CFSR Indian Weather Data Model

unknownda

2022-08-28

Setting up the data frame

Installing and loading *readr*, *dplyr*, *readxl*, *caTools*, *ggplot2* and *car* packages. Then import the weather data from the *yearwise_raw.xlsx* file.

The file contains the values of different weather parameters of 215 cities across India averaged annually from 1979-2014. The different parameters include **Year**, **Minimum Temperature** (**Tmin** in degree Celsius), **Maximum Temperature** (**Tmax** in degree Celsius), **Solar Radiation** (**Solar** in MJ/mˆ2), **Wind Speed** (**Wind** in m/s), **Relative Humidity**, **Precipitation** (in mm). Data collated from CFSR Global Weather Data for SWAT 1979-2014. SWAT Data

The objective is to prepare a model to predict Precipitation on the basis of all other parameters. For the sake of analysis, the differences between cities has been disregarded and hence the data has been considered to represent India as a whole.

Analyzing Input Data

```
## tibble [7,740 x 9] (S3: tbl_df/tbl/data.frame)
##  $ State            : chr [1:7740] "Kerala" "Kerala" "Kerala" "Kerala" ...
##  $ City             : chr [1:7740] "Kottayam" "Kottayam" "Kottayam" "Kottayam" ...
##  $ Year             : num [1:7740] 1979 1980 1981 1982 1983 ...
##  $ Tmin             : num [1:7740] 20.5 20.2 20.2 20 20.1 ...
##  $ Tmax             : num [1:7740] 28.4 29.3 29 29.8 30.7 ...
##  $ Solar            : num [1:7740] 18.3 18 17 18.6 16.9 ...
##  $ Wind             : num [1:7740] 2.33 2.44 2.39 2.35 2.46 ...
##  $ Relative_Humidity: num [1:7740] 0.843 0.8 0.813 0.783 0.758 ...
##  $ Precipitation    : num [1:7740] 10.98 9.91 11.1 7.51 8.41 ...
```

```
##     State               City                Year           Tmin        
##  Length:7740        Length:7740        Min.   :1979   Min.   :-23.00  
##  Class :character   Class :character   1st Qu.:1988   1st Qu.: 17.77  
##  Mode  :character   Mode  :character   Median :1996   Median : 19.45  
##                                        Mean   :1996   Mean   : 16.55  
##                                        3rd Qu.:2005   3rd Qu.: 20.95  
##                                        Max.   :2014   Max.   : 25.58  
##       Tmax            Solar             Wind        Relative_Humidity
##  Min.   :-9.426   Min.   : 8.355   Min.   :0.8167   Min.   :0.2535   
##  1st Qu.:30.490   1st Qu.:18.494   1st Qu.:2.3008   1st Qu.:0.4668   
##  Median :32.636   Median :19.150   Median :2.7303   Median :0.5570   
##  Mean   :29.282   Mean   :19.103   Mean   :2.7731   Mean   :0.5624   
##  3rd Qu.:34.007   3rd Qu.:19.778   3rd Qu.:3.2202   3rd Qu.:0.6518   
##  Max.   :38.207   Max.   :24.343   Max.   :5.6257   Max.   :0.9486
```

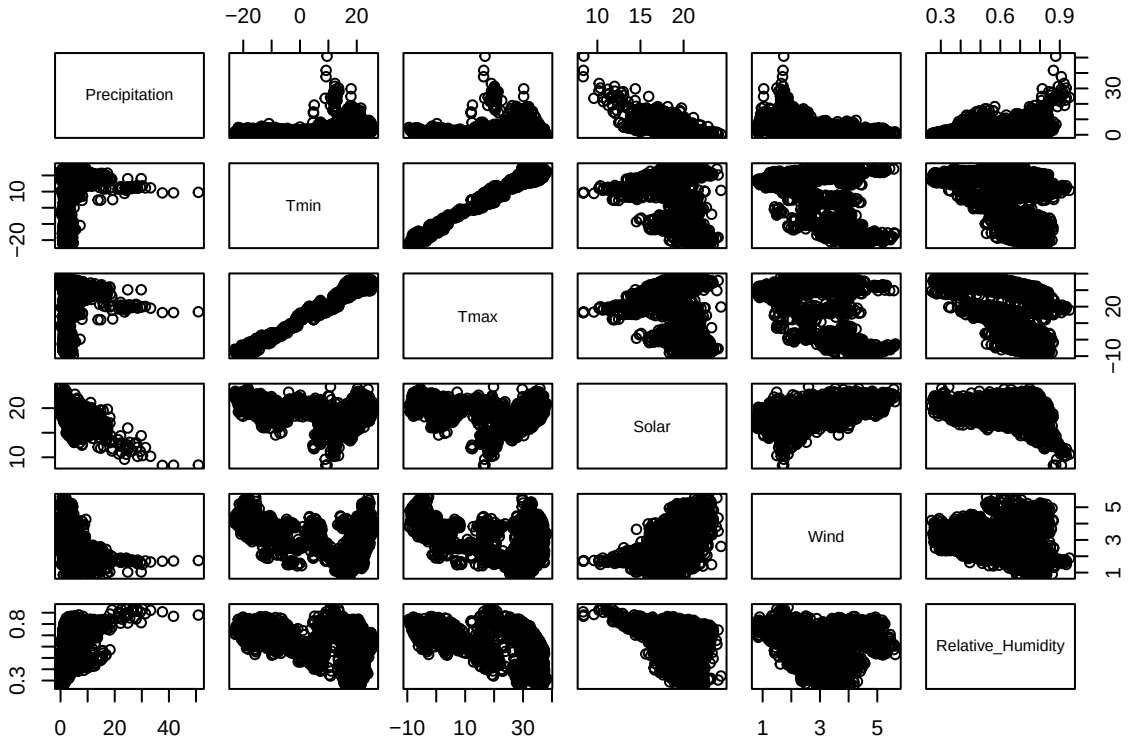
```
##  Precipitation
##  Min.   : 0.00484
##  1st Qu.: 1.98080
##  Median : 3.03571
##  Mean   : 3.65172
##  3rd Qu.: 4.42884
##  Max.   :50.88406
```

