Turbulence Analysis

Tingnan Hu, Peter Liu, Islina Shan, Ken Ye, Nancy Zhang

2023-10-28

Introduction

Turbulence is one of the fascinating topics in the research in fluid dynamics. It is characterized by its chaotic motion, rapid fluctuations and lack of predictable patterns. Yet, there have been numerous attempts in scientific literature trying to model the behavior of turbulent flows, as turbulent flows are prevalent in our world and are the underlying forces that drive plenty of the physical processes, from wisps of smoking swirling up from the cigarette to mixing of chemicals in industrial processes. A better understanding and prediction of turbulent flow will help us gain a deeper insight into a wide range of applications, such as improved aerodynamics in airplane designs and better climatic modelling.

A subdomain in turbulent flow research deals with particle clustering in turbulent flow focusing on small particles" behavior in turbulent fluids. For our project, we are provided with a set of simulation results on small particle probability distribution. The outcome variable was originally a probability distribution for particle cluster volumes, but it was converted into its first four raw moments, $E[X]$ to $E[X^4]$, to facilitate analysis. The predictor set contains three variables:

- Reynolds number, `Re`, which provides information on the type of flow a fluid is experiencing. A low `Re` corresponds with laminar flow (smooth and orderly), while a high `Re` corresponds with turbulent flow.
- Gravitational acceleration, `Fr`, which measures the gravitational forces particles are experiencing.
- Stokes number, `St`, where larger value corresponds with larger particle size.

The main research objective of our project will be to build a viable statistical model to predict the response variable (first four raw moments of particle probability distribution) using the three predictors at hand and the provided training set. Specifically, we are interested in the following:

- Does there exist a significant linear relationship between the predictors and the raw four moments?
- Is there any significant interaction effects between predictors on the response variables?
- Does a linear regression model suffice? Do we need a more complex model to better explain the relationship between the predictors and responses?
- Do the identified effects of the predictors vary for the four moments?

Ultimately, we aim for our model to capture adequate trends in our training data, so that for a new parameter setting of (`Re`, `Fr`, `St`), we can accurately predict its particle cluster volume distribution in terms of its four raw moments, as well as make inference on how each parameter affects the probability distribution for particle cluster volumes.

Methodology and Results

First, we examine the predictor and response variables and perform adequate transformations. For predictor variables, we first noticed that `Fr` only takes on 0.052, 0.3, and Inf in both our training and testing data set, and directly using these values as they are is not viable as they contain infinity. Therefore, we create a new categorical variable called `gravity` using the following categorization:

Fr	Gravity
$\mathbf{Fr < 0.1}$	low gravity
$\mathbf{0.1 < Fr < 1}$	moderate gravity
$\mathbf{Fr > 1}$	high gravity

We also noticed that the predictor variable `Re` only takes on 90, 224, and 398 in both our training and testing data set. We thus create a new categorical variable called `flow` using the following categorization:

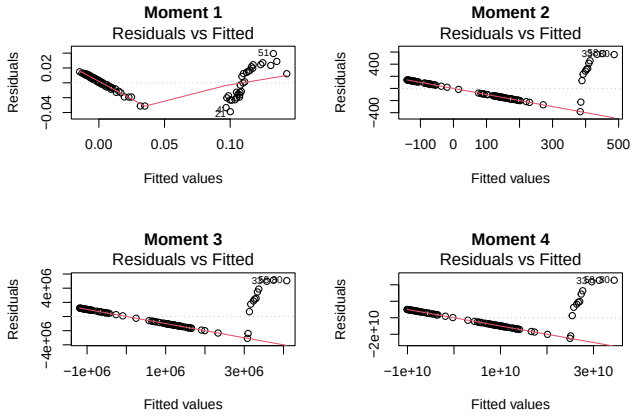
Re	Flow
$\mathbf{Re < 100}$	low flow
$\mathbf{100 < Re < 300}$	moderate flow
$\mathbf{Re > 300}$	high flow

Simple Linear Regression

We begin with simple linear regression, yielding adjusted R-squared values of 0.9251, 0.4175, 0.4023, and 0.3906 for moments 1 to 4. In all four models, the p-values are close to zero, indicating significant linear relationships between the predictors and the raw moments.

