

HarvardX PH125.9x Data Science Capstone Choose Your Own Project: Use of Machine Learning to Predict Monthly Residential Water Demand in California

Dawn Flores

2025-11-17

Section 1: Introduction and Overview

This project has been completed as part of the HarvardX PH125.9x Data Science course series. Forecasting water use (also referred to as water demand) is becoming critical as a water supply management tool as urban areas grow and face challenges such as climate change causing water supplies to be less reliable. In the state of California in the United States (US), water management is particularly challenging due to the varying climates and availability of water across the state where the water is used across various sectors including urban use, agricultural use, and environmental use. Urban water use can be divided into a number of categories such as residential, commercial, institutional, industrial, irrigation, and other uses. Per the California State Water Resources Control Board, residential water use can be affected by numerous factors, including: rainfall, temperature, evaporation rates, population growth, population density, socio-economic measures, and water prices (State Water Resources Control Board, 2015). This analysis explores residential water demand and supply patterns across California using the CaRDS dataset (Gross et al, 2024a) and readily available supplementary socioeconomic and climate data. The goal is to understand factors influencing water use, normalize demand metrics, and evaluate predictive models for estimating residential water demand.

The following key steps were used to accomplish this:

1. Download, extract and apply basic formatting to dataset
2. Review dataset
3. Develop rating prediction models
4. Select a prediction model
5. Generate a set of rating predictions for the final holdout dataset

Section 2: Methods and Analysis

The following sections explain the process and techniques used in the analysis, including data import and formatting, data exploration, and the modeling approach and analysis. The analysis was completed using R code in the RStudio software (version 2025.05.0+496 “Mariposa Orchid” Release for windows).

2.1 Data Import and Formatting

The following datasets were used for this analysis:

- California Residential Water Demand and Supply (CaRDS) open dataset, which includes separate files with water use and supply data, and public water supplier data (Gross et al, 2024a)

- 2022 Risk Assessment Results for the California State Water Resources Control Board (SWRCB) Safe and Affordable Funding for Equity and Resilience (SAFER) program (SWRCB, 2022)

The CaRDS dataset contains nine years (2013 to 2021) of monthly water supply and demand time series for 404 water suppliers in California, USA, precipitation data and temperature data. This data was compiled from different open-access data sources, including the SWRCB Electronic Annual Reports (eAR), the PRISM Climate Data from the PRISM Climate Group, and the Climate Division Data (ClimDiv) from the National Oceanic and Atmospheric Administration (Gross et al, 2024b). The primary CaRDS dataset is structured with columns for public water system identification number (PWSID), Variable, and columns representing years and months with the data for each variable structured as xYYYY.MM.01. Variables include demand (measured in liters), supply (measured in liters), temperature (degrees C), precipitation (millimeters (mm)), and PDSI (unitless, stands for Palmer Drought Severity Index). The creators of the CaRDS dataset processed the data by ensuring that key information was present, including PWSID, measurement units, records for every year, no missing values, and reporting of both supply and demand data.

The secondary dataset, supplier_info, contains the following fields: PWSID, supplier name, zip code, whether the supplier is a large water system or small water system (size), county, hydrologic region, climate zone, 2021 population, and PDSI division.

These datasets, which were previously downloaded, were loaded and examined using the below code:

```
# Load CaRDS data: Monthly water supply and demand time series for 404 water suppliers
# in California (2013-2021)
water_data <- read.csv("CaRDS.csv")
supplier_data <- read.csv("Supplier_Info.csv")

dim(water_data)
head(water_data)
unique(water_data$PWSID)
unique(water_data$Variable)

dim(supplier_data)
head(supplier_data)
```

This dataset was reshaped to a format that can be readily used in machine learning using the code shown below, and the CaRDS data set joined to the supplier information table.

```
# Reformat water data so that variables are columns and dated records are rows
water_data_long <- pivot_longer(water_data, "X2013.01.01":"X2021.12.01")
Water_data_wide <- pivot_wider(water_data_long, names_from = Variable,
                                values_from = value)

# Adjust date column to remove "X" and day
Water_data_wide <- Water_data_wide |>
  mutate(name = str_remove(name, "^\u00d7")) |> # Remove the leading "X"
  separate(name, into = c("year", "month", "day"), sep = "\u00d7.")

# Recreate date column that is recognized by R as a date
Water_data_wide$date <- as.Date(
  paste(Water_data_wide$year, Water_data_wide$month, "01", sep = "-"))

# Join supplier data to demand data
water_data_supplier <- left_join(Water_data_wide, supplier_data, by = "PWSID")
```

Demand was normalized to account for large differences in demand by water system and for varying days per month. The below code was used to create a new column for residential demand measured in gallons per capita per day (gpcd), which is a common measure of residential demand in the United States. Water

systems that were missing population or with negative population were removed. For the purposes of this analysis, only large water systems (LWS) were included.

```
# Create column for normalized demand (gallons per capita per day)
water_data_supplier <- water_data_supplier |>
  mutate(days_in_month = days_in_month(ymd(Date)),
    gpcd = (demand * 0.264172) / (Population_21 * days_in_month)
  ) |> ungroup()

# Filter out rows with negative or 0 demand or population values, and include
# only large water systems (size is LWS)
data_filtered <- water_data_supplier |>
  filter(demand > 0, Population_21 > 0, size == "LWS")

data_filtered <- data_filtered |>
  mutate(gpcd = as.numeric(gpcd)) # Ensure gpcd is numeric

min(data_filtered$gpcd)
max(data_filtered$gpcd)
```

The resulting gpcd column has a range of 5.9344e-06 to 51878519, which is outside the expected range of gpcd values. In California, gpcd values for water systems typically range from 40 gpcd to 500 gpcd (Heberger, 2014). GPCD values outside of this range were filtered out using the code below. Per the data documentation provided for the CaRDS dataset, a small percentage of demand outliers are possible due to misreporting of data by water suppliers to the SWRCB (Gross et al, 2024b).

```
# Filter out gpcd values outside of lowest and highest residential gpcds
# identified by the Pacific Institute (rounded to the nearest 10 gpcd)
# https://pacinst.org/new-data-show-residential-per-capita-water-use-across-california/
lower_bound <- 40
upper_bound <- 590

data_filtered_clean <- data_filtered |>
  filter(gpcd >= lower_bound & gpcd <= upper_bound)
max(as.numeric(data_filtered_clean$gpcd), na.rm = TRUE)
min(as.numeric(data_filtered_clean$gpcd), na.rm = TRUE)
```

Next, median household income (MHI) for each water system was loaded from the SWRCB's 2022 SAFER data workbook. The workbook, in Microsoft Excel format, contains several spreadsheets of data. For the purposes of this project, the XXXX sheet was imported then filtered to include only PWSID and MHI using the code below. For cases where the MHI was missing, the average California MHI, \$78,672, was used (United States Census Bureau, 2020).

```
# Load California State Water Resources Control Board (SWRCB) 2022 SAFER dataset
SAFER <- read.xlsx("2022risk.xlsx", sheet = "Afford. Raw Data Summary (1)",
  startRow = 2)

# Filter SAFER file for median household income (MHI).
MHI <- SAFER |> select(PWSID, Weighted.Average.MHI2) |>
  rename(MHI = Weighted.Average.MHI2)

# Join MHI dataset to the data_filtered_clean dataset. Replace missing values
# with California statewide MHI of $78,672
# (US Census, 2020 American Community Survey 5-Year Estimates)
data_filtered_clean <- left_join(data_filtered_clean, MHI, by = "PWSID")
```

```

data_filtered_clean <- data_filtered_clean |>
  mutate(MHI = na_if(MHI, "N/A")) |>
  mutate(MHI = as.numeric(MHI)) |>
  mutate(MHI = replace_na(MHI, 78672))

```

Finally, during the COVID pandemic, the State of California issued stay at home orders that were in place from March 2020 to June 2021 (Executive Department State of California, 2021). This order resulted in residential water use increasing as much of the population in the state was required to stay at home when businesses were closed. For the period of time this order was in place, a factor of 1 was assigned while all other months were assigned a 0, as shown in the code below.

```

# Incorporate the State of California COVID stay at home order period as a factor
# (March 2020 to June 2021)
# https://www.gov.ca.gov/wp-content/uploads/2021/06/6.11.21-EO-N-07-21-signed.pdf
data_filtered_clean <- data_filtered_clean |>
  group_by(year, month) |>
  mutate(COVID = if_else(
    (year == 2020 & month >= 3) | (year == 2021 & month <= 6),
    1, 0
  )) |>
  ungroup()

```

Once the dataset was compiled as described above, columns that were expected to be used in the analysis were converted from characters to either factors or numerical data types to facilitate the analysis, and was applied to the following columns using the below code: Hydrologic.Region, Climate.Zone, County, and month. In addition, given that precipitation can be extremely variable across the state, a new column of log transformed precipitation was created to better manage extreme precipitation values.

```

# Convert Hydrologic.Region to a factor.
data_filtered_clean$Hydrologic.Region <- as.factor(
  data_filtered_clean$Hydrologic.Region)

# Convert Climate.Zone to a factor.
data_filtered_clean$Climate.Zone <- as.factor(
  data_filtered_clean$Climate.Zone)

# Convert County to a factor.
data_filtered_clean$County <- as.factor(
  data_filtered_clean$County)

# Convert month to a factor.
data_filtered_clean$Month <- as.factor(
  data_filtered_clean$Month)

# Transform precipitation to reduce potential impacts of extreme precipitation periods
data_filtered_clean$precipitation_log <- log(data_filtered_clean$precipitation+1)

```

The resulting data set contains monthly records from 2013 to 2022 for 165 large water systems, with data for demand, supply, precipitation, temperature, PDSI, zip code, county, hydrologic region and climate zone.

2.2 Data Exploration

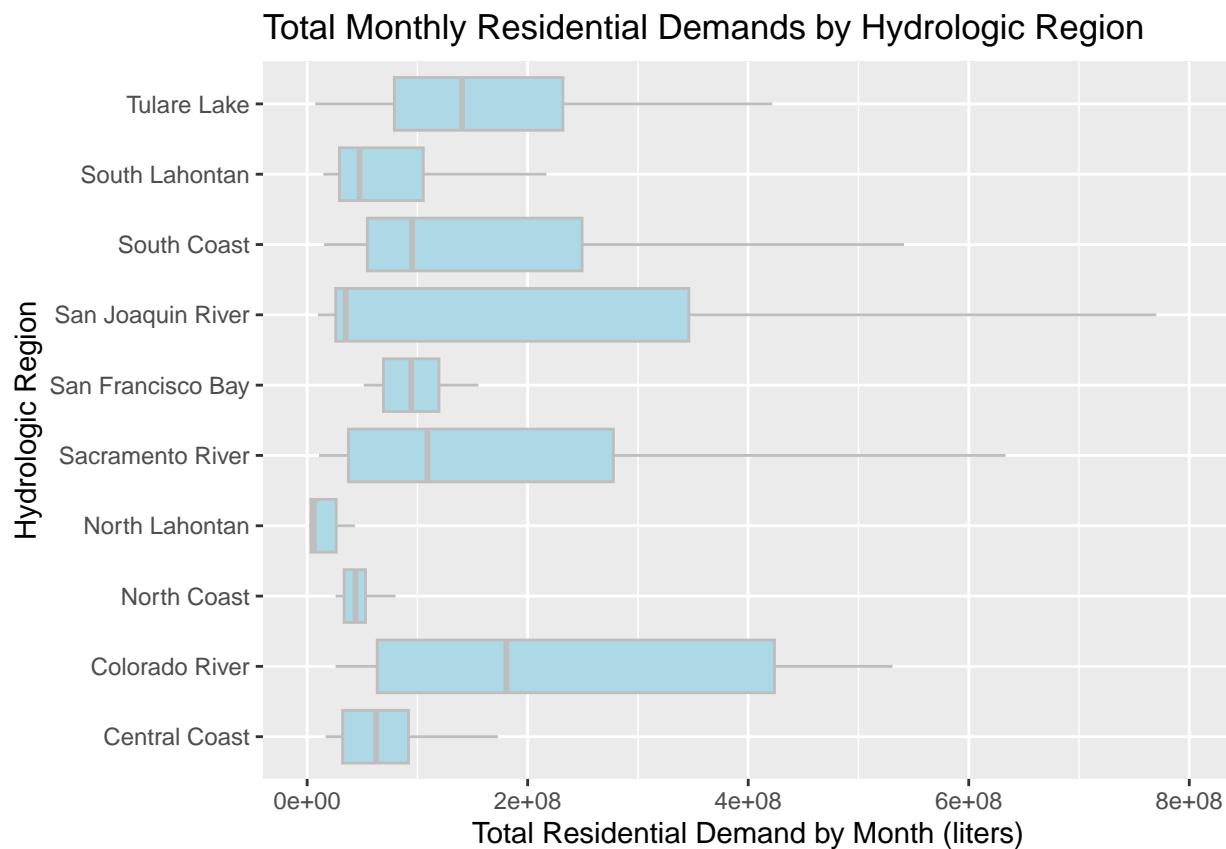
Once the dataset was created as described in Section 2.1, the data was explored to determine potential ranges and patterns.

