Pset 1 - Water usage

425/625

Spring 2024

Introduction

Water scarcity is a major issue in many parts of the world. According to the United Nations, "About two billion people worldwide don't have access to safe drinking water today (SDG Report 2022), and roughly half of the world's population is experiencing severe water scarcity for at least part of the year (IPCC). These numbers are expected to increase, exacerbated by climate change and population growth (WMO)."

In this problem set, we will investigate water usage estimates by crop in the United States. The `.csv` for this data set comes from here (by checking Select All and clicking Get Custom Zip) and the associated academic journal article is here. See this thread on X for a summary.

Read the academic article to familiarize yourself with the basics of the water usage data. You don't need to know how these water usage levels were estimated, so you can skip over those parts. We are going to focus on visualizing the water levels using the estimates that they generated.

Data preparation

The `.zip` file `rawdata/DOI-10-13012-b2idb-4607538_v1.zip` contains one `.csv` file per source (SWW, GWW, GWD) per year from 2008 to 2020. There are also a couple of `.txt` files in the folder. We can use `unzip` with `list = TRUE` to see what's in the `.zip` file.

```
unzip(zipfile = 'rawdata/DOI-10-13012-b2idb-4607538_v1.zip',
      list = TRUE) ## this lists the filename, but does not unzip the file
```

```
##                                          Name  Length                Date
## 1        DOI-10-13012-b2idb-4607538_v1/readme.txt    1053 2023-10-29 14:08:00
## 2      DOI-10-13012-b2idb-4607538_v1/gwa_2008.csv 2274812 2023-10-29 14:08:00
## 3      DOI-10-13012-b2idb-4607538_v1/gwa_2009.csv 2274812 2023-10-29 14:08:00
## 4      DOI-10-13012-b2idb-4607538_v1/gwa_2010.csv 2200859 2023-10-29 14:08:00
## 5      DOI-10-13012-b2idb-4607538_v1/gwa_2011.csv 2274812 2023-10-29 14:08:00
## 6      DOI-10-13012-b2idb-4607538_v1/gwa_2012.csv 2274812 2023-10-29 14:08:00
## 7      DOI-10-13012-b2idb-4607538_v1/gwa_2013.csv 2274812 2023-10-29 14:08:00
## 8      DOI-10-13012-b2idb-4607538_v1/gwa_2014.csv 2274812 2023-10-29 14:08:00
## 9      DOI-10-13012-b2idb-4607538_v1/gwa_2015.csv 2200859 2023-10-29 14:08:00
## 10     DOI-10-13012-b2idb-4607538_v1/gwa_2016.csv 2275517 2023-10-29 14:08:00
## 11     DOI-10-13012-b2idb-4607538_v1/gwa_2017.csv 2275517 2023-10-29 14:08:00
## 12     DOI-10-13012-b2idb-4607538_v1/gwa_2018.csv 2275517 2023-10-29 14:08:00
## 13     DOI-10-13012-b2idb-4607538_v1/gwa_2019.csv 2275517 2023-10-29 14:08:00
## 14     DOI-10-13012-b2idb-4607538_v1/gwa_2020.csv 2275517 2023-10-29 14:08:00
## 15     DOI-10-13012-b2idb-4607538_v1/gwd_2008.csv  211884 2023-10-29 14:08:00
## 16     DOI-10-13012-b2idb-4607538_v1/gwd_2009.csv  208249 2023-10-29 14:08:00
## 17     DOI-10-13012-b2idb-4607538_v1/gwd_2010.csv  214546 2023-10-29 14:08:00
## 18     DOI-10-13012-b2idb-4607538_v1/gwd_2011.csv  213608 2023-10-29 14:08:00
## 19     DOI-10-13012-b2idb-4607538_v1/gwd_2012.csv  210157 2023-10-29 14:08:00
```

```
## 20     DOI-10-13012-b2idb-4607538_v1/gwd_2013.csv  207564 2023-10-29 14:08:00
## 21     DOI-10-13012-b2idb-4607538_v1/gwd_2014.csv  209619 2023-10-29 14:08:00
## 22     DOI-10-13012-b2idb-4607538_v1/gwd_2015.csv  208683 2023-10-29 14:08:00
## 23     DOI-10-13012-b2idb-4607538_v1/gwd_2016.csv  206644 2023-10-29 14:08:00
## 24     DOI-10-13012-b2idb-4607538_v1/gwd_2017.csv  206188 2023-10-29 14:08:00
## 25     DOI-10-13012-b2idb-4607538_v1/gwd_2018.csv  206429 2023-10-29 14:08:00
## 26     DOI-10-13012-b2idb-4607538_v1/gwd_2019.csv  208246 2023-10-29 14:08:00
## 27     DOI-10-13012-b2idb-4607538_v1/gwd_2020.csv  208252 2023-10-29 14:08:00
## 28      DOI-10-13012-b2idb-4607538_v1/sw_2008.csv 2274792 2023-10-29 14:08:00
## 29      DOI-10-13012-b2idb-4607538_v1/sw_2009.csv 2274792 2023-10-29 14:08:00
## 30      DOI-10-13012-b2idb-4607538_v1/sw_2010.csv 2200839 2023-10-29 14:08:00
## 31      DOI-10-13012-b2idb-4607538_v1/sw_2011.csv 2274792 2023-10-29 14:08:00
## 32      DOI-10-13012-b2idb-4607538_v1/sw_2012.csv 2274792 2023-10-29 14:08:00
## 33      DOI-10-13012-b2idb-4607538_v1/sw_2013.csv 2274792 2023-10-29 14:08:00
## 34      DOI-10-13012-b2idb-4607538_v1/sw_2014.csv 2274792 2023-10-29 14:08:00
## 35      DOI-10-13012-b2idb-4607538_v1/sw_2015.csv 2200839 2023-10-29 14:08:00
## 36      DOI-10-13012-b2idb-4607538_v1/sw_2016.csv 2275497 2023-10-29 14:08:00
## 37      DOI-10-13012-b2idb-4607538_v1/sw_2017.csv 2275497 2023-10-29 14:08:00
## 38      DOI-10-13012-b2idb-4607538_v1/sw_2018.csv 2275497 2023-10-29 14:08:00
## 39      DOI-10-13012-b2idb-4607538_v1/sw_2019.csv 2275497 2023-10-29 14:08:00
## 40      DOI-10-13012-b2idb-4607538_v1/sw_2020.csv 2275497 2023-10-29 14:08:00
## 41 DOI-10-13012-b2idb-4607538_v1/dataset_info.txt    3894 2023-10-29 14:08:00
```

