Sagehen Groundwater Data Processing

Jennifer Natali, with help from chatgpt for learning R, dplyr, stats

2024 October 24

Load and Summarize Data

1. Upload data (.csv file)
2. Examine the following properties:

- length and frequency of the time series
- completeness of each time series
- descriptive statistics
  - basics: mean, CV, ACF for each variable
  - normality: histogram, qqplot, skewness, kurtosis

```
# Load libraries
library(dplyr)
library(astsa)
library(lubridate)
library(moments) # for skewness and kurtosis testing

# Setup directories and filepaths
home_dir='/Volumes/SANDISK_SSD_G40/GoogleDrive/GitHub/'
repository_dir = paste(home_dir,'sagehen_meadows/', sep='')
groundwater_filepath = paste(repository_dir,'data/field_observations/groundwater/biweekly_manual/groundw
observation_filepath = paste(repository_dir,'data/field_observations/groundwater/biweekly_manual/groundw

# Load groundwater data
groundwater <- read.csv(groundwater_filepath)

# Manage dates and times
groundwater$timestamp <- ymd_hms(groundwater$timestamp)

# Check timestamp formatting
str(groundwater$timestamp)
```

```
##  POSIXct[1:1116], format: "2018-06-01 07:45:00" "2018-06-18 08:32:00" "2018-06-30 08:55:00" ...
```

```
# Create columns for date and isoweek (starts on Monday)
groundwater <- groundwater %>% mutate(
  date = as.Date(timestamp),
  year = year(timestamp),
  isoweek = isoweek(date),
  day_of_year = yday(date))

# summarize the full times series
summary(groundwater)
```

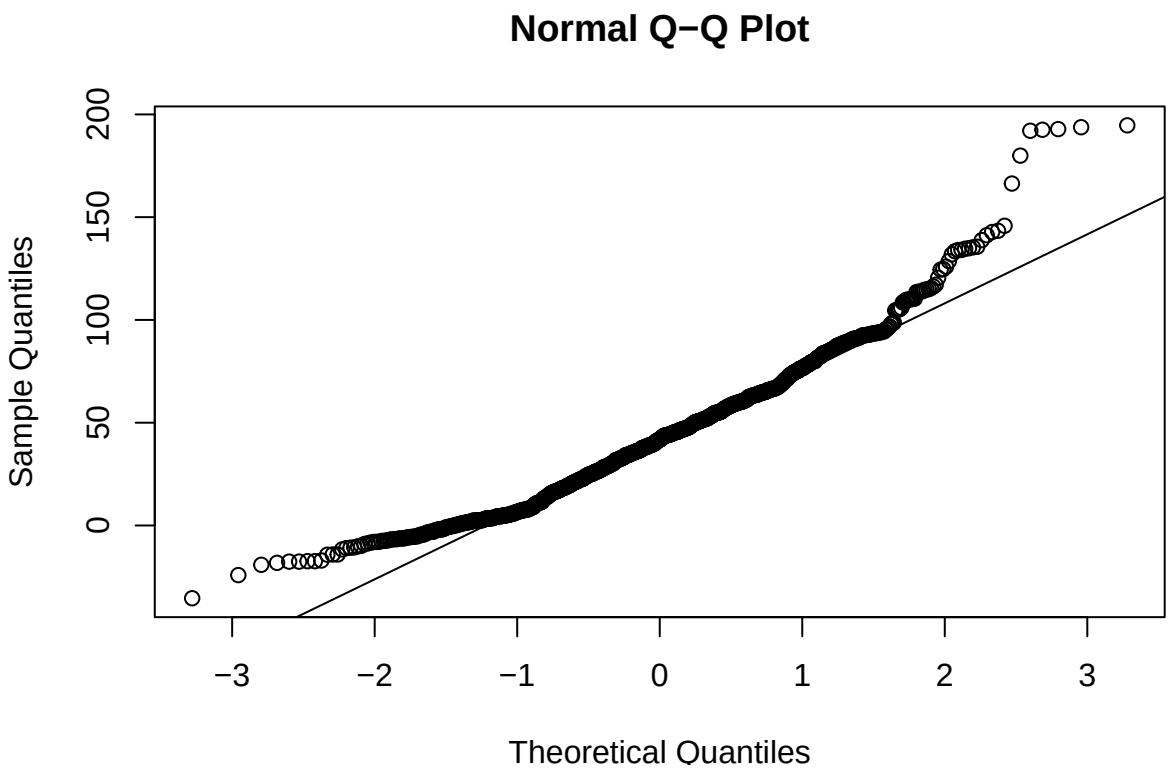
```
##    well_id            timestamp                     ground_to_water_cm
```

```
##  Length:1116        Min.   :2018-05-31 08:30:00.00   Min.   :-35.41
##  Class :character   1st Qu.:2018-10-14 08:55:45.00   1st Qu.: 18.30
##  Mode  :character   Median :2019-09-21 07:50:30.00   Median : 41.86
##                     Mean   :2020-03-02 18:05:30.48   Mean   : 43.85
##                     3rd Qu.:2021-07-20 06:53:30.00   3rd Qu.: 63.59
##                     Max.   :2021-11-14 10:14:00.00   Max.   :194.67
##                                                      NA's   :151
##       date                 year         isoweek       day_of_year
##  Min.   :2018-05-31   Min.   :2018   Min.   :20.00   Min.   :140.0
##  1st Qu.:2018-10-14   1st Qu.:2018   1st Qu.:26.00   1st Qu.:185.0
##  Median :2019-09-21   Median :2019   Median :31.00   Median :217.0
##  Mean   :2020-03-02   Mean   :2020   Mean   :31.94   Mean   :222.2
##  3rd Qu.:2021-07-20   3rd Qu.:2021   3rd Qu.:38.00   3rd Qu.:264.0
##  Max.   :2021-11-14   Max.   :2021   Max.   :46.00   Max.   :322.0
##
```

