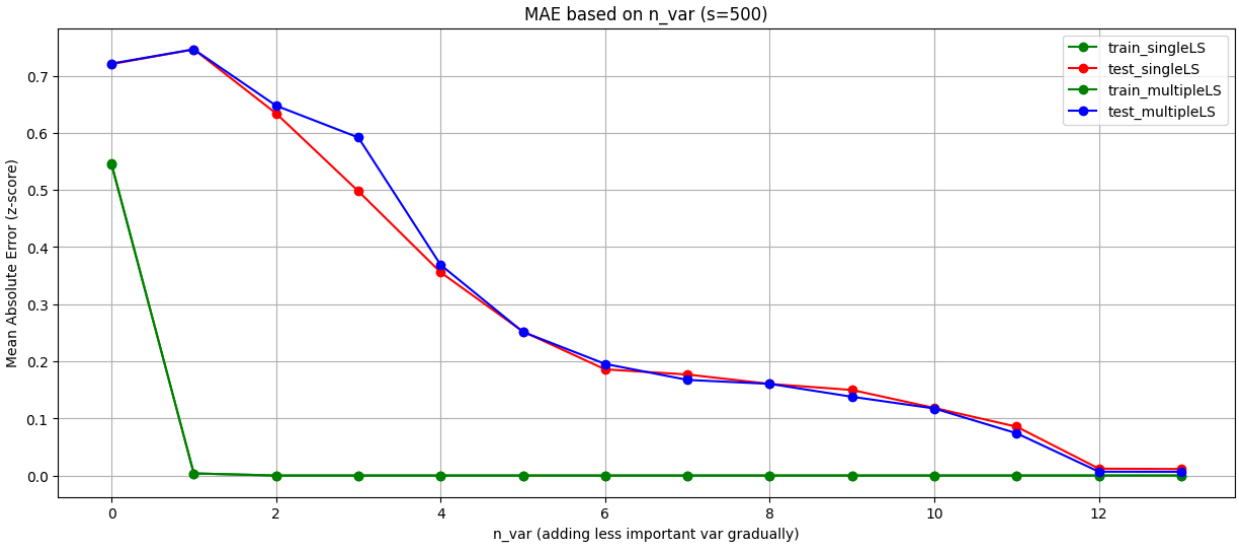
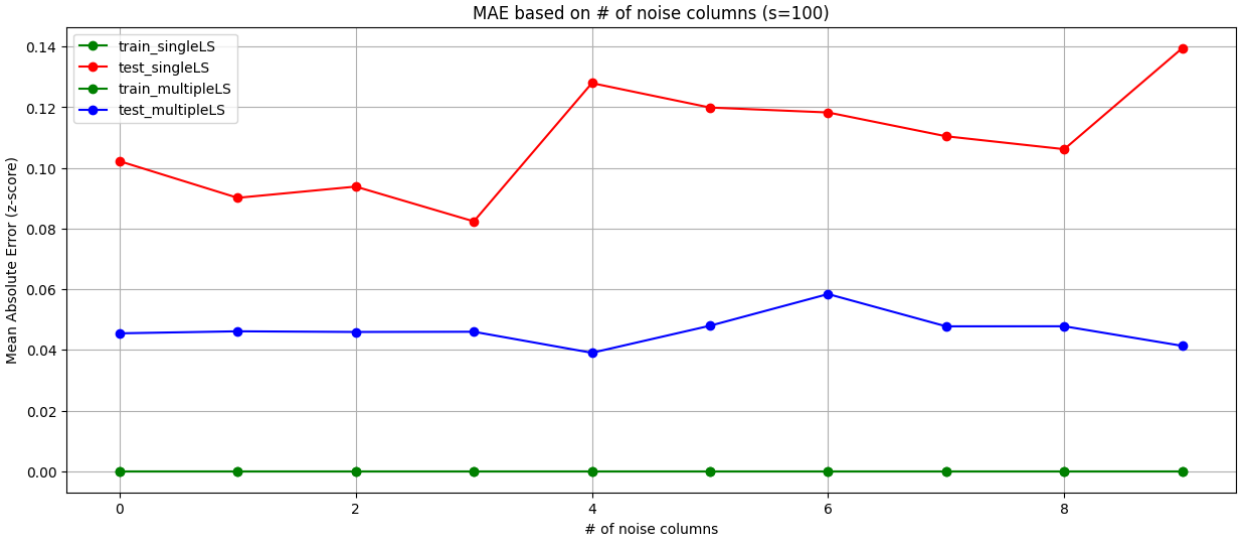
Length Scales



