Particle Clustering in Turbulence Case Report

Jack Lichtenstein, Abbey List, Jingxuan Liu, Linda Tang, Mary Wang, and Justin Zhao

1 Introduction

Turbulence is a fundamental concept in fluid mechanics. Irregular, unpredictable, and energy-dissipating, turbulent flow enhances mixing, which leads to non-uniform distribution of particles and cluster formation. Understanding and predicting turbulence has great practical significance in many scientific areas including aeronautical engineering and environmental science.

Existing research suggests that the clustering of particles subject to turbulent flow is mainly determined by three main parameters: Reynolds (Re) number, Froude (Fr) number and Stokes (St) number, which correspond to the intensity of turbulence, impact of gravitational acceleration, and particle size properties (Chanson, 2009; Ireland et al., 2016). These parameters may influence clustering individually; a large St, for example, correlates with large particle size, which tends to form relatively loose clusters (Ireland et al., 2016). The parameters may also interact with each other to impact cluster formation.

Despite the importance of turbulence, current understanding of *how* Re, Fr, and St contribute to clustering remains rudimentary. Simulation methods like the Direct Numerical Simulation (DNS) of Navier-Stokes equations have been applied, but progress is limited by the time-consuming and computationally-expensive nature of such methodologies (Moin & Mahesh, 1998). Leveraging on generated data, the present study *aims* to build a statistical model that investigates how Re, Fr and St influence particle cluster volume distribution. We hope that our model will 1) enhance our understanding of the relative influence of these parameters in turbulent flow, and 2) enable an efficient and accurate prediction of a particle cluster volume distribution without the need for simulation.

2 Methods

2.1 Data The data ($n = 89$) for the present study come from Direct Numerical Simulation. Voronoi Tessellation, a technique that examines general features of individual clusters in the underlying turbulence, was applied to generate a distribution of cluster volumes. The original data contains information regarding the generated distributions in the form of the first four raw moments $\mathbb{E}(X)$, $\mathbb{E}(X^2)$, $\mathbb{E}(X^3)$, and $\mathbb{E}(X^4)$, as well as Re, Fr, and St values.

For better interpretability, we transformed the 2nd, 3rd, and 4th raw moments into central moments (Moment 1 is untouched as it already signifies the mean), which are related to the variance (how flow varies over time), skewness (indication of symmetric properties of the flow), and kurtosis of the distribution (how particular cluster volumes deviate). In addition, despite their numerical natures, Fr and Re only contain three levels ($Fr \in \{0.052, 0.3, \text{inf}\}$ and $Re \in \{90, 224, 398\}$, see Figure S1 for distribution) in the training data. We decided to model them as categorical variables because 1) 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 2) we believe that these categories could carry some real life significance. For instance, Fr = 0.3 is representative of cumulonimbus clouds and 0.052 is representative of cumulus clouds. Interpreting their effect in categories can offer insight on practical examples (?).

2.2 Model Construction 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 (Figures S1-S4 in the Appendix). These relationships vary between roughly linear to noticeably curved, potentially quadratic or cubic. In addition, such relationships diverge across different moments, and appear highly influenced by the combined levels of Re and Fr. From the curved shape of the

relationship between St and the responses, we also noticed that we may benefit from taking certain roots of St to increase linearity.

Given the paucity of theoretical background in related fields to guide model building, we decided to adopt a k-fold cross validation approach. This approach would allow us to explore different combinations of polynomial patterns and the interactions observed in our EDA and select the best fit model. Specifically, we used K = 5 given the small sample size at hand to avoid overfitting.

To implement the k-fold cross validation, for each moment, we trained candidate models to predict the moment with the general formula $Moment_i \sim poly(St_i^{(1/root)}, degree) * (Fr, Re)$, varying the *degree* parameter from 1 to 3, *root* from 1 to 6, and testing a log transformation of the response. Literature supports a logarithmic dependence of acceleration on the Reynolds number (Zamansky, 2013); thus, we combined Fr and Re to form a new interaction variable (Fr, Re) with 9 levels representing all combinations of Fr and Re. We tested the model on each fold, and chose the parameters that gave lower root mean squared errors (See Figure 1), with preference for less complexity if the error is similar. After selecting the features for each moment's model, we then fit the final models on the full training dataset using standard least-squares for interpretation.

