

Multivariate analysis of phytoplankton spectral *in vivo* fluorescence: estimation of phytoplankton biomass during a mesocosm study in the Baltic Sea

Jukka Seppälä^{1,3,*}, Kalle Olli²

¹Finnish Institute of Marine Research, Erik Palménin aukio 1, PO Box 2, 00561 Helsinki, Finland

²Institute of Botany and Ecology, University of Tartu, Lai st. 40, 51005 Tartu, Estonia

³Present address: Finnish Environment Institute, PO Box 140, 00251 Helsinki, Finland

ABSTRACT: Multivariate calibration methods were compared for the retrieval of phytoplankton biomass in different taxonomic groups from the spectral fluorescence signal of a living phytoplankton community. During a mesocosm experiment in the northern Gulf of Finland (Baltic Sea), the natural phytoplankton community was manipulated by additions of nutrients and fish. The changes in phytoplankton biomass and species structure were followed using traditional microscopic methods and examination of the spectral fluorescence of living cells. Based on the pigmentation of photosystem II, phytoplankton were divided into 4 groups: (1) cyanobacteria with phycobilins, (2) cryptophytes with phycobilins, chlorophyll *a* (chl *a*), chlorophyll *c* (chl *c*) and carotenoids, (3) chromophytes with chl *a*, chl *c* and carotenoids, and (4) chlorophytes with chl *a* and chlorophyll *b* (chl *b*) and a small amount of carotenoids. The phytoplankton biomass in these groups was predicted from the spectral fluorescence signal using classical least squares, principal component regression, and partial least squares (PLS) regression. The prediction ability of the models was compared using the root mean square error of prediction during full cross validation, partial cross validation and external validation. Regarding relevancy for the operational monitoring of phytoplankton community dynamics using spectral data, the PLS model gave the closest predictions for all taxonomic groups and with the accuracy needed for phytoplankton bloom detection.

KEY WORDS: Multivariate calibration · Partial least squares · Spectral fluorescence signal · Phytoplankton pigments · Baltic Sea

—Resale or republication not permitted without written consent of the publisher—

INTRODUCTION

The differences in pigmentation, and consequently in spectral properties, between different phytoplankton groups are well established and can be used as a starting point in the chemotaxonomical discrimination of phytoplankton (Cullen et al. 1997, Schofield et al. 1999, Millie et al. 2002). Signature pigments are best detected by chromatographic methods (Jeffrey et al. 1997), but less laborious and more rapid methods are required for phytoplankton bloom detection and monitoring systems. Various optical approaches are currently available for the automated detection of phyto-

plankton biomass and chemotaxonomy (Cullen et al. 1997, Schofield et al. 1999). Absorption and reflectance signals are, however, influenced by other optically active compounds like chromophoric dissolved organic matter (CDOM). For this reason, especially in coastal and inland waters, phytoplankton-related information is not easily retrieved from the spectra. The fluorescence of living phytoplankton, on the other hand, takes place at wavelengths that are not disturbed much by the other compounds (Millie et al. 2002). Non-photosynthetic carotenoids with low taxonomic specificity do modify spectral absorption and reflectance, but do not influence the spectral fluorescence signal (Johnsen &

*Email: jukka.seppala@fimr.fi

Sakshaug 1996). Furthermore, the detection limit for fluorescence is lower than that for absorption or reflectance. As the motivation for the current study, we feel that the power of multivariate calibration in chemotaxonomic analyses using the spectral fluorescence of phytoplankton populations has not been fully utilised, or, alternatively, not adequately reported.

The fluorescence emission of living phytoplankton cells arises from chlorophyll *a* (chl *a*) associated with photosystem II (PSII). PSII is excited through pigments in light-harvesting complexes and consequently the fluorescence excitation spectra are determined by the absorption properties of these pigments (Johnsen & Sakshaug 2007). Additionally, fluorescence emission arises from phycobiliproteins that are found in cyanobacteria, cryptomonads and a few other phytoplankton groups.

A single wavelength *in vivo* fluorometric method (Lorenzen 1966) is widely used for the estimation of chl *a* concentrations. Currently this method has a key role in the detection of bulk phytoplankton biomass, e.g. by ship-of-opportunity systems, even though a large variability in the chl *a* specific fluorescence (*R*) has been observed. Briefly, *R* is affected by phytoplankton community structure, cell size, photoacclimation, nutrient stress and diurnal cycles (e.g. Kiefer 1973, Alpine & Cloern 1985, Cleveland & Perry 1987, Guo & Dunstan 1995, Dandonneau & Neveux 1997). At the cellular level, the main sources for the variability in *R* are the composition of the pigment complexes, and photochemical and non-photochemical fluorescence quenching (Falkowski & Kiefer 1985). The energy of absorbed quanta that cannot be used for photochemistry by the reaction centres will be lost as fluorescence, via triplet formation, or emitted as heat.

Yentsch & Yentsch (1979) introduced spectral fluorescence as a tool for the detection of the chemotaxonomic characteristics of a phytoplankton community. This technique is based on the distinct spectral features of the major phytoplankton groups due to their main PSII accessory pigments: cyanobacteria with various phycobilins, cryptophytes with phycobilins, chlorophyll *c* (chl *c*) and xanthophylls, chromophytes with chl *c* and xanthophylls, and chlorophytes with chlorophyll *b* (chl *b*) (see Poryvkina et al. 1994). The shape or index of spectral fluorescence has been used in revealing information about the phytoplankton community structure at a spatial resolution not easily accessible by other methods (e.g. Yentsch & Phinney 1985, Cowles et al. 1993, Babichenko et al. 1999). The fluorescence signal of phycobilins, in particular, has been used to monitor the distribution of cyanobacteria (Wood et al. 1998, Seppälä et al. 2007).

In the analysis of spectral data to discriminate between various phytoplankton pigment groups or to

detect single species, several techniques have been used, including Fourier-transform-based pattern recognition (Oldham et al. 1985), discriminant analysis (Johnsen et al. 1994), a similarity index algorithm (Millie et al. 1997, 2002) and deconvolution using spectral libraries (Gerhardt & Bodemer 2000, Beutler et al. 2002). Multivariate calibration methods are found to be applicable for many spectroscopic applications when signals from different constituents are overlapping, the background noise is high and variable, and not all the optically active compounds are known (Martens & Næs 1989). These methods, therefore, would seem relevant for extracting the information carried by the spectral fluorescence of natural phytoplankton samples (Kaitala et al. 1994). Recently these methods have also been used for studies of pigment extracts (Moberg et al. 2000) and phytoplankton absorption (Moberg et al. 2002, Stæhr & Cullen 2003).

In this paper, we measured spectral fluorescence during a mesocosm study carried out at the entrance to the Gulf of Finland, in the northern Baltic Sea. Different levels of phytoplankton biomass and changes in the species structure were obtained by manipulating the natural plankton community by additions of nutrients and fish. The phytoplankton responses to different manipulations were quite contrasting, providing a wide range of qualitatively and quantitatively varying natural phytoplankton assemblages. The main purpose of the study was to examine the feasibility of the spectral fluorescence method, when combined with chemometric multivariate analyses, for detection of the changes in phytoplankton community structure and bloom dynamics in the Baltic Sea.

MATERIALS AND METHODS

Experimental setup and sampling. During a mesocosm experiment in the northern Gulf of Finland (Baltic Sea) the natural brackish-water plankton community was manipulated by additions of inorganic nutrients (NP; i.e. PO₄-P, NO₃-N and NH₄-N) and fish at the fry stage (F). Floating transparent polyethylene enclosures, with a depth of 14 m, a diameter of 2.3 m, and a total volume of about 50 m³, were filled with surface water from the study site, Tvärminne Storfjärden (59° 51' 20" N, 23° 15' 45" E; for a description of the study area see Niemi 1975). The experiment lasted for 3 wk, from 8 to 28 July in 1993. Four differently manipulated enclosures (Control, NP, F and NPF [containing both added inorganic nutrients and fish]) were examined for this study. Nutrients were added on experimental Days 0, 7 and 14 to the NP and NPF enclosures (the additions were 0.48 µM phosphate, 2.14 µM nitrate and 2.14 µM ammonium), which aimed to

induce a series of phytoplankton blooms. Fry-stage fish (stickleback *Gasterosteus aculeatus*; 400 individuals per enclosure) were added at the beginning of the experiment to the F and NPF enclosures. The additions of fish aimed to suppress mesozooplankton and, thus, reduce the grazing loss on phytoplankton. One enclosure served as a control (Ctrl) and was not manipulated. The experimental design and the effects of top-down and bottom-up manipulations on phytoplankton biomass and community structure are discussed by Olli et al. (1996).

Samples for fluorescence measurements were taken every second day at 14:00 h local time from experimental Day 7 (15 July) to Day 19 (27 July). Samples were taken with a Ruttner-type sampler at 6 depths (0, 2, 4, 6, 8 and 10 m) and for logistical reasons were left for 1 to 4 h in darkness to avoid sun-induced photoinhibition and heating until measurements were carried out in the laboratory. The dark period also allowed relaxation of non-photochemical fluorescence quenching, which is partly completed during 15 to 30 min dark acclimation. An underwater photosynthetically available radiation (PAR) sensor (PNF-300, Biospherical Instruments) was used to determine the attenuation of downward scalar irradiance at the time of sampling. Vertical profiles of temperature were obtained with a SIS CTD 100Plus.

Fluorescence, chl *a* and biomass analyses. *In vivo* excitation spectra were recorded with a spectrofluorometer (Shimadzu RFPC-5001). The excitation wavelength was scanned from 400 to 670 nm with a step of 1 nm, and fluorescence emission intensity was measured at the maximum wavelength of chl *a* (682 nm). The excitation and emission slits were 5.0 nm. It took about 30 s to perform one scan. Fluctuations in the fluorescence intensity, assessed by daily triplicate spectra from 2 samples, were minor. The resulting coefficient of variation in the spectral fluorescence intensity was on average below 10% and of a random nature.

Fluorescence spectra were corrected for signals caused by Raman scattering and dissolved organic matter by subtracting the fluorescence of a fraction of enclosure water filtered through Whatman GF/F filters (nominal pore size, 0.7 µm). The Raman and background filtrate signals were constant for all days and enclosures (data not shown). Further, fluorescence spectra were corrected for instrument optics using Basic Blue 3 (Kopf & Heinze 1984).

Chl *a* concentrations were measured from the same samples as used for fluorescence. Subsamples of 100 ml were filtered onto Whatman GF/F filters and extracted in 96% ethanol for 24 h at room temperature and in darkness (Arvola 1981). Concentrations were measured with a spectrofluorometer, which was calibrated using a chl *a* standard (Sigma).

