Date of last README edit: 2020-12-06

A unified dataset of co-located sewage pollution, periphyton, and benthic macroinvertebrate community and food web structure from Lake Baikal (Siberia)

Overview

Authors

Michael F. Meyer[1]*
Ted Ozersky[2]
Kara H. Woo[3]
Kirill Shchapov[2]
Aaron W. E. Galloway[4]
Julie B. Schram[4]
Daniel D. Snow[5]
Maxim A. Timofeyev[6]
Dmitry Yu. Karnaukhov[6]
Matthew R. Brousil[3]
Stephanie E. Hampton[3]

[1.]School of the Environment, Washington State University, Pullman, WA, USA
[2.]Large Lakes Observatory, University of Minnesota - Duluth, Duluth, MN, USA
[3.]Center for Environmental Research, Education, and Outreach, Washington State University, Pullman, WA, USA
[4.]Oregon Institute of Marine Biology, University of Oregon, Charleston, OR, USA
[5.]School of Natural Resources, University of Nebraska-Lincoln, Lincoln, NE, USA
[6.]Biological Research Institute, Irkutsk State University, Irkutsk, Irkutsk Oblast, Russia

*corresponding author: michael.f.meyer@wsu.edu

Abstract

Sewage released from lakeside development can introduce nutrients and micropollutants that can restructure aquatic ecosystems. Lake Baikal, the world's most ancient, biodiverse, and voluminous lake, has been experiencing localized sewage pollution from lakeside settlements. Nearby increasing filamentous algal abundance suggests benthic communities are responding to this localized pollution. We surveyed 40-km of Lake Baikal's southwestern shoreline 19-23 August 2015 for sewage indicators, including pharmaceuticals, personal care products, and microplastics with co-located periphyton, macroinvertebrate, stable isotope, and fatty acid samplings. Unique identifiers corresponding to sampling locations are retained throughout all data files to facilitate interoperability among the dataset's 150+ variables. The data are structured in a tidy format (a tabular arrangement familiar to limnologists) to encourage reuse. For Lake Baikal studies, these data can support continued monitoring and research efforts. For global studies of lakes, these data can help characterize

sewage prevalence and ecological consequences of anthropogenic disturbance across spatial scales.

The data product can be cited as:

Meyer, M.F., Ozersky, T., Woo, K.H., Shchapov, K., Galloway, A.W.E., Schram, J.B., Snow, D.D., Timofeyev, M.A., Karnaukhov, D.Yu., Brousil, M.R., Hampton, S.E. A unified dataset of co-located sewage pollution, periphyton, and benthic macroinvertebrate community and food web structure from Lake Baikal (Siberia). Environmental Data Initiative.

Directory information and structure

This main directory contains three subdirectories, each with their own version of the dataset.

1. `original_data` : Raw, unaggregated data that may have misspellings, some taxa that were poorly preserved, and some misidentified taxa. These data are cleaned and aggregated by script `00_disaggregated_data_cleaning.R` .
2. `clean_disaggregated_data` : Data are cleaned, standardized, and aggregated to the replicate-level. These are available on the Environmental Data Initiative (DOI) portal.
3. `cleaned_data` : Site-level aggregated data that have been cleaned and aggregated with script `01_data_cleaning.R` .

Scripts and workflow

Scripts in this repository

Scripts that are central to the linear workflow of the dataset's build process are numbered consecutively.

1. `00_disaggregated_data_cleaning.R` : Aggregate datasets to replicate-level, remove poorly preserved or misidentified taxa, and correct spelling
   - Inputs: Disaggregated, replicate level raw CSVs
   - Outputs: Replicate-level CSVs for each type of data collected within a new directory
2. `01_data_cleaning.R` : Aggregate datasets to site-level for analyses
   - Inputs: Disaggregated, replicate level CSVs and KML file derived from Google Earth project
   - Outputs: Site-level CSVs for each type of data collected within a new directory
3. `02_sewage_indicator_analysis.R` : Performs sewage indicator analyses
   - Inputs: Site-level aggregated data from `01_data_cleaning.R`
   - Outputs: Individual and combined plots for each regression analysis.
4. `03_community_composition_analysis.R` : Performs univariate and multivariate periphyton and benthic macroinvertebrate community composition analyses
   - Inputs: Site-level aggregated data from `01_data_cleaning.R`
   - Outputs: Individual and combined univariate and multivariate plots
5. `04_fatty_acid_analysis.R` : Performs univariate and multivariate analyses pertaining to primary producer and benthic macroinvertebrate fatty acid data
   - Inputs: Site and species-level aggregated data from `01_data_cleaning.R`

