Predicting Particle Cluster Volume Distributions in Turbulence

STA 325: Data Expedition Case Study

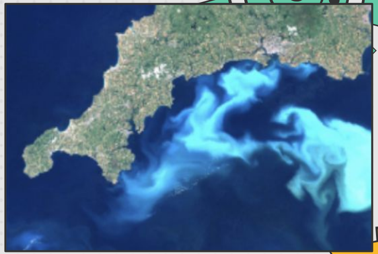
Joanne Chae | Brian Kim

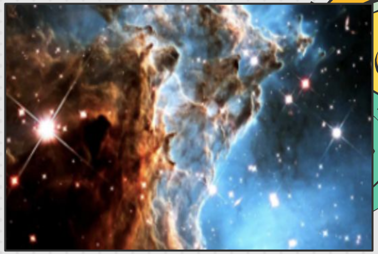
01

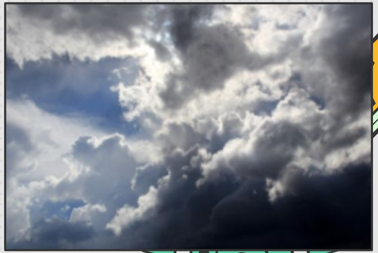
Introduction

Motivation

- Turbulence drives how small particles behave in natural & engineered flows
- Particle clustering influences... (etc.,)
  - Cloud formation
  - Pollutant dispersion
  - Planetary accretion
- Modeling these clusters helps predict how particle concentration changes across flow regimes







Introduction

CHALLENGES

- Each simulation yields a distribution of cluster volumes—complex to model directly
- We summarize with 4 **central statistical moments**:
  - Mean (μ) → central tendency
  - Standard deviation (σ) → spread
  - Skewness (γ) → asymmetry
  - Kurtosis (κ) → tail heaviness

GOAL

- Predict (μ, σ, γ, κ) from 3 turbulence parameters:
  - **Re** (Reynolds #) → turbulence intensity
  - **Fr** (Froude #) → gravitational acceleration
  - **St** (Stokes #) → particle characteristic

Modeling Approach Overview

Approach:

Use **Generalized Additive Models (GAMs)** (to predict each moment)

Combines:

- ***Interpretability*** of linear regression
- ***Flexibility*** of nonlinear smoothing

Which...

Allows **direct interpretation** of how each turbulence parameter **affects** the cluster-volume distribution

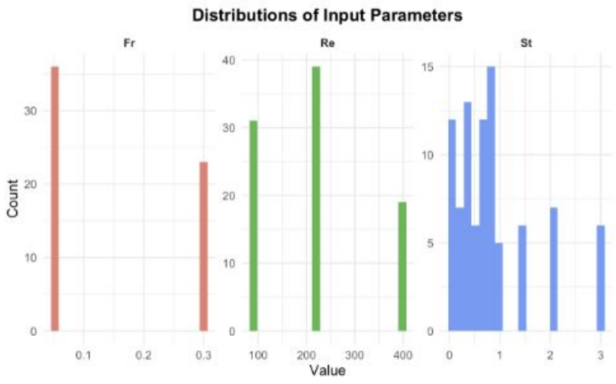
02

Methodology

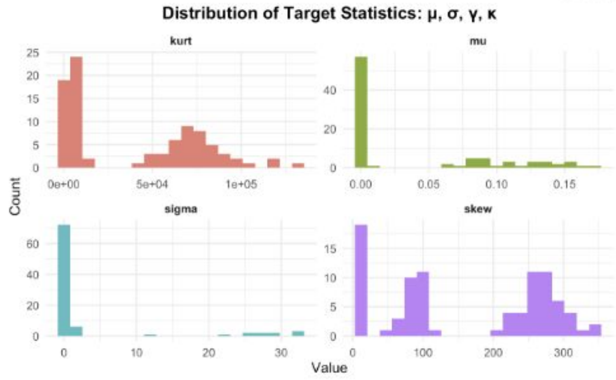
Exploratory Data Analysis (pt.1)

Findings:

- Re takes on only three discrete values
- Fr assumes two levels
- St varies continuously between 0 and 3
- Mix of categorical and continuous predictors -> Re as factor and model smooth nonlinear effects for Fr and St
- Response statistics are heavily right skewed



