Predicting cases of dengue feaver based on evnivornemtnal data using a random forest model.

Madison Hobbs and Jenn Havens

Introduction:

Cases of dengue fever are related to the current and past climate. Environmental data can be used for predicting patterns in rates of dengue fever. Patterns of cases known in advance can used as an early warning system to help local authorities prepare for unusually high number of cases as well as informing which areas may need the most outside assistance.

The Model

We created two (2) predictive models, one for each of the cities, San Juan, Puerto Rico and Iquitos, Peru. Our model is a random forest algorithm which was trained on data from each city individually.

Missing Values:

We impute missing values using imputation via bagging (from the caret package). According to their documentation: " Imputation via bagging fits a bagged tree model for each predictor (as a function of all the others). This method is simple, accurate and accepts missing values, but it has much higher computational cost. "

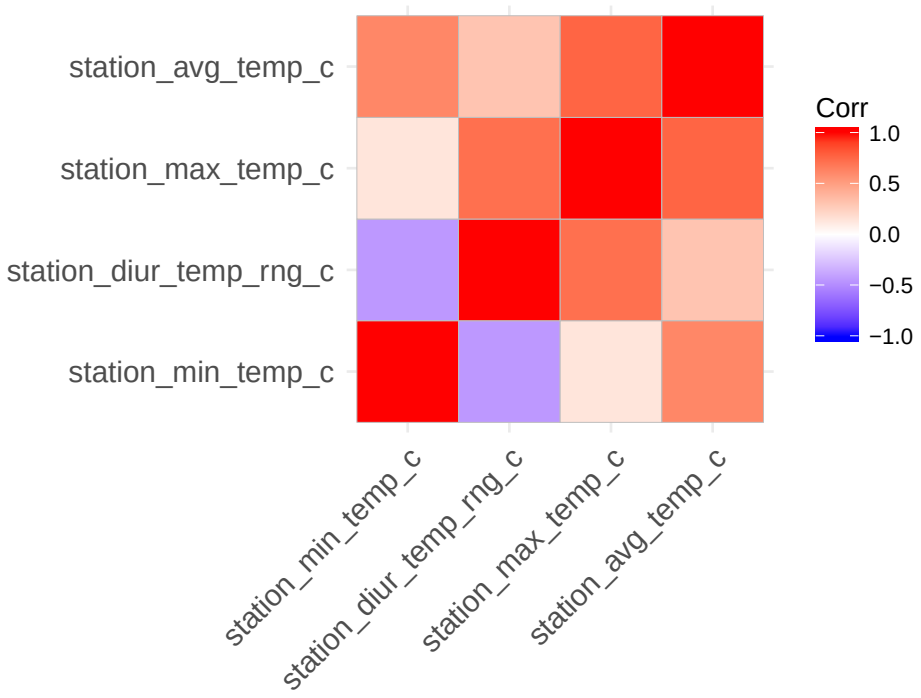
Variable Selection

Temperature

We think that the station-measured temperature would be preferable, because these are not measured from satellite or estimated from a model.

We should also think about whether to choose minimum temperature, maximum temperature, average temperature, or diurnal temperature range. We suspect that these variables are redundant.

In fact, all pairwise correlations between the four measures of temperature are significant. We can see these strong correlations represented in the correlation plot below:



Because of this correlation and because past studies have found that mean temperature was significantly associated with dengue rates, but maximum temperature and minimum were not always significant, we want to use just mean temperature for all of the correlated temperature variables.

Precipitation

Variable descriptions https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/page/82/.

There are three measures of precipitation. One is station_precip_mm which is the total daily precipitation as measured by NOAA's GHCN weather stations (https://www.ncdc.noaa.gov/ghcn-daily-description). Another is precipitation_amt_mm which represents total precipitation as measured by PERSIANN satellites. The third and forth, reanalysis_sat_precip_amt_mm and reanalysis_precip_amt_kg_per_m2 are both generated by NOAA's NCEP Climate Forecast System Reanalysis (https://climatedataguide.ucar.edu/climate-data/climate-forecast-system-reanalysis-cfsr).

