Application of functional data analysis to investigate seasonal progression with interannual variability in plankton abundance in the Bay of Fundy, Canada

Takayoshi Ikeda[a,*], Michael Dowd[b], and Jennifer L. Martin[c]

[a] Graduate School of Environmental Earth Science, Hokkaido University, North-10 West-5, Kita-ku, Sapporo 060-0810, Japan

[b] Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia B3H 3J5, Canada

[c] Fisheries and Oceans Canada, Biological Station, 531 Brandy Cove Road, St. Andrews, New Brunswick E5B 2L9, Canada

∗Corresponding author: *E-mail address*: tak@ees.hokudai.ac.jp

Abstract

The statistical technique of functional data analysis (FDA) is applied to a time series analysis of plankton monitoring data. The analysis is focused on revealing patterns in the seasonal cycle to assess interannual variability of several different taxonomic groups of plankton. Cell concentrations of diatom, dinoflagellate and zooplankton abundances from the Bay of Fundy, Canada provide the observations for analysis. FDA was performed on the log-transformed abundance data as a new approach for treating such types of sparse and noisy data. Differences in the seasonal progression were seen, with peak numbers, timings and abundance levels varying for the three groups as determined by curve registration and higher order derivatives using the objectively fit FDA curves. Nonmetric multidimensional scaling was used to capture seasonal variation among years. These results were further assessed in terms of dominant species and the relationships between groups for different seasons and years. It is anticipated that the easy to use, general and flexible technique of FDA could be applied to a wide variety of marine ecological data that are characterized by missing values and non-Gaussian distributions.

Keywords: statistical analysis; diatoms; algae blooms; temporal variations; interannual variability; abundance estimation

1. Introduction

Observations of plankton abundance and species composition from environmental monitoring programs provide foundational information on marine ecology. These data are derived from long-term collection and analysis of water samples (Smayda, 1978). Plankton patchiness generally results in a large amount of environmental noise in such observations (Weibe and Holland, 1968). To interpret these data, a central problem to be addressed is the extraction of the underlying abundance signal from these noisy data (Wyatt, 1995). Various statistical methods have been used to handle sparse and noisy abundance data, for example, time series analysis (Li and Smayda, 2001) Licandro et al. 2001), spline fitting (Wood and Horwood, 1995), objective analysis (Zhou, 1998), and also more complex methods focused on forecasting future blooms, such as artificial neural networks (Teles et al. 2006; Velo-Su*á*rez and Guti*é*rrez-Estrada, 2007) and population dynamics models involving population viability analysis (Holmes et al. 2007).

In this study, we apply the statistical method of functional data analysis (hereafter, FDA) to time series of plankton abundance. The nonparametric method demonstrates how noisy monitoring data can be fit to continuous and smooth functions capturing the main features of plankton variability, without the need for explicit distributional assumptions or filters. Another common problem that arises when handling monitoring data is sparseness caused by frequent occurrences of missing values. The FDA method satisfies conditions of continuity and smoothness without abrupt alterations in the fitted curves, and is not influenced by unequally spaced and missing observations. FDA is still relatively new method and has recently been applied in other fields of study with longitudinal type data, such as in neurological experiments (Long et al., 2005), 3D simulations

of human motion (Ormoneit et al., 2005), recognition of asymmetry in facial characteristics of infants (Bock and Bowman, 2006), or examining cash flow from ATM withdrawals (Laukaitis et al., 2005).

The motivation for the data analysis technique introduced here follows from Dowd et al. (2003; 2004), which was concerned with extracting the abundance signal from sparse and noisy monitoring data. This time series analysis method relied on a cyclic model at the annual period, and used the state space framework with the Kalman filter/smoother. Although statistically rigorous, it is fairly complex and likely difficult to apply for non-statistical practitioners. As formulated, it is also somewhat restrictive for use with general plankton abundance data since it only considers an adaptive sinusoidal cycle at a single frequency (corresponding to the annual period), so that spring and fall blooms are not supported by the analysis. This general state space method has been extended by Godsill et al. (2004) to the non-linear and non-Gaussian case, in which a Monte Carlo approach to smoothing was established, improving the overall ability to allow for more realistic distributions in the data. Other studies have examined the seasonal variation in taxonomic composition through cluster analysis based on nonmetric multidimensional scaling (Salmaso 1996) and also by a multivariate approach with principal response curves (Willis et al. 2004). There is also interest in describing trends of plankton biomass (Li et al. 2006). With FDA, one can handle data with both missing values and observational noise with no distributional assumptions necessary, providing more flexibility in applying the method. It has the ability to include as many terms as needed to relate to any behavior seen in the data without over complicating the model. In addition, FDA can handle data with high dimensions, and deforming or phase shifting of replicate curves can be done for alignment purposes. The main goal of this paper is to introduce a robust, flexible, objective and easily applied analysis method compared to those considered previously. It is applied to the problem of abundance estimation for interannual and seasonal variability in taxonomic groups, by efficiently treating noisy and sparse monitoring data that are often troublesome for practitioners to handle.

2. Materials and procedures

2.1. Monitoring data

The plankton monitoring data considered for this analysis were collected by personnel at the St. Andrews Biological Station of Fisheries and Oceans Canada. They have been conducting an environmental and phytoplankton monitoring program since 1987, focusing on the occurrence of several types of phytoplankton species and smaller zooplankton occurring in the Bay of Fundy in eastern Canada. Technical reports by Martin et al. (1999; 2001; 2006) outline plankton species abundances for each year and will be used as the basis for the analysis. One of the purposes of the monitoring program is to collect baseline information on phytoplankton populations that could be used for establishing temporal patterns and trends and to better understand the marine ecology of the region. Practical goals include understanding and predicting future occurrences of harmful algae blooms (HABs). We demonstrate how the data analysis method of FDA can be used to interpret these data to help achieve these goals. The data set consists of concentrations (cells/L) for three taxonomic groups; diatoms, dinoflagellates

and zooplankton, following the observation set used for analyses used in Dowd et al. (2004) and detailed in the reports of Martin et al. (1999; 2001; 2006). Samples were collected at a station near the Wolves Islands in the southwestern region of the Bay of Fundy, Canada (45° 59.57'N, 66° 44.36'W) at the surface and depths of 10, 25 and 50 m, from June 1988 to December 1999. Sampling took place at an irregularly spaced frequency: weekly during summer, bi-weekly during spring and fall, and monthly during winter, as a result of consistent low densities of plankton and the lesser likelihood of problems of HABs. The Bay of Fundy, having a large tidal range, is overall well mixed over most of its region. The sampling station varies in the degree of mixing, depending mainly on the season, as in the fall and winter, stronger winds cause more wind-induced vertical mixing (Meadows and Campbell, 1988). For this reason, plankton abundances were summed over depths. Furthermore, the log of the abundance was taken to stabilize overall variability and facilitate model fitting. Zero sums were incremented by one to avoid infinite values. The log values for each plankton type are plotted in Fig. 1 for the entire period.

(Figure 1 here.)

2.2. Analysis procedure

When collecting data, measurements are usually made discretely at different instances of time, locations or a combination of the two. In many scientific studies, these discrete measurements are better conceptualized as smooth and continuous curves. Treating the observations as such provides a reasonable estimation of any desired segment of the curve, even if the actual values were not collected. It also does not require observations to be taken at equally spaced time intervals, as is the case for monitoring data. This is the central motivation and idea that underlies the statistical method of FDA, which is described in detail below.

The fitting of continuous curves in FDA is done by first selecting a basis function making sure that the most desirable feature of the data is captured. A wide variety of basis functions can be applied in different cases based on previous knowledge of the data being handled. For instance, a B-spline basis is suitable for monotonous non-periodical behavior, and an exponential basis would be advantageous when dealing with organisms growing or decaying at an exponential rate. In many environmental analyses, such as the plankton data with a seasonal cycle, a Fourier basis is appropriate, since it captures oscillating movements and supplies continuity and infinite differentiation. Such a model takes the form,

$$x(t) = \sum_{k=0}^{n} A_k \sin k\omega t + B_k \cos k\omega t \tag{1}$$

where plankton abundance or concentration x is a function of time t, the fundamental frequency is ω, and A_k and B_k are constants. To estimate the order of the model, n, a balance must be struck between having a small number of components to avoid over-fitting of the data, against sufficient flexibility to capture the important features in the variability. In the case of plankton behavior in the North Atlantic, strong spring blooms and less distinct but significant fall blooms occur, making the

distinction between single and double peaks dominantly taking place in a given year for each group an important feature to be considered, and can be done with $n = 2$; a five-component sinusoidal function. The resulting fitted curve from the model shows a reasonable fit with approximately Gaussian-distributed residual values (Fig. 2).

(Figure 2 here.)

