# ST117 FINAL Written Report - Phase 2 Task B

Report Pod 041

Today's date in the format 2025-05-01

## Contributions

This submission was created by:

1. WARWICK ID 5600761 Alex Bannister:
2. WARWICK ID 5604173 Pratham Bhargava: Phase 1: q5, q7, q8, q9
3. WARWICK ID 5627113 Yanbo Dong: Phase 2: task C
4. WARWICK ID 5645242 Daniel Guo: Phase 2: task B, Phase 1: q6
5. WARWICK ID 5650102 Jules Reinaud: Phase 2: task A

## Setting up data frames before for phase 2 part B

```
#Import the dataframes from phase 1
df_stream_wide <- readRDS("/Users/danielguo/Desktop/University/Year 1/ST117/5645242_041_WR/df_stream_wide.rds")
df_precipitation_wide <- readRDS("/Users/danielguo/Desktop/University/Year 1/ST117/5645242_041_WR/df_precipitati
df_soil_wide <- readRDS("/Users/danielguo/Desktop/University/Year 1/ST117/5645242_041_WR/df_soil_wide.rds")

#Averaging the repeated sampling by the variable LCODE
df_stream_avg <- df_stream_wide %>%
  group_by(SDATE, SITECODE, LCODE) %>%
  summarise(across(where(is.numeric), \(x) mean(x, na.rm = TRUE)), .groups = "drop")

#weighted averaging considering the sampling volume (VOLUME) by the variable RID
df_soil_avg <- df_soil_wide %>%
  group_by(SDATE, SITECODE, RID) %>%
  summarise(across(where(is.numeric), \(x) weighted.mean(x, w = VOLUME, na.rm = TRUE)), .groups = "drop")

#STAGE, VACUUM, and VOLUME are only contained in some of the datasets and may be dropped unless needed.
df_stream_B <- df_stream_avg %>%
  select(-STAGE)
df_precipitation_B <- df_precipitation_wide %>%
  select(-VOLUME)
df_soil_B <- df_soil_avg %>%
  select(-VACUUM, -VOLUME)
```

