Available online at www.sciencedirect.com
ScienceDirect
Journal of Marine Systems 69 (2008) 328–338
JOURNAL OF MARINE SYSTEMS
www.elsevier.com/locate/jmarsys

# Towards a seascape topology II: Zipf analysis of one-dimensional patterns

James G. Mitchell [a,*], Laurent Seuront [a,b]

[a] School of Biological Sciences, The Flinders University of South Australia, GPO Box 2100, Adelaide 5001 South Australia
[b] Ecosystem Complexity Research Group, Station Marine de Wimereux, CNRS UMR 8013 ELICO,
Université des Sciences et Technologies de Lille, 28 avenue Foch, F-62930 Wimereux, France

Received 11 June 2005; received in revised form 22 December 2005; accepted 11 March 2006
Available online 22 February 2007

## Abstract

Large temporal and spatial data series are increasingly available and easy to produce. This paper uses Zipf analysis to evaluate serial data sets from the HOTS, BATS, EquaPac and high-resolution vertical profiles of FluoroMAP. Zipf analysis produced Zipf exponents from best-fit lines that permitted comparison among data sets. It allows comparison of one-dimensional series despite differences in scale and missing data. Zipf exponents ranged from 0.043 to 0.83. Serial data with sampling intervals of milliseconds and months showed exponents that ranged around 0.3. To the extent that Zipf exponents measure structure and variation, the indication is that structure of distributions is similar over millimeters and hundreds of kilometers. Zipf analysis provides a means to quantify similarities and differences, and suggests that variation is linked across many length scales for phytoplankton.

*Keywords:* Chlorophyll *a*; Fluorescence; Zipf; Microscale; Phytoplankton patchiness

## 1. Introduction

The production of large amounts of data in marine ecology dates back to the Challenger expedition volumes. However, only relatively recently have those data accumulated rapidly enough to be common and pose some challenge to analysis and meaningful interpretation. For phytoplankton this began with fluorometry data during the early 1970s (e.g. Platt, 1972) and later satellite data, beginning with the coastal zone color scanner (Feldman et al., 1984). The buildup of extensive fluorometry and chlorophyll data sets continues today with projects such as the Hawaiian Ocean Time Series (HOTS) and Bermuda Atlantic Time Series (BATS) not only accumulating large data sets, but also making them available on the internet. In these cases there are a large number of scientists and technicians involved in the data stream. Their contributions range from hardware and software design, e.g. smoothing filters, through the physical process of sampling to analysis and writing. This prevents an individual from drowning in data preparation, reduction, analysis and interpretation by sharing the data analysis.

The decreasing price of electronic data gathering devices and the increasing power of personal computers are increasingly enabling individuals to rapidly gather data sets containing tens of thousands to tens of millions of pieces of information. In addition, to the pervasive availability of satellite data, fluorometers and other rapid sampling or measuring devices now permit rapid data

* Corresponding author. Tel.: +61 882013684; fax: +33 882013015.
*E-mail address:* Jim.Mitchell@flinders.edu.au (J.G. Mitchell).

0924-7963/$ - see front matter © 2007 Published by Elsevier B.V.
doi:10.1016/j.jmarsys.2006.03.026

accumulation. Fluorometry series data have been invaluable in assessing the response to iron limitation (Behrenfield et al., 1996), finding structure associated with fronts and ocean biomass structure (Platt, 1972; Strutton et al., 1996; Strutton et al., 1997a,b; Chavez et al., 1999), and getting various glimpses at microscale phytoplankton distributions (Mitchell and Fuhrman, 1989; Cowles et al., 1993; Cowles and Desiderio, 1998; Franks and Jaffe, 2001). Demand for analysis of spatial series is also growing in nutrient analysis, benthic microbe distributions and with optical plankton counters (Currie et al., 1998; Seuront et al., 2002; Seuront and Spilmont, 2002).

The accumulation of fluorescence data, in particular, is likely to accelerate as collecting devices become more rapid and more widely available due the advent of high-frequency profilers (Wolk et al., 2002), rapidly processing the data to discern whether there are features of interest, will be necessary for accurate and timely interpretation. There already exists standard software that rapidly provides data analysis, such as power spectra and variogram analyses. However, most of these procedures implicitly assume a Gaussian distribution (Chatfield, 1989). This assumption is seldom tested or met. Additionally, these data often need processing (e.g. despiking and detrending) prior to the analysis, and exactly how the data is handled in these processes requires careful consideration (Chatfield, 1989). Often data sets are non-stationary and have variable sampling intervals. The presence of variable sampling interval, in particular, often makes a data set unsuitable for these methods and requires further processing (e.g. interpolation) and interpretation.

Near, real-time analysis may be particularly important in research of microscale processes, where ‘on the spot’ sampling scale decisions might need to be made to help unite information on the scale at which phytoplankton interact with the scale at which biological oceanographers usually make measurements on plankton. Attempts in this area have been made by Seymour et al. (2000), Franks and Jaffe (2001) and Waters and Mitchell (2002), as well as others before that. The results of these studies were deficient in sample size, resolution or sampler design. With the advent of high-resolution fluorometers that can take approximately 1 million of measurements per hour (Wolk et al., 2002), the problem of under sampling a distribution switches to one of being overwhelmed with data by beginning to representatively sample at appropriate scales.

The focus of this paper is the initial analysis of large fluorescence time and spatial series. Specifically, the interest is in the rapid assessment of the extent to which fluorescence series contain meaningful information that is worth spending time on. Implicit in this approach are two features, the ability to rapidly process an entire data set and the understanding to make equally rapid and reliable interpretation of the data. This also implies that incomplete or flawed data can be discarded. Such luxury is a bonus of electronic data collection and allows improved data quality by moving away from the concept that all data must be used and is useful, no matter how much time, interpolation, extrapolation, transformation and *post-hoc* assumption must be done. Even if data cannot be discarded, what is still needed is a clear understanding of the quality of the data.

Here, we use the Zipf analysis presented in Seuront and Mitchell (companion paper 1) on a wide range of fluorescence and chlorophyll time and spatial series. Most of the data sets are comparatively small and taken at the common oceanographic intervals of meters and kilometers. Zipf analysis is used on these data because they are familiar and a useful standard against which to compare high-resolution sample sets where the measurement rate is roughly 2 million points per hour. Zipf analysis rapidly shows and quantifies hidden structure in most data. These results are useful in demonstrating the utility of Zipf analysis and in furthering our understanding of phytoplankton distributions in the sea, particularly at the microscale. To achieve this, data series are presented from large to small scale. Each example was chosen to reflect a particular area or problem in the analysis of phytoplankton distributions. The analysis provides valuable insight into the investigated data sets, in particular, it shows, at least for the data analysed that there is strong heterogeneity at the millimeter to centimeter scale, in contrast to what can be inferred from theoretical work (Siegel, 1998).