Before summarizing/visualizing this data, we'll want to join these data sets. We could certainly unzip the file manually. We can also do this in R using `unzip`.

```
unzip(zipfile = 'rawdata/DOI-10-13012-b2idb-4607538_v1.zip',
      junkpaths = TRUE,
      exdir = 'rawdata') ## gets rid of paths, keeps only filenames
```

1. Join data First, let's create a data set with all years/crops together in one data frame. Below is some code to help you get started. Add comments to each place there is `##` to explain what the chunk of code is doing. Then add code to the `Tranforming data` Section to transform the data into a data frame with 5 columns: `GEOID`, `crop`, `source`, `year`, and `value` (indicating km^3 of water).

Note that `eval = F` at the start of the chunk will prevent this chunk from evaluating when you knit the document. You can temporarily remove it if you'd like, but you'll want to add it back before knitting the document so that knitting takes less time.

```
sources = c('gwd', 'sw', 'gwa')
years = 2008:2020
d = NULL

for(s in sources){
  cat(s, '') ## show progress

  for(year in years){
    cat(year, '') ## show progress

    ## The code is reading in the data from the .csv file by combining directory,
    ## the source, year, and .csv. The csv is then stored as a dataframe called df.
    filename = paste0('rawdata/', s, '_', year, '.csv')
    df = read.csv(filename)
    head(df)
```

```
    ## Tranform data ########################################
    ## Use `pivot_longer`, `separate`, and/or other functions to transform this
    ## data frame into a data frame with 5 columns:
    ## GEOID, crop, source, year, and value (indicating km^3 of water)

    df = df %>%
      pivot_longer(cols = -GEOID, names_to = "crop_info", values_to = "value") %>%
      separate(col = 'crop_info', into = c("src", "crop", 'year'), sep = "\\.")


    ## end of transforming data ##############################

    ## We then row bind the dataframe df to the master dataframe d.
    d = rbind(d, df)
  }

  cat('\n') ## start a new line before showing progress for the next source
}
head(d)
tail(d)
```

Data exploration and summaries

Let's load the data we'll use for the rest of the assignment. This is the data set created in #1, so if you were unable to finish #1, you can still do the rest of the assignment.

```
d = readRDS('data/water.usage.rds')
head(d)
```

```
## # A tibble: 6 x 5
##   GEOID crop        src   year  value
##   <int> <chr>       <chr> <chr> <dbl>
## 1  1001 barley      gwd   2008      0
## 2  1001 corn        gwd   2008      0
## 3  1001 cotton      gwd   2008      0
## 4  1001 millet      gwd   2008      0
## 5  1001 oats        gwd   2008      0
## 6  1001 other_sctg2 gwd   2008      0
```

2. Summaries of data Find the mean, the change from 2008 to 2020, and the percent change from 2008 to 2020, for each crop and each source (SWW, GWW, GWD).

```
dd = d %>%
  group_by(crop, src, year) %>%
  summarise(value = sum(value)) %>%
  group_by(crop, src) %>%
  mutate(mean = mean(value)) %>%
  filter(year %in% c(2008, 2020)) %>%
  pivot_wider(names_from = year,
              values_from = value,
              names_prefix = 'y') %>%
  mutate(diff=y2020-y2008,
         perc = diff/y2008*100) %>%
  ungroup()
```

```
## `summarise()` has grouped output by 'crop', 'src'. You can override using the
## `.groups` argument.
```

```
head(dd)
```