(a) Input distributions

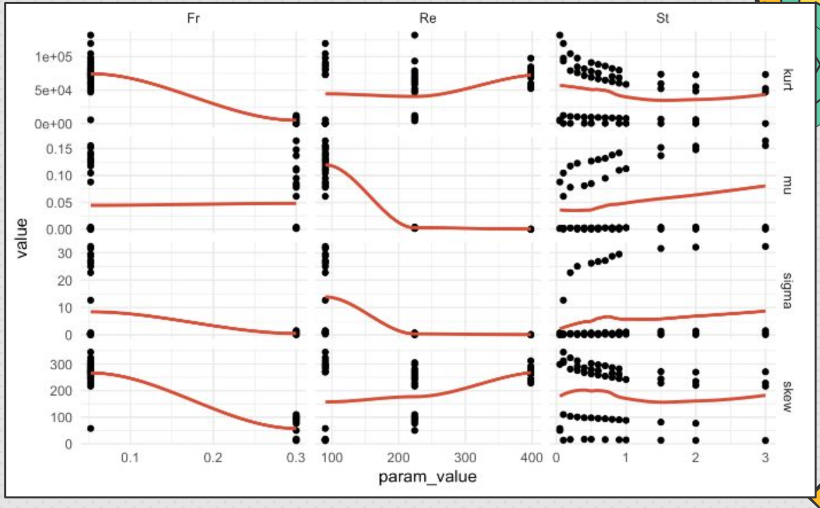


(b) Target moment distributions

Exploratory Data Analysis (pt.2)

Findings:

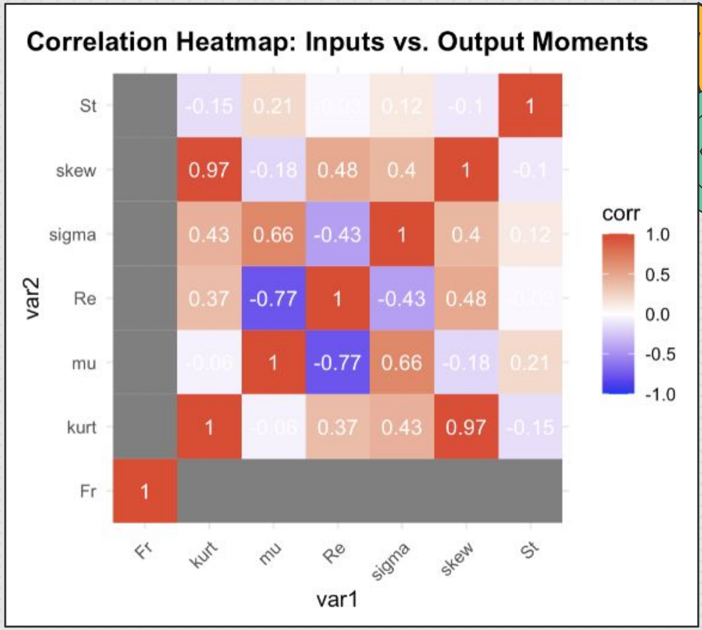
- Increasing St generally raises all moments
- Re shows decreasing mu and sigma
- Fr exhibits mild decline across all moments



Exploratory Data Analysis (pt.3)

Pairwise Correlations:

- Fr has strongest positive correlation with output moments
- Re exhibits negative correlation with mu and sigma
- St shows weak correlations



Data Processing & Feature Transformation

Step 1

Compute 4 moments (μ, σ, γ, κ) from $E[X]$, $E[X^2]$, $E[X^3]$, $E[X^4]$

Step 2

Replace non-finite values with $\varepsilon = 1e\text{-}10$

Step 3

Apply transformations:

- **Fr, St:** logit-transform → (0, 1) range → $\mathbb{R}$
- **Re:** treated as categorical to capture discrete turbulence regimes

Purpose: Mitigate scale differences & stabilize model estimation

Model Specification

Each response variable fitted with separate GAM:

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

Interpretation:

- f1(Re): categorical regime effects
- s1, s2: smooth nonlinear functions

Choice of Smoothing Parameters

Parameter Selection:

- Basis dimensions k = 3 for Fr, k = 5 for St
- Smaller k → interpretability; larger k → flexibility
- Estimated using **REML** for automatic smoothness control
- Balanced to avoid overfitting given small sample size (n = 89)

Model Fitting & Evaluation

Key Steps:

- Trained 4 GAMs on 89 simulations from data-train.csv
- Compared to linear baselines (factor(Re), (Fr), (St))
- Linear models capture only additive trends; GAMs add smooth nonlinear effects
- Evaluated with **10-fold CV** using **RMSE** across folds

Findings:

- GAMs → lower RMSE for all moments, especially γ & κ
- Cross-validation mitigates split bias