To implement the FDA procedure, the data are transformed from discrete to functional form, with concentrations being represented as continuous functions with the Fourier basis as in (1) with $n = 2$ and corresponding dates arranged consecutively ranging from day 1 to 365 (366 for leap years). FDA also determines the most appropriate coefficient values A_k and B_k by minimizing the least squares fit while controlling the "roughness" in the curves (explained later), as well as checking the efficiency in the chosen basis or improving the number of basis functions in a model. With the resulting curves, which efficiently extract the underlying signal in a noisy environment, one can then compare abundances within or between taxonomic groups. Three central elements of functional data analysis are outlined below.

2.2.1. Higher derivatives

Once a basis is chosen, a number of tasks can be done based on the model's derivative. One can use it as a method in verifying whether the chosen basis is appropriate for the data. This can be done by applying a linear operator to the fitted curve combining several higher derivatives. Say, for the dinoflagellate concentration, we would like to check whether the periodicity of the data can be captured using a fitted curve with a five-component Fourier basis and known frequency of $\omega_o = 2\pi/365$. Suppose we apply the following linear operator L to (1),

$$Lx(t) = \omega_o^2 Dx(t) + D^3 x(t) \tag{2}$$

where D^m is the m^{th} derivative with respect to t. This combination of higher derivatives was chosen as a measure of the fitted model to the data. By verifying the values of (2), one can see how much information could not be represented by the model. If values are large, or curves show another type of behavior other than sinusoidal, then this indicates an inappropriate choice of basis. In Fig. 3, the resulting curves from (2) are shown. Values on the vertical axis show that $Lx(t)$ is close to zero indicating only small oscillations that could not be captured by the model, but can be considered as negligible in terms of the general behavior.

(Figure 3 here.)

2.2.2. Roughness penalty approach

Data over-fitting can occur in many cases when too many basis functions are included in the model, as the fitted curves capture even the small characteristics of the data, thus making it impossible to see the underlying natural behavior. To avoid this from happening, FDA uses a technique called the roughness penalty approach, which preserves the features of the basis function while smoothing the curves by penalizing the weight on a certain higher derivative. The function $x(t)$ is then determined by minimizing the penalized residual sum of squares ($PENSSE$),

$$PENSSE = \int [y - x(t)]^2 dt + \lambda \int \left[D^2 x(t)\right]^2 dt \tag{3}$$

where y is the raw data and the penalizing parameter λ controls the degree of smoothness, which acts as a weight on higher derivatives in curve fitting. When λ is zero, there is no smoothing, and PENSSE does not change. As λ increases, the curve becomes less strict in fitting each point and relaxes the overall fit of the curve, hence creating a smoother fit. The value of λ is best calculated by the generalized maximum likelihood approach stated in Wahba (1985), or determined using the degrees of freedom (Hastie and Tibshirani, 1990). Furthermore, note that the second derivative can also be smoothed by penalizing the fourth derivative in the same manner. To illustrate the effect of λ, consider including 10 components in the basis function for the dinoflagellate concentration (years 1997–1999). This is obviously too many components since we can clearly see that almost all data points are connected by the fitted curve (left panels in Fig. 4). By penalizing the roughness, and taking the optimal λ, a better fit of yearly dinoflagellate abundance can be achieved with only one peak around the summer of each year and all other minor oscillations omitted (right panels in Fig. 4).

(Figure 4 here.)

2.2.3. Curve registration

Another useful procedure in FDA is that of the registration of curves. This involves adjusting the dependent or independent variable giving the effect of a shift or deforming of the curves in either a horizontal or vertical direction. It is to be applied when the physical times of occurrence are not of significance, or when the actual shifts and shape of the curves are of primary concern. For the plankton data, oceanographic conditions determine the onset of spring blooms and it is known that the timing of these events varies from year to year. By registering the curves, one can compare their general shape of the abundance curve without making reference to the absolute time of year. In general, a group of curves $x_i(t)$ can be shifted with a time lag component δ_i added to the time variable t,

$$x_i^*(t) = x_i(t + \delta_i) \tag{4}$$

where $x^*(t)$ is the registered curve. This type of registration is common when the curves are of the same shape and magnitude, say for instance, the dinoflagellate concentrations that have one peak during the summer. Another way of registration consists

of deforming the curve by another function $h(\cdot)$, such that

$$x_i^*(t) = x_i(h_i(t)). \tag{5}$$

The function h would match the timings of a certain feature of the curves, possibly being local maximum and minimum points and intercepts, as well as recognition of the levels of higher derivatives. In the next section, registration will be used to contrast the interannual cycles for the three plankton groups, by shifting each curve according to their local maximum, i.e., the moments at which derivative values are zero and the second derivative is negative.

For further details on these elements of FDA, the interested reader is referred to Ramsay and Silverman (2002; 2005) and Clarkson et al. (2005). Most notably, software packages for S-PLUS, R and Matlab can be used for this analysis and their manuals are available online at the website[1].

3. Results

Based on the above preliminary analysis, we implemented the FDA procedure with the model given in (1) with five Fourier basis components to represent the different behaviors of the three taxonomic groups. The roughness penalty λ was chosen based on previously introduced information that optimized the fit and smoothness of the curves. Fig. 5 shows the resulting objectively fitted curves for the three taxonomic groups from 1988 to 1999. The fitted curves for contrasting groups indicate how individual years differ from the mean curve as well as how they vary from each other.

(Figure 5 here.)

Fig. 6 shows the 11 yearly curves (since 1988 is partial) overlaying one another for each plankton group, with the bold line representing the mean curve. We can see that the fitted curves (left hand panels, Fig. 6) vary to a different degree for each taxonomic group, with diatom concentrations varying the most, and having different numbers of peaks occurring earlier and later than that of its mean curve. For dinoflagellate and zooplankton concentrations, only the timings of the peaks differed. For a better comparison of shape magnitudes, one can examine the peak-registered curves (right hand panels, Fig. 6. Note that for diatoms, only years with early and double peaks were registered, due to those with only fall events being of a different nature). The mean curve for diatoms maintains a relatively longer period of high values indicating that, on average, concentrations start to increase early in the year, and gradually decrease near the end. Some years show greater initial peaks, whereas others achieve another peak after the first. Registered dinoflagellate and zooplankton curves tend to be more similar among all years, with the exception of certain years having an earlier peak than the others.

[1] `http://www.psych.mcgill.ca/faculty/ramsay/ramsay.html`

Making use of the objectively fitted curves obtained by FDA (Fig. 5), higher derivatives and registration were used to calculate peak timings of the three groups for the twelve years (Fig. 7). Curve peaks were recognized by moments of maximum rate of change in which its derivative equals zero and the sign of the second derivative was verified to distinguish between maxima and minima. The timings of the peaks were obtained by shift registration of the local maximum. Interannual variability in peak abundance can be clearly seen in yearly diatom concentrations, whereas dinoflagellate and zooplankton abundance peaks showed lesser variation. Dowd et al. (2004) have shown that each year, the taxonomic groups deviated from their respective mean curves, although interannual variability could only be detected by controlling for the error variance for the abundance measurements.

(Figure 6 here.)

(Figure 7 here.)

Trends in the abundance are also of interest in terms of interannual variability (Li et al. 2006). The FDA curve for the entire sampling period was plotted with the maximum and minimum points of both spring and fall, corresponding to the seasonal peaks and troughs, respectively (Fig. 8). Sign-switching trends were seen in linear regression lines between spring and fall maxima for diatoms, but for both seasons, peaks and troughs were highly variable among years. The regression line for minimum spring abundance in dinoflagellates showed a slight increase over time, but no trends were seen for the other cases. For dinoflagellates and zooplankton, all fall peaks and troughs were higher than those of spring, indicating an occurrence of single blooms being somewhat stronger in the fall as expected. However, occasionally, some spring peaks for zooplankton were higher than those of fall (1990, 1994, 1999), possibly due to the variability in the peaks in diatoms. It is also worth noting that identification of these trends would be nearly impossible without the noise removal via FDA.

(Figure 8 here.)

Using the differences in peak number, timing and magnitude for each of the groups from FDA, comparisons are now made by grouping the years to see relationships between the plankton taxonomic groups. Based on Fig. 6, it is clear that for diatoms, the main difference separating them from the other groups is the number of distinct peaks, either being one (1990, 1992–1996, 1998, 1999) or two (1989, 1991, 1997). Also, within the single peak years, 1992, 1993 and 1996 had an early peak, whereas 1990, 1994, 1995, 1998 and 1999 had a later one. The overall concentration is relatively higher than those of the dinoflagellates and zooplankton in all years (Fig. 8). Dinoflagellate curves had single events occurring in the summer, with larger concentration seen in years 1989–1991, 1993 and 1995, reaching values of around 10 on the log scale or higher. Dinoflagellate peaks play the role of indicating whether diatom events, either occurring before or after the peak, are to be categorized as spring or fall events. Zooplankton peaks seemed to occur at, or after, those of the spring diatoms.

