Case Study: Turbulence

Julia Rosner, Emily Mittleman, Yuzhe Gu, Ashley Chen

2022-11-3

Introduction

Understanding and predicting turbulence in fluid motion is incredibly important to a vast range of problems, such as air pollution, population dynamics, and weather. However, turbulence looks random, irregular, unpredictable, making it difficult to understand. Thus, our goals are as follows: For a new parameter setting of (Re, Fr, St), predict its particle cluster volume distribution in terms of its four raw moments. Inference: Investigate and interpret how each parameter affects the probability distribution for particle cluster volumes.

Methodology

Inference Our univariate exploratory data analysis of (Re, Fr, St) and each moment revealed that the R moments are heavily right skewed, which poses a problem to linear regression. We applied log transformations on each moment to obtain more normally distributed variables. We did not change the first raw moment to the central moment, because the mean of the distribution conveys more meaningful information than the central moment, which is always zero. However, we converted the second, third, and fourth raw moments to central moments for inference so that our results are more interpretable.

In addition, despite their numerical natures, Fr and Re only contain three levels ($Fr \in \{0.052, 0.3, \infty\}$ and $Re \in \{90, 224, 398\}$). We decided to make them categorical variables for our inference model for 2 main reasons. Firstly, treating them as numeric variables puts our model at risk for extrapolation due to lack of data at many levels of Re and Fr, making our model unable to learn the trends around such regions; and secondly, we believe that these categories could carry real life significance. For instance, Fr = 0.3 is representative of cumulonimbus clouds and 0.052 is representative of cumulus clouds. Focusing on observations collected at such specific levels may lend unique insights into practical problems.

A closer examination of the data revealed interesting interactive patterns among the independent and dependent variables. Specifically, St appears to assume a strong, non-linear relationship with each of the moments. This relationship between St and R moments appeared logarithmic, so we log-transformed St. Also, while treating the Fr variable as categorical, we observed that there is a linear and decreasing relationship between Fr (gravitational acceleration) and R moments. Lastly, there is also a linear and decreasing relationship between Re (Reynolds number) and R moments. We also explored multicollinearity through VIFs for each model, which were very low.

Our final models for each of the moments for inference is a cubic polynomical with interaction terms between Re and Fr:

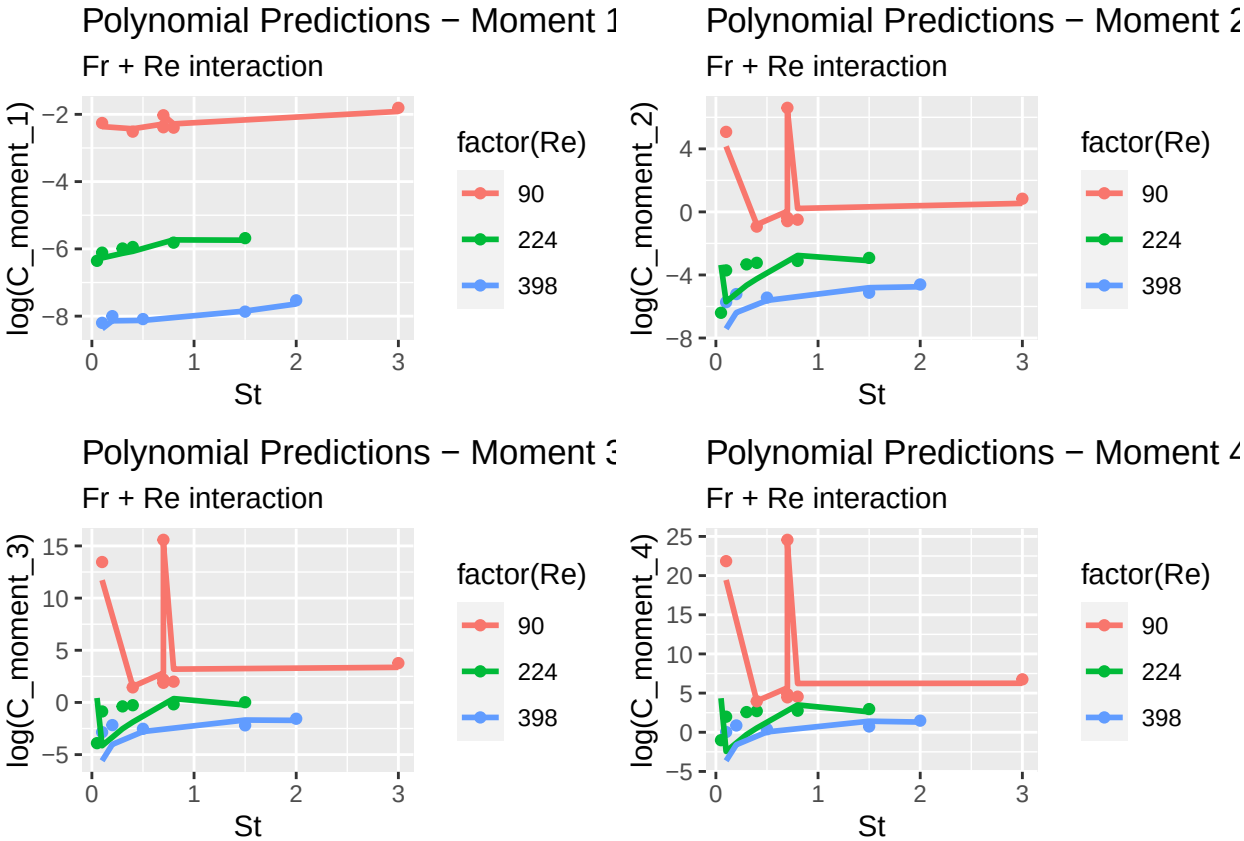
$$\begin{aligned} log(R_{moment1}) = &-2.048 + 2.002(log(St)) + 0.161(log(St))^2 + 0.446(log(St))^3 \\ &-3.824(Re_{224}) - 6.015(Re_{398}) - 0.286(Fr_{0.3}) - 0.317(Fr_{\infty}) \\ &+0.238(Re_{224} * Fr_{0.3}) + 0.389(Re_{224} * Fr_{\infty}) + 0.504(Re_{398} * Fr_{\infty}) \end{aligned}$$

$$log(C_{moment2}) = 6.005 + 9.517(log(St)) - 4.885(log(St))^2 + 3.907(log(St))^3$$
$$-7.520(Re_{224}) - 11.673(Re_{398}) - 6.679(Fr_{0.3}) - 6.562(Fr_{\infty})$$
$$+4.612(Re_{224} * Fr_{0.3}) + 4.629(Re_{224} * Fr_{\infty}) + 7.157(Re_{398} * Fr_{\infty})$$

$$log(C_{moment3}) = 14.567 + 14.025(log(St)) - 8.234(log(St))^2 + 6.315(log(St))^3$$
$$-11.292(Re_{224}) - 17.490(Re_{398}) - 12.888(Fr_{0.3}) - 12.625(Fr_{\infty})$$
$$+8.726(Re_{224} * Fr_{0.3}) + 8.632(Re_{224} * Fr_{\infty}) + 13.498(Re_{398} * Fr_{\infty})$$

$$log(C_{moment4}) = 23.168 + 18.0035(log(St)) - 22.202(log(St))^2 + 8.469(log(St))^3$$
$$-15.075(Re_{224}) - 23.325(Re_{398}) - 19.018(Fr_{0.3}) - 18.619(Fr_{\infty})$$
$$+12.755(Re_{224} * Fr_{0.3}) + 12.576(Re_{224} * Fr_{\infty}) + 19.746(Re_{398} * Fr_{\infty})$$

Our goal for inference was to build a model that was both accurate and interpretable. In testing different models, we found that the linear model, while easy to interpret, performed quite poorly. On the other hand, the spline performed quite well, and ever better than the cubic polynomial, but was difficult to interpret. We compared different degrees of polynomial regression models using the validation set approach, and found that for all four moments, the test MSE began to flatten at the third degree. While higher degree polynomial models performed slightly better, we choose degree three avoid the issue of overfitting. Plots of our model predictions on the test set show that the predicted curves fit well to the testing points. We also performed 5-fold cross validation, which showed fairly low RMSE values and high R^2 values, further increasing confidence in the model's performance.