Micro- and nanophytoplankton cells were counted from the samples taken from 2 integrated layers of 0 to 5 m (surface) and 6 to 12 m (deep layer) on even-numbered experimental days (Days 6, 8, 10, etc.) as described by Olli et al. (1996). Picocyanobacteria and pico-sized (<2 µm) eukaryotes and chl *a* in the <2 µm size fraction (pico-chl *a*) were measured from the same samples, but only for the surface layer. Total chl *a* was measured daily from the surface layer and, on even experimental days, also from the deeper layer. We assume that the chl *a* to phytoplankton biomass ratio, the share of various species in the total biomass and the pico-chl *a* to total chl *a* ratio, for the days on which these variables were not measured, could be estimated as the average values determined on the preceding and subsequent days. The phytoplankton biomass, the share of different species and the pico-chl *a* at the times of fluorescence measurements could then be estimated by adjusting the ratios with measurements of chl *a* concentrations at the surface (average of 0, 2 and 4 m) and in deeper (average of 6, 8 and 10 m) layers. For the comparison of fluorescence and biomass data, spectral fluorescence measurements from 0, 2 and 4 m were averaged to represent the surface, while those from 6, 8 and 10 m were averaged to represent the deeper layer.

Statistical analysis of the relation between spectral fluorescence and phytoplankton chemotaxonomy. Several taxonomic phytoplankton groups, but not species, can be differentiated by their photosynthetic pigments or by spectral properties (Poryvkina et al. 1994, Johnsen & Sakshaug 2007). In our study, the Chlorophyta group includes chl *b*-containing species classified in Chlorophyceae, Prasinophyceae and Euglenophyta (with *Eutreptiella gymnastica* as the dominant species in this experiment). The Chromophyta group consisted of phytoplankton belonging to groups containing chl *c* (Dinophyceae and Chrysophyceae) and pico-eukaryotes with an unknown taxonomic position. Chl *c* and phycobilin-containing Cryptophyta formed a third group. Picocyanobacteria and filamentous cyanobacteria were grouped together as cyanobacteria, though in the Baltic Sea they differ in their pigmentation (Seppälä et al. 2007). Additional tests to predict separately the biomass of Dinophyceae, filamentous cyanobacteria or picocyanobacteria were also performed, as well as the retrieval of pico-chl *a*.

Analytical spectrofluorometric methods for multi-component samples are based on Beer's law given that (1) the fluorescence intensity at a given excitation-emission wavelength combination is the sum of the fluorescence intensities of each component present, and (2) the fluorescence intensity of a single component is linearly related to its concentration. These

can be considered true for dilute solutions of many compounds, like phytoplankton pigments in extracts (Neveux & Lantoiné 1993). For living phytoplankton samples, the second assumption is not valid, as the fluorescence properties of phytoplankton cells are highly variable, depending on their physiological conditions. Thus, strictly, phytoplankton biomass is not linearly related to the fluorescence intensity. However, we tested different chemometric techniques to retrieve phytoplankton biomass in various chemotaxonomic groups using spectral fluorescence data. All the calculations were carried out by MATLAB v. 6.5. software using custom built programs.

Calibration models were made using the data from the surface layer samples, and the prediction ability of the models was assessed by 3 independent methods. First, in full cross validation, one observation at a time was removed from the calibration data set and the phytoplankton biomass of the different chemotaxonomic groups in that sample was estimated by a model made with the remaining data. Secondly, in partial cross validation, data from one mesocosm at a time was removed during calibration; the aim was to reveal the robustness of resulting models to slight changes in phytoplankton taxonomy and biomass levels not included in the calibration samples. Thirdly, in external validation, the models were made by using all the surface layer samples, and were then used to reconstruct the biomass in the deep layer samples. The latter method may not have been as effective as the others as we had no estimates of picoplankton abundance in the deep layer.

Different multivariate models were compared using the root-mean-square error of prediction (RMSEP) obtained in full cross validation. RMSEP is computed as the square root of the average squared difference between measured (c_i) and predicted (\hat{c}_i) biomass values for I samples (Martens & Næs 1989).

$$\text{RMSEP} = \sqrt{I^{-1} \sum_{i=1}^I (c_i - \hat{c}_i)^2} \quad (1)$$

For all models a systematic error, the bias, was estimated as the average difference between predicted and observed biomass values;

$$\text{Bias} = I^{-1} \sum_{i=1}^I (\hat{c}_i - c_i) \quad (2)$$

Least-squares approximation. From the 2 assumptions of Beer's law presented in the previous section, it follows that

$$\text{SFS}(\lambda) = \sum_{i=1}^n c_i k_i(\lambda) \quad (3)$$

where $\text{SFS}(\lambda)$ is the observed spectral fluorescence signal of the multicomponent sample, c_i is the concentra-

tion of the i th compound, $k_i(\lambda)$ is a spectral biomass specific fluorescence coefficient for the i th compound, and λ is the wavelength. To separate the number (n) of spectrally different compounds by least-squares approximation (LS) requires at least n equations, i.e. measurements from n fixed excitation–emission wavelength combinations. To enhance the separation of spectrally overlapping components, the number of equations, and, thus, wavelengths, must be increased; these will be noted as m . We used all measured wavelengths, thus m equals 271 wavelengths and the same number of equations.

Using matrix algebra, Eq. (3) can be written as:

$$\text{SFS} = CK + E \quad (4)$$

where SFS is a (l by m) matrix of the sample spectra, C is a (l by n) matrix of the concentrations of constituents, and K is a (n by m) matrix of fluorescence coefficients. The number of samples is l and the error matrix is represented by E . The concentration of the different components in a sample (C_S) can be estimated from the SFS of that sample (SFS_S) when K is known:

$$C_S = (\text{SFS}_S K') (K K')^{-1} \quad (5)$$

where K' is a transpose of K . Two major problems arise when using such an approach for natural phytoplankton. First, the LS method requires knowledge of all the components present in the sample. Secondly, direct measurement of K for any component in a living phytoplankton sample is hard to perform (even when excluding variable fluorescence at this point). An estimate for K may be obtained from a spectral library consisting of the species apparently present in the samples or, as in this study, by the least-squares method from samples with a known phytoplankton biomass and species composition:

$$K = (C' C)^{-1} (C' \text{SFS}) \quad (6)$$

As the classical least-squares approximation technique may result in negative concentrations for some components, and, thus, be unacceptable, the non-negative least-squares technique was used for the calculations.

Factor-based regression: principal component regression and partial least squares. It is seldom possible to fully discover the composition of phytoplankton samples, as some of the phytoplankton species may be unknown, ignored or misclassified. Further, constituents in the calibration samples may be linearly related, i.e. collinear. In such cases LS yields erroneous predictions. Multivariate factor-based regression methods (principal component regression [PCR] and partial least squares [PLS]), instead, do not require information about all components, but only that of the preferred ones. Collinearity and noise in spectral data

is avoided by factor analysis and subsequent selection of relevant factors (Martens & Næs 1989).

In the principal component analysis of spectral multi-component data, the intercorrelated spectral data are reduced to a few, linearly independent variables called eigenvectors or factors, F (f by m matrix, f presents the number of factors). The score matrix, S (l by f), of the principal component solution reflects the concentrations of the different constituents, and the original spectral data can be written as:

$$\text{SFS} = SF \quad (7)$$

We used a 2-way weighted summation algorithm (presented for example in Legendre & Legendre 1998) for computing F and S . In PCR, the calculated scores are subsequently related to the observed concentrations using a matrix of regression coefficients, B (f by n),

$$C = SB \quad (8)$$

Having a set of calibration samples with known concentrations of components, B can be solved as:

$$B = (S'S)^{-1} (S'C) \quad (9)$$

Finally, the concentrations of the constituents in the unknown sample can now be estimated as:

$$C_S = \text{SFS}_S F'B \quad (10)$$

The principle of PLS is similar to PCR in that spectral data are reduced to fewer factors, but PLS uses the correlation between spectral data and concentrations to build a calibration model, and a separate regression step (Eqs. 8 & 9 in PCR) is not needed. This operation reduces the effects of possibly large, but irrelevant, spectral variations in resulting model. In PLS, separate scores (S_{SFS} , S_C) and factors (F_{SFS} , F_C) are created for both spectral data and component concentrations:

$$\text{SFS} = S_{\text{SFS}} F_{\text{SFS}} \quad (11a)$$

$$C = S_C F_C \quad (11b)$$

The basic form of PLS (PLS-1) applies for a single predicted variable, but several modifications are available (Martens & Næs 1989, Wold et al. 2001). Calibration for several components simultaneously can be carried out using PLS-2. The algorithm for PLS-2 is more complex and is iterative for each factor, but it is faster to compute than separate PLS-1 analyses for each of the components. PLS-2 overrides PLS-1 if the components are intercorrelated, but nonlinearities in the concentration–fluorescence relationship may cause overfitting when using PLS-2. In a preliminary test, using the surface samples, we noted that PLS-1 and PLS-2 gave identical results for our data (not shown), for which reason we used the orthogonalized PLS-2 algorithm (as given in Martens & Næs 1989) in preference. Additionally, PLS-1 was used to test the prediction of

separate phytoplankton groups (filamentous cyanobacteria, picocyanobacteria, dinoflagellates and picoplankton).

The number of eigenvectors used in PCR and PLS analyses must be optimized. Too few eigenvectors will result in underfitting problems, as some of the patterns in the data are not then included in the model; too many eigenvectors, on the other hand, will result in overfitting problems, as the prediction error increases due to the increased complexity of calibration. To optimise the number of eigenvectors and minimise prediction error, we used the method of full cross validation and calculated RMSEP (Eq. 1) for each phytoplankton group. Models with a range of from 1 to 15 eigenvectors were inspected; the predictive ability of the model with the lowest RMSEP was considered to be the highest. Additionally, in the selection of models, a visual inspection of loadings was made to be sure that they did not show an unexpected spectral shape or random noise.

Phytoplankton cultures. We compared the spectral shapes of group-specific fluorescence coefficients in the LS method (K estimated by Eq. 6) with the spectra obtained from pure cultures. For this, we cultured one representative from each pigment group: *Synechococcus* sp. (Strain CCY9202) for cyanobacteria, *Rhodomonas* sp. (TV22) for Cryptophyta, the diatom *Thalassiosira pseudonana* (TV5) for Chromophyta, and *Chlamydomonas* sp. (TV44) for Chlorophyta. Cultures were grown in approximately $30 \mu\text{mol q m}^{-2} \text{s}^{-1}$ in a 16 h light:8 h dark cycle. Samples were taken in the exponential growth phase, and fluorescence excitation spectra were measured as was described for natural samples.

