The Sewage-Indicator-Baikal-Ecological-Response (SIBER) dataset: Co-located sewage pollution, periphyton, and benthic macroinvertebrate community and foodweb dataset from Lake Baikal (Siberia)

Michael F. Meyer1\*

Stephanie E. Hampton2

Ted Ozersky3

Kara H. Woo2

Kirill Shchapov3

Daniel D. Snow4

Emma J. Rosi5

Maxim A. Timofeyev6

Yulia M. Zaitseva7

Dmitry Yu. Karnaukhov6

Nina A. Bondarenko7

Aaron Galloway8

Julie Schram8

Matthew R. Brousil2

1. School of the Environment, Washington State University, Pullman, WA, USA

2. Center for Environmental Research, Education, and Outreach, Washington State University,

Pullman, WA, USA

3. Large Lakes Observatory, University of Minnesota - Duluth, Duluth, MN, USA

4. School of Natural Resources, University of Nebraska-Lincoln, Lincoln, NE, USA

5. Cary Institute of Ecosystem Studies, Millbrook, NY, USA

6. Biological Research Institute, Irkutsk State University, Irkutsk, Irkutsk Oblast, Russia

7. Limnological Institute SB RAS, Irkutsk, Irkutsk Oblast, Russia

8. Oregon Institute of Marine Biology, University of Oregon, Charleston, OR, USA

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

**Author Contribution Statement**

Conceptualized the project:

Collected samples in the field:

Processed samples:

Wrote and Reviewed R scripts:

Data management:

Wrote and edited the manuscript:

Approved the final manuscript:

**Grant sponsor information:**

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 (No. GR 01201461929; 1354-2014/51).

**Key Words:** pharmaceuticals, microplastics, fatty acids, stables isotopes, amphipod, mollusk, diatom, spirogyra

**URL of the Dataset and Metadata with permanent identifier:**

EDI will provide the DOI

**Code URL with permanent identifier:**

EDI will provide the DOI

**Measurement(s):** Chlorophyll a, Fatty Acids, Pharmaceuticals and Personal Care Products, Microplastics, Periphyton community abundance, benthic macroinvertebrate abundance, Stable Isotopes, nitrate, ammonium, total phosphorus

**Technology Type(s): NA**

**Temporal range:** 19 – 23 August 2015

**Frequency or sampling interval:** single snapshot in time

**Spatial scale:** site-based

**Abstract (144 of 150 words)**

Treated and untreated sewage released from lakeside development can introduce nutrients and micropollutants into aquatic systems. Lake Baikal, the world’s most ancient, biodiverse, and voluminous lake, has been experiencing localized sewage pollution from lakeside settlements, near which increasing filamentous algal abundance has suggested that littoral benthic communities are responding. We surveyed 40-km of Baikal’s southwestern shoreline 19-23 August 2015 for sewage indicators, including pharmaceuticals and microplastics with co-located periphyton and macroinvertebrate relative abundance and indicators of food web structure (stable isotopes and fatty acids). Unique identifiers are retained throughout all data files to facilitate interoperability among the dataset’s 125 variables. The data are structured in a tidy format (a tabular arrangement familiar to limnologists) to encourage future reuse. For Baikal, these data can supplement continued monitoring efforts. For other lakes, these data can help synthesize sewage prevalence, intensity, and ecological consequences across spatial scales.

**Background and Motivation**

Globally, sewage pollution is a common and often concentrated source of nitrogen and phosphorus inputs that can reshape aquatic ecosystems. Sewage inputs are often associated with increased primary production (Edmondson 1970; Moore et al. 2003), which can eventually cause nuisance algal blooms (Hall et al. 1999; Lapointe et al. 2015). Even in instances where sewage pollution is mitigated, restoring systems can be complicated and necessitate system-specific (Jeppesen et al. 2005), long-term mitigation strategies (Hall et al. 1999; Tong et al. 2020). As such, effective sewage monitoring can require merging a suite of chemical, biological, and ecological data to synthesize locations and timing of inputs as well as associated shifts in local ecological communities (Rosenberger et al. 2008; Hampton et al. 2011).

While ecological consequences can be stark, definitively identifying sewage pollution can be challenging. Nutrients can originate from multiple sources, such as agriculture (Powers et al. 2016) or melting permafrost (Turetsky et al. 2000), which can obfuscate wastewater signals. Unlike nutrients, sewage-specific micropollutants, such as enhanced δ15N stable isotope signatures (Costanzo et al. 2001; Camilleri and Ozersky 2019), pharmaceuticals and personal care products (PPCPs) (Bendz et al. 2005; Rosi-Marshall and Royer 2012) and microplastics (Barnes et al. 2009), have garnered global attention for their usefulness in identifying presence and quantifying magnitude of wastewater inputs. While indicators may accumulate preferentially in certain taxa (Gartner et al. 2002; Green 2016; Vendel et al. 2017; Richmond et al. 2018) and even cause deleterious effects, acutely dangerous concentrations are not common in most environmental systems, although chronic exposure to microplastics and PPCPs at minute concentrations (e.g., ug/L) can disrupt ecological processes (Richmond et al. 2017), such as oxazepam increasing feeding rate and decreasing sociability of river perch (Brodin et al. 2013) or microplastics releasing dissolved organic carbon for microbial communities (Romera-Castillo et al. 2018).

When assessing biological responses to increased nutrient loading, littoral, benthic algal and macroinvertebrate communities often respond most markedly, as their physical proximity to the shoreline puts them closer to the source (Rosenberger et al. 2008; Hampton et al. 2011). Filamentous algae, for example, can quickly increase abundance near sewage loading (Rosenberger et al. 2008; Hampton et al. 2011). As algal communities change, foodwebs can also restructure to include different feeding groups, such as detritivores, or altered nutritional value. Among the suite of food quality metrics, availability of essential fatty acids (EFAs) offers a nuanced understanding of food quality as primary producers usually maintain a consistent EFA signatures (Taipale et al. 2013) and consumers acquire EFAs by grazing (Dalsgaard et al. 2003) or trophic upgrading (Sargent and Falk-Petersen 1988; Dalsgaard et al. 2003).

Together, food web structure, community composition, and sewage indicator data can be powerful tools to assess biological impacts of sewage pollution. Despite their utility, these data are not often available for many limnological systems. PPCPs, for example, have historically been less frequently observed in lake environment (Meyer et al. 2019). In instances where data are available, efficiently merging disparate data into a single, analytically-friendly format can be challenging and require extensive computational workflows (Meyer et al. 2020).

