Diatom Phosphorus Index: An index for nutrient enrichment impact assessment in wadeable streams based on diatom species composition

Nina Desianti

5/29/2024

Abstract

Wisconsin Department of Natural Resources has developed an index to assess the impact of phosphorus enrichment in wadeable streams based on diatom species composition (Diatom Phosphorus Index, DPI) in response to the required water quality standards assessment for the Wisconsin State under the terms of the Clean Water Act. DPI employed the weighted averaging statistical method on a suite of diatom species composition and total phosphorus (TP) training datasets collected in 2001 to 2003 and 2011 from 197 streams in Wisconsin to infer streams' TP values. In this study, we developed a refined DPI model using larger datasets collected from 2016 to 2023 that represent updates of previous phosphorus levels in Wisconsin and applied the weighted averaging partial least square method (WA-PLS). The refined DPI model has a higher accuracy in predicting TP values compared to the previous model. It shows that larger training datasets have improved the performance of DPI, aside from the implementation of WA-PLS. TP levels of streams in Wisconsin varied throughout the years but exhibited a consistent spatial distribution pattern that may be related to historical nutrient legacies, consistent nutrient sources, and inherent differences in the nutrient removal capacity in each stream. A combined approach for determining phosphorus concentration criteria that incorporates DPI may have an advantage over the general phosphorus criteria as it reflects not only the nutrient status of the streams but also their biological integrity.

Introduction

Eutrophication or nutrient enrichment in lakes, streams, and rivers from either point or diffuse sources has caused excessive plant and algal growth, as well as low dissolved oxygen that leads to fish kills [1-2]. With the current projection of climate change impacts, such as a reduction in summer flows and higher water temperatures, the risk of algal blooms, including harmful algae that potentially release toxins in waterbodies may increase [3]. Excess phosphorus as the main driver of eutrophication is the most widespread stressor in streams, rivers, and reservoirs within the United States [1,4]. In Wisconsin, the impaired waterbodies with total phosphorus concentration (TP) above the criteria for the maximum limit or allowable maximum according to the Water Quality Standards rule (75 µg/L or within 75 to <150 µg/L range for "combined approach") [5] reached about seven percent of the total assessed water bodies; often accompanied by excess algal growth, particularly in lakes and reservoirs [6].

The Wisconsin Department of Natural Resources (WDNR) has implemented phosphorus assessment procedures that incorporate biological metrics which is termed the "combined assessment" approach [5]. Under this procedure, the measured phosphorus concentration in a waterbody is evaluated alongside its phosphorus response indicators, such as algae and plant metrics. These metrics are then used to determine if a waterbody is showing a biological response to phosphorus. In the case of a waterbody exceeding the statewide phosphorus criterion but does not show a biological or recreational use impairment, it would not be considered impaired as specified under section 303 (d) of the Clean Water Act, 33 USC 1313 [5,7]. For stream assessment, the phosphorus response indicators include benthic algal biomass screening and Diatom Phosphorus Index (DPI). DPI was developed using the weighted averaging statistical method on a suite of diatom

species composition and TP training data sets from the 197 stream sites in Wisconsin to infer TP values of streams, as part of the Nutrient Impacts study conducted by Robertson, et al. (2008) [5,8]. For purposes of assessing the attainment of the phosphorus response indicator, DPI results are then compared to the stream phosphorus criterion of 75 µg.L-1 phosphorus. If only one diatom sample per site is available, the confidence interval approach described under s. NR 102.52 (2) (c) is applied. If the DPI is below 75 µg/L as specified under s. NR 102.52 (2) (c) 1., the phosphorus response indicator is attained. If more than one diatom sample is available from the most recent 5 years, the mean score of the surveys is calculated and compared to the threshold of 75 µg/L without applying confidence intervals.

Wisconsin has continued to make progress in minimizing nutrient losses to water and improving the water quality of lakes, rivers, streams, and groundwater within the state through Wisconsin's Nutrient Reduction Strategy [9].The use of the "Wisconsin Pollutant Discharge Elimination System" permits led to a 70% reduction of phosphorus loads between 1995 and 2018; "Water Quality Trading" has achieved phosphorus reduction of an average of roughly 800 lbs/year, and the "adaptive management" program has set a goal to curtail about 30,000 lbs/year of phosphorus loading within permittees' watersheds over the next five-year permit term [9]. These programs along with the ongoing Multi-Discharger Variance, Nutrient Management Plan, Nine Key Element Watershed Plan, state financial assistance/grants, and various collaborations with private and public sectors resulted in a significant reduction of total phosphorus in flowing waters for all years since 2013, and most of the reduction occurs in sites that drain to the Mississippi River basin [9]. However, phosphorus continues to be a main cause of impairments in rivers, lakes, and streams; and nutrient concentrations in most lakes have not changed over time [6,9].

Improvement of water quality in Wisconsin will require an understanding of both point and diffuse nutrient sources and sinks across different ecoregions within the State. Excess nutrients stored in catchment terrestrial soils due to decades of agriculture practices as we refer to nutrient legacies [10], can create time lags between changes in contemporary nutrient inputs and the response of waterbodies' ecosystems to these changes [11]. The analysis of nearly two decades of nutrient concentration monitoring in the U.S. revealed various temporal changes of nutrients in streams, with an increased trend of TP from 2000 to 2013, followed by a decreased trend from 2013 to 2018 as observed in the majority of the streams, yet spatial patterns were persistent across the U.S.; this trends potentially caused by the nutrient legacies, consistent nutrient sources, and inherent differences in nutrient removal capacity for various ecosystems [12]. In this study, we focused on developing refined DPI using larger training datasets collected from 2016 to 2023 that represent updates of previous phosphorus levels in Wisconsin.

Study Area and Field sampling

This study was conducted in the wadable streams across Wisconsin State (Fig. 1). Wisconsin State is bordered by two Great Lakes, Lake Michigan and Lake Superior, and an interior of forests and farms. The state has abundant water resources with approximately 1.2 million lake and impoundment acres, 1000 miles of Great Lakes shoreline, over 88,000 river and stream miles, and 5.3 million wetland acres [6]. These water bodies are present across four primary level III ecoregions in Wisconsin: Northern Lakes and Forests (NLF), North Central Hardwood Forests (NCHF), Southeastern Wisconsin Till Plains (SWTP), and Driftless Area (DFA), and also in small pieces of the Western Cornbelt Plains and the Central Cornbelt Plains ecoregions (Fig. 1.) [13]. Each ecoregion is defined by similar environmental characteristics, such as soil type, vegetation, climate, geology, water quality, and land use within its boundary [13]. Such a classification scheme allows environmental assessments, water quality, biological criteria setting, and non-point source pollution management at a regional scale. Diatoms were sampled following the WDNR Monitoring Protocols and Study Design [14]. Benthic diatoms were sampled from two pieces of natural substrates, 5 – 25 cm diameter rocks, particularly the upper surface of each piece from each of the three fast-water habitat units (i.e. riffles) for a total of six samples; or three rocks from one riffle if rocks and/or riffles are rare in the reach. Samples were taken from a depth of approximately 15 - 20 cm (if the stream is <15cm deep sample at the deepest points). Immediately after collection samples are preserved using glutaraldehyde with a final concentration of 3-5%. Diatom samples were then stored in a refrigerator until ready to be shipped to the Wisconsin State Laboratory of Hygiene (WSLH) during the summer and fall for sample preparation. In total, 479 samples were collected from wadable streams near the 248 Surface Water Integrated Monitoring System (SWIMS) stations across

different Ecoregions and analyzed for diatoms in this study. Environmental characteristics were measured from each sample location, a detailed account of these parameters, physical attributes including photographs, and GPS coordinates are publicly accessible at https://dnr.wi.gov/topic/surfacewater/swdv. Some wadeable streams were sampled multiple times from 2016 to 2023.

Dataset and sources for producing map of study area

- **wi_eco_l3.shp** is a shapefile downloaded from (https://www.epa.gov/eco-research/ecoregion-download-files-state-region-5#pane-47)
- **sampling_loc.csv** is a csv file containing GPS coordinates of each closest SWIMS ID of diatom sampling locations. *Notes: Make sure that you have installed the packages we are going to use here before you call them. You can install them using the command install.packages("sf"), install.packages("ggplot2"), and so on. Also, make sure to set the working directory to where this file is saved.You can use the command setwd("use the path to your file instead of this text").*

```
library(sf) #Call the package
```

```
## Linking to GEOS 3.12.1, GDAL 3.8.4, PROJ 9.3.1; sf_use_s2() is TRUE
```

```
#The coordinate system used in shapefile isn't lat-long.
shp <- st_read("wi_eco_l3.shp")#read shapefile into R using "sf" package.
```

```
## Reading layer `wi_eco_l3' from data source
##   `C:\Users\Nina\Documents\WDNR DPI\wi_eco_l3.shp' using driver `ESRI Shapefile'
## Simple feature collection with 6 features and 13 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: 242558.7 ymin: 2180085 xmax: 718516.7 ymax: 2688850
## Projected CRS: USA_Contiguous_Albers_Equal_Area_Conic_USGS_version
```

```
#Transform shapefile into the World Geodetic System 1984 (WGS84)
shpfile <- st_transform(shp, "+proj=longlat +ellps=WGS84 +datum=WGS84")
sampling_loc <- read.csv("sampling_loc.csv", row.names = 1) #read csv file
library(ggplot2) #Call the package
library(ggspatial) #Call the package
# Plot shapefile in ggplot as a polygon using geom_sf function
eco_map <- ggplot() +
  geom_sf(aes(fill = US_L3NAME), #color Ecoregion polygon area
              data = shpfile, #specify which data
              color = "dark red", #color Ecoregion polygon edges
              alpha = .2, #determine the opacity
              linewidth = .2) + #determine the width of the polygon edges
  labs(fill = "Ecoregion") + #rename legend title
  geom_point(data = sampling_loc, aes(long, lat)) + #plot sampling location
  theme_bw() + #uses a white background and thin grey grid lines
  theme(panel.grid = element_blank()) + #remove all the thin grey grid lines
  theme(axis.title.x = element_blank(), axis.title.y = element_blank()) + #remove axis title
  annotation_north_arrow(height = unit(0.8, "cm"), #annotate north arrow on map
  width = unit(0.8, "cm"), # specify width of the north arrow
  pad_x = unit(0.8, "cm"), # specify location of the north arrow on x axis
  pad_y = unit(1.2, "cm"), # specify location of the north arrow on y axis
  rotation = NULL, #determine the rotation of the north arrow
  style = north_arrow_orienteering) + #determine the style of the north arrow
  annotation_scale(location = "bl", width_hint = 0.3) + #annotate the scale of the map
  coord_sf(crs = 4326) #using World Geodetic System 1984 (WGS84) to determine the map scale
eco_map
```

