Pool 4 Lower Exploration

Amber Lee

6/3/2021

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.1     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
library(RColorBrewer)
```

pictures, geographical information about the pool some data information

- for example, casey only has main channel and side channel data. (what are the strata that we have?)
- statistical summary of 4 variables: `TP`, `CHLcal`, `SS`, `TURB`
- number of observations per strata
- future directions / ideas
- an interesting paper or article from USGS website relating to pool 4 lower
- missing data
- extreme outliers
- `TP` and `TN` may have some missing values

(next friday: learn to pose own questions within an ecosystem framework, explain the project in our own words)

```
veg19 <- read.csv(file = "../pool data/ltrm_vegsrs_data_lat_long.csv")
water20 <- read.csv(file = "../pool data/ltrm_water_data_lat_long.csv")
# lter use data through 2020 that has lat/lng
```

Vegetation data

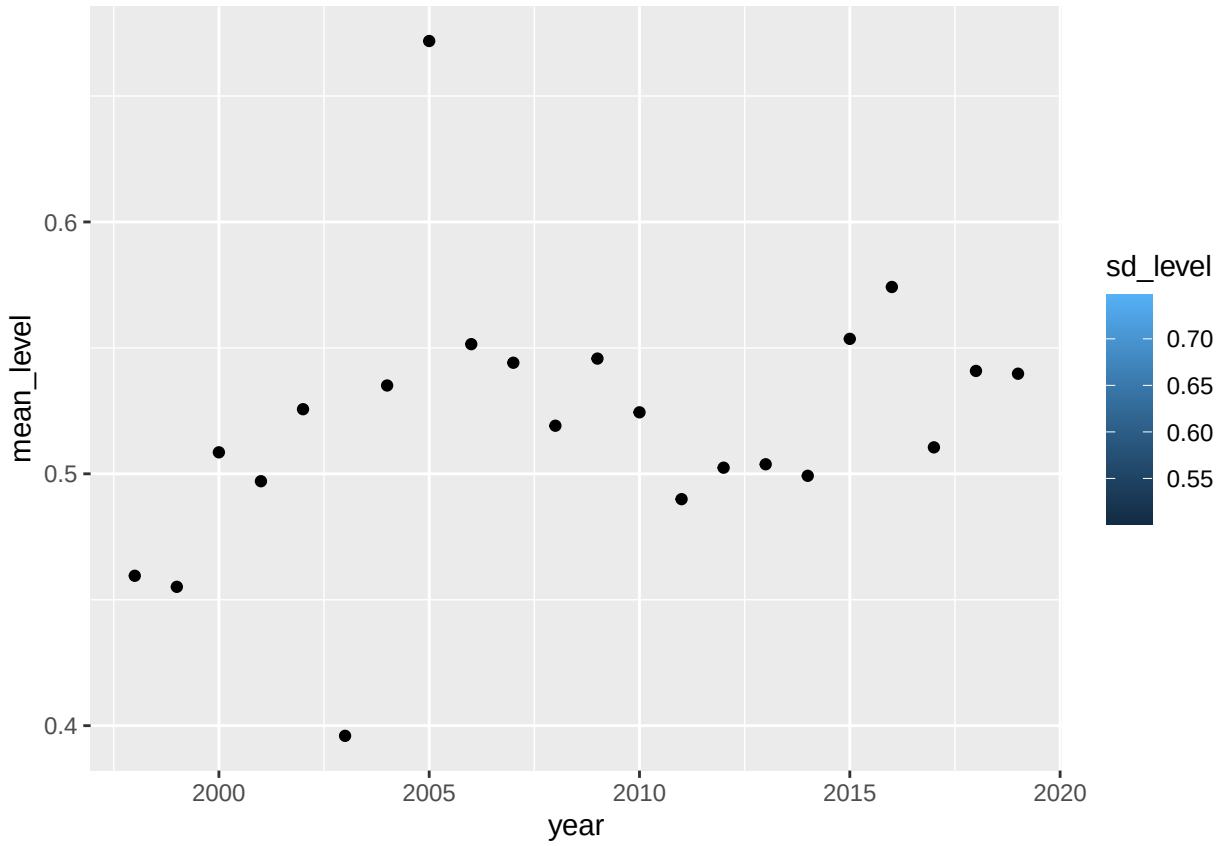
```
veg19 %>%
  filter(POOL == "04",
         str_detect(MSTRATUM, "-L")) %>%
  select(DATE, DETRITUS, SBSTRATE, VEG_S, VEG_RF,
         VEG_E, VEG_A, VEG_NRF, VEG_U,
         COV_NRF, COV_RF, COV_E,
         RAKE1, RAKE2, RAKE3, RAKE4, RAKE5, RAKE6) %>% head()
```

```
##         DATE DETRITUS SBSTRATE VEG_S VEG_RF VEG_E VEG_A VEG_NRF VEG_U COV_NRF
## 1 07/08/1998        1        6     S     NA     E     A       N             1
## 2 07/08/1998        1        6     S     NA     E     A       N             1
## 3 07/08/1998        1        6     S     NA     E     A       N             1
## 4 07/08/1998        1        6     S     NA     E     A       N             1
## 5 07/08/1998        1        6     S     NA     E     A       N             1
## 6 07/08/1998        1        6     S     NA     E     A       N             1
##   COV_RF COV_E RAKE1 RAKE2 RAKE3 RAKE4 RAKE5 RAKE6
## 1      0     1     0     0     1     0     1     0
## 2      0     1     0     0     1     1     0     0
## 3      0     1     0     0     0     0     0     0
## 4      0     1     0     0     0     0     0     0
## 5      0     1     0     0     0     0     0     0
## 6      0     1     0     0     0     1     0     0
```

```
veg_4L <- veg19 %>%
  filter(POOL == "04",
         str_detect(MSTRATUM, "-L")) %>%
  mutate(DATE = mdy(DATE))
```

Rake distribution