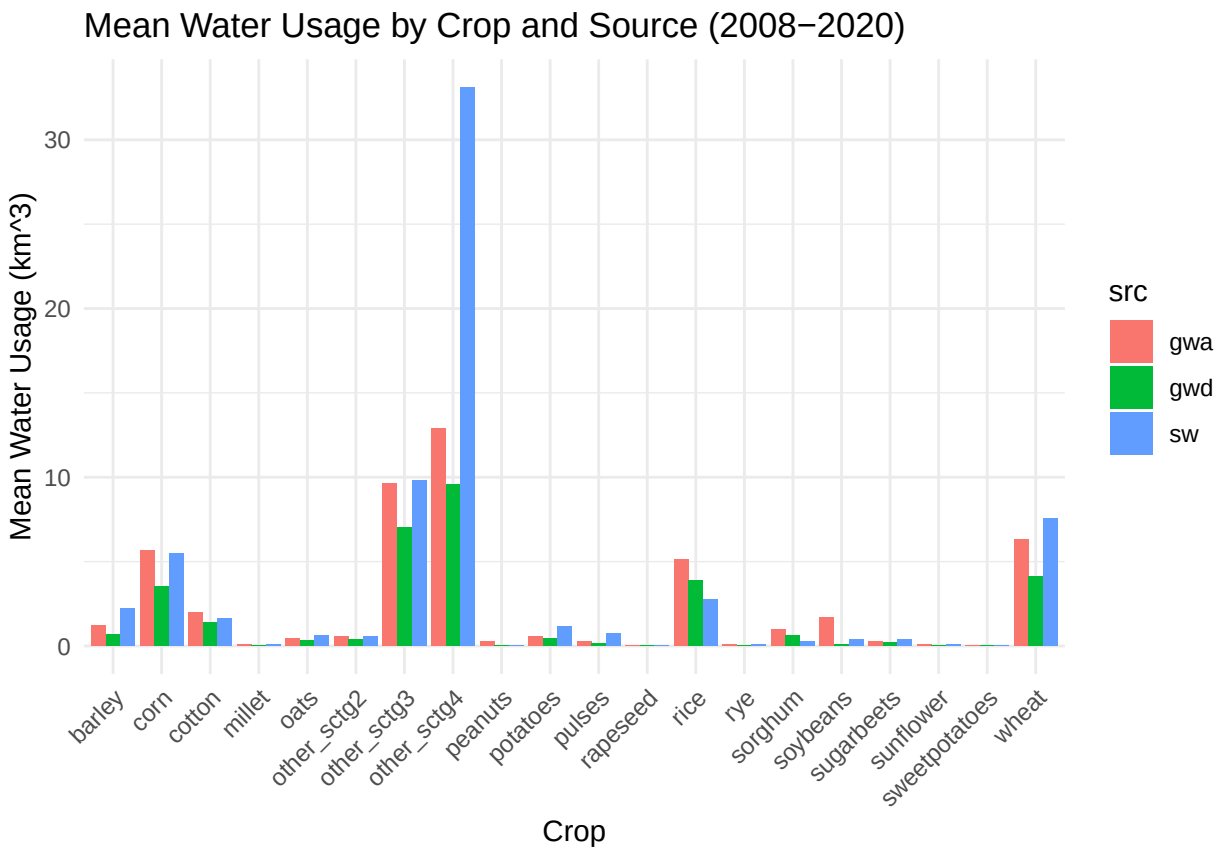
```
## # A tibble: 6 x 7
##   crop   src    mean y2008 y2020    diff   perc
##   <chr>  <chr> <dbl> <dbl> <dbl>   <dbl>  <dbl>
## 1 barley gwa   1.19  1.21  1.28   0.0631   5.21
## 2 barley gwd   0.711 0.677 0.559 -0.118  -17.4
## 3 barley sw    2.20  2.38  1.87  -0.508  -21.4
## 4 corn   gwa   5.65  5.38  5.99   0.617   11.5
## 5 corn   gwd   3.52  3.63  3.46  -0.167   -4.61
## 6 corn   sw    5.50  7.12  4.94  -2.19   -30.7
```

```
write.csv(dd, 'water_usage_summary.csv')
```

