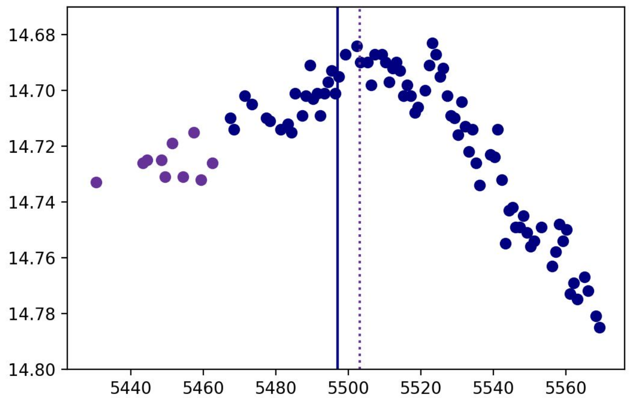
https://github.com/lenatreiber/LXP69.5
- LXP69.5FlareModelling
- Flare.py

Latest summary at the end of this doc

I. Simple Gaussian Fit

The reference model is the gaussian fit to the I mag data for a flare subtracted from the max I mag value for that flare. From this model we get 13 flare centers: [5497, 56557, 5799, 5960, 6318, 6637, 6989, 7314, 7477, 7682, 8155, 8559, 8756]. The model returns an error of at most a few days for each fit. The fits look reasonable (although more uncertain that just a few days) for all flares except for the first. The solid navy line shows the gaussian fit center. We can also see right away that the center changes by more than the given error if we change the data included in the fit. For example, if we take out the first ten points (purple), we get the dashed purple center.

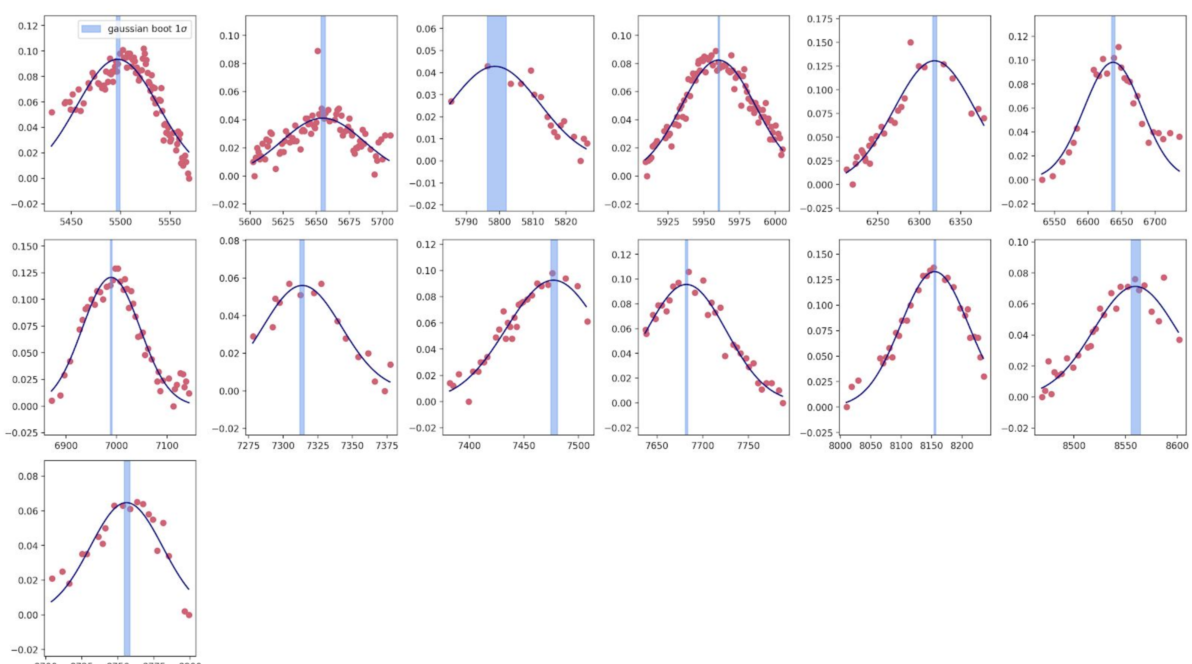


A. *Changing the Index Ranges*

Starting with the first flare, we see that changing the data used in the fit can alter the best fit center by up to six days. However, there’s no real justification of cutting that beginning data or adding the data beyond what’s shown above, which turns brighter, ascending towards the following flare. The test that makes sense, though, is altering the cutoff at the end by a couple days in each direction, which only shifts the center by half a day. Likewise, if we use reasonable ranges within which the cutoff of the flare would be, the center generally changes by at most a few days.