The data set contains no missing values. Removing the *State* and *City* variables as explained earlier. The **Correlation Matrix** is shown below:

```
##                          Year        Tmin        Tmax      Solar        Wind
## Year               1.000000000  0.02083892  0.02491832  0.1186939 -0.07470837
## Tmin               0.020838918  1.00000000  0.98200354 -0.2002940 -0.31743443
## Tmax               0.024918317  0.98200354  1.00000000 -0.1354049 -0.33681424
## Solar              0.118693887 -0.20029404 -0.13540489  1.0000000  0.47178242
## Wind              -0.074708366 -0.31743443 -0.33681424  0.4717824  1.00000000
## Relative_Humidity -0.004805462 -0.37961429 -0.49313804 -0.3460251 -0.15032073
## Precipitation      0.037955736  0.10513579  0.02733605 -0.6251573 -0.45606184
##                   Relative_Humidity Precipitation
## Year                     -0.004805462    0.03795574
## Tmin                     -0.379614288    0.10513579
## Tmax                     -0.493138041    0.02733605
## Solar                    -0.346025057   -0.62515726
## Wind                     -0.150320728   -0.45606184
## Relative_Humidity         1.000000000    0.51150741
## Precipitation             0.511507411    1.00000000
```

- *Year*, *Tmin*, *Tmax* & *Relative_Humidity* are positively correlated while *Solar* & *Wind* are negatively correlated with *Precipitation*.
- There is a high positive correlation between *Tmin* & *Tmax*.

The **Scatter Plots** of the weather parameters are displayed next.



Training & Test Set

```
# Dividing data into Training and Testing set

set.seed(0)
split <- sample.split(df$City, SplitRatio = 0.8)
training_set <- subset(df_cleaned,split==TRUE)
test_set <- subset(df_cleaned, split==FALSE)
```

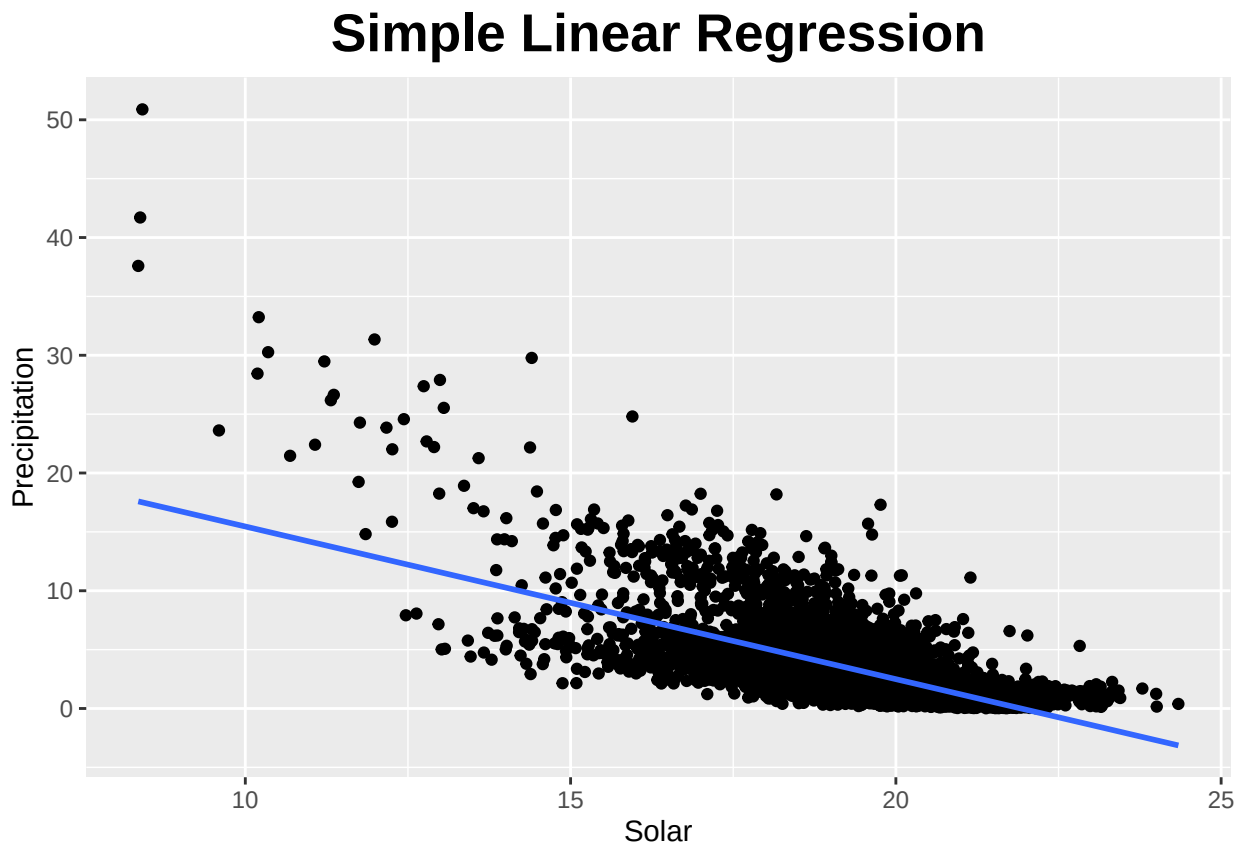
Simple Linear (SL) Regression of Precipitation onto Solar

First, simple linear regression of *Precipitation* is performed with *Solar* as it has the maximum correlation coefficient.

```
##
## Call:
## lm(formula = Precipitation ~ Solar, data = training_set)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.989 -1.408 -0.402  0.934 33.377
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 28.42053    0.39315   72.29   <2e-16 ***
## Solar       -1.29644    0.02053  -63.15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.347 on 6240 degrees of freedom
## Multiple R-squared:  0.3899, Adjusted R-squared:  0.3898
## F-statistic:  3988 on 1 and 6240 DF,  p-value: < 2.2e-16
```

```
## `geom_smooth()` using formula 'y ~ x'
```



The Adjusted R-squared is quite low and hence a better model is required.