It is of interest to further examine the FDA taxonomic group abundance results in relation to the kinds of species that were present. Tables summarizing the technical reports by Martin et al. (1999; 2001; 2006) covering for years 1993–1999 will be used to determine which species had been dominant in specific seasons and any underlying species relationships, while verifying whether the FDA curves reproduced each case. For our purpose, diatoms at counts greater than 10^4 cells/L will represent dominance in a particular species. As for the other two groups, there were not as many instances in which counts reached this high since overall concentrations were lower. In defining a dominance scale for dinoflagellates and zooplankton, we will consider taking exponential values of the lowest of the maximum log values from their resulting fitted FDA curves leading to concentrations of 5×10^3 and 3×10^3 cells/L defining dinoflagellates and zooplankton as being dominant, respectively. From Fig. 7, we can then divide diatom behavior into the following three general groups from the resulting FDA curves: (1) strong spring peak, (2) strong fall peak, and (3) spring and fall events.

For 1993 and 1996, a spring peak was seen in diatoms. Table 1 shows that during spring, the *P. delicatissima* group as well as some other species were dominant, and also through to the summer (Table 2), but *Guinardia delicatula* was solely dominant in fall. All three dinoflagellate species *A. fundyense, Heterocapsa triquetra* and *Scrippsiella trochoidea* were dominant (Table 4). Although the zooplankton peak occurred after that of the dinoflagellates in 1993, the initial increase and the lengthy period of *Mesodinium rubrum* being dominant (not shown) may be the result of the high number and/or density of diatom and dinoflagellate species during the year, providing plenty of prey for the zooplankton. Conversely, in 1996, when the zooplankton peak occurred before that of the dinoflagellates, this may have been due to the earlier increase in diatoms compared to that of 1993, resulting in a lower chance for other diatom species to grow in spring (Table 1).

For 1994, 1995, 1998 and 1999, a fall peak was seen in diatoms. The *P. delicatissima* group was dominant throughout the year. Furthermore during fall, *Ditylum brightwellii* was also dominant (Table 3). The zooplankton peak happening before that of the dinoflagellates, except in 1995, may be due to the earlier initial increase of diatoms in 1994 than that in 1995, as well as the high densities of dominant dinoflagellate species *A. fundyense, H. triquetra* and *S. trochoidea*. As for 1998 and 1999, less dinoflagellate species were found dominant for only 7 to 14 days (Table 4), but instead, four or more additional diatom species were dominant in the spring compared to other years, with 1999 having the most number of dominant diatom species in summer. It is also interesting to note that the diatom and zooplankton curves are similar in shape for 1995, the year with the highest density of the *P. delicatissima* group (Table 3) and longest spanning of dominant dinoflagellate species (Table 4). The cause for this can be speculated based on the relationship among taxonomic groups and the greater dominance in dinoflagellates during 1995 that differentiates it from other years. 1999 can be due to the early increase in zooplankton which happens earlier in that year than in other years (Fig. 7), possibly due to the high number of dominant diatom species in spring and summer, and is comparable to what is usually seen during the fall season.

1997 was the only year with an evident double peak in diatoms, which are similar to both spring and fall event years,

with a noticeable difference in summer. In spring, both the *P. delicatissima* group and *Thalassiosira nordenskioeldii* were dominant, being the same as the spring event years (Table 1). During summer, only *G. delicatula*, the species dominant for every year, was dominant (Table 2). Lastly, in fall, the *P. delicatissima* group and *D. brightwellii* were noticeably dominant as in the fall event years (Table 3). *A. fundyense* was also not dominant in this year as in other years with short spanning dominant dinoflagellate species (Table 4). The zooplankton peak, which occurred before that of the dinoflagellates is possibly due to the early increase in diatoms as in other similar years.

4. Discussion

This study has introduced the method of FDA and applied it to analyze sparse and noisy plankton monitoring data to better interpret and understand fluctuations in plankton abundance in a noisy environment. Furthermore, results were examined in conjunction with information on species composition. FDA allows considerable flexibility in the types of problems to be tackled since the approach can readily be adjusted or tailored to the kind of data being considered, and the goals of the analysis procedure. It has been widely used in other fields of study, but to the authors' knowledge has not been applied to data from the marine sciences.

Another goal of this study was to examine seasonal progressions related to interannual variabilities within the three taxonomic groups, diatoms, dinoflagellates and zooplankton, from plankton monitoring data in the Bay of Fundy. FDA was applied to the time series data by using a Fourier basis of five components and reproduced abundance estimates for each group and showed distinguishable events (Fig. 5). With higher derivative calculation and registration, peak timings were estimated, distinguishing numbers of strong events, which differed for diatoms among years (Fig. 7). These were categorized into (1) strong spring event, (2) strong fall event and (3) both spring and fall events. Dinoflagellate and zooplankton curves slightly shifted in timing, each somewhat dependent on the other groups' behavior. Seasonal peaks and troughs showed variable trends for diatom, but rather flat trends for dinoflagellate and zooplankton (Fig. 8). An in-depth investigation of the species composition data in the technical reports (Martin et al. 1999; 2001; 2006) was carried out linking the abundance data to shifts in dominant species.

Many candidate statistical methods are available and were briefly mentioned above, each with advantages and disadvantages, according to the type of data being analyzed and the objectives of the study. The method in Dowd et al. (2003; 2004) was based on a rather complex model involving state space processes with the Kalman filter/smoother. Since only a single frequency were set for the adaptive sinusoidal cycle, spring and fall blooms could not be supported by the analysis. The extended non-linear and non-Gaussian method by Godsill et al. (2004) has not yet been applied to monitoring data, and is considerably more complex to apply than the Gaussian case. These methods have a slight disadvantage concerning their complexity, allowing FDA to be a more appealing and less restrictive method for non-statistical practitioners to apply. Studies on

seasonal variation in taxonomic composition have been carried out (Salmaso 1996; Willis et al. 2004), however both studies conducted sampling over a course of only one year, therefore yearly comparisons could not be carried out. Finally, long term trends could be recognized in the same manner as in Li et al. (2006), but their analysis relies on very basic methods, hence this study suggests the use of more modern statistical approaches for trend analysis.

Previous results with time series analysis methods involving monitoring data have been demonstrated for long-term sampling, in which investigations of chlorophyll (Li and Smayda 2001) and zooplankton (Licandro et al. 2001) were attempted. Li and Smayda (2001) experienced favorable conditions during the sampling period, resulting an equally spaced weekly sampling interval. On the other hand, in the study by Licandro et al. (2001), the 30-year data included missing gaps of up to 35 months, and so the eigen-vector filtering method was used to treat the missing values. However, it was explained in Mars et al. (1999) that the reliability of the eigen-vector filter is highly dependent on the ratio of undesired to desired signal amplitudes, indicating that the method could only be applied in which sampling takes place on a long timescale. However, the FDA method did not rely on equally spaced points or the availability of data in other years, being applicable to a shorter term of sampled data. Other methods have been considered in further investigating monitoring data focusing on spatial distribution, such as spline fitting (Wood and Horwood 1995) and objective analysis with Lagrangian-Eulerian interpolation (Zhou, 1998), but the methods were only applied to a few months in a single year. Recently, data assimilation with ecosystem models is being used to reproduce spatiotemporal distributions on a longer timescale (e.g. Zhaoa et al. 2005). Although it was not attempted, it may be worthwhile to test the FDA method on a spatial scale using b-spline basis functions. As for forecasting future blooms, recent studies are focusing on the use of methods such as population viability analysis (Holmes et al. 2007) and artificial neural networks (Teles et al. 2006; Velo-Su*á*rez and Guti*é*rrez-Estrada, 2007), where results from FDA could provide a starting point in carrying out such analyses.

In summary, FDA is a useful statistical approach that can readily be applied to a wide variety of marine ecological data characterized by being sparse, noisy and non-Gaussian, while allowing seasonal trends to be identified. It can be readily applied by non-statistical practitioners using existing statistical software. A number of choices must be made in terms of model selection and smoothness, but objective methods are available to do so (such as the derivative based approaches outlined). Future work should highlight aspects associated with statistical inference and error bar estimation (by bootstrapping), in order to better show the reliability of the curves. Extensions of FDA analysis to facilitate the interpretation of monitoring data in terms of other oceanographic variables relies on adaptations of FDA to regression and principal component analysis (see Ramsay and Silverman, 2005). Such multivariate FDA methods would also be useful for an examination of plankton community structure using monitoring data similar to the type presented here.

Acknowledgments

Many thanks to Dr. B. Smith of Dalhousie University, Department of Mathematics and Statistics, for introducing T. Ikeda to this work and for his support during and after his degree. Also, we are grateful to M. LeGresley for analyzing the phytoplankton samples. Much appreciation to Captain W. Miner and the crew of the Pandalus III for assisting in data collection. M. Dowd was supported by an NSERC Discovery Grant.