The linear regression model on the first moment suggests that `St`, `low gravity`, and `low flow` are statistically significant. Linear regression on the second moment shows that `St` is non-statistically significant, which is consistent with linear regression for the third and fourth moment. The regression model fits the first moment well, since it has an adjusted R-squared of 0.9251, but it appears that the underlying model for the other moments are not linear: the adjusted R-squared for the second, the third, and the fourth moment are: 0.4175, 0.4023, 0.3906. The AIC and BIC values suggest that the the linear regression model fits the first moment well, since the values are low. Nevertheless, other responses all have high AIC and BIC values (> 1200), which indicates that the linear regression model is not a good fit for these response variables. Therefore, linear regression model might have a good predictive power on the first moment but not for the other moments. This is reasonable, since the first moment represents the expectation of particle cluster volumes. Even if the predictors have a linear relationship with the mean of the response, it does not necessarily follow that the the predictors have a linear relationship with the second moment, which represents variance and does not have a linear relationship with the expected value of particle cluster volumes, the third moment, which represent skewness, or the fourth moment, which is a measure of the heaviness of the tail of the distribution.

Explore Response Variable Transformation

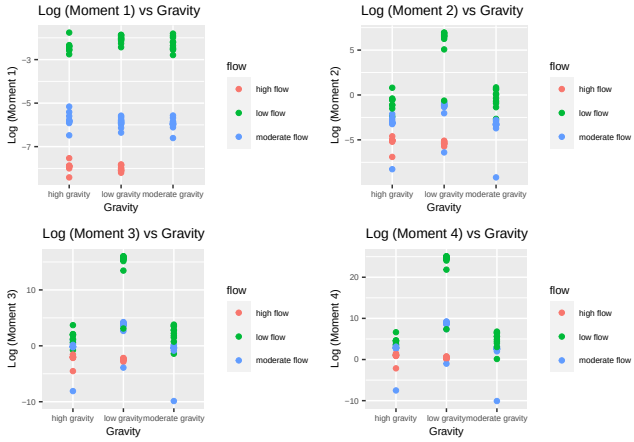


Looking at the Residuals vs Fitted plots, we observe clear patterns, which suggests that the linearity assumption is violated.

However, log-transforming the response variables not only results in more favorable Residuals vs. Fitted plots but also leads to improved adjusted R-squared values of 0.9949, 0.7633, 0.6802, and 0.6518 for moments 1 to 4.

Explore Interaction Effect

We then investigate the possible interaction effects between the predictor variables. The plot below suggests a possible interaction effect between `gravity` and `flow`.



The plot reveals that fitting a linear regression line between `gravity` and the response variables for low, moderate, and high `flow` would result in different slopes. These varying slopes serve as indicators of an interaction between these two variables. This interpretation is further validated by incorporating the interaction term into our simple linear regression models, which, in turn, yields significant p-values for the interaction terms.

Incorporating this insight into our simple linear regression models by including the interaction term `gravity:flow` results in adjusted R-squared values of 0.9966, 0.8909, 0.8770, and 0.8809 for moments 1 through 4.

We consider the interaction effects of `Re`, `St`, and `Fr` on each of the four moments. Out of the three singular variables, low flow (`Re`) has the most significant effect on the first moment, while both low flow (`Re`) and `St`

have significant effects on the second, third, and fourth moments. All of the three pair-wise interaction terms have a significant effect on all four moments. For the first moment, the interaction between `St` and low `Re` has the greatest effect. For the second, third, and fourth moments, interaction between low `Fr` and low `Re` has the most significant effect. For the first moment, this linear model with all three pairwise interaction terms is decent with $aR^2 = 0.987$ and small AIC, BIC values around -600. However, the AIC, BIC values for the linear models with interactions for second, third, and fourth moments are too large, which warrants fitted models with better predictive performance.

Explore Polynomial Term

Considering Stokes number are relatively continuously distributed, we decide to explore higher degrees of stokes number in our model, on top of the interaction effect we just explored. Using analysis of variance of nested models, we discovered that up to eight degrees of stokes number are significant (except for the first moment, which only has up to 7 degrees of Stokes Number as significant terms). We displayed the anova summary for the fourth moment result in the table below.