Multiple Linear Regression Models

1. Multiple Linear (ML) including all variables

```
##
## Call:
## lm(formula = Precipitation ~ ., data = training_set)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.933 -1.021 -0.171  0.773 34.666
```

```
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -23.402126   5.058491  -4.626  3.8e-06 ***
## Year                0.021379   0.002538   8.423  < 2e-16 ***
## Tmin                0.216319   0.021526  10.049  < 2e-16 ***
## Tmax               -0.193718   0.023003  -8.421  < 2e-16 ***
## Solar              -0.772620   0.023051 -33.518  < 2e-16 ***
## Wind               -0.913457   0.050014 -18.264  < 2e-16 ***
## Relative_Humidity   6.684372   0.356250  18.763  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.034 on 6235 degrees of freedom
## Multiple R-squared:  0.542,  Adjusted R-squared:  0.5415
## F-statistic:  1230 on 6 and 6235 DF,  p-value: < 2.2e-16
```

```
##             Year             Tmin             Tmax            Solar
##         1.043445        64.395946        77.139525         1.677934
##             Wind Relative_Humidity
##         2.065393         3.314531
```

- Residual Standard Error is 2.034.
- Adjusted R-squared is 0.5415.
- VIF of *Tmin* & *Tmax* are very high.

2. Multiple Linear without Tmin

```
##
## Call:
## lm(formula = Precipitation ~ . - Tmin, data = training_set)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.941 -1.049 -0.183  0.771 34.773
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -30.054705   5.055029  -5.946 2.91e-09 ***
## Year                0.022737   0.002555   8.899  < 2e-16 ***
## Tmax                0.034757   0.003525   9.860  < 2e-16 ***
## Solar              -0.837502   0.022305 -37.548  < 2e-16 ***
## Wind               -0.655076   0.043241 -15.149  < 2e-16 ***
## Relative_Humidity   9.096779   0.265322  34.286  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.05 on 6236 degrees of freedom
## Multiple R-squared:  0.5345, Adjusted R-squared:  0.5342
## F-statistic:  1432 on 5 and 6236 DF,  p-value: < 2.2e-16
```

```
##             Year             Tmax            Solar             Wind
##         1.040490         1.782901         1.546296         1.519532
```

```
## Relative_Humidity
##          1.809462
```

- Residual Standard Error is 2.05.
- Adjusted R-squared is 0.5342.
- All VIF are within acceptable range.

3. Multiple Linear without Tmax

```
##
## Call:
## lm(formula = Precipitation ~ . - Tmax, data = training_set)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.852 -1.041 -0.183  0.777 34.829
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -29.150356   5.040247  -5.784 7.67e-09 ***
## Year                0.022417   0.002550   8.793  < 2e-16 ***
## Tmin                0.037147   0.003291  11.288  < 2e-16 ***
## Solar              -0.820607   0.022461 -36.536  < 2e-16 ***
## Wind               -0.679371   0.041811 -16.249  < 2e-16 ***
## Relative_Humidity   8.887163   0.243208  36.541  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.046 on 6236 degrees of freedom
## Multiple R-squared:  0.5367, Adjusted R-squared:  0.5364
## F-statistic:  1445 on 5 and 6236 DF,  p-value: < 2.2e-16
```

```
##              Year              Tmin             Solar              Wind
##          1.040988          1.488363          1.575397          1.427417
## Relative_Humidity
##          1.527653
```

- Residual Standard Error is 2.046.
- Adjusted R-squared is 0.5364.
- All VIF are within acceptable range.

4. Multiple Linear with Interaction term

```
##
## Call:
## lm(formula = Precipitation ~ . + Tmin:Tmax, data = training_set)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.098 -1.038 -0.185  0.779 34.531
##
## Coefficients:
```

```
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -1.695e+01  5.034e+00  -3.367 0.000766 ***
## Year               2.002e-02  2.514e-03   7.964 1.97e-15 ***
## Tmin               2.293e-01  2.132e-02  10.754  < 2e-16 ***
## Tmax              -3.091e-01  2.480e-02 -12.468  < 2e-16 ***
## Solar             -8.358e-01  2.343e-02 -35.669  < 2e-16 ***
## Wind              -1.243e+00  5.691e-02 -21.843  < 2e-16 ***
## Relative_Humidity  5.184e+00  3.749e-01  13.827  < 2e-16 ***
## Tmin:Tmax          4.141e-03  3.533e-04  11.721  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.012 on 6234 degrees of freedom
## Multiple R-squared:  0.5518, Adjusted R-squared:  0.5513
## F-statistic:  1097 on 7 and 6234 DF,  p-value: < 2.2e-16
```

```
## GVIFs computed for predictors
```

```
##                          GVIF Df GVIF^(1/(2*Df)) Interacts With
## Year                1.045675  1        1.022583           --
## Tmin                6.665854  3        1.371858         Tmax
## Tmax                6.665854  3        1.371858         Tmin
## Solar               1.771577  1        1.331006           --
## Wind                2.732899  1        1.653148           --
## Relative_Humidity 3.751651  1        1.936918           --
##                                                Other Predictors
## Year                Tmin, Tmax, Solar, Wind, Relative_Humidity
## Tmin                      Year, Solar, Wind, Relative_Humidity
## Tmax                      Year, Solar, Wind, Relative_Humidity
## Solar              Year, Tmin, Tmax, Wind, Relative_Humidity
## Wind               Year, Tmin, Tmax, Solar, Relative_Humidity
## Relative_Humidity              Year, Tmin, Tmax, Solar, Wind
```

- Residual Standard Error is 2.012.
- Adjusted R-squared is 0.5513.
- All GVIF^(1/(2*Df)) are less than 2 i.e. all VIF are less than 4, hence acceptable.

Multiple Non-Linear Regression Models

1. Multiple Non-Linear (MNL) including all variables

```
##
## Call:
## lm(formula = Precipitation ~ Year + Tmin + Tmax + I(Solar^-1) +
##     I(Wind^-1) + Relative_Humidity, data = training_set)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.7041 -0.9334 -0.1387  0.7551 23.8598
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)       -59.694153   4.787738 -12.468  < 2e-16 ***
## Year                0.021847   0.002361   9.252  < 2e-16 ***
## Tmin                0.171884   0.020147   8.532  < 2e-16 ***
## Tmax               -0.137433   0.021375  -6.430 1.37e-10 ***
## I(Solar^-1)       299.219834   6.224803  48.069  < 2e-16 ***
## I(Wind^-1)          4.810133   0.256249  18.771  < 2e-16 ***
## Relative_Humidity   5.805376   0.351655  16.509  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.91 on 6235 degrees of freedom
## Multiple R-squared:  0.5963, Adjusted R-squared:  0.5959
## F-statistic:  1535 on 6 and 6235 DF,  p-value: < 2.2e-16
```

```
##              Year              Tmin              Tmax       I(Solar^-1)
##          1.024488         64.002637         75.575269          1.457507
##        I(Wind^-1) Relative_Humidity
##          1.877222          3.664399
```

- Residual Standard Error is 1.91.
- Adjusted R-squared is 0.5959.
- VIF of *Tmin* & *Tmax* are very high.

