Demystifying Kepler Data: A Primer for Systematic Artifact Mitigation

K. Kinemuchi

Bay Area Environmental Research Institute & NASA-Ames Research Center

Mail Stop 244-30, Moffett Field, CA 94035

karen.kinemuchi@nasa.gov

T. Barclay

Bay Area Environmental Research Institute & NASA-Ames Research Center

Mail Stop 244-30, Moffett Field, CA 94035

thomas.barclay@nasa.gov

M. Fanelli

Bay Area Environmental Research Institute & NASA-Ames Research Center

Mail Stop 245-1, Moffett Field, CA 94035

michael.n.fanelli@nasa.gov

J. Pepper

Vanderbilt University

Department of Physics and Astronomy, Nashville, TN 37235

joshua.pepper.v@gmail.com

M. Still

Bay Area Environmental Research Institute & NASA-Ames Research Center

Mail Stop 244-30, Moffett Field, CA 94035

martin.still@nasa.gov

and

Steve B. Howell

time-series photometry (Borucki et al. 2010; Koch et al. 2010; Caldwell et al. 2010). Transit durations typically last a few hours, and they are separated by intervals of days to years. The Kepler mission collects data mostly on a 30-minute cadence near-continuously with > 92% completeness. Detection of transits by small planets requires parts-per-million photometric precision (Jenkins et al. 2002). The primary objective will be realized through the combination of space-based photometric sensitivity, regular observing cadence, a 116 square degree field-of-view containing a large number of target stars, and high duty cycle (Koch et al. 2010).

One of the primary Kepler legacies is a community archive containing the multi-year, time-series photometry of 1.5×10^5 astrophysical targets observed for the planetary survey (Batalha et al. 2010) and community-nominated targets of other astrophysical interest. In addition to exoplanet science, Kepler provides unique datasets and raises scientific potential in fields such as asteroseismology (e.g. Bedding et al. 2011; Chaplin et al. 2011; Antoci et al. 2011; Beck et al. 2011), gyrochronology (Meibom et al. 2011) , stellar activity (e.g. Basri et al. 2011; Walkowicz et al. 2011), binary stars (e.g. Carter et al. 2011; Derekas et al. 2011; Slawson et al. 2011; Thompson et al. 2012), and active galactic nuclei (Mushotzky et al. 2011). Our aim for this paper is that it is used as a reference guide for the astronomical community who wishes to use Kepler data. This paper will address how the community can mitigate and exploit the public data for stellar and extragalactic science. Many of the techniques presented here are applied to archived light curves and target pixel files and will help optimize the data for astrophysical research.

An understanding of the nature of archived data is critical for the effective exploitation of the Kepler legacy. Kepler provides high-precision photometry on the 1 and 30 minute timescales. The data also contain artifacts that occur through spacecraft operation events and systematic trends over longer timescales as a natural consequence of mission design (Van Cleve & Caldwell 2009; Christiansen et al. 2011). Artifacts can both mask astrophysical signal and be misinterpreted as astrophysical in origin (Christiansen et al. 2011). Furthermore, cavalier approaches to artifact mitigation (e.g. using simple function fits to time-series data) can destroy astrophysical signal. Archived Kepler data have been processed through a data reduction pipeline (Jenkins et al. 2010b). The pipeline functions include pixel-level calibration (Quintana et al. 2010), simple aperture photometry (Twicken et al. 2010a) and artifact mitigation (Twicken et al. 2010b). This third step of artifact removal will always be a subjective process. An archive user can either work with the default pipeline correction to be of suitable quality to enable their scientific objectives or choose to perform artifact removal themselves, starting from either the calibrated pixels or aperture photometry.

In this paper we provide a guide to Kepler data, insight into the nature of systematic

6.54 s. Exposures are summed onboard and stored at either long cadence (29.4 min; see Jenkins et al. (2010a)) or short cadence (58.85 s; see Gilliland et al. (2010)). Science data are downloaded approximately once per month when Kepler leaves the field temporarily to point its high-gain antenna towards Earth (Haas et al. 2010). The number of pixels collected and transmitted is a trade-off between maximizing the number of targets delivered and minimizing the length of the data gaps. For long cadence observations, a maximum of 5.4×10^6 pixels are stored onboard in the spacecraft Solid State Recorder (Jenkins et al. 2010b), and the number of targets can range from 150,000 to 170,000. Short cadence data are limited to 512 targets. Download requires approximately a 24-hour hiatus in data collection.

The long and short cadence pixels equate to less than 6% of the detector plane, and the remaining pixel data are not stored. The stored pixels are chosen strategically to provide postage stamp images centered on the positions of Kepler targets (Batalha et al. 2010). The size of a postage stamp increases with target brightness, and the yield lies typically between 163,000–170,000 targets per month. The critical concept for understanding instrumental artifacts is that in order to maximize the number of targets collected, the postage stamp sizes and shapes are chosen to maximize the per-target photometric signal-to-noise on the 3–12 hour timescales of exoplanet transits. The postage stamps do not contain all the flux from a target because the collection of the target's faint PSF wings degrades signal-to-noise by the inclusion of more sky background. The pixels within a postage stamp are defined by a calculation that combines the photometry and astrometry within the Kepler Input Catalog (KIC; Brown et al. 2011) and an analytical pixel response model for the detector and optics (Bryson et al. 2010a). The postage stamps are fixed within the pixel array; they do not evolve over a 93-day observation period, or "quarter". A new target list is uploaded to the spacecraft after each quarterly roll. Changes in the target list occur due to new detector geometry, operational developments to the exoplanet survey, and community-led science programs. Any time-dependent variation in the position of the target or the size of the point spread function will result in a redistribution of flux within the postage stamp pixels. The spacecraft pointing stability is good to typically 20 milliarcsec over 6.5 hours but the high precision light curves can contain systematic noise that manifests from thermally-driven focus variations, pointing offsets, and differential velocity aberration (Christiansen et al. 2011; Van Cleve & Caldwell 2009). Many of the systematics within the archived light curves are the result of time-dependent light losses as the target wings fall out of the pixel apertures and time-dependent contamination by neighboring sources falling into the pixel apertures. All collected data are stored and propagated to the community by the MAST. Technical manuals and reference material describing the mission, its data products and defining mission-specific terminology are also archived at the MAST.

- **Kepler Instrument Handbook** - describes the design, operation, and in-flight per-

A more detailed examination of these files and the consequences of the systematic effects within the TPFs are provided in Section 4.

3. Kepler Archived Target Pixel Files (TPF)

For each individual Kepler target, light curve files (discussed in Section 4) are accompanied by a TPF. The TPF is the single-most informative resource in the archive for understanding the instrumental, non-astrophysical features within a target light curve. Consequently, it is recommended that users always examine a TPF in conjunction with its light curve. They provide the detector pixel and celestial coordinate mapping of a target mask and its subset containing the optimal aperture. TPFs reveal the motion of a target across the optimal aperture, and the motion of nearby contaminating sources.

As with the light curve file, the TPF content is organized by timestamp. The timestamps are the barycentric Julian date at the midpoint of each accumulated exposure. Pixel data are presented as a time-series of photometrically calibrated images, one image per timestamp, all located within the first extension of the FITS file. Typical Kepler targets ($K_p > 12$) will contain 10–50 pixels, but bright sources will include a larger pixel set. A typical quarter will contain approximately 4,300 collected images in long cadence for each target, while a typical month will yield 43,200 images in short cadence mode. For each timestamp within the TPF, there is an image of the uncalibrated pixels collected around a target. This group of pixels is referred to as the *target mask*, which contain pixels assigned either to the optimal aperture or a halo around it. The *optimal aperture* is the set of pixels over which the collected flux is summed by the Kepler pipeline to produce a light curve. The *halo pixels* are those pixels surrounding the optimal aperture pixels, which are used for calibration purposes and provide operational margin.

