Inference & Prediction of Fluid Turbulence Case Study

Khushmeet Chandi, Athena Ru, Jason Ren, Zaid Muqsit

10/28/2023

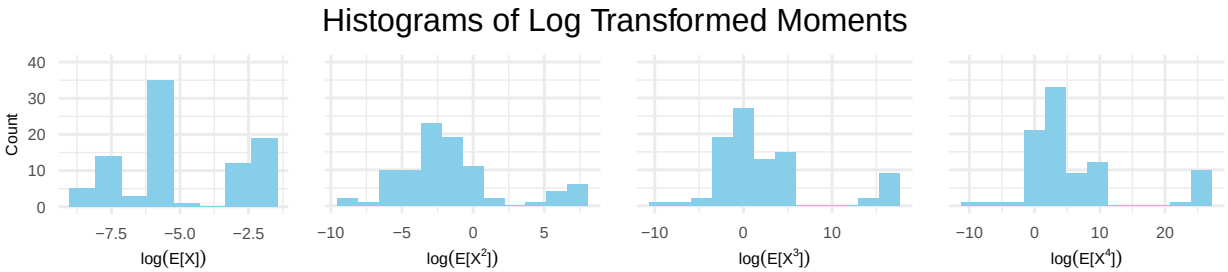
Introduction

The primary focus revolves around unraveling the intricate dynamics of the distributions particle clusters in turbulence. The dataset, comprising 89 simulations, is characterized by three predictors: Reynolds number (Re), gravitational acceleration (Fr), and particle characteristic (St). The response variables are the four raw moments: the first raw moment ($E[X]$) is mean or average volume; the second raw moment ($E[X^2]$) is the variance; the third raw moment ($E[X^3]$) is the skewness; and the fourth raw moment ($E[X^4]$) is kurtosis which gives insights into the distribution's tail behavior.

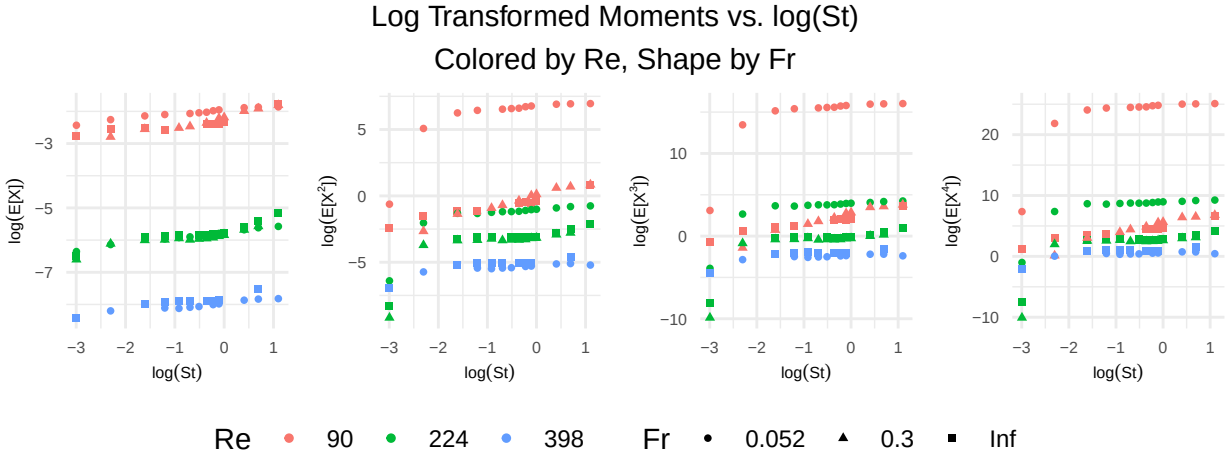
In terms of physical meaning, the greater the Reynolds number, the higher the particle tracking resolution. If gravitational acceleration increases, particles may fall faster and cover a larger spatial area within the given domain. Notably, cumulonimbus (high-level clouds) correspond to `Fr = 0.3` or `Inf` and cumulus (low-level clouds) correspond to `Fr = 0.052`. Lastly, the greater the particle property, the more difficult it is to solve for the particle's equation.