## 2. Materials and methods

### 2.1. Zipf analysis of time series: years to seconds

The data examined consisted of vertical profiles from the BATS (Bermuda Atlantic Time Series), vertical and horizontal profiles from the HOTS (Hawaiian Ocean Time Series) programs, vertical and horizontal profiles from the Cooperative Survey of the Pacific Equatorial Zone (EquaPac) and, at the microscale, vertical profiles from a FluoroMAP (Alec Electronics) fluorescence profiler in Sagami Bay, Japan, 5 km from the coast. The BATS and HOTS data were obtained from internet archives at http://www.bbsr.edu/cintoo/bats/bats.html and http://hahana.soest.hawaii.edu/hot/hot_jgofs.html, respectively. The EquaPac data are at www.crseo.ucsb.

edu/seawifs/8D_global/SWxtrct_0N95W.txt, 10N95W.txt, 10S95W.txt, 12N95W.txt, 20S85W.txt, 2N95W.txt, 2S95W.txt, 3.5N95W.txt, 5N95W.txt, 5S95W.txt, 8N95W.txt and 8S95W.txt. The EquaPac data is a multistation chlorophyll *a* transect taken across the equator as illustrated in Fig. 5. The BATS profiles consisted of CTD fluorometer measurements averaged to a sampling interval of 20 cm and taken according to Knap et al. (1994). The BATS profiles were from 1988 to 2000. The HOTS data are chlorophyll measurements from bottle samples taken at approximately 5, 25, 50, 75, 100, 125, 150, 175 and 200 m. Since a 9-point vertical profile was insufficient for our purposes here, all profiles from October 1988 through December 2000 were used. Using all of the data produced 9 depth-specific time series of approximately 400 data points each.

FluoroMAP is a 60 cm long free-falling cylinder that measures fluorescence and pressure. The sampling rate was 512 Hz. The nominal falling speed was around 10 cm/s. The excitation beam was a blue diode laser (Nichia electronics, Japan) 1 mm in diameter from baseline to baseline. Ten-micrometer resolution measurements of the beam showed about 75% of the intensity was confined to an inner 0.4 mm core. The sensor window had a 5 mm radius, but lost sensitivity in the outer 1 mm.

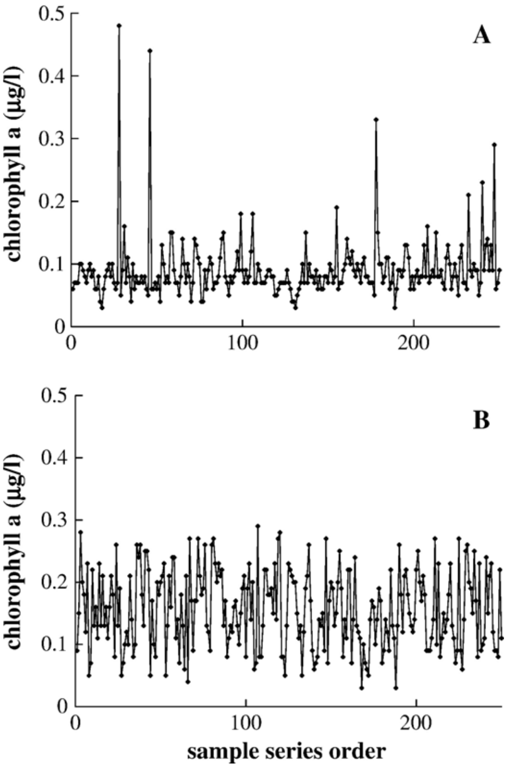

Fig. 1. Samples of the Hawaiian Ocean Time Series showing the maximum (A) and minimum (B) variation patterns.

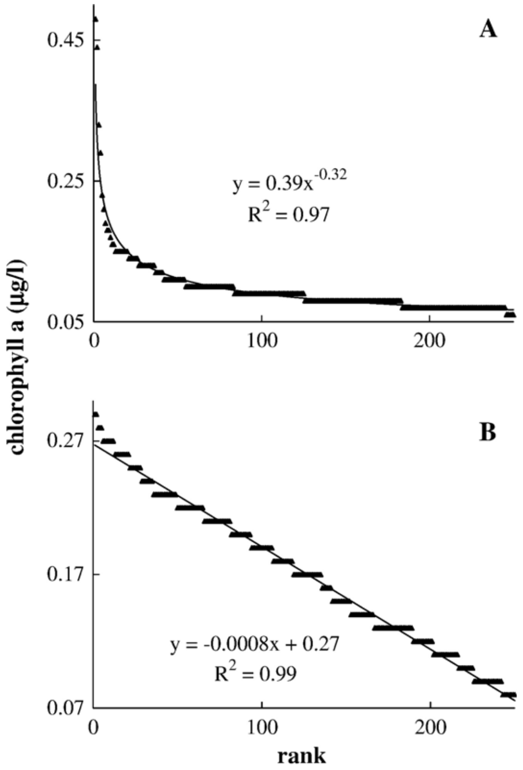

Fig. 2. Zipf plot of the time series from Fig. 1. A and B are plotted on a linear scale. Note that the chlorophyll range for A is 0.4 µg l$^{-1}$ while the range for B is 0.2 µg l$^{-1}$. The constant interval between each stepped groupings of points is the measurement resolution, i.e. the smallest interval the machine can detect.

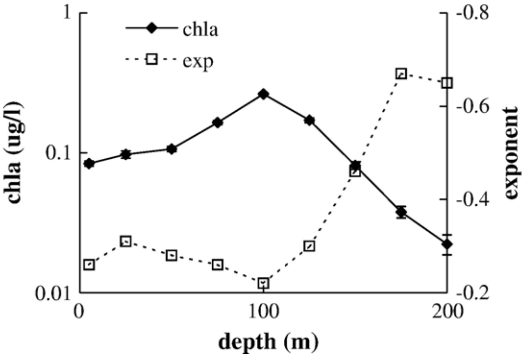

Fig. 3. The mean chlorophyll profile for the Hawaiian Ocean Time Series and Zipf α exponents plotted to show their inverse relation. The Zipf α exponents are obtained by ordering all of the time series values at a given depth. The solid line indicates the mean chlorophyll *a* concentration (µg l$^{-1}$). The dashed line is the Zipf power law exponent α. The error bars are 95% confidence intervals. The chlorophyll axis is a log scale. The depth and exponent axes are linear scales.

BATS, HOTS and EquaPac data were used as they appeared on the websites. FluoroMAP profiles were taken by hand releasing FluoroMAP over the side of the *RV Tansei Maru* in Sagami Bay, Japan. A data cable was played out after it for use as a recovery tether. The tether was left loose so that FluoroMAP fell under its own mass until recovery, usually at about 15 m.

