Final Report

Conner Byrd, Eric Han, Ky Hyun, Sara Shao, Alex Shen, Mona Su, Dani Trejo, Steven Yuan

10/12/2021

Introduction

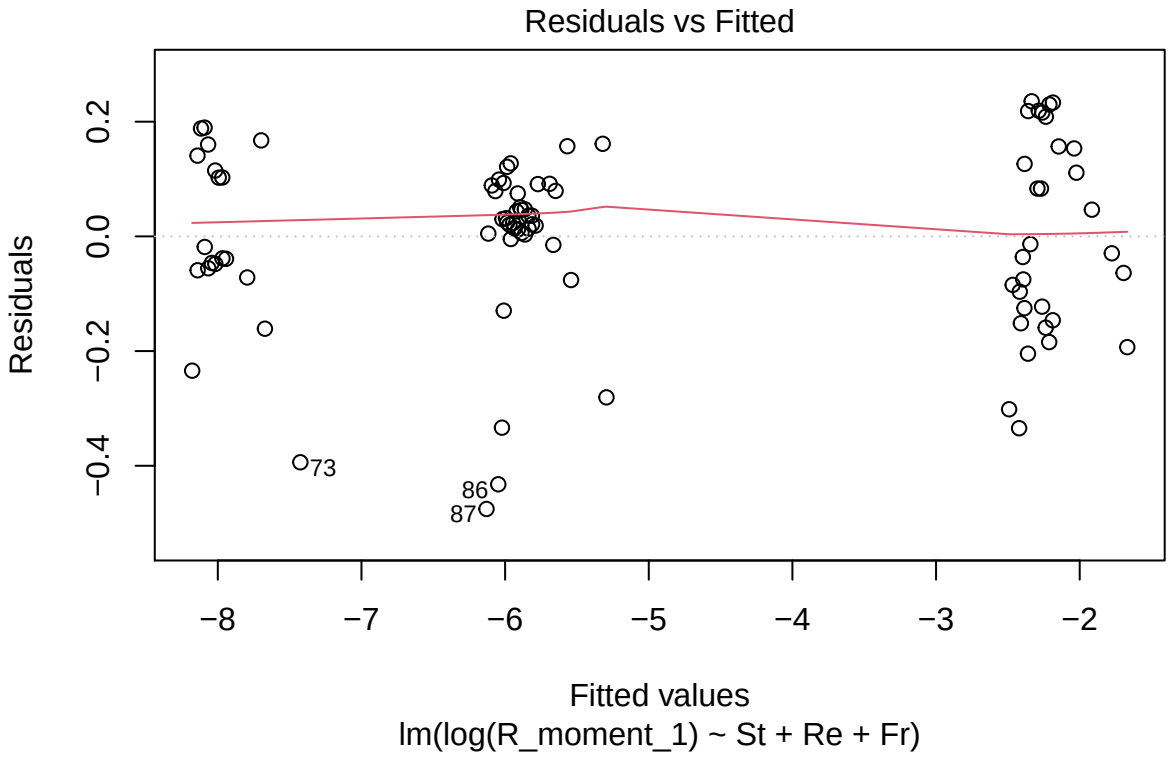
Our key research objectives include understanding and predicting how turbulence affects the dynamics of water droplets and ice crystals (how they collide and mix) in clouds. With our machine learning model, we are trying to infer the volume distribution of clusters within clouds.