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
80	1.3763506	NA	NA	NA	NA
79	1.0282599	1	0.3480907	47.505854	0.0000000
78	0.9619689	1	0.0662910	9.047093	0.0036081
77	0.8122765	1	0.1496924	20.429338	0.0000234
76	0.6608893	1	0.1513872	20.660646	0.0000213
75	0.5830262	1	0.0778631	10.626409	0.0016949
74	0.5518522	1	0.0311740	4.254490	0.0427051
73	0.5348946	1	0.0169576	2.314294	0.1325095

We choose the model including up to the 7th term of Stokes number, to keep it consistent across all 4 moments, but we do recognize that higher order terms and their associated coefficients are harder to be interpreted.

Ridge Regression

To further improve the predictive performance of our model for moment 2, 3, and 4, we conducted ridge regression on each of the four moments, using 4-fold cross-validation to determine the optimal λ and MSE to evaluate the performance of our fitted model. The MSE of ridge regression models greatly improved from simple linear regression models, and are consistent across the four moments around 0.001.

Splines

We use splines to model non-linear data and capture the complex relationship between three parameters (`Re`, `Fr`, `St`) and four moments. We first find the degree of freedom by cross validation, then plug the degree moment back to the natural spline models to measure the effect of `St`, and factor `Re` and `Fr` on the four raw moments.

As the scientific inference, we can observe the factor Re and Fr have the largest effect on the fourth raw moment. This makes sense because the fourth moment is the expected value of the largest polynomial x^4. We can also observe the consistent ratio effect of Re and Fr on the change from second to third (around 8150 times), third to fourth moment (around 8200 times).

Similarly, the effects are consistency for the continuous parameters St. The magnitude of the change from smaller moment to large moment increases a lot.

Natural Splines

```
##        moment 1 moment 2 moment 3   moment 4
## Re224 -0.108230  -261.03 -2126058 -1.745e+10
## Re398 -0.111706  -316.77 -2579988 -2.117e+10
## Fr0.3 -0.007752  -269.58 -2201059 -1.807e+10
## FrInf -0.010200  -214.17 -1744529 -1.430e+10
```

```
##       moment 2/moment 1 moment 3/moment 2 moment 4/moment 3
## Re224          2411.808          8144.880          8207.678
## Re398          2835.747          8144.673          8205.465
## Fr0.3         34775.542          8164.771          8209.685
## FrInf         20997.059          8145.534          8197.055
```

Comparing the MSE values of the natural spline models for the four moments, we conclude that natural splines is most suitable for the first moment. However, natural splines perform very poorly for the other three moments, which suggests that it might not yield the best model for consistent predictive performance across the four moments.

Final selected model and predictive performance

Considering the relative performances of simple linear regression models with interaction effects and polynomial terms, ridge regression model and splines, we decided to use the following model to minimize prediction error and optimize model consistency across all four moments:

$$log(moment) = \beta_0 + \beta_1 St + \beta_2 St^2 + \beta_3 St^3 + \beta_4 St^4 + \beta_5 St^5 +$$
$$\beta_6 St^6 + \beta_7 St^7 + \beta_8 Fr_{low} + \beta_9 Fr_{moderate} + \beta_{10} Re_{low} +$$
$$\beta_{11} Re_{moderate} + \beta_{12} Fr_{low} * Re_{low} + \beta_{13} Fr_{moderate} * Re_{low} + \beta_{14} Fr_{low} * Re_{moderate}$$

Our selected model performs consistently well on the provided test data, with adjusted R^2 values above 0.95. The model performs particularly exceptionally for moment 4, which has an adjusted R^2 value of 0.99.