```
veg_4L %>%
  mutate(year = year(DATE)) %>%
  select(year, RAKE1, RAKE2, RAKE3, RAKE4, RAKE5, RAKE6) %>%
  pivot_longer(!year, names_to = "subsample_rake",
               values_to = "level") %>%
  group_by(year) %>%
  summarize(mean_level = mean(level),
            sd_level = sd(level)) %>%
  ggplot(aes(x = year, y = mean_level,
             fill = sd_level)) +
  geom_point()
```

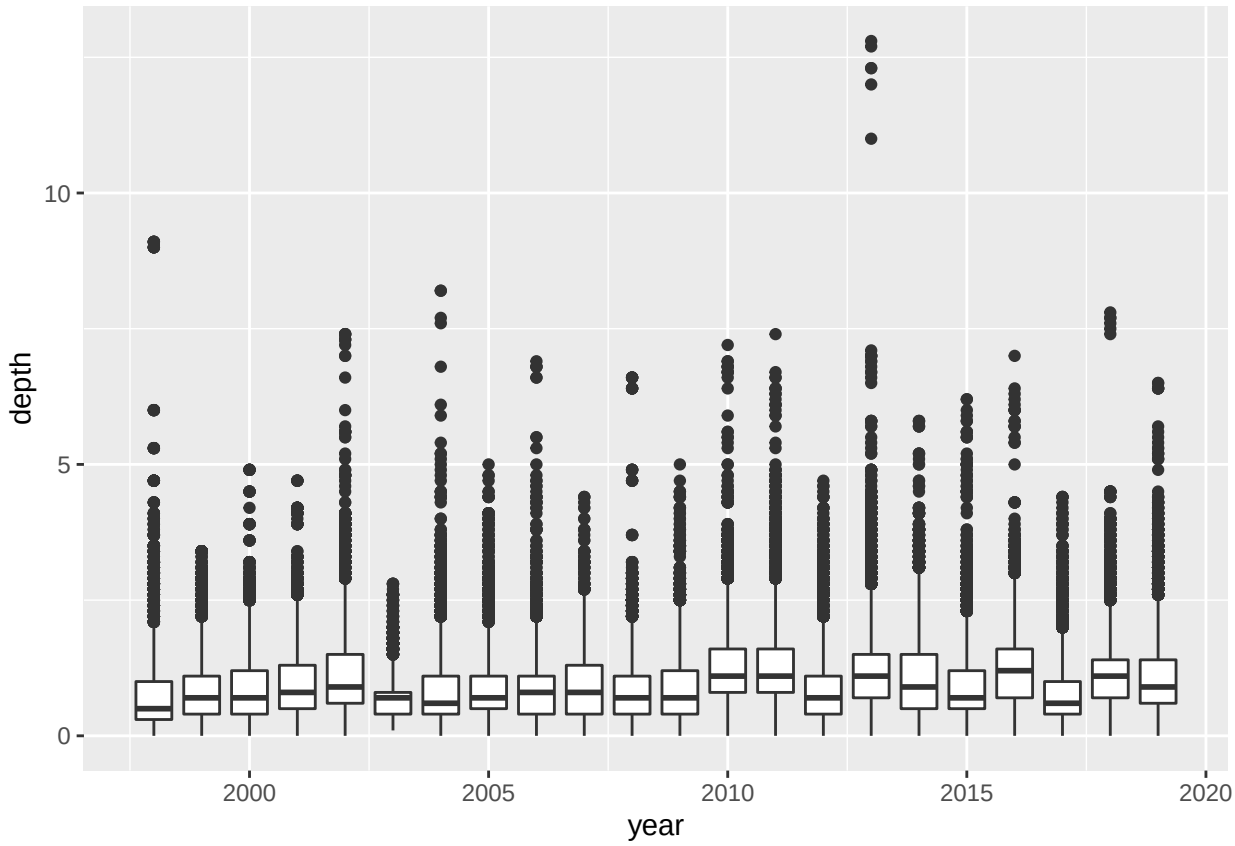



```
  #
  # ggplot(aes(x = RAKE1)) +
  # geom_histogram()
```

Water depth

```
veg_4L %>%
  mutate(year = year(DATE)) %>%
  select(year, starts_with("DEPTH")) %>%
  pivot_longer(!year, names_to = "subsample",
               values_to = "depth") %>%
  ggplot(aes(x = year, y = depth, group = year)) +
  geom_boxplot()
```

```
## Warning: Removed 132 rows containing non-finite values (stat_boxplot).
```

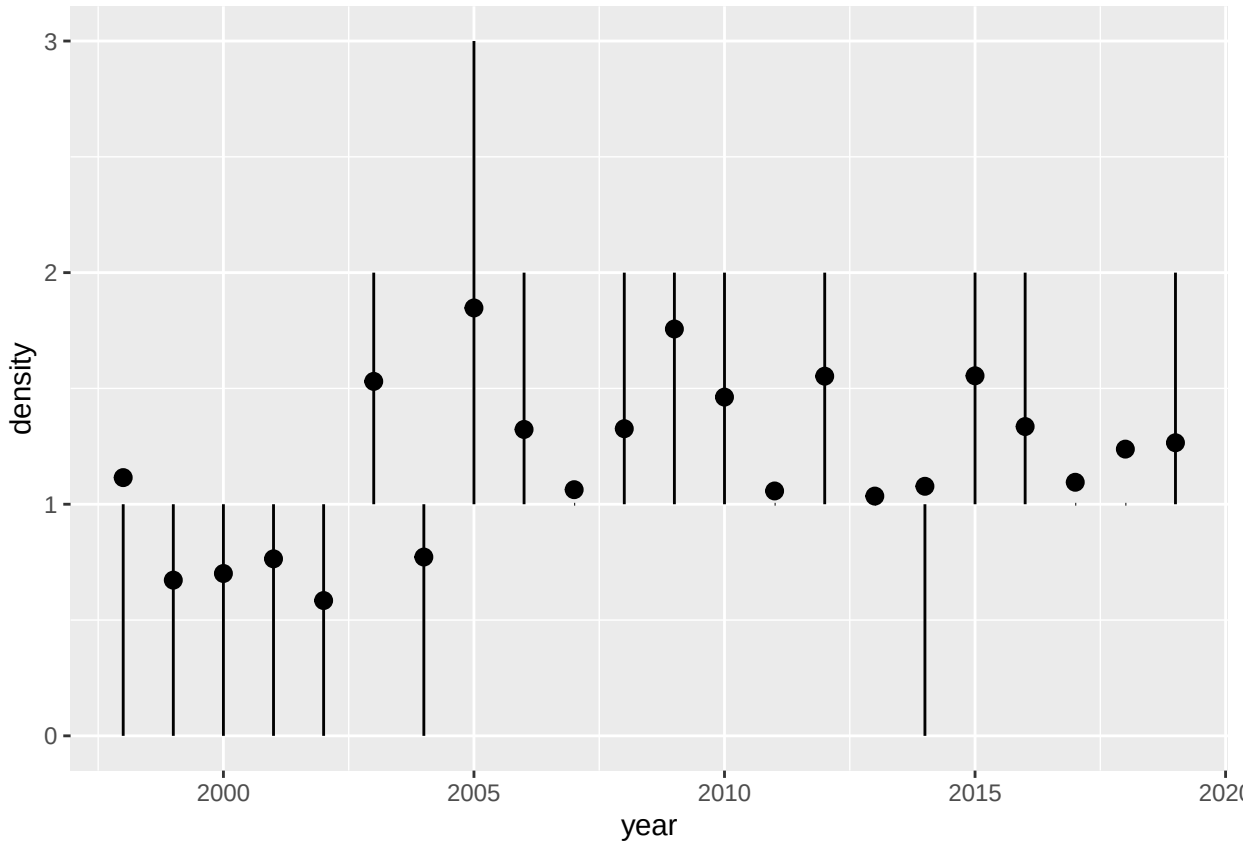