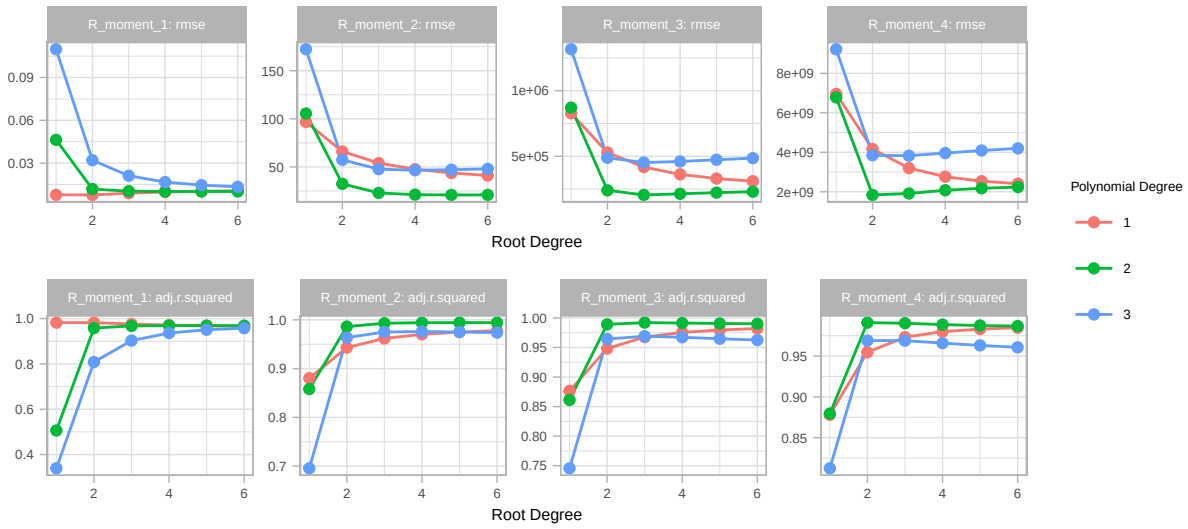


Figure 1. Root mean squared error and mean adjusted r-squared value of models for each moment with varying root and polynomial degrees on St. The combination that lead to the lowest root mean squared error was selected as the final model for the moment.

Based on the k-fold cross-validation approach, the final models for each moment are as below:

$$\text{Raw Moment}_1 \sim St * (Fr, Re)$$
$$\text{Central Moment}_2 \sim \text{poly}(St^{1/4}, 2) * (Fr, Re)$$
$$\text{Central Moment}_3 \sim \text{poly}(St^{1/3}, 2) * (Fr, Re)$$
$$\text{Central Moment}_4 \sim \text{poly}(\sqrt{St}, 2) * (Fr, Re)$$

The linearity, constant variance, and independence assumptions for linear regression are satisfied, but we observed some violation of normality. There are a few outliers and influential points. Furthermore, the variance inflation factors indicated no serious concerns of multicollinearity (Model diagnostics section in the Appendix).

2.3 Model Extension 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. To address the limitations of our current models, we provide the following related models using natural splines that can be used for extrapolation:

$$\text{Raw Moment}_1 \sim \text{ns}(St, df = 1) * Re * Fr'$$

$$\text{Raw Moment}_{2,3,4} \sim \text{ns}(\sqrt{St}, df = 1) * Re * Fr'$$

In these models, Re is numeric and Fr is transformed to a numeric variable on $[0, 1]$ using $Fr' = \frac{2}{\pi} * arctan(Fr)$. The form is similar in keeping the significant three-way interaction, except with natural splines to address the poor fits of polynomials at the tails, a location that is especially important in learning about particle behavior in high turbulence. Again, the root and degree are chosen through 5-fold cross validation. See Appendix for fitted coefficients.

Results

Selected outputs from the final models of each moment are displayed in Table 1. Please refer to Table S1 for the error values for the final models and Tables S2-S5 for full outputs.

Term	β	p value	95 CI Lower	95 CI Upper
Moment 1				
interaction0.052: 90	0.127	<0.001	0.122	0.131
interaction0.3: 90	0.098	<0.001	0.093	0.103
interactionInf: 90	0.093	<0.001	0.089	0.098
poly(St, 1):interaction0.052: 90	0.137	<0.001	0.097	0.177
poly(St, 1):interaction0.3: 90	0.26	<0.001	0.218	0.301
poly(St, 1):interactionInf: 90	0.247	<0.001	0.204	0.29
Moment 2				
interaction0.052: 90	698.156	<0.001	683.192	713.119
poly(St, 2)1:interaction0.052: 90	2603.304	<0.001	2468.448	2738.161
poly(St, 2)2:interaction0.052: 90	-792.335	<0.001	-927.47	-657.2
Moment 3				
interaction0.052: 90	5693362.973	<0.001	5550918.321	5835807.624
poly(St, 2)1:interaction0.052: 90	23598512.694	<0.001	22314031.293	24882994.094
poly(St, 2)2:interaction0.052: 90	-6676295.814	<0.001	-7962838.718	-5389752.911
Moment 4				
interaction0.052: 90	46674298604.957	<0.001	45374438727.779	47974158482.135
poly(St, 2)1:interaction0.052: 90	208144184726.544	<0.001	196414359316.744	219874010136.345
poly(St, 2)2:interaction0.052: 90	-66729237684.868	<0.001	-78494438928.089	-54964036441.647

Table 1. Outputs from final models of each moment. Only the significant terms are displayed here given limited space. Please refer to Tables S2-S5 in Appendix for full output.

Firstly, all models possess an adjusted R^2 of at least 0.98, indicating that at least 98% of the variation in each centralized moment is explained by the predictors in the models (Table S1 in the Appendix). The test error (modeled by root mean squared error) from k-fold cross validation is only 0.00764 and 20.86 for models of centralized moments 1 and 2, respectively. Although the values are larger for models of centralized moments 3 and 4 (above 200,000 and above 1 million, respectively), this is expected given the extremely large variation in these moments. Overall, these metrics indicate that our model has good predictive power.