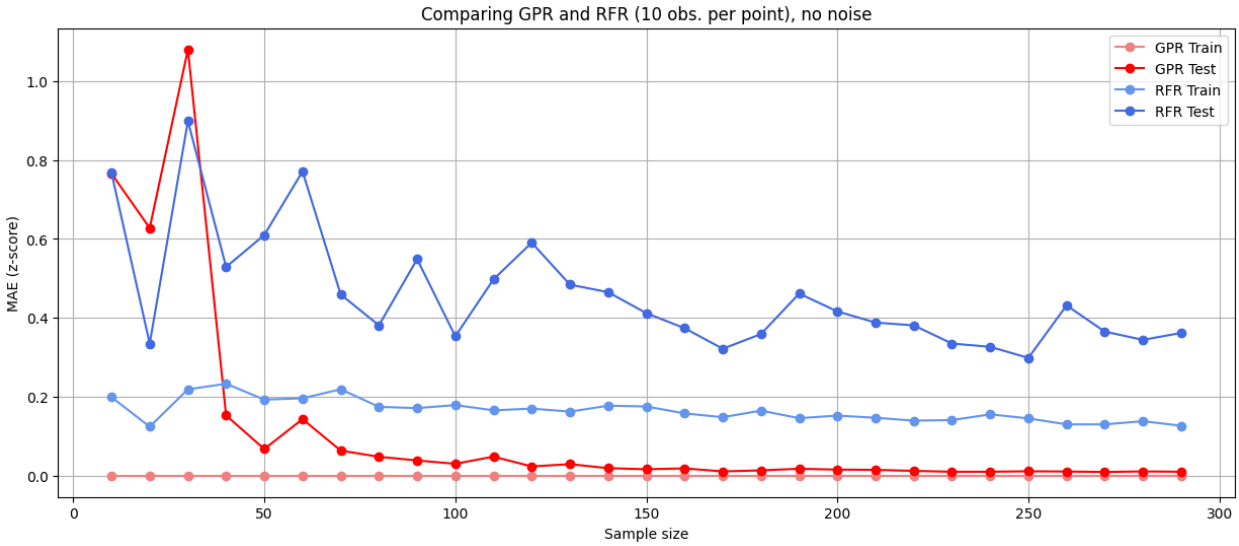
Increased dimensionality does not seem to negatively impact the model. Varying length_scale parameters do not make a noticeable difference when used with non-noise variables.

Noise Columns

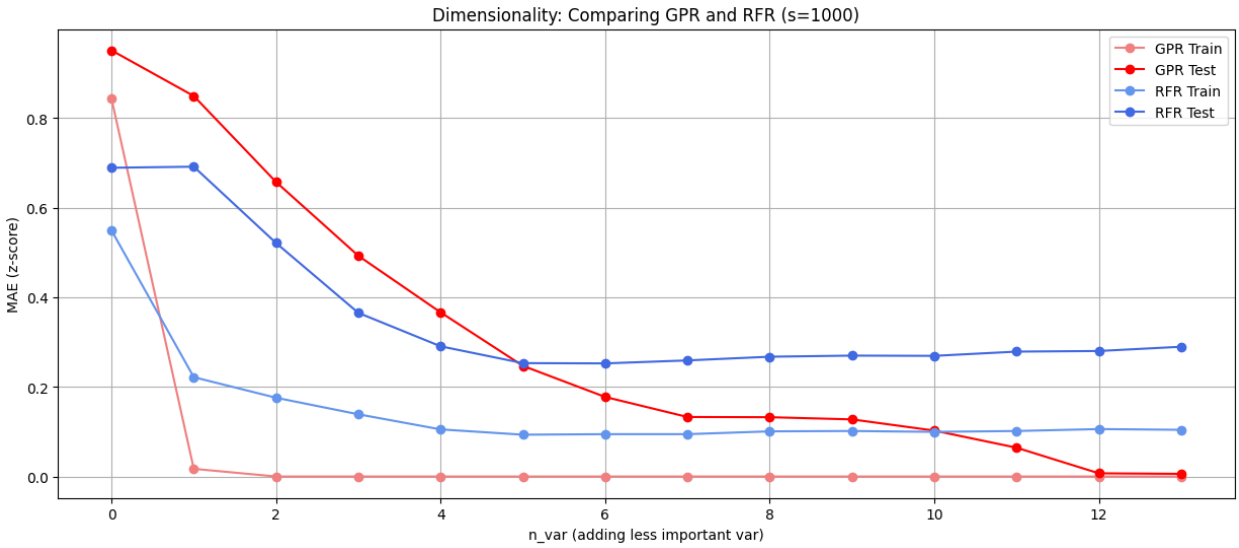


By passing an array of 1's into the length_scale parameter, it makes the model robust to additional noise columns, while the constant length_scale model suffers from noise/dimensionality.