Additionally, for each timestamp, the TPF includes the raw counts (FITS column labeled "RAW_CNTS") a fully calibrated postage stamp pixel image (FLUX) which incorporates bias correction, dark subtraction, flat fielding, cosmic ray removal, gain and nonlinearity corrections, smear correction (the Kepler photometer has no shutter), and local detector electronic undershoot (i.e. sensitivity of the pixel response to bright objects). The Data Processing Handbook (Fanelli et al. 2011) and Caldwell et al. (2010) contain further details of these corrections. The TPFs also provide images for each timestamp containing the 1-σ flux uncertainties of each pixel (FLUX_ERR), a calibrated sky background (FLUX_BKG), 1-σ uncertainties to the sky background (FLUX_BKG_ERR), and cosmic rays incidences (COSMIC_RAYS). All of these postage stamp images are found in the first extension of the FITS file. Sky background generally cannot be well-estimated from the TPFs themselves be-

within the TPF (the FITS columns labelled POS_CORR1 and POS_CORR2). Predictions of the target motion trace many of the Kepler systematics, and this measured astrometric deviation from the reference star predictions are likely to be astrophysical in nature. For example, as a target's brightness varies, the centroid of the flux distribution across the pixels will move if there are contaminating sources in the target mask's pixels. The change in the flux centroid position is a method that can help detect faint, unresolved background binaries or other variable stars, which can appear as a false detection of planetary transit. (e.g. see Appendix A).

Figure 1 provides a typical example of a Kepler long cadence light curve, specifically the quarter 2 simple aperture photometry of the eclipsing binary star V1950 Cyg. In this example, the flux affected by short-term systematics are dominated by the high-amplitude intrinsic variability of the target, but the effects of DVA still clearly manifest as a 3% decrease in target flux over the duration of the quarter. The situation is clarified by inspection of the associated TPF. Figure 2 reconstructs the photometry for each individual pixel within the target mask of the source of interest. The mask includes both the star itself and the halo pixels around the star. The pixels in Figure 2 with gray backgrounds are chosen by the Kepler pipeline to be the photometric optimal aperture for the archived light curve. The other pixels in the halo are associated with the target mask but they play no part in the calculation of the archived time-series photometry. The wings of the point spread function extend into many pixels surrounding the optimal aperture. Kepler apertures, in general, are designed to maximize signal-to-noise which usually undercaptures target flux. Without full capture of a source, motion and focus-induced artifacts in the summed light curve are inevitable. In Figure 2, as the quarter proceeds, note how the flux captured in each pixel can increase or decrease as the target moves across the pixel array. This example shows that the flux is continuously being redistributed between neighboring pixels of the optimal aperture, and different amounts of flux will fall outside this aperture over time. The times and nature of events related to the discontinuities seen in the TPF light curves are recorded under the QUALITY column in the TPF FITS headers.

The example of V1950 Cyg illustrates how Kepler light curves can easily be affected by instrument systematics (e.g., DVA). There are instances where extracting target pixels over a larger detector area compared to the optimal aperture chosen by the pipeline minimizes or removes the instrument effects from the light curves. However, the artifact mitigation of including more pixels comes at a cost. More pixels within the flux summation decreases the photometric signal-to-noise by adding more sky background from the additional pixels. The user must decide whether adding more pixels is an acceptable mitigation for the systematics. Many Kepler targets exist within crowded fields, and multiple nearby sources may contribute to the flux within the target mask. While it is sometimes beneficial to include more pixels

users need to be aware that a SAP light curve can be contaminated by astrophysics from neighboring sources. One can inspect the concurrent TPF to identify contamination. A new SAP light curve can be extracted from the TPF using a custom selection of pixels, as shown in Appendix B.

Archive users must expect, *a posteriori*, that SAP photometry is contaminated by the systematics discussed in Section 3. To continue using SAP data for scientific exploitation, one must decide whether the artifacts will impact their results and conclusions. There is "low-hanging fruit" that has dominated Kepler astrophysics activity in the early phases of the mission because the SAP data has proved to be adequate for specific science goals without artifact mitigation. For example, asteroseismology of solar-like oscillations, δ Scuti and γ Doradus pulsations have been hugely successful because signals of frequency $\gtrsim$ 1-d^{-1} are mostly unaffected by the majority of artifacts (Balona & Dziembowski 2011; Uytterhoeven et al. 2011; Balona et al. 2011). Some high frequency artifacts that could prove problematic to these programs can be filtered out of the time-series using the quality flags provided. Data analysis of cataclysmic variables, RR Lyr stars and Cepheids are just as successful. While many of the astrophysical frequencies of interest in these pulsators can be longer than a few days and similar to the thermal resettling times of the spacecraft after a pointing maneuver, the large amplitude of target variability dominate over systematics that can consequently be neglected (e.g. Still et al. (2010); Nemec et al. (2011); Szabó et al. (2011)).

There are many astrophysical applications that are less likely to benefit from direct employment of SAP data. These include any science relying on more subtle light curve structures and periods longer than a few days, in which case the systematics discussed are more likely to be significant. Investigations of magnetic activity, gyrochronology, binary stars and long period variables must scrutinize the SAP data with great care before proceeding and will most likely benefit from one of three available artifact mitigation methods. The three methods are to use archived PDCSAP photometry (Section 4.2), to re-extract the SAP light curve over a larger set of pixels (Appendix B), or to perform a custom correction on the archived SAP data using cotrending basis vectors (Appendix C). These methods and their precise application are very subjective. The highest quality Kepler research will in most cases result from the experience and understanding gained by applying all three of these methods to the archived data.

4.2. Pre-search data conditioning simple aperture photometry (PDCSAP)

The PDCSAP data included within the archived light curve files are produced by a pipeline module that remains under continuing development at the time of writing (Smith

time-series. If artifact mitigation is required subsequent to light curve extraction, then the only viable option is to manually fit the CBVs. Manual CBV fitting and subtraction is the subject of Section 5.

5. Removing systematic artifacts with cotrending basis vectors

The 16 most significant CBVs per channel for each quarter are calculated by the Kepler pipeline and packaged as FITS binary files. These CBVs are available for download at the MAST[3]. File content and format are described in Fraquelli & Thompson (2011). With the provision of CBVs, the responsibility rests with the archive user to either improve upon the existing artifact mitigation done by the pipeline or perform manual artifact removal from photometry re-extracted from a TPF. Basis vectors are usually fit to the SAP light curve linearly, i.e. each basis vector is scaled by a coefficient and subtracted from the flux time-series. Computationally, the most efficient method is a linear least-squares fit. In Figure 4 we plot 16 SAP light curves from channel 50 in quarter 5. In Figure 5, we plot the same light curves of Figure 4 but with the most significant CBVs fit and subtracted. We can see that in all the light curves in Figure 5, the systematic trends have been greatly reduced and the astrophysics is more clearly delivered. Appendix C provides an example of how to apply the CBVs to data.

An important decision for the CBV user is how many basis vectors to fit and remove from the SAP data. Fitting too few will capture instrumental artifacts less effectively. However, using too many can overfit the data, removing real astrophysical features. A further consideration is that no basis vector is perfect. The inclusion of each additional CBV to the fit adds a noise component to the data. The choice of CBV number is in reality a trade between maximizing the removal of systematics on the one hand, and avoiding the removal of real astrophysics and minimizing the effects of CBV noise on the other. A minimum of two basis vectors should be fit to the data because, instead of strictly enforcing a constant first or second basis vector, CBVs are created by mixing a constant offset with the strongest non-constant basis vector. We find that an interative method is the most effective approach, starting with two basis vectors and increasing the number of vectors monotonically until deciding upon a subjective optimal fit. Appendix C shows example fits of 2, 5, and 8 CBVs to demonstrate the iterative method of determining the number of CBVs to use.