```
  #
  # ggplot(aes(x = RAKE1)) +
  # geom_histogram()
```

plant density

```
veg_4L %>%
  mutate(year = year(DATE)) %>%
  select(year, starts_with("DENSITY")) %>%
  pivot_longer(!year, names_to = "subsample",
               values_to = "density") %>%
  ggplot(aes(x = year, y = density, group = year)) +
  stat_summary(
    mapping = aes(x = year, y = density),
    fun.min = function(z) { quantile(z,0.25) },
    fun.max = function(z) { quantile(z,0.75) },
    fun = mean)
```

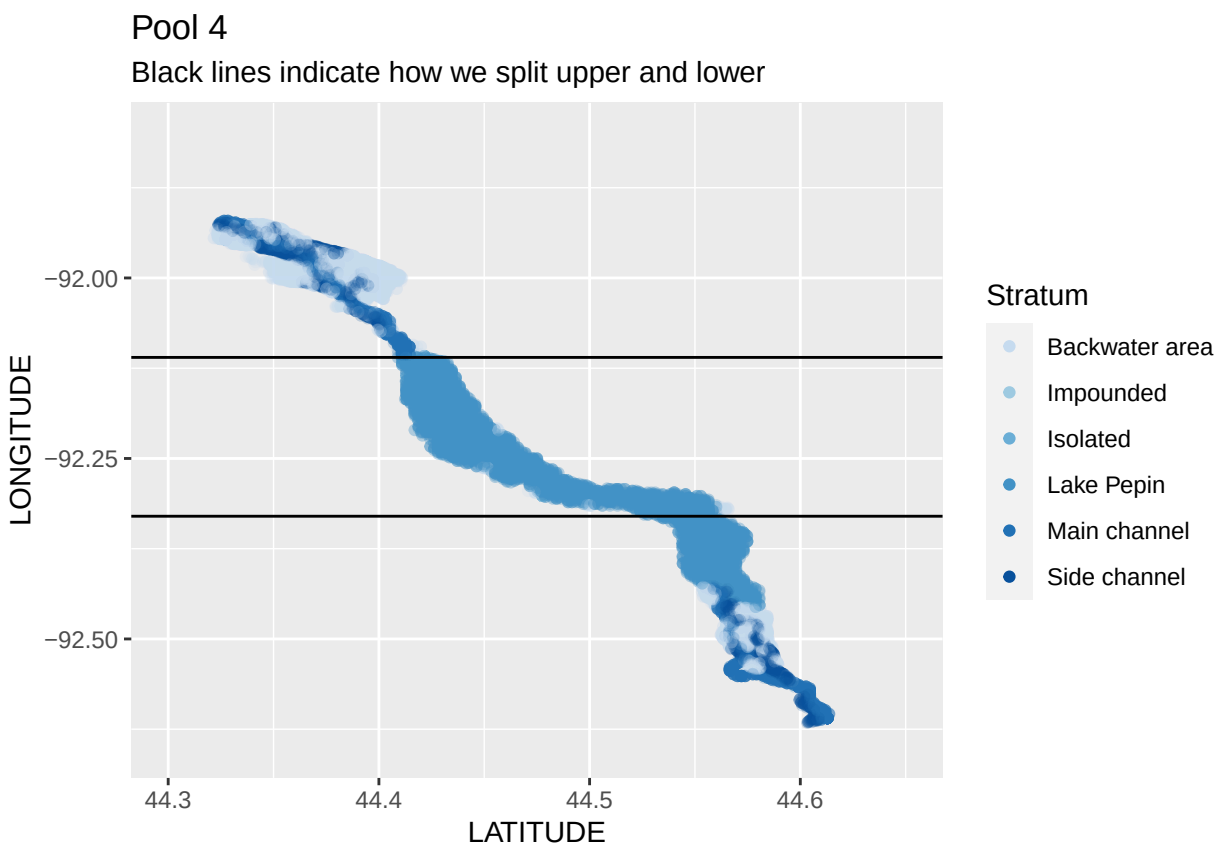
```
## Warning: Removed 204 rows containing non-finite values (stat_summary).
```



Water data

```
water20 %>%
  filter(FLDNUM == "1", !is.na(STRATUM)) %>%
  mutate(`Pool 4 Regions` = case_when(LONGITUDE <= -92.33 ~ "Lower",
                                      LONGITUDE >= -92.11 ~ "Upper",
                                      TRUE ~ "A lake"),
         Stratum = case_when(STRATUM == 1 ~ "Main channel",
                             STRATUM == 2 ~ "Side channel",
                             STRATUM == 3 ~ "Backwater area",
                             STRATUM == 4 ~ "Lake Pepin",
                             STRATUM == 5 ~ "Impounded",
                             STRATUM == 6 ~ "Isolated",
                             STRATUM == 7 ~ "New Terrestrial"),
         DATE = mdy(DATE),
         year = year(DATE)) %>%
  ggplot(aes(x = LATITUDE, y = LONGITUDE)) +
  geom_point(aes(color = Stratum), alpha = 0.2) +
  scale_color_manual(values = brewer.pal(9,'Blues')[3:10]) +
  geom_abline(intercept = -92.11, slope = 0) +
  geom_abline(intercept = -92.33, slope = 0) +
  ggtitle("Pool 4",
          subtitle = "Black lines indicate how we split upper and lower") +
  guides(colour = guide_legend(override.aes = list(alpha = 1))) +
  scale_x_continuous(limits = c(44.3, 44.65)) +
```

```
  scale_y_continuous(limits = c(-92.65, -91.8))
```




```
ggsave("splitting_pool4.png")
```

```
## Saving 6.5 x 4.5 in image
```

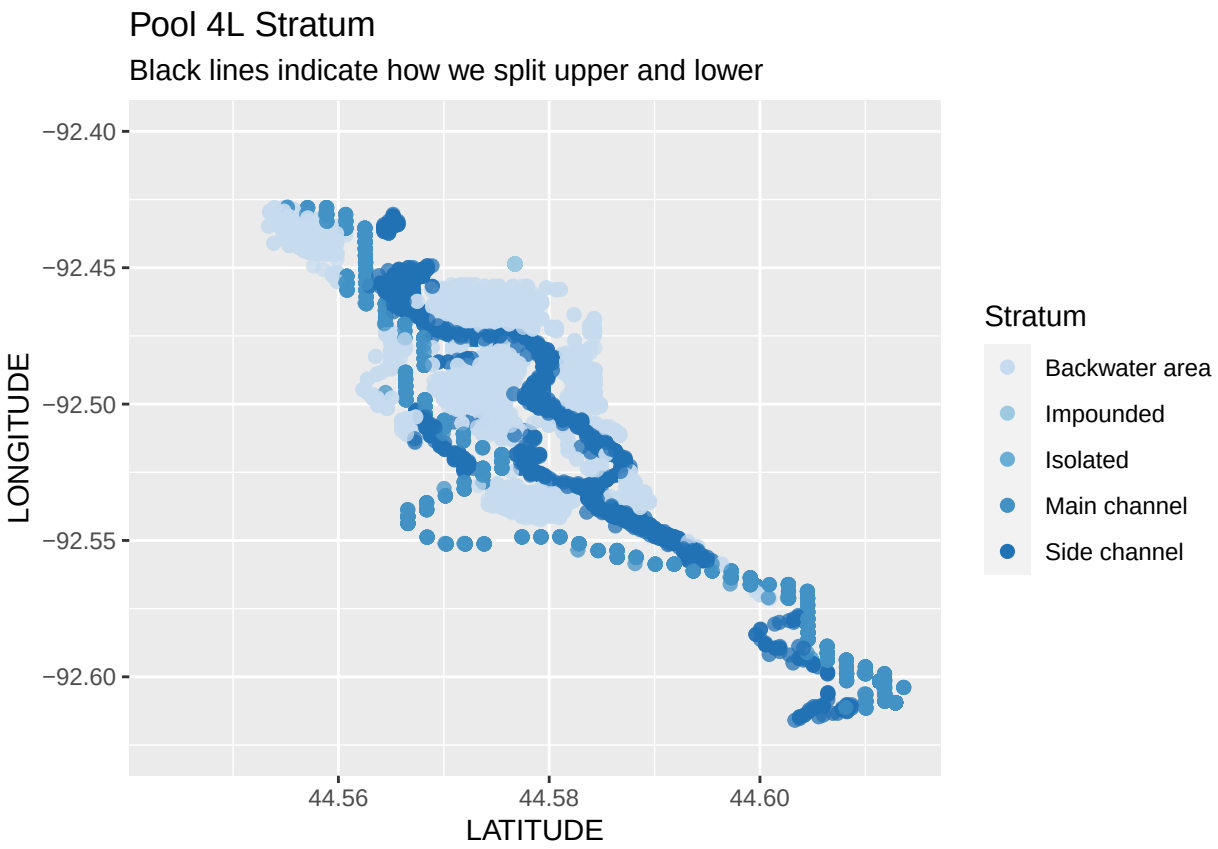
```
#3000 out of 204k observations removed
```

```
water_var <- c('TN','TP','TEMP','DO','TURB',
               'COND','VEL','SS','WDP','CHLcal','SECCHI')

