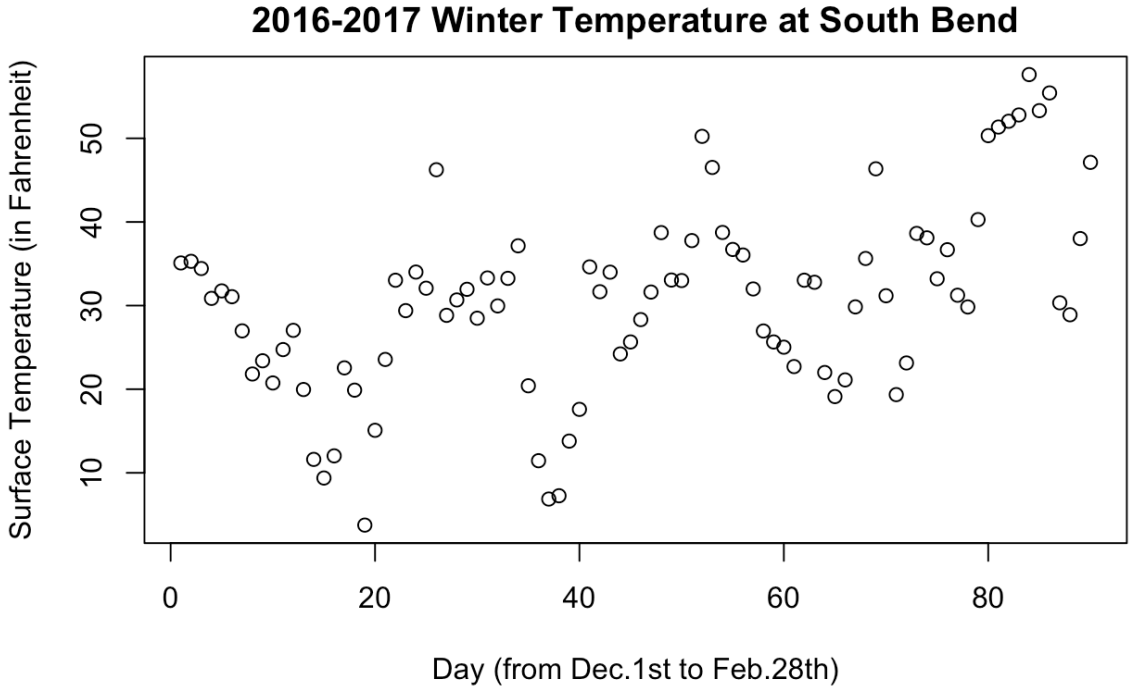
Weather Prediction Game Report

Objective:

Given a set of data from MERRA2(NASA) with 9 seemed related variables such as albedo, precipitation, humidity, snowfall rate, and temperature of all the locations during the winter of 2016. The goal is to build the best and simplest model for temperature prediction for the winter of 2017.

Method:

I started with extracting the necessary information that I need, such as all the 9 predictors located at South Bend (lon: 119, lat: 84). Then I plot the data to see if any patterns or unusual things appearing.