Demand and Hydrologic Regions

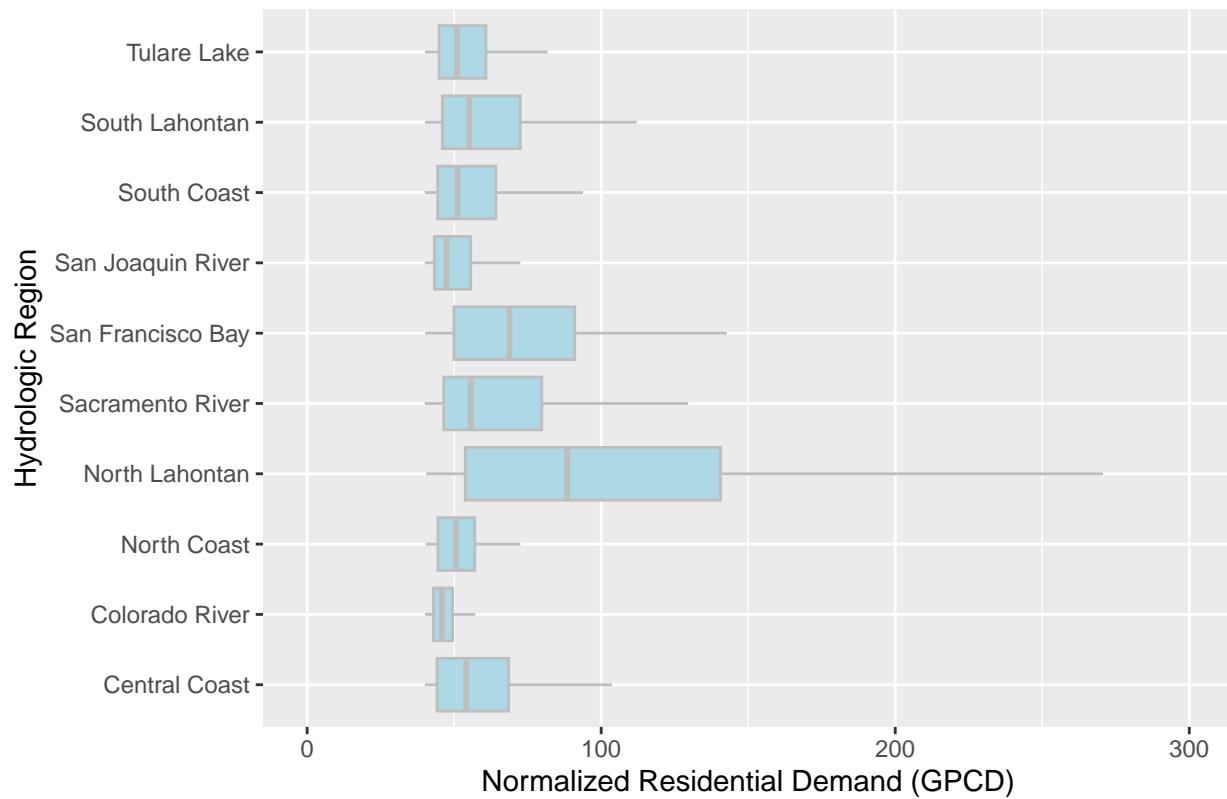
The range of demands within each hydrologic region was examined to visualize whether demand may vary across region. Hydrologic regions are geographic divisions created by the SWRCB based on regional watershed boundaries. There are ten regions named as follows: North Coast, Sacramento River, North Lahontan, San Francisco Bay, San Joaquin River, Central Coast, Tulare Lake, South Lahontan, South coast and Colorado River. A map of these zones along with the number of suppliers per region can be found in Gross et al, 2024b.

As shown in the box plot below titled “Range of Monthly Residential Demands by Hydrologic Region”, total demand varies significantly among hydrologic regions. This is expected given that population varies widely within each region, ranging from 0 liters to nearly 800,000,000 liters per month per water system. The chart titled “Normalized Residential Demand by Hydrologic Region” better reflects patterns of water use across hydrologic region. For example, while the North Lahontan Hydrologic Region has low total demand, normalized demand shows a much higher water use per capita than other regions, while the opposite can be said for the Colorado River Hydrologic Region. Note that to better visualize the interquartile range, outliers are not shown on the charts.

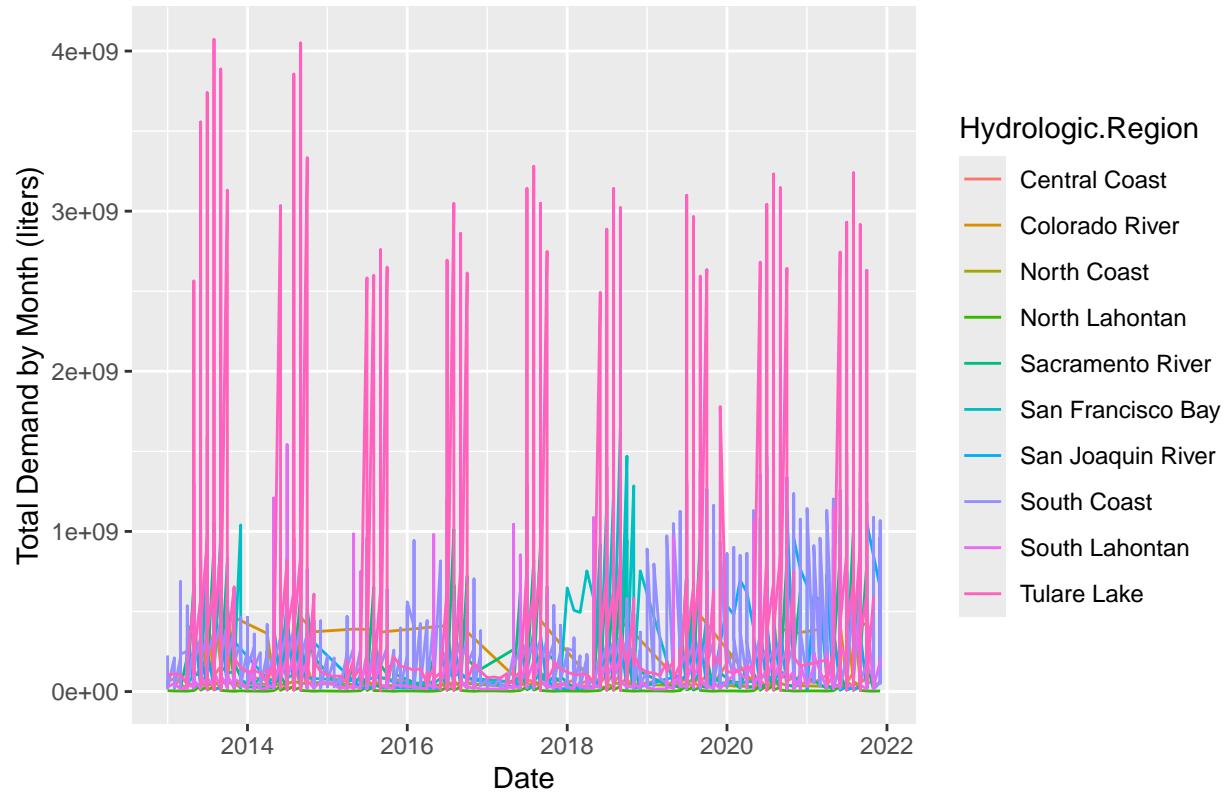
The variation of water demand by month can be seen in the chart “Monthly Residential Demand by Hydrologic Region”. This chart reflects that during certain months of the year, demand is higher. This is likely due to increased temperatures and decreased precipitation during summer months. The variation of annual demand can be seen in “Annual Residential Demand by Hydrologic Region”, which sums demand for each year. This chart shows that there is annual variation in water, which is expected given that precipitation in California can vary from year to year. This chart also reflects increased residential demands in 2020 due to the COVID stay at home order issued in California.



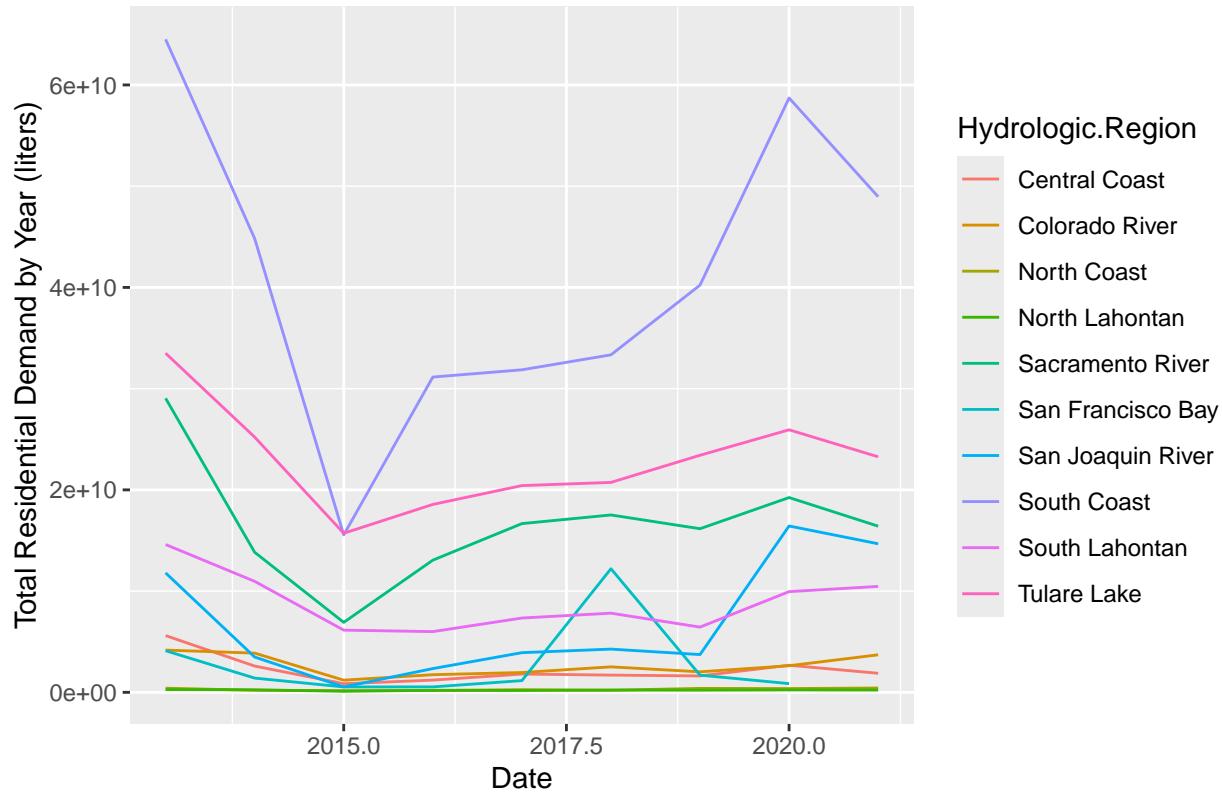
Normalized Residential Demand by Hydrologic Region



