Predicting Particle Cluster Volume Distributions in Turbulence

Brian Kim, Joanne Chae

1 Introduction

Turbulence is a fundamental phenomenon in fluid dynamics with broad applications across astrophysics, climatology, and engineering. Understanding how particles cluster within turbulent flows provides key insights into transport, mixing, and aggregation processes. In this case study, we develop a statistical model that predicts the distribution of particle cluster volumes based on three dimensionless parameters: the Reynolds number (Re), gravitational acceleration (Fr), and the particle characteristic (St).

Because each simulation produces a full probability distribution of cluster volumes, direct modeling of these distributions is computationally intensive and analytically complex. To address this, the distributions are summarized by their first four central moments — the mean (μ), standard deviation (σ), skewness (γ), and kurtosis (κ) — which describe the location, spread, asymmetry, and tail behavior of the distribution, respectively. The primary research objective is therefore to construct a predictive model that maps the input parameters (Re, Fr, St) to these four summary statistics.

Our modeling approach uses Generalized Additive Models (GAMs), which combine the interpretability of linear regression with the flexibility of nonlinear smoothing. By modeling smooth functions of logit(Fr) and logit(St), and treating Re as a categorical factor, the GAM framework captures nonlinear effects while remaining interpretable. We evaluate model performance using 10-fold cross-validation and interpret parameter effects using the fitted smooth functions from the GAMs. Our model aims to accurately predict the four statistical moments of particle cluster distributions under new input settings and provide interpretable insights into the physical relationships between flow parameters and particle clustering behavior.

2 Methodology

2.1 Exploratory Data Analysis

We first examined the empirical distributions of the three input parameters (Re, Fr, St) and the four response moments ($\mu, \sigma, \gamma, \kappa$) (Figure 1).

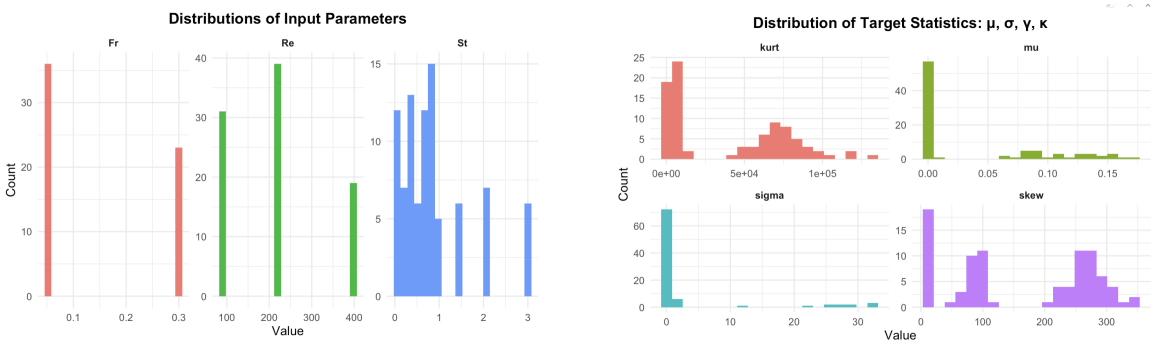
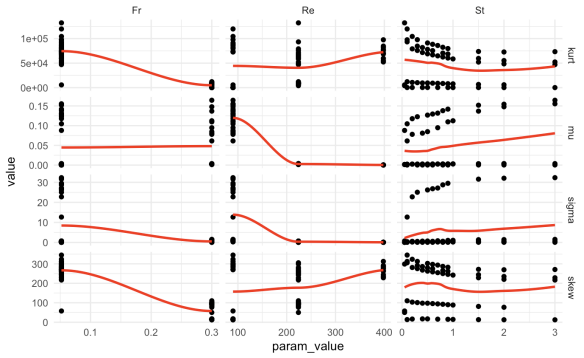


Figure 1: Distributions of input parameters and target statistics.

