Canadian Journal of Fisheries and Aquatic Sciences
Journal canadien des sciences halieutiques et aquatiques

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

Journal:	<i>Canadian Journal of Fisheries and Aquatic Sciences</i>
Manuscript ID	cjfas-2016-0151.R1
Manuscript Type:	Article
Date Submitted by the Author:	10-Aug-2016
Complete List of Authors:	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
Keyword:	BIODIVERSITY < General, Dirichlet-multinomial, continuous plankton recorder, ZOOPLANKTON < Organisms, Gulf of Maine

SCHOLARONE™
Manuscripts

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.¹ 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 inference (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 ecostate with a very diverse mix of other taxa comprising about a third of the community (#7), an ecostate where the three large-bodied genera (*Paraeuchaeta*, *Calanus*, *Metridia*) make up 96% of the community (#8), and an ecostate where *Calanus* alone makes up 96% of the community (#9). Among the low-*Calanus* ecostates, 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 ecostates.

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* ecostates made up 56% of all ecostates. Beginning in late 1990 there was a diversification of ecostates toward a more rapid and irregular succession, as well as higher variability along the transect. *Calanus* ecostates reduced to 19% of the total ecostates, while ecostate #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* ecostates comprising 44% of the total. The other decades were sparsely sampled, but taken together, high *Calanus* ecostates comprised 46% of the total. Over the duration of the time series, the diverse *Calanus* ecostate (#7) comprised a decreasing proportion of all *Calanus* ecostates, from $\sim 40\%$ to $\sim 15\%$ ($p < 0.01$). Among the low *Calanus* ecostates, the most prevalent was the *Centropages*-dominated ecostate #2, with a consistent autumn timing throughout the dataset. Ecostate #4, a diverse ecostate 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. Ecostates #3 and #4 were mostly confined to the ends of the transect, inshore of the 100 m isobath.

Many of the ecostates had similar realized niches in PCI-temperature space (Fig. 4). The three *Calanus*-dominated ecostates had a strong mode at low temperatures—around 5 °C—spanning a range of PCI values, with ecostate #9 also extending into higher temperatures. Realized niches for ecostates #1, 4, and 6 had a similar shape. The common *Centropages* ecostate (#2) had a mode around 10 °C, but with a realized niche that spanned all temperatures and PCI values ≤ 1 . Ecostate #5 had a mode around 15 °C at PCI values ranging from 0-2, higher than any other ecostate at this temperature. Ecostate #3,

135 while comparatively rare through the time series, had a realized niche that appeared to include nearly
136 all temperature-PCI pairs, excepting the highest PCI values.

137 4. Discussion

138 There is a substantial body of work analyzing the copepod community in this CPR dataset (Jossi
139 & Goulet 1993, Greene & Pershing 2000, Conversi et al 2001, Licandro et al 2001, Piontkovski &
140 Hameed 2002, Greene et al 2003, Pershing et al 2005, Kane 2009, Record et al 2010, Johnson et al
141 2011, Greene et al 2012, Stamieszkin et al 2015). These studies have looked at the Gulf of Maine zoo-
142 plankton community through a variety of lenses, including conventional statistical analysis like corre-
143 lation (Pershing et al 2010) and principal component analysis (Pershing et al 2005), as well as commu-
144 nity properties like taxonomic richness (Record et al 2010) and size structure (Stamieszkin et al 2015).
145 These studies have identified decadal and interannual community shifts associated with climate signals
146 and other forcing, with a notable shift that lasted throughout the 1990s driven by stratification and the
147 intensification of an autumn phytoplankton bloom (Greene et al 2012). Interestingly, described changes
148 in the copepod community structure across all of these studies boil down to fluctuations in abundance
149 of *Calanus finmarchicus*, whose prominence in the Gulf of Maine has been recognized since the earliest
150 plankton surveys (Bigelow 1924). Essentially, these studies describe a bimodal system: there is either
151 a comparatively high abundance of *C. finmarchicus*, dominating the copepod community and driving
152 the size structure and species richness, or a comparatively low abundance of *C. finmarchicus*, giving
153 way to a consistent suite of smaller copepods (i.e. higher diversity and steeper size spectral slope). This
154 dichotomy encapsulates the conventional ecological analyses described above, whether they analyze
155 richness, size structure, community structure, or individual species dynamics. The analysis here repro-
156 duces this main finding of previous work. That is, there is a marked dynamic between high-*Calanus*
157 and low-*Calanus* communities, with the main signals manifesting as a shift during the 1990s as well as
158 within seasonal succession dynamics. In this sense, the DMM inference reproduced previous findings
159 in the Gulf of Maine CPR dataset: the dynamics between high-*Calanus* and low-*Calanus* communities
160 showed the same pattern.