Monthly Residential Demand by Hydrologic Region



Annual Residential Demand by Hydrologic Region



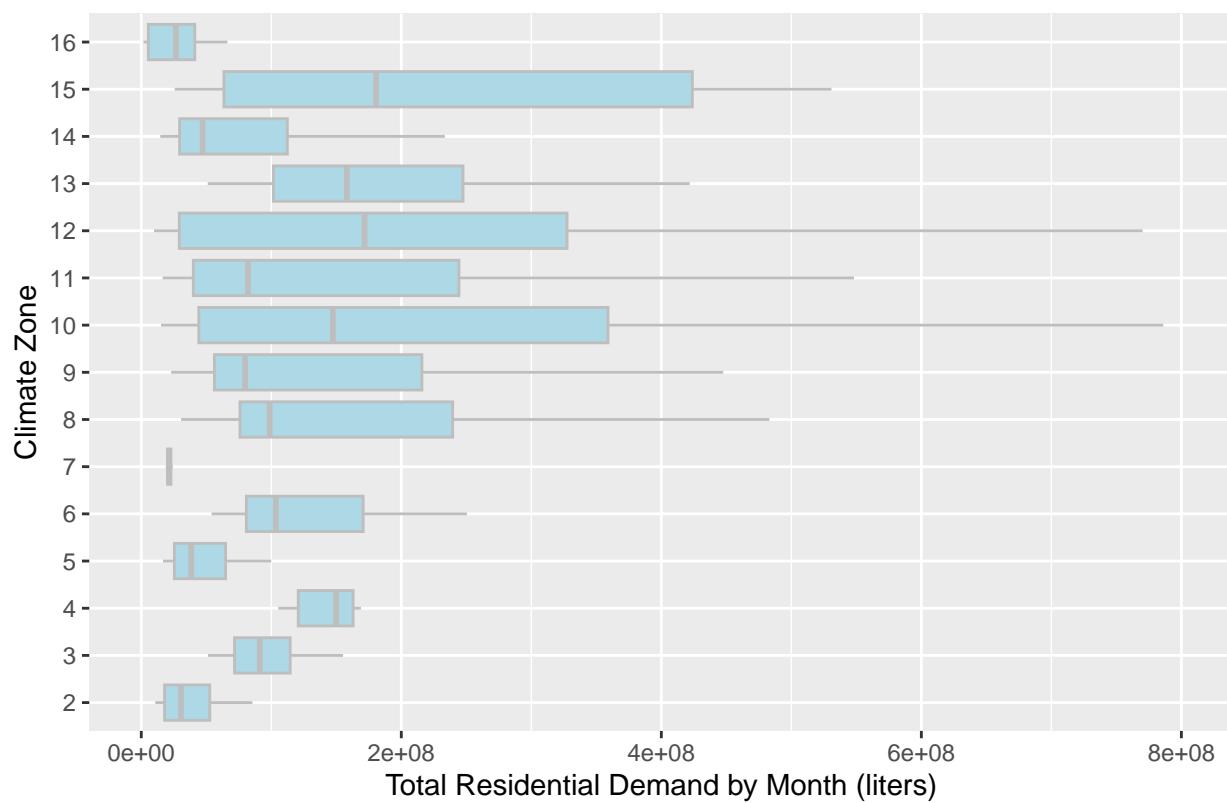
Demand and Climate Zones

The range of demands within each climate zone was also examined. Climate zones are geographic divisions created by the California Energy Commission and are “geographical regions that are defined by unique weather patterns, average temperatures, and energy demands” (Stillitano, 2025). While there is some similarity between hydrologic regions and climate zones, the 16 climate zones have somewhat different boundaries. A map of climate zones is provided in Gross et al, 2024b.

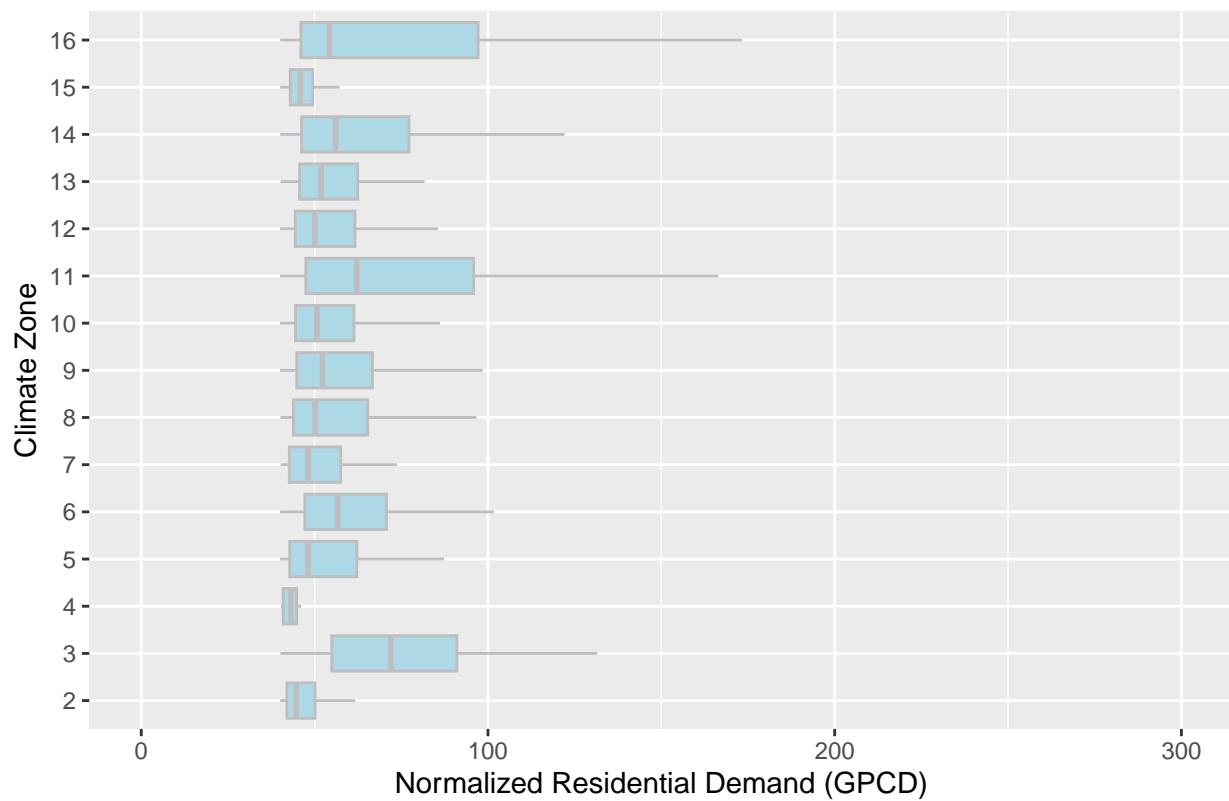
Similar to demand by hydrologic region, there is significant variation in total demand among climate zones, as shown in the box plot charts titled “Total Monthly Residential Demands by Climate Zone”. Normalized demand shows more similarity among climate zones as shown in the chart titled “Normalized Residential Demand by Climate Zone”. Note that outliers have been removed to better show the interquartile range.

The charts titled “Monthly Residential Demand by Climate Zone” and “Annual Residential Demand by Climate Zone” show the variation among months and years, similar to the patterns seen by hydrologic region.

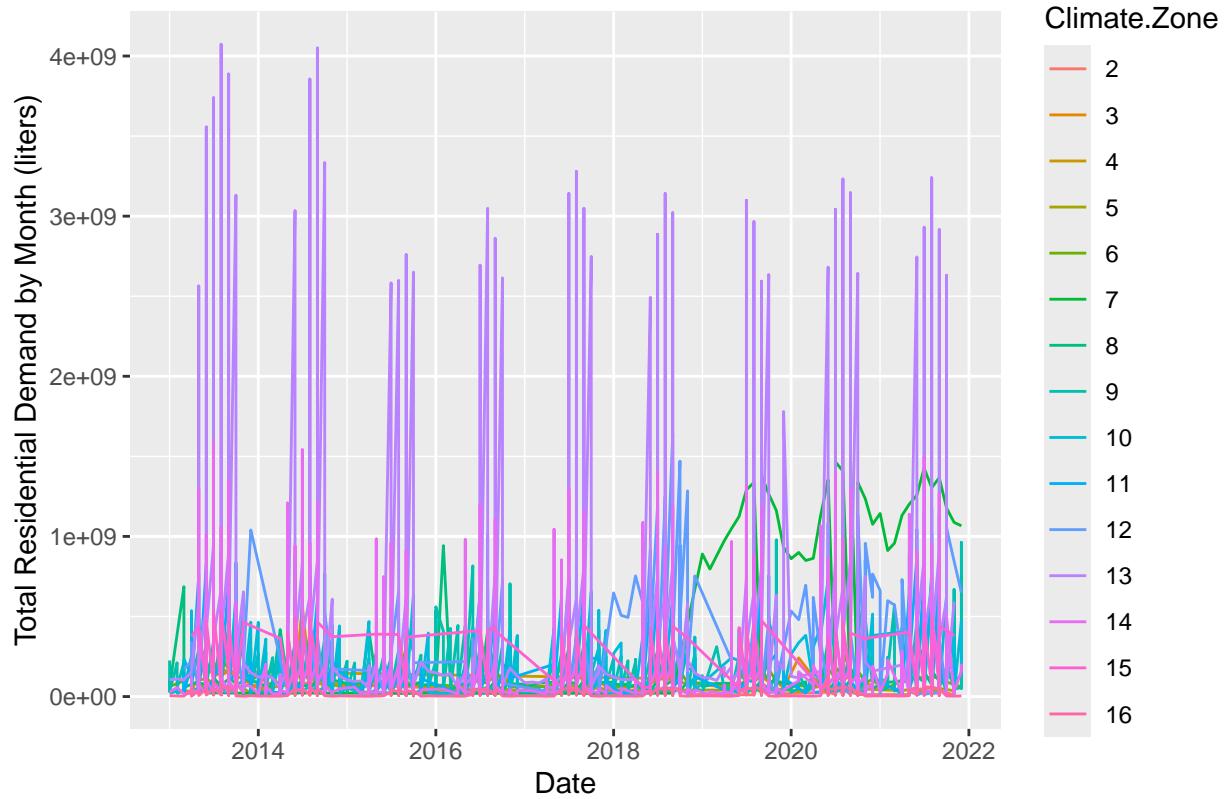
Total Monthly Residential Demands by Climate Zone



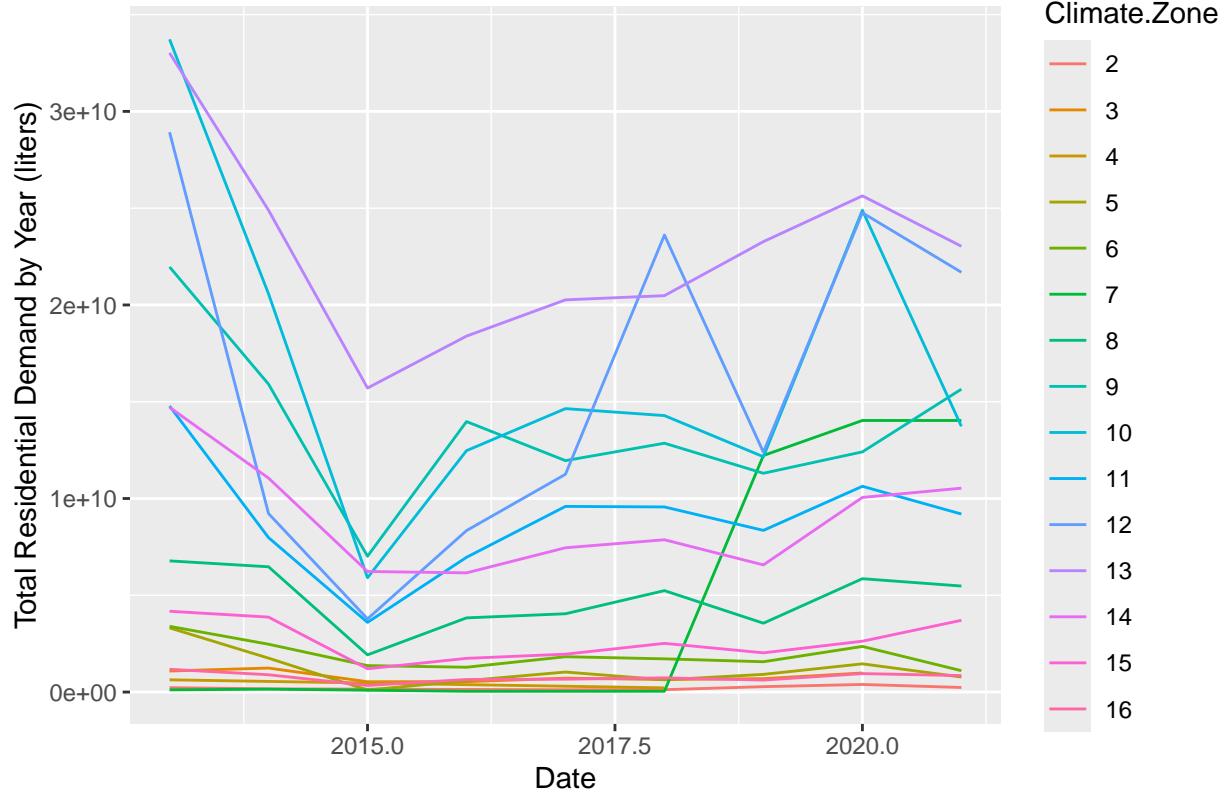
Normalized Residential Demand by Climate Zone