GPR vs. RFR

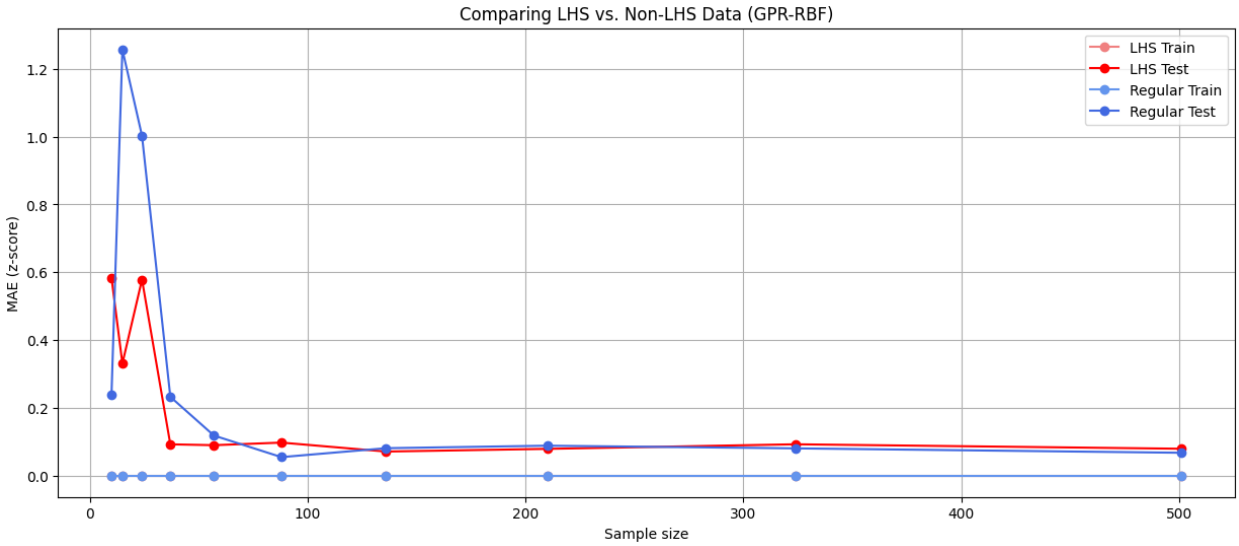


With all the variables, GPR beats out RFR.



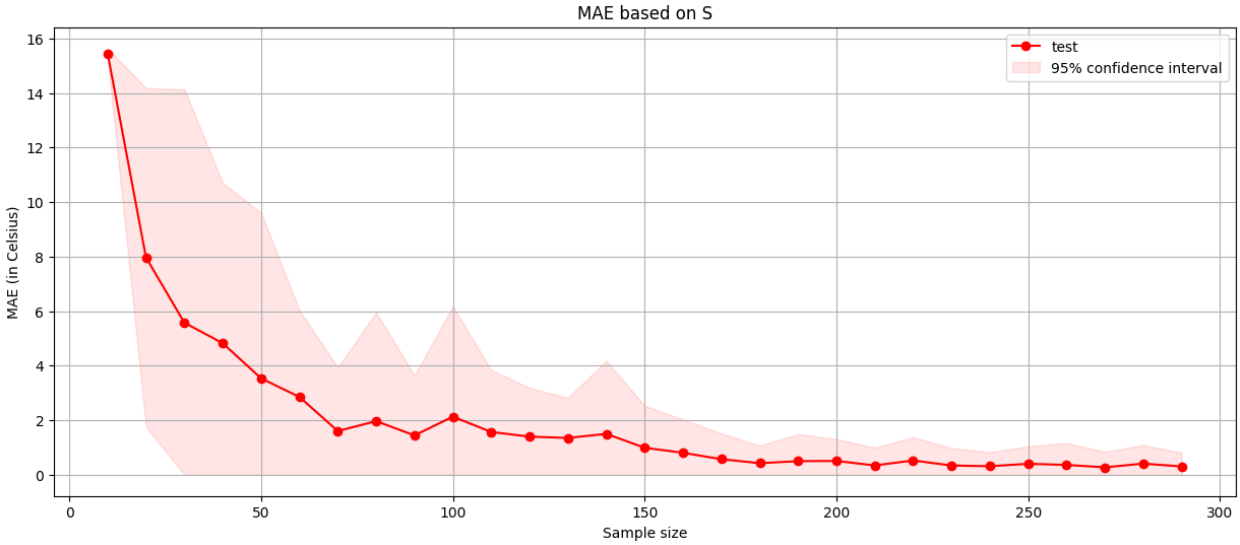
Past 5 variables, GPR beats out RFR.

Latin Hypercube Sampling



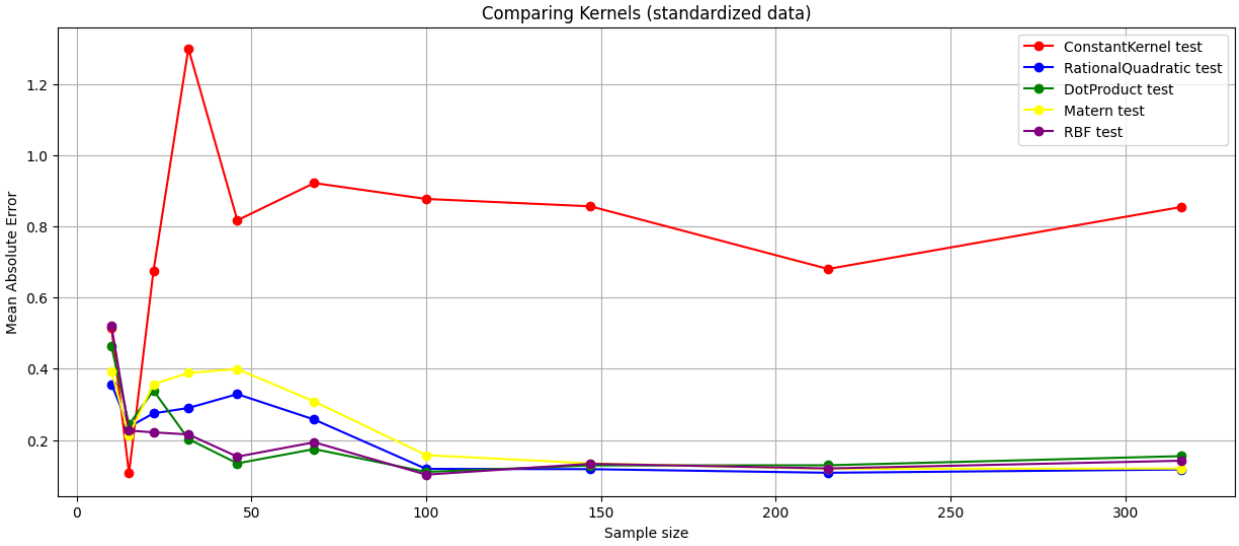
Aligning with theory, LHS requires slightly less samples to reach a better MAE, but not by much.

