Vol. 473: 53–66, 2013
doi: 10.3354/meps10065

MARINE ECOLOGY PROGRESS SERIES
Mar Ecol Prog Ser

Published January 21

Harmonizing large data sets reveals novel patterns in the Baltic Sea phytoplankton community structure

Kalle Olli[1,*], Olga Trikk[1], Riina Klais[1], Robert Ptacnik[2], Tom Andersen[3], Sirpa Lehtinen[4], Timo Tamminen[4]

[1]Institute of Ecology and Earth Sciences, University of Tartu, 51005 Tartu, Estonia
[2]ICBM, University of Oldenburg, 26382 Wilhelmshaven, Germany
[3]Department of Biology, University of Oslo, 0316 Oslo, Norway
[4]Marine Research Centre, Finnish Environment Institute, 00251 Helsinki, Finland

ABSTRACT: Phytoplankton forms the basis of aquatic food webs, and shifts in community composition reflect changes in environmental conditions. Despite the accepted importance, the processes behind maintaining spatial and temporal community structure and biodiversity at the base of the aquatic food web remain poorly described and understood. A recognized challenge hampering validation of ecological models and meta-analysis is the scarcity of large phytoplankton data sets. Compared to other aquatic data, harmonization of quantitative phytoplankton data sets from different sources and academic institutions has remained a major challenge. Here we demonstrate and examine processes used to compile and harmonize a large multi-sourced phytoplankton data set covering 40 yr of monitoring and over 15 000 quantitative samples from the Baltic Sea. We show differences in the quality of data among countries and analyze autocorrelation scales in field data. Phytoplankton community composition showed positive autocorrelation at a temporal scale of less than 30 d and had a recurrent pattern at a yearly interval. Both total biomass and community composition showed a positive spatial autocorrelation, but the extent of the data determines the autocorrelation scale and strength. We introduce a new strategy to select the best performing model to assess regional taxon richness in phytoplankton field data. The Weibull 4-parameter model showed both the best fit with data and robust parameter estimates at varying sample size.

KEY WORDS: Meta-analysis · Phytoplankton · Monitoring · Time-series · Autocorrelation · Species richness

INTRODUCTION

Marine ecosystems are changing rapidly in response to natural processes, human activities, and climate change (Beaugrand et al. 2010). Phytoplankton are ecologically and biogeochemically relevant indicators of marine ecosystem change because they conduct a large proportion of system-scale primary production and C cycling and are sensitive to environmental pressures (Boyce et al. 2010). Phytoplankton are dominant marine primary producers; they mediate nutrient flux and cycling and transfer organic matter to the benthos via vertical flux (Wassmann et al. 2003), as well as to higher trophic levels. As key primary producers, phytoplankton reflect the immediate effects of changes in the input of nutrients in coastal ecosystems. Demographic traits of phytoplankton make them particularly suitable for comparative analysis of ecosystem changes across local to regional to global scales.

Phytoplankton can be quantified by relatively simple and intercomparable sampling methods (HELCOM 1988). Researchers and governmental agencies around the world have relied on phytoplankton as a key indicator of water-quality monitoring programs. Many data sets are available now and have been presented in conferences and research papers (Moe et al. 2008, Jurgensone et al. 2011). However, our

[*]Email: kalle.olli@ut.ee

understanding of marine ecosystem change is incomplete because we have not adequately explored, inventoried, nor compared data sets collected by various independent researchers, agencies and countries. This is unfortunate because producing quantitative phytoplankton data is costly in terms of time and expertise required.

We believe the fragmentation of data is largely a methodological issue. Often, unifying and harmonizing phytoplankton data from a variety of sources has methodologically proven to be more challenging than anticipated (Moe et al. 2008). The problems stem from the independent and widely different formats in which the data are stored by various institutions. Equally challenging are the different traditions in taxonomic conventions, the ever-changing nomenclature and taxonomic knowledge.

Notes on phytoplankton abundance and even algal blooms in the Baltic Sea date back to the mid 19th century (Eichwald 1847, Olli 1996, Finni et al. 2001). Data sets with species abundances and biomasses, taken and processed with modern methods, are available since the mid-1960s (Wasmund et al. 2008, Jurgensone et al. 2011). Since then, a variety of national agencies and academic institutions around the Baltic Sea have engaged in monitoring phytoplankton composition and biomass and auxiliary chemical and physical properties of seawater. The Baltic Marine Environment Protection Commission (HELCOM, est. 1974) has been coordinating phytoplankton sampling schedules and binding methods through the Baltic Sea Monitoring Program (BMP, later COMBINE) since 1979.

The objective of this paper is to communicate our experience and potential pitfalls in the procedures of combining and harmonizing different phytoplankton data sets from several countries and institutions around the Baltic Sea. Next, we analyze the overall spatial and temporal distribution of the joint data set samples and the comparability of the phytoplankton data from different countries. In particular, we analyze whether differences in the sample-counting practices in different countries, e.g. the taxonomic resolution used, can be revealed from the data. Using the Baltic Sea phytoplankton data set as a case study, we encourage a wider use of historic data sets to tackle present-day ecological questions, like changing biodiversity or adaptation of communities to climate change. Through re-analysis of historic observational data, our second objective is to resolve some of the long-standing challenges in phytoplankton community structure that cannot be efficiently dealt with by using limited-scale local data sets. Firstly, we focus on autocorrelation to reveal scales of repeating patterns in phytoplankton observational field data. Common statistical tests make the assumption of independence of observations. This assumption cannot be rigorously assessed and is often violated in field studies with modest sample size. Statistical inference from temporally and/or spatially separated points can thus be incorrect, unless autocorrelation structures are used in the models. Second, predicting ongoing changes in biodiversity is one of the pressing issues in ecology and conservation biology (Hooper et al. 2012, Reich et al. 2012). Estimates of local or regional species richness depend on sample size and sampling effort and have commonly been evaluated by extrapolating from species accumulation curves (Colwell & Coddington 1994, Gotelli & Colwell 2001, Tjørve 2003, Ugland et al. 2003, Chapman & Underwood 2009). The form and model to describe species accumulation curve varies between organism types and habitats. Here we use the extensive data set to test the performance of common species accumulation curve models with phytoplankton filed samples. In model testing we go beyond the commonly used assessment of goodness of fit with data. We analyze the stability of model parameters with increasing sampling effort, which adds confidence to extrapolated species richness estimates.

MATERIALS AND METHODS

Data compilation procedure

Nine academic institutions around the Baltic Sea (Table 1) provided the quantitative phytoplankton data. The historic phytoplankton data were counted from Lugol fixed samples under an inverted microscope after settling for 24 h (Edler 1979, HELCOM 1988). In the late 1960s and early 1970s, Keefe's solution (Keefe 1926) was used as fixative by the City of Helsinki Environment Centre (Finni et al. 2001). Older samples (1975 to 1992) provided by the Institute of Aquatic Sciences, University of Latvia, were formalin fixed and counted using a settling method (for details see Jurgensone et al. 2011). Phytoplankton sampling involved pooling discrete surface samples from pre-defined depths or taking an integrated sample with a sampling hose. Species-specific cell volumes were used to calculate the total phytoplankton biovolume (Edler 1979).

Prior to data manipulation, the original data tables from providers were transformed into read-only text format source files. All subsequent data manipula-

Table 1. Data providers with acronyms used in the figures and the number of phytoplankton samples and data records from each provider

Data provider	Acronym	No. of samples	No. of records
Finnish Environment Institute	SYKE	2395	70 379
Finnish Institute of Marine Research	FIMR	461	19 417
City of Helsinki Environment Centre, Finland	HEL	4018	122 591
Institute of Aquatic Sciences, Latvia	LV	1534	35 831
Stockholm University, Sweden	SWE	1415	31 828
Institut für Ostseeforschung Warnemünde, Germany	IOW	1195	40 472
National Environmental Research Institute, Denmark	NERI	3296	113 153
Estonian Marine Institute, Estonia	EST	1047	27 567
Terttu Finni (private researcher), Finland	TERT	504	24 100
Total		15 865	485 338

tion was sequentially implemented and documented in a hierarchical system of Perl language scripts. The central data manipulation policy was to implement all modifications in a single linear scripted workflow, while keeping the source files strictly unchanged. The essential information (Table 2) from each provider's data was extracted into a rectangular table. Data entries were supplemented with a code identifying the data provider for later tracking. First, we identified and removed duplicate samples, which arose either from erroneous inclusions of more than one entry of a sample into the file by the providers, or from reporting the same sample both by national agencies and HELCOM.