### 2.2. Zipf analysis of one-dimensional patterns

Following Seuront and Mitchell (companion paper 1), chlorophyll *a* concentration (μg l$^{-1}$) and *in vivo* fluorescence will all be referred to as the variable $X_r$. The rank-size behavior of the variable $X_r$, the variable value at a given rank, will thus be expressed as a power law behavior:

$$X_r \propto r^{-\alpha} \qquad (1)$$

where $r$ is the rank of the variable $X_r$. The Zipf exponent α is subsequently estimated as the slope of the best linear regression of $X_r$ vs. $r$ in a log–log plot. However, because Zipf plots do not necessarily exhibit a power law behavior over the whole range of available values of $r$ (Seuront and Mitchell, companion paper 1; their Fig. 14), an objective criterion is needed for deciding upon an appropriate range

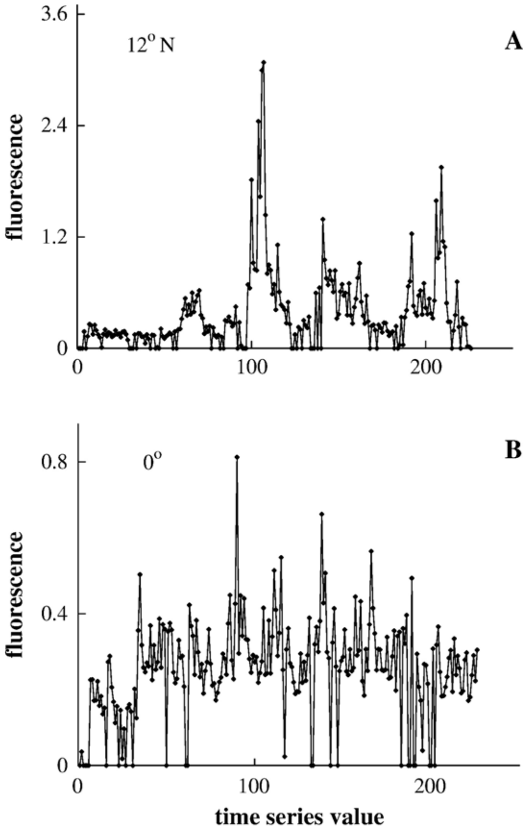

Fig. 4. EquaPac time series showing extremes in chlorophyll variation at specific latitudes. The Zipf analysis and best fit, power law regressions are shown in Fig. 6. The series were chosen because of their different ranges and numerous zero values.

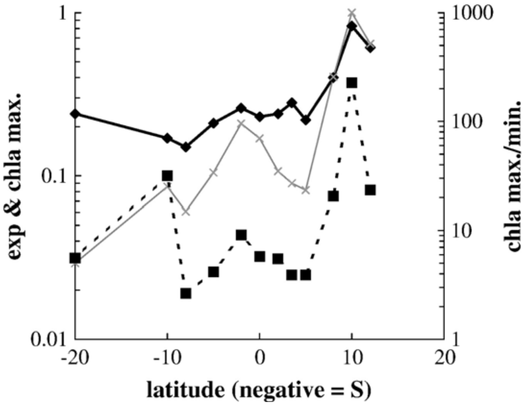

Fig. 5. Transequatorial maximum chlorophyll values (grey line), Zipf exponents (black line) and the ratio between maximum and minimum chlorophyll values (right axis, dashed line). The maximum chlorophyll values are plotted for comparison, rather than the means, because the numerous and unequal number of zeros at each station produces small values showing no pattern. For clarity, the chlorophyll values are normalised such that the maximum value of 4.78 μg l$^{-1}$ at 10° N is 1. Vertical axes are plotted on a log scale.

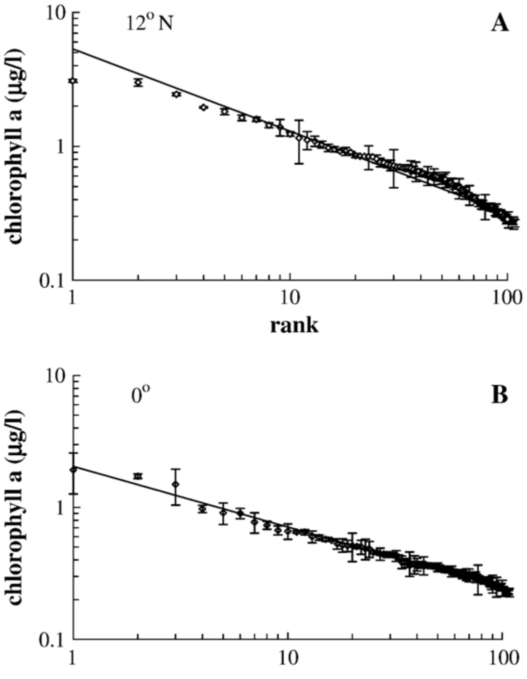

Fig. 6. Rank ordering of time series from Fig. 4. The best line fits are $y=5.33x^{-0.61}$, $r^2=0.96$ (A) and $y=2.05x^{-0.46}$, $r^2=0.99$ (B). Error bars are standard deviations.

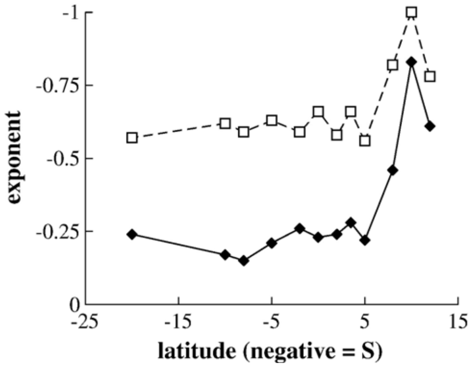

Fig. 7. Comparison of the Zipf exponents for chlorophyll data at each latitude (solid line) to the Zipf exponents for the standard deviation (dashed line) of the chlorophyll at each latitude. Zipf exponents for the standard deviation are consistently closer to −1 than for the chlorophyll values themselves. The data set at each latitude ranged from 149 to 210 data points. All data sets were terminated at 100 data points to calculate the exponent to avoid over sampling bias (Seuront and Mitchell, companion paper 1).

of $r$ to include in the regressions. We used the values of $r$ which maximized the coefficient of determination ($r^2$) and minimised the total sum of the squared residuals for the regression (Seuront and Lagadeuc, 1997; Seuront et al., 2004). None of the data was binned.