Conclusion

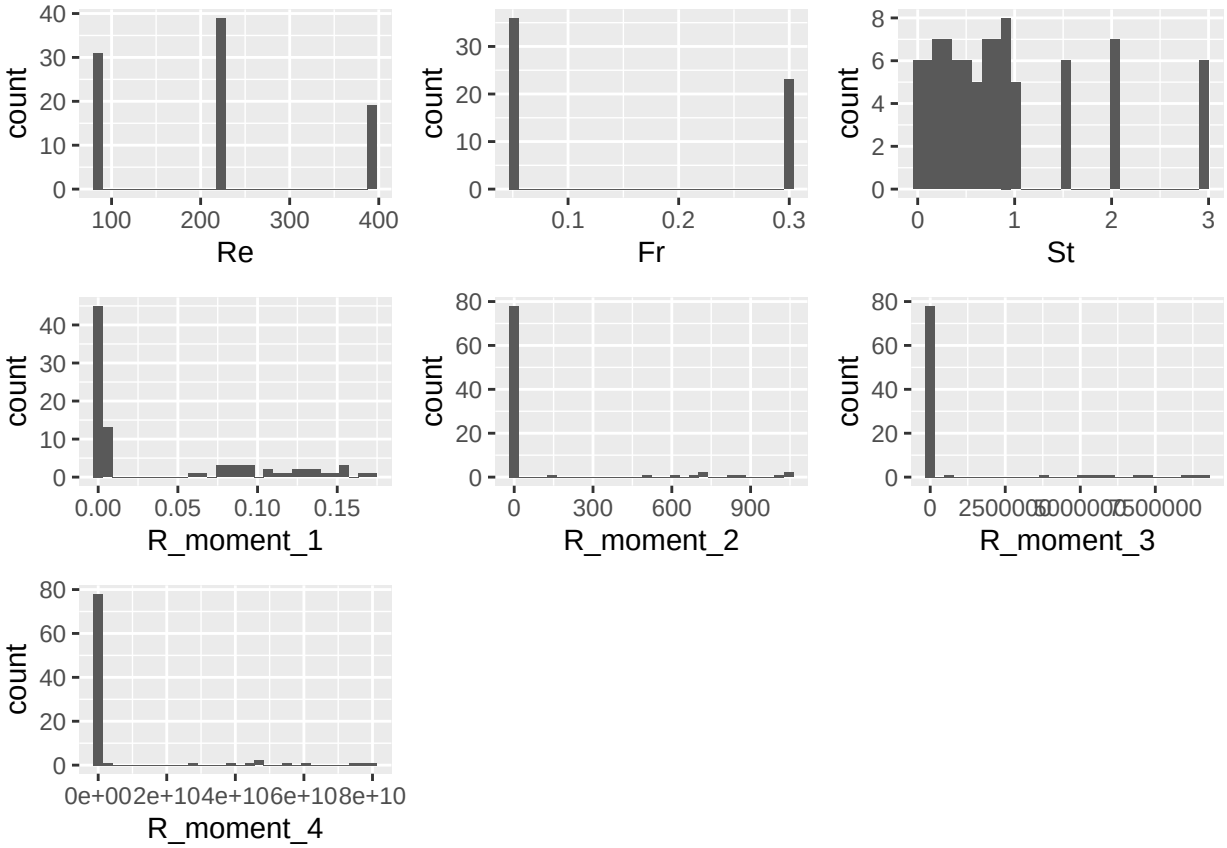
In conclusion, our analysis contains the examination of both linear and nonlinear regression models. These models included simple linear regression, log-transformation, interactions, polynomial terms, ridge regression, and natural splines. Our goal was to construct a predictive model for the particle cluster volume distribution in terms of its four raw moments. After a comprehensive assessment involving the residual analysis, Mean Squared Error (MSE), Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC), we determined that employing a seventh-degree polynomial for the parameters St, Re, Fr, and the interaction between Re and Fr provided the best fit for predicting the natural logarithm of the response variable (four raw moments).

Regarding inference, our findings indicated that the continuous parameter St and the categorical parameters Re and Fr show a consistency influence on all moments, with the largest effect observed on the 4th raw moment.