RESULTS

Environmental conditions and responses of the phytoplankton community to manipulations

The enclosures were filled after an upwelling event as indicated by low temperatures; 11°C at a depth of 1 m and 5°C at 12 m (see Olli et al. 1996). The temperature of the surface layer increased smoothly and reached 14.5°C by the time of our first sampling on Day 7 and 16.7°C by the end of the experiment. By the end of the experiment a slight thermocline had formed at a depth of 8 to 10 m. The attenuation of downward scalar irradiance (PAR) had a similar pattern in all the enclosures. The depth of the euphotic layer (1% of the surface irradiance) ranged from 7 to 10 m.

At the beginning of the experiment (Day 0), the phytoplankton biomass was at a moderate level for the

study area, $4.2 \mu\text{g chl } a \text{ l}^{-1}$, and the share of pico-chl *a* out of the total chl *a* was 14%. The phytoplankton community was dominated by a phytoflagellate, *Eutreptiella gymnastica* (Euglenophyceae), which constituted 82 to 92% of the total phytoplankton biomass on Day 0. For a complete discussion of the dynamics of *E. gymnastica* during the experiment, see Olli et al. (1996).

In the control unit (Ctrl), the share of *Eutreptiella gymnastica* declined smoothly and had practically disappeared by the end of the experiment (Fig. 1). It was displaced by picocyanobacteria, filamentous cyanobacteria *Aphanizomenon* sp., dinoflagellates *Dinophysis acuminata* and cryptophytes *Plagioselmis prolunga*. At the end of the experiment, micro- and nanophytoplankton were dominated by dinoflagellates in the surface layer and by cryptophytes in the deep layer (data not shown), contributing 45% and 50%, respectively, of the total phytoplankton biomass (excluding picoplankton). The phytoplankton size structure changed considerably, and picocyanobacteria became dominant after Day 11. In the enclosure manipulated with fish (F), the share of *E. gymnastica* declined below 10% during the experi-

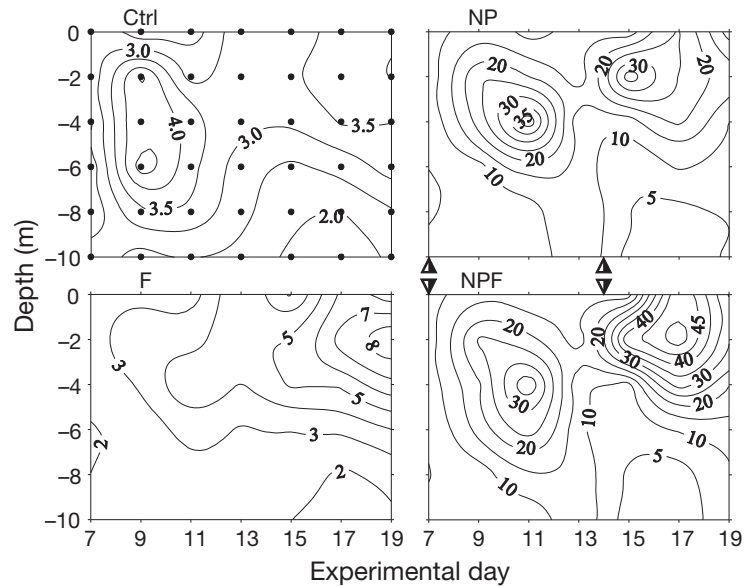


Fig. 2. Development of extracted chl *a* ($\mu\text{g l}^{-1}$) in the different enclosures (Ctrl, NP, F, NPF). The sampling grid is indicated for the control unit (Ctrl). Arrowheads indicate the times of pulsed nutrient additions

ment. It was replaced partly by the same species as in the Ctrl unit, but mostly by *Uroglena americana* (Chrysophyceae) with a share of up to 25% of the total biomass.

The phytoplankton biomass and chl *a* in the NP and NPF units peaked 2 to 3 d after each weekly nutrient addition. Three consecutive blooms—2 of them considered in this study—consisted mainly of *Eutreptiella gymnastica* (Figs. 1 & 2). The proportion of other phytoplankton species remained very low in these units, and it was only during the last days of the experiment that a decline in *E. gymnastica* took place in the NP unit, with small cryptophytes (*Plagioselmis prolunga* and *Teleaulax acuta*) concurrently increasing their share. At this phase the picocyanobacteria, *Aphanizomenon flos-aquae*, and *Dinophysis acuminata* also had a biomass comparable to *E. gymnastica*. The change in the community structure in the NP unit was also seen as an increase in the pico-chl *a* share of the total chl *a*. In the NPF unit, *E. gymnastica* remained dominant throughout the experiment, and accordingly the pico-chl *a* share of the total chl *a* was low (<27%).

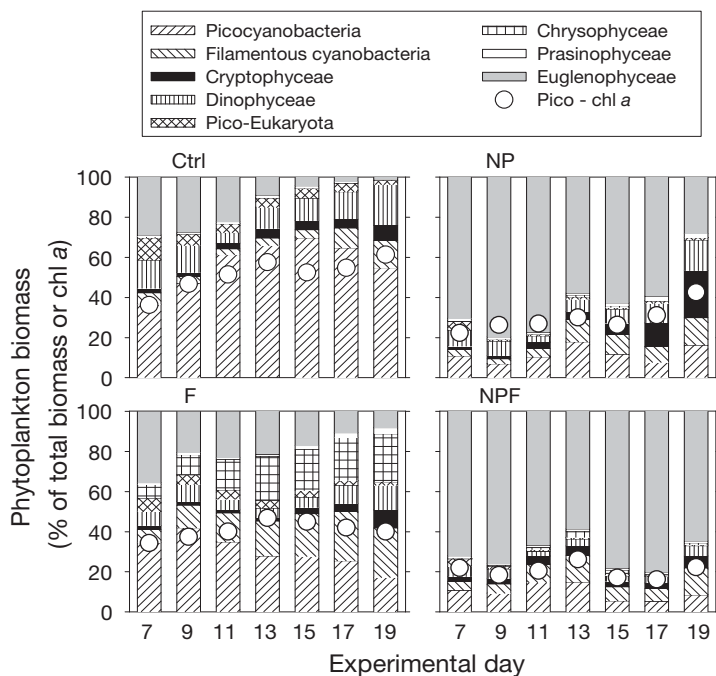


Fig. 1. Phytoplankton community structure (bars, % of total fresh weight obtained by microscopic identification and counting of the cells) and pico-chl *a* (circles, % of total chl *a* obtained by fluorometric detection of extracted samples) in surface layer (0 to 6 m) of enclosures. See 'Materials and methods' for abbreviations

Spectral *in vivo* fluorescence

The relationships between the fluorescence intensities at the chl *a* peak (at 442 nm) and

chl *a* concentrations or total phytoplankton biomass were linear ($r = 0.97$, $n = 168$ and $r = 0.96$, $n = 28$, respectively). This linearity, however, stems from a wide range of concentrations, and therefore does not imply a low variability in *R*. In single enclosures, the correlation coefficients were lower, but only in the Ctrl unit were fluorescence and total phytoplankton biomass not related (not shown), partly due to a narrower range of concentrations in this unit.

Quantum-corrected fluorescence spectra had some distinct peaks and shoulders (Fig. 3). In the blue region, the chl *a* peak was observed around 442 nm; the shoulders at 464 nm and 487–491 nm are due to chl *c* and chl *b*, respectively (Rabinowitch & Govindjee 1969, Anderson & Barrett 1986). All of these, however, are overlapped by spectra of the different photosynthetic carotenoids. Phycoerythrin, possibly phycoerythrocyanin (Bryant 1982, Seppälä et al. 2005) and photosynthetic carotenoids are responsible for the peak at 570–586 nm. The shoulder at 623–625 nm is due to phycocyanin (Sidler 1994) and different chlorophylls (Anderson & Barrett 1986).

Some of the main patterns in the phytoplankton community structure were clearly reflected in the shape of the spectral fluorescence (Fig. 3). First, the ratio of the phycoerythrin peak at 570 nm (F_{570}) to the chl *a* peak at 442 nm (F_{442}) is higher for the Ctrl unit than for the other units (Tukey's test, $p < 0.01$). Consequently, the share of cyanobacteria and cryptophytes—the main phycoerythrin-containing species—of the total biomass was highest in the Ctrl unit. For all spectra, the ratio $F_{570}:F_{442}$ was linearly related to the picocyanobacteria share of the total biomass ($r = 0.77$, $n = 28$, $p < 0.01$). Secondly, the NP and NPF units with a higher share of the chl *b*-containing species *Eutryptiella gymnastica* had a relatively higher shoulder at 480 nm ($F_{480}:F_{442}$) than did the Ctrl and F units (Tukey's test, $p < 0.01$). Clearly, $F_{480}:F_{442}$ was related to the *E. gymnastica* share of the total biomass ($r = 0.68$, $n = 28$, $p < 0.01$).

K matrix in LS models

The estimation of the biomass specific spectral fluorescence coefficients for different phytoplankton groups (*K* in Eq. 6) is the starting point in the prediction of phytoplankton biomass from spectral fluorescence

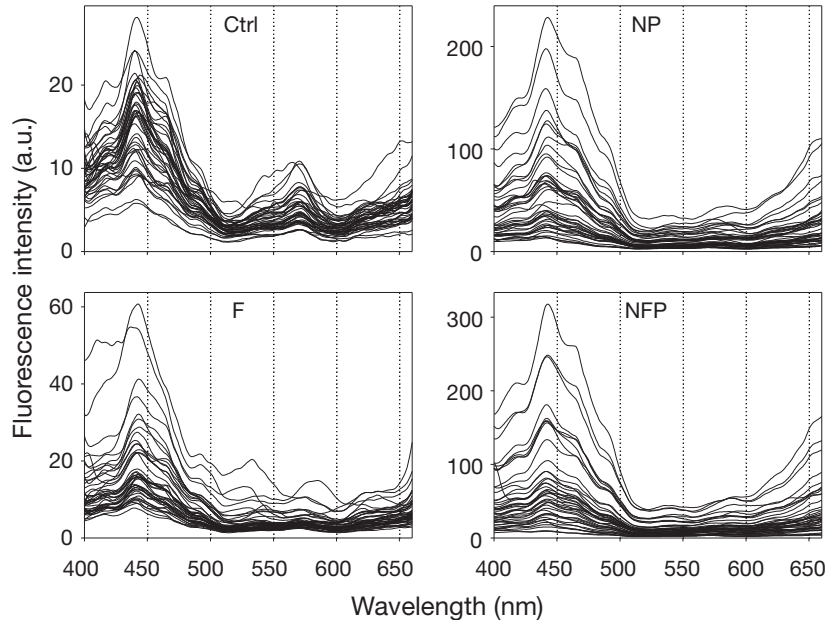


Fig. 3. Phytoplankton *in vivo* fluorescence spectra for the different enclosures (Ctrl, NP, F, NPF) during the mesocosm experiment ($n = 42$ for each unit). Note the different scales on the y-axis. a.u.: arbitrary units

when using LS models. These coefficients, when calculated using all upper layer samples, are shown in Fig. 4. For each taxonomic group they are compared with the spectra obtained for cultured species. To make the comparison easier, spectra for the cultures were scaled to the mean values of the corresponding *K* spectra.