Table 2. Variables in the final data set system of 4 inter-linked tables, which are joined by ID variables. S: sample table, SP: species table, C: count table, E: environment table

Variable	Format (unit)	Tables
Sample ID	integer	S, C, E
Sampling ID	integer	S, C, E
Latitude	decimal degree	S
Longitude	decimal degree	S
Station name	character	S
Sampling date	yyyy-mm-dd	S
Minimum sampling depth	integer (m)	S
Maximum sampling depth	integer (m)	S
Data provider	character	S
Species ID	integer	SP, C
Rank of the taxon	character	SP
Genus	character	SP
Species	character	SP
Subspecies	character	SP
Variety	character	SP
Order	character	SP
Class	character	SP
Division	character	SP
Corrected species name	character	SP, C
Original species name	character	C
Species descriptor	character	C
Cell biovolume	integer (µm^3)	C
Units counted	integer	C
Abundance	real (cells ml^{-1})	C
Wet weight	real (µg l^{-1})	C
Carbon biomass	real (µg C l^{-1})	C
Parameter name	character	E
Parameter value	real	E
Data source	character	E

Data set structure

The final data set was made up of 4 interlinked tables: sample, species, count and environment tables (Table 2). The sample table contained the essential information about individual samples (coordinates, date, depths, provider), with a primary key being the sample ID code. The number of records corresponded to the total number of samples in the data set. The species table incorporated information about the recorded taxa, with a species ID code as the primary key. The number of records was equal to the number of recorded taxa in the data set. The count table incorporated the actual abundances and biomasses of taxa in each sample and was linked to the sample table by sample ID code and to the species table by species ID code. The environment table contained all the physical and chemical parameters and was linked to the sample table by the sample ID code.

Sample table

Sample was the main operational unit in later analysis. The inconsistency in sample coding conventions rendered the unique sample codes by providers less useful, and we created new sample ID codes for operational purposes but retained provider codes for back reference. Samples were defined as a unique combination of coordinates, date, and sampling depth range. In addition we inserted sampling ID, a code that could encompass more than one sample in case the phytoplankton sample was taken from different depths (e.g. 0 to 10 m and 10 to 20 m). A vertical profile of phytoplankton samples shares the same sampling ID but has unique sample ID codes for every sampled layer.

Station-naming conventions also varied between providers and were retained only for back reference. We did not explicitly create station codes into the sample table. In the analysis phase, stations were identified as a unique geographic location with an arbitrarily defined precision to account for wind drift of the research ship and GPS precision.

Species table and taxonomic harmonization

The purpose of the species table was to provide a final list of taxa at the lowest identifiable level and the higher-rank taxonomic categories for summary purposes. Taxonomic harmonization was one of the most demanding tasks and started with a raw list of unique taxon names as provided. Correcting typing errors and common abbreviations narrowed the list. Non-nomenclatural comments, e.g. size class, were cut but retained as a separate variable. Narrative comments about the presence or absence of flagella, photo- or heterotrophy, and presence or absence of theca in dinoflagellates were used to narrow down the distinctiveness of unidentified or partly identified taxa. In the worst case, unidentified eukaryote was the closest acceptable identity. The original naming convention of the data providers was preserved in the count table for later reference. The rank of identification (subspecies, species, genus, order, higher taxon or NA for unidentified taxa) was added as an additional variable for summary purposes.

Tracing the plethora of taxonomic synonyms in the data was based on the Baltic Sea phytoplankton checklist by Hällfors (2004), and only on rare occasions the Catalogue of Life (www.catalogueoflife.org/) was used. Changes in nomenclature over time, caused by advances in taxonomic knowledge, have often led to recognition of earlier names as synonyms. The multiple historic synonyms were always lumped to the latest legal taxon name. The opposite process, splitting a single name historically denoting a species complex into multiple modern legal names, could not be resolved. For example, *Microcystis reinboldii* sensu Pankow (1976) has been split into 6 species of *Aphanocapsa* and 1 species of *Microcystis* (Hällfors 2004). As the only alternative, we suggest using the modern names when possible, but also keeping the old name with the historic records in the data set. However, attention must be exercised in the later analysis phase, where splitting may influence statistical inference, like long-term changes in diversity.

Count table

The count table was the largest, containing 1 row for each recorded taxon within each sample. The essential information in the count table was cell abundance (cells ml^{-1}), the biovolume of the cells or other counting units (µm^3), and wet weight biomass (µg l^{-1}). If only 2 were given, the third could be calculated. If only wet weight or abundance was given, the other could not be calculated without knowledge of the cell volume. In this case we used the cell volume of the same taxon by the same provider, or if not available, by other providers. The number of cells counted (or colonies, filaments, other counting units) is a vitally important variable that defines the statistical precision of a count but was frequently neglected by the analysts. Species and sample ID codes were added to each row by joining from the respective record in the species and sample tables. For back reference, the original unchanged taxon name was retained as a separate variable in the count table.

Environment table

The environment table contained all the environmental variables, which were recorded by the provider together with the phytoplankton data, augmented with variables extracted from publicly available data sets (e.g. Sokolov et al. 1997). Data from external sources were linked with the phytoplankton samples using common geographic coordinates, sampling time and depth intervals. To account for a modest wind drift of the research ship, we permitted a 0.5 km discrepancy in all directions for a positive match.

Statistical analysis

Frequency distributions were used to describe the long-term temporal, seasonal and spatial sampling efforts, and the commonness of taxa. For spatial frequency distribution, latitude and longitude coordinates were truncated to 0.1 and 0.05°, respectively. One of our concerns was the comparability of data sets between providers. To reveal differences in the taxonomic resolution between the data providers we analyzed (1) the number of recorded species per genus (graphically and with linear regression models) and (2) the proportion of sample biomass that was identified to at least species and genus levels. Identity of the most commonly observed taxa and also the long tail of rare taxa in the joint data set were revealed with taxon frequency distribution analysis.

Spatial autocorrelation was analyzed to describe scales of spatially repeating patterns in the phytoplankton data. First, log transformed wet weight biomass as a univariate descriptor was analyzed with a correlogram based on a Moran *I* statistic (Legendre & Legendre 1998). Second, multivariate spatial autocorrelation in the community composition domain was analyzed by means of a Mantel correlogram. A normalized Mantel statistic was computed between a Bray-Curtis community dissimilarity matrix among samples and a matrix where pairs of samples belonging to the same distance class received a value of 0 and the other pairs a value of 1 (Legendre & Legendre 1998). To reveal autocorrelation patterns, the Mantel statistic was calculated for a range of randomly generated distance classes. For the analysis we created a whole Baltic Sea test data set of 2234 samples from June and August to avoid excessive seasonal variability and from 1990 to 2008 to restrict the long-term temporal trend, with the additional exclusion of coastal samples (<3 km from coast). The spatial scale of autocorrelation depended on the extent of the data, i.e. the total area covered by the analysis. To demonstrate the scale dependency, both correlograms were done with the full extent of the Baltic Sea data, as well as with a subset geographically restricted to the Gulf of Finland (779 samples).

Temporal autocorrelation was analyzed with a Mantel correlogram as described above, but the distance classes in this case were time intervals. To avoid confounding by spatial variability we used a subset of samples from the central Gulf of Finland provided by the City of Helsinki Environment Centre, taken from a compact geographic area within ca. 30 km distance. By further restricting the data to years 1990 to 2008, we obtained a community matrix of 1439 samples and 213 taxa (excluding rare taxa present in less than 10 samples). The maximum temporal difference between the samples was set to 920 d, which covers 3 yr and thus shows the seasonal pattern in sample similarity.

Taxon richness was estimated by first making the assumption that the total phytoplankton taxon richness in a water body is a finite entity and that this entity can be approximated with the asymptote of the species accumulation models. Thus, only asymptotic models of species accumulation curves are considered in the following.