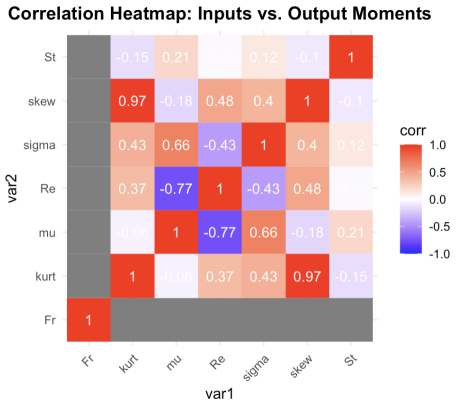
The input parameters vary in both scale and type: Re takes on only three discrete values representing distinct turbulence regimes, Fr assumes two levels corresponding to different gravitational intensities, and St varies continuously between 0 and 3. This mix of categorical and continuous predictors motivates treating Re as a factor and modeling smooth nonlinear effects for Fr amd St.

The response statistics are strongly right-skewed, suggesting the presence of heavy tails in the cluster volume distributions and motivates the use of log-transformations and nonlinear modeling approaches.

To examine dependencies among variables, we computed pairwise correlations between all inputs and response moments and visualized nonlinear relationships (Figure 2).



(a) Relationships between inputs and output moments.



(b) Correlation heatmap of inputs and response moments.

Figure 2: Nonlinear relationships and correlations among input parameters and response moments.

The correlation heatmap shows that Fr has the strongest positive associations with all output moments, especially μ, γ, and κ, indicating gravity's dominant influence on cluster behavior. Re exhibits moderate negative correlations with μ and σ, suggesting higher turbulence reduces cluster size and spread. In contrast, St shows weak correlations, implying limited direct effect on the distribution's shape.

The scatterplots with fitted smooths reveal distinct nonlinear trends: increasing St generally raises all moments, indicating that particle inertia amplifies both cluster size and variability. Re shows decreasing μ and σ with possible upturns at higher turbulence levels, while Fr exhibits mild declines across moments, suggesting a weaker, stabilizing influence of gravity on clustering behavior.

These findings motivated the use of transformations to stabilize variable ranges and the categorical treatment of Re to capture regime-specific effects. Given the nonlinear trends observed, we employed a Generalized Additive Model (GAM) framework with smooth terms for the continuous predictors and a factor term for Re, aligning model structure with the empirical patterns identified in the exploratory analysis

2.2 Data Processing and Feature Transformation

Following the exploratory analysis, we prepared the dataset by computing the four target moments— $\mu, \sigma, \gamma, \kappa$ —from the provided raw moments $(E[X], E[X^2], E[X^3], E[X^4])$. To ensure numerical stability, we replaced non-finite values and small variances with a minimal constant $\varepsilon = 10^{-10}$ before computing higher-order moments.

Given that the input parameters (Re, Fr, St) differ in scale and exhibit strong skewness, we applied transformations to stabilize variance and linearize their relationships with the target variables. Specifically, Fr and St were logit-transformed to map their bounded range $(0, 1)$ onto the real line, while Re was treated as a categorical variable to capture regime-dependent effects that were evident in the exploratory plots. These transformations mitigate scale differences and help stabilize model estimation.

2.3 Model Specification

To model the relationships between input parameters and the four target statistics, we used a *Generalized Additive Model (GAM)* for each response variable:

$$\begin{aligned}
\text{Mean:} \quad & \mu = \beta_0 + f_1(Re) + s_1(Fr) + s_2(St) + \varepsilon_\mu, \\
\text{Standard deviation:} \quad & \sigma = \beta_0 + f_1(Re) + s_1(Fr) + s_2(St) + \varepsilon_\sigma, \\
\text{Skewness:} \quad & \gamma = \beta_0 + f_1(Re) + s_1(Fr) + s_2(St) + \varepsilon_\gamma, \\
\text{Kurtosis:} \quad & \kappa = \beta_0 + f_1(Re) + s_1(Fr) + s_2(St) + \varepsilon_\kappa,
\end{aligned}$$

where $f_1(Re)$ represents categorical effects for discrete turbulence regimes, and $s_1(\cdot)$ and $s_2(\cdot)$ are smooth spline functions capturing nonlinear effects of the pre-transformed Fr and St variables.