The spectral comparison (Fig. 4) shows that for our data the shapes of *K* are realistic, except for cyanobacteria, and contain information about the light-harvesting pigments specific to each group (for wavelengths see Bryant 1982, Anderson & Barrett 1986, Sidler 1994, Lohrenz et al. 2003). In the *K* for cyanobacteria, the blue chl *a* peak is at 440 nm, and the phycoerythrin peak at 570 nm. Peaks at about 615 nm and 650 nm are due to other phycobilins, but the peak at 462 nm is not realistic, and indicates an incomplete solution of the LS model. Moreover, the spectral shape, especially the phycoerythrin-to-chl *a* peak ratio for the cultured *Synechococcus* sp., differs distinctly from that estimated by Eq. (6) for cyanobacteria. For cyanobacteria the *K* values were lower than for the other groups because in cyanobacteria most of the chl *a* is located in the non-fluorescing photosystem I.

For Cryptophyta, the shape of *K* and the spectra of cultured *Rhodomonas* sp. were similar, with a phycoerythrin peak around 570 nm (with an overlap with chl *c*) and a shoulder around 540 nm. Additionally these spectra have a broad maximum in the blue, with peaks or shoulders at 443 nm (chl *a*), 460 nm (chl *c*) and 493 nm (carotenoid). The *Rhodomonas* sp. culture has

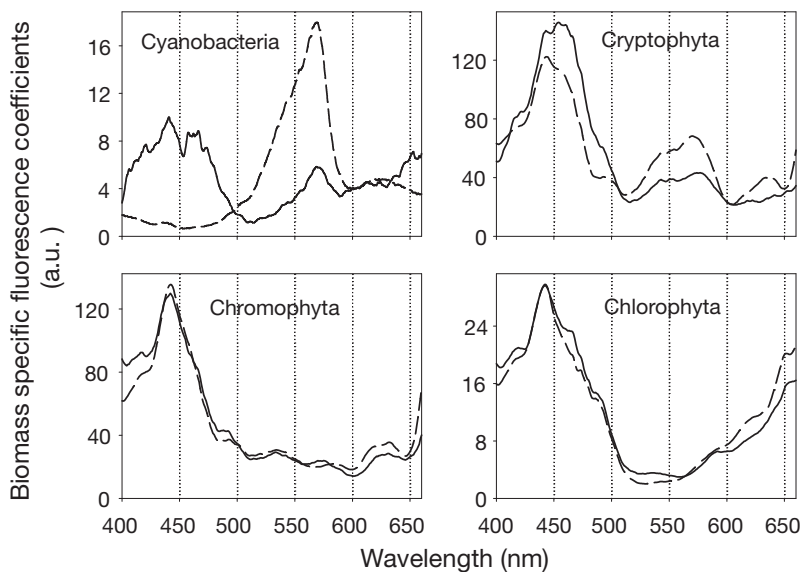


Fig. 4. Biomass specific spectral fluorescence coefficients, K , for the different phytoplankton groups as estimated with Eq. (6) (solid lines). For each group a typical spectra of a cultured species (*Synechococcus* sp. for cyanobacteria, *Rhodomonas* sp. for Cryptophyta, *Thalassiosira pseudonana* for Chromophyta and *Chlamydomonas* sp. for Chlorophyta; dashed lines) is shown for comparison. For better comparison, species spectra are scaled to the mean values of the corresponding K . a.u.: arbitrary units

an additional peak at 632 nm (chl *a*/chl *c*), not seen in K for Cryptophyta. The estimate of K for Chromophyta had a similar shape as the spectra for the diatom, *Thalassiosira pseudonana*, with the peaks and shoulders being located almost identically at 441 nm (chl *a*), 460 nm (chl *c*), 490 nm (carotenoid), 532 nm (carotenoid), 577 nm (chl *c*) and 630 nm (chl *a*/chl *c*). For Chlorophyta our estimate of K shows peaks and shoulders at 443 nm (chl *a*), 466 nm (chl *b*), 486 nm (chl *b*), 589 nm (unknown), 630 nm (chl *a*) and 651 nm (chl *b*). The spectrum of *Chlamydomonas* sp. has a similar

shape. Chlorophyta, which consisted mainly of *Eutreptiella gymnastica*, had lower K -values than did other eukaryotic groups; whether this is due to differences in their photosystems or due to a lower cellular pigment content of *E. gymnastica* cannot be concluded from our data.

Selection of PCR- and PLS-models

The first eigenvector of the principal component analysis had the shape of the average sample spectra (Fig. 5) and already contained 99.47% of the spectral variability, and the scores correlated with the biomass of Chlorophyta ($r = 0.94$, $n = 28$, $p < 0.01$). The second eigenvector had a sharp peak at 460 nm and a wide peak from 540 to 580 nm, the scores being best correlated with Cryptophyta ($r = 0.50$, $n = 28$, $p < 0.01$). The spectral shapes of the additional eigenvectors were not explicitly related to species spectra. As seen in the RMSEP curves (Fig. 6), for Chlorophyta and Cryptophyta the best prediction, i.e. the lowest RMSEP values, was obtained with 5 eigenvectors. For other groups or total biomass the optimal number of eigenvectors was from 3 to 6. As a result, 5 eigenvectors were selected for the final model. These explained 99.95% of the spectral variability and the remaining eigenvectors showed somewhat pure spectral noise (data not shown).

Similarly, in PLS the factors are selected top-down, until the model is optimal. Again, the first factor had

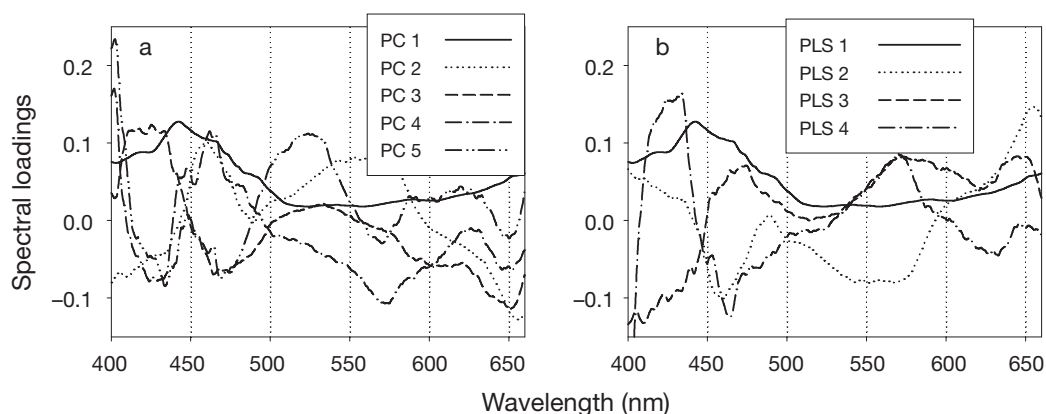
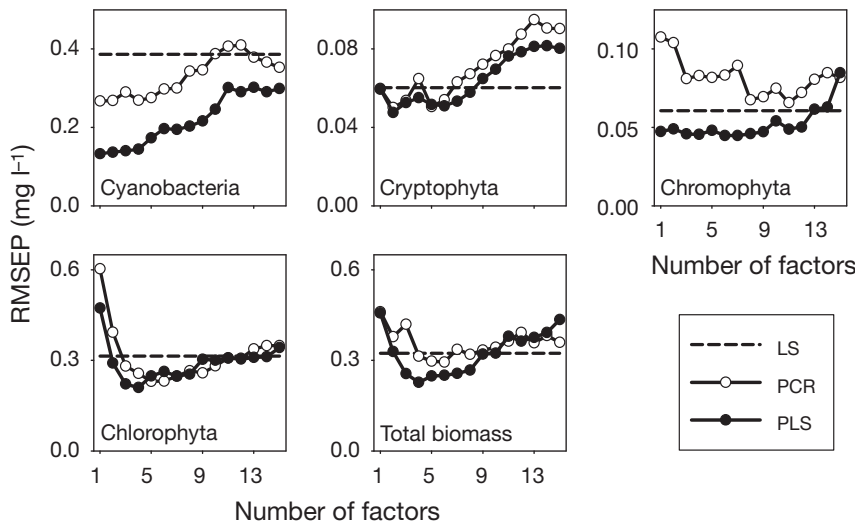


Fig. 5. Loading spectrum (A) for the 5 largest eigenvectors (PC1 to PC5) used in the PCR model and (B) for the 4 factors used in the PLS model (PLS1 to PLS4)



the shape of average spectra, and the second factor had the inverse shape of the second principal component of the PCR model ($r = -0.99$, $n = 281$, $p < 0.01$). The shapes of the 3rd and 4th factors are more difficult to interpret. The predictions for cyanobacteria, Cryptophyta or Chromophyta were not much affected by the addition of the first few factors, but a clear minimum in RMSEP was noted for Chlorophyta when 4 factors were used (Fig. 6). Consequently 4 factors were used in the final PLS models.

