Additional Results, Simulations and Data Analysis Details

A supplement to: Detecting and modeling changes in a time series of proportions

Thomas J. Fisher and Jing Zhang
Department of Statistics, Miami University
and
Stephen Colgate*
Department of Biostatistics, University of Cincinnati
and Michael J. Vanni
Department of Biology, Miami University

January 31, 2021

Abstract

This document provides some additional results to the manuscript, "Detecting and modeling changes in a time series of proportions." This includes a discussion (with links) of phytoplankton in the news, additional discussion on the model and implementation, simulations and details from the data analysis section of the main paper. The supplement also describes the three change point hypothesis tests implemented in the main article for comparison to the proposed methods.

1 Phytoplankton in Society

Phytoplankton are microscopic, autotrophic organisms found in oceans, seas and freshwater basin ecosystems. Phytoplankton are a key component of the food web as they transform energy via photosynthesis from sunlight to organic matter. That provides food for herbivores and indirectly for organisms at the top of the food chain. However certain phytoplankton can be dangerous. In 2014, a bloom of toxic cyanobacteria (or blue-green

*Stephen Colegate was a Master's student at Miami University during his contribution to this work

algae) contaminated Lake Erie near Toledo, Ohio, USA shutting down the city's supply of drinking water and forcing roughly 500,000 residents to use bottle water. A news article describing this event can be found here:

`https://www.nytimes.com/2014/08/04/us/toledo-faces-second-day-of-water-ban.html`

Similarly, a bloom of diatoms in the Pacific northwest affected fisheries in 2015. Diatoms can produce domoic acid in shellfish and other small marine animals, and can be toxic for larger invertebrates. Although rare in humans, Amnesic Shellfish Poisoning caused by domoic acid, can be life threatening. This algae bloom was in the news in 2015:

`https://oceanservice.noaa.gov/news/sep15/westcoast-habs.html`

2 Modifications to the Proposed Model

In the primary paper, we propose a hidden Markov model (HMM) where the state of the Markov chain controls the coefficients in a set of generalized linear models that model the location and scale of a Dirichlet distributed random variable. Mathematically, we assume the observed p-dimensional composition at time t, $\mathbf{Y}_t$, comes from a Dirichlet distribution where the shape parameter is $\boldsymbol{\alpha} = \tau\boldsymbol{\theta}$ with $\boldsymbol{\theta}$ as the location of $\mathbf{Y}_t$, i.e., $E[\mathbf{Y}_t] = \boldsymbol{\theta}$ and τ is a term that influences the scale, variance or dispersion; note $\boldsymbol{\theta}'\mathbf{1}_p = 1$, where $\mathbf{1}_p$ is a p-dimensional vector of ones and $'$ is the transpose operation. Further, we allow the location parameter to be modeled by

$$\boldsymbol{\theta} = \boldsymbol{\eta}/(\boldsymbol{\eta}'\mathbf{1}_p), \quad \text{where} \quad \log(\eta_i) = \beta_{i0} + \beta_{i1}X_1 + \beta_{i2}X_2 + \ldots + \beta_{ik}X_k, \tag{1}$$

where X_j, $j = 1, \ldots, k$ are available predictor variables with β_{ij} as the coefficient on the j^{th} covariate for component i. The scale parameter can be further modeled by

$$\log(\tau) = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \ldots + \gamma_k X_k, \tag{2}$$

where the γ_j terms are the coefficients on the j^{th} predictor.

We further assume that S_t, a latent *state* of the t^{th} observation, where $S_t \in \{1, 2, \ldots, m\}$

influences the distribution of the response; namely,

$$P\left(\mathbf{Y}_t \in A | S_1, \ldots, S_t = s_t\right) = P\left(\mathbf{Y}_t \in A | S_t = s_t\right).$$

Further assume the process controlling the states S_t satisfy the Markov property

$$P\left(S_{t+1} = j | S_t = i, S_{t-1}, S_{t-2}, \ldots, S_1\right) = P\left(S_{t+1} = j | S_t = i\right) = p_{ij}. \tag{3}$$

That is, the probability distribution at time t only depends on the underlying state at time t. This setup is an m state hidden Markov model with a Dirichlet response. Using the Dirichlet formulation above, the latent Markov process effectively determines the regression coefficients in (1) and (2). Following Chib [1998] we add a constraint to the transition matrix such that a Markov chain in state i can only jump to state $i+1$ or remain in state i at the next transition; i.e., $p_{ij} = 0$ for all $j \neq i, i+1$. The Viterbi algorithm [Cappé et al., 2005] is a process to find the most likely sequence of hidden states that result in a sequence of observations. The Viterbi states, or paths, from the estimated HMM can be used to determine if a change point occurred – a change in Viterbi state indicates a change in the observed distribution.

Several natural modifications exists for this model, we outline some of them here.

2.1 Shape Parameter Model

For one, the model could be modified so the shape of the Dirichlet distribution, $\boldsymbol{\alpha}$, is directly modeled rather than decomposing into a location and scale parameter. For example, one could use a log-link function; i.e.,

$$\log(\alpha_i) = \beta_{i0} + \beta_{i1} X_1 + \ldots + \beta_{ik} X_k$$

for predictor variable X_j, $j = 1, \ldots, k$. This variant is effectively a standard Dirichlet regression with a hierarchical hidden Markov process controlling the coefficients.

2.2 Location or Scale only Model

The proposed model can be further modified by considering only changes in the location or the scale. Specifically, we can construct a model where a HMM controls the coefficients in (1) while the coefficients in (2) are constant for all states in the HMM; we call this the *Location HMM* below. Likewise, the HMM can control the values in (2) while the values in (1) are constant across all states. This model is known as the *Scale HMM* below. These two models can be considered special cases of the more general model proposed in the primary manuscript, called the *Location & Scale HMM*.

3 Bayesian Implementation

We implement our proposed model in the Bayesian framework using the `rstan` package [Stan Development Team, 2018] in R. Source code of the `.stan` and `.R` files are available here: `https://github.com/tjfisher19/hmmDirichletModel`.

In practice fitting models using computational Bayesian methods requires some tuning of the Markov chain Monte Carlo (MCMC) method. In particular, convergence to the steady state distribution of the MCMC chain is critical for any inference. Some tuning of the MCMC parameters (e.g., number of iterations or initial values) and prior values may be required. In our simulation study, we need a more uniform approach as to compare methods (tuning of the MCMC algorithm for each simulated dataset is infeasible). Unless otherwise specified, for each dataset explored in the simulation study we fit two MCMC chains based on 100,000 iterations (50,000 burn-in iterations with thinning by 200) of the no-U-turn sampler (NUTS) algorithm in `rstan`, resulting in a sample from the posterior distribution of size 500. These values were chosen as to make the computation and simulation results feasible (posterior samples for 200 synthetic dataset grow quickly and create memory management issues). In the analysis of the phytoplankton data, the MCMC specific was similar but we thinned by 100 so the resulting posterior sample was of size 1000.

Bayesian inference, of course, requires specification of prior distributions on all model parameters. In our setting, the parameters of interest include all β_{ij}-terms in (1), γ_j-terms in (2) and the transition probabilities, p_{ij} in (3). We use a $N(0, \sigma = 2)$ prior on all model

coefficients (β_{ij} and γ_j) but the supplied `stan` code allows for a user to easily specify the standard deviation. Although this prior is subjective we argue it is fairly non-informative in the sense that our coefficients are on an exponential scale.

To explain, in the case of no covariate information present and working with the parameter controlling the location, $\boldsymbol{\eta}$, $\eta_i = \exp\{\beta_{i0}\}$, for $i = 1, \ldots, p$. By the empirical rule one would expect β_{i0} to take on values in $(-6, 6)$ concentrated near zero a priori, which corresponds to η_i values in the interval $(0.00247, 403.4288)$, a substantially large range. Furthermore, the mean, median and mode of the prior distribution is the value zero, which corresponds to $\eta_i = 1$ for all $i = 1, \ldots, p$; thus, a priori, the distribution of location of $\mathbf{Y}_t$ is a p-dimensional uniform distribution, on average.

The values of γ_j follow the same prior distribution in our setting. Again we highlight that the model is based on logarithmic-link function and that although this prior is subjective, it is not particularly informative. In our experience, typical β_{ij} and γ_j values tend to fall within this prior distribution. Of course, this specification can easily be modified in practice if a practitioner has subjective matter expertise.

The transition probabilities, p_{ii} and $p_{i(i+1)}$, must sum to one and estimation may require some tuning. A too restrictive prior on p_{ii} may prevent the HMM from ever changing states. Likewise, something too flexible may encourage non-existent jumps, isolating single outliers, or create convergence issues. In our setting we utilize a $Beta(9.5, 0.5)$ prior for p_{ii}. The expectation of p_{ii}, a priori, is 0.95 with a standard deviation of approximately 0.066. This results in a HMM that generally is hesitant to jump from state i to state $i + 1$ unless there is conclusive evidence. This provides a posterior that allows us to properly assess whether a jump occurred.

With these MCMC settings and priors, we achieved MCMC convergence in our application and in most simulated datasets.