# filter for pool 4 lower
water_4L <- water20 %>%
  filter(FLDNUM == "1") %>%
  mutate(DATE = mdy(DATE),
         year = year(DATE),
         is_lower = case_when(LONGITUDE <= -92.33 ~ TRUE,
                              TRUE ~ FALSE),
         decade = case_when(year <= 2000 ~ "1993-2000",
                            year >= 2001 & year <= 2014 ~ "2001-2014",
                            year >= 2014 ~ "2014-2020"),
         Stratum = case_when(STRATUM == 1 ~ "Main channel",
                             STRATUM == 2 ~ "Side channel",
                             STRATUM == 3 ~ "Backwater area",
                             STRATUM == 4 ~ "Lake Pepin",
                             STRATUM == 5 ~ "Impounded",
                             STRATUM == 6 ~ "Isolated",
                             STRATUM == 7 ~ "New Terrestrial")) %>%
```

```
  filter(is_lower & STRATUM != 4)
```

```
ggplot(water_4L, aes(x = LATITUDE, y = LONGITUDE)) +
  geom_point(aes(color = Stratum), alpha = 0.75, size = 2) +
  scale_color_manual(values = brewer.pal(7,'Blues')[2:7]) +
  ggtitle("Pool 4L Stratum",
          subtitle = "Black lines indicate how we split upper and lower") +
  guides(colour = guide_legend(override.aes = list(alpha = 1))) +
  scale_y_continuous(limits = c(-92.625, -92.4))
```

```
## Warning: Removed 840 rows containing missing values (geom_point).
```




```
ggsave("stratum_pool4L.png")
```

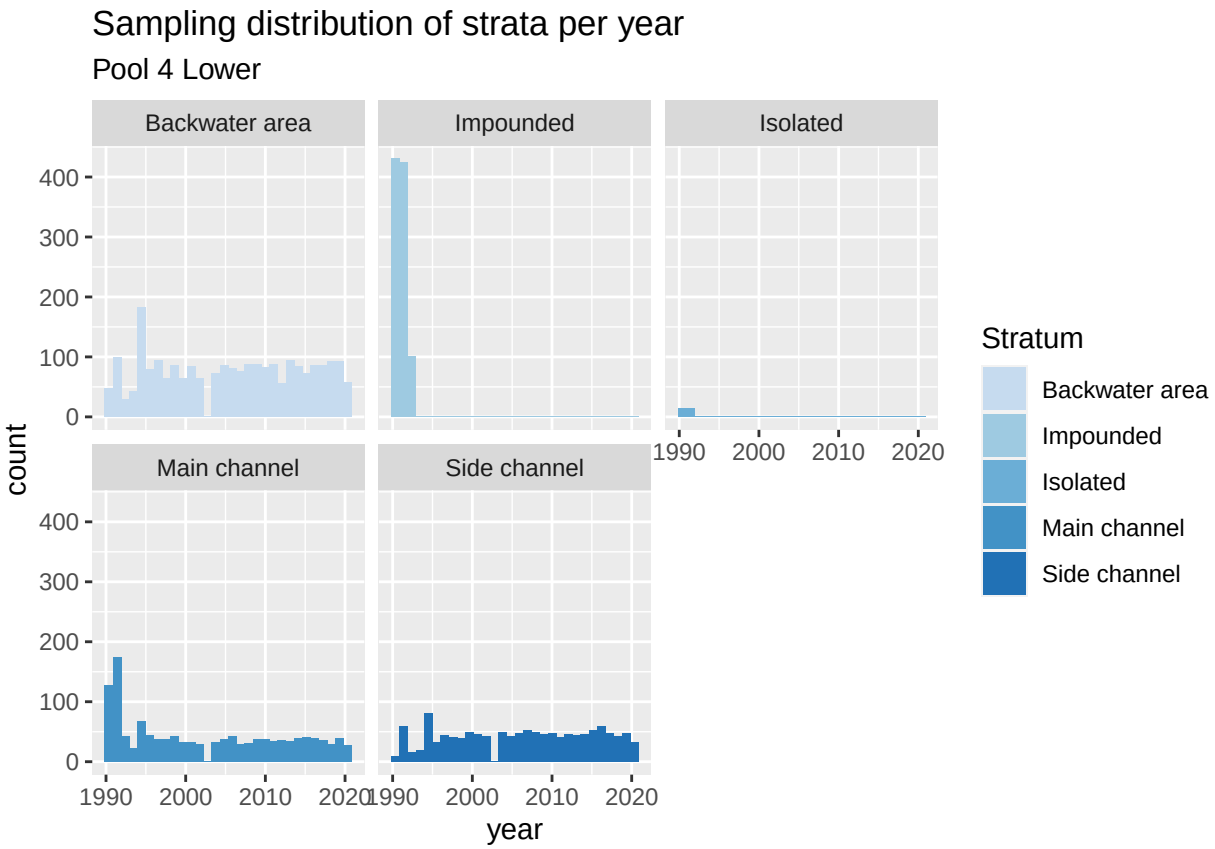
```
## Saving 6.5 x 4.5 in image
```

```
## Warning: Removed 840 rows containing missing values (geom_point).
```

Sampling distribution

```
ggplot(water_4L) +
  geom_histogram(aes(x = year, fill = Stratum)) +
  facet_wrap(~ Stratum) +
  scale_fill_manual(values = brewer.pal(7,'Blues')[2:7]) +
  ggtitle("Sampling distribution of strata per year",
          subtitle = "Pool 4 Lower")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggsave("sampling_distribution_year_strata.png")
```

```
## Saving 6.5 x 4.5 in image
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

