Spatio-temporal variation in benthic primary productivity along multiple lake depth profiles

Kirsten Bevandick* Kate Colson† Stefano Mezzini* William Ou†

Julee Stewart‡

2021-10-06

1 Study Information

1.1 Title

Spatio-temporal variation in benthic primary productivity along multiple lake depth profiles

1.2 Description

Whole-lake ecosystem processes are complex and hard to quantify. Carbon and nitrogen cycling depend on the composition and abundance of a aquatic organisms, which, in turn, can vary spatio-temporally due to many physical and chemical factors. While most studies on whole-lake ecosystem processes focused focused primarily on pelagic production, Gushulak *et al.* (2021) presented the first study on nutrient cycling using depth profiles. Using depth profiles of soil δ^{13}C, δ^{15}N, percent C, percent N, and the C to N ratio, they determined that intermediate lake depths offered the best conditions for maximum phytobenthic production. The objective of our study is to replicate the study by Gushulak *et al.* (2021) while also accounting for the variation between different locations of the lake.

*University of British Columbia (Okanagan campus), British Columbia, Canada

†University of British Columbia (Vancouver campus), British Columbia, Canada

‡University of Regina, Saskatchewan, Canada

1.3 Hypothesis

Since light availability and disturbance due to currents and mixis affect phytobenthic production, we hypothesize that productivity will be highest at intermediate lake depths, where disturbance is low but the lake is shallow enough to allow large quantities of light to penetrate to the benthic region.

2 Design Plan

2.1 Study type

The proposed study is an observational study. Surface layer soil samples will be collected from various transects within the bed of Gall Lake. Isotopic signatures (δ^{13}C, δ^{15}N) and elemental composition (%C, %N, C:N) will be measured in each sample. For further details, see the sampling plan and description of variables below.

2.2 Blinding

Sample will be labeled with a non-descriptive code to remove all information regarding the samples' locations and depths.

2.3 Study design

Superficial soil samples (0-1 cm of depth) will be collected at 1-meter depth intervals (up to 16 m) in Gall Lake, Ontario, Canada, following four orthogonal transects (North, East, South, West). Four independent replicates will be performed along each transect, for a total of $16 \times 4 \times 4 = 256$ soil samples.

2.4 Randomization

No randomization will be present in the study. To reduce sources of anthropogenic of variation, samples will be taken by a single individual on the same week and analyzed by a single technician.

3 Sampling Plan

3.1 Existing data

Registration prior to creation of data. As of the date of submission of this research plan for preregistration, the data have not yet been collected, created, or realized. All data presented here was simulated

based on the data from Gushulak *et al.* (2021).

3.2 Explanation of existing data

Existing data will not be used in our replication study; all data will be sampled after submitting the preregistration.

3.3 Data collection procedures

Data will be collected at the Gall Lake study site (50°11′ N, 90°42′ W) located in Ontario, Canada. This undisturbed study site is surrounded by boreal forest with a high abundance of black spruce, jack pine, and poplar and a lower abundance of birch, balsam fir, and larch (Kingsbury, Laird & Cumming, 2012). Data collection will take place in a single week during the summer of 2022.

Surface sediment samples (0-1cm) will be collected using a mini-Glew gravity coring apparatus (Glew, 1991) as was done in Gushulak *et al.* (2021) at sequential depth intervals, determined using depth sounders (Gushulak *et al.*, 2021). The surface sediment sampling will be along four depth transects in each of the basins of Gall Lake, to expand upon the area covered in Gushulak *et al.* (2021) and account for any differences in deposition to the surface sediments due to currents, input, and mixing in Gall Lake. This expansion of coverage will allow us to account for spatial variation among basins, while the replicates along each transect will account for variation within basins to be considered during the statistical analysis.

The stable isotope analysis from the surface sediment samples will be conducted using standard methods (Savage, Leavitt & Elmgren, 2004; Bunting *et al.*, 2010) using Thermo Finnigan Delta V isotope ratio mass spectrometer that is equipped with a ConFlow IV dilution inlet system, to ensure consistency with the methods of Gushulak *et al.* (2021).

3.4 Sample size

There will be $n = 16$ surface sediment samples taken at one-meter depth intervals (from 1 m to 16 m) along the four depth transects in each basin of Gall Lake. Each transect will be replicated four times. This will result in a sample size of $N = 16 \times 4 \times 4 = 256$ for our replication study.