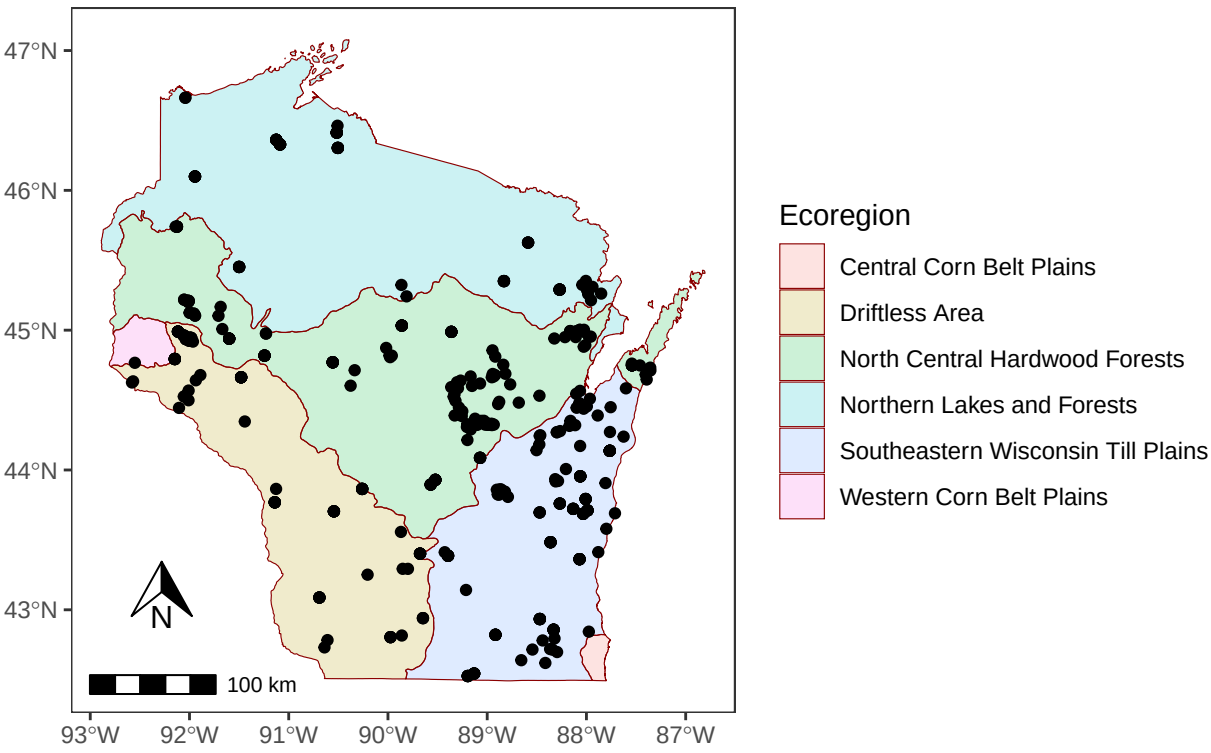

Fig. 1. Map of the study area showing closest SWIMS ID of diatom sampling locations.

Diatom sample analysis

About 40 mL of diatom slurry was subsampled and digested in 70% nitric acid using microwave heating to remove organic matter content, following WSLH's diatom analysis standard operating procedure (EHD Env Tox Method 4730). Samples were then repeatedly rinsed with reverse osmosis (RO) water until near-neutral pH. Approximately 20 to 1000 µL of the cleaned sample containing diatom frustules were dripped onto coverslips and air-dried. Coverslips were then mounted on microscope slides using Naphrax™ mounting medium (Brunel Microscopes, U.K.). Diatoms were counted and identified using a Zeiss AxioImager microscope equipped with DIC. Six hundred valves were counted for each sample at 1000x magnification. All identifications were made to species/variety level whenever possible using primarily the Diatom Atlas of Wisconsin Streams and Wetlands identification guidebook (Desianti 2024) and Freshwater Benthic Diatoms of Central Europe (Lange-Bertalot, et al. 2017). Several samples were observed under scanning electron microscopy (SEM) Zeiss Gemini 450 to identify some morphologically similar taxa.

Data analysis

Temporal and spatial change in nutrients

We assessed temporal and spatial changes in nutrient concentrations in streams within the sampling period of 2016 to 2022 across Driftless Area (DFA), North Central Hardwood Forests (NCHF), Northern Lakes and Forests (NLF), and Southeastern Wisconsin Till Plains (SWTP) ecoregions in Wisconsin. One-way Analysis of variance (One-way ANOVA) tests were conducted to test the statistical differences in total phosphorus (TP), total nitrogen (TN), nitrate (NO3) + nitrite (NO2), and ammonia (NH3) concentration across different years of sampling in each ecoregion. Post-hoc Tukey's HSD pairwise comparisons were used to test the significant differences among years of sampling in each ecoregion. To investigate the spatial distribution of nutrients, one-way ANOVA tests were also conducted to test the statistical differences in TP, TN, NO3+NO2, and NH3 across different ecoregions regardless of the year of sampling. Post-hoc Tukey's HSD pairwise comparisons were used to test the significant difference across different ecoregions.

Gradient analysis of diatom assemblages

The diatom distribution pattern was investigated using a non-metric multidimensional scaling (NMDS) ordination technique based on diatom relative abundance data using a subset of data with available pH, conductivity, TP, TN, NO3+NO2, and NH3. This subset of data consists of 436 diatom taxa from 121 samples of 70 SWIMS stations. A rank index analysis was performed prior to NMDS (Faith et al. 1987) and Gower distance was selected as a suitable dissimilarity index to be applied in the subsequent NMDS analysis. A stress value was calculated which reflects the amount of error in the correlation between pairwise distances in the original matrix and a matrix calculated with the NMDS. Stress values of 0.1 indicate excellent representation in reduced dimensions, 0.2 good, and values 0.3 provide a poor representation (Clarke & Warwick 1994). The effect of singleton removal was tested by comparing NMDS for datasets with and without singletons. Singletons are defined as species that occur (1) in only one sample (in this case 93 taxa were removed), (2) in less than 5% of samples (245 taxa were removed), and (3) in less than 10% of samples (301 taxa were removed). Procrustes analysis (999 permutations) was used to identify the significance of the congruence between the ordinations with singleton removal and the ordination on the complete data set (Peres-Neto and Jackson 2001). Only the singleton removal of taxa in less than 10% of the sample produced a somewhat different ordination (Procrustes correlation coefficient = 0.71, p-value $<$ 0.01), and this configuration has a slightly higher stress value (0.22) compared to all taxa included (0.19) or removal of taxa occurred in less than 5% (0.22); therefore, all taxa were included in the NMDS analysis.

Diatom Phosphorus Index development

Transfer functions also known as inference models were developed for inferring TP from the composition of diatom assemblages using training datasets of diatom and TP collected from 2016 to 2023. Data are publicly available at https://dnr.wi.gov/topic/surfacewater/swdv. We subsetted the data to examine the correlation between diatom assemblages and TP data collected on the same day, one to seven days before, eight to fourteen days before diatom sampling, average of six months TP data, and selected TP data. We used the weighted averaging partial least squares (WA-PLS) method (Birks et al. 2012) for constructing the transfer functions that produced better-performing models compared to the classical weighted averaging method. Bootstrapping was chosen as a cross-validation procedure. The models' performance was estimated by coefficients of determination (R2 and R2 boot) and the root square mean errors of prediction (RMSEP and RMSEPboot). Average of TP with selected data (sample removal with TP values detected as outliers) has the best performance and we used this model to develop DPI. Diatom species optima and tolerances for TP were calculated as weighted averages and weighted standard deviations (Birks et al. 2012). All numerical analyses were conducted and graphics were generated using R 4.4.0 software (http://www.r-project.org) using packages "vegan" (Oksanen et al. 2013), "Rioja" (Juggins 2017), and "ggplot2" (Wickham 2009).

Results

Change in nutrient concentrations from 2016 to 2022

The temporal analysis of nutrient concentrations showed differing patterns among TP, TN, NO3+NO2, and NH3. There were negligible changes in NH3 concentrations ($p > 0.05$) across ecoregions during 2016-2022. Other nutrient parameters, such as NO3+NO2, TN, and TP showed variability in concentration throughout the year, particularly in North Central Hardwood Forests and Southern Wisconsin Till Plains for NO3+NO2 and TP; and in the Driftless Area, North Central Hardwood Forests, and Northern Lakes and Forests for TN. There was a decrease in NO3+NO2 from 2016 to 2019 and then an increase from 2019 to 2022 in North Central Hardwood Forests and Southern Wisconsin Till Plains. TN followed the same pattern as NO3+NO2, especially in the North Central Hardwood Forests ecoregion, whereas in the Driftless Area, a significant increase was observed in 2020. TP steadily decreased, the pronounced trend observed especially in the Southern Wisconsin Till Plains ecoregion.

