Effect of Turbulent Intensity, Gravity, and Particle Charateristic on Particle Clustering Distribution

David Jeong

2025/10/28

Introduction

Particle-laden turbulent flows appear in various engineering systems and natural phenomena. Yet understanding the interaction of inertial particles in turbulence may be extremely difficult to understand. In this study, we aim to explore the effect of fluid turbulence (quantified by Reynolds number, Re), gravitational acceleration (quantified by Froud number, Fr), and particle characteristics (quantified by Stokes number, St) on spatial distribution and clustering of particles in an idealized turbulence.

Specifically, we develop machine learning models to understand the relationship with two key objectives in mind:

- *Inference*: Investigate and interpret how each parameter (Re, Fr, St) affects the probability distribution for particle cluster volumes.
- *Prediction*: For unseen values of (Re, Fr, St), predict its particle cluster volume distribution in terms of mean, standard deviation, skewness, and kurtosis.

For the purpose of inference, we present that the following multiple linear regression models for mean (μ), standard deviation (σ), skewness (γ), and kurtosis (κ) with appropriate variable transformations show a good fit and model interpretability on the training data:

$$\hat{\mu_t} = 2.846 - 0.0678 \cdot \mathrm{Re} + 0.113 \log(\mathrm{St}) - 1.402 \log(\mathrm{Fr}^*) + 0.00320 \cdot \mathrm{Re} \cdot \log(\mathrm{St}) + 0.00857 \cdot \mathrm{Re} \cdot \log(\mathrm{Fr}^*) \tag{1}$$

$$\hat{\sigma_t} = 1.476 - 0.0140 \cdot \mathrm{Re} + 0.647 \log(\mathrm{St}) - 1.213 \log(\mathrm{Fr}^*) \tag{2}$$

$$\begin{aligned}\hat{\gamma_t} = {} & 6.228 \times 10^2 - 1.534 \times \mathrm{Re} + 1.264 \times \log(\mathrm{St}) - 1.712 \times 10^3 \mathrm{Fr}^* + 1.089 \times 10^3 (\mathrm{Fr}^*)^2 \\ & -7.837 \times 10^{-3} \times \mathrm{Re} \times \log(\mathrm{St}) + 4.432 \times \mathrm{Re} \times \mathrm{Fr}^* - 2.825 \times \mathrm{Re} \times (\mathrm{Fr}^*)^2\end{aligned} \tag{3}$$

$$\begin{aligned}\hat{\kappa_t} = {} & 1.240 \times 10^3 - 3.032 \times \mathrm{Re} + 1.832 \times \log(\mathrm{St}) - 3.407 \times 10^3 \mathrm{Fr}^* + 2.167 \times 10^3 (\mathrm{Fr}^*)^2 \\ & -1.475 \times 10^{-2} \times \mathrm{Re} \times \log(\mathrm{St}) + 8.758 \times \mathrm{Re} \times \mathrm{Fr}^* - 5.580 \times \mathrm{Re} \times (\mathrm{Fr}^*)^2\end{aligned} \tag{4}$$

where $\hat{\mu_t} = \dfrac{\hat{\mu}^{-0.25} - 1}{-0.25}$, $\hat{\sigma_t} = \dfrac{\hat{\sigma}^{-0.25} - 1}{-0.25}$, $\hat{\gamma_t} = \dfrac{\hat{\gamma}^{0.5} - 1}{0.5}$, $\hat{\kappa_t} = \dfrac{\hat{\kappa}^{0.25} - 1}{0.25}$, and $\mathrm{Fr}^* = \dfrac{1}{1 + e^{-Fr}}$.

For the purpose of prediction, we present that a random forest model with number of trees chosen from k-fold cross validation may be more suitable for prediction, given that the parameters Re, St, Fr suggest a complex non-linear relationship with the particle clustering distribution, and the model is robust to noise due to averaging over many trees.

Methodology

Inference

To derive the inference models (1), (2), (3), and (4), we transformed Fr, which had a value of infinity, by applying a sigmoid function so that its domain can be mapped to $[0, 1]$ and the parameter can be used to fit the linear model.

Then, we considered the distribution of μ, σ, γ, and κ through their histograms, which suggested a heavy right-skew. Hence box-cox transformations of the summary statistics were computed to reduce the heavy tails in the distributions, specifically μ_t, σ_t, γ_t, and κ_t as shown earlier.

Next, the correlation between the predictors and the transformed responses was taken into account by looking at the pairwise plots. While Re seem to have a linear relationship with the summary statistics, St seem to have a logarithmic relationship with the response. On the other hand, Fr seem to share a logarithmic curve with μ and σ but reasonably a quadratic relationship with γ, κ. Based on these non-linear relationships, appropriate predictor transformations were made to ensure linearity for fitting linear regression.