## Question 1 and 2

```
# Function to convert the dataframes into averaged data by weeks for each sites
convert_to_weekly <- function(df, date_col = "SDATE", site_col = "SITECODE") {
  df <- df %>%
    mutate(
      DATE = as.Date(.data[[date_col]]),  #Convert date column to Date type
      YEAR = format(DATE, "%Y"), # Extract year
      WEEK = format(DATE, "%W") # Extract week
    ) %>%
    group_by(.data[[site_col]], YEAR, WEEK) %>% # Group by different site, year, and week
    summarise(across(where(is.numeric), ~ mean(.x, na.rm = TRUE)), .groups = "drop")  #Average numeric columns b

  return(df)
}

# function to merge three datasets (stream, precipitation, soil)
merge_datasets <- function(stream, precipitation, soil) {
  # Rename columns to Varieble_dataset abreviations
  stream <- stream %>% rename_with(~ paste0(., "_str"), -c(SITECODE, YEAR, WEEK))
  precipitation <- precipitation %>% rename_with(~ paste0(., "_prec"), -c(SITECODE, YEAR, WEEK))
  soil <- soil %>% rename_with(~ paste0(., "_soil"), -c(SITECODE, YEAR, WEEK))
# merge and full_join all the datasets
  merged <- full_join(stream, precipitation, by = c("SITECODE", "YEAR", "WEEK")) %>%
    full_join(soil, by = c("SITECODE", "YEAR", "WEEK"))

  return(merged)
}

# Function to remove outliers from the applicable numeric columns using interquartile range formula
remove_outliers <- function(df) {
  df %>% mutate(across(where(is.numeric), ~ {
    Q1 <- quantile(.x, 0.25, na.rm = TRUE)
    Q3 <- quantile(.x, 0.75, na.rm = TRUE)
    IQR <- Q3 - Q1
    low <- Q1 - 1.5 * IQR
    high <- Q3 + 1.5 * IQR
    ifelse(.x < low | .x > high, NA, .x)
  }))
}

# function to compute correlation matrix of one chemical between different datasets for given site and year rang
variable_correlation <- function(merged_df, variable, site, year_range, method = "pearson") { #setting pearson a
  df <- merged_df %>%
    filter(SITECODE == site, as.numeric(YEAR) %in% year_range) %>%
    select(matches(paste0("^", variable, "_"))) %>%
    na.omit()

  if (nrow(df) < 2) {
    stop("Not enough data for correlation.")  #hault if insufficient data
  }

  if (sum(complete.cases(df)) / nrow(df) < 0.5) {
    warning("More than 50% of data is missing for this site/variable") # Warning if there exist many NAs
  }

  # Use cat instead of message to prevent double output
  cat("Correlation matrix for", variable,
      "from stream, precipitation, and soil solution (Site:", site,
      "; Years:", paste(range(year_range), collapse = "-"), ")\n")

  round(cor(df, method = method), 2) #rounded correlation matrix to 2 decimal places
}

# function to summarise the correlation data in a table formate
summary_correlation_across_sites <- function(merged_df, variable, sites, year_range, method = "pearson") {
  result <- lapply(sites, function(site) {
    tryCatch({
      corr <- variable_correlation(merged_df, variable, site, year_range, method) # computing correlation
      data.frame(Site = site,
                 Str_vs_Prec = corr[1, 2],
                 Str_vs_Soil = corr[1, 3],
                 Prec_vs_Soil = corr[2, 3],
                 stringsAsFactors = FALSE)
    }, error = function(e) { #if fail, return NA
      data.frame(Site = site,
                 Str_vs_Prec = NA,
                 Str_vs_Soil = NA,
                 Prec_vs_Soil = NA,
                 stringsAsFactors = FALSE)
    })
  })

  do.call(rbind, result) #combining all the results into one dataframe
}

#a function to plot the correlation summary as a heatmap for visualisatio.
plot_correlation_summary <- function(summary_df, variable) {
  summary_df_long <- summary_df %>%
    pivot_longer(cols = -Site, names_to = "Comparison", values_to = "Correlation") #reshape the dataframe for ea

  ggplot(summary_df_long, aes(x = Comparison, y = Site, fill = Correlation)) +
    geom_tile() + # coloured tile heatmap
    geom_text(aes(label = round(Correlation, 2)), size = 3) + # add correlation coefficient over the heatmap for
    scale_fill_gradient2(low = "darkblue", high = "maroon",  #colour coding
                         midpoint = 0, limit = c(-1, 1)) +
    theme_minimal() +
    labs(
      title = paste("Correlation Summary for", variable, "Across Sites"),
      x = "Comparison", y = "Site"
    ) +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1)) # rotate x-axis labels for better readability
}
```

# Question 3

### 1. convert_to_weekly

**Purpose:** Converts any water chemistry dataset into a weekly data by averaging repeated measurements within the week.

**Arguments:**

- df (data.frame): Input dataframes (e.g., df_stream_B, df_precipitation_B, df_soil_B).
- date_col (character): Column name for date variable (default: "SDATE": no need to input)
- site_col (character): Column name for site code variable (default: "SITECODE": no need to input)

