# Predicting Turbulence Simulations

Enzo Moraes Mescall, Martin Olarte, Charlotte Coudert

2022-11-05

## Introduction

Turbulence has been described as "the last great unsolved problem in classical physics" by renowned physicist Richard Feynman. It has not only been crucial in mathematics but has important applications in astrophysics, climatology, and engineering.

In this project, we explored the distribution and clustering of particles in an idealized turbulence (homogeneous isotropic turbulence or HIT) as affected by three independent predictors representing the properties of turbulent flow. Reynold number (Re), Froude number (Fr), and Stokes number (St) quantify the fluid turbulence, gravitational acceleration, and particle characteristics, respectively. We used data from Direct Numerical Simulations (DNS) of Re, Fr, and St and their associated clusters produced, quantifying these distributions by their first 4 moments.

The purpose of this paper was twofold: scientific inference and predictive modeling. Our final models described below will help scientists understand how each parameter influences the probability distribution function (PDF) of particle cluster volumes while also being able to predict the PDF for a new parameter setting of (Re, Fr, St), in terms of its four raw moments.

For moment 1, we were able to find a generalized linear model incorporating the interaction of Fr and St values as factors and log(Re). This model balances the trade-off of complexity and interpretability, as it gives us a moderate predictive accuracy while being interpretable in the context of turbulence with respect to the available parameters.

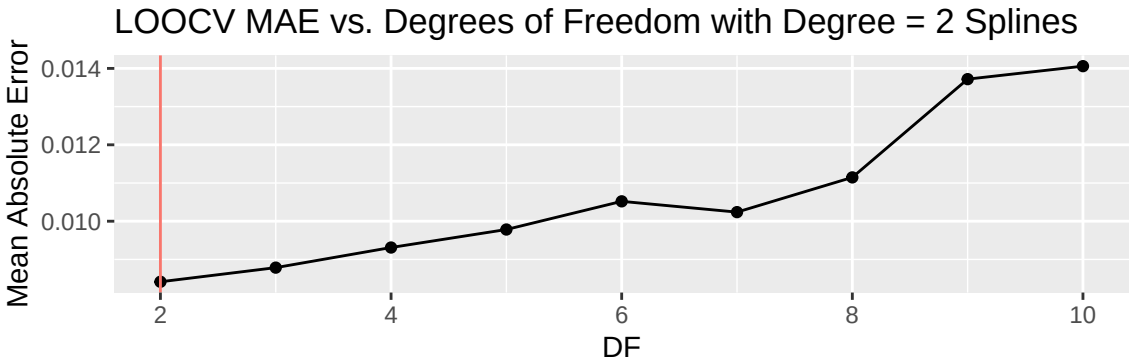
## Methodology

The first noticeable thing about the data is that despite both Fr and Re being numerical predictors, they only have three levels each. Therefore, we decided to treat them as discrete predictors instead of interpolating or extrapolating values outside the given set for Re ($\{\mathbf{90}, \mathbf{224}, \mathbf{398}\}$) and Fr ($\{\mathbf{0.052}, \mathbf{0.3}, \infty\}$). This transformation was appropriate in the context of the problem since the testing data would also have values within the specified sets only. The caveat is that this analysis should not be directly used to understand other real-world problems like oceanic and atmospheric turbulent flow which have Re within the order of magnitude of $\mathbf{10^7}$.