Species accumulation curves are plots of the cumulative number of species recorded with increasing levels of sampling effort. To assess the taxon richness based on a set of phytoplankton samples, we first used rarefaction to calculate a smooth sample based species accumulation curve. Rarefaction procedure re-samples randomly an increasing sub-set of samples multiple times and calculates the average number of taxa as a function of the number of samples. Then we fitted an asymptotic non-linear model to the species accumulation curve and used the asymptote as an estimate of total taxon richness.

Recent literature provides various non-linear models to analyze species accumulation curves (Tjørve 2003, Dengler 2009, Williams et al. 2009). Here we assessed the accuracy and performance of 9 commonly used models to obtain asymptotic estimators of taxon richness from phytoplankton monitoring samples. To select the best performing model, we first followed the common approach and assessed the goodness of fit of the models to the empirical data. We used Akaike's and Bayesian information criteria (AIC, BIC), which both measure the likelihood function, and introduced a penalty for the number of parameters in the model to avoid over-fitting (Table 3).

Table 3. Asymptotic non-linear functions examined in this study. In the formulae, a is the asymptote, and b, c, and d are the curve shape parameters of the models. Akaike (AIC) and Bayesian (BIC) information criteria describe the fit of the model to the smoothed species accumulation curve of 2200 samples (the terminal points in Fig. 10)

Model	Parameters	Formula	AIC	BIC
Negative exponential	2	$\mathrm{a}\,[1 - \exp(-\mathrm{b}x)]$	23 161	23 179
Monod	2	$\mathrm{a}x \,/\, (\mathrm{b} + x)$	20 442	20 459
Rational function	3	$(\mathrm{a} + \mathrm{b}x) \,/\, (1 + \mathrm{c}x)$	17 138	17 162
Logistic	3	$\mathrm{a} \,/\, \{1 + \exp[(\mathrm{b} - x) \,/\, \mathrm{c}]\}$	20 549	20 572
Gompertz	3	$\mathrm{a}\exp\,[-\exp(-\mathrm{bc}^{x})]$	20 071	20 093
Lomolino	3	$\mathrm{a} \,/\, [1 + \mathrm{b}^{\log(\mathrm{c}/x)}]$	8167	8189
Weibull3	3	$\mathrm{a}\,[1 - \exp(-\mathrm{b}x^{\mathrm{c}})]$	10 513	10 536
Weibull4	4	$\mathrm{a} - \mathrm{b}\exp[-\exp(\mathrm{c})\, x^{\mathrm{d}}]$	2978	3006
Asymptotic regression	2	$\mathrm{a}\,\{1 - \exp[-\exp(\mathrm{b})\, x]\}$	23 161	23 179

Estimating total taxon richness with an asymptotic model involves the inherent danger of extrapolation beyond the actual data. To provide extra confidence to the results beyond the overall fit of the model, we analyzed if and to what extent the asymptote of the models depend on the sampling effort (number of samples used in the analysis). An unsatisfactory property of many species accumulation models is that the asymptote tends to increase with the number of available samples. For a robust asymptotic estimator of taxon richness one would expect the asymptote to be insensitive to sampling effort. With the large number of samples available, we tested the stability of the model asymptotes and their sensitivity to sample size.

We used our previous test data set to construct a 2234 × 731 samples to species community composition matrix. From this matrix we drew N random samples, calculated the rarefaction curve, fitted the 9 non-linear models to the rarefaction curve and saved the parameter estimates for later reference. The procedure was repeated 60 times. N was increased from 50 to 2200 samples with an increment of 50. As a result we generated a relationship between the sample size and the model asymptote. The final selection of the best performing model was based on both goodness of fit to the data (AIC and BIC, the smaller the better) and the stability of the asymptote as the sample size was increased.

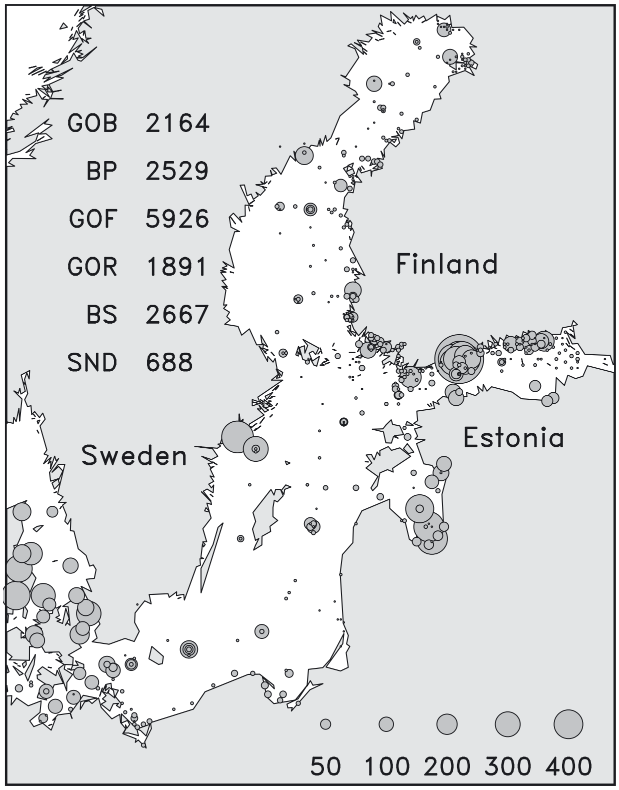

Fig. 1. Spatial distribution of phytoplankton samples in the data set. The size of the symbols (lower right scale) is proportional to the number of samples in each location. Top left list shows the total number of samples in the major basins: GOB: Gulf of Bothnia, BP: Baltic Proper, GOF: Gulf of Finland, GOR: Gulf of Riga, BS: Belt Sea and Kattegat, SND: the Sound

RESULTS

Spatio-temporal sample distribution

The spatial distribution of samples was strongly aggregated, but the overall basin coverage was reasonably good (Fig. 1). Much of the sampling effort was concentrated into coastal areas; 20 and 38% of the samples were taken from less than 3 and 5 km from the shore, respectively. The temporal range of the samples was from 1966 to 2008, and the sampling effort almost doubled in the mid 1980s (Fig. 2A). Summer months were more frequently sampled (Fig. 2B).

Data quality and comparability between providers

We extracted data records indentified to taxonomic level of species or lower, and plotted the number of species against the number of genera per sample (Fig. 3). Most of the provider specific linear regression slopes were relatively similar and varied from 1.11 to 1.35 species per genus, although, due to the large sample size, the slopes were statistically different. However, NERI data had a substantially higher slope of 1.74 (Fig. 3).

The proportion of the total wet weight biomass that was identified to at least species (Fig. 4A) or genus levels (Fig. 4B) indicated provider-specific differences. On average, the proportion of biomass indentified to species level was ca. 0.7, while the mean proportion was considerably higher (0.99) in the Latvian data set. The average proportion of at least genus level biomass was 0.92, while Estonian and Latvian data sets had consistently higher proportions.

Frequency distribution of taxa

The total taxon count of the whole data set was 1270. Most taxa were identified to species level or below (972) or genus level (251), the rest were identi-