## 3. Results

The chlorophyll $a$ values for the HOTS data ranged from 0.01 to 0.5 μg $l^{-1}$. Depth-specific time series showed a variety of patterns. Fig. 1 shows the two extreme patterns observed among all of the time series, corresponding to the best and worst power law fits to the data after ranking, respectively. The intermittent behavior (i.e. a few dense patches over a wide range of low density values) is fully compatible with a power law behavior (Fig. 2A). In contrast, the regular, symmetric fluctuations suggest a uniform distribution (Fig. 1B) and then a linear Zipf behavior (Fig. 2B). More generally, for all the depths investigated, the power law fits ($r^2 \in [0.71–0.97]$) were significantly higher ($p<0.01$) than the linear ones ($r^2 \in [0.43–0.87]$). The best fits were used to extract the α exponents of the power law for each depth. The relationship between the Zipf exponent α and chlorophyll concentration increased from 0 to 100 m, but was always less than 0.3. Below the chlorophyll maximum, α rose to 0.67 as the chlorophyll $a$ values fell (Fig. 3).

The mean chlorophyll values at each site for the EquaPac transect data ranged from 0.01 to 5 μg $l^{-1}$, a factor of 10 higher than the HOTS data for the upper limit. By using repeated measurements from the same station, site-specific time series were generated. These showed a variety of patterns, ranging from intermittent (Fig. 4A) to more uniform (Fig. 4B) as determined by the best and worst power law fits to the data after ranking. In contrast, the data was also looked at as a transequatorial transect of biomass and Zipf exponents (Fig. 5). The $r^2$ values for Fig. 5 were between 0.95 and 0.99. All $r^2$ values were significant at $p<0.01$. The α

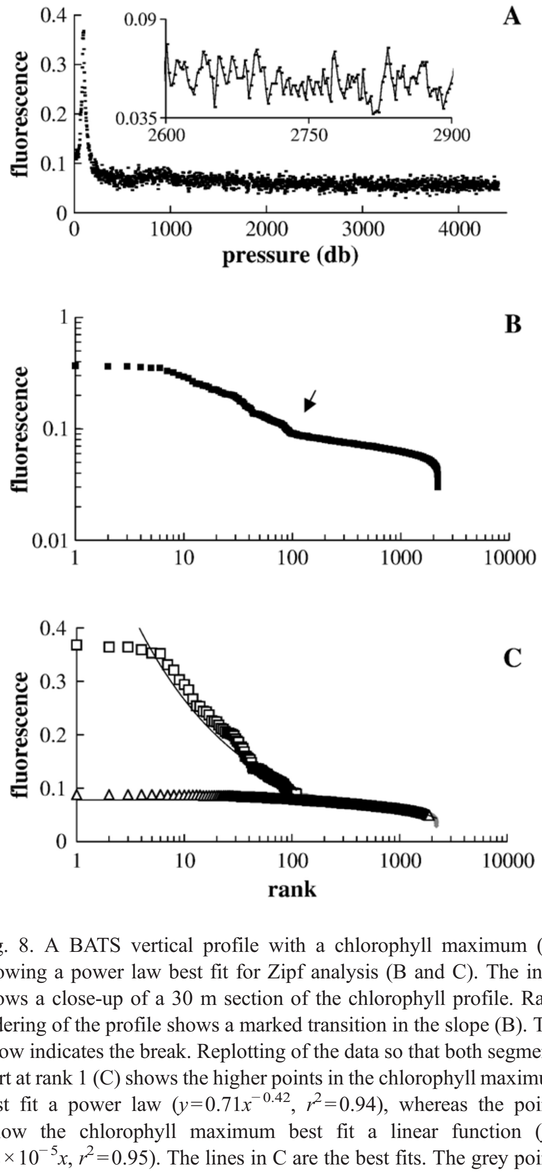

Fig. 8. A BATS vertical profile with a chlorophyll maximum (A) showing a power law best fit for Zipf analysis (B and C). The inset shows a close-up of a 30 m section of the chlorophyll profile. Rank ordering of the profile shows a marked transition in the slope (B). The arrow indicates the break. Replotting of the data so that both segments start at rank 1 (C) shows the higher points in the chlorophyll maximum best fit a power law ($y=0.71x^{-0.42}$, $r^2=0.94$), whereas the points below the chlorophyll maximum best fit a linear function ($y=-2\times10^{-5}x$, $r^2=0.95$). The lines in C are the best fits. The grey points at the end of the bottom line in C show the beginning of the sharp drop from over sampling. These and subsequent grey points were not used in calculating the best fit. This Figure should be compared with Fig. 9, where the best fit for the shallow, high chlorophyll is linear and the best fit for the deep, low chlorophyll values is a power function.

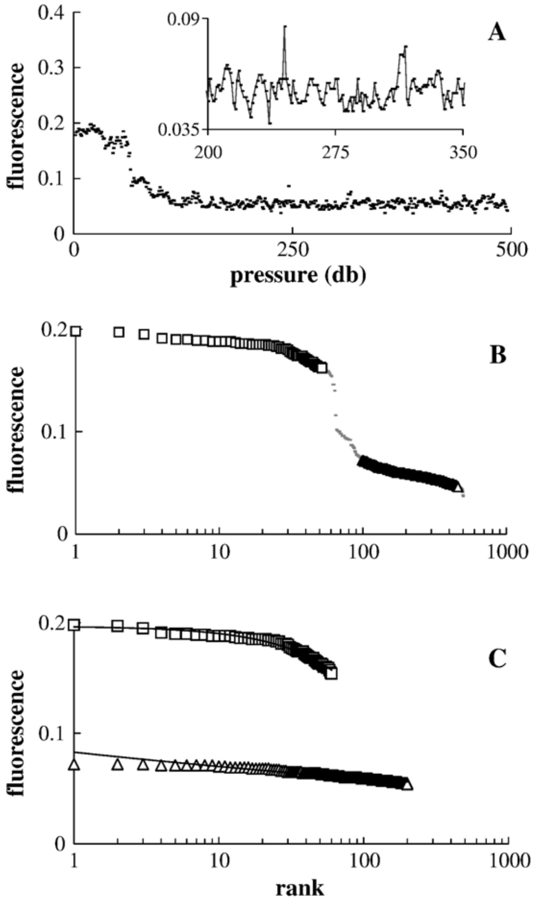

Fig. 9. A BATS vertical profile with a chlorophyll maximum (A) showing a linear best fit for Zipf analysis (B and C). The inset shows a close-up of a 30 m section of the chlorophyll profile. The inset scale is the same as the inset in Fig. 8A for ease of comparison. Rank ordering of the profile shows a marked transition in the slope (B). The *y*-axis is a linear scale to highlight the difference between the two sections of the data (cf. Fig. 8B). The small grey dots are transition points that fall on neither line. The squares and triangles, again emphasize the two different slopes in the data. Replotting of the data so that both segments start at rank 1 (C) shows the higher points in the chlorophyll maximum best fit a linear function ($y=-7\times10^{-4}x+0.2$, $r^2=0.97$), whereas the points below the chlorophyll maximum best fit a power law ($y=8.3\times10^{-2}x^{-0.074}$, $r^2=0.93$). In contrast to Fig. 8, this figure shows the power law fit in the chlorophyll values below the maximum.

exponents showed variation across the transect that ranged from 0.17 to 0.86. At each site, the mean chlorophyll value was an average of 6.9 replicates (replicate $\sigma=2.7$, number of replicated stations = 1859). This replication allowed error bars to be plotted for each Zipf point. Representative graphs are shown in Fig. 6. The error estimates, standard deviations, of these means also have Zipf distributions. The α exponents from Zipf of the standard deviations closely followed that of the α exponents for the chlorophyll *a* values (Fig. 7).