We should choose one of these precipitation measures for the model, since these four precipitation measures all measure approximately the same thing. According to the NOAA Dengue Forecasting recommendations (http://dengueforecasting.noaa.gov), "remotely sensed observations are generally an excellent observation of precipitation and vegetation conditions for a location." With the multiple sources for total precipitation, we decide to use only the satellite measured total precipitation for each city to build our model.

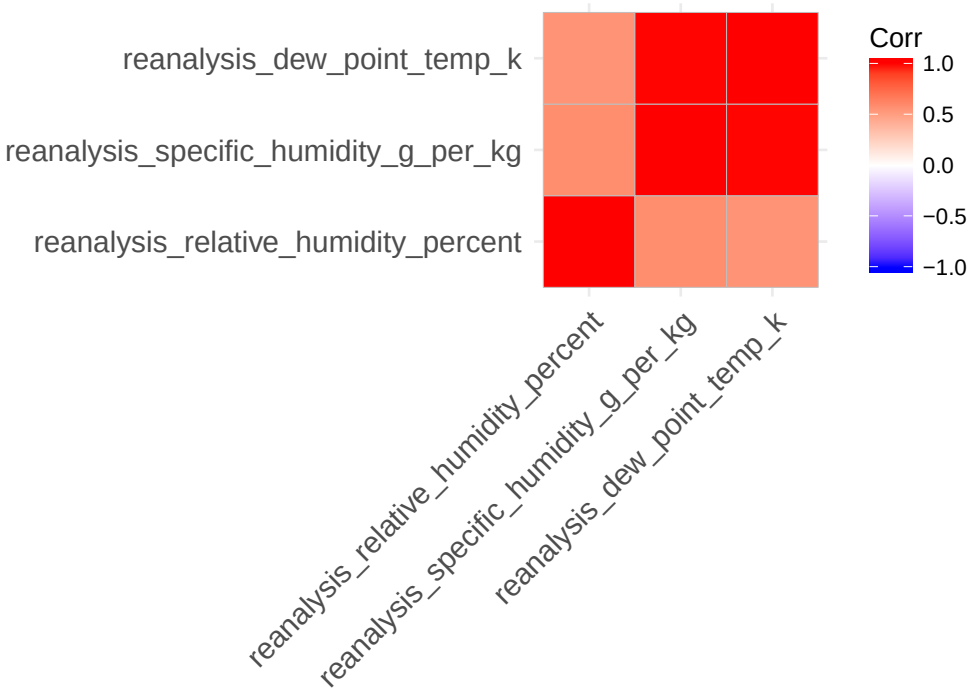
Which Air Temperature Measures Should Be Included?

Note that Choi et. al say : "Mean temperature was significantly associated with dengue incidence in all three provinces, but incidence did not correlate well with maximum temperature in Banteay Meanchey, nor with minimum temperature in Kampong Thom at a lag of three months in the negative binomial model." I'm inclined to only use mean temperature.

Should humidity and Dewpoint be included?

Specific humidity, relative humidity, and dew point are all provided and are all measured using NOAA's NCEP Climate Forecast System Reanalysis. All three measures are significantly correlated, as seen below.

```
## [1] TRUE
```



Since these measure roughly the same information, we opt to include only one in our model. Relative humidity is the measure most often used in literature we have read. It is also an easy-to-find measure, making our model more user-friendly. Therefore, we use only relative humidity in the model.

Creating Data for Model and Lagged Data

We separate San Juan and Iquitos data to produce two models, one to predict weekly dengue fever cases in San Juan and the other in Iquitos. This is because weather and vegetation will behave differently in relation to time between both locations, because these are locations separated by distance, climate, and ecosystem.

