Omic-style statistical clustering reveals old and new patterns in the Gulf of Maine ecosystem

<table>
<thead>
<tr>
<th>Journal:</th>
<th>Canadian Journal of Fisheries and Aquatic Sciences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID:</td>
<td>cfas-2016-0151.R1</td>
</tr>
<tr>
<td>Manuscript Type:</td>
<td>Article</td>
</tr>
<tr>
<td>Date Submitted by the Author:</td>
<td>10-Aug-2016</td>
</tr>
<tr>
<td>Complete List of Authors:</td>
<td>Record, Nicholas; Bigelow Laboratory for Ocean Sciences, O’Brien, John; Bowdoin College, Mathematics Stamieszkin, Karen; University of Maine, School of Marine Sciences Runge, Jeffrey; University of Maine, School of Marine Sciences; Gulf of Maine Research Institute</td>
</tr>
<tr>
<td>Keyword:</td>
<td>BIODIVERSITY &lt; General, Dirichlet-multinomial, continuous plankton recorder, ZOOPLANKTON &lt; Organisms, Gulf of Maine</td>
</tr>
</tbody>
</table>
Omic-style statistical clustering reveals old and new patterns in the Gulf of Maine ecosystem

Nicholas R. Record, John D. O’Brien, Karen Stamieszkin, and Jeffrey A. Runge

Abstract: The burgeoning of omic technology has spawned a new subfield of statistics aimed at interpreting the complex information contained in omic data. Some of these statistical methods can be applied to any dataset with taxonomic counts, and they have the potential to provide additional insights over traditional approaches. We test this potential by reanalyzing a well-studied zooplankton dataset—the Gulf of Maine continuous plankton recorder series—using a modified Dirichlet-multinomial mixture (DMM) model. The dataset has ∼50 years of ∼monthly samples along a transect from Boston, USA to Yarmouth, Canada. The results from the DMM analysis were largely consistent with previous analyses, but also provided new insights. Notably, the *Calanus*-dominated communities that returned following a reduction in the 1990s showed a loss of background diversity, suggesting a shift in sources and possibly higher vulnerability of these communities. The DMM analysis also revealed a breakdown of seasonal ecological succession in the 1990s. These changes could be a precursor to similar changes in other *Calanus*-dominated systems. The approach demonstrates a path toward linking traditional analyses with recent omic-style analyses.

Key words: biodiversity, Dirichlet-multinomial, continuous plankton recorder, zooplankton, Gulf of Maine.

Nicholas R. Record. Bigelow Laboratory for Ocean Sciences, East Boothbay ME, USA
John D. O’Brien. Bowdoin College, Brunswick ME, USA
Karen Stamieszkin. University of Maine, Orono ME, USA
Jeffrey A. Runge. University of Maine, Orono ME, USA

1Corresponding author (e-mail: nrecord@bigelow.org).
1. Introduction

Ocean ecology is at a crossroads. There is a tension between the need to simplify systems down to their dominant species or groups on the one hand and the need to understand the dynamics of complex aggregations of species on the other hand. In ocean ecology, the former perspective has its roots in the pioneering work of Gordon Riley (Riley et al 1949), whose dynamical systems equations form the core of modern biogeochemical ocean models. The latter perspective represents a primary challenge in ecology, with origins in information theory (Shannon 1948), encapsulated by the term “biodiversity.” While understanding of the complexity of biodiverse communities has long been a goal of ecology, the explosion of omics data has brought this question to the fore, driving the development of new methods, and challenging conventional paradigms.

Explanations for the properties and dynamics of diverse ecological communities range from the empirical (Currie 1991, McGrady-Steed et al 1997, Peterson et al 1998) to the theoretical (Hubbell 2001, Allen et al 2002, Cropp & Norbury 2012, Record et al 2013). Historically, the struggle to understand and simplify ecological complexity can be seen in the “diversity of diversities” phenomenon—that is, the multifarious definitions and indices attempting to describe biodiversity (Purvis & Hector 2000, Magurran 2013, e.g.). The omics revolution, in providing new tools for characterizing and quantifying biological molecules, has taken this line of research in new directions. The revolutionary capacities of metagenomic techniques in particular, where researchers can measure the abundance of tens of thousands of microbial taxa simultaneously and unbiased by a culturing step—has led to an explosion of data on microbial ecosystems, requiring new statistical interpretation methods. Such “omic-style” statistical methods initially focused on metrics for determining shifts of relative abundance (Lozupone & Knight 2005, Huson et al 2007, White et al 2009). The past few years have seen an assortment of different approaches for analyzing these data, including regression (Chen et al 2012, Chen & Li 2013), co-occurrence networks (Biswas et al 2015), and ecological consortia inferral (Holmes et al 2012). These methods have proven essential to detailing new aspects of human health, environmental science, and agriculture.