Dataset of Water Quality Parameters

All_WQ_2016_2022 is a csv file containing all water quality parameters

```r
library(multcompView) #Call the package
library(tidyverse) #Call the package
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
wq <- read.csv("All_WQ_2016_2022.csv", header = T) #read csv file
#removing Ecoregion that has only 1 sampling site.
wq2 <- wq[-which(wq$Ecoregion=='Western Corn Belt Plains'),]
wq2$Year <- as.character(wq2$Year)

#Subsetting data based on Parameter Name
tp <- wq2[wq2$PramName=="TP",]
tn <- wq2[wq2$PramName=="TN",]
nh3 <- wq2[wq2$PramName=="NH3_N",]
no3 <- wq2[wq2$PramName=="NO2+NO3",]

#Subsetting TP data based on Ecoregion
tp_DA <- tp[tp$Ecoregion=="Driftless Area",]
tp_SWTP <- tp[tp$Ecoregion=="Southeastern Wisconsin Till Plains",]
tp_NLF <- tp[tp$Ecoregion=="Northern Lakes and Forests",]
tp_NCHF <- tp[tp$Ecoregion=="North Central Hardwood Forests",]

#Subsetting TN data based on Ecoregion
tn_DA <- tn[tn$Ecoregion=="Driftless Area",]
tn_SWTP <- tn[tn$Ecoregion=="Southeastern Wisconsin Till Plains",]
tn_NLF <- tn[tn$Ecoregion=="Northern Lakes and Forests",]
tn_NCHF <- tn[tn$Ecoregion=="North Central Hardwood Forests",]

#Subsetting NO2+NO3 data based on Ecoregion
no3_DA <- no3[no3$Ecoregion=="Driftless Area",]
no3_SWTP <- no3[no3$Ecoregion=="Southeastern Wisconsin Till Plains",]
no3_NLF <- no3[no3$Ecoregion=="Northern Lakes and Forests",]
no3_NCHF <- no3[no3$Ecoregion=="North Central Hardwood Forests",]

#Subsetting NH3 data based on Ecoregion
nh3_DA <- nh3[nh3$Ecoregion=="Driftless Area",]
nh3_SWTP <- nh3[nh3$Ecoregion=="Southeastern Wisconsin Till Plains",]
nh3_NLF <- nh3[nh3$Ecoregion=="Northern Lakes and Forests",]
nh3_NCHF <- nh3[nh3$Ecoregion=="North Central Hardwood Forests",]

#One-way ANOVA for TP data
aov_tp_DA <- aov(result ~ Year, data = tp_DA)
aov_tp_SWTP <- aov(result ~ Year, data = tp_SWTP)
aov_tp_NLF <- aov(result ~ Year, data = tp_NLF)
aov_tp_NCHF <- aov(result ~ Year, data = tp_NCHF)
```

```r
#One-way ANOVA for TN data
aov_tn_DA <- aov(result ~ Year, data = tn_DA)
aov_tn_SWTP <- aov(result ~ Year, data = tn_SWTP)
aov_tn_NLF <- aov(result ~ Year, data = tn_NLF)
aov_tn_NCHF <- aov(result ~ Year, data = tn_NCHF)

#One-way ANOVA for NO2+NO3 data
aov_no3_DA <- aov(result ~ Year, data = no3_DA)
aov_no3_SWTP <- aov(result ~ Year, data = no3_SWTP)
aov_no3_NLF <- aov(result ~ Year, data = no3_NLF)
aov_no3_NCHF <- aov(result ~ Year, data = no3_NCHF)

#One-way ANOVA for NH3 data
aov_nh3_DA <- aov(result ~ Year, data = nh3_DA)
aov_nh3_SWTP <- aov(result ~ Year, data = nh3_SWTP)
aov_nh3_NLF <- aov(result ~ Year, data = nh3_NLF)
aov_nh3_NCHF <- aov(result ~ Year, data = nh3_NCHF)

#Tukey posthoc for TP data
tp_DA_tukey <- TukeyHSD(aov_tp_DA)
tp_SWTP_tukey <- TukeyHSD(aov_tp_SWTP)
tp_NLF_tukey <- TukeyHSD(aov_tp_NLF)
tp_NCHF_tukey <- TukeyHSD(aov_tp_NCHF)

#Tukey posthoc for TN data
tn_DA_tukey <- TukeyHSD(aov_tn_DA)
tn_SWTP_tukey <- TukeyHSD(aov_tn_SWTP)
tn_NLF_tukey <- TukeyHSD(aov_tn_NLF)
tn_NCHF_tukey <- TukeyHSD(aov_tn_NCHF)

#Tukey posthoc for NO2+NO3 data
no3_DA_tukey <- TukeyHSD(aov_no3_DA)
no3_SWTP_tukey <- TukeyHSD(aov_no3_SWTP)
no3_NLF_tukey <- TukeyHSD(aov_no3_NLF)
no3_NCHF_tukey <- TukeyHSD(aov_no3_NCHF)

#Tukey posthoc for NH3 data
nh3_DA_tukey <- TukeyHSD(aov_nh3_DA)
nh3_SWTP_tukey <- TukeyHSD(aov_nh3_SWTP)
nh3_NLF_tukey <- TukeyHSD(aov_nh3_NLF)
nh3_NCHF_tukey <- TukeyHSD(aov_nh3_NCHF)

# Extract labels and factor levels from Tukey post-hoc on TP ANOVA
tp_DA_cld <- multcompLetters4(aov_tp_DA, tp_DA_tukey)
tp_DA_cld <- as.data.frame.list(tp_DA_cld$Year)
tp_DA_cld$Year <- rownames(tp_DA_cld)
tp_DA_cld['Ecoregion'] <- 'Driftless Area'
tp_SWTP_cld <- multcompLetters4(aov_tp_SWTP, tp_SWTP_tukey)
tp_SWTP_cld <- as.data.frame.list(tp_SWTP_cld$Year)
tp_SWTP_cld$Year <- rownames(tp_SWTP_cld)
tp_SWTP_cld['Ecoregion'] <- 'Southeastern Wisconsin Till Plains'
tp_NLF_cld <- multcompLetters4(aov_tp_NLF, tp_DA_tukey)
tp_NLF_cld <- as.data.frame.list(tp_NLF_cld$Year)
```

```
tp_NLF_cld$Year <- rownames(tp_NLF_cld)
tp_NLF_cld['Ecoregion'] <- 'Northern Lakes and Forests'
tp_NCHF_cld <- multcompLetters4(aov_tp_NCHF, tp_NCHF_tukey)
tp_NCHF_cld <- as.data.frame.list(tp_NCHF_cld$Year)
tp_NCHF_cld$Year <- rownames(tp_NCHF_cld)
tp_NCHF_cld['Ecoregion'] <- 'North Central Hardwood Forests'

tp_DA_cld <- tp_DA_cld[c('Letters', 'Year', 'Ecoregion')]
tp_SWTP_cld <- tp_SWTP_cld[c('Letters', 'Year', 'Ecoregion')]
tp_NLF_cld <- tp_NLF_cld[c('Letters', 'Year', 'Ecoregion')]
tp_NCHF_cld <- tp_NCHF_cld[c('Letters', 'Year', 'Ecoregion')]

tp_tukey <- rbind(tp_DA_cld, tp_SWTP_cld, tp_NLF_cld, tp_NCHF_cld)
tp_tukey$PramName <- 'TP'

# Extract labels and factor levels from Tukey post-hoc on TN ANOVA
tn_DA_cld <- multcompLetters4(aov_tn_DA, tn_DA_tukey)
tn_DA_cld <- as.data.frame.list(tn_DA_cld$Year)
tn_DA_cld$Year <- rownames(tn_DA_cld)
tn_DA_cld['Ecoregion'] <- 'Driftless Area'
tn_SWTP_cld <- multcompLetters4(aov_tn_SWTP, tn_SWTP_tukey)
tn_SWTP_cld <- as.data.frame.list(tn_SWTP_cld$Year)
tn_SWTP_cld$Year <- rownames(tn_SWTP_cld)
tn_SWTP_cld['Ecoregion'] <- 'Southeastern Wisconsin Till Plains'
tn_NLF_cld <- multcompLetters4(aov_tn_NLF, tn_DA_tukey)
tn_NLF_cld <- as.data.frame.list(tn_NLF_cld$Year)
tn_NLF_cld$Year <- rownames(tn_NLF_cld)
tn_NLF_cld['Ecoregion'] <- 'Northern Lakes and Forests'
tn_NCHF_cld <- multcompLetters4(aov_tn_NCHF, tn_NCHF_tukey)
tn_NCHF_cld <- as.data.frame.list(tn_NCHF_cld$Year)
tn_NCHF_cld$Year <- rownames(tn_NCHF_cld)
tn_NCHF_cld['Ecoregion'] <- 'North Central Hardwood Forests'

tn_DA_cld <- tn_DA_cld[c('Letters', 'Year', 'Ecoregion')]
tn_SWTP_cld <- tn_SWTP_cld[c('Letters', 'Year', 'Ecoregion')]
tn_NLF_cld <- tn_NLF_cld[c('Letters', 'Year', 'Ecoregion')]
tn_NCHF_cld <- tn_NCHF_cld[c('Letters', 'Year', 'Ecoregion')]

tn_tukey <- rbind(tn_DA_cld, tn_SWTP_cld, tn_NLF_cld, tn_NCHF_cld)
tn_tukey$PramName <- 'TN'

# Extract labels and factor levels from Tukey post-hoc on NO2+NO3 ANOVA
no3_DA_cld <- multcompLetters4(aov_no3_DA, no3_DA_tukey)
no3_DA_cld <- as.data.frame.list(no3_DA_cld$Year)
no3_DA_cld$Year <- rownames(no3_DA_cld)
no3_DA_cld['Ecoregion'] <- 'Driftless Area'
no3_SWTP_cld <- multcompLetters4(aov_no3_SWTP, no3_SWTP_tukey)
no3_SWTP_cld <- as.data.frame.list(no3_SWTP_cld$Year)
no3_SWTP_cld$Year <- rownames(no3_SWTP_cld)
no3_SWTP_cld['Ecoregion'] <- 'Southeastern Wisconsin Till Plains'
no3_NLF_cld <- multcompLetters4(aov_no3_NLF, no3_DA_tukey)
no3_NLF_cld <- as.data.frame.list(no3_NLF_cld$Year)
no3_NLF_cld$Year <- rownames(no3_NLF_cld)
```

```r
no3_NLF_cld['Ecoregion'] <- 'Northern Lakes and Forests'
no3_NCHF_cld <- multcompLetters4(aov_no3_NCHF, no3_NCHF_tukey)
no3_NCHF_cld <- as.data.frame.list(no3_NCHF_cld$Year)
no3_NCHF_cld$Year <- rownames(no3_NCHF_cld)
no3_NCHF_cld['Ecoregion'] <- 'North Central Hardwood Forests'

no3_DA_cld <- no3_DA_cld[c('Letters', 'Year', 'Ecoregion')]
no3_SWTP_cld <- no3_SWTP_cld[c('Letters', 'Year', 'Ecoregion')]
no3_NLF_cld <- no3_NLF_cld[c('Letters', 'Year', 'Ecoregion')]
no3_NCHF_cld <- no3_NCHF_cld[c('Letters', 'Year', 'Ecoregion')]

no3_tukey <- rbind(no3_DA_cld, no3_SWTP_cld, no3_NLF_cld, no3_NCHF_cld)
no3_tukey$PramName <- 'NO2+NO3'

# Extract labels and factor levels from Tukey post-hoc on NH3 ANOVA
nh3_DA_cld <- multcompLetters4(aov_nh3_DA, nh3_DA_tukey)
nh3_DA_cld <- as.data.frame.list(nh3_DA_cld$Year)
nh3_DA_cld$Year <- rownames(nh3_DA_cld)
nh3_DA_cld['Ecoregion'] <- 'Driftless Area'
nh3_SWTP_cld <- multcompLetters4(aov_nh3_SWTP, nh3_SWTP_tukey)
nh3_SWTP_cld <- as.data.frame.list(nh3_SWTP_cld$Year)
nh3_SWTP_cld$Year <- rownames(nh3_SWTP_cld)
nh3_SWTP_cld['Ecoregion'] <- 'Southeastern Wisconsin Till Plains'
nh3_NLF_cld <- multcompLetters4(aov_nh3_NLF, nh3_DA_tukey)
nh3_NLF_cld <- as.data.frame.list(nh3_NLF_cld$Year)
nh3_NLF_cld$Year <- rownames(nh3_NLF_cld)
nh3_NLF_cld['Ecoregion'] <- 'Northern Lakes and Forests'
nh3_NCHF_cld <- multcompLetters4(aov_nh3_NCHF, nh3_NCHF_tukey)
nh3_NCHF_cld <- as.data.frame.list(nh3_NCHF_cld$Year)
nh3_NCHF_cld$Year <- rownames(nh3_NCHF_cld)
nh3_NCHF_cld['Ecoregion'] <- 'North Central Hardwood Forests'

nh3_DA_cld <- nh3_DA_cld[c('Letters', 'Year', 'Ecoregion')]
nh3_SWTP_cld <- nh3_SWTP_cld[c('Letters', 'Year', 'Ecoregion')]
nh3_NLF_cld <- nh3_NLF_cld[c('Letters', 'Year', 'Ecoregion')]
nh3_NCHF_cld <- nh3_NCHF_cld[c('Letters', 'Year', 'Ecoregion')]

nh3_tukey <- rbind(nh3_DA_cld, nh3_SWTP_cld, nh3_NLF_cld, nh3_NCHF_cld)
nh3_tukey$PramName <- 'NH3_N'

#Combine Tukey posthoc results
all_tukey <- rbind(tp_tukey, tn_tukey, no3_tukey, nh3_tukey)

#Add Tukey posthoc to dataset
new_wq <- wq2 %>% inner_join(all_tukey, by=c("PramName", "Ecoregion", "Year"))
```

```r
library(scales)
```

```
##
## Attaching package: 'scales'
```

```
## The following object is masked from 'package:purrr':
##
##     discard
```

```
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
library(ggpubr)
library(ggtext)
new_wq$PramName <- as.factor(as.character(new_wq$PramName))
levels(new_wq$PramName) <- c("NH<sub>3</sub> (mg.L<sup>-1</sup>)",
                             "NO<sub>3</sub> +NO<sub>2</sub> (mg.L<sup>-1</sup>)",
                             "TN(mg.L<sup>-1</sup>)",
                             "TP (mg.L<sup>-1</sup>)")
STATS = new_wq %>% group_by(Year, Ecoregion, PramName) %>%
  summarize(Q75 = quantile(result, probs = 0.75),
            Q25 = quantile(result, probs = 0.25),
            MaxVal = max(result), .groups = "keep") %>%
  mutate(WhiskUp = 1.05* (Q75 + 1.5 * (Q75 - Q25)))
new_wq <- new_wq %>% inner_join(STATS, by=c("PramName", "Ecoregion", "Year"))

#Creating plot for WQ data
wqplot <- ggplot(data = new_wq, aes(x = Year, y = result, fill = Year)) +
  geom_boxplot(outlier.colour = alpha("black", 0.2)) +
  scale_fill_brewer(palette = "Pastel1") +
  scale_y_continuous(expand = expansion(mult = .2)) +
  theme_bw() +
  theme(legend.position = "bottom",
        axis.title.x = element_blank(),
        axis.title.y = element_blank(),
        axis.text.x = element_blank(),
        axis.ticks.x = element_blank(),
        strip.text = ggtext::element_markdown(),
        strip.background = element_blank(),
        strip.placement = "outside",
        panel.border = element_blank(),
        panel.grid.major.x = element_blank(),
        panel.grid.minor.x = element_blank(),
        panel.grid.minor.y = element_blank(),
        strip.text.y.left = element_text(face = "bold", size = 5),
        strip.text.x = ggtext::element_textbox_simple( width  = unit(1, "npc"),
                                                       height = unit(3, "lines"),
                                                       face = 'bold',
                                                       size = 10,
                                                       hjust = 0.5,
                                                       vjust = 0.5,
                                                       halign = 0.5,
                                                       valign = 0.5)) +
  facet_grid(PramName~Ecoregion,
             scales = "free",
             switch = "y") +
  guides(fill = guide_legend(title = "Year of sampling",
                             nrow = 1)) +
  stat_compare_means(method = "anova", size = 2, label.x = 1.5) +
  stat_summary(fun = mean,
               geom = "line",
               aes(group = 1),
```

```
                col = "red",
                linewidth = 1.1) +
  stat_summary(geom = "crossbar",
                width=0.65,
                fatten=0,
                color="white",
                fun.data = function(x){return(c(y=median(x),
                                                ymin=median(x),
                                                ymax=median(x)))}) +
  geom_text(aes(y = WhiskUp,
                label = Letters),
            stat = 'identity',
            check_overlap = T,
            color = "black")
wqplot
```