For each smooth term $s(Fr)$ and $s(St)$, we specified basis dimensions of $k = 3$ and $k = 5$, respectively. These values balance model flexibility and interpretability given the limited sample size $(n = 89)$. Preliminary trials with larger k did not meaningfully improve cross-validation performance but increased the risk of

overfitting, particularly for higher moments (γ and κ). The smaller basis for Fr reflects its narrow range and weaker influence observed in exploratory analysis, while the larger k for St allows greater flexibility to capture its stronger nonlinear behavior. All models were estimated using restricted maximum likelihood (REML), which optimizes the smoothing penalty and likelihood jointly to ensure appropriate smoothness without manual tuning.

2.4 Model Fitting and Evaluation

Each GAM was trained using the 89 available simulation runs from `data-train.csv`. To assess the benefit of nonlinear smooths, we compared each GAM against a simpler multiple linear regression model using the same transformed predictors: `factor(Re)`, `Fr`, and `St`. The linear models capture only additive linear effects, whereas the GAMs allow flexible nonlinear relationships for `Fr` and `St` through penalized spline smooths.

Model performance was evaluated using 10-fold cross-validation, computing the average root mean squared error (RMSE) across folds as the primary performance metric. This approach balances bias and variance by repeatedly partitioning the data into training and validation subsets.

To quantify predictive uncertainty, we used the built-in standard error estimates from the `mgcv::predict(..., se.fit = TRUE)` function. From these, we constructed approximate 95% confidence intervals for each predicted moment:

$$\hat{y}_{i,\text{lower}} = \hat{y}_i - 1.96\,\text{SE}(\hat{y}_i), \qquad \hat{y}_{i,\text{upper}} = \hat{y}_i + 1.96\,\text{SE}(\hat{y}_i).$$

These intervals provide interpretable bounds on model uncertainty while maintaining computational efficiency, allowing both prediction accuracy and confidence to be communicated when applying the model to new parameter settings.

Model diagnostics for the fitted GAMs (Figure 3) indicate that residuals are approximately centered around zero with no strong heteroscedasticity or systematic bias across fitted values. The QQ-plots for the lower-order moments (μ, σ) show that deviance residuals closely follow the theoretical quantiles, suggesting that the models capture most of the explainable variation. For higher-order moments (γ, κ), mild departures from normality and heavier tails are visible, consistent with the increased variability and skewness observed in the data. Histograms confirm roughly symmetric residual distributions, while the residual–fitted relationships show no discernible trends. Overall, these diagnostics indicate that the GAMs provide an adequate fit to the data and that nonlinear smooths effectively model the systematic structure without overfitting. Each GAM assumes smoothness in Fr and St and independence of residuals, which were verified via diagnostic plots.

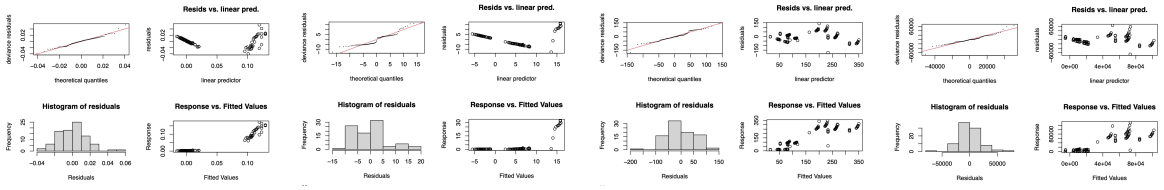


Figure 3: Residual diagnostics for μ, σ, γ, κ respectively. Each panel summarizes QQ-plots, residual distributions, and fitted relationships.

2.5 Model Justification

The choice of Generalized Additive Models (GAMs) is well-suited to this problem for three key reasons. First, the exploratory analysis revealed strong nonlinearities in the effects of Re, Fr, and St on all four response statistics, particularly for higher-order moments. GAMs naturally capture such nonlinear relationships through smooth functions without imposing rigid parametric assumptions, allowing flexible modeling of gradual transitions in turbulence behavior.

Second, treating Re as a categorical variable enables the model to represent distinct turbulence regimes that may produce qualitatively different clustering patterns. Third, GAMs offer high interpretability—partial dependence plots for the smooth terms of Fr and St directly visualize how each parameter influences distributional characteristics, aligning statistical inference with physical intuition about turbulence and particle inertia.