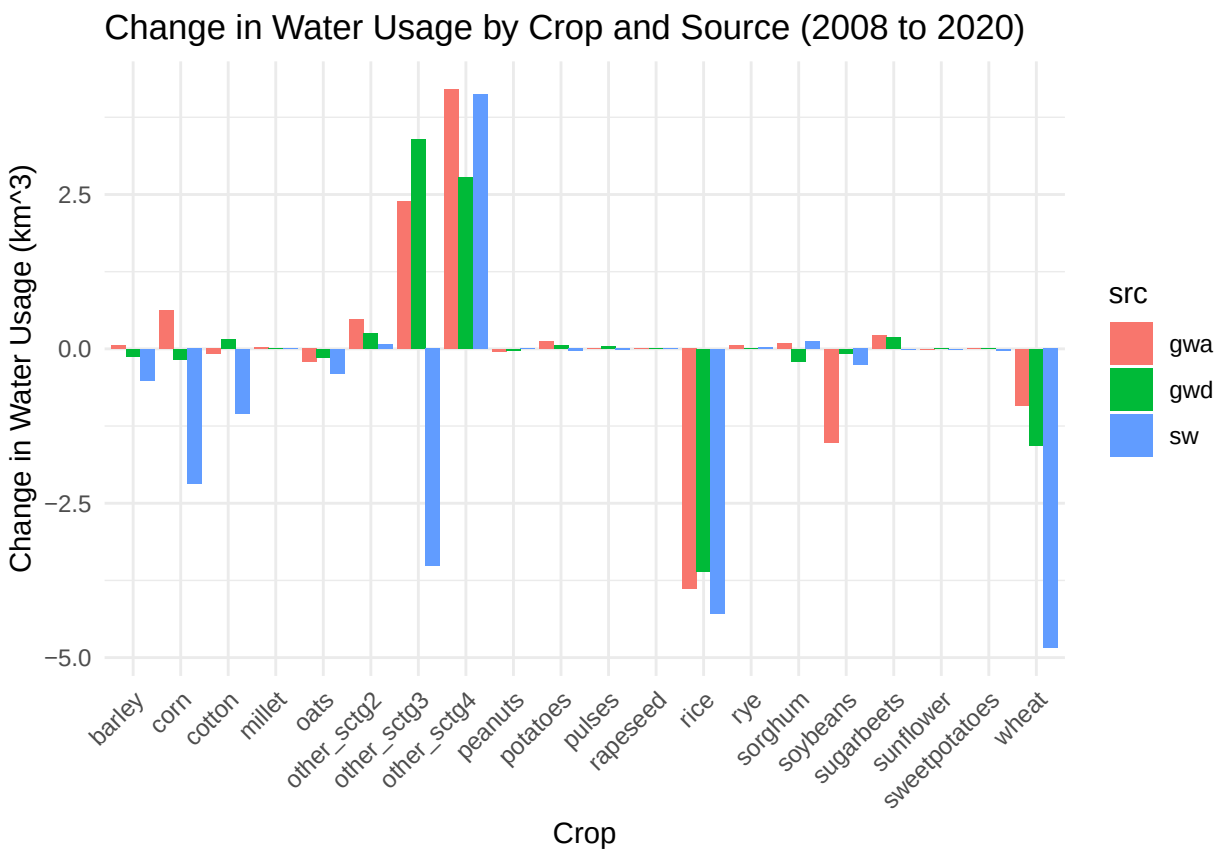
3. Convert Table 2 to a visualization

Create a visual representation of the information in Table 2. Create a visualization (or visualizations) that contains mean, change, and percent change in water usage from each crop and source. Pivot longer Title labeled axes text? Visualization best practices

```
ggplot(dd, aes(x = crop, y = mean, fill = src)) +
  geom_bar(stat = "identity", position = position_dodge()) +
  labs(title = "Mean Water Usage by Crop and Source (2008-2020)",
       x = "Crop",
       y = "Mean Water Usage (km^3)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
# Visualization for Change in Water Usage
ggplot(dd, aes(x = crop, y = diff, fill = src)) +
  geom_bar(stat = "identity", position = position_dodge()) +
  labs(title = "Change in Water Usage by Crop and Source (2008 to 2020)",
       x = "Crop",
       y = "Change in Water Usage (km^3)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```r
# Visualization for Percent Change in Water Usage
ggplot(dd, aes(x = crop, y = perc, fill = src)) +
  geom_bar(stat = "identity", position = position_dodge()) +
  labs(title = "Percent Change in Water Usage by Crop and Source (2008 to 2020)",