In many ways, the problem of interpreting metagenomic counts is the same as the problem of interpreting traditional species counts. There is the same challenge of pulling the most meaningful or important signals out of the complexity of biodiverse systems. It is reasonable to ask whether these
new statistical analyses are consistent with traditional analyses and whether they add any new insights.

To answer these questions, we have reanalyzed a thoroughly studied zooplankton dataset—the Gulf of Maine continuous plankton recorder (CPR) time series—using an infinite dimensional Dirichlet mixture model (DMM) designed to analyze metagenomic count data. The Gulf of Maine is a semi-enclosed basin, divided between the northeast United States and Atlantic Canada (Fig 1), with a rich history of scientific study (Bigelow 1924). On both the US and Canadian sides, the region supports culturally and economically important fisheries, a growing aquaculture industry, and an ecotourism industry. Over the past 10 years, this region has experienced some of the fastest warming on the planet, with the rate of warming exceeding that of 99.9% of the global ocean, punctuated by a year-long ocean heat wave with temperatures consistently 2 - 3 °C above the climatological mean (Pershing et al 2015). This time period has also been one of profound changes in the physical properties (Smith et al., 2012), precipitation patterns, and productivity (Balch et al., 2012) in the Gulf of Maine, with direct consequences for species throughout the food web, including those of commercial interest (Mills et al., 2013; Hunter et al., 2013). At the same time, changes in the zooplankton community that had been expected with warm temperatures have not yet come to pass (Runge et al. 2015). This rapid and extreme change provides new motivation to understand the basic ecology of this system. The Gulf of Maine CPR dataset has previously been analyzed in terms of principal component analysis (Pershing et al 2005), species richness (Record et al 2010), community size structure (Stamieszkin et al 2015), and correlation statistics (Pershing et al 2010), and makes for an interesting benchmark for these new statistical methods.

With this study, we aim to answer the questions: 1) does the DMM reproduce patterns from previous analyses, and 2) does the DMM provide any new insights into the community structure and dynamics? In addressing these questions, we hope to help bridge the gap between conventional ecological approaches and the new omics-based statistical approaches.

2. Materials and Methods

The continuous plankton recorder (CPR) program, initiated in 1931 by Sir Alister Hardy (Richardson et al 2006), has produced a uniquely long-term, standardized database of marine zooplankton counts. The CPR device is towed behind ships of opportunity, continuously sampling the surface zooplankton community through a 270 µm spooling silk mesh. The Gulf of Maine CPR transect traces 450 km of the Gulf of Maine, approximately west-east from Massachusetts to Nova Scotia (Fig 1), with
4791 samples beginning in 1961. There are ~120 zooplankton taxa represented in the data, identified to varying taxonomic levels. The zooplankton community is dominated by copepods, and most of the 51 copepod taxa are identified to genus or species, so we focused on copepods for this study. In the CPR dataset, some copepods are identified to genus and some to species. Just as with microbe data, the line distinguishing species is still a matter of debate (Parent et al 2012), so we used the taxonomic classification as-is. We binned the data by month and by 0.5-degree longitudinal section.

We tested a version of a widely-used analytic method for microbiome data: the Dirichlet-multinomial mixture model (DMM). Starting with taxa count data from a set of samples, the model uses a Markov chain Monte Carlo (MCMC) technique with a Dirichlet process mixture model to infer the number of components, where each component is a Dirichlet-multinomial distribution. This distribution is similar to a multinomial distribution generally suitable for count data but allows for additional dispersion among observations over what a strict multinomial permits. The specific inference methodology has been described recently (O’Brien et al 2015, 2016), while the original variational approach for microbiome data and a more recent extension to test for niche neutrality lay out the generative model in greater detail (Holmes et al 2012, Harris et al 2014). All scripts were implemented in the R computing environment. Scripts implementing the MCMC as well as visualizations routines are available in an open repository: https://github.com/jacobian1980/ecostates.