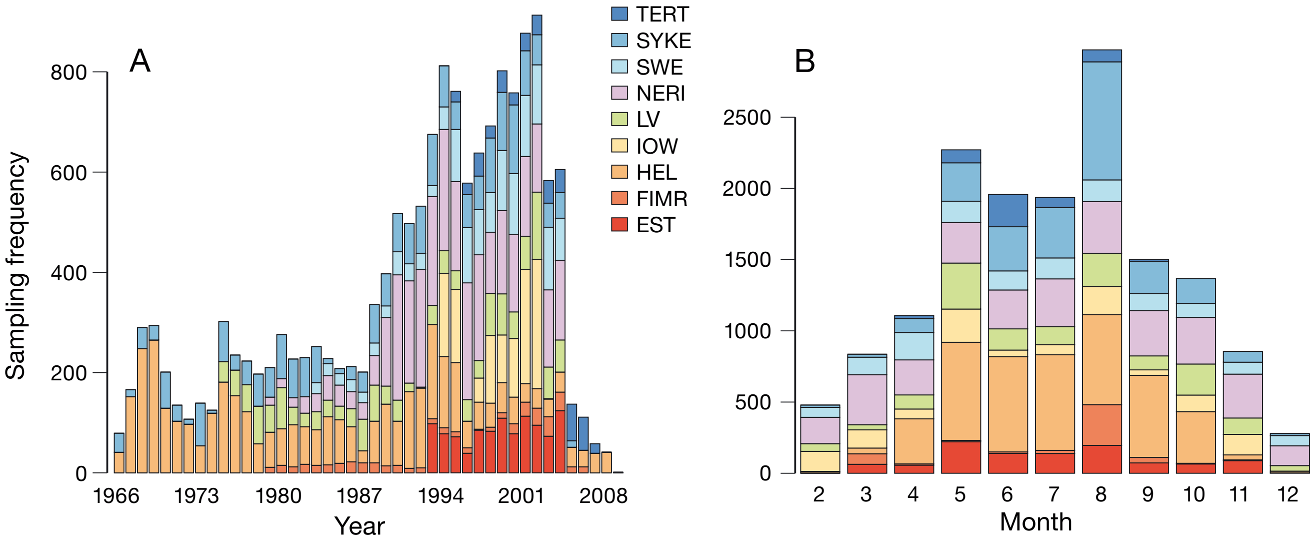

Fig. 2. Temporal distribution of phytoplankton samples indicating the long-term (A) annual and (B) seasonal sampling frequency. Frequency distribution is color-coded according to data provider (see Table 1)

fied to a higher taxonomic level. The most diverse algal classes were diatoms (344 taxa), followed by chlorophytes (308, including 24 prasinophytes and 63 desmids), dinoflagellates (193) and cyanobacteria (183). All other major groups had less than 100 taxa. The most diverse phytoplankton genera were *Chaetoceros* (43 spp.), *Protoperidinium* (31 spp.), *Anabanea* (26 spp.), *Gymnodinium* (25 spp.), *Desmodesmus* (22 spp.) and *Thalassiosira* (21 spp.).

The most common taxa, *Pyramimonas* spp., unidentified cryptophytes, and *Ebria tripartita*, were encountered in 64, 53 and 55% of the samples, respectively (Fig. 5). At the rare end, 174 taxa were encountered in only one sample and further 99, 62 and 49 taxa were encountered in 2, 3 and 4 samples respectively. Overall, 502 taxa, i.e. ca. 40% of the total taxon count, were recorded in less than 10 samples. Reichert et al. (2010) classified taxa in an assemblage as rare, intermediate or common based on occurrence in <1, 1 to 10 or >10% of the samples, respectively. Using this classification, rare taxa (i.e. occurring in <1% of the samples) constituted 74% of

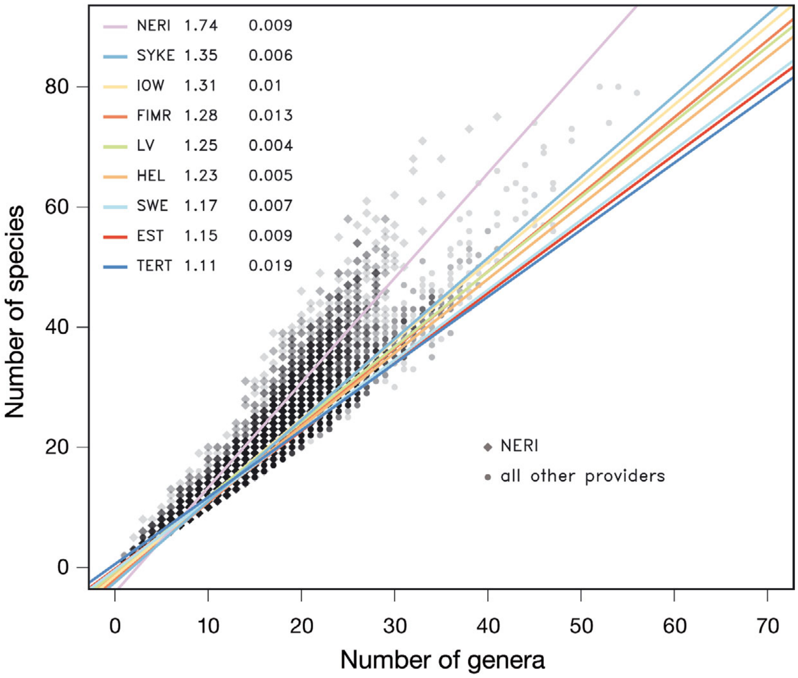

Fig. 3. Differences in the taxonomic resolution between data providers (see Table 1), expressed as the number of species per genera in a sample. Only records identified to species level or below are considered. The inscription shows the provider-specific slope estimates with the standard errors, listed in descending order. National Environmental Research Institute, Denmark (NERI) has a substantially higher slope than the other providers. The gray symbols use transparency to show dark color when overlapping

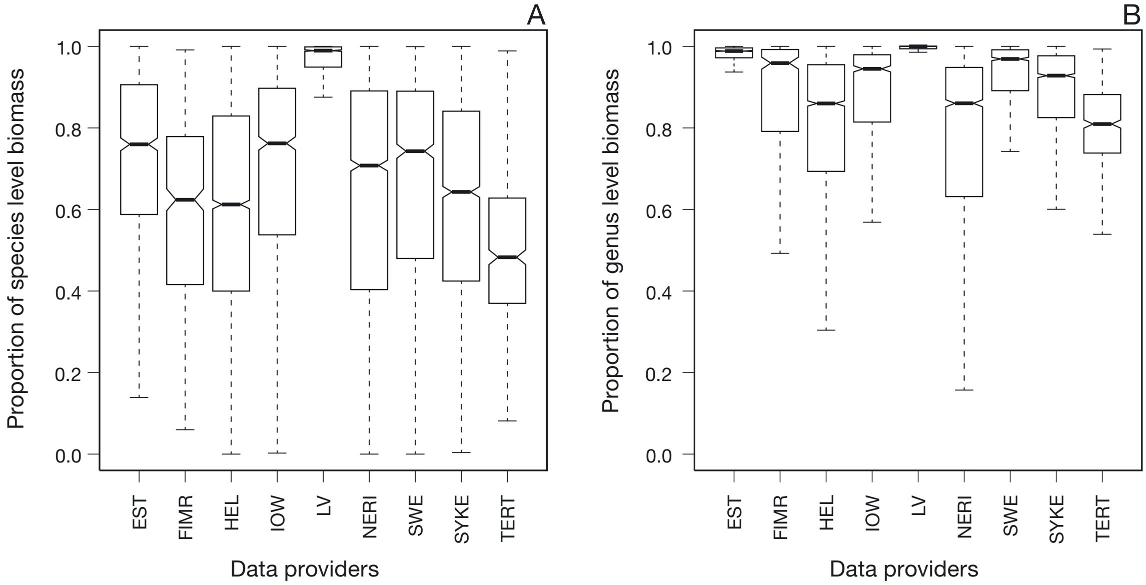

Fig. 4. Median and inter-quartile range of the proportion of (A) species-level and (B) genus-level biomass from the total sample biomass, split by the data providers (see Table 1). Whiskers extend to extreme data points, which are up to 1.5 times the inter-quartile range from the box

the total taxon pool, and taxa with intermediate and common occurrence made up 20 and 6% of the taxon pool, respectively.

Spatial and temporal autocorrelation

Any analysis of field data, or design of a sampling program, should consider autocorrelation. Here we utilized our extensive data set to analyze the scales and extent of spatial and temporal autocorrelation in phytoplankton field data.

The univariate Moran correlogram of the Baltic Sea phytoplankton biomass revealed positive autocorrelation up to 400 km distance and a negative autocorrelation at larger distances (Fig. 6A). The scale of positive autocorrelation was in the order of 50 km when the extent of the analysis was geographically restricted to the Gulf of Finland only (Fig. 6B). When the Moran correlogram was calculated from original sample data, a high degree of temporal variability resulted in a relatively damped range of spatial autocorrelation. When the temporal variability was averaged out by calculating a mean biomass for every unique location (defined by spatial coordinates), the strength of the spatial autocorrelation at short distances increased, but the overall pattern remained unchanged.