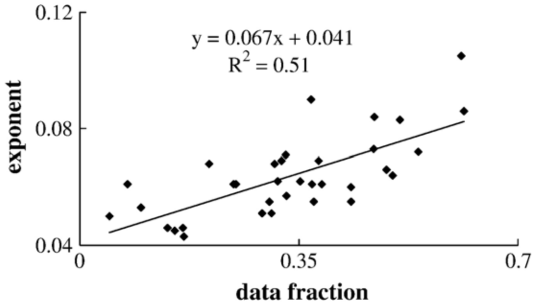

Fig. 10. The value of the Zipf exponents α, shown as a function of the fraction of the data where the best fit was a power law.

For BATS, fluorescence profiles showed classical chlorophyll maxima and low-level variability below the maxima (Figs. 8A and 9A). Ranking the data showed discrete breaks between the maxima and the deeper fluorescence values (Figs. 8B and 9B). The fluorescence maxima and submaxima fluorescence showed best fits that were either linear or power law (Figs. 8C and 9C). A

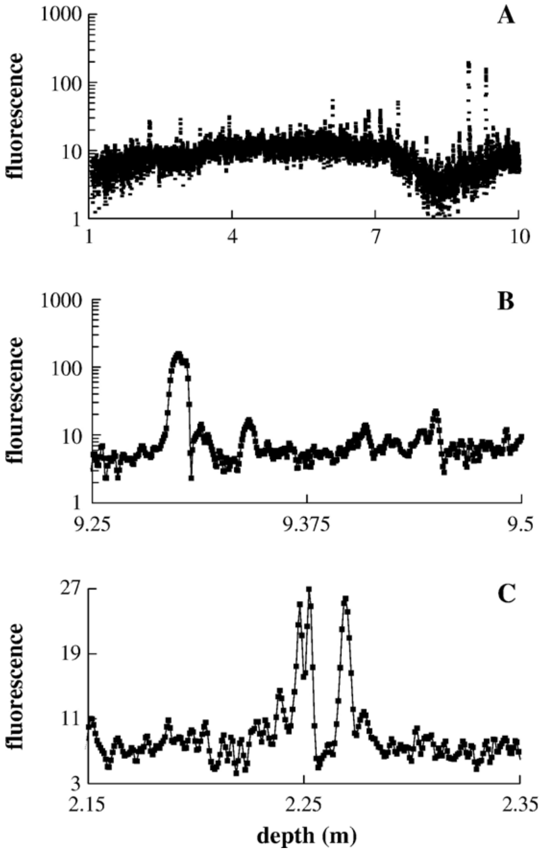

Fig. 11. A FluoroMAP profile shown at 3 scales, 10 m (A), 25 cm (B) and 20 cm (C). Each point represents one measurement. Note that for A and B, fluorescence is plotted on a log scale and spans more than 2 orders of magnitude.

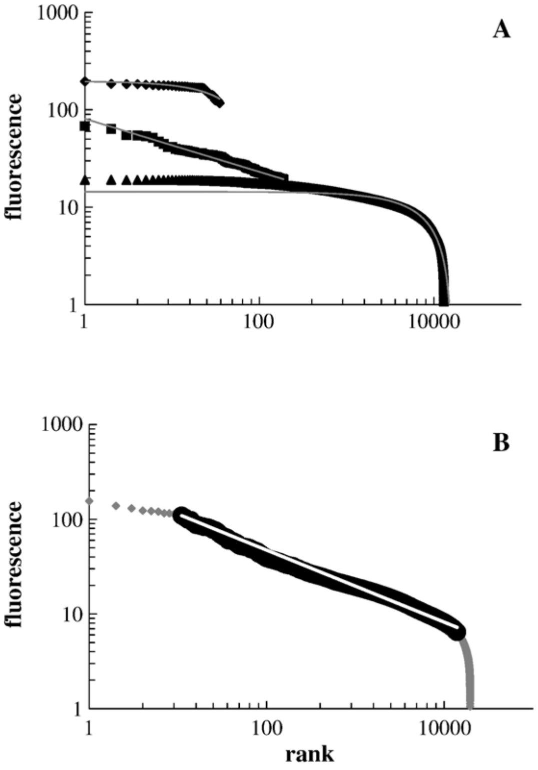

Fig. 12. Rank ordering of FluoroMAP profiles. The data in Fig. 11 is plotted in A as 3 sets of points. The top and bottom sets are linear best fits ($y=1.9x+197$, $r^2=0.88$ and $y=-9\times10^{-4}x+14.41$). The middle set is a power law best fit ($y=80.8x^{-0.37}$, $r^2=0.99$). The Zipf analysis for a second FluoroMAP profile (not shown) at the same location 5 min later is shown in B. The linear fit portions are in grey. The white line fits a power law ($y=272x^{-0.38}$, $r^2=0.98$).

power law was the best fit for submaxima fluorescence 32 out of 35 times, with two ties and one linear best fit. The power law α exponents ranged from 0.043 to 0.105. The 32 data sets ranged in size from 88 to 2037 values per set and the best-fit α exponent occurred over a fraction of the total data set that ranged from 0.07 to 0.61. To determine the extent to which the size of the Zipf α exponent was a function of how much of the data set the best-fit α exponent included, Fig. 10 plots α exponent size against the fractional length of the total data set that gave the best fit.