The research objectives are inference and prediction. Through statistical methods such as cross validation and regression, our models achieve these objectives by interpreting the influence of individual parameters (Re, Fr, St) on the probability distribution of particle cluster volumes and by predicting the particle cluster volume distribution for novel parameter settings (Re, Fr, St). Ultimately, we gain a deeper comprehension of the underlying mechanisms governing particle cluster formation in diverse environmental conditions.

Cleaning & EDA:



The distributions of the moments are all very right skewed. Turbulence data often tends to exhibit long tails and extreme values due to the chaotic nature of the phenomenon. Log transforming them yields a slightly more normal distribution - a necessary condition for our regression. This stabilizes variance and makes the data more conducive to statistical analysis and interpretation. Similarly, a log transformation was applied to `St` to make it more normal. A Box-cox transformation yielded similar results, but log transformation was chosen for easier interpretability. We also double checked for multicollinearity with a heatmap. The correlations between variables were relatively low. As such, we can go ahead and use all of these variables in our models.



In the plots above, we see greater mixing between `St`, `Fr`, and `Re` as we go from the first moment to the fourth moment. **The plot of the log(First Moment)** shows clear distinctions amongst Re values, hinting at a particular relationship.

Additionally, we can see a trend upwards as log(St) increases. There doesn't seem to be significance towards the Fr values from the plot, but there is a possible interaction between Fr and Re, as for certain values of Re, we see slightly different slopes when altering the Fr values (specifically for Infinity, the square, which increases more dramatically towards the seconds half of the graph). **The plot of log(Second Moment)** shows that the red circles are separated from the cluster. Thus, an indicator variable for when `Re` ≤ 100 and `Fr`=0.052 will be created. In general, as log(St) increases, so does $\log(E[X^2])$. A similar trend was found in **the log(Third Moment) plot**, where most data points followed a general pattern except for when `Re`=90 and `Fr`=0.052. As such, we consider interaction effects in our final model when predicting `log_R_moment_3`. **The fourth plot** shows the greatest overlap in points. The relationship between the log-transformed fourth moment and the predictors might not be entirely linear. There could be a threshold effect, meaning that the relationship changes differently for different ranges of the parameter values.

Methodology and Modeling:

For each moment, we tried three different methods, specifically linear regression, ridge regression, and lasso regression. Because of lack of certainty on which predictors are significant out of the three (transformed) predictors and the interaction, we used lasso to do variable selection.

For the second, third, and fourth models, `Fr=0.052` was chosen as the baseline level for `Fr` because it contains the most data points (36) compared to the other two levels, allowing a coefficient estimate with a smaller variance.

To compare models, we used cross validation for linear regression, and then nested cross validation for ridge regression and lasso regression. Nested cross validation nests the hyperparameter optimization within the model selection procedure, providing a more accurate and robust evaluation of performance for models that require hyperparameter tuning like lasso and ridge regression. The lack of and inclusion of interaction effects was also considered based on the moment. Average RMSE across all folds was used as the performance metric.

In fear of overfitting with such a small dataset, we chose to use the 1se lambda, the largest value of lambda such that the average RMSE is within 1 standard error of the minimum average cross validation RMSE, instead of the lambda that minimizes RMSE. In this way, we get a parsimonious model that performs well.

First Moment Model

For the first moment, we treat Reynolds number (Re) and Froude's number (Fr) as continuous variables, even though they only have three possible values for extrapolation (even if it is with high uncertainty) for the mean of particle cluster volumes. We use a logarithmic transformation of Fr to normalize and properly scale the variables because we have two values below one, and the other is Infinity.

From LASSO, we see that all the coefficients are significant, except log(Fr). However, because of the hierarchy principle, we decided to still include log(Fr) in the model for interpretability. We then ran a linear model after variable selection. While the linear model's coefficients are significant, we know it cannot do variable selection, so we will still assume log(Fr) is the only one that is not significant, and it is only present for interpretability. Otherwise, all other p values are under 0.05.

$$log(RMoment1) = 15.636633 + 0.196321log(St) - 3.943795log(Re) - 0.131090log(Fr) + 0.0251log(Fr) * log(Re)$$