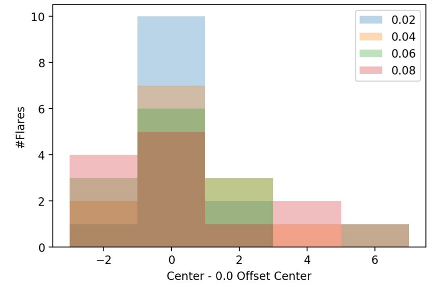
A. *Bootstrapping*

Next, bootstrapping the data addresses the sampling within the range. How much does a center change with fewer observations? Each flare is plotted below with the gaussian fit in navy and the blue shading showing one sigma from the mean bootstrap result (using 100 iterations). I don’t think the gaussian bootstrapping is particularly helpful. We see that it mostly speaks to the number of data points.



II. Gaussian + Vertical Offset

We can look at the effect on the center when we add a vertical offset (so use max - I data + offset). The effect of such a change on the other gaussian parameters is shown in the Overleaf. The change on the center is mostly within the original error, but there are a couple flare center values that change by more than three days. The remaining gaussian models maintain an offset of zero.

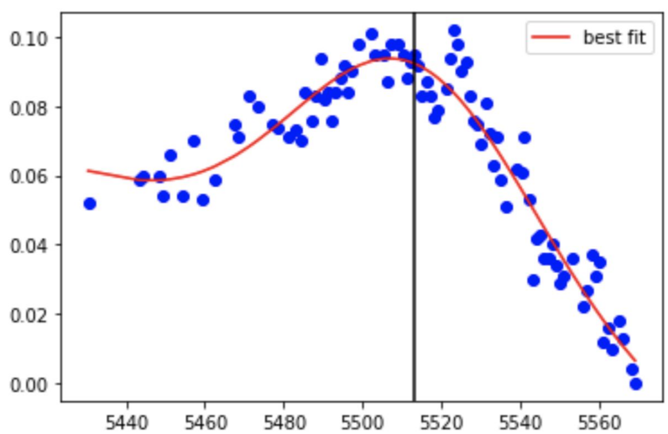


III. Gaussian With Detrended Data

Because of the long-term trend, it may seem best to fit the gaussians to the detrended rather than original data. However, the change in centers is quite small -- the standard deviation is 1.5 days and the maximum change is three days (for the first flare). Furthermore, the choice of detrending window affects the result.

IV. Gaussian+Line

Combining the gaussian fit with a line (essentially a base with a nonzero slope) is another way to address the asymmetry of the flares and the underlying trend. We see with this first example of a different model that the main source of error in fitting comes from the model assumption.



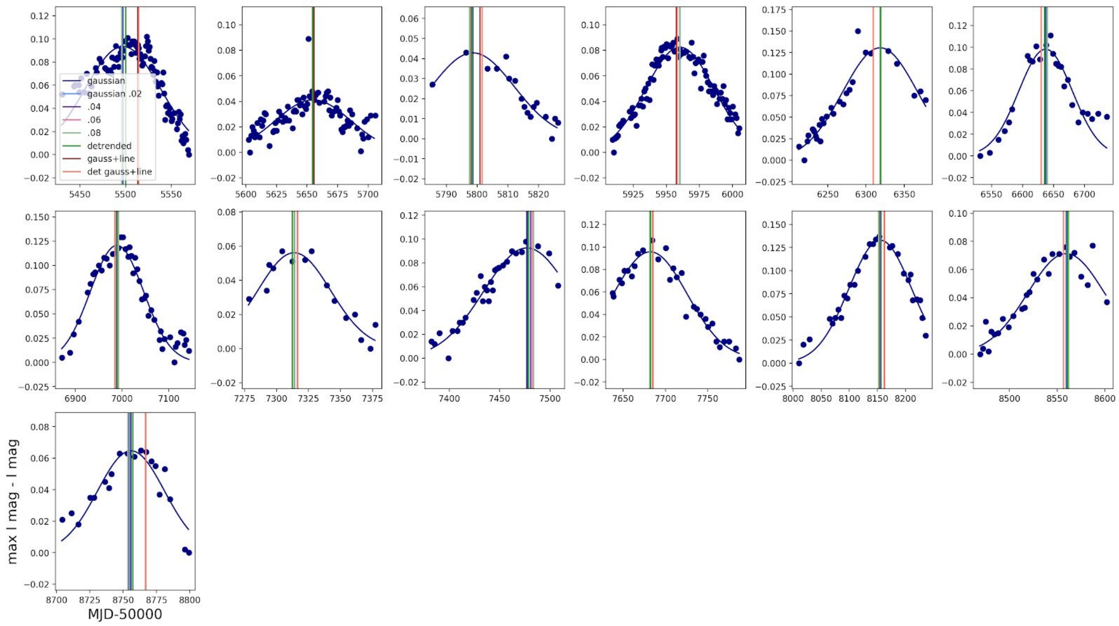
On the left is the gaussian+line fit to the first flare. The center has shifted right by 16 days to a more believable place. The fit also looks better by-eye. The reduced chi-squared of the g+line fit is 5.79e-05 and 1.2e-04 for the simple gaussian.

