OVERVIEW OF THE *KEPLER* SCIENCE PROCESSING PIPELINE

Jon M. Jenkins[1], Douglas A. Caldwell[1], Hema Chandrasekaran[1], Joseph D. Twicken[1], Stephen T. Bryson[2],
Elisa V. Quintana[1], Bruce D. Clarke[1], Jie Li[1], Christopher Allen[3], Peter Tenenbaum[1], Hayley Wu[1], Todd C.
Klaus[3], Christopher K. Middour[3], Miles T. Cote[3], Sean McCauliff[3], Forrest R. Girouard[3], Jay P. Gunter[3],
Bill Wohler[3], Jeneen Sommers[3], Jennifer R. Hall[3], AKM K. Uddin[3], Michael S. Wu[5], Paresh A. Bhavsar[2],
Jeffrey Van Cleve[1], David L. Pletcher[2], Jessie A. Dotson[2], Michael R. Haas[2], Ronald L. Gilliland[4], David G.
Koch[2], and William J. Borucki[2]

[1]SETI Institute/NASA Ames Research Center, M/S 244-30, Moffett Field, CA 94035, USA
[2]NASA Ames Research Center, M/S 244-30, Moffett Field, CA 94035, USA
[3]Orbital Sciences Corporation/NASA Ames Research Center, M/S 244-30, Moffett Field, CA 94035, USA
[4]Space Telescope Science Institute, Baltimore, MD 21218, USA and
[5]Bastion Technologies/NASA Ames Research Center, M/S 244-30, Moffett Field, CA 94035, USA

ABSTRACT

The *Kepler Mission* Science Operations Center (SOC) performs several critical functions including managing the ∼156,000 target stars, associated target tables, science data compression tables and parameters, as well as processing the raw photometric data downlinked from the spacecraft each month. The raw data are first calibrated at the pixel level to correct for bias, smear induced by a shutterless readout, and other detector and electronic effects. A background sky flux is estimated from ∼4500 pixels on each of the 84 CCD readout channels, and simple aperture photometry is performed on an optimal aperture for each star. Ancillary engineering data and diagnostic information extracted from the science data are used to remove systematic errors in the flux time series that are correlated with these data prior to searching for signatures of transiting planets with a wavelet-based, adaptive matched filter. Stars with signatures exceeding 7.1σ are subjected to a suite of statistical tests including an examination of each star's centroid motion to reject false positives caused by background eclipsing binaries. Physical parameters for each planetary candidate are fitted to the transit signature, and signatures of additional transiting planets are sought in the residual light curve. The pipeline is operational, finding planetary signatures and providing robust eliminations of false positives.

Subject headings: techniques: photometric — methods: data analysis

1. INTRODUCTION

The *Kepler Mission* seeks to detect Earth-like planets transiting solar-like stars by performing photometric observations of ∼156,000 carefully selected target stars in *Kepler's* 115 deg^2 field of view (FOV), as reviewed in Borucki et al. (2010) and Koch et al. (2010). These Long Cadence (LC) targets are sampled every 29.4 minutes and include all the planetary targets for which we seek signatures of transiting planets. In addition, a total of 512 Short Cadence (SC) targets are sampled at 58.85 s intervals permitting further characterization of the planet-star systems for the brighter ($Kp <12$) stars via asteroseismology, and more precise transit timing. The *Kepler Mission* Science Operations Center (SOC) at NASA Ames Research Center performs nine major functions:

1. Manage target aperture and definition tables specifying which 5.4×10^6 of the 95×10^6 pixels in the CCD array are processed and stored on the Solid State Recorder for later downlink[1].

2. Manage the science data compression tables and parameters, including the length-limited Huffman coding table, and the requantization table.

3. Report on the *Kepler* photometer's health and status semiweekly after each X-band contact and monthly after each Ka-band science data downlink.

4. Monitor the pointing error and compute pointing tweaks when necessary to adjust the spacecraft pointing to ensure the validity of the uplinked science target tables,

5. Process the science data each month to obtain calibrated pixels for all LC and SC targets, raw flux time series, and systematic error-corrected flux time series,.