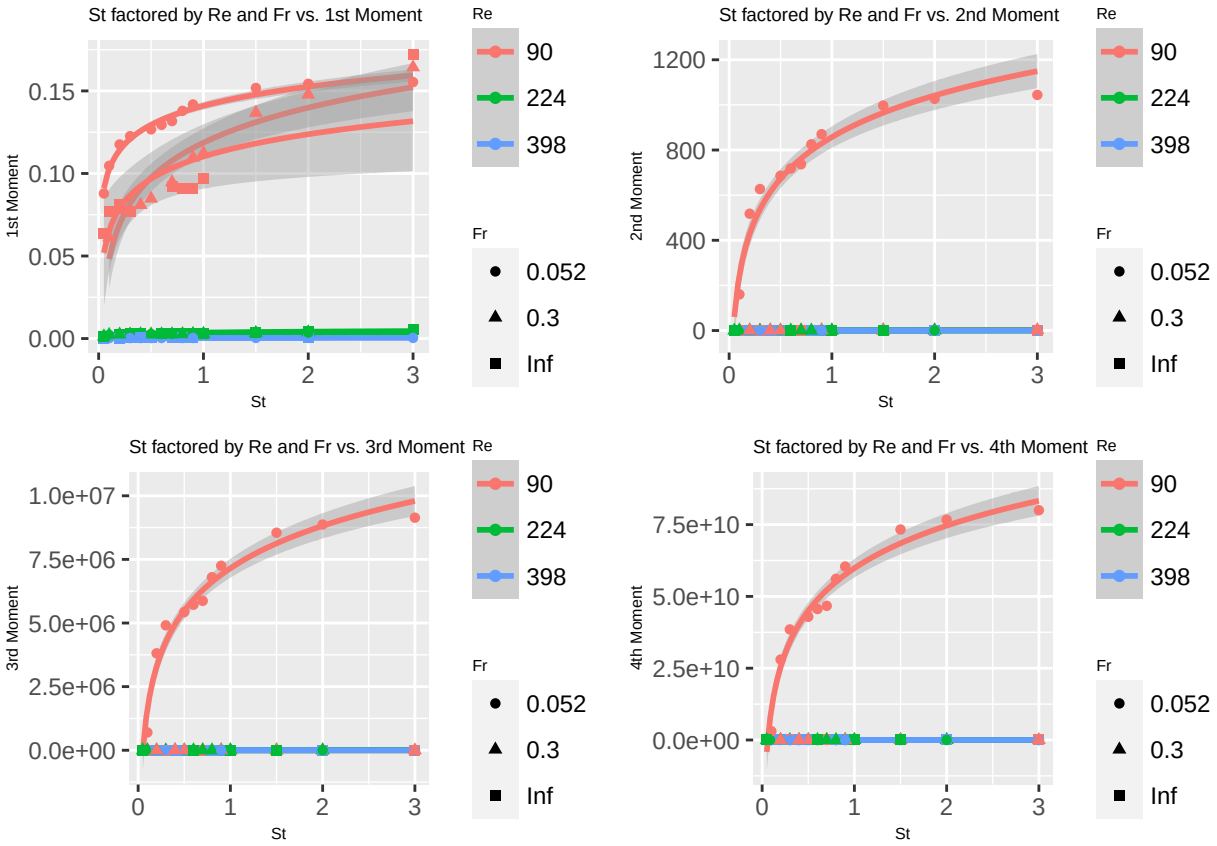
Since we have relatively few data points, n = 89, we decided to rely on LOOCV for our model selection process. We also vetted multiple different cost functions, from MSE to root MSE to root MSlogE to MAE. The main reason for this was that MSE was over-inflating the negative of outliers and making some models seem as if they were performing worse than expected. Root MSE, or rMSE, and MAE helped to taper this by producing more interpretable values that were easier to visualize whilst still preserving the minimum and maximum of the MSE function. One step further was looking at root MSlogE, or RMSLE, which heavily scaled MSE and allowed us to compare the performance of one model to another across differently scaled response variables. For example if the 1st moment's model has an MSE of 0.1 and the 2nd moment's model has an MSE of 0.2, one would be inclined to assume the first model is superior, but since the values of the second moment tend to be orders of magnitude higher the latter model is actually performing a much more impressive estimation. This relative difference is much better captured in RMSLE.