**Returns:** A data frame sorted by site, year, and week, with averaged numeric variables for each weeks.

```
# Example: (set the data into another dataframe for easier use in later functions)
df_stream_weekly <- convert_to_weekly(df_stream_B)
df_precipitation_weekly<- convert_to_weekly(df_precipitation_B)
df_soil_weekly<- convert_to_weekly(df_soil_B)
head(df_stream_weekly)
```

```
## # A tibble: 6 x 23
##   SITECODE YEAR  WEEK  LCODE ALUMINIUM TOTALN CHLORIDE   DOC    IRON MAGNESIUM
##   <chr>    <chr> <chr> <dbl>     <dbl>  <dbl>    <dbl> <dbl>    <dbl>     <dbl>
## 1 T02      1993  19        1     0.053    NaN     6.89   6.8   0.14       1.38
## 2 T02      1993  20        1   NaN        NaN   NaN    NaN   NaN        NaN
## 3 T02      1993  21        1   NaN        NaN   NaN    NaN   NaN        NaN
## 4 T02      1993  22        1     0.282    NaN     5.84  13.3   0.369      1.02
## 5 T02      1993  23         1   NaN        NaN   NaN    NaN   NaN        NaN
## 6 T02      1993  24         1   NaN        NaN   NaN    NaN   NaN        NaN
## # i 13 more variables: NH4N <dbl>, NO3N <dbl>, PH <dbl>, PO4P <dbl>,
## #   POTASSIUM <dbl>, SO4S <dbl>, SODIUM <dbl>, CALCIUM <dbl>, PHAQCS <dbl>,
## #   PHAQCU <dbl>, ALKY <dbl>, CONDY <dbl>, TOTALP <dbl>
```

```r
head(df_precipitation_weekly)
```

```
## # A tibble: 6 x 22
##   SITECODE YEAR  WEEK  ALUMINIUM SODIUM  SO4S POTASSIUM  PO4P    PH  NO3N  NH4N
##   <chr>    <chr> <chr>     <dbl>  <dbl> <dbl>     <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 T01      1993  50          NaN      0   0           0   NaN   5.3   0     1.7
## 2 T01      1993  01          NaN      0   0           0   NaN   5.2   0.5   1.9
## 3 T01      1993  02            0      0   0           0     0   5.3   0.3   0.4
## 4 T01      1993  03            0      0  15.5         0     0   5     0     0.5
## 5 T01      1993  04            0      0  41.3         0     0   5.1   0.7   0.6
## 6 T01      1993  05          NaN    NaN NaN         NaN   NaN   4.6   0.7   1
## # i 11 more variables: MAGNESIUM <dbl>, IRON <dbl>, DOC <dbl>, CHLORIDE <dbl>,
## #   CALCIUM <dbl>, TOTALN <dbl>, PHAQCS <dbl>, PHAQCU <dbl>, CONDY <dbl>,
## #   ALKY <dbl>, TOTALP <dbl>
```

```r
head(df_soil_weekly)
```

```
## # A tibble: 6 x 22
##   SITECODE YEAR  WEEK   SO4S SODIUM TOTALN ALUMINIUM CALCIUM CHLORIDE   DOC
##   <chr>    <chr> <chr> <dbl>  <dbl>  <dbl>     <dbl>   <dbl>    <dbl> <dbl>
## 1 T01      1993  15      NaN    NaN    NaN       NaN     NaN      NaN   NaN
## 2 T01      1993  43      NaN    NaN    NaN       NaN     NaN      NaN   NaN
## 3 T01      1993  45      NaN    NaN    NaN       NaN     NaN      NaN   NaN
## 4 T01      1993  47      NaN    NaN    NaN       NaN     NaN      NaN   NaN
## 5 T01      1993  49      NaN    NaN    NaN       NaN     NaN      NaN   NaN
## 6 T01      1993  51      NaN    NaN    NaN       NaN     NaN      NaN   NaN
## # i 12 more variables: IRON <dbl>, MAGNESIUM <dbl>, NH4N <dbl>, NO3N <dbl>,
## #   PH <dbl>, PO4P <dbl>, POTASSIUM <dbl>, PHAQCS <dbl>, PHAQCU <dbl>,
## #   ALKY <dbl>, CONDY <dbl>, TOTALP <dbl>
```

### 2. merge_datasets

**Purpose:** Combines stream water, precipitation, and soil solution dataframes into a large merged dataframe.

**Arguments:**

- stream (data.frame): Weekly stream dataset. (df_stream_weekly)
- precipitation (data.frame): Weekly precipitation dataset.(df_precipitation_weekly)
- soil(data.frame): Weeklysoil dataset.(df_soil_weekly)

(df_stream_weekly, df_precipitation_weekly, df_soil_weeklyfrom last function's example)