To offer a template for harmonizing sewage indicator and biological data, we present the Sewage-Indicator-Baikal-Ecological-Responses (SIBER) dataset, which contains disparate data collected from 14 littoral and 3 pelagic sites at Lake Baikal between 19 through 23 August 2015 (Figure 1). Located in Siberia, Lake Baikal is the oldest, most voluminous, and deepest freshwater lake in the world (Hampton et al. 2018), with the majority of Baikal’s biodiversity occurring in the littoral zone (Kozhova and Izmest’eva 1998). While Lake Baikal’s pelagic zone is generally ultra-oligotrophic (Yoshida et al. 2003; O’Donnell et al. 2017), littoral areas abutting lakeside settlements have recently shown distinct signs of eutrophication (Timoshkin et al. 2016; Volkova et al. 2018). SIBER consists of over 125 variables collected at 14 littoral and 3 pelagic sampling sites. The dataset was structured in a tidy format, where each row is a sample, each column is a variable, and each CSV file is an observable unit (Wickham 2014). Independent CSV files can be merged using unique locational identifiers, enabling future researchers to customize analyses around a particular suite of variables. As a result of SIBER’s interoperability, reproducibility, and extensive variable content, the dataset is well poised for future reuse as supporting evidence of low-level sewage pollution in Lake Baikal. Additionally, SIBER’s flexibility and consistent structure enables it to be merged with similar datasets, so as to synthesize biological responses to sewage across systems and scales.

**Data Description**

The final SIBER data are openly accessible from the Environmental Data Initiative (EDI). Data are provided as 11 separate CSV files, each structed in a tabular format and containing a “site” column that can be used to merge tables. The repository also contains a compressed folder of R scripts, which were used in the main analysis of the dataset (analysis\_scripts.tar.gz) and described in detail in the companion manuscript (Meyer et al., 20XX). Future users wishing to use the data are encouraged to cite both the EDI data product as well as the data paper.

chlorophylla.csv

This file contains chlorophyll a concentrations as well as fluorometric corrections for each littoral and pelagic sampling location.

*site*

Unique alphanumeric identifier for a sampling location.

*replicate*

Replicate number.

*filtered\_volume\_ml*

Lake water volume filtered in milliliters for a given replicate.

*sample\_volume\_ml*

Sample volume filtered for chlorophyll a extraction.

*raw\_fluo*

Raw, uncorrected fluorometric reading for chlorophyll analysis.

*adjusted\_raw*

Corrected fluorometric reading for chlorophyll analysis.

*chl\_conc*

Chlorophyll a concentration in milligrams per liter.

distance\_weighted\_population\_metrics.csv

This file contains human population data for each sampled location. Although the majority of sites do not have adjacent shoreline human developments, we calculated inverse distance weighted (IDW) population for each sampling location. IDW population is a generalized representation of the size of and proximity to a sampling location’s neighboring human settlements. A full description of the methods used to calculate IDW population can be found in the companion manuscript Meyer et al. (20XX).

*site*

Unique alphanumeric identifier for a sampling location.

*distance\_weighted\_population*

Inverse distance weighted population for a given sampling location and estimated as number of people. Because this interpolation process is a function of the size of and proximity to neighboring developed sites, values can contain decimal values.

fatty\_acid.csv

This file contains fatty acid concentrations for various benthic macroinvertebrate genera, periphyton, and endemic *Drapa spp.* collected from the 14 littoral sampling locations.

*site*

Unique alphanumeric identifier for a sampling location.

*Genus*

Genus of the analyzed organism.

*Species*

Species of the analyzed organism. When organism was identified solely to genus, the Species value is NA.

*c12\_0*

Concentration of 12:0 fatty acid as micrograms of fatty acid per milligram of tissue.

*i\_14\_0*

Concentration of i-14:0 fatty acid as micrograms of fatty acid per milligram of tissue.

*c14\_0*

Concentration of 14:0 fatty acid as micrograms of fatty acid per milligram of tissue.

*c14\_4n5*

Concentration of 14:4n-5 fatty acid as micrograms of fatty acid per milligram of tissue.

*i\_15\_0*

Concentration of i-15:0 fatty acid as micrograms of fatty acid per milligram of tissue.

*a\_15\_0*

Concentration of a-15:0 fatty acid as micrograms of fatty acid per milligram of tissue.

*c15\_1w7*

Concentration of 15:1ω7 fatty acid as micrograms of fatty acid per milligram of tissue.

*i\_16\_0*

Concentration of i-16:0 fatty acid as micrograms of fatty acid per milligram of tissue.

*c16\_0*

Concentration of 16:0 fatty acid as micrograms of fatty acid per milligram of tissue.

*c16\_1w9*

Concentration of 16:1ω9 fatty acid as micrograms of fatty acid per milligram of tissue.

*c16\_1w8*

Concentration of 16:1ω8 fatty acid as micrograms of fatty acid per milligram of tissue.

*c16\_1w7*

Concentration of 16:1ω7 fatty acid as micrograms of fatty acid per milligram of tissue.

*c16\_1w6*

Concentration of 16:1ω6 fatty acid as micrograms of fatty acid per milligram of tissue.

*c16\_1w5*

Concentration of 16:1ω5 fatty acid as micrograms of fatty acid per milligram of tissue.

*i\_17\_0*

Concentration of i-17:0 fatty acid as micrograms of fatty acid per milligram of tissue.

*a\_17\_0*

Concentration of a-17:0 fatty acid as micrograms of fatty acid per milligram of tissue.

*c17\_0*

Concentration of 17:0 fatty acid as micrograms of fatty acid per milligram of tissue.

*c17\_1n7*

Concentration of 17:1n-7 fatty acid as micrograms of fatty acid per milligram of tissue.

*c16\_2w7*

Concentration of 16:2ω7 fatty acid as micrograms of fatty acid per milligram of tissue.

*c16\_2w6*

Concentration of 16:2ω6 fatty acid as micrograms of fatty acid per milligram of tissue.

*c16\_2w4*

Concentration of 16:2ω4 fatty acid as micrograms of fatty acid per milligram of tissue.

*c16\_3w6*

Concentration of 16:3ω6 fatty acid as micrograms of fatty acid per milligram of tissue.

*c16\_3w4*

Concentration of 16:3ω4 fatty acid as micrograms of fatty acid per milligram of tissue.

*c16\_3w3*

Concentration of 16:3ω3 fatty acid as micrograms of fatty acid per milligram of tissue.

*c16\_4w3*

Concentration of 16:4ω3 fatty acid as micrograms of fatty acid per milligram of tissue.

*c16\_4w1*

Concentration of 16:4ω1 fatty acid as micrograms of fatty acid per milligram of tissue.

*c18\_0*

Concentration of 18:0 fatty acid as micrograms of fatty acid per milligram of tissue.

