Online Supplementary Material for

A Trait-Based Clustering for Phytoplankton Biomass Modeling and Prediction

Crispin M. Mutshinda[1,*], Zoe V. Finkel[2], Claire E. Widdicombe[3] and Andrew J. Irwin[1]

[1]Department of Mathematics and Statistics, Dalhousie University, Halifax, NS B3H 4R2, Canada
[2]Department of Oceanography, Dalhousie University, Halifax, NS B3H 4R2, Canada
[3]Plymouth Marine Laboratory, Prospect Place, Plymouth, PL1 3DH, UK
*Correspondence: crispin.mutshinda@dal.ca

1. Synopsis

These supplementary notes present the R code used in the paper "A Trait-Based Clustering for Phytoplankton Biomass Modeling and Prediction" to generate a trait-based clustering of phytoplankton as an alternative to the prevalent functional type categorization for trait value characterization and biomass prediction.

The data consist of weekly biomass for 74 species (57 diatoms and 17 dinoflagellates) at Station L4 in the Western English Channel, UK between April 2003 and December 2009, and coincident measurements of environmental variables describing water conditions and resource availability.

The species-specific occurrence trait values estimated from the Bayesian logistic model linking species presence-absence to the environmental variables (equations 5-6 in the paper) and serving as basis for the trait-based clustering reported in the paper are provided at the end of this document in the 74×6 data frame "betaPresAbs". Therein, columns 1 through 6 represent respectively the irradiance, temperature, salinity, nitrogen, silicate, and phosphate effects. The first 57 species (rows 1-57) are diatoms, whereas the last 17 species (rows 58-74) are dinoflagellates.

It worth emphasizing that the focus here is on the trait-based clustering procedure and the computation of cluster biomass time series under the soft clustering resulting from the Gaussian mixture model (GMM) adopted in the study. The Bayesian model of group-level biomass dynamics (equations 1-4 in the paper) and the Bayesian logistic model linking species presence-absence to potential environmental drivers (equations 5-6 in the paper) were easily fitted to data by MCMC simulation through OpenBUGS [2]. The biomass dynamics model applies to group-level biomass data, irrespective of the clustering approach. As a result, the same model applies to all three clustering schemes considered here namely, biomass aggregated according to the trait-based clusters, biomass aggregated by functional types, and total biomass.

2. R code for performing the trait-based clustering and computing cluster biomass time series

2.1. Principal component Analysis (PCA) of the matrix betaPresAbs of occurrence trait value estimates

```
pc<-prcomp(betaPresAbs, scale=F)
# scale=F since trait estimates are divided by standard deviations
gmDat=pc$x[,1:3] # Species scores on PC1, PC2, and PC3
```

2.2 Fitting the GMM to species scores on the first 3 PCs

The GMM is fitted using the EM algorithm implemented in R by the function `mvnormalmixEM` from package `mixtools`. Since the GMM is a proper probability distribution, we selected the number of components by maximizing the likelihood of the data.

```
library(mixtools)
gmModel<-mvnormalmixEM(gmDat, k=3)
```

```
# Cluster responsibilities(p) and mixing weights (lambda)
p=round(gmModel$posterior,2)
gmModel$lambda
```

```
# Plotting the log-likelihood
plot(gmModel$all.loglik[3:65], type="l", lwd=2, col=" salmon1",
xlab="EM iteration number", ylab="log-likelihood", axes=F,
xlim=c(1,40))
axis(1, at=c(1,10,20,30,40), labels= c(1,10,20,30,40))
axis(2)
box()
```

Since each iteration of the EM increases the likelihood, the plot of the log-likelihood against the iteration number helps decide whether the algorithm has converged when the plot reaches a plateau and does not increase further. For our analysis, the EM algorithm converged after about 15 iterations (Fig. S1).

```
# Maximum a posteriori clustering solution
clust<-dput(as.numeric (apply(gmModel$posterior, 1, function(row)
which.max(row))))

FunctType=c(rep("diatom", 57)), rep("dinoflagellate",17))

sppNames=rownames(betaPresAbs)

resClust=data.frame(clust, sppNames, FunctType)
```