For every 1% increase in St, the first moment is expected to increase by 0.196% on average while holding all other variables constant. For every 1% increase in Re, the first moment is expected to change by $-3.94\% + (0.0251 * log(Fr))\%$ while holding all else constant. This term is more complicated because of the interaction term. However, all it means is that, if everything else is constant, then an increase in Re causes the first moment to fall, but that fall can be lessened by (or completely negated by) the value of Fr. This also makes sense because Re represents the turbulence around the system. As there is more turbulence in the system, fewer clusters will stick together, and thus cluster volumes will be lower. The gravity term (Fr) plays a role here because higher gravity might negate some of the turbulence's effect and hold clusters together, depending on how high the value is. For every 1% increase in Fr, the first moment is expected to change by $-0.131090\% + 0.251 * log(Re)\%$ if all else is held constant. Here we see that, when everything else is constant an increase in Fr leads to a lower first moment, and this can be counteracted by the Reynolds number. The interpretation of the interaction term would be similar as above.

We then looked into the assumptions. The residual plot for the most part seemed properly distributed across, with just slightly larger variances towards the latter values. The points on the Q-Q Residuals plot follow a straight line, indicating that the residuals are roughly normally distributed. We also don't see extreme outliers or high leverage points.

Second Moment Model

$$log(M2) = 2.5540 + 0.9078 * log(St) - 1.3585 * Fr_{0.3} - 0.9475 * Fr_{Inf} - 0.0176Re + 5.5442 * RedCircle$$

All the predictors in this model are statistically significant, as seen by their p-value which is less than 0.05.

The coefficient for `log_St` is 0.9078, and this means that for any 10% increase in `St`, we expect a 9% increase ($1.10^{0.9078} = 1.09$) in `moment_2`, holding all else constant. This makes sense because a higher Stokes number means larger or denser particles which tend to contribute to increased turbulence variance. Both coefficients for `Fr=0.3,Inf` are negative, so we expect a 74.3% ($e^{-1.3585} = 0.257$) and 61.2% ($e^{-0.9475} = 0.388$) decrease in the geometric mean of `moment_2` when we go from `Fr=0.3` or `Fr=Inf` to `Fr=0.052`, respectively, holding all else constant. This implies when we go from cumulonimbus clouds (strong gravitational effects) to cumulus clouds (weaker gravitational effects), we see less turbulent fluid flow, resulting in a reduction in the variability. The coefficient for `Re` is -0.0176, and this means for a one-unit increase in Reynolds number, we expect to see about 1.7% decrease ($e^{-0.0176} = 0.983$) in the geometric mean of `moment_2`, holding all else constant. This is surprising since a higher Reynolds number typically indicates more turbulent flow, but the model here says that the variance decreases correspondingly. This could be due to the limited size of the data.

The residual and scale-location plot shows a slight tapering as the fitted values increase, but for the most part, the errors are mostly homoscedastic. Similarly, most of the Q-Q Residuals plot follow a straight line, indicating that the residuals are roughly normally distributed. The residuals vs leverage plot does show that there are a few high leverage and high residual points due to the very right skewed nature of the original data, but again for the most part, the points are clustered around the center thanks to the log transformation. The model fit for moment 2 is not perfect, but given the low RMSE and high R^2, it works decently and should be used with caution especially if the new data contains extreme or unusual values.

Third Moment

Optimal Lambda: 0.021

$$\begin{aligned} log(M3) = {} & 15.340 + 1.3831 * log(St) - 12.687 * Fr_{0.3} - 12.611 * Fr_{Inf} \\ & - 11.233 * Re_{224} - 17.266 * Re_{398} \\ & + 8.5632 * Fr_{0.3} * Re_{224} + 8.5185 * Fr_{Inf} * Re_{224} + 13.4058 * Fr_{Inf} * Re_{398} \end{aligned}$$

We performed cross validation on linear regression on its own, and then linear regression after using lasso regression to perform feature selection. With both methods, we considered the inclusion and exclusion of interaction effects.

For lasso regression, we chose to use the 1se lambda to get a parsimonious model that performs well. Lasso regression seemed to remove only `Fr_catInf` and `Fr_cat0.3:Re_cat398`. Because this is difficult to interpret, we decided to keep all the variables and levels of interaction effects.

Linear regression with interaction effects performed the best. This does make sense, as predictors on their own may not predict skewness well, but with interaction effects, we get a better idea of the real effects of our variables.