161 The second aim of this study was to determine whether a DMM analysis can provide new insights.
162 There were a few signals that revealed new characteristics of change in the copepod community. For
163 example, one aspect of the 1980s-1990s transition that has not been described is the breakdown of a

164 consistent seasonal succession. In the 1980s, *Calanus* ecostates dominated in the winter, spring, and
165 summer, giving way to a *Centropages* ecostate in the autumn. In the 1990s, in addition to a transition
166 away from *Calanus* ecostates, there was a transition toward higher entropy—often an indication of a
167 disturbance or perturbation. The PCI data also consistently showed a regular spring bloom pattern in
168 the 1980s, across the full transect and including all years but one, followed by an erratic pattern in the
169 1990s, where blooms occurred at many different times of year (Fig. 5). This erratic pattern included
170 a new high-temperature ($\sim 15^{\circ}\text{C}$), comparatively high-PCI (≥ 1) niche that corresponded with the
171 appearance of ecostate #5. There was a partial return in the 2000s to the regularity of the seasonal
172 succession in ecostates, but the PCI patterns continued to be erratic. In seasonal environments like the
173 Gulf of Maine, species often have life cycles tuned to phytoplankton bloom periods. More erratic bloom
174 timing, as inferred from the PCI, could lead to mismatches during important life history stages (Cushing
175 1969, 1990). Haddock recruitment (Leaf & Friedland 2014), zooplankton biovolume (Friedland et al
176 2015), and *C. finmarchicus* production (Durbin et al 2003, Runge et al 2015) all appear to respond to
177 phenological changes in primary production. The strong alignment between phytoplankton phenology
178 and the ecostate patterns suggests that the influence of match-mismatch dynamics can extend to the
179 structure of the copepod community.

180 It is notable that within the *Calanus*-dominated ecostates, the diverse ecostate #7 has become pro-
181 gressively rarer. In addition to having a higher diversity among the dominant genera, this ecostate also
182 had a long tail of rare taxa. In other words, in the early part of the time series, although *Calanus*
183 dominated, that dominance was associated with a higher richness of other taxa. The return of *Calanus*
184 communities following the 1990s low did not include a return of this background richness. This tran-
185 sition poses some interesting questions. Moving northward from the Gulf of Maine, copepod com-
186 munities are increasingly *Calanus*-dominated and decreasingly diverse. It is reasonable to ask why,
187 during recent warming, the *Calanus* ecostate has shifted toward one of a more northerly character. One
188 hypothesis is that prior to the 1990s, the *Calanus* ecostates had greater richness because the copepod
189 community, including *Calanus*, was more locally supported. In contrast, warmer recent conditions may
190 favor advected *Calanus* communities but not a locally supported population. One of the major sources
191 of advection to the Gulf of Maine is across the Scotian Shelf, which has a more subarctic character,
192 and changes in currents can drive changes in the zooplankton community (Greene et al 2003, Kane
193 2007, Runge et al 2015). If the *Calanus*-dominated ecostates are more predominantly supplied from

194 the Scotian Shelf than previously, this would explain a shift toward less diverse *Calanus* ecostates. If
195 the hypothesis is true, one implication is a greater vulnerability of this key species, with an increased
196 dependence on advective supply over local production.

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

216 The question of whether a diapausing copepod like *C. finmarchicus* will continue to dominate in
217 the Gulf of Maine as temperatures warm is a key ecological question (Runge et al 2015). Through
218 lipid storage, diapausing copepods convert brief pulsed blooms of primary production into an energy
219 resource available to higher trophic levels throughout much of the year. Many fish stocks valuable to
220 Canadian and US fisheries rely on this energy resource during critical times in their life histories, as
221 do protected and endangered species. While *Calanus* appears to have unexpectedly persisted, and even
222 thrived, during this period of warming, the apparent persistence disguises a change in the background
223 diversity that could indicate a fundamental shift in the origin and fate of *Calanus* communities in the

224 Gulf of Maine. Understanding such a shift is important in the context of the more northerly *Calanus*-
225 dominated, lipid-based food webs throughout the Subarctic that are likely to follow a similar
226 transition as temperatures increase.