**Returns:** A data.frame merged by SITECODE, YEAR, and WEEK, with columns named like ALUMINIUM_str, ALUMINIUM_prec, ALUMINIUM_soil

```r
# Example: (set the data into another dataframe for easier use in later functions)
df_merged <- merge_datasets(df_stream_weekly, df_precipitation_weekly, df_soil_weekly)
head(df_merged)
```

```
## # A tibble: 6 x 61
##   SITECODE YEAR  WEEK  LCODE_str ALUMINIUM_str TOTALN_str CHLORIDE_str DOC_str
##   <chr>    <chr> <chr>     <dbl>         <dbl>      <dbl>        <dbl>   <dbl>
## 1 T02      1993  19            1         0.053        NaN         6.89     6.8
## 2 T02      1993  20            1       NaN            NaN       NaN      NaN
## 3 T02      1993  21            1       NaN            NaN       NaN      NaN
## 4 T02      1993  22            1         0.282        NaN         5.84    13.3
## 5 T02      1993  23            1       NaN            NaN       NaN      NaN
## 6 T02      1993  24            1       NaN            NaN       NaN      NaN
## # i 53 more variables: IRON_str <dbl>, MAGNESIUM_str <dbl>, NH4N_str <dbl>,
## #   NO3N_str <dbl>, PH_str <dbl>, PO4P_str <dbl>, POTASSIUM_str <dbl>,
## #   SO4S_str <dbl>, SODIUM_str <dbl>, CALCIUM_str <dbl>, PHAQCS_str <dbl>,
## #   PHAQCU_str <dbl>, ALKY_str <dbl>, CONDY_str <dbl>, TOTALP_str <dbl>,
## #   ALUMINIUM_prec <dbl>, SODIUM_prec <dbl>, SO4S_prec <dbl>,
## #   POTASSIUM_prec <dbl>, PO4P_prec <dbl>, PH_prec <dbl>, NO3N_prec <dbl>,
## #   NH4N_prec <dbl>, MAGNESIUM_prec <dbl>, IRON_prec <dbl>, DOC_prec <dbl>, ...
```

### 3. remove_outliers

**Purpose:** Removes outliers from the dataframe using the interquartile range method. Values outside $[Q1 - 1.5 \times \text{IQR},\ Q3 + 1.5 \times \text{IQR}]$ are replaced with NA.

**Arguments:**

- df_merged (data.frame): Merged dataset. (df_merged from last function, merge_datasets(), example)

**Returns:** A data.frame with outlier values replaced by NA.

```r
# Example: (set the data into another dataframe for easier use in later functions)
df_merged_no_outliers <-remove_outliers(df_merged)
head(df_merged_no_outliers)
```

```
## # A tibble: 6 x 61
##   SITECODE YEAR  WEEK  LCODE_str ALUMINIUM_str TOTALN_str CHLORIDE_str DOC_str
##   <chr>    <chr> <chr>     <dbl>         <dbl>      <dbl>        <dbl>   <dbl>
## 1 T02      1993  19            1         0.053         NA          6.89     6.8
## 2 T02      1993  20            1        NA              NA         NA        NA
## 3 T02      1993  21            1        NA              NA         NA        NA
## 4 T02      1993  22            1        NA              NA          5.84    13.3
## 5 T02      1993  23            1        NA              NA         NA        NA
## 6 T02      1993  24            1        NA              NA         NA        NA
## # i 53 more variables: IRON_str <dbl>, MAGNESIUM_str <dbl>, NH4N_str <dbl>,
## #   NO3N_str <dbl>, PH_str <dbl>, PO4P_str <dbl>, POTASSIUM_str <dbl>,
## #   SO4S_str <dbl>, SODIUM_str <dbl>, CALCIUM_str <dbl>, PHAQCS_str <dbl>,
## #   PHAQCU_str <dbl>, ALKY_str <dbl>, CONDY_str <dbl>, TOTALP_str <dbl>,
## #   ALUMINIUM_prec <dbl>, SODIUM_prec <dbl>, SO4S_prec <dbl>,
## #   POTASSIUM_prec <dbl>, PO4P_prec <dbl>, PH_prec <dbl>, NO3N_prec <dbl>,
## #   NH4N_prec <dbl>, MAGNESIUM_prec <dbl>, IRON_prec <dbl>, DOC_prec <dbl>, ...
```

### 4. variable_correlation

**Purpose:** Generate a correlation matrix for a given variable across the three water types for a given site and year range.

**Arguments:**

- df_merged_no_outliers (data.frame): Cleaned merged dataset.
- variable (character): Name of the variable ("ALUMINIUM").
- site (character): Site code ("T02").
- year_range (numeric vector): Range of years to include (1992:1993). (data from 1992 to 2005)
- method (character): "pearson" (default) or "spearman". (no need to input)

**Returns:** A 3×3 correlation matrix between the variable's values in _str, _prec, and _soil.

```r
#For example: Aluminuim at site T02 for 1992-1993 (do not set to dataframe)
variable_correlation(df_merged_no_outliers, "ALUMINIUM", "T02", 1992:1993)
```

```
## Correlation matrix for ALUMINIUM from stream, precipitation, and soil solution (Site: T02 ; Years: 1992-1993
##                ALUMINIUM_str ALUMINIUM_prec ALUMINIUM_soil
## ALUMINIUM_str           1.00          -0.59          -0.75
## ALUMINIUM_prec         -0.59           1.00          -0.07
## ALUMINIUM_soil         -0.75          -0.07           1.00
```

### 5. summary__correlation__across__sites

**Purpose:** Generates a summary table of pairwise correlations for a given variable across different data, for a given site and given year range.