At the whole Baltic Sea extent the multivariate Mantel correlogram showed a positive spatial autocorrelation up to a scale of 100 km (Fig. 7A). At the geographically restricted Gulf of Finland extent, the positive autocorrelation of community composition was on a 50 km spatial scale (Fig. 7B).

There was a strong temporal autocorrelation between samples taken within a less than 30 d time window, but overall the similarity between samples decreased rapidly with time (Fig. 8). However, the positive temporal autocorrelation pattern was repeated once the samples were taken a year or two apart, but within the same seasonal time frame.

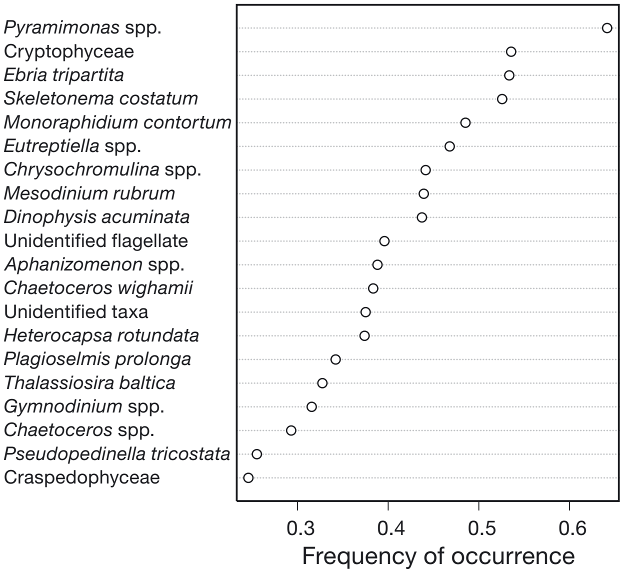

Fig. 5. The most frequent taxa in the phytoplankton monitoring samples. The *x*-axis shows the proportion of samples where the taxon was encountered

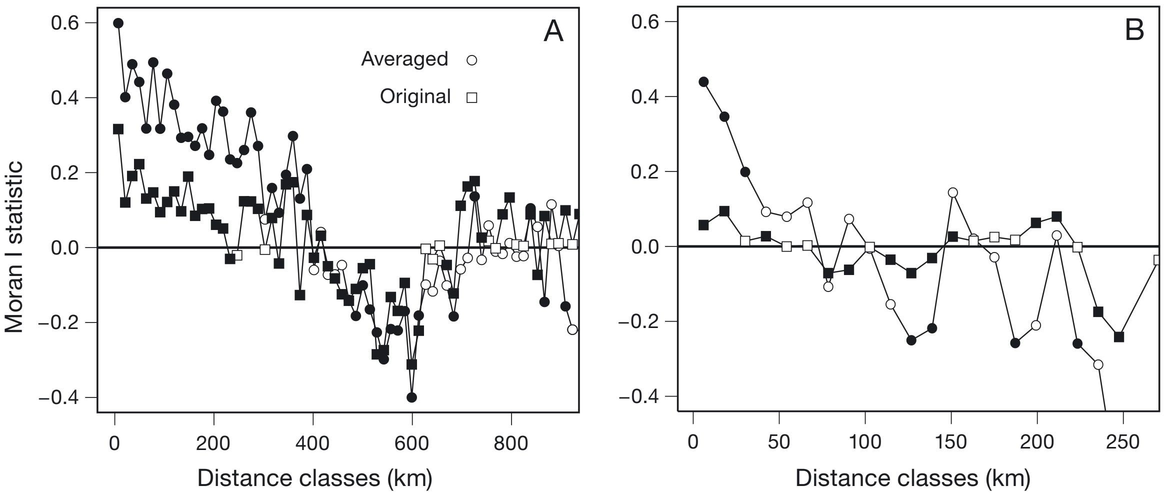

Fig. 6. Univariate spatial correlograms of (A) the Baltic Sea and (B) the Gulf of Finland phytoplankton biomass. The data are restricted to June and August samples from 1990 to 2008, excluding coastal samples less than 3 km from the shore. Filled symbols are significantly ($p < 0.05$) different from zero, based on a permutation test. The significance of the Moran *I* statistic depends also on the number of sample pairs per distance class, which is always higher with the original samples compared to the averaged ones. Therefore the original samples show higher significance at lower absolute value of the statistic compared to the averaged samples

How many phytoplankton taxa are in the Baltic Sea?

The species accumulation curve of the whole data set did not reach an asymptote. The high end of the curve revealed an appearance of a new taxon after every 93 samples (Fig. 9). Three of the models (Weibull4, Weibull3 and Lomolino) fitted the complete observational data set extremely well, while the others revealed clearly unsatisfactory performance (Fig. 9).

With most models the asymptote continued to increase as a function of sampling effort, when fitted to the smoothed species accumulation curves (Fig. 10). Further, 4 of the models converged at an asymptote below the actual taxon count, which indicates poor fit of the models (Fig. 10). Notable exceptions were the Lomolino and the Weibull4 models, which both had a relatively stable asymptote at a substantially higher level than the actual taxon count (Fig. 10). Both

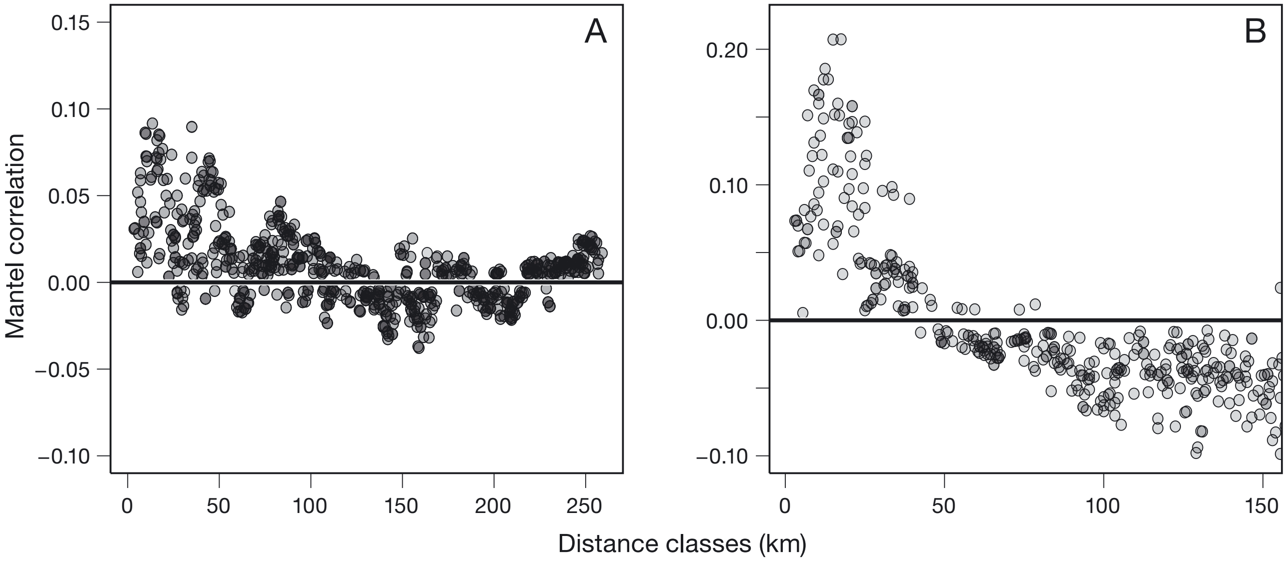

Fig. 7. Multivariate spatial Mantel correlograms of (A) the Baltic Sea and (B) the Gulf of Finland phytoplankton community composition. The data set is the same as in Fig. 6. The distance classes are set randomly to show the overall pattern of the autocorrelatation. The gray symbols use transparency to show dark color when overlapping

