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.

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 of these variables between the 3 distinct values in the data. To continue refining the models, additional data with different values for these models 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, we applied different transformations to some of the response variables: log transforming the Mean, log transforming the Standard Deviation, and square rooting Skew. These transformations were determined by analyzing the distributions of these response variables as well as results from the Box-Cox transformation. Since we fit initial models with linear regression as well as using linear regression for coefficient interpretation for certain models, transforming Mean and Skew improved the distribution of the noise to closer resemble a normal distribution with the residuals randomly scattered around 0, satisfying the assumption that the error is normally distributed in linear regression.

Results

Mean

To predict the mean particle cluster volume, several models were assessed including simple linear regression, interaction terms, polynomial models, ridge regression, and spline. After performing 5-fold CV on each, the polynomial model on log(Mean) performed best (assessed using residual plots, R^2, AIC, and BIC). The log transformation was necessary because without it, the residuals showed clear patterns and heteroscedasticity.

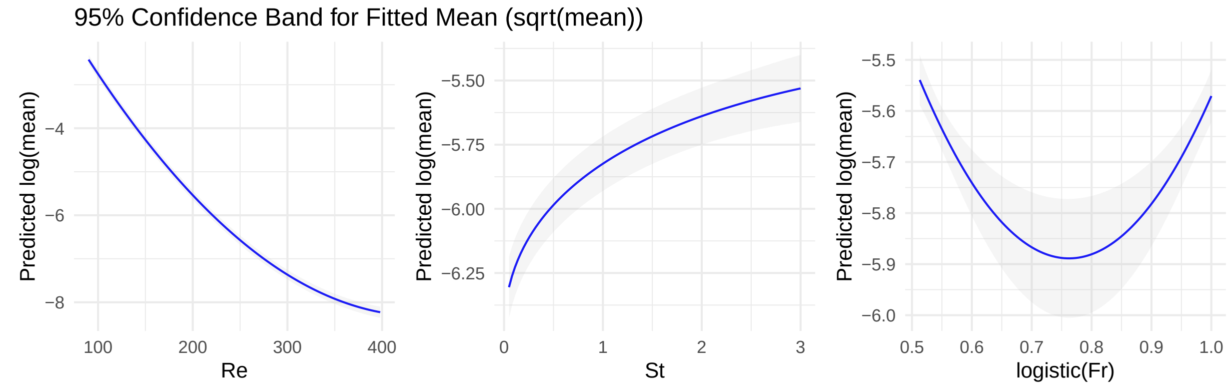
The final model includes second-degree polynomial terms for all three predictors (√St, Re, and the logistic-transformed Fr) along with three two-way interactions: √St:Re, √St:logistic_Fr, and Re:logistic_Fr. This model achieved great fit, with an adjusted R^2 of 0.9978, meaning it explains over 99% of 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. This close agreement between training and validation performance indicates that the model generalizes well to unseen data and does not appear to overfit the data despite its polynomial and interaction terms.