The FluoroMAP vertical profiles showed similar trends to the BATS submaxima profiles, but with the differences that measurements had a 0.7 mm sampling interval (0.708±0.140 mm; $\bar{x}$±SD, $n=20{,}000$), rather than a meter-scale sampling interval and with all profiles having power laws as the best fit. Example profiles are shown in Figs. 11 and 12. The salient features of the profiles (Figs. 11A and 12A) are that they are composed of 20,000 points and that fluorescence values change by up to 1000 times over distances less than 10 cm. Despite sub-centimeter to supra-meter changes in fluorescence distributions (Fig. 11), ranking all values in each profile still provided a single unbroken power law (Fig. 12B). The small deviations at low ranks did not change the α exponent or $r^2$ values and there were no clear breaks in the slope of the line (Fig. 12B), so the higher ranks were not plotted separately as was done for the BATS data (Figs. 8 and 9). Some FluoroMAP profiles produced Zipf slopes with 4 sections, characteristically the highest and lowest ranks showed a linear best fit, while intermediate ranks showed two distinct α best fits (Fig. 13B). The change in slope appears to be caused by

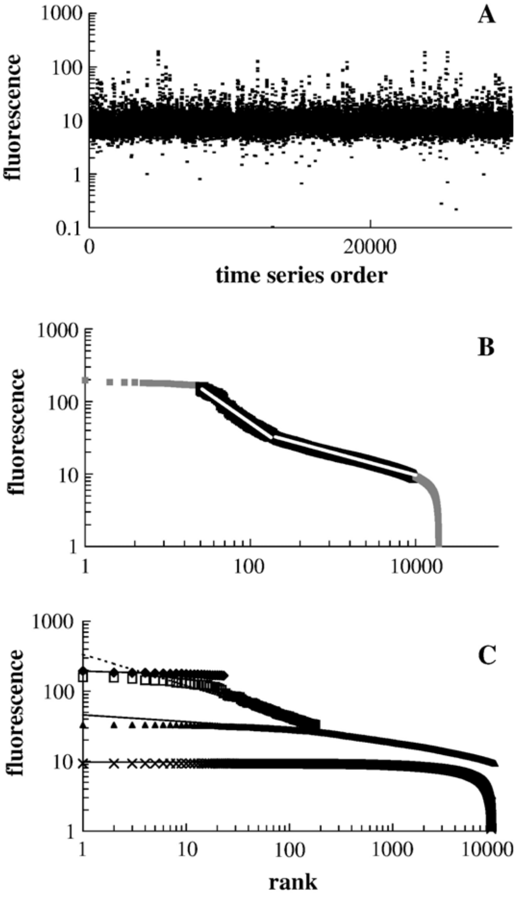

Fig. 13. A FluoroMAP time series at 15 m, where the variation in fluorescence spans 3.5 orders of magnitude. The 30,000 points in the time series were taken over 59 s (A). A rank ordering of the data (B). The distribution has a break and flattens from ranks 1 to 10 and so line fitting is best done on replotted data. Replotting of the 4 subsections (C). The best fits from top to bottom of the graph are power law ($y=194x^{-0.046}$, $r^2=0.97$), power law ($y=336x^{-0.44}$, $r^2=0.95$), power law ($y=101x^{-0.25}$, $r^2=0.96$) and linear ($y=-7\times10^{-4}x+9.71$, $r^2=0.94$).

the presence of sharp peaks in the fluorescence distributions (Fig. 13A).

## 4. Discussion

### *4.1. From time series analysis to Zipf analysis*

Time series sampling has been used to support or reveal many of the basic paradigms of phytoplankton ecology. The classical examples are vertical migration of dinoflagellates and the seasonal changes in phytoplankton abundance and species composition (Veldhuis et al., 1997; Kamykowski et al., 1998, 1999; Sin et al., 2000). Similarly, time series have been used to follow the development of blooms from upwelling and other nutrient inputs (Aristegui et al., 1997; Malej et al., 2003). Common to all of these uses is the detection of episodic events. However, temporal and spatial series are also used for detecting community structure and characteristic time and length scales for both descriptive and dynamical analysis (Platt, 1972; Sugihara and May, 1990; Strutton et al., 1997a,b; Seuront et al., 1999, 2002). Here, we combined the latter use with Zipf analysis to reveal structure that is not otherwise apparent from examination of the data. The series data examined range up to 30,000 points.

### *4.2. Zipf and partial Zipf laws: towards a seascape topology*

Ranking time and spatial series erases the temporal and spatial information implicit in the sampling order, leaving just a distribution of magnitudes. The slope of the rank indicates the distribution of the measured values. Time series from HOTS provide examples of how ranking of distributions vary (Figs. 1 and 2). Most of the distributions shown here contained linear and power law components. We claim that the information lies in the relative number of values that contribute to each component and the goodness of fit for each component (Fig. 10). In particular, eliminating the temporal and spatial relationship in the data permits comparison of distributions sampled at different scales. Measuring the variation in those distributions through the Zipf slopes then reveals structure in the variation. This is an improvement over standard measures of variation, such as the standard deviation or the confidence interval because there is the capacity to dissect the signal into ranges where all values are equally likely and ranges where values are unequally likely based on best fit (Fig. 13B; Seuront and Mitchell, companion paper 1). In particular, comparing the ranges over which values have unequal likelihoods in different samples indicates that there can be a comparable or greater amount of distributional variation over a few centimeters as compared to the variation over tens of meters vertically or a few thousand kilometers horizontally (Figs. 3, 7 and 11)). This quantitative comparison is particularly useful when the data sets include intermittent time series with missing values. These can give the visual impression of being highly variable due to zero values (Fig. 4).

Examining all of the series, from the multiple year BATS series to the high-resolution spatial series from the FluoroMAP profiles, indicates that the fluorescence and chlorophyll distributions are described by power functions. These power functions can arise from rare, multipoint peaks, such as in the BATS data, where the first 100 points out of 3000 give a steep, straight slope followed by a shallower, straighter slope (Fig. 8). More commonly, as in the HOTS and FluoroMAP data, the larger majority of the data fit a single power function (Figs. 2, 12 and 13B), revealing order across all scales and providing a way of quantifying large data sets. Rapid analysis and quantification, along with being a metric to compare disparate profiles and time series, are the strengths of one-dimensional Zipf. Stepped data distributions can produce multiple slopes in Zipf analysis (Fig. 9). The slopes are useful intra time series comparisons and help emphasize that there can be non-random variation at multiple scales, even in regions traditionally considered homogeneous, such as some sub-chlorophyll maximum regions (Fig. 8, note 8A inset). Achieving the best fit requires truncating the data set (Seuront and Mitchell, companion paper 1). In general, the larger the fraction of data explained by the best fit, the larger the exponent (Fig. 10), which may be a quantitative way of determining the extent to which non-random variation permeates a time series.