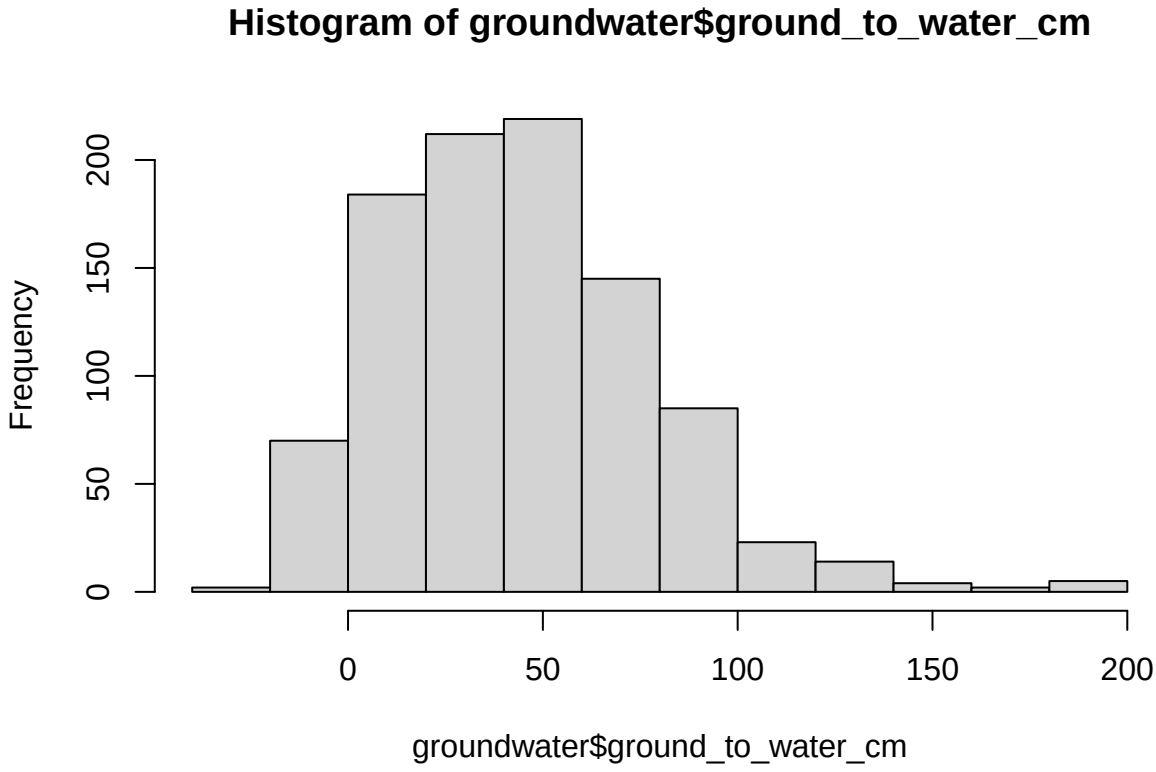
```r
# use z-score? test if data is normally distributed

# shapiro-wilk test; data is likely non-normal if p-value < 0.05
shapiro.test(groundwater$ground_to_water_cm)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  groundwater$ground_to_water_cm
## W = 0.96037, p-value = 1.653e-15
```

```r
# results: p << 0.05, data is non-normal

# Q-Q plots; data is normal if falls on a straight line
qqnorm(groundwater$ground_to_water_cm)
qqline(groundwater$ground_to_water_cm)
```




```
# results: mostly normal but some outliers

# Histogram; check for bell-shaped curve
hist(groundwater$ground_to_water_cm)
```



```r
# Skewness; test if near 0 (symmetric), >0 (positive skew), <0 (neg skew)
skewness(groundwater$ground_to_water_cm, na.rm = TRUE)
```

```
## [1] 0.8304517
```

```r
# result: 0.83; positive skewed, >0.5 so moderately skewed

# Kurtosis; test for heavy tails (if ~3 normal, if >3 heavy tails + sharp peak, if <3 light tails, flat
kurtosis(groundwater$ground_to_water_cm, na.rm = TRUE)
```

```
## [1] 4.468805
```

```r
# result 4.5; heavy tail and sharp peak

# YES, USE Z-SCORE!
# ---TODO: Add ground_to_water_zscore column

# summarize the span of full time series by date
groundwater %>% summarize(
  start_date = min(date, na.rm=TRUE),
  stop_date = max(date, na.rm=TRUE),
  timespan = difftime(stop_date, start_date, units="days"),
  unique_dates_count = n_distinct(date)
)
```

```
##   start_date  stop_date  timespan unique_dates_count
## 1 2018-05-31 2021-11-14 1263 days                 53
```

```r
# summarize the weekly data
groundwater_weekly_summary <- groundwater %>%
  group_by(isoweek) %>%
  summarize(
      n_week = n()  # Number of entries in each week
  )
groundwater_weekly_summary
```

```
## # A tibble: 24 x 2
##    isoweek n_week
##      <dbl>  <int>
##  1      20     27
##  2      22     84
##  3      23     17
##  4      24     51
##  5      25     55
##  6      26     58
##  7      27     72
##  8      28      3
##  9      29     85
## 10      30     83
## # i 14 more rows
```

```r
# summarize spacing of the full time series
unique_observations <- groundwater %>%
  select(-well_id, -ground_to_water_cm, -timestamp) %>%
  distinct(date, .keep_all=TRUE) %>%
  group_by(year) %>%
  arrange(date) %>%
```

```
  mutate(day_diff = as.numeric(difftime(lead(date), date, units="days")))