Model predictions

Fig. 6. Comparison of calibration models using RMSEP obtained in full cross validation. A smaller RMSEP indicates better predictive ability. For LS, only one RMSEP can be calculated for each component. For PCR and PLS, several models, with variable numbers of eigenvectors (from 1 to 15), were calculated; optimal models were obtained with 5 and 4 eigenvectors, respectively

Using the full cross validation, the phytoplankton biomass in each sample was predicted (Fig. 7). For cyanobacteria, the predictions by LS or PCR had a

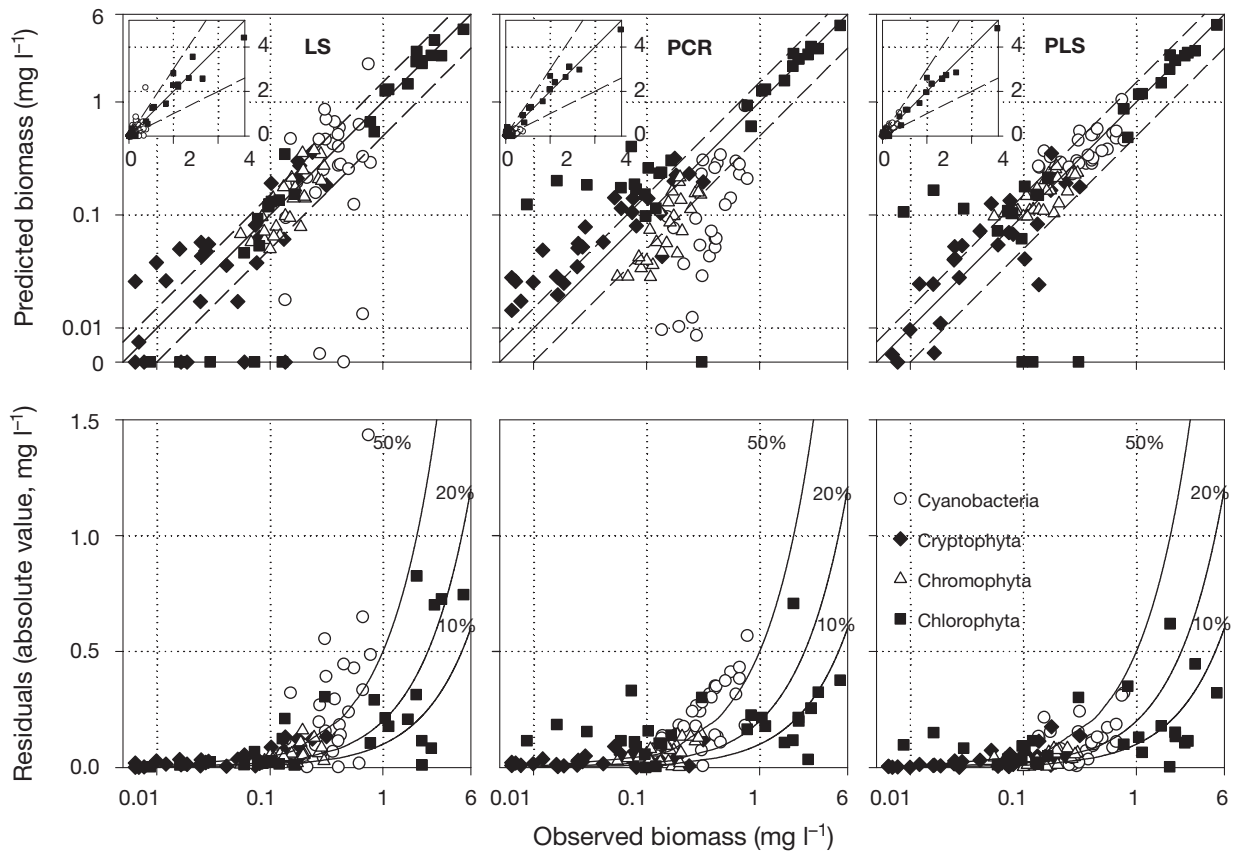


Fig. 7. Relationship between observed and predicted phytoplankton biomass for the different models using logarithmic (upper panels) and linear scales (inserts). Solid lines show the 1:1 relation, while dashed lines show 50% prediction errors. Predictions are carried out using full cross validation. The absolute prediction residuals (lower panels) are calculated for each model as the absolute difference between observed and predicted values. Levels for prediction errors at 10, 20 and 50% are shown

Table 1. Regression statistics between observed and predicted phytoplankton biomass. In the prediction, 3 different regression models were compared and 3 different validation methods were used. Regression intercept and slope (95% confidence levels in parentheses) are given for full cross validation. Coefficient of determination (r^2), root mean square error for prediction (RMSEP, mg l^{-1}) and bias (mg l^{-1}) are given for each method

	Intercept	Slope	Full cross validation			Partial cross validation			External validation		
			r^2	RMSEP	Bias	r^2	RMSEP	Bias ²	r^2	RMSEP	Bias
LS											
Cyanobacteria	0.09 (0.41)	0.79 (0.97)	0.12	0.387	0.013	0.06	0.602	0.011			
Cryptomonads	0.00 (0.04)	0.99 (0.34)	0.63	0.060	0.002	0.53	0.072	0.018	0.05	0.024	-0.015
Chromophyta	-0.01 (0.07)	1.04 (0.40)	0.58	0.061	-0.007	0.42	0.071	-0.017	0.01	0.066	0.052
Chlorophyta	0.05 (0.18)	0.95 (0.11)	0.94	0.314	-0.004	0.96	0.264	0.031	0.95	0.372	-0.274
Total biomass	-0.03 (0.22)	1.02 (0.10)	0.96	0.323	0.006	0.90	0.574	0.043			
PCR											
Cyanobacteria	-0.10 (0.15)	0.66 (0.35)	0.42	0.275	-0.229	0.51	0.267	-0.228			
Cryptomonads	0.03 (0.03)	0.81 (0.25)	0.68	0.051	0.019	0.39	0.079	0.031	0.11	0.020	-0.002
Chromophyta	-0.01 (0.04)	0.63 (0.24)	0.58	0.082	-0.069	0.47	0.086	-0.067	0.04	0.028	0.009
Chlorophyta	0.13 (0.12)	0.96 (0.07)	0.97	0.230	0.084	0.97	0.290	0.161	0.93	0.243	-0.115
Total biomass	-0.18 (0.15)	0.99 (0.07)	0.98	0.297	-0.196	0.96	0.378	-0.103			
PLS											
Cyanobacteria	0.16 (0.13)	0.61 (0.32)	0.43	0.144	0.012	0.47	0.145	0.009			
Cryptomonads	0.02 (0.03)	0.79 (0.30)	0.59	0.055	0.000	0.44	0.072	0.009	0.52	0.017	-0.009
Chromophyta	0.06 (0.04)	0.63 (0.25)	0.57	0.046	-0.001	0.44	0.057	-0.001	0.02	0.075	0.070
Chlorophyta	0.03 (0.12)	0.96 (0.07)	0.97	0.210	-0.010	0.97	0.251	0.045	0.92	0.290	-0.193
Total biomass	0.02 (0.15)	0.99 (0.07)	0.98	0.227	-0.005	0.96	0.352	0.053			

lot of scatter, and PCR predictions were clearly biased (Table 1). Consequently the RMSEP value for cyanobacteria was lowest in PLS. The scatter in the prediction of Cryptophyta was smaller, and not biased with any method. For this group the observed biomass was sometimes very low, and on some occasions the LS model especially predicted their total absence. The biomass range for Chromophyta was narrow, from 0.06 to 0.30 mg l^{-1} , and the predictions by PCR were negatively biased. The LS and PLS predictions for this group were comparable, the latter having a slightly lower RMSEP and bias. For Chlorophyta and total biomass, the relationship between observed and predicted biomass was linear in each model, with a slope close to 1, an intercept close to 0 and with a high coefficient of determination ($r^2 > 0.94$). PCR tended to systematically overestimate the biomass of Chlorophyta and underestimate total biomass.

For all methods the relative prediction error (absolute model residuals divided by the observed concentrations) was higher when the biomass was lower. Using this relationship we may determine rough detection limits. If, for example, we consider the limit of acceptable model error as 20% for observed biomass, we get thresholds of 1.0, 0.9 and 0.3 mg l^{-1} for the LS, PCR and PLS methods, respectively. For LS and PCR, the relative error seems higher for cyanobacteria than for Chlorophyta, while the errors are similar when the PLS model is used. In the low concentration part, the relative error in bio-

mass prediction is not the most relevant measure of model success. In this case, the absolute residuals should be considered. For PLS 71% of the absolute residuals were smaller than 0.1 mg l^{-1} and 92% were smaller than 0.2 mg l^{-1} . For LS the same cumulative frequencies were 64% and 80%, and for PCR were 50% and 78%, respectively. Absolute residuals larger than 0.4 mg l^{-1} were rare: 1.8%, 8.9% and 3.6% of the residuals in PLS (2 cases), LS (10 cases) and PCR (4 cases), respectively.

A more robust validation was obtained using partial cross validation, where one unit at a time was removed from the data set used for calibration. As expected, the predictive ability of all the models somewhat decreased, except for cyanobacteria when using PCR or PLS, and for Chlorophyta when using LS. However, there were no major changes in prediction errors, or in the relative success of the models (Table 1).

In the last validation test, the models made using the upper layer samples were used to predict biomass in the external set of samples taken from the deep part of mesocosms. As there were no observations of the picocyanobacterial biomass, predictions for cyanobacteria or total biomass were not made. Here, only the biomass of Chlorophyta was accurately predicted, and PCR gave the closest matchup. The observed biomass for other groups was low ($<0.11 \text{ mg l}^{-1}$) and predictions were highly scattered with all models. Due to the low biomass, RMSEP and bias values remained low.

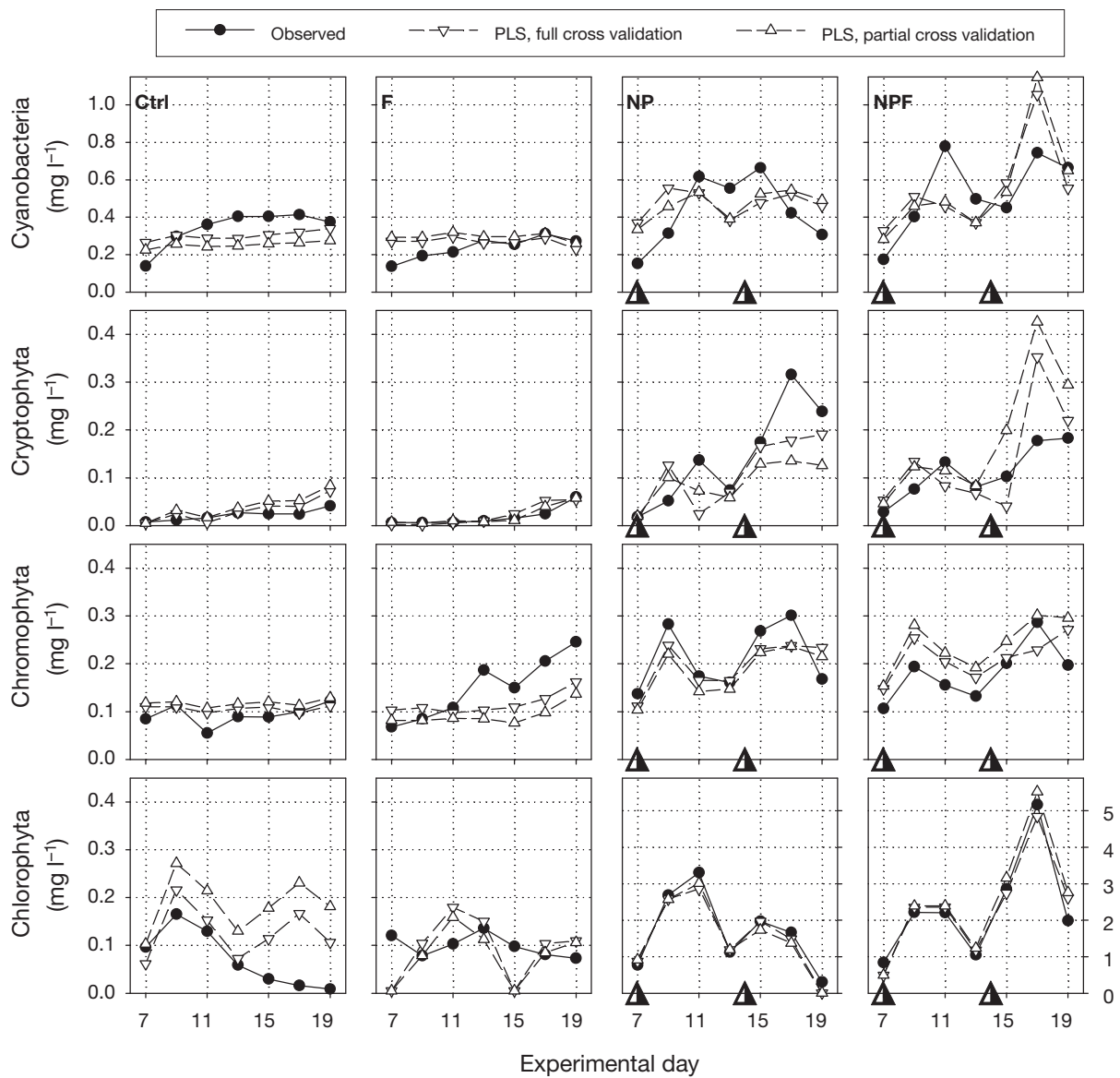


Fig. 8. Development of phytoplankton biomass in different pigment groups as observed by microscopy and estimated by the PLS model using full cross validation and partial cross validation. Arrowheads indicate the times of pulsed nutrient additions. The y-axis values for the NP and NPF units for Chlorophyta are given on the right-hand side