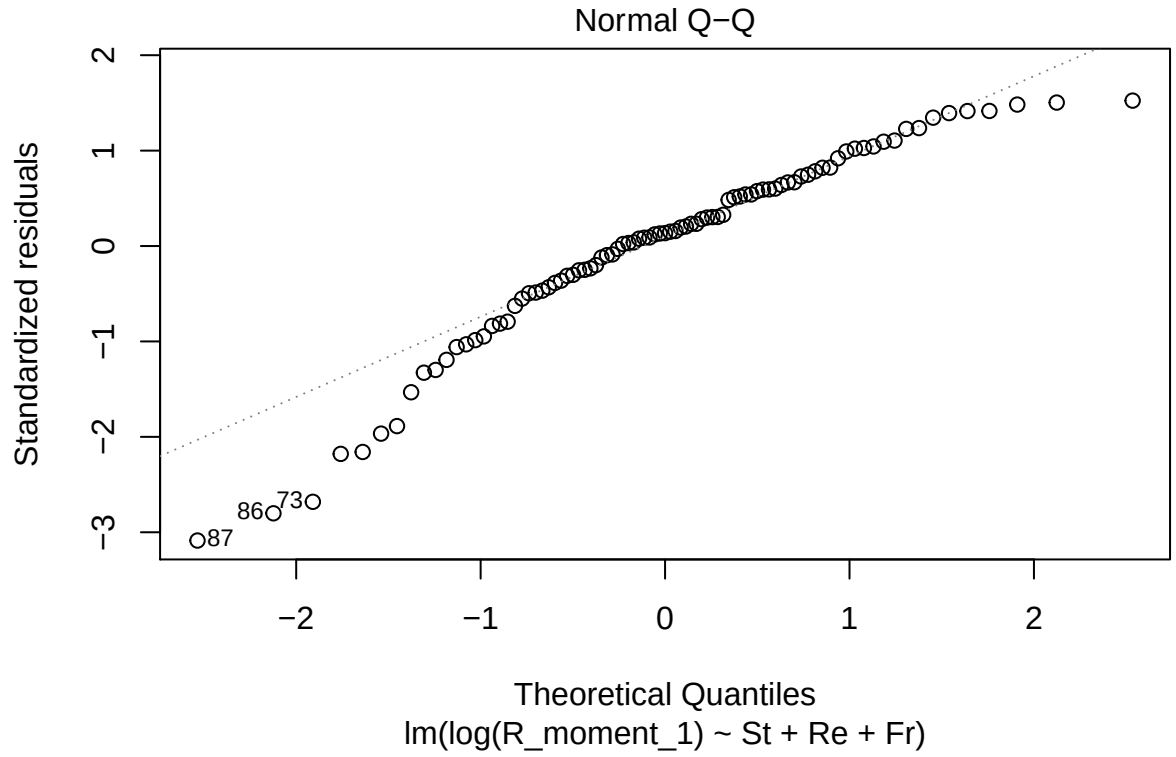
To do this, we began by doing some basic Exploratory Data Analysis on the three predictor variables: Reynolds number (`Re`), gravitational acceleration (`Fr`), and particle characteristic (`St`). (Our graphs are included in Appendix: Section 1)