       x = "Crop",
       y = "Percent Change in Water Usage (%)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

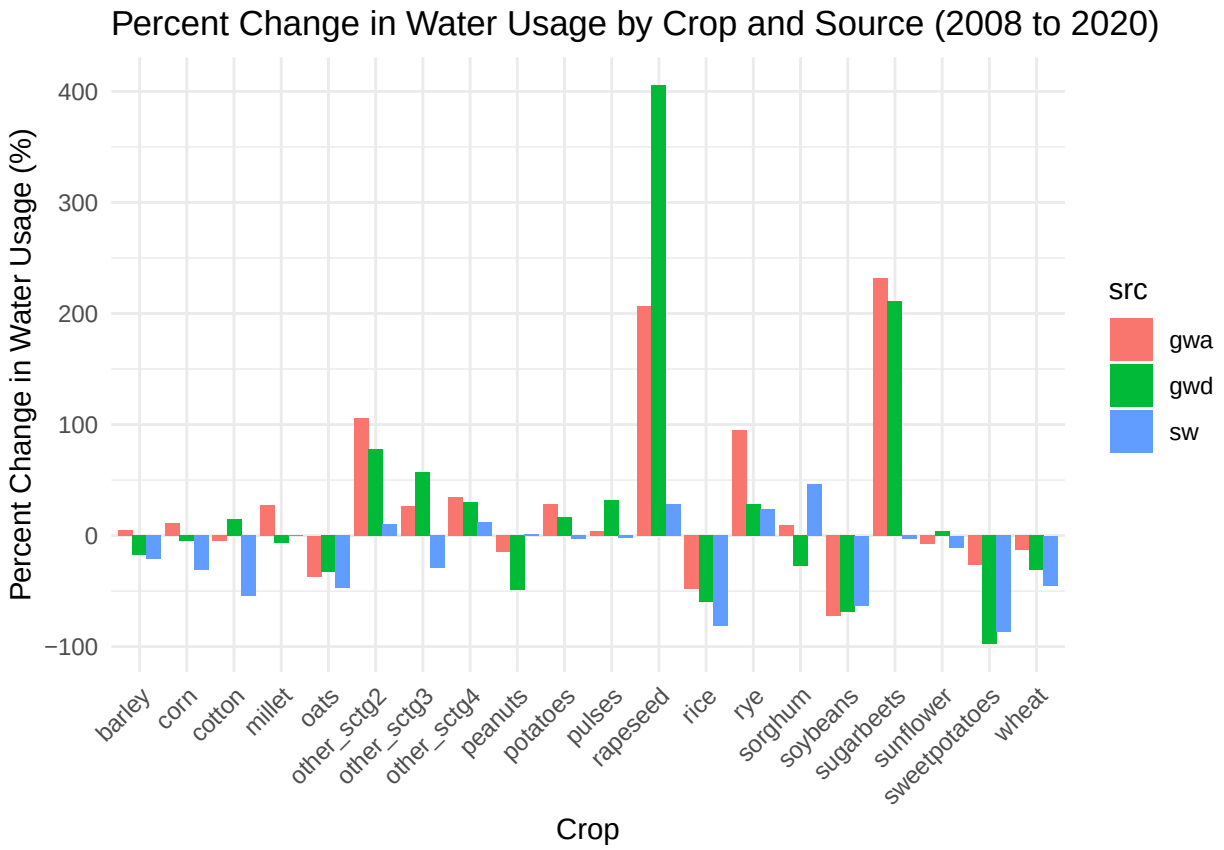


Figure 4

Figure 4 shows the average water usage by crop and source.

- A. average irrigation water usage by source, colored by crop,
- B. average irrigation water usage by crop, colored by source

Two other options for visualizing a numeric variable broken down by two different categorical variable would be a tile plot/grid plot (e.g. https://github.com/bmacGTPM/pubtheme?tab=readme-ov-file#grid-plot) and a mosiac plot (https://haleyjeppson.github.io/ggmosaic/).

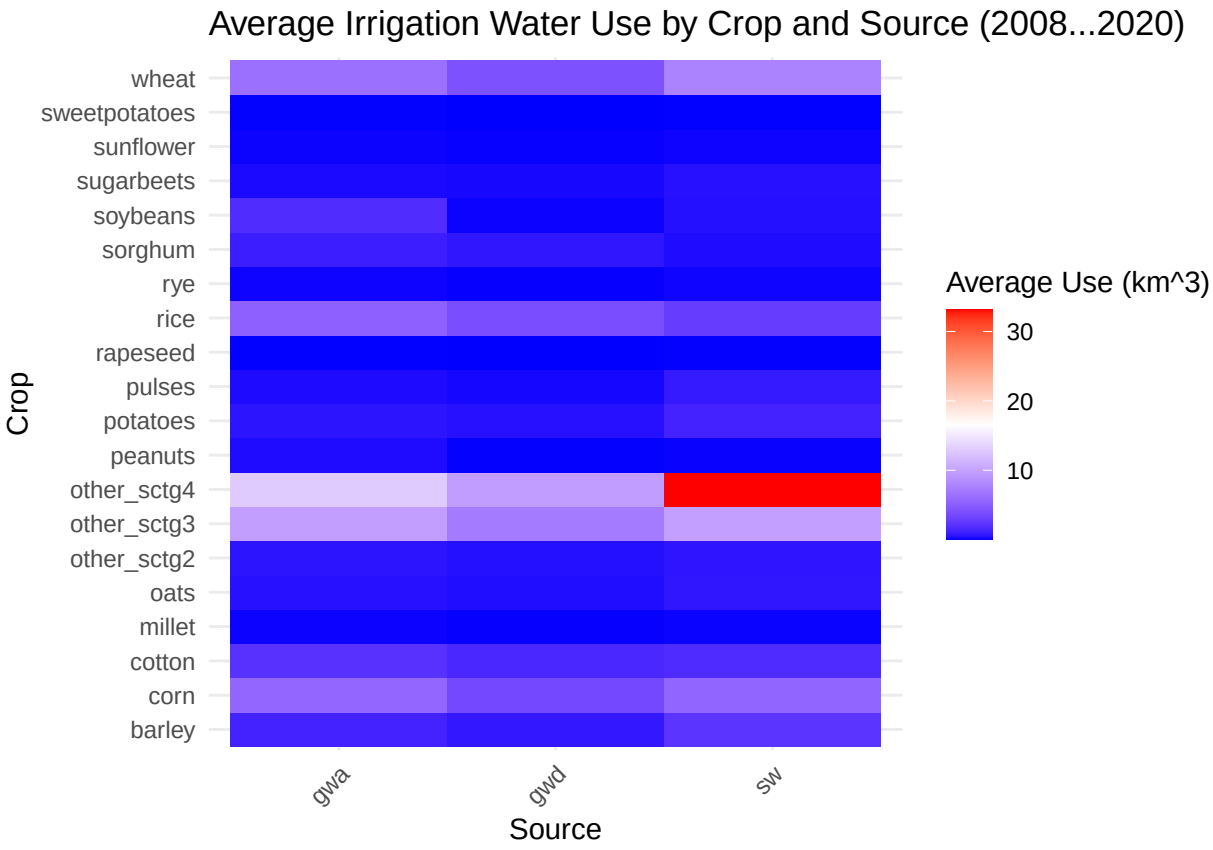
4. Create a tile plot/grid plot of the data in Figure 4.

```
df_melted <- melt(dd, id.vars = c("crop", "src"), measure.vars = "mean")
mean_center <- mean(c(min(df_melted$value), max(df_melted$value)))

ggplot(df_melted, aes(x = src, y = crop, fill = value)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = mean_center) +
  labs(title = "Average Irrigation Water Use by Crop and Source (2008-2020)",
       x = "Source",
       y = "Crop",
       fill = "Average Use (km^3)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
```

