STA 325 Case Study

Abby L., Ava E., Ella T., Grady P. Laura C.

October 26, 2025

Introduction

Whether it's stirring a morning cup of coffee or feeling your stomach lurch during a bumpy airplane ride, nearly everyone has experienced some form of turbulence. Beyond these everyday encounters, turbulence plays a crucial role in many complex natural and industrial processes, from air pollution and chemical reactions to heat transfer and weather systems. Despite its ubiquity, turbulence remains notoriously difficult to predict and has long been regarded as "the last great unsolved problem in classical physics."

While we may not be physicists ourselves, we have undertaken the challenge of developing a predictive model to better understand this elusive phenomenon in collaboration with our Professor, Simon Mak. Given observations for fluid turbulence (quantified by Reynolds number Re), gravitational acceleration (quantified by Froud number Fr), and the particle's characteristics (quantified by Stokes number St), we have explored models designed to predict the mean, standard deviation, skewness, and kurtosis of the spatial distribution and clustering of particles in clouds in a state of idealized turbulence. Our goal is to create a model that balances predictive accuracy with interpretability; one that not only performs well statistically but also provides clear, meaningful insights into how different conditions influence turbulent behavior.

Methodology

Observing the distributions of the predictor variables, the square root of St was used as well as the log of Re to improve the symmetry of the distribution to address the constant variance assumption. Additionally, the logit function was applied to Fr so that all values are bound between 0 and 1 as a way of handling infinite Fr values.

One concern with the data is that Re and Fr each only have 3 unique values. Hence, using these variables to model the distribution of the response variable requires a large amount of uncertainty and interpolation at values between the 3 values in the data. To continue refining the models, additional data with different values for these variables will reduce uncertainty and improve model fit.

Additionally, we analyzed the distributions of each response variable before building the model. To correct for asymmetries and heteroskedasticity amongst the residuals, transformations were applied to all of the response variables: log transforming the Mean, log transforming the Standard Deviation, square rooting Skew, and log transforming Kurtosis. These transformations were chosen by analyzing the distributions of these variables as well as the optimal lambda results from the Box Cox transformation.

For all central moments, we first fit several different models including simple linear regression, interaction terms, polynomial models, ridge regression, and spline models. We compared initial diagnostics such as residual plots, adjusted R^2 values, AIC/BIC, and Cp. Then, 5-fold cross-validation was used to ensure that our chosen model was not overfitting the data by selecting models with low RMSE across the folds.

Results

Mean

The polynomial model on log(Mean) was chosen as the best model including second-degree polynomial terms for all three predictors (√St, Re, and the logistic-transformed Fr) along with three two-way interactions: √St:Re, √St:logistic_Fr, and Re:logistic_Fr. This model achieved great fit, with an adjusted R^2 of 0.9978, meaning it explains over 99% of the variation in log(mean) cluster volume. The model's RMSE of 0.098 on the log scale translates to predictions that are typically within about 10% of the true mean cluster volume when un-transformed, although there is still uncertainty regarding the predictions due to the presence of categorical predictors. Performing 5-fold cross-validation yielded a RMSE only slightly higher than the training RMSE with a comparable R^2 value, providing evidence the model generalizes well to unseen data.

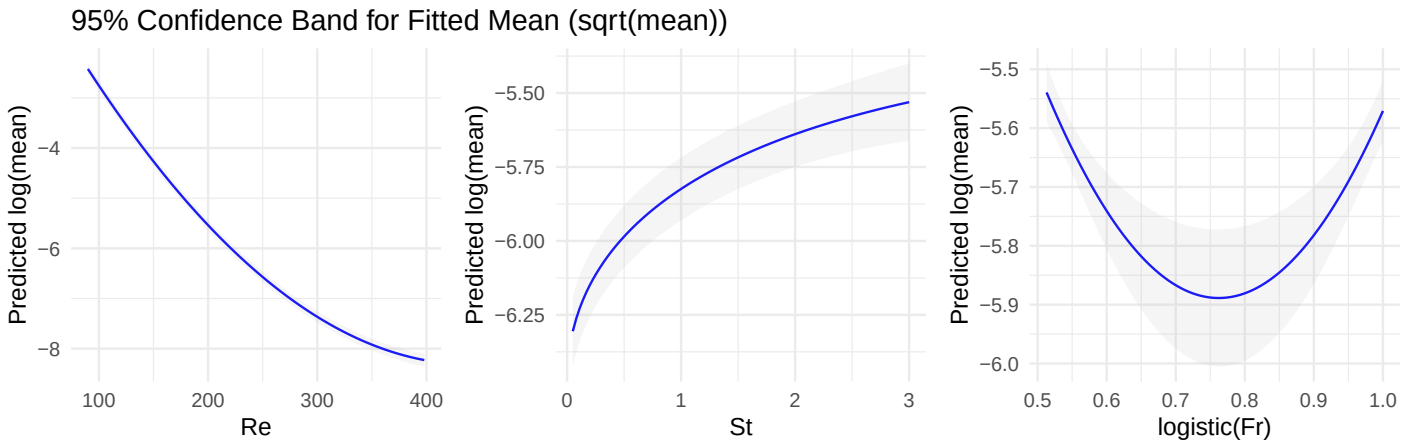
Below is the output for a few of the statistically significant coefficients in the model:

