Predicting the Distribution of Gastroenteritis-Causing Agents and its Physicochemical Drivers

Antonette Tadle

2023-12-22

I. Introduction

Ecologically, rivers provide significant habitats for a range of plants and wildlife. However, they are also prone to pollution due to a variety of human-induced hazards such as effluents from industries and sewage discharge, runoffs, lack of treatment and sewage systems, sedimentation, and rapid urbanization (Lin et al., 2022). The Butuanon River is a 23-kilometer river nestled between Mandaue and Cebu City that is defined as Cebu's "dead river" because of extreme pollution which hinders floral and faunal growth (Parilla et al., 2019). High levels of contaminants, such as industrial effluents and untreated sewage, disrupt the river's ecosystem. This pollution can lead to a decline in aquatic plant populations, affecting the oxygen levels and nutrient cycling essential for a healthy ecosystem (Garg et al., 2022). Additionally, it poses a threat to various fauna species by contaminating their habitats, disrupting food chains, and causing long-term harm to biodiversity (Kaur and Brraich, 2022). Being one of the seven major rivers of Cebu and as a crucial water source, Butuanon river plays a vital role in the lives of the local population, making it imperative to understand the epidemiological drivers of water-borne diseases in this context. Discerning the environmental factors supporting the presence of a disease agent elucidates the disease's ecology and geography (Peterson, 2007).

Gastroenteritis is among the highest causes of morbidity in Cebu province and is the one of the main causes of infant mortality. It is associated with a variety of identified pathogens in the Butuanon river including Salmonella, Enterococcus, Enterobacter, E. coli, Klebsiella, Citrobacter, Serratia, Pseudomonas, Proteus, Providencia, and Acinetobacter which causes viral diarrhea. Short-term health responses to water quality investigations have revealed effects on infant mortality and gastroenteritis incidence related to biological contamination. These responses are significantly connected because fecal coliform contamination resulting in repeated bouts of diarrhea has been proven to impair the child's ability to absorb nutrition, leading to mortality or growth stunting (Gadgil, 1998).

Total coliforms, a group of bacteria found in the environment, are widely used as indicators of water quality, especially in assessing the safety of drinking water and recreational waters. They include a range of bacteria, some of which originate from the intestines of warm-blooded animals, including humans. Their presence in water indicates fecal contamination and the potential presence of pathogens that cause gastroenteritis, such as E. coli, Salmonella, and Shigella. Several studies have established a correlation between elevated total coliform counts in water sources and increased cases of gastroenteritis within affected communities. However, variability in coliform levels due to seasonal changes, geographical factors, and human activities requires sophisticated modeling techniques for accurate predictions.

Currently, disease biogeography is poorly addressed in biodiversity research (Peterson et al., 2011). While substantial research has been conducted on the distribution of species and their interactions in ecosystems, the study of diseases within this context has received insufficient attention. As a result, there is a substantial gap in our understanding of how illnesses spread across different geographical regions and interact with biodiversity patterns. Species distribution modeling is used in epidemiology for mapping spatial disease patterns, the prediction of disease introduction risks through pathogen-host interactions, and the prediction of exposure changes owing to future environmental changes (Martínez-Minaya et al., 2018). While historical records of infections and vectors were limited, the introduction of monitoring systems and media sources has resulted in new online data sources on their occurrences. Furthermore, distribution modeling tools are

becoming more ecologically realistic by taking dispersal, biotic interactions, and evolutionary restrictions into account. These factors, together with abiotic circumstances and recording biases, jointly influence disease, vector, and wildlife species distributions.

Ecological niche modeling (ENM) is an advanced computational approach that is used to explain the occurrence of infectious agents with the influence of environmental factors (Escobar and Craft, 2016). Specifically, correlative ENM is a convenient approach in predicting site-specific biogeography since they only require environmental data associated with species occurrence locations to predict the spatial distribution of species (Graham et al., 2004; Guisan and Thuiller, 2005, as cited in Sebes, 2023). It is preferred over mechanistic ENM modeling because it can capture environmental and socioeconomic factors for predicting and understanding patterns in transmission (Hay et al., 2013, as cited in Purse and Golding, 2015). Consequently, species distribution modeling (SDM) is often associated with ENM to correlate species occurrence data with environmental data, such as MaxEnt (Phillips et al., 2006, as cited in Sebes, 2023). These modeling approaches are supplemented with GAMs and ANNs for generating nonlinear statistical models, species distribution modeling, environmental monitoring, and image classification; GCMs to model the spread of species and disease changes over time under different future climatic conditions; and Ensemble to combine multiple algorithms or models to improve predictive accuracy and robustness.

This study addresses diverse challenges related to these approaches including better data organization, such as quantification of pixel count changes and modeling at increased resolutions. Applying correlative ENM and SDM to the context of water-borne diseases provides a novel perspective on disease ecology. This study can identify pollution sources and vulnerable areas by generating species distribution maps of gastroenteritis-causing pathogens and physicochemical factors. Furthermore, the study's findings can inform targeted interventions and regulations for pollution sources, recommend effective water treatment methods, assess health risks associated with pathogens, promote community awareness for water safety, and contribute to evidence-based policymaking and climate change adaptation for better public health outcomes.

II. Data

Existing data from Project REHAB containing identified species and physicochemical data in Butuanon River water samples was used as input data points in this project. This data was obtained from sampling periods in July and September 2023 and was wrangled and transformed into a uniform format. The data includes latitude and longitude values for both species occurrence and environmental data, as well as values for each physicochemical parameter including temperature, oxidation-reaction potential, pH, DOP, DO, EC, TDS, salinity, BGA-PC, altitude, and barometer.

```
library(readxl)
excel_file <- read_excel("~/Desktop/bio 118-class/exercises/bio 118 project/bio-118-project-attadle/Data
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

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

```
glimpse(excel_file)
```

```
## Rows: 188
## Columns: 5
## $ Site    <chr> "S6", "S14", "S10", "S5", "S6", "S7", "S9", "S10", "S11", "S12~
## $ Lon     <dbl> 123.9191, 123.9724, 123.9349, 123.9174, 123.9203, 123.9197, 12~
## $ Lat     <dbl> 10.38379, 10.34038, 10.35251, 10.41272, 10.39078, 10.38156, 10~
```

```
## $ Date    <dttm> 2023-07-24, 2023-07-24, 2023-07-24, 2023-07-24, 2023-07-24, 2~
## $ Species <chr> "Acinetobacter", "Acinetobacter", "Aeromonas", "Citrobacter", ~
```

```
head(excel_file)
```

```
## # A tibble: 6 x 5
##   Site    Lon   Lat Date                Species
##   <chr> <dbl> <dbl> <dttm>              <chr>
## 1 S6     124.  10.4 2023-07-24 00:00:00 Acinetobacter
## 2 S14    124.  10.3 2023-07-24 00:00:00 Acinetobacter
## 3 S10    124.  10.4 2023-07-24 00:00:00 Aeromonas
## 4 S5     124.  10.4 2023-07-24 00:00:00 Citrobacter
## 5 S6     124.  10.4 2023-07-24 00:00:00 Citrobacter
## 6 S7     124.  10.4 2023-07-24 00:00:00 Citrobacter
```