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

237 Acknowledgements

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

242 References

- 243 Allen, A.P., Brown, J.H., Gillooly, J.F. 2002. Global biodiversity, biochemical kinetics, and the
244 energetic-equivalence rule. *Science*. 297(5586):1545-1548.
- 245 Bigelow, H.B. 1924. Plankton of the offshore waters of the Gulf of Maine. *Bull. Bur. Fish.*, XL, Part
246 II, No. 968: 1-509.
- 247 Biswas, S., McDonald, M., Lundberg, D. S., Dangl, J. L., Jojic, V. 2015. Learning microbial interaction
248 networks from metagenomic count data. *Research in Computational Molecular Biology*, pages 32-
249 43. Springer, 2015.

- 250 Caron, D. A., Countway, P. D. 2009. Hypotheses on the role of the protistan rare biosphere in a chang-
251 ing world. *Aquatic Microbial Ecology*, 57(3), 227-238.
- 252 Chen, J., Bittinger, K., Charlson, E. S., Hoffmann, C., Lewis, J., Wu, G.D., Collman, R. G., Bush-
253 man, F. D., Li, H. 2012. Associating microbiome composition with environmental covariates using
254 generalized UniFrac distances. *Bioinformatics*, 28(16):2106–2113, 2012.
- 255 Chen, J., Li, H. 2013. Variable selection for sparse Dirichlet-multinomial regression with an application
256 to microbiome data analysis. *The Annals of Applied Statistics*, 7(1):418–442, 2013.
- 257 Conversi, A., Piontkovski, S., Hameed, S. 2001. Seasonal and interannual dynamics of *Calanus fin-*
258 *marchicus* in the Gulf of Maine (Northeastern US shelf) with reference to the North Atlantic Oscil-
259 lation. *Deep Sea Research Part II: Topical studies in Oceanography*. 48(1): 519-530.
- 260 Countway, P. D., Gast, R. J., Savai, P., Caron, D. A. 2005. Protistan diversity estimates based on 18S
261 rDNA from seawater incubations in the western North Atlantic. *Journal of Eukaryotic Microbiology*,
262 52(2), 95-106.
- 263 Cropp, R., Norbury, J. 2012. The mechanisms of coexistence and competitive exclusion in complex
264 plankton ecosystem models. *Ecosystems*. 15(2): 200-212.
- 265 Currie, D.J. 1991. Energy and large-scale patterns of animal- and plant-species richness. *The American*
266 *Naturalist*. 137(1):27-49
- 267 Cushing, D.H. 1969. The regularity of the spawning season of some fishes. *J. Conseil Permanent In-*
268 *ternational pour l'Exploration de la Mer (Cons Int Explore Mer)*. 33: 81-92.
- 269 Cushing, D.H. 1990. Plankton production and year-class strength in fish populations: an update of the
270 match/mismatch hypothesis. *Advances in Marine Biology*. 26: 249-293.
- 271 Durbin, E.G., R.G. Campbell, M.S. Casas, M.D. Ohman, B. Niehoff, J. Runge, M. Wagner. 2003.
272 Interannual variation in phytoplankton blooms and zooplankton productivity and abundance in the
273 Gulf of Maine during winter. *Marine Ecology Progress Series* 254: 81-100.
- 274 Friedland K.D., R.T. Leaf, J. Kane, D. Tommasi, R.G. Asch, N. Rebeck, R. Ji, S.I Large, C. Stock,
275 V.S. Saba. 2015. Spring bloom dynamics and zooplankton biomass on the US Northeast Continental
276 Shelf. *Continental Shelf Research* 102: 47-61.