Confidence Intervals



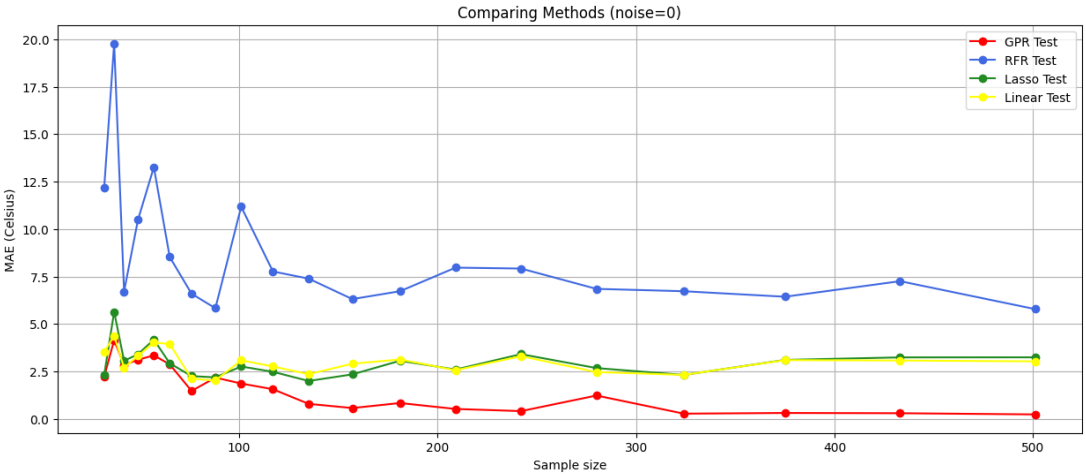
I learned that with GPR (because of the length_scales) you need to standardize the dataset before computations. Here I was able to translate the MAE in z-score back to Celsius, which confirmed that the predictions are within the 20°C range.

Comparing GPR Kernels



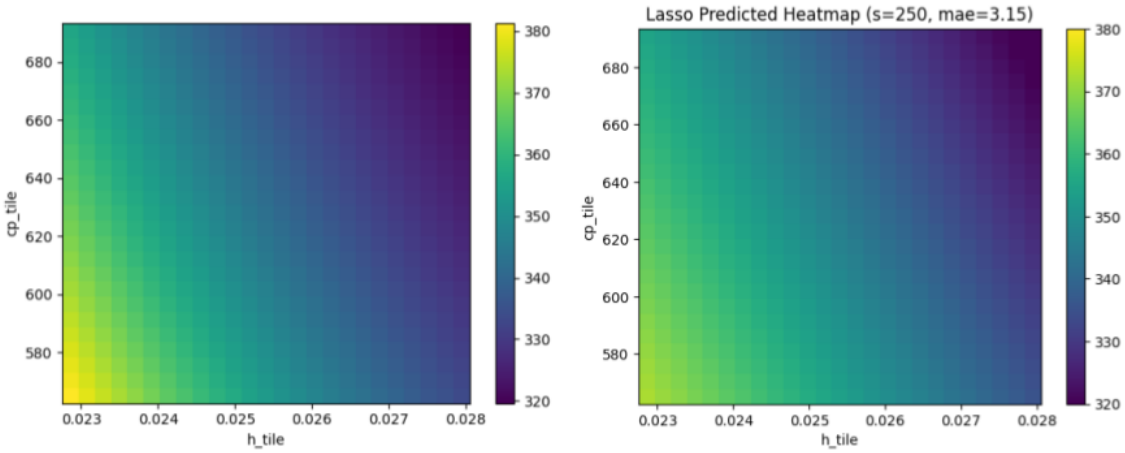
Comparing the different GPR kernels head-to-head was a goal of mine. Here we see that the RBF is about equal to the others (besides Constant). Since RBF is the standard, then there's no need to change it.