```
str(excel_file)
```

```
## tibble [188 x 5] (S3: tbl_df/tbl/data.frame)
##  $ Site   : chr [1:188] "S6" "S14" "S10" "S5" ...
##  $ Lon    : num [1:188] 124 124 124 124 124 ...
##  $ Lat    : num [1:188] 10.4 10.3 10.4 10.4 10.4 ...
##  $ Date   : POSIXct[1:188], format: "2023-07-24" "2023-07-24" ...
##  $ Species: chr [1:188] "Acinetobacter" "Acinetobacter" "Aeromonas" "Citrobacter" ...
```

III. Data Analysis

The central question in this study is to assess the distribution of gastroenteritis-causing agents by correlating the spatial distribution of species with physicochemical data. The predictor variables include the different physicochemical factors while the outcome variable is the abundance values of the species.

Since the data lacks abundance values of the species, a distribution map was created to plot the distribution of the species (per site) and the concentration values of each physicochemical variable in every site. A correlation matrix between the variables was generated to test the relationship between the physicochemical variables and determine the correlated variables which may pose a great impact on species presence.

IV. Results and Discussion

The data was filtered to include only the Gastroenteritis-causing pathogens in a species distribution map. The map features longitude values on the x-axis and latitude values on the y-axis.

```
# load the required libraries
library(tidyverse)
```

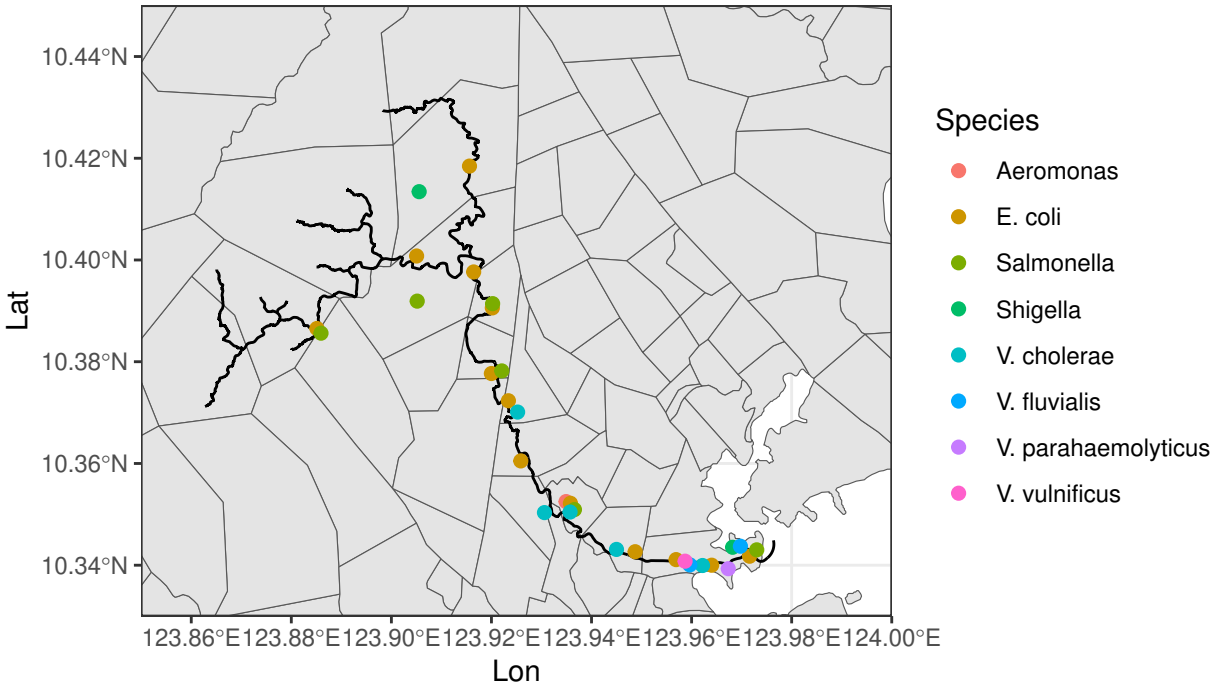
```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v readr     2.1.4
## v ggplot2   3.4.3     v stringr   1.5.0
## v lubridate 1.9.2     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.0
## -- Conflicts ------------------------------------------ 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
```

```
library(sp)
library(leaflet)
library(readxl)
library(ggplot2)
library(dplyr)
library(readxl)
```

```r
library(mapview)
library(spdep)
```

```
## Loading required package: spData
## To access larger datasets in this package, install the spDataLarge
## package with: `install.packages('spDataLarge',
## repos='https://nowosad.github.io/drat/', type='source')`
## Loading required package: sf
## Linking to GEOS 3.11.0, GDAL 3.5.3, PROJ 9.1.0; sf_use_s2() is TRUE
```

```r
library(patchwork)
library(sf)

# Species Occurrence Distribution Map
excel_file <- "~/Desktop/bio 118-class/exercises/bio 118 project/bio-118-project-attadle/Data/combined_
data <- read_excel(excel_file, sheet = "1. Species Occurrence")
selected_species <- c("E. coli", "Salmonella", "Shigella", "V. cholerae", "V. fluvialis", "V. parahaemol

# Filter data
filtered_data1 <- data %>%
  filter(Species %in% selected_species,
         !Species %in% c("Unknown", "Unknown (Vibrio)"))

# Convert to a regular data frame
filtered_data_df <- fortify(filtered_data1, region = "Species")

# Create an sp object
coordinates(filtered_data1) <- c("Lon", "Lat")

# River Shape Files
rivers <- sf::read_sf("~/Desktop/bio 118-class/exercises/bio 118 project/bio-118-project-attadle/Data/SH
ph_shp <- sf::read_sf("~/Desktop/bio 118-class/exercises/bio 118 project/bio-118-project-attadle/Data/SH

# Plot the map
ggplot(data = filtered_data_df) +
  geom_sf(data = ph_shp) +
  geom_sf(data = rivers) +
  coord_sf(xlim = c(123.85, 124), ylim = c(10.33, 10.45), expand = FALSE) +
  geom_point(aes(x = Lon, y = Lat, color = Species), size = 2) +
  theme_bw()
```



The graph shows the distribution of specific gastroenteritis-causing agents along the Butuanon river including Aeromonas, E. coli, Salmonella, Shigella, V. cholerae, V. fluvialis, V. parahaemolyticus, and V. vulnificus. It is evident that some points in the graph did not fall exactly within the Butuanon River course due to errors during the first sampling period. It can be inferred that E. coli had the largest distribution across all sites from upstream to downstream, followed by Salmonella, V. cholerae, and Shigella. Meanwhile, V. fluvialis, V. parahaemolyticus, V. vulnificus, and aeromonas had the least distribution across all sites.

For every physicochemical parameter, a distribution map was created, depicting longitude values on the x-axis and latitude values on the y-axis. Each parameter value was assigned a color gradient for visual representation.

