Introduction

Predicting fluid motion is a challenging task because turbulence produces complex and chaotic behavior that plays a critical role in many natural and industrial processes. Understanding and predicting these clustering dynamics is important in fields such as astrophysics, climatology, and engineering, where they influence processes like cloud formation, sediment transport, and dust aggregation.

Direct Numerical Simulation (DNS) provides the most accurate representation of turbulence by calculating fluid motion using the Navier–Stokes equations. However, as the Reynolds number (dimensionless metric that indicates whether flow will be laminar or turbulent) increases, the range of turbulent scales grows rapidly, which makes DNS computationally prohibitive. Machine learning provides a potential solution because it can learn complex nonlinear relationships between flow parameters and turbulent statistics directly from data.

The objectives of this case study are twofold. First, the model aims to predict the four summary statistics of the particle cluster volume distribution for new combinations of Reynolds numbers (Re, how laminar to turbulent the flow is), Froude numbers (Fr, ratio of inertia to gravitational forces), and Stokes number (St, particle deviation from flow), thereby enabling accurate prediction of clustering behavior across different flow regimes. Second, the analysis seeks to interpret how each parameter influences the shape and variability of the particle cluster distribution, including potential nonlinear and interaction effects.

Methods

Data Wrangling

Our first step was to convert the raw moment data in the metrics we are interested in predicting with our model, the mean, variance, skewness, and kurtosis. Next, we realized that the Fr number had several infinity values (suggesting negligible gravity in those simulations). To utilize these observations in our analysis, we applied an inverse logit transformation, which compresses values between 0 and 1:

$$Fr_{\text{invlogit}} = \frac{1}{1 + e^{-Fr}}$$

Despite Re and Fr having only three distinct values, we opted to keep them as quantitative variables in our model because, in future prediction, the model should be able to account for a range of values for these two variables.

Exploratory Data Analysis

Next, we were interested in examining the data to conduct exploratory data analysis and understand the relationship between the variables at play. We fit a ggpairs() plot to visualize the correlations and relationships between variables, and then we created individual histograms to visualize the distributions of all variables. We decided to apply a log transformation to the following variables to reduce skewness: St, mean, variance, skewness, and kurtosis. Correlations between predictors wasn't a major issue based on our EDA.

Modeling Approach

With the need to fit four separate models, one for each metric of the cluster distribution, we employed the same workflow across all four models. We employ regression models to maintain solid predictive performance while not sacrificing interpretability. For each model:

Fit the full linear regression model with all two-way interaction effects, then perform best subset selection to select the best model using AIC as the selection metric. We use best subset selection over forward vs backward selection because we do not have a large number of variables to try out, so the fit is computationally not excessively expensive. We use AIC as the selection metric because model simplicity is not of utmost concern; there aren't many variables to start with, so the penalty that BIC induces might be too harsh. The hierarchy principle is imposed to ensure interpretability.

Fit all variables and interactions on a lasso model, allowing lasso to select variables of interest by zeroing out unimportant terms. We were interested in lasso as lasso can often offer better predictive performance over OLS in cases where overfitting may be of concern (In our case, we do not seek to simply optimize interpretability/fit on the data at hand; we are also interested in future prediction).

Compare the best subset, AIC-selected model, with the lasso model using ten-fold cross-validation. Then we select the one with the lower CV using RMSE as our evaluation metric.

Prediction and Uncertainty Quantification

For each of the four separate models, we must make predictions using the model on the testing dataset. To quantify uncertainty in predictions, we obtain prediction intervals to understand and account for prediction errors and potential noise. Extracting 95% prediction intervals for OLS is straightforward using their built-in functions, but when lasso is selected as a model, we use a proxy to estimate the interval: we extract the prediction from lasso, and obtain the prediction interval by fitting the lasso-selected variables on OLS to obtain the intervals from OLS as an imperfect estimation.

Results

Mean Model

$$\log(\text{Mean}) = \beta_0 + \beta_1 \log(St) + \beta_2 Re + \beta_3 Fr_{\text{invlogit}} + \beta_4 (Re \times Fr_{\text{invlogit}}) + \varepsilon$$

F-statistic p-value of almost 0 indicates that the model provides better explanatory power than a null model. The R^2 value of 93% indicates that the model is able to explain a high proportion of the variability in log(mean). These two metrics suggest that the model does fairly well from an explanatory power perspective (looking at training data). However, the residuals vs fitted plot shows a systematic pattern (instead of the residuals being randomly scattered around 0), suggesting a dimension of the relationship that isn't captured by linear predictors. This can be evidence of model misspecification. Additionally, the Q-Q plot shows an S-shaped curve around the diagonal line, suggesting that the normality of residuals assumption is not satisfied. These two violations likely didn't worsen explanatory power; however, the reported coefficient standard errors and potentially even the p-values may lose accuracy, and we may get more uncertain prediction intervals.