6. Archive calibrated pixels, raw and corrected flux time series and centroid measurements to the Data Management Center (DMC).

7. Search each flux time series for signatures of transiting planets,.

8. Fit physical parameters and calculate error estimates for planetary signatures, and .

9. Perform statistical tests to reject false positives and establish accurate statistical confidence in each detection. .

Kepler's observations are organized into three month intervals called quarters defined by the roll maneuvers

Jon.Jenkins@nasa.gov

[1] *Kepler's* pointing stability requirement is 0".009, 3 σ, allowing us to preselect the pixels of interest for each star (Haas et al. 2010).

the spacecraft executes about its boresight to keep the solar arrays pointed toward the Sun (Haas et al. 2010). Once each month, the accumulated science data are transmitted via the Deep Space Network[2] (DSN) to the Mission Operations Center[3], which forwards them to the DMC[4]. The DMC packages them into FITS files and pushes them to the SOC. A selected set of ancillary engineering data are also delivered with the science data, containing any parameters likely to have a bearing on the quality of the science data, such as temperature measurements of the focal plane and readout electronics.

The Science Pipeline is divided into several components in order to allow for efficient management and parallel processing of the data, as shown in Figure 1. Raw pixel data downlinked from the *Kepler* photometer are calibrated by the Calibration module (CAL) to produce calibrated target and background pixels and their associated uncertainties. The calibrated pixels are processed by Photometric Analysis (PA) to fit and remove sky background and extract simple aperture photometry from the background-corrected, calibrated target pixels. PA also measures the centroid locations of each star on each frame. The final step to produce light curves happens in Pre-search Data Conditioning (PDC) where signatures in the light curves correlated with systematic error sources such as pointing drift, focus changes, and thermal transients are removed. Output data products include raw and calibrated pixels, raw and systematic error-corrected flux time series, centroids and associated uncertainties for each target star, which are archived to the DMC and eventually made available to the public through the Multimission Archive at STScI.[5]

In Transiting Planet Search (TPS) a wavelet-based, adaptive matched filter is applied to identify transit-like features with durations in the range of 1–16 hours. Light curves with transit-like features whose combined (folded) transit detection statistic exceeds 7.1σ for some trial period and epoch are designated as Threshold Crossing Events (TCEs) and subjected to further scrutiny by Data Validation (DV). This threshold ensures that no more than one false positive will occur due to random fluctuations over the course of the mission, assuming non-white, non-stationary Gaussian observation noise (Jenkins, Caldwell & Borucki 2002; Jenkins 2002). DV performs a suite of statistical tests to evaluate the confidence in the detection, to reject false positives by background eclipsing binaries, and to extract physical parameters of each system (along with associated uncertainties and covariance matrices) for each planet candidate. After the planetary signature has been fitted, it is removed from the light curve and the residual is subjected to a search for additional transiting planets. This process repeats until no further TCEs are identified. The DV results and diagnostics are furnished to the Science Team to facilitate disposition by the Follow-up Observing Program (FOP; Gautier et al. 2010).

2. PIXEL LEVEL CALIBRATIONS

The Pipeline module CAL corrects the raw *Kepler* photometric data at the pixel level prior to the extraction of photometry and astrometry. Several of the processing steps given in Figure 2 are familiar to ground-based photometrists. However, a few are peculiar to *Kepler* due to the lack of a shutter and unique features in its analog electronics chains. Details of these instrument characteristics and how they were determined and updated in flight are discussed in Caldwell et al. (2010) and are comprehensively documented in the *Kepler* Instrument Handbook (**?**).

The sequence of processing steps in CAL that produce calibrated pixels and associated uncertainties is as follows. (1) The two-dimensional black level (CCD bias voltage) structure (fixed pattern noise) is removed, followed by fitting and removing a dynamic estimate of the black level. (2) Gain and nonlinearity corrections are applied. (3) The analog electronics chain exhibits memory, necessitating the application of a digital filter to remove this effect, called Local Detector Electronics (LDE) undershoot. (4) Cosmic ray events in the black and smear measurements are removed prior to subsequent corrections. (5) The smear signal caused by operating in the absence of a shutter and the dark current for each CCD readout channel are estimated from the masked and virtual smear collateral data measurements. (6) A flat field correction is applied.