Monthly Residential Demand by Climate Zone



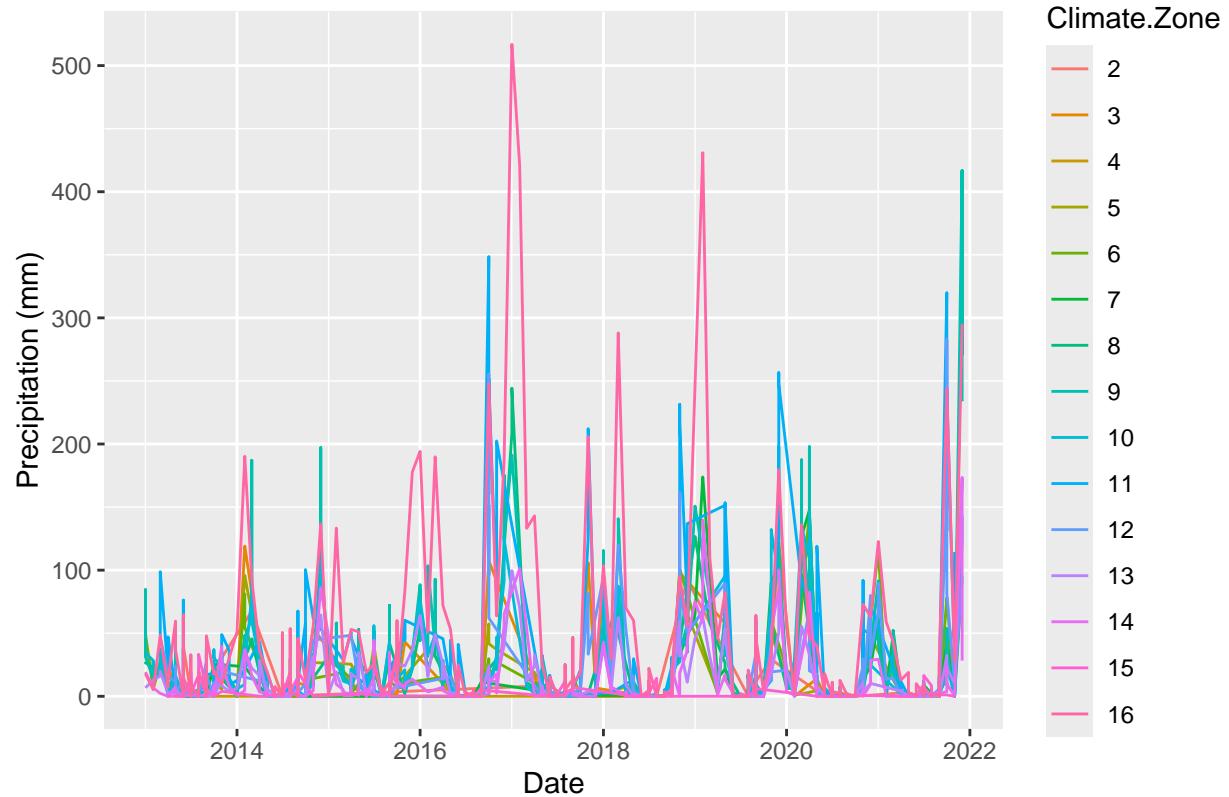
Annual Residential Demand by Climate Zone



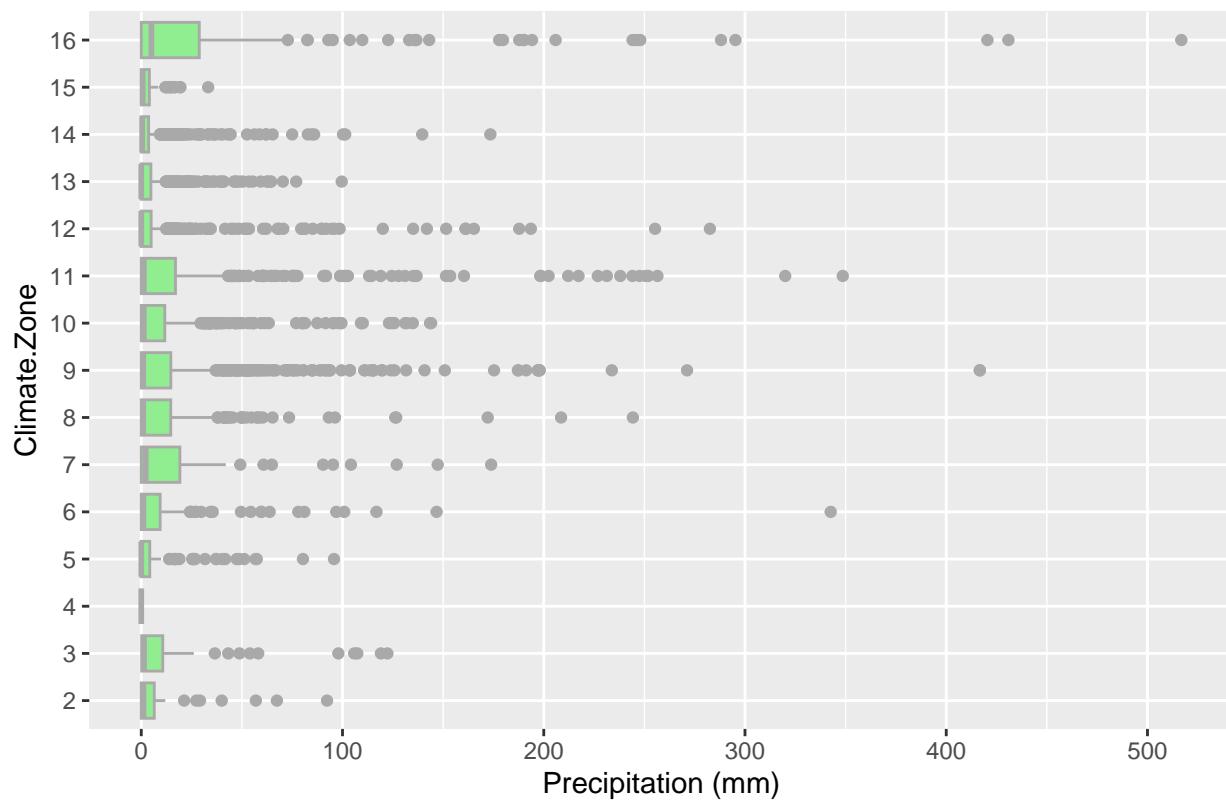
Climate Factors

As described previously, the dataset contains multiple climate factors, including precipitation, temperature and PDSI. Charts were created to show the variation of these climate factors by climate zone. The chart titled “Precipitation by Climate Zone Time Series” shows the monthly total of precipitation by climate zone across the time series, and reflects how the monthly and annual variability of precipitation. The boxplot titled “Precipitation by Climate Zone” reflects the variation of precipitation by geography, showing that some areas receive significantly more precipitation than other areas.

Precipitation by Climate Zone Time Series

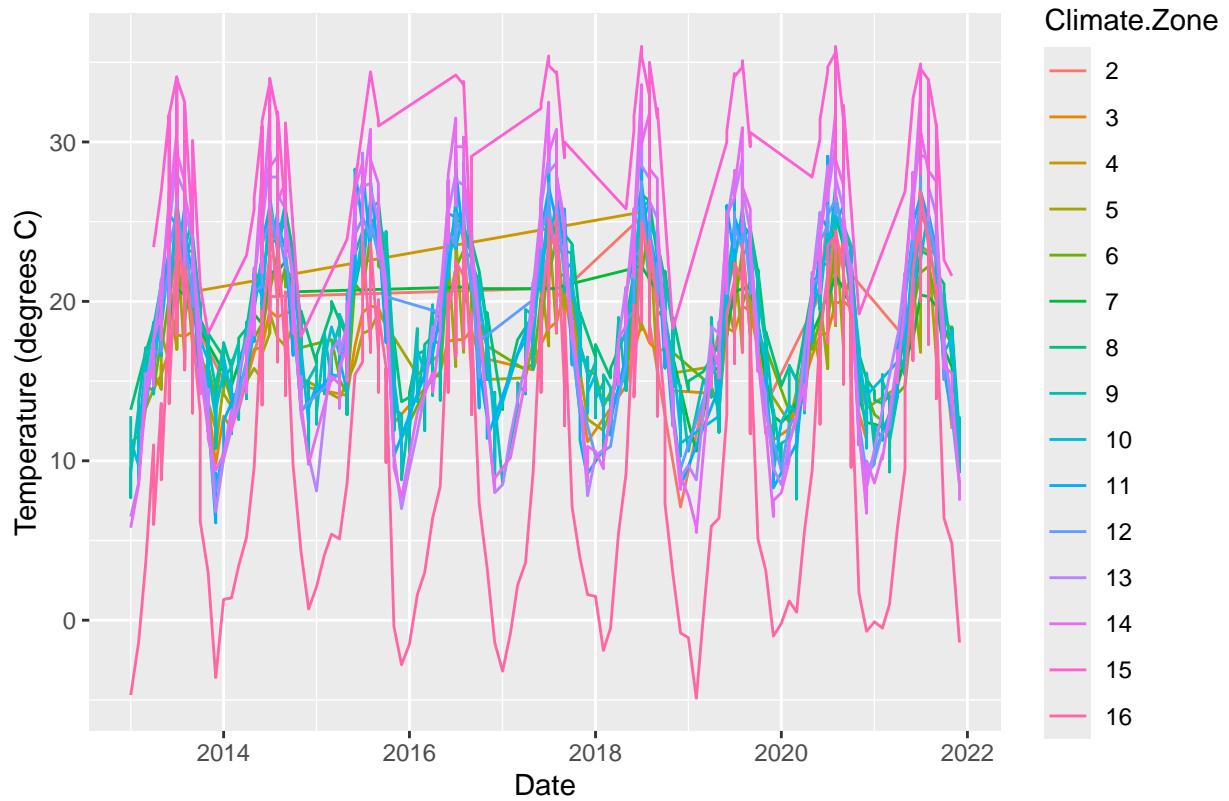


Precipitation by Climate Zone

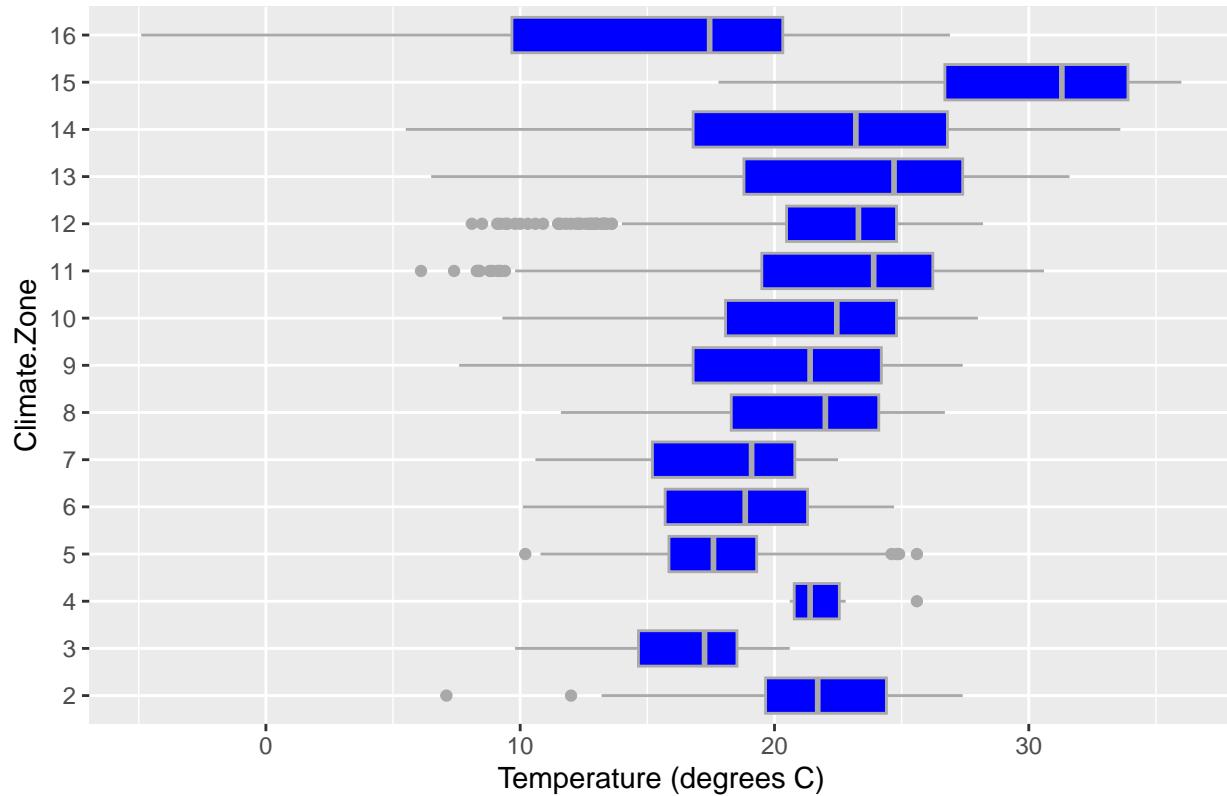


Temperature was similarly examined by time series in the chart titled “Temperature by Climate Zone Time Series”, which shows the monthly average temperature by climate zone across the time series. This chart reflects the monthly variability of temperature, as well as the significant variability among climate zones. The boxplot titled “Temperature by climate zone reflects the variation of precipitation by geography, showing that some areas receive significantly more precipitation than other areas.