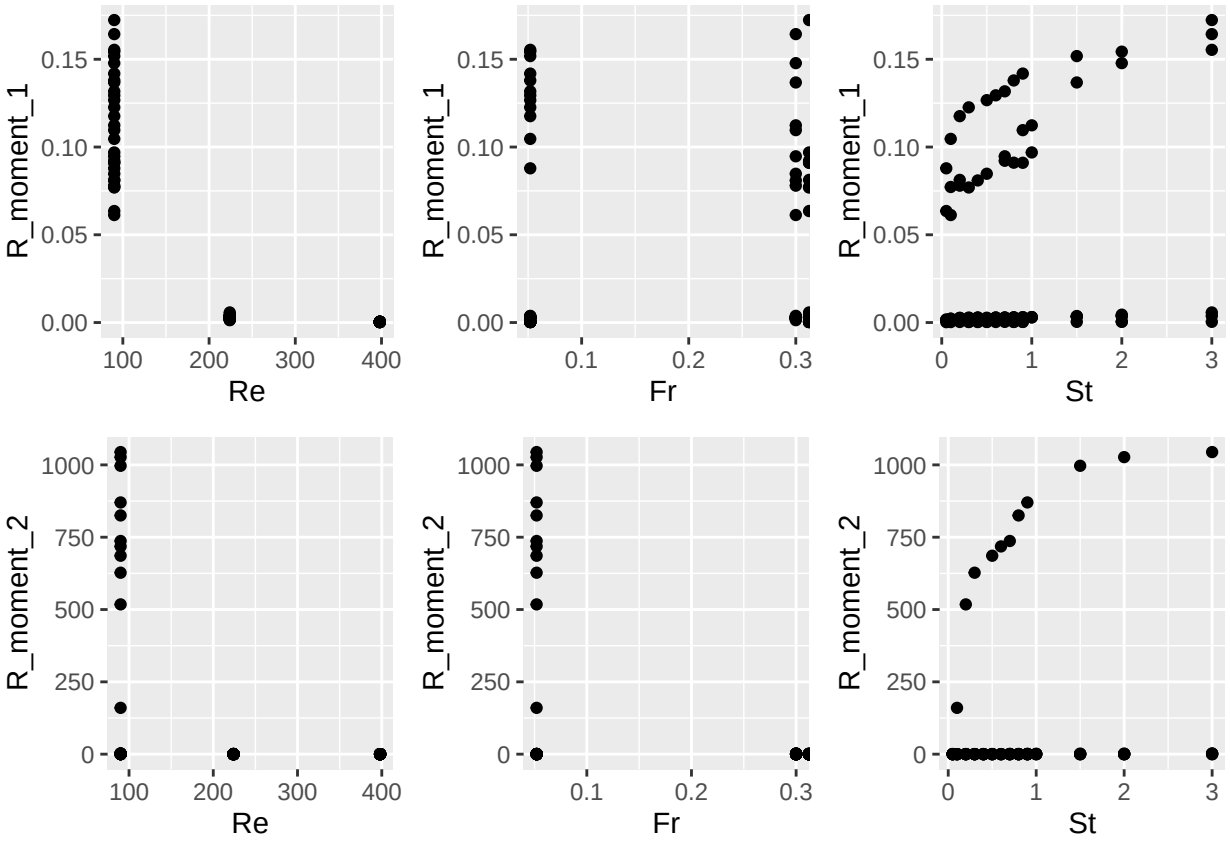
Appendix

EDA

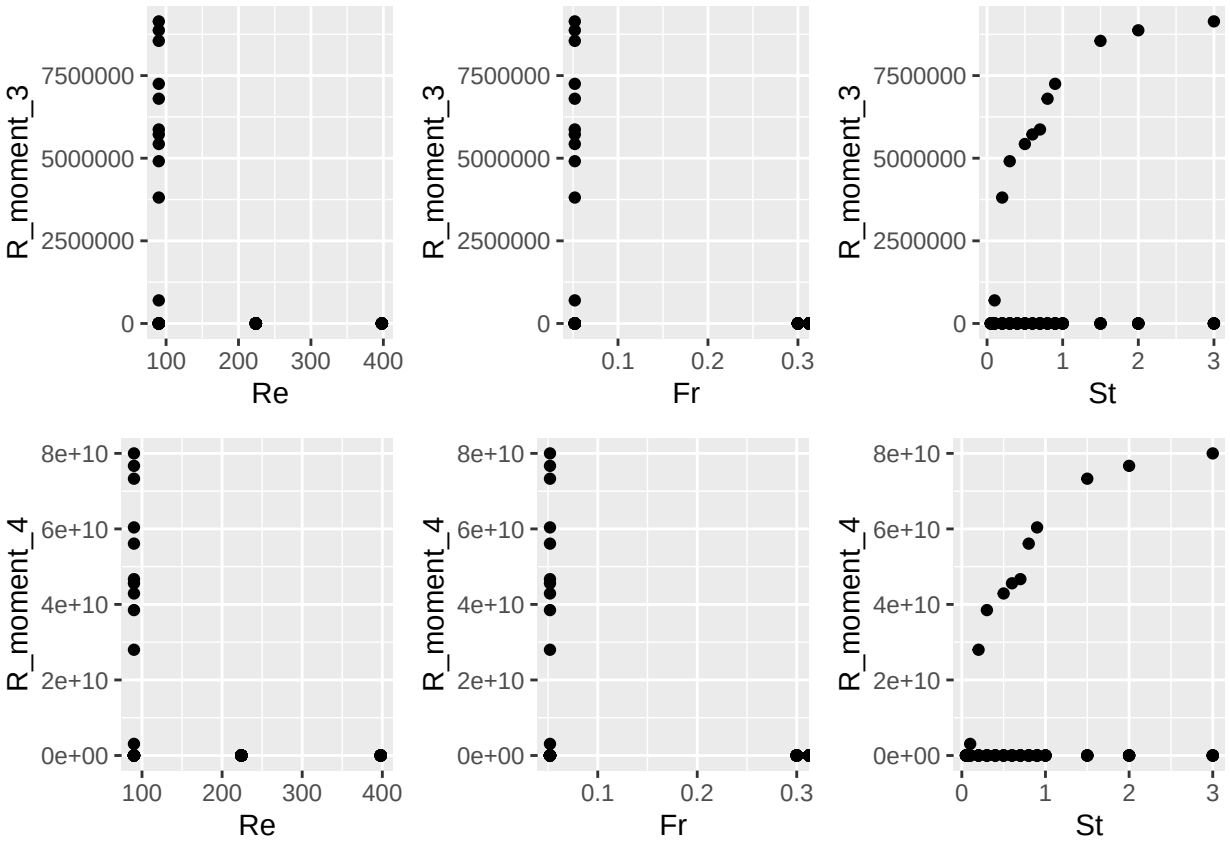
```
##        St               Re               Fr          R_moment_1
##  Min.   :0.0500   Min.   : 90.0   Min.   :0.052   Min.   :0.000222
##  1st Qu.:0.3000   1st Qu.: 90.0   1st Qu.:0.052   1st Qu.:0.002157
##  Median :0.7000   Median :224.0   Median :0.300   Median :0.002958
##  Mean   :0.8596   Mean   :214.5   Mean   :  Inf   Mean   :0.040394
##  3rd Qu.:1.0000   3rd Qu.:224.0   3rd Qu.:  Inf   3rd Qu.:0.087868
##  Max.   :3.0000   Max.   :398.0   Max.   :  Inf   Max.   :0.172340
##    R_moment_2          R_moment_3        R_moment_4          gravity
##  Min.   :   0.0001   Min.   :      0   Min.   :0.000e+00   Length:89
##  1st Qu.:   0.0245   1st Qu.:      0   1st Qu.:3.000e+00   Class :character
##  Median :   0.0808   Median :      1   Median :2.100e+01   Mode  :character
##  Mean   :  92.4902   Mean   : 753370   Mean   :6.194e+09
##  3rd Qu.:   0.5345   3rd Qu.:     40   3rd Qu.:5.345e+03
##  Max.   :1044.3000   Max.   :9140000   Max.   :8.000e+10
##      flow
##  Length:89
##  Class :character
##  Mode  :character
##
##
##
```