	Moment 1	Moment 2	Moment 3	Moment 4
RMSE	0.098	0.742	1.154	1.827
Rsquared	0.998	0.968	0.951	0.940
MAE	0.069	0.451	0.699	1.026

From these models, we can gain insights into how Re, Fr, and St affect the probability distributions for particle cluster volumes. We can see that a larger Reynolds number would lead to lower predictions for all four moments. This means that as the ratio of fluid momentum force to viscous shear force increases, the mean of the distribution would decrease, the variance would decrease, the distribution will skew more to the right, and the tail will be lighter. This is in line with our knowledge that higher Reynolds numbers correspond to greater turbulent behavior. Likewise, an increase in gravitational acceleration decreases all four moments.

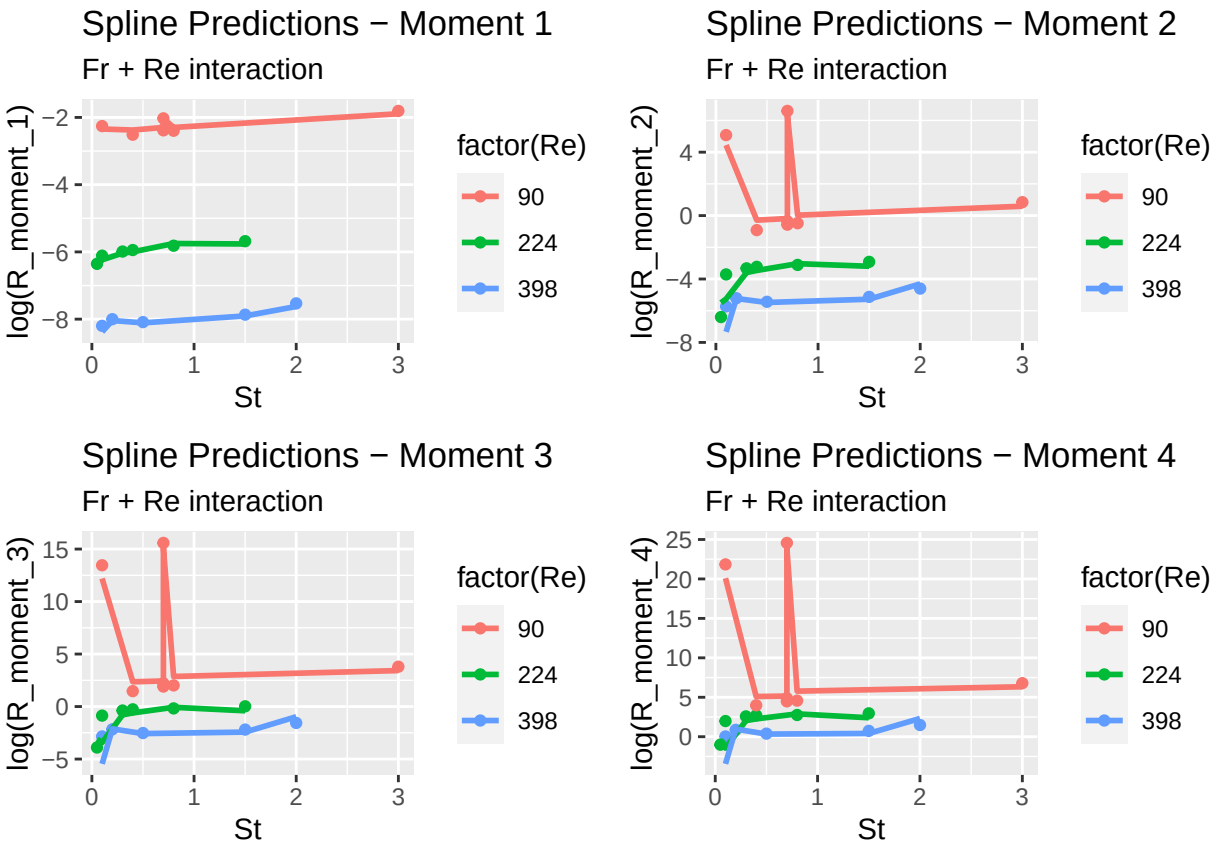
Even though a higher Reynold's number or gravitational acceleration alone have a negative effect on all four log expected moments, interaction effects with the two variables indicate that as both take on larger values at the same time, the coefficients take a positive effect, partially opposing the coefficient of the original negative effect. Thus, this may indicate that the proportion of Re and Fr together, with them either being both high or both low, mitigates the negative association on the log moment of a negative Re or a negative Fr on its own. While interaction effects are usually less interpretable, in this case it makes sense to keep them in the final model. This is because high speeds cause turbulent flow, so the effect of turbulence (Re) on the distribution is different depending on the gravitational acceleration.