In addition to the transformation of predictors, we also consider potential interaction effects between parameters in our inference model, which was systematically chosen by fitting a lasso model to shrink the insignificant predictor terms. The predictors turned out to be significant by the lasso models were then re-fitted into a multiple linear regression model, so that any main effects considered insignificant but its interaction effects considered significant can be included with respect to the hierarchy principle. Model conditions were checked using diagnostic plots to ensure that linearity, normal residual distribution, constant variance, and independence were satisfied.

Given that the pairwise plots already suggest a complex, non-linear relationship between the parameters and summary statistics, using a linear model for inference may be considered appropriate as it provides an intuitive insight into how each predictor, Re, St, Fr and their interactions, affects the particle clustering distribution. As long as model conditions are satisfied through variable transformations, it is reasonable to use a linear model to simplify our understanding of turbulence.

Prediction

Given that prediction is focused on increasing predictive accuracy rather than interpretability, we suggest that a random forest model may be suitable for our problem. This is because random forest uses partition space to model the complex, non-linear relationship between Re, Fr, St and summary statistics of particle clustering distribution, which has been observed in the pairwise plots when building the model for inference. Also, it is robust to outliers and noise, which may exist in our tiny dataset of $n = 89$ observations, given that it is averaged out over many trees.

While there are other non-linear model options such as generalized additive model (GAM), the fact that only three distinct values are given for the Reynolds and Froude number render GAM unsuitable in our context. GAM's smooth function expects a predictor that wiggles across a continuous range, so it will struggle in this case as there are too few unique data points to smooth.

In order to find the model that minimizes the test MSE value of our training data, we conducted a 5-fold cross validation over a list of `ntree` values ranging from 100 to 2000 by an interval of 50, where `ntree` represents the number of decision trees in the random forest. The folds were created with stratified random sampling in terms of the combination of Re, St, and Fr values to prevent an imbalanced data between train and validation split.

Results

Model Conditions

Fig. 1. represents diagnostic plots for Model (3) used to predict skewness. We see that linearity and constant variance are roughly satisfied given that most data points seem to be randomly scattered in the residuals vs. fitted plot, and the red trendline in the scale-location plot is fairly horizontal throughout. Except a little deviation at the tails, the Q-Q residuals plot shows that the residuals have a fairly normal distribution. In the residuals vs. leverage plot, there may be some points with high residuals, but most points seem to be scattered around the zero-residual, nor a curvature or a distinct funnel shape is observed. Assuming independence in data generation, we may expect our inference model (while not shown, other models have similar diagnostic plots as well) to satisfy the model conditions, and it may be reasonable to present our scientific insights based on these models.

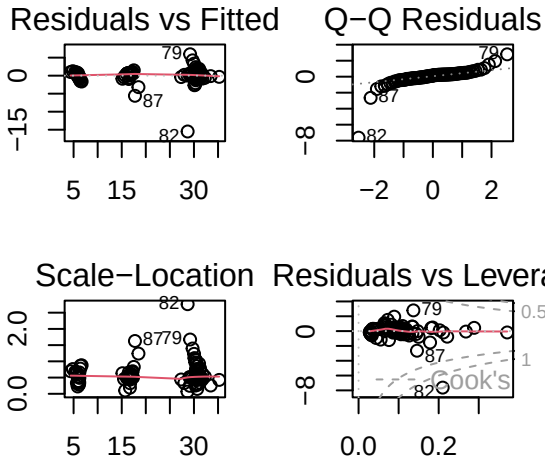

Figure 1: Diagnostic Plots for Model (3), Skewness

Scientific Insights

Model (1) reveals that μ_t, the transformed mean, is strongly influenced by the Froude number given the large negative coefficient of -1.402. This suggests that gravity suppresses clustering, pulling particles down and preventing large aggregates from forming, which decreases the cluster volume on average. The positive coefficient (0.113) of log(St) shows that larger Stokes number increases cluster volume. This indicates that particles with greater inertia is less prone to small-scale turbulent fluctuations and more likely to coalesce into bigger clusters. As for Re, the small negative coefficient of -0.0678 signifies that greater turbulent intensity may break up clusters. The positive coefficients of interaction effects suggest that at higher values of Re, the influence of St on cluster volume is amplified, and the negative effect of gravity is slightly offset.