For our $\log(St)$ term, our coefficient is 0.2 (p<0.001, SE = 0.058). We can interpret the coefficient as for a 10% increase in St, we expect the mean particle cluster volume to be multiplied by a factor of about $1.1^{0.2} = 1.02$. This positive correlation implies that higher St (particle inertia dominates fluid flow) leads to increased clustering in particles. Higher turbulence (Re = -0.02, p<0.001) and higher ratios of inertia-to-gravity force (Fr_{invlogit} = -0.947, p>0.05) both reduce mean cluster size, as more chaotic and irregular flow patterns and reduced gravitational pull disperse particles into smaller volumes. For our interaction term, our coefficient is 0.005 (p<0.001, SE = 0.002). We see that the positive (albeit small) value indicates that the effect of Re on mean particle cluster volume is increased by larger values of Fr_{invlogit}, indicating that higher turbulence will have a greater impact on mean cluster volumes when there are also large inertial forces at play relative to gravitational impact.

For model selection, our CV RMSE for OLS is about 0.6, whereas the lasso yielded .62 at the best lambda value. The marginal difference suggests the induced bias from lasso is not worth it in the overall bias/variance tradeoff.

Variance Model

$$\log(\text{Variance}) = \beta_0 + \beta_1 \log(St) + \beta_2 Re + \beta_3 Fr_{\text{invlogit}} + \beta_4 (Re \times Fr_{\text{invlogit}}) + \varepsilon$$

We have a R^2 value of 0.714, meaning that our model explains roughly 71.4% of the variance in log(variance) and indicates that have a reasonably good fit, though not as strong as our mean model. Our adjusted R^2 of 0.7 is very close to our normal R^2, and our f-statistic of 52.545 ($p < 0.001$) means that our predictors collectively have an extremely significant relationship with variance. The residuals vs fitted plot shows clear heteroscedasticity, in which the residuals clearly form a fan shape at higher values. Hence, our constant variance assumption is violated. Moreover, the Q-Q plot shows relatively heavy tails with some slight initial deviating points (87, 29, 20), but the skew isn't too extreme. Since all our effects are extremely significant, these limitations won't muddy our main conclusions, but they're important to note.

For our $\log(St)$ coefficient of 0.89 ($p < 0.001$, SE = 0.201), we can see a power-law relationship where doubling St multiplies variance by about $2^{0.89}$ = ~1.85x. This means that particles where inertia dominates flow (larger Stokes number) create more variable cluster sizes on average. For our Re coefficient of -0.049 ($p < 0.001$, SE = 0.006) we see that each unit increase in Reynolds number decreases log(variance) by 0.049, meaning variance is multiplied by $e^{-0.049}$ = ~0.952. This indicates that higher turbulence intensity reduces the variability in cluster sizes. This implies that higher amounts of turbulence would break up clusters and make them more homogeneous/uniform. Additionally, our Fr_{invlogit} of -11.911 ($p < 0.001$, SE = 2.140) is our baseline effect, meaning that increasing inertial force relative to gravitational force dramatically reduces variance in cluster sizes. Lastly, our interaction term with $Re \times Fr_{\text{invlogit}} = 0.036$ ($p < 0.001$, SE = 0.008) is our key interaction and shows the effect of Fr on variance changes depending on turbulence. More specifically, at $Re = 224$, our total effect is $-11.911 + 0.036(224) = -3.85$ (variance multiplies by $e^{-3.85}$ = ~0.021) and at $Re = 398$, our total effect is $-11.911 + 0.036(398) = -3.58$ (variance multiplies by $e^{-3.58}$ = ~0.028). This indicates that as turbulence increases, Fr's negative effect on variance decreases.

For our model selection, our CV RMSE for OLS is 2.092 while for LASSO we have 2.111, a minor difference of 0.019. It makes sense that OLS was slightly better, as LASSO generally adds bias through regularization to reduce variance, but since it wasn't really a problem, it just hurt performance slightly instead.