```
## TableGrob (3 x 3) "arrange": 7 grobs
##   z     cells    name           grob
## 1 1 (1-1,1-1) arrange gtable[layout]
## 2 2 (1-1,2-2) arrange gtable[layout]
## 3 3 (1-1,3-3) arrange gtable[layout]
## 4 4 (2-2,1-1) arrange gtable[layout]
## 5 5 (2-2,2-2) arrange gtable[layout]
## 6 6 (2-2,3-3) arrange gtable[layout]
## 7 7 (3-3,1-1) arrange gtable[layout]
```




```
## TableGrob (2 x 3) "arrange": 6 grobs
##   z     cells    name           grob
## 1 1 (1-1,1-1) arrange gtable[layout]
## 2 2 (1-1,2-2) arrange gtable[layout]
## 3 3 (1-1,3-3) arrange gtable[layout]
## 4 4 (2-2,1-1) arrange gtable[layout]
## 5 5 (2-2,2-2) arrange gtable[layout]
## 6 6 (2-2,3-3) arrange gtable[layout]
```




```
## TableGrob (2 x 3) "arrange": 6 grobs
##   z     cells    name           grob
## 1 1 (1-1,1-1) arrange gtable[layout]
## 2 2 (1-1,2-2) arrange gtable[layout]
## 3 3 (1-1,3-3) arrange gtable[layout]
## 4 4 (2-2,1-1) arrange gtable[layout]
## 5 5 (2-2,2-2) arrange gtable[layout]
## 6 6 (2-2,3-3) arrange gtable[layout]
```