```
## conversion failure on 'Average Irrigation Water Use by Crop and Source
## (2008-2020)' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Average Irrigation Water Use by Crop and Source
## (2008-2020)' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Average Irrigation Water Use by Crop and Source
## (2008-2020)' in 'mbcsToSbcs': dot substituted for <93>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Average Irrigation Water Use by Crop and Source
## (2008-2020)' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Average Irrigation Water Use by Crop and Source
## (2008-2020)' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Average Irrigation Water Use by Crop and Source
## (2008-2020)' in 'mbcsToSbcs': dot substituted for <93>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Average Irrigation Water Use by Crop and Source
## (2008-2020)' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Average Irrigation Water Use by Crop and Source
## (2008-2020)' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Average Irrigation Water Use by Crop and Source
## (2008-2020)' in 'mbcsToSbcs': dot substituted for <93>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Average Irrigation Water Use by Crop and Source
## (2008-2020)' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Average Irrigation Water Use by Crop and Source
## (2008-2020)' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Average Irrigation Water Use by Crop and Source
## (2008-2020)' in 'mbcsToSbcs': dot substituted for <93>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Average Irrigation Water Use by Crop and Source
## (2008-2020)' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Average Irrigation Water Use by Crop and Source
## (2008-2020)' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Average Irrigation Water Use by Crop and Source
## (2008-2020)' in 'mbcsToSbcs': dot substituted for <93>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Average Irrigation Water Use by Crop and Source
```

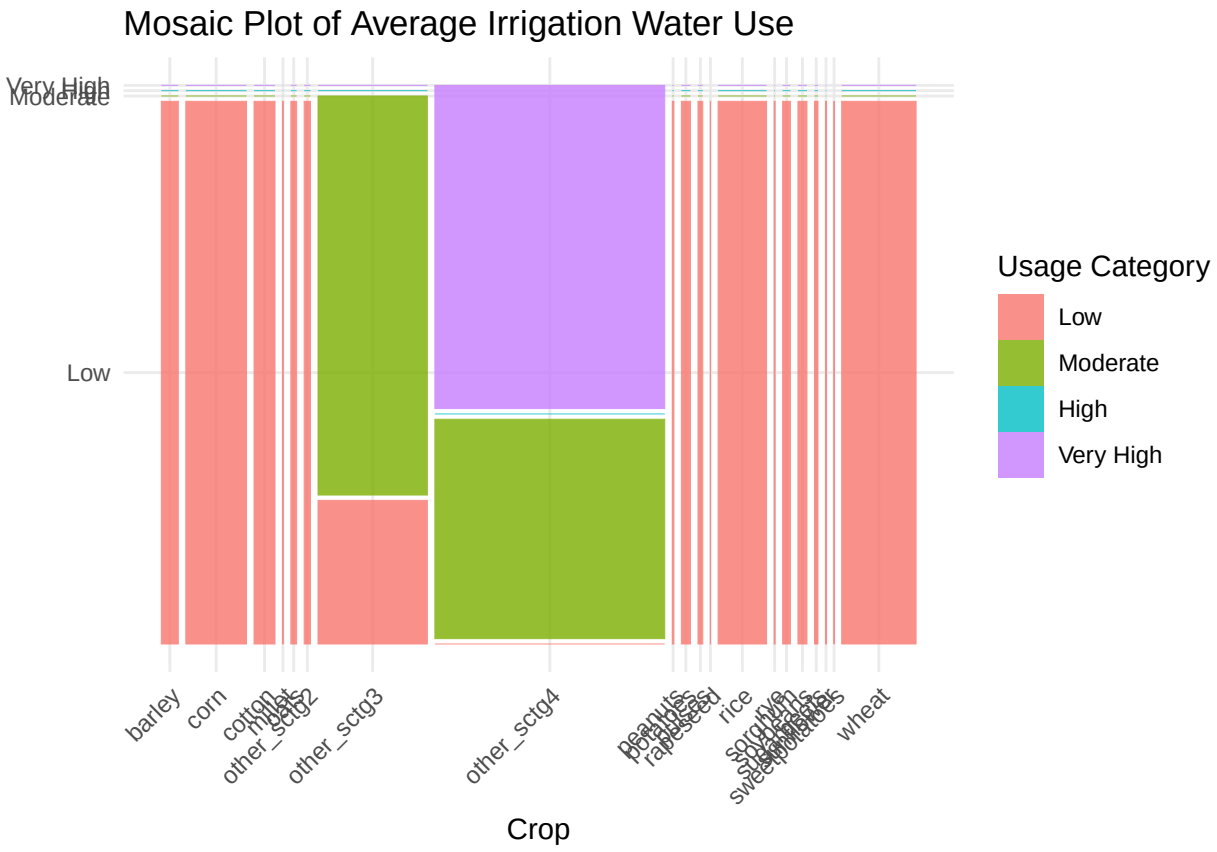
```
## (2008-2020)' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Average Irrigation Water Use by Crop and Source
## (2008-2020)' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Average Irrigation Water Use by Crop and Source
## (2008-2020)' in 'mbcsToSbcs': dot substituted for <93>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Average Irrigation Water Use by Crop and Source
## (2008-2020)' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Average Irrigation Water Use by Crop and Source
## (2008-2020)' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Average Irrigation Water Use by Crop and Source
## (2008-2020)' in 'mbcsToSbcs': dot substituted for <93>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Average Irrigation Water Use by Crop and Source
## (2008-2020)' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Average Irrigation Water Use by Crop and Source
## (2008-2020)' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Average Irrigation Water Use by Crop and Source
## (2008-2020)' in 'mbcsToSbcs': dot substituted for <93>
```



5. Create a mosiac plot of the data in Figure 4.

```
dd$mean_category <- cut(dd$mean, breaks = 4, labels = c("Low", "Moderate", "High", "Very High"))

ggplot(data = dd) +
  geom_mosaic(aes(weight = mean, x = product(crop), fill = mean_category)) +
  labs(title = "Mosaic Plot of Average Irrigation Water Use",
       x = "Crop",
       y = "Mean Water Usage Category",
       fill = "Usage Category") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        axis.title.y = element_blank())
```

```
## Warning: `unite_()` was deprecated in tidyr 1.2.0.
## i Please use `unite()` instead.
## i The deprecated feature was likely used in the ggmosaic package.
##   Please report the issue at <https://github.com/haleyjeppson/ggmosaic>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



6. What are the benefits (other than it fits on one plot) and drawbacks of these two plots?

T

7. Figure 6

Figure 6 uses a different color scale for each plot. Discuss the benefits and drawbacks of this choice. What was the main purposes of this figure? Given the main purpose, would you recommend using the same color scale, or different color scales, for each plot?