3. PHOTOMETRIC ANALYSIS

Before photometry and astrometry can be extracted from the calibrated pixel time series, the Pipeline detects so-called "Argabrightening" events in the background pixel data. These mysterious transient increases in the background flux were identified early in Commissioning. The current hypothesis is that these transient events are due to small dust particles from *Kepler* achieving escape velocity after micrometeorite hits and reflecting sunlight into the barrel of the telescope as they drift across the FOV. Argabrightenings that affect 10 or more CCD readout channels occur ~15 times per month, but the rate is dropping over time. The most egregious of these events cannot be perfectly corrected by the current background correction. We gap the data when the excess background flux exceeds the 100 Median Absolute Deviation (MAD) level.

PA then robustly fits a two-dimensional surface to ~4500 background pixels on each channel to estimate the sky background, which is evaluated at each target star pixel location and subtracted from the calibrated pixel values. Each target pixel time series is scanned for cosmic rays by first detrending the time series with a moving median filter with a width of five cadences (time steps) and examining the residuals for outliers compared to the MAD of the residuals for each pixel. Care is taken not to remove clusters of outliers that might be due to astrophysical signatures such as flares or transits that are intrinsic to the target star.

The photocenters of the 200 brightest, unsaturated stars on each channel are fitted using the pixel response functions (PRF; Bryson et al. 2010) and then used to define the ensemble motion of the stars over the observations. The aggregate star motion is used along with the PRFs reconstructed from Commissioning data to define the optimal aperture as the collection of pix-

els that maximizes the mean signal-to-noise ratio of the flux measurement for each star (Bryson et al. 2010). The background-corrected, cosmic ray-corrected pixels are then summed over the optimal aperture to define a flux estimate for each cadence frame.

4. SYSTEMATIC ERROR CORRECTIONS

PDC's task is to remove systematic errors from the raw flux time series. These include pointing errors, focus changes, and thermal effects on instrument properties. PDC co-trends each flux time series against ancillary engineering data such as temperatures of the focal plane and electronics, reconstructed pointing and focus variations to remove signatures correlated with these proxy systematic error measurements. A Singular Value Decomposition (SVD) is applied to the design matrix containing the ancillary data to identify the most significant, independent components and to stabilize the matrix inversion inherent in the fit to the data. Additionally, PDC identifies residual isolated outliers, and fills intra-quarter gaps so that the data for each quarterly segment are contiguous when presented to TPS. Finally, PDC adjusts the light curves to account for the excess flux in the optimal apertures due to starfield crowding in order to make apparent transit depths uniform from quarter to quarter as the stars move from detector to detector with each roll maneuver. This is achieved by estimating the mean excess flux in each photometric aperture from sources other than the target star itself from knowledge of the PRF and background star positions and magnitudes and subtracting this value from each point in the time series (see Bryson et al. 2010).

Significant effort has been applied to PDC in order to achieve good results with flight data. There are a number of phenomena that were significantly different than expected, including focus variations, and the amount of pointing drift observed during the first two quarters of operation. The systematic errors observed in flight exhibit a range of different time scales, from a few hours to several days to many days and weeks. Such phenomena include the intermittent modulation of the focus by $\sim$1 μm every 3.2 hr by a heater on one of the reaction wheel assemblies. One of the Fine Guidance Sensors' guide stars through the first quarter (Q1) of observations was an eclipsing binary whose 30% eclipses induced a 1 mpix pointing excursion lasting $\sim$8 hr every 1.7 days (Haas et al. 2010). By far the strongest systematic effects in the data so far have occurred after each of two safe mode events (Haas et al. 2010) during which the photometer was shut off, the telescope cooled and the focus changed by $\sim$2.2 μm per $^\circ$C. One of these occurred at the end of Q1 and the second $\sim$2 weeks into Q2. Thermal effects can be observed in the science data for $\sim$5 days after each safe mode recovery. The fact that most systematics such as these affect all the science data simultaneously, and that there is a rich amount of ancillary engineering data and science diagnostics available provides significant leverage in dealing with these effects.