3.5 Sample size rationale

The sample size was determined to maximize statistical power while remaining within financial and personnel constraints. In addition, the project will be done in a Bayesian framework, so no Frequentist hypothesis

testing will be performed.

3.6 Stopping rule

The entirety of the dataset will be collected as described above. If the entire dataset cannot be collected, such as if Gall Lake is shallower than 16 m at the time of sampling (Kingsbury, Laird & Cumming, 2012), the sample size will be reduced accordingly. If the depth of Gall lake is below 15 m at the time of sampling, additional samples will be taken at intermediate locations (e.g. 1.5 m, 2.5 m, ...) until the sample size of $N = 256$ is reached.

4 Variables

4.1 Manipulated variables

No variables will be manipulated in this study, as it is an observational study.

4.2 Measured variables

All measured variables will be reported as was done by Gushulak *et al.* (2021).

Depth: The depth of Gall Lake will be measured along four transects in the lake's basins using a depth sounder as Gushulak *et al.* (2021) did. Four sediment samples will be collected at every 1-m depth interval along each transect from 1 m to 16 m of depth. In all `R` scripts and data files, the variable will be named `depth_m`.

δ^{13}C: The ^{13}C isotopic signatures will be measured from freeze-dried sediment samples using a Thermo Finnigan Delta V isotope ratio mass spectrometer with a ConFlow IV dilution inlet system, as described by Gushulak *et al.* (2021). The values will be recorded using the standard δ^{13}C notation (expressed in parts per thousand relative to the Vienna Pee Dee Belemnite standard, VPDB). The ^{13}C values will be referred to as `d13c` in all `R` scripts and datasets.

δ^{15}N: The ^{15}N isotopic signature will be measured using freeze-dried sediment samples (as above), calibrated using atmospheric gas (Savage, Leavitt & Elmgren, 2004; Bunting *et al.*, 2010), and recorded using the standard δ^{15}N notation (expressed in parts per thousand relative to the atmospheric abundance of ^{15}N). The ^{15}N values will be referred to as `d15n` in all `R` scripts and datasets.

Nitrogen fraction: The fraction of nitrogen in the freeze-dried sample, f_N, will be calculated using the simple formula $f_N = m_N/m$, where m_N is the mass of nitrogen in the sample and m is the sample's mass.

The variable will be named `frac_n` in all datasets and `R` scripts. The percentage of nitrogen in the sample will be used in the figure(s) for consistency with Gushulak *et al.* (2021), and it will be indicated as `perc_n` in all datasets and `R` scripts (when used).

Carbon fraction: The fraction of carbon in the freeze-dried sample will be calculated similarly to the nitrogen fraction (see above). The fractions and percentages of carbon in the freeze-dried samples will be indicated as `frac_c` and `perc_c`, respectively, in all `R` scripts and datasets.

Carbon to Nitrogen ratio: The C:N ratio will be expressed as a unit-less number.

4.3 Indices

Carbon to Nitrogen ratio: The ratio of carbon to nitrogen will be produced using the simple formula $R = f_C/f_N$, where f_C and f_N are the fractions of carbon and nitrogen in the sample, respectively. In `R` scripts and the dataset, the variable will be named `c_n_ratio`.

5 Analysis Plan

All analysis will be performed in `R` (R Core Team, 2021). Data wrangling will be performed with packages from the `tidyverse` set of packages (Wickham *et al.*, 2019). All supporting `R` scripts, simulated data, and supporting figures are available in the public GitHub repository at https://github.com/jiaangou/TheBestGroup/.

5.1 Statistical models

Similarly to Gushulak *et al.* (2021), the isotope data will be analyzed using Generalized Additive Models (GAMs) via the `mgcv` package using and Empirical Bayesian approach, so priors will be estimated by the data (Wood, 2011, 2017). Although we could estimate priors from the data presented by Gushulak *et al.* (2021) and fit fully Bayesian models via the `brms` package (Bürkner, 2017), we will use the `mgcv` package to produce results which are independent from the results of Gushulak *et al.* (2021).

The amount of ^{15}N and ^{1}3C (`d15n` and `d13c` in the `R` scripts and datasets, respectively) will be modeled using a Gaussian conditional distribution with an *identity* link function, since both parameters can be either positive or negative. The fractions of nitrogen and carbon in the samples (`frac_n` and `frac_c`) will be modeled using GAMs with a beta conditional distribution with a *logit* link function, as there is no distribution in the `mgcv` package for numbers strictly between 0 and 100. Finally, the proportion of carbon to nitrogen (`c_n_ratio`) will be modeled using a GAM with a gamma conditional distribution and a *log* link function.