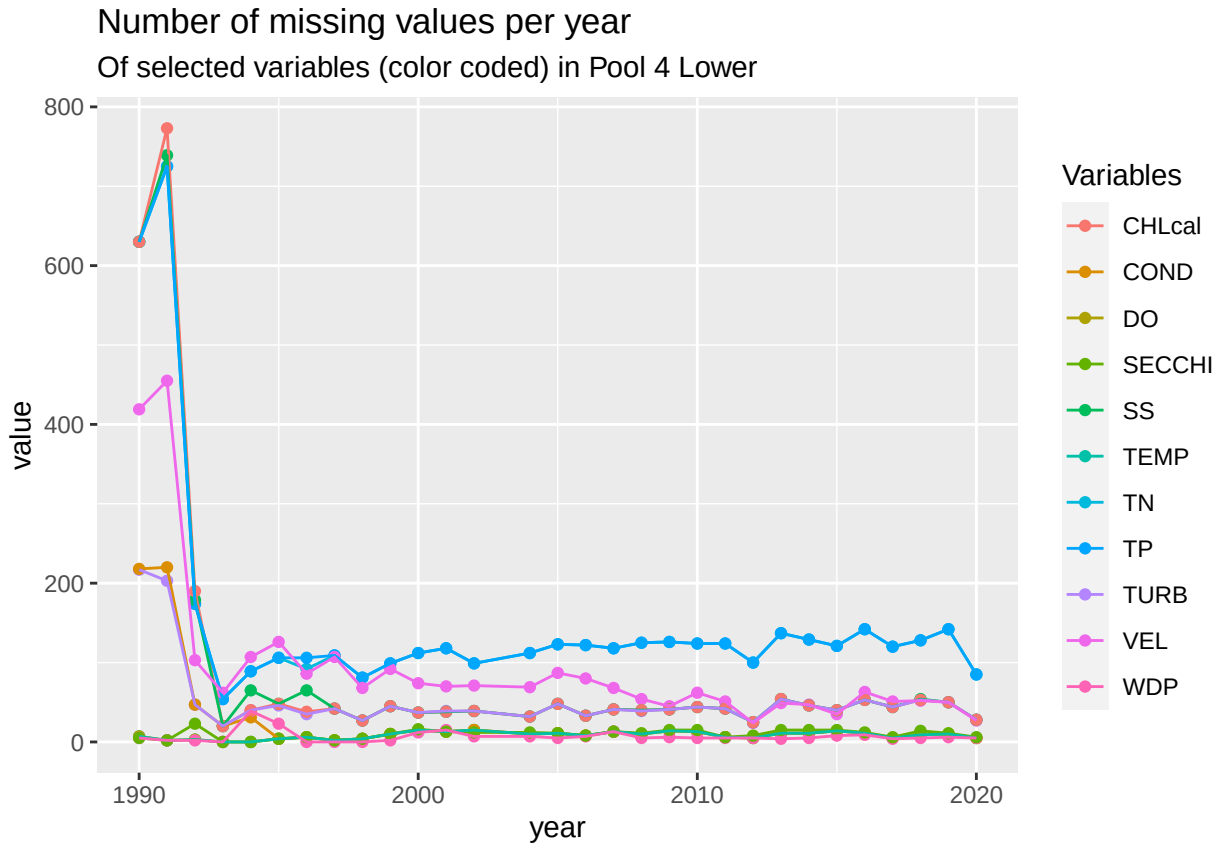
Missing values

- number of missing values per variable per year
- Missingness by decade (zooming out)

```
water_4L <- bind_cols(water_4L, missing = rowSums(is.na(water_4L)))

# data.frame(Variables = names(water_4L), Missing = colSums(is.na(water_4L))) %>%
#   filter(Variables %in% water_var) %>%
#   ggplot(aes(x = Variables, y = Missing)) +
#   geom_bar(stat = "identity")

water_4L %>%
  group_by(year) %>%
  summarise(across(water_var, ~ sum(is.na(.x)))) %>%
  pivot_longer(water_var, names_to = "Variables") %>%
  ggplot(aes(x = year, y = value)) +
  geom_point(aes(color = Variables)) +
  geom_line(aes(color = Variables)) +
```

```
  ggtitle("Number of missing values per year",
          subtitle = "Of selected variables (color coded) in Pool 4 Lower")
```

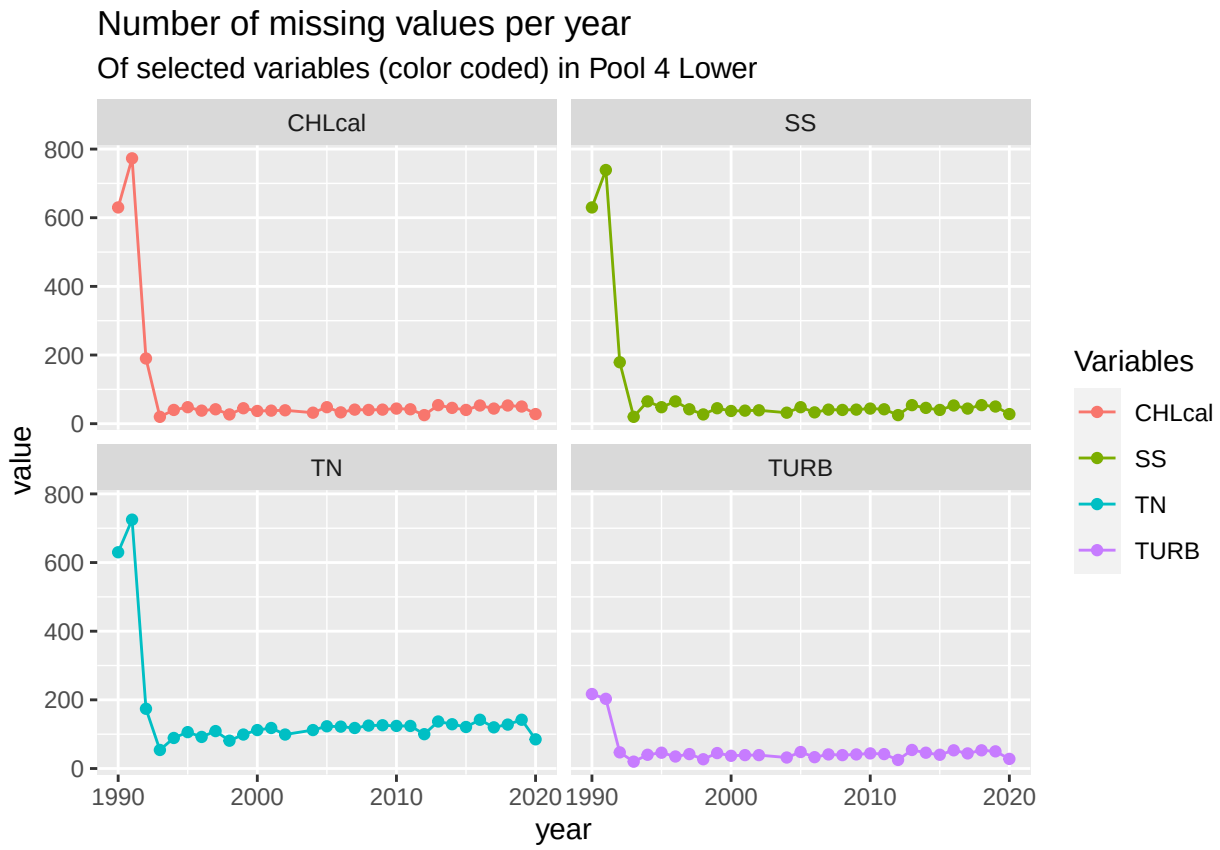
```
## Note: Using an external vector in selections is ambiguous.
## i Use `all_of(water_var)` instead of `water_var` to silence this message.
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This message is displayed once per session.
```




```
ggsave("missing_count_selected_pool4L.png")
```

```
## Saving 6.5 x 4.5 in image
```