Prediction One of our goals for this case study was to make a model that is specialized in making accurate predictions, meaning that we are not concerned with inference for this particular model. Because we are not concerned with interpreting this model, we explored more complex models because they tend to have better predictive performance than simple models.

Despite certain advantages in modeling Re and Fr as categorical variables, we recognize the need to extrapolate beyond the three levels of Re and Fr to model real life circumstances and enable a wider range of prediction. In order to predict values for new observations between and/or outside the ranges of values we have seen, we had to make a model that could both interpolate and extrapolate. Instead of using categorical values for our predictor variables Fr and Re, we used numerical values so that our model is more generalizable to accept numerical values between or outside the ranges of Fr and Re that were given in the training data. In order to make Fr a numerical variable, we needed to perform a transformation on it to make Fr take on finite values for modeling. To do this, we transformed Fr so that the new value is equal to $1 - 2^{-Fr}$, so that the new Fr takes on values between 0 and 1.

After exploring linear and polynomial models both with and without interaction terms, we were still not satisfied with the accuracy of these models. Since the performance of our simpler models still showed much room for improvement, we further explored cubic splines with an interaction between Fr and Re. We expected the splines to perform better compared to polynomial regression because instead of fitting a single polynomial function to the entire dataset, spline interpolation fits a piecewise continuous function composed of many polynomials to model the data set.

We did analysis on cubic spline models with and without interaction terms, and as we expected, the model including the interaction between Fr and Re outperformed the model with no interaction. To analyze predictive performance of our spline model, we performed 5-fold cross validation and selected the final model output containing the best parameters selected by cross validation. The cubic spline with the interaction term for Re and Fr outperformed any other model we created thus far, as it had extremely high R^2 values and it produced lower test MSEs for all four R moments as compared to any other models.