GPR vs. RFR vs. Lasso vs. Linear

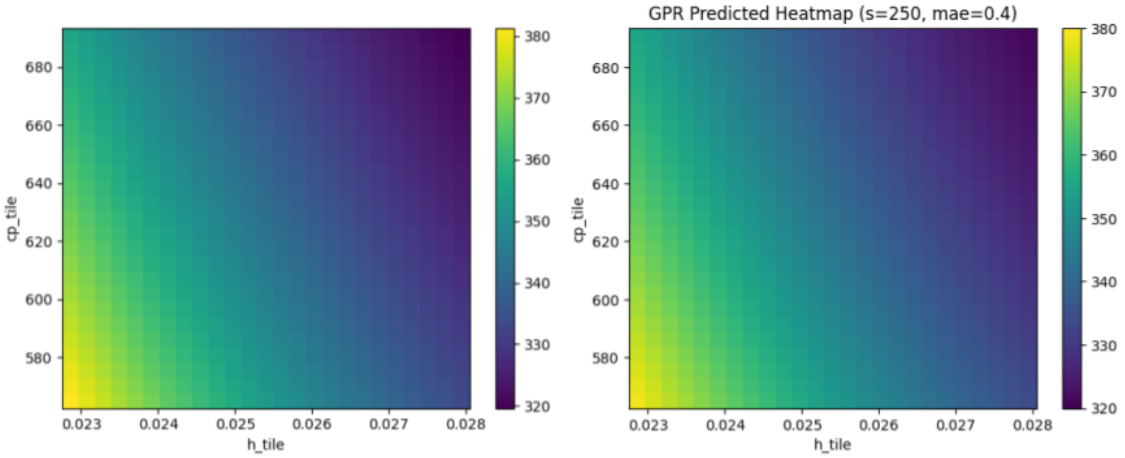


Linear and Lasso perform relatively well which indicate a linear data shape.

Heatmaps: Lasso

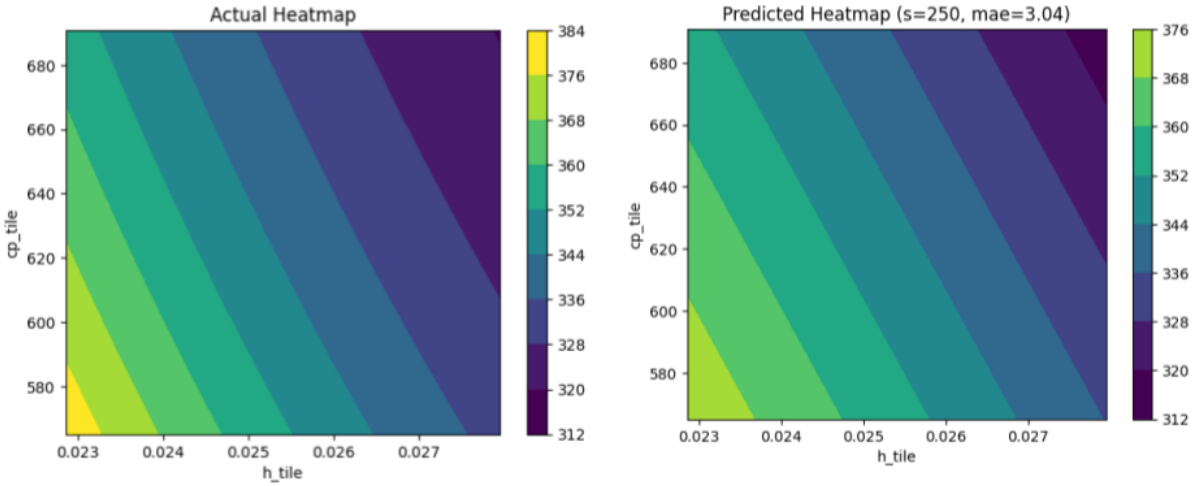


Heatmaps: GPR

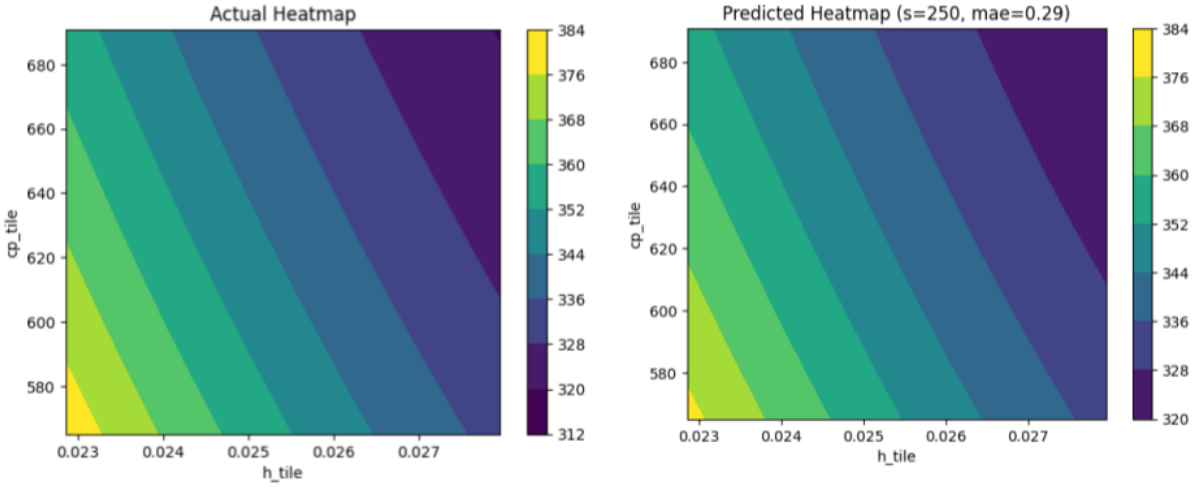


The predicted heatmaps show slight differences from the truth in the bottom left and upper right.