At the same time, we create two lag variables for each site: temperature lag and precipitation lag. The lagged temperature and precipitation variables record, respectively, what the temperature or precipitation was 12 weeks prior for the observation at hand.

Time laged environmental variables have been shown to be a signifigant predictors. Specifically we consider temperature and relative humidity.

Model Evaluation

A [time series Random Forest model] (https://link.springer.com/article/10.1186/1471-2105-15-276) was used for prediction of avian influenza H5N1 outbreaks. The model was assessed by taking data from the last 30 weeks to predict for the next week. The simulation steps forward interativly adding new data for each week, the new model using the updated datat to predict the next week, and comparing to the true number of cases.

Moving in steps of 30 weeks, we train the model on 30 weeks, then predict the next week.

We'll assess our model's performance by computing the error (MSE) between the model's predicted number of cases and the actual number of cases for each week. The Driven Data team uses Mean Absolute Error to measure error, so we will, too.

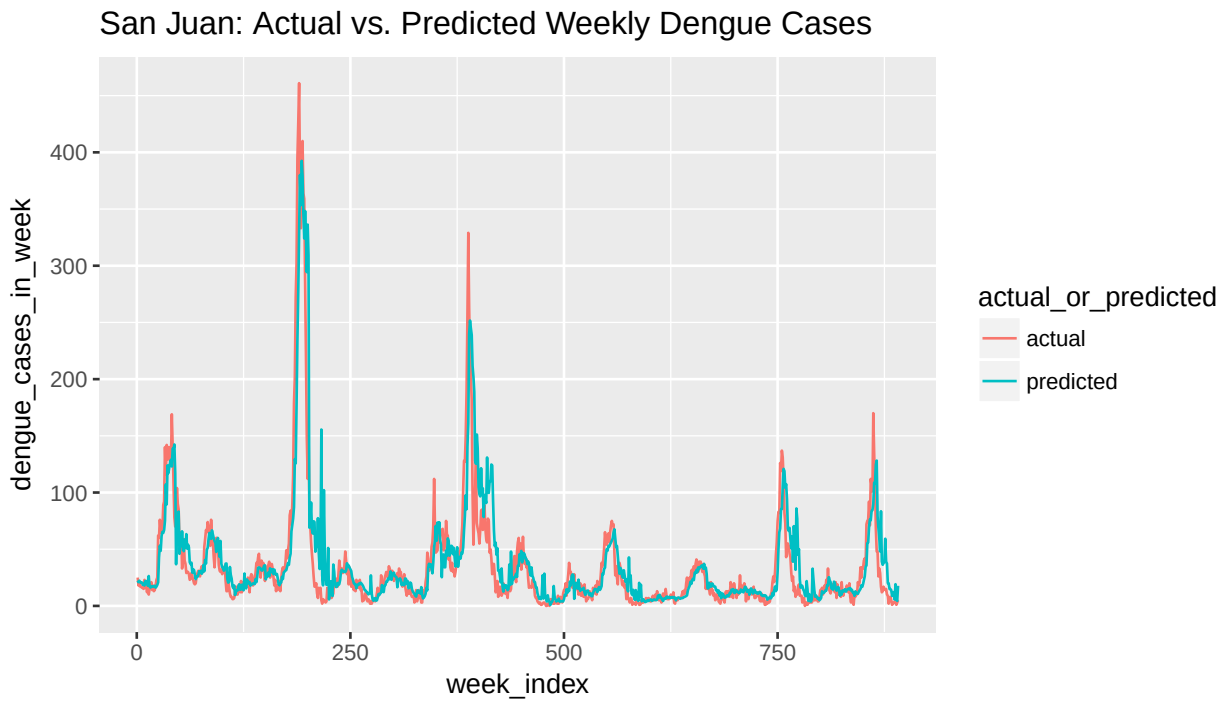
San Juan

Testing with current environemntal data, lagged environemntal data and date information

```
## [1] 26.98486
```

```
## [1] 14.44279
```

```
## [1] 51.4
```