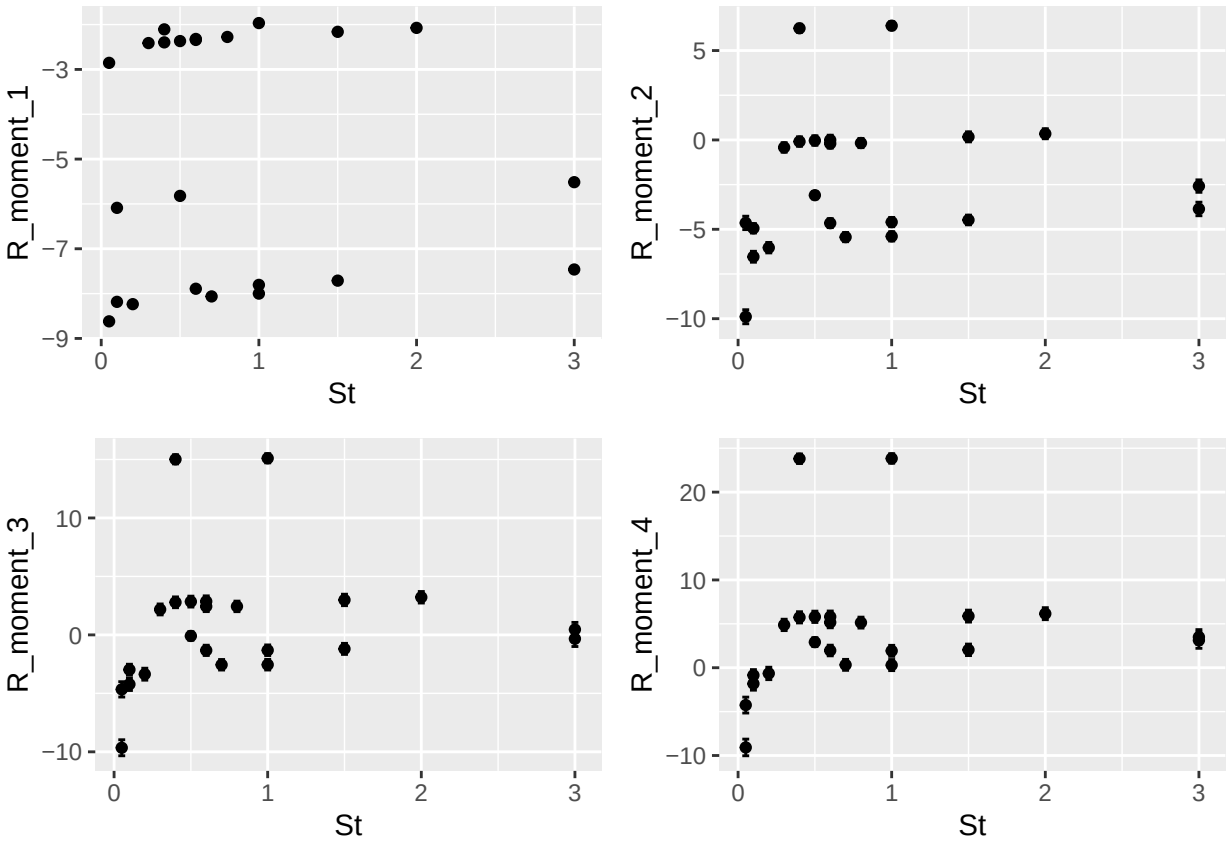
Moment 1	Moment 2	Moment 3	Moment 4
0.036	0.261	0.456	0.631
0.043	0.313	0.546	0.755
0.048	0.346	0.604	0.836
0.038	0.278	0.485	0.672
0.029	0.212	0.371	0.513
0.032	0.235	0.410	0.567
0.035	0.252	0.439	0.608
0.035	0.256	0.448	0.620
0.032	0.235	0.410	0.567
0.040	0.291	0.508	0.703
0.040	0.291	0.509	0.704
0.035	0.257	0.449	0.621
0.038	0.275	0.480	0.664
0.040	0.294	0.513	0.709
0.030	0.221	0.387	0.535
0.037	0.270	0.472	0.653
0.035	0.258	0.450	0.623
0.048	0.346	0.605	0.837

	Moment 1	Moment 2	Moment 3	Moment 4
MSE	0.010	0.548	1.334	3.337
Ajd R^2	0.999	0.971	0.964	0.963

RMSE	0.098	0.740	1.827	0.098
Rsquared	0.998	0.968	0.940	0.998
MAE	0.069	0.450	1.026	0.069

Prediction

Results



Conclusion

All three predictor variables display some significant statistical effect on particle clustering. In general, larger particle characteristics (i.e. size, density) lead to higher first moments. Higher fluid turbulence (Re) and gravitational acceleration lead to lower values of all four moments, but this is partially mitigated if they are both high as evidenced by the coefficients of the interaction effects, rather than one of these two values being individually high. Furthermore, in higher moments we see evidence of non-linearities between the variables.