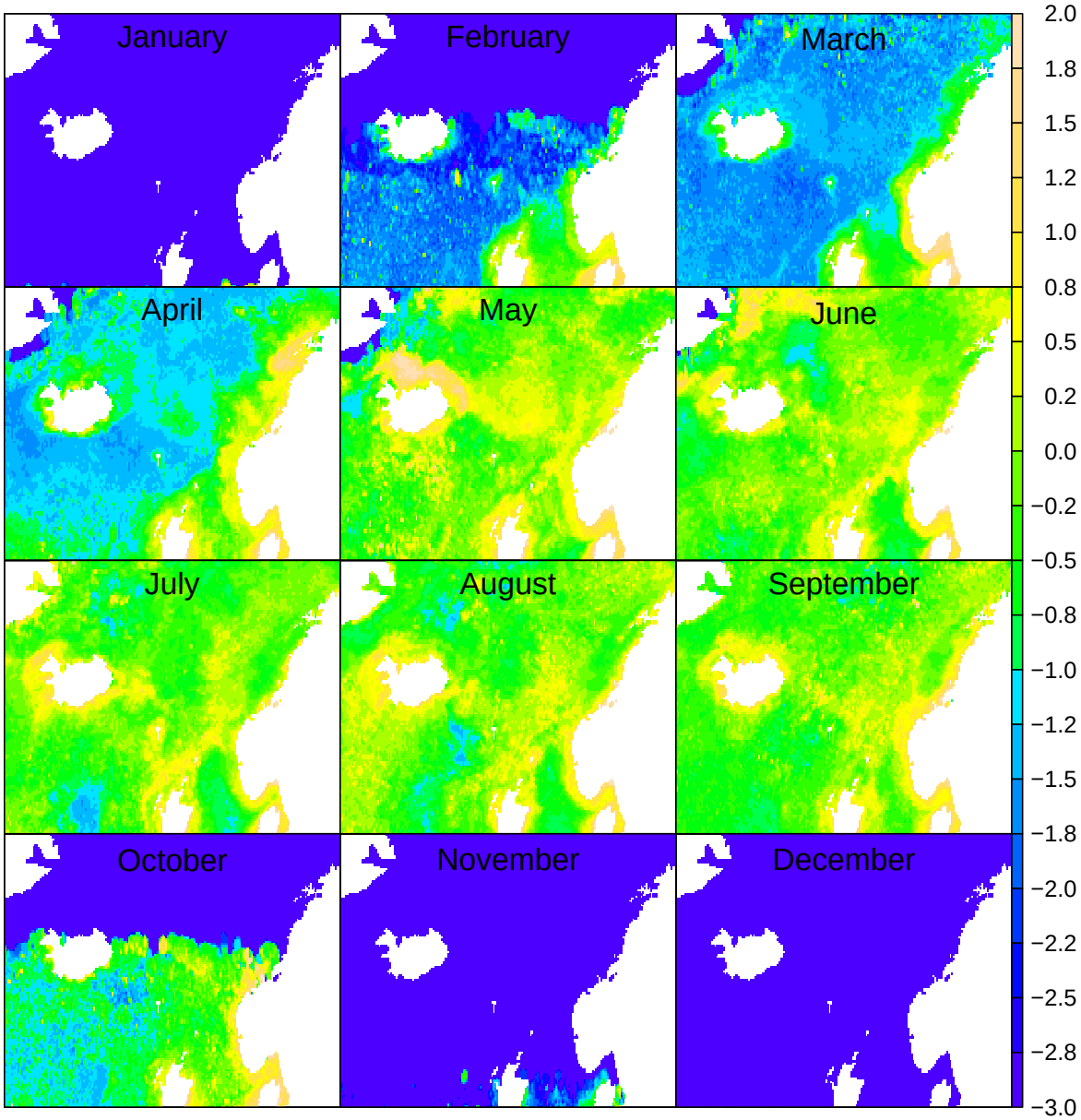

Figure 2: Estimates of log(chlorophyll *a* concentration), measured in log (mg m^{-3}) over the North East Atlantic (56°N–75°N, 30°W–20°E) for selected monthly midpoints throughout the year. These estimates have been obtained from SeaWiFS data and are plotted on a log-scale to show small changes at low levels more clearly.