Occasionally the linear least-squares fit is not sufficient, and a more robust fitting method must be utilized. One option is to fit the CBVs to the SAP time-series using

[3]http://archive.stsci.edu/kepler/cbv.html

6.1. Suggested stitching methods

There are several methods available to attempt correcting for photometric discontinuities across the quarter gaps. None of the methods are well-suited to all occasions. Individually they can perform good corrections under specific conditions. We discuss the three methods below.

6.1.1. Crowding and aperture flux loss adjustment

The first method is to align different quarters using the time-invariant approximations for crowding and aperture flux losses stored within the light curve FITS file keywords. These unitless keywords are stored in all Kepler data processed after September 2011. For earlier data the same quantities are populated in the meta-tables of the data search and retrieval page at MAST. The FITS keyword FLFRCSAP contains the fraction of target flux falling within the optimal aperture. The keyword CROWDSAP contains the ratio of target flux relative to flux from all sources within the photometric aperture. Both quantities are the average value over each quarter and are estimated using point-spread function and spacecraft jitter models (Bryson et al. 2010b) combined with source characteristics found within the KIC (Brown et al. 2011). The PDC time-series data archived within the FITS light curve files have both of these corrections applied by default. Both corrections can be applied to SAP data manually using the *keparith* task within the PyKE package (see Appendices for a description of PyKE).

The limitations of this first method are two-fold. The corrective factors supplied are model-dependent. The characterization of the Kepler point-spread function does not provide the same order of photometric precision as aperture photometry. Furthermore, PSF modeling of the pixels within an optimal aperture is only as complete as the KIC, which is complete only at $K_P < 17$. Secondly, the corrective factors are averaged over time, whereas in reality they vary from timestamp to timestamp as the field moves within the aperture. Furthermore, as the target and neighboring stars vary independently, the crowding correction in reality is an additive one rather than multiplicative. Therefore, while these two keywords collectively provide the simplest method of quarter stitching, the solution is often inadequate. Given that target images for each timestamp are provided in the archive within the Target Pixel Files, there is some scope for Kepler users to improve upon these corrections by fitting a field and point-spread function model to the individual images within these files. The Pixel Response Function (the combination of point-spread function and spacecraft jitter) information is available at the MAST archive within the focal plane characteristics download

ues to increase with time. Several approaches have been developed to manually improve the fidelity of aperture photometry. In order to minimize the impact of aperture photometry artifacts and obtain the highest quality time-series data, archive users are recommended to explore all of the methods discussed in this paper. The most correct approach for individual targets will be subjective and will be achieved through experimentation and hands-on experience.

To date, exploitation of Kepler data has naturally been dominated by its primary mission goals of exoplanet transit detection (Borucki et al. 2010) and asteroseismology of solar-like stars (Chaplin et al. 2011). By the nature of the mission design, both research areas continue reaping a rich harvest from the Kepler archive. However, there is a sensitivity threshold for both of these disciplines that will require more state-of-the art artifact mitigation before reaching their full potential. Similarly Kepler has the potential through multi-year, highly-regular monitoring to provide a startling legacy archive for active galactic nuclei, stellar activity, gyrochronology, and, perhaps most-compellingly, stellar cycles, for example. For much of the detailed astrophysics with Kepler data, and for Kepler to reach its broader scientific potential, the challenges of removing aperture photometry systematics must be met.

Acknowledgements

Funding for the Kepler mission is provided by NASA's Science Mission Directorate. The Kepler Guest Observer Office is funded through NASA co-operative agreement NNX09AR96A. All of the data presented in this paper were obtained from the Multimission Archive at the Space Telescope Science Institute (MAST). STScI is operated by the Association of Universities for Research in Astronomy, Inc., under NASA contract NAS5-26555. Support for MAST for non-HST data is provided by the NASA Office of Space Science via grant NNX09AF08G and by other grants and contracts.

Appendices: Data analysis examples

PyKE[5] is a suite of python software tools developed to reduce and analyze Kepler light curves, TPFs, and FFIs. PyKE was developed as an add-on package to PyRAF - a python wrapper for IRAF[6] that provides for a new tool development to occur entirely in the python

[5]http://keplergo.arc.nasa.gov/PyKE.shtml

[6]http://iraf.noao.edu

This object shows regular, low amplitude dips in brightness every 4.9 days that, at face value, are suggestive of a planetary transit of the target star. Figure 9 shows a calibrated flux time series of each target mask pixel collected over Q3. This figure was produced with the PyKE task *keppixseries*:

```
keppixseries infile=kplr002449074-2009350155506_lpd-targ.fits
outfile=keppixseries.fits plotfile=keppixseries.png plottype=local filter=n
```

Figure 9 reveals unambiguously that the target star is not the source of the "transit" features. A background eclipsing binary star is situated 10 arcsec from the target star (2.5 pixels to the left of KIC 2449074 on the figure) and is leaking into the optimal aperture. By extracting the light curve manually using different pixels, one can either reduce the contaminating flux from the eclipsing binary or, alternatively, extract flux from the eclipsing binary. New mask files are created interactively using the *kepmask* tool:

```
kepmask infile=kplr002449074-2009350155506_lpd-targ.fits maskfile=mask1.txt
plotfile=kepmask.png tabrow=2177 iscale=linear cmap=bone
kepmask infile=kplr002449074-2009350155506_lpd-targ.fits maskfile=mask2.txt
plotfile=kepmask.png tabrow=2177 iscale=linear cmap=bone
```

The image associated with the 2,177th timestamp in the Target Pixel File is plotted over a linear intensity scale using the *bone color* lookup table. Users of the *kepmask* tool define a new aperture interactively by moving the mouse over a pixel. One press of the 'X' keyboard key selects a pixel for inclusion within the new aperture, a second press deselects the pixel. The aperture is stored by clicking the 'DUMP' button on the interactive GUI. The task *kepextract* is employed to sum the pixels without weights within the newly defined aperture:

In each case a new mask was defined interactively using the method described in Appendix B. Figure 10 plots the mask stored in the file mask1.txt, while in Figure 11 plots the mask stored in file mask2.txt. The commands used to extract new SAP light curve from the TPF are:

```
kepextract infile=kplr002449074-2009350155506_lpd-targ.fits maskfile=mask1.txt
outfile=kepextract1.fits
kepextract infile=kplr002449074-2009350155506_lpd-targ.fits maskfile=mask2.txt
outfile=kepextract2.fits
```

The photometric aperture that yielded the light curve in Figure 14a and the individual pixel photometry are provided in Figure 15. This image can be replicated using the PyKE task *keppixseries*:

```
keppixseries infile=kplr008703536-2009259160929_lpd-targ.fits
outfile=keppixseries.fits plotfile=keppixseries.png plottype=global
filter=n
```

The file kplr008703536-2009259160929_lpd-targ.fits is the archived Target Pixel File coupled to the archived light curve. Two new files will be created – keppixseries.fits contains the tabulated data for each individual pixel light curve, and keppixseries.png contains the requested plot. The parameter *plot type=global* requests that all light curves are plotted on the same photometric scale and the data plotted are not bandpass filtered to reduce the low frequency effects of differential velocity aberration.