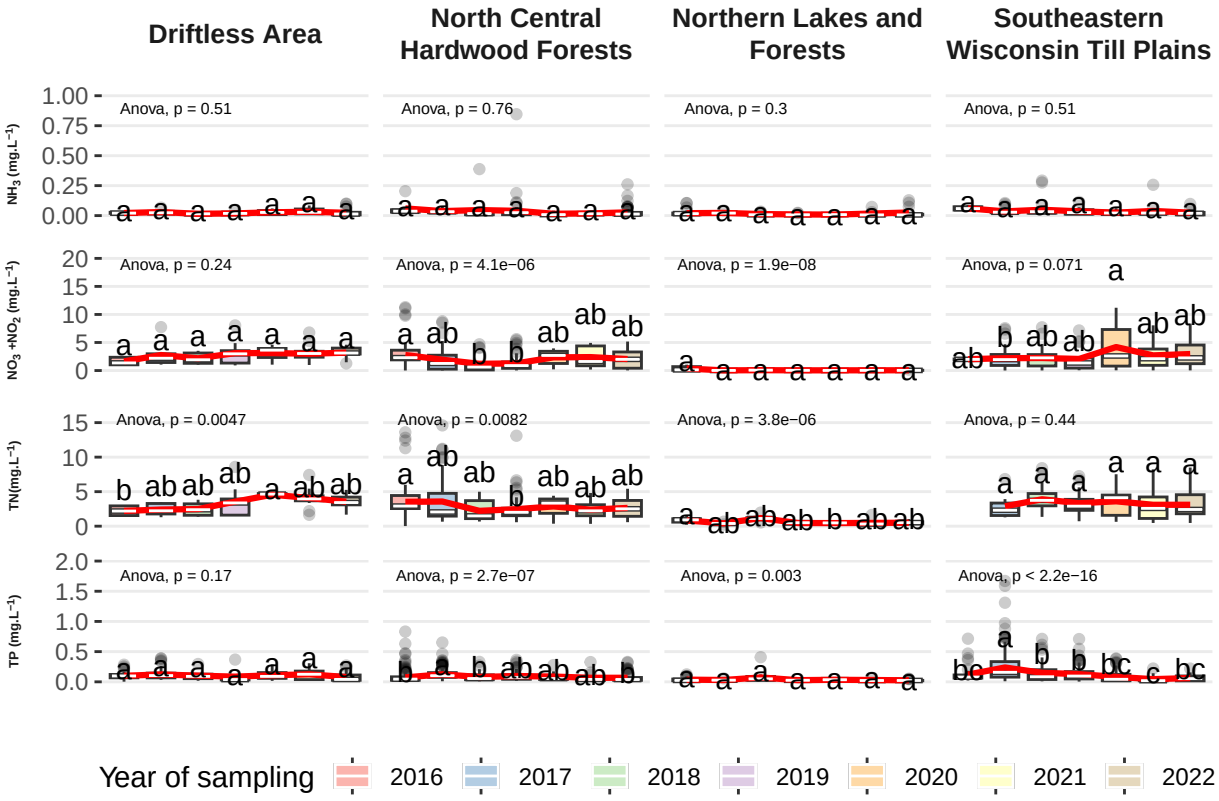

Fig. 2. Boxplots show changes in nutrient concentrations across ecoregions sampled from 2016 to 2022. The middle white horizontal line in the boxplots shows the median and the boxes represent the interquartile range. Red lines represent mean concentration. Post-hoc Tukey's HSD pairwise comparisons are shown in letters.

Spatial pattern of nutrients

Nutrient concentrations across ecoregions are significantly different ($p < 0.05$) with the lowest concentration of TP, TN, NO3+NO2, and NH3 observed in the Northern Lakes and Forests ecoregion (Fig. 3). TP was found to be highest in the Southern Wisconsin Till Plains ecoregion.

```
#One-way ANOVA for Ecoregion
aov_tp <- aov(result ~ Ecoregion, data = tp)
```

```r
aov_tn <- aov(result ~ Ecoregion, data = tn)
aov_nh3 <- aov(result ~ Ecoregion, data = nh3)
aov_no3 <- aov(result ~ Ecoregion, data = no3)

#Tukey posthoc for Ecoregion
tp_tukey_eco <- TukeyHSD(aov_tp)
tn_tukey_eco <- TukeyHSD(aov_tn)
nh3_tukey_eco <- TukeyHSD(aov_nh3)
no3_tukey_eco <- TukeyHSD(aov_no3)

# Extract labels and factor levels from Tukey post-hoc on ANOVA for Ecoregion
tp_cld_eco <- multcompLetters4(aov_tp, tp_tukey_eco)
tn_cld_eco <- multcompLetters4(aov_tn, tn_tukey_eco)
nh3_cld_eco <- multcompLetters4(aov_nh3, nh3_tukey_eco)
no3_cld_eco <- multcompLetters4(aov_no3, no3_tukey_eco)

tn_cld_eco <- as.data.frame.list(tn_cld_eco$Ecoregion)
tn_cld_eco$Ecoregion <- rownames(tn_cld_eco)
tn_cld_eco['PramName'] <- 'TN'

tp_cld_eco <- as.data.frame.list(tp_cld_eco$Ecoregion)
tp_cld_eco$Ecoregion <- rownames(tp_cld_eco)
tp_cld_eco['PramName'] <- 'TP'

no3_cld_eco <- as.data.frame.list(no3_cld_eco$Ecoregion)
no3_cld_eco$Ecoregion <- rownames(tn_cld_eco)
no3_cld_eco['PramName'] <- 'NO3+NO2'

nh3_cld_eco <- as.data.frame.list(nh3_cld_eco$Ecoregion)
nh3_cld_eco$Ecoregion <- rownames(nh3_cld_eco)
nh3_cld_eco['PramName'] <- 'NH3_N'

tn_cld_eco <- select(tn_cld_eco, c(Letters, Ecoregion, PramName))
tp_cld_eco <- select(tp_cld_eco, c(Letters, Ecoregion, PramName))
no3_cld_eco <- select(no3_cld_eco, c(Letters, Ecoregion, PramName))
nh3_cld_eco <- select(nh3_cld_eco, c(Letters, Ecoregion, PramName))

#Combine Tukey posthoc results
all_tukey_eco <- rbind(tp_cld_eco, tn_cld_eco, no3_cld_eco, nh3_cld_eco)

#Calculating sample size
sample_size = new_wq %>% group_by(PramName) %>% summarize(num=n())

#Creating plot for WQ data
library(viridis)
```

```
## Loading required package: viridisLite
```

```
##
## Attaching package: 'viridis'
```

```
## The following object is masked from 'package:scales':
##
##     viridis_pal
```

```r
wqplot_eco <- ggplot(data = new_wq, aes(x = Ecoregion,
                                        y = result,
                                        fill = PramName)) +
  facet_grid(PramName ~ Ecoregion,
             scales = "free",
             switch = "y") +
  geom_violin() +
  geom_boxplot(width = 0.2, color="grey") +
  scale_fill_viridis(discrete = TRUE) +
  theme_bw() +
  theme(axis.title.x = element_blank(),
        axis.title.y = element_blank(),
        axis.text.x = element_blank(),
        axis.ticks.x = element_blank(),
        strip.text = ggtext::element_markdown(),
        strip.background = element_blank(),
        strip.placement = "outside",
        panel.border = element_blank(),
        panel.grid.major.x = element_blank(),
        panel.grid.minor.x = element_blank(),
        panel.grid.minor.y = element_blank(),
        strip.text.y.left = element_text(face = "bold", size = 5),
        strip.text.x = ggtext::element_textbox_simple( width  = unit(1, "npc"),
                                                       height = unit(3, "lines"),
                                                       face = 'bold',
                                                       size = 10,
                                                       hjust = 0.5,
                                                       vjust = 0.5,
                                                       halign = 0.5,
                                                       valign = 0.5)) +
  stat_summary(fun = mean,
               geom = "point",
               col = "red",
               size = 3) +
  guides(fill = "none")
wqplot_eco
```

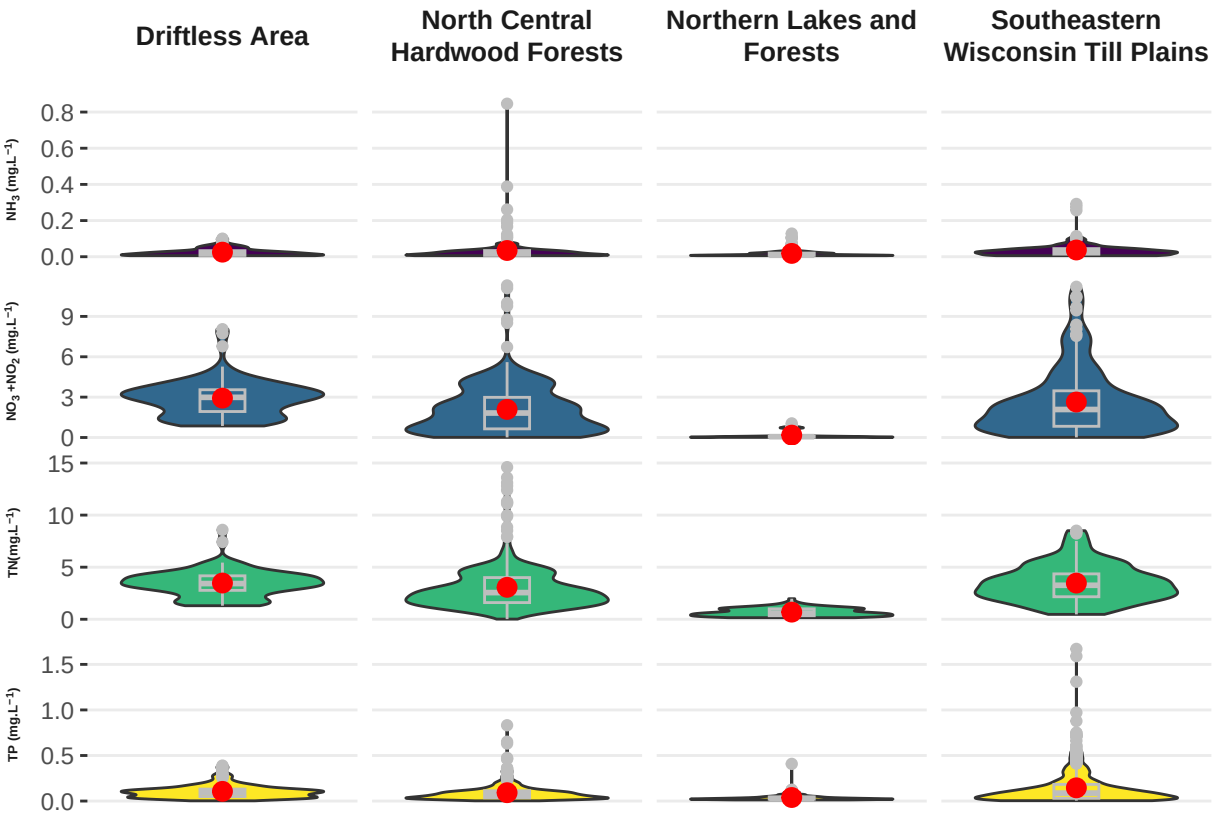

Fig. 3. Violin plot shows a comparison of nutrient concentrations across ecoregion with grey boxplots shows the median and the interquartile range and red dots represents the mean.

