QUARTUS ENGINEERING

Starshade Data Challenge

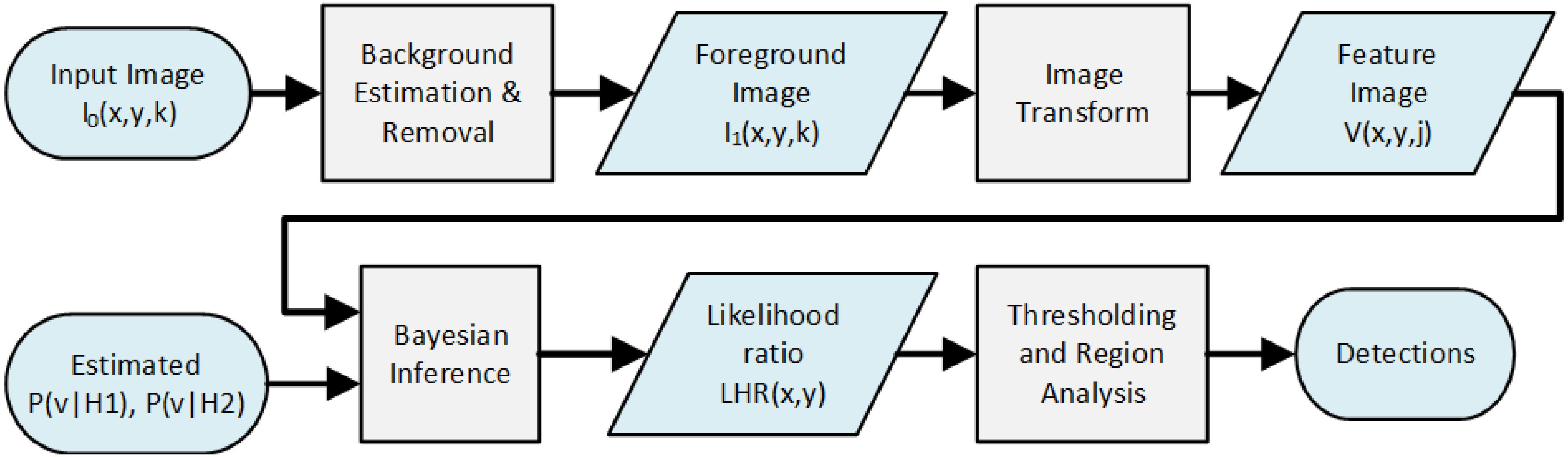
Release 2 update

Brian Dunne
brian.dunne@quartus.com

6/16/2021

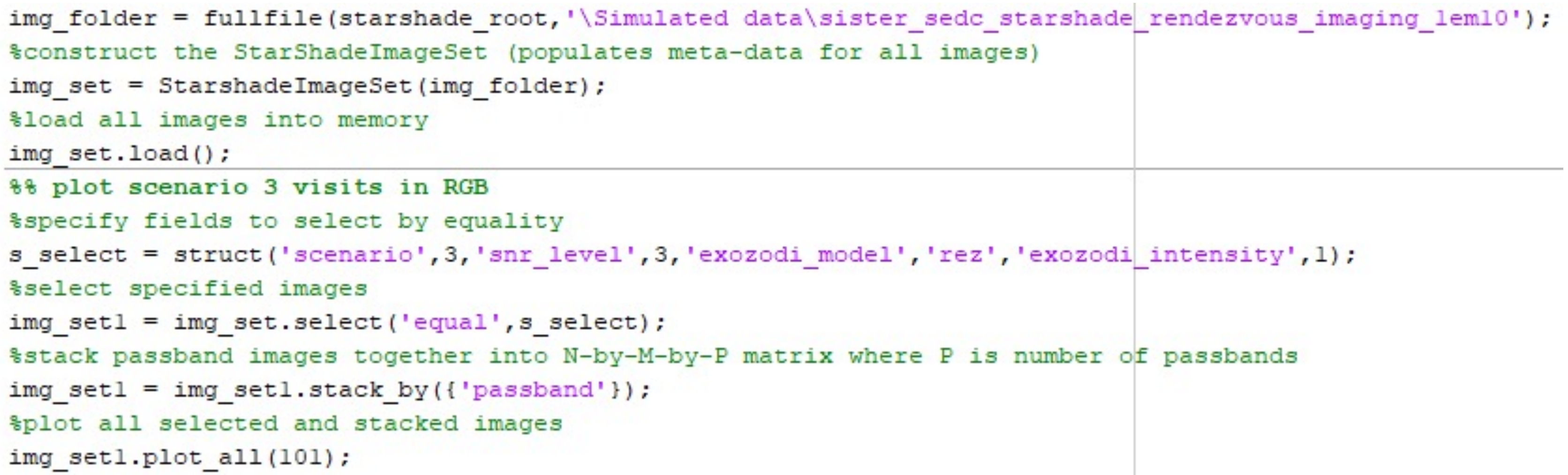
PROPOSED APPROACH

- The proposed approach includes three high level steps:
  1. Background estimation and removal.
  2. Transformation of the multispectral foreground image pixels into a feature space based on matched filtering.
  3. Perform Bayesian inference on the feature space to produce a likelihood ratio which can be thresholded for planet detection.