Mosaic plots excel in showing proportional relationships and comparisons between categorical data, but they can oversimplify continuous data like average water usage, leading to potential interpretative challenges, especially in crowded plots. Tile plots are more suited for visualizing continuous data, using color gradients to intuitively convey variations in values like water usage across categories. However, tile plots can sometimes lose finer details and depend heavily on color perception, which might be challenging for color vision-deficient viewers. In the context of irrigation water use by crop and source, a tile plot is more appropriate due to its effectiveness in displaying continuous variables.

8. Figure 8

Figure 8 also uses a different color scale for each plot. Discuss the benefits and drawbacks of this choice. What was the main purposes of this figure? Given the main purpose, would you recommend using the same color scale, or different color scales, for each plot?

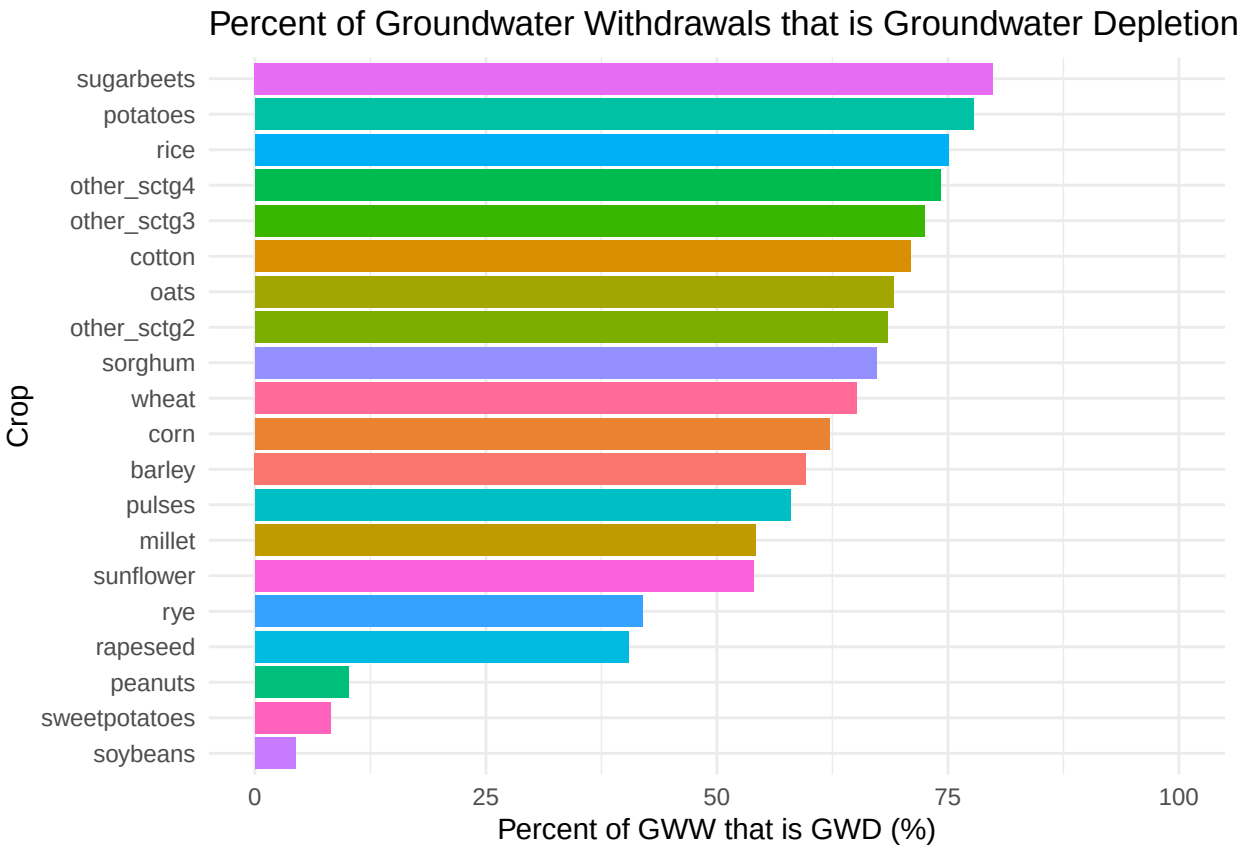
The purpose of figure 8 is to illustrate the spatial differences between the irrigation water use estimates obtained from the PCR-GLOBWB 2 model and the values reported by the U.S. Geological Survey. The use of different color scales enhances the detail and contrast within each dataset, allowing for a tailored interpretation

that emphasizes specific ranges and nuances. However, this approach can hinder direct comparability across plots, as variations in scales may lead to misunderstandings or inaccurate assessments of relative differences. The inconsistency in visual representation can also add complexity to the interpretation process, potentially causing confusion for the reader. In the context of comparing model estimates to USGS data across different water sources and years, a consistent color scale might be more beneficial for easy comparison and coherent understanding. Given the context and purpose of the figure, if the intention is to allow readers to quickly gauge and compare the magnitude of discrepancies between model estimates and USGS data across different water sources and years, a consistent color scale across all plots would be more beneficial.