To better understand the relative influence of turbulence, gravitational acceleration and particle characteristics on cluster formation, we examined the model coefficients. For centralized moment 1, we see that interactions between all settings of Fr and the lowest setting of Re, along with their interactions with St,

yield an expected increase in mean particle cluster volume size. This indicates that smoother flow at any level of gravitational acceleration and particle size may allow larger clusters to form, as they are not broken up by chaotic turbulence. These coefficients are statistically significant at the $\alpha = 0.01$ level with small standard errors. For centralized moments 2, 3, and 4, it appears that the interaction between lowest Fr and lowest Re, along with its interaction with St, has a positive association with each moment. The interaction with squared St has a negative association with each moment (note that the exact value of this squared transformation differs based on the root transformation applied to St for each moment) (?????? ask simon about interpreting positive non-squared, negative squared). This indicates that smoother flow and low gravitational acceleration are associated with greater variance, skew, and kurtosis (distributional "tailedness") of particle cluster size. The effect of particle size at these levels of Re and Fr may differ from the effect at other levels, as it is only statistically significant at these low settings of flow intensity and gravitational acceleration. These coefficients are statistically significant at the $\alpha = 0.01$ level.

Discussion

In this study, we constructed four linear models with polynomial terms and interaction terms between Reynolds' number, Froude's number, and Stokes' number to predict the first four centralized moments (related to mean, variance, skew, and kurtosis) of particle cluster distribution. Our models were able to explain a large percentage of variation in each of the four moments with reasonable mean squared prediction error for each model. We found that smoother flow at any level of gravitational acceleration and particle size may be associated with larger mean cluster size, and smoother flow with low gravitational acceleration seemed to have a positive association with variance, skew, and kurtosis. Also, the effects of particle size on each moment seemed to be significantly different at lower flow intensity and lower gravitational acceleration than at other settings. Finally, it appears that for centralized moments 2, 3, and 4, the main effect of St (at low Fr and Re) has a positive association with the moments, while the squared effect has a negative association (how to explain this ?).

One limitation of our analysis is that the coefficient estimates may not be accurate for our models since the 95% confidence intervals are relatively wide. This likely results from having a small training set with less than 100 data points. In addition, the normality assumption of linear regression is violated due to the extreme skewness/variation of the moments and several outliers for each model. Our training set may not be a representative sample that allows our model to learn the underlying trend and model the highly complex distribution of the moments. We recommend using caution in interpreting the exact coefficient estimates.

Furthermore, Fr and Re were used as categorical variables in our models despite that they take on numeric values in practice. This approach limits the generlizability of our model since it cannot be applied on values outside the categories. We attempt to address this limitation by building a natural spline model while treating Fr and Re as numeric variables to allow for extrapolation. However, the predictive power of this model has not been thoroughly assessed, and this model likely requires a much larger training set to fit well.

In addition, given the training set, our model was only fit on Reynolds numbers up to 398, although numbers in the thousands are common in practice. Large Reynolds numbers are highly relevant to real life situations as a measure of turbulence intensity in atmospheric, oceanic turbulence flow (J. den Toonder et al., 1997). Although our modeling approach avoids the high computational cost of simulation at large Reynolds numbers, it may not be suitable for extrapolation at these settings. Using numerical variables and higher Reynolds numbers may be topics for further investigation.

In summary, our models assessed the effects of Reynolds, Froude, and Stokes numbers on particle cluster size and can be used in predicting particle cluster volume distribution given certain restrictions. We hope our results could inform further research on understanding particle clustering in turbulence and improve the efficiency of prediction.

References

- J. M. J. den Toonder, & Nieuwstadt, F. T. M. (1997, November 1). Reynolds number effects in a turbulent pipe flow for low to moderate re. AIP Publishing. Retrieved October 12, 2021, from https://aip.scitation.org/doi/pdf/10.1063/1.869451.
- Schlichting, H., Gersten, K., Krause, E., & Oertel, H. (2017). Boundary-layer theory. Springer.
- Chanson, Hubert (2009). "Development of the Bélanger Equation and Backwater Equation by Jean-Baptiste Bélanger (1828)" (PDF). Journal of Hydraulic Engineering. 135 (3): 159–63. doi:10.1061/(ASCE)0733-9429(2009)135:3(159).
- Ireland, P. J., Bragg, A. D., & Collins, L. R. (2016, May 11). The effect of Reynolds number on inertial particle dynamics in isotropic turbulence. part 2. simulations with gravitational effects: Journal of Fluid Mechanics. Cambridge Core. Retrieved October 12, 2021, from https://www.cambridge.org/core/journals/journal-of-fluid-mechanics/article/effect-of-reynolds-number-on-inertial-particle-dynamics-in-isotropic-turbulence-part-2-simulations-with-gravitational-effects/67C55CDC28B1B1C7868B7A402E279AF9.
- Moin, P., & Mahesh, K. (n.d.). Direct numerical simulation: A tool in turbulence research. Annual Reviews. Retrieved October 12, 2021, from https://www.annualreviews.org/doi/full/10.1146/annurev.fluid.30.1.539.
- Zamansky, R., Vinkovic, I., & Gorokhovski, M. 2013. Acceleration in turbulent channel flow: universalities in statistics, subgrid stochastic models and an application. Journal of Fluid Mechanics. Retrieved October 14, 2021, from https://hal.archives-ouvertes.fr/hal-00931506/document.
- slides: https://sakai.duke.edu/access/content/group/e1e1b166-17bd-4efc-bdfb-f3909d696910/Case%20Study/Data_Expedition_F2020_Reza_Jon.pdf