```
library(leaflet)
library(mapview)
library(htmlwidgets)
library(readxl)
library(ggplot2)
library(dplyr)
library(patchwork)

# Read data
excel_file <- "~/Desktop/bio 118-class/exercises/bio 118 project/bio-118-project-attadle/Data/combined_
data2 <- read_excel(excel_file, sheet = "2. Environmental Factors")
data2$Parameter <- as.character(data2$Parameter)

# Temperature
temp <- dplyr::filter(data2, Parameter == "Temperature")
plot_temp <- ggplot(data = data2)+
  geom_sf(data = ph_shp)+
  geom_sf(data = rivers)+
  geom_tile(data = temp, aes(x = Lon, y = Lat, fill = Value), height = .005, width = .005)+
  coord_sf(xlim = c(123.85, 124), ylim = c(10.33, 10.45), expand = FALSE)+
  #geom_point(aes(x=Lon, y=Lat, color= Species), size= 2)+
  theme_bw()+
```

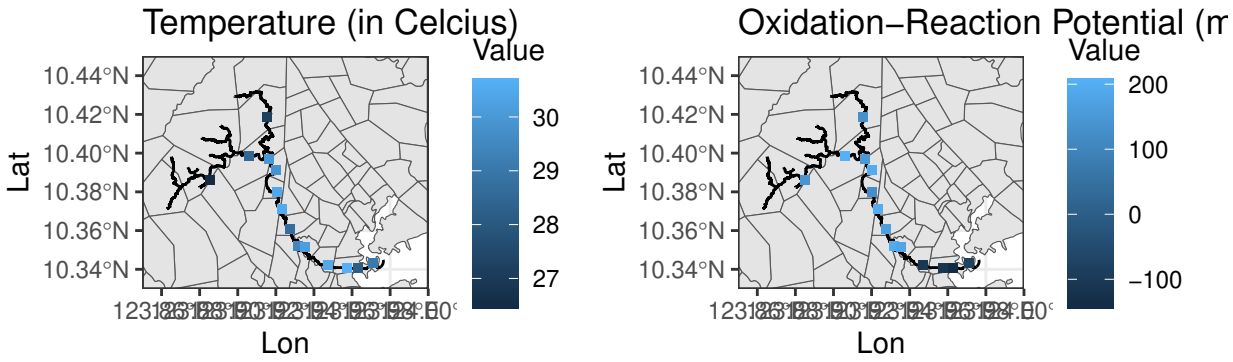
```
  ggtitle("Temperature (in Celcius)")

ggsave(filename = paste0("./Outputs/", "Temperature.png"), device = "png")
```

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

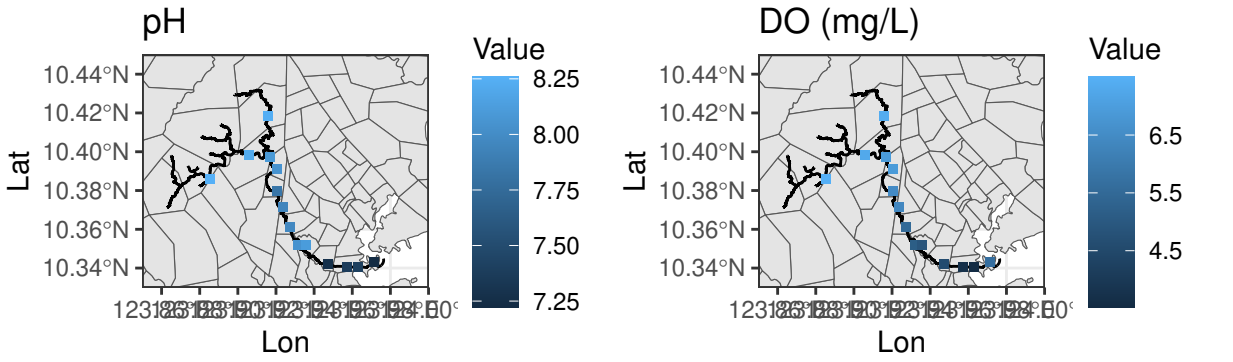
```
# Oxidation-Reaction Potential
OR <- dplyr::filter(data2, Parameter == "Oxidation-Reaction Potential")
plot_OR <- ggplot(data = data2)+
  geom_sf(data = ph_shp)+
  geom_sf(data = rivers)+
  geom_tile(data = OR, aes(x = Lon, y = Lat, fill = Value), height = .005, width = .005)+
  coord_sf(xlim = c(123.85, 124), ylim = c(10.33, 10.45), expand = FALSE)+
  #geom_point(aes(x=Lon, y=Lat, color= Species), size= 2)+
  theme_bw()+
  ggtitle("Oxidation-Reaction Potential (mV)")

# pH
pH <- dplyr::filter(data2, Parameter == "pH")
plot_pH <- ggplot(data = data2)+
  geom_sf(data = ph_shp)+
  geom_sf(data = rivers)+
  geom_tile(data = pH, aes(x = Lon, y = Lat, fill = Value), height = .005, width = .005)+
  coord_sf(xlim = c(123.85, 124), ylim = c(10.33, 10.45), expand = FALSE)+
  #geom_point(aes(x=Lon, y=Lat, color= Species), size= 2)+
  theme_bw()+
  ggtitle("pH")

# DO(mg/L)
DO <- dplyr::filter(data2, Parameter == "DO")
plot_DO <- ggplot(data = data2)+
  geom_sf(data = ph_shp)+
  geom_sf(data = rivers)+
  geom_tile(data = DO, aes(x = Lon, y = Lat, fill = Value), height = .005, width = .005)+
  coord_sf(xlim = c(123.85, 124), ylim = c(10.33, 10.45), expand = FALSE)+
  #geom_point(aes(x=Lon, y=Lat, color= Species), size= 2)+
  theme_bw()+
  ggtitle("DO (mg/L)")

# EC
EC <- dplyr::filter(data2, Parameter == "EC")
plot_EC <- ggplot(data = data2)+
  geom_sf(data = ph_shp)+
  geom_sf(data = rivers)+
  geom_tile(data = EC, aes(x = Lon, y = Lat, fill = Value), height = .005, width = .005)+
  coord_sf(xlim = c(123.85, 124), ylim = c(10.33, 10.45), expand = FALSE)+
  #geom_point(aes(x=Lon, y=Lat, color= Species), size= 2)+
  theme_bw()+
  ggtitle("Electrical Conductivity (uS/cm)")

# TDS
TDS <- dplyr::filter(data2, Parameter == "TDS")
plot_TDS <-ggplot(data = data2)+
  geom_sf(data = ph_shp)+
```

```
  geom_sf(data = rivers)+
  geom_tile(data = TDS, aes(x = Lon, y = Lat, fill = Value), height = .005, width = .005)+
  coord_sf(xlim = c(123.85, 124), ylim = c(10.33, 10.45), expand = FALSE)+
  #geom_point(aes(x=Lon, y=Lat, color= Species), size= 2)+
  theme_bw()+
  ggtitle("Total Dissolved Solids (mg/L)")