Temperature by Climate Zone Time Series

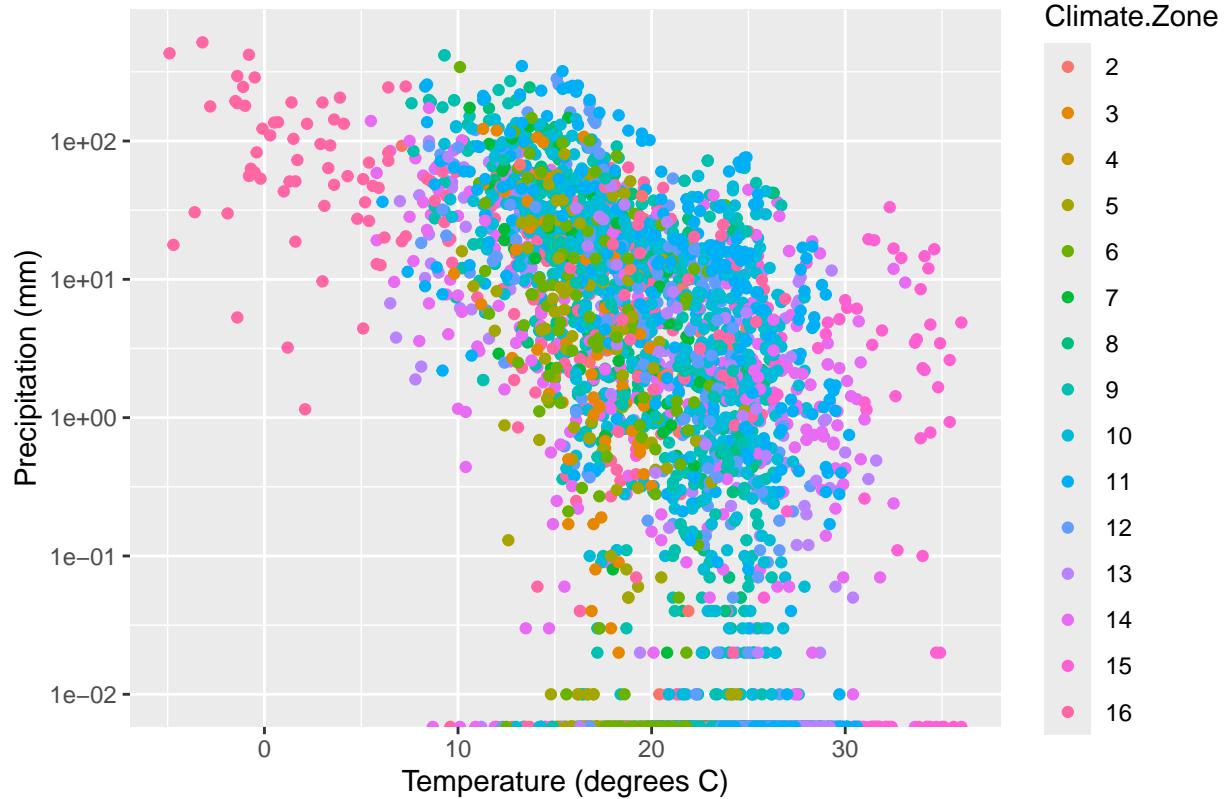


Temperature by Climate Zone



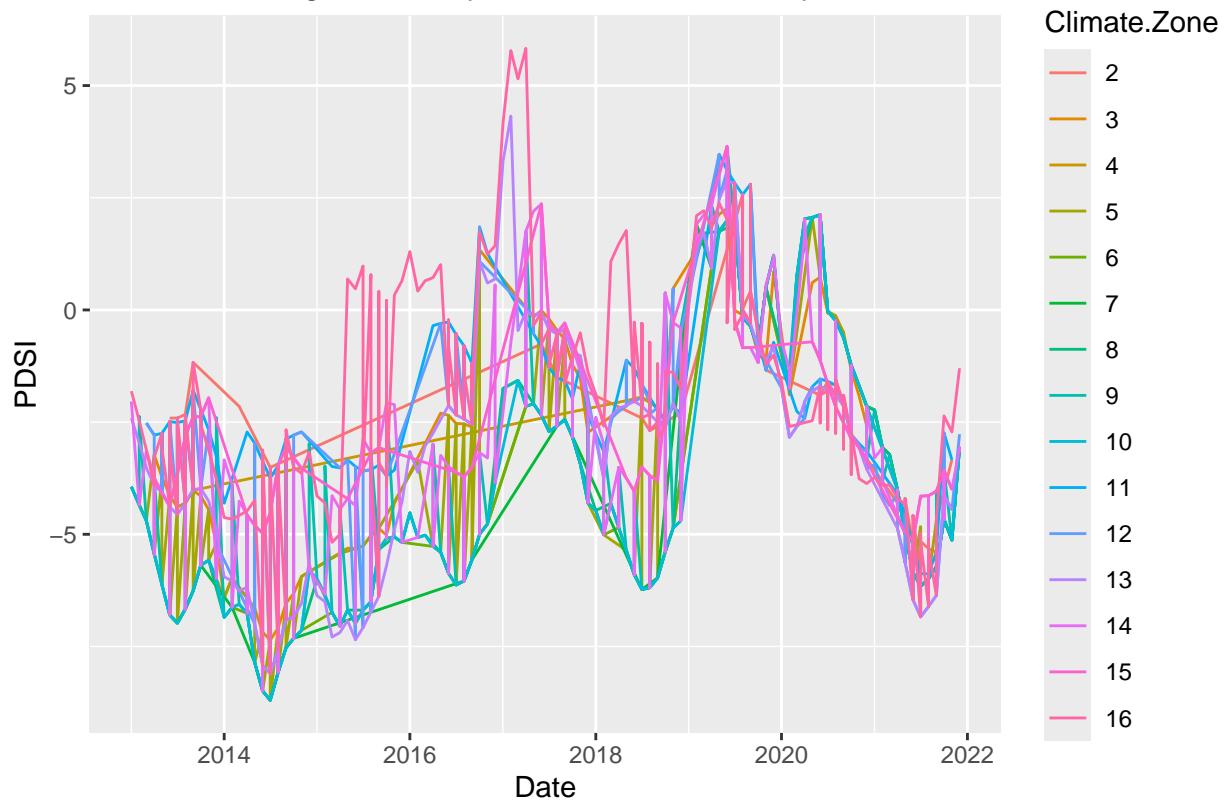
Because precipitation and temperature are both seasonal, temperature was plotted against precipitation (shown below). Although there is a slight correlation between the two, the variability is large enough to warrant including both factors in the machine learning model.

Temperature versus Precipitation by Climate Zone

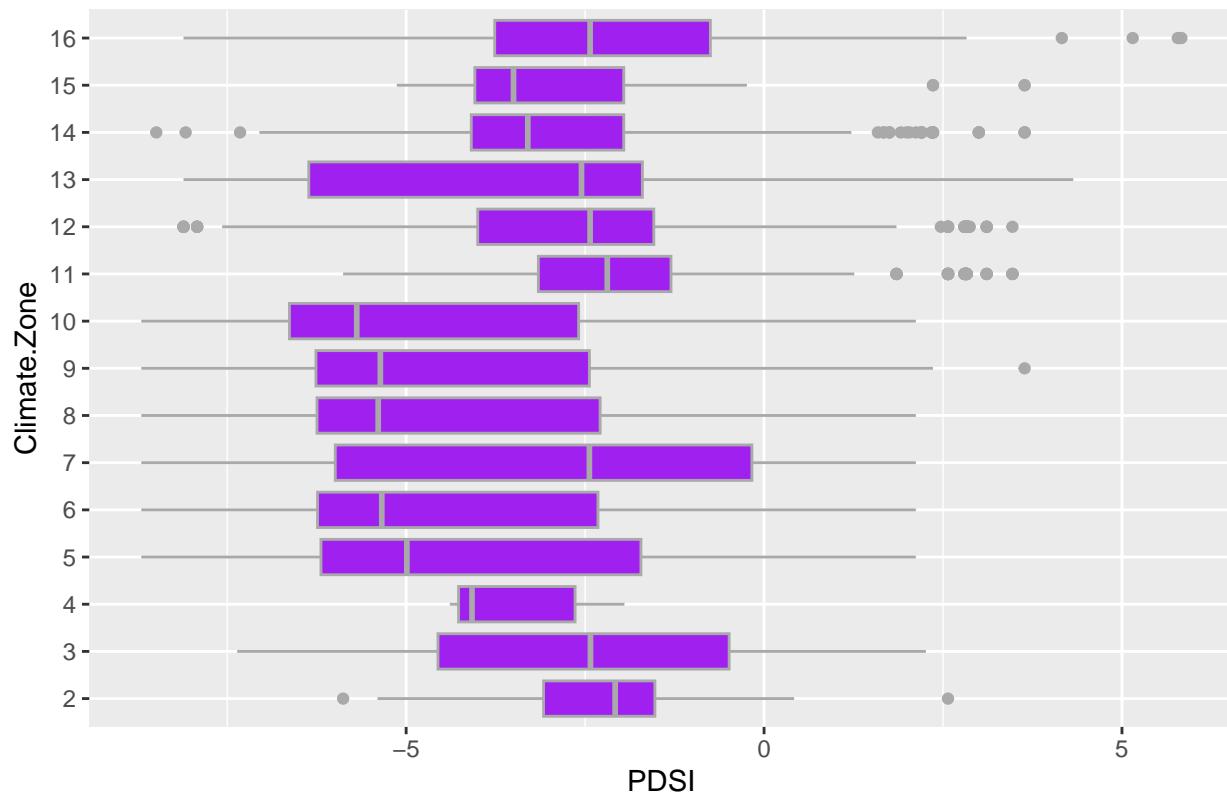


The PDSI was examined by plotting PDSI as a time series and as a box plot by climate zone. PDSI is “a measure to estimate relative dryness and it is very effective in accounting for long-term drought conditions, taking the basic effects of global warming into account. The PDSI is provided by NOAA and calculated for large areas, roughly following the division of hydrologic regions” (Gross et al, 2024b). As shown in the time series, the PDSI follows a similar pattern across all climate zones, though varies in magnitude across climate zones. The boxplot shows the variation of PDSI across climate zones, showing that some zones have a higher magnitude of drought severity than other zones. Given that PDSI is a long term measure of drought, it may not be a suitable measure of monthly demand and therefore is not included in the machine learning model discussed in this project.