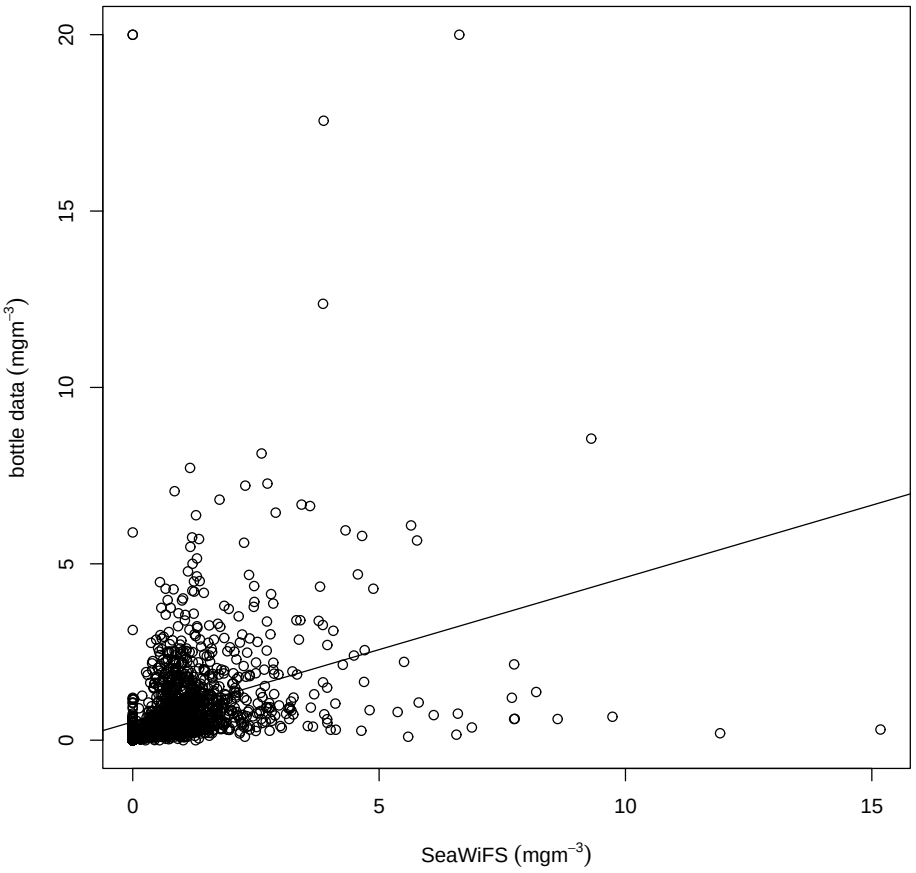

Figure 3: Chlorophyll *a* concentrations for a subset of the bottle data (the fitting data: see Fig. 5) plotted against the corresponding satellite predictions, with a regression line superimposed. R^2 is only 10%, indicating a poor correlation between SeaWiFS and bottle data.