unique_observations
```

```
## # A tibble: 53 x 5
## # Groups:   year [3]
##    date        year isoweek day_of_year day_diff
##    <date>     <dbl>   <dbl>       <dbl>    <dbl>
##  1 2018-05-31  2018      22         151        1
##  2 2018-06-01  2018      22         152       17
##  3 2018-06-18  2018      25         169        1
##  4 2018-06-19  2018      25         170       11
##  5 2018-06-30  2018      26         181        1
##  6 2018-07-01  2018      26         182       15
##  7 2018-07-16  2018      29         197        1
##  8 2018-07-17  2018      29         198        8
##  9 2018-07-25  2018      30         206        2
## 10 2018-07-27  2018      30         208       14
## # i 43 more rows
```

```
#write.csv(unique_observations, observation_filepath)

unique_observations %>% summarize(
  max_days = max(day_diff, na.rm=TRUE),
  mean_days = mean(day_diff, na.rm=TRUE)
)
```

```
## # A tibble: 3 x 3
##    year max_days mean_days
##   <dbl>    <dbl>     <dbl>
## 1  2018       35     10.1
## 2  2019       19      7.05
## 3  2021       39     13.7
```

```
# filter measurements, only before 11a
am_time_limit <- 10
groundwater_filter_by_time <- groundwater %>%
  filter(hour(timestamp) < am_time_limit)

# number of rows lost from this filter
nrow(groundwater) - nrow(groundwater_filter_by_time)
```

```
## [1] 193
```

```
# filter for duplicate entries (same well, same week)
groundwater_filter_duplicates <- groundwater_filter_by_time %>%
  group_by(well_id, isoweek) %>%
  distinct(well_id, isoweek, .keep_all = TRUE) %>%
  ungroup()

# compare and validate results
summary(groundwater)
```

```
##    well_id            timestamp                      ground_to_water_cm
##  Length:1116        Min.   :2018-05-31 08:30:00.00   Min.   :-35.41
##  Class :character   1st Qu.:2018-10-14 08:55:45.00   1st Qu.: 18.30
```

```
##  Mode  :character   Median :2019-09-21 07:50:30.00   Median : 41.86
##                     Mean   :2020-03-02 18:05:30.48   Mean   : 43.85
##                     3rd Qu.:2021-07-20 06:53:30.00   3rd Qu.: 63.59
##                     Max.   :2021-11-14 10:14:00.00   Max.   :194.67
##                                                      NA's   :151
##       date                 year         isoweek       day_of_year
##  Min.   :2018-05-31   Min.   :2018   Min.   :20.00   Min.   :140.0
##  1st Qu.:2018-10-14   1st Qu.:2018   1st Qu.:26.00   1st Qu.:185.0
##  Median :2019-09-21   Median :2019   Median :31.00   Median :217.0
##  Mean   :2020-03-02   Mean   :2020   Mean   :31.94   Mean   :222.2
##  3rd Qu.:2021-07-20   3rd Qu.:2021   3rd Qu.:38.00   3rd Qu.:264.0
##  Max.   :2021-11-14   Max.   :2021   Max.   :46.00   Max.   :322.0
##
```