Of the distributions available in `mgcv`, a gamma distribution is most appropriate since the ratio is strictly positive but not right-bound.

The effects' Bayesian credible intervals (CIs) will be estimated on the models' link scales using normal approximation (± 1.96 standard deviations) and back-transformed to the response scale using the models' inverse-link functions. For example, the CIs for the mean C:N will be estimated as the estimated mean ± 1.96 standard deviations on the *log* scale, and then they will be back-transformed to ratios by exponentiating the values. See the example code in the *Other* section and the following section for more information on GAMs and transformations.

All five models accounted for a shared global trend between replicates and trends within-replicate. For simplicity, the smoothness parameter was assumed to be the same between replicates (see model "GS" in Pedersen *et al.*, 2019).

Each model will have three smooth predictors: (1) `s(depth_m, k = 15)` will estimate the mean effect of lake depth, (2) `s(depth_m, basin, k = 10, bs = 'fs')` will account for the variation between basins, and (3) `s(depth_m, replicate, k = 10, bs = 'fs')` will account for the variation within replicates. All three predictors will use thin plate regression splines (the default for the `s()` function in `mgcv`).

The dimension of the basis of each spline was selected to allow the global term (term (1)) to have the greatest complexity (`k = 15`), while the predictors which account for the effect of `basin` and `replicate` on `depth` will have lower complexity. Predictors (2) and (3) can be viewed as estimates of the average deviation of each group from the global trend (see model "GS" in Pedersen *et al.*, 2019). Predictors (2) and (3) are defined as *factor smooths* (`bs = 'fs'`) since it seems reasonable to assume a common smoothness parameter between basins and between replicates and include the effect of `basin` and `replicate` as random effects. If residual diagnostics show strong autocorrelation due to different levels of smoothness between `basin`s, the second term will be changed to a `by` smooth with the syntax `s(depth_m, by = basin, k = 10)` (see model "GI" in Pedersen *et al.*, 2019). (In such case, the effect of `basin` will become fixed effect instead of a random effect.)

The models' will be fit via penalized maximum likelihood with a penalty on the smooths' curvatures to avoid over-fitting (Wood, 2011; Simpson, 2018). The smoothness parameters will be optimized using Restricted Marginal Likelihood (`method = 'REML'`), rather than the default Generalized Cross Validation via Mallow's C_p (`'GCV.Cp'`), since REML does not over-fit as often as GCV (Reiss & Ogden, 2009; Wood, 2011).

Although credible intervals will be estimated, this study will be performed in a purely Bayesian framework, so statistical significance and Frequentist null-hypothesis testing will not be considered. (Also note that the

p-values for smooth terms returned by `summary` are approximate and highly uncertain, see Wood, 2017.) Rather, the posterior distributions of the models will be used to assess the likelihood of the results of Gushulak *et al.* (2021).

5.2 Transformations

The only data transformation that was performed was the conversion of percent carbon and nitrogen (between 0% and 100%) to proportions (between 0 and 1). Note that since the transformation $a = \frac{b}{100}$ is a linear transformation, Jensen's inequality does not apply here (Jensen, 1906).

Modeling the data with GAMs removes the need for transforming data that violates the normality assumptions which linear models depend on. Since GAMs estimate transformed mean responses ($g[\mathbb{E}(Y)]$, where $g(\cdot)$ is the link function) rather than the mean transformed response ($\mathbb{E}[g(Y)]$), Jensen's inequality (Jensen, 1906) does not apply here.

5.3 Inference criteria

The likelihood of the results by Gushulak *et al.* (2021) will be assessed based on the 95% credible intervals from the model's posterior distributions. Additional credible intervals (75%, 90%) will be also calculated to provide more information on the posterior distributions.

5.4 Data exclusion

All data will be checked with simple sanity checks. Any unreasonable values (e.g. proportions outside the interval $[0, 1]$, and negative C:N ratios) will be checked for errors during transcription, calculations, and coding. All calculations will be done in `R` scripts, and the original raw data will not be modified unless to correct errors. If no reasonable error source can be detected, the outlier(s) will be removed, but we will present the models fit to the full dataset and the models fit to the cleaned dataset.