# Sal
sal <- dplyr::filter(data2, Parameter == "SAL")
plot_sal <- ggplot(data = data2)+
  geom_sf(data = ph_shp)+
  geom_sf(data = rivers)+
  geom_tile(data = sal, aes(x = Lon, y = Lat, fill = Value), height = .005, width = .005)+
  coord_sf(xlim = c(123.85, 124), ylim = c(10.33, 10.45), expand = FALSE)+
  #geom_point(aes(x=Lon, y=Lat, color= Species), size= 2)+
  theme_bw()+
  ggtitle("Salinity (PSU)")

# BGA-PC
BGAPC <- dplyr::filter(data2, Parameter == "BGA-PC")
plot_BGAPC <- ggplot(data = data2)+
  geom_sf(data = ph_shp)+
  geom_sf(data = rivers)+
  geom_tile(data = BGAPC, aes(x = Lon, y = Lat, fill = Value), height = .005, width = .005)+
  coord_sf(xlim = c(123.85, 124), ylim = c(10.33, 10.45), expand = FALSE)+
  #geom_point(aes(x=Lon, y=Lat, color= Species), size= 2)+
  theme_bw()+
  ggtitle("Phycocyanin Blue-Green Algae Sensor (nm)")

# Altitude
alt <- dplyr::filter(data2, Parameter == "Altitude")
plot_alt <- ggplot(data = data2)+
  geom_sf(data = ph_shp)+
  geom_sf(data = rivers)+
  geom_tile(data = temp, aes(x = Lon, y = Lat, fill = Value), height = .005, width = .005)+
  coord_sf(xlim = c(123.85, 124), ylim = c(10.33, 10.45), expand = FALSE)+
  #geom_point(aes(x=Lon, y=Lat, color= Species), size= 2)+
  theme_bw()+
  ggtitle("Altitude (PSI)")

# Barometer
baro <- dplyr::filter(data2, Parameter == "Barometer")
plot_baro <-ggplot(data = data2)+
  geom_sf(data = ph_shp)+
  geom_sf(data = rivers)+
  geom_tile(data = baro, aes(x = Lon, y = Lat, fill = Value), height = .005, width = .005)+
  coord_sf(xlim = c(123.85, 124), ylim = c(10.33, 10.45), expand = FALSE)+
  #geom_point(aes(x=Lon, y=Lat, color= Species), size= 2)+
  theme_bw()+
  ggtitle("Barometer (PSI)")

# Depth
depth <- dplyr::filter(data2, Parameter == "Depth")
```

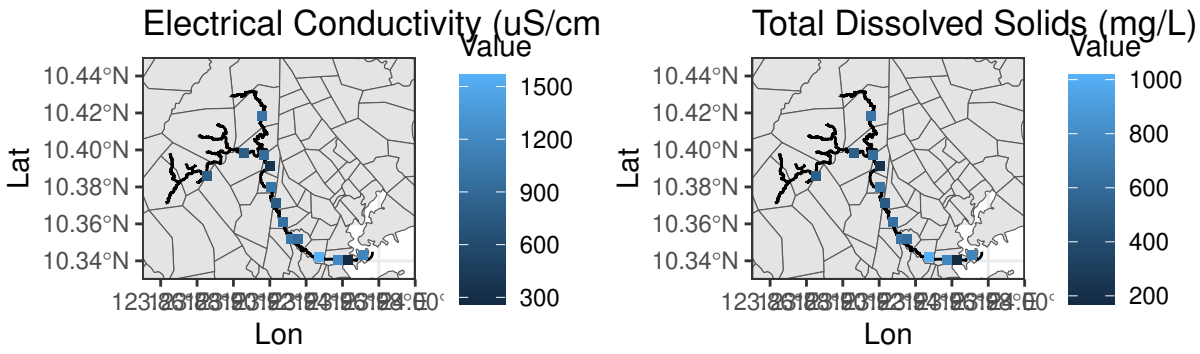
```
plot_depth <-ggplot(data = data2)+
  geom_sf(data = ph_shp)+
  geom_sf(data = rivers)+
  geom_tile(data = depth, aes(x = Lon, y = Lat, fill = Value), height = .005, width = .005)+
  coord_sf(xlim = c(123.85, 124), ylim = c(10.33, 10.45), expand = FALSE)+
  #geom_point(aes(x=Lon, y=Lat, color= Species), size= 2)+
  theme_bw()+
  ggtitle("Depth (m)")

combined_plot1 <- plot_temp + plot_OR
combined_plot2 <- plot_pH + plot_DO
combined_plot3 <- plot_EC + plot_TDS
combined_plot4 <- plot_sal + plot_BGAPC
combined_plot5 <- plot_alt + plot_baro + plot_depth