Palmder Drought Severity Index Time Series by Climate Zone



Palmder Drought Severity Index by Climate Zone



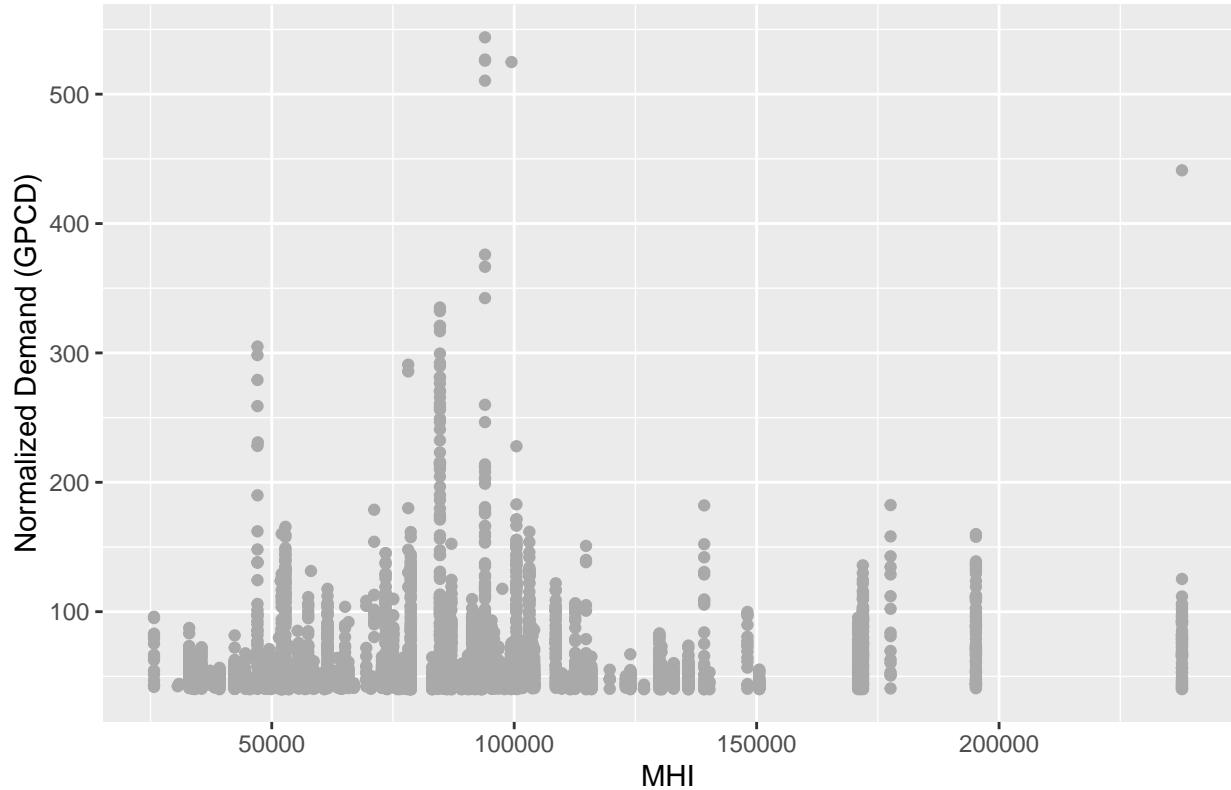
Months

Normalized demand (GPCD) by month was examined to see whether GPCD may noticeably change over the course of a year. As shown in the below boxplot titled “Normalized Demand by Month”, there is some variation in GPCD from month to month, indicating that month could account for some variation in demand.

Median Household Income

Median household income is a socioeconomic factor that could help to explain some variation in water use. MHI was plotted against normalized demand in the below chart to see whether any patterns emerge. Based on the chart, it is difficult to tell what type of impact MHI might have on GPCD, therefore it will be examined further through ANOVA testing to determine its potential for inclusion in the machine learning model.

Normalized Demand versus MHI



Divide Dataset into Training and Test Sets

To further analyze the data through ANOVA analysis, the dataset set was split into training and test sets using the below code which splits 80% of the records into the training set and 20% of the records into the test set. A seed of 1999 was set to allow for reproducibility of results.

```
set.seed(1999, sample.kind="Rounding")
test_index <- createDataPartition(
  y = data_filtered_clean$demand, times = 1, p = 0.2, list = FALSE)
train_set <- data_filtered_clean[-test_index,]
test_set <- data_filtered_clean[test_index,]
```

ANOVA Analysis

A series of one-way ANOVAs were used to assess how each of the above described factors contribute to variations in normalized demand in gpcd for consideration as predictors in the machine learning model using the training data set. The ANOVA results shown in the below table indicate that regional and climatic factors have the strongest influence on water use, with County, Hydrologic Region and Climate Zone showing highly significant effects. Socioeconomic conditions (MHI) and the COVID period also contribute significantly, alongside temporal factors such as Year, Month and Temperature exhibit weaker but still significant effects, suggesting some seasonal and weather-related variability. In contrast, precipitation does not show a significant impact, implying it adds little explanatory power in this model.

Table 1: ANOVA Results

variable	df	sumsq	meansq	f_value	p_value	signif
County	42	1191555.7	28370.4	27.0328	0.00000	***
Hydrologic.Region	9	567371.1	63041.2	51.8463	0.00000	***
Climate Zone	14	354885.3	25348.9	19.8392	0.00000	***
Year	8	49137.7	6142.2	4.5102	0.00002	***
COVID	1	33480.8	33480.8	24.5537	0.00000	***
MHI	1	28175.2	28175.2	20.6401	0.00001	***
Month	11	27235.2	2475.9	1.8083	0.04728	*
Temperature	1	8033.1	8033.1	5.8604	0.01554	*
Precipitation	1	2207.3	2207.3	1.6084	0.20480	

2.3 Modeling Approach and Analysis

Three machine learning models were trained to predict GPCD using the hydrologic, climatic, temporal, socioeconomic, and weather-related variables described above by using the following three methods:

- Generalized linear model (GLM)
- K-nearest neighbors (KNN)
- Random forest

GLM (linear regression)

The generalized linear model (GLM) was trained using the hydrologic, climatic, temporal, socioeconomic, and weather-related predictors described above using the below code.

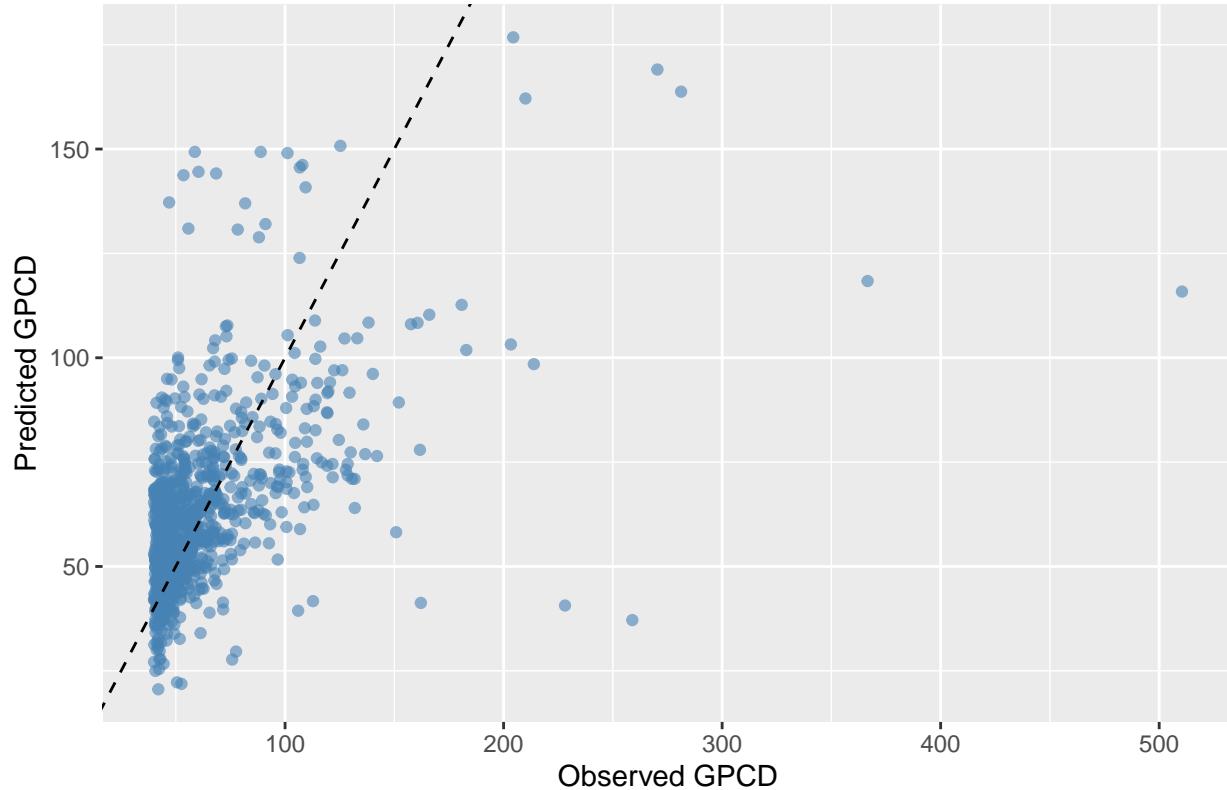
```
# Train and test glm model
train_glm <- train(
  gpcd ~ Hydrologic.Region + Climate.Zone + year + COVID + MHI + month + County + temperature,
  method = "glm",
  data = train_set)

y_hat_glm <- predict(train_glm, test_set, type = "raw")
```

On the test set, the model produced an RMSE of 29.3, an MAE of 16.9, and an R² of 0.26, indicating very low predictive accuracy. This is reflected in the Predicted vs Observed GPCD scatter plot which shows predicted GPCD values from the model against observed values. The black dashed line represents the ideal 1:1 relationship where predictions perfectly match observations. Most points deviate from this line, indicating prediction errors. A warning was generated about a rank-deficient fit suggests multi-collinearity among predictors, which likely contributed to poor performance. Overall, the GLM model does not adequately capture variability in GPCD.

```
##      RMSE Rsquared      MAE
## 29.3126  0.2639 16.8811
```

Predicted vs Observed GPCD (GLM Model)



KNN (k-nearest neighbors)

A k-nearest neighbors regression model was trained to predict GPCD using the same hydrologic, climatic, temporal, socioeconomic, and weather-related variables used for the GLM model.

```
# Train and test knn model
train_knn <- train(
  gpcd ~ Hydrologic.Region + Climate.Zone + year + COVID + MHI + month + County + temperature,
  method = "knn",
  data = train_set,
  tuneGrid = data.frame(k = seq(2, 18, 2)))
train_knn$bestTune

##   k
## 3 6

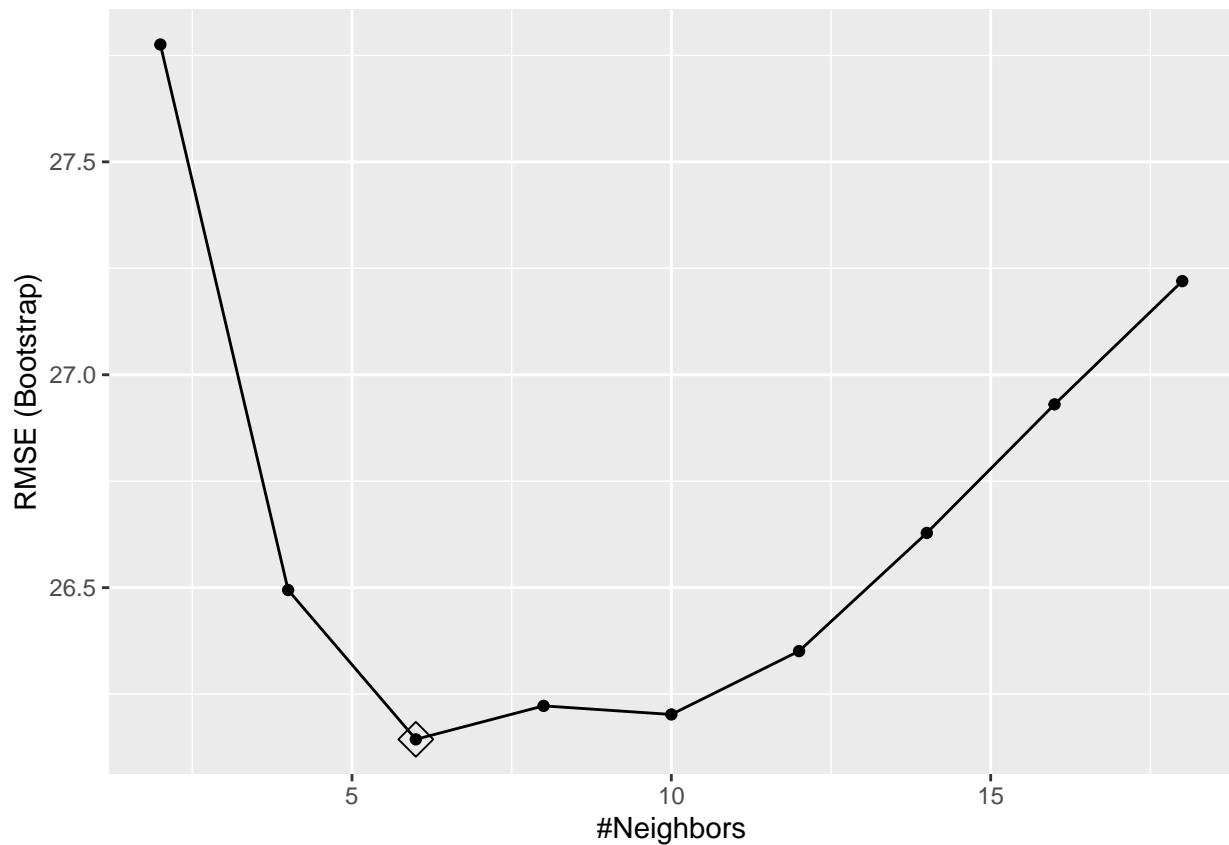
train_knn$finalModel

## 6-nearest neighbor regression model
y_hat_knn <- predict(train_knn, test_set, type = "raw")
```