Dataset and sources for producing TP map

- **TPmap2** is a csv file contain GPS coordinates of each closest SWIMS ID of diatom sampling locations and their TP data.

```
library(dplyr)
library(ggnewscale)
library(hexbin)
TPmap <- read.csv("TPmap2.csv", row.names = 1)
TPmap_new <- TPmap %>% mutate(tp_bin = cut(TP,
                                breaks=c(0, 0.3, 0.6, 0.9, 1.2, 1.5, 1.8)))
tp_map <- ggplot() +
geom_sf(aes(fill = US_L3NAME), #color Ecoregion polygon area
               data = shpfile, #specify which data
               color = "dark red", #color Ecoregion polygon edges
               alpha = .2, #determine the opacity
               linewidth = .2) + #determine the width of the polygon edges
  labs(fill = "Ecoregion") + #rename legend title
  theme_bw() +
  theme(panel.grid = element_blank()) +
  theme(axis.title.x = element_blank(), axis.title.y = element_blank()) +
  new_scale_fill() +
  geom_hex(data = TPmap_new, aes(x=long, y=lat, fill=tp_bin, alpha = 0)) +
  scale_fill_viridis_d(direction = -1, option = "inferno",
  labels = paste(c("0 - 0.3", "0.3 - 0.6", "0.6 - 0.9", "0.9 - 1.2",
                   "1.2 - 1.5", "1.5 - 1.8"))) +
```

```
  guides(alpha="none", fill=guide_legend(title="TP (mg/L)"))
tp_map
```

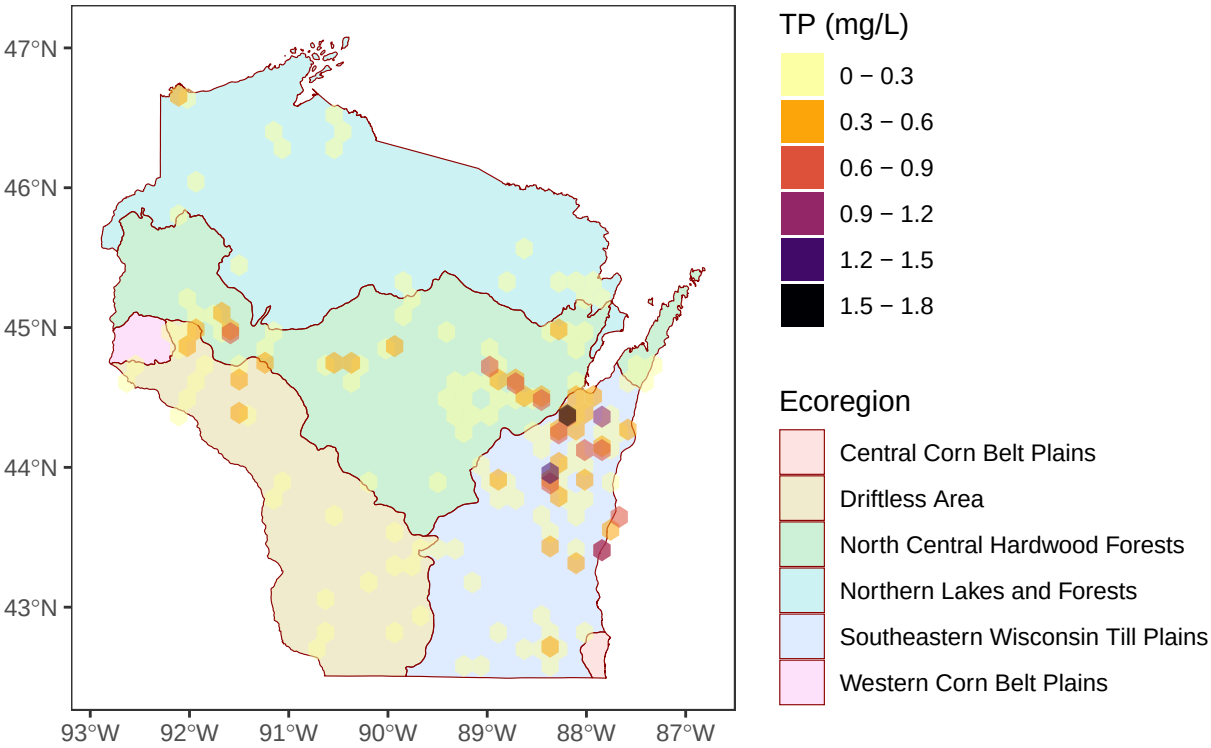

Fig. 4. Map of total phosphorus concentration (TP) distribution in Wisconsin

Dataset to investigate the diatom distribution across five Ecoregions in Wisconsin

- **nmds_data.csv** is environmental variables (pH, conductivity, TP, TN, NO3+NO2, and NH3) and diatom relative abundance data of 436 diatom taxa from 121 samples of 70 Surface Water Integrated Monitoring System (SWIMS) stations at different Ecoregion. Some wadeable streams were sampled multiple times within the time span of 2016 to 2023.

rank index

```
library(vegan)
```

```
## Loading required package: permute
```

```
## Loading required package: lattice
```

```
## This is vegan 2.6-6.1
```

```
nmds_data <- read.csv("nmds_dataset.csv", row.names = 1, check.names = F)
rankindex(nmds_data[,439:444], nmds_data[,1:436],
          indices = c("euc", "man", "gow", "bra", "kul"),
          stepacross = FALSE, method = "spearman")
```

```
##         euc         man         gow         bra         kul
## -0.05433653  0.03776997  0.07131486  0.03780464  0.03780727
```

Gower distance has the best good rank order relation to distance along environmental gradients.

Procrustes

- **Diat1.csv** is a csv file containing diatom relative abundance after the removal of singleton.
- **Diat5.csv** is a csv file containing diatom relative abundance after the removal of taxa occur in less than 5% of samples.
- **Diat10.csv** is a csv file containing diatom relative abundance after the removal of taxa occur in less than 10% of samples.

```
d1 <- read.csv("Diat1.csv", header = T, row.names = 1)
d5 <- read.csv("Diat5.csv", header = T, row.names = 1)
d10 <- read.csv("Diat10.csv", header = T, row.names = 1)
mdsall <- metaMDS(nmds_data[,1:436], distance = "gow")
```

```
## Square root transformation
## Wisconsin double standardization
## Run 0 stress 0.1978154
## Run 1 stress 0.1981658
## ... Procrustes: rmse 0.007661042  max resid 0.08189861
## Run 2 stress 0.1985683
## Run 3 stress 0.2018787
## Run 4 stress 0.2054546
## Run 5 stress 0.2023382
## Run 6 stress 0.2084179
## Run 7 stress 0.2102204
## Run 8 stress 0.2001299
## Run 9 stress 0.1992331
## Run 10 stress 0.2054411
## Run 11 stress 0.2019371
## Run 12 stress 0.1988692
## Run 13 stress 0.2027748
## Run 14 stress 0.2001032
## Run 15 stress 0.2008401
## Run 16 stress 0.2054196
## Run 17 stress 0.2056187
## Run 18 stress 0.2015568
## Run 19 stress 0.2026834
## Run 20 stress 0.2118981
## *** Best solution was not repeated -- monoMDS stopping criteria:
##      6: no. of iterations >= maxit
##     14: stress ratio > sratmax
```

```
md1 <- metaMDS(d1[,1:343], distance = "gow")
```

```
## Square root transformation
## Wisconsin double standardization
## Run 0 stress 0.229416
## Run 1 stress 0.2299308
## Run 2 stress 0.237756
## Run 3 stress 0.2345134
## Run 4 stress 0.2393019
## Run 5 stress 0.2323563
## Run 6 stress 0.2332618
## Run 7 stress 0.233164
## Run 8 stress 0.2293624
## ... New best solution
## ... Procrustes: rmse 0.008702699  max resid 0.06562812
```

```
## Run 9 stress 0.2410159
## Run 10 stress 0.2319713
## Run 11 stress 0.2411153
## Run 12 stress 0.24685
## Run 13 stress 0.238234
## Run 14 stress 0.2333365
## Run 15 stress 0.2292863
## ... New best solution
## ... Procrustes: rmse 0.00570393  max resid 0.05210241
## Run 16 stress 0.2352577
## Run 17 stress 0.2294949
## ... Procrustes: rmse 0.008100868  max resid 0.06552709
## Run 18 stress 0.2323574
## Run 19 stress 0.2415701
## Run 20 stress 0.2421483
## *** Best solution was not repeated -- monoMDS stopping criteria:
##      1: no. of iterations >= maxit
##     19: stress ratio > sratmax
```

```
md5 <- metaMDS(d1[,1:191], distance = "gow")
```

```
## Square root transformation
## Wisconsin double standardization
## Run 0 stress 0.2431645
## Run 1 stress 0.2431533
## ... New best solution
## ... Procrustes: rmse 0.003897333  max resid 0.03082083
## Run 2 stress 0.2635511
## Run 3 stress 0.251072
## Run 4 stress 0.2697389
## Run 5 stress 0.2462591
## Run 6 stress 0.2566759
## Run 7 stress 0.243995
## Run 8 stress 0.2449212
## Run 9 stress 0.2573795
## Run 10 stress 0.2495581
## Run 11 stress 0.2434126
## ... Procrustes: rmse 0.006930084  max resid 0.04518116
## Run 12 stress 0.2488711
## Run 13 stress 0.2531741
## Run 14 stress 0.2431708
## ... Procrustes: rmse 0.002043903  max resid 0.01306085
## Run 15 stress 0.2529276
## Run 16 stress 0.250991
## Run 17 stress 0.2463086
## Run 18 stress 0.2613286
## Run 19 stress 0.2502803
## Run 20 stress 0.2440929
## *** Best solution was not repeated -- monoMDS stopping criteria:
##      3: no. of iterations >= maxit
##     17: stress ratio > sratmax
```

```
md10 <- metaMDS(d1[,1:135], distance = "gow")
```

```
## Square root transformation
```

```
## Wisconsin double standardization
## Run 0 stress 0.228448
## Run 1 stress 0.2284865
## ... Procrustes: rmse 0.01709491  max resid 0.1713435
## Run 2 stress 0.2292802
## Run 3 stress 0.2422535
## Run 4 stress 0.2303549
## Run 5 stress 0.2284877
## ... Procrustes: rmse 0.01716732  max resid 0.1713813
## Run 6 stress 0.2414073
## Run 7 stress 0.241811
## Run 8 stress 0.2297449
## Run 9 stress 0.2292516
## Run 10 stress 0.239054
## Run 11 stress 0.2285183
## ... Procrustes: rmse 0.01696455  max resid 0.1711331
## Run 12 stress 0.2356268
## Run 13 stress 0.2353803
## Run 14 stress 0.231158
## Run 15 stress 0.2324441
## Run 16 stress 0.2289433
## ... Procrustes: rmse 0.02707207  max resid 0.1693656
## Run 17 stress 0.2439509
## Run 18 stress 0.230547
## Run 19 stress 0.2291335
## Run 20 stress 0.2280009
## ... New best solution
## ... Procrustes: rmse 0.01479043  max resid 0.1097173
## *** Best solution was not repeated -- monoMDS stopping criteria:
##      6: no. of iterations >= maxit
##     14: stress ratio > sratmax
```

```
protest(mdsall, md1)
```

```
##
## Call:
## protest(X = mdsall, Y = md1)
##
## Procrustes Sum of Squares (m12 squared):        0.09598
## Correlation in a symmetric Procrustes rotation: 0.9508
## Significance:  0.001
##
## Permutation: free
## Number of permutations: 999
```

```
protest(mdsall, md5)
```

```
##
## Call:
## protest(X = mdsall, Y = md5)
##
## Procrustes Sum of Squares (m12 squared):        0.1377
## Correlation in a symmetric Procrustes rotation: 0.9286
## Significance:  0.001
##
```

```
## Permutation: free
## Number of permutations: 999
```

```
protest(mdsall, md10)
```

```
##
## Call:
## protest(X = mdsall, Y = md10)
##
## Procrustes Sum of Squares (m12 squared):        0.4722
## Correlation in a symmetric Procrustes rotation: 0.7265
## Significance:  0.001
##
## Permutation: free
## Number of permutations: 999
```

```
mdsall$stress
```

```
## [1] 0.1978154
```

```
md1$stress
```

```
## [1] 0.2292863
```

```
md5$stress
```

```
## [1] 0.2431533
```

```
md10$stress
```

```
## [1] 0.2280009
```