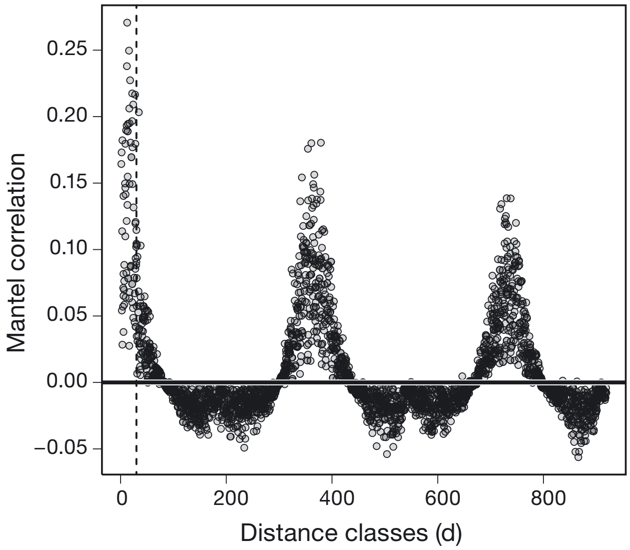

Fig. 8. Multivariate temporal Mantel correlogram of the phytoplankton community composition off the city of Helsinki (central Gulf of Finland). The data cover all the seasons from 1990 to 2008 (1439 samples). The vertical dashed line shows a distance class of 30 d, which approximately demarks the scale of a strong positive autocorrelation. The distance classes are set randomly. The gray symbols use transparency to show dark color when overlapping

showed a steeply increasing asymptote at low sampling effort, but thereafter the performance of the 2 models departed. The asymptote of the Lomolino model stabilized at a higher sample size (Fig. 10). Model evaluation revealed substantially better performance of the Weibull4 model, as it had lower AIC and BIC values compared to the Lomolino model (Table 3). On these grounds we considered Weibull4 as the best out of the 9 tested models to estimate the asymptotic taxon richness of the phytoplankton field data. Using the Weibull4 model, the asymptotic taxon richness of phytoplankton in the Baltic Sea, which could be revealed by routine microscopy, was estimated to be 1824 (Fig. 9).

DISCUSSION

Joining a large part of the publicly available data to a joint data set is a demanding, but revealing exercise. The Baltic Sea phytoplankton monitoring data show a considerable spatial and temporal aggregation of samples. Many statistical analysis methods, e.g. spatial interpolation, benefit from random sample distribution and special care must be exercised to overcome the observed uneven sampling effort. On the other hand, time series analysis benefit from a large number of samples from a geographically restricted region, like the high sample density in the Helsinki archipelago, provided by the City of Helsinki Environmental Centre. Overall, a compromise between the temporal and spatial coverage is never perfect, depending on questions asked and the intended analysis.

The various academic backgrounds and traditions in different countries add a degree of heterogeneity

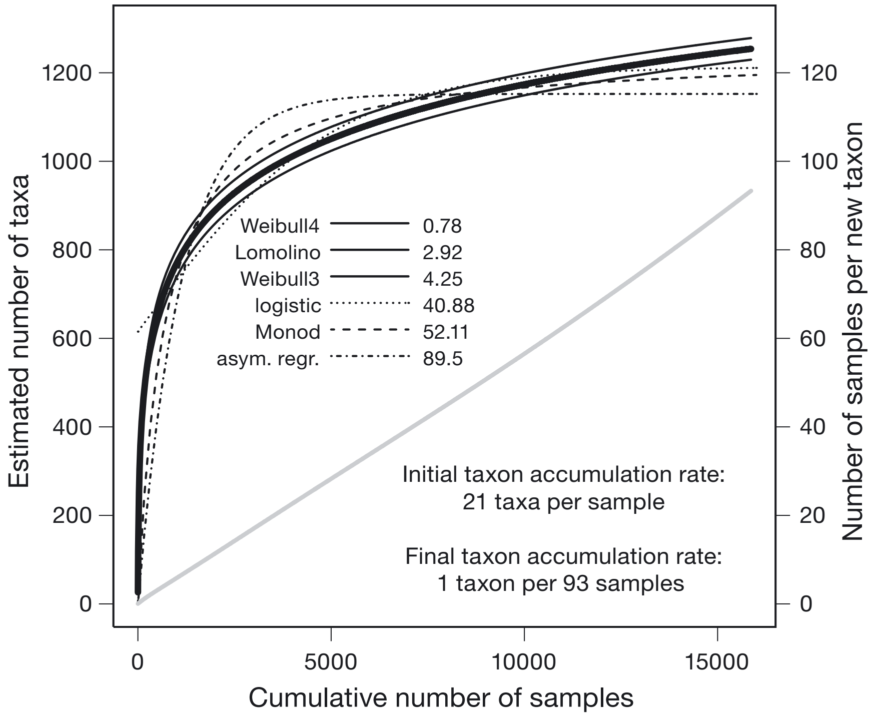

Fig. 9. Species accumulation curve of the Baltic Sea joint phytoplankton data set (15 865 samples, thick solid line) and standard deviation boundaries (thin solid lines). Weibull3, Weibull4 and Lomolino models fit the observational species accumulation curve extremely well and cannot be distinguished graphically (solid lines in the inscription). The dashed curves show the poor fit of the logistic, Monod and asymptotic regression models. The gray solid line shows the taxon discovery effort with a slight upward curvature (right *y*-axis). The Gompertz model (not shown) almost overlaps with the logistic model and the negative exponential (not shown) with the asymptotic regression model. The numbers in the inscription show the residual standard error of model fit

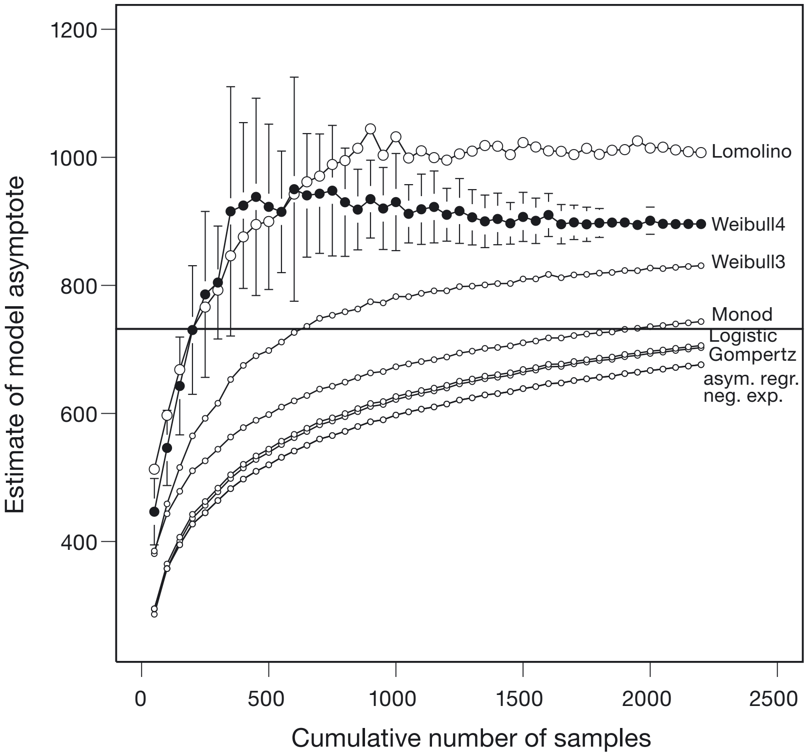

Fig. 10. Robustness of model asymptote estimates when sample size increases from 50 to 2200 random samples with an increment of 50 (*x*-axis). The observed taxon richness of the 2234 sample test data set was 731 taxa (horizontal line). The initial trend of increasing asymptote, as the sample size increases, is obvious for all models, but with the Lomolino and Weibull4 models we see a leveling off at ca. 800 and 400 samples, respectively. Other models show an undesirable monotonic increase of asymptote estimate as the sample size increases. The error bars show the standard deviation of Weibull4 model asymptote estimates based on 60 random sample sets

when compiling and harmonizing the data. It is encouraging that the taxonomic resolution, approximated as the average number of species per genus, appears reasonably similar between the data providers. NERI data had a higher mean number of species per genus, but we do not consider this a human factor, as the Danish coastal waters with high salinity have a very different biodiversity pattern compared to the rest of the Baltic Sea. On the other hand, we found notable data provider specific differences in the proportion of total phytoplankton biomass identified to species or genus levels. This indicates the tendency to ignore unknown and unidentified taxa and may justify careful corrections of the total phytoplankton biomass data when basin-wide comparisons are made. We could also identify an opposite subjective tendency — to readily record species which are very easy to identify under light microsopy. The most frequent taxa, *Pyramimonas* spp., cyptophytes indentified to class level, and the enigmatic heterotrophic flagellate *Ebria tripartita*, are all easy to recognize by morphology even to an inexperienced analyst. We believe that top frequency of morphologically easily recognized taxa is not by chance, but reflects a true bias in light microscopic phytoplankton monitoring. This implies a major qualitative difference when microscopy-based phytoplankton monitoring will be compared to or replaced by metagenomics-based approaches in the future (Cuvelier et al. 2010, Yan & Yu 2011).