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

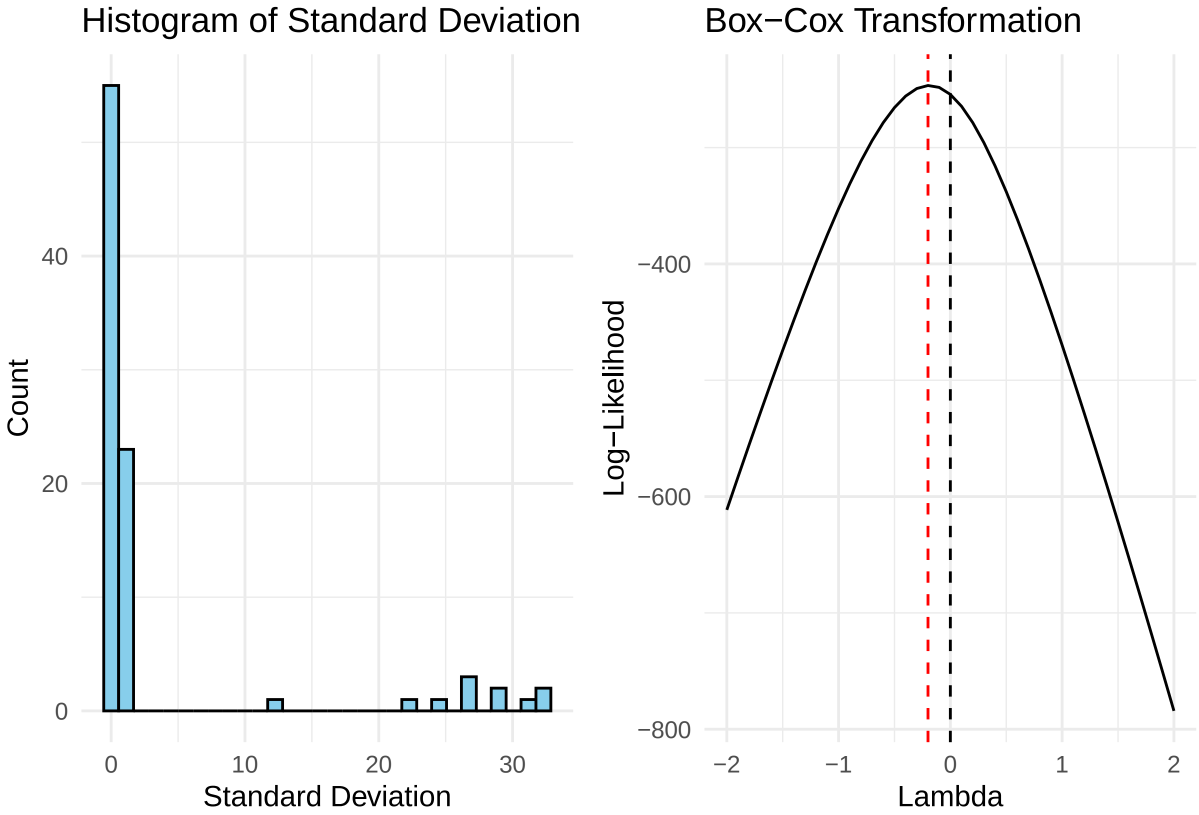
- Reynolds number shows a strong negative linear effect but a positive quadratic effect, indicating that the log(mean) cluster volume initially decreases as the intensity of turbulence of the flow initially increases, but then as the intensity of the flow increases, the mean cluster volume begins increasing.
- The size of the particles (St) has a similar effect. Increasing particle size initially decreases the mean of the cluster volume, but then begins increasing the mean as the particles become much larger.
- Lastly, the Froude number's negative linear and positive quadratic effects suggest that gravitational acceleration initially reduces 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 physical 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. This interaction means that predicting particle behavior requires considering both turbulence and gravity together, rather than treating them as independent effects.


95% Confidence Band for Fitted Mean (sqrt(mean))

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 appears to be a nonlinear and positive relationship between St and log(mean) yet the wide confidence intervals at the boundaries convey uncertainty. The relationship between logistic(Fr) and log(mean) appears to be 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 would hope to continue refining our model after receiving additional data points for Fr.

Standard Deviation



The Box Cox transformation shows that the optimal lambda is around 0, so the log transformation on sd is the optimal transformation to address skewness of the response variable.

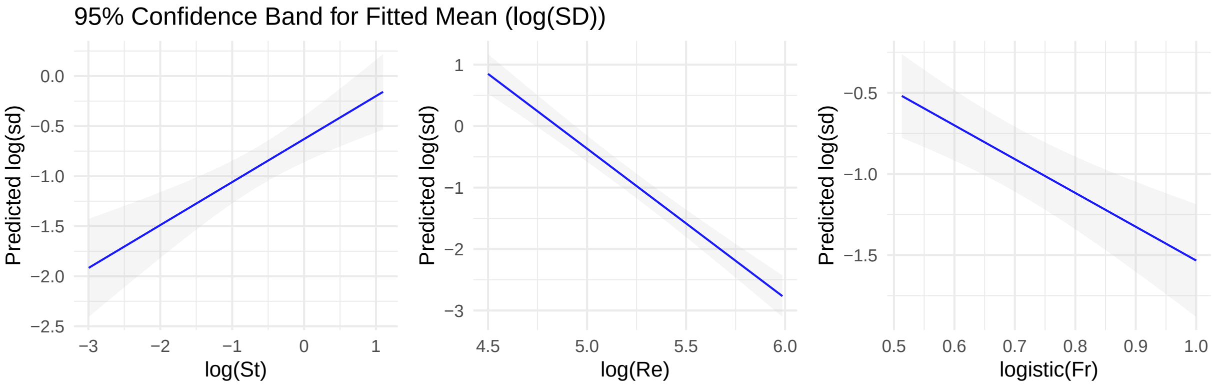
We were given a training and testing dataset, so I trained different linear (with and without ridge and lasso), polynomial and spline regression models comparing the adjusted R-squared the p-values, and used 5-fold CV to compare the top model to determine which model had the best prediction accuracy while working on new, unseen data.