References

Bock, M.T., and Bowman, A.W., 2006. On the measurement and analysis of asymmetry with applications to facial modelling. Applied Statistics 55, 77–91.

Bray, J.R., and Curtis, J.T., 1957. An ordination of the upland forest communities of Southern Wisconsin. Ecological Monograph 27, 325–349.

Clarkson, D.B., Fraley, C., Gu, C.C., Ramsay, J.O., 2005. S+ Functional Data Analysis: User's Manual for Windows. Springer, New York, 192 pp.

Dowd, M., Martin, J.L., LeGresley, M.M., Hanke, A., Page, F.H., 2003. Interannual variability in a plankton time series. Environmetrics 14, 73–86.

Dowd, M., Martin, J.L., LeGresley, M.M., Hanke, A., Page, F.H., 2004. A statistical method for the robust detection of interannual changes in plankton abundance: analysis of monitoring data from the Bay of Fundy, Canada. Journal of Plankton Research 26, 509–523.

Godsill, S.J., Doucet, A., West, M., 2004. Monte Carlo Smoothing for Nonlinear Time Series. Journal of the American Statistical Association 99, 156–168.

Hastie, T.J., Tibshirani R.J., 1990. Generalized Additive Models. Chapman & Hall, London, 336 pp.

Holmes, E.E., Sabo, J.L., Viscido, S.V., Fagan, W.F., 2007. A statistical approach to quasi-extinction forecasting. Ecology Letters 10, 1–17.

Laukaitis, A., Račkauskas, A., 2005. Functional data analysis for clients segmentation tasks. European Journal of Operational Research 163, 210–216.

Li, W.K.W., Harrison, W.G., Head, E.J.H., 2006. Coherent Sighn Switching in Multiyear Trends of Microbial Plankton. Science 311, 1157–1160.

Li, Y., Smayda, T.J., 2001. A chlorophyll time series for Narragansett Bay: assessment of the potential effect of tidal phase on measurement. Estuaries 24, 328–336.

Licandro, P., Conversi, A., Ibanez, F., Jossi, J., 2001. Time series analysis of interrupted long term data set (1961–1991) of zooplankton abundance in the Gulf of Maine (northern Atlantic, USA). Oceanologica Acta 24, 453–466.

Long, C.J., Brown, E.N., Triantafyllou, C., Aharon, I., Wald, L.L., Solo, V., 2005. Nonstationary noise estimation in functional MRI. NeuroImage 28, 890–903.

Mars, J., Rector, I., James, W., Lazaratos, S.K., 1999. Filter formulation and wavefield separation of cross-well seismic data.

Geophysical Prospecting 47, 611–636.

Martin, J.L., LeGresley, M.M., Strain, P.M., Clement, P., 1999. Phytoplankton monitoring in the Southwest Bay of Fundy during 1993–96. Canadian Technical Report of Fisheries and Aquatic Sciences 2265.

Martin, J.L., LeGresley, M.M., Strain, P.M., 2001. Phytoplankton Monitoring in the Western Isles Region of the Bay of Fundy during 1997–98. Canadian Technical Report of Fisheries and Aquatic Sciences 2349.

Martin, J.L., LeGresley, M.M., Strain, P.M., 2006. Phytoplankton Monitoring in the Western Isles Region of the Bay of Fundy during 1999–2000. Canadian Technical Report of Fisheries and Aquatic Sciences 2629.

Meadows, P.S., Campbell, J.I., 1988. An Introduction to Marine Science, Blackie, Glasgow, 285 pp.

Ormoneit, D., Black, M.J., Hastie, T., Kjellström, H., 2005. Representing cyclic human motion using functional analysis. Image and Vision Computing 23, 1264–1276.

Ramsay, J.O., and Silverman, B.W., 2002. Applied Functional Data Analysis. Springer, New York, 190 pp.

Ramsay, J.O., and Silverman, B.W., 2005. Functional Data Analysis, Springer, New York, 430 pp.

Salmaso, N., 1996. Seasonal variation in the composition and rate of change of the phytoplankton community in a deep subalpine lake (Lake Garda, Northern Italy). An application of nonmetric multidimensional scaling and cluster analysis. Hydrobiologia 337, 49–68.

Smayda, T.J., 1978. Estimating cell numbers. In: Sournia, A. (Ed.) Phytoplankton Manual. UNESCO Publications, Paris, pp. 273–279.

Teles, L.O., Pereira, E., Saker, M., Vasconcelos, V., 2006. Time Series Forecasting of Cyanobacteria Blooms in the Crestuma Reservoir (Douro River, Portugal) Using Artificial Neural Networks. Environmental Management 38, 227–237.

Vallino, J.J., 2000. Improving marine ecosystem models: use of data assimilation and mesocosm experiments. Journal of Marine Research 58, 117–164.

Velo-Suárez, L. and Gutiérrez-Estrada, J.C., 2007. Artificial neural network approaches to one-step weekly prediction of Dinophysis acuminata blooms in Huelva (Western Andalucía, Spain). Harmful Algae 6, 361–371.

Wahba, G.A., 1985. Comparison of GCV and GML for choosing the smoothness parameter in the generalized spline smoothing problem. Annals of Statistics 13, 1378–1402.

Weibe, P.H., Holland, W.R., 1968. Plankton patchiness effects on repeated net tows. Limnology and Oceanography 13, 315–321.

Willis, K., Van Den Brink, P., Green, J., 2004. Seasonal Variation in Plankton Community Responses of Mesocosms Dosed with Pentachlorophenol. Ecotoxicology 13, 707–720.

Wood, S.N., Horwood, J.W., 1995. Spatial distribution functions and abundances inferred from sparse noisy plankton data: an application of constrained thin-plated splines. Journal of Plankton Research 17, 1189–1208.

Wyatt, T., 1995. Global spreading, time series, models and monitoring. In: Lassus, P., Arzul, G., Erard, E., Gentien, P., Marcallou, C. (Eds.), Harmful Marine Algal Blooms. Lavoisier, Paris, pp. 755–764.

Zhaoa, L., Weia, H., Xub, Y., Fenga, S., 2005. An adjoint data assimilation approach for estimating parameters in a three-

dimensional ecosystem model. Ecological Modelling 186, 235–250.

Zhou, M., 1998. An objective interpolation method for spatiotemporal distribution of marine plankton. Marine Ecology - Progress Series 174, 197–206.

Figure legends

Fig. 1. *Log*-transformed concentrations (cells/L) for (a) diatoms, (b) dinoflagellates and (c) zooplankton from June 1988 to December 1999.

Fig. 2. Diagnostic plots for fitted models with a five-component basis for diatoms, dinoflagellates and zooplankton (left, middle and right columns, respectively), consisting of scatter plot of fitted values vs. *log*-transformed data (top row), histogram of residuals (middle row) and QQ-plot (bottom row).

Fig. 3. Linear operator $L = \omega_o^2 D + D^3$ applied on dinoflagellate concentrations for all years. Small fluctuating values close to zero indicate that the basis function of five components is a suitable choice capturing the general behavior of the data.

Fig. 4. Dinoflagellate concentrations and fitted curves for years 1997 to 1999: a) with 10 basis functions and b) after smoothing with the roughness penalty approach.

Fig. 5. Fitted curves from FDA with Fourier basis of five components. Y-axis is the *log*-transformed concentration data for diatoms, dinoflagellates and zooplankton from 1988 to 1999.

Fig. 6. Unregistered (left panels) and shift registered curves (right panels) for diatoms, dinoflagellates and zooplankton.

Fig. 7. Predicted event timings for diatoms, dinoflagellates and zooplankton for the 12 years obtained by registration and derivatives of the fitted curves.

Fig. 8. Time series of diatom, dinoflagellate and zooplankton log depth-integrated concentration from FDA. Seasonal peaks (black) and troughs (white) for spring (square) and fall (triangle) are fit with linear regression lines (black: peak, gray: trough; solid: spring, dotted: fall).