2.3. Computing cluster biomass time series under the trait-based clustering

In the following code, the, the 74 columns of the species-level biomass dataset Bio74 (not shown for copyright restrictions) are the weekly biomass time series of the 74 species in the same order as the rows of betaPresAbs (57 first columns for diatoms and 17 last columns for dinoflagellates)

```
# Biomass for cluster1
 C1=Bio74
 for(j in 1:ncol(Bio74)){
  C1[,j]<-C1[,j]*p[j,1]
 }
 G1=rowSums(C1)
 Biomass_G1=G1+min(G1[G1!=0]/2)

# Biomass for cluster2
 C2=bio74
 for(j in 1:ncol(Bio74)){
  C2[,j]<-C2[,j]*p[j,2]
 }
 G2=rowSums(C2)
 Biomass_G2=G2+min(G2[G2!=0]/2)


# Biomass for cluster3
 C3=bio74
 for(j in 1:ncol(Bio74)){
  C3[,j]<-C3[,j]*p[j,3]
 }

 G3=rowSums(C3)

 Biomass_G3=G3+min(G3[G3!=0]/2)

# Total biomass
 totBiomass=apply(cbind(Biomass_G1, Biomass_G2, Biomass_G3),1,sum))
```

The cluster biomass data and the total biomass are to be log-transformed for use in the Bayesian model of cluster biomass dynamics. The model is the same under the any clustering scheme. In our study, we considered three clustering schemes namely, the trait-based clustering, functional type and total biomass, where the latter aggregate the biomass of all species into a single time series.

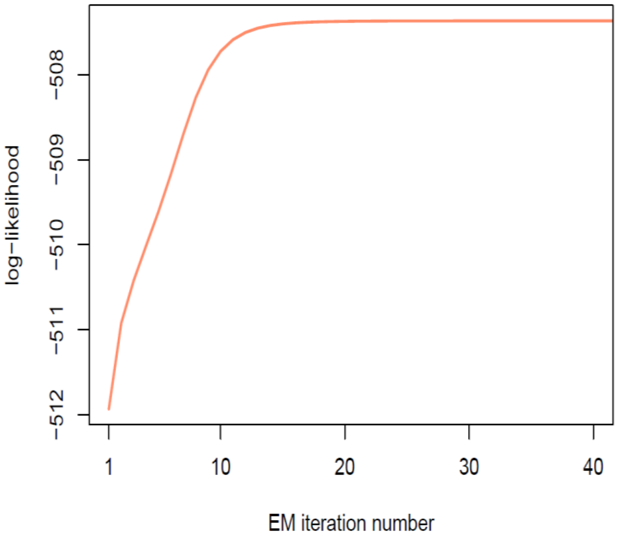

Figure S1. *Log-likelihood of the EM algorithm against the EM iteration number. The log-likelihood reached a plateau after approximately 15 iterations indicating convergence of the EM algorithm*