*c18\_1w9*

Concentration of 18:1ω9 fatty acid as micrograms of fatty acid per milligram of tissue.

*c18\_1w7*

Concentration of 18:1ω7 fatty acid as micrograms of fatty acid per milligram of tissue.

*c18\_2w6t*

Concentration of 18:2ω6t fatty acid as micrograms of fatty acid per milligram of tissue.

*c18\_2w6*

Concentration of 18:2ω6 fatty acid as micrograms of fatty acid per milligram of tissue.

*c18\_3w6*

Concentration of 18:3ω6 fatty acid as micrograms of fatty acid per milligram of tissue.

*c18\_3w3*

Concentration of 18:3ω3 fatty acid as micrograms of fatty acid per milligram of tissue.

*c18\_4w4*

Concentration of 18:4ω4 fatty acid as micrograms of fatty acid per milligram of tissue.

*c18\_4w3*

Concentration of 18:4ω3 fatty acid as micrograms of fatty acid per milligram of tissue.

*c18\_5w3*

Concentration of 18:5ω3 fatty acid as micrograms of fatty acid per milligram of tissue.

*c20\_0*

Concentration of 20:0 fatty acid as micrograms of fatty acid per milligram of tissue.

*c20\_1w9*

Concentration of 20:1ω9 fatty acid as micrograms of fatty acid per milligram of tissue.

*c20\_1w7*

Concentration of 20:1ω7 fatty acid as micrograms of fatty acid per milligram of tissue.

*c20\_2\_5\_11*

Concentration of 20:2-5-11 fatty acid as micrograms of fatty acid per milligram of tissue.

*c20\_2\_5\_13*

Concentration of 20:2-5-13 fatty acid as micrograms of fatty acid per milligram of tissue.

*c20\_2w6*

Concentration of 20:2ω6 fatty acid as micrograms of fatty acid per milligram of tissue.

*c20\_3w6*

Concentration of 20:3ω6 fatty acid as micrograms of fatty acid per milligram of tissue.

*c20\_4w6*

Concentration of 20:4ω6 fatty acid as micrograms of fatty acid per milligram of tissue.

*c20\_3w3*

Concentration of 20:3ω3 fatty acid as micrograms of fatty acid per milligram of tissue.

*c20\_4w3*

Concentration of 20:4ω3 fatty acid as micrograms of fatty acid per milligram of tissue.

*c20\_5w3*

Concentration of 20:5ω3 fatty acid as micrograms of fatty acid per milligram of tissue.

*c22\_0*

Concentration of 22:0 fatty acid as micrograms of fatty acid per milligram of tissue.

*c22\_1w9*

Concentration of 22:1ω9 fatty acid as micrograms of fatty acid per milligram of tissue.

*c22\_1w7*

Concentration of 22:1ω7 fatty acid as micrograms of fatty acid per milligram of tissue.

*c22\_2w6*

Concentration of 22:2ω6 fatty acid as micrograms of fatty acid per milligram of tissue.

*c22\_4w6*

Concentration of 22:4ω6 fatty acid as micrograms of fatty acid per milligram of tissue.

*c22\_5w6*

Concentration of 22:5ω6 fatty acid as micrograms of fatty acid per milligram of tissue.

*c22\_3w3*

Concentration of 22:3ω3 fatty acid as micrograms of fatty acid per milligram of tissue.

*c22\_4w3*

Concentration of 22:4ω3 fatty acid as micrograms of fatty acid per milligram of tissue.

*c22\_5w3*

Concentration of 22:5ω3 fatty acid as micrograms of fatty acid per milligram of tissue.

*c22\_6w3*

Concentration of 22:6ω3 fatty acid as micrograms of fatty acid per milligram of tissue.

*c24\_0*

Concentration of 24:0 fatty acid as micrograms of fatty acid per milligram of tissue.

*comments*

Quality flag column. Two samples spilled during fatty acid extraction. These samples are flagged as such. Although concentrations are lower than other samples, proportions between fatty acids are consistent.

invertebrates.csv

This file contains abundance for benthic macroinvertebrates collected at each of the 14 littoral sampling locations. Only amphipod taxa were identified to species. Mollusks and isopods were identified to genus.

*site*

Unique alphanumeric identifier for a sampling location.

*replicate*

Replicate for sampling location. While three replicates were collected in the field, some samples were poorly preserved and not enumerated so as to prevent potential errors.

*Acroloxidae*

Mollusk genus

*Asellidae*

Endemic isopod genus

*Baicaliidae*

Mollusk genus, most of which are endemic

*Benedictidae*

Mollusc genus, most of which are endemic

*Brandtia\_latissima*

Endemic amphipod species

*Brandtia\_parasitica\_parasitica*

Endemic amphipod species

*Caddisflies*

General grouping; were not identified to species.

*Cryptoropus\_inflatus*

Endemic amphipod species

*Cryptoropus\_pachytus*

Endemic amphipod species

*Cryptoropus\_rugosus*

Endemic amphipod species

*Eulimnogammarus\_capreolus*

Endemic amphipod species

*Eulimnogammarus\_cruentes*

Endemic amphipod species

*Eulimnogammarus\_cyaneus*

Endemic amphipod species

*Eulimnogammarus\_grandimanus*

Endemic amphipod species

*Eulimnogammarus\_juveniles*

Endemic amphipod genus. Identification kept at genus level so as to prevent misclassification.

*Eulimnogammarus\_maackii*

Endemic amphipod species

*Eulimnogammarus\_marituji*

Endemic amphipod species

*Eulimnogammarus\_verucossus*

Endemic amphipod species

*Eulimnogammarus\_viridis\_viridis*

Endemic amphipod species

*Eulimnogammarus\_vittatus*

Endemic amphipod species

*Flatworms*

Not identified beyond order.

*Leeches*

Not identified beyond order, although 12 endemic species do exist.

*Maackia*

Mollusk genus, most of which are endemic

*Pallasea\_brandtia\_brandita*

Endemic amphipod species

*Pallasea\_brandtii\_tenera*

Endemic amphipod species

*Pallasea\_cancelloides*

Endemic amphipod species

*Pallasea\_cancellus*

Endemic amphipod species

*Pallasea\_viridis*

Endemic amphipod species

*Planorbidae*

Mollusk genus, most of which are endemic

*Poekilogammarus\_crassimus*

Endemic amphipod species

*Poekilogammarus\_ephippiatus*

Endemic amphipod species

*Poekilogammarus\_juveniles*

Endemic amphipod genus. Identifying to species introduced risk of misclassification.