Some systematic phenomena are specific to individual stars and cannot be corrected by co-trending against ancillary data. The first issue is the occasional, abrupt drop in pixel sensitivity that introduces a step discontinuity in an affected star's light curve (and associated centroids). This is often preceded immediately by a cosmic ray event, and is sometimes followed by an exponential recovery over a few hours, but usually not to the same flux level as before. The typical drop in sensitivity is 1%, which is unmistakable in the flux time series. Such step discontinuities are identified separately from those due to operational activities, such as safe modes and pointing tweaks, and are mended by raising the light curve after the discontinuity for the remainder of the quarter. These events do not mimic transits since they do not recover to the same pre-event flux level, and few transits, if any, are affected by this correction. The second issue is that many stars exhibit coherent or slowly evolving oscillations that interfere with systematic error removal. The approach taken is to identify and remove strong coherent components in the frequency domain prior to co-trending, and then to restore these components to the residuals after co-trending.

Figure 3 shows the results of running two flux time series obtained during Quarter 2 through PDC on schedule for release early in 2010, demonstrating PDC's effectiveness. We expect that learning to deal with the various systematic errors will consume a great deal of effort over the lifetime of the mission as we push the detection limit to smaller and smaller planets.

5. TRANSITING PLANET SEARCH

TPS searches for transiting planets by "stitching" the quarterly segments of data together to remove gaps and edge effects and then applies the wavelet-based, adaptive matched filter of Jenkins (2002). This approach is a time-adaptive approach that estimates the power spectrum of the observational noise as a function of time. This approach was developed specifically for solar-like stars with colored, broadband power spectra. Some modifications to the original approach have been developed to accommodate target stars that exhibit coherent structure in the frequency domain. Similar to the approach adopted in PDC, we fit and remove strong harmonics that are inconsistent with transit signatures prior to applying the wavelet-based filter. This significantly increases the sensitivity of the transit search for such stars and also provides photometric precision estimates (as byproducts of the search) that are more realistic for such targets. If the transit-like signature of a given target star exceeds 7.1 σ then a TCE is recorded and sent to DV for additional scrutiny.

6. DATA VALIDATION

DV performs a suite of tests to establish or break confidence in each TCE flagged by TPS, as well as to fit physical parameters to each transit-like signature. DV is currently under development and we anticipate its release in early 2010 to support the next FOP observing season.

The statistical confidence in the TCE is examined by performing a bootstrap test (Jenkins 2002; Jenkins, Caldwell & Borucki 2002) to take into account non-Gaussian statistics of the individual light curves. A transiting planet model is fitted to the transit signature as a joint noise characterization/parameter estimation problem. That is, the observation noise is *not* assumed to be white and its characteristics are estimated using the wavelet-based approach employed in TPS, but as an estimator, rather than as a detector. This process yields a

set of physical parameters and an associated covariance matrix.

To eliminate false positives due to eclipsing binaries, the planet model fit is performed again only to the even transits, and then only to the odd transits, and the resulting odd/even depths and epochs are compared in order to see if the results indicate the presence of secondary eclipses. After the multi-transiting planet search is complete, the periods are compared to detect eclipsing binaries with significant eccentricity causing TPS to detect two transit pulse trains at essentially the same period, but at a phase other than 0.5.

To guard against background eclipsing binaries, a centroid motion test is performed to determine whether the centroids shifted during the transit sequence. If so, the source right ascension and declination can be estimated by the measured in- versus out-of-transit centroid shift normalized by the fractional change in brightness of the system (i.e., the transit "depth"; Batalha et al. 2010b; Monet et al. 2010).