- The initial focus is on background estimation and removal, since backgrounds must be successfully estimated and removed before later steps in the pipeline can be used. Initial findings during the background estimation development may also lead to revised ideas for the later steps.

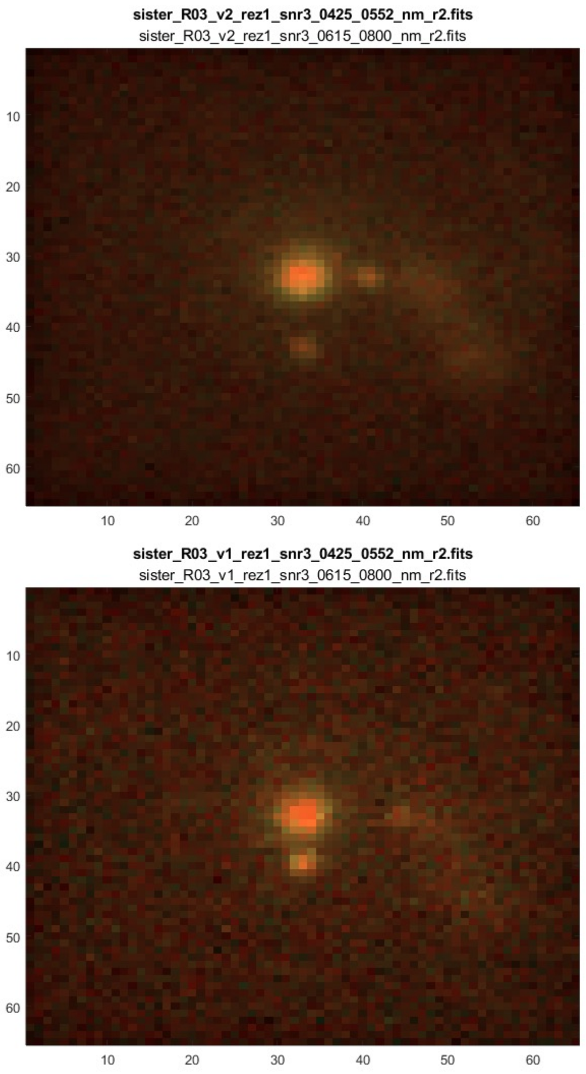

Input Image $I_0(x,y,k)$
Background Estimation & Removal
Foreground Image $I_1(x,y,k)$
Image Transform
Feature Image V(x,y,j)
Estimated P(v|H1), P(v|H2)
Bayesian Inference
Likelihood ratio LHR(x,y)
Thresholding and Region Analysis
Detections