```
summary(groundwater_filter_by_time)
```

```
##    well_id            timestamp                     ground_to_water_cm
##  Length:923         Min.   :2018-05-31 08:30:00.00   Min.   :-19.12
##  Class :character   1st Qu.:2018-09-29 08:25:00.00   1st Qu.: 19.30
##  Mode  :character   Median :2019-09-03 07:55:00.00   Median : 41.97
##                     Mean   :2020-02-26 00:58:58.04   Mean   : 44.37
##                     3rd Qu.:2021-07-05 08:14:00.00   3rd Qu.: 63.71
##                     Max.   :2021-11-14 09:59:00.00   Max.   :194.67
##                                                      NA's   :133
##       date                 year         isoweek      day_of_year
##  Min.   :2018-05-31   Min.   :2018   Min.   :20.0   Min.   :140.0
##  1st Qu.:2018-09-29   1st Qu.:2018   1st Qu.:26.0   1st Qu.:182.0
##  Median :2019-09-03   Median :2019   Median :30.0   Median :207.0
##  Mean   :2020-02-25   Mean   :2020   Mean   :30.9   Mean   :214.8
##  3rd Qu.:2021-07-05   3rd Qu.:2021   3rd Qu.:36.0   3rd Qu.:251.0
##  Max.   :2021-11-14   Max.   :2021   Max.   :45.0   Max.   :318.0
##
```

```
summary(groundwater_filter_duplicates)
```

```
##    well_id            timestamp                     ground_to_water_cm
##  Length:688         Min.   :2018-05-31 08:30:00.00   Min.   :-19.12
##  Class :character   1st Qu.:2018-08-10 08:31:30.00   1st Qu.: 17.84
##  Mode  :character   Median :2019-08-04 07:59:00.00   Median : 39.80
##                     Mean   :2019-12-02 23:25:09.50   Mean   : 42.73
##                     3rd Qu.:2021-06-20 07:12:30.00   3rd Qu.: 63.38
##                     Max.   :2021-11-14 09:59:00.00   Max.   :193.76
##                                                      NA's   :109
##       date                 year         isoweek       day_of_year
##  Min.   :2018-05-31   Min.   :2018   Min.   :20.00   Min.   :140.0
##  1st Qu.:2018-08-10   1st Qu.:2018   1st Qu.:26.00   1st Qu.:181.0
##  Median :2019-08-04   Median :2019   Median :30.00   Median :208.0
##  Mean   :2019-12-02   Mean   :2019   Mean   :31.02   Mean   :215.8
##  3rd Qu.:2021-06-20   3rd Qu.:2021   3rd Qu.:36.00   3rd Qu.:251.0
##  Max.   :2021-11-14   Max.   :2021   Max.   :45.00   Max.   :318.0
##
```