The standard deviation of the difference between the gaussian+line and gaussian centers is seven days. Again, detrended the data first does not change much.

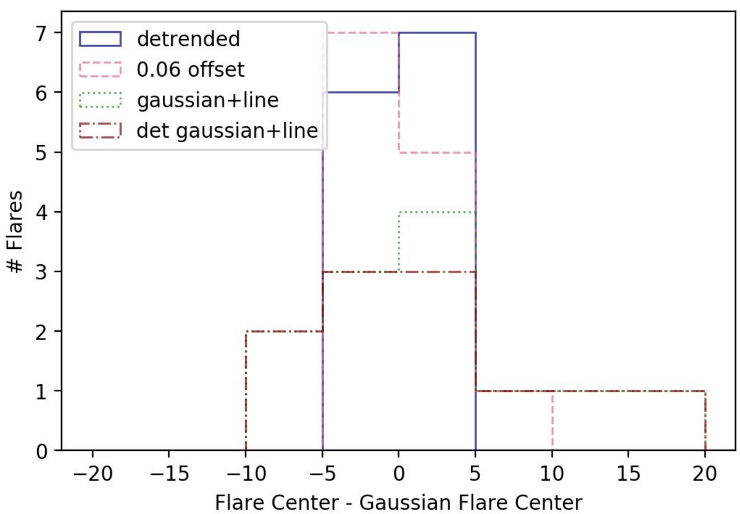
As may be suspected, this model is more sensitive to the inclusion of start and end data, since it will change the slope of the line.

V. Comparing Gaussian-Based Models

We can visualize the effect of using these different models by comparing the centers on the flare scatterplots and by making a histogram using the original model as a reference point. We will repeat these comparisons once we add other models.



The histogram only includes one additional offset (0.06 magnitudes). The differences above ten days come from the gauss+line models for the first and last flares. The binning is set to have bounds every five days.