### *4.3. Ubiquity of Zipf structure: a phenomenological framework*

In this paper, applying Zipf analysis shows that power law distributions occur from sampling intervals of kilometers down to millimeters. The $\alpha$ exponents for these one-dimensional distributions were less than $-1$ and generally greater than $-0.1$. This is less than the $-1$ value found for the original Zipf's law (Zipf, 1949; Seuront and Mitchell, companion paper 1) that is interpreted to indicate aggregative behavior, such as human populations in cities (Marsili and Zhang, 1998), but that is probably not surprising given that plankton are not fixed, nor, by definition able to control their position extensively. If anything, the unexpected result is that there is a non-zero power law at all, particularly at the millimeter sampling interval. One interpretation of these non-zero slopes is that plankton do aggregate and possibly control their distribution over these short distances. Physical and biological aggregation processes are well established over these distances. One spin-off of the Zipf analysis here is that it may provide a means of detecting aggregation from one-dimensional distributions. However, if this is going to be claimed, a close look at errors associated with the method is necessary, particularly at the smaller sampling intervals.

Error at the smallest sampling interval will be the focus here, as the measuring device and sampling interval are not commonly used. Fig. 11B and C shows a strong autocorrelation observed over 5 to 10 sequential points. This is probably the result of FluoroMAP containing low pass filters on its circuit board that smooths signal fluctuations (Alec Electronics, personal communication). Further smoothing occurs because the laser beam in FluoroMAP is oriented vertically as it falls. This vertical orientation maximizes the amount of time that phytoplankton will spend in the beam. Sampling every 500 μm with a 1.2 mm long sampling cylinder (Mitchell et al., unpublished data) means that a given bit of water can remain in the beam for up to 25 sampling points. Given that some lateral advection is likely to occur and the beam's radius of 1 mm, a given cell is unlikely to traverse the entire long axis of the sampling volume. The conclusion is that because of the configuration of FluoroMAP variability is under estimated due to smoothing and the highest resolution of the instrument is lost.

In Zipf analysis, smoothing reduces the magnitude of the extremes and so can decrease the α exponent. The important ecological implications are that Zipf analysis underestimates aggregation, that is it is a conservative measure of aggregation, and that it applies as far down as we are able to discern in the data presented for FluoroMAP. Furthermore, there is no evidence of homogenisation or decrease in variation at the smallest scale. This is in conflict with Siegel's hypothesis on plankton distributions (Siegel, 1998). Siegel proposed continuously changing variation as a function of length scale. Most notably there was a variation minimum when the population size was small, on the order of thousands of individuals. For phytoplankton, the distance over which such a population size can be found ranges from millimeters for coastal cyanobacteria to many centimeters for large, eukaryotic phytoplankton in oligotrophic waters. Although the presence of variation has been quantified before, the classical statistical techniques of variance, coefficient of variation, standard deviation and 95% confidence intervals tend to deemphasize the extremes, which in some circumstances are expected to be the most ecologically relevant values. Zipf emphasizes these and allows them to be quantified. This permits the comparison of real and theoretical distributions and provides a mean of testing field observations, laboratory and modelling experiments against proposed mechanisms. Ultimately, this should provide insight into the balance among behavioral, chemical, and physical mechanisms that control phytoplankton distributions. The influence of all three processes, particularly behavior and chemistry, is most direct and intuitive over distances relevant to individual cells or small populations. However, as Young et al. (2001) have shown, such small process can, in theory, cascade upwards to the maximum dimension in the system.

### 4.4. Zipf analysis as an index of patchiness

If Zipf analysis of phytoplankton emphasizes extremes, it is worth exploring what those extremes represent. Departures from background concentrations that occur in a confined area are traditionally labelled as patches (Okubo and Mitchell, 2001). Implicit in the term patch is the assumption that the change in chlorophyll concentration is positive. To accommodate the increasing accuracy of measurements and our developing view of the ocean environment, the terms 'hotspot' and 'coldspot' have been introduced to discriminate between volumes of increased biomass and volumes of decreased biomass relative to background (Azam, 1998). Zipf analysis, by visually emphasizing the hot and cold spots is effectively highlighting patchiness.

The classic use of Zipf for city aggregations (Marsili and Zhang, 1998), business size (Axtell, 2001) as well as standard plankton patch theory assumes a background of zero (Okubo and Mitchell, 2001). Here, however, there is an average background concentration and the fluctuations can be positive (local clusters from aggregation, swarming, particle disintegration) or negative (local grazing, viral lysis, particle scavenging). While Zipf analysis implicitly emphasizes hotspots responsible for the observed power law distributions and that has been the focus here, in the future it may be worth exploring Zipf as a descriptor and quantifier of cold spots.

No matter the direction of the departure from background, it is important to keep in mind that the process of rank ordering in Zipf analysis does not necessarily erase all spatial information. Particularly where power function fits are good and the α exponent is steep, the few points that contribute to this process may all belong to the same extremely high patch. In this hypothetical case, a single patch may bias the slope upwards. This is what might be intuitively predicted. However, the data presented here tends to show the opposite, that the few high points form a shallower line than the medium rank values (Figs. 6B, 8C, 12B, 13B). This is explained in Seuront and Mitchell (companion paper 1) and arises as a consequence of under sampling of rare intense patches. In a qualitative sense, the emphasis Zipf places on the high values is balanced by the under sampling of those rare values. Bias from the extremes, in any case, can be excluded, as was done here (Figs. 12 and 13; Seuront and Mitchell, companion paper 1), by applying the best fit to the medium rank values. This medium ranks still encompassed a large fraction of the data sets (Figs. 12 and 13).

Zipf analysis of chlorophyll and fluorescence series data permits comparison of data sets from disparate scales and sampling intervals that contain missing values. The comparison is quantitative, rapid and visual. From the Zipf analysis of the series presented here it appears subtle variation in phytoplankton distributions can be detected. More importantly, it appears that variation across scales of millimeters and seconds is similar to that found across kilometers and years. Variations at the small scales may be important to individual plankters.

## Acknowledgements

The crew and staff of the University of Tokyo's *RV Tansei Maru* provided extensive support for obtaining the FluoroMAP data. Thank you to the BATS, HOTS and EquaPac programs for making their data publicly available online. The Australian Research Council, Flinders University of South Australia and the University of Tokyo's Ocean Research Institute supported this work. The Flinders Faculty of Science and Engineering provided a visiting scholarship to LS.

## References

Aristegui, J., Tett, P., Hernandez-Guerra, A., Basterrretxea, G., Montero, M.F., Wild, K., Sangra, P., Hernandez-Leon, S., Canton, M., Garcia-Braun, J.A., Pacheco, M., Barton, E.D., 1997. The influence of island-generated eddies on chlorophyll distribution: a study of mesoscale variation around Gran Canaria. Deep-Sea Res. 44, 71–96.

Axtell, R.L., 2001. Zipf distribution of U.S. firm sizes. Science 293, 1818–1820.

Azam, F., 1998. Microbial control of oceanic carbon flux: the plot thickens. Science 280, 694–696.

Behrenfield, M.J., Bale, A.J., Kolber, Z.S., Aiken, J., Falkowski, P.G., 1996. Confirmation of iron limitation of phytoplankton photosynthesis in the equatorial Pacific Ocean. Nature 383, 508–511.