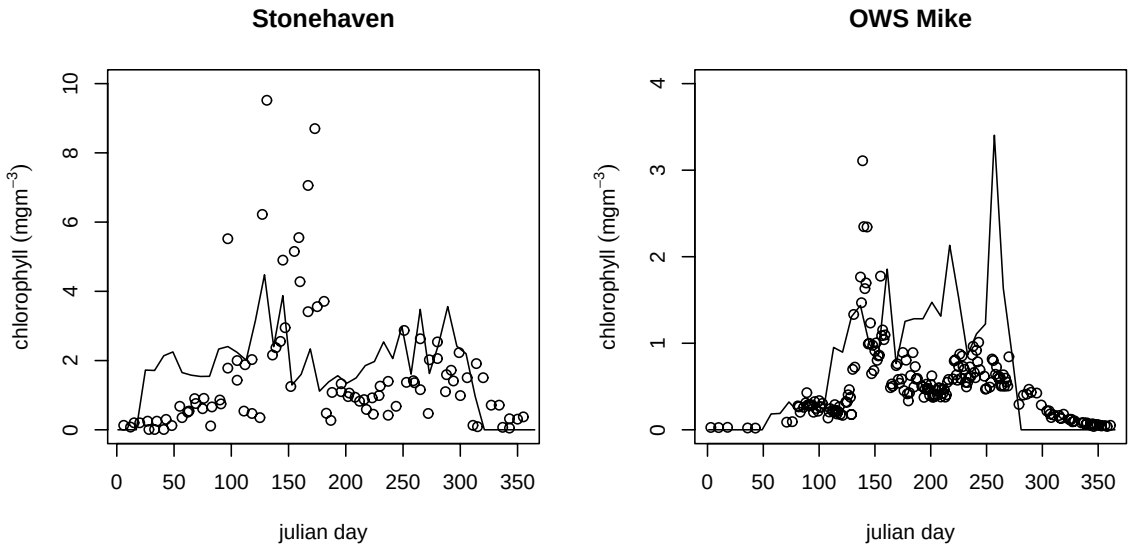

Figure 4: Time series of bottle data (circles) with SeaWiFS predictions superimposed as a line for the two locations marked in Fig 5(a).