Our coefficient estimates seem to be statistically significant under the alpha level of 0.05. This excludes one of the factor levels `Fr_cat0.3:Re_cat398`, which didn't have a coefficient due to singularity, meaning the column can be linearly predicted from the other columns. Holding all else constant, for every 10% increase in `St`, we expect a 0.1318 (`1.3831*log(1.1)`) increase in `log_R_moment_3`, and hence a 1.141 ($e^{0.1318} = 1.09$) times higher `R_moment_3` value, on average. This suggests that a higher `St` tends to increase the third moment, or skewness of the velocity distribution. Considering interaction variables, it is difficult to interpret the effects of Fr and Re and as such won't be explicitly written out. For example, an observation with `Fr`=0.3 has $-12.687 + 8.563 * Re_{224}$ difference in `log_R_moment_3` compared to the baseline observation's (`log(St)`=0, `Fr`=0.052, `Re`=90) `log_R_moment_3` value on average. Just considering `Fr` being either 0.3 or Inf, it has a large negative effect on `R_moment_3` compared to the baseline. But when `Re`=398 and `Fr`=Inf, `Fr` and the interaction effect have a positive effect on skewness compared to the baseline, which makes sense as higher values of `Fr` and `Re` tend to contribute to turbulence. `Re` is similar in that when it increases in value, it is predicted to decrease `R_moment_3` compared to the baseline. However, with `Fr`, the decrease in `R_moment_3` can potentially be much less.

There are some outliers with particularly high residuals. Most Q-Q Residuals follow a straight line, so our residuals are approximately normally distributed. However, there may be some trend in the residuals, indicating heteroscedasticity, which can make our predictions biased. This could be due to many reasons, such as the size of the dataset, or the fact that we factorized `Re` and `Fr`, treating the levels equally.

Fourth Moment

$$\begin{aligned} & log(M4) = 94.2320 - 15.6260 * log(Re) - 76.7147 * Fr_{0.3} - 78.2808 * Fr_{Inf} + 1.7702 * log(St) \\ & + 12.9348 * log(Re) * Fr_{0.3} + 13.2817 * log(Re) * Fr_{Inf} \end{aligned}$$

94.230 is the unconditional expected mean of log of the fourth moment. The exponentiated value exp(94.230), 8.3863e+40, is the geometric mean of the fourth moment. For every 1% increase in Reynolds Number (Re), the fourth moment is

predicted to increase by 15.6260%, holding all else constant. For every 1% increase in Particle Characteristic (St), the fourth moment is predicted to increase by 1.7702%, holding all else constant. When the Gravitational Acceleration (Fr) is 0.3, the log-transformed fourth moment is predicted to be 76.7147 less than when Fr is 0.052, holding all else constant. When the Gravitational Acceleration (Fr) is Inf, the log-transformed fourth moment is predicted to be 78.2808 less than when Fr is 0.052, holding all else constant. This makes sense since higher gravitational acceleration may induce more pronounced fluid movements and fluctuations, leading to lower log-transformed fourth moments. For every 1% increase in Reynolds Number (Re) when Fr is 0.3, the fourth moment is predicted to increase by 12.938%, holding all else constant. For every 1% increase in Reynolds Number (Re) when Fr is Inf, the fourth moment is predicted to increase by 13.2817%, holding all else constant. These interaction effects indicate that the combination of higher Reynolds Numbers with varying levels of gravitational forces contributes to a more pronounced increase in the fourth moment. This suggests that the interaction between these parameters amplifies the turbulence effects, resulting in the observed increase in the fourth moment.

The residual plot shows a cluster of points towards the left two thirds of the plot, but there is not a jarring pattern. The Q-Q plot has initial curvature followed by a straight line, indicating that there is possible non-linearity or some skew as observed in the EDA.