DATA MANAGEMENT TOOLS

- Some library level tools were developed to organize image data and meta-data from the .fits files into a standard class, and provide methods to filter, reorganize, and plot the data.
- These tools are built around two classes:
  - StarShadeImage: encapsulates all the data associated with a single .fits file.
  - StarshadeImageSet: encapsulates all the data associated with a full simulation data release, including all per-image data and instrument level data.
    - Properties:
      - instrument_meta – all calibration data and instrument meta data that is constant across all images
      - images – an array of StarShadeImage
    - Methods
      - load() – Load the image set from .fits into memory.
      - select() – Selects a subset of the data based on any meta data filters, i.e. scenario, snr, etc.
      - unique() – Finds subsets that are unique in specified parameters, for example to group the data by scenario, SNR, etc.
      - stack_by() – Stacks M-by-N-by-1 images into M-by-N-by-P image stacks where P is the number of unique images for the specified parameters.
      - plot_all() – Plot the current set in a tiled layout.



```
img_folder = fullfile(starshade_root,'\Simulated data\sister_sedc_starshade_rendezvous_imaging_lem10');
%construct the StarShadeImageSet (populates meta-data for all images)
img_set = StarshadeImageSet(img_folder);
%load all images into memory
img_set.load();
%% plot scenario 3 visits in RGB
%specify fields to select by equality
s_select = struct('scenario',3,'snr_level',3,'exozodi_model','rez','exozodi_intensity',1);
%select specified images
img_set1 = img_set.select('equal',s_select);
%stack passband images together into N-by-M-by-P matrix where P is number of passbands
img_set1 = img_set1.stack_by({'passband'});
%plot all selected and stacked images
img_set1.plot_all(101);
```



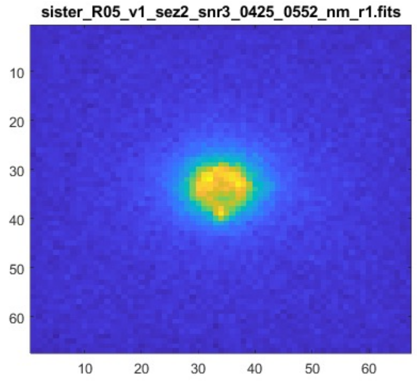
Example plot generated by:
StarshadeImageSet.plot_all()

Release 2 Data Observations

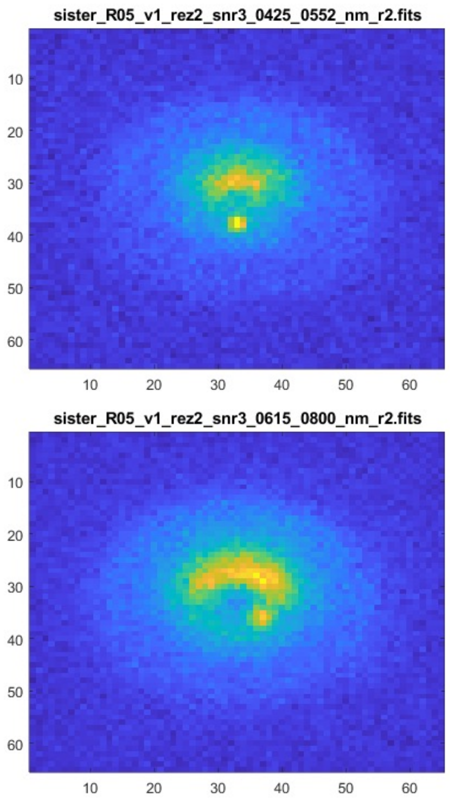
- The updated exozodi model including resonant structures significantly complicates the requirements for a successful parametric background model.
- The parametric background estimation approach demonstrated on release 1 data gives good results in some cases but is lacking some major effects, particularly forward scattering and resonant structures.
- While it seems feasible to continue the parametric model approach and explicitly fit a fairly complex model, some additional thought is warranted on whether a simpler approach might be successful.