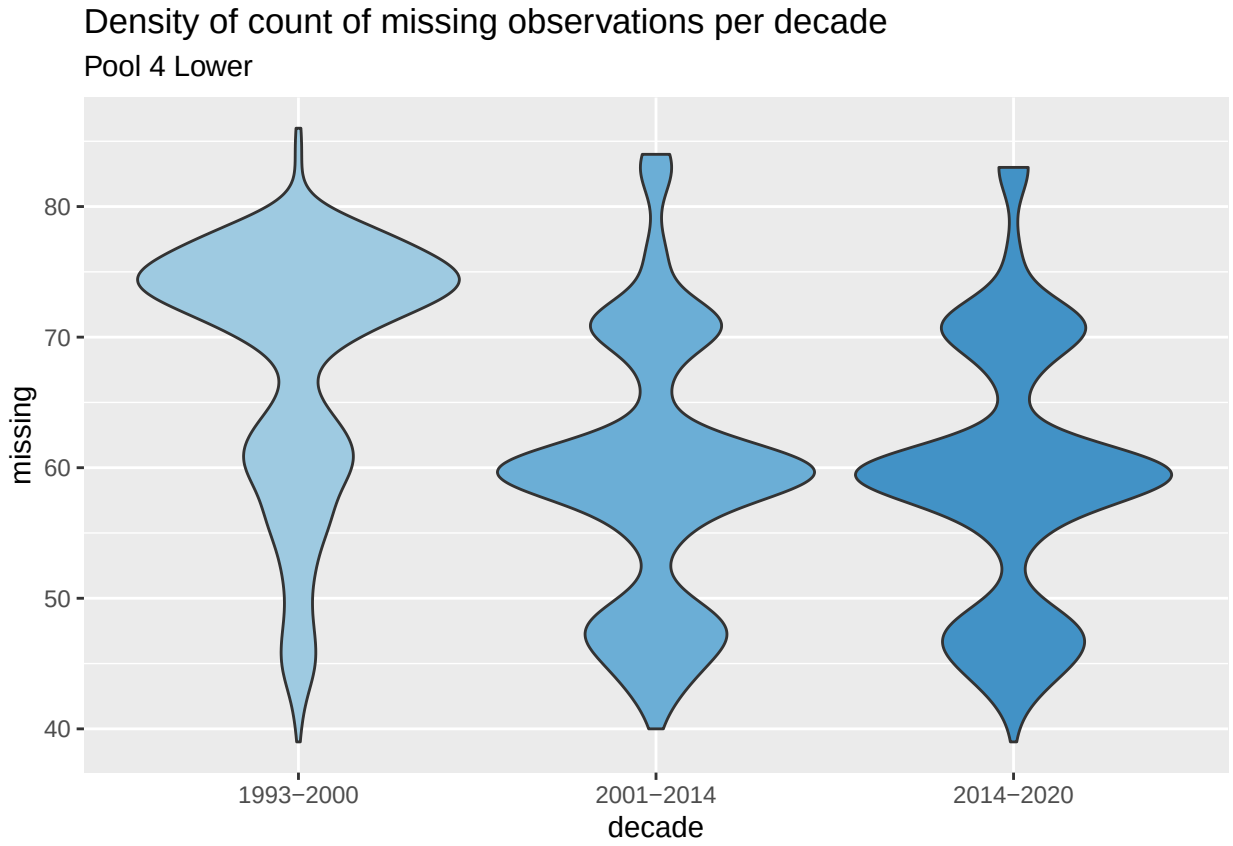
```
water_4L %>%
  select(year, all_of(c("TN", "TURB", "SS", "CHLcal"))) %>%
  group_by(year) %>%
  summarise(across(c("TN", "TURB", "SS", "CHLcal"), ~ sum(is.na(.x))) ) %>%
  pivot_longer(c("TN", "TURB", "SS", "CHLcal"), names_to = "Variables") %>%
  ggplot(aes(x = year, y = value)) +
  geom_point(aes(color = Variables)) +
  geom_line(aes(color = Variables)) +
  facet_wrap(~ Variables) +
  ggtitle("Number of missing values per year",
          subtitle = "Of selected variables (color coded) in Pool 4 Lower")
```




```
ggsave("missing_count_4_pool4L.png")
```

```
## Saving 6.5 x 4.5 in image
```

```
ggplot(water_4L, aes(x = decade, y = missing)) +
  geom_violin(aes(group = decade, fill = decade)) +
  ggtitle("Density of count of missing observations per decade",
          subtitle = "Pool 4 Lower") +
  scale_fill_manual(values = brewer.pal(7,'Blues')[3:6]) +
  theme(legend.position = "none")
```




```
ggsave("missing_violin_pool4L.png", height = 4, width = 6)
```

Continuous variable box plots

```
plotter_box_by_year <- function(var_str, data, facet_bool){

  # facet_bool gives if you should facet by STRATUM type

  title <- paste("boxplot_by_year", "4L", var_str, sep = "_")

  if (facet_bool){title <- paste(title, "_facet")}

  title <- paste(title, "png", sep = ".")

  if (facet_bool){
    data %>%
      filter(!is.na(!!sym(var_str))) %>%
      ggplot(aes(x = year, y = !!sym(var_str), group = year)) +
      geom_boxplot() +
      facet_wrap(~ STRATUM)
  } else {
    data %>%
      filter(!is.na(!!sym(var_str))) %>%
      ggplot(aes(x = year, y = !!sym(var_str), group = year)) +
      geom_boxplot()
  }
```

```
  ggsave(title)

}

sapply(c("TN", "TURB", "SS", "CHLcal"), plotter_box_by_year, water_4L, F)
```

```
## Saving 6.5 x 4.5 in image
## Saving 6.5 x 4.5 in image
## Saving 6.5 x 4.5 in image
## Saving 6.5 x 4.5 in image
```

```
## $TN
## NULL
##
## $TURB
## NULL
##
## $SS
## NULL
##
## $CHLcal
## NULL
```

```
sapply(c("TN", "TURB", "SS", "CHLcal"), plotter_box_by_year, water_4L, T)
```

```
## Saving 6.5 x 4.5 in image
## Saving 6.5 x 4.5 in image
## Saving 6.5 x 4.5 in image
## Saving 6.5 x 4.5 in image
```

```
## $TN
## NULL
##
## $TURB
## NULL
##
## $SS
## NULL
##
## $CHLcal
## NULL
```

```
# SS, turbidity, and secchi all three measure water quality

# velocity graph is new

# dissolved oxygen is variable by the time of day

# temperature is important per season

# habitat class for these water quality data is all NA
# unique(water_4L$HABCLASS)

unique(water_4L$STRATUM)
```

```
## [1] 5 6 1 2 3
```

```
table(water_4L$STRATUM)
```

```
##
##    1    2    3    5    6
## 1299 1274 2322  956   30
```

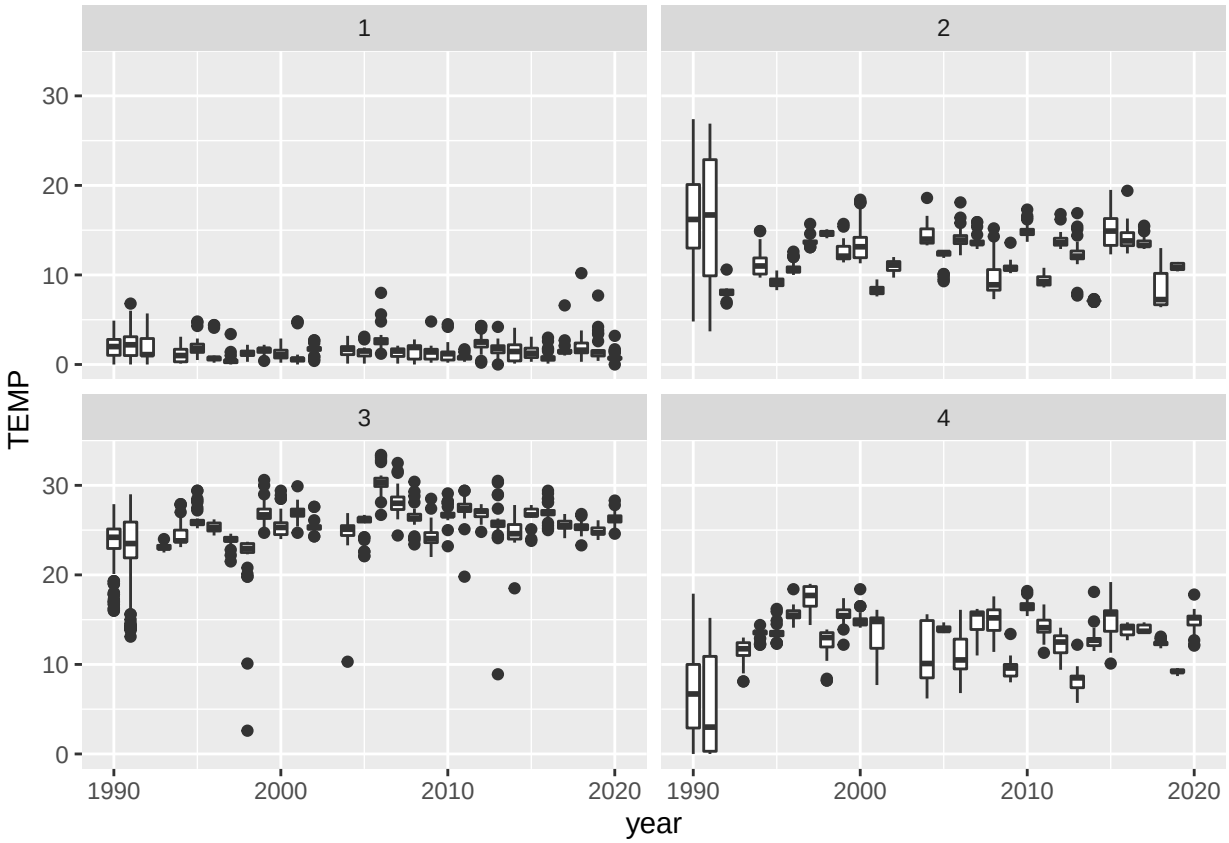
Note: stratum types 1 = Main channel 2 = Side channel 3 = Backwater area contiguous to the main channel 4 = Lake Pepin or Swan Lake