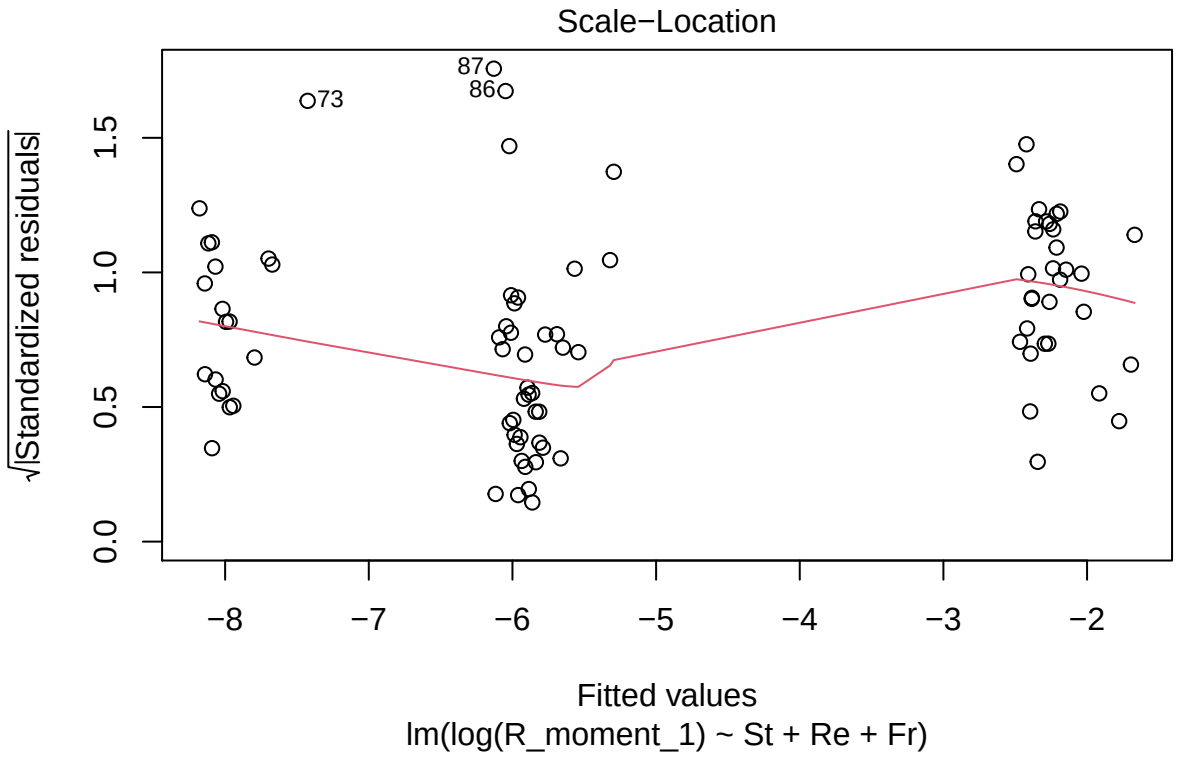
```
##     St  Re    Fr R_moment_1 R_moment_2 R_moment_3 R_moment_4
## 1 0.10 224 0.052 0.00215700  0.1303500   14.37400  1586.5000
## 2 3.00 224 0.052 0.00379030  0.4704200   69.94000 10404.0000
## 3 0.70 224   Inf 0.00290540  0.0434990    0.82200    15.5510
## 4 0.05  90   Inf 0.06352800  0.0906530    0.46746     3.2696
## 5 0.70 398   Inf 0.00036945  0.0062242    0.12649     2.5714
## 6 2.00  90 0.300 0.14780000  2.0068000   36.24900   671.6700
```