2. Multiple Non-Linear without Tmin

```
##
## Call:
## lm(formula = Precipitation ~ Year + Tmax + I(Solar^-1) + I(Wind^-1) +
##     Relative_Humidity, data = training_set)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.1399 -0.9645 -0.1477  0.7656 23.5795
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -64.817181   4.777195 -13.568   <2e-16 ***
## Year                0.022486   0.002374   9.473   <2e-16 ***
## Tmax                0.042966   0.003148  13.649   <2e-16 ***
## I(Solar^-1)       312.097882   6.073676  51.385   <2e-16 ***
## I(Wind^-1)          3.702373   0.222186  16.663   <2e-16 ***
## Relative_Humidity   7.878131   0.255696  30.811   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.921 on 6236 degrees of freedom
## Multiple R-squared:  0.5916, Adjusted R-squared:  0.5913
## F-statistic:  1807 on 5 and 6236 DF,  p-value: < 2.2e-16
```

```
##              Year              Tmax       I(Solar^-1)        I(Wind^-1)
##          1.023458          1.620315          1.371803          1.395248
## Relative_Humidity
##          1.915334
```

- Residual Standard Error is 1.921.
- Adjusted R-squared is 0.5913.
- All VIF are within acceptable range.

3. Multiple Non-Linear without Tmax

```
##
## Call:
## lm(formula = Precipitation ~ Year + Tmin + I(Solar^-1) + I(Wind^-1) +
##     Relative_Humidity, data = training_set)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2510 -0.9568 -0.1411  0.7672 23.7317
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -63.478636   4.766762 -13.317   <2e-16 ***
## Year                0.022265   0.002368   9.403   <2e-16 ***
## Tmin                0.043745   0.002959  14.781   <2e-16 ***
## I(Solar^-1)       307.654804   6.104637  50.397   <2e-16 ***
## I(Wind^-1)          3.918144   0.216142  18.128   <2e-16 ***
## Relative_Humidity   7.495564   0.234328  31.987   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.916 on 6236 degrees of freedom
## Multiple R-squared:  0.5936, Adjusted R-squared:  0.5933
## F-statistic:  1822 on 5 and 6236 DF,  p-value: < 2.2e-16
```

```
##              Year              Tmin       I(Solar^-1)        I(Wind^-1)
##          1.023709          1.372200          1.392769          1.326988
## Relative_Humidity
##          1.616655
```

- Residual Standard Error is 1.916.
- Adjusted R-squared is 0.5933.
- All VIF are within acceptable range.

4. Multiple Non-Linear with Interaction Term

```
##
## Call:
## lm(formula = Precipitation ~ Year + Tmin + Tmax + I(Solar^-1) +
##     I(Wind^-1) + Relative_Humidity + Tmin:Tmax, data = training_set)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.2761 -0.9514 -0.1488  0.7304 22.9404
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)       -5.820e+01  4.726e+00 -12.314   <2e-16 ***
## Year               2.079e-02  2.331e-03   8.916   <2e-16 ***
## Tmin               1.721e-01  1.988e-02   8.657   <2e-16 ***
## Tmax              -2.339e-01  2.236e-02 -10.462   <2e-16 ***
## I(Solar^-1)        3.238e+02  6.428e+00  50.382   <2e-16 ***
## I(Wind^-1)         6.357e+00  2.795e-01  22.749   <2e-16 ***
## Relative_Humidity  4.331e+00  3.651e-01  11.863   <2e-16 ***
## Tmin:Tmax          4.093e-03  3.147e-04  13.005   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.884 on 6234 degrees of freedom
## Multiple R-squared:  0.607,  Adjusted R-squared:  0.6065
## F-statistic:  1375 on 7 and 6234 DF,  p-value: < 2.2e-16
```

```
## GVIFs computed for predictors
```

```
##                          GVIF Df GVIF^(1/(2*Df)) Interacts With
## Year                 1.025740  1        1.012788           --
## Tmin                 5.448396  3        1.326513         Tmax
## Tmax                 5.448396  3        1.326513         Tmin
## Solar             1280.553577  0             Inf           --
## Wind              1280.553577  0             Inf           --
## Relative_Humidity    4.055620  1        2.013857           --
##                                                Other Predictors
## Year                    Tmin, Tmax, Solar, Wind, Relative_Humidity
## Tmin                          Year, Solar, Wind, Relative_Humidity
## Tmax                          Year, Solar, Wind, Relative_Humidity
## Solar             Year, Tmin, Tmax, Solar, Wind, Relative_Humidity
## Wind              Year, Tmin, Tmax, Solar, Wind, Relative_Humidity
## Relative_Humidity                    Year, Tmin, Tmax, Solar, Wind
```

- Residual Standard Error is 1.884.
- Adjusted R-squared is 0.6065.
- All GVIF^(1/(2*Df)) are less than or near 2 i.e. all VIF is less than or near 4, hence acceptable.

Multiple Logarithmic Regression Models

1. Multiple Logarithmic (MLG) including all variables

```
##
## Call:
## lm(formula = Precipitation ~ Year + Tmin + Tmax + log(Solar) +
##     log(Wind) + Relative_Humidity, data = training_set)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.8108 -0.9713 -0.1445  0.7563 30.5596
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         7.531096   4.854445   1.551    0.121
```

```
## Year                0.021436   0.002444   8.771   <2e-16 ***
## Tmin                0.222556   0.021075  10.560   <2e-16 ***
## Tmax               -0.198626   0.022468  -8.840   <2e-16 ***
## log(Solar)        -15.359997   0.389821 -39.403   <2e-16 ***
## log(Wind)          -2.503994   0.124848 -20.056   <2e-16 ***
## Relative_Humidity   5.656737   0.358143  15.795   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.966 on 6235 degrees of freedom
## Multiple R-squared:  0.5722, Adjusted R-squared:  0.5718
## F-statistic:  1390 on 6 and 6235 DF,  p-value: < 2.2e-16
```

```
##              Year              Tmin              Tmax        log(Solar)
##          1.035885         66.091742         78.801366          1.592621
##         log(Wind) Relative_Humidity
##          2.039372          3.586881
```

- Residual Standard Error is 1.966.
- Adjusted R-squared is 0.5718.
- VIF of *Tmin* & *Tmax* are very high.