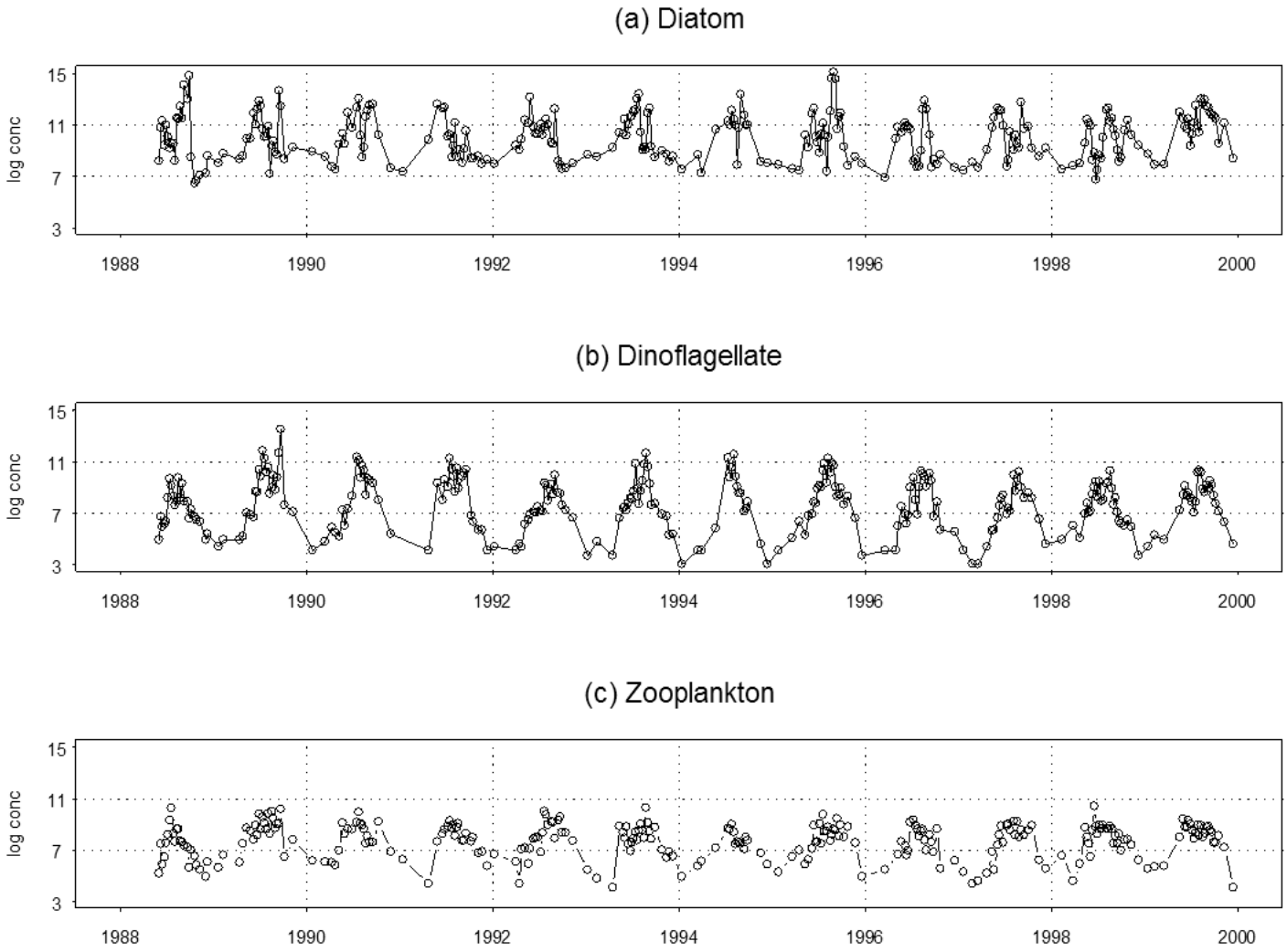

Fig. 1. *Log*-transformed concentrations (cells/L) for (a) diatoms, (b) dinoflagellates and (c) zooplankton from June 1988 to December 1999.

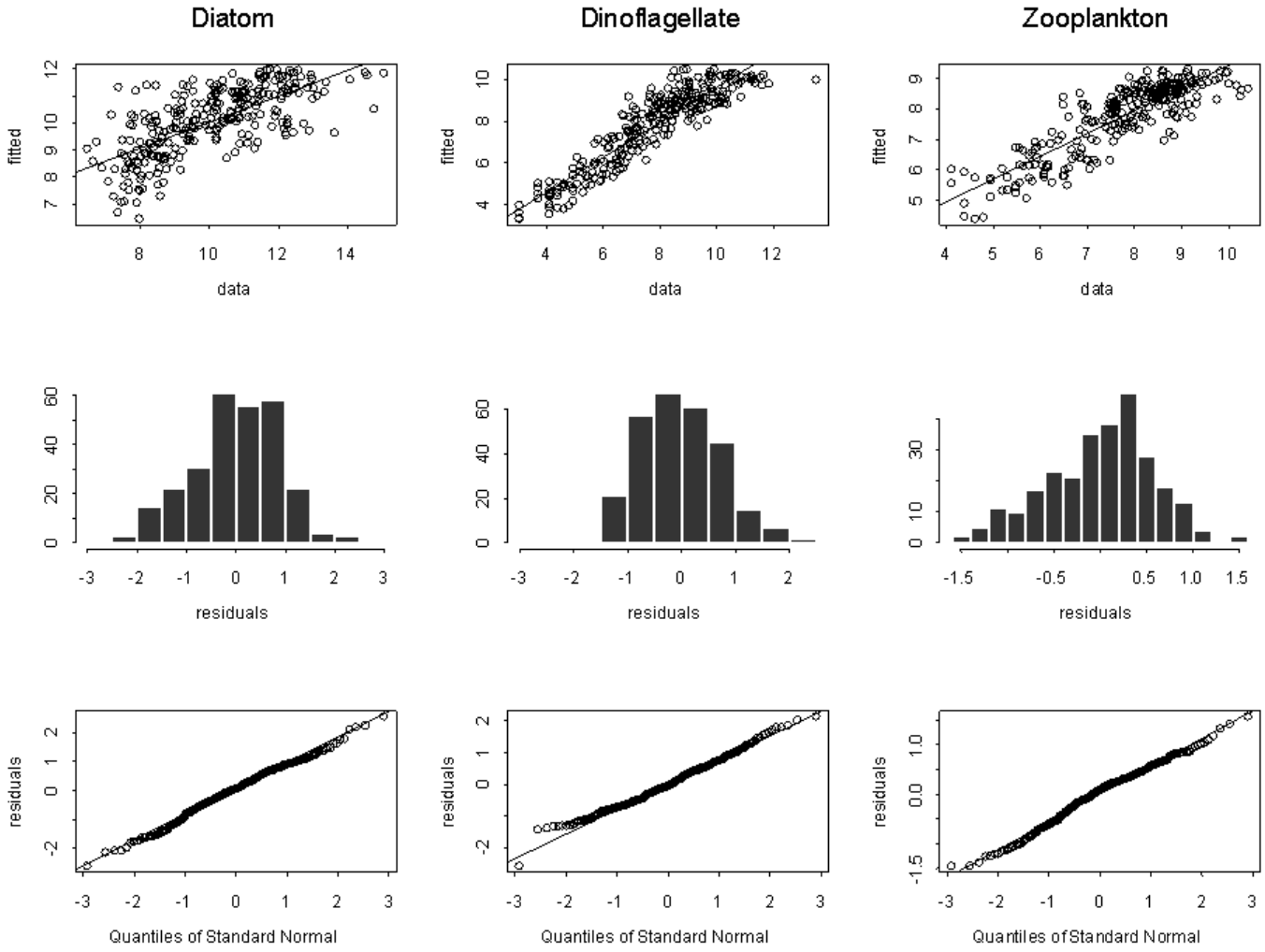

Fig. 2. Diagnostic plots for fitted models with a five-component basis for diatoms, dinoflagellates and zooplankton (left, middle and right columns, respectively), consisting of a scatter plot of fitted values vs. *log*-transformed data (top row), histogram of residuals (middle row) and QQ-plot (bottom row).