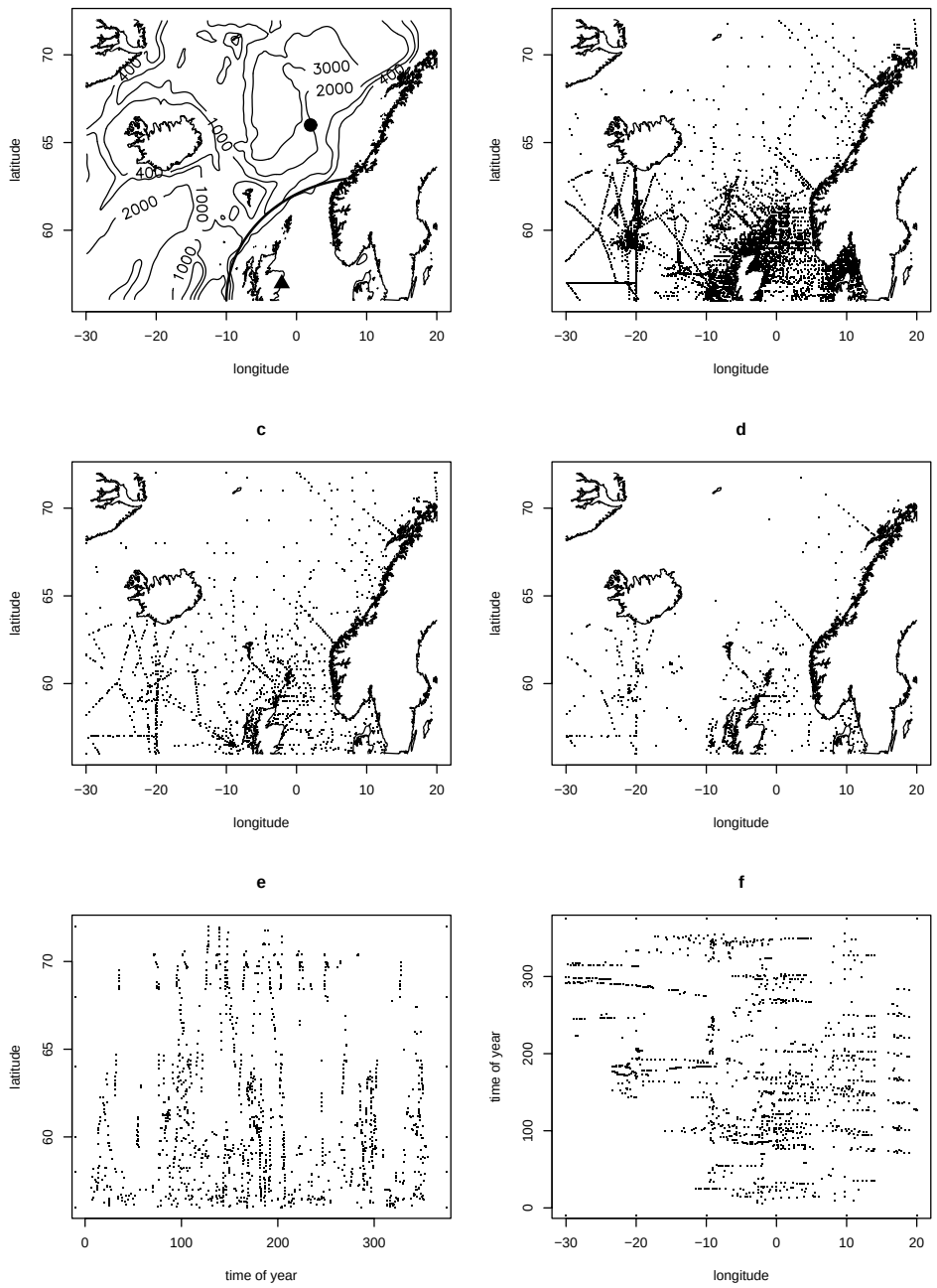

Figure 5: (a) Contour plot of the bathymetry (measured in metres) of the northeast Atlantic, with the boundary between southern shelf and ocean indicated by the thick line. The positions of the TASC monitoring stations at Stonehaven (circle) and Ocean Weather Ship Mike (triangle) are also indicated. (b) The locations of all the available bottle data. (c) The locations of the fitting data. (d) The locations of the validation data. (e) Latitude against time of year (julian days) for the fitting data. (f) Time of year (julian days) against longitude for the fitting data.