- total nitrogen is variable by year, but within year, not too variable
- velocity has increased, water volume is greater. increase in velocity reflects collection in data? collect

Temperature

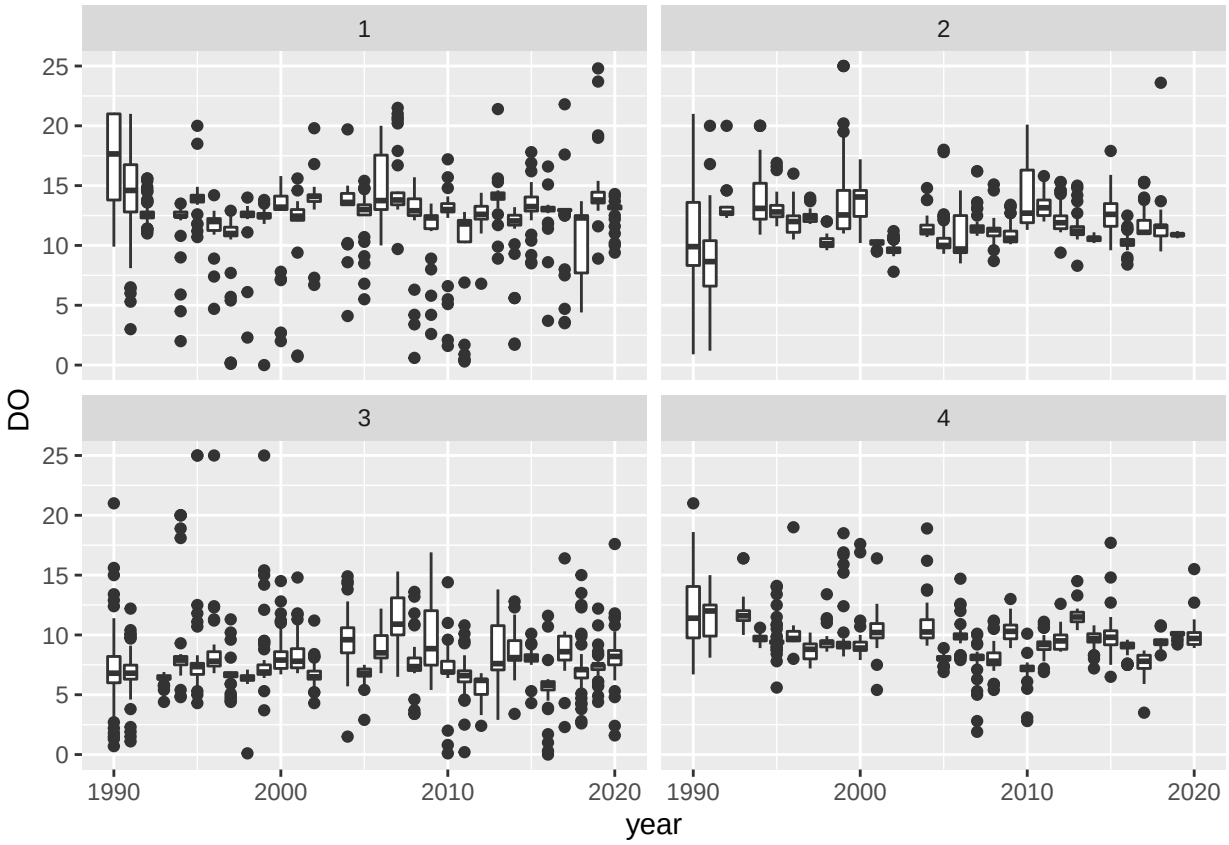
```
water_4L %>%
  mutate(season = quarter(DATE)) %>%
  ggplot(aes(x = year, y = TEMP, group = year)) +
  geom_boxplot() +
  facet_wrap(~ season)
```

```
## Warning: Removed 247 rows containing non-finite values (stat_boxplot).
```



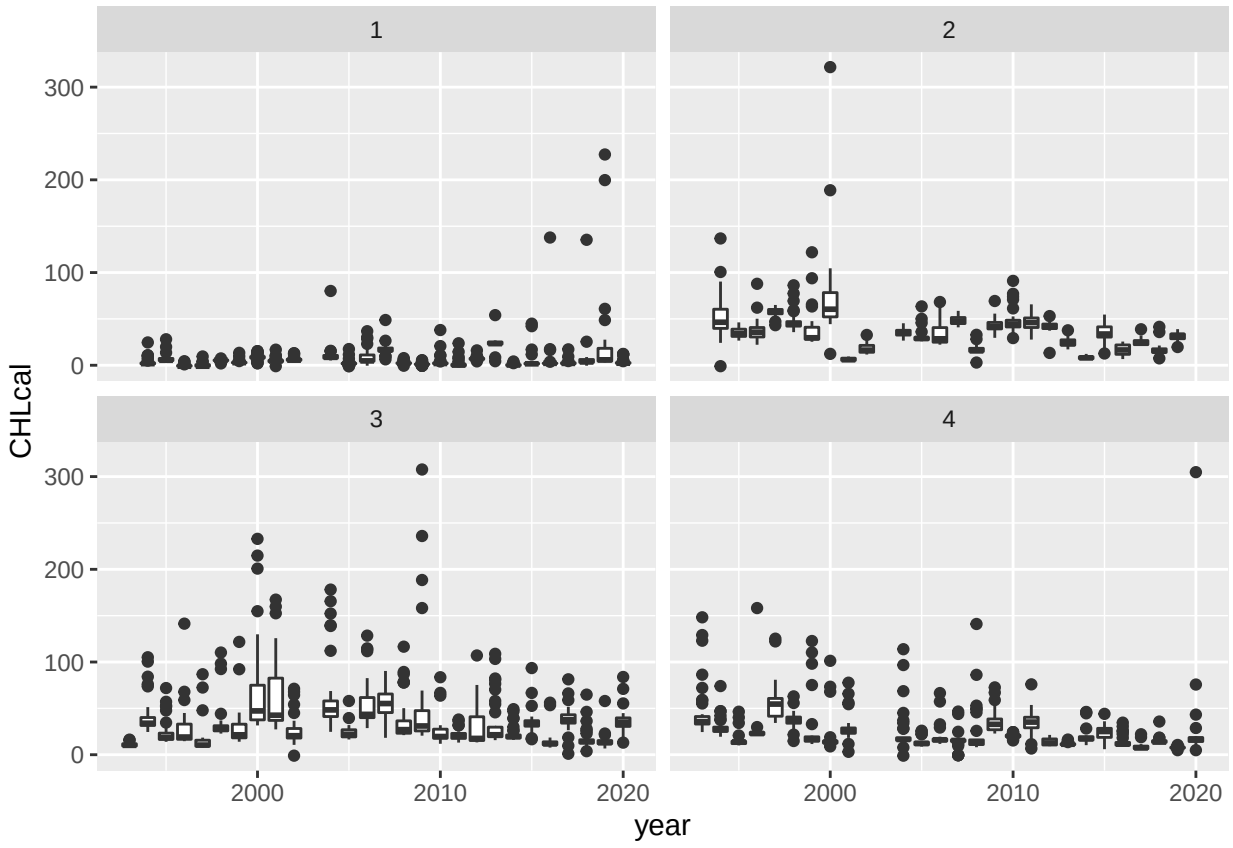

```
water_4L %>%
  mutate(season = quarter(DATE)) %>%
  ggplot(aes(x = year, y = DO, group = year)) +
  geom_boxplot() +
  facet_wrap(~ season)
```

```
## Warning: Removed 247 rows containing non-finite values (stat_boxplot).
```




```
water_4L %>%
  mutate(season = quarter(DATE)) %>%
  ggplot(aes(x = year, y = CHLcal, group = year)) +
  geom_boxplot() +
  facet_wrap(~ season)
```