Heatmaps: Lasso



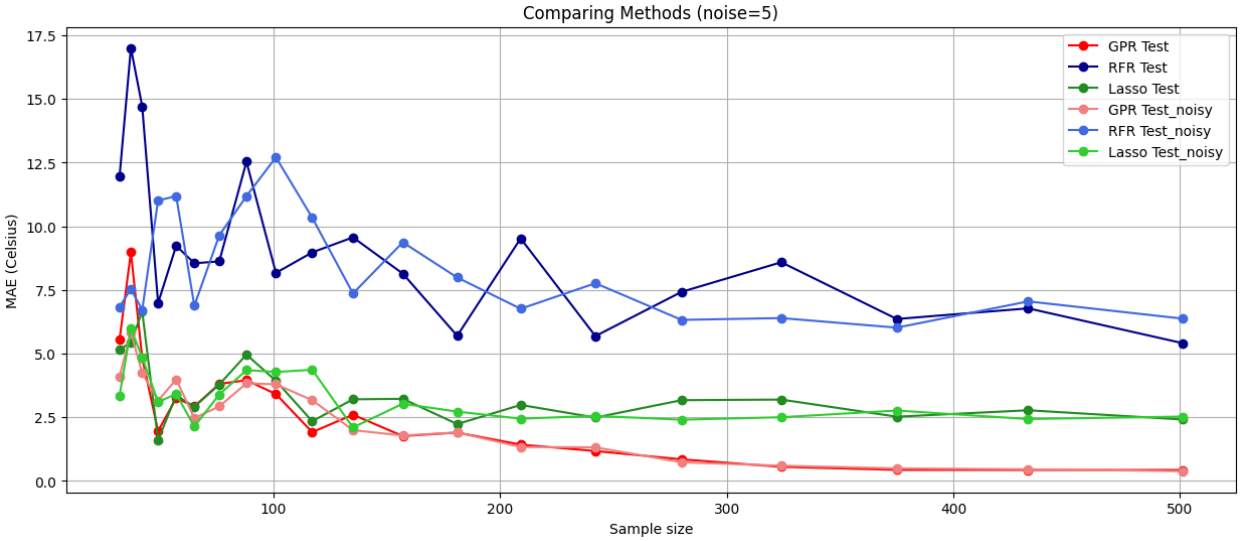
Heatmaps: GPR



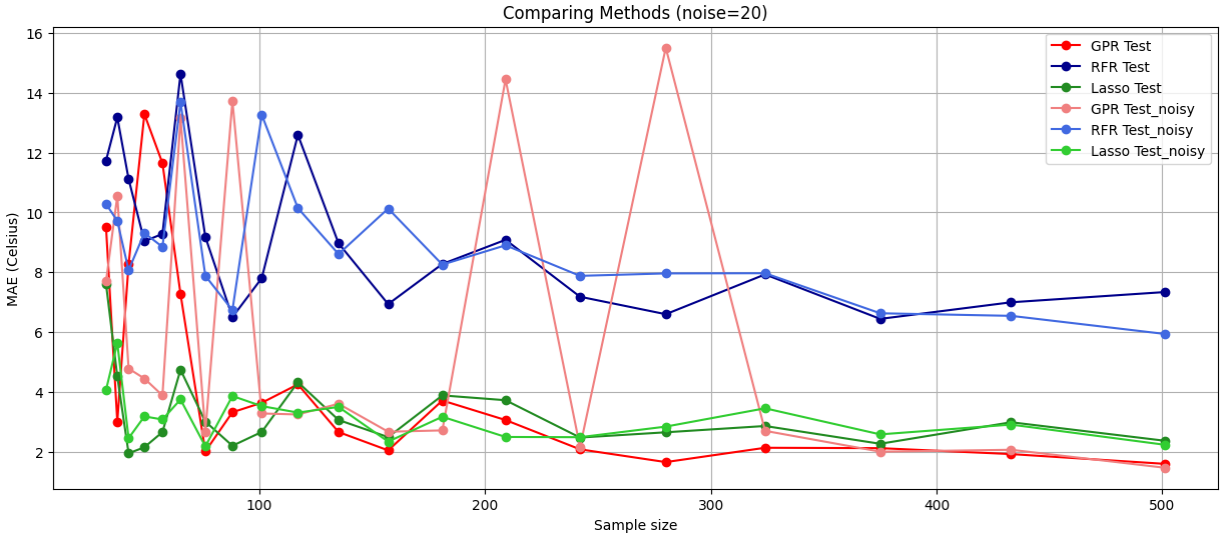
The contour plots give a better view of the ‘slice’ of yellow at the bottom left that both models miss out on, albeit GPR has a small slice.

The defined lines also show a very linear relationship with these two significant variables.

GPR vs. RFR vs. Lasso (noise)