Major drivers of diatom distribution

There was a gradual change in diatom species composition from Northern Lakes and Forests to Driftless Area and Southeastern Wisconsin Till Plains ecoregions, as shown in NMDS plots (Fig. 5). Most samples collected from the Northern Lakes and Forests ecoregion have negative scores on NMDS axis 1 and 2, whereas most samples collected within North Central Hardwood Forest, Driftless Area, and Southeastern Wisconsin Till Plains ecoregions display negative scores on NMDS axis 1 and 2. Samples with positive scores on NMDS Axis 1 and 2 are mostly dominated by Mayamaea permitis, Navicula gregaria, Nitzschia palea var. debilis, Nitzschia oligotraphenta, Nitzschia sociabilis, Nitzschia soratensis, and Sellaphora nigri that are indicators of organic pollution; and these samples are correlated with high nutrient concentration (TP, TN, and NO3+NO2) (Fig. 5). Majority of samples collected from Northern Lakes and Forests ecoregion have negative scores on NMDS Axis 1 and 2; they are dominated by diatoms characteristic of oligotrophic waters, such as Achnanthidium atomoides, Achnanthidium rivulare, Fragilaria tenera, Pseudostaurosira brevistriata, and Staurosira construens var. venter.

```
library(vegan)
library(ggrepel)
library(ggalt)
```

```
## Registered S3 methods overwritten by 'ggalt':
##   method                  from
##   grid.draw.absoluteGrob  ggplot2
##   grobHeight.absoluteGrob ggplot2
##   grobWidth.absoluteGrob  ggplot2
##   grobX.absoluteGrob      ggplot2
##   grobY.absoluteGrob      ggplot2
```

```r
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
```

```
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```r
mdsall <- metaMDS(nmds_data[,1:436], distance = "gow")
```

```
## Square root transformation
## Wisconsin double standardization
## Run 0 stress 0.1978154
## Run 1 stress 0.2000721
## Run 2 stress 0.1989494
## Run 3 stress 0.2032245
## Run 4 stress 0.202524
## Run 5 stress 0.200411
## Run 6 stress 0.2018012
## Run 7 stress 0.2076967
## Run 8 stress 0.2058242
## Run 9 stress 0.2018627
## Run 10 stress 0.2123053
## Run 11 stress 0.19972
## Run 12 stress 0.208225
## Run 13 stress 0.1998817
## Run 14 stress 0.2026782
## Run 15 stress 0.2089557
## Run 16 stress 0.2100347
## Run 17 stress 0.2041852
## Run 18 stress 0.1981306
## ... Procrustes: rmse 0.007390181  max resid 0.07459936
## Run 19 stress 0.2092358
## Run 20 stress 0.2013904
## *** Best solution was not repeated -- monoMDS stopping criteria:
##      2: no. of iterations >= maxit
##     18: stress ratio > sratmax
```

```r
mx <- apply(nmds_data[,1:436], 2, max)
sumsp <- as.data.frame(colSums(nmds_data[,1:436]))
meansp <- as.data.frame(colMeans(nmds_data[,1:436]))
scr <- as.data.frame(mdsall$species)
scrsp <- cbind.data.frame(scr, abund = meansp$`colMeans(nmds_data[, 1:436])`)
tscrsp <- as.data.frame(t(scrsp))
mxsc <- apply(tscrsp[3,],2,max)
spsel <- as.data.frame(t(as.data.frame(tscrsp[,mxsc>0.5])))
mdspoints <- as.data.frame(mdsall$points)
env.fit_all <- envfit(mdspoints, nmds_data[,439:444], perm=999, na.rm = T)
vec.mds_all <- as.data.frame(env.fit_all$vectors$arrows*sqrt(env.fit_all$vectors$r)/16)
MDS <- data.frame(MDS1 = mdsall$points[,1],
                  MDS2 = mdsall$points[,2],
                  Ecoregion = as.factor(nmds_data$Ecoregion))
nmds_plot <- ggplot(data = MDS, aes(x = MDS1, y = MDS2)) +
  stat_ellipse(geom = "polygon", aes(fill = Ecoregion, alpha = 0.2)) +
```

```
  scale_fill_manual(values = c("#B79F00", "#00BA38", "#00BFC4", "#619CFF")) +
  guides(alpha = "none") +
  geom_label_repel(data = spsel, aes(x = MDS1, y = MDS2),
                   label=rownames(spsel), size=4, inherit.aes = F) +
  geom_segment(data = vec.mds_all,
               aes(x = 0,xend = MDS1, y = 0, yend = MDS2),
               arrow = arrow(length = unit(0.4, "cm")),
               colour = "red",
               inherit.aes = F, linewidth = 1.2) +
  geom_text_repel(data = vec.mds_all,
                  aes(x = MDS1, y = MDS2,
                      label=rownames(vec.mds_all)),
            size=5,
            fontface = "bold",
            inherit.aes = F) +
  coord_fixed() + theme_bw() + theme(panel.grid = element_blank())
nmds_plot
```

```
## Warning: ggrepel: 31 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```

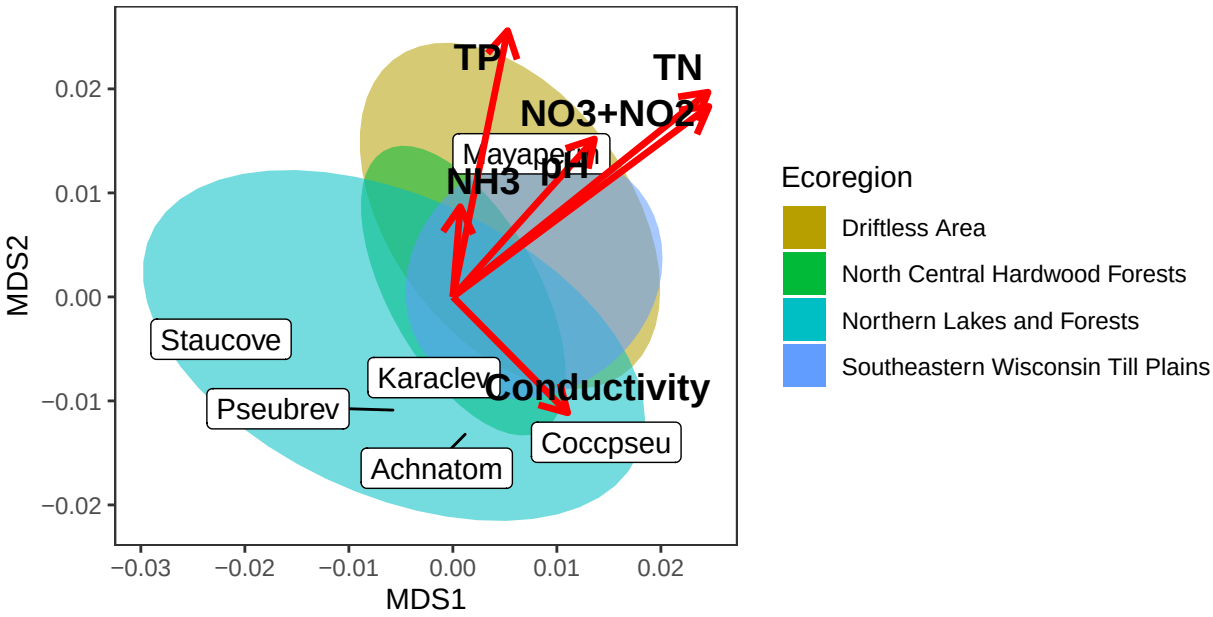

Fig. 5. Non-metric Multidimensional Scaling ordination plots of 121 samples from Wisconsin's streams showing 95% confidence ellipses for Ecoregion and the fitted environmental variables. Vectors indicate the direction and strength of each environmental variable to the overall distribution. Achnatom = Achnanthidium atomoides, Achminu = Achnanthidium minutissimum, Achnpyre = Achnanthidium pyreniacum, Achnrivu = Achnanthidium rivulare, Amphpedi = Amphora pediculus, Coccplac = Cocconeis placentula, Coccpseu = Cocconeis pseudothumensis, Fragtene = Fragilaria tenera, Karaclev = Karayevia clevei, Mayaperm = Mayamaea permitis, Navicryp = Navicula cryptocephala, Navigreg = Navicula gregaria, Nitzpade = Nitzschia palea var. debilis, Nitzolig = Nitzschia oligotraphenta, Nitzsoci = Nitzschia sociabilis, Nitzsora = Nitzschia soratensis, Planfreq = Planothidium frequentissimum, Planlac = Planothidium lanceolatum, Platcons = Platessa conspicua, Pseubrev = Pseudostaurosira brevistriata, Sellatom = Sellaphora atomus, Sellnigr = Sellaphora nigri, Sellsaug = Sellaphora saugerresii.