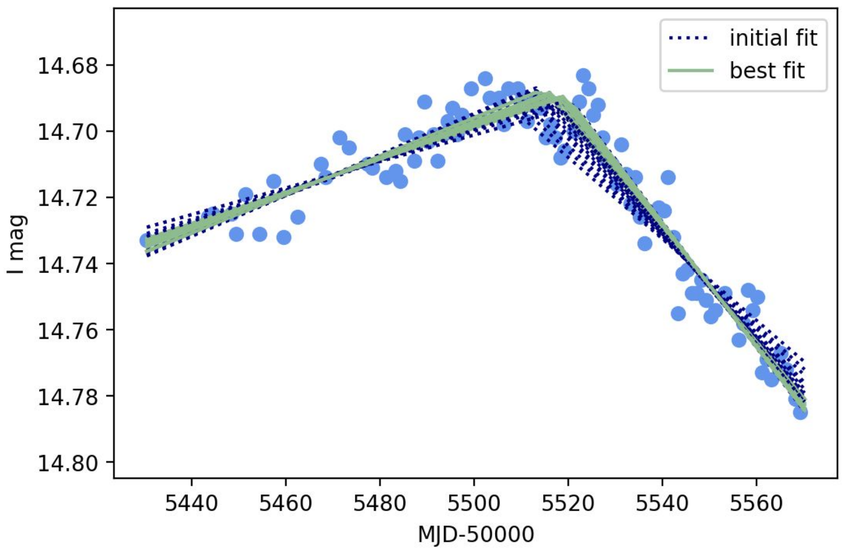


VI. Sinusoidal Model

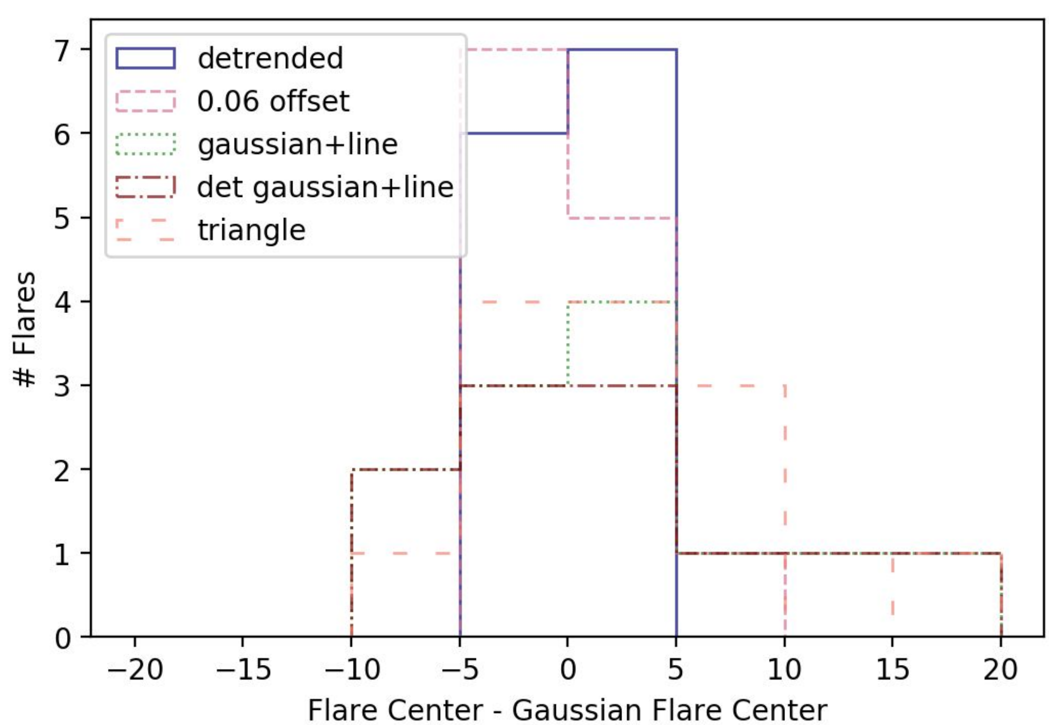
Using a sine function to fit the flares resulted in a maximum center change of three days; only one other difference was more than a day.

VII. Triangle Model

The final model is the triangle. The triangle function is parametrized by the two slopes and two intercepts of the lines that intersect at the "peak." There are other possibilities, such as using the peak as a parameter, but this version worked. The intersection is then found and the first line is used before the x location of the peak and the second after. A decent initial guess was needed for the fit to succeed, so the fitting function first splits the flare in half and fits a line to each side. The resulting parameters are used as an initial guess. The concern is then, of course, how much the location of the split matters for the final fit.

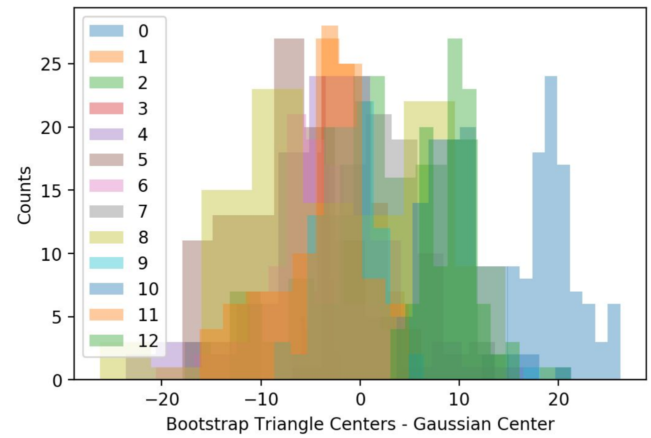


The split location should be reasonable -- somewhere in the middle of the peak -- but the fit converges even with various split points. The third flare, thanks to its asymmetry in the number of data points, needs the split to be shifted a little bit in order to find a reasonable peak with the best fit triangle. If the split location is drastically varied (by 20 points in one direction and 14 in the other), we see the first flare's best fit peak change by four days (left). As one might expect when working with linear models, the triangle fit is sensitive to the cutoff of points used in each flare. We can look at each of these three variables (split location, start index, and end index) independently or in a 3D grid to get a sense of the uncertainty on each triangular fit. If we look by flare at the effect of changing the end index used in the fit, we standard deviations between zero and nine days. The changes in start index alter the resulting center largely by less than two days. The determination of these ranges is not rigorous, but still gives a sense of what can affect the fit. The histogram below is an update from the gaussian histogram, with the triangle values being the means from the 3D grids.