Some general options for proceeding are:

1. Continue to add effects to the explicit parametric model demonstrated previously. This seems feasible but requires some work. The improved model would start to look more like an astrophysical model rather than a rough 'eyeballed' empirical model.
2. Try some nonparametric approaches that may be able to exploit additional information in the release 2 data such as multiple passbands and multiple visits per scenario. A nonparametric approach, if it worked, could be simpler and more general than an explicit model fit. The down side is that a nonparametric approach will not estimate disk parameters.
3. Some combination of a simple parametric model to estimate basic disk parameters, with a nonparametric model which can account for background effects not accounted for by the simple disk model.


sister_R05_v1_sez2_snr3_0425_0552_nm_r1.fits

Release 1 scenario 5 with smooth exozodi

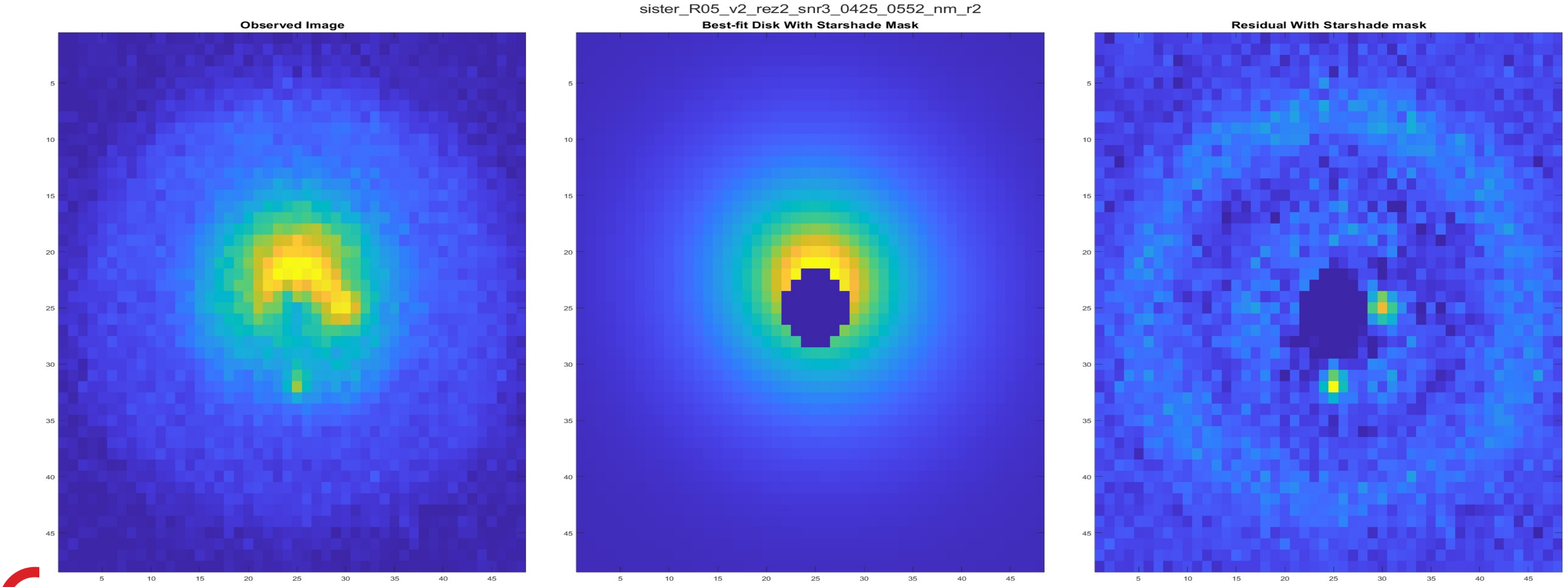

sister_R05_v1_rez2_snr3_0425_0552_nm_r2.fits

sister_R05_v1_rez2_snr3_0615_0800_nm_r2.fits