Appendix

Section A: Figures

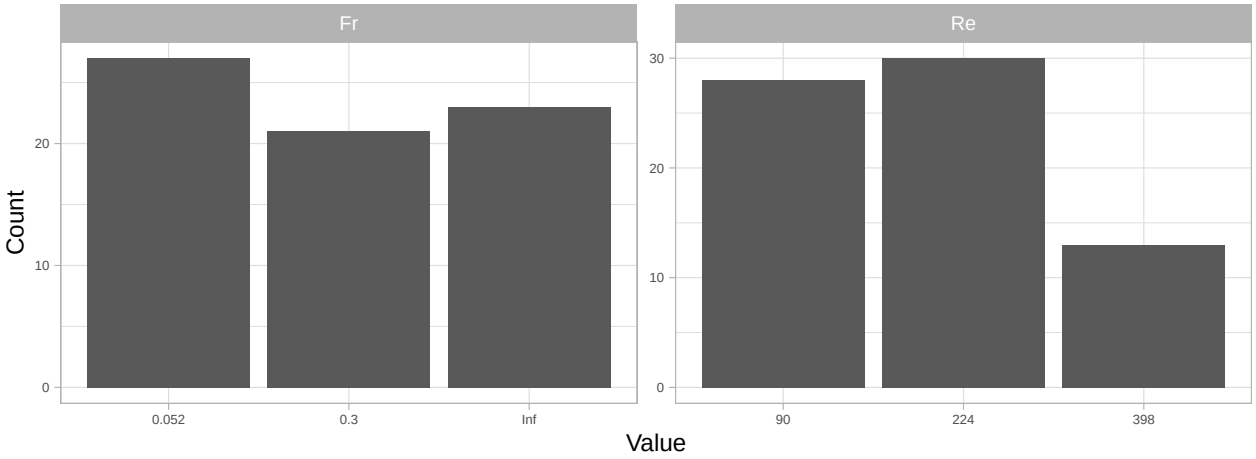


Figure S1. Distribution of the three levels of Re and Fr. The values are treated as factor levels in our main analysis.

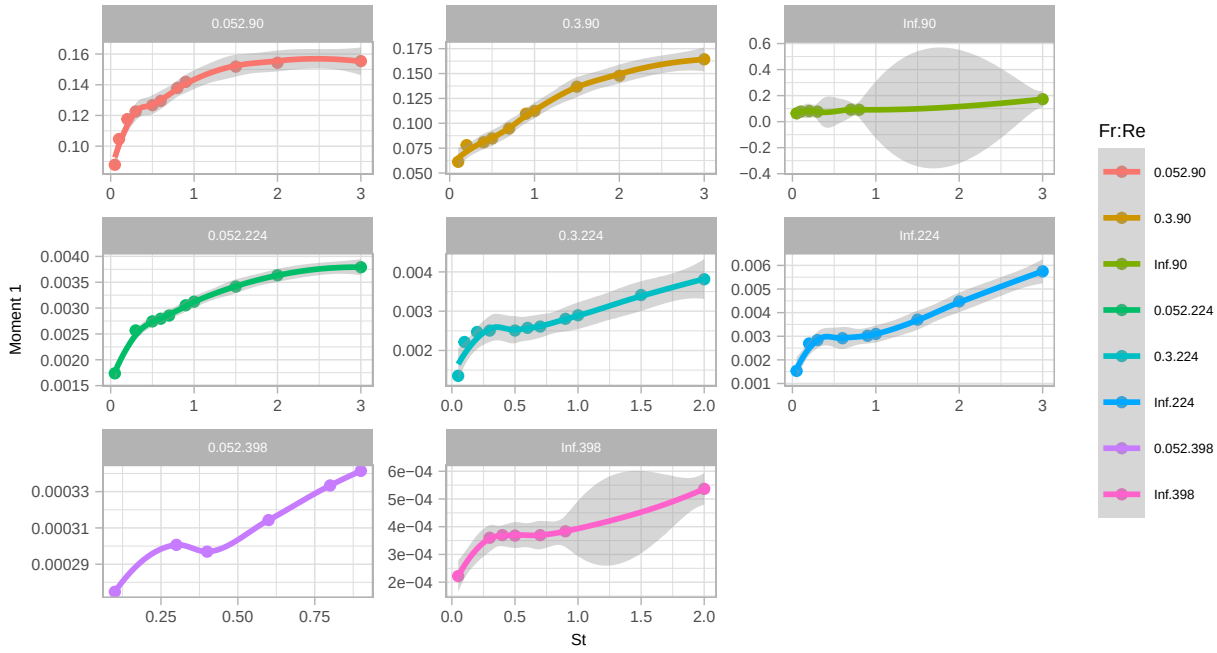


Figure S2. Moment 1 values as a function of St at different levels of interaction between Fr and Re.