Overall, this modeling framework balances predictive performance with interpretability, enabling accurate forecasts of cluster-distribution statistics and meaningful scientific inference about the underlying flow dynamics.

3 Results

3.1 Predictive Performance

As shown in Figure 4, the Generalized Additive Models (GAMs) consistently outperform the linear baselines across all four target statistics. For the lower-order moments (μ and σ), both models perform similarly, consistent with the near-linear relationships observed in the exploratory analysis. However, for the higher-order moments (γ and κ), the GAMs achieve substantially lower prediction error and higher explanatory power. Specifically, cross-validated RMSE decreases from 87.7 to 55.1 for skewness and from 29,858 to 19,804 for kurtosis, while R^2 increases from 0.38 to 0.73 and from 0.35 to 0.71, respectively.

These improvements confirm that nonlinear smooths capture complex dependencies between turbulence parameters and the shape of particle cluster volume distributions, particularly in modeling asymmetry and tail behavior. Overall, the results highlight the importance of including nonlinear terms for accurate and physically interpretable predictions of turbulence-induced clustering phenomena.

Target <chr>	RMSE_Linear <dbl>	RMSE_GAM <dbl>	MAE_Linear <dbl>	MAE_GAM <dbl>	R2_Linear <dbl>	R2_GAM <dbl>
mu	0.017	0.017	0.013	0.013	0.843	0.841
sigma	7.465	6.732	5.610	5.826	0.548	0.589
skew	87.688	55.078	75.579	44.760	0.377	0.729
kurt	29858.128	19803.580	25328.453	15519.667	0.348	0.710

Figure 4: Comparison of 10-fold cross-validated RMSE, MAE, and R^2 across linear and GAM

3.2 Prediction Uncertainty

The 95% confidence intervals derived from the GAM standard errors quantify how prediction reliability varies across the four target statistics. The intervals are narrow for the lower-order moments (μ, σ), indicating stable, well-constrained predictions. In contrast, the intervals widen substantially for the higher-order moments (γ, κ), where the model captures greater nonlinear variability and sensitivity to the input parameters.

These patterns are consistent with the physical intuition that tail-related features of the cluster-volume distribution are more difficult to estimate under changing turbulence regimes. Overall, the confidence intervals provide a transparent measure of model reliability, allowing uncertainty in predicted cluster-distribution characteristics to be communicated clearly when applying the model to new flow conditions.

Table 1: Summary of predicted moments and average 95% confidence interval widths.

	Mean ($\hat{y}$)	SD of $\hat{y}$	Mean CI Width	CI Range
μ (Mean)	0.05	0.04	0.018	[-0.02, 0.12]
σ (Std. dev.)	3.44	5.8	7.2	[-8.9, 12.3]
γ (Skewness)	150.0	85.1	59.1	[12.5, 381.8]
κ (Kurtosis)	37128	30400	21514	[-16437, 91778]

4 Conclusion

This study developed and evaluated a predictive framework for modeling the distribution of particle cluster volumes in turbulent flow, using Re, Fr, St as explanatory variables. Through exploratory analysis, we identified strong nonlinear relationships between turbulence parameters and the higher-order moments of the particle cluster distribution, motivating the use of a Generalized Additive Model (GAM) framework.

Cross-validated results demonstrated that the GAM models outperformed linear baselines, particularly for the higher-order statistics—skewness (γ) and kurtosis (κ)—where nonlinear smooth terms captured complex turbulence effects that linear models could not. For lower-order moments, mean (μ) and standard deviation (σ), both models achieved similar accuracy, consistent with their near-linear dependence on Re and Fr.

Predictive uncertainty was quantified using 95% confidence intervals derived from model standard errors, providing interpretable bounds on prediction reliability across new parameter settings. Uncertainty was lowest for μ and σ and highest for γ and κ, reflecting greater variability in the tails of the cluster-volume distributions. Overall, our findings highlight the dominant influence of particle inertia (St) in shaping asymmetry and tail behavior in turbulence-induced clustering. The GAM framework balances predictive performance, interpretability, and physical plausibility, offering a flexible foundation for future extensions to more complex turbulence regimes or additional physical parameters.