Despite the potential biases, compiling a large multi-source data set enables questions to be asked and analyzed, which would not be possible with regional or smaller-scale data. In a recent analysis we demonstrated a conspicuous long-term change in the Baltic Sea phytoplankton community structure over the last 4 decades (Olli et al. 2011). We suggest that, partly, the overall trend reflects the response of the ecosystem to human-induced eutrophication and changes in species composition, but also shifts in the dominance of major functional groups (e.g. Wasmund & Uhlig 2003, Suikkanen et al. 2007, Jurgensone et al. 2011). On the other hand, dramatic changes in the proportion of diatoms and dinoflagellates in the Baltic Sea spring bloom are difficult to relate to direct human impacts, and rather reflect climatic fluctuations combined with pronounced expansion of dominant spring bloom dinoflagellates in parts of the Baltic Sea, e.g. in the Gulf of Finland (Klais et al. 2011). Whenever conventional statistical tests are used to detect phytoplankton trends, the independence of samples is not a safe assumption. Here we demonstrate the utility of a large compiled data set to detect the strength and scales of temporal and spatial autocorrelation in phytoplankton data. Thereafter we discuss how an extensive data set can be utilized to assess phytoplankton species richness, which is a fundamental concept in conservation biology.

Spatial and temporal autocorrelation

Dispersal limitation and local selective processes are the main mechanisms that determine patterns of biological diversity and community composition (Ricklefs 2004). Community-level autocorrelation depends on whether dispersal rates are high relative to local demographic rates. Large body size often scales with poor dispersal ability, leading to more clumped distribution and greater spatial autocorrelations in community composition (Shurin et al. 2009). In contrast, the high dispersal capacity of small aquatic organisms is frequently used to explain the distribution of small-bodied organisms: 'everything is everywhere, but the environment selects' (Green et al. 2008). This is in line with the modest spatial autocorrelation in our data, both at community composition and biomass levels. The positive autocorrelation of total phytoplankton biomass had a larger spatial scale compared to the autocorrelation of the community composition, confirming that the overall productivity determines the total biomass, which can be made up by a community with different composition at any given productivity and biomass level. Notably the highest correlation coefficients of the multivariate Mantel correlogram, both spatial and temporal, remain below 0.25, suggesting a relatively low correlation between the community composition of samples taken from close proximity in time or space.

The evidence from our analysis points to statistically significant, but ecologically modest, spatial and temporal autocorrelation patterns in the phytoplankton biomass and community composition. With large data sets even small departures from randomness will be detected, and statistical significance often is not a significant concept. In the following we therefore focus more on the ecological interpretation. Low autocorrelation suggests low predictability of the local community structure from a single phytoplankton sample. Low predictability is often related to high noise level in the data (Attayde & Bozelli 1998) and with scale effects (Beisner et al. 2006, Soininen et al. 2007). Temporal scales deserve special attention in this context because of the importance of short-term environmental variation on the structure of phytoplankton communities (Acuña et al. 2007). The difference between the autocorrelation of phytoplankton biomass between original samples and time-averaged values (Fig. 6) further highlights the short-term temporal variability in the phytoplankton field data, which largely masks autocorrelation and results in low autocorrelation coefficients.

The temporal autocorrelation of community composition indicated highest similarity between samples taken within a ca. 30 d time period. Interestingly, community similarity was repeated between samples taken at the same time of different years, which shows repeatability of the seasonal succession pattern. Also, the similarity between samples taken 1 yr apart was higher than between samples taken 2 yr apart, indicating a memory effect of the system.

The extent of autocorrelation is important to consider when designing sampling programs and in the later data analysis phase. The extent of the planned sampling scheme determines also the range of the spatial autocorrelation. For example, with a sampling scheme covering the whole Baltic Sea, samples from 1 sub-basin all show positive autocorrelation. When the sampling program covers just 1 sub-basin, autocorrelation scales are smaller. Thus autocorrelation depends on the context and scope of the study, and no universal scale exists. Regular statistical tests assume independence of samples and autocorrelation violates this assumption, inflating the significance levels of the tests. Usually appropriate autocorrelation structures can be used in the statistical models to correct for autocorrelation. However, simultaneous consideration of temporal and spatial autocorrelation is not well resolved in statistics, and requires some understanding of characteristic spatial and temporal scales (Gurarie & Ovaskainen 2011).

Taxon richness of the Baltic Sea phytoplankton community

Obtaining complete information about species composition in an area or ecosystem is a difficult task. A number of methods are available to estimate taxon richness from a limited collection of samples (Colwell & Coddington 1994, Gotelli & Colwell 2001). Most of these fall into 2 categories: (1) estimates based on the extrapolations from models fitted to randomized species accumulation curves and (2) analytical expressions of nonparametric estimators of total richness using either presence absence or abundance data (Borges et al. 2009).

Assessment of the performance of the methods generally depends on how the 'true' species richness, against which the estimations are compared, is obtained. Usually no inventory is so exhaustive that all the species present are recorded, although exceptions exist for taxonomic groups where species can be identified easily, such as birds or higher plants (Colwell & Coddington 1994). Therefore the common justification of the chosen model relies on the R^2 values close to 1 while fitting the curves, or some other goodness of fit parameter; the assumption being that, if the model approximates the observed accumulation curve, its extrapolation must be quite close to the true species richness. Ugland et al. (2003) argued that extrapolation of the randomized accumulation curves will in general not provide reliable information of the true species richness. They developed a modified method that accounts for heterogeneity in species diversity in sub-areas. However, the validation of their method depends on *a priori* knowledge of the true species list (Reichert et al. 2010).

Here we use a completely different approach to assess the performance of commonly used models, which, to the best of our knowledge, has not been used before. As the number of phytoplankton samples can be increased infinitely, we expect a good model to have relatively robust parameter estimates, which are independent of the sampling effort. Thus a good model is circumscribed not only with a perfect fit to the existing data but also stability of parameter estimates, particularly the asymptote, when new data are added. We found 3 models (Weibull4, Weibull3 and Lomolino) to have a very close fit to the data, but when also considering the stability of the asymptote as a function of sampling effort, Weibull4 clearly outperformed the others. Several models showed unacceptable fit and gave asymptote estimates, which were in fact lower than the number of taxa in the data table. Conspicuously, most models gave a strong positive relationship between the asymptote and sampling effort (Fig. 10), which is a serious problem in assessing total biodiversity.

Our approach to select the best performing model does not assume prior knowledge of 'true' species richness, which, in the case of phytoplankton filed data, cannot be assessed with conventional methods. Our data indicate that, even with almost 16 000 samples, the taxon list continues to increase, though with a decreasing rate. Thus, at the high end, increasingly more samples need to be analyzed to discover a new taxon. We interpreted it as a taxon discovery effort, which increases close to linearly in the range of our actual amount of data, with only a minor upward curvature (Fig. 9). The curvature becomes readily visible when the curve is extrapolated to several orders of higher sampling effort (not shown). Therefore we conclude that increasing sampling effort within any realistic ranges does not give a true species richness. We believe this is at least partly due to the very nature of phytoplankton community structure, which has a relatively high proportion of rare taxa, as was revealed by the long tail of the taxon frequency distribution.

Ugland et al. (2003) developed an analytical method which gives exact cumulative number of species and obviates the need for randomization and curve fitting. However, their analytical method is sensitive to the proportion of rare taxa (Reichert et al. 2010). Also, their estimate of total taxon richness in an area makes use of the (usually very small) proportion of the area covered by sampling, which is a concept not readily applicable to phytoplankton monitoring data. Similarly, nonparametric estimators, e.g. Chao and the Jackknives, perform poorly in terms of precision and bias under circumstances of numerous rare species (Chapman & Underwood 2009), and were not considered in this study.