Release 1 scenario 5 with resonant exozodi

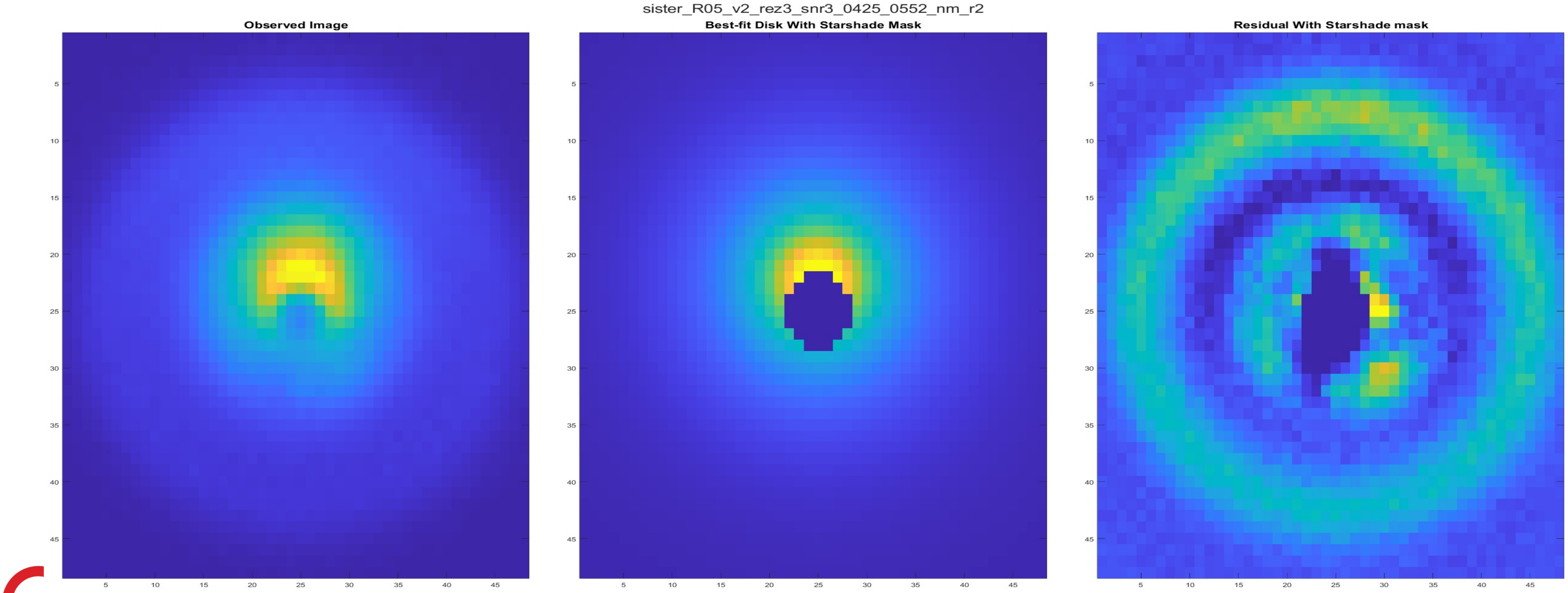
PARAMETRIC BACKGROUND SUBTRACTION

- The parametric background subtraction approach developed for release 1 data was tested on release 2 data without modification. Fairly good results are achieved in some circumstances. Results for scenario 5, with the resonant exozodi with intensity level 2 are shown below.(R05_v2_rez2_snr3) Resonant structures are apparent in the residual, but the signal from the planets is clear with the resonant structure component of the residual being relatively insignificant from the standpoint of planet detection.



Parametric Background Subtraction

- The parametric background subtraction approach starts to break down significantly for higher intensity resonant exozodi. Results are shown below for the same scenario as the previous slide, but with the exozodi intensity increased to level 3. In this case a similar residual error is present but scaled up in intensity such that background residual is of similar intensity to planet residual.