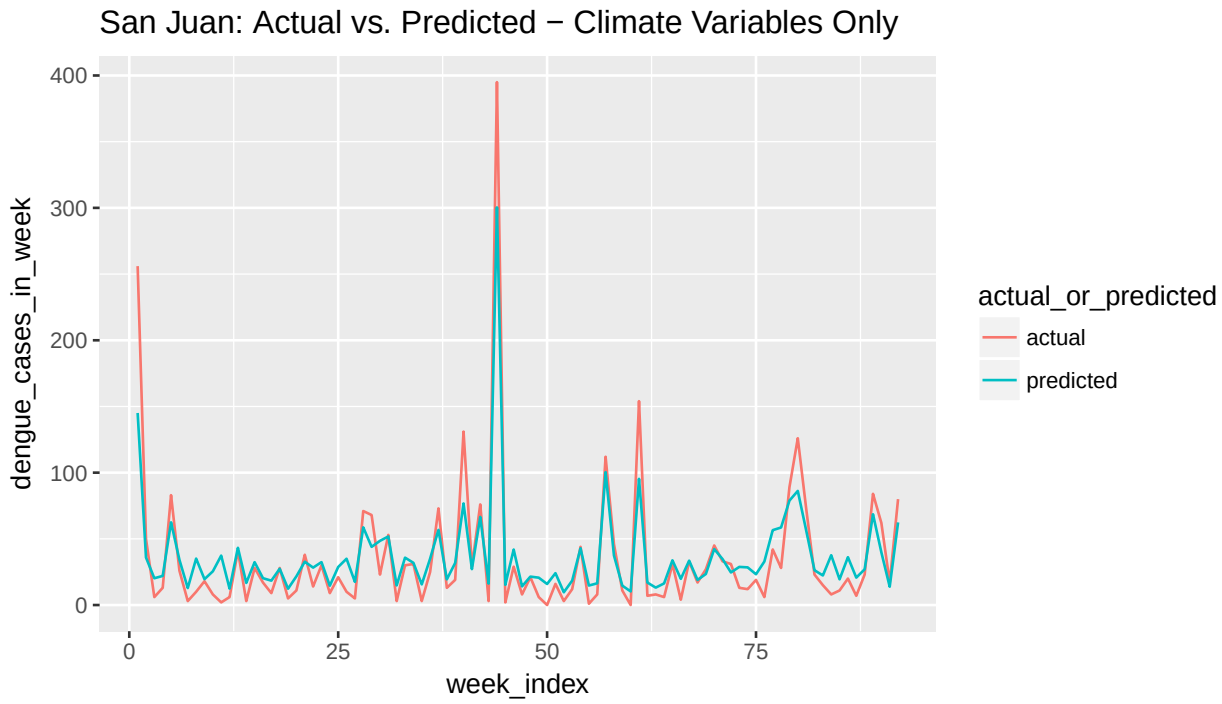
With environmental varibles only

```
## Random Forest
##
## 829 samples
##   7 predictor
##
## No pre-processing
## Resampling results across tuning parameters:
##
##   mtry  RMSE      Rsquared
##   1     42.97660  0.2996247
##   2     41.83573  0.3363162
##   3     42.27996  0.3221468
```

```
##   4     42.26559  0.3226075
##   5     42.56941  0.3128337
##   6     43.42063  0.2850777
##   7     44.43886  0.2511541
##
## RMSE was used to select the optimal model using  the smallest value.
## The final value used for the model was mtry = 2.
```

```
## Random Forest
##
## 921 samples
##   7 predictor
##
## No pre-processing
## Resampling results:
##
##   RMSE    Rsquared
##   42.665  0.3174114
##
## Tuning parameter 'mtry' was held constant at a value of 1
```

```
## [1] 21.89844
```