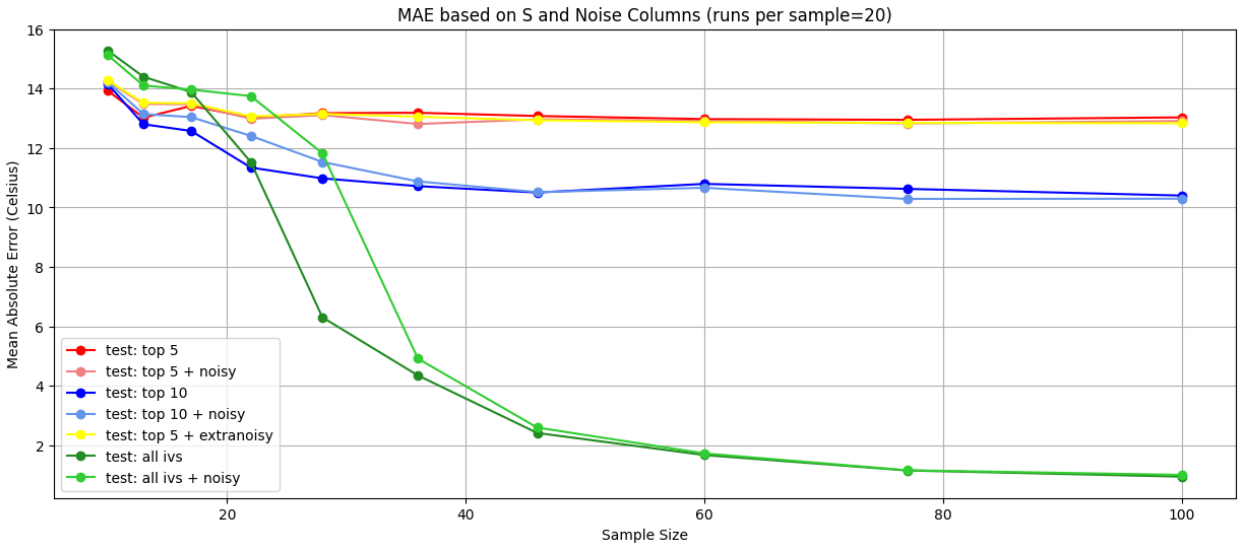


Low noise does not seem to impact any of the methods, when all IV columns are being used.



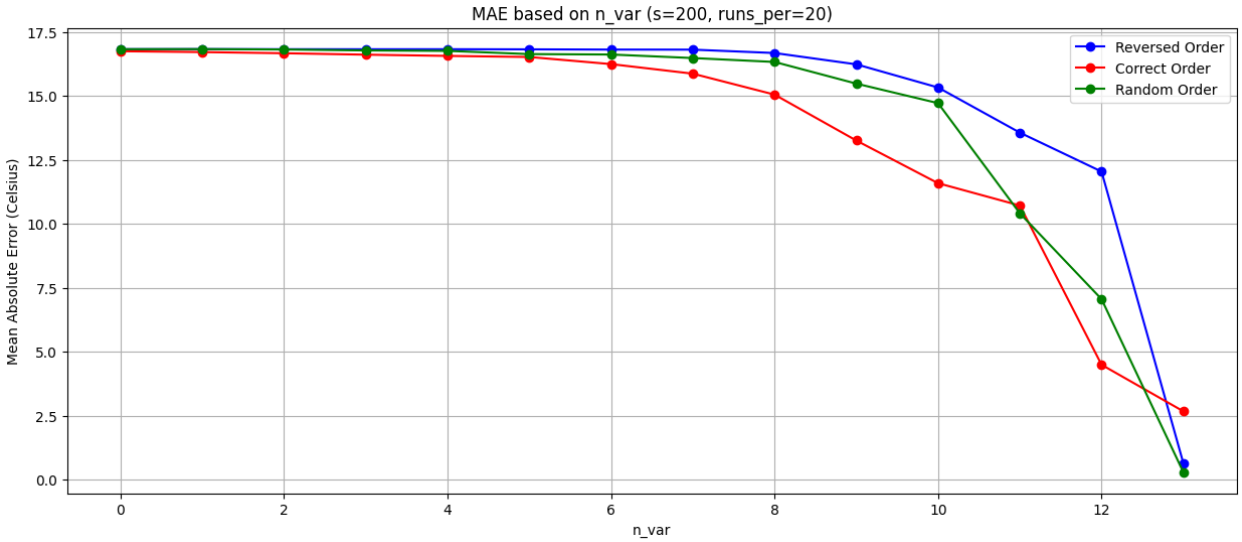
High noise only hurts GPR, but this is only with 1 run per sample size. Having multiple length columns lowers the impact of useless variables.