Our task is to reduce the systematic artifacts by extracting a new light curve from a more strategic choice of pixel aperture. The righthand panel of Figure 15 contains the pixel image from one specific timestamp in the Quarter 2 Target Pixel File of KIC 8703536. The green, transparent pixels represent a new custom photometric aperture defined using the PyKE task *kepmask*. An interactive image is called with the command:

```
kepmask infile=kplr008703536-2009259160929_lpd-targ.fits maskfile=mask.txt
plotfile=kepmask.png tabrow=2177 iscale=linear cmap=bone
```

As described in Appendix A, in this example, the new aperture is stored in a filed called mask.txt, and the task *kepextract* is called for this target pixel file.

```
kepextract infile=kplr008703536-2009259160929_lpd-targ.fits maskfile=mask.txt
outfile=kepextract.fits
```

The light curve in the lower panel of Figure 14 is the target data re-extracted from the new aperture and plot, as before with *kepdraw*:

```
kepdraw infile=kepextract.fits outfile=kepextract.png datacol=SAP_FLUX
ploterr=n errcol=SAP_FLUX_ERR quality=y
```

the time-series. We will strive to obtain a correction where the heights of each event are identical. Performing another fit using five basis vectors with the following command yields the result shown in Figure 18:

```
kepcotrend infile=kplr003749404-2009350155506_llc.fits
outfile=kplr003749404-2009350155506_cbv.fits
cbvfile=kplr2009350155506-q03-d04_lcbv.fits vectors='1 2 3 4 5' method=llsq
iterate=n
```

This new result is a qualitative improvement compared to the two-CBV fit, but the solution is still not optimal. We would like to improve the fit to the thermal event at BJD 2,455,156.5 and also further flatten the structure occurring after BJD 2,455,145. We fit the SAP data again, this time using eight basis vectors. The plot is shown in Figure 19 but the result appears to be less optimal than the 5 basis vector fit. Anomalous structure has very likely been added to the time series by the CBVs. One possible reason for the less than optimal solution is that eight basis vectors are over-fitting the periodic brightenings and adding new systematic noise to the intervals between them. To test this hypothesis we masked out light curve segments containing the large amplitude brightenings and fit the CBV to the remaining data. We employed the PyKE task *keprange* to define the masked regions in the time series:

```
keprange infile=kplr003749404-2009350155506_llc.fits outfile=keprange.txt
column=SAP_FLUX
```

This will plot the SAP_FLUX column data within the light curve file over time. Ranges in time can be defined by selecting start and stop times with the mouse and 'X' keyboard key. We masked four ranges in this example, as illustrated in Figure 20, and these ranges will be saved to a text file after clicking the 'SAVE' button on the interactive GUI. We performed the eight basis vector fit one last time, excluding from the fit the regions defined in Figure 20, again using the *kepcotrend task*:

```
kepcotrend infile=kplr003749404-2009350155506_llc.fits
outfile=kplr003749404-2009350155506_cbv.fits
cbvfile=kplr2009350155506-q03-d04 lcbv.fits vectors='1 2 3 4 5 6 7 8'
method=llsq iterate=n maskfile=keprange.txt
```

In Figure 21 we see the final version of the light curve. Systematic effects still remain, e.g. the thermal settling event is not totally removed but, subjectively, the data are much

Gilliland, R. L., et al. 2010, ApJ, 713, L160

Haas, M. R., et al. 2010, ApJ, 713, L115

Horne, J. D. 2008, Society for Astronomical Sciences Annual Symposium, 27, 55

Jenkins, J. M., Caldwell, D. A., & Borucki, W. J. 2002, ApJ, 564, 495

Jenkins, J. M., et al. 2010a, ApJ, 713, L120

—. 2010b, ApJ, 713, L87

Koch, D. G., et al. 2010, ApJ, 713, L79

Meibom, S., et al. 2011, ApJ, 733, L9

Mushotzky, R. F., Edelson, R., Baumgartner, W., & Gandhi, P. 2011, ApJ, 743, L12

Nemec, J. M., et al. 2011, MNRAS, 417, 1022

Oke, J. B. 1974, ApJS, 27, 21

Pence, W. D., Chiappetti, L., Page, C. G., Shaw, R. A., & Stobie, E. 2010, A&A, 524, A42

Quintana, E. V., et al. 2010, Proc. SPIE, 7740, 77401X

Slawson, R. W., et al. 2011, AJ, 142, 160

Smith, J. C., et al. 2012, ArXiv e-prints

Still, M., Howell, S. B., Wood, M. A., Cannizzo, J. K., & Smale, A. P. 2010, ApJ, 717, L113

Stumpe, M. C., et al. 2012, ArXiv e-prints

Szabó, R., et al. 2011, MNRAS, 413, 2709

Thompson, S. E., et al. 2012, ArXiv e-prints

Twicken, J. D., et al. 2010a, Proc. SPIE, 7740, 774023

—. 2010b, Proc. SPIE, 7740, 77401U

Uytterhoeven, K., et al. 2011, A&A, 534, A125

Van Cleve, J. E., & Caldwell, D. A. 2009, Kepler Instrument Handbook, Tech. rep., NASA-Ames Research Center

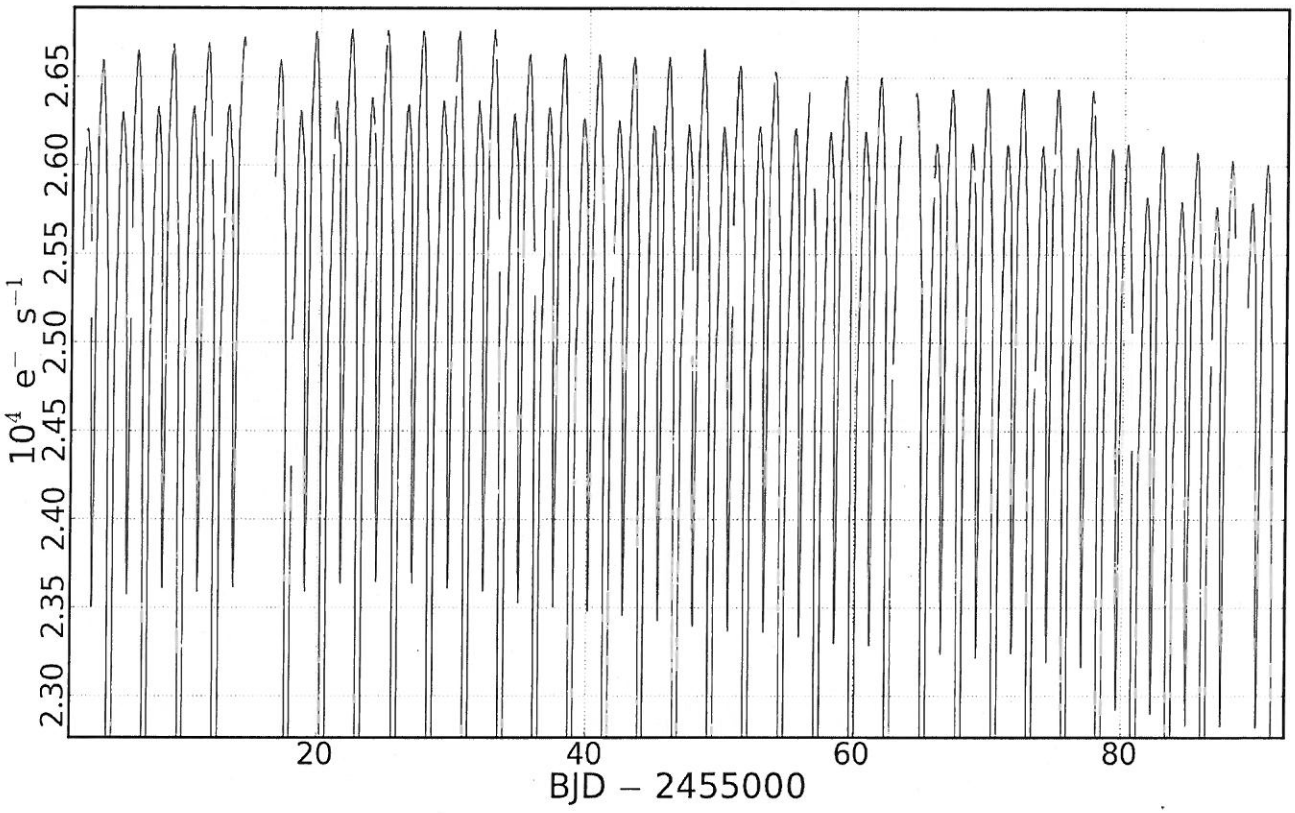

Fig. 1.— Quarter 2 long cadence SAP light curve of the eclipsing binary star V1950 Cyg (KIC 12164751; Horne (2008)), produced by the PyKE tool *kepdraw*. The most-prominent systematic effect in this light curve is the long-term decay in flux from the target which falls by 3% over the duration of the quarter. This drop is a consequence of differential velocity aberration pushing the under-captured target position across the fixed pixel aperture over time.