We tackled this problem sequentially, first focusing on models that provide information about the 1st

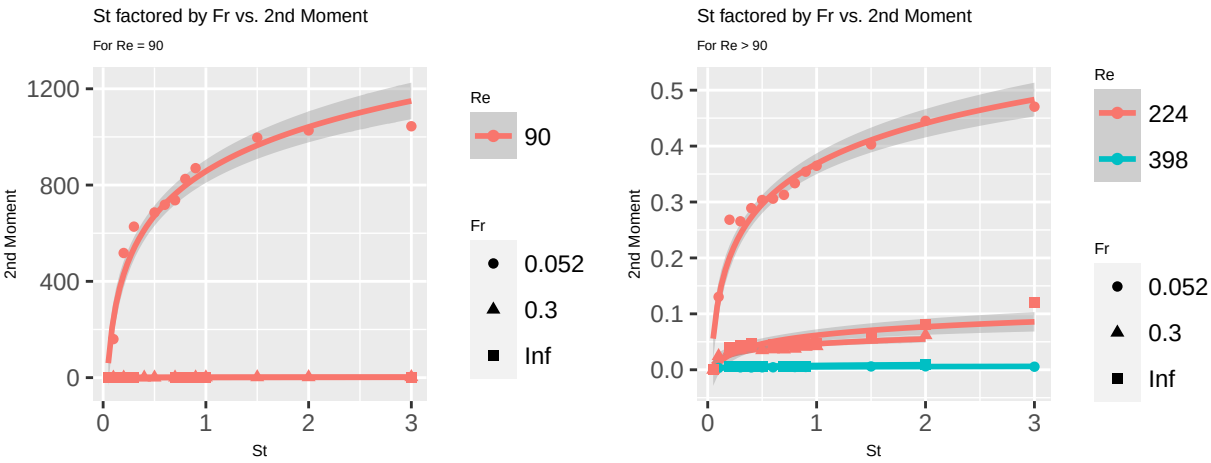
moment, and moving forward until the 4th moment. To create a proficient model we estimated the test error performance of various different degrees of polynomials and splines with various combinations of degrees of freedom. An example is shown in the graph below of the process looking at the difference in LOOCV MAE between different numbers of knots for a quadratic spline model.




Notably, the simplest model has won out. Then, the next step was to look at what kind of interactions we may need. To identify these interactions, we plotted the data and performed some exploratory data analysis.




As we get to further moments, it becomes very clear that the majority of our error came from the much larger values all associated with an Re of 90 and an Fr of 0.052. These looked to follow a logarithmic path of St and when we applied that to all the interaction terms, we found our best model yet.




Our model selection process began with a very complex model containing a complete set of interactions terms, which included triple interaction terms across the three predictors (and abiding by the strong hierarchy principle all component interactions and main effects), and included log, square root, and inverse terms. We then performed backwards model selection using LOOCV, MSE, and MAE to select optimal predictors and all metrics agreed on the following models for both the first and second moments.

Below we have the final models we derived. They are all similar in selections of predictors, which follows from a similar shape of functions observed in the exploratory data analysis. After running these models we make sure that all the moments are positive by applying `max(predict, 0)` since, according to the slides, the distribution should follow a log-normal distribution.

```
predictions = data.train

lm1 = glm(R_moment_1 ~ log(St)*Fr*Re + St*Fr*Re, data = data.train)
lm2 = glm(R_moment_2 ~ log(St)*Fr*Re + St*Fr*Re, data = data.train)

predictions$moment_1 = predict(lm1, predictions)

lm3 = glm(R_moment_3 ~ moment_1*Fr*Re, data = predictions)
lm4 = glm(R_moment_4 ~ log(St)*Fr*Re, data = data.train)
```

Notably, the third moment uses our prediction for the first moment. We found that our predicted value for the mean was accurate enough to be reliable used in predicting other moments.