combined_plot1
```



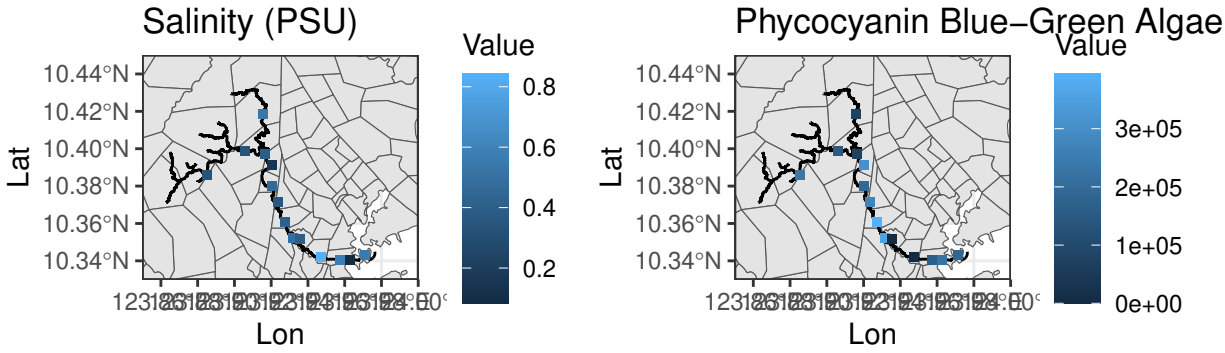

```
combined_plot2
```



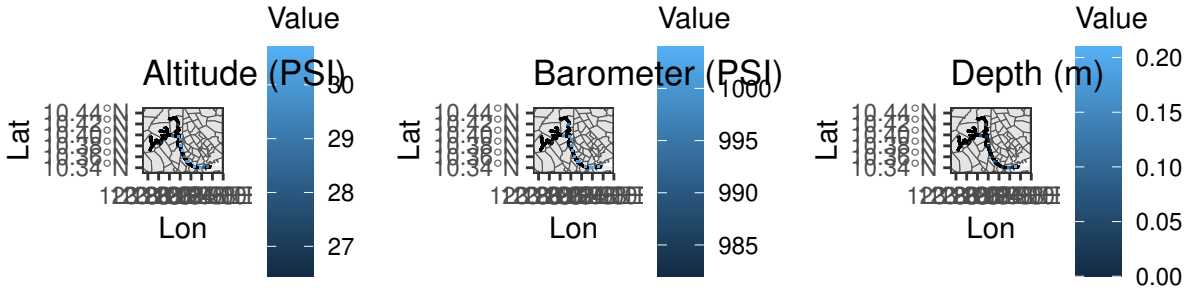

```
combined_plot3
```




```
combined_plot4
```




```
combined_plot5
```



The observed species occurrence data reveals interesting correlations with the visualized physicochemical parameters across the sampled sites. E. coli exhibits the broadest distribution, spanning from upstream to downstream, suggesting its potential in surviving through various environments and a broader environmental niche. Salmonella also demonstrates a widespread presence, albeit less extensive than E. coli. On the other hand, V. cholerae and Shigella show narrower distributions, hinting at specific ecological niches or preferences within the water system. The lower distribution of V. fluvialis, V. parahaemolyticus, V. vulnificus, and aeromonas across all sites may indicate a more localized or selective habitat for these species.

Examining the physicochemical parameters, variations in barometric pressure and altitude reveal distinct patterns. Barometer values increase downstream, suggesting potential atmospheric influences, while higher altitude in mid to downstream areas indicates topographical changes. Depth peaks at Site 14 in the bay, reflecting the specific hydrological conditions of this location. BGA-PC shows variability across sites, with the highest concentrations in the lower midstream areas at sites 8 and 9, indicating potential nutrient enrichment. Salinity also increases downstream, possibly influencing the distribution of certain species.

Temperature exhibits a gradient with higher values in mid to downstream regions, aligning with the broader distribution of E. coli and Salmonella. Oxidation-reduction potential (OR) is highest upstream and midstream, potentially influencing the occurrence of specific species adapted to these conditions. pH and dissolved oxygen levels are greatest upstream, suggesting a preference for more oxygenated and neutral environments by certain species. Electrical conductivity peaks downstream, possibly influencing the distribution of species adapted to

higher conductivity levels. Total dissolved solids (TDS) also show an increasing trend downstream, which may impact the habitat preferences of certain species.

```
# Correlation heatmap
merged_data <- merge(filtered_data_df, data2, by = "Site")

# Pivot the data2 data frame to have parameters as columns
pivoted_data2 <- data2 %>%
  pivot_wider(names_from = Parameter, values_from = Value)

# Merge the pivoted data2 with the presence-absence data
merged_data <- merge(filtered_data_df, pivoted_data2, by = "Site")

# Extract relevant columns for correlation analysis
cor_data <- merged_data[, c("Species", "Temperature", "Oxidation-Reaction Potential", "pH", "DOP", "DO"

# 1. Convert Species column to a factor
cor_data$Species <- as.factor(cor_data$Species)

# Calculate correlation matrix
cormat <- cor(cor_data[, -1], method = "pearson")

# Print the correlation matrix
print(cormat)
```

```
##                              Temperature Oxidation-Reaction Potential
## Temperature                   1.00000000                   0.05846874
## Oxidation-Reaction Potential  0.05846874                   1.00000000
## pH                           -0.15310117                   0.93073051
## DOP                          -0.21894111                   0.68384155
## DO                           -0.29629028                   0.66317999
## EC                            0.31302752                  -0.20949971
## TDS                           0.30960190                  -0.20560285
## SAL                           0.27980859                  -0.26137881
## BGA-PC                       -0.17828414                   0.05785166
## Altitude                     -0.44895621                   0.49286303
## Barometer                     0.54243073                  -0.44151538
## Depth                        -0.16825891                  -0.50082213
##                                       pH         DOP           DO           EC
## Temperature                  -0.15310117 -0.21894111 -0.296290276  0.313027518
## Oxidation-Reaction Potential  0.93073051  0.68384155  0.663179987 -0.209499714
## pH                            1.00000000  0.59507985  0.594771347 -0.317460756
## DOP                           0.59507985  1.00000000  0.996552214  0.012937940
## DO                            0.59477135  0.99655221  1.000000000 -0.007250255
## EC                           -0.31746076  0.01293794 -0.007250255  1.000000000
## TDS                          -0.31492542  0.01839693 -0.001733534  0.999912264
## SAL                          -0.35400540  0.01444820 -0.003636814  0.985145082
## BGA-PC                       -0.04778145  0.12394400  0.131460327 -0.437698597
## Altitude                      0.58528094  0.68649878  0.705980242  0.033491680
## Barometer                    -0.61069587 -0.50970604 -0.539476337  0.134378677
## Depth                        -0.60821433 -0.04589661 -0.025175372  0.417617017
##                                     TDS          SAL      BGA-PC    Altitude
## Temperature                  0.309601902  0.279808589 -0.17828414 -0.44895621
## Oxidation-Reaction Potential -0.205602850 -0.261378806  0.05785166  0.49286303
## pH                           -0.314925424 -0.354005404 -0.04778145  0.58528094
```

```
## DOP                            0.018396933  0.014448199  0.12394400  0.68649878
## DO                            -0.001733534 -0.003636814  0.13146033  0.70598024
## EC                             0.999912264  0.985145082 -0.43769860  0.03349168
## TDS                            1.000000000  0.985258908 -0.43895271  0.03495298
## SAL                            0.985258908  1.000000000 -0.42800518  0.03184924
## BGA-PC                        -0.438952709 -0.428005178  1.00000000  0.02209216
## Altitude                       0.034952983  0.031849236  0.02209216  1.00000000
## Barometer                      0.132884532  0.132989469  0.01218365 -0.95108398
## Depth                          0.416444656  0.403212221  0.10820796 -0.19391286
##                                 Barometer       Depth
## Temperature                    0.54243073 -0.16825891
## Oxidation-Reaction Potential -0.44151538 -0.50082213
## pH                           -0.61069587 -0.60821433
## DOP                          -0.50970604 -0.04589661
## DO                           -0.53947634 -0.02517537
## EC                            0.13437868  0.41761702
## TDS                           0.13288453  0.41644466
## SAL                           0.13298947  0.40321222
## BGA-PC                        0.01218365  0.10820796
## Altitude                     -0.95108398 -0.19391286
## Barometer                     1.00000000  0.36331754
## Depth                         0.36331754  1.00000000
```