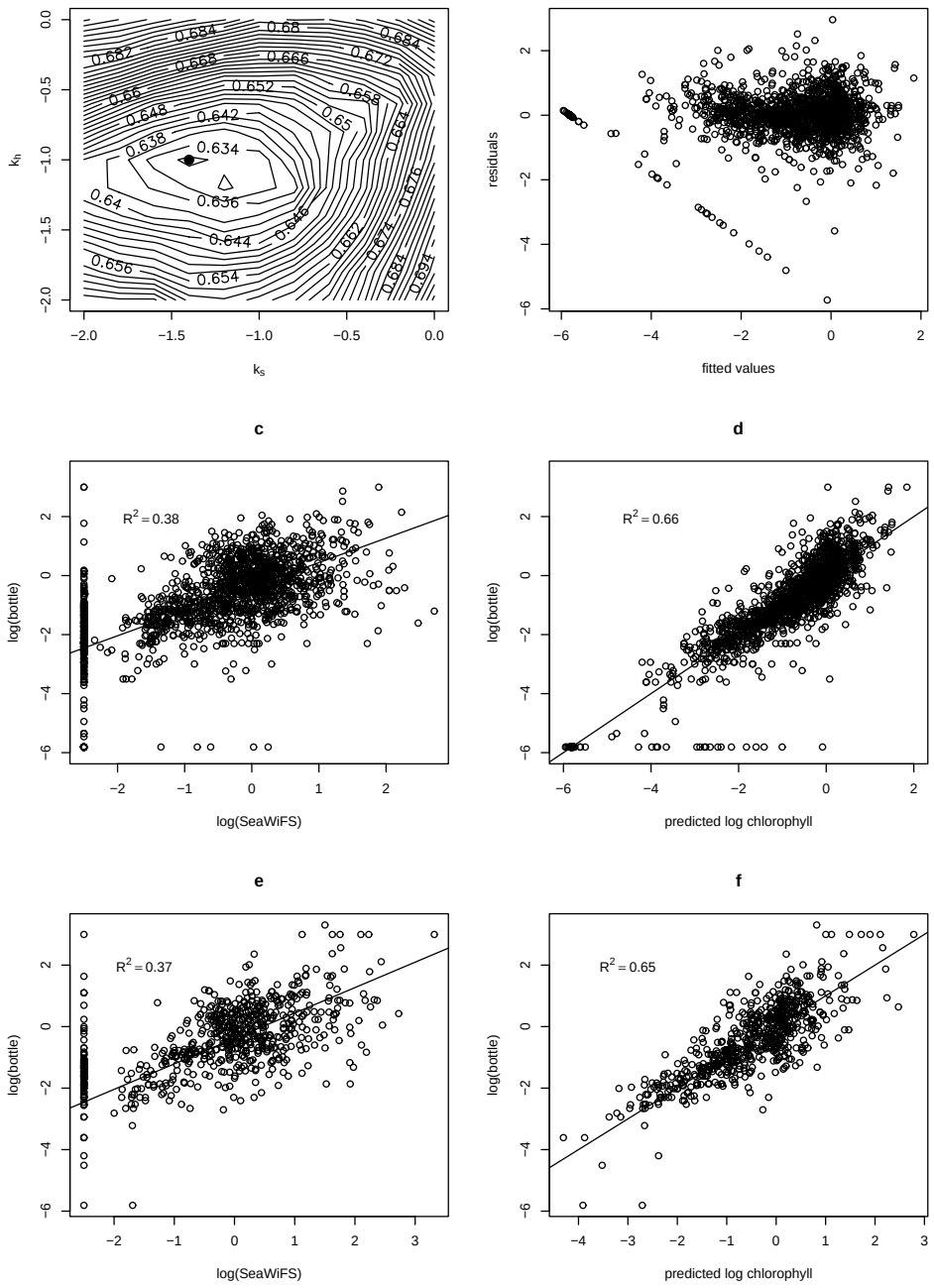

Figure 6: The fitted model: (a) Contour plot of the GCV score as it changes with scaling parameters for SeaWiFS and depth (k_s and k_h, respectively). The minimum GCV score is indicated by a solid circle. (b) Residual plot of the chosen model. (c) Observed vs SeaWiFS, both measured in log(mg m^{-3}), for the fitting data. (d) Observed vs fitted values, both measured in log(mg m^{-3}), for the fitting data. (e) Observed vs SeaWiFS, both measured in log(mg m^{-3}), for the validation data. (f) Observed vs fitted values, both measured in log(mg m^{-3}), for the validation data.

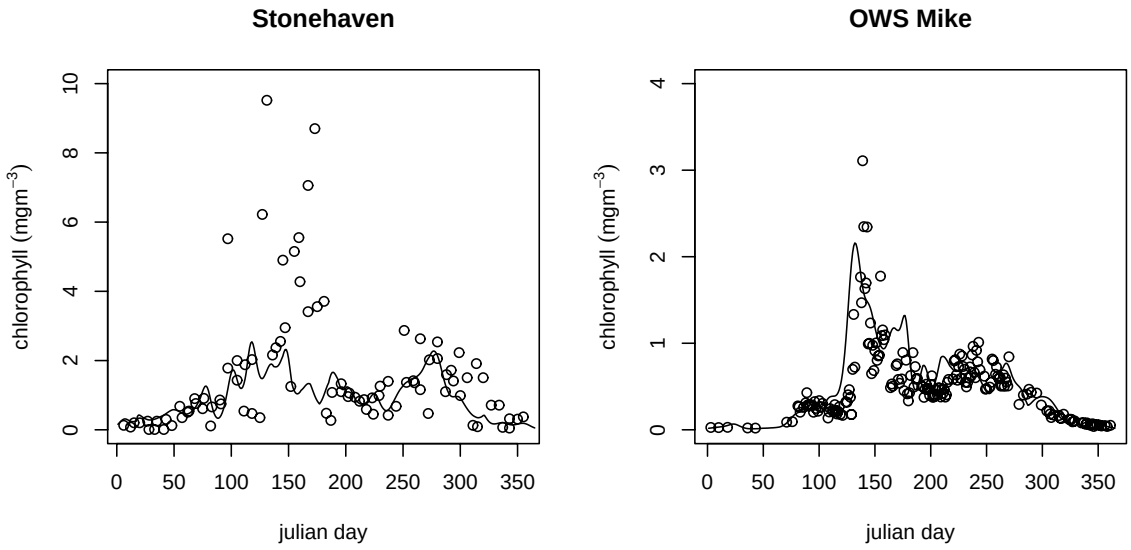

Figure 7: Time series of bottle data (circles), with model predictions overlaid as lines, for the locations marked in Fig 5(a).