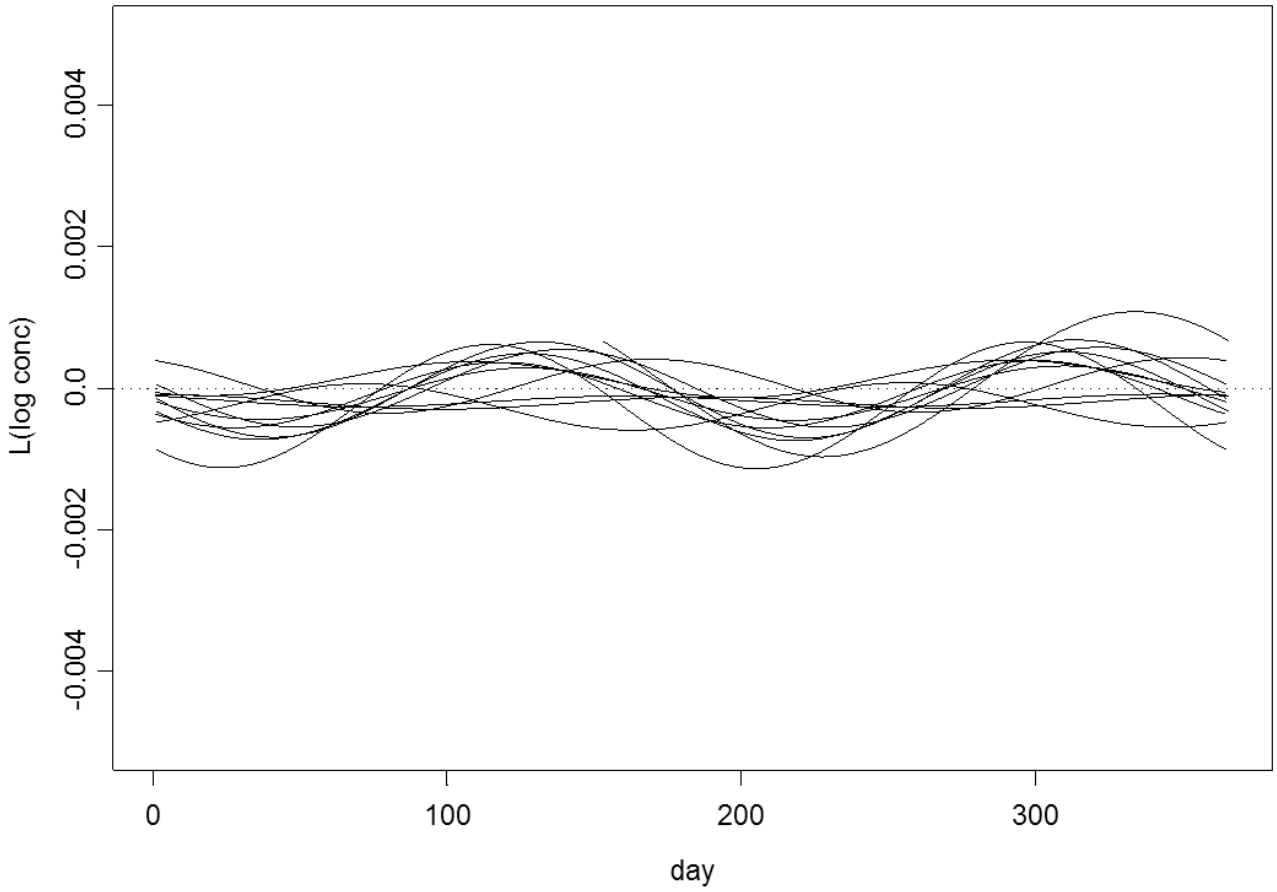

Fig. 3. Linear operator $L = \omega_o^2 D + D^3$ applied on dinoflagellate concentrations for all years. Small fluctuating values close to zero indicate that the basis function of five components is a suitable choice capturing the general behaviour of the data.

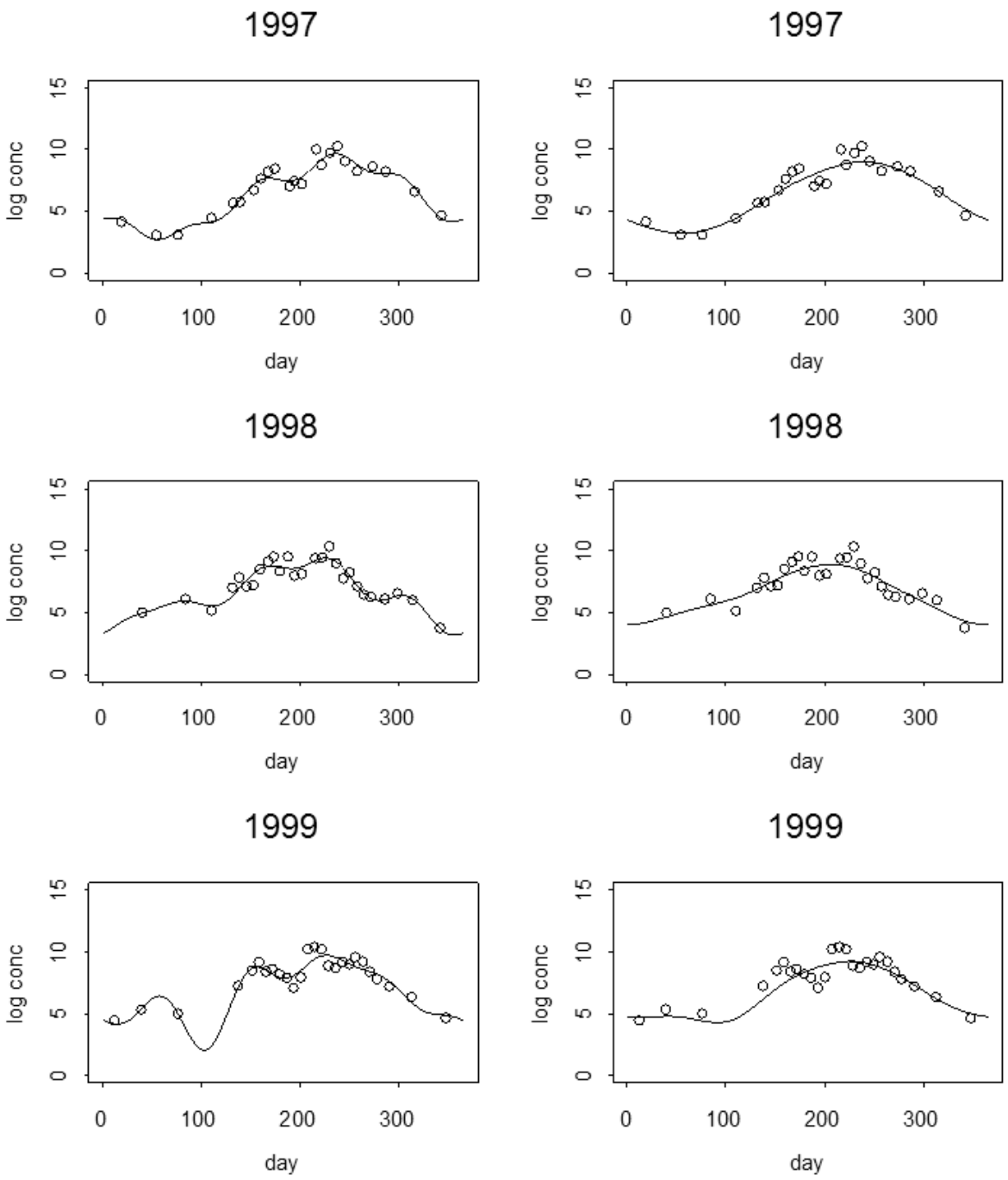

Fig. 4. Dinoflagellate concentrations and fitted curves for years 1997 to 1999: a) with 10 basis functions and b) after smoothing with the roughness penalty approach.