```
# 2. create correlation heatmap
library(reshape2)
```

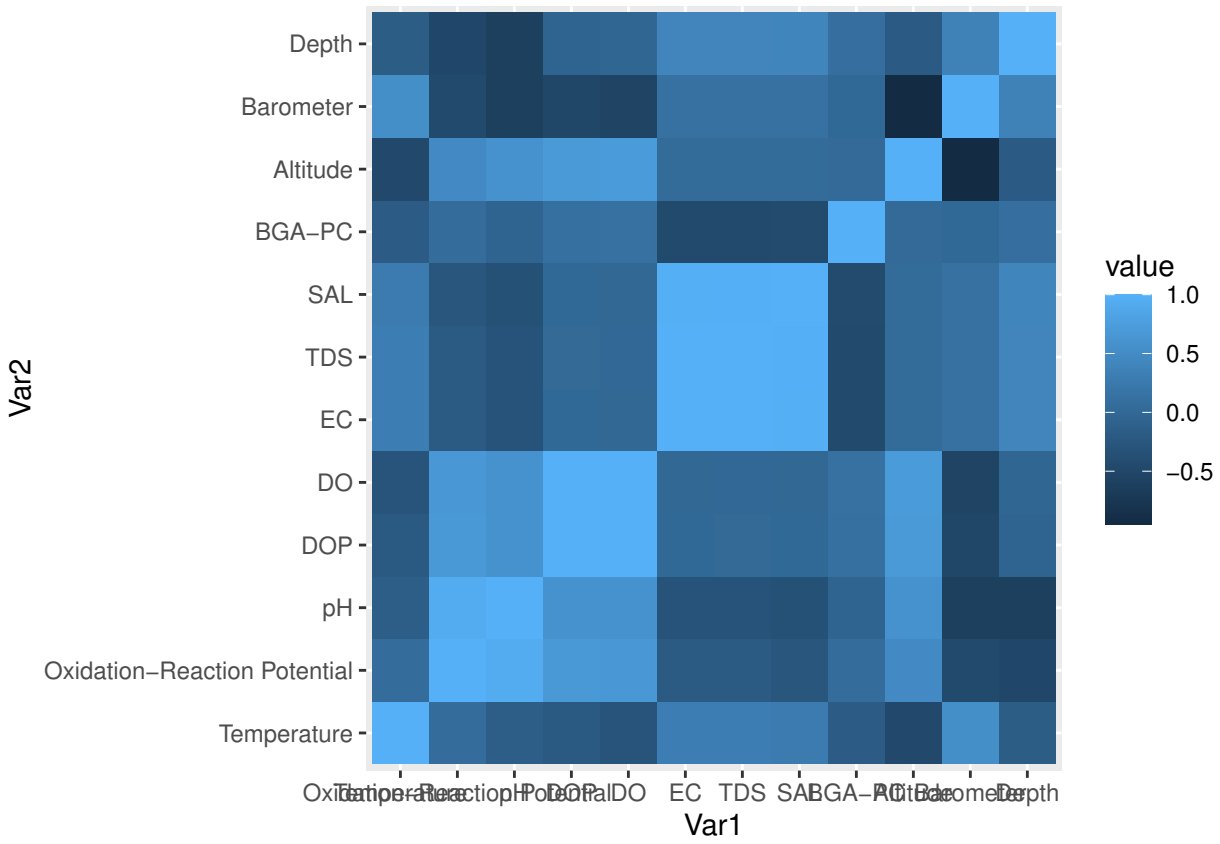
```
##
## Attaching package: 'reshape2'
```

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

```
melted_cormat <- melt(cormat)
head(melted_cormat)
```

```
##                           Var1        Var2       value
## 1                  Temperature Temperature  1.00000000
## 2 Oxidation-Reaction Potential Temperature  0.05846874
## 3                           pH Temperature -0.15310117
## 4                         DOP Temperature -0.21894111
## 5                          DO Temperature -0.29629028
## 6                          EC Temperature  0.31302752
```

```
library(ggplot2)
ggplot(data = melted_cormat, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile()
```



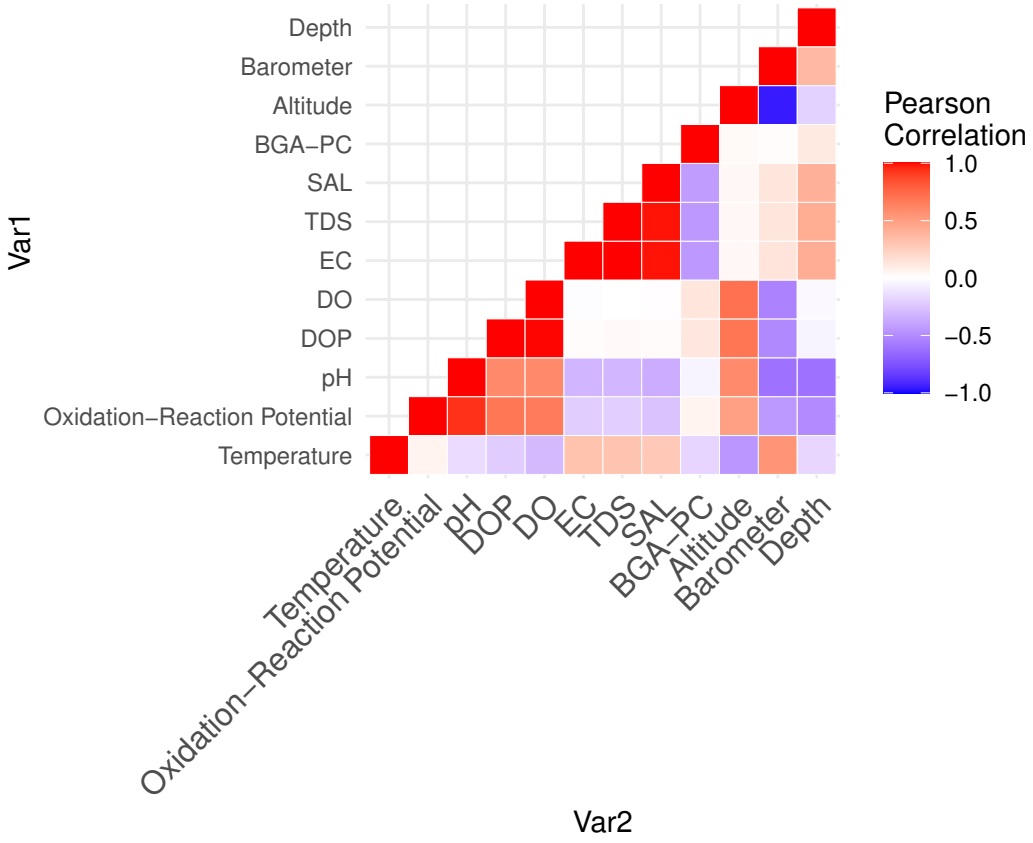

```
# 3. upper and lower triangles
# Get lower triangle of the correlation matrix
get_lower_tri <- function(cormat) {
  cormat[upper.tri(cormat)] <- NA
  return(cormat)
}