With Lagged Temperature, Humidity, and Precipitation

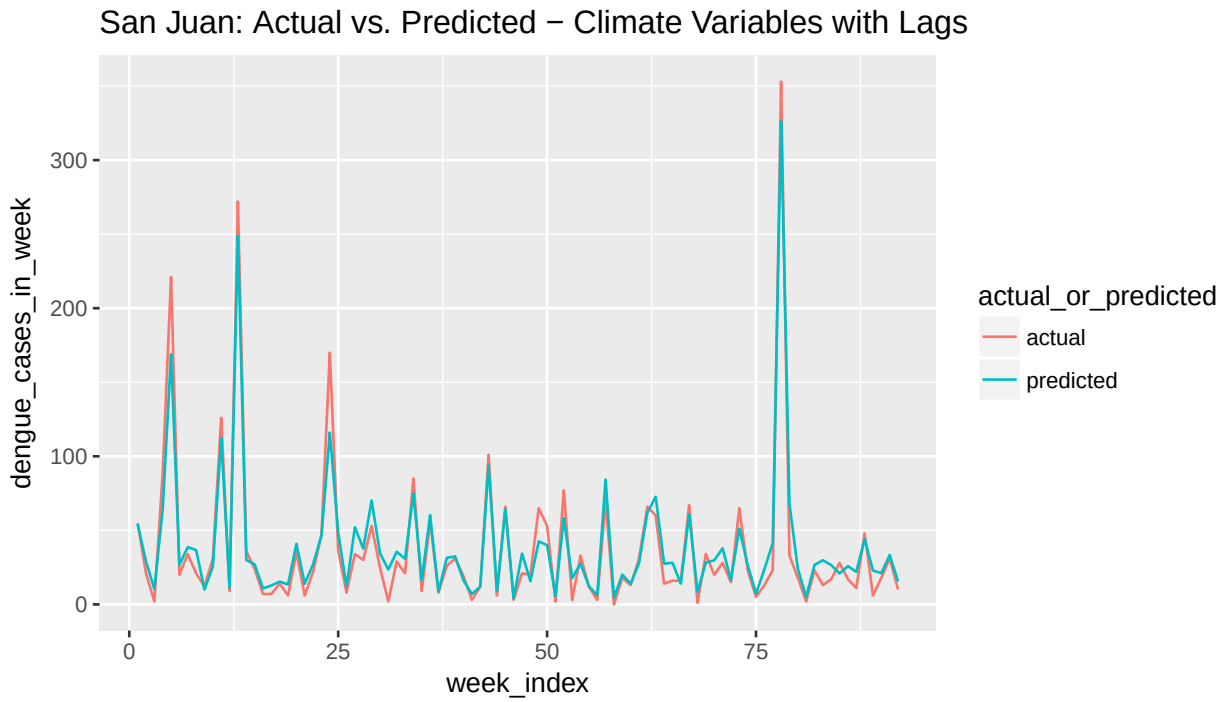
```
## Random Forest
##
## 829 samples
##  10 predictor
##
```

```
## No pre-processing
## Resampling results across tuning parameters:
##
##   mtry  RMSE      Rsquared
##    1    42.25831  0.3232115
##    2    41.53855  0.3460698
##    3    41.37540  0.3511966
##    4    40.70834  0.3719482
##    5    40.75764  0.3704260
##    6    40.13203  0.3896050
##    7    40.68963  0.3725256
##    8    40.09861  0.3906212
##    9    39.61952  0.4050959
##   10    39.46100  0.4098467
##
## RMSE was used to select the optimal model using  the smallest value.
## The final value used for the model was mtry = 10.
```

```
## Random Forest
##
## 921 samples
##  10 predictor
##
## No pre-processing
## Resampling results:
##
##   RMSE      Rsquared
##   37.59175  0.4700918
##
## Tuning parameter 'mtry' was held constant at a value of 10
```

```
## [1] 13.08583
```