We use the term ecostate to refer to DMM components, which represent consortia of taxa that cluster together consistently, as well as their respective abundances (O’Brien et al 2016). To visualize the separation between ecostates, we plotted them in realized-niche space. We used sea surface temperature (Reynolds et al 2007) and the CPR phytoplankton color index (PCI, Richardson et al 2006) as environmental dimensions, and plotted the occurrence of each ecostate.

3. Results

With 10,000 iterations, the DMM found a strong mode at 9 ± 0.4 ecostates, with extrema at 7 and 10 over the full set of iterations. To condense this information, we combined a subset 10 randomly chosen iterations and used the modal ecostate at each time/location. The system is generally Calanus-dominated, so we ordered the ecostates by abundance of Calanus taxa (Fig 2). Ecostates #1-5 are composed of < 26% Calanus, and ecostates #6-9 are composed of > 50% Calanus. Within this grouping, there is some interesting variability. For example, within the high-Calanus communities, there is...
an ecosate with a very diverse mix of other taxa comprising about a third of the community (#7), an ecosate where the three large-bodied genera (*Paraeuchaeta, Calanus, Metridia*) make up 96% of the community (#8), and an ecosate where *Calanus* alone makes up 96% of the community (#9). Among the low-*Calanus* ecosates, there is a distinction between those for which *Centropages* has a dominant relative abundance (#2, #5) and those for which *Centropages* has a small relative abundance (#1, #3, #4). Other taxa, such as the predatory *Paraeuchaeta* or the very small *Oithona*, as well as the evenness of taxa, mark the other distinctions between ecosates.

Ecostate variability in space and time showed some notable patterns (Fig. 3). Through most of the 1980s there was a stable repeating pattern alternating seasonally between *Calanus*-dominated communities in the winter-summer (#7-9) and a diverse *Centropages*-dominated community in autumn (#2). This pattern spanned most of the transect over the deep basins and broke down at the ends of the transect near the coasts (inshore of the 100 m isobath). During the 1980s, excluding the coastal endpoints, *Calanus* ecosates made up 56% of all ecosates. Beginning in late 1990 there was a diversification of ecosates toward a more rapid and irregular succession, as well as higher variability along the transect. *Calanus* ecosates reduced to 19% of the total ecosates, while ecosate #5 was nearly absent in all decades except the 1990s. In the early 2000s, there was a partial shift back toward the pattern observed in the 1980s, with the *Calanus* ecosates comprising 44% of the total. The other decades were sparsely sampled, but taken together, high *Calanus* ecosates comprised 46% of the total. Over the duration of the time series, the diverse *Calanus* ecosate (#7) comprised a decreasing proportion of all *Calanus* ecosates, from ~ 40% to ~ 15% (*p* < 0.01). Among the low *Calanus* ecosates, the most prevalent was the *Centropages*-dominated ecosate #2, with a consistent autumn timing throughout the dataset. Ecosate #4, a diverse ecosate characterized by *Pseudocalanus* and a long tail of rare taxa, was almost nonexistent until 1980, and then maintained a steady low presence of around 3% of the total. Ecosates #3 and #4 were mostly confined to the ends of the transect, inshore of the 100 m isobath.

Many of the ecosates had similar realized niches in PCI-temperature space (Fig. 4). The three *Calanus*-dominated ecosates had a strong mode at low temperatures—around 5 °C—spanning a range of PCI values, with ecosate #9 also extending into higher temperatures. Realized niches for ecosates #1, 4, and 6 had a similar shape. The common *Centropages* ecosate (#2) had a mode around 10 °C, but with a realized niche that spanned all temperatures and PCI values ≤ 1. Ecosate #5 had a mode around 15 °C at PCI values ranging from 0-2, higher than any other ecosate at this temperature. Ecosate #3,
while comparatively rare through the time series, had a realized niche that appeared to include nearly all temperature-PCI pairs, excepting the highest PCI values.