# Get upper triangle of the correlation matrix
get_upper_tri <- function(cormat) {
  cormat[lower.tri(cormat)] <- NA
  return(cormat)
}

upper_tri <- get_upper_tri(cormat)
upper_tri
```

```
##                                Temperature Oxidation-Reaction Potential
## Temperature                              1                    0.05846874
## Oxidation-Reaction Potential            NA                    1.00000000
## pH                                      NA                            NA
## DOP                                     NA                            NA
## DO                                      NA                            NA
## EC                                      NA                            NA
## TDS                                     NA                            NA
## SAL                                     NA                            NA
## BGA-PC                                  NA                            NA
## Altitude                                NA                            NA
```

```
## Barometer                              NA                          NA
## Depth                                  NA                          NA
##                                        pH        DOP         DO           EC
## Temperature                    -0.1531012 -0.2189411 -0.2962903  0.313027518
## Oxidation-Reaction Potential    0.9307305  0.6838416  0.6631800 -0.209499714
## pH                              1.0000000  0.5950798  0.5947713 -0.317460756
## DOP                                    NA  1.0000000  0.9965522  0.012937940
## DO                                     NA         NA  1.0000000 -0.007250255
## EC                                     NA         NA         NA  1.000000000
## TDS                                    NA         NA         NA           NA
## SAL                                    NA         NA         NA           NA
## BGA-PC                                 NA         NA         NA           NA
## Altitude                               NA         NA         NA           NA
## Barometer                              NA         NA         NA           NA
## Depth                                  NA         NA         NA           NA
##                                       TDS          SAL      BGA-PC    Altitude
## Temperature                    0.309601902  0.279808589 -0.17828414 -0.44895621
## Oxidation-Reaction Potential -0.205602850 -0.261378806  0.05785166  0.49286303
## pH                           -0.314925424 -0.354005404 -0.04778145  0.58528094
## DOP                           0.018396933  0.014448199  0.12394400  0.68649878
## DO                           -0.001733534 -0.003636814  0.13146033  0.70598024
## EC                            0.999912264  0.985145082 -0.43769860  0.03349168
## TDS                           1.000000000  0.985258908 -0.43895271  0.03495298
## SAL                                    NA  1.000000000 -0.42800518  0.03184924
## BGA-PC                                 NA           NA  1.00000000  0.02209216
## Altitude                               NA           NA          NA  1.00000000
## Barometer                              NA           NA          NA          NA
## Depth                                  NA           NA          NA          NA
##                                Barometer       Depth
## Temperature                    0.54243073 -0.16825891
## Oxidation-Reaction Potential -0.44151538 -0.50082213
## pH                           -0.61069587 -0.60821433
## DOP                          -0.50970604 -0.04589661
## DO                           -0.53947634 -0.02517537
## EC                            0.13437868  0.41761702
## TDS                           0.13288453  0.41644466
## SAL                           0.13298947  0.40321222
## BGA-PC                        0.01218365  0.10820796
## Altitude                     -0.95108398 -0.19391286
## Barometer                     1.00000000  0.36331754
## Depth                                 NA  1.00000000
```

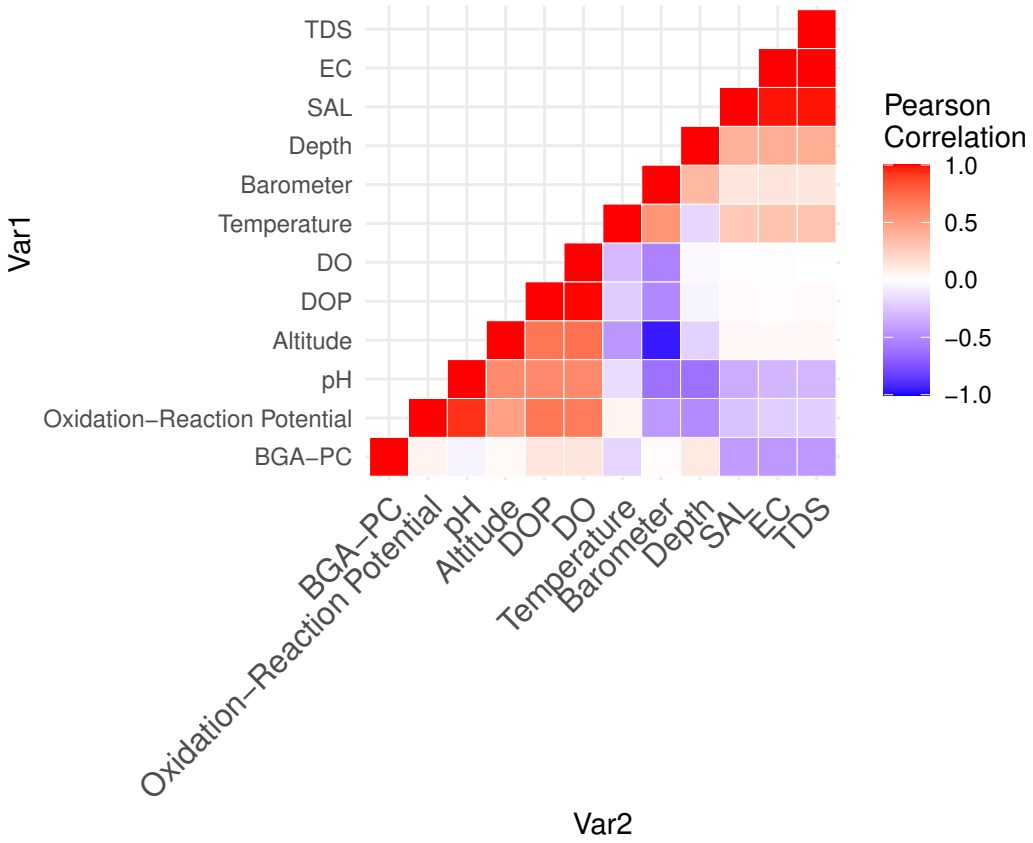
```
# 4. Finished correlation matrix heatmaps
# Melt the correlation matrix
melted_cormat <- melt(upper_tri, na.rm = TRUE)

# Heatmap
ggplot(data = melted_cormat, aes(Var2, Var1, fill = value))+
  geom_tile(color = "white")+
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limit = c(-1,1), space = "Lab",
                       name="Pearson\nCorrelation") +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 45, vjust = 1,
```