4 Comparative Methods: Hypothesis Testing

In the manuscript we compare the proposed methods to three different change point hypothesis testing methods, here we describe the implementation of each.

4.1 Nonparametric Tests

We utilize two multivariate nonparametric tests in the manuscript. One is the test of Holmes et al. [2013] implemented in the `npcp` package in R [see Kojadinovic, 2020], denoted as *Nonparametric Test* in the main manuscript. The test uses a Cramér-von Mises statistic to detect a shift in distribution of a continuous multivariate distribution.

The other is the hierarchical clustering approach in Matteson and James [2014], implemented in the `ecp` package [James and Matteson, 2014], denoted *ecp Test*, which anticipates p-dimensional data in $\mathbb{R}^d$.

These two test are applied to transformed Dirichlet data (transformed from the simplex of dimension 4 to the trivariate real number system). Specifically, we transform $\mathbf{Y}_t = (Y_1, Y_2, Y_3, Y_4)'_t$ from the 4-dimensional simplex to 3-dimensional continuous space via an additive log ratio (alr):

$$Z_i = \log\left(\frac{Y_i}{Y_4}\right) \quad \text{for } i = 1, 2, 3.$$

For data that was generated to be seasonal (or in the case of the phytoplankton data), we then calculate the mean vector response of the the multivariate continuous data $\mathbf{Z}_t = (Z_1, Z_2, Z_3)'_t$ for each of the three seasons, Spring, Summer and Fall. All $\mathbf{Z}_t$ terms are demeaned by the appropriate seasonal average (each $\mathbf{Z}_t$ in the spring season is differenced with $\bar{\mathbf{Z}}_{spring}$). The resulting sequence is a 3-dimensional zero mean series, call it $\mathbf{Z}^*_t$, with the same length as the original series. If a change exists in the underlying distribution of the original compositional data $\mathbf{Y}_t$ (say, in the shape parameter $\boldsymbol{\alpha}$), it should be present in the distribution of new sequence $\mathbf{Z}^*_t$.

We then apply the nonparametric change point test in the `cpDist` function in the `npcp` package in R. Note that this particular function includes a bandwidth parameter labeled `b` that controls the level of serial correlation present in the data. We chose to implement this function using both `b=1`, which assumes independence of the observations, and `b=NULL` where an optimal value of `b` is found and can handle some level of serial correlation. Although one may argue the `b=NULL` is more appropriate and robust, we chose to report the `b=1` results in the main manuscript as the reported powers are greater and compares more readily to our method (the Dirichlet regression model ignores potential serial correlation).

We also implement the moving-window algorithm in Prabuchandran et al. [2019] to detect multiple change points where the `cpDist` test is the underlying hypothesis testing. This test is applied to look for multiple change points in either our simulated data or real data, although the original test was designed for the single change point problem.

We also apply the test in the `e.divisive` function in the `ecp` package to our transformed data. Given our small-to-moderate sample size, we specified `min.size=18` as the minimum number of observations between any change points. This method is designed to test for multiple change points and does not require additional modifications.

4.2 Dirichlet Likelihood Permutation test

We implement a modification to the test of Prabuchandran et al. [2019]. There, the observed data $\mathbf{Y}_t$ are assumed to be from the p-dimensional Dirichlet distribution with shape parameter $\boldsymbol{\alpha}$. The log-likelihood under the null hypothesis is found, L_0. Under the alternative the observations are segmented into two regimes (time points $1, \ldots, c$ are a regime, and points $c+1, \ldots, n$ are another) and the log-likelihood is found (a different $\boldsymbol{\alpha}$ is found for each regime), call it $L_a^{(c)}$. The maximum difference between the log-likelihoods is found as a test statistic:

$$\max_{1<c<n} \left\{ L_a^{(c)} - L_0 \right\} \quad \text{with change point} \quad \hat{c} = \arg\max_{1<c<n} \left\{ L_a^{(c)} - L_0 \right\} \tag{4}$$

The distribution of the statistic is found through permutation methods.

The test of Prabuchandran et al. [2019] does not directly apply to our scenario as the phytoplankton compositions are clearly seasonal. We modify the log-likelihood approach by estimating the $\boldsymbol{\alpha}$ within each season. Thus, under the null hypothesis a different $\boldsymbol{\alpha}$ is found for the spring, for the summer and for the fall (thus 3×4 parameters are estimated). We do the same under the alternative hypothesis; thus $2 \times 3 \times 4$ parameters are found. To make finding a likelihood tractable (several observations are needed to estimate $\boldsymbol{\alpha}$), we only consider potential changes in time points 12 to 51 when 63 observations are available, thus guaranteeing a minimum of 4 years in each regime. The location, $\hat{c}$, of the maximum difference in likelihoods is found (4) as the perspective change point.

Approximating the distribution of statistic is found through a random permutation approach. We implement a modified version of that in Prabuchandran et al. [2019]. To retain the seasonality in the response, permutation is performed within season (that is, *spring* measurements are permuted with other *spring* observations). The location of maximum difference is found using the permuted data and the process is repeated a large number of times (we use 1000 permuted samples in our implementation). The approximate p-value of the original statistic is found based on the random permutation.

We also implement this modified test with the moving-window algorithm presented in Prabuchandran et al. [2019] to detect multiple change points in a given series.

5 Simulation Studies

We now present some additional simulation studies highlighting our proposed model. We retain all the notation and methods described in the primary manuscript. The results herein are meant to supplement that article, and not constitute a standalone document.

5.1 Changes in Location & Scale

Our first additional study is designed to demonstrate that by separately modeling the location and scale of the Dirichlet distribution we may be able to detect more change points than the other approaches.

A sample of length $n = 63$ is generated under two regimes, of length 30 and 33, respectively. Data in the first regime is generated from a Dirichlet distribution with location $\boldsymbol{\theta}^{(1)} = (0.3, 0.3, 0.2, 0.2)'$ and scale $\tau^{(1)} = 6$. Data in the second regime has location $\boldsymbol{\theta}^{(2)} = (0.25, 0.25, 0.25, 0.25)'$ with scale $\tau^{(2)} = 4$. Note there is a change in both the location and the scale but the shape of each Dirichlet distributions are $\boldsymbol{\alpha}^{(1)} = (1.8, 1.8, 1.2, 1.2)'$ and $\boldsymbol{\alpha}^{(2)} = (1, 1, 1, 1)'$, respectively. This process is repeated 200 time creating replicates for our study.

We fit the proposed HMM modeling the Location & Scale terms separately to the 200 datasets. For comparison, we also perform the studied approaches from the literature. Table 1 reports the results of the study including the proportion of time the MCMC algo-

rithm converged when fitting the HMM using the parameters in Section 3 along with the proportion of times a change point was detected in 100 randomly selected *converged* fits. We see the Location & Scale HMM reports the highest number of change points detected.

Table 1: Proportion of models where the MCMC algorithm converged and proportion of times a change point was detected for each model fit/method when a difficult-to-detect change point exists.

Model	MCMC Converged	Change Point Detected
Location & Scale HMM	0.56	0.48
Nonparametric Test	—	0.21
ecp Test	—	0.27
Likelihood Permutation	—	0.37

5.2 Simulation Parameters in Manuscript

The simulation studies in the manuscript have been structured to reasonably mimic the motivating phytoplankton dataset. As such, our synthetic data was of length $n = 63$ and had a similar seasonality as the observed phytoplankton taxa. The parameter values were chosen based on maximum likelihood estimation values from the motivating phytoplankton data. For example, in the *No Change Data* in the primary manuscript, and used below, the parameters (Table 2) for the seasonal Dirichlet response variables are the MLE of the phytoplankton when estimating effects within season.

Table 2: Parameter values used for simulated datasets when no change point is present.

MLEs for Phytoplankton Data

Season	θ_1	θ_2	θ_3	θ_4	τ
Spring	0.14	0.21	0.46	0.18	7
Summer	0.65	0.11	0.15	0.08	11
Fall	0.55	0.14	0.22	0.09	11

The parameter values used for simulations when a change point is present are also based on the motivating phytoplankton datasets. When there is a change in location, we calculated the MLEs for the first seven years of data and set that as $\boldsymbol{\theta}^{(1)}$ in Table 3. For the second regime we used the MLE values from data over the last seven years as seen in Table 3. Similarly, for the scale, $\tau^{(s)}$, values in Table 3 are based on ± 6 of the MLE values in Table 2.

Table 3: Parameter values used for simulated datasets when a change point occurred.

Season	State (s)	$\theta_1^{(s)}$	$\theta_2^{(s)}$	$\theta_3^{(s)}$	$\theta_4^{(s)}$	$\tau^{(s)}$
Spring	1	0.10	0.12	0.40	0.38	13
	2	0.24	0.22	0.47	0.07	1
Summer	1	0.51	0.15	0.16	0.18	5
	2	0.70	0.11	0.16	0.04	17
Fall	1	0.46	0.11	0.23	0.20	5
	2	0.60	0.15	0.21	0.04	17

And lastly, in our simulation study involving a covariate influence on a Dirichlet response, we simulated with regression coefficients in Table 4. Those coefficient values are based on some of the values from the fitted Dirichlet regression with epilimnion temperature as a covariate. Summary statistics of the posterior samples from that fitted model are reported in Table 9 in the below data analysis section. The coefficients have been selected such that a change point is present in the simulated data.