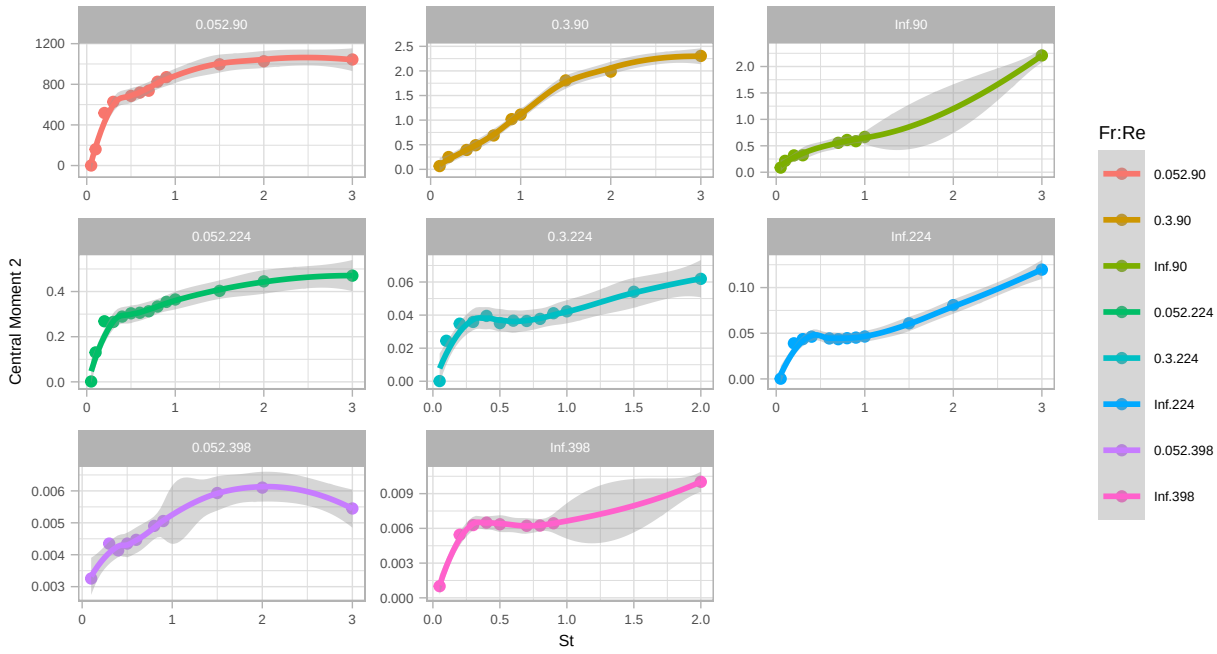
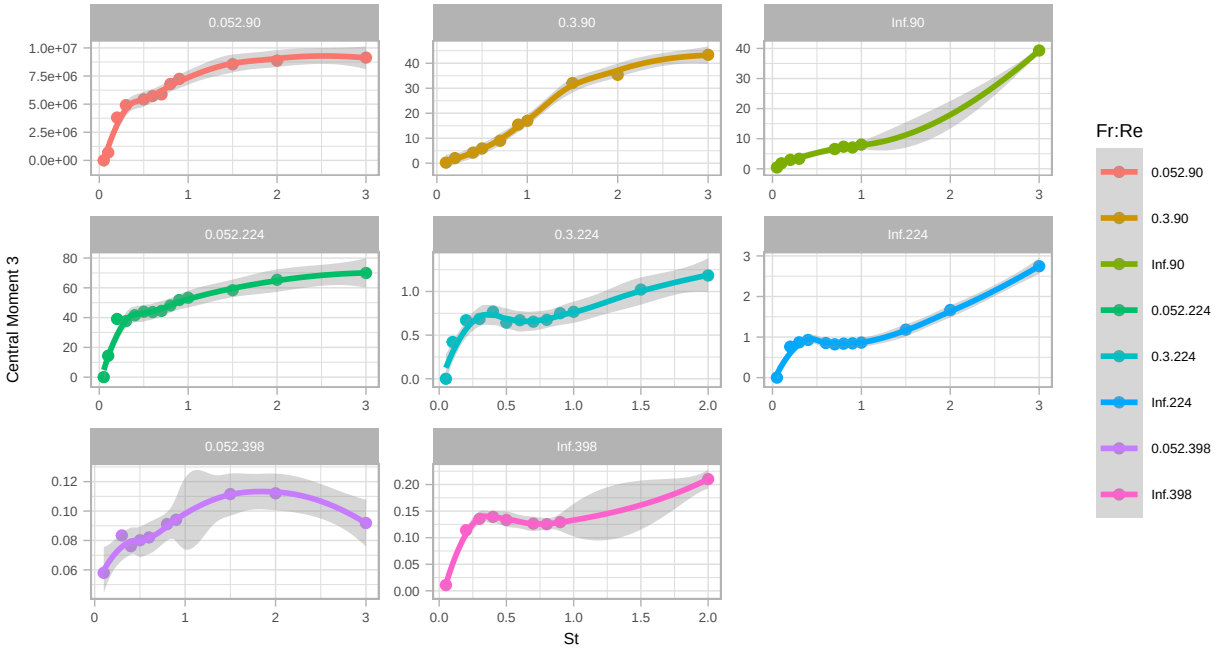


Figure S3. Central Moment 2 values as a function of St at different levels of interaction between Fr and Re.