For each phytoplankton group, time series were plotted for biomass values obtained with PLS using full and partial cross validation together with the observed values (Fig. 8). The differences between the 2 validation methods were minor, and both predicted correctly the overall biomass levels for each group and for each experimental unit. The *Eutroptiella gymnastica* blooms in the NP and NPF units were accurately predicted by PLS. On the other hand, in the Ctrl and F units the temporal trends in Chlorophyta biomass, which was very low, were not correctly predicted. For the other groups, the prediction was somewhat better in the Ctrl and F

units than in the NP and NPF units. Simply, in the latter units the *E. gymnastica* biomass dominated and obviously masked the fluorescence signals from the other groups, challenging the model behaviour.

Dinophyceae contributed from 16 to 93% of the biomass of Chromophyta, and from 2 to 20% of total phytoplankton biomass. Despite its low biomass, Dinophyceae was accurately predicted by PLS-1 (Fig. 9). The PLS-1 retrieval of filamentous cyanobacteria and pico-chl *a* were acceptable as well. Surprisingly, picocyanobacteria were not correctly retrieved by PLS-1.

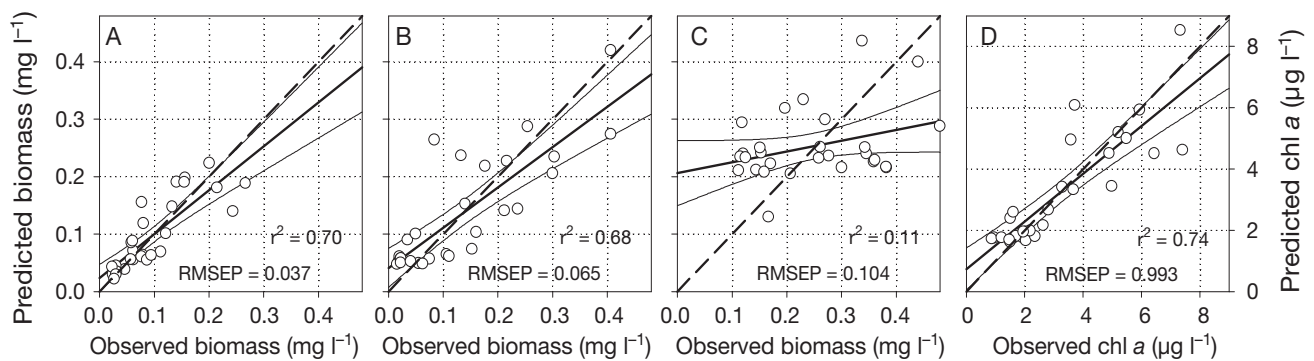


Fig. 9. Relationship between observed and PLS-1 predicted biomass for (A) Dinophyceae, (B) filamentous cyanobacteria, (C) picocyanobacteria and (D) pico-chl *a*. Predictions are carried out using full cross validation. Dashed lines show the 1:1 relation, bold solid lines show the linear least-squares regression fit and thin solid lines show the 95% confidence limits

DISCUSSION

Collinearity problems

A fundamental problem in multicomponent spectral calibration models is the possible collinearity of component concentrations in matrix *C* (Næs & Mevik 2001). Primarily collinearity is to be avoided by careful design of the calibration samples, and the resulting *C* should preferably be orthogonal. When the calibration dataset is based on observations from a natural system, as in our case, there is a risk that some of the components are collinear. For the application described here, collinearity in *C* would mean that, despite changes in the phytoplankton biomass, there are no notable changes in the community structure.

The goodness of the calibration design should be evaluated. Effectively, for a small number of components, multicollinearity can be visually checked by plotting concentrations in *n*-dimensional space. When $n \geq 4$, these plots are not available and other methods have been described (Martens & Næs 1989, Kalivas & Lang 1994). First, for a usable calibration design, *C* must have full rank. This means that the number of calibration samples must be larger than the number of compounds to be predicted, and that concentration data of different compounds are not linearly related but contribute unique pieces of information. Secondly, measurement of the distance of *C* from rank deficiency should be performed. Kalivas & Lang (1994) noted that for a good calibration design the condition number of *C* should be small relative to the largest singular value, indicating that *C* is nearly orthogonal. Further, the value of the smallest singular value should be 'large,' indicating that *C* is not close to rank deficient. In our case *C* had a full rank of 4 and the largest and smallest singular values were 9013 and 284.5, yielding a condition number of 31.68. According to criteria by Kalivas

& Lang (1994), our calibration set seemed to be nearly orthogonal and suitable for modelling.

Our samples originated from manipulated natural phytoplankton communities, and the overall variability in the community structure was large, thereby minimising the risk of collinearity. In future studies with natural systems, special attention must be paid to the collection of calibration samples. One possibility is to collect a large set of samples from which the subset for microscopy is selected.

Poor spectral orthogonality in the *K* matrix, meaning that the intensities at fixed wavelengths are dependent or nearly dependent, should seriously be taken into account (Kalivas & Lang 1994). In a really unfortunate case the poor orthogonality arises because the spectra of one component can be represented as the sum of the other components. For example, such a case would occur if the spectra of Cryptophyta can be represented as the sum of the spectra of cyanobacteria and Chromophyta (which was not the case in our study). The obvious way to treat poor spectral orthogonality is to use factorial regressions, like PCR and PLS, where the spectral data is compressed into a new set of orthogonal variables (Martens & Næs 1989, Næs & Mevik 2001).

Methods to discriminate phytoplankton spectral groups

The methods described here are based on the spectral variability between different phylogenetic phytoplankton groups; typically, 4 to 5 classes can be discriminated (Poryvkina et al. 1994, Millie et al. 2002). For the discrimination of phytoplankton spectral groups, several statistical approaches are available. First, spectral libraries using pure cultures of representative species can be collected, especially using envi-

ronmental conditions close to those found in the water body under study. The algorithms used to derive taxonomic information from the spectra, with the help of libraries, include the similarity index (e.g. Millie et al. 1997) and least-squares techniques with modifications (Gerhardt & Bodemer 2000, Beutler et al. 2002). For analytical spectroscopy, e.g. in the analysis of pure pigments in organic solvent (Neveux & Lantoiné 1993), the least-squares technique is practical as long as all of the components are included in the model. If this assumption is violated, or if the spectral library is not representative, erroneous predictions are to be expected.

The second method, our LS approach, is operationally similar to the first one, but the spectra for the groups (K) are derived during a separate calibration step with samples of known taxonomic composition. This reduces possible problems caused by pigment acclimation in nature relative to cultures and by the fact that some species abundant in nature are difficult to culture. The workload of this method is a careful taxonomic examination of calibration samples followed by a grouping based on the pigmentation of the various taxonomic classes. In our study, the calculated K spectra for each pigment group yielded comparable spectra to those of the cultures, with the exception of cyanobacteria. Occasionally LS methods may yield a negative concentration for a certain component, due to noisy signals, a variable background or unknown components. This can be avoided by using a non-negative least-squares method, which assigns zero to negative results, and makes a new calculation with fewer components involved. The most obvious problem in LS, when used in spectroscopic applications, is the inability to separate signals due to a variable background and unknown components.

In the third method, PCR, most of the variability in the sample spectra is represented by a smaller number of variables, eigenvectors. The concentration data and the PCR scores are then related to create regression coefficients, which are subsequently used in the prediction. To avoid under- or overfitting, the number of eigenvectors must be optimised. PCR does not require knowledge of all the components present in a sample, but, on the other hand, there is no guarantee that the principal component scores are related to the concentrations of the preferred components.

The 4th alternative, PLS, differs from PCR as it uses the correlation between concentration and spectral data when building up the predictive model. The model results in decomposing both concentration data and spectral data into separate eigenvectors and scores, and the predictive power of the model is optimised by maximising the covariance of the scores. In

other words, PLS tries to find those eigenvectors that are most relevant for predicting the concentrations. The power of PLS is that it can deal with collinear spectral data. The resulting PLS models are not trivial, and the parsimony principle, i.e. low dimensionality, should be preferred to avoid overfitting (Martens & Næs 1989).

In our study, the prediction errors of the PLS method were generally smaller than those of LS and PCR, though in some cases all the methods seem to predict equally well. When comparing the similarity index and PLS, Stæhr & Cullen (2003) indicated PLS as the preferred method for single phytoplankton species discrimination from absorption data. The good performance of PLS over the other methods has been demonstrated for other applications as well (e.g. Reis & Saraiva 2004, Trygg 2004). However, there is no *a priori* knowledge as to which method gives the most accurate prediction, and the model outcomes should be compared analytically, as in our study.

Prediction error may arise from several sources. One reason may be that the assumptions of the model are not fulfilled. As an example of this, LS methods may fail due to existence of components not included in the model. Another source of prediction error is noisy data. In our study, major problems in prediction took place when retrieving biomass for groups making only a minor contribution to the total phytoplankton biomass. It is clear that minor groups make a low and noisy contribution to the total fluorescence signal, challenging a reliable prediction. Moreover, microscopically derived biomass estimates are not very trustworthy for species or groups with low abundance.