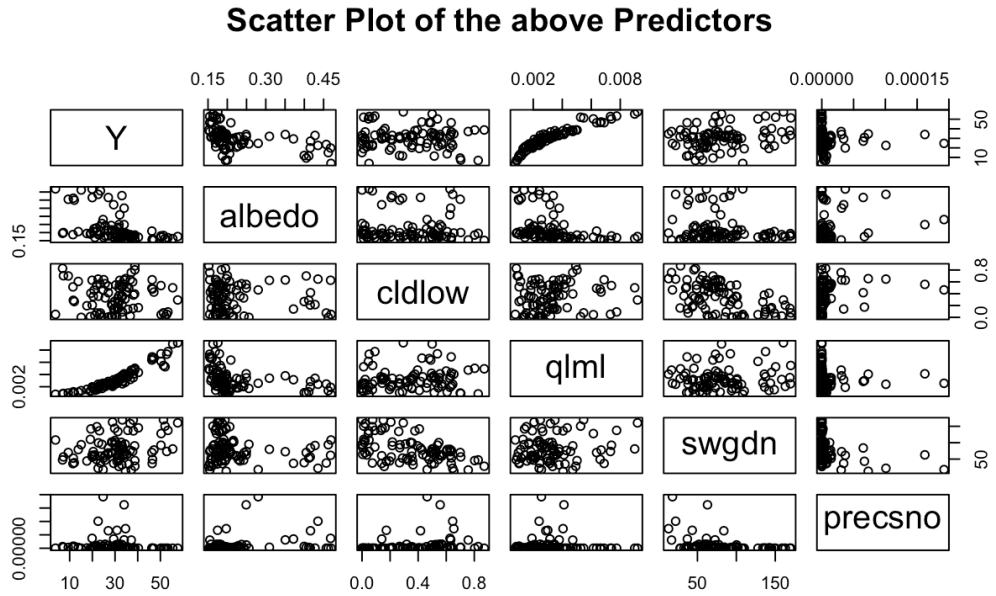
This plot of the winter temperature at South Bend makes sense because the temperature is arranging from $0°F - 60°F$, and there is no substantial temperature drop in successive day. Although the tendency is going up as a whole, intuitively, the end of Feb. can be viewed as early spring, so I am not surprised with this trend. One can take a log to clear out the trend. However, consider there has only 90 points, which is not a really big data set, I decide to ignore such minor trend effects.

To detect if there is a linear relationship associate with other available data, I plot a scatter plot which contains several predictors based on their physical meaning that I think it would be useful, which explained as the following:

1. Albedo is a measure of the diffuse reflection of sunlight out of the total sunlight received by earth. My intuition is to expect a negative effect with temperature because more diffuse reflection means less receiving of sunlight, which will make the temperature drop.
2. CLDLOW is a measure of clouds such as stratus clouds. I choose it because it makes sense that the temperature would tend to be low in a cloudy day.

3. QLML, essentially the humidity, as inference from Midterm 2.
4. SWGDN is the surface incoming shortwave flux. From physics, we know that all objects emit electromagnetic radiation, and the hot objects emit more of their light at short wavelengths, visa versa for cold object.
5. PRECSNO here measures the snowfall rate. My intuition is that snowfall came from cold day.

From the scatter plot below, I see some obviously linear relationship such as temperature(Y) against humidity(qlml), and others might or might not be linear, which I can run several tests to see.



I ran a global F-test to confirm the existence of linear relation among 9 predictors:

$$H0: \beta 1 = \beta 2 = \cdots = \beta 9 = 0$$
$$Ha: at\ least\ 1\ \beta s\ is\ not\ 0$$

I got a F-statistics, 113, on 9 and 80 degree of freedom, and this corresponds with a significant p-value <2e-16. It confirms that there is a linear relationship. We see that on average the temperature will increase/decrease when one of the significant predictors increasing/decreasing by one unit, while keeping the other predictors constant. However, with 9 predictors, we also see that some of them are not significant, this means I can build a better model that uses less predictors.