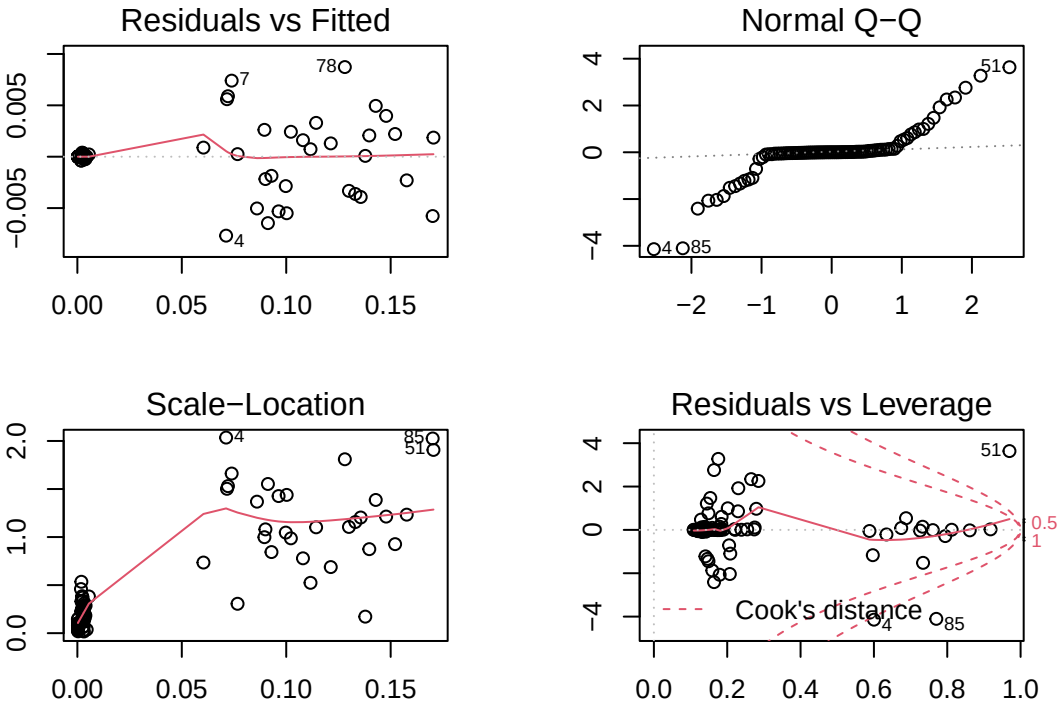
## Results

Our final models had highly accurate predictions of the training set for moments 1 and 2 but got further from the true data into moments 3 and 4. The model is more interpretable than a spline or other models may as the linearity allows us to look at coefficients for specific interactions and variables explicitly. However, there are a lot of interaction terms so the individual parameter effects are not obvious to see.

The main scientific finding is that, generally as can be see in our graphs, there is a logarithmic relationship between Stokes number and each moment with its coefficient dependent on its values of Froude number and Reynolds number. All moments increase with St with the moments of Re = 90, the smallest Re, and Fr = 0.052, the smallest Fr, creating much largest moments of all degrees. This means that, comparing these levels of Re and Fr, a smaller fluid turbulence and gravitational acceleration are correlated with larger particle clusters. For all interactions, cluster size increases logarithmically with Stokes, which quantifies particle density and size.

## Conclusion

The most significant weakness in the project is that our final regression models all had rather poor model diagnostics. Analyzing the residuals reveals a multitude of outliers, non-constant variance, and a relationship that is unlikely to be normal. However, this can all be rapidly fixed by re-fitting the models with identical predictors but fitting them to the natural log of the response variables. The reason we decided to rely on the former is because the log fit models have significantly higher MSE than their regular counterparts, ranging from a 14% increase for the first moment to a multiple order of magnitude increase for the later moments.




Another absence in our model is the NA values for the interactions of Fr = 0.3 and Re = 398. These do not have values in our data set because there is no data point with that value. Fortunately, there is not a data point with those values either in the test data so there will be no gaps in predicting this specific data. Generally, this is just one example of how this model may not be accurate to the real world situation as we are fitting off of a very small data set. This limitation is also due to the fact that we represented the Froude numbers and Reynolds numbers as factors in the model, which completely ignores the quantitative difference between values.

A possible extension to these models if we had more time would be to simulate more data. With this we could have more accurate prediction, a full model without NAs for the points we have currently nothing to train on, and enough to separate more training and testing data for fitting. Additionally, with more time and beyond the scope of this course, we would be interested to train a neural network to improve prediction on the test set.