*Poekilogammarus\_megonychus\_perpolitus*

Endemic amphipod species

*Poekilogammarus\_pictus*

Endemic amphipod species

*Valvatidae*

Mollusk genus, most of which are endemic

metadata.csv

This file contains metadata for each of the pelagic and littoral sampling locations. Missing data are assigned as NA.

*site*

Unique alphanumeric identifier for a sampling location.

*year*

Year sampling occurred.

*month*

Month sampling occurred.

*day*

Day sampling occurred.

*time*

Time sampling occurred as Hours:Minutes.

*lat*

Latitude of sampling location in decimal degrees.

*long*

Longitude of sampling location in decimal degrees.

*site\_description*

Researchers’ description of sampling location at the time of sampling.

*distance\_to\_shore\_m*

Distance from *in situ* sampled location to the shoreline in meters.

*depth\_m*

Depth at *in situ* sampling location in meters.

*air\_temp\_celsius*

Temperature of air at sampling location in Celsius.

*surface\_temp\_celsius*

Temperature of water’s surface at sampling location in Celsius.

*mid\_temp\_celsius*

Temperature of water midway between surface and bottom at sampling location in Celsius.

*bottom\_temp\_celsius*

Temperature of water near sediment at sampling location in Celsius.

*comments*

Notes in the field describing sampling conditions.

*shore\_photo*

Whether or not photos of the shoreline were taken. Photos are available on the project’s Open Science Framework page (DOI).

*substrate\_photo*

Whether or not photos of the substrate were taken.

*sponges*

Whether or not sponges were present at a sampling location.

*brandtia*

Whether or not *Brandtia spp.* (endemic amphipod) was present at a sampling location.

microplastics.csv

This file contains microplastics counts for each of the pelagic and littoral sampling locations.

*site*

Unique alphanumeric identifier for a sampling location.

*replicate*

Replicate for a given sampling location.

*fragments*

Number of microplastic fragments observed.

*fibers*

Number of microplastic fibers observed.

*beads*

Number of microplastic beads observed.

*comments*

Observer comments while enumerating microplastics

*volume\_filtered\_ml*

Volume in milliliters for a given replicate filtered.

nutrients.csv

This file contains nutrient concentrations for each of the associated sampling locations. Nutrient samples were not filtered prior to analysis, meaning that nitrogen concentrations have the potential to be spurious.

*site*

Unique alphanumeric identifier for a sampling location.

*replicate*

Replicate for a given sampling location.

*nh4\_mg\_dm3*

Ammonium concentration in milligrams of ammonium per cubic decimeter.

*no3\_mg\_dm3*

Nitrate concentration in milligrams of nitrate per cubic decimeter

*tp\_mg\_dm3*

Total phosphorus concentration in milligrams of phosphorus per cubic decimeter.

*tpo43\_mg\_dm3*

Total phosphate concentration as phosphate in milligrams per cubic decimeter.

periphyton.csv

This file contains periphyton abundance data for each of the sampled littoral locations. For poorly preserved samples, counts are detailed as NA for each taxonomic grouping, and a note in the “comments” column is provided.

*site*

Unique alphanumeric identifier for a sampling location.

*replicate*

Replicate number for a given sampling site.

*subsamples\_counted*

Number of 10 microliter subsamples counted for a given replicate.

*diatom*

Number of diatom cells counted for a given replicate.

*spirogyra*

Number of *Spirogyra spp.* cells counted for a given replicate.

*spirogyra\_filaments*

Number of *Spirogyra spp.* filaments counted for a given replicate.

*ulothrix*

Number of *Ulothrix spp.* cells counted for a given replicate.

*ulothrix\_filaments*

Number of *Ulothrix spp.*filaments counted for a given replicate.

*tetrasporales*

Number of *Tetrasporales* spp. cells counted for a given replicate

*pediastrum*

Number of *Pediastrum spp.*cells counted for a given replicate.

*desmidales*

Number of *Desmidales* spp. cells counted for a given replicate.

*lyngbya*

Number of *Lyngbya spp.* cells counted for a given replicate.

*comments*

Notes from the observer.

ppcp.csv

This file contains Pharmaceutical and Personal Care Product (PPCP) concentrations for each littoral and pelagic sampling location.

*site*

Unique alphanumeric identifier for a sampling location.

*paraxanthine*

Concentration of paraxanthine, also known as 1,7-dimethylxanthine, in nanograms per liter. Paraxanthine is the main human metabolite of caffeine

*acetaminophen*

Concentration of acetaminophen, also known as paracetamol, in nanograms per liter.

*amphetamine*

Concentration of amphetamine in nanograms per liter.

*caffeine*

Concentration of caffeine in nanograms per liter.

*carbamezapine*

Concentration of carbamezapine in nanograms per liter.

*cimetidine*

Concentration of cimetidine in nanograms per liter.

*cotinine*

Concentration of cotinine, which is the main human metabolite of nicotine, in nanograms per liter.

*diphenhydramine*

Concentration of diphenhydramine in nanograms per liter.

*mda*

Concentration of mda in nanograms per liter.

*mdma*

Concentration of mdma in nanograms per liter.

*methamphetamine*

Concentration of methamphetamine in nanograms per liter.

*morphine*

Concentration of morphine in nanograms per liter.

*phenazone*

Concentration of phenazone in nanograms per liter.

sulfachloropyridazine

Concentration of sulfachloropyridazine in nanograms per liter.

sulfamethoxazole

Concentration of sulfamethoxazole in nanograms per liter.

*thiabendazole*

Concentration of thiabendazole in nanograms per liter.

*trimethoprim*

Concentration of trimethoprim in nanograms per liter.

*collection\_year*

Year sample was collected in the field.

*collection\_month*

Month sample was collected in the field.

*collection\_day*

Day sample was collected in the field.

*analysis\_year*

Year sample was analyzed.

*analysis\_month*

Month sample was analyzed.

*analysis\_day*

Day sample was analyzed.

stable\_isotopes.csv

This file contains carbon (δ13C) and nitrogen (δ15N) values for various benthic macroinvertebrate genera, periphyton, and endemic *Drapa spp* collected from the 14 littoral sampling locations.

*site*

Unique alphanumeric identifier for a sampling location.

*C13*

Carbon (δ13C) stable isotope values in parts per thousand.

*N15*

Nitrogen (δ15N) stable isotope values in parts per thousand.

*Genus*

Genus of the analyzed organism.

*Species*

Species of the analyzed organism. When organism was identified solely to genus, the Species value is NA.

total\_lipid.csv

*site*

Unique alphanumeric identifier for a sampling location.

*Genus*

Genus of the analyzed organism.

*Species*

Species of the analyzed organism. When organism was identified solely to genus, the Species value is NA.