4. Discussion

There is a substantial body of work analyzing the copepod community in this CPR dataset (Jossi & Goulet 1993, Greene & Pershing 2000, Conversi et al 2001, Licandro et al 2001, Piontkovski & Hameed 2002, Greene et al 2003, Pershing et al 2005, Kane 2009, Record et al 2010, Johnson et al 2011, Greene et al 2012, Stamieszkin et al 2015). These studies have looked at the Gulf of Maine zooplankton community through a variety of lenses, including conventional statistical analysis like correlation (Pershing et al 2010) and principal component analysis (Pershing et al 2005), as well as community properties like taxonomic richness (Record et al 2010) and size structure (Stamieszkin et al 2015). These studies have identified decadal and interannual community shifts associated with climate signals and other forcing, with a notable shift that lasted throughout the 1990s driven by stratification and the intensification of an autumn phytoplankton bloom (Greene et al 2012). Interestingly, described changes in the copepod community structure across all of these studies boil down to fluctuations in abundance of *Calanus finmarchicus*, whose prominence in the Gulf of Maine has been recognized since the earliest plankton surveys (Bigelow 1924). Essentially, these studies describe a bimodal system: there is either a comparatively high abundance of *C. finmarchicus*, dominating the copepod community and driving the size structure and species richness, or a comparatively low abundance of *C. finmarchicus*, giving way to a consistent suite of smaller copepods (i.e. higher diversity and steeper size spectral slope). This dichotomy encapsulates the conventional ecological analyses described above, whether they analyze richness, size structure, community structure, or individual species dynamics. The analysis here reproduces this main finding of previous work. That is, there is a marked dynamic between high-*Calanus* and low-*Calanus* communities, with the main signals manifesting as a shift during the 1990s as well as within seasonal succession dynamics. In this sense, the DMM inference reproduced previous findings in the Gulf of Maine CPR dataset: the dynamics between high-*Calanus* and low-*Calanus* communities showed the same pattern.

The second aim of this study was to determine whether a DMM analysis can provide new insights. There were a few signals that revealed new characteristics of change in the copepod community. For example, one aspect of the 1980s-1990s transition that has not been described is the breakdown of a
consistent seasonal succession. In the 1980s, *Calanus* ecostates dominated in the winter, spring, and summer, giving way to a *Centropages* ecostate in the autumn. In the 1990s, in addition to a transition away from *Calanus* ecostates, there was a transition toward higher entropy—often an indication of a disturbance or perturbation. The PCI data also consistently showed a regular spring bloom pattern in the 1980s, across the full transect and including all years but one, followed by an erratic pattern in the 1990s, where blooms occurred at many different times of year (Fig. 5). This erratic pattern included a new high-temperature (∼15°C), comparatively high-PCI (≥1) niche that corresponded with the appearance of ecostate #5. There was a partial return in the 2000s to the regularity of the seasonal succession in ecostates, but the PCI patterns continued to be erratic. In seasonal environments like the Gulf of Maine, species often have life cycles tuned to phytoplankton bloom periods. More erratic bloom timing, as inferred from the PCI, could lead to mismatches during important life history stages (Cushing 1969, 1990). Haddock recruitment (Leaf & Friedland 2014), zooplankton biovolume (Friedland et al 2015), and *C. finmarchicus* production (Durbin et al 2003, Runge et al 2015) all appear to respond to phenological changes in primary production. The strong alignment between phytoplankton phenology and the ecostate patterns suggests that the influence of match-mismatch dynamics can extend to the structure of the copepod community.

It is notable that within the *Calanus*-dominated ecostates, the diverse ecostate #7 has become progressively rarer. In addition to having a higher diversity among the dominant genera, this ecostate also had a long tail of rare taxa. In other words, in the early part of the time series, although *Calanus* dominated, that dominance was associated with a higher richness of other taxa. The return of *Calanus* communities following the 1990s low did not include a return of this background richness. This transition poses some interesting questions. Moving northward from the Gulf of Maine, copepod communities are increasingly *Calanus*-dominated and decreasingly diverse. It is reasonable to ask why, during recent warming, the *Calanus* ecostate has shifted toward one of a more northerly character. One hypothesis is that prior to the 1990s, the *Calanus* ecostates had greater richness because the copepod community, including *Calanus*, was more locally supported. In contrast, warmer recent conditions may favor advected *Calanus* communities but not a locally supported population. One of the major sources of advection to the Gulf of Maine is across the Scotian Shelf, which has a more subarctic character, and changes in currents can drive changes in the zooplankton community (Greene et al 2003, Kane 2007, Runge et al 2015). If the *Calanus*-dominated ecostates are more predominantly supplied from...
the Scotian Shelf than previously, this would explain a shift toward less diverse *Calanus* ecostates. If the hypothesis is true, one implication is a greater vulnerability of this key species, with an increased dependence on advective supply over local production.