Section B: Model Diagnoistics

Modeling assumptions for linear regression:

- **Linearity**: Linearity is satisfied for all 4 models since in the predicted vs. standardized residuals plots (Panel A), there's no obvious pattern in the residuals are the value of the predictors increase. The

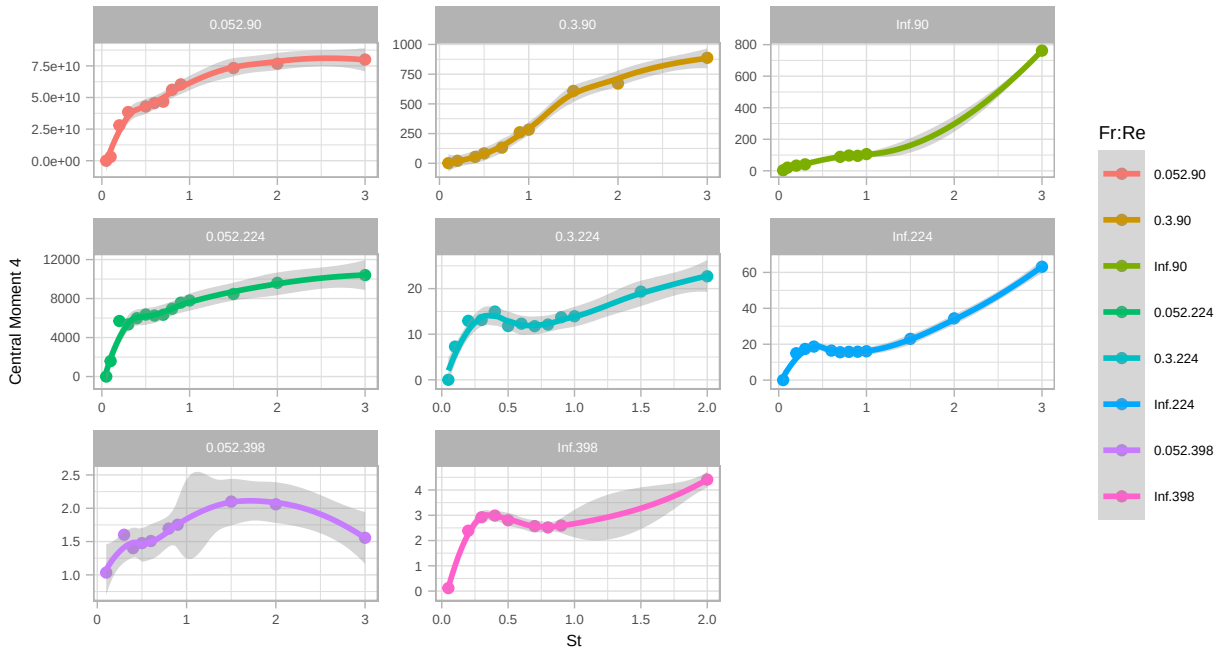


Figure S5. Central Moment 4 values as a function of particle size at different levels of interaction between Fr and Re.

residuals are randomly scattered around 0, supporting that there's a linear relationship between the predictors and the response (after variable transformation).

- **Constant variance**: In the residuals vs. predicted plot for all 4 models, the vertical spread of the residual remain relatively constant as the predicted values increases, suggesting that the variance of the error is constant along all predicted values (Panel A).
- **Independence**: The data was generated from Direct Numerical Stimulation (DNS) of the Navier-Stokes equations, where each trial was conducted independently using different values of the parameter (Re, Fr, St). Based on the information about data collection, we believe the independence assumption is satisfied.
- **Normality**: Normality may be violated since the distribution of the residuals doesn't follow a normal distribution and the points do not fall along a straight diagonal line on the normal quantile plot. The flat region in the normal quantile plot indicates there's a lot of nearly identical values, and the curved shape suggests that the distribution may be heavy-tailed. This violation makes sense in the context of the dataset since we observed that both the raw moments and central moments are highly skewed to the right with most values close to 0 with some extreme outliers (Panel B).

Influential points and outliers: We assessed influential points in our dataset using Cook's distance (Panel C). Besides having a few influential points for model 1, all observations had Cook's distance less than 0.5 for the rest of the models. In addition, from the predicted vs. standardized residuals plots (Panel A), there are only a few outliers with standardized residual greater than 5.

Multicolinearity: We assessed multicolinearity using generalized variance inflation factor (GVIF) accounting for the degree of freedom, which was below 5 for all predictors in all 4 models, suggesting no serious issue of multicolinearity.

The variable inflation factor (VIF) for all model coefficients

	GVIF	DF	GVIF (adjusted for DF)	Model
poly(St, 1)	6.228	1	2.496	model 1

	GVIF	DF	GVIF (adjusted for DF)	Model
interaction	1.323	7	1.02	model 1
poly(St, 1):interaction	8.024	7	1.16	model 1
poly(St, 2)	38.226	2	2.487	model 2
interaction	1.539	7	1.031	model 2
poly(St, 2):interaction	56.57	14	1.155	model 2
poly(St, 2)	38.253	2	2.487	model 3
interaction	1.577	7	1.033	model 3
poly(St, 2):interaction	57.982	14	1.156	model 3
poly(St, 2)	38.348	2	2.488	model 4
interaction	1.687	7	1.038	model 4
poly(St, 2):interaction	62.142	14	1.159	model 4

Section C: Tables

Model outputs are shown below:

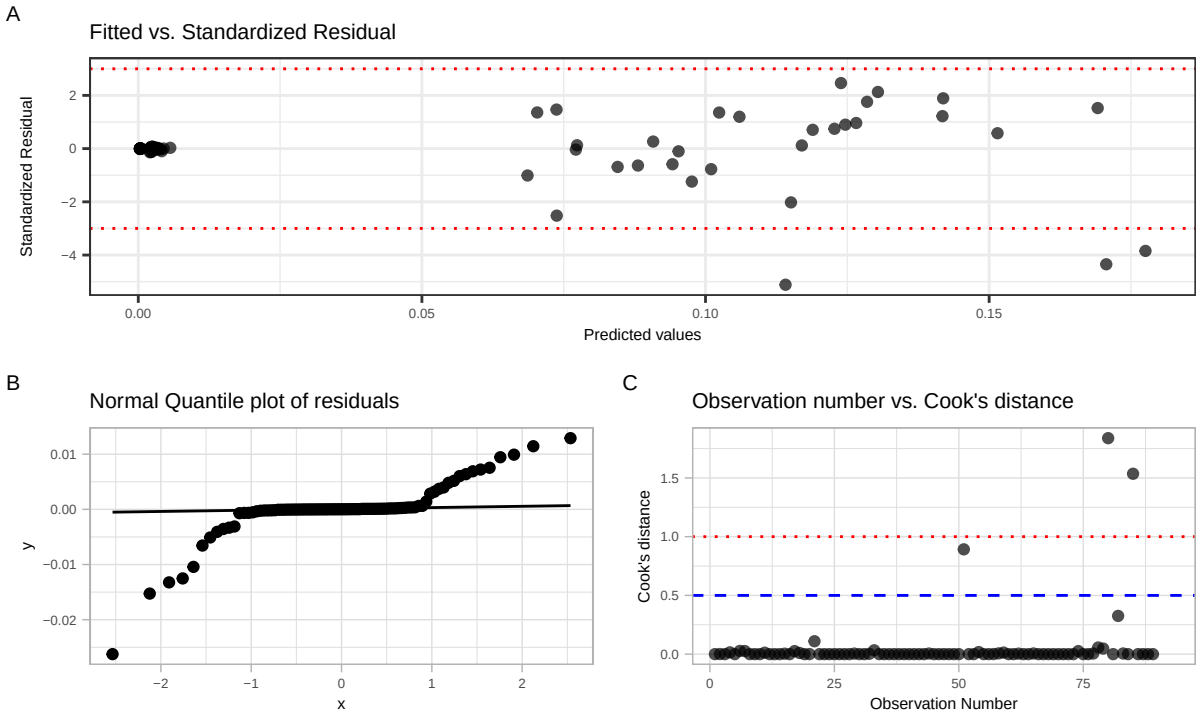


Figure S6. Diagnoistic plots for Model 1.

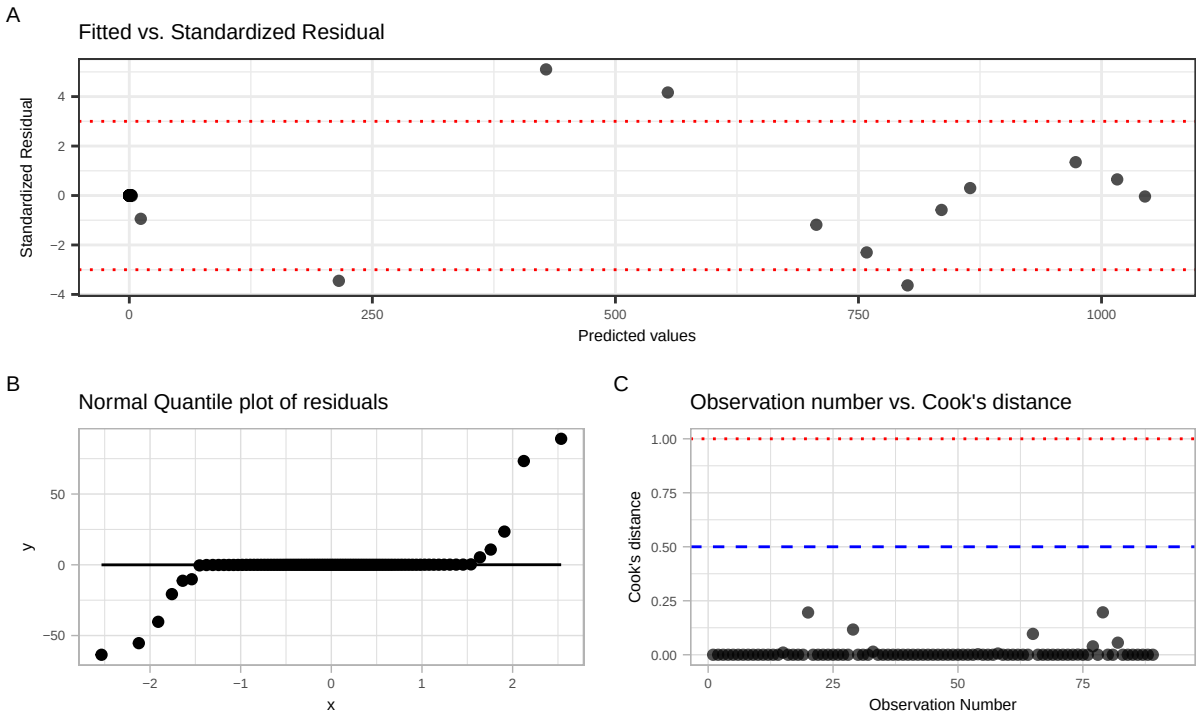


Figure S7. Diagnoistic plots for Model 2.