For the methods described, there are 3 sources of error related to pigmentation. First, some of the pigment groups contain several phytoplankton classes, and all of them contain numerous species. Consequently, the diversity of pigmentation inside each pigment group is large (Jeffrey & Vesik 1997). Secondly, the amount of light-harvesting pigments is regulated as a response to environmental variables, especially to light and nutrients (e.g. Johnsen & Sakshaug 1996, Stæhr et al. 2002). Adjustment of the pigmentation is especially large for cyanobacteria that may use their phycobilins as nitrogen reserves. Third, a variable fluorescence yield, related to photosynthetic activity and the physiological state of cells (Falkowski & Kiefer 1985) may introduce an additional error source. This could be avoided in future studies by using an inhibitor of photosynthesis like 3-(3,4-dichlorophenyl)-1,1-dimethylurea (DCMU). As a conclusion, no single constant spectra exist for any group, making the all-purpose solution unattainable.

For the reliable prediction of future samples, outside the validation dataset, it is important to have proper

calibration data. It is clear that changes in the phytoplankton community structure or pigment acclimation, not included in the spectral library or the calibration set, may result in poor prediction. This was seen in our study partly as a decrease of predictive ability when partial cross validation was used. A further reduction was noted when external validation was carried out. Then, a possible reason for the poor predictions for Cryptophyta and Chromophyta was their differences in light acclimation between the calibration (surface) and prediction (deep layer) data sets. The prediction of Chlorophyta was not that much altered, probably as they dominated the overall biomass and had a much lower relative measurement error. Additionally, the major species, *Eutreptiella gymnastica*, is a fast-swimming, vertically-migrating flagellate and possibly is not much influenced by different light conditions.

Model interpretation, besides prediction, is another important property of calibration models (Trygg 2004). LS models are simple, and the spectral coefficients obtained from the model can be compared with the measured ones. Further, during the prediction of future samples, the spectral fit can be evaluated. For more abstract models like PLS, the spectral loadings do not directly quantify the spectral coefficients. Spectral coefficients can, however, be estimated from the PLS regression coefficients as described by Trygg (2004).

Application of the method in the Baltic Sea

The high variability in phytoplankton biomass and community composition created by the manipulations provided us with comprehensive test data for the optical discrimination of phytoplankton groups. The predictive ability of the PLS model was clearly adequate for monitoring the bloom of *Eutreptiella gymnastica*. For the other groups, or generally during cases with lower biomass, prediction was less accurate. If we set requirements for the prediction that errors must be below 0.1 mg l^{-1} or, alternatively, that the residuals should not exceed $\pm 20\%$ of the observed biomass, 79% of the PLS predicted biomass values are acceptable. Restricting these constraints down to 0.05 mg l^{-1} for errors and residuals at $\pm 10\%$ of the observed biomass will diminish the acceptable predictions down to 57%. These prediction error levels should be compared with the errors in the determination of cell abundances by microscopy, problems in the determination of cell volumes for various species, variability in the biomass-to-pigment ratio and changes in the variable fluorescence. Different water volumes for fluorescence measurements and microscopy may introduce further errors for rare and large species.

The overall biomass levels in our calibration samples ranged from 2.5 to $45 \mu\text{g chl } a \text{ l}^{-1}$ or from 0.3 to $6.4 \text{ mg fresh weight l}^{-1}$. During the phytoplankton growth season, this level of biomass is representative for the various Baltic Sea sub-basins (HELCOM 2002, Thamm et al. 2004, Raateoja et al. 2005, Seppälä et al. 2007). At some coastal locations, higher biomass levels can be found, especially during the spring bloom (e.g. Seppälä & Balode 1999). In spring, diatoms and dinoflagellates typically dominate the community (Wasmund et al. 1998, Spilling et al. 2006), and their pigmentations are similar, with the chl *a*/chl *c*/carotenoid complex harvesting light for PSII (Jeffrey & Vesik 1997). In this case, discrimination of these 2 groups using optical methods is not straightforward, but experiments with pure cultures are promising (J. Seppälä & K. Spilling unpubl.). Further, during this study we demonstrated that dinoflagellates could be discriminated from the rest of the groups using PLS, though diatoms were not present. After the spring bloom, during the early summer minimum period, low biomass values can be found especially in the open sea (HELCOM 2002, Seppälä et al. 2007), and then it would be a challenge to determine the optical discrimination of phytoplankton with the method described here. Later in the summer, intensive blooms of filamentous cyanobacteria are typical (Sellner 1997, Kahru et al. 2000, Seppälä et al. 2007). Mass occurrence of other species, like *Heterocapsa triquetra* (Dinophyta) (Kononen et al. 1999, Lindholm & Nummelin 1999), *Dinophysis* spp. (Dinophyta) (Meyer-Harms & Pollehne 1998) and *Chlamydomonas* sp. (Chlorophyta) (HELCOM 2002), occasionally takes place as well. Moreover, some groups not forming blooms, like picocyanobacteria and cryptophytes (Kuparinen & Kuosa 1993, Suikkanen et al. 2007), could, at times, be the main constituents of the phytoplankton community.

The poor prediction of cyanobacteria as a single group was due to problems with picocyanobacteria. In our experiment the abundance of filamentous cyanobacteria was low, >10 -fold less than during bloom situations. However, PLS-1 was able to predict this group accurately. On the other hand, PLS-1 failed in the prediction of picocyanobacteria. This is surprising, considering that picocyanobacteria were more abundant than filamentous cyanobacteria in our experiment, and due to their smaller size they should have been evenly distributed in samples with sufficiently high biomass levels. Considering that picocyanobacteria contributed $>71\%$ of the picophytoplankton biomass, this result is in contradiction with the accurate prediction of picochl *a*. However, picocyanobacteria is a non-homogeneous group with variable cell sizes and colony types, which presents a challenge when using quantitative microscopy that uses an average cell size in biomass

calculations. Further, the phycobilin pigmentation of picocyanobacteria is dynamic, reflecting the light and nutrient conditions.

In our experiment, the background filtrate signal, due to CDOM, was subtracted from the sample spectra and, thus, did not affect the outcome of any discrimination method. The background was noted as a constant during the experiment. This reflects the domination of allochthonous CDOM in the Baltic Sea (Kowalczyk et al. 2006), and as a consequence of which phytoplankton-derived CDOM has a minor effect for optical signals. If samples are taken from different sea areas with variable CDOM content, the background will be variable presenting problems especially for LS methods.

Future perspectives

Monitoring of the phytoplankton community structure using traditional water sampling and microscopy is time-consuming, and the spatial and seasonal variations are not easily resolved at the scales required for thorough analyses of phytoplankton dynamics. An alternative method for phytoplankton chemotaxonomy, based on pigment analyses by HPLC has also been used in the Baltic Sea (Wänstrand & Snoeijis 2006), but only a limited number of samples can be analysed with this method. Advanced flow cytometer methods have recently been developed, but these are in the exploratory phase (Sosik & Olson 2007). The chemotaxonomic method based on *in vivo* spectral fluorescence, as presented here, should not be considered as an alternative to these methods. Rather, it is a supplementary tool allowing detection of the more detailed spatio-temporal dynamics of phytoplankton.

We used the excitation spectra of chl *a* using a fixed emission wavelength at the chl *a* maximum. A better discrimination between cyanobacteria and eukaryotes, and between different cyanobacterial groups can be achieved if 3-dimensional fluorescence spectroscopy is used, including the emission wavelengths of various phycobilins. The resulting excitation–emission matrix cannot be adequately analysed by the PCR and PLS methods described here, but more sophisticated multi-way models like parallel factor analysis (PARAFAC) and n-PLS could be used (Bro 1996, 1998, Andersen & Bro 2003).

On the whole, the biomass levels as well as the chemotaxonomic composition of the phytoplankton were predicted by PLS at the level needed for bloom detection. The next step with the method described is to make measurements in truly natural systems, by measuring the seasonal succession and following bloom events. Some of these activities have already started, including spectral absorption and reflectance measurements.

Acknowledgements. We thank the members of the PELAG III research group and the staff of the Tvärminne Zoological Station for their assistance during the experiment, Dr. P. Kuoppo for picophytoplankton counts, Drs. S. and G. Hällfors (Finnish Institute of Marine Research) and the working group of Dr. L. Stal (NIOO-KNAW, Centre for Estuarine and Marine Ecology) for providing the phytoplankton cultures, and R. King for improving the language. Critical comments by Dr. S. Kaitala and the anonymous reviewers are acknowledged. The Academy of Finland, the Maj and Tor Nessling foundation (J.S.), NorFa and the Estonian Science Foundation (ESF 6470; K.O.) supported this study financially.

LITERATURE CITED

- Alpine AE, Cloern JE (1985) Differences in *in vivo* fluorescence yield between three phytoplankton size classes. *J Plankton Res* 7:381–390
- Andersen CM, Bro R (2003) Practical aspects of PARAFAC modeling of fluorescence excitation-emission data. *J Chemometr* 17:200–215
- Anderson JM, Barrett J (1986) Light-harvesting pigment-protein complexes of algae. In: Staehelin LA, Arntzen CJ (eds) *Photosynthesis III*. *Encl Plant Phys Vol 19*, Springer-Verlag, Berlin, p 269–285
- Arvola L (1981) Spectrophotometric determination of chlorophyll *a* and phaeopigments in ethanol extractions. *Ann Bot Fenn* 18:221–227
- Babichenko S, Kaitala S, Leeben A, Poryvkina L, Seppälä J (1999) Phytoplankton pigments and dissolved organic matter distribution in the Gulf of Riga. *J Mar Syst* 23: 69–82
- Beutler M, Wiltshire KH, Meyer B, Moldaenke C and others (2002) A fluorometric method for the differentiation of algal populations *in vivo* and *in situ*. *Photosynth Res* 72: 39–53
- Bro R (1996) Multi-way calibration. Multi-linear PLS. *J Chemometr* 10:47–62
- Bro R (1998) Multi-way analysis in the food industry, theory, algorithms and application. PhD dissertation, University of Amsterdam
- Bryant DA (1982) Phycoerythrocyanin and phycoerythrin: properties and occurrence in cyanobacteria. *J Gen Microbiol* 128:835–844
- Cleveland JS, Perry MJ (1987) Quantum yield, relative specific absorption and fluorescence in nitrogen-limited *Chaetoceros gracilis*. *Mar Biol* 94:489–497
- Cowles TJ, Desiderio RA, Neuer S (1993) In situ characterization of phytoplankton from vertical profiles of fluorescence emission spectra. *Mar Biol* 115:217–222
- Cullen JJ, Ciotti AM, Davis RF, Lewis MR (1997) Optical detection and assessment of algal blooms. *Limnol Oceanogr* 42:1223–1239
- Dandonneau Y, Neveux J (1997) Diel variations of *in vivo* fluorescence in the eastern equatorial Pacific: an unvarying pattern. *Deep-Sea Res II* 44:1869–1880
- Falkowski P, Kiefer DA (1985) Chlorophyll *a* fluorescence in phytoplankton: relationship to photosynthesis and biomass. *J Plankton Res* 7:715–731
- Gerhardt V, Bodemer U (2000) Delayed fluorescence excitation spectroscopy: a method for determining phytoplankton composition. *Arch Hydrobiol Spec Issue Adv Limnol* 55:101–119
- Guo C, Dunstan WM (1995) Depth-dependent changes in chlorophyll fluorescence number at a Sargasso Sea station. *Mar Biol* 122:333–339