Skewness Model

$$\log(\text{Skewness}) = \beta_0 + \beta_1 Re + \beta_2 Fr_{\text{invlogit}} + \beta_3 (Re \times Fr_{\text{invlogit}}) + \varepsilon$$

This model explains approximately 55% of the variation in log-skewness ($R^2 = 0.55$), suggesting a moderate fit that, while capturing key physical trends, does not explain as much variability compared to our mean and variance models. Residual and Q–Q plots indicate approximate normality, with some heteroscedasticity—reasonable given that skewness (a third-order statistic) may fluctuate strongly due to variations in particle simulations.

Three predictors were retained: Froude number (Fr_{invlogit}), Reynolds number (Re), and their interaction ($Re \times Fr_{\text{invlogit}}$). For our Re coefficient of –0.00224 (p = 0.348, SE = 0.00237), each unit increase in Reynolds number decreases log(skewness) by 0.00224, which corresponds to a multiplicative factor of $e^{-0.00224} \sim 0.998$. This means that turbulence slightly reduces skewness, though the effect is statistically insignificant, aligning with physical expectations that stronger turbulence helps homogenize particle distribution. For our Fr_{invlogit} coefficient of –4.62 (p < 0.001, SE = 0.813), we can see that weaker gravity dramatically reduces skewness by about 99% ($e^{-4.62} \sim 0.0099$). Physically, this suggests that when gravitational effects are small relative to inertial force, particle cluster distributions become far less skewed, as gravity no longer pulls particles into dense, asymmetric formations. The interaction term ($Re \times Fr_{\text{invlogit}}$ = 0.0119, p < 0.001, SE = 0.00317) implies that the effect of Fr on skewness depends on turbulence intensity. For example, at $Re = 90$, the total effect is $-4.62 + 0.0119(90) = -3.55$, meaning skewness is multiplied by $e^{-3.55} \sim 0.028$, while at $Re = 398$, the effect becomes $-4.62 + 0.0119(398) = 0.11$, which corresponds to $e^{0.11} \sim 1.12$, a slight 12% increase in skewness. This means that at low turbulence, Fr may reduce asymmetry in particle distributions, but as turbulence grows stronger, Fr may increase asymmetry in particle distribution.

Cross-validation favored OLS over LASSO by a small margin (CV RMSE: 0.775 vs. 0.788). Although the CV difference is marginal, this could suggest that the induced bias offered by lasso did not overall improve predictive power.

Kurtosis Model

$$\log(\text{Kurtosis}) = \beta_0 + \beta_1 \log(St) + \beta_2 Fr_{\text{invlogit}} + \beta_3 (\log(St) \times Re) + \beta_4 (Re \times Fr_{\text{invlogit}}) + \varepsilon$$

The kurtosis model achieved an R-squared of 0.549, indicating that approximately 54.9% of the variation in log(kurtosis) is explained by the selected predictors. The F-statistic of 20.64 (p < 0.001) indicates that the model predictors collectively have a significant relationship with log(kurtosis). The residuals vs fitted plot shows a clear non-random pattern, with residuals peaking around fitted values near 9. This suggests possible nonlinearity and heteroscedasticity, meaning the model may not fully capture the relationship between variables or that error variance changes with fitted values. The Q-Q plot shows that residuals deviate from the straight diagonal line, especially in the lower tail, indicating that they are not normally distributed. This suggests the model's residuals may exhibit skewness or heavy tails, violating the normality assumption.

The LASSO model with a lambda value of 0.026 reduced the coefficients of Re and $\log(St) : Fr_{\text{invlogit}}$ to zero, suggesting these variables did not significantly improve predictive performance. The LASSO model also achieved a slightly lower cross-validated RMSE (1.562) than the OLS model (1.572), suggesting marginally better generalization. Interpreting the coefficients on the original kurtosis scale, a one-unit increase in $\log(St)$ multiplies the predicted kurtosis by $\exp(-0.129)$ (p = 0.568, SE = 0.320), indicating that inertial-dominated particles produce cluster distributions with smaller tails, though this effect is statistically insignificant. Likewise, a one-unit increase in Fr_{invlogit} multiplies kurtosis by $\exp(-7.583) \sim 0.0005$ (p < 0.001, SE = 1.622), representing a dramatic decrease in tail sizes (lower probability of observing extreme particle cluster volumes). The positive interaction between Re and Fr_{invlogit} ($\beta = 0.0171$) (p < 0.001, SE = 0.006) means that for each one-unit increase in Re, the effect of Fr_{invlogit} on kurtosis is multiplied by $\exp(0.0171) \sim 1.017$, suggesting that as turbulence intensity rises, inertial domination over gravity's effect on kurtosis increases. The $\log(St) \times Re$ interaction (–0.000021, p = 0.922) is negligible. Because the LASSO model does not produce standard errors or p-values, these statistics were obtained from an unpenalized OLS proxy model refitted using the predictors selected by the LASSO.