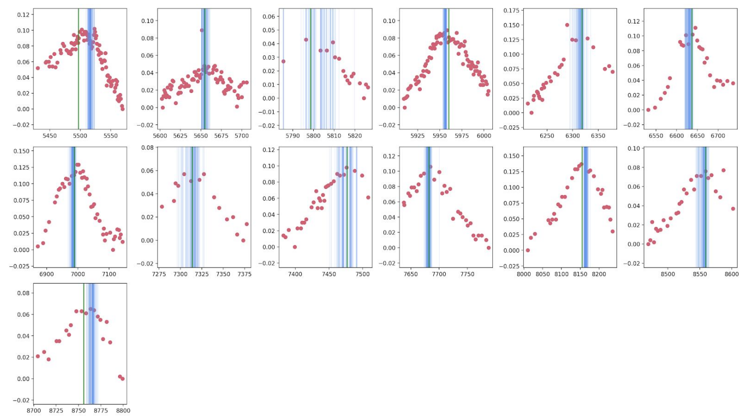


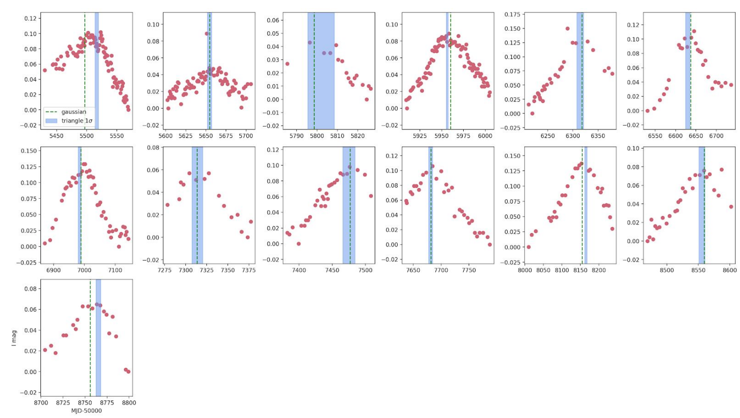
The greatest differences come from gaussian+line and triangle.

Using bootstrapping with 100 iterations again gives the histogram on the right. The scatter plots below may be more informative, but the histogram shows the distribution for each flare. The distributions have different widths of ~10 days, and some triangle fits are systematically different from the gaussian result (particularly for the first and last flares, which also showed systematic change between the gaussian and gaussian+line.