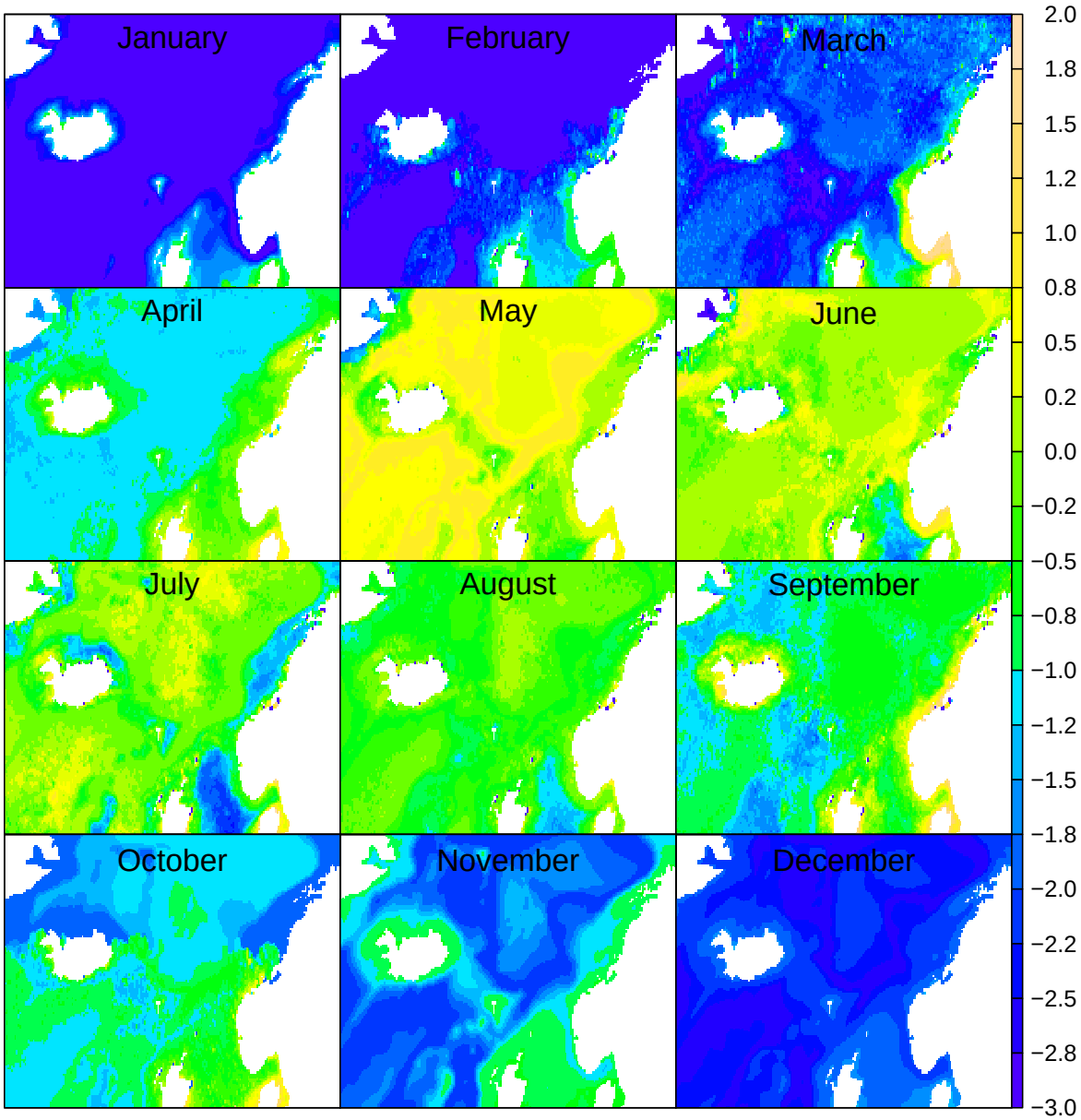

Figure 8: Monthly midpoint predictions of log(chlorophyll concentration), measured in log(mg m^{-3}), for the best model with SeaWiFS, depth and time of year as covariates.