- 277 Greene, C. H., Pershing, A. J. 2000. The response of *Calanus finmarchicus* populations to climate vari-
278 ability in the Northwest Atlantic: basin-scale forcing associated with the North Atlantic Oscillation.
279 ICES Journal of Marine Science: Journal du Conseil. 57(6): 1536-1544.
- 280 Greene, C. H., Pershing, A. J., Conversi, A., Planque, B., Hannah, C., Sameoto, D., Head, E., Smith,
281 P. C., Reid, P. C., Jossi, J., Mountain, D., Benfield, M. C., Wiebe, P. H., Durbin, E. 2003. Trans-
282 Atlantic responses of *Calanus finmarchicus* populations to basin-scale forcing associated with the
283 North Atlantic Oscillation. Progress in Oceanography. 58(2): 301-312.
- 284 Greene, C. H., Monger, B. C., McGarry, L. P., Connelly, M. D., Schnepf, N. R., Pershing, A. J., Belkin,
285 I. M., Fratantoni, P. S., Mountain, D. G., Pickart, R. S., Ji, R., Bisagni, J. J., Chen, C., Hakkinen,
286 S. M. A., Haidvogel, D. B., Wang, J., Head, E., Smith, P., Conversi, A. 2012. Recent Arctic climate
287 change and its remote forcing of Northwest Atlantic shelf ecosystems. Oceanography. 25(3):208-
288 213.
- 289 Harris, K., Parsons, T. L., Ijaz, U. Z., Lahti, L., Holmes, I., Quince, C. 2014. Linking statistical and
290 ecological theory: Hubbell's unified neutral theory of biodiversity as a hierarchical Dirichlet process.
291 Proceedings of the IEEE. PP(99): 1-14.
- 292 Holmes, I., Harris, K. Quince, C. 2012. Dirichlet multinomial mixtures: generative models for micro-
293 bial metagenomics. PLoS One, 7(2):e30126, 2012.
- 294 Hubbell S. 2001. The Unified Neutral Theory of Biodiversity and Biogeography. Princeton University
295 Press.
- 296 Huson, D. H., Richter, D. C., Mitra, S., Auch, A. F., Schuster, S. C. 2007. Methods for comparative
297 metagenomics. BMC bioinformatics, 10(Suppl 1):S12, 2009.
- 298 Johnson, C. L., Runge, J. A., Curtis, K. A., Durbin, E. G., Hare, J. A., Incze, L. S., Link, J. S., Melvin,
299 G. D., O'Brien, T. D., Van Guelpen, L. 2011. Biodiversity and ecosystem function in the Gulf of
300 Maine: pattern and role of zooplankton and pelagic nekton. PLoS One. 6(1): e16491.
- 301 Jossi, J. W., Goulet, J. R. 1993. Zooplankton trends: US north-east shelf ecosystem and adjacent regions
302 differ from north-east Atlantic and North Sea. ICES Journal of Marine Science: Journal du Conseil.
303 50(3): 303-313.

- 304 Kane, J. 2007. Zooplankton abundance trends on Georges Bank, 1977-2004. *ICES Journal of Marine*
305 *Science*, 64: 909-919.
- 306 Kane, J. 2009. A comparison of two zooplankton time series data collected in the Gulf of Maine.
307 *Journal of Plankton Research*. 31(3): 249-259.
- 308 La Rosa, P. S., Brooks, J. P., Deych, E., Boone, E. L., Edwards, D. J., Wang, Q., Sodergren, E., We-
309 instock, G. Shannon, W. D. 2012. Hypothesis testing and power calculations for taxonomic-based
310 human microbiome data. *PLoS ONE*. 7(12): e52078.
- 311 Leaf R.T., K.D. Friedland. 2014. Autumn bloom phenology and magnitude influence haddock recruit-
312 ment on George's Bank. *ICES Journal of Marine Science* 71: 2017-2025.
- 313 Licandro, P., Conversi, A., Ibanez, F., Jossi, J. 2001. Time series analysis of interrupted long-term data
314 set (1961-1991) of zooplankton abundance in Gulf of Maine (northern Atlantic, USA). *Oceanologica*
315 *Acta*, 24(5), 453-466.
- 316 Lozupone, C., Knight, R. 2005. Unifrac: a new phylogenetic method for comparing microbial commu-
317 nities. 2005. *Applied and environmental microbiology*, 71(12):8228-8235, 2005.
- 318 Magurran, A.E., 2013. *Measuring biological diversity*. John Wiley & Sons.
- 319 McGrady-Steed, J., Harris, P. M., Morin, P. J. 1997. Biodiversity regulates ecosystem predictability.
320 *Nature* 390:162-165
- 321 Mills, K. E., Pershing, A. J., Brown, C. J., Chen, Y., Chiang, F. S., Holland, D. S., Lehuta, S., Nye, J.A.,
322 Sun, J.C., Thomas, A.C., Wahle, R. A. 2013. Fisheries management in a changing climate lessons
323 from the 2012 ocean heat wave in the Northwest Atlantic. *Oceanography*, 26(2), 191-195.
- 324 O'Brien, J.D., Lin, K., MacEachern, S. 2015. Mixture model of pottery distributions from Lake Chad
325 Basin archaeological sites reveals ancient segregation patterns. arXiv preprint arXiv:1511.05185.
- 326 O'Brien, J. D., Record, N. R., Countway, P. D. 2016. The power and pitfalls of Dirichlet-multinomial
327 mixture models for ecological count data. bioRxiv:10.1101/045468.