The importance of rare taxa in the ocean is an open question. The “rare biosphere” has been studied in some depth for microbial organisms, including microzooplankton (Caron & Countway 2009), but there is little work on metazoans. The null hypothesis is that the rare biosphere is ecologically inactive, essentially representing noise. There is evidence, however, that the rare biosphere is important to ecological function, and that rare taxa can have a magnified importance when conditions change, sometimes representing a reservoir of alternative strategies (Countway et al 2005). Under this hypothesis, a decline in rare taxa would represent a loss of potential resilience. Although the CPR has a small sampling opening, long-term CPR datasets contain a sufficient sample abundance to represent a long tail of rare taxa, allowing us to examine the rare biosphere question for copepods. Time series of rare taxa are zero-inflated, so the previous analyses reviewed here (i.e. principle component analysis, species richness, and size spectrum analysis) do not capture their dynamics. The DMM offers an alternative analysis that can extract signals from the time series of these taxa. The distributional assumptions underlying the DMM derive from those of the unified theory of biodiversity (Harris et al 2014). Rare taxa still pose challenges under this framework, but when they pool together into the same ecostate, that signal can be detected (La Rosa et al 2012). The non-random pattern in the dynamics of rare taxa shown here implies that there is a connection between the rare copepod biosphere and the functioning of the ecosystem. Understanding the mechanisms behind this connection requires an investigation of rare taxa across multiple systems, at the basin or global scale, which should be possible by aggregating existing datasets.

The question of whether a diapausing copepod like *C. finmarchicus* will continue to dominate in the Gulf of Maine as temperatures warm is a key ecological question (Runge et al 2015). Through lipid storage, diapausing copepods convert brief pulsed blooms of primary production into an energy resource available to higher trophic levels throughout much of the year. Many fish stocks valuable to Canadian and US fisheries rely on this energy resource during critical times in their life histories, as do protected and endangered species. While *Calanus* appears to have unexpectedly persisted, and even thrived, during this period of warming, the apparent persistence disguises a change in the background diversity that could indicate a fundamental shift in the origin and fate of *Calanus* communities in the
Gulf of Maine. Understanding such a shift is important in the context of the more northerly *Calanus*-dominated, lipid-based food food webs throughout the Subarctic that are likely to follow a similar transition as temperatures increase.

In the case of the Gulf of Maine, the DMM is consistent with conventional analyses and appears to provide some new insight as well. There are some qualifiers to this conclusion however. Statistical methods like the DMM model provide a new entry point into understanding ecological problems at the community level, which can give us new insights. These analyses are, however, still basically empirical, whereas the mechanism-based dynamical equations of Riley (Riley et al 1949) and his successors have predictive potential because they derive from mechanisms. An important open question is whether we can link community-level changes to their underpinning mechanisms, so that we might leverage the new insights they provide to predict ocean dynamics. This is the work that is needed to bridge the mechanistically-oriented approach of Gordon Riley and the empirically-oriented methods borne with the omics revolution. The analysis here takes a step toward bridging this gap.

**Acknowledgements**

Funding for this research came from the National Science Foundation (NRR, OCE 1459092; JAR, OCE1459087) and Bigelow Laboratory institutional funds (NRR). We thank the Northeast Fisheries Science Center for collecting and maintaining the Gulf of Maine CPR data and making the dataset publicly available. We also thank two anonymous reviewers for helpful insights.

**References**


Fig. 1. The Gulf of Maine CPR sampling points. Light grey line indicates the Hague Line separating the United States and Canada.
Fig. 2. The nine ecostates based on the DMM inference, shown with taxa grouped by abundant genera (a) and for the full data (b), and their proportional occurrence in the full data set (c). Shading shows relative abundance within ecostate on a log scale. Ecostate order is based on *Calanus* abundance. Vertical lines in (b) show the grouping of genera used in (a).
Fig. 3. Dynamics of the nine ecostates through time and space. Each window shows ecostates for one year, binned by month on the x-axis and longitudinal distance along transect on the y-axis. Color indicates ecostate, based on numbering in Fig. 2. The histograms at the left show the relative proportion of each ecostate represented in the decade corresponding with that row.
Fig. 4. Realized niches for ecostates in PCI-temperature space. Shading of scatter plots indicates the concentration of points.
Fig. 5. Histogram showing month of peak PCI in the years 1980-1989 (A) and 1990-1999 (B).