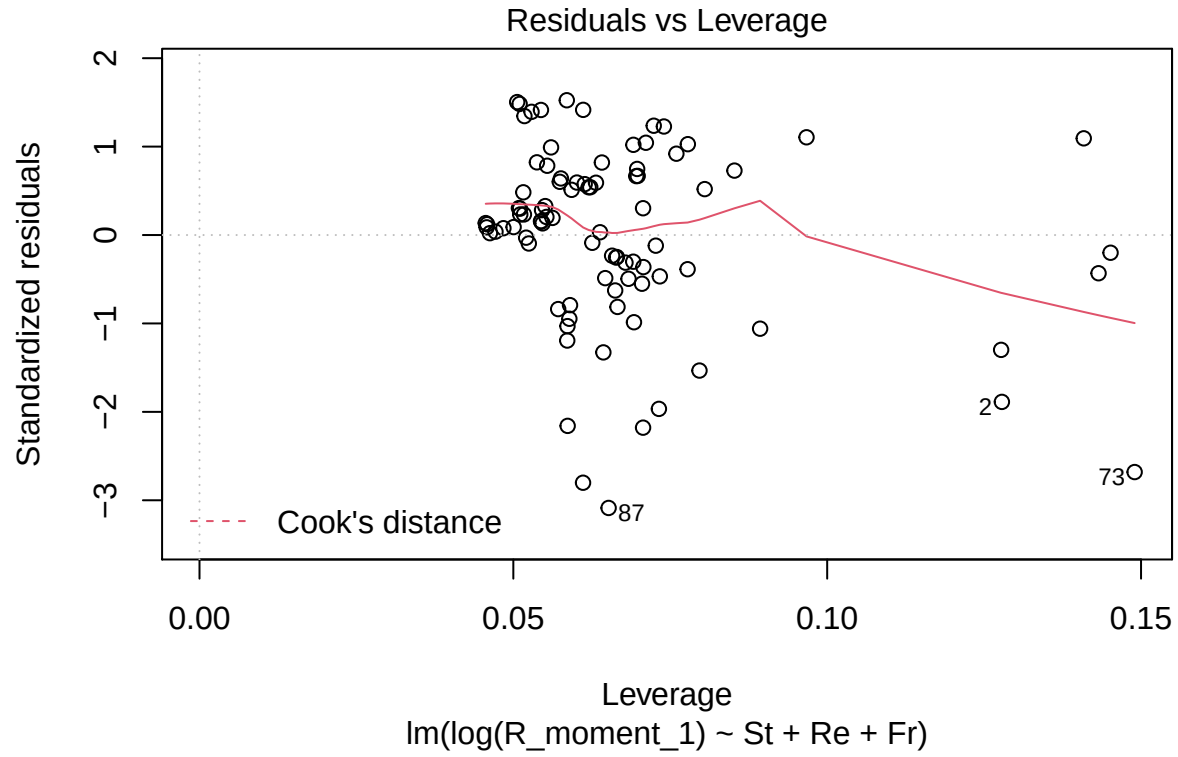
Methodology

BAD LINEAR MODEL SHOWS TRANSFORMATIONS NEEDED AND GAM MIGHT BE A GOOD IDEA


Residuals vs Fitted
Residuals
0.2
0.0
−0.2
−0.4
73
86
87
−8
−7
−6
−5
−4
−3
−2
Fitted values
lm(log(R_moment_1) ~ St + Re + Fr)


Normal Q−Q
Standardized residuals
2
1
0
−1
−2
−3
86
73
87
−2
−1
0
1
2
Theoretical Quantiles
lm(log(R_moment_1) ~ St + Re + Fr)





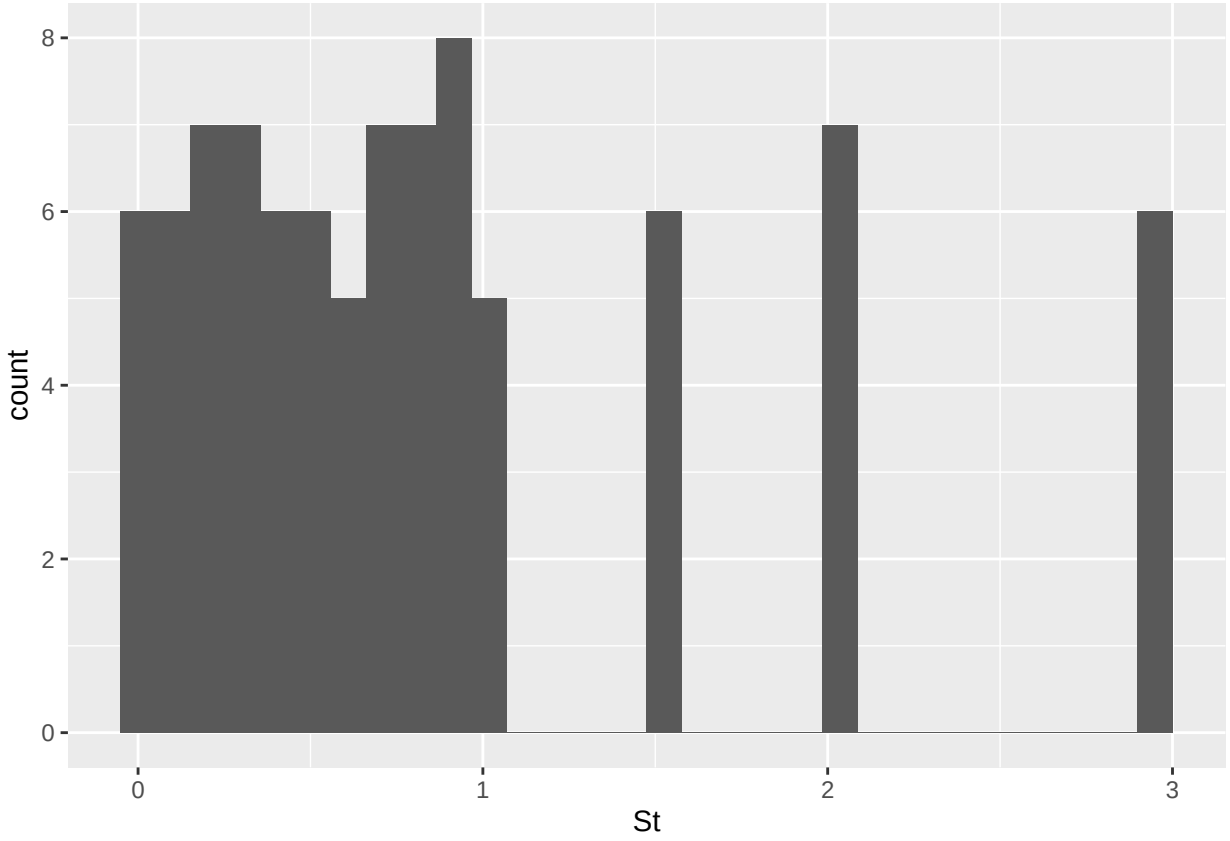
We decided to initially fit the most basic linear model to see what it would look like. We see that there may be an issue with heteroskedasticity in the residuals vs fitted values plot. In addition, looking at the Normal Q-Q plot, the normality assumption also seems to be violated. This information, combined with the seemingly nonlinear relationship between our only continuous predictor St and the 4 moments as seen in the EDA, leads us to try to use a GAM to model the relationship between the predictors and the 4 moments due to the increased flexbility GAMs provide.