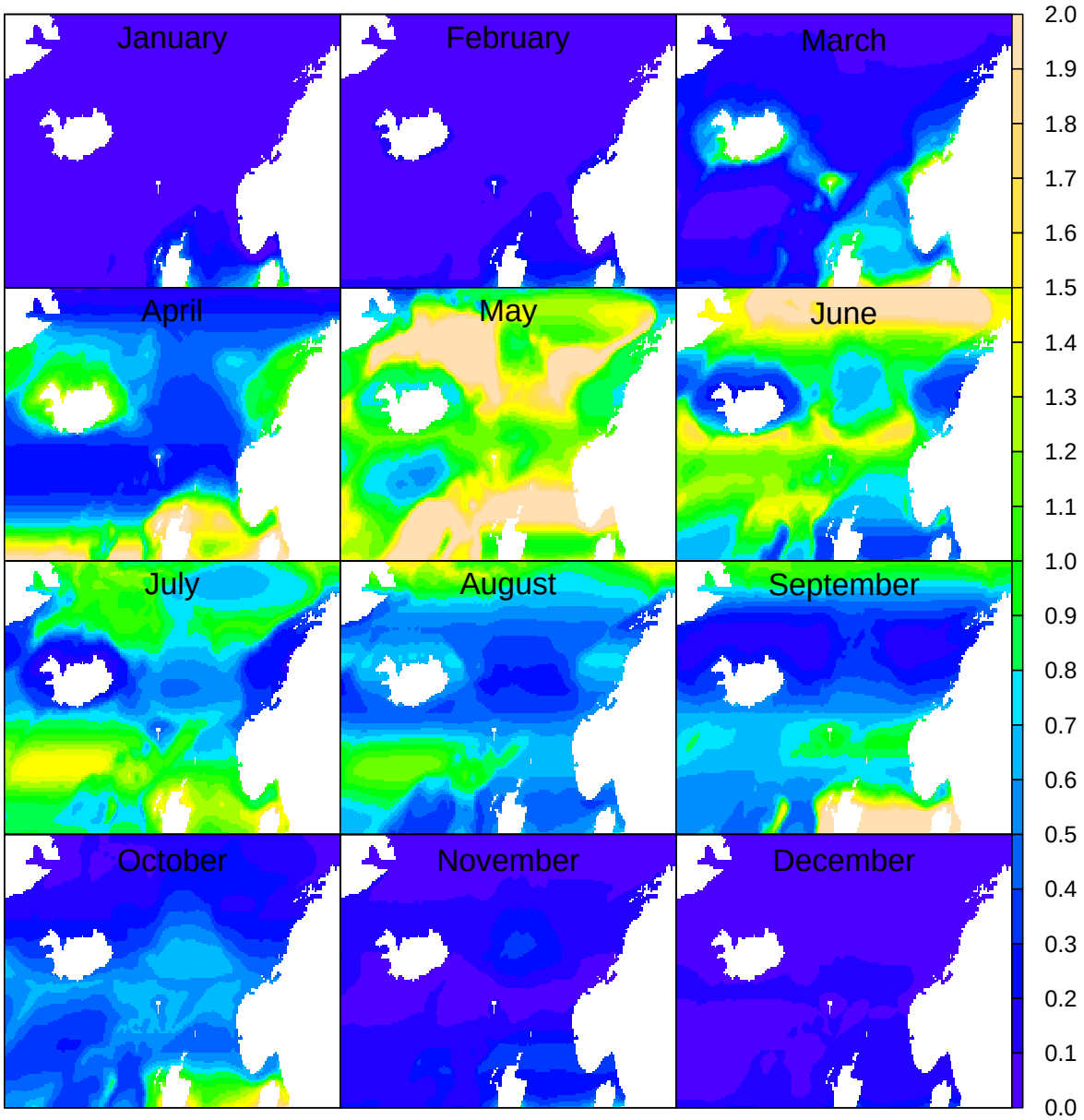

Figure 9: Monthly midpoint predictions of log(chlorophyll concentration), measured in log(mg m^{-3}), for the best model with latitude, depth and time of year as covariates.