- HELCOM (Helsinki Commission) (2002) Environment of the Baltic Sea area 1994–1998. Baltic Sea Environ Proc 82B
- Jeffrey SW, Vesik M (1997) Introduction to marine phytoplankton and their pigment signatures. In: Jeffrey SW, Mantoura RFC, Wright SW (eds) Phytoplankton pigments in oceanography: guidelines to modern methods. UNESCO, Paris, p 37–84
- Jeffrey SW, Mantoura RFC, Wright SW (eds) (1997) Phytoplankton pigments in oceanography: guidelines to modern methods. UNESCO, Paris
- Johnsen G, Sakshaug E (1996) Light harvesting in bloom-forming marine phytoplankton: species-specificity and photoacclimation. In: Figueroa FL, Jiménez C, Pérez-Lloréns JL, Niell FX (eds) Underwater light and algal photobiology. Sci Mar 60 (Suppl 1):47–56
- Johnsen G, Sakshaug E (2007) Biooptical characteristics of PSII and PSI in 33 species (13 pigment groups) of marine phytoplankton, and the relevance for pulse-amplitude-modulated and fast-repetition-rate fluorometry. J Phycol 43:1236–1251
- Johnsen G, Samset O, Granskog L, Sakshaug E (1994) *In vivo* absorption characteristics in 10 classes of bloom-forming phytoplankton: taxonomic characteristics and responses to photoadaptation by means of discriminant and HPLC analysis. Mar Ecol Prog Ser 105:149–157
- Kahru M, Leppänen JM, Rud O, Savchuk OP (2000) Cyanobacteria blooms in the Gulf of Finland triggered by saltwater inflow into the Baltic Sea. Mar Ecol Prog Ser 207:13–18
- Kaitala S, Babichenko S, Poryvkina L, Leeben A (1994) Fluorescent analysis of pigment composition of natural phytoplankton. Mar Technol Soc J 28:50–58
- Kalivas JH, Lang PM (1994) Mathematical analysis of spectral orthogonality. Marcel Dekker, New York
- Kiefer DA (1973) Chlorophyll a fluorescence in marine centric diatoms: responses of chloroplasts to light and nutrient stress. Mar Biol 23:39–46
- Kononen K, Huttunen M, Kanoshina I, Laanemets J, Moisaner P, Pavelson J (1999) Spatial and temporal variability of a dinoflagellate–cyanobacterium community under a complex hydrodynamical influence: a case study at the entrance to the Gulf of Finland. Mar Ecol Prog Ser 186:43–57
- Kopf U, Heinze J (1984) 2,7-Bis-(diethylamino) phenazonium chloride as a quantum counter for emission measurements between 240 and 700 nm. Anal Chem 56:1931–1935
- Kowalczyk P, Stedmon CA, Markager S (2006) Modeling absorption by CDOM in the Baltic Sea from season, salinity and chlorophyll. Mar Chem 101:1–11
- Kuparinen J, Kuosa H (1993) Autotrophic and heterotrophic picoplankton in Baltic Sea. Adv Mar Biol 29:73–128
- Legendre P, Legendre L (1998) Numerical ecology. Elsevier, Amsterdam
- Lindholm T, Nummelin C (1999) Red tide of the dinoflagellate *Heterocapsa triquetra* (Dinophyta) in a ferry-mixed coastal inlet. Hydrobiologia 393:245–251
- Lohrenz SE, Weidemann AD, Tuel M (2003) Phytoplankton spectral absorption as influenced by community size structure and pigment composition. J Plankton Res 25:35–61
- Lorenzen CJ (1966) A method for continuous measurement of *in vivo* chlorophyll concentration. Deep-Sea Res 13: 223–227
- Martens H, Næs T (1989) Multivariate calibration. John Wiley & Sons, Chichester
- Meyer-Harms B, Pollehne F (1998) Alloxanthin in *Dinophysis norvegica* (Dinophysiales, Dinophyceae) from the Baltic Sea. J Phycol 34:280–285
- Millie DF, Schofield OM, Kirkpatrick GJ, Johnsen G, Tester PA, Vinyard BT (1997) Detection of harmful algal blooms using photopigments and absorption signatures: a case study of the Florida red tide dinoflagellate, *Gymnodinium breve*. Limnol Oceanogr 45:1240–1251
- Millie DF, Schofield OME, Kirkpatrick GJ, Johnsen G, Evens TJ (2002) Using absorbance and fluorescence spectra to discriminate microalgae. Eur J Phycol 37:313–322
- Moberg L, Karlberg B, Blomqvist S, Larsson U (2000) Comparison between a new application of multivariate regression and current spectroscopy methods for the determination of chlorophylls and their corresponding pheopigments. Anal Chim Acta 411:137–143
- Moberg L, Karlberg B, Sørensen K, Källqvist T (2002) Assessment of phytoplankton class abundance using absorption spectra and chemometrics. Talanta 56:153–160
- Næs T, Mevik BH (2001) Understanding the collinearity problem in regression and discriminant analysis. J Chemometr 15:413–426
- Neveux J, Lantoin F (1993) Spectrofluorometric assay of chlorophylls and pheophytins using the least squares approximation technique. Deep-Sea Res 40:1747–1765
- Niemi Å (1975) Ecology of phytoplankton in the Tvärminne area, SW coast of Finland. II. Primary production and environmental conditions in the archipelago and the sea zone. Acta Bot Fenn 105:1–73
- Oldham PB, Zilliox EJ, Warner IM (1985) Spectral ‘fingerprinting’ of phytoplankton populations by two-dimensional fluorescence and Fourier-transform-based pattern recognition. J Mar Res 43:893–906
- Olli K, Heiskanen AS, Seppälä J (1996) Development and fate of *Eutreptialla gymnastica* bloom in nutrient enriched enclosures in the coastal Baltic Sea. J Plankton Res 18: 1587–1604
- Poryvkina L, Babichenko S, Kaitala S, Kuosa H, Shalapjonok A (1994) Spectral fluorescence signatures in the characterization of phytoplankton community composition. J Plankton Res 16:1315–1327
- Raateoja M, Seppälä J, Kuosa H, Myrberg K (2005) Recent changes in the productive state of the SW Finnish coast in the Baltic Sea. Ambio 34:188–191
- Rabinowitch E, Govindjee (1969) Photosynthesis. John Wiley & Sons, New York
- Reis MS, Saraiva PM (2004) A comparative study of linear regression methods in noisy environments. J Chemometr 18:526–536
- Schofield O, Grzymiski J, Bissett WP, Kirkpatrick GJ, Millie DF, Moline M, Roesler CS (1999) Optical monitoring and forecasting systems for harmful algal blooms: possibility or pipe dream. J Phycol 35:1477–1496
- Sellner KG (1997) Physiology, ecology, and toxic properties of marine cyanobacterial blooms. Limnol Oceanogr 42: 1089–1104
- Seppälä J, Balode M (1999) Spatial distribution of phytoplankton in the Gulf of Riga during spring and summer stages. J Mar Syst 23:51–67
- Seppälä J, Ylöstalo P, Kuosa H (2005) Spectral absorption and fluorescence characteristics of phytoplankton in different size fractions across a salinity gradient in the Baltic Sea. Int J Remote Sens 26:387–414
- Seppälä J, Ylöstalo P, Kaitala S, Hällfors S, Raateoja M, Maunula P (2007) Ship-of-opportunity based phycocyanin fluorescence monitoring of the filamentous cyanobacteria bloom dynamics in the Baltic Sea. Estuar Coast Shelf Sci 73:489–500
- Sidler WA (1994) Phycobilisome and phycobiliprotein structures. In: Bryant DA (ed) The molecular biology of cyanobacteria. Kluwer, Amsterdam, p 139–216

- Sosik HM, Olson RJ (2007) Automated taxonomic classification of phytoplankton sampled with imaging-in-flow cytometry. *Limnol Oceanogr Methods* 5:204–216
- Spilling K, Kremp A, Tamelander T (2006) Vertical distribution and cyst production of *Peridiniella catenata* (Dinophyceae) during a spring bloom in the Baltic Sea. *J Plankton Res* 28:659–665
- Stæhr PA, Cullen JJ (2003) Detection of *Karenia mikimotoi* by spectral absorption signatures. *J Plankton Res* 25:1237–1249
- Stæhr PA, Henriksen P, Markager S (2002) Photoacclimation of four marine phytoplankton species to irradiance and nutrient availability. *Mar Ecol Prog Ser* 238:47–59
- Suikkanen S, Laamanen M, Huttunen M (2007) Long-term changes in summer phytoplankton communities of the open northern Baltic Sea. *Estuar Coast Shelf Sci* 71:580–592
- Thamm R, Schernewski G, Wasmund N, Neumann T (2004) Spatial phytoplankton pattern in the Baltic Sea. *Coastline Rep* 4:85–109
- Trygg J (2004) Prediction and spectral profile estimation in multivariate calibration. *J Chemometr* 18:166–172
- Wänstrand I, Snoeijs P (2006) Phytoplankton community dynamics assessed by ships-of-opportunity sampling in the northern Baltic Sea: a comparison of HPLC pigment analysis and cell counts. *Estuar Coast Shelf Sci* 66:135–146
- Wasmund N, Nausch G, Matthäus W (1998) Phytoplankton spring blooms in the southern Baltic Sea—spatio-temporal development and long-term trends. *J Plankton Res* 20:1099–1117
- Wold S, Sjöström M, Eriksson L (2001) PLS-regression: a basic tool of chemometrics. *Chemom Intell Lab Syst* 58:109–130
- Wood AM, Phinney DA, Yentsch CS (1998) Water column transparency and the distribution of spectrally distinct forms of phycoerythrin-containing organisms. *Mar Ecol Prog Ser* 162:25–31
- Yentsch CS, Phinney DA (1985) Spectral fluorescence: an ataxonomic tool for studying the structure of phytoplankton populations. *J Plankton Res* 7:617–632
- Yentsch CS, Yentsch CM (1979) Fluorescence spectral signatures: the characterization of phytoplankton populations by the use of excitation and emission spectra. *J Mar Res* 37:471–483

Editorial responsibility: Hans Heinrich Janssen, Oldendorf/Luhe, Germany

*Submitted: November 29, 2007; Accepted: July 11, 2008
Proofs received from author(s): October 15, 2008*