Final Model Prediction

```
##                                          2.5 %      97.5 %
## (Intercept)                         -7.939298493 -7.82334046
## poly(St, 7)1                         1.672594667  2.02138313
## poly(St, 7)2                        -0.770337453 -0.42356993
## poly(St, 7)3                         0.084144441  0.42909869
## poly(St, 7)4                        -0.563427748 -0.21448271
## poly(St, 7)5                         0.215518236  0.56229114
## poly(St, 7)6                        -0.456138492 -0.10796955
## poly(St, 7)7                         0.004516756  0.35052136
## gravitylow gravity                  -0.258368242 -0.09863447
## gravitymoderate gravity             -0.187254325 -0.04881787
## flowlow flow                         5.437037723  5.60278442
## flowmoderate flow                    2.000444750  2.15369892
## gravitylow gravity:flowlow flow      0.381020025  0.60226857
## gravitymoderate gravity:flowlow flow 0.041197175  0.25347856
```

```
## gravitylow gravity:flowmoderate flow       0.006915519  0.21672947
## gravitymoderate gravity:flowmoderate flow           NA          NA
```

```
##                                              2.5 %     97.5 %
## (Intercept)                               -5.5752087 -4.6535949
## poly(St, 7)1                               4.8083007  7.5804095
## poly(St, 7)2                              -6.1293490 -3.3733023
## poly(St, 7)3                               2.6261641  5.3677992
## poly(St, 7)4                              -5.5113647 -2.7380115
## poly(St, 7)5                               2.3989737  5.1550631
## poly(St, 7)6                              -4.5309360 -1.7637510
## poly(St, 7)7                               1.1574799  3.9074631
## gravitylow gravity                        -1.1319182  0.1376174
## gravitymoderate gravity                   -0.6502088  0.4500596
## flowlow flow                               3.8931698  5.2104951
## flowmoderate flow                          1.0227884  2.2408253
## gravitylow gravity:flowlow flow            6.1711169  7.9295611
## gravitymoderate gravity:flowlow flow      -0.7906288  0.8965460
## gravitylow gravity:flowmoderate flow       1.6464146  3.3139786
## gravitymoderate gravity:flowmoderate flow         NA         NA
```

```
##                                              2.5 %     97.5 %
## (Intercept)                               -2.9023575 -1.3259034
## poly(St, 7)1                               6.3578425 11.0996358
## poly(St, 7)2                              -9.6374271 -4.9231086
## poly(St, 7)3                               4.0540163  8.7436833
## poly(St, 7)4                              -8.9795124 -4.2355906
## poly(St, 7)5                               3.6892908  8.4036824
## poly(St, 7)6                              -7.4234123 -2.6900414
## poly(St, 7)7                               1.8068859  6.5108325
## gravitylow gravity                        -1.8036893  0.3678978
## gravitymoderate gravity                   -1.0597234  0.8223258
## flowlow flow                               2.9091137  5.1624468
## flowmoderate flow                          0.2982245  2.3817211
## gravitylow gravity:flowlow flow           11.8337475 14.8416305
## gravitymoderate gravity:flowlow flow      -1.4699681  1.4160059
## gravitylow gravity:flowmoderate flow       3.3668070  6.2192361
## gravitymoderate gravity:flowmoderate flow         NA         NA
```

```
##                                               2.5 %     97.5 %
## (Intercept)                                -0.1998051  1.9620528
## poly(St, 7)1                                7.6769984 14.1796192
## poly(St, 7)2                              -12.6893827 -6.2244394
## poly(St, 7)3                                5.2650087 11.6961465
## poly(St, 7)4                              -12.0537247 -5.5481850
## poly(St, 7)5                                4.8120653 11.2771090
## poly(St, 7)6                               -9.9502700 -3.4591992
## poly(St, 7)7                                2.3578671  8.8085872
## gravitylow gravity                         -2.4058629  0.5721258
## gravitymoderate gravity                    -1.4365013  1.1444319
## flowlow flow                                2.0951842  5.1852747
## flowmoderate flow                          -0.3347007  2.5224859
## gravitylow gravity:flowlow flow            17.4645997 21.5894361
```

```
## gravitymoderate gravity:flowlow flow       -2.0699674  1.8876903
## gravitylow gravity:flowmoderate flow        5.1131865  9.0248426
## gravitymoderate gravity:flowmoderate flow          NA         NA
```