2. Multiple Logarithmic without Tmin

```
##
## Call:
## lm(formula = Precipitation ~ Year + Tmax + log(Solar) + log(Wind) +
##     Relative_Humidity, data = training_set)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.3843 -1.0039 -0.1537  0.7816 30.4328
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3.013052   4.878216   0.618    0.537
## Year                 0.022705   0.002463   9.220   <2e-16 ***
## Tmax                 0.036060   0.003337  10.805   <2e-16 ***
## log(Solar)         -16.500421   0.377868 -43.667   <2e-16 ***
## log(Wind)           -1.804076   0.106737 -16.902   <2e-16 ***
## Relative_Humidity    8.270922   0.261099  31.677   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.983 on 6236 degrees of freedom
## Multiple R-squared:  0.5646, Adjusted R-squared:  0.5642
## F-statistic:  1617 on 5 and 6236 DF,  p-value: < 2.2e-16
```

```
##              Year              Tmax        log(Solar)         log(Wind)
##          1.033380          1.708236          1.470394          1.464628
## Relative_Humidity
##          1.873207
```

- Residual Standard Error is 1.983.
- Adjusted R-squared is 0.5642.
- All VIF are within acceptable range.

3. Multiple Logarithmic without Tmax

```
##
## Call:
## lm(formula = Precipitation ~ Year + Tmin + log(Solar) + log(Wind) +
##     Relative_Humidity, data = training_set)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.4276 -1.0006 -0.1586  0.7735 30.5457
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         3.370143   4.861368   0.693    0.488
## Year                0.022405   0.002457   9.120   <2e-16 ***
## Tmin                0.038279   0.003122  12.261   <2e-16 ***
## log(Solar)        -16.204697   0.380259 -42.615   <2e-16 ***
## log(Wind)          -1.876698   0.103356 -18.158   <2e-16 ***
## Relative_Humidity   8.027968   0.238778  33.621   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.978 on 6236 degrees of freedom
## Multiple R-squared:  0.5669, Adjusted R-squared:  0.5665
## F-statistic:  1632 on 5 and 6236 DF,  p-value: < 2.2e-16
```

```
##              Year              Tmin        log(Solar)         log(Wind)
##          1.033800          1.432720          1.496933          1.380586
## Relative_Humidity
##          1.574907
```

- Residual Standard Error is 1.978.
- Adjusted R-squared is 0.5665.
- All VIF are within acceptable range.

4. Multiple Logarithmic with Interaction Term

```
##
## Call:
## lm(formula = Precipitation ~ Year + Tmin + Tmax + log(Solar) +
##     log(Wind) + Relative_Humidity + Tmin:Tmax, data = training_set)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.1606 -0.9817 -0.1514  0.7638 30.0631
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
```

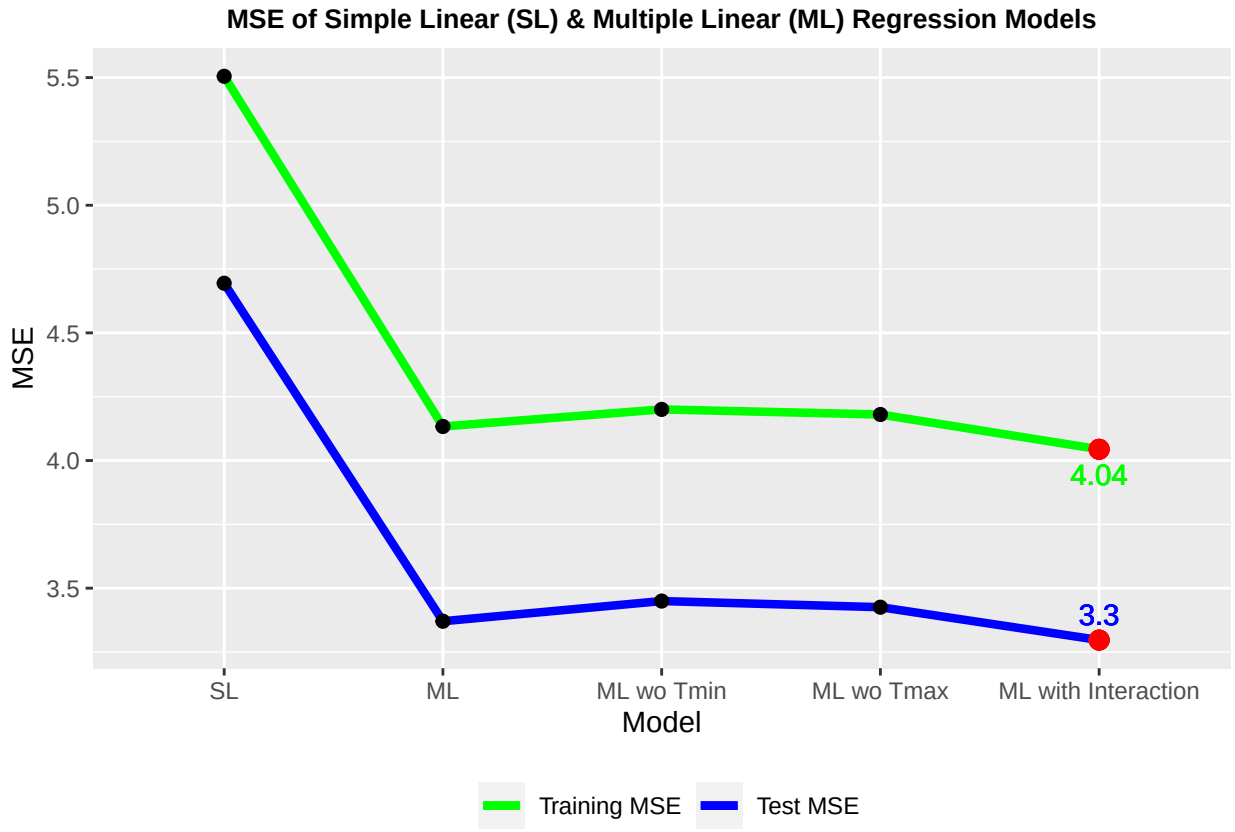
```
## (Intercept)        1.754e+01  4.831e+00   3.631 0.000285 ***
## Year               1.994e-02  2.408e-03   8.281  < 2e-16 ***
## Tmin               2.371e-01  2.077e-02  11.415  < 2e-16 ***
## Tmax              -3.285e-01  2.395e-02 -13.717  < 2e-16 ***
## log(Solar)        -1.680e+01  3.969e-01 -42.313  < 2e-16 ***
## log(Wind)         -3.448e+00  1.399e-01 -24.655  < 2e-16 ***
## Relative_Humidity  3.794e+00  3.764e-01  10.080  < 2e-16 ***
## Tmin:Tmax          4.743e-03  3.353e-04  14.146  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.935 on 6234 degrees of freedom
## Multiple R-squared:  0.5855, Adjusted R-squared:  0.5851
## F-statistic:  1258 on 7 and 6234 DF,  p-value: < 2.2e-16
```