```
# test completeness of full time series
# ---TODO: create a ground_to_water_greater_than column for NO WATER readings
#          to capture the depth of the well; this is more info than nothing.
```

```r
# completeness in terms of entries with NA
# ---NOTE: expect 153 NA entries due to NO WATER readings

groundwater %>%
  summarize(
    na_sum = sum(is.na(ground_to_water_cm)),
    na_percent = 100 * sum(is.na(ground_to_water_cm)) / n())
```

```
##   na_sum na_percent
## 1    151   13.53047
```

```r
# basic descriptive statistics across all groundwater readings: mean, CV, ACF
groundwater %>%
  summarize(
    mean_value = mean(ground_to_water_cm, na.rm = TRUE),
    sd_value = sd(ground_to_water_cm, na.rm = TRUE),
    var_value = var(ground_to_water_cm, na.rm = TRUE),
    cv_value = 100 * sd_value / mean_value
  )
```

```
##   mean_value sd_value var_value cv_value
## 1   43.84532 34.49195  1189.695 78.66735
```

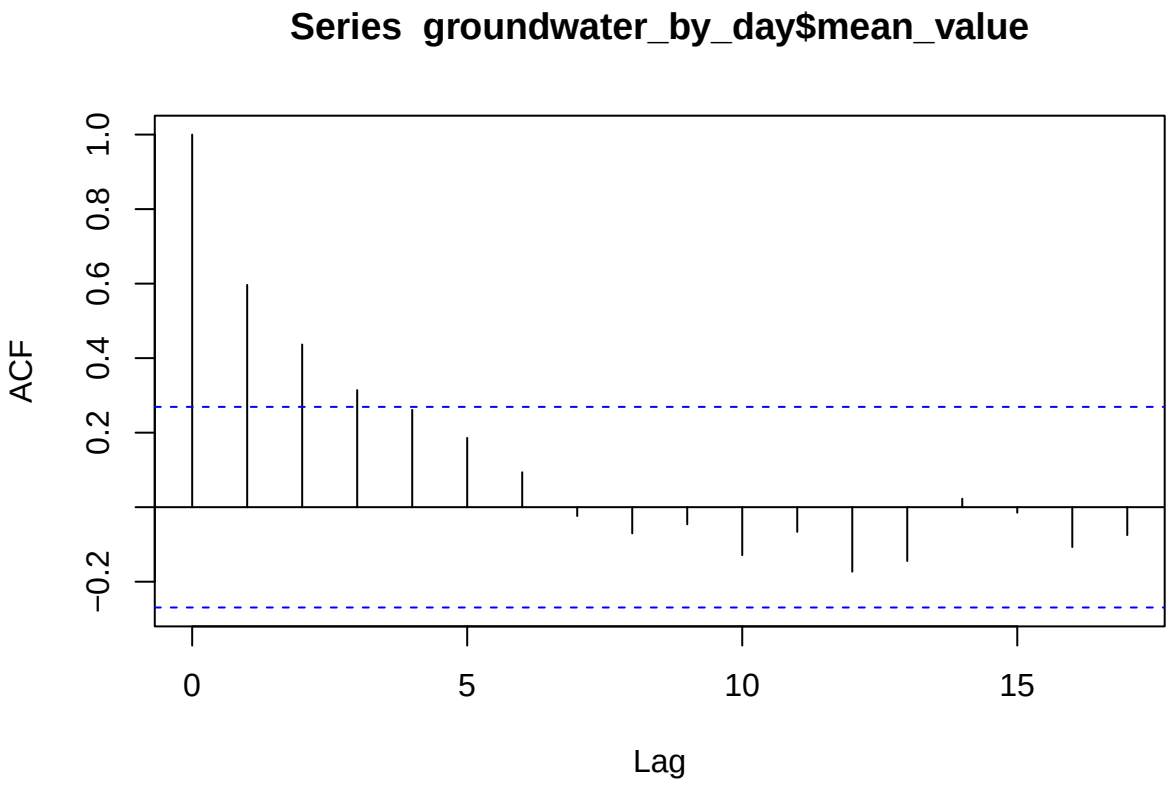
```r
groundwater_by_day <- groundwater %>%
  group_by(date) %>%
  summarise(
    mean_value = mean(ground_to_water_cm, na.rm = TRUE),
    sd_value = sd(ground_to_water_cm, na.rm = TRUE),
    var_value = var(ground_to_water_cm, na.rm = TRUE)
  )

# basic descriptive statistics by groupings: mean, CV, ACF for each variable
groundwater_by_well <- groundwater %>%
  group_by(well_id) %>%
  summarise(
    mean_value = mean(ground_to_water_cm, na.rm = TRUE),
    sd_value = sd(ground_to_water_cm, na.rm = TRUE),
    var_value = var(ground_to_water_cm, na.rm = TRUE)
  )
groundwater_by_well
```