Top 5 variables vs. Top 10 vs. All 14

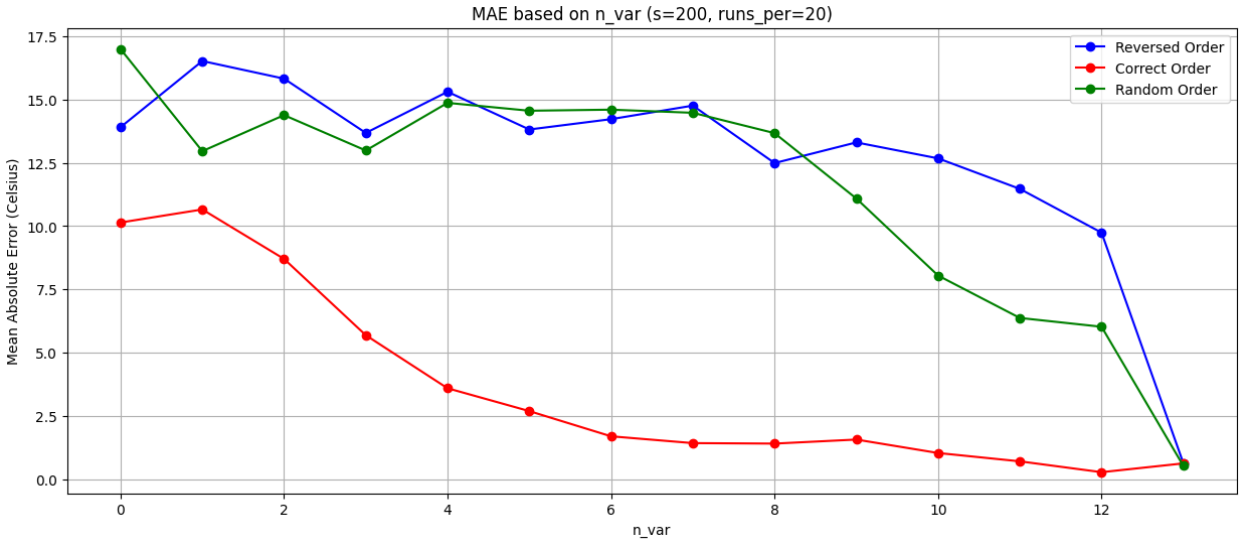


The higher the dimensionality, the greater the impact of noise (at s<100). This is even with using multiple length scales. GPR needs a certain amount of data to determine what is useful and what isn't. It's also interesting just how much better the green curve becomes as s increases.

Impact of Input Ordering:

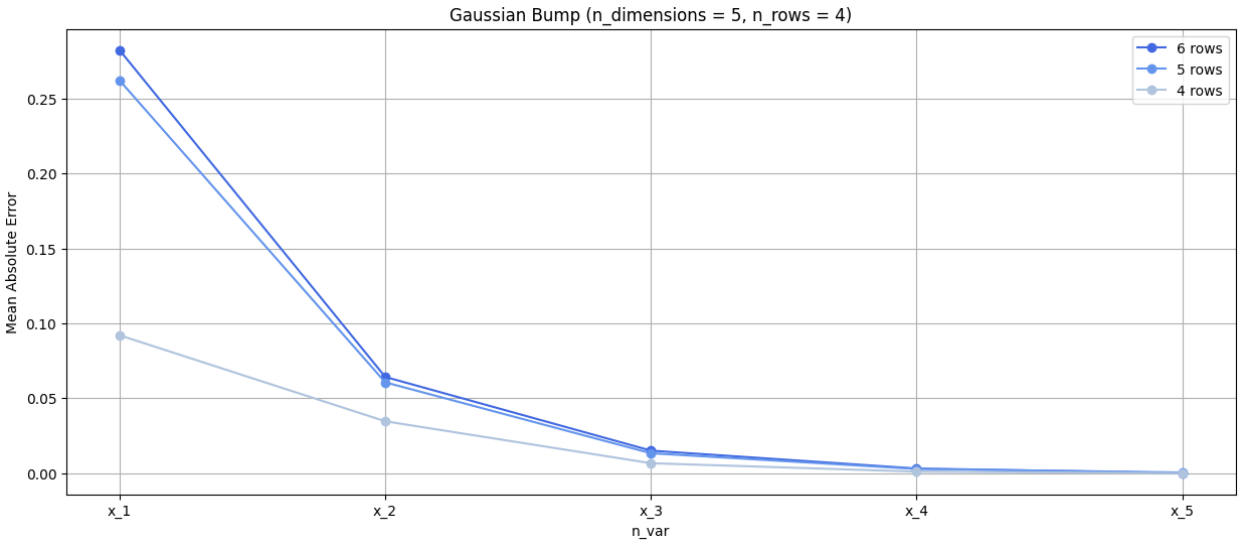


This graph is with each point being compared to a 14-dimensional test set. For 2 variables, there would be 12 zero columns to extrapolate on. These curve shapes are roughly what we would expect from

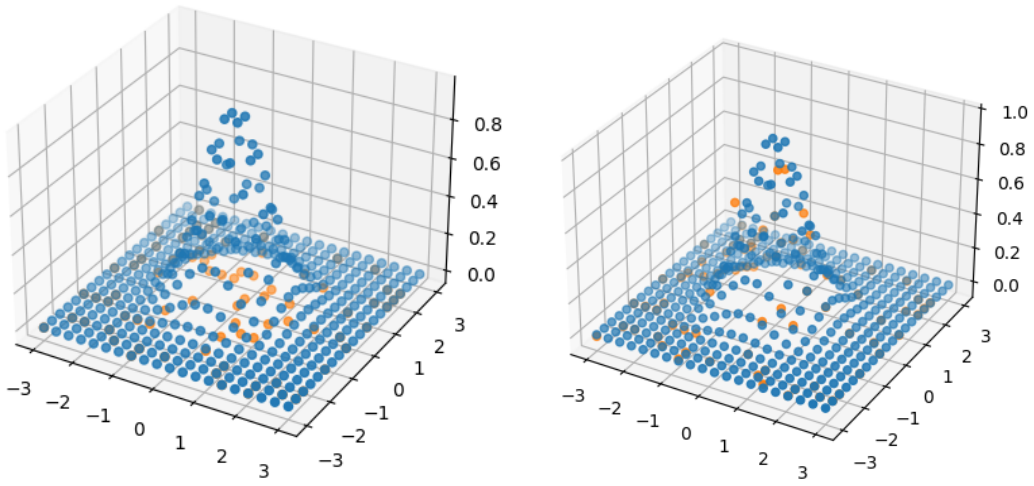


This graph compares n_var with an n_dimensional test set.

Gaussian Bump



As we reveal each additional dimension, we see the expected convex shape. The model performs pretty well on 5 dimensions only seeing 4.



Left graph: 2-D estimation based off 1-D test set. Right graph: 2-D test set.