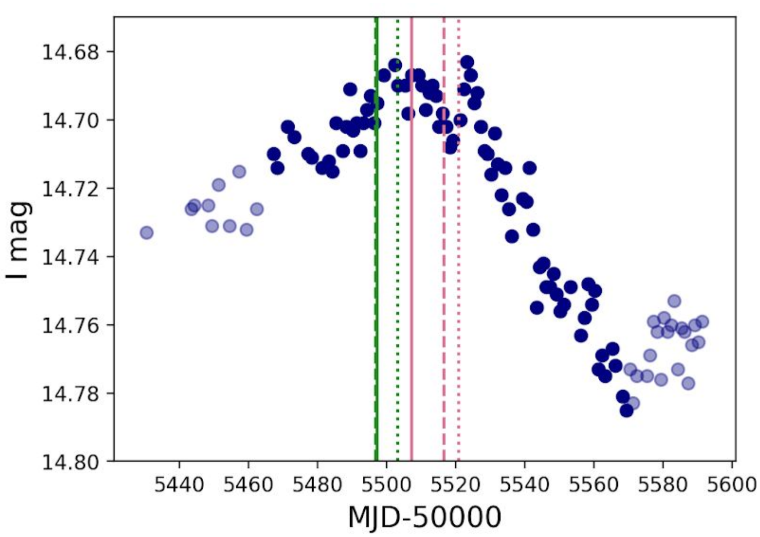


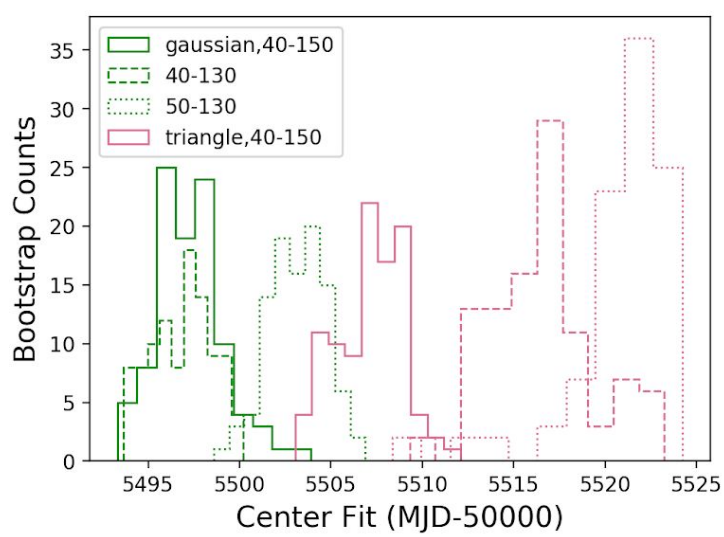
The two plots below are also on the Overleaf. The first shows all bootstrap results. One problem is that the bootstrap results of course change each time they are run. The second just shows a shaded region of the standard deviation from the bootstrapping. The standard deviations are generally higher than they were from the gaussian bootstrapping. The gaussian centers are shown in green.





If we again focus on the first flare, we can look at the effects of different indices and models on the center result:

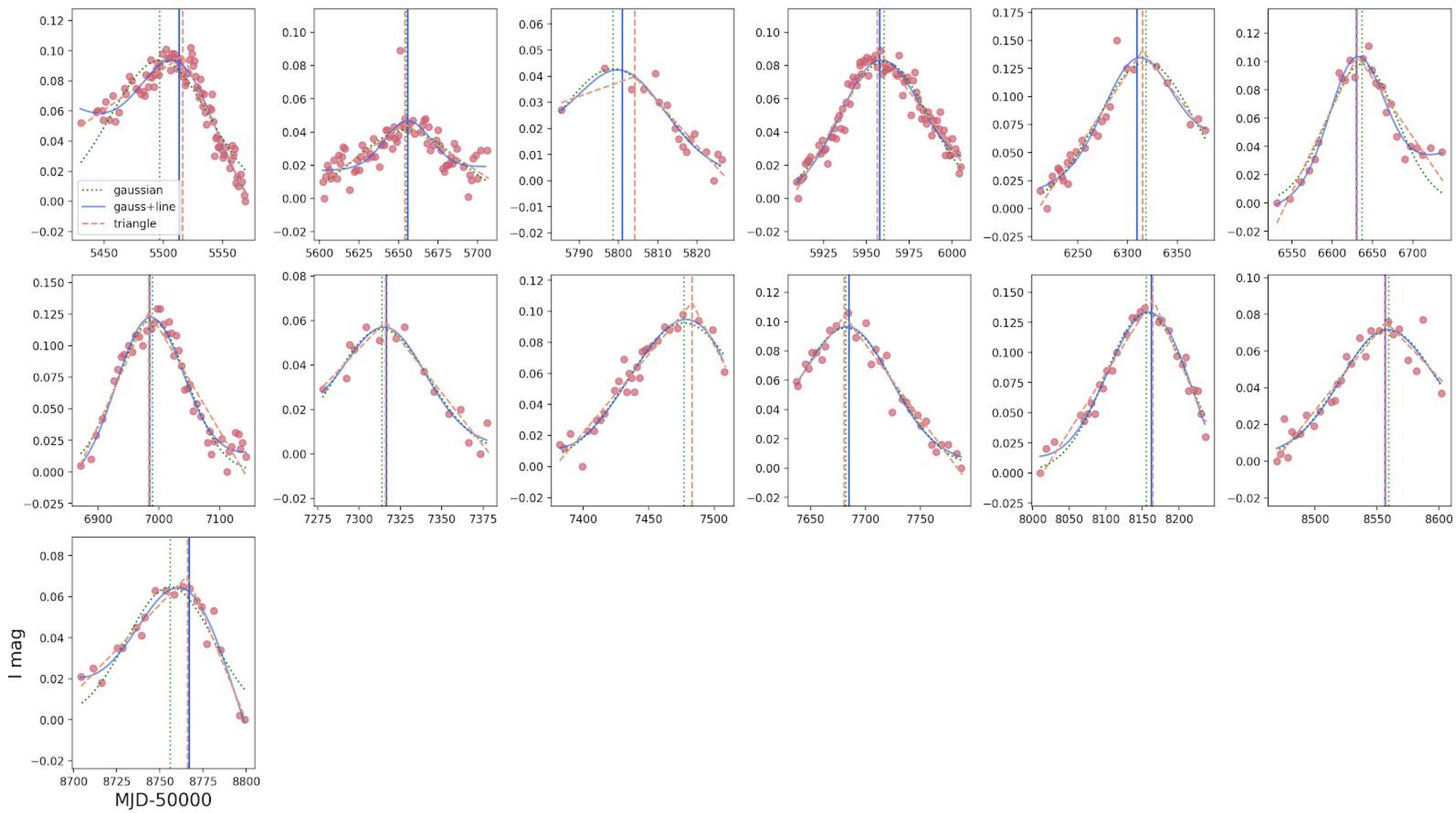




The green lines in each subplot show the gaussian results, while the pink shows the triangle fits. The line styles represent the indices used in the fit. The navy points are index 50-130; the first purple ones are 40-50 and the last are 130-150. 40-130 were generally used for fitting. The histograms on the right are produced from bootstrapping each model. The triangle results are systematically higher. They are most sensitive to the end data inclusion (which makes sense given the linear fitting), while the gaussian fit is most affected by the inclusion of the early points.

VIII. Final Comparison

We can now compare the three models. The choice of model is quite important, but other choices (for the triangle model in particular) can shift the center. The models and resulting centers are shown below. The simplest triangle model (effective middle split, same data range as other models) is shown.



Latest Summary

The modeling jupyter notebook shows everything I've done, but here's a quick summary:

The choice of model for finding the flare centers is non-trivial. The resulting systematic error dominates the errors on the centers. We used gaussians, a gaussian + line composite model (which allows for a continuum with nonzero slope), and triangles. There is no single answer per flare using each model. For example, the choice of start and end point for each flare can shift the center result by differing degrees using each model. We also tried fitting using the detrended rather than original data, and varied the offset used for the gaussian model. Because lower magnitude values are brighter, we had to flip each flare somehow before fitting. The default method involved replacing each flare point with the maximum I mag value from that flare minus the data point. This method assumes that the maximum value is the minimum of the gaussian. An offset allows for the gaussian to extend further, which may make sense since the data is not dense enough to show us every minimum.

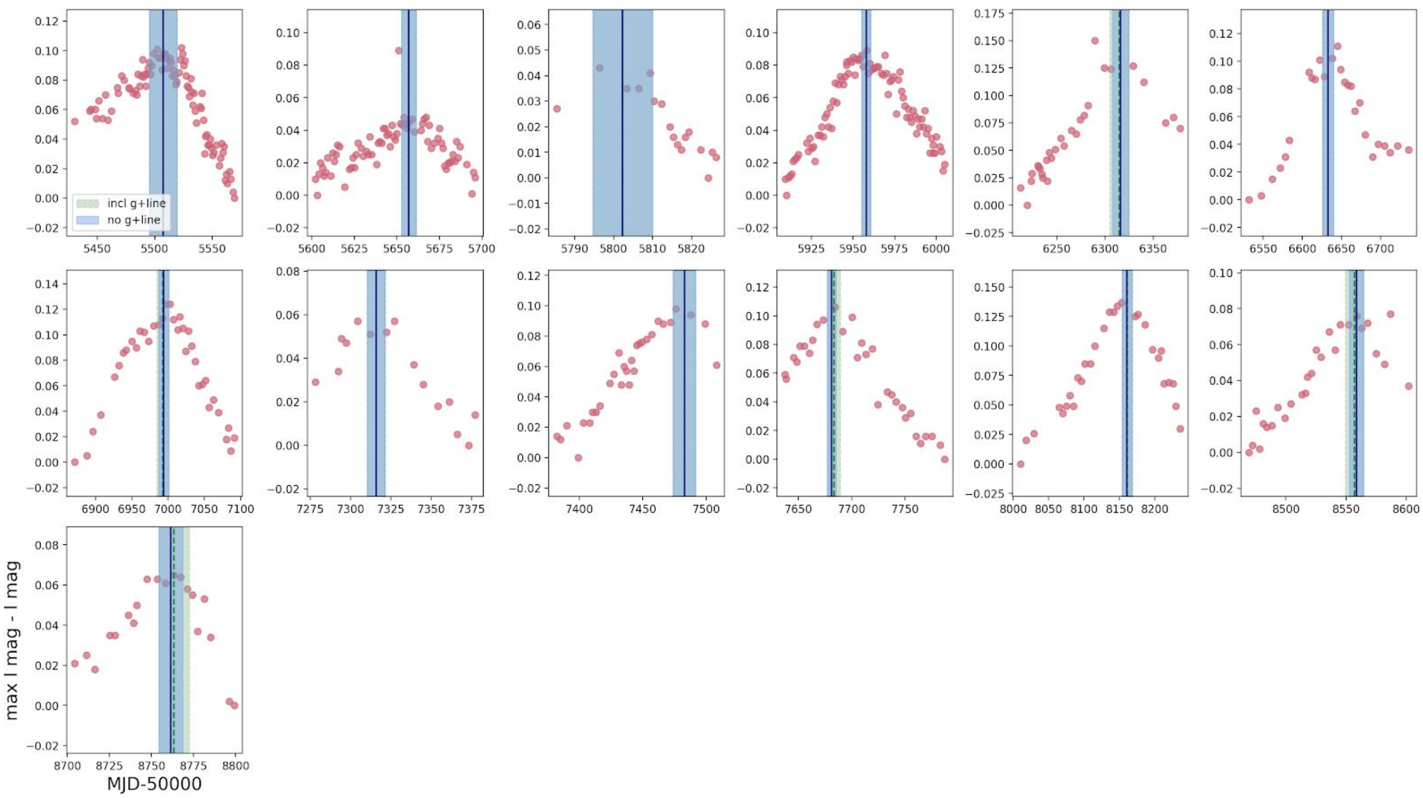
Gaussian
- Detrended vs. original
- Offset
- Start and end points
- Small variations with bootstrapping