3. Occurrence trait values for the 74 species under consideration

```
betaPresAbs=structure(.Data=c(-1.7804, -3.3985, -0.9553, -3.3017, -4.6143, -0.913, -
3.9706, 0.2541, -3.9386, -0.8467, -3.3948, -0.8613, -5.216, 1.2042, -1.5271, -1.0645, -
3.0575, -5.0847, -3.8777, -3.5276, -2.6594, -3.4337, -4.0988, -2.3227, -1.761, -1.9429,
-1.7882, -4.4765, -0.1549, -2.7757, -2.8001, -0.254, -5.6402, -2.9366, -5.2134, -2.3857,
-3.3872, -0.4646, -3.2497, -2.7591, -1.4039, 0.9388, -1.0245, -2.3736, -1.7159, -4.024,
-5.1915, -0.2813, -5.6714, -1.2517, -4.2411, -3.1646, -3.5243, -1.3897, -0.9836, -
0.2551, -4.282, 1.0114, 2.3744, 0.2159, 2.6663, 2.7818, 4.6187, 1.8405, 1.4967, 3.1075,
3.4975, 1.9648, -1.6284, 3.3773, -0.6049, -2.164, 3.8184, -0.3598, 0.7395, -1.5339,
0.9599, -3.3392, -3.5744, 0.6017, -2.5201, 3.1398, -0.7818, 0.7403, -1.2601, -1.6011, -
3.005, -1.9843, 3.7159, -3.4616, 2.4238, -0.3936, 0.456, -3.6889, -2.6073, -1.6753, -
0.3039, 1.6336, -4.1608, -1.1272, -4.6249, -0.6035, -3.5953, -4.1946, -1.7987, 1.0409,
-4.7419, -0.3685, -0.552, 0.3437, -0.0851, 1.8895, -1.2578, -2.045, 1.9628, 2.6537, -
1.9244, -3.5454, -3.673, 0.1594, -3.2286, -0.6222, -3.2884, -2.4181, -0.2293, 2.3696,
1.5711, -1.7666, -2.2756, 0.1544, -3.2906, 0.9742, 3.3775, 1.4281, 4.8336, 0.4363,
3.9934, 1.0711, 0.331, 1.662, 3.3096, -0.6149, -1.2966, 1.7187, 3.5392, 2.7204, 4.0562,
0.3868, 1.1353, 0.4253, 2.1873, -0.8954, 2.3027, 1.817, 0.4272, -0.1727, -0.3549, -
0.435, 0.3848, -0.453, 1.2762, 0.3788, -0.7983, 1.4031, 1.5708, 0.9315, 0.7879, -0.0226,
4.4986, -2.108, 0.4411, -0.7583, 4.163, 0.6375, 1.4684, -1.0362, -0.1055, -0.1408,
1.8273, -0.5575, -2.1447, 1.2632, 1.6627, 1.819, 0.1754, 0.0251, 1.0962, -1.7865, -
2.9985, 2.3285, -2.1464, 1.7101, 2.2795, -1.7103, -0.911, -2.6259, 0.1163, 1.9046,
1.5747, 2.5197, -4.7137, 0.0149, 1.6448, 1.7346, -1.2939, 1.161, -2.8731, 1.0056, -
0.2309, 3.2278, 1.1206, 0.2882, 1.0474, 0.0662, 3.0801, 0.0707, -2.8912, 1.7626, -
1.1496, -1.6043, -1.1104, -0.5768, 4.8278, -5.0819, 0.3675, 5.2247, -6.42, 4.1182, -
7.3498, -1.6833, 5.4448, -2.1967, -1.3697, -2.5, 1.6821, -4.6621, -3.2013, -0.4862, -
4.4853, -1.9304, -2.0543, -3.7109, -0.0487, -3.8126, 1.3937, -1.9217, -2.4488, 0.2698,
-0.2659, -0.9579, 1.7902, -3.6302, -1.7792, -3.5458, -3.6109, -2.3181, 1.6374, 0.5882,
2.5113, 1.515, 3.1349, 0.0195, 0.9058, 2.1182, -0.3746, -2.8558, 0.3578, -2.5913, -
2.9484, -4.1801, -1.3984, -1.5187, -4.3928, 3.8878, -2.319, -2.6252, 0.3831, 3.8862, -
3.6841, -1.6164, -1.4201, 1.2517, -0.0614, -3.8044, -0.159, -2.2885, -1.5097, -2.0067,
-0.0886, -2.7, -2.2762, -2.4815, -1.672, 0.0901, -1.0598, -2.3694, 1.0326, -0.0317, -