Table 4: Coefficient values for the covariate model where X is a univariate seasonal time series and the response is generated based on equations (1) and (2).

	Regime 1					Regime 2				
	θ_1	θ_2	θ_3	θ_4	τ	θ_1	θ_2	θ_3	θ_4	τ
β_{i0}/γ_0	-3	0.8	1.6	0.7	-0.2	-1	0.8	1.4	0.3	-0.15
β_{i1}/γ_1	0.2	-0.02	-0.20	0.05	0.10	0.2	-0.05	-0.05	0.10	0.15

5.3 Variant Models

We also report some simulation results when we implement the variant versions of our model: a HMM where only the parameters in (1) may change, and a HMM where only the parameters in (2) may change.

5.3.1 No Change point present

Table 5: Proportion of models where the MCMC algorithm converged and proportion of times a jump occurred in the HMM for each combination of data and model fit.

No Change Data

Model	MCMC Converged	Change Point Detected
Location Shift HMM	0.50	0.00
Scale Shift HMM	0.95	0.04

Our first study corresponds to the "No Change Data" in the manuscript. The Location Shift HMM struggles to achieve MCMC algorithm convergence likely due to model misspecification. The Scale Shift HMM model is also an intentional misspecification but only three terms (the three coefficients γ_j, one for each of the three seasons) can change compared to 12 and 15 model coefficients in the Location Shift HMM and Location & Scale HMM, respectively. Even with a high rate of convergence, in only 4% of our simulated datasets did the Viterbi algorithm suggest any sort of jump had occurred.

5.3.2 Single Change Point

Next we consider the two scenarios where a single change point is present. In the synthetic data *Location Change Data* a single change point is present at time 31 and the shift is only present in the location parameters based on $\boldsymbol{\theta}^{(1)}$ and $\boldsymbol{\theta}^{(2)}$ in Table 3. In the *Scale Change Data* a single change point is present but only changes in the underlying scale are present, $\tau^{(1)}$ and $\tau^{(2)}$. In Table 6 we report the proportion of times the MCMC algorithm converged along with the proportion of times a jump occurred in the underlying Markov chain. For comparison we also include the results presented in the primary manuscript.

Table 6: Proportion of models where the MCMC algorithm converged and proportion of times a jump occurred in the HMM for the two simulation sets with a change point present.

	Location Change Data		**Scale Change Data**	
Model	MCMC Converged	Change Point Detected	MCMC Converged	Change Point Detected
Dirichlet Regression	1.00	—	1.00	—
Location & Scale HMM	0.98	1.00	0.99	1.00
Nonparametric Test	—	1.00	—	0.14
ecp Test	—	1.00	—	0.22
Likelihood Permutation	—	1.00	—	1.00
Location Shift HMM	0.99	1.00	0.50	0.09
Scale Shift HMM	0.97	0.13	1.00	1.00

We see that the MCMC algorithm converges for both variant models when a shift in the location parameter is present. Not surprisingly, in only 13% of 100 randomly selected fits did the Scale Shift HMM jump. Similarly, in the presence of a shift in the scale parameter, the Scale Shift HMM reports a high rate of convergence and detection of the change point. Given it is a dramatic case of model misspecification, it is not too surprising that Location Shift HMM struggles to achieve convergence and does not detects the shift in the scale.

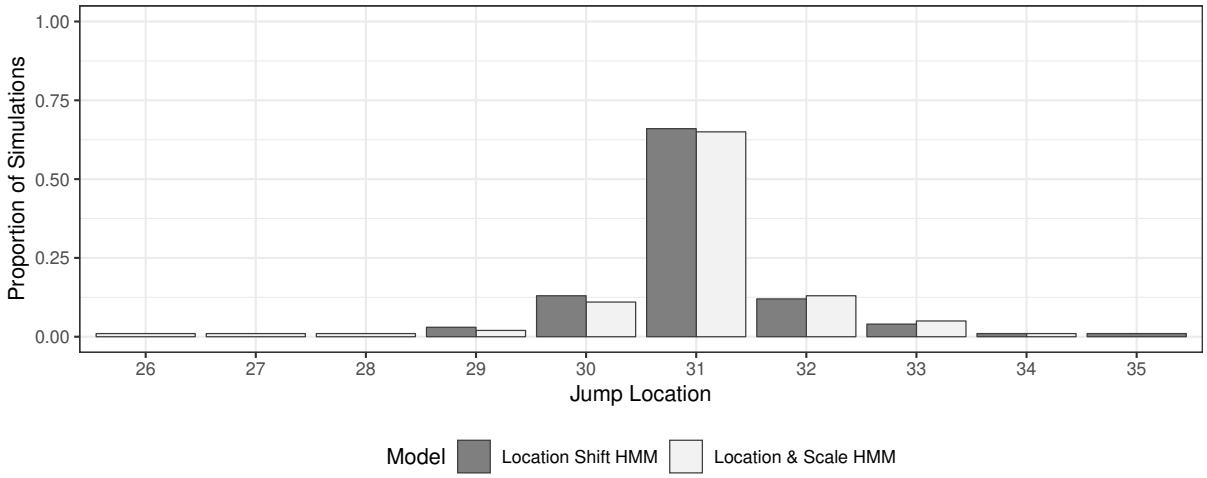

Figure 1: Distribution of HMM jump locations for 100 randomly selected converged model fits using Location Change Data.

We also compare the location of the change point locations for the "correct" models, and compare it with the proposed Location & Scale HMM presented in the primary article. In Figure 1 we see that both the Location Shift HMM and Location & Scale HMM report a similar distribution in the estimated location of the change point, located at time point 31.

In Figure 2 we report the estimated location of the change point locations for the Scale Shift HMM and Location & Shift HMM techniques. Overall the estimated jump locations do appear to occur a little before the true value (true shift occurs at time point 30-31), this is likely explained by the seasonality in the series.

5.3.3 Model Goodness-of-fit

Lastly, we explore the goodness-of-fit of the models against the simulated data. For the "No change data", the Dirichlet regression is the most appropriate model. For the location change data, we would expect the Location Shift HMM to be the most appropriate and for the spread change data, we would expect the Scale Shift HMM to be best fitting. Of course, knowledge of the presence and nature of a change point would require oracle-type knowledge, and is not tenable in practice. Here use those as baselines to demonstrate the proposed Location & Scale HMM does a good job of closely matching the "correct" model.

To compare the Dirichlet regression-based models, we use the approximate leave-one-

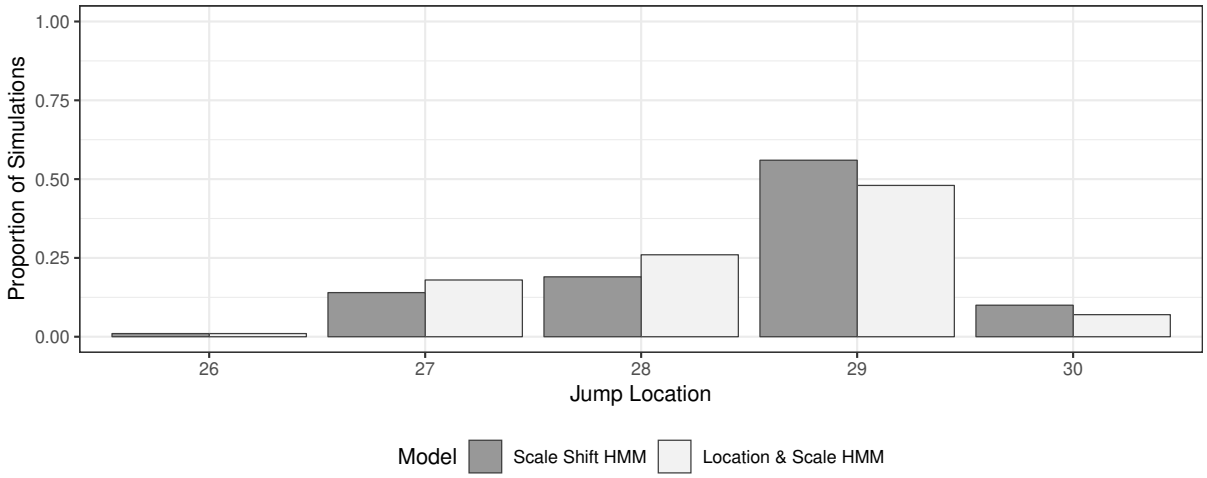

Figure 2: Distribution of HMM jump locations for 100 randomly selected converged model fits using Scale Change Data.

out (LOO) cross-validation [Vehtari et al., 2017] implemented in the `loo` package [Vehtari et al., 2018] to assess goodness-of-fit. Figure 3 provides boxplots of the LOO values for each model and simulation scenario based on the the randomly sampled 100 converged fitted models.