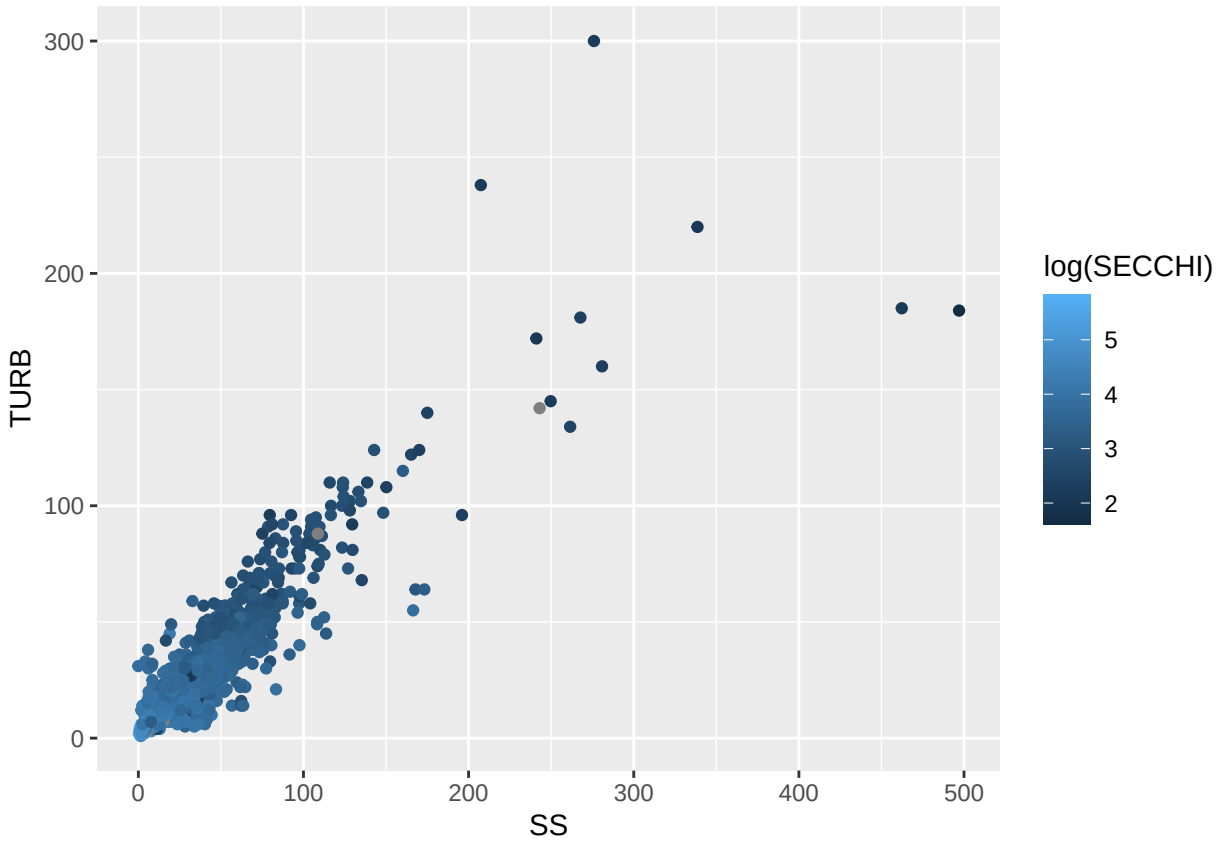
```
## Warning: Removed 2681 rows containing non-finite values (stat_boxplot).
```



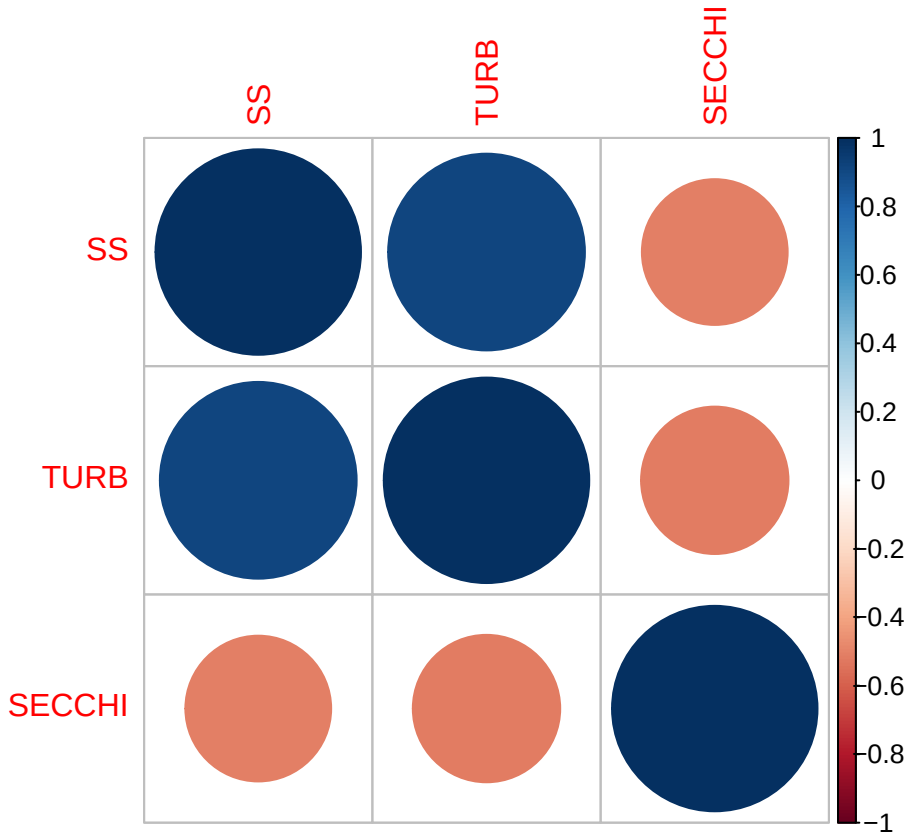
Water quality

```
waterqual_4L <- water_4L %>%
                    select(SS, TURB, SECCHI)

waterqual_4L <- waterqual_4L %>%
  bind_cols(missing = rowSums(is.na(waterqual_4L))) %>%
  filter(missing == 0) %>% # remove missing
  select(-missing)

ggplot(data = water_4L, aes(x = SS, y = TURB)) +
  geom_point(aes(color = log(SECCHI)))
```

```
## Warning in log(SECCHI): NaNs produced

## Warning in log(SECCHI): NaNs produced

## Warning: Removed 2691 rows containing missing values (geom_point).
```




```
corrplot(round(cor(waterqual_4L ), digits = 4))
```




```
ggsave("waterquality_corrplot.png")
```

```
## Saving 6.5 x 4.5 in image
## Warning in log(SECCHI): NaNs produced
## Warning in log(SECCHI): NaNs produced
## Warning: Removed 2691 rows containing missing values (geom_point).
```

website