Conclusions

This study met its objectives by building and comparing interpretable regression models (OLS, LASSO) and flexible GAMs to predict four summary statistics of particle-cluster volume distributions (mean, variance, skewness, kurtosis) from Reynolds number (Re), Froude number (Fr), and Stokes number (St). Overall, the models capture clear, physically interpretable relationships: gravitational-dominating forces (low Fr) generally increases clustering intensity and reduces variability and skewness, increasing particle inertia's dominating effect on its flow (higher St) tends to weaken sharp clustering and increase variance, and turbulence (Re) modulates these effects. Quantitatively, the fitted models explain a large share of variability for mean ($R^2 \approx 0.93$) and variance ($R^2 \approx 0.71$), with more moderate performance for skewness ($R^2 \approx 0.55$) and kurtosis ($R^2 \approx 0.55$).

Model selection shows a trade-off between bias and variance. OLS had slightly lower CV RMSE for mean, variance, and skewness, suggesting that regularization was not necessary for those responses given the available predictors and sample size. LASSO performed best for kurtosis, where penalization helped by simplifying the model and reducing overfitting risk. For inference, we relied on an OLS proxy fitted to the LASSO-selected predictors because LASSO does not provide valid standard errors or p-values; this yields approximate inference but should be interpreted with caution. Uncertainty quantification revealed significant practical limitations across all models. For the mean model, 95% prediction intervals have an average width of about 0.1, which in the context of the values we typically see for the mean (in the training data) is not particularly helpful, as most of our mean values are already extremely small, though upon inspecting individual datapoints, many have reasonable widths of prediction. For the variance model, the 95% prediction interval has a median width of 11.77, with a median relative width of 6685% of the predicted variance, a much better measure than our mean width of 189.36, which is inflated by outliers in the back-transformed scale. This is obviously quite wide and is effectively useless for precise predictions, as the typical interval is about 67 times wider than the predicted value. That said, it's somewhat expected as we have only 3 observed values for Re, and so predictions outside this set would be understandably uncertain. Additionally, the heteroscedasticity and log scale SE measure is not fully indicative of the full range, as we will eventually need to back-transform with exp(). For the skewness model, 95% prediction intervals had a mean width of approximately 185 units and a median width of 142 units, corresponding to an average relative width of about 70–90% of the predicted value. These are very wide, reflecting both the limited range of observed Reynolds numbers (90, 224, and 398) and the inherently high variability of turbulent systems. For the kurtosis model, the 95% prediction interval width is 1,066,339, which is exceptionally wide. This large interval likely arises from the exponential back-transformation of a log-scaled model, where small differences on the log scale translate into large absolute differences on the original scale. Additionally, the underlying log(kurtosis) values span a wide range (from 5.014 to 11.792), amplifying uncertainty when converted back to the original scale. Together, these factors make the prediction interval appear disproportionately large even though the model performs reasonably well on the log scale. Overall, these models are good for understanding how variables affect clustering statistics, but would not be very reliable for predicting exact values at new conditions given the limited Re coverage.

Diagnostics indicate some model assumption violations including nonlinearity, heteroscedasticity, and departures from normality for residuals in some models. These issues may degrade the accuracy of standard errors and p-values and suggest the potential value of more flexible parametric forms, GAMs, or other nonparametric approaches where appropriate. The dataset's sparse coverage of Re and the presence of infinite Fr values (handled with an inverse-logit transform) limit extrapolation and increase predictive uncertainty for unobserved regimes. To improve predictive reliability and scientific insight we recommend: (1) augmenting the dataset with additional simulations covering a wider range of Re and Fr, (2) performing targeted DNS or high-fidelity simulations to validate model predictions, (3) using simulation-based prediction intervals for final reporting, and (4) investigating physics-informed or hybrid learning approaches that incorporate known mechanistic constraints.

The models provide interpretable relationships linking Re, Fr, and St to particle clustering statistics and can serve as computationally cheap alternatives for DNS. However, users should be cautious about predictions outside of the training data range and should rely on proper uncertainty quantification when using model outputs for decision-making.