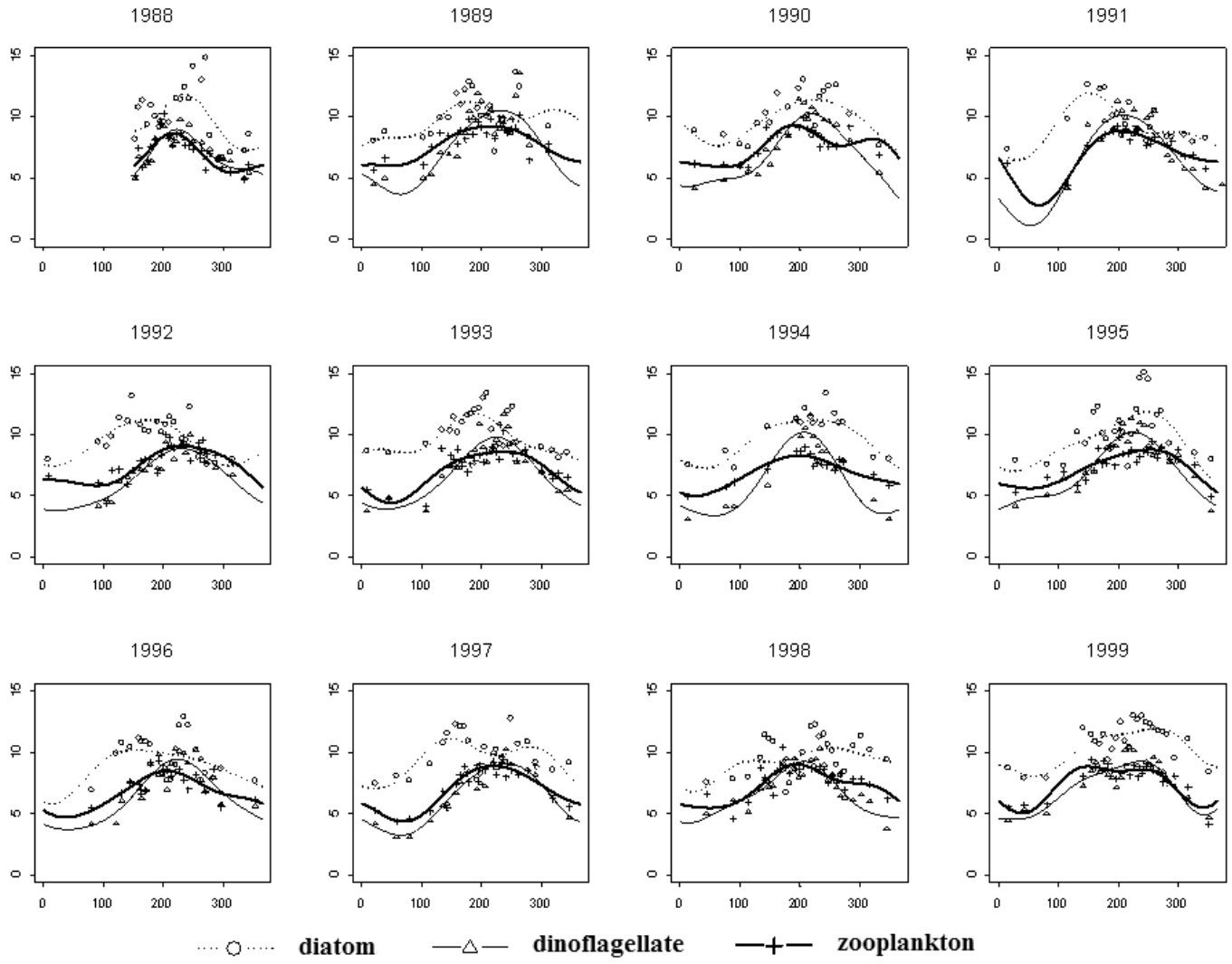

Fig. 5. Fitted curves from FDA with Fourier basis of five components. Y-axis is the *log*-transformed concentration data for diatoms, dinoflagellates and zooplankton from 1988 to 1999.

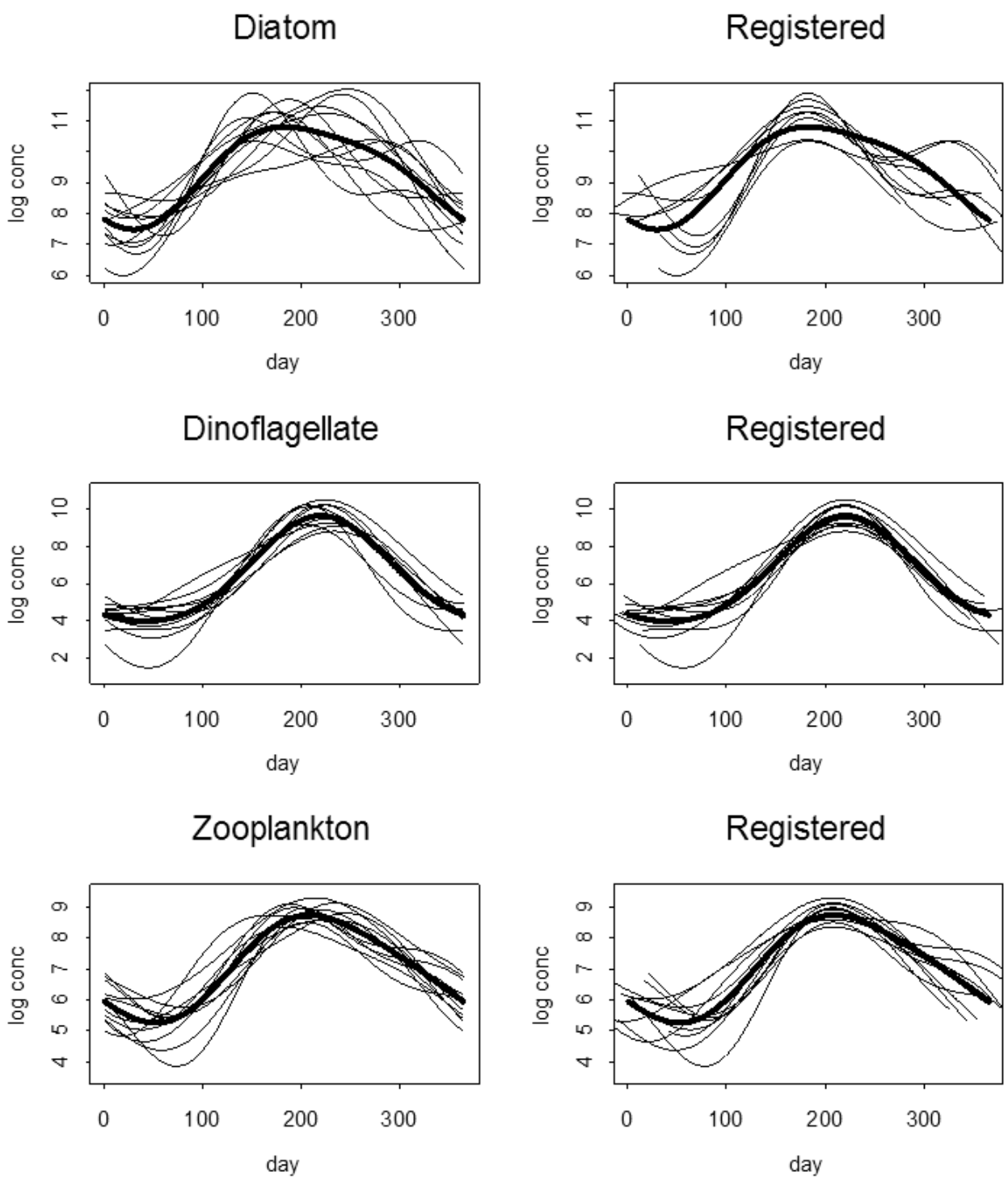

Fig. 6. Unregistered (left panels) and shift registered curves (right panels) for diatoms, dinoflagellates and zooplankton.

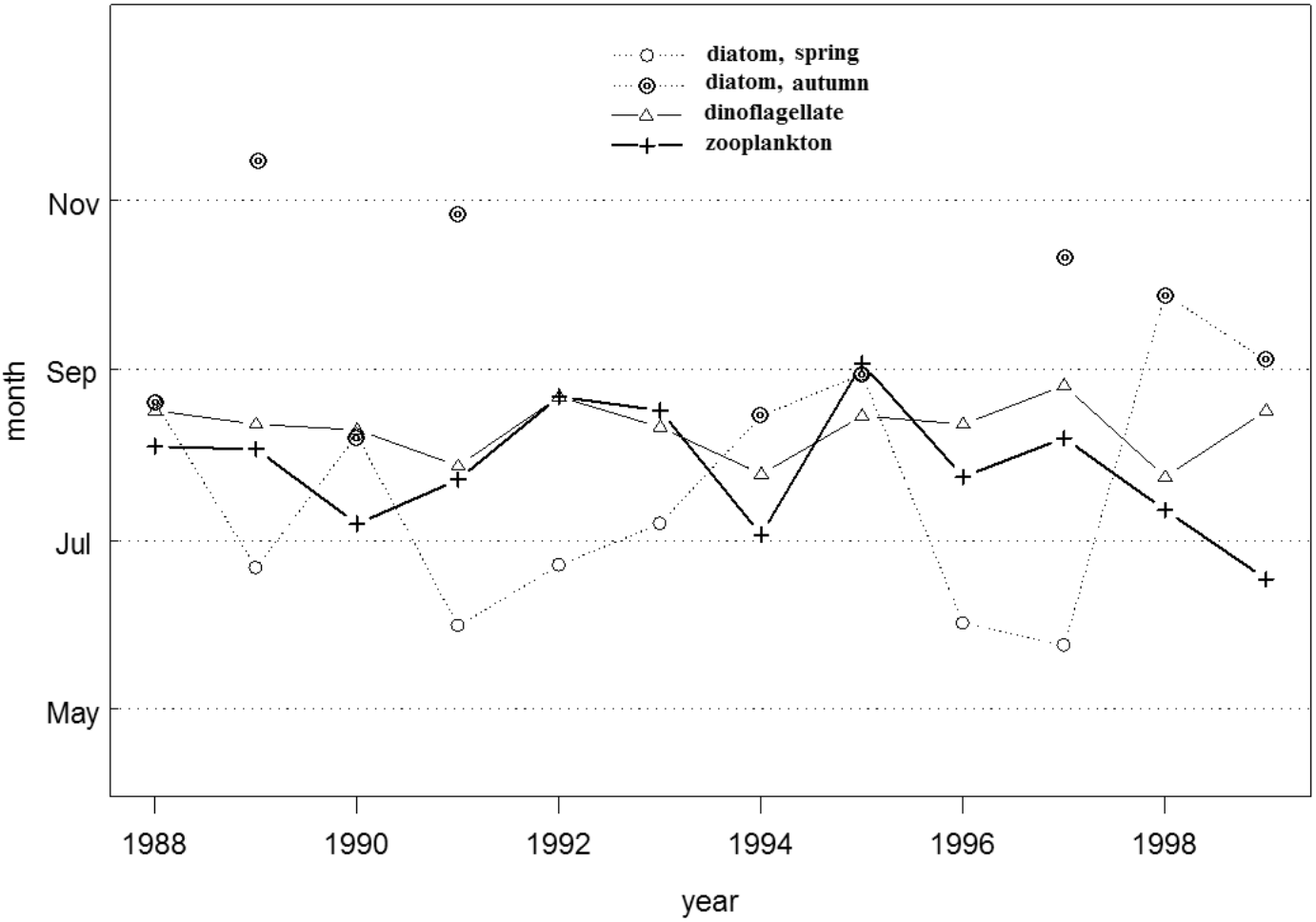

Fig. 7. Predicted event timings for diatoms, dinoflagellates and zooplankton for the 12 years obtained by registration and derivatives of the fitted curves.