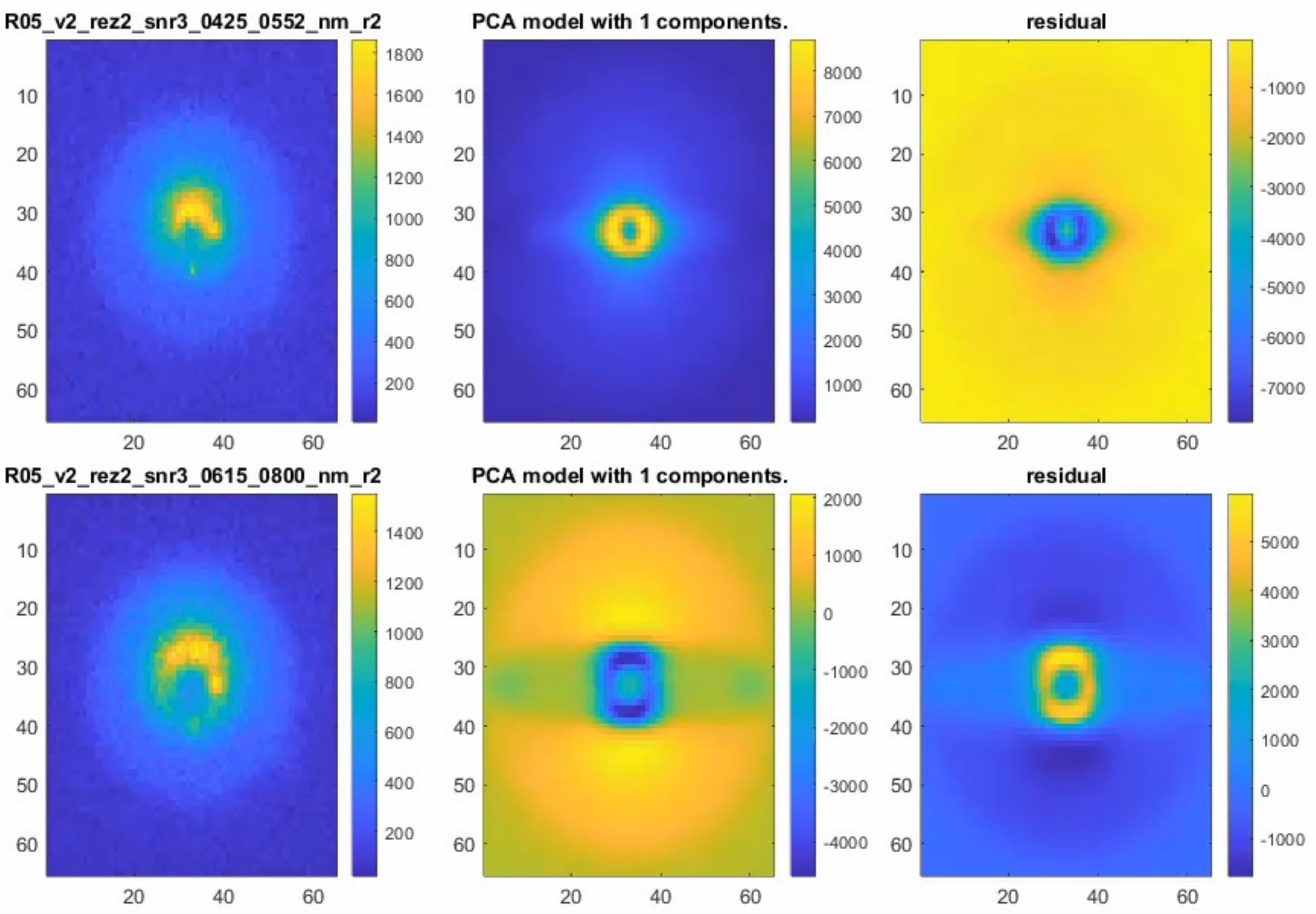
Background Subtraction Update Ideas

- With increased complexity of backgrounds due to resonant structures, and previous problems with unmodeled affects such as forward scattering, it seems that an ‘eyeballed’ empirical model of smooth exozodi is inadequate. In order to proceed with an explicit parametric model of the background it would be sensible to use something resembling more of an astrophysical model that explicitly models the dominant effects. Such a model could be simplified as appropriate to capture the dominant effects in the image with as few free parameters as possible.
- Before proceeding with an increasingly complex parametric model approach, it seems appropriate to try explore nonparametric approaches. In additional to featuring more complex backgrounds, the release 2 data also includes additional information that can be exploited to separate exoplanet signals from background. These additional aspects of the data are:
  - Multispectral images. The appearance of the exozodi seems to vary as a function of passband in a way that is different from how the exoplanet signals vary. The exozodi appearance tends to broaden with increasing wavelength for example,(perhaps a Mie scattering effect?) whereas the planets tend to appear as point sources in roughly the same position. This difference in how exozodi and exoplanets appear as a function of passband may be exploitable.
  - Multiple visits. Whereas the release 1 data only included a single visit per astrophysical scenario, the release two data provides two visits per scenario, all else being equal. Between visits the background appears to stay the same, whereas the exoplanets move. Similarly to the multispectral differences, the decoupling of how exoplanets change in time vs background can potentially be exploited.
- Some initial ideas for nonparametric background estimation approaches to try:
  1. Median filtering or other non-linear filtering approach. The idea is that the planets are small sharp signals that can be removed by a median filter or bilateral filter while the background structure is preserved. This approach was attempted briefly using median filtering and bilateral filtering. Median filtering gave some limited success but only in scenarios where exoplanets are already somewhat visible an significant relative to background.
  2. Principal Component Analysis (PCA). Because the nature of the variance of the exozodi and background content is different compared to how the planets vary, in that the planets change position between visits, and the exozodi tend to broaden with increasing passband, a principal component analysis approach may be expected to exploit these differences in variance. It is also the case that because the backgrounds dominate the signal energy in the image, the lower principal components in a PCA model will tend to model the backgrounds and not the planets. The planets may tend to be ignored by the lower components in the PCA similarly to how measurement noise is not incorporated.

PCA Background Model

- A PCA model was trained on the data in an attempt to learn a non-parametric model of the background.
- Quick explanation of PCA:
  - PCA takes N samples of input data(in this case a ‘stack’ of images) and finds ‘principal components’ of the samples, which represent an orthonormal basis set of the input data. The principal components are ordered by descending percent of variance explained.
  - PCA is often called a dimensionality reduction technique, in that it finds a representation of multivariate data samples in fewer dimensions.(latent variables) In this case we are finding a lower dimensionality representation of the stack of images, such that we can give an approximate representation of any image by taking linear combinations of only a few principal components.

- The video to the right shows the PCA model for an increasing number of components. This can be thought of as a cumulative sum of components, where the first few components represent most of the image intensity and later components add less significant details.
- Results are shown for R05_v2_rez2_snr3 for comparison with the previous slide.