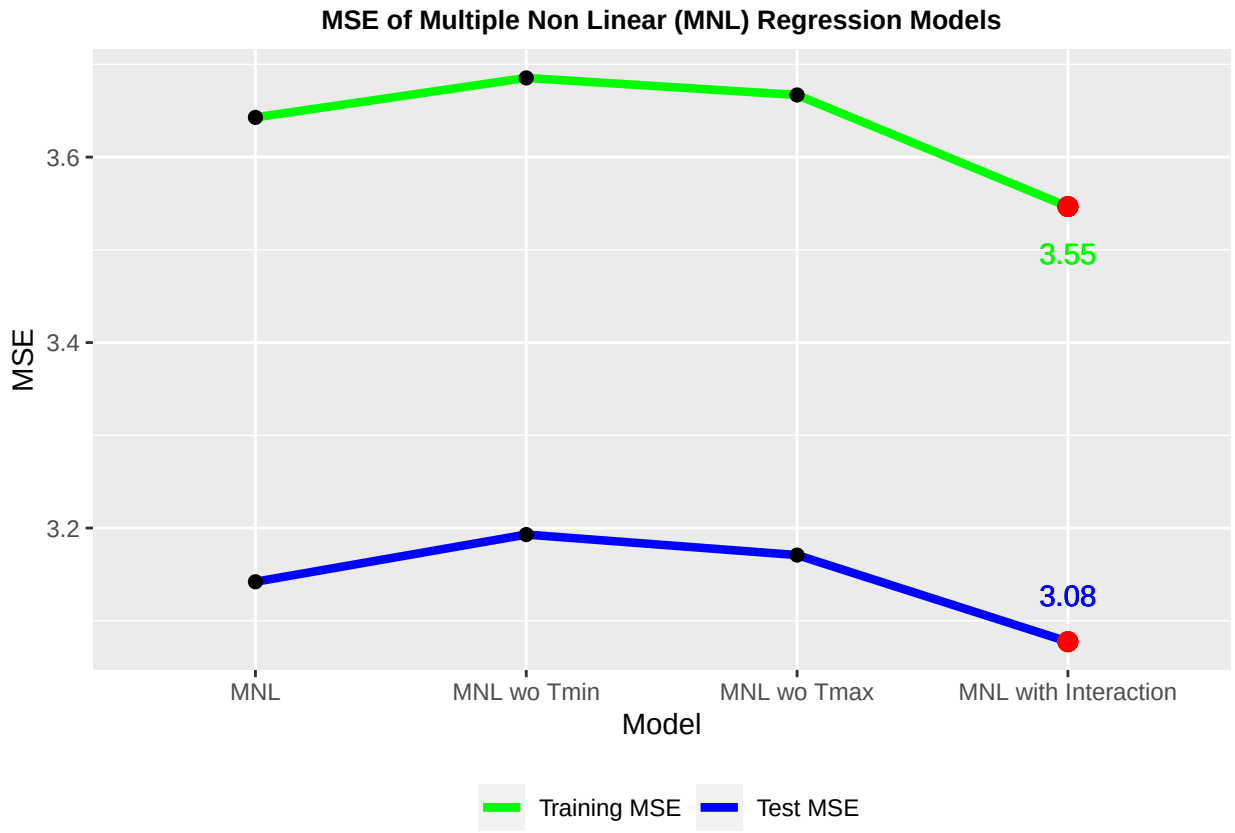
```
## GVIFs computed for predictors
```