Predictor	Estimate
poly(Re, 2)1	-22.26000
poly(Re, 2)2	5.20600
poly(sqrt(St), 2)1	-5.68900
poly(sqrt(St), 2)2	1.09800

Predictor	Estimate
poly(logistic_Fr, 2)1	-2.06500
poly(logistic_Fr, 2)2	0.57900
Re:logistic_Fr	0.00295

- Reynolds number shows a strong negative linear effect but a positive quadratic effect, indicating that the log(mean) cluster volume decreases as the intensity of turbulence of the flow initially increases, but then as the intensity of the flow increases, the mean cluster volume begins increasing.
- The size of the particles (St) has a similar effect. Increasing particle size decreases the mean of the cluster volume initially, but then begins increasing the mean as the particles become much larger.
- Fr's negative linear and positive quadratic effects suggest that gravitational acceleration initially reduces the mean clustering but this effect diminishes at extreme gravity levels.
- The positive Re:logistic_Fr interaction indicates that gravity's effect on clustering becomes stronger in more turbulent flows, which makes sense as turbulence and gravity compete in determining particle settling behavior.

Overall, Reynolds number (turbulence intensity) is the most dominant factor controlling mean cluster size, but its effect depends significantly on gravitational conditions, specifically gravity's impact on clustering strengthens in more turbulent flows. Thus, predicting particle behavior requires understanding the effects of turbulence and gravity together, rather than independently.



The plots depict the 95% CI for the predicted log(mean) as each predictor varies, given that the other two predictors are held at their average values. There appears to be a strong negative relationship between Re and log(mean) which is supported by the narrow confidence interval. For St, there seems to be a nonlinear and positive relationship between St and log(mean) yet the wide confidence intervals at the boundaries convey uncertainty. Lastly, the relationship between logistic(Fr) and log(mean) looks parabolic, yet there is great uncertainty for logistic(Fr) values around 0.65-0.85. This uncertainty is attributable to the data only having 3 unique values for Fr, so we will continue refining our model after receiving additional data points for Fr.

Standard Deviation

A linear model predicting the log(standard deviation) including all interaction terms was selected as the best model based on the balance between predictive accuracy (RMSE = 2.01 and adjusted $R^2 = 0.74$). The interaction terms were included to capture the potential for one variable effecting log(sd) to depend on another variable. The linear model is interpretable and has higher bias and lower variance, meaning it is less likely to overfit the training data.

Although across other model $\sqrt{}$St was the transformation for St that had the largest improvement on the model,we used the log(St) in this model since the log transformation performed better. Below are a few significant findings:

Predictor	Estimate
log_Re	-4.915
logistic_Fr	-20.704
log(Re):logistic_Fr	3.557

- Similarly to the effect of the Reynolds number on mean, increasing the flow turbulence decreases the standard deviation of the clustering distribution causing a narrower distribution as the particles are clustered closer together. Although initially surprising, it makes sense that particles flowing quicker are more likely to be clustered since they collide more often. This is a finding to further discuss with our collaborator to better understand the physics behind this negative relationship.
- We also observe a negative relationship between the gravitational acceleration of the cloud particles and the standard deviation of the clustering distribution. Based on our results above, as clouds decrease in height, the model predicts that the clustering distribution will get narrower.
- Lastly, the interaction between gravitational acceleration and turbulence flow is positive. Given that both effects are negative independently, this is a finding warranting further examination. The effects of Fr and Re on their own negatively impact the standard deviation of the clustering, yet combining these effects yields a positive relationship indicating there might be correlation between these two variables which causes Fr and Re to have a less negative the standard deviation when the other variable is controlled for, although we would like to further examine this relationship

Overall, more turbulent conditions with more dense particles lead to narrower particle cluster distributions. However, the negative effect of the flow's turbulence is reduced in conditions with higher gravitational acceleration, as seen with the interaction effect between log(Re) and logistic(Fr).

Skew

The model best fitting √(Skew) is a ridge model fit using 5-fold cross validation. To evaluate the significance of coefficients, the same predictors are fit with OLS to develop intuition regarding the magnitude and significance of coefficients.