*total\_lipid\_mg\_per\_g*

Total amount of lipids in a sample in milligrams of Lipid per gram of tissue.

*deviation*

Samples were weighed three times and deviation in measurement was calculated. All values are reported in milligrams of Lipid per gram of tissue.

*comments*

Quality flag column. Two samples spilled during fatty acid extraction. These samples are flagged as such.

**Methods**

*Inverse distance weighted (IDW) population calculation for each sampling location*

Our workflow for calculating IDW population can be described in five main steps. First, we traced polygons and shorelines from satellite imagery for each developed site in Google Earth. Second, polygon and line shapefiles were downloaded from Google Earth as a .kml file. Third, the .kml file was imported into the R statistical environment (R Core Team, 2019), where using the sf package (Pebesma, 2018) we calculated shoreline length, polygon area, and centroid location for each developed site. Fourth, we joined point locations of each sampling site with the spatial polygons to calculate the distance from each sampling location to each developed site’s centroid. Fifth, we calculated IDW population for each location, using formula (1).

(1)

where *I* is the IDW population at sampling location *j*, *P* is the population at each of the three developed sites Listvyanka (LI), Bolshie Koty (BK), Bolshoe Goloustnoe (BGO), *A* is the area of a developed site in km2, *L* is the shoreline length at a developed site in km, and *D* is the distance from developed site *j* to each developed site in km.

*Nutrients*

For ammonium (2016a) and nitrate (2017) concentrations, samples were analyzed with a spectrophotometer following the addition of Nessler’s reagent and disulfuric acid respectively. Phosphate and total phosphorus concentration was measured with a spectrophotometer following the addition of persulfate (2016b).

*Chlorophyll a*

Chlorophyll samples were processed in a manner similar to that of Parson (1963) and Lorenzen (1967). Nitrocellulose filters were ground in 90% acetone, in which chlorophyll extraction was allowed to proceed overnight. Samples were then centrifuged for 15-20 minutes. After centrifugation, absorbance of the chlorophyll extract was measured in a spectrophotometer at 630, 645, 665, and 750 nm. Concentrations were calculated using the formula: ; where A is the absorbance value of a particular wavelength, V1 is the volume of the filtered water, and V2 is the volume of extract.

*Pharmaceuticals and Personal Care Products (PPCPs)*

Within 12 h of collection, samples were filtered directly from the amber glass bottle using a single-stream 25-mm GF/F SPE cartridge setup (Waters Corporation, Milford, MA). Lab personnel wore gloves and face masks to minimize contamination. Prior to filtration, GF/F filters and SPE cartridges were primed with at least 5 mL of either methanol or acetone and then washed with at least 5 mL of sample water. Rate of SPE occurred at approximately 1 drop per second. Extraction proceeded until water could no longer pass through the SPE cartridge or until all collected water was filtered. Cartridges were stored in whirlpacks at -20°C until analysis for PPCPs following methods of Lee et al. (2016).

*Microplastics*

Each sample was vacuum filtered on to a 47-mm diameter GF/F filter. During filtration, aluminum foil was used to cover the filtration funnel to prevent potential contamination from airborne microplastic particles. After filtration, filters were dried under vacuum pressure and then stored in 50-mm petri dishes. Following filtration of all three replicates, the filtrate was collected and then re-filtered through a GF/F filter as a control for contamination.

Microplastic counting involved visual inspection of the entire GF/F according to guidelines proposed in Van Cauwenberghe et al. (2015). Visual enumeration was conducted under a stereo microscope at approximately 100x magnification, and microplastics were classified into one of three categories: fibers, fragments, or beads. For all categories, plastics were defined as observed objects with apparently artificial colors. Fibers were defined as smooth, long plastics with consistent diameters. Fragments were defined as plastics with irregularly sharp or jagged edges. Beads were defined as spherical plastics. During enumeration, GF/Fs remained in the petri dish to minimize potential for contamination from the air. Following enumeration of both experimental and control samples, fibers, fragments, and beads enumerated in the controls were subtracted from the experimental microplastic densities for each plastic type and from each replicate.

*Periphyton abundance*

We haphazardly selected three rocks representative of local substrate. A plastic stencil was used to define a surface area of each rock from which we scraped a standardized 14.5 cm2 patch of periphyton. Samples were preserved with Lugol’s solution and stored in plastic scintillation vials. Periphyton taxonomic identification and enumeration was performed by subsampling 10 μL aliquots from each preserved sample. Cells, filaments, and colonies were counted for each taxonomic group until at least a total of 300 cells were identified. For all subsamples, the entire aliquot was counted. In instances where 300 cells were counted prior to completing the aliquot, the entire aliquot was still counted.

*Benthic macroinvertebrate abundance*

Three kick-net samples were collected for assessment of benthic community composition and abundance. Using a D-net, we collected macroinvertebrates by flipping over 1-3 rocks, and then sweeping five times in a left-to-right motion across approximately 1 m. After the series of sweeps, the catch was rinsed into a plastic bucket. For each replicate, bucket contents were concentrated using a 64 μm mesh and placed in glass jars with 40% ethanol (vodka; the only preservative available to us at the time) for preservation and refrigerated at 4°C aboard the research vessel. The 40% ethanol preservative was replaced with ~80% ethanol upon return to the lab within 24 to 48 hours, and samples were stored at ~4°C. For enumeration, entire samples were sorted into taxonomic groupings under a stereomicroscope.

*Stable Isotope Analysis*

Prior to isotopic analysis, samples were freeze dried for ~24 hours, homogenized to powder, and then weighed. Measurements of δ15N and δ13C were performed on an elemental analyzer-isotope ratio mass spectrometer (EA-IRMS; Finnigan DELTAplus XP, Thermo Scientific) at the Large Lakes Observatory, University of Minnesota Duluth.

*Fatty Acid Analysis*

Fatty acid extractions generally involved four phases: (1) freeze dried for ~24 hours, homogenized to powder, and then weighed, (2)100% chloroform extraction, (3) chloroform-methanol extraction, and (4) fatty acid methylation. Fatty acid extraction methods were adapted from similar methods developed in Schram et al. (2018). Final fatty acid quantification was performed with a Shimadzu QP2020 GC/MS in a similar method as described in Schram et al. (2018).

**Technical Validation**

The SIBER dataset had three main validation procedures: taxonomic, analytical, and reproducible.

For taxonomic validation, all phylogenetic groupings were based off most recent identification keys. Amphipods were identified according to Taakhteev & Didorenko (2015). Mollusks were identified according to Sitnikova (2012). Algal taxa were identified according to Izhboldina (2007). For consistency, all taxa were identified by one person (Michael F. Meyer), who was trained by experts in Baikal taxonomy.