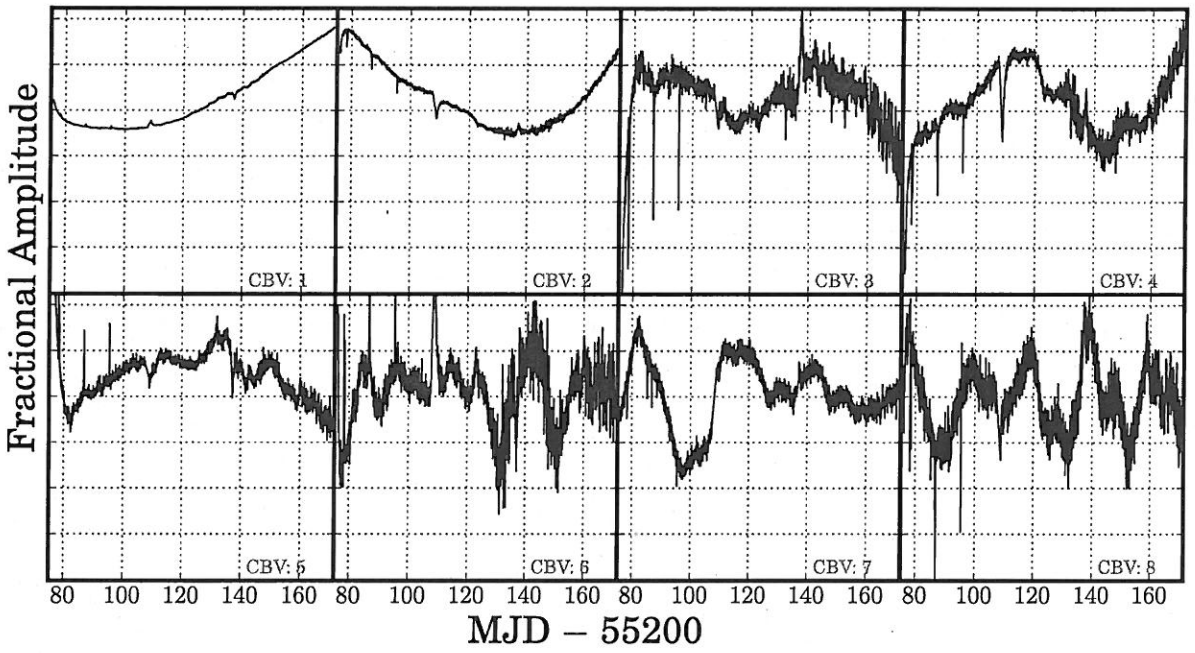

Fig. 3.— An example of eight cotrending basis vectors with the highest principle values, or contribution to systematic variability, from channel 50 over operational quarter 5. Each basis vector is on the same relative flux scale, centered about 0.0 e^- s^{-1}. Basis vectors can be linearly-fit to a light curve and subtracted off to mitigate for systematic effects. The fit coefficients can either be positive or negative. They run from left-to-right, top-to-bottom, in order of significance.

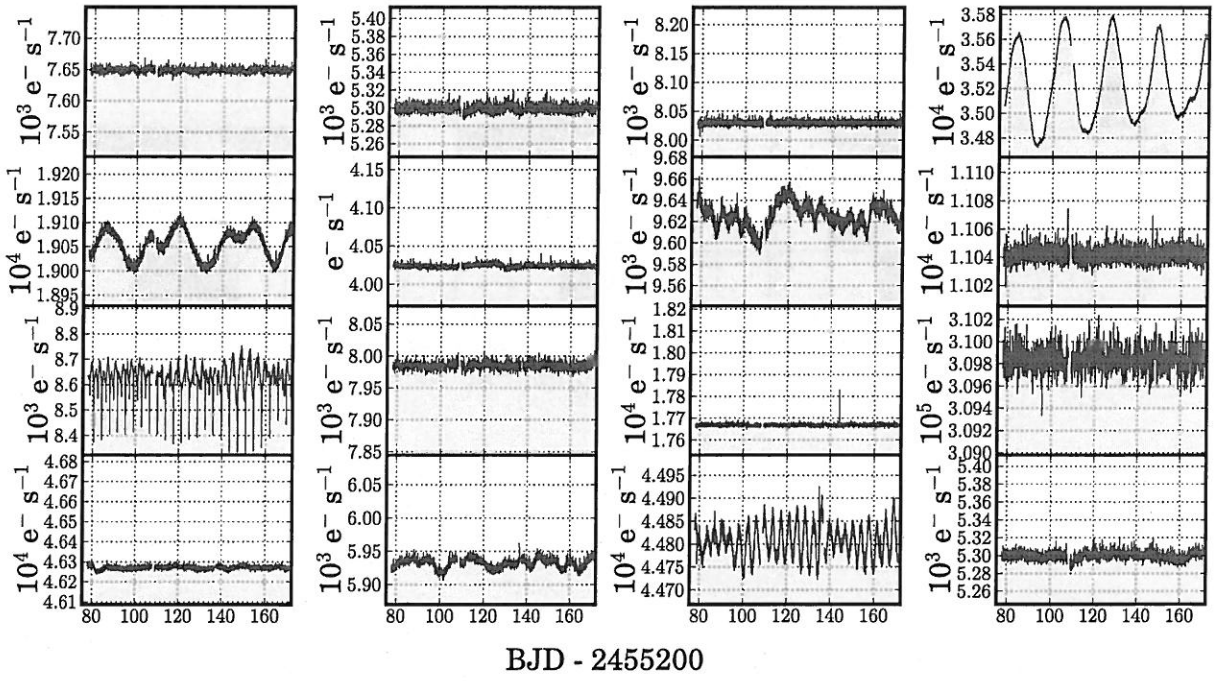

Fig. 5.— The sixteen quarter 5 SAP light curves presented in Figure 4 after the best-fit CBV ensemble has been subtracted.