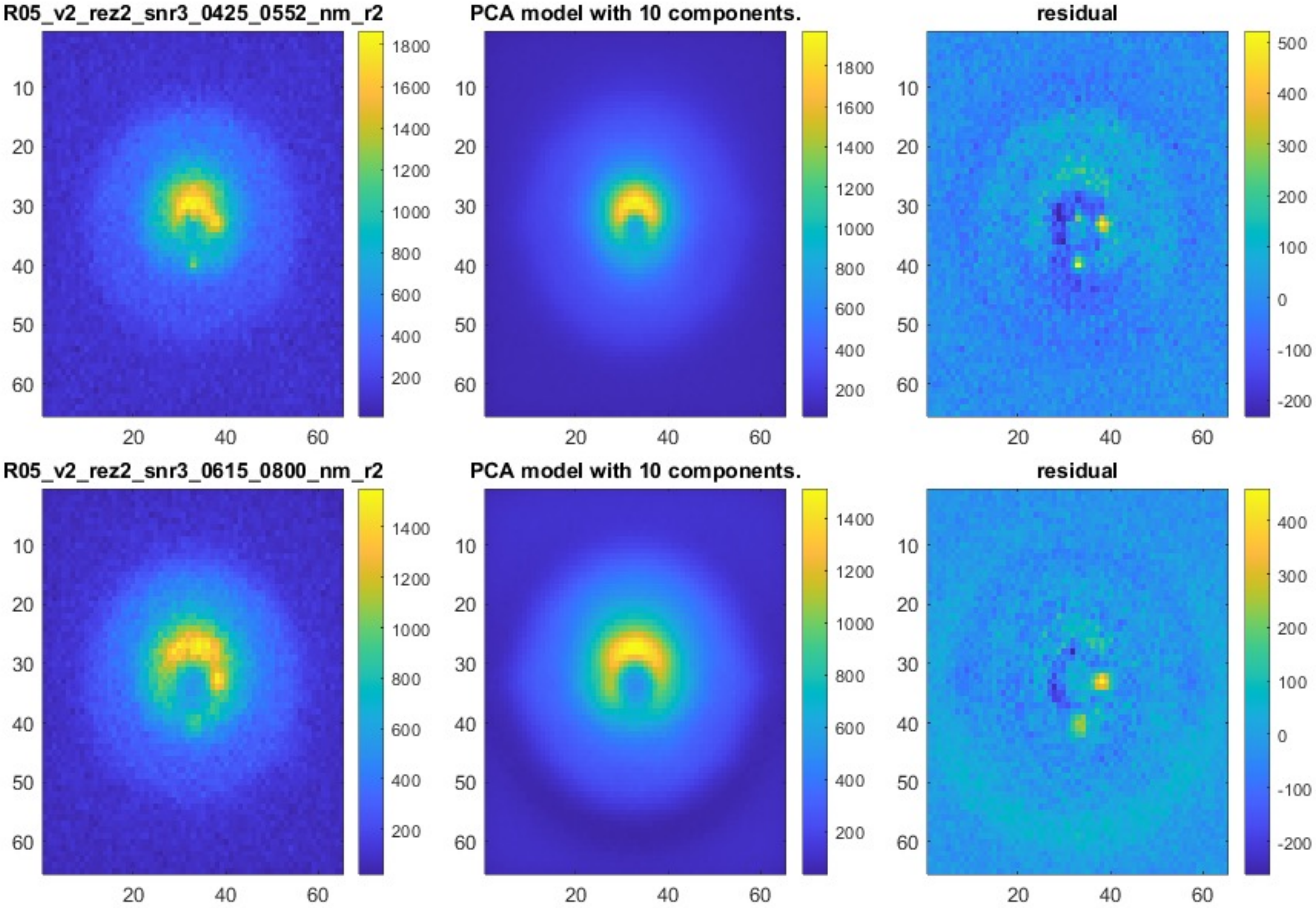
QUARTUS ENGINEERING

PCA Background model

- Example PCA model results are shown for R05_v2_rez2_snr3 for comparison with the previous slide. This shows a 10-component model which does a reasonable job of modelling background but not planets.

- Tuning to get this result included:
  - Data augmentation with 90 degree rotations.
  - Mean centering and normalization.
  - Combining or separating subsets, i.e. res vs sez combined or as separate models.
  - Choosing the number of principal components to use to model backgrounds but exclude planets.

- Performance on levels 1 and 2 exozodi intensity is generally reasonable but results are poor for level 3 exozodi intensity.



NEXT STEPS

- Look into related dimensionality reduction techniques that may be able to ‘learn’ low order background structure nonparametrically.
  - ICA : Similar to PCA but the components are statistically ‘independent’ and not necessarily orthogonal.
  - Autoencoders : The neural network analogue of PCA/ICA. Uses a neural network to find a low dimensional representation of a set of images.
- Think about how to combine a general nonparametric approach with a simple disk model to yield disk basic disk parameter estimates while accounting for unmodelled effects with the nonparametric component.
- Proceed from background detection to planet detection and parameter and spectral estimation.
  - Now that some background subtraction approaches that are basically functional have been established, the next steps in the processing pipeline can be developed. Background subtraction can be improved further if it is still the limiting factor after developing the downstream steps.