We note that the expected best fitting model reports the smallest LOO in each case. We also note that the Location & Scale HMM is the second best fitting in both simulation studies where a change point is present. In the No Change Data case, it is the worst fitting but this is not unexpected; it has the largest number of parameters to estimate of the four models. Even then, the model never suggested a change point was present.

Overall the simulations in the primary manuscript, and those included here, suggest the Location & Scale HMM is effective in determining the presence of change points and can provide a value tool to determine the nature of any changes (shifts in location and/or scale of a Dirichlet distribution).

6 Data Analysis

Next we provide some additional results regarding the analysis of the phytoplankton taxa in the primary manuscript.

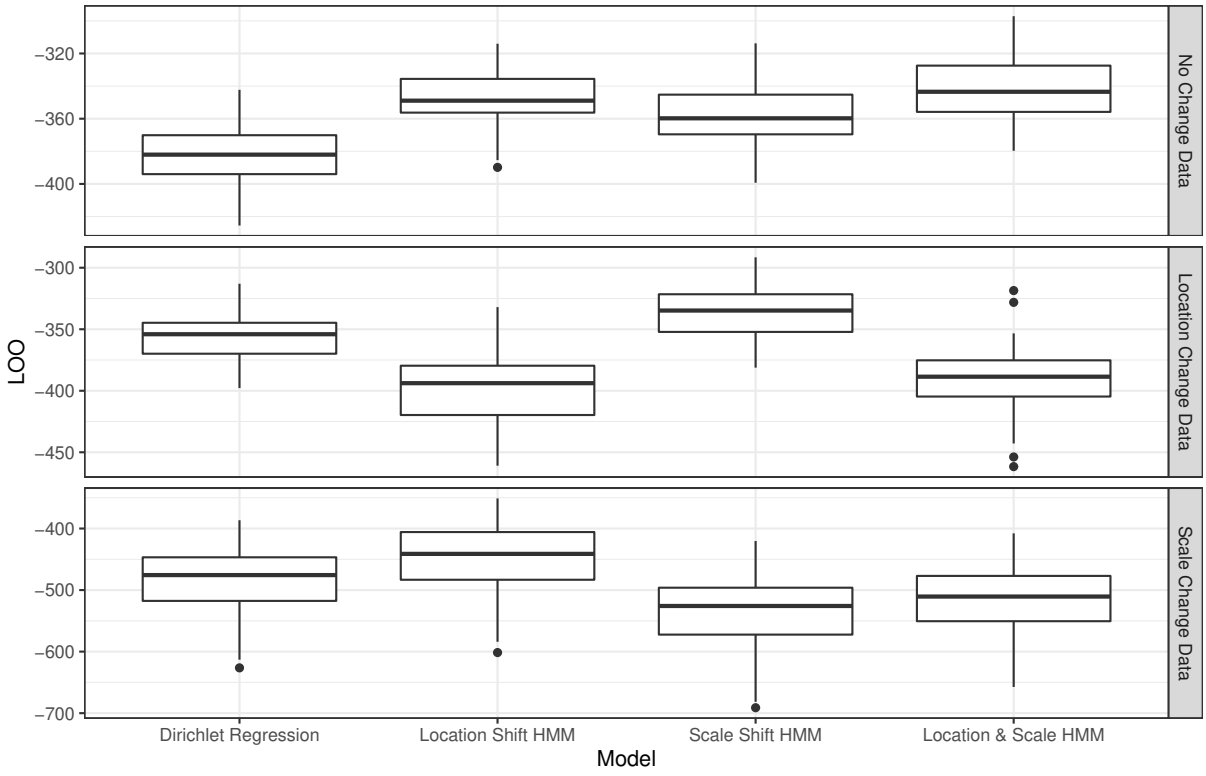

Figure 3: Distribution of LOO values for each model and 100 realizations of simulated data.

6.1 Model with seasonality

Through the analysis of the phytoplankton we eventually chose a 2-state Location & Scale HMM to best explain the nature of the behavior of phytoplankton taxa. This particular model included dummy variables (or indicators) to model the seasonality.

6.1.1 Location & Scale HMM fit

The chosen model was fit using the parameters outlined in section 3. By standard metrics, the MCMC algorithm converged; the Gelman & Rubin statistics for all parameters is near 1. As a further demonstration of convergence of the MCMC algorithm, Figure 4 displays the trace plots for the β_{0j} coefficients (*spring* coefficients) for both regime 1 and regime 2. The index in R output `b_loc[k,i,j]` is read as such: the `k` term defines the HMM state (1 or 2), the `i` determines the β_{i-1} coefficient, here $i = 1$ in all cases which coincides to the β_0 terms (spring) and the `j` determines which component (1 = blue-green algae, 2= green

algae, 3=flagellates and 4=diatoms).

Similarly, Figure 5 displays correlograms of the autocorrelation functions for the posterior samples. Between Figures 4 and 5 we have further evidence of the MCMC chains converged and that we have an uncorrelated posterior sample, leading to reasonable estimations of the parameter standard errors.

Table 7 provides summary statistics (posterior mean, standard deviation, 10^{th}, 50^{th} and 90^{th} percentiles) for all the β coefficients in (1) for the fitted Location & Scale HMM on the phyotplankton data.

Table 7: Posterior distribution summary statistics for β coefficients in (1) when modeling seasonality with a dummy variable.

		Regime	Mean	Std. dev.	10%	50%	90%
Blue green algae	Spr.	1	-0.7229	1.0541	-2.0372	-0.7453	0.6194
		2	-0.2601	1.0084	-1.6310	-0.2292	1.0474
	Sum.	1	1.6222	1.0237	0.3423	1.6633	2.8915
		2	1.9540	1.0635	0.6438	1.9568	3.3137
	Fall	1	1.4575	1.0344	0.0846	1.4669	2.7967
		2	1.5256	1.0376	0.1730	1.5054	2.8233
Green algae	Spr.	1	-0.4098	1.0555	-1.7292	-0.4344	0.9346
		2	0.1120	1.0021	-1.2183	0.1154	1.4197
	Sum.	1	-0.1071	1.0724	-1.5178	-0.1054	1.2768
		2	-0.5281	1.0696	-1.8615	-0.5506	0.9033
	Fall	1	-0.2669	1.0575	-1.6340	-0.2810	1.0315
		2	-0.3293	1.0217	-1.5881	-0.3357	1.0083
Flagellates	Spr.	1	0.6139	1.0406	-0.7169	0.6069	1.9058
		2	0.7935	1.0054	-0.5323	0.7862	2.1258
	Sum.	1	-1.0057	1.0531	-2.4202	-0.9852	0.3119
		2	-0.8077	1.0638	-2.1257	-0.8502	0.5541
	Fall	1	-0.5918	1.0203	-1.8909	-0.6008	0.6642
		2	-0.6856	1.0236	-1.9563	-0.7225	0.6409
Diatoms	Spr.	1	0.4449	1.0499	-0.9134	0.4502	1.7923
		2	-0.6557	1.0063	-1.9776	-0.6736	0.6780
	Sum.	1	-0.7283	1.0507	-2.1049	-0.7149	0.6279
		2	-0.4618	1.0771	-1.8353	-0.4845	0.9384
	Fall	1	-0.5391	1.0366	-1.9462	-0.5553	0.7646
		2	-0.5927	1.0402	-1.9015	-0.5892	0.7236

Note, in the main article we report the corresponding $\boldsymbol{\theta}$ values in Figure 7. The table is structured in such a way that it is easy to compare the coefficients within the HMM states.