```
## # A tibble: 54 x 4
##    well_id  mean_value sd_value var_value
##    <chr>         <dbl>    <dbl>     <dbl>
##  1 EEF-1          28.8    31.7     1003.
##  2 EER-1          34.1    10.8      117.
##  3 EET-1          59.9    32.8     1076.
##  4 EET-2         101.     41.2     1699.
##  5 EET-XB4S       31.6    20.0      400.
##  6 EFF-XA1N       40.1    11.3      127.
##  7 EFF-XA2N       27.0    11.6      134.
##  8 EFF-XB7S       30.5    36.8     1357.
##  9 EFR-XB1S       45.2     4.96      24.6
## 10 EFR-XB2N       28.7    11.3      127.
## # i 44 more rows
```

```r
groundwater_by_well_year <- groundwater %>%
  mutate(year = year(timestamp)) %>%
  group_by(well_id, year) %>%
  summarise(
    mean_value = mean(ground_to_water_cm, na.rm = TRUE),
    sd_value = sd(ground_to_water_cm, na.rm = TRUE),
    var_value = var(ground_to_water_cm, na.rm = TRUE),
    .groups = "keep"
  )
groundwater_by_well_year
```

```
## # A tibble: 141 x 5
## # Groups:   well_id, year [141]
##    well_id  year mean_value sd_value var_value
##    <chr>   <dbl>      <dbl>    <dbl>     <dbl>
##  1 EEF-1    2018       6.21     7.70      59.3
##  2 EEF-1    2019      19.2     21.6      465.
##  3 EEF-1    2021      61.0     29.8      888.
##  4 EER-1    2018      34.5      8.50      72.3
##  5 EER-1    2019      27.6     12.4      155.
##  6 EER-1    2021      39.1      8.76      76.7
##  7 EET-1    2018      65.4     31.7     1007.
##  8 EET-1    2019      57.0     41.7     1736.
##  9 EET-1    2021      56.4     31.1      968.
## 10 EET-2    2019     107.      20.4      418.
## # i 131 more rows
```