We do not pretend our selected model, or even the method of selection, to be the most appropriate for other organism types, other habitats or even other scales. He & Legendre (2002) analyzed a species-rich assemblage and concluded that there is no model that is universally best, all depending on sampling scales. We suggest the best model depends in addition on the type of organisms and ecosystems studied, the sampling methods and the habitat heterogeneity, which determines the variability in biodiversity. Our test data set of 2234 samples included the habitat heterogeneity from the whole Baltic Sea salinity range but was seasonally restricted to late summer. The analysis revealed that about 400 random phytoplankton samples were needed before the extrapolated total taxon richness estimate leveled off. This is a conservative, high-end estimate, as most phytoplankton data sets have lower internal heterogeneity and likely require smaller sample size to reliably estimate taxon richness.

Acknowledgements. This study was supported by the Estonian Science Foundation (grant 7787), the Academy of Finland (grant 128987), the 6th FP project THRESHOLDS (GOCE-003900), and the European Regional Development Fund (Center of Excellence FIBIR). The constructive comments of 3 anonymous reviewers improved the manuscript.

LITERATURE CITED

➤ Acuña P, Vila I, Marín VH (2007) Short-term responses of phytoplankton to nutrient enrichment and planktivorous fish predation in a temperate South American mesotrophic reservoir. Hydrobiologia 600:131–138

➤ Attayde JL, Bozelli RL (1998) Assessing the indicator properties of zooplankton assemblages to disturbance gradients by canonical correspondence analysis. Can J Fish Aquat Sci 55:1789–1797

➤ Beaugrand G, Edwards M, Legendre L (2010) Marine biodiversity, ecosystem functioning, and carbon cycles. Proc Natl Acad Sci USA 107:10120–10124

➤ Beisner BE, Peres-Neto PR, Lindström ES, Barnett A, Longhi ML (2006) The role of environmental and spatial processes in structuring lake communities from bacteria to fish. Ecology 87:2985–2991

➤ Borges PAV, Hortal J, Gabriel R, Homem N (2009) Would species richness estimators change the observed species area relationship? Acta Oecol 35:149–156

➤ Boyce DG, Lewis MR, Worm B (2010) Global phytoplankton decline over the past century. Nature 466:591–596

➤ Chapman MG, Underwood AJ (2009) Evaluating accuracy and precision of species–area relationships for multiple estimators and different marine assemblages. Ecology 90:754–766

➤ Colwell RK, Coddington JA (1994) Estimating terrestrial biodiversity through extrapolation. Philos Trans R Soc Lond B 345:101–118

➤ Cuvelier ML, Allen AE, Monier A, McCrow JP and others (2010) Targeted metagenomics and ecology of globally important uncultured eukaryotic phytoplankton. Proc Natl Acad Sci USA 107:14679–14684

➤ Dengler J (2009) Which function describes the species–area relationship best? A review and empirical evaluation. J Biogeogr 36:728–744

Edler L (1979) Recommendations on methods for marine biological studies in the Baltic Sea. Phytoplankton and chlorophyll. Baltic Mar Biol Publ 5:1–38

Eichwald E (1847) Erster Nachtrag zu Infusorienkunde Russlands. Bull Soc Impér Nat Moscou 20:285–366

➤ Finni T, Laurila S, Laakkonen S (2001) The history of eutrophication in the sea area of Helsinki in the 20th century. Long-term analysis of plankton assemblages. Ambio 30:264–271

➤ Gotelli NJ, Colwell RK (2001) Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. Ecol Lett 4:379–391

➤ Green JL, Bohannan BJM, Whitaker RJ (2008) Microbial biogeography: from taxonomy to traits. Science 320: 1039–1043

➤ Gurarie E, Ovaskainen O (2011) Characteristic spatial and temporal scales unify models of animal movement. Am Nat 178:113–123

Hällfors G (2004) Checklist of Baltic Sea phytoplankton species, Baltic Sea Environmental Proceedings, Vol 95. HELCOM, Helsinki

➤ He F, Legendre P (2002) Species diversity patterns derived from species–area models. Ecology 83:1185–1198

HELCOM (Helsinki Commission) (1988) Guidelines for the Baltic monitoring programme for the third stage; Part D. Biological determinands. Baltic Sea Environmental Proceedings, Vol 27D. HELCOM, Helsinki

➤ Hooper DU, Adair EC, Cardinale BJ, Byrnes JEK and others (2012) A global synthesis reveals biodiversity loss as a major driver of ecosystem change. Nature 486:105–108

➤ Jurgensone I, Carstensen J, Ikauniece A, Kalveka B (2011) Long-term changes and controlling factors of phytoplankton community in the Gulf of Riga (Baltic Sea). Estuaries Coasts 34:1205–1219

➤ Keefe AM (1926) A preserving fluid for green plants. Science 64:331–332

➤ Klais R, Tamminen T, Kremp A, Spilling K, Olli K (2011) Decadal-scale changes of dinoflagellates and diatoms in the anomalous Baltic Sea spring bloom. PLoS ONE 6: e21567

Legendre P, Legendre LFJ (1998) Numerical ecology, Vol 20, 2nd edn. Elsevier Science, Amsterdam

➤ Moe SJ, Dudley B, Ptacnik R (2008) REBECCA databases: experiences from compilation and analyses of monitoring data from 5000 lakes in 20 European countries. Aquat Ecol 42:183–201

➤ Olli K (1996) Mass occurrences of cyanobacteria in Estonian waters. Phycologia 35:156–159

Olli K, Klais R, Tamminen T, Ptacnik R, Andersen T (2011) Long term changes in the Baltic Sea phytoplankton community. Boreal Env Res 16:3–14

Pankow H (1976) Algenflora der Ostsee. II. Plankton (einschl. benthischer Kieselalgen). Gustav Fischer Verlag, Jena

➤ Reich PB, Tilman D, Isbell F, Mueller K, Hobbie SE, Flynn DFB, Eisenhauer N (2012) Impacts of biodiversity loss escalate through time as redundancy fades. Science 336: 589–592

➤ Reichert K, Ugland KI, Bartsch I, Hortal J, Bremner J, Kraberg A (2010) Species richness estimation: estimator performance and the influence of rare species. Limnol Oceanogr Methods 8:294–303

➤ Ricklefs RE (2004) A comprehensive framework for global patterns in biodiversity. Ecol Lett 7:1–15

➤ Shurin JB, Cottenie K, Hillebrand H (2009) Spatial autocorrelation and dispersal limitation in freshwater organisms. Oecologia 159:151–159

Soininen J, McDonald R, Hillebrand H (2007) The distance decay of similarity in ecological communities. Ecography 30:3–12

Sokolov A, Andrejev O, Wulff F, Medina MR (1997) The data assimilation system for data analysis in the Baltic Sea, Vol 3. Stockholm University, Stockholm

➤ 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

➤ Tjørve E (2003) Shapes and functions of species–area curves: a review of possible models. J Biogeogr 30:827–835

➤ Ugland KI, Gray JS, Ellingsen KE (2003) The species–accumulation curve and estimation of species richness. J Anim Ecol 72:888–897

➤ Wasmund N, Uhlig S (2003) Phytoplankton trends in the Baltic Sea. ICES J Mar Sci 60:177–186

➤ Wasmund N, Göbel J, Bodungen BV (2008) 100-years-changes in the phytoplankton community of Kiel Bight (Baltic Sea). J Mar Syst 73:300–322

Wassmann P, Olli K, Wexels Riser C, Svensen C (2003) Ecosystem function, biodiversity and vertical flux regulation in the twilight zone. In: Wefer G, Lamy F, Mantoura F (eds) Marine science frontiers for Europe. Springer-Verlag, Berlin, p 279–287

➤ Williams MR, Lamont BB, Henstridge JD (2009) Species–area functions revisited. J Biogeogr 36:1994–2004

➤ Yan Q, Yu Y (2011) Metagenome-based analysis: a promising direction for plankton ecological studies. Sci China Life Sci 54:75–81

Editorial responsibility: Martin Solan, Southampton, UK

Submitted: December 1, 2011; Accepted: September 18, 2012
Proofs received from author(s): December 20, 2012