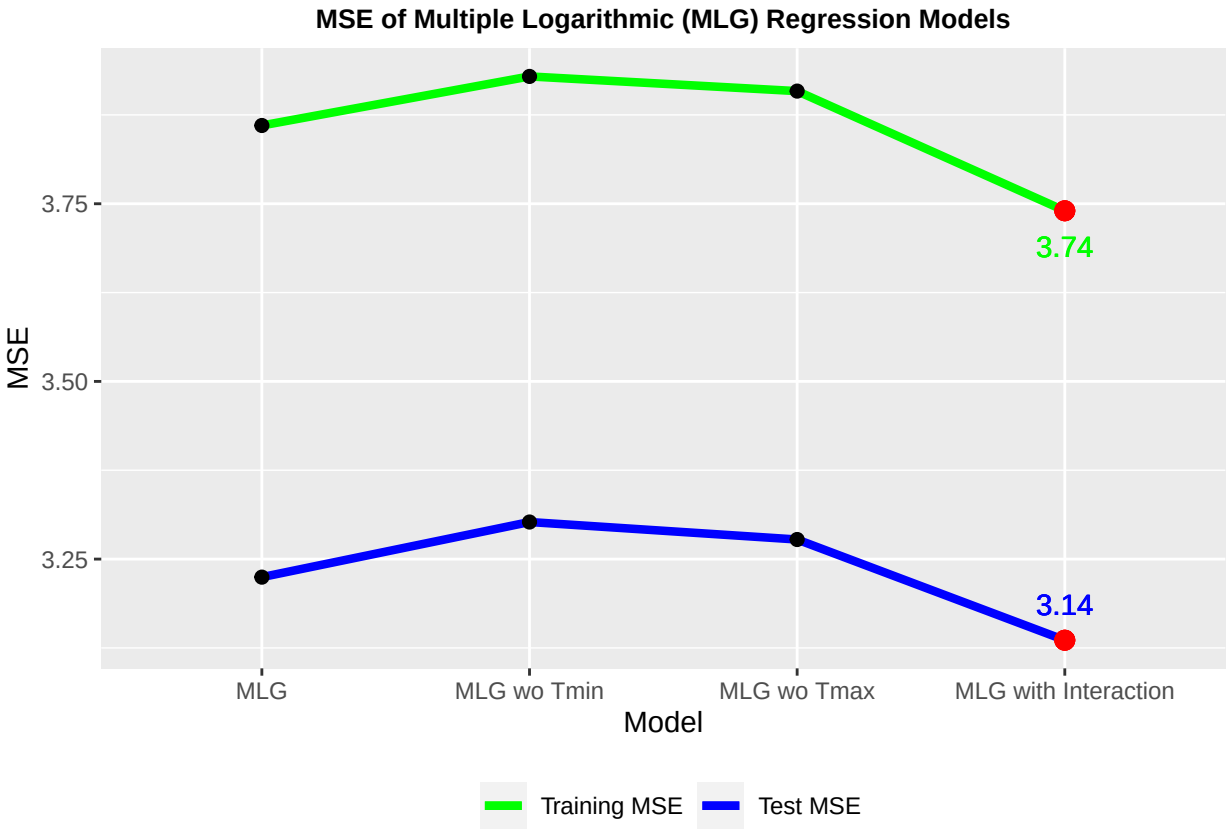
```
##                          GVIF Df GVIF^(1/(2*Df)) Interacts With
## Year                 1.037879  1        1.018763           --
## Tmin                 6.383258  3        1.361989           Tmax
## Tmax                 6.383258  3        1.361989           Tmin
## Solar             1537.971802  0             Inf           --
## Wind              1537.971802  0             Inf           --
## Relative_Humidity    4.087530  1        2.021764           --
##                                                Other Predictors
## Year                    Tmin, Tmax, Solar, Wind, Relative_Humidity
## Tmin                          Year, Solar, Wind, Relative_Humidity
## Tmax                          Year, Solar, Wind, Relative_Humidity
## Solar             Year, Tmin, Tmax, Solar, Wind, Relative_Humidity
## Wind              Year, Tmin, Tmax, Solar, Wind, Relative_Humidity
## Relative_Humidity                    Year, Tmin, Tmax, Solar, Wind
```

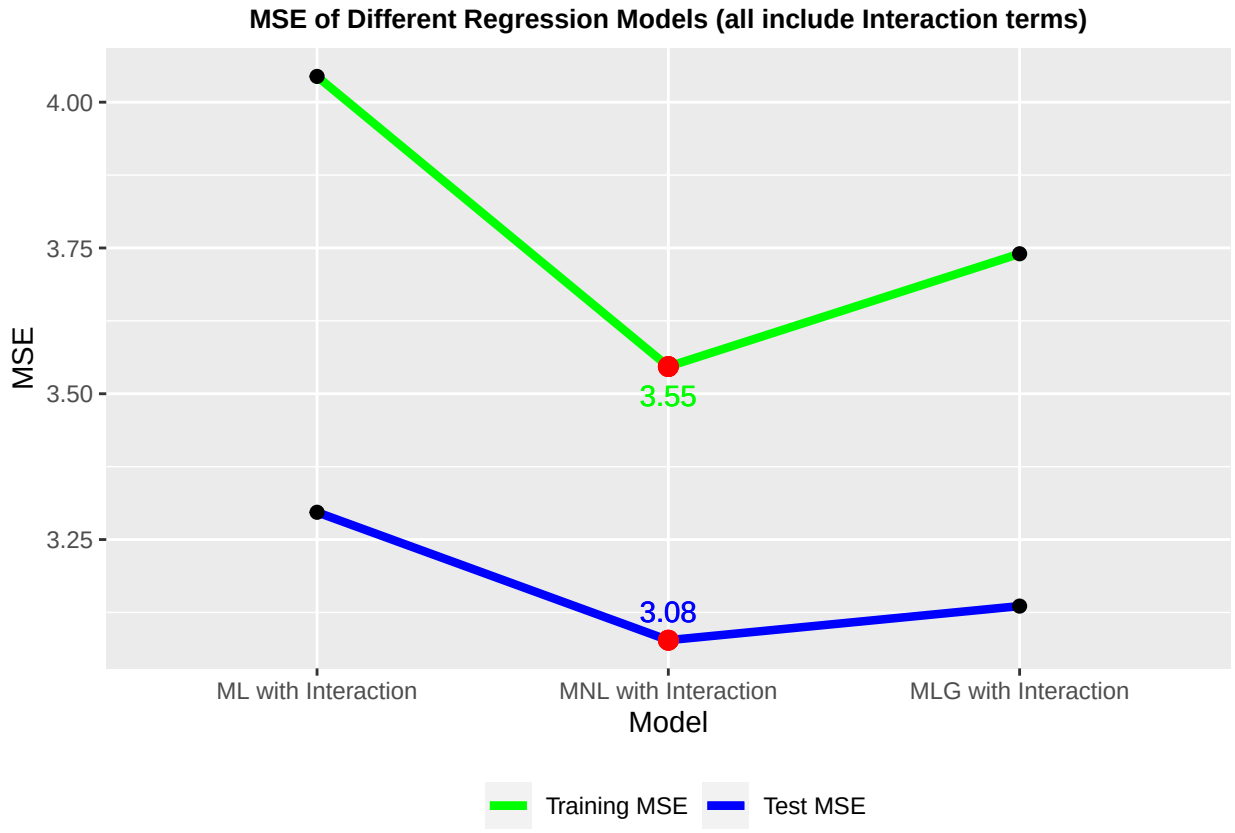
- Residual Standard Error is 1.935.
- Adjusted R-squared is 0.5851.
- All GVIF^(1/(2*Df)) are less than or near 2 i.e. all VIF is less than or near 4, hence acceptable.

Plotting Training & Test Set MSE of different Models






MSE of Multiple Logarithmic (MLG) Regression Models
MSE
3.75
3.50
3.25
3.74
3.14
MLG
MLG wo Tmin
MLG wo Tmax
MLG with Interaction
Model
Training MSE
Test MSE



- Multiple Non-Linear Regression Model including Interaction term gives the lowest Training & Test MSE.