For analytical validation, internal standards were used for all mass-spectroscopy analyses. PPCPs analyses involved an internal nicotinamide standard to assess yield. Stable isotope values were calibrated against certified reference materials including L-glutamic acid (NIST SRM 8574), low organic soil and sorghum flour (standards B-2153 and B-2159 from Elemental Micro-analysis Ltd., Okehampton, UK) and in-house standards (acetanilide and caffeine). Replicate analyses of external standards showed a mean standard deviation of 0.06 ‰ and 0.09 ‰, for δ13C and δ15N, respectively. Finally, fatty acid estimations used a known internal 19:0 standard to assess oxidation of fatty acids during extraction, methylation, and quantification.

For data reproducibility, we systematically reviewed all R code to format and aggregate source data. An initial cleaning script (00\_disaggregated\_data\_cleaning.R) removed incorrect spellings, erroneous data values, and inconsistent column names from raw data. This step created the standardized .csv files detailed above, which are available on the EDI repository. Raw data files are available on the project’s Open Science Framework page (DOI) but are not included in the EDI repository to prevent confusion or incorrect usage. Data hosted on EDI are at the replicate-level but can be aggregated to the sampling-site-level using script 01\_data\_cleaning.R. In addition to aggregation scripts, five R scripts used for analyses in Meyer et al. (20XX) are also available on the EDI repository within the compressed entity “scripts.tar.gz”. All R code for data aggregation was written by one person (Michael F. Meyer) and then independently review by two others (Kara H. Woo and Matthew R. Brousil) to confirm that code performed as intended, was well documented, and annotations were complete.

**Data Use and Recommendations for Reuse**

The SIBER dataset can be applied to a suite of research questions pertaining to ecological responses to human disturbance. We highlight two main areas for immediate application.

First, SIBER can be harmonized with other littoral sampling efforts throughout Lake Baikal, so as to enhance spatial coverage and data diversity. Since 2010, Lake Baikal has experienced increasing filamentous algae, especially near larger lakeside developments (Kravtsova et al. 2014; Timoshkin et al. 2016, 2018; Volkova et al. 2018). Recent benthic algal surveys throughout Baikal’s entirety have suggested that cosmopolitan filamentous algae, such as *Spirogyra spp.*, tend to be more abundant near larger lakeside developments (Timoshkin et al. 2016; Volkova et al. 2018). For example, Listvyanka is a small town located at the beginning of the Angara River, Baikal’s only outflow. While Listvyanka’s permanent population is approximately 2,000 persons, the town is a growing tourism hub, which hosts over 1.2 million tourists per year. Surveys conducted near Listvyanka have suggested increased *Spirogyra spp.* abundance is associated with wastewater release (Timoshkin et al. 2016). Although wastewater inputs are likely low and are diluted to negligible concentrations offshore, combining monitoring efforts across spatial and temporal scales are necessary to evaluate the spatial and temporal extent of wastewater entering Baikal. As such, SIBER could compliment previous, current, and future monitoring efforts, where data may be missing.

Second, SIBER contains data useful to expanding freshwater PPCP, microplastic, and associated biological responses across large spatial scales. Recent syntheses of PPCPs’ have reported that studies involving lakes are less abundant in the literature relative to lotic systems (Meyer et al. 2019), and likewise microplastics studies have noticed freshwater environments being less represented in the literature relative to marine ecosystems (Horton et al. 2017). For both PPCPs microplastics, toxic responses to even minute concentrations can be expansive and often uncertain, thereby necessitating synthesis across systems. As a result of PPCPs and microplastics garnering increasing attention worldwide, expansive sampling of PPCPs and microplastics with co-located biological data across multiple spatial scales would be helpful to synthesize biotic responses to micropollutants across systems. Although SIBER constitutes a small number of such data that exist, SIBER data are highly structured and flexible for merging with similar datasets. Additionally, SIBER’s sequential, integrated harmonization workflow could be adopted by similar monitoring efforts, thereby facilitating ease of data harmonization and interoperability. Through integration with similar monitoring efforts, SIBER can contribute to global synthesis of emerging contaminant consequences, especially in a region of the world that is often not easily accessible to many researchers.

Given its potential for future applications, SIBER serves not only as a resource for monitoring wastewater entering Lake Baikal, but also provides data useful to synthesizing spatial extents of emerging contaminants across systems and spatial scales.

***Optional section:* Comparison with existing datasets**

[Authors are encouraged to compare the dataset with existing datasets if there are others that are similar. It is recommended that the authors describe the components of this dataset that set it apart from already published datasets. This section can also serve to point future researchers to other similar types of data that some may be looking for.]

**References**

Barnes, D. K. A., F. Galgani, R. C. Thompson, and M. Barlaz. 2009. Accumulation and fragmentation of plastic debris in global environments. Philos Trans R Soc Lond B Biol Sci **364**: 1985–1998. doi:10.1098/rstb.2008.0205

Bendz, D., N. A. Paxéus, T. R. Ginn, and F. J. Loge. 2005. Occurrence and fate of pharmaceutically active compounds in the environment, a case study: Höje River in Sweden. Journal of Hazardous Materials **122**: 195–204. doi:10.1016/j.jhazmat.2005.03.012

Brodin, T., J. Fick, M. Jonsson, and J. Klaminder. 2013. Dilute Concentrations of a Psychiatric Drug Alter Behavior of Fish from Natural Populations. Science **339**: 814–815. doi:10.1126/science.1226850

Camilleri, A. C., and T. Ozersky. 2019. Large variation in periphyton δ13C and δ15N values in the upper Great Lakes: Correlates and implications. Journal of Great Lakes Research **45**: 986–990. doi:10.1016/j.jglr.2019.06.003

Costanzo, S. D., M. J. O’Donohue, W. C. Dennison, N. R. Loneragan, and M. Thomas. 2001. A New Approach for Detecting and Mapping Sewage Impacts. Marine Pollution Bulletin **42**: 149–156. doi:10.1016/S0025-326X(00)00125-9

Dalsgaard, J., M. St. John, G. Kattner, D. Müller-Navarra, and W. Hagen. 2003. Fatty acid trophic markers in the pelagic marine environment, p. 225–340. *In* Advances in Marine Biology. Elsevier.

Edmondson, W. T. 1970. Phosphorus, Nitrogen, and Algae in Lake Washington after Diversion of Sewage. Science **169**: 690–691.