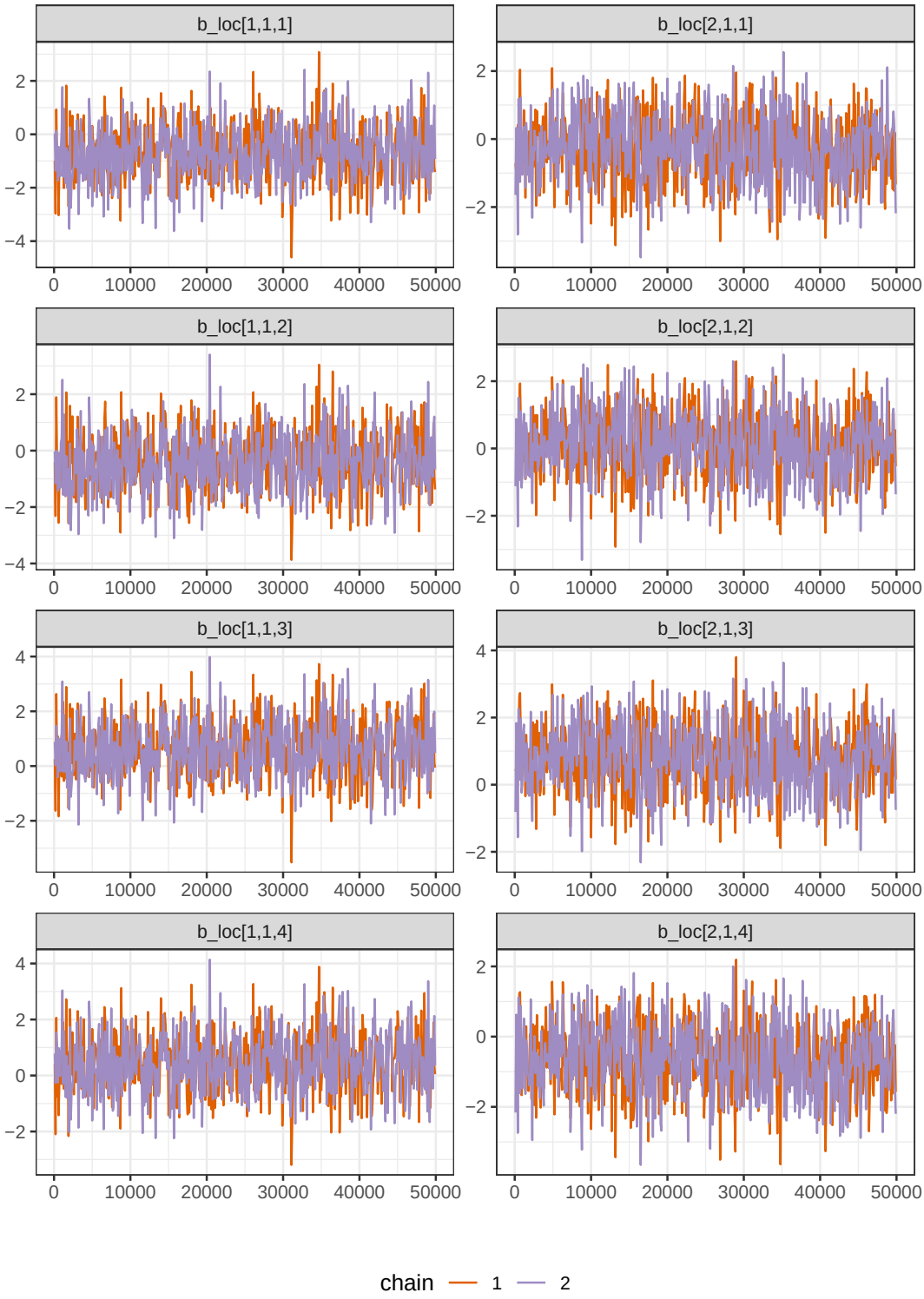

Figure 4: Trace plots of the β_{0j} coefficients (spring) for each of the four phytoplankton types, blue-green algae, green algae, flagellates and diatoms, respectively. Left side is regime 1 trace plots while right side is regime 2.

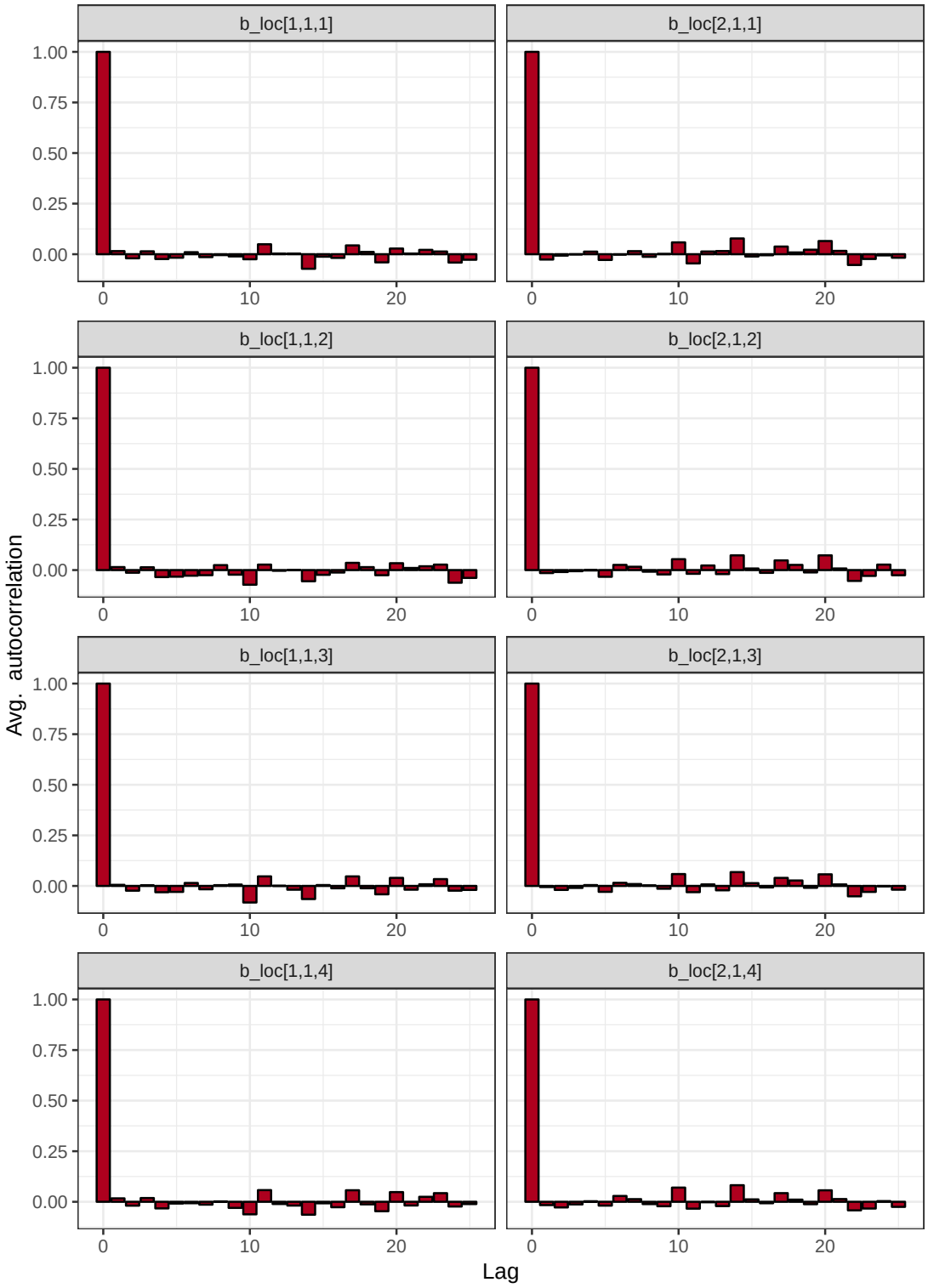

Figure 5: Correlograms (ACF) of the β_{0j} coefficients (spring) for each of the four phytoplankton types, blue-green algae, green algae, flagellates and diatoms, respectively. Left side is regime 1 ACF plots while right side is regime 2.

6.1.2 Variant model fits

In addition the Location & Scale HMM fit in the primary manuscript, we also considered fitting the variant models, the Location Shift HMM and the Scale Shift HMM. In the primary manuscript we see that the 2-state Location & Scale HMM was best of the models considered (the 3-state HMM never jumped to the third state). Here, we summarize the LOO CV values for the single state Dirichlet regression, the 2-state Location & Scale HMM, and the two variant models.

Table 8: LOO values for phytoplankton fits

Dirichlet Regression	Location Shift HMM	Scale Shift HMM	Location & Scale HMM
-384.9429	-403.6335	-371.5268	-407.9545

We see that the Location & Scale HMM is the best fitting of those considered here suggesting there is a change in both location and scale. Even though Location Shift HMM and Scale Shift HMM were not as good, they still provide some valuable information regarding the behavior of the phytoplankton.

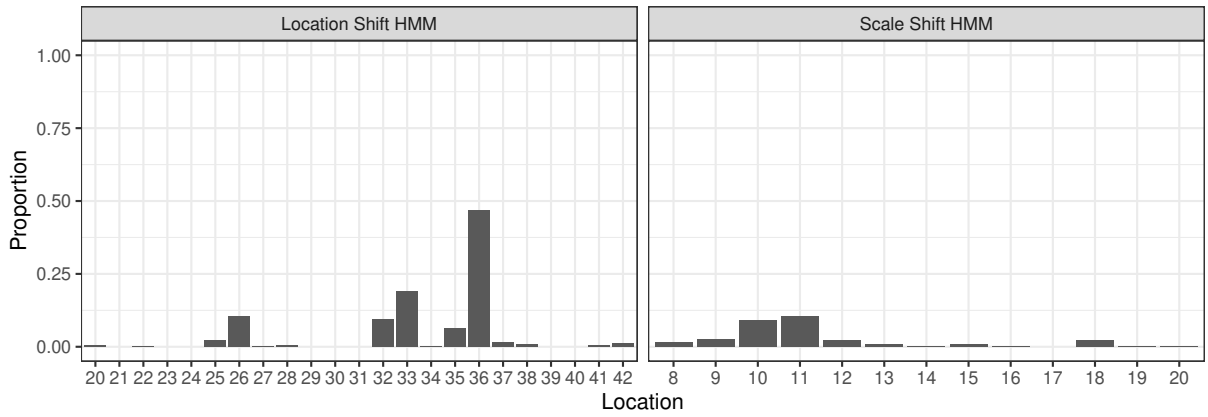

Figure 6: Distribution of jump locations based on the Location Shift HMM and Scale Shift HMM (only reporting the 30.8% posterior samples where a jump occurred) for the phytoplankton data.