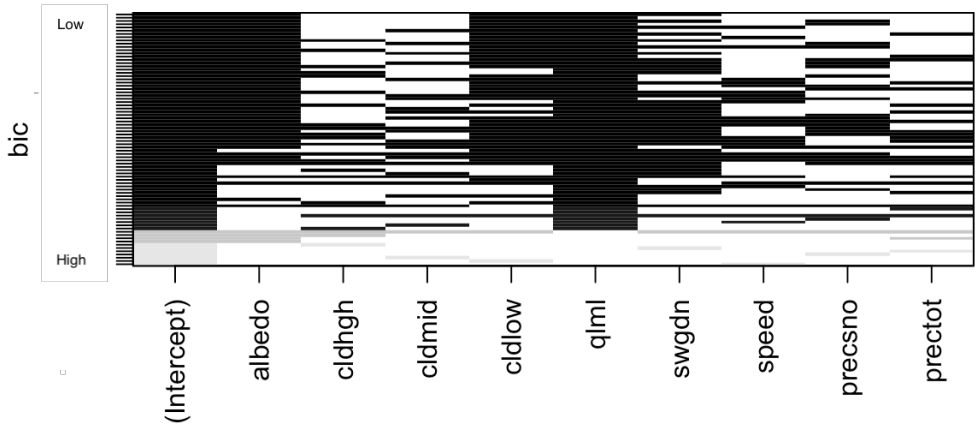
```
Call:
lm(formula = Y ~ ., data = table)

Residuals:
    Min      1Q  Median      3Q     Max
-10.633  -1.885   0.717   1.842   5.561

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.84e+01   2.95e+00    6.23  2.1e-08 ***
albedo      -2.06e+01   4.62e+00   -4.46  2.7e-05 ***
cldhgh       5.58e-01   1.72e+00    0.32    0.747
cldmid      -1.14e+00   2.50e+00   -0.45    0.650
cldlow      -5.58e+00   2.15e+00   -2.60    0.011 *
qlml         5.42e+03   2.51e+02   21.56  < 2e-16 ***
swgdn        2.49e-02   1.35e-02    1.85    0.069 .
speed       -1.41e-01   1.56e-01   -0.90    0.369
precsno      2.52e+04   1.62e+04    1.55    0.124
prectot     -3.41e+03   1.06e+04   -0.32    0.749
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.21 on 80 degrees of freedom
Multiple R-squared:  0.927,     Adjusted R-squared:  0.919
F-statistic:  113 on 9 and 80 DF,  p-value: <2e-16
```

Further investigate in dropping some predictors from full model is necessary. Since there are 10 predictors (including interception), by consider each predictor “to be or not to be” in a model, there would be $2^{10} = 1024$ possible choose for model selection. A brute force method using “regsubsets” function in R with bic as criteria by default indicates that “albedo”, “coldlow”, and “qlml” should keep in the model.



As suggested from the above R commend, I build a multiple regression model with three predictors, namely, “albedo”, “cldlow”, and “qlml”. I got the results as the following:

```
Call:
lm(formula = Y ~ albedo + cldlow + qlml, data = table)

Residuals:
    Min      1Q  Median      3Q     Max
-10.303  -1.844   0.543   1.785   6.786

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    19.65       1.50   13.10  < 2e-16 ***
albedo        -20.05       4.33   -4.63  1.3e-05 ***
cldlow         -7.51       1.47   -5.12  1.9e-06 ***
qlml         5499.62     206.91   26.58  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.27 on 86 degrees of freedom
Multiple R-squared:  0.918,     Adjusted R-squared:  0.916
F-statistic:  323 on 3 and 86 DF,  p-value: <2e-16
```

Here, besides all three predictors together are significant, each individual is also significant. The adjusted R-squared indicates that the model has captured 91.6% of the data, comparing to the full model (91.9%), only lost 0.3% data unexplained, so I think the trade-off is okay here by dropping 6 predictors.

Multicollinearity Issue:

However, among these three predictors, they might be not independent (i.e. they might combine to have an effect on temperature). To see if there is an interaction effect, I checked the correlation matrix among these three predictors.

```
         albedo  cldlow   qlml
albedo  1.0000 -0.0508 -0.387
cldlow -0.0508  1.0000  0.107
qlml   -0.3871  0.1067  1.000
```

fig. correlation matrix

It seems like I need to investigate the interaction of “albedo” and “qlml” because comparing other correlation entries around 0, the number -0.387 signals me a moderate correlation between these two predictors.

3Predictor+1Interaction Model:

```
Call:
lm(formula = Y ~ albedo + cldlow + qlml + albedo * qlml, data = table)

Residuals:
   Min     1Q Median     3Q    Max
-7.616 -1.361  0.267  1.636  7.540

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    27.11       1.77   15.32  < 2e-16 ***
albedo        -59.23       7.46   -7.94  7.5e-12 ***
cldlow         -8.01       1.24   -6.47  6.1e-09 ***
qlml         2140.93     584.50    3.66  0.00043 ***
albedo:qlml 18484.55    3070.51    6.02  4.3e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.76 on 85 degrees of freedom
Multiple R-squared:  0.943, Adjusted R-squared:  0.94
F-statistic:  350 on 4 and 85 DF,  p-value: <2e-16
```

I got same results as before with interaction term, which is also significant, and I have noticed that the adjusted R-squared got improved up to 94%, whereas before is 91.6%. This means adding interaction term, the model can explain additional 2% of the data.