Gartner, A., P. Lavery, and A. J. Smit. 2002. Use of delta N-15 signatures of different functional forms of macroalgae and filter-feeders to reveal temporal and spatial patterns in sewage dispersal. Mar. Ecol.-Prog. Ser. **235**: 63–73. doi:10.3354/meps235063

Green, D. S. 2016. Effects of microplastics on European flat oysters, Ostrea edulis and their associated benthic communities. Environmental Pollution **216**: 95–103. doi:10.1016/j.envpol.2016.05.043

Hall, R. I., P. R. Leavitt, R. Quinlan, A. S. Dixit, and J. P. Smol. 1999. Effects of agriculture, urbanization, and climate on water quality in the northern Great Plains. Limnology and Oceanography **44**: 739–756. doi:10.4319/lo.1999.44.3\_part\_2.0739

Hampton, S. E., S. C. Fradkin, P. R. Leavitt, and E. E. Rosenberger. 2011. Disproportionate importance of nearshore habitat for the food web of a deep oligotrophic lake. Marine and Freshwater Research **62**: 350. doi:10.1071/MF10229

Hampton, S. E., S. McGowan, T. Ozersky, and others. 2018. Recent ecological change in ancient lakes. Limnology and Oceanography **63**: 2277–2304. doi:10.1002/lno.10938

Horton, A. A., A. Walton, D. J. Spurgeon, E. Lahive, and C. Svendsen. 2017. Microplastics in freshwater and terrestrial environments: Evaluating the current understanding to identify the knowledge gaps and future research priorities. Science of The Total Environment **586**: 127–141. doi:10.1016/j.scitotenv.2017.01.190

Jeppesen, E., M. Søndergaard, J. P. Jensen, and others. 2005. Lake responses to reduced nutrient loading – an analysis of contemporary long-term data from 35 case studies. Freshwater Biology **50**: 1747–1771. doi:10.1111/j.1365-2427.2005.01415.x

Kozhova, O. M., and L. R. Izmest’eva. 1998. Lake Baikal: Evolution and Biodiversity, Backhuys Publishers.

Kravtsova, L. S., L. A. Izhboldina, I. V. Khanaev, and others. 2014. Nearshore benthic blooms of filamentous green algae in Lake Baikal. Journal of Great Lakes Research **40**: 441–448. doi:10.1016/j.jglr.2014.02.019

Lapointe, B. E., L. W. Herren, D. D. Debortoli, and M. A. Vogel. 2015. Evidence of sewage-driven eutrophication and harmful algal blooms in Florida’s Indian River Lagoon. Harmful Algae **43**: 82–102. doi:10.1016/j.hal.2015.01.004

Meyer, M. F., S. G. Labou, A. N. Cramer, M. R. Brousil, and B. T. Luff. 2020. The global lake area, climate, and population dataset. Scientific Data **7**: 174. doi:10.1038/s41597-020-0517-4

Meyer, M. F., S. M. Powers, and S. E. Hampton. 2019. An Evidence Synthesis of Pharmaceuticals and Personal Care Products (PPCPs) in the Environment: Imbalances among Compounds, Sewage Treatment Techniques, and Ecosystem Types. Environ. Sci. Technol. **53**: 12961–12973. doi:10.1021/acs.est.9b02966

Moore, J. W., D. E. Schindler, M. D. Scheuerell, D. Smith, and J. Frodge. 2003. Lake eutrophication at the urban fringe, Seattle region, USA. AMBIO: A Journal of the Human Environment **32**: 13–18.

O’Donnell, D. R., P. Wilburn, E. A. Silow, L. Y. Yampolsky, and E. Litchman. 2017. Nitrogen and phosphorus colimitation of phytoplankton in Lake Baikal: Insights from a spatial survey and nutrient enrichment experiments. Limnology and Oceanography **62**: 1383–1392. doi:10.1002/lno.10505

Powers, S. M., T. W. Bruulsema, T. P. Burt, and others. 2016. Long-term accumulation and transport of anthropogenic phosphorus in three river basins. Nature Geoscience **9**: 353–356. doi:10.1038/ngeo2693

Richmond, E. K., M. R. Grace, J. J. Kelly, A. J. Reisinger, E. J. Rosi, and D. M. Walters. 2017. Pharmaceuticals and personal care products (PPCPs) are ecological disrupting compounds (EcoDC). Elem Sci Anth **5**: 66. doi:10.1525/elementa.252

Richmond, E. K., E. J. Rosi, D. M. Walters, J. Fick, S. K. Hamilton, T. Brodin, A. Sundelin, and M. R. Grace. 2018. A diverse suite of pharmaceuticals contaminates stream and riparian food webs. Nature Communications **9**: 4491. doi:10.1038/s41467-018-06822-w

Romera-Castillo, C., M. Pinto, T. M. Langer, X. A. Álvarez-Salgado, and G. J. Herndl. 2018. Dissolved organic carbon leaching from plastics stimulates microbial activity in the ocean. Nat Commun **9**: 1–7. doi:10.1038/s41467-018-03798-5

Rosenberger, E. E., S. E. Hampton, S. C. Fradkin, and B. P. Kennedy. 2008. Effects of shoreline development on the nearshore environment in large deep oligotrophic lakes. Freshwater Biology **53**: 1673–1691. doi:10.1111/j.1365-2427.2008.01990.x

Rosi-Marshall, E. J., and T. V. Royer. 2012. Pharmaceutical Compounds and Ecosystem Function: An Emerging Research Challenge for Aquatic Ecologists. Ecosystems **15**: 867–880. doi:10.1007/s10021-012-9553-z

Sargent, J. R., and S. Falk-Petersen. 1988. The lipid biochemistry of calanoid copepods. Hydrobiologia **167–168**: 101–114. doi:10.1007/BF00026297

Taipale, S., U. Strandberg, E. Peltomaa, A. W. E. Galloway, A. Ojala, and M. T. Brett. 2013. Fatty acid composition as biomarkers of freshwater microalgae: analysis of 37 strains of microalgae in 22 genera and in seven classes. Aquatic Microbial Ecology **71**: 165–178. doi:10.3354/ame01671

Timoshkin, O. A., M. V. Moore, N. N. Kulikova, and others. 2018. Groundwater contamination by sewage causes benthic algal outbreaks in the littoral zone of Lake Baikal (East Siberia). Journal of Great Lakes Research. doi:10.1016/j.jglr.2018.01.008

Timoshkin, O. A., D. P. Samsonov, M. Yamamuro, and others. 2016. Rapid ecological change in the coastal zone of Lake Baikal (East Siberia): Is the site of the world’s greatest freshwater biodiversity in danger? Journal of Great Lakes Research **42**: 487–497. doi:10.1016/j.jglr.2016.02.011