The Location Shift HMM converged and a jump was found in all 1000 posterior samples, the location of those jumps can be found in the left panel of Figure 6. The Scale Shift HMM provides inconclusive findings. In the 1000 posterior samples, only in 30.8% of the Viterbi paths did a jump occur. The right panel of Figure 6 provides the locations of the

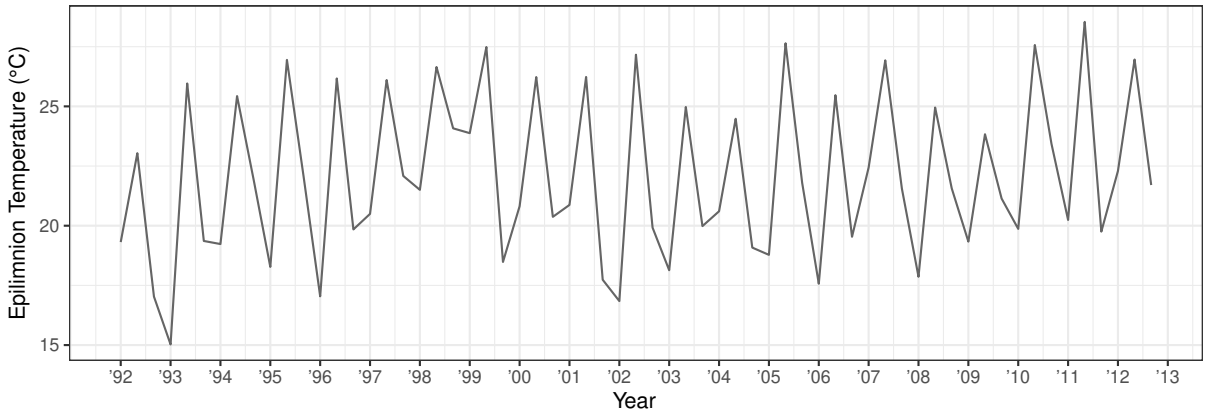

Figure 7: Epilimnion temperature (degrees Celsius), aggregated into three measures per year, at our study study Lake Acton.

posterior jumps for the Scale Shift HMM. It suggests a shift in the location at time point 36 (posterior mode, fall 2003), 33 or 26 and there is some weak evidence of a potential scale shift at time point 10 or 11 (posterior mode, summer 1995).

In the primary manuscript we report that the Location & Scale HMM suggests a change in regimes in all 1000 posterior samples. The point of that shift is suggested to occur at time point 26 (posterior mode, summer 2000). Note that this same time point was suggested in the Location Shift HMM in roughly 10% of posterior samples and that time point 26 is essentially a compromise between the shifts occurring at time 36 (or 33) in the Location Shift HMM and that of time point 11 suggested by the Scale Shift HMM.

6.2 Time varying Covariate

In the primary paper, we used the epilimnion temperature as a time varying covariate to predict the proportions of phytoplankton using (1) and (2). Figure 7 displays the epilimnion temperature at Acton Lake for the 21 years of study.

The data is highly seasonal with an expected increase in temperature from spring to summer and a decrease into the fall. Thus, using this as a covariate in the analysis of the phytoplankton compositions in the proposed model should help address the seasonality. After removing the seasonal effects, the epilimnion temperature time series exhibits some slight autocorrelation and weak evidence of an underlying trend. We fit a basic trend model (trend coefficient: 0.0732 with SE 0.0445) with an AR(1) ($\hat{\phi} = 0.2543$, $SE_{\hat{\phi}} = 0.1208$) to

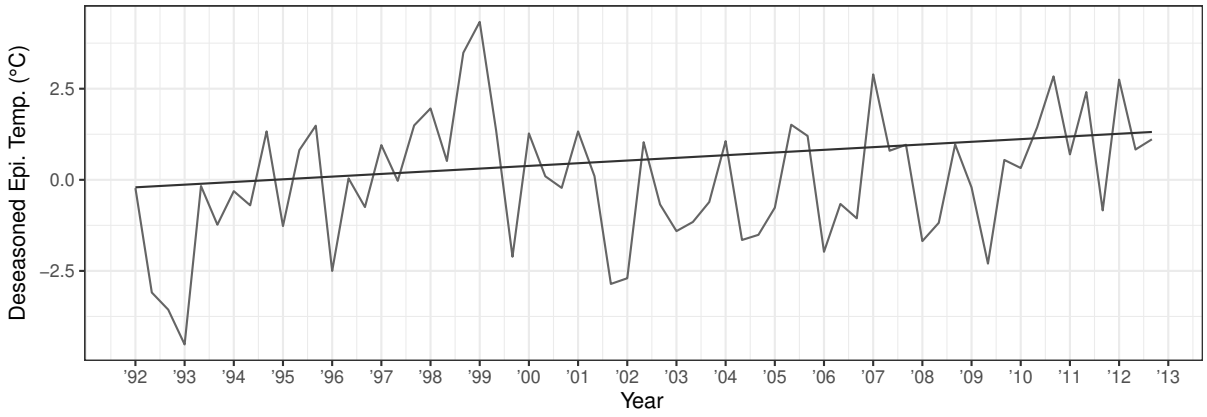

Figure 8: Deseasoned epilimnion temperature (°C) in time with suggested trend based on a trend regression plus AR(1) error model.

the deseasoned series. Figure 8 displays the deseasoned epilimnion temperature for the lake along with a modeled underlying trend. We note that the trend is not statistically significant (based on historical practice) but the line is suggestive. Further, the deseasoned data contains 31 negative values and 32 positive values (as expected), but 13 of these negative values occur before time point 26 (the suggested regime shift in our previous modeling) and 20 occur before the time point 36 (99.6% of the posterior paths had a regime shift at or before this point). Overall, there is at least suggestive evidence that the epilimnion temperature increased during the period of study and may help explain the shift in phytoplankton phenology.

We then fit a Location & Scale HMM where epilimnion temperature is the only predictor variable present in the model. From the fitted Dirichlet regression model, we calculated the marginal residuals for each phytoplankton group (difference between the observed proportions and the expected proportion) and are displayed in Figure 9.

From the residuals it is clear that epilimnion temperature is not adequately modeling the seasonality for the blue-green algae proportions, nor the flagellates. The expected proportion of green algae appears to be over predicted (mostly negative residuals) and epilimnion temperature does not appear to adequately model the shifts in diatoms.

The fitted model appears worse than the HMM that model seasonality with dummy variables by several metrics. We also note that the LOO value is worse than that when modeling the seasonality with dummy variables.

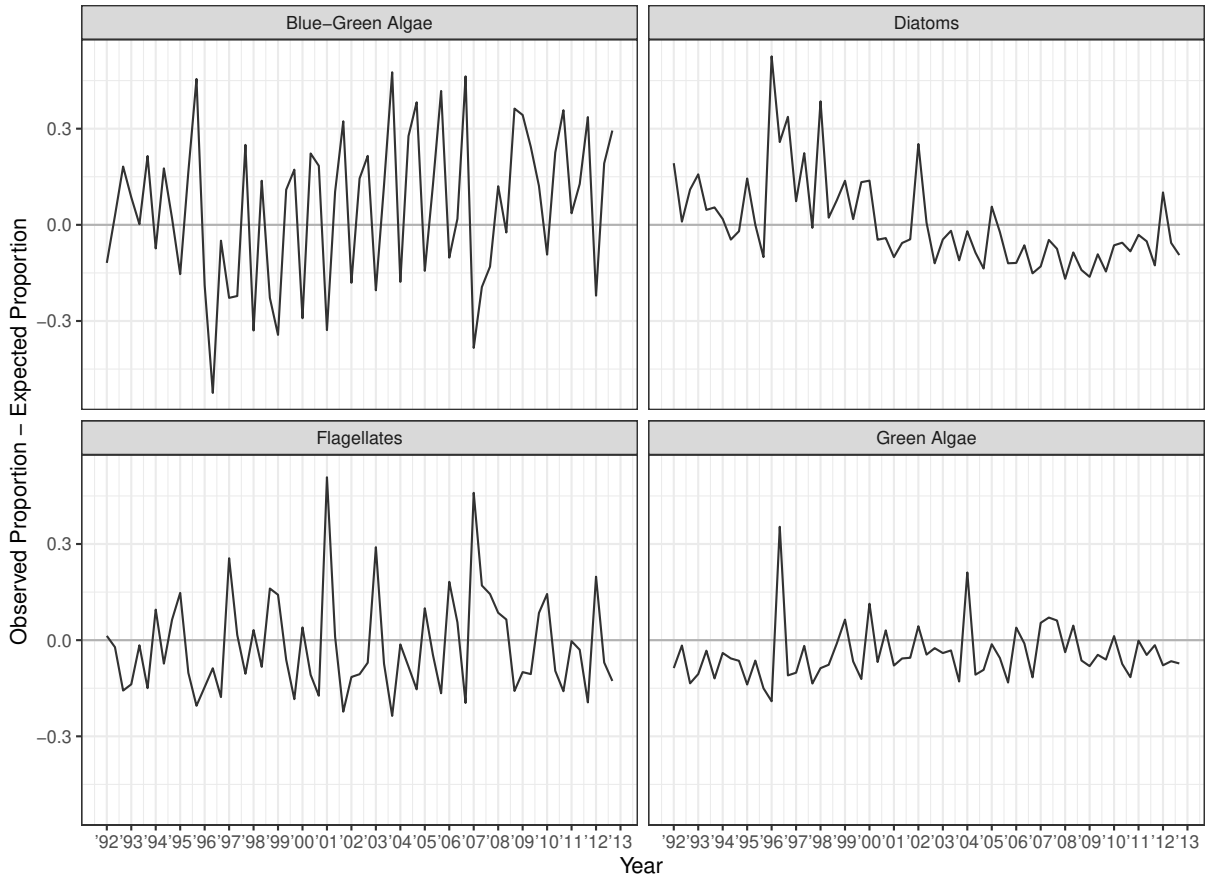

Figure 9: Residuals in time for each phytoplankton taxa based on a Dirichlet regression with epilimnion temperature as the covariate.