- Parent, G.J., Plourde, S., Turgeon, J., 2012. Natural hybridization between *Calanus finmarchicus* and *C. glacialis* (Copepoda) in the Arctic and Northwest Atlantic. *Limnology and oceanography*, 57(4), pp.1057-1066.
- Pershing, A. J., Greene, C. H., Jossi, J. W., O'Brien, L., Brodziak, J. K., Bailey, B. A. 2005. Interdecadal variability in the Gulf of Maine zooplankton community, with potential impacts on fish recruitment. *ICES Journal of Marine Science: Journal du Conseil*. 62(7), 1511-1523.
- Pershing, A. J., Head, E. H., Greene, C. H., Jossi, J. W. 2010. Pattern and scale of variability among Northwest Atlantic Shelf plankton communities. *Journal of Plankton Research*. fbq058.
- Pershing, A. J., Alexander, M. A., Hernandez, C. M., Kerr, L. A., Le Bris, A., Mills, K. E., Nye, J.A., Record, N.R., Scannell, H.A., Scott, J.D., Sherwood, G. D. 2015. Slow adaptation in the face of rapid warming leads to collapse of the Gulf of Maine cod fishery. *Science*, 350(6262), 809-812.
- Peterson, G., Allen, C.R., Holling, C.S. 1998. Ecological resilience, biodiversity, and scale. *Ecosystems*. 1(1):6-18
- Piontkovski, S., Hameed, S. 2002. Precursors of copepod abundance in the Gulf of Maine in atmospheric centers of action and sea surface temperature. *The Global atmosphere and ocean system*, 8(4), 283-291.
- Purvis, A., Hector, A., 2000. Getting the measure of biodiversity. *Nature*, 405(6783), pp.212-219.
- Record, N.R., Pershing, A.J., Jossi, J.W. 2010. Biodiversity as a dynamic variable in the Gulf of Maine continuous plankton recorder transect. *J. Plank. Res.* 32(12): 1675-1684.
- Record, N. R., Pershing, A. J., Maps, F. 2013. The paradox of the "paradox of the plankton". *ICES Journal of Marine Science: Journal du Conseil*, fst049.
- Reynolds, R. W., Smith, T. M., Liu, C., Chelton, D. B., Casey, K. S., Schlax, M. G. 2007. Daily high-resolution-blended analyses for sea surface temperature. *Journal of Climate*, 20, 5473-5496.
- Richardson, A.J., Walne, A.W., John, A.W.G., Jonas, T.D., Lindley, J.A., Sims, D.W., Stevens, D., Witt, M. 2006. Using continuous plankton recorder data. *Progress in Oceanography*. 68(1): 27-74.

- 353 Riley, G.A., Stommel, H., Bumpus, D.F. 1949. Quantitative ecology of the plankton of the western
354 North Atlantic. Bulletin of the Bingham Oceanography Collection. 12(3):1-169.
- 355 Runge, J.A., Ji, R., Thompson, C., Record, N.R., Maps, F., Chen, C., Salisbury, J., Vandemark, D. 2015.
356 Persistence of *Calanus finmarchicus* in the western Gulf of Maine during recent extreme warming.
357 J. Plank. Res. doi:10.1093/plankt/fbu098
- 358 Shannon, C.E. 1948. A mathematical theory of communication. Bell System Technical Journal.
359 27:379-423 and 623-656
- 360 Stamieszkin, K., Pershing, A. J., Record, N. R., Pilskaln, C. H., Dam, H. G., Feinberg, L. R. 2015.
361 Size as the master trait in modeled copepod fecal pellet carbon flux. Limnology and Oceanography.
362 60(6): 2090-2107.
- 363 White, J. R., Nagarajan, N., Pop, M. 2009. Statistical methods for detecting differentially abundant
364 features in clinical metagenomic samples. PLoS computational biology, 5(4):e1000352, 2009.

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).

Draft