Additional tests include checking whether the transit signature is consistent in the target pixels, whether the transit signature is correlated with any ancillary engineering data or any collateral data, and whether the distribution of cosmic ray events during transit is significantly different than that out-of-transit.

7. FUTURE DEVELOPMENT

Future development for the SOC includes implementing difference image analysis photometry, completing DV, and developing and implementing mitigations for the instrument artifacts described in Caldwell et al. (2010). These artifacts affect a small portion of the *Kepler* FOV at any one time. The development schedule calls for delivery of these features to allow us to discover long period Earths by late 2010 and recover at least 90% of the FOV in 2011 in time to characterize the frequency of Earth-size transiting planets in the habitable zones of solar-like stars.

The instrumental artifacts consist of two categories of phenomena: (1) temperature-dependent two-dimensional bias image structure, and other temperature-dependent electronics effects, and (2) Moiré patterns caused by an unstable circuit with an operational amplifier oscillating at ∼1.5 GHz. Normally this latter feature appears as a high-frequency oscillation on each readout row whose frequency changes with row number as the readout electronics heat up during readout. When this signal aliases to the sample rate of the CCD readout, a transient band appears in an affected channel and slowly rolls across the frame as the temperature changes. The Moiré pattern can interact with bright, saturated star signals and generate scene-dependent effects. The typical amplitude of these image artifacts is <1 ADU per pixel per read, comparable to or smaller than the typical readout noise. It is important to note that not all of these op amps are oscillating and that the perturbations to the images are very small. Our mitigation plan consists of two approaches, one for the temperature-dependent effects, and one for the Moiré pattern effects, and most of the effort takes place prior to pixel level calibrations.

Before launch, we added pixels to the target table that allow us to sample and trend the artifacts simultaneously with the science data. The *Kepler* Science Office and SOC are developing and prototyping algorithms that use these image artifact pixels, together with other science data, to reconstruct the underlying temperature-dependent two-dimensional bias structure as a function of time over each quarter. The resulting dynamic model will allow the temperature-dependent bias signals to be removed directly from the data. Moreover, the thermal environment of the *Kepler* photometer is very stable and changes slowly during nominal operations. Thus any residual thermal two-dimensional bias effects will be small after the corrections are in place and can be co-trended out of the data by PDC like other thermally driven, instrumental effects.

Given that the Moiré pattern noise exhibits both high spatial frequencies and high temporal frequencies, the prospect of reconstructing a high fidelity model of the effects at the pixel level with an accuracy sufficient to *correct* the affected data appears unlikely. We are developing algorithms that identify when these Moiré patterns are present and mark the affected CCD regions as suspect on each affected LC. These suspect data flags will then be used to inform downstream modules so that the affected data can be appropriately weighted, and so that, for example, TPS can selectively ignore time intervals that are potentially contaminated with electronics-induced transients when searching for transit signatures. DV will produce a contamination report for each TCE indicating the fraction and severity of the Moiré pattern effect. The Pipeline will track and trend diagnostic metrics reflecting the prevalence and severity of the Moiré pattern as a diagnostic of this aspect of the photometer performance.

In spite of the presence of these image artifacts, *Kepler* is already achieving photometric precision sufficient to detect Earth-size planets transiting solar-like stars for the majority of the FOV at any given time (Jenkins et al. 2010). We are confident that these efforts will enable us to minimize the impact of these artifacts on exoplanet detection, and produce high-quality photometric and astrometric time series for other scientific investigations by the greater astronomical community.

8. CONCLUSIONS

We have presented an overview of the *Kepler* SOC science pipeline processing. The output products include raw and calibrated pixel time series, raw and systematic error-corrected flux time series, centroid time series for each star, and associated uncertainties. These products will permit the detection and characterization of transiting planets in the *Kepler* FOV as well as enabling astrophysical investigations and serendipitous discoveries not contemplated in *Kepler's* driving design requirements.

REFERENCES

Batalha, N. B., et al. 2010, ApJ, this issue

Borucki, W. J., et al. 2010, Science, in