Table 9: Posterior distribution summary statistics for β coefficients in (1) when using the Epilimnion Temperature as a time-varying covariate.

	Component	Mean	Std. dev.	10%	50%	90%
Intercept	Blue-Green	-2.9634	1.1087	-4.4167	-2.9657	-1.6361
	Green algae	0.5163	1.1337	-0.9779	0.5039	1.9681
	Flagellates	1.5771	1.0919	0.1850	1.5544	2.9949
	Diatoms	0.7568	1.1394	-0.6322	0.7041	2.2719
Epi. Temp.	Blue-Green	0.1876	1.0050	-1.0761	0.2093	1.5038
	Green algae	-0.0076	1.0053	-1.2809	0.0089	1.3126
	Flagellates	-0.0354	1.0054	-1.2939	-0.0235	1.2827
	Diatoms	-0.0309	1.0070	-1.3062	-0.0137	1.2762

Table 9 provides summary statistics of the posterior distribution for the estimated β coefficients in (1) when epilimnion temperature is used as a predictor variable. Note that the coefficient on the epilimnion temperature is centered around the value zero for all four components. This highlights some of the reasons we see such a poor fit when using this model.

6.3 Phytoplankon: 5 measures per year

The raw data collected at Acton Lake is done so in irregular intervals. In the primary manuscript we aggregated the phytoplankton biomass, by taxonomic group, into three measures per year, roughly corresponding to the spring mix period, summer stratification and fall mixing periods.

Here, we briefly consider a different aggregation where the data are broken into 5 segments per year: data collected in the spring (before June 16), an early summer period (from June 16 to July 15), a Midsummer period (from July 16 to August 15), a late summer period (from August 16 to Sep 15) and a fall season (after September 15). This data can be seen in Figure 10.

We note there are now $21 \times 5 = 105$ observations of 4-dimensional compositions. Further, using an m-state Location & Scale HMM with dummy variables to handle seasonality in the response requires estimation of $m \times 5 \times 4$ regression parameters for the location and $m \times 5$ regression parameters for the scale. As such, we needed to tune our HMM to achieve MCMC convergence. We modified the implementation of the HMM so that the initial state

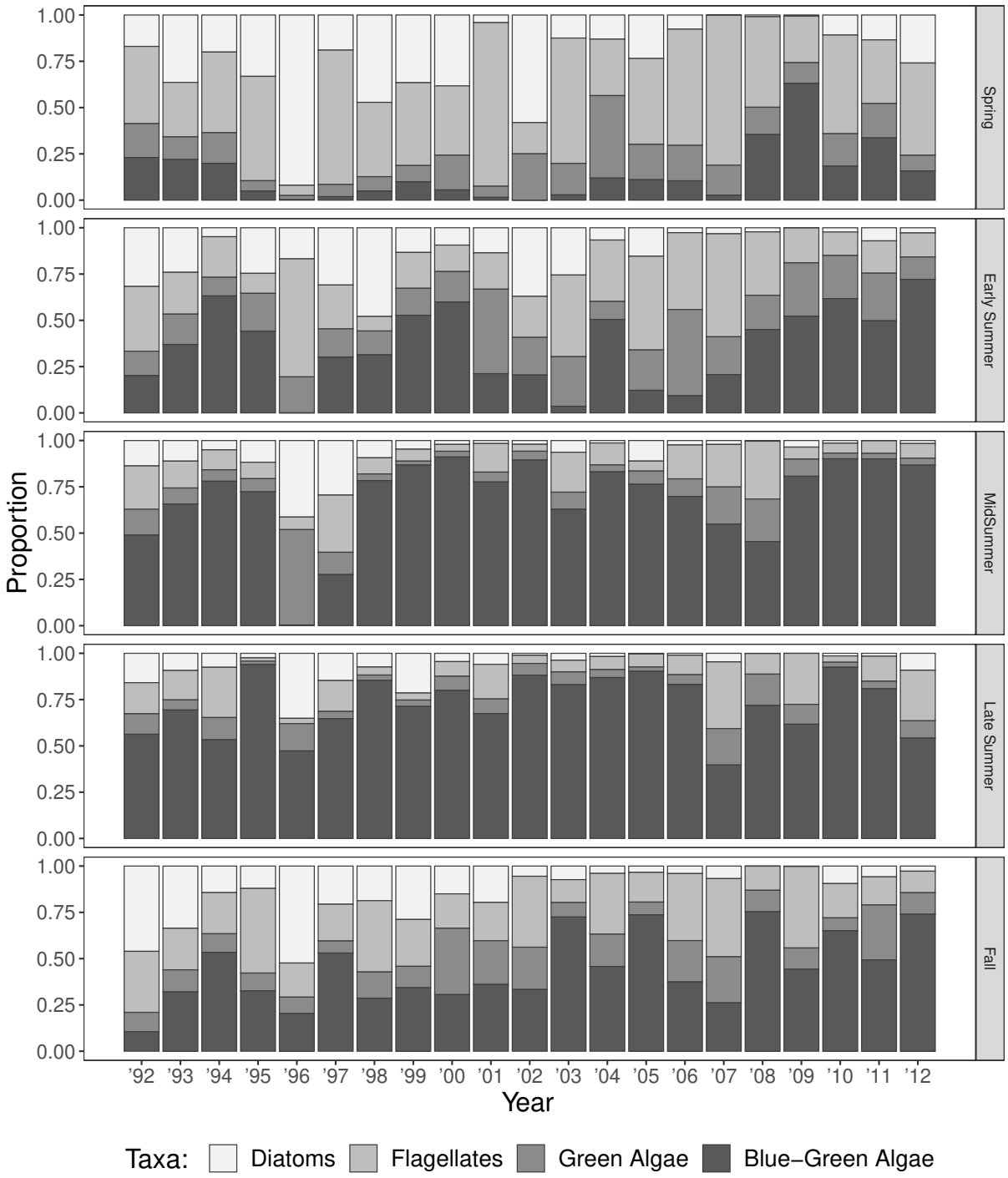

Figure 10: Composition of phytoplankton in time, faceted by seasons. There are 5 measurements per year over 21-years of 4-dimensional data. Note the time series has seasonal effects and there appears to be a change in the proportions during the length of study.

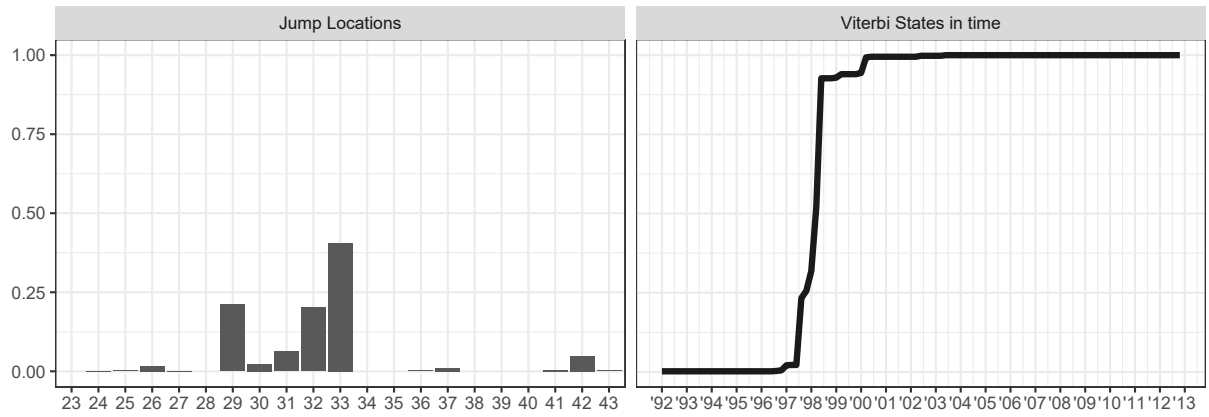

Figure 11: Posterior distribution of jump locations (left panel) and mean posterior Viterbi states (right panel) for the best fitting Location & Scale HMM for the phytoplankton data when aggregated to 5 measurements per year.

of the chain for the 3 years of data (or $3 \times 5 = 15$ observation) was state 1. This effectively allowed the MCMC to converge quickly and generate posterior samples of the regression coefficients to more accurately detect any changes.