Chatfield, C., 1989. The Analysis of Time Series: An Introduction, 4th ed. Chapman and Hall, London.

Chavez, F.P., Strutton, P.G., Friederich, G.E., Feely, R.A., Feldman, G.C., Foley, D.G., McPhaden, M.J., 1999. Biological and chemical response of the equatorial Pacific Ocean to the 1997–98 El Niño. Science 286, 2126–2131.

Cowles, T.J., Desiderio, R.A., 1998. Small-scale planktonic structure: persistence and trophic consequences. Oceanography 11, 4–9.

Cowles, T.J., Desiderio, R.A., Neuer, S., 1993. In situ characterization of phytoplankton from vertical profiles of fluorescence emission spectra. Mar. Biol. 115, 217–222.

Currie, W.J.S., Claereboudt, M.R., Roff, J.C., 1998. Gaps and patches in the ocean: a one-dimensional analysis of planktonic distributions. Mar. Ecol., Prog. Ser. 171, 15–21.

Feldman, G.C., Clark, D., Halpern, D., 1984. Satellite color observations of the phytoplankton distribution in the eastern Equatorial Pacific during the 1982–83 El Niño. Science 226, 1069–1071.

Franks, P.J.S., Jaffe, J.S., 2001. Microscale distributions of phytoplankton: initial results from a two-dimensional imaging fluorometer, OSST. Mar. Ecol., Prog. Ser. 220, 59–72.

Kamykowski, D.E.J., Milligan, E.J., Reed, R.E., 1998. Relationships between taxis responses and diel vertical migration in autotrophic dinoflagellates. J. Plankton Res. 20, 1781–1796.

Kamykowski, D.E.J., Milligan, E.J., Reed, R.E., Liu, W., 1999. Geotaxis/phototaxis and biochemical patterns in *Heterocapsa* (Cachonina) *illdefina* (Dinophyceae) during diel vertical migrations. J. Phycol. 35, 1397–1403.

Knap, A. et al. (eds.) 1994. Protocols for the Joint Global Ocean Flux Study (JGOFS) Core Measurements. JGOFS Report 19 as IOC Manual 29, UNESCO, Cat. No. 99739.

Malej, A., Mozetiè, P., Turk, V., Terziè, S., Ahel, M., Cauwet, G., 2003. Changes in particulate and dissolved organic matter in nutrient-enriched enclosures from an area influenced by mucilage: the northern Adriatic Sea. J. Plankton Res. 25, 949–966.

Marsili, M., Zhang, Y.C., 1998. Intercating individuals leading to Zipf's law. Phys. Rev. Lett. 80, 2741–2744.

Mitchell, J.G., Fuhrman, J.A., 1989. Centimeter scale vertical heterogeneity in bacteria and chlorophyll *a*. Mar. Ecol., Prog. Ser. 54, 141–148.

Okubo, A., Mitchell, J.G., 2001. Patchy Distribution and Diffusion, In: Okubo, A., Levin, S.A. (Eds.), Diffusion and Ecological Problems – Modern Perspectives, 2nd edition. Springer-Verlag, New York.

Platt, T., 1972. Local phytoplankton abundance and turbulence. Deep-Sea Res. 19, 183–187.

Seuront, L., Lagadeuc, Y., 1997. Characterisation of space–time variability in stratified and mixed coastal waters (Baie des Chaleurs, Québec, Canada): application of fractal theory. Mar. Ecol., Prog. Ser. 159, 81–95.

Seuront, L., Spilmont, N., 2002. Self-organized criticality in intertidal microphytobenthos patch patterns. Physica A 313, 513–539.

Seuront, L., Schmitt, F., Lagadeuc, Y., Schertzer, D., Lovejoy, S., 1999. Multifractal analysis as a tool to characterize multiscale inhomogeneous patterns. Example of phytoplankton distribution in turbulent coastal waters. J. Plankton Res. 21, 877–922.

Seuront, L., Gentilhomme, V., Lagadeuc, Y., 2002. Small-scale nutrient patches in tidally mixed coastal waters. Mar. Ecol., Prog. Ser. 232, 29–44.

Seuront, L., Schmitt, F.G., Brewer, M.C., Strickler, J.R., Souissi, S., 2004. From random walk to multifractal random walk in zooplankton swimming behavior. Zool. Stud. 43, 8–19.

Seymour, J.R., Mitchell, J.G., Pearson, L., Waters, R.L., 2000. Heterogeneity in bacterioplankton abundance from 4.5 millimetre resolution sampling. Aquat. Microb. Ecol. 22, 143–153.

Siegel, D.A., 1998. Resource competition in a discrete environment: why are plankton distributions paradoxical? Limnol. Oceanogr. 43, 1133–1146.

Sin, Y., Wetzel, R.L., Anderson, I.C., 2000. Seasonal variations of size-fractionated phytoplankton along the salinity gradient in the York River estuary, Virginia (USA). J. Plankton Res. 22, 1945–1960.

Strutton, P.G., Mitchell, J.G., Parslow, J.S., 1996. Non-linear analysis of chlorophyll a transects as a method of quantifying spatial structure. J. Plankton Res. 18, 1717–1726.

Strutton, P.G., Mitchell, J.G., Parslow, J.S., Greene, R.M., 1997a. Phytoplankton patchiness: quantifying the biological contribution via fast repetition rate fluorometry. J. Plankton Res. 19, 1265–1274.

Strutton, P.G., Mitchell, J.G., Parslow, J.S., 1997b. Using non-linear analysis to compare the spatial structure of chlorophyll with passive tracers. J. Plankton Res. 19, 1553–1564.

Sugihara, G., May, R.M., 1990. Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. Nature 344, 734–741.

Veldhuis, M.J.W., Kraay, G.W., Van Bleijswijk, J.D.L., Baars, M.A., 1997. Seasonal and spatial variation in phytoplankton biomass, productivity and growth in the northwestern Indian Ocean: the southwest and northeast monsoon, 1992–1993. Deep Sea Res. 44, 425–449.

Waters, R.L., Mitchell, J.G., 2002. Centimeter-scale spatial structure of estuarine in vivo fluorescence profiles. Mar. Ecol., Prog. Ser. 237, 51–63.

Wolk, F., Yamazaki, H., Seuront, L., Lueck, R.G., 2002. A new free-fall profiler for measuring biophysical microstructure. J. Atmos. Ocean. Technol. 19, 780–793.

Young, W.R., Roberts, A.J., Stuhne, G., 2001. Reproductive pair correlations and the clustering of organisms. Nature 412, 328–331.

Zipf, G.K., 1949. Human Behavior and the Principle of Least Effort. Hafner, New York.