Another Approach:

To check if this is the “best” model, I also used several other model selection criteria such as Mallow’s Cp value to see if I still get the same model selection results:

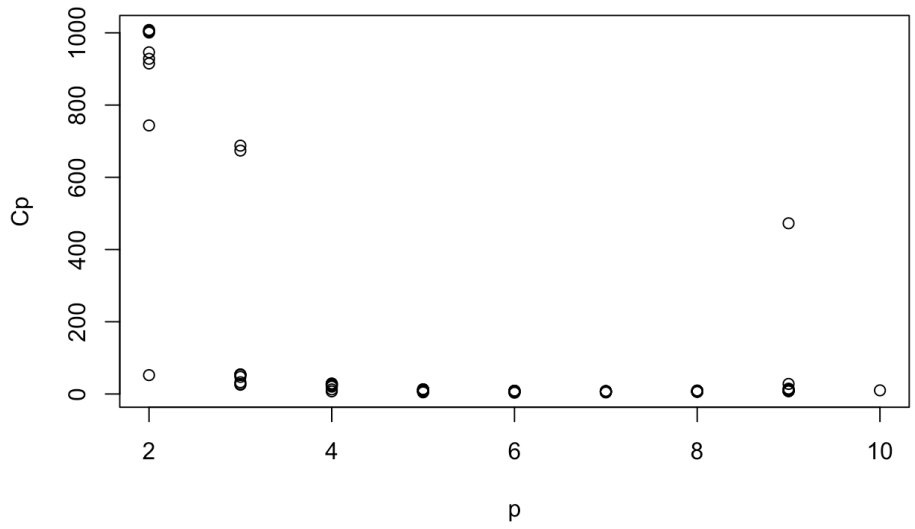


fig. Cp values of 10 predictors

1	2	3	4	5	6	7	8	9
TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	TRUE	FALSE

The mallow’s Cp indicates that “albedo”, “cldlow”, “qlml”, “swgdn”, “precsno’ are good indicators in predicting temperature in South Bend. To confirm this finding, besides using

such brute force method in 10 predictors, I can use another test such as forward stepwise via AIC in both direction, which is locally optimal rather globally optimal the results.

```
Call:
lm(formula = Y ~ qlml + swgdn + albedo + cldlow + precsno, data = table)

Residuals:
    Min      1Q  Median      3Q     Max
-10.513  -2.113   0.516   1.863   5.914

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.69e+01   1.93e+00    8.77  1.7e-13 ***
qlml         5.43e+03   2.02e+02   26.92  < 2e-16 ***
swgdn        2.62e-02   1.05e-02    2.50  0.01448 *
albedo      -1.98e+01   4.36e+00   -4.54  1.8e-05 ***
cldlow      -5.84e+00   1.71e+00   -3.41  0.00099 ***
precsno      2.13e+04   1.20e+04    1.78  0.07817 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.16 on 84 degrees of freedom
Multiple R-squared:  0.926,     Adjusted R-squared:  0.921
F-statistic:  209 on 5 and 84 DF,  p-value: <2e-16
```

Above through stepwise method via AIC, we see that these 5 predictors all together are significant, but “precsno” individually is not significant, so I decided to drop this insignificant predictor for the model.