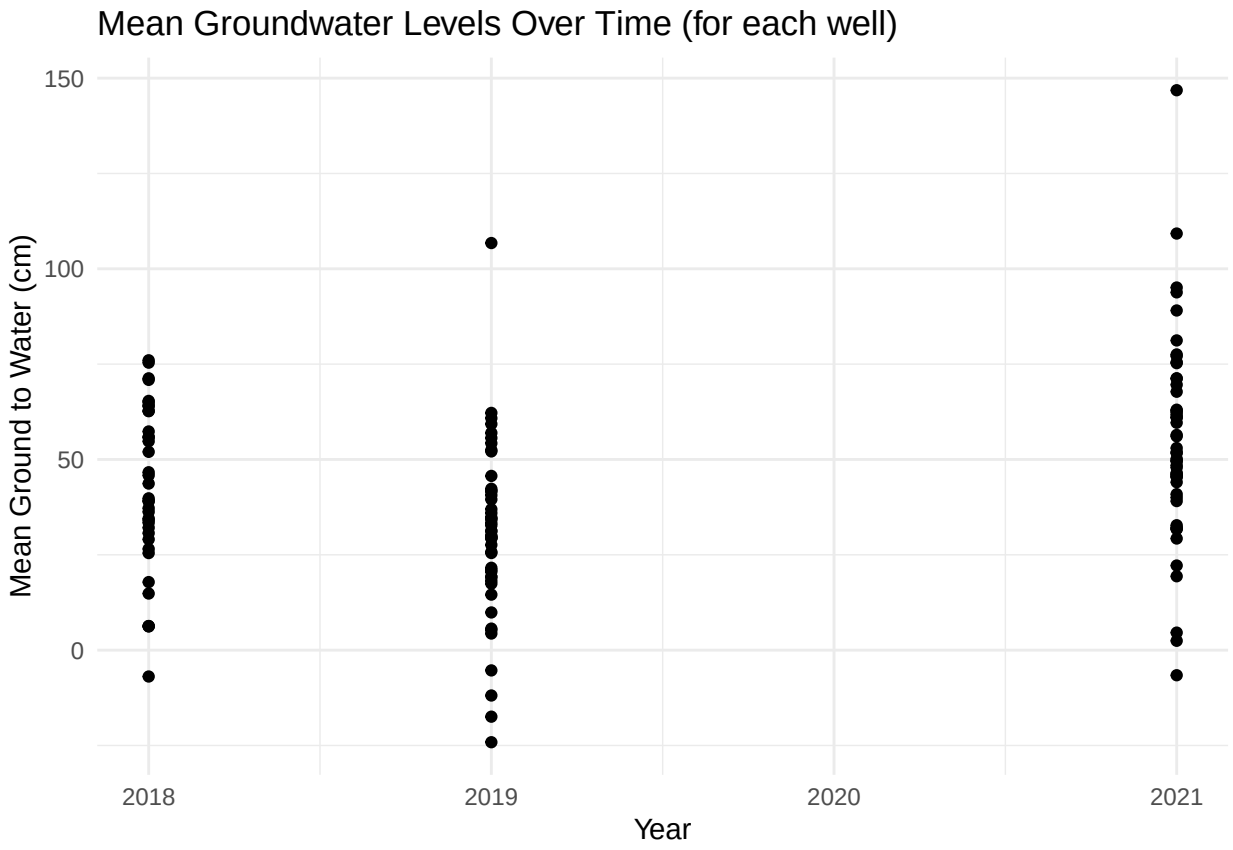
```r
# Trying to get acf, but not sure if this summarized data means anything
acf(groundwater_by_day$mean_value)
```



```
# TODO Next step summary statistics
# ------wells in each meadow group: kiln, east, low
# ------wells from each plant functional type: sedge, willow, mixed herbaceous, pine
# ------wells from each hydrogeomorphic zone: riparian, terrace, fan

# TODO: Next time series to consider
# ------discharge (at one point)
# ------daily precipitation (at one point)
# ------sunlight, aka PAR (at one point)
# ------max, mean daily temperature (at each meadow)
```

Plots

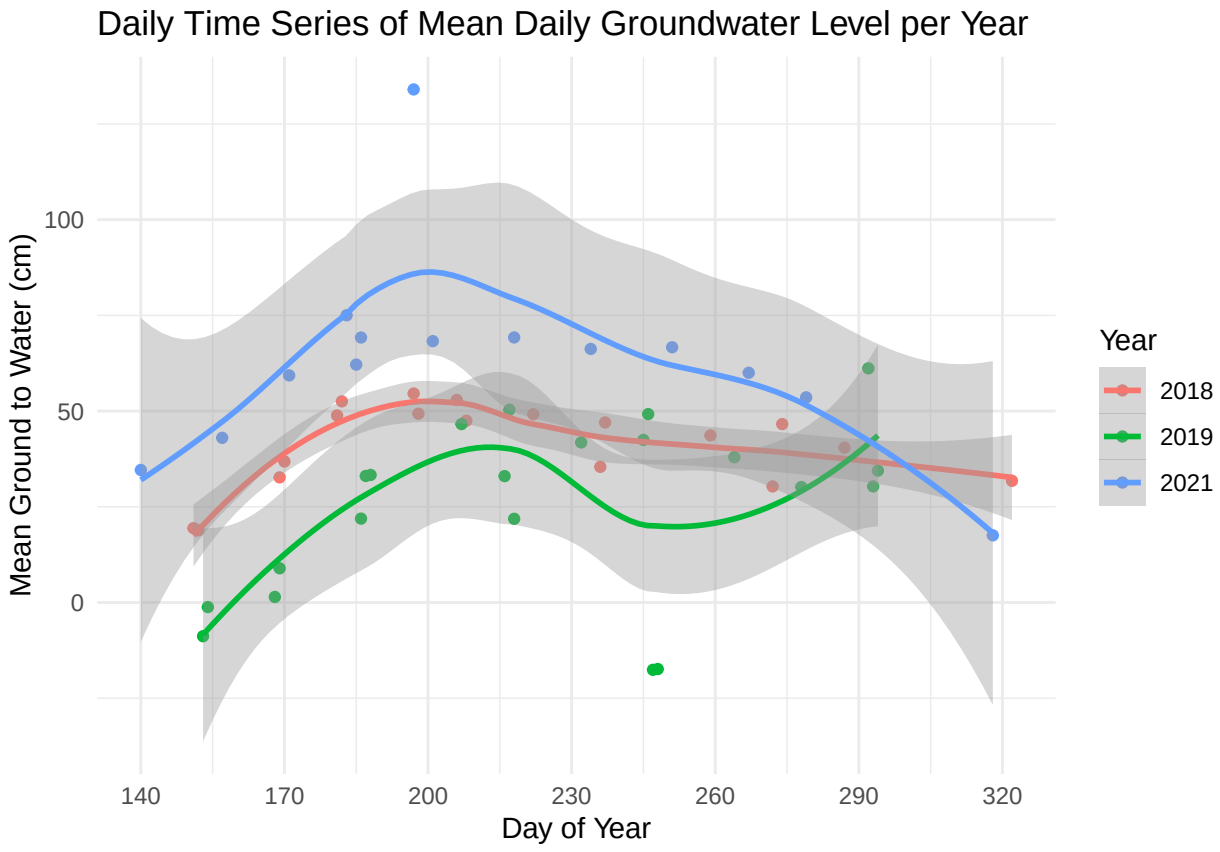
```
# Load libraries
library(ggplot2)
library(tidyr)

# Plot (1): the mean annual groundwater level for all wells for each year
ggplot(groundwater_by_well_year, aes(x = year, y = mean_value, group = well_id)) +
  geom_point() +  # Optional: add points at each data point
  labs(title = "Mean Groundwater Levels Over Time (for each well)",
       x = "Year", y = "Mean Ground to Water (cm)") +
  theme_minimal()
```




```
# Plot (2): year-over-year daily time series of mean groundwater level

# ---Setup dataframe with new columns for year and day_of_year
```

```r
groundwater_by_day <- groundwater_by_day %>%
  mutate(
    year = year(date),                # Extract the year from Date
    day_of_year = yday(date)          # Extract the day of year (1-365/366)
  )

# ---Add NA values for days with no measurement (or mean_value)
complete_groundwater_by_day <- groundwater_by_day %>%