After multiple models were compared from their adjusted R-squared and 5-fold CV, the linear model with all of the interaction terms was selected based on the balance between predictive accuracy (RMSE = 2.01, Adjusted R-squared = 0.74). The interaction terms were included to capture the potential effects of predictors coupled together. The linear model is interpretable and less risky to overfit, which is helpful in our goal to predict the standard deviations of the distributions given St, Re, and Fr.

Model	RMSE
lin	2.077942
lin_interactions	2.013816
lin_interactions_subset	2.166138
poly	2.072878
spline	2.115911
lasso	2.068582
ridge	2.080013

Model	RMSE

When running 5-fold CV, the RMSE for the selected model is 2.0138, which is lower compared to other linear, ridge, lasso, spline and polynomial models.



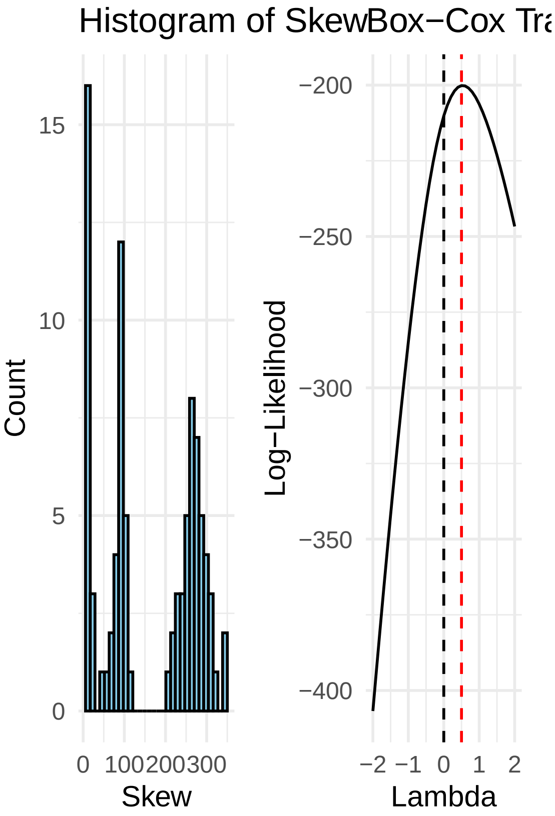
The plots depict the 95% CI for the predicted log(sd) as each predictor varies, given that the other two predictors are held at their average values. For Re, we are fairly confident that there appears to be a positive relationship between log(St) and log(SD), with greater variability near the extremes. For Re, we are very confident that there is a strong negative relationship between log(Re) and log(SD), such that at larger Re values, there is less variability in log(Re). For Fr, we have some confidence that the relationship between logistic(Fr) and log(SD) is negative; however, the confidence band is much wider compared to the other predictors, which means that there does appear to be more uncertainty.

```
## 
## Call:
## lm(formula = log(sd) ~ (log_St + log_Re + logistic_Fr)^2, data = data)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -2.35706 -0.43689  0.09339  0.41670  1.90516 
## 
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)    
## (Intercept)          26.49639    3.06941   8.632 3.88e-13 ***
## log_St                0.64222    0.89622   0.717    0.476    
## log_Re               -4.91459    0.58036  -8.468 8.22e-13 ***
## logistic_Fr         -20.70407    4.25749  -4.863 5.50e-06 ***
## log_St:log_Re        -0.02716    0.16702  -0.163    0.871    
## log_St:logistic_Fr   -0.10234    0.41962  -0.244    0.808    
## log_Re:logistic_Fr    3.55720    0.79835   4.456 2.62e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.9463 on 82 degrees of freedom
## Multiple R-squared:  0.7577, Adjusted R-squared:   0.74 
## F-statistic: 42.75 on 6 and 82 DF,  p-value: < 2.2e-16
```

From our model, we can see that log_Re, logistic_Fr, and the interaction between log_Re and logistic_Fr has significant effects on the standard deviation of the particle cluster volume distribution. For example, holding other factors constant, for a 1% increase in the Reynold's Number, there would be a 4.915% in decrease standard deviation. Also, holding other factors constant, for a 1% increase in the Stokes number, there would be a 20.704% decrease in standard deviation. In context, this means that more turbulent regimes and regimes with more dense particles lead to narrower particle cluster distributions. However, the effect with Reynold's Number is reduces in regimes with higher gravitational acceleration, or Fr, as seen with the interaction effect between log(Re) and logistic(Fr).

Skew

Looking at the distribution of a Skew and a Box-Cox transformation, the square root transformation is selected to address the skewness of the response variable.