9. Breakdown of GWW

The paper notes in Section 3.1 that $GWW = GWW_{sustainable} + GWW_{unsustainable}$, and that $GWD = GWW_{unsustainable}$. Create a visualization showing the percent of GWW that is GWD for each crop. Use the mean values for water usage.

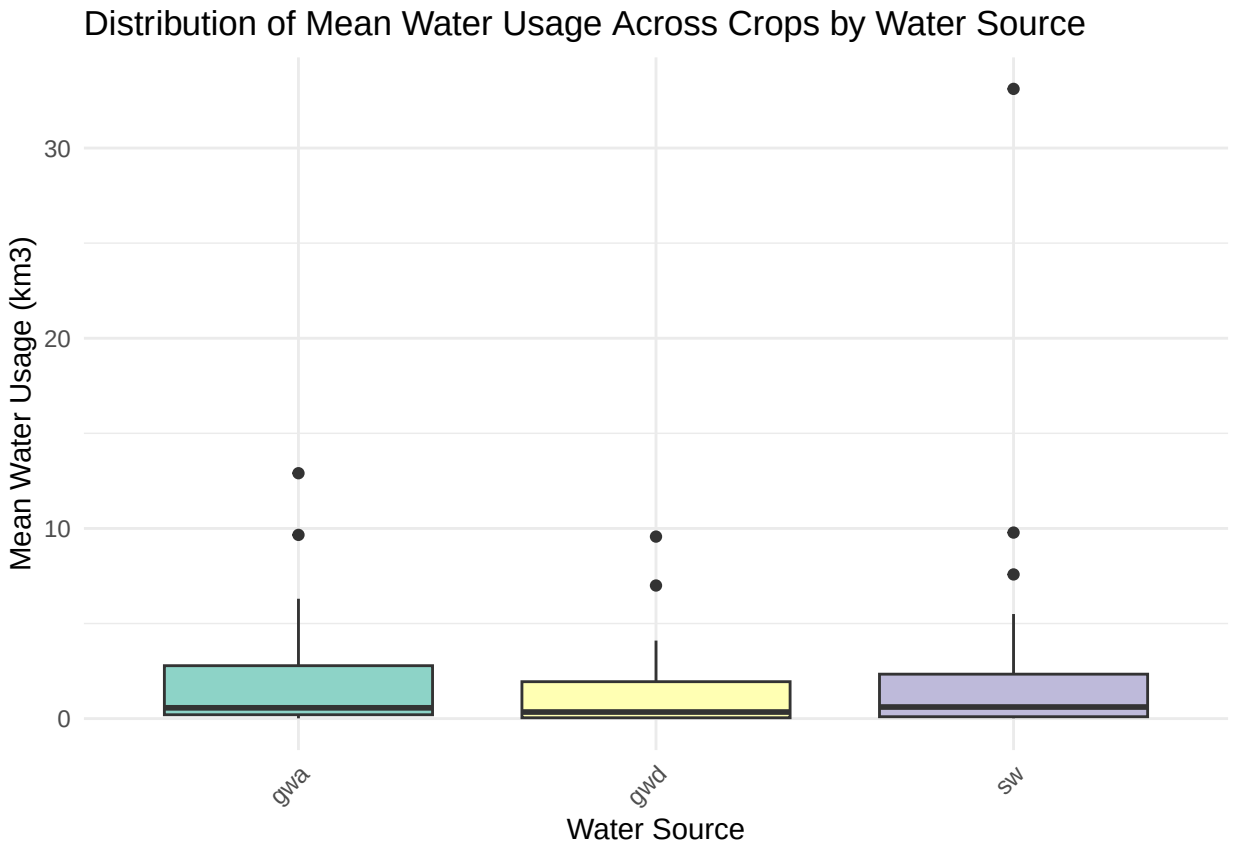
```
data_gww <- dd %>% filter(src == "gwa") %>% select(crop, mean) %>% rename(GWW = mean)
data_gwd <- dd %>% filter(src == "gwd") %>% select(crop, mean) %>% rename(GWD = mean)

merged_data <- merge(data_gww, data_gwd, by = "crop")

# Calculating the percent of GWW that is GWD for each crop
merged_data$Percent_GWD_of_GWW <- (merged_data$GWD / merged_data$GWW) * 100

ggplot(merged_data, aes(x = reorder(crop, Percent_GWD_of_GWW), y = Percent_GWD_of_GWW, fill = crop)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  ylim(0,100) +
  labs(title = "Percent of Groundwater Withdrawals that is Groundwater Depletion by Crop",
       x = "Crop",
       y = "Percent of GWW that is GWD (%)") +
  theme_minimal() +
  theme(legend.position = "none")
```



10. Custom visualization

What is another question you have about this data? Create a visualization that attempt to answer your question. What is the distribution of mean water usage across crops for each source? What are the ranges, median values? Are there outliers?

```
ggplot(dd, aes(x = src, y = mean, fill = src)) +
  geom_boxplot() +
  labs(title = "Distribution of Mean Water Usage Across Crops by Water Source",
       x = "Water Source",
       y = "Mean Water Usage (km3)") +
  scale_fill_brewer(palette = "Set3", name = "Water Source") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        legend.position = "none")
```


Distribution of Mean Water Usage Across Crops by Water Source
Mean Water Usage (km3)
30
20
10
0
gwa
gwd
sw
Water Source