Tong, Y., M. Wang, J. Peñuelas, and others. 2020. Improvement in municipal wastewater treatment alters lake nitrogen to phosphorus ratios in populated regions. Proc Natl Acad Sci USA **117**: 11566–11572. doi:10.1073/pnas.1920759117

Turetsky, M. R., R. K. Wieder, C. J. Williams, and D. H. Vitt. 2000. Organic matter accumulation, peat chemistry, and permafrost melting in peatlands of boreal Alberta. Écoscience **7**: 115–122. doi:10.1080/11956860.2000.11682608

Vendel, A. L., F. Bessa, V. E. N. Alves, A. L. A. Amorim, J. Patrício, and A. R. T. Palma. 2017. Widespread microplastic ingestion by fish assemblages in tropical estuaries subjected to anthropogenic pressures. Marine Pollution Bulletin **117**: 448–455. doi:10.1016/j.marpolbul.2017.01.081

Volkova, E. A., N. A. Bondarenko, and O. A. Timoshkin. 2018. Morphotaxonomy, distribution and abundance of *Spirogyra* (Zygnematophyceae, Charophyta) in Lake Baikal, East Siberia. Phycologia **57**: 298–308. doi:10.2216/17-69.1

Wickham, H. 2014. Tidy Data. Journal of Statistical Software **59**: 1–23. doi:10.18637/jss.v059.i10

Yoshida, T., T. Sekino, M. Genkai-Kato, and others. 2003. Seasonal dynamics of primary production in the pelagic zone of southern Lake Baikal. Limnology **4**: 53–62. doi:10.1007/s10201-002-0089-3

2016a. Methods for determination of nitrogen-containing matters (with corrections) (Методы определения азотсодержащих веществ (с Поправками)).

2016b. Methods for determination of phosphorus-containing matters (with corrections) (Методы определения фосфорсодержащих веществ).

2017. Nitrate concentration in waters: Photometric methods with Giress reagent following stabilization in a cadmium reducer (Массовая концентрация нитратного азота в водах: Методика измерений фотометрическим методом с реактивом Грисса после восстановления в камиевом редукторе).

**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.

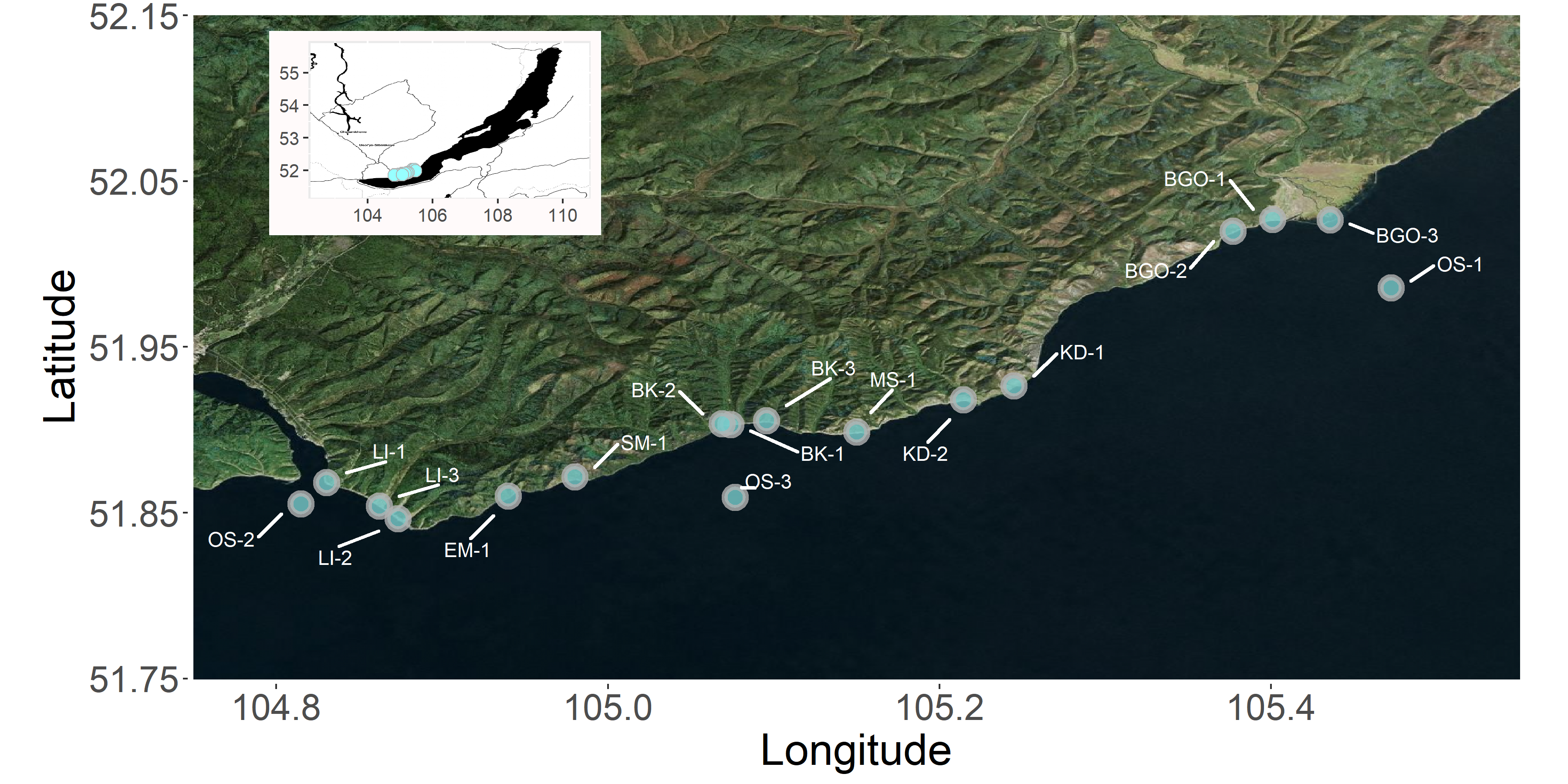


Figure 1: Map of all sampling locations with sites labeled with alphanumeric code. The entire transect included three developed sites (i.e., Listvyanka (LI), Bolshie Koty (BK), Bolshoe Goloustnoe (BGO)). Three offshore sites (OS) were also sampled to compare pelagic sewage signals to those in the littoral. Sites without adjacent lakeside development included Emelyanikha Bay (EM), Maloe Kadilnoe (KD), Mys Soboliny (MS), Sredny Mys (SM). Littoral sampling locations were all 8.9-20.75 m from shore and at a depth approximately of 0.75 m, whereas pelagic sites were approximately 2-5 km from shore and ranged in depth from 900 to 1300 m.