```
library(akima)
tp_mds <- with(mdsall,
```

```
              interp(x=mdsall$points[,1],
                     y=mdsall$points[,2],
                     z=nmds_data$TP,
                     xo=seq(min(mdsall$points[,1]),
                            max(mdsall$points[,1]),
                            length=60)))
tp_mds2=melt(tp_mds$z, na.rm = T)
names(tp_mds2)=c("x", "y", "TP")
tp_mds2$MDS1=tp_mds$x[tp_mds2$x]
tp_mds2$MDS2=tp_mds$y[tp_mds2$y]
ggplot(data = tp_mds2, aes(x=MDS1, y=MDS2, fill=TP, z=TP)) +
  geom_tile() + scale_fill_continuous(high = "red", low = "white") +
  theme_bw() + theme(panel.grid = element_blank()) +
  geom_label_repel(data=spsel,
                   aes(x=MDS1, y=MDS2),
                   label=rownames(spsel),
                   size=3, colour="black",
                   inherit.aes = F) +
  scale_size_continuous(range = c(1,10))
```

```
## Warning: ggrepel: 31 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```

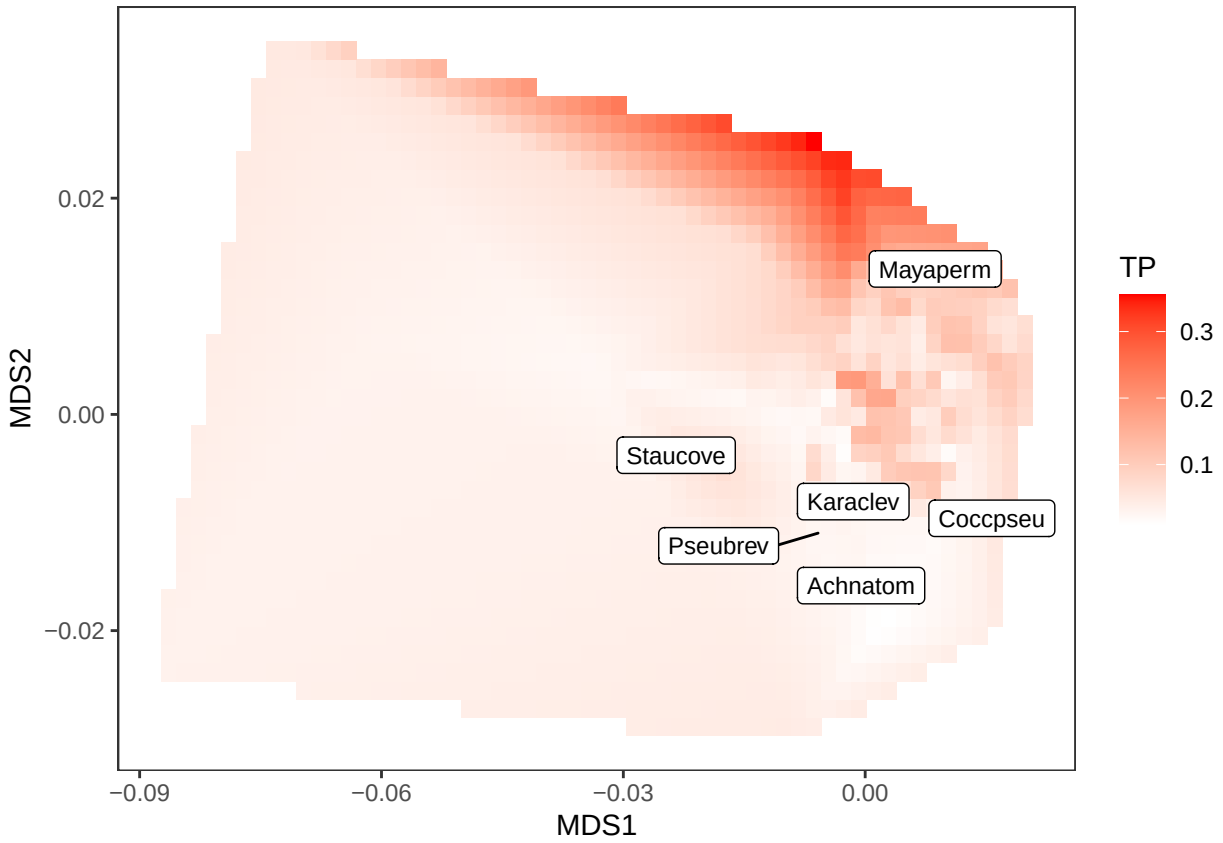

Fig. 6.Non-metric Multidimensional Scaling ordination plots of 121 samples from Wisconsin's streams with a contour map of the diatom-based inferred values of total phosphorus concentration (TP, mg/L) generated via interpolating values of TP. The color key indicates how the color is related to TP. Achnatom = Achnanthidium atomoides, Achminu = Achnanthidium minutissimum, Achnpyre = Achnanthidium pyreniacum, Achnrivu = Achnanthidium rivulare, Amphpedi = Amphora pediculus, Coccplac = Cocconeis

placentula, Coccpseu = Cocconeis pseudothumensis, Fragtene = Fragilaria tenera, Karaclev = Karayevia clevei, Mayaperm = Mayamaea permitis, Navicryp = Navicula cryptocephala, Navigreg = Navicula gregaria, Nitzpade = Nitzschia palea var. debilis, Nitzolig = Nitzschia oligotraphenta, Nitzsoci = Nitzschia sociabilis, Nitzsora = Nitzschia soratensis, Planfreq = Planothidium frequentissimum, Planlac = Planothidium lanceolatum, Platcons = Platessa conspicua, Pseubrev = Pseudostaurosira brevistriata, Sellatom = Sellaphora atomus, Sellnigr = Sellaphora nigri, Sellsaug = Sellaphora saugerresii.

DPI development and its dataset

Develop DPI using weighted averaging regression model (WA) and cross-validate the resulting model using the crossval function, which allows internal cross-validation using bootstrapping method. - **AllDiatTP.csv** is a TP and diatom relative abundance data of 565 diatom taxa from 479 samples of 248 Surface Water Integrated Monitoring System (SWIMS) stations at different Ecoregion. Some wadeable streams were sampled multiple times within the time span of 2016 to 2023. - **AllDiatTP_sameday.csv** is a diatom relative abundance subset data of 491 diatom taxa from 207 samples with TP that were sampled on the same day of diatom sampling. - **AllDiatTP_d8.csv** is a diatom relative abundance subset data of 416 diatom taxa from 92 samples with TP that were sampled on a day to seven days before the diatom sampling. - **AllDiat_TP_d8_14.csv** is a diatom relative abundance subset data of 395 diatom taxa from 85 samples with TP that were sampled on eight to fourteen days before the diatom sampling. - **AllDiatTP_highresrem.csv** is a diatom relative abundance subset data of 557 diatom taxa from 418 samples with removal of samples with TP values detected as outliers.

```
library(rioja)
```

```
## This is rioja  1.0-6
```

```
alldata <- read.csv("AllDiatTP.csv", row.names = 1, check.names = F)
dpiwa <- WA(alldata[,1:565],alldata$TP, tolDW = T)
dpiwa.cv <- crossval(dpiwa, nboot = 1000)
```

```
## Cross-validating:
##   |                                                                              |
```

```
rand.t.test(dpiwa.cv)
```

```
##                  RMSE        R2     Avg.Bias  Max.Bias     Skill delta.RMSE
## WA.inv     0.09765459 0.3125680 0.0005015733 0.8081895  30.93169         NA
## WA.cla     0.14594363 0.3152443 0.0012433702 0.7576068 -54.26372         NA
## WA.inv.tol 0.09669374 0.3258692 0.0032404380 0.7993024  32.28417 -0.9839316
## WA.cla.tol 0.13773664 0.3287255 0.0073861559 0.7466861 -37.40181 -5.6234005
##                 p
## WA.inv         NA
## WA.cla         NA
## WA.inv.tol 0.358
## WA.cla.tol 0.074
```

```
same_data <- read.csv("AllDiatTP_sameday.csv", row.names = 1, check.names = F)
dpiwa_same <- WA(same_data[,1:491],same_data$TP, tolDW = T)
dpiwa_same.cv <- crossval(dpiwa_same, nboot = 1000)
```

```
## Cross-validating:
##   |                                                                              |
```

```
rand.t.test(dpiwa_same.cv)
```

```
##                  RMSE        R2      Avg.Bias  Max.Bias     Skill delta.RMSE
## WA.inv     0.08171070 0.2685375 -0.0003506063 0.6289200  26.30939         NA
## WA.cla     0.12391165 0.2756161 -0.0008666386 0.6300272 -69.46443         NA
```

```
## WA.inv.tol 0.07988826 0.2986535  0.0026388493 0.6065122  29.55986  -2.230359
## WA.cla.tol 0.10579434 0.3053151  0.0057393717 0.5857637 -23.53191 -14.621148
##                p
## WA.inv        NA
## WA.cla        NA
## WA.inv.tol 0.133
## WA.cla.tol 0.002
```

```
eight_data <- read.csv("AllDiatTP_d8.csv", row.names = 1, check.names = F)
dpiwa_eight <- WA(eight_data[,1:416], eight_data$TP, tolDW = T)
eight.cv <- crossval(dpiwa_eight, nboot = 1000)
```

```
## Cross-validating:
##   |                                                                      |
```

```
rand.t.test(eight.cv)
```

```
##                 RMSE        R2    Avg.Bias  Max.Bias    Skill delta.RMSE     p
## WA.inv     0.1358438 0.3742458 0.002274671 0.7950720 33.54209         NA    NA
## WA.cla     0.1510016 0.4238989 0.004923365 0.5444456 17.88359         NA    NA
## WA.inv.tol 0.1440755 0.3165954 0.023380775 1.0904171 25.24383   6.059648 0.664
## WA.cla.tol 0.1464962 0.3132199 0.044026318 1.0095659 22.71065  -2.983668 0.454
```

```
day8_14_data <- read.csv("AllDiatTP_d8_14.csv", row.names = 1, check.names = F)
dpiwa_8_14 <- WA(day8_14_data[,1:395],day8_14_data$TP, tolDW = T)
day8_14.cv <- crossval(dpiwa_8_14, nboot = 1000)
```

```
## Cross-validating:
##   |                                                                      |
```

```
rand.t.test(day8_14.cv)
```

```
##                 RMSE        R2    Avg.Bias  Max.Bias        Skill delta.RMSE
## WA.inv     0.09815276 0.2716095 0.003739594 0.4721131  25.58636485         NA
## WA.cla     0.11966548 0.2872018 0.007121009 0.3822256 -10.60768131         NA
## WA.inv.tol 0.11383469 0.1430029 0.010415098 0.4874879  -0.09140322   15.97707
## WA.cla.tol 0.15385707 0.1570211 0.019850168 0.4102967 -82.84472519   28.57264
##                p
## WA.inv        NA
## WA.cla        NA
## WA.inv.tol 0.986
## WA.cla.tol 0.996
```

```
rem_data <- read.csv("AllDiatTP_highresrem.csv", row.names = 1, check.names = F)
dpiwa_rem <- WA(rem_data[,1:557],rem_data$TP, tolDW = T)
dpiwa_rem.cv <- crossval(dpiwa_rem, nboot = 1000)
```

```
## Cross-validating:
##   |                                                                      |
```

```
rand.t.test(dpiwa_rem.cv)
```

```
##                 RMSE        R2     Avg.Bias  Max.Bias    Skill delta.RMSE
## WA.inv     0.03940961 0.5487149 0.0005527796 0.1872970 54.85873         NA
## WA.cla     0.04771489 0.5529548 0.0008656916 0.1638644 33.82754         NA
## WA.inv.tol 0.03969815 0.5425494 0.0014304907 0.1846980 54.19530  0.7321578
## WA.cla.tol 0.04665171 0.5467610 0.0022260747 0.1613926 36.74360 -2.2281998
##                p
```

```
## WA.inv        NA
## WA.cla        NA
## WA.inv.tol 0.468
## WA.cla.tol 0.463
```

```
#WAPLS
dpiwapls_rem <- WAPLS(rem_data[,1:557],rem_data$TP)
dpiwapls_rem.cv <- crossval(dpiwapls_rem, nboot = 1000)
```

```
## Cross-validating:
##   |                                                                              |
```

```
rand.t.test(dpiwapls_rem.cv)
```

```
##              RMSE        R2     Avg.Bias  Max.Bias    Skill delta.RMSE     p
## Comp01 0.03940961 0.5487149 0.0005527796 0.1872970 54.85873 -32.812750 0.001
## Comp02 0.03766781 0.5893875 0.0004151990 0.1773948 58.76081  -4.419744 0.003
## Comp03 0.03860349 0.5722840 0.0009941152 0.1633679 56.68658   2.484022 0.863
## Comp04 0.04079090 0.5365424 0.0008538798 0.1575666 51.63893   5.666354 0.983
## Comp05 0.04467369 0.4766592 0.0012852172 0.1566552 41.99397   9.518786 0.997
```