**Arguments:**

- df__merged__no__outliers (data.frame): Cleaned merged dataset.
- variable (character): Name of the variable ("ALUMINIUM").
- site (character): Site code ("T02").
- year__range (numeric vector): Range of years to include (1992:1993). (data from 1992 to 2005)
- method (character): "pearson" (default) or "spearman". (don't need to input)

**Returns:** A data frame with correlations between water types.

```
#For example: Aluminuim at site T02 for 1992-1993 (set the data into another dataframe for easier use in the las
summary_alum_T02_9293 <- summary_correlation_across_sites(df_merged_no_outliers, "ALUMINIUM", "T02", 1992:1993)
```

```
## Correlation matrix for ALUMINIUM from stream, precipitation, and soil solution (Site: T02 ; Years: 1992-1993
```

```
summary_alum_T02_9293
```

```
##   Site Str_vs_Prec Str_vs_Soil Prec_vs_Soil
## 1  T02       -0.59       -0.75        -0.07
```

### 6. plot__correlation__summary

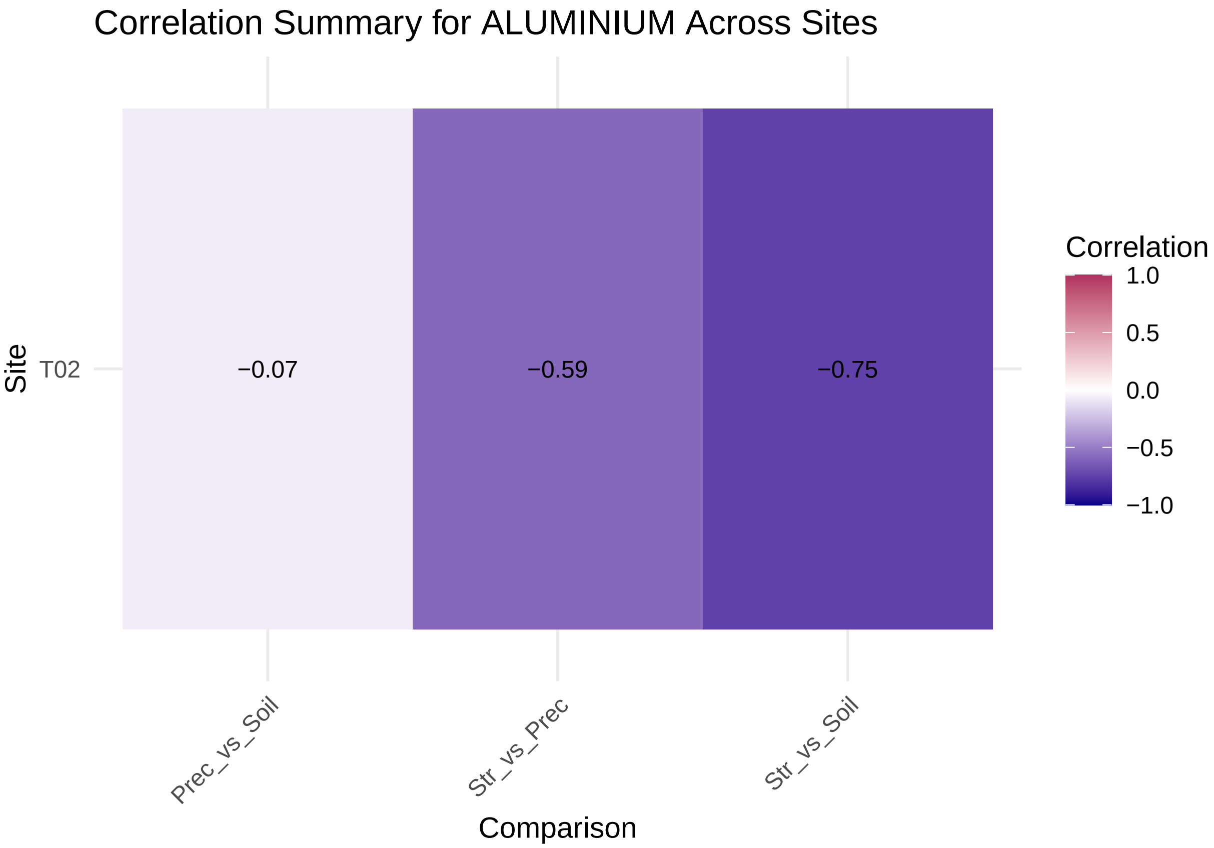
**Purpose:** Generates a heatmap that provides visualisation for a given variable for a given year range.