Multicollinearity Issue:

Next step is to detect multicollinearity in my new model with four predictors (“albedo”, “cldlow”, “qlml”, “swgdn”). I checked the correlation matrix of these four predictors, and I found that “albedo” with “qlml” and “cldlow” with “swgdn” are moderately correlated. Using the same reasoning as before, I think these two groups of predictors should have combine effect in the temperature to improve my new model.

	albedo	cldlow	qlml	swgdn
albedo	1.0000	-0.0508	-0.387	-0.192
cldlow	-0.0508	1.0000	0.107	-0.520
qlml	-0.3871	0.1067	1.000	0.118
swgdn	-0.1922	-0.5203	0.118	1.000

fig. correlation matrix

4Predictor+2Interaction Model:

To test out my thought, I run a multiple linear regression model with two interactions. According to the principal of marginality, assuming no interaction, the main effect of “cldlow” is insignificant. However, the interaction of “swgdn” and “cldlow” is significant, this means I cannot explain the interaction of “swgdn” and “cldlow” while leaving “cldlow” absent, despite its insignificant main effect, so I decided to include “cldlow” predictor here.

```
Call:
lm(formula = Y ~ albedo + qlml + swgdn + cldlow + albedo * qlml +
    cldlow * swgdn, data = table)

Residuals:
   Min     1Q Median     3Q    Max
-7.990 -0.665  0.434  1.441  5.921

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.31e+01   2.09e+00   11.05  < 2e-16 ***
albedo       -5.84e+01   7.16e+00   -8.16  3.1e-12 ***
qlml          1.95e+03   5.61e+02    3.47  0.00083 ***
swgdn         3.96e-02   1.21e-02    3.26  0.00160 **
cldlow       -9.97e-01   2.86e+00   -0.35  0.72803
albedo:qlml   1.93e+04   2.95e+03    6.53  4.9e-09 ***
swgdn:cldlow -7.25e-02   3.61e-02   -2.01  0.04762 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.62 on 83 degrees of freedom
Multiple R-squared:  0.949,     Adjusted R-squared:  0.946
F-statistic:  260 on 6 and 83 DF,  p-value: <2e-16
```

Model Comparison:

Comparing with all the models I have built so far, using R_adj, AIC, and BIC as judgments, the model with four predictors and two interactions indicates the “best model” because it has the highest R_adj value, meaning 94.6% and lowest AIC and BIC values among all other models.

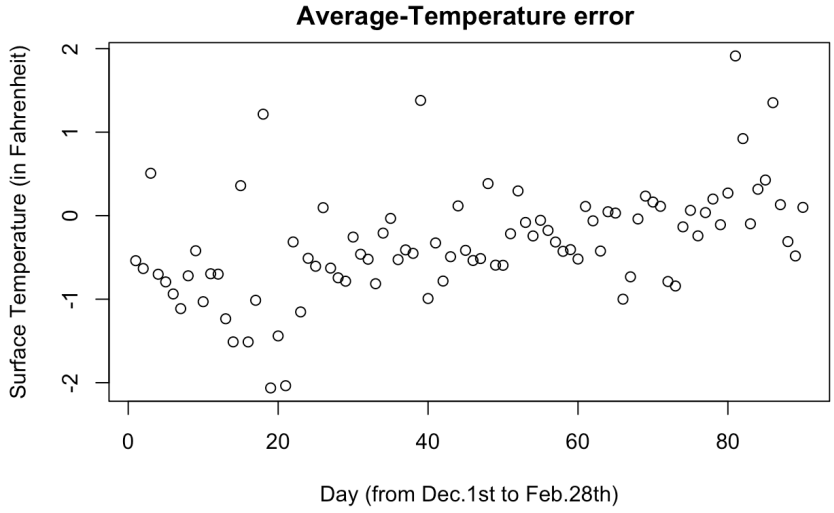
	Full model	3 Predictors Model	3Pred.+1Iner. Model	5 Predictors Model	4Pred.+2 Inter. Model
R_adj	91.9%	91.6%	94%	91.6%	94.6%
AIC	477	475	445	475	438
BIC	504	487	460	490	440

One might argue that the model with three predictors and two interaction is the “simplest” model which also did a good job in explaining 94% of the data. By solely introducing one predictor, “swgdn”, R_adj did not improve at all but reduce, and one has to also adding its interaction with “cldlow” to beats the “simplest” model just by capturing 0.6% of the data. It seems to be regardless to adding two extra terms while only improves minor.