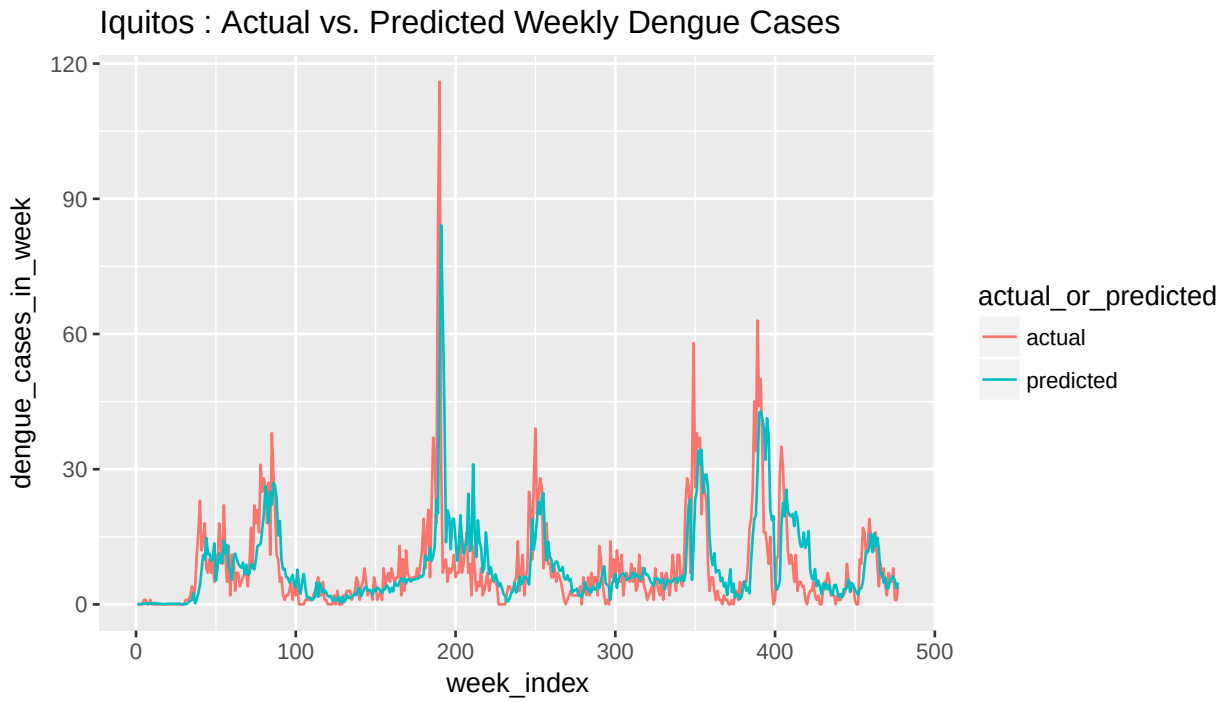
When taking ndvi out of the model, the RMSE rises to 24.66074.



Iquitos

Testing with current environemntal data, lagged environemntal data and date information

```
## [1] 9.101376
## [1] 5.180046
## [1] 82.7
```