Results:

Table 1: Moment 1-4 MLR Models

Metric	Moment_1_Value	Moment_2_Value	Moment_3_Value	Moment_4_Value
5-Fold RMSE	1.0287	1.0802	1.5083	2.2082
Adj. R2	0.9967	0.9213	0.9228	0.9229
F-Statistic	6627.0776	207.0287	132.4843	176.6640
Numerator DF	4.0000	5.0000	8.0000	6.0000
Denominator DF	84.0000	83.0000	80.0000	82.0000
F-Statistic p-value	0.0000	0.0000	0.0000	0.0000

The 5-Fold RMSE across all 4 models is very low. The adjusted R-squared of the models are all at least 92%, meaning that the majority of the variability in the data is accounted for by the model. The F-Statistic of each model has a p-value less than 0.05 so there is a relationship between each log transformed moment and its corresponding predictors. However, we have relatively high uncertainty in our models due to lack of data points, and since some of the predictors were turned into factors in our models, we can't easily extrapolate to values that are not in these levels. When the next batch of data is collected, they should be processed into these levels of the corresponding factor of the appropriate model.

Conclusion:

We investigated characteristics of turbulence data, applying log transformations to stabalize the variance and normalize the data for regression analysis. We identified significant relationships between the moments and the three parameters. The log transformation of moments improved the normality of the data, making it more suitable for statistical analysis. Distinct trends and relationships were observed between the log-transformed moments and Re, St, and Fr, suggesting complex interactions between these variables. Multicollinearity between the variables was low, allowing for the inclusion of all relevant parameters in the regression models. We used cross-validation and regularization methods to ensure the reliability of the models and mitigate overfitting concerns.

All moments demonstrated varying degrees of sensitivity to the parameters, yet each moment demonstrated nonlinearity in their relationships with the parameters based on the data provided. Among the moments, we also observed similarities in terms of the significance of interaction effects, emphasizing the need to consider the combined influences of the parameters. There were clear differences in terms of the patterns observed within the data for the moments with some moments exhibiting clear trends in relation to what we know physically, while others exhibited deviations from general patterns.

Overall, we observed the significance of interaction effects and nonlinear relationships in understanding the behavior of particle clustering in turbulence.

Table 2: Lasso Results on Left, and Linear Model on Right

	moment_one_lasso.coef
(Intercept)	15.2410257
log(Re)	-3.8684912
log(Fr)	0.0000000
log(St)	0.1802590
log(Re):log(Fr)	0.0000013

term	estimate	std.error	statistic	p.value
(Intercept)	15.6366325	0.1325191	117.995332	0e+00
log(Re)	-3.9437950	0.0253124	-155.805167	0e+00
log(St)	0.1963208	0.0127422	15.407095	0e+00
log(Fr)	-0.1310900	0.0229010	-5.724209	2e-07
log(Re):log(Fr)	0.0250996	0.0043136	5.818688	1e-07

Appendix

```
save_moment_1_model
```

```
save_moment_2_model
```

term	estimate	std.error	statistic	p.value
(Intercept)	2.5540	0.4272	5.9781	0.0000
log_St	0.9078	0.1031	8.8045	0.0000
Fr0.3	-1.3585	0.3444	-3.9449	0.0002
FrInf	-0.9475	0.2954	-3.2073	0.0019
Re	-0.0176	0.0012	-14.1915	0.0000
Red_Circle_Indicator1	5.5442	0.4484	12.3640	0.0000

```
save_moment_3_model
```

term	estimate	std.error	statistic	p.value
(Intercept)	15.3400	0.4669	32.8552	0
log_St	1.3831	0.1579	8.7616	0
Fr_cat0.3	-12.6866	0.6769	-18.7413	0
Fr_catInf	-12.6106	0.6965	-18.1057	0
Re_cat224	-11.2330	0.6207	-18.0968	0
Re_cat398	-17.2660	0.6770	-25.5052	0
Fr_cat0.3:Re_cat224	8.5632	0.9109	9.4006	0
Fr_catInf:Re_cat224	8.5185	0.9340	9.1202	0
Fr_cat0.3:Re_cat398	NA	NA	NA	NA
Fr_catInf:Re_cat398	13.4058	1.0055	13.3321	0

```
save_moment_4_model
```

term	estimate	std.error	statistic	p.value
(Intercept)	94.2320	3.2241	29.2278	0
log_Re	-15.6260	0.6066	-25.7606	0
Fr0.3	-76.7147	5.9568	-12.8784	0
FrInf	-78.2808	4.8201	-16.2405	0
log_St	1.7702	0.2134	8.2968	0
log_Re:Fr0.3	12.9348	1.1661	11.0923	0
log_Re:FrInf	13.2817	0.9055	14.6672	0