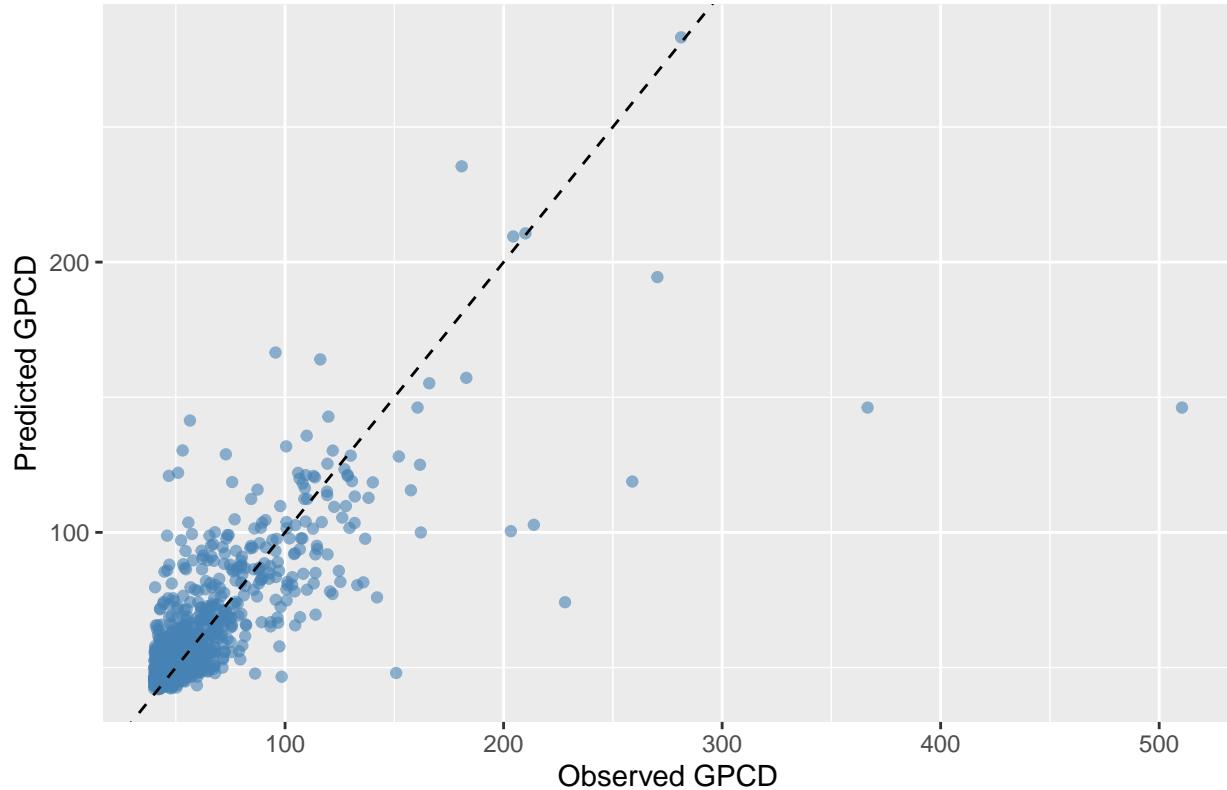
The optimal number of neighbors was $k = 6$, as determined through tuning. On the test set, the model achieved an RMSE of 22.6, an MAE of 10.8, and an R^2 of 0.56, indicating moderate predictive accuracy. The predictive accuracy is also reflected in the Predicted vs Observed GPCD scatter plot which shows predicted GPCD values from the model against observed values. When compared to the scatter plot of predicted versus observed GPCD values for the GLM, points are closer to the 1:1 line indicating improved accuracy.

```
##      RMSE Rsquared      MAE
```

```
## 22.6004 0.5567 10.8150
```



Predicted vs Observed GPCD (KNN Model)



Random Forest

A random forest model was trained to predict GPCD using the same variables used for the GLM and KNN models using the code shown below.

```
# Train and test random forest model
tune_grid <- expand.grid(mtry = seq(30, 50, 5))

train_rf <- train(
  gpcd ~ Hydrologic.Region + Climate.Zone + year + COVID + MHI + month + County + temperature,
  method = "rf",
  data = train_set,
  trControl = trainControl(method = "cv", number = 5, allowParallel = TRUE),
  tuneGrid = tune_grid,
  ntree = 1000
)

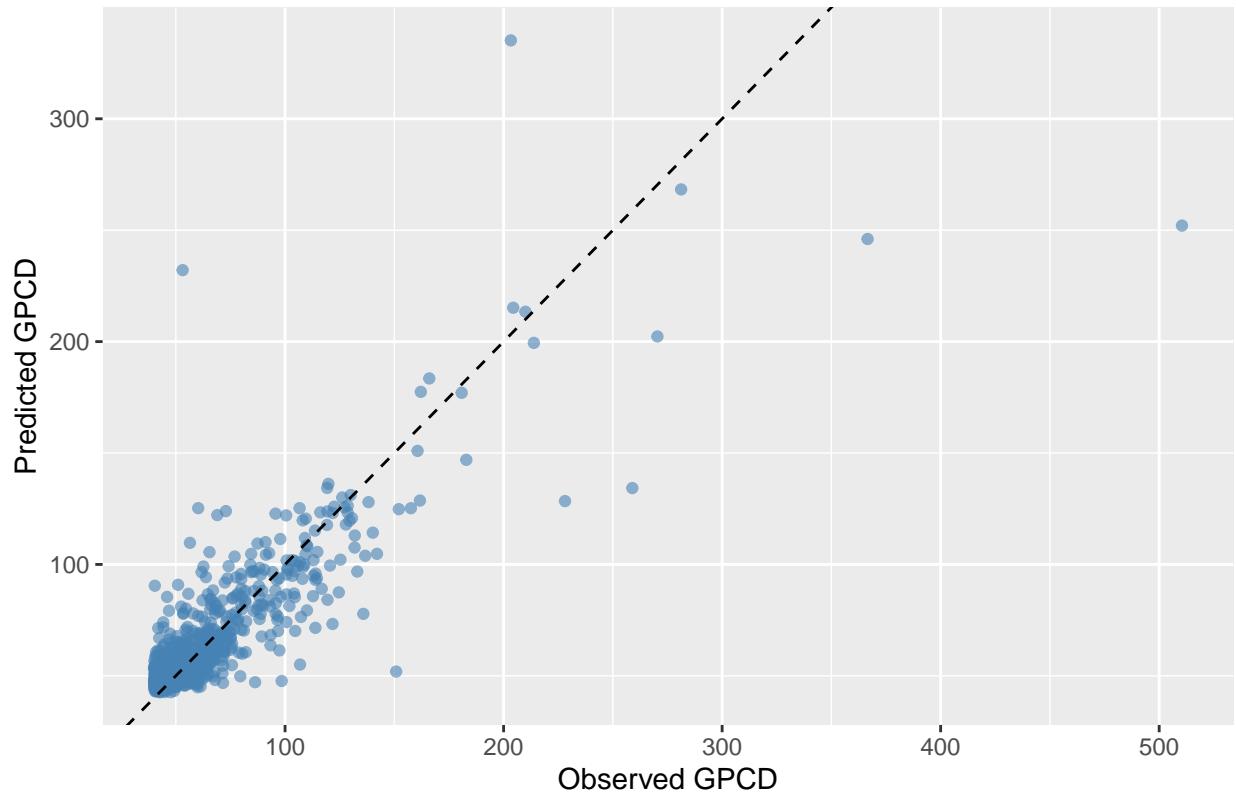
y_hat_rf <- predict(train_rf, test_set, type = "raw")
```

The random forest model outperformed both GLM and KNN, achieving an RMSE of 18.22, MAE of 9.2, and R² of 0.71, explaining about 71% of variance in GPCD. The predictive accuracy is reflected in the Predicted vs Observed GPCD scatter plot which shows points much closer to the line than seen with the GLM and KNN models, indicating higher prediction accuracy. The scatter plot also shows that at higher gpcd values, the model tends to under-predict gpcd. Variable importance analysis identified temperature, MHI, and hydrologic region as key predictors.

Detailed results and a summary of the random forest model are provided below.

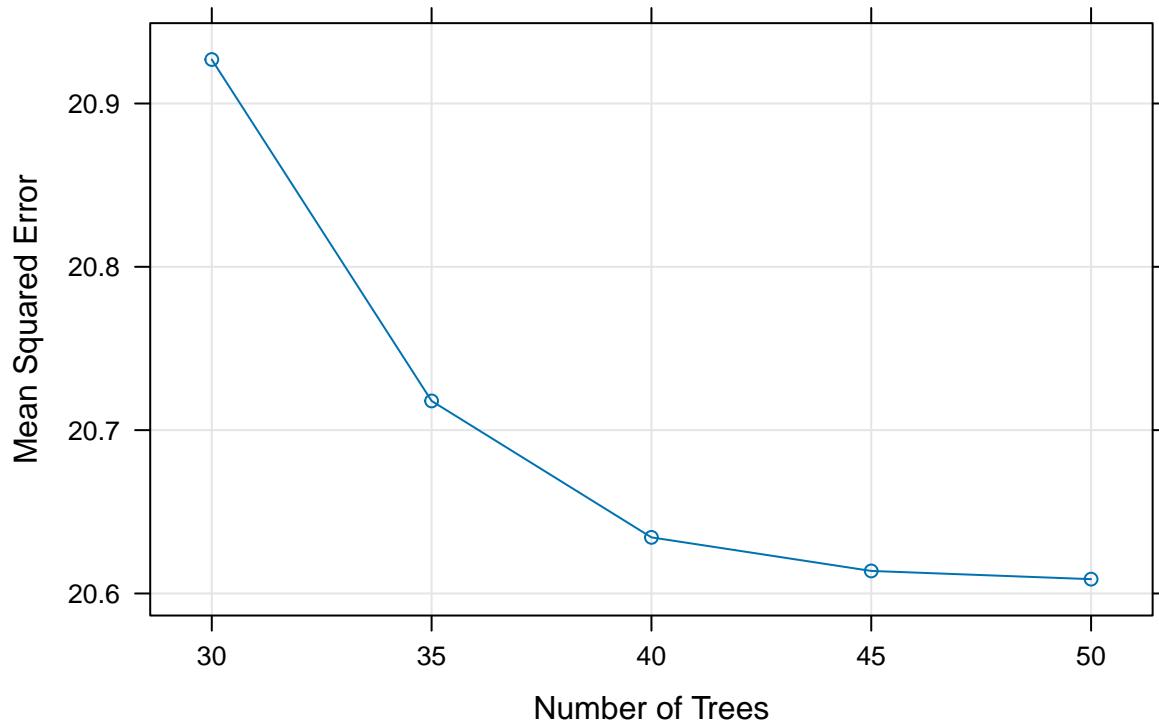
```
##      RMSE Rsquared      MAE
## 18.2193  0.7127 9.1974
```