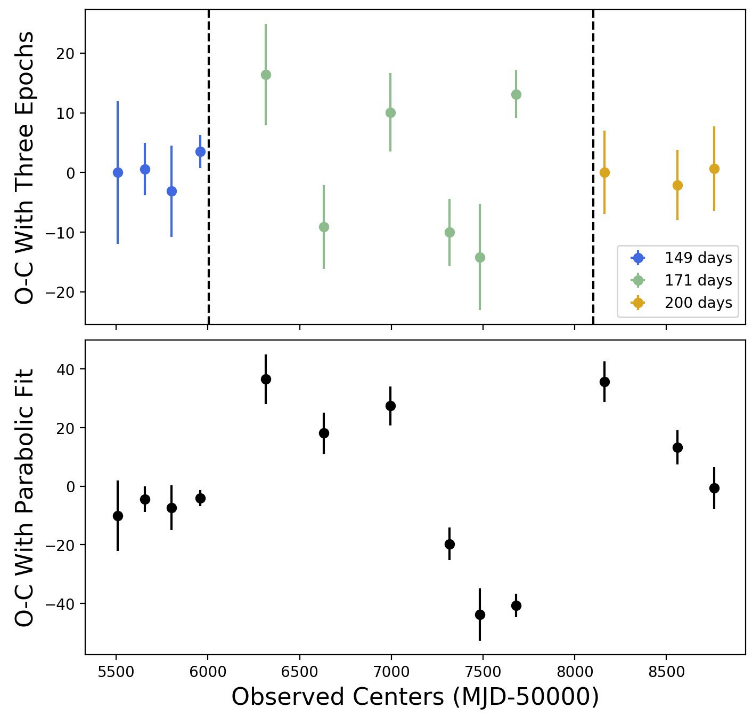
Gaussian+line
- Limits on center determination
- Detrended vs. original
- Offset
- Start and end points
- Most sensitive to bootstrapping

Triangle
- Initial fit
- Start and end points
- Bootstrapping

To find the final center and error for a given flare, I found the minimum and maximum and used the midway point as the center and the difference between the center and the min (or max) as the error. To find the minimum, I subtracted the gaussian standard deviation from bootstrapping (100 iterations) from the initial gaussian result. I did the same for each model, and the equivalent to find the maximum. The overall minimum was the minimum of the three values. I found centers and errors with and without the use of gaussian+line bootstrapping results, since those required extra constraints to cut out outliers. I checked the results from varying other parameters (offset, detrending, etc.) and very few values fell outside of the min-max ranges. Those values were often unreasonable or within a day of the range, so I kept this method.

Using the centers and errors for the 13 flares, I redid the analysis of the best period values for each epoch, which ended up being in agreement with the periodogram results (149, 171, and 200 days). I looked at both the sum of the squares of the O-C values as well as that value divided by the square of the error. Minimizing these values gave the best period in an epoch. The version using error switched the middle epoch result to 170 days. Once again, the chi-squared (including reduced) from three epochs was lower than the value from the linearly increasing period model.





The residuals for the first flare of the first and third epochs are set to 0, since those centers are used to predict the remaining flares in those epochs. The fourth flare, which is predicted by the 149 day period in the first epoch, is used for the predictions in the second epoch.

The first flare seems most affected by systematic error and the choice of the start and end point. Only the first flare's gaussian result changed significantly by using the method of fitting the symmetrical flare (so cutting off any extra data on a side so that the minimum mags are about equal at the start and end.