We knew that using GAMs made interpretability an issue, because interpreting a complex smooth function of a continuous predictor is very hard. As a result, we decided to use variable transformations and interaction effects to make linear models with suitable model diagnostics for all 4 moments for the purpose of inference, but we would also compare these models with 4 GAM models (one for each moment) in order to find the best models for prediction.

Results

Appendix

Figure 1.1



We will try using a log transform on the St variable since the distribution for the St variable is not normally distributed.

Figure 1.2

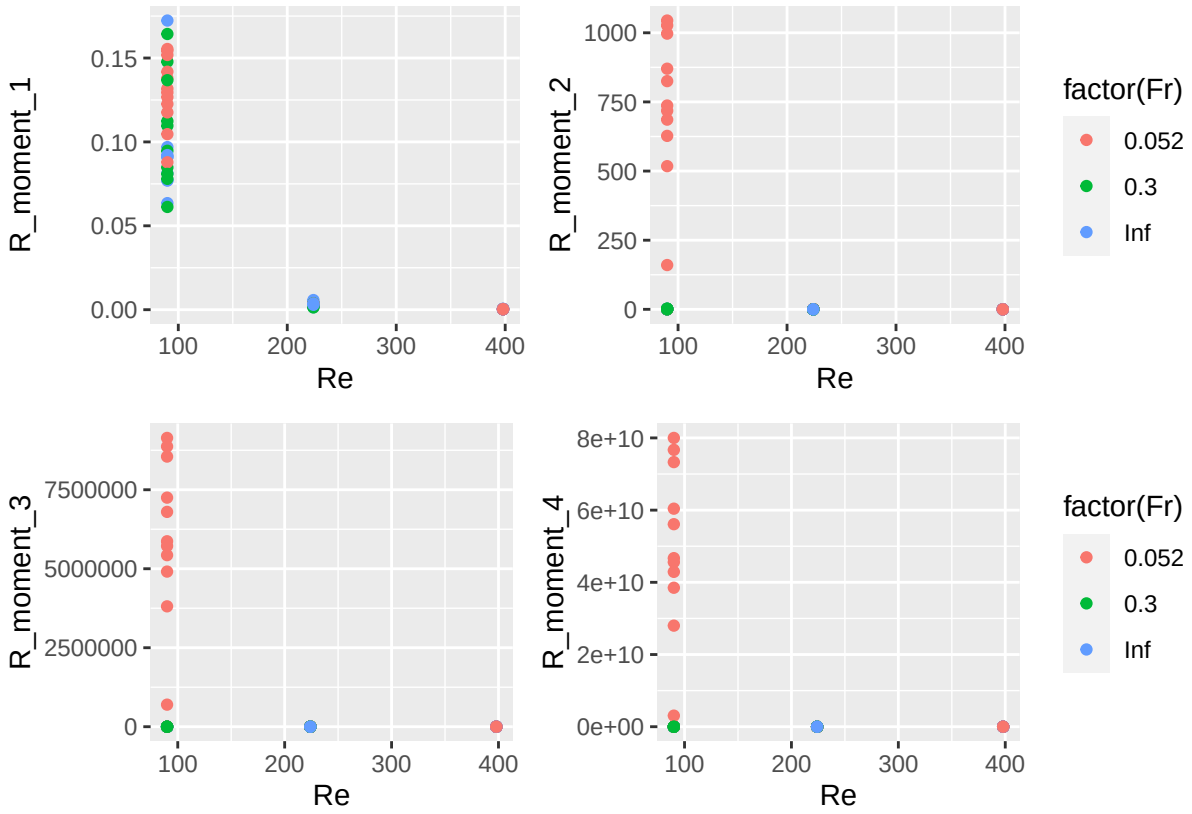
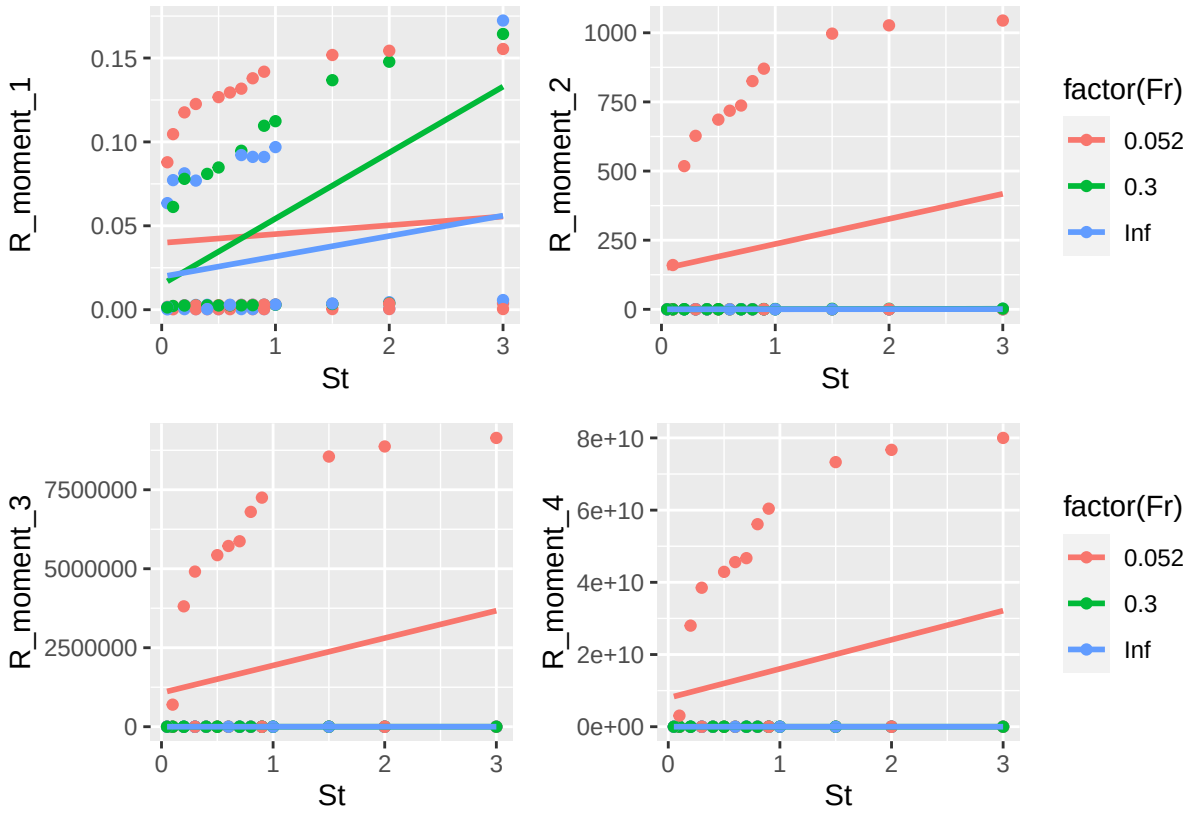

R_moment_1
0.15
0.10
0.05
0.00
100
200
300
400
Re
R_moment_2
1000
750
500
250
0
R_moment_3
7500000
5000000
2500000
0
R_moment_4
8e+10
6e+10
4e+10
2e+10
0e+00
factor(Fr)
0.052
0.3
Inf

Figure 1.3



The graphs above show some evidence of interactions, so we will explore interaction terms in our model.