The model detected a shift in the distribution of the phytoplankton; the timing of that shift can be seen in Figure 11. There, we see that the change point is detected circa 1998-1999.

The model found a fairly substantial change in both the location and scale of the distribution. In Figure 12 we see the mean posterior values for the location of phytoplankton proportions in the two regimes. Similar to the results in the primary manuscript, there is a noticeable increase in Blue-Green Algae between the two regimes, and largely to the detriment in the proportion of diatoms.

In Figure 13 we report the posterior distribution of the scale parameters. We note that overall the scale appears to have increased in Regime 2 (since approximately 1999). Recall that the variability of the Dirichlet distribution is inversely related to the scale parameter. So not only have the proportion of Blue-Green algae increased, they have done so more consistently since circa 1999.

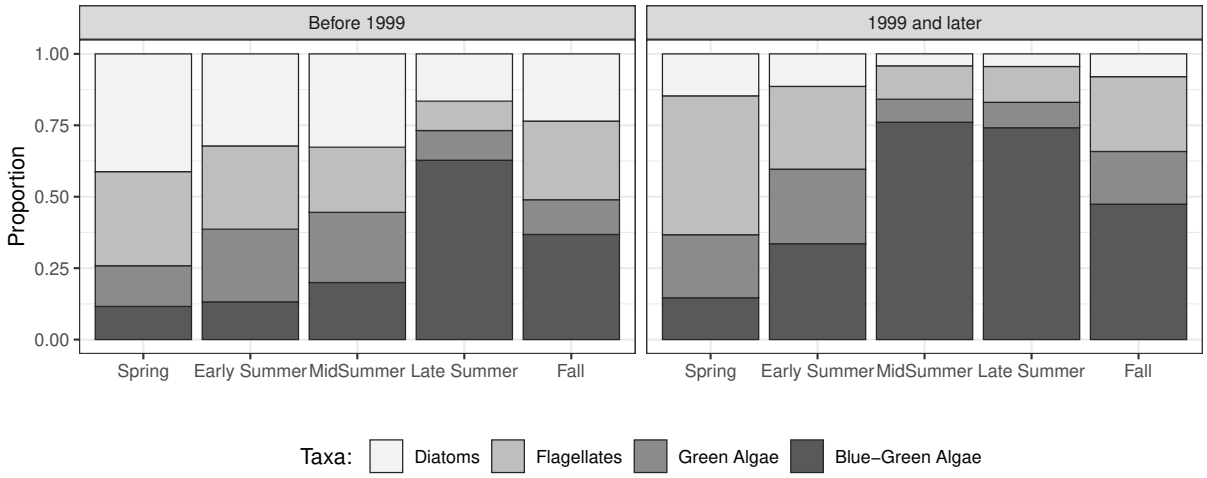

Figure 12: Expected compositions of phytoplankton based on the Location & Scale Shift hidden Markov model when phytoplankton data has been aggregated to 5 measurements per year.

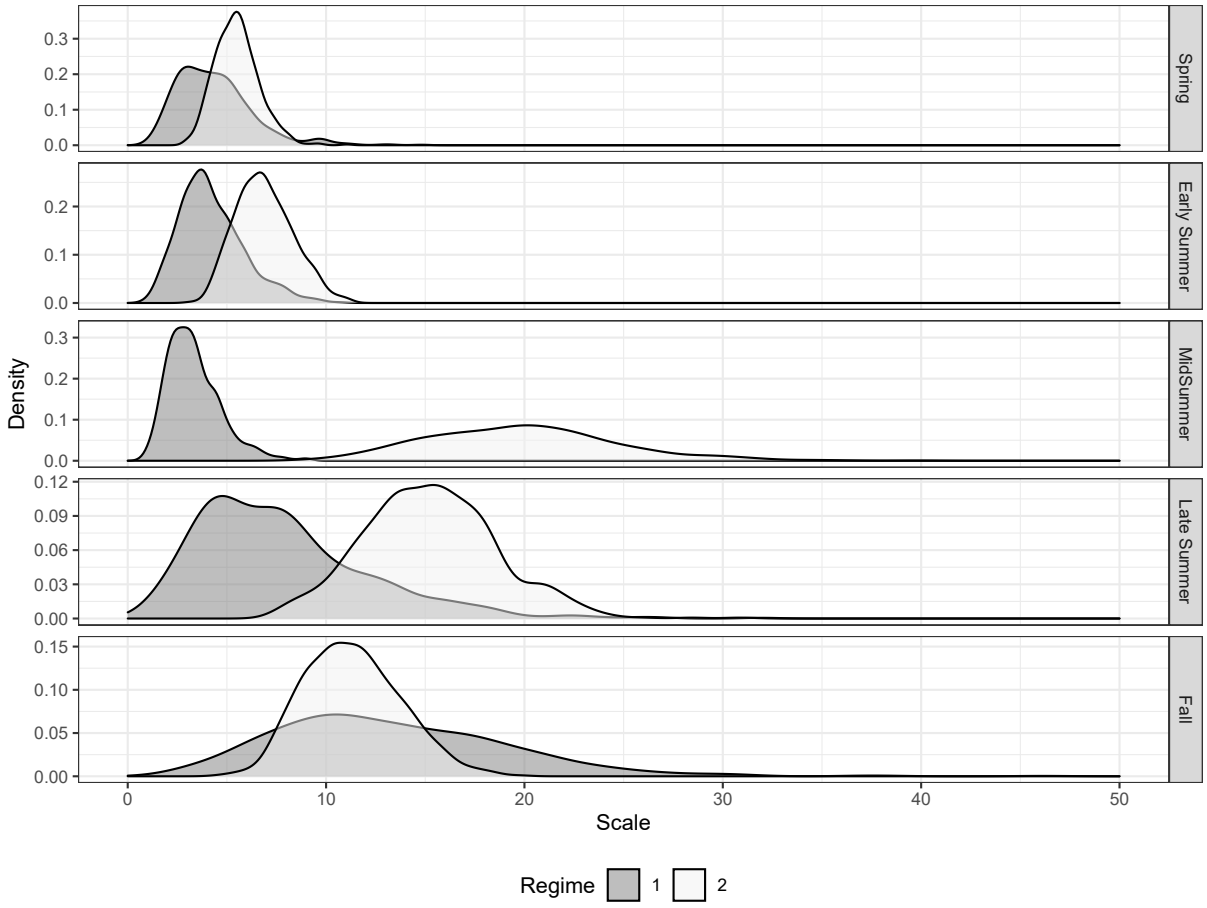

Figure 13: Posterior distribution of estimation phytoplankton scale parameters by season.

References

Olivier Cappé, Eric Moulines, and Tobias Ryden. *Inference in Hidden Markov Models (Springer Series in Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2005. ISBN 0387402640.

Siddhartha Chib. Estimation and comparison of multiple change-point models. *J. Econometrics*, 86(2):221–241, 1998. ISSN 0304-4076. doi: 10.1016/S0304-4076(97)00115-2. URL http://dx.doi.org/10.1016/S0304-4076(97)00115-2.

Mark Holmes, Ivan Kojadinovic, and Jean-François Quessy. Nonparametric tests for change-point detection à la Gombay and Horváth. *Journal of Multivariate Analysis*, 115(C):16–32, 2013. doi: 10.1016/j.jmva.2012.10.00. URL https://ideas.repec.org/a/eee/jmvana/v115y2013icp16-32.html.

Nicholas A. James and David S. Matteson. ecp: An R package for nonparametric multiple change point analysis of multivariate data. *Journal of Statistical Software*, 62(7):1–25, 2014. URL https://www.jstatsoft.org/v62/i07/.

Ivan Kojadinovic. *npcp: Some Nonparametric CUSUM Tests for Change-Point Detection in Possibly Multivariate Observations*, 2020. URL https://CRAN.R-project.org/package=npcp. R package version 0.2-0.

David S. Matteson and Nicholas A. James. A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association*, 109 (505):334–345, 2014. doi: 10.1080/01621459.2013.849605.

K.J. Prabuchandran, Nitin Singh, Pankaj Dayama, and Vinayaka Pandit. Change point detection for compositional multivariate data. *Preprint*, 2019.

Stan Development Team. RStan: the R interface to Stan, 2018. URL http://mc-stan.org/. R package version 2.18.2.

Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, 27(5):1413–

1432, 2017. ISSN 1573-1375. doi: 10.1007/s11222-016-9696-4. URL https://doi.org/10.1007/s11222-016-9696-4.

Aki Vehtari, Jonah Gabry, Yuling Yao, and Andrew Gelman. loo: Efficient leave-one-out cross-validation and waic for bayesian models, 2018. URL https://CRAN.R-project.org/package=loo. R package version 2.0.0.