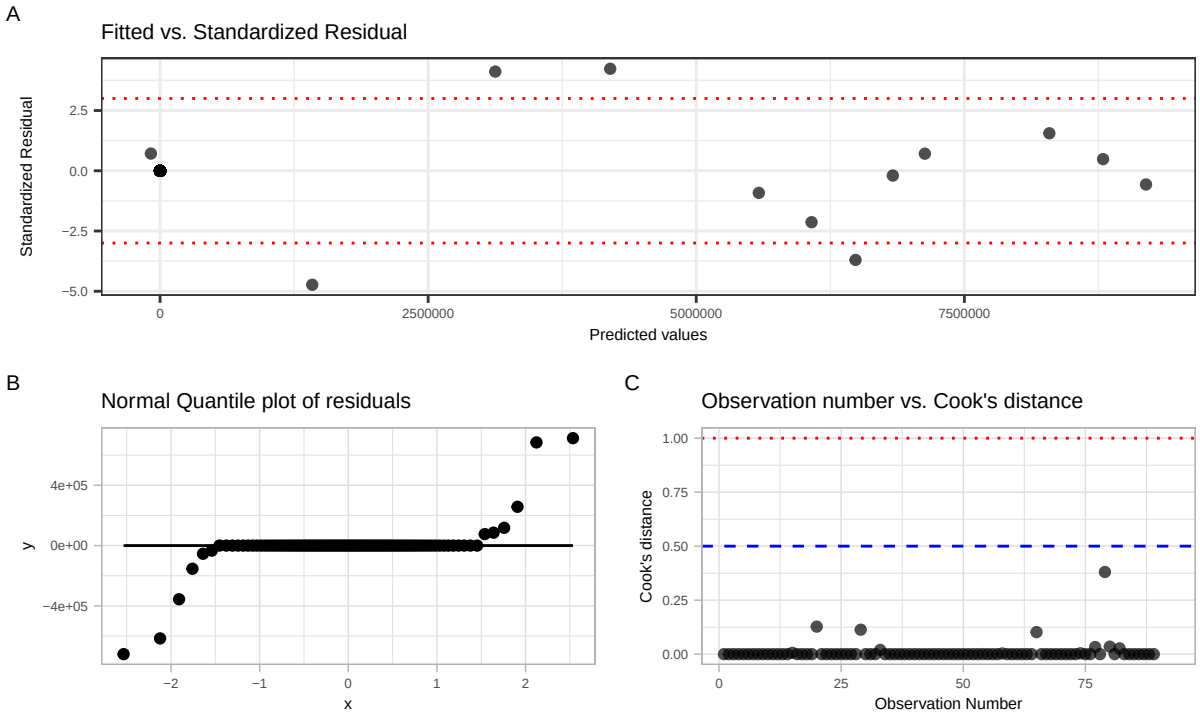


Figure S8. Diagnoistic plots for Model 3.

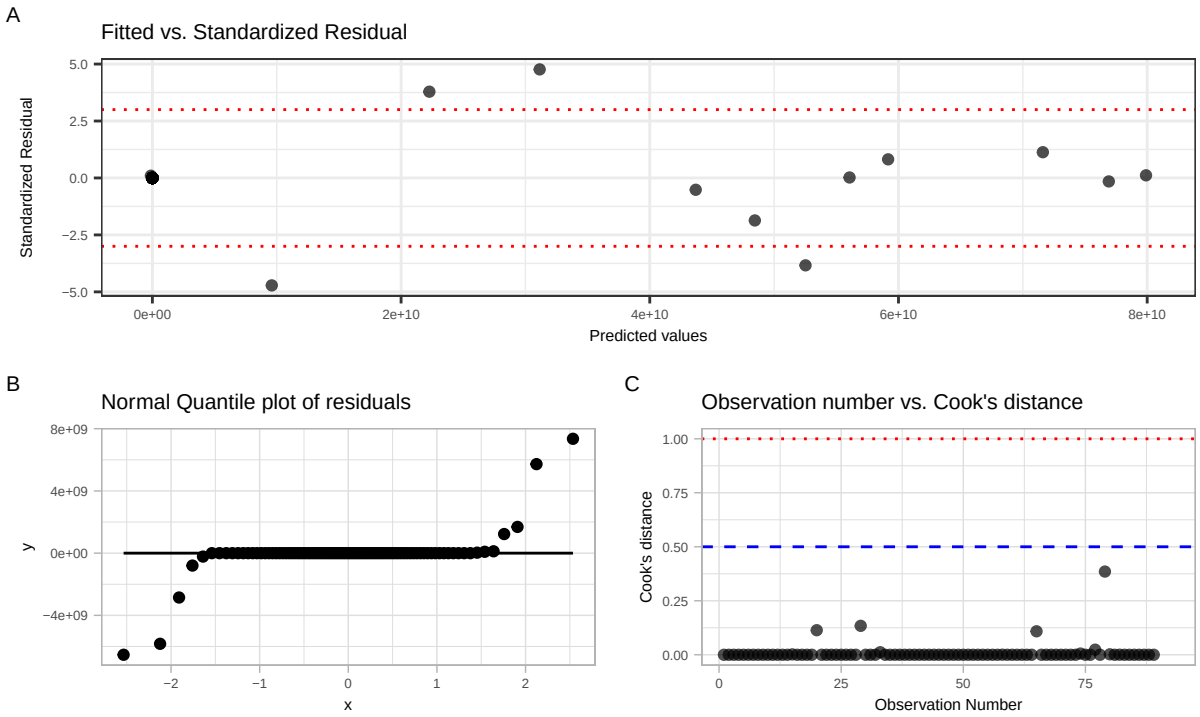


Figure S9. Diagnoistic plots for Model 4.