  group_by(year) %>%
  complete(day_of_year =
           min(groundwater_by_day$day_of_year):max(groundwater_by_day$day_of_year),
           fill = list(mean_value = NA))  # Fill missing days with NA

# ---Plot it!
ggplot(complete_groundwater_by_day, aes(x = day_of_year, y = mean_value, color = factor(year), group = y
  geom_point() +
  geom_smooth() +
  theme_bw() +
  labs(title = "Daily Time Series of Mean Daily Groundwater Level per Year",
       x = "Day of Year",
       y = "Mean Ground to Water (cm)",
       color = "Year") +
  scale_x_continuous(breaks = seq(min(groundwater_by_day$day_of_year),
                                  max(groundwater_by_day$day_of_year), by = 30)) +  # Customize x-axis
  theme_minimal()
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



Research Questions and Hypotheses

1. How does meadow groundwater vary by season and climate as influenced by elevation, hydrogeomorphic zones, and evapotranspiration rates of plant functional types?

- Hypothesis: I expect evapotranspiration to drive daily and seasonal groundwater levels with sensitivity to meteorology and day length.

2. What controls plant functional type phenology?

- Hypothesis: peak productivity and senescence will correlate to groundwater levels as governed by meteorology but moderated by hydrogeomorphic zones and elevation.

3. Does topography or subsurface character influence groundwater reliability?

- Hypothesis: I expect that groundwater reliability will correlate to topographic convergence or subsurface boundaries (i.e. differing conductivity).