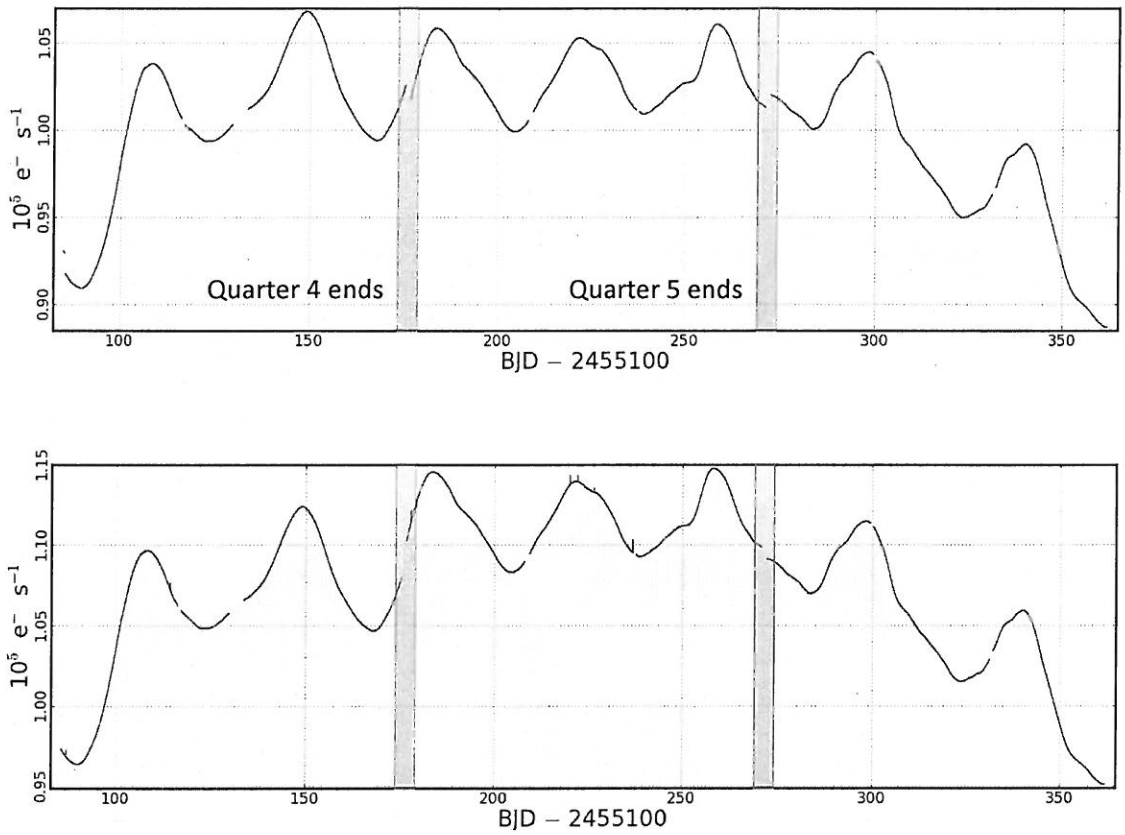

Fig. 7.— Upper: The archived SAP quarter 4, 5 and 6 light curves for StHA 169 (KIC 9603833). Each light curve was extracted from the optimal apertures (grey pixels) defined in Figure 6. Photometric discontinuities occur at each quarterly roll due to the redistribution of target flux over a new CCD (indicated by the blue bars for clarity), and the redefinition of the optimal aperture. Lower: Quarterly roll discontinuities are reduced by re-extracting the three light curves over all available pixels in the target masks. In this specific example, the redefined optimal aperture collects more of the target's flux without introducing significant contaminating flux from nearby sources. Extraction was performed using the PyKE *kepextract* task with the *maskfile=all* parameter. Some further examples of pixel extraction are provided in Appendices A and B.

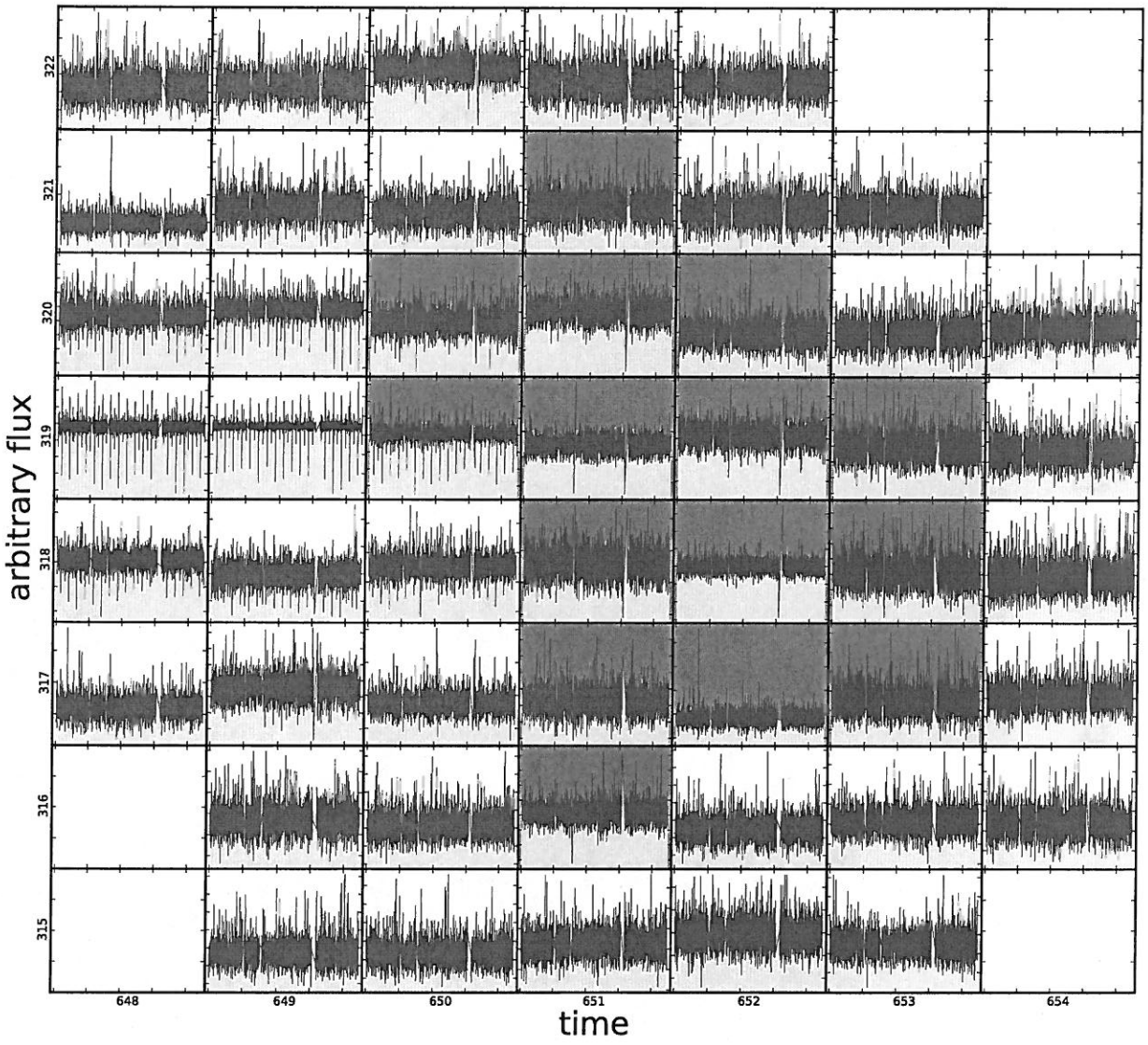

Fig. 9.— Calibrated time-series photometry for every pixel within the mask surrounding the quarter 3 target KIC 2449074. Unlike Figure 2, each light curve is plotted on an independent flux scale in order to identify if any background sources with respect to the brighter target appears. Source confusion with a background binary within the optimal aperture (gray pixels) is evident in pixel x=650, y=319. The source of the regular dips found in the PDCSAP light curve of Figure 8 is not the target that the mask was designed for. The source of the dips is a faint, background binary star on the left-hand side of the mask.