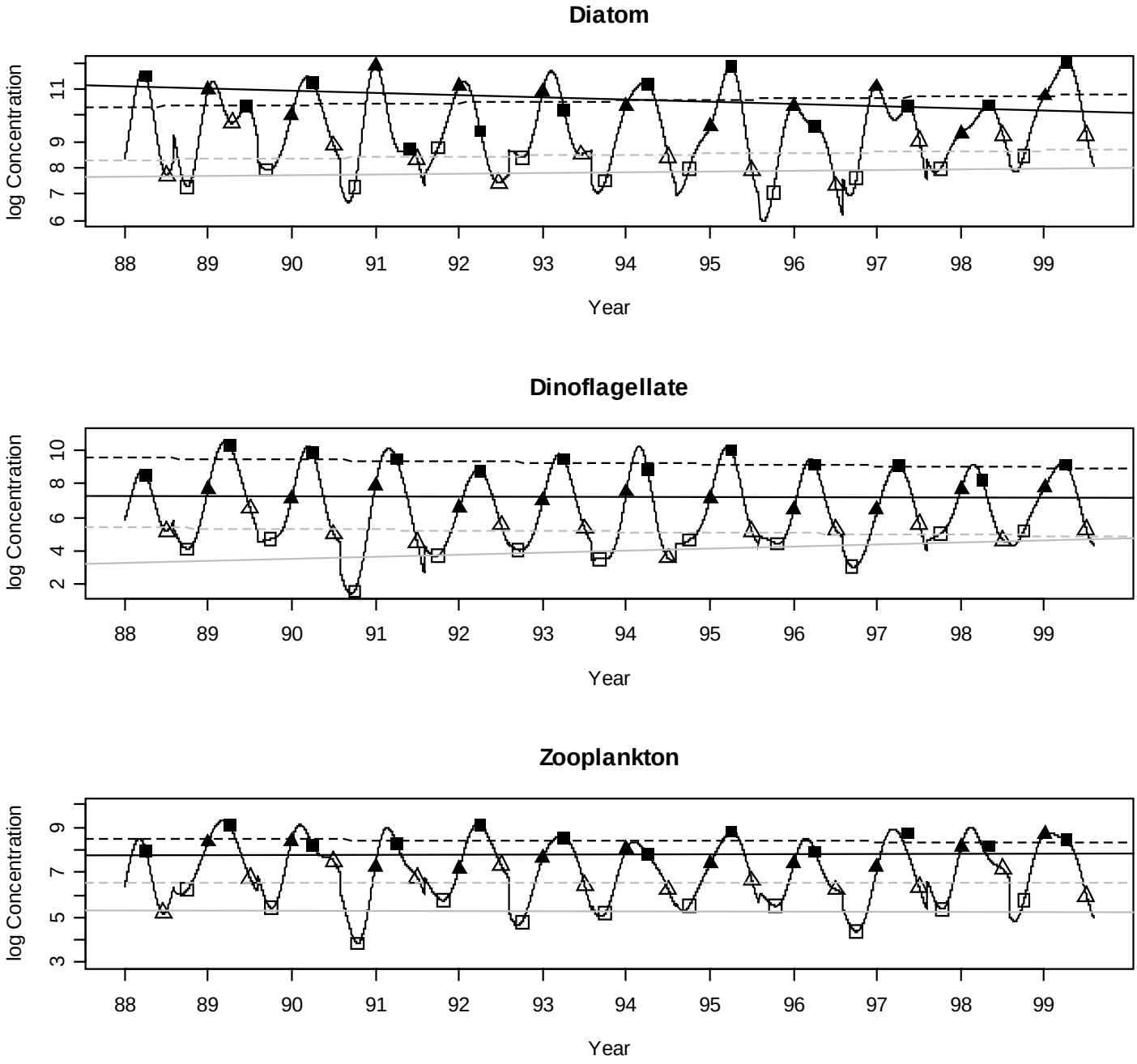

Fig. 8: Time series of diatom, dinoflagellate and zooplankton log depth-integrated concentration from FDA. Seasonal peaks (black) and troughs (white) for spring (square) and fall (triangle) are fit with linear regression lines (black: peak, grey: trough; solid: spring, dotted: fall).

Tables

Table 1: Dominant diatom species (greater than 10^4 cells/L) in spring (April to May) from 1993–1999. '$*$' denotes greater than 10^5 cells/L.

Species	1993	1994	1995	1996	1997	1998	1999
Chaetoceros compressus	x	-	-	-	-	x	-
Chaetoceros debilis	-	-	-	-	-	x	-
Chaetoceros socialis	-	-	-	-	-	-	x
Chaetoceros spp.	x	-	-	-	-	x	x
Guinardia delicatula -	-	-	-	-	-	-	
Leptocylindrus danicus	-	-	-	-	-	-	x
Pseudo-nitzschia delicatissima g.	x	-	$*$	x	$*$	x	-
Thalassiosira angulata	-	-	x	-	-	-	-
Thalassiosira decipiens	-	-	-	-	x	-	-
Thalassiosira nordenskioeldii	x	x	x	x	x	-	x
Thalassiosira spp.	-	-	-	x	-	x	$*$

Table 2: Same as Table 1 but in summer (June to August).

Species	1993	1994	1995	1996	1997	1998	1999
Chaetoceros compressus	x	-	-	-	-	-	-
Chaetoceros debilis	-	-	-	-	-	x	x
Chaetoceros socialis	-	-	-	-	-	x	$*$
Chaetoceros spp.	x	x	-	x	-	-	x
Cerataulina pelagica	-	-	-	-	-	-	-
Guinardia delicatula	x	x	x	x	x	x	$*$
Guinardia flaccida	-	-	-	-	-	x	-
Leptocylindrus danicus	-	-	-	-	-	-	$*$
Leptocylindrus minimus	x	x	-	-	-	-	-
Pseudo-nitzschia delicatissima g.	x	x	x	x	-	x	$*$
Pseudo-nitzschia seriata g.	-	-	x	-	-	-	-
Skeletonema costatum	x	x	x	x	-	x	x
Thalassiosira auguste-lineata	-	x	-	-	-	-	-
Thalassiosira oestrupii	-	-	-	-	-	-	x
Thalassiosira spp.	-	-	-	-	-	-	x

Table 3: Same as Table 1 but in fall (September to November). '$\star$' denotes being greater than 10^6 cells/L.

Species	1993	1994	1995	1996	1997	1998	1999
Asterionellopsis glacialis	-	-	-	-	$*$	-	-
Chaetoceros debilis	-	-	x	-	-	-	x
Chaetoceros socialis	-	-	-	-	-	x	x
Ditylum brightwellii	-	x	x	-	x	x	x
Eucampia zodiacus	-	-	-	-	-	-	x
Guinardia delicatula	$*$	-	-	x	-	-	-
Guinardia striata	-	x	-	-	-	-	-
Leptocylindrus minimus	-	-	-	-	-	x	-
Pseudo-nitzschia delicatissima g.	-	$*$	$\star$	-	x	x	-
Pseudo-nitzschia seriata g.	-	-	x	-	-	-	-
Skeletonema costatum	-	$*$	-	-	$*$	-	-
Thalassiosira gravida	-	x	x	-	-	-	x
Thalassiosira spp.	-	x	-	-	-	-	-

Table 4: Dominant dinoflagellate species (greater than 5 x 10^3 cells/L) in summer (June to August). '$*$' and '$\star$' denote being greater than 3 x 10^4 and 6 x 10^4 cells/L, respectively. '$\sharp$ days' is the total number of days from starting to ending dates for high counts of dominant species.

Species	1993	1994	1995	1996	1997	1998	1999
Armoured dinoflagellate	-	-	-	-	x	-	x
Alexandrium fundyense	$*$	$\star$	x	x	-	-	-
Ceratium lineatum	-	-	x	x	-	x	-
Gonyaulax spinifera	-	x	-	-	-	-	-
Heterocapsa triquetra	$*$	x	$*$	x	x	-	x
Scrippsiella trochoidea	$\star$	$\star$	x	x	-	x	x
$\sharp$ days	48	29	87	74	21	7	14