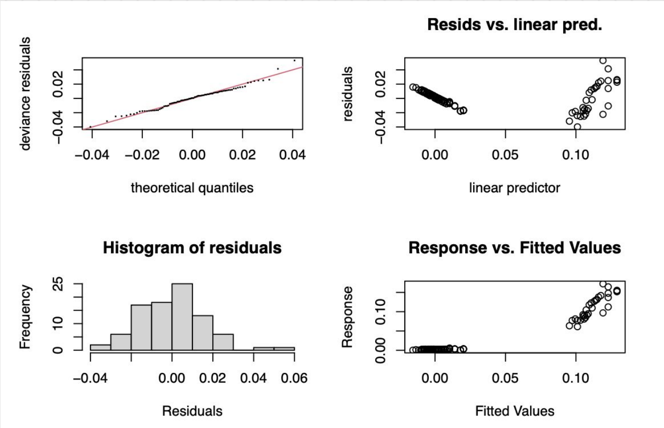
Model Fitting and Evaluation

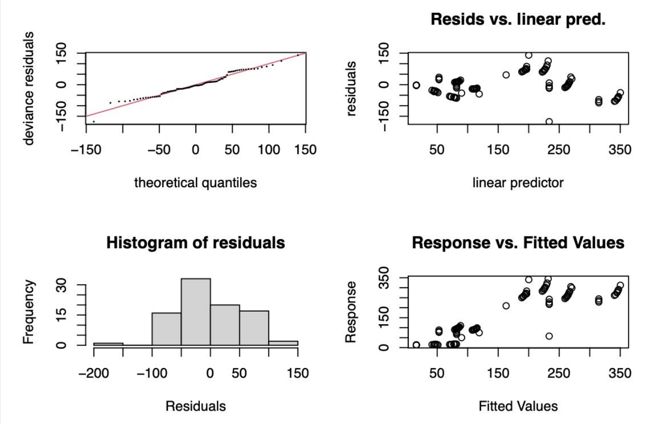
Method:

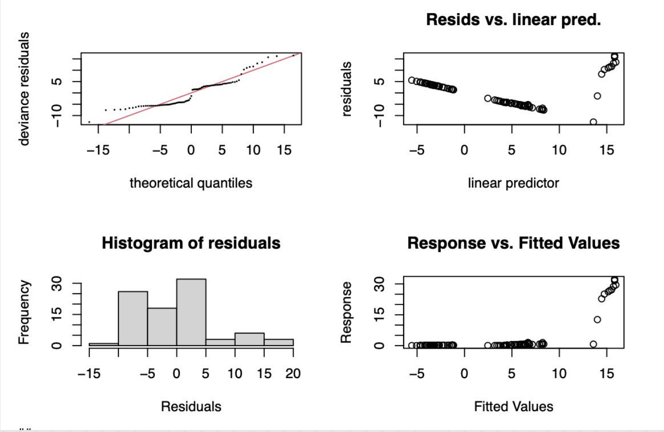
$$\hat{y}_{i,\text{lower}} = \hat{y}_i - 1.96\,\mathrm{SE}(\hat{y}_i), \qquad \hat{y}_{i,\text{upper}} = \hat{y}_i + 1.96\,\mathrm{SE}(\hat{y}_i)$$

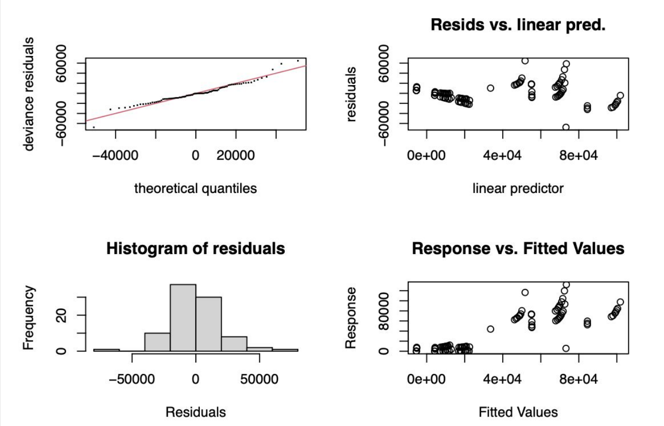
- Built 95 % confidence intervals using `predict(..., se.fit = TRUE)`
- Provides interpretable uncertainty bounds for new parameter settings

Model Diagnostics









Model Justification

Why GAMs work here:

- Strong nonlinearities among Re, Fr, St → GAMs capture them naturally
- Re as factor → distinct turbulence regimes
- Smooth terms visualize physical effects of gravity and particle inertia
- Framework balances accuracy + interpretability

03

Results

Predictive Performance

Model Comparison (10-Fold CV):

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

Highlights:

- Major RMSE drop for γ and $\kappa \rightarrow$ captures complex nonlinear patterns
- Similar performance for μ and $\sigma \rightarrow$ nearly linear behavior
- Supports EDA findings of nonlinear Re & St effects

Prediction Uncertainty

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]

- 95% confidence intervals are narrow for lower order moments of mu and sigma
- Intervals widen substantially for higher order moments
- Consistent with physical intuition that tail related features of cluster volume distributions are harder to estimate under changing turbulence regimes

04

Conclusion

Key Takeaways:

- Strong nonlinear relationships between turbulence parameters and higher order moments of particle cluster distribution -> use of GAM
- Cross validated results show GAM models outperformed linear baselines
- 95% confidence intervals showed uncertainty was lowest for mu and sigma and highest for gamma and kappa
- GAM framework balances predictive performance, interpretability, and physical plausibility