```
                                size = 12, hjust = 1))+
  coord_fixed()
```

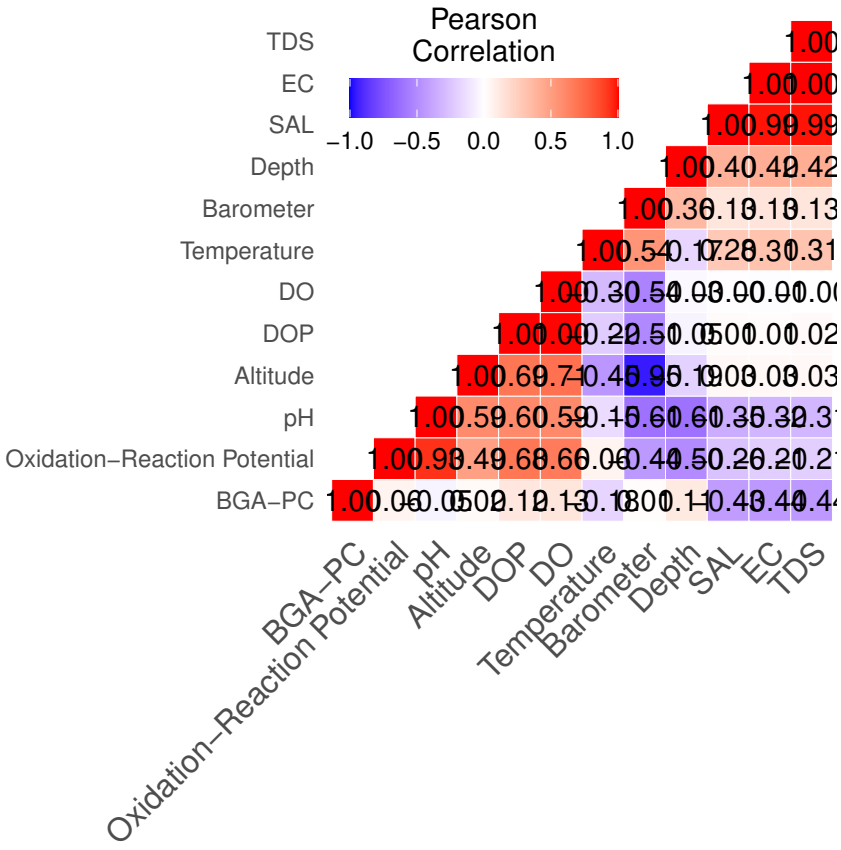



```
# 5. Reorder correlation matrix
reorder_cormat <- function(cormat){
# Use correlation between variables as distance
dd <- as.dist((1-cormat)/2)
hc <- hclust(dd)
cormat <-cormat[hc$order, hc$order]
}

# Reorder the correlation matrix
cormat <- reorder_cormat(cormat)
upper_tri <- get_upper_tri(cormat)
# Melt the correlation matrix
melted_cormat <- melt(upper_tri, na.rm = TRUE)
# Create a ggheatmap
ggheatmap <- ggplot(melted_cormat, aes(Var2, Var1, fill = value))+
  geom_tile(color = "white")+
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limit = c(-1,1), space = "Lab",
                       name="Pearson\nCorrelation") +
  theme_minimal()+ # minimal theme
  theme(axis.text.x = element_text(angle = 45, vjust = 1,
                                   size = 12, hjust = 1))+
  coord_fixed()
# Print the heatmap
```

```r
print(ggheatmap)
```




```r
# Add correlation coefficients
ggheatmap +
  geom_text(aes(Var2, Var1, label = sprintf("%.2f", value)), color = "black", size = 4) +
  theme(
    axis.title.x = element_blank(),
    axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    panel.border = element_blank(),
    panel.background = element_blank(),
    axis.ticks = element_blank(),
    legend.justification = c(1, 0),
    legend.position = c(0.6, 0.7),
    legend.direction = "horizontal")+
  guides(fill = guide_colorbar(barwidth = 7, barheight = 1,
                               title.position = "top", title.hjust = 0.5))
```



The correlation heatmap shows the strengths of the relationships between the physicochemical variables. A pearson correlation value close to 1 or -1 indicates a strong correlation while the positive and negative signs indicates its direction. Darker gradients represent a greater strength of correlation compared to light colored tiles. Electrical conductivity (EC), salinity, and total dissolved solids (TDS) are have the strongest correlation, followed by pH, BGA-PC, altitude, and dissolved oxygen (DO% and DO mg/L). These are the variables that greatly influenced the presence of the species across the sampling sites from upstream to downstream. The other variable relationship combinations contain a pearson correlation value that is close to 0 and negative numbers and had cooler colors which indicates a weak and negative correlation respectively.

In conclusion, this project underscores the potential factors influencing gastroenteritis-causing agents. Nevertheless, the precision of predictions is constrained by the robustness and completeness of data, notably the absence of species abundance data, which could yield more insightful correlations with physicochemical variables. Despite this limitation, the study successfully identified variables contributing to the prevalence of gastroenteritis-causing agents and identified disease hotspots.

References:

Lin, L., Yang, H., Xu, X., 2022. Effects of Water Pollution on Human Health and Disease Heterogeneity: A Review. Front. Environ. Sci. 10.

Parilla, R.B., Cañedo, L.F., Amores, K.D.L., Jr, G., Lawas, R.W.S., 2019. The DisappearingFish Community in Butuanon River, Cebu, Philippines: The Ignored Impact of Pollution.

Peterson, A.T., 2007. Ecological niche modelling and understanding the geography of disease transmission. Vet Ital 43.

Gadgil, A., 1998. Drinking Water in Developing Countries. Annu. Rev. Energy Environ. 23. https://doi.org/10.1146/annurev.energy.23.1.253

Peterson, A.T., Soberón, J., Pearson, R.G., Anderson, R.P., Martínez-Meyer, E., Nakamura, M., Araújo,

M.B., 2011. Ecological Niches and Geographic Distributions (MPB-49), in: Ecological Niches and Geographic Distributions (MPB-49). Princeton University Press. https://doi.org/10.1515/9781400840670

Martínez-Minaya, J., Cameletti, M., Conesa, D., Pennino, M.G., 2018. Species distribution modeling: a statistical review with focus in spatio-temporal issues. Stoch. Environ. Res. Risk Assess. 32, 3227–3244. https://doi.org/10.1007/s00477-018-1548-7

Escobar, L.E., Craft, M.E., 2016. Advances and Limitations of Disease Biogeography Using Ecological Niche Modeling. Front. Microbiol. 7.

Graham, C.H., Ferrier, S., Huettman, F., Moritz, C., Peterson, A.T., 2004. New developments in museum-based informatics and applications in biodiversity analysis. Trends Ecol. Evol. 19, 497–503. https://doi.org/10.1016/j.tree.2004.07.006

Guisan, A., Thuiller, W., 2005. Predicting species distribution: Offering more than simple habitat models. Ecol. Lett. 8, 993–1009. https://doi.org/10.1111/j.1461-0248.2005.00792.x Sebes, R., 2023. Predicting Biogeographical Responses of Marine Coccolithophores to Climate Change Using Correlative Ecological Niche Modeling.

Hay, S.I., Battle, K.E., Pigott, D.M., Smith, D.L., Moyes, C.L., Bhatt, S., Brownstein, J.S., Collier, N., Myers, M.F., George, D.B., Gething, P.W., 2013. Global mapping of infectious disease. Philos. Trans. R. Soc. B Biol. Sci. 368, 20120250. https://doi.org/10.1098/rstb.2012.0250

Purse, B.V., Golding, N., 2015. Tracking the distribution and impacts of diseases with biological records and distribution modelling. Biol. J. Linn. Soc. 115, 664–677. https://doi.org/10.1111/bij.12567

Phillips, S.J., Anderson, R.P., Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. Ecol. Model. 190, 231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026