    - Outputs: CSV tables for particular fatty acid analyses as well as plots resulting from various univariate and multivariate analyses
6. `05_table_formatting.R` : Aggregate metadata and sewage indicators for accompanying manuscripts
    - Inputs: Site-level aggregated data from `01_data_cleaning.R`
    - Outputs: Metadata and sewage indicator tables for accompanying manuscripts
7. `06_map_making.R` : Create map of study site based on metadata and inverse distance weighted population
    - Inputs: Site-level aggregated data from `01_data_cleaning.R`
    - Outputs: Map of study region
8. `07_stable_isotope_biplot.R` : Make biplot based of stable isotope values
    - Inputs: Site-level aggregated data from `01_data_cleaning.R`
    - Outputs: Biplot of stable isotope data, aggregated by site and grouped IDW population values
9. `panel_cor_function.R` : Sourced script that performs pairwise correlations between variables and calculates R^2 as well as p-values.

R session info:

The following R packages are essential to produce the dataset:

- tidyverse
- lubridate
- stringr
- janitor
- sf
- spdplyr

Detailed R session info is below:

```
- Session info -------------------------------------------------------------------------------------------------------
 setting  value
 version  R version 3.6.2 (2019-12-12)
 os       Windows 10 x64
 system   x86_64, mingw32
 ui       RStudio
 language (EN)
 collate  English_United States.1252
 ctype    English_United States.1252
 tz       America/Los_Angeles
 date     2020-10-15

- Packages -----------------------------------------------------------------------------------------------------------
 package     * version date       lib source
 assertthat    0.2.1   2019-03-21 [1] CRAN (R 3.6.2)
 backports     1.1.8   2020-06-17 [1] CRAN (R 3.6.2)
 broom         0.5.3   2019-12-14 [1] CRAN (R 3.6.2)
 cellranger    1.1.0   2016-07-27 [1] CRAN (R 3.6.2)
 class         7.3-15  2019-01-01 [2] CRAN (R 3.6.2)
 classInt      0.4-2   2019-10-17 [1] CRAN (R 3.6.2)
 cli           2.0.2   2020-02-28 [1] CRAN (R 3.6.3)
 colorspace    1.4-1   2019-03-18 [1] CRAN (R 3.6.1)
 crayon        1.3.4   2017-09-16 [1] CRAN (R 3.6.2)
 DBI           1.1.0   2019-12-15 [1] CRAN (R 3.6.2)
 dbplyr        1.4.2   2019-06-17 [1] CRAN (R 3.6.2)
 dplyr       * 1.0.0   2020-05-29 [1] CRAN (R 3.6.3)
 e1071         1.7-3   2019-11-26 [1] CRAN (R 3.6.2)
 ellipsis      0.3.1   2020-05-15 [1] CRAN (R 3.6.3)
 fansi         0.4.1   2020-01-08 [1] CRAN (R 3.6.2)
 forcats     * 0.4.0   2019-02-17 [1] CRAN (R 3.6.2)
 fs            1.3.1   2019-05-06 [1] CRAN (R 3.6.2)
 generics      0.0.2   2018-11-29 [1] CRAN (R 3.6.2)
 ggplot2     * 3.3.2   2020-06-19 [1] CRAN (R 3.6.3)
 glue          1.4.1   2020-05-13 [1] CRAN (R 3.6.3)
 gtable        0.3.0   2019-03-25 [1] CRAN (R 3.6.2)
```

```
haven         2.2.0   2019-11-08 [1] CRAN (R 3.6.2)
hms           0.5.3   2020-01-08 [1] CRAN (R 3.6.2)
httr          1.4.1   2019-08-05 [1] CRAN (R 3.6.3)
janitor     * 1.2.1   2020-01-22 [1] CRAN (R 3.6.2)
jsonlite      1.7.0   2020-06-25 [1] CRAN (R 3.6.3)
KernSmooth    2.23-16 2019-10-15 [2] CRAN (R 3.6.2)