Preparing & Plotting the Selected Model on complete data set

```
##
## Call:
## lm(formula = Precipitation ~ Year + Tmin + Tmax + I(Solar^-1) +
##     I(Wind^-1) + Relative_Humidity + Tmin:Tmax, data = df_cleaned)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.2473 -0.9487 -0.1524  0.7216 23.1874
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -5.410e+01  4.185e+00  -12.93   <2e-16 ***
## Year               1.884e-02  2.063e-03    9.13   <2e-16 ***
## Tmin               1.760e-01  1.760e-02   10.00   <2e-16 ***
## Tmax              -2.346e-01  1.984e-02  -11.83   <2e-16 ***
## I(Solar^-1)        3.198e+02  5.785e+00   55.28   <2e-16 ***
## I(Wind^-1)         6.264e+00  2.495e-01   25.10   <2e-16 ***
## Relative_Humidity  4.454e+00  3.240e-01   13.75   <2e-16 ***
## Tmin:Tmax          3.973e-03  2.799e-04   14.20   <2e-16 ***
## ---
```

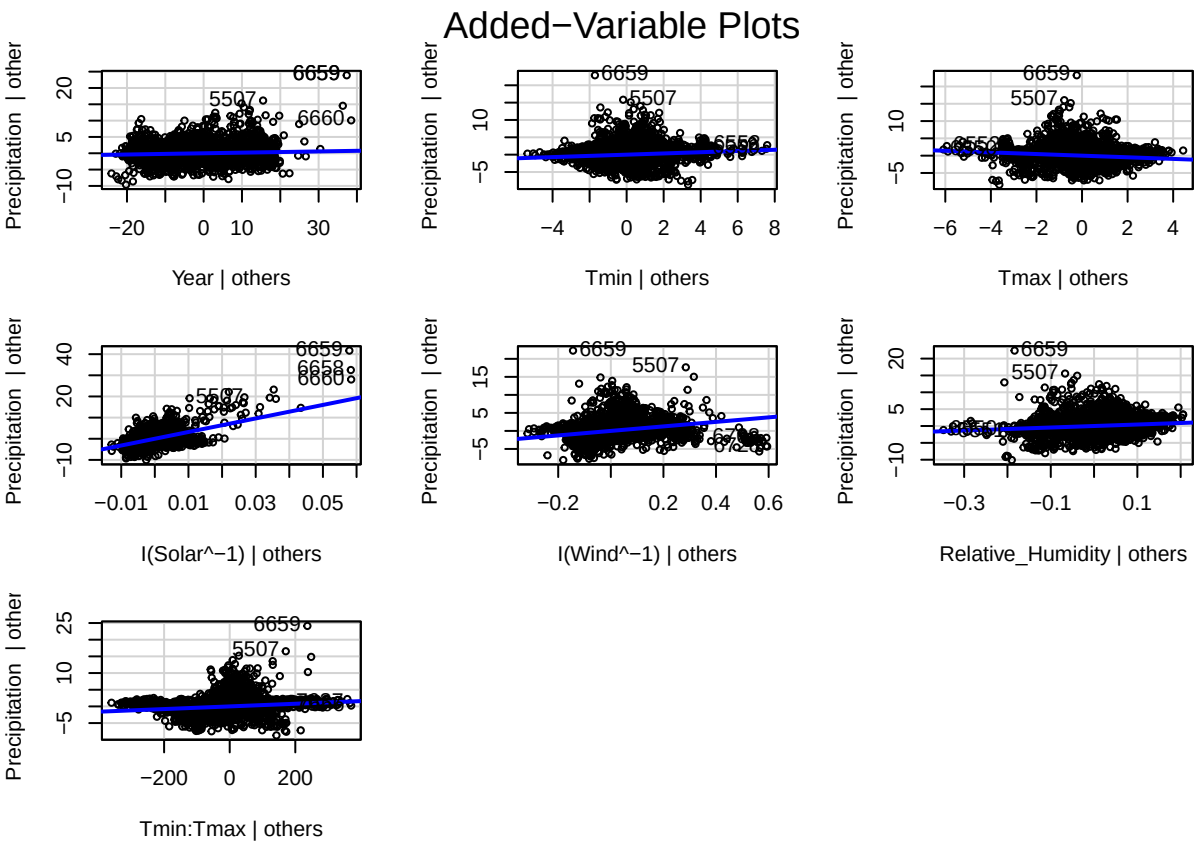
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.86 on 7732 degrees of freedom
## Multiple R-squared:  0.6065, Adjusted R-squared:  0.6061
## F-statistic:  1702 on 7 and 7732 DF,  p-value: < 2.2e-16
```

```
## GVIFs computed for predictors
```

```
##                          GVIF Df GVIF^(1/(2*Df)) Interacts With
## Year                 1.028099  1        1.013952           --
## Tmin                 5.476219  3        1.327639         Tmax
## Tmax                 5.476219  3        1.327639         Tmin
## Solar             1288.006703  0             Inf           --
## Wind              1288.006703  0             Inf           --
## Relative_Humidity    4.044735  1        2.011153           --
##                                                 Other Predictors
## Year                     Tmin, Tmax, Solar, Wind, Relative_Humidity
## Tmin                           Year, Solar, Wind, Relative_Humidity
## Tmax                           Year, Solar, Wind, Relative_Humidity
## Solar             Year, Tmin, Tmax, Solar, Wind, Relative_Humidity
## Wind              Year, Tmin, Tmax, Solar, Wind, Relative_Humidity
## Relative_Humidity                    Year, Tmin, Tmax, Solar, Wind
```

- All model coefficients are significant.
- The overall model is also significant.
- Mean Squared Error is 3.45.
- Residual Standard Error is 1.86.
- Adjusted R-squared is 0.6061.
- All GVIF^(1/(2*Df)) are less than or near 2 i.e. all VIF is less than or near 4, hence acceptable.

The final model plots are shown below:


Added−Variable Plots
Precipitation | others
Year | others
Tmin | others
Tmax | others
I(Solar^−1) | others
I(Wind^−1) | others
Relative_Humidity | others
Tmin:Tmax | others
6659
5507
6660
6658