Predictor	Estimate
logistic_Fr	-21.350
Re:logistic_Fr	0.055

- Transforming logistic(Fr) back to Fr yields a coefficient extremely close to 0. This indicates that changes in the gravitational acceleration of particles (height of the cloud) do not greatly change the asymmetry of the clustering distribution.
- Although Fr by itself has a basically negligible impact on the distribution of √(Skew), the interaction between Fr and Re is statistically significant. A positive interaction between Re and Fr indicates that either as a cloud drops lower (Fr) or as the flow's turbulence increases (Re), √(Skew) is increasing. Given that logistic(Fr) has a negative coefficient yet the interaction between Re and logistic(Fr) is positive, this finding is one that should be investigated further

In a more interpretable model with the same linear combinations of predictors, logistic(Fr) and Re were treated categorically. Despite a categorical model being an unrealistic fit for data with continuous values, the model had better performance and provides insights into certain interactions:

Predictor	Estimate
sqrt(St):as.factor(Re)224	-1.4173
sqrt(St):as.factor(Re)398	-2.5683

- When treated as categorical variables, the relationships which are most significant both involve √(St) and the levels of Re.
- Interpretation: When the turbulence within the flow (Re) increases from the baseline level to 224, the expected square root of Skew decreases holding √(St) constant. An even larger decrease is observed when the Re increases to 398 and √(St) is held constant. As randomness increases and the size of the particle remains the same, we expect to see a decrease in the asymmetry of the distribution of the turbulence so the distribution of the clustering of the particles appears more symmetrical.

Kurtosis

The best model for Kurtosis is a polynomial model for the log transformed values of Kurtosis where $\sqrt{}$(St) is to the first degree, Re and logistic(Fr) are squared, and two-way interactions terms are included. Many of the coefficients are extremely significant in this model. For the final polynomial model, logistic(Fr) is again statistically significant for both degrees which was also observed for Skewness above. However, we do notice some additional statistically significant interactions which are worth mentioning:

Predictor	Estimate
sqrt(St)	-0.402
poly(Re, 2)1	14.711
poly(Re, 2)1:poly(logistic_Fr, 2)1	48.397

- The negative sign on sqrt(St) coefficient indicates that as the size of the particles in the environment increase, the kurtosis of the distribution decreases. Simulating from larger particles causes turbulence distribution to become less peaked as it's not as heavily concentrated around a single value. One possible conclusion is that increasing the size the particles increases the range of the clustering distribution leading to skinnier distribution tails.
- The first degree for Re is largely positive indicating an increase in the turbulence of the flow of the particles increases how heavy the tails of the clustering distribution are.
- There is a significant interaction between the first degrees for both Re and Fr which can be interpreted in two ways. Firstly, increasing the turbulence of the flow while holding the gravitational acceleration constant is expected to increase the Kurtosis of the model as the clustering distribution becomes more peaked. Alternatively, increasing the gravitational acceleration while holding turbulence constant, is expected to increase the Kurtosis of the model. Intuitively, one would imagine that first interpretation is more likely given that increasing flow turbulence seems to have a more significant impact (given the bullet above), but further discussion with the collaborator is needed to determine the true interpretation.

Limitations & Next Steps

One limitation to our results is that the interpretations may be less intuitive due to the response transformations using log and sqrt. For example, the log transform must be interpreted on a multiplicative scale; however, the strength with using these transformed responses is that the distributions are more symmetrical to allow for more normally distributed residuals. Another limitation is that the confidence bands near the boundaries of St, Re, and Fr widen at the edges, indicating more uncertainty in predictions as values are being extrapolated. Furthermore, the confidence bands are wider for values in between the distinct values for Re & Fr, so more observations are needed to address this uncertainty. Furthermore, all four models rely on the assumption of linearity or smooth continuity which may not capture the nonlinear turbulent particle patterns that other models such as GAMs and tree-based methods could potentially address.

Conclusion

Some of the key findings are listed below:

- Across the models for Mean, Standard Deviation, and Kurtosis, Re has a decreasing statistically significant first degree so increasing the turbulence of the flow negatively impacts these central moments. It is surprising to find a coefficient with the same directional effect across central moments so this is worth further investigation.
- Additionally, logistic(Fr) has a decreasing statistically significant coefficient across the models for Mean, Standard Deviation, and Kurtosis. This implies that an increase in the gravitational acceleration of the cloud particles (a decrease in how high the clouds hangs), decreases three of the central moments.

Moving forward, we wish to collect more data with a greater number of unique values for Fr and Re so we reduce the need to interpolate for both of these models. We would also like to further discuss some of our model assumptions with our collaborators and work to develop greater comprehension regarding how the central moment predictions can be used for understanding the entire cluster distribution. We are confident that with additional data, domain knowledge, and resources, we can continue improving our models to improve predictive performance for modeling the clustering distribution of cloud particles.

We would like to thank our collaborators in Professor Simon Mak and the Duke Civil & Environmental Engineering department. One can access the code repository for this project via this link: Github Repository.