lattice       0.20-38 2018-11-04 [2] CRAN (R 3.6.2)
lazyeval      0.2.2   2019-03-15 [1] CRAN (R 3.6.2)
lifecycle     0.2.0   2020-03-06 [1] CRAN (R 3.6.3)
lubridate   * 1.7.4   2018-04-11 [1] CRAN (R 3.6.2)
magrittr      1.5     2014-11-22 [1] CRAN (R 3.6.2)
modelr        0.1.5   2019-08-08 [1] CRAN (R 3.6.2)
munsell       0.5.0   2018-06-12 [1] CRAN (R 3.6.2)
nlme          3.1-142 2019-11-07 [2] CRAN (R 3.6.2)
pillar        1.4.4   2020-05-05 [1] CRAN (R 3.6.3)
pkgconfig     2.0.3   2019-09-22 [1] CRAN (R 3.6.2)
purrr       * 0.3.4   2020-04-17 [1] CRAN (R 3.6.3)
R6            2.4.1   2019-11-12 [1] CRAN (R 3.6.2)
Rcpp          1.0.5   2020-07-06 [1] CRAN (R 3.6.3)
readr       * 1.3.1   2018-12-21 [1] CRAN (R 3.6.2)
readxl        1.3.1   2019-03-13 [1] CRAN (R 3.6.2)
reprex        0.3.0   2019-05-16 [1] CRAN (R 3.6.2)
rlang         0.4.6   2020-05-02 [1] CRAN (R 3.6.3)
rstudioapi    0.11    2020-02-07 [1] CRAN (R 3.6.3)
rvest         0.3.5   2019-11-08 [1] CRAN (R 3.6.2)
scales        1.1.1   2020-05-11 [1] CRAN (R 3.6.3)
sessioninfo   1.1.1   2018-11-05 [1] CRAN (R 3.6.2)
sf          * 0.9-5   2020-07-14 [1] CRAN (R 3.6.3)
sp          * 1.3-2   2019-11-07 [1] CRAN (R 3.6.2)
spbabel       0.5.0   2019-01-08 [1] CRAN (R 3.6.2)
spdplyr     * 0.3.0   2019-05-13 [1] CRAN (R 3.6.2)
stringi       1.4.6   2020-02-17 [1] CRAN (R 3.6.2)
stringr     * 1.4.0   2019-02-10 [1] CRAN (R 3.6.2)
tibble      * 3.0.1   2020-04-20 [1] CRAN (R 3.6.3)
tidyr       * 1.1.0   2020-05-20 [1] CRAN (R 3.6.3)
tidyselect    1.1.0   2020-05-11 [1] CRAN (R 3.6.3)
tidyverse   * 1.3.0   2019-11-21 [1] CRAN (R 3.6.3)
units         0.6-5   2019-10-08 [1] CRAN (R 3.6.2)
vctrs         0.3.1   2020-06-05 [1] CRAN (R 3.6.3)
withr         2.2.0   2020-04-20 [1] CRAN (R 3.6.3)
xml2          1.3.2   2020-04-23 [1] CRAN (R 3.6.3)
```

Funding

Funding was provided by the National Science Foundation (NSF-DEB-1136637) to S.E.H., a Fulbright Fellowship to M.F.M., a NSF Graduate Research Fellowship to M.F.M. (NSF-DGE-1347973), and the Russian Ministry of Education and Science Research Project (N FZZE-2020-0026).

Acknowledgements

We would like to thank the faculty, students, staff, and mariners of the Irkutsk State University's Biological Research Institute Biostation for their expert field, taxonomic, and laboratory support; Marianne Moore and Bart De Stasio for helpful advice; the researchers and students of the Siberian Branch of the Russian Academy of Sciences Limnological Institute for expert taxonomic and logistical assistance; Oleg A. Timoshkin, Tatiana Ya. Sitnikova, Irina V. Mekhanikova, and Vadim V. Takhteev for offering insights and taxonomic training throughout the development of this project. Funding was provided by the National Science Foundation (NSF-DEB-1136637) to S.E.H., a Fulbright Fellowship to M.F.M., a NSF Graduate Research Fellowship to M.F.M. (NSF-DGE-1347973), and Russian Ministry of Science and Education (N FZZE-2020-0026). This work serves as one

chapter of M.F.M.’s doctoral dissertation in Environmental and Natural Resource Sciences at Washington State University.