5.5 Missing data

Any rows with missing data will be removed from the dataset for the respective model(s) only. For example, a missing δ^{15}N value will affect the δ^{15}N model, but the row will not be removed from the datasets used in the other four models. The removal of missing values is automated by the `gam()` function in the `mgcv` package.

6 Other

6.1 Model definition

```
# stable nitrogen isotope
m_d15n <-
  gam(d15n ~
        # average effect of depth across all basins and replicates
        s(depth_m, k = 15) +
        # effect of basin (keep complexity < that of global smooth)
        s(depth_m, basin, k = 10, bs = 'fs') +
        # effect of replicates (keep complexity < that of basin)
        s(depth_m, replicate, k = 5, bs = 'fs'),
      family = gaussian(link = 'identity'),# conditional distribution
      data = isotopes,
      # optimize the smoothness parameter via REML (see ?mgcv::gam)
      method = 'REML')

# stable carbon isotope
m_d13c <-
  gam(d13c ~
        s(depth_m, k = 15) +
        s(depth_m, basin, k = 10, bs = 'fs') +
        s(depth_m, replicate, k = 5, bs = 'fs'),
      family = gaussian(link = 'identity'),
      data = isotopes,
      method = 'REML')

# fraction of carbon in the samples
m_frac_c <-
  gam(frac_c ~
        s(depth_m, k = 15) +
        s(depth_m, basin, k = 10, bs = 'fs') +
        s(depth_m, replicate, k = 5, bs = 'fs'),
      # logit(0) = -Inf, logit(1) = Inf
      family = betar(link = 'logit'),
      data = isotopes,
      method = 'REML')

# fraction of nitrogen in the samples
m_frac_n <-
  gam(frac_n ~
        s(depth_m, k = 15) +
        s(depth_m, basin, k = 10, bs = 'fs') +
        s(depth_m, replicate, k = 5, bs = 'fs'),
      family = betar(link = 'logit'),
      data = isotopes,
      method = 'REML')
```

```
# C:N ratio in the samples
m_c_n <-
  gam(c_n_ratio ~
        s(depth_m, k = 15) +
        s(depth_m, basin, k = 10, bs = 'fs') +
        s(depth_m, replicate, k = 5, bs = 'fs'),
      # log(0) = -Inf, log(Inf) = Inf
      family = Gamma(link = 'log'),
      data = isotopes,
      method = 'REML')
```

6.2 Prediction example

```
# predicting using m_c_n model (given a dataset `new_data`)
bind_cols(
  new_data, # bind data used for predictions to the predictions
  predict(object = m_c_n, # model from which to predict
          newdata = new_data, # "new" data for predictions
          type = 'link', # predict on the log scale
          se.fit = TRUE)) %>% # include the standard error
  mutate(mu = exp(fit), # move mean to the response scale
         lwr = exp(fit - 1.96 * se.fit), # create 95% CIs, then
         upr = exp(fit + 1.96 * se.fit)) # move to the response scale
```