**Arguments:**

- summary__df (summary__alum__T02__9293 from last function's example) (data.frame): Summary with site-wise correlations of an varianle between water types.
- variable (character): (Aluminium for example): Name of the variable (e.g., "ALUMINIUM").

**Returns:** A data frame with site-wise correlations between water types for a given variablea nd year range.

```
#For example: Aluminuim at site T02 for 1992-1993
plot_correlation_summary(summary_alum_T02_9293, "ALUMINIUM")
```



# Question 4

## Hypotheses 1:

```
sites_1 <- c("T02", "T04", "T05", "T06", "T07", "T11") # sites we want to examine
summary_correlation_across_sites(df_merged_no_outliers, variable = "POTASSIUM", sites = sites_1, 2010:2015, meth
```

```
## Correlation matrix for POTASSIUM from stream, precipitation, and soil solution (Site: T02 ; Years: 2010-2015
## Correlation matrix for POTASSIUM from stream, precipitation, and soil solution (Site: T04 ; Years: 2010-2015
## Correlation matrix for POTASSIUM from stream, precipitation, and soil solution (Site: T05 ; Years: 2010-2015
## Correlation matrix for POTASSIUM from stream, precipitation, and soil solution (Site: T06 ; Years: 2010-2015
## Correlation matrix for POTASSIUM from stream, precipitation, and soil solution (Site: T07 ; Years: 2010-2015
## Correlation matrix for POTASSIUM from stream, precipitation, and soil solution (Site: T11 ; Years: 2010-2015
```

```
##   Site Str_vs_Prec Str_vs_Soil Prec_vs_Soil
## 1  T02        0.56        0.32         0.35
## 2  T04        0.35        0.22         0.26
## 3  T05        0.16        0.34         0.16
## 4  T06        0.38        0.38        -0.17
## 5  T07        0.16        0.18         0.27
## 6  T11        0.29        0.32         0.01
```

**The abundance of Potassium (essential for fertile land) in soil solution is weakly positively correlated with stream water. This is supported by Str__vs__Soil Pearson correlations consistently falling within [0.15, 0.5] between 2010–2015 across sites T02, T07, T11, and T12, as shown in the summary table.**

## Hypotheses 2:

```
sites_2 <- c("T02", "T04", "T05", "T07", "T12")
summary_correlation_across_sites(df_merged_no_outliers, variable = "NO3N", sites = sites_2, 1992:2015, method =
```

```
## Correlation matrix for NO3N from stream, precipitation, and soil solution (Site: T02 ; Years: 1992-2015 )
## Correlation matrix for NO3N from stream, precipitation, and soil solution (Site: T04 ; Years: 1992-2015 )
## Correlation matrix for NO3N from stream, precipitation, and soil solution (Site: T05 ; Years: 1992-2015 )
## Correlation matrix for NO3N from stream, precipitation, and soil solution (Site: T07 ; Years: 1992-2015 )
## Correlation matrix for NO3N from stream, precipitation, and soil solution (Site: T12 ; Years: 1992-2015 )
```

```
##   Site Str_vs_Prec Str_vs_Soil Prec_vs_Soil
## 1  T02       -0.04        0.20        -0.01
## 2  T04        0.25        0.31         0.13
## 3  T05        0.05        0.22         0.08
## 4  T07       -0.07        0.34         0.06
## 5  T12        0.26        0.61         0.16
```

**TThere is no relationship of the abundance of Nitrate Oxygen (abundant in acid rain) in stream water and precipitation. The absolute values of the Str_vs_Prec Spearman correlations exceed 0.2 only twice during the full study period, as shown in the summary table across selected sites.**