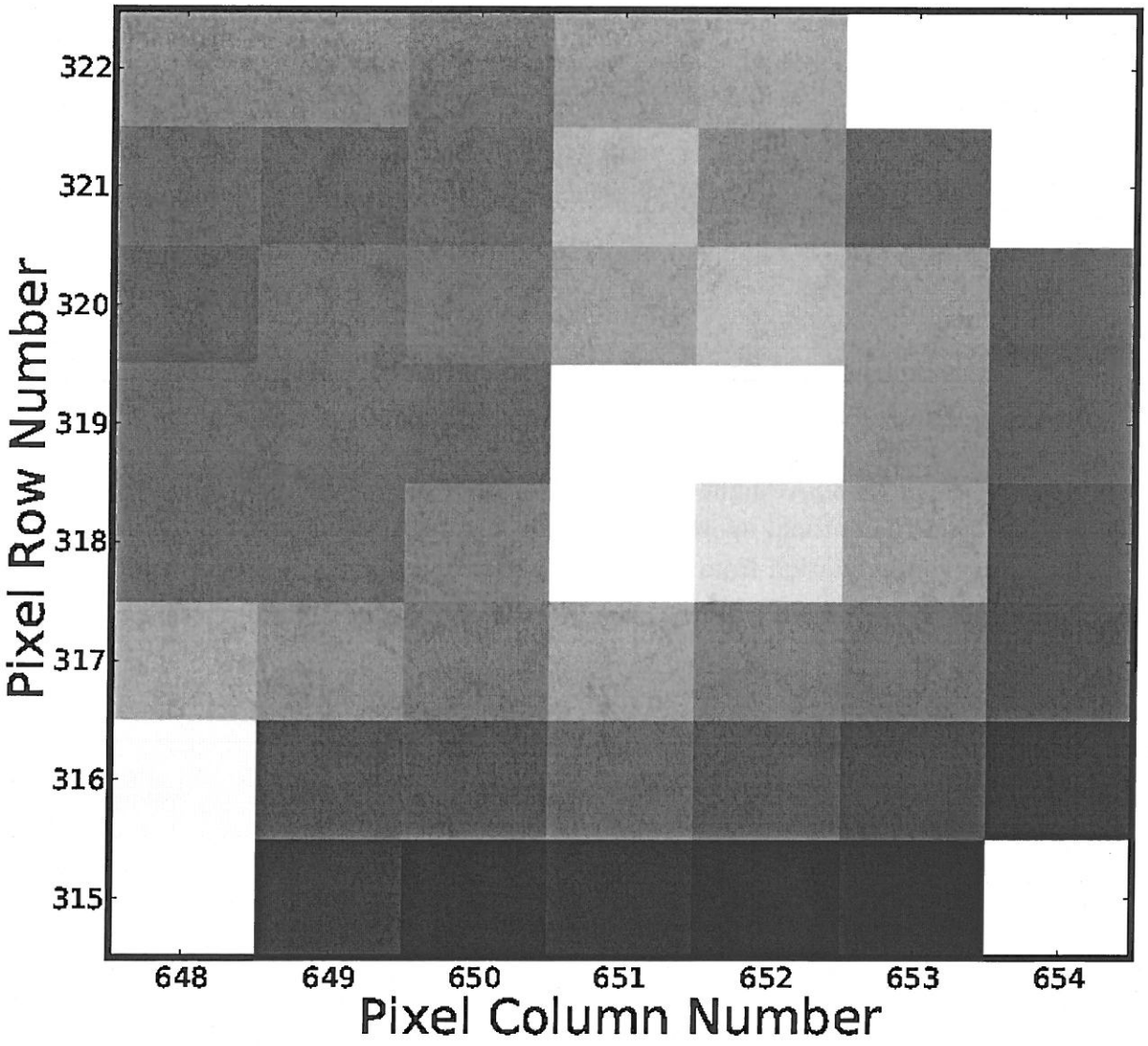

Fig. 11.— As for Figure 10, except the green region is a manually-defined optimal aperture that maximizes the signal from the background binary star within the mask. The selected green pixels minimize the contamination from the target star. When summed, the pixels within the new optimal aperture produce the light curve in Figure 13.

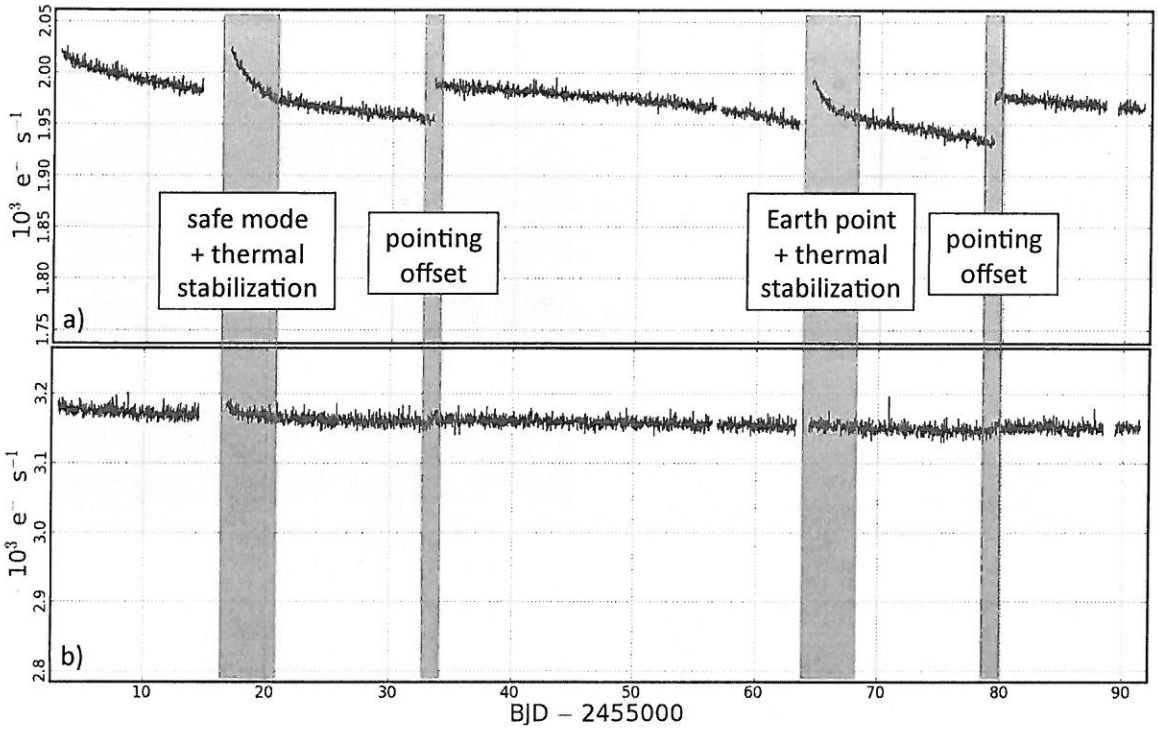

Fig. 14.— a) The archived quarter 2 SAP light curve for the Kepler target KIC 8703536. Blue regions identify specific events adding conspicuous systematic noise to the light curve, as labelled within the white boxes. b) A version of the light curve, mitigated for systematics by a different choice of optimal aperture pixels. A recipe for creating this new light curve from the archived Target Pixel File is provided in Appendix B.