After evaluating linear regression, polynomial models, ridge regression, and splines, the model which best fit the √(Skew) response is a ridge model fit using 5-fold cross validation. To evaluate the significance of coefficients, the same predictors were fit with OLS to develop better intuition regarding the magnitude and significance of coefficients.

Predictor	Estimate
logistic_Fr	-21.350
Re:logistic_Fr	0.055

From the OLS model, the above coefficients were determined to be the most significant:

- 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) does not greatly change the asymmetry of the turbulence distribution.

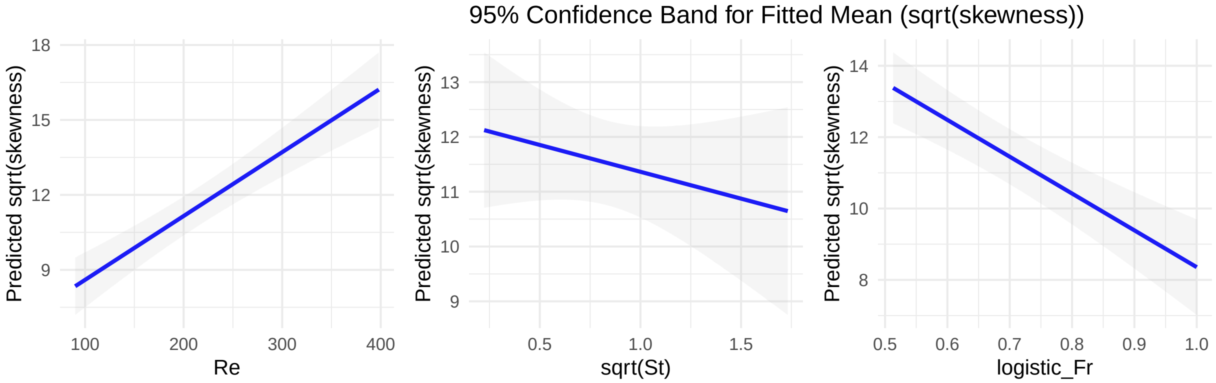
- Although Fr by itself has a basically negligible impact on the distribution of √(Skew), the interaction between Fr and Re is statistically significant. Lower-hanging clouds have a higher Fr value so a positive interaction between Re and logistic_Fr indicates that either as a cloud's Fr increases (the cloud get lower) or Re increases (the flow is more turbulent), √(Skew) is increasing. Given that logistic_Fr has a negative coefficient yet the interaction between Re and logistic_Fr is positive, exploring the relationship between the height of the cloud and the turbulence of the flow warrants further investigation.

In a separate model with the same linear combinations of predictors, logistic_Fr and Re were treated categorically. Using categorical variables greatly improved the model performance (quantified by R^2, AIC, BIC, and Cp). Despite a categorical model being an unrealistic fit for data with continuous values, the model still provided insights into the interactions between these variables:

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 Reynolds' Number 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 Reynolds' Number increases to 398 and √(St) is held constant. As randomness increases and the size of the particles remains the same, we expect to see a decrease the asymmetry of the distribution of the turbulence so the distribution of the clustering of the particles appears more symmetrical with an increase in randomness.

Given the nature of the three levels for both Re and Fr, there is a large degree of uncertainty for these predictors when predicting the response which is especially evident in the wide nature of the 95% confidence intervals near the extremes of both of these predictors. The below plots show slices of each of the predictors for the response at the mean of the other two predictors which aren't on the x-axis. Receiving additional data in the future would address the high amount of uncertainty in the model as we continue to refine the model.



Kurtosis

The best model for Kurtosis is a polynomial model for the untransformed values of Kurtosis where √(St) is not used in a polynomial term while both Re and logistic_Fr are raised to the second power. There are also interactions terms included between all three variables and their polynomial terms.

This polynomial model performed well when evaluated using 5-fold CV with a fairly high R^2 value and fairly low RMSE, providing us with confidence that this model is not overfit to the training data. 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)	-16383
poly(Re, 2)1	204158
poly(Re, 2)1:poly(logistic_Fr, 2)1	888507

- 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. When we simulate from larger particles, the turbulence distribution becomes less sharp as it's not as concentrated around a single value. One conclusion from this may be that increasing the size the particles increases the range of the distribution of particle turbulence resulting in less heavy tails for the distribution.
- 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.
- Lastly, 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 also increase the Kurtosis of the model. Intuitively, one would imagine that first interpretation is more intuitive given that increasing flow turbulence seems to have a more significant impact given the bullet above, but further discussion with the collaborator and model investigation would be worthwhile to determine the true effects.

We see a similar statistical significance between $\sqrt{}$(St) and the categorical values of Re for the polynomial terms.

Conclusion

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.