6.3 Supporting figure

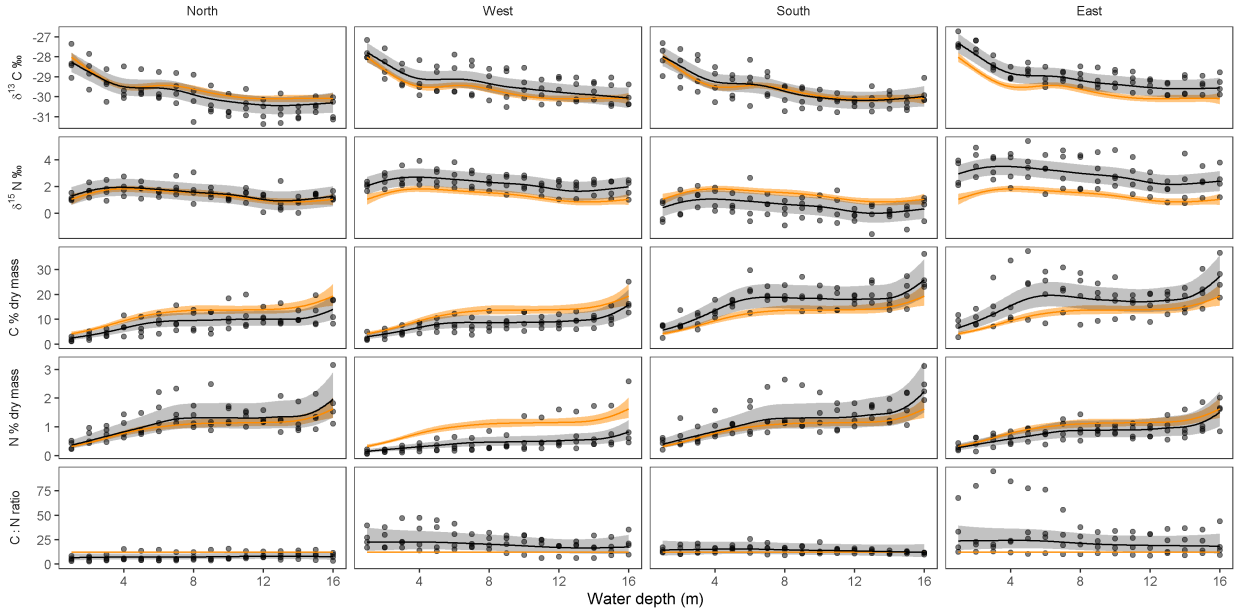

Figure 1: Simulated data and resulting smooths with 95% Bayesian credible intervals. The data was simulated starting from the data presented by Gushulak *et al.* (2021) and adding Brownian motion to each basin and each sample. The smooths and 95% CIs from the models of Gushulak *et al.* (2021) are superimposed in orange.

7 References

Bunting L., Leavitt P.R., Weidman R.P. & Vinebrookeb R.D. (2010). Regulation of the nitrogen biogeochemistry of mountain lakes by subsidies of terrestrial dissolved organic matter and the implications for climate studies. *Limnology and Oceanography* **55**, 333–345. https://doi.org/10.4319/lo.2010.55.1.0333

Bürkner P.-C. (2017). Brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software* **80**. https://doi.org/10.18637/jss.v080.i01

Glew JohnR. (1991). Miniature gravity corer for recovering short sediment cores. *Journal of Paleolimnology* **5**. https://doi.org/10.1007/BF00200351

Gushulak C.A.C., Haig H.A., Kingsbury M.V., Wissel B., Cumming B.F. & Leavitt P.R. (2021). Effects of spatial variation in benthic phototrophs along a depth gradient on assessments of whole-lake processes. *Freshwater Biology*, fwb.13820. https://doi.org/10.1111/fwb.13820

Jensen J.L.W.V. (1906). Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica* **30**, 175–193. https://doi.org/10.1007/BF02418571

Kingsbury M.V., Laird K.R. & Cumming B.F. (2012). Consistent patterns in diatom assemblages and diversity measures across water-depth gradients from eight Boreal lakes from north-western Ontario (Canada): Diatom assemblages and diversity in relation to lake depth in boreal lakes. *Freshwater Biology* **57**, 1151–1165. https://doi.org/10.1111/j.1365-2427.2012.02781.x

Pedersen E.J., Miller D.L., Simpson G.L. & Ross N. (2019). Hierarchical generalized additive models in ecology: An introduction with mgcv. *PeerJ* **7**, e6876. https://doi.org/10.7717/peerj.6876

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reiss P.T. & Ogden T.R. (2009). Smoothing parameter selection for a class of semiparametric linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71**, 505–523. https://doi.org/10.1111/j.1467-9868.2008.00695.x

Savage C., Leavitt P.R. & Elmgren R. (2004). Distribution and retention of effluent nitrogen in surface sediments of a coastal bay. *Limnology and Oceanography* **49**, 1503–1511. https://doi.org/10.4319/lo.2004.49.5.1503

Simpson G.L. (2018). Modelling Palaeoecological Time Series Using Generalised Additive Models. *Frontiers in Ecology and Evolution* **6**, 149. https://doi.org/10.3389/fevo.2018.00149

Wickham H., Averick M., Bryan J., Chang W., McGowan L., François R., *et al.* (2019). Welcome to the Tidyverse. *Journal of Open Source Software* **4**, 1686. https://doi.org/10.21105/joss.01686

Wood S.N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semi-parametric generalized linear models. *Journal of the Royal Statistical Society (B)* **73**, 3–36

Wood S.N. (2017). *Generalized Additive Models: An Introduction with R*, 2nd edn. Chapman; Hall/CRC.