As for Model (2), it suggests that gravity not only reduces average cluster volume but also limits the spread, as evidenced by the negative coefficient of the main effect (-1.213). The Stokes number has a moderate positive effect with a coefficient of 0.647 on its log transformation, meaning that higher particle inertia increases the spread of cluster volumes. Reynolds number has a slight negative effect. Physically, greater turbulence may homogenize cluster sizes slightly.

Model (3) is concerned with the skewness of particle cluster volume. Re has a small negative effect, indicating that more chaotic flows reduce asymmetry in cluster volume. St has a small positive effect; higher particle inertia increases asymmetry, enabling some clusters to grow much larger than others. Froude number is interesting in this model equation, as the linear term greatly reduces skewness at small values of Fr, but at very high values, the effect may taper or reverse due to the quadratic term. The interaction terms show that some interplays between particle inertia, turbulent intensity, and gravity promote extremely large clusters, while others suppress them.

Model (4) represents the physics of kurtosis and the effect of Re, St, and Fr on it. The negative coefficient of -3.032 for Reynolds number hints at how greater turbulence slightly tends towards a more normal distribution of cluster sizes. The positive effect of 1.832 on log transformation of Stokes number shows that particles with high inertia are more likely to form extreme clusters, producing heavy tails in the distribution. The Froude number is also an interesting case here in the model equation: at low values, it leads to a strong decrease in kurtosis with a negative coefficient of -3407, but at higher values, the quadratic term offsets this with a strong positive coefficient of 2167. The interaction terms suggest that there is a complex nonlinear coupling between kurtosis and gravity, while at higher values of Re, the positive effect of Stokes number is slightly dampened.

Prediction Results

Figure 2 shows predicted summary statistics from the random forest model with the optimal number of decision trees found from 5-fold cross validation, against the Reynolds number. With the increase in the values of Re, we see that the predicted mean of cluster volume decreases, which coincides with our scientific insight that greater turbulent intensity breaks up cluster sizes, contributing to the decrease. So is the case for predicted standard deviation, where only a slight decline is observed, as we have concluded in our inference model that greater turbulence homogenizes cluster sizes slightly. On the other hand, we

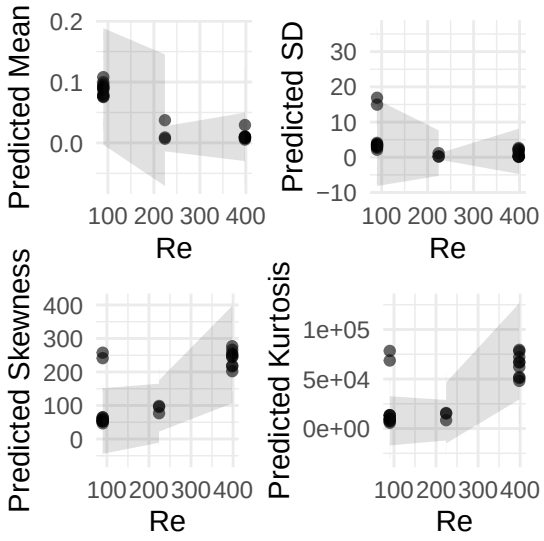

Figure 2: Predicted Summary Statistics vs. Re

see that the change is more drastic for predicted skewness and kurtosis with higher values of Re, and the positive relationship is observed in the original pairwise plots between Re and transformed skewness and kurtosis values.

One thing to note, which also can be improved, is the 95% prediction interval shown in the plots. The interval is wide, which can be attributed to small data size and only three distinct values were observed for the Reynolds and Froude numbers. Hence more data may be desired for the random forest model to learn this complex particle turbulence relationship.

Conclusion

It is easy to associate that the properties of turbulence and particles within the turbulent flow are correlated. Yet the relationship tends to be complex and non-linear that there aren't many pervasive models today that can predict the particle clustering distribution reliably. This study aims to go one step further towards that milestone by presenting two different models for inference and prediction: a multiple linear regression model to simplify our understanding, and a random forest model to increase our predictive accuracy with the given inputs: Reynolds number (turbulent intensity), Froude number (gravitational acceleration), and Stokes number (particle inertia).

The inference model has given us insight into how gravity suppresses clusters, and so does greater turbulent intensity in breaking clusters and slightly homogenizing cluster sizes. The model for skewness and kurtosis have shown that higher inertial particles have the tendency to form extreme clusters, contributing to the increase in asymmetry and heaviness of the distribution tail.

The random forest prediction model has shown predictive results that closely coincide with our scientific insights from inference and the original relationship between the parameters and the response variables.

However, the wide prediction interval from the prediction model may be an area of improvement arising from small data used to train the model, as only three distinct values were observed for Reynolds and Froude numbers. Hence a further study with greater data size and other complex models such as neural networks for prediction may be desired.