0.0865, -0.9288, -0.9044, -1.2203, -0.9371, -0.8732, -2.4214, -2.1623, -1.3172, -2.1758,
0.3228, 2.5705, -0.4122, -0.554, 0.3453, -2.1714, 0.3394, 1.4702, 1.6039, 1.3111, -
1.4815, -0.4735, -2.376, -1.3784, 0.2193, -0.8861, -2.097, 0.6051, -2.2315, 0.6377, -
2.2749, 0.1472, 0.5074, -1.9099, 1.4767, 1.3341, -1.8107, 2.2175, -1.0907, -1.8925, -
2.378, 1.8482, -0.4944, 1.0401, -0.6884, -2.659, 1.332, 1.0247, -0.6522, -2.8289, 2.3857,
1.5086, -0.029, -1.8034, 1.4775, 1.614, -1.2942, 2.2604, 2.4554, 2.2381, 2.2747, -
1.2557, -1.3189, -0.3369, 2.0862, 0.6612, 0.2964, 1.5578, 0.9364, -0.0879, 0.6671,
0.3726, -2.7634, -1.3228, -2.836, -2.9884, 0.9977, -0.6052, 2.2188, -1.8147, -3.1825, -
0.2542, 0.0371, -2.7301, -1.3749, -3.5535, -2.1568, 0.2325, 4.2499, 0.366, -0.5492, -
1.5655, 1.3613, -2.3621, 0.2167, -3.3304, -2.0645, -1.1004, -0.7491, 2.0104, -3.5807, -
3.6936, -3.1603, -2.8523, -4.6696, -1.1794, 1.937, -1.4851, -0.8839, 1.1368, -0.5223, -
0.3726, 2.6465, 0.8076, -0.775, -2.3085, -0.8969, 1.4361, -1.951, 1.2919, 3.7678, 3.7227,
2.0886, 1.0931, 3.0334, 0.8949, 1.2544, 0.3863, -1.1673, 1.3015, 2.608, 0.3413, 1.4305,
0.0825, 0.8561, -0.5248, 1.7528, -0.4091, 1.0045, -1.8824, -1.3967, -0.575, 1.3541, -
1.8486, 0.8975, -0.8737), .Dim = c(74, 6), .Dimnames = list(c("N.clost", "G.delic", "P-
n.delic", "P.sulcata", "M.membr", "Pleurosigma", "C.pelagica", "P-n.seriata", "D.cabro",
"L.annulata", "Thalass.10µm", "Ch.debilis", "N.distans", "P.alata.5µm", "L.danicus",
"C.danicus", "E.zodiacus", "T.nitzsch", "G.striata", "G.flaccida", "N.sigmoidea",
"D.fragil", "R.tesselata", "R.setigera.5µm", "C.densus", "Navicula.sp.",
"C.criophilum", "D.brightwel", "S.costatum", "R.imbric.5µm", "R.imbric.15µm",
"L.minimus", "T.rotula", "C.affinis", "O.mobil", "C.decipiens", "Pennate.50µm",
"R.imbric.10µm", "P.stelligera", "P.plancton", "Thalass.20µm", "Thalass.4µm",
"C.simplex", "R.stylif", "B.paradoxa", "R.setigera.25µm", "P-n.pungens", "C.socialis",
"T.punctigera", "Small.Pennate", "P.truncata", "Pennate.30µm", "L.mediterr", "P.alata",
"C.radiatus", "P.pandurif", "D.pumila", "C.fusus", "C.horridum", "C.lineatum",
"C.tripos", "D.acuminata", "K.mikimotoi", "G.spinifera", "Gymnod.sp", "G.cf.pygmaeum",
"M.perforatus", "Micranthodinium.sp.", "P.balticum", "P.micans", "Ptrum.minimum",
"P.triestinum", "S.trochoidea", "Scrip.sp.cyst"), c("PAR", "Temperature", "Salinity",
"Nitrogen", "Silicate", "Phosphate")))
```

References

Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data *via* the EM algorithm. *J. Roy. Stat. Soc. B* **1977**, 39, 1–38.

Thomas, A.; O'Hara, R.B.; Ligges, U.; Sturtz, S. Making BUGS open. *R News* **2006,** 6, 12–17.