Predicted vs Observed GPCD (RF Model)



```
##    mtry
## 5    50
```

Random Forest Model Error Rate



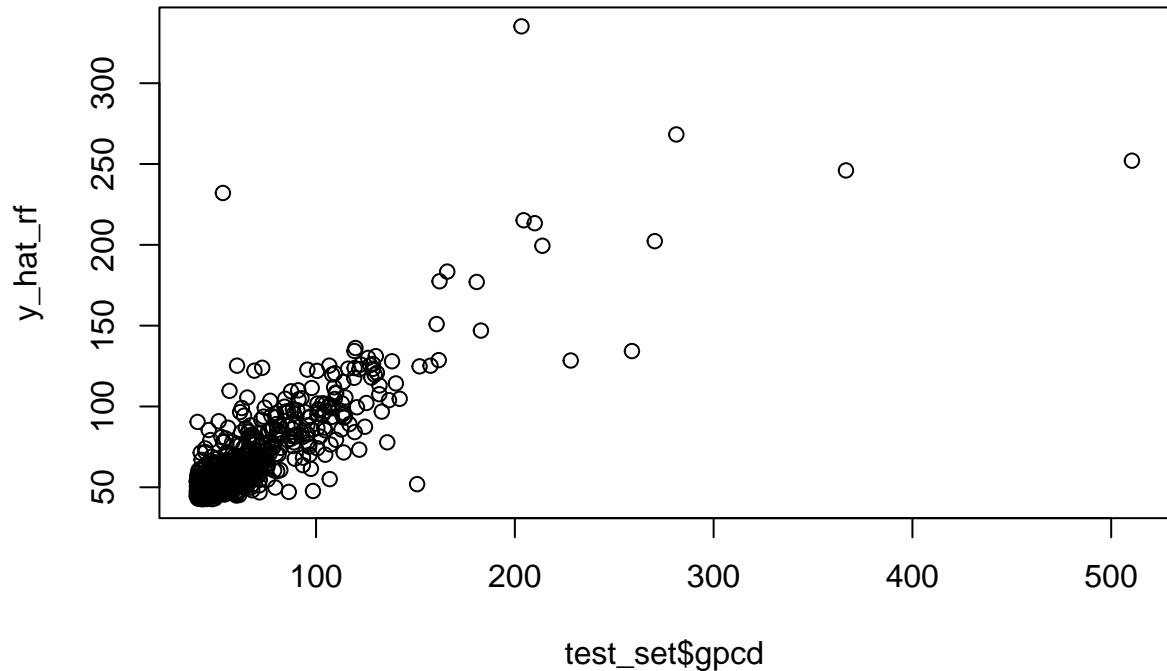
```
##  
## Call:  
##   randomForest(x = x, y = y, ntree = 1000, mtry = param$mtry)  
##           Type of random forest: regression  
##           Number of trees: 1000  
## No. of variables tried at each split: 50  
##  
##           Mean of squared residuals: 395.37  
##                         % Var explained: 71.19  
  
## rf variable importance  
##  
##   only 20 most important variables shown (out of 87)  
##  
##                                     Overall  
## temperature                      100.00  
## MHI                             51.39  
## CountyNevada                     45.80  
## Hydrologic.RegionSouth Lahontan  20.84  
## Hydrologic.RegionSouth Coast     17.47  
## year2021                        14.88  
## COVID                           14.82  
## Hydrologic.RegionNorth Lahontan 12.06  
## year2020                        10.90  
## Climate.Zone14                  10.57  
## month08                         9.87  
## month07                         9.84
```

```

## CountyKings           9.83
## CountyShasta          8.17
## month09               6.19
## CountyPlacer.Sacramento 5.92
## year2014              5.07
## CountySan Diego        4.96
## CountySan Bernardino  4.60
## year2015              4.52

##             Length Class      Mode
## call            5    -none-   call
## type            1    -none-   character
## predicted       3552   -none-  numeric
## mse             1000   -none-  numeric
## rsq              1000   -none-  numeric
## oob.times       3552   -none-  numeric
## importance      87    -none-  numeric
## importanceSD     0    -none-  NULL
## localImportance  0    -none-  NULL
## proximity        0    -none-  NULL
## ntree            1    -none-  numeric
## mtry             1    -none-  numeric
## forest           11   -none-  list
## coefs            0    -none-  NULL
## y                3552   -none-  numeric
## test             0    -none-  NULL
## inbag            0    -none-  NULL
## xNames           87    -none-  character
## problemType      1    -none-  character
## tuneValue         1    data.frame list
## obsLevels         1    -none-  logical
## param            1    -none-  list

```



Section 3: Results

The data exploration and ANOVA analysis yielded the following results:

- Demand Patterns: Residential water demand in California shows substantial spatial and temporal variability.
- Regional variation: Hydrologic regions differ significantly in normalized demand (GPCD).
- Seasonality: Demand peaks during summer months, correlating with higher temperatures and lower precipitation.
- Annual trends: Yearly demand fluctuates with climate conditions, with a notable increase during the COVID stay-at-home period (March 2020–June 2021).
- ANOVA Findings: One-way ANOVA tests identified the most influential factors on normalized demand
 - Highly significant: County, Hydrologic Region, Climate Zone, Year, COVID period, and Median Household Income (MHI).
 - Moderately significant: Month and Temperature.
 - Not significant: Precipitation, suggesting limited explanatory power for monthly demand.

Three models were evaluated on the test set:

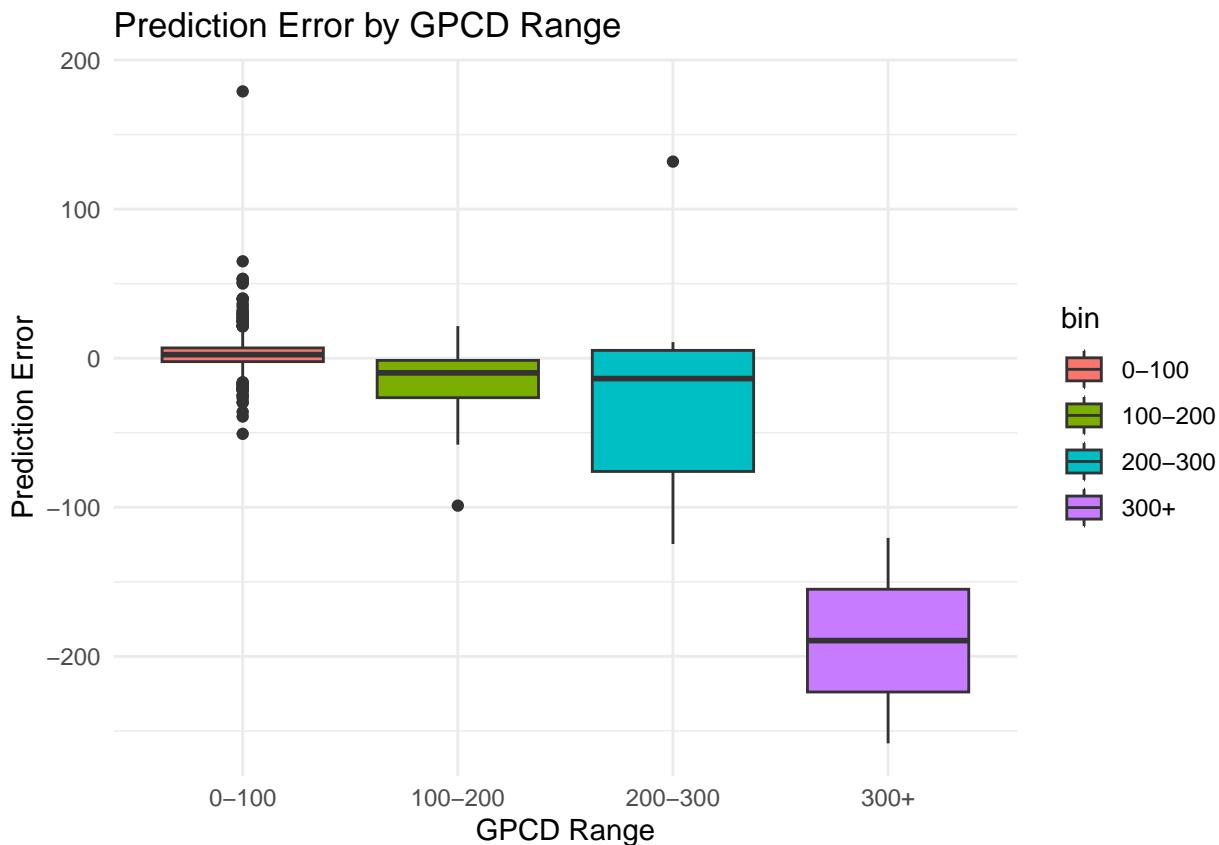
- GLM (Linear Regression): RMSE = 29.3, R² = 0.26, MAE = 16.9 — poor predictive accuracy
- KNN (k = 6): RMSE = 22.6, R² = 0.56, MAE = 10.8 — moderate performance.

- Random Forest: RMSE = 18.2, R² = 0.71, MAE = 9.2 — best performance, explaining ~71% of variance.

An analysis of the variable importance in the random forest model revealed:

- Top predictors: Temperature, MHI, and county-level factors (e.g. Nevada County).
- Secondary factors: Hydrologic Regions (South Lahontan, South Coast and North Lahontan), COVID period, year indicators (2020 and 2021)
- Seasonal indicators: Summer months (July and August) contributed modestly.
- Other notable predictors: specific climate zones (e.g. Climate Zone 14)
- Additional counties including Kings, Shasta, and San Diego

Box plots examining error at different GPCD levels were created to see how well the random forest model predicts gpcd as gpcd increases. This analysis, shown below, reveals higher prediction errors for higher GPCD ranges (200+), suggesting that stratified modeling or additional features could further enhance accuracy.



```
## # A tibble: 4 x 4
##   bin      RMSE      MAE      R2
##   <fct>    <dbl>    <dbl>    <dbl>
## 1 0-100    12.271   7.4319  0.43600
## 2 100-200   24.141  17.788   0.37209
## 3 200-300   77.455  58.223   0.035120
## 4 300+     201.60   189.45    1
```

Section 4: Conclusion

California's residential water demand is shaped by regional, climatic, and socioeconomic factors. Machine learning models, particularly Random Forest, provide robust predictive capability for GPCD. Additional factors could further improve these results, such as water prices, population density, and water conservation program data. In addition, one representative year was used to estimate population and MHI for each water system, but in reality these values vary by year. Incorporating annual changes in population and MHI could further improve the model. These insights can inform future water supply needs based on population growth, water conservation strategies, policy decisions, and the potential impacts of climate change on demands.

Section 5: References

Executive Department State of California (2021). Executive Order N-07-21. <https://www.gov.ca.gov/wp-content/uploads/2021/06/6.11.21-EO-N-07-21-signed.pdf>

Gross, M., A. Escriva-Bou, E. Porse, A. Cominola (2024a). CaRDS – the Statewide California Residential Water Demand and Supply open dataset, HydroShare (data download), <https://doi.org/10.4211/hs.4ec7019fe63944bf87d40d2cdfa0d686>

Gross, MP., Escriva-Bou, A., Porse, E. *et al.* (2024b). CaRDS - the statewide California Residential water Demand and Supply open dataset. *Sci Data* **11**, 632 (2024). <https://doi.org/10.1038/s41597-024-03474-y>

Heberger, Matthew (2014). New Data Show Residential Water Use across California. Pacific Institute. <https://pacinst.org/new-data-show-residential-per-capita-water-use-across-california/>

State Water Resources Control Board (2015). Factors that can affect per capita water. <https://www.waterboards.ca.gov/drought/docs/factors.pdf>

State Water Resources Control Board (2022). 2022 Risk Assessment Results. Prepared as part of the Safe and Affordable Funding for Equity and Resilience (SAFER) Program. https://www.waterboards.ca.gov/drinking_water/certlic/drinkingwater/documents/needs/2022risk.xlsx

Stillitano, Caroline (2025). Understanding California's climate Zones. San Diego Regional Climate collaborative. 5-6-2025. <https://digital.sandiego.edu/cgi/viewcontent.cgi?article=1043&context=npi-sdclimate>

United States Census Bureau (2020). US Census, 2020 American Community Survey 5-Year Estimates (2016-2020). <https://dof.ca.gov/reports/demographic-reports/american-community-survey/#ACS2020x5>