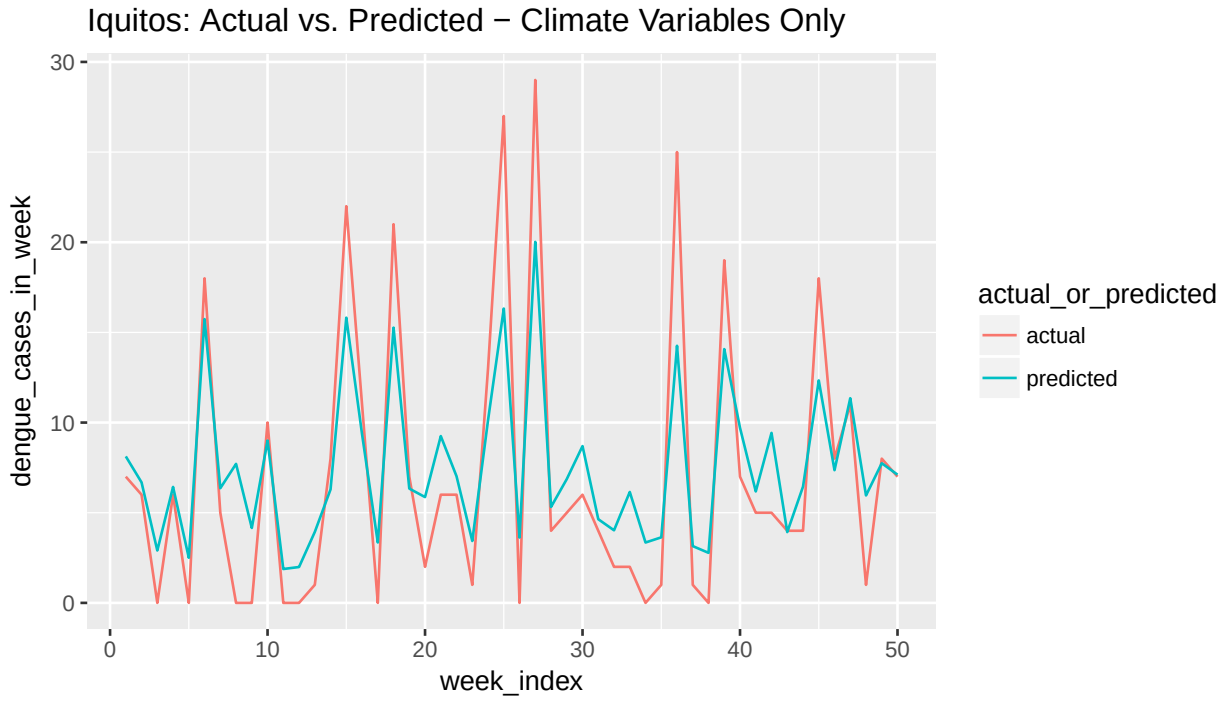
With environmental varibles only

```
## Random Forest
##
## 457 samples
##   7 predictor
##
## No pre-processing
## Resampling results across tuning parameters:
##
##   mtry  RMSE      Rsquared
##   1     11.24609  -0.02231511
##   2     11.41885  -0.05396601
##   3     11.37250  -0.04542754
##   4     11.45515  -0.06067885
##   5     11.46841  -0.06313446
##   6     11.48320  -0.06587799
##   7     11.61092  -0.08972003
##
## RMSE was used to select the optimal model using  the smallest value.
## The final value used for the model was mtry = 1.
```

```
## Random Forest
##
## 507 samples
##   7 predictor
##
## No pre-processing
## Resampling results:
##
```

```
##   RMSE      Rsquared
##   11.03027  -0.03742801
##
## Tuning parameter 'mtry' was held constant at a value of 1
```

```
## [1] 3.873413
```



With Lagged Temperature, Humidity, and Precipitation

```
## Random Forest
##
## 457 samples
##  10 predictor
##
## No pre-processing
## Resampling results across tuning parameters:
##
##   mtry  RMSE      Rsquared
##    1    10.98754  -0.01232690
##    2    11.03654  -0.02137667
##    3    11.08305  -0.03000240
##    4    11.11931  -0.03675408
##    5    11.15284  -0.04301610
##    6    11.20729  -0.05322424
##    7    11.17540  -0.04723947
##    8    11.26984  -0.06501480
##    9    11.33203  -0.07679967
##   10    11.20328  -0.05247058
##
## RMSE was used to select the optimal model using  the smallest value.
```

```
## The final value used for the model was mtry = 1.
```

```
## Random Forest
##
## 507 samples
##  10 predictor
##
## No pre-processing
## Resampling results:
##
##   RMSE      Rsquared
##   10.79078  0.007131944
##
## Tuning parameter 'mtry' was held constant at a value of 1
```

```
## [1] 4.910832
```