AllDiatTP_highresrem.csv has the best performance and hence will be used to calculate DPI

TP transfer function

```
segmentwise.rmse <- function(dpiwapls_rem.cv, ng = 5, k = 2, plot = TRUE, ...){
  if(is.null(dpiwapls_rem.cv$residuals.cv)){
    if(class(dpiwapls_rem.cv) == "WAPLS"){
      r <- dpiwapls_rem.cv$fitted.values[,k]-dpiwapls_rem.cv$x
      perf<-performance(dpiwapls_rem.cv)$object
    }else{
      stop("Need cross-validated model to calculate RMSEP")
    }
  }else{
    r <- dpiwapls_rem.cv$residuals.cv[,k]
    perf<-performance(dpiwapls_rem.cv)$crossval
  }
  breaks <- seq(min(dpiwapls_rem.cv$x),max(dpiwapls_rem.cv$x), length=ng)
  envcut <- cut(dpiwapls_rem.cv$x,breaks=breaks, include.lowest=TRUE)
  segRMSEP <- tapply(r,envcut,function(x)sqrt(mean(x^2)))
  allsegRMSEP <- sqrt(mean(segRMSEP^2))

  if(plot){
    hist(dpiwapls_rem.cv$x, breaks=breaks, col="grey70", border=NA, ...)
    par(new=T)
    mid<-((c(breaks,NA)+c(NA,breaks))/2)
    mid<-mid[!is.na(mid)]
    plot(mid,segRMSEP, type="n", xlim=par()$usr[1:2],xaxs="i", yaxt="n", ylab="", xlab="",
    col="black")
    lines(breaks,c(segRMSEP[1],segRMSEP), type="S", col="black")
    axis(4)
    mtext("RMSEP", side=4, line=1.5)
  }
  list(breaks = breaks, segRMSEP = segRMSEP, allsegRMSEP = allsegRMSEP)
}
segmentwise.rmse(dpiwapls_rem.cv, k=1, main="", xlab="TP (mg/L)")
```

```
## $breaks
## [1] 0.00400 0.10475 0.20550 0.30625 0.40700
##
## $segRMSEP
## [0.004,0.105] (0.105,0.205] (0.205,0.306] (0.306,0.407]
##    0.02826961    0.04274452    0.13795870    0.12073090
##
## $allsegRMSEP
## [1] 0.09517723
```

```
plot(dpiwapls_rem.cv, xval=T, xlab="Observed TP (mg/L)", ylab="Inferred TP (mg/L)",
col="darkgreen", pch=19)
mtext("WA-PLS", 3, adj=0)
mtext(expression("R"^2*"-boot=0.55, RMSEP=0.047"), adj=1)
```

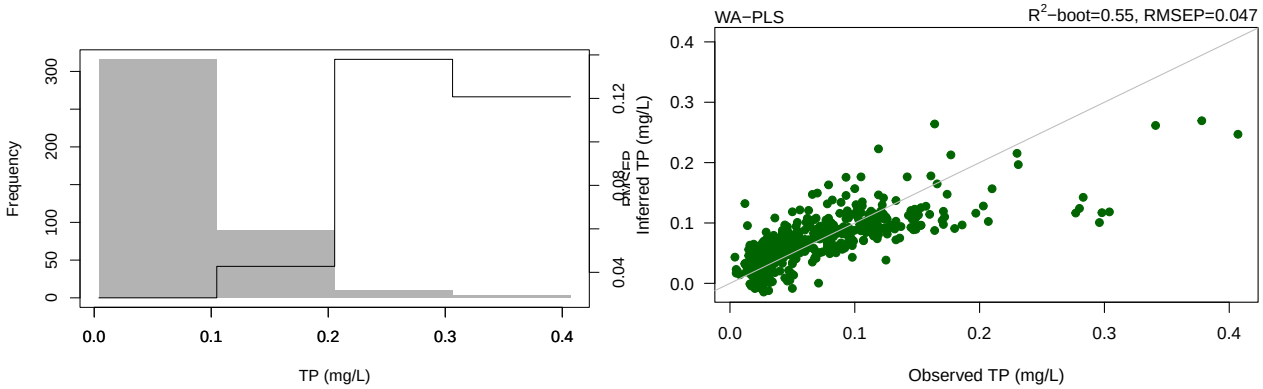

Figure 7. WA-PLS transfer function for total phosphorus concentration (TP). Left panel shows a comparison of the overall RMSEPboot(dashed line) and segment-wise RMSEPboot (solid line). Gray bars are the numbers of samples within four segments of TP values. Right panel shows relationships between observed and inferred TP values.

TP prediction with new diatom data

newdata is an example of diatom data of 34 taxa. *Notes: Before calculating DPI make sure that the diatom taxa names match with names used in the AllDiatTP_highresrem.csv training dataset and file is in the csv format*

```
#read csv file of example of new diatom data
newdata <- read.csv("newdata.csv", row.names = 1, check.names = F)
#infer TP value from new diatom data
newdata_dpi <- predict(dpiwapls_rem, newdata[,1:34], sse = T, nboot = 1000)
```

```
## Bootstrapping for SSE:
##   |                                                                      |
```

```
newdata_dpi$fit.boot
```

```
##             Comp01     Comp02     Comp03     Comp04     Comp05
## 2016-07 0.07828565 0.07229879 0.07249916 0.07408739 0.07444665
```

```
#DPI result is the WA.cla with inferred TP value of 0.0809 mg/L
```

Literature

1. Schindler, D. W. 2006. Recent advances in the understanding and management of eutrophication. Limnology and Oceanography, 51(1part2), 356-363. https://doi.org/10.4319/lo.2006.51.1_part_2.0356
2. Hilton, J., O'Hare, M., Bowes, M. J., & Jones, J. I. 2006. How green is my river? A new paradigm of eutrophication in rivers. Science of The Total Environment, 365(1-3), 66-83. https://doi.org/10.1016/j.scitotenv.2006.02.055
3. Brian Moss, Sarian Kosten, Mariana Meerhoff, Richard W. Battarbee, Erik Jeppesen, Néstor Mazzeo, Karl Havens, Gissell Lacerot, Zhengwen Liu, Luc De Meester, Hans Paerl & Marten Scheffer 2011. Allied attack: climate change and eutrophication, Inland Waters, 1:2, 101-105, DOI: 10.5268/IW-1.2.359
4. U.S. Environmental Protection Agency. 2017. National Water Quality Inventory: Report to Congress. (EPA 841-R-16-011). Washington D.C.: US EPA.
5. Water Quality Standards for Wisconsin Surface Waters. Wisconsin Administrative Code § NR. 102.03 (2022). https://docs.legis.wisconsin.gov/code/admin_code/nr/100/102.pdf#page=28
6. Wisconsin Department of Natural Resources. 2024. Wisconsin Water Quality Report to Congress 2024. (EGAD Number: 3200-2024-03). Madison: WDNR.
7. 33 U.S.C §§ 1251 et seq. Code Chapter 26 - WATER POLLUTION PREVENTION AND CONTROL. Federal Water Pollution Control Act, on June 30, 1948 (Clean Water Act became the common name with the 1972 amendments)
8. Robertson, D.M., Weigel, B.M., and Graczyk, D.J., 2008, Nutrient concentrations and their relations to the biotic integrity of nonwadeable rivers in Wisconsin: U.S. Geological Survey Professional Paper 1754, 81 p.
9. Wisconsin Department of Natural Resources. 2020. Implementation Progress Report 2017-2019. (EGAD Number: 3200-2020-15). Madison: WDNR.
10. Bennett, E., Reed-Andersen, T., Houser, J. et al. A Phosphorus Budget for the Lake Mendota Watershed. Ecosystems 2, 69–75 (1999). https://doi.org/10.1007/s100219900059
11. Hanson, P. C., Ladwig, R., Buelo, C., Albright, E. A., Delany, A. D., & Carey, C. C. 2023. Legacy phosphorus and ecosystem memory control future water quality in a eutrophic lake. Journal of Geophysical Research: Biogeosciences, 128, e2023JG007620. https://doi.org/10.1029/2023JG007620
12. Frei RJ, Lawson GM, Norris AJ, Cano G, Vargas MC, Kujanpa¨a¨ E, et al. 2021. Limited progress in nutrient pollution in the U.S. caused by spatially persistent nutrient sources. PLoS ONE 16(11): e0258952. https://doi.org/10.1371/journal. pone.0258952
13. Omernik, J.M., Chapman, S.S., Lillie, R.A., and Dumke, R.T. 2000. Ecoregions of Wisconsin: Transactions of the - Wisconsin Academy of the Wisconsin Sciences, Arts, and Letters, v. 88, p. 77–103.
14. Wisconsin State Laboratory of Hygiene. 2021. EHD ENV TOX Method 4730 Diatom Analysis.
15. Desianti, N. 2024. Diatom Atlas of Wisconsin Streams and Wetlands. Madison: WDNR, Final Reports.
16. Lange-Bertalot, H., Hofmann, G., Werum, M. and Cantonati, M. 2017. Freshwater Benthic Diatoms of Central Europe: Over 800 Common Species Used in Ecological Assessment. English edition with updated taxonomy and added species. Koeltz Botanical Books, Schmitten-Oberreifenberg, 942 pp.
17. Faith, D.P., P.R. Minchin, and L. Belbin. 1987. Compositional dissimilarity as a robust measure of ecological distance. Vegetatio 69 (1-3): 57–68. https://doi.org/10.1007/BF00038687
18. Clarke, K.R. and R.M. Warwick. 1994. An approach to statistical analysis and interpretation. 1st edition. Plymouth Marine Laboratory, Plymouth, U.K.
19. Peres-Neto, P.R., and D.A. Jackson. 2001. How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test. Oecologia 129 (2): 169–178. https://doi.org/10.1007/s004420100720.
20. Birks, H.J.B. 1995. Quantitative palaeoenvironmental reconstructions. In Statistical modelling of quaternary science data. Technical guide 5, ed. D. Maddy and J.S. Brew, 161–254. Cambridge: Quaternary Research Association https://www.uib.no/en/rg/EECRG/57871/quantitative-palaeoenvironmental-reconstructions
21. Juggins, S. 2020. Rioja: Analysis of Quaternary Science Data, R Package Version (0.9-15.1) (http:

//Cran.r-Project.Org/Package=rioja).
22. Oksanen, Author Jari, F Guillaume Blanchet, Roeland Kindt, Pierre Legen-, Peter R Minchin, R B O Hara, Gavin L Simpson, Peter Soly-, M Henry H Stevens, and Helene Wagner. 2020. "Package ' Vegan ."' https://doi.org/ISBN 0-387-95457-0.