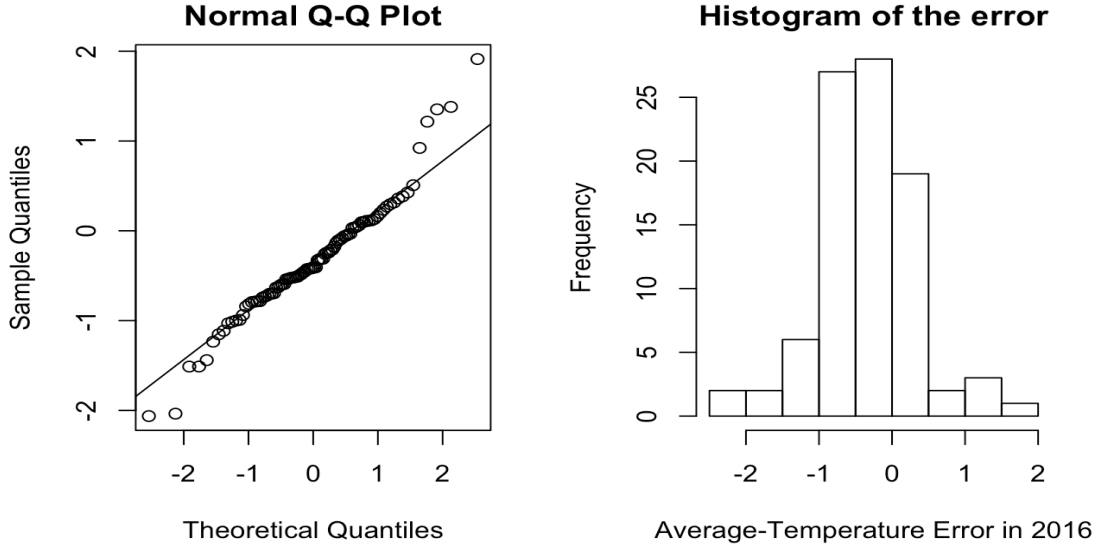
To test out that if it is necessary to having this extra predictor, and decide the “best and simplest” model, I would like to put these model into the prediction of next year (2017) winter data. Although the temperature of South Bend is unavailable, but we know the temperature around its location. I can use these known temperatures to approximate the temperature in South Bend. See below diagram as how I use locations to estimates South Bend temperature:

(118, 85)	(119, 85)	(120, 85)
(118, 84)	South Bend ? ? ?	(120,84)
(118, 83)	(119, 83)	(120, 83)

Since the temperature of these locations around South Bend might be highly correlated due to their close related geographical location. Without giving more information such as the direction of wind, it is subtle to judge which location would contribute more in temperature, so I choose to take the average of these eight locations temperature around South Bend (Left, TopLeft, Top, TopRight, Right, BottomRight, Bottom, and BottomLeft) through training data, and compare the temperature average with the true temperature of South Bend in 2016 winter. I plot the error below, and found it is almost around zero mean with some minor fluctuation.



The qq plot and histogram further indicates the normality of errors producing by taking the average temperatures.



By demonstrating the "standard error", $\sqrt{\sum(\ fit - Y)^2/90}$, of the two models with average temperature, Y=Y_avg, versus with true temperature, Y, we see that their "standard error" are very close together. Hence, average temperature can be used as a good indicator for the following year temperature reference.

	3Pred.+1Iner. Model	4Pred.+2 Inter. Model
Standard Error of Y	2.68	2.54
Standard Error of Y_avg	2.62	2.52