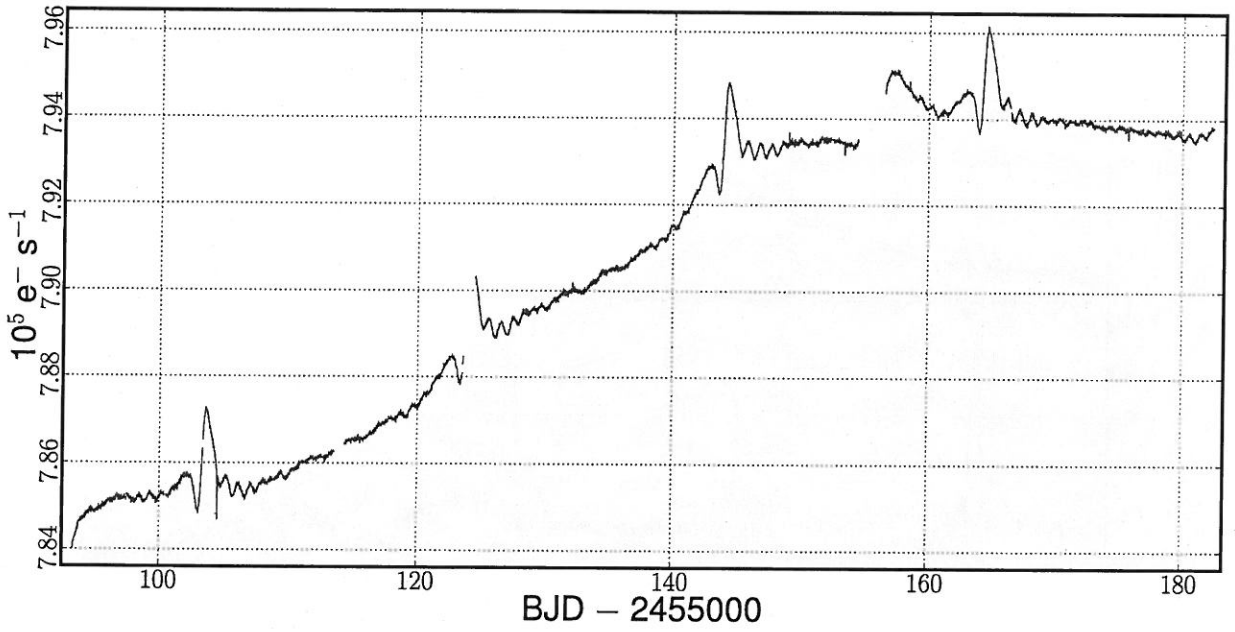

Fig. 16.— The quarter 3 long cadence SAP light curve of KIC 3749404. The long-term trend of increasing flux, 1% in amplitude, is most-likely caused by differential velocity aberration. A 5-d interval of thermal settling after an Earth point at BJD 2,455,156.5 stands out as a likely systematic feature over the astrophysical signal. Cotrending Basis Vectors can be employed to remove or reduce all the undesirable systematic effects.

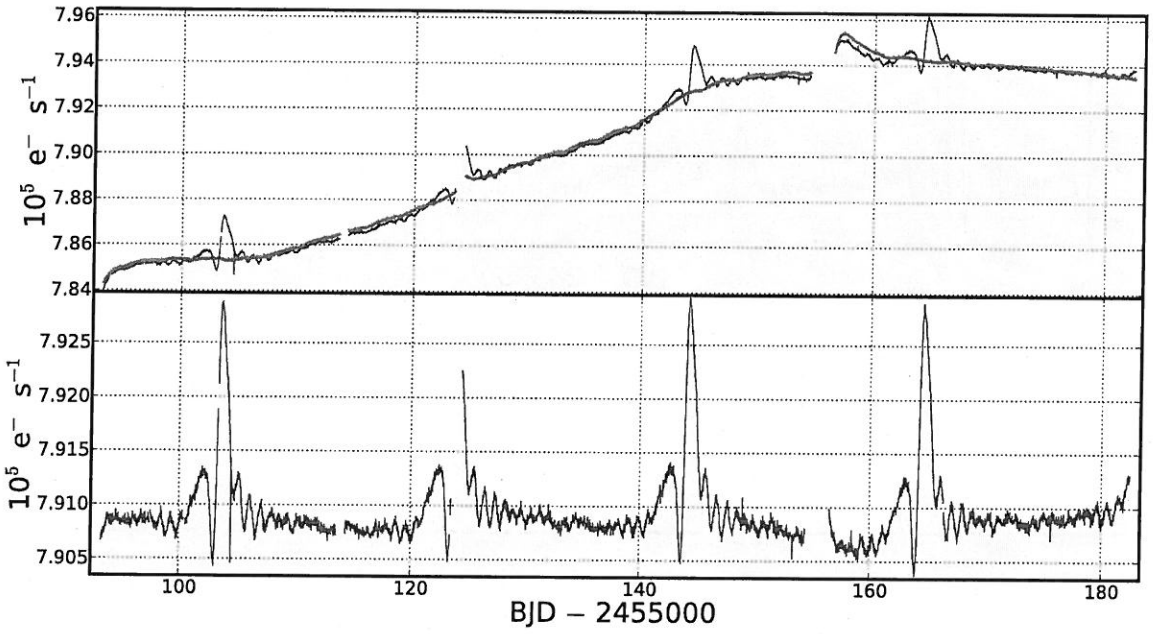

Fig. 18.— As for Figure 17. This time, we fit five CBVs to the quarter 3 SAP light curve of KIC 3749404. The systematics now appear to be much reduced but there are still some effects in the second half of the quarter that can be mitigated further (e.g. the poor fit to the thermal settling event around BJD 2,455,156.5.)

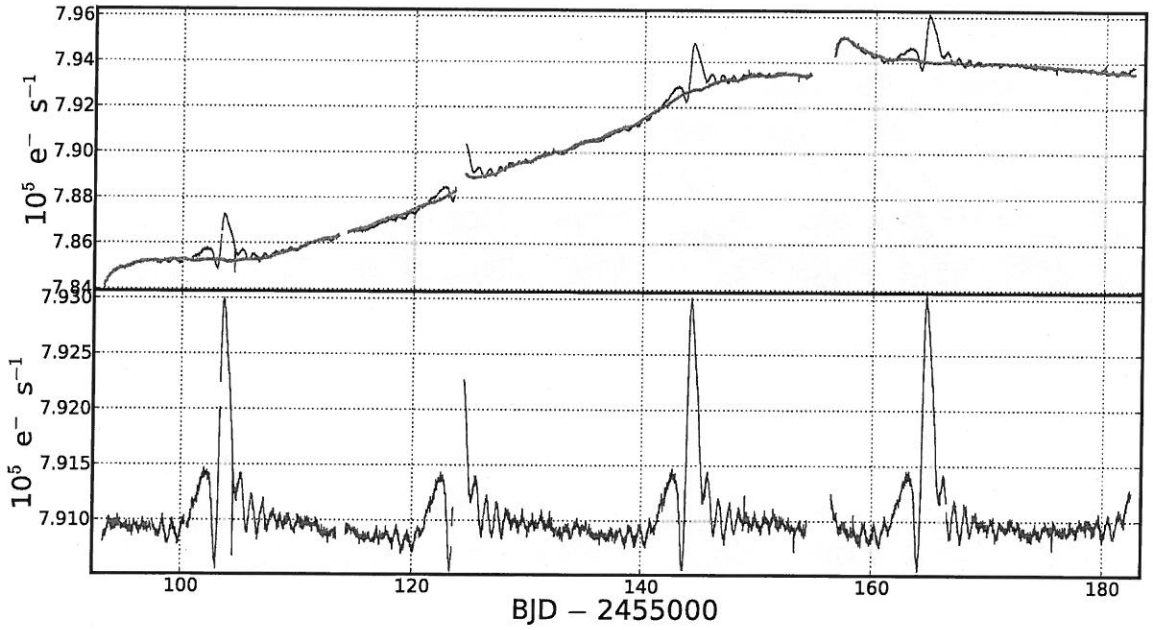

Fig. 21.— The final iteration of the CBV fit to the quarter 3 SAP light curve of KIC 3749404. We used the PyKE tool *kepcotrend* and fit eight basis vectors to the light curve. We did not fit the regions of the light curve highlighted in Figure 20.