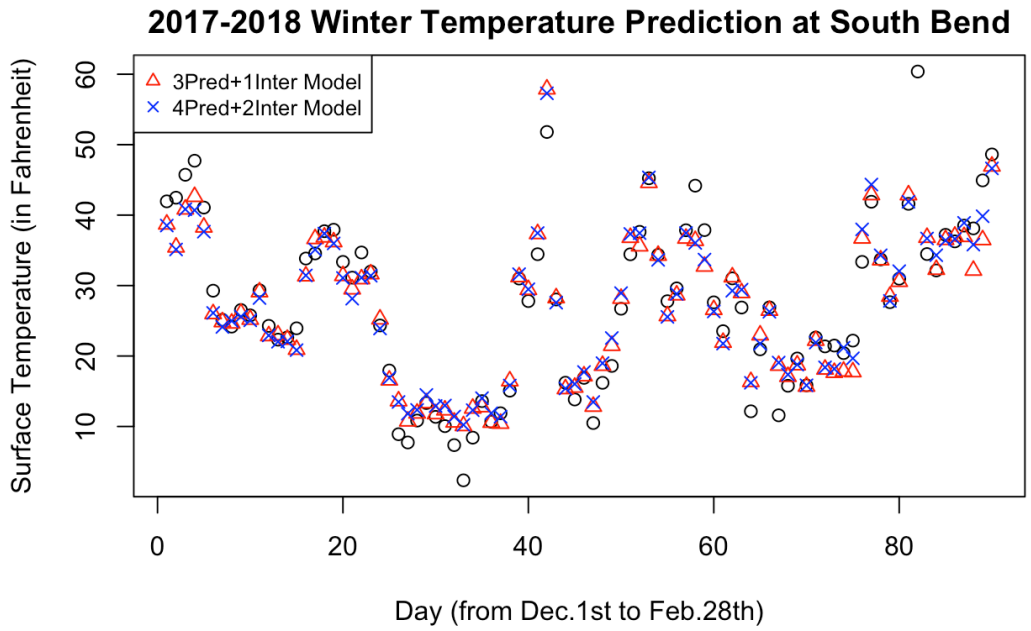
I take the average of the same eight locations through the prediction data set, use it as the temperature approximation of South Bend in the winter of 2017. To test out which of the above model can produce with the least "standard errors" with respected to the approximated temperature of South Bend by taking its surrounding temperature average. See below for the first 15 days of prediction table:

Day	fit.3Pred_1Inter	fit.4Pred_2Inter	Y_new_avg
1	38.7	38.5	41.96
2	35.4	35.1	42.47
3	40.8	40.8	45.73
4	42.6	40.7	47.72
5	38.3	37.7	41.07
6	26.0	26.1	29.27
7	24.8	24.1	25.14
8	24.6	25.0	24.18
9	26.0	25.6	26.53
10	25.2	25.1	25.77
11	29.1	28.2	29.33
12	22.9	22.9	24.28
13	23.0	22.0	22.32
14	22.3	22.1	22.63
15	20.9	20.8	23.94
…			

Table: Prediction Comparison with average-temperature

From the table, we see that both models do not have much fluctuation around the average temperature. Both models have 2.96 "standard error". I think this is a small "standard error" in temperature, because no one would feel the difference if there are a few degrees variation in Fahrenheit.

For visualization, I plot the prediction points as below, noticed that the "black circle" is the average temperature, the "red triangle" is the prediction from three predictors and one interaction model, and the "blue cross" is the prediction from four predictors and two interactions model. We see that the "red triangle" almost coincide with "blue cross", despite they almost can capture the black circles, except few points. The plots also match up with the general pattern of 2016 temperature plot on the first page.



"Best and Simplest" Model:

Since these two models have the same "standard error", almost identical R_adj. Hence, I would say that the "best and simplest" model is to use "albedo", "cldlow", and "qlml three predictors and one interaction, "albedo:qlml". However, one might argue that we should fully use the information available here, and with nowadays' technology, adding two extra terms in lm commend does not cost that much, especially for such small data set. I would say, with only three predictors, if we can do almost the same thing as with four predictors, there might come to a substantial cost in collecting that additional predictor. As for now, data is not that big, if we are doing prediction in a much bigger data set, the advantages of taking the model that uses less predictors while preserve accuracy would be more obvious. Although, for prediction purpose, multicollinearity might not be a big issue, one can just simply take a model as such: $E(Y) = \beta 0 + \beta 1 * albedo + \beta 2 * cldlow + \beta 3 * qlml$. I think adding the interaction term as: $E(Y) = \beta 0 + \beta 1 * albedo + \beta 2 * cldlow + \beta 3 * qlml + \beta 4 * albedo{:}\ qlml$ improves the model by explaining more than 2.4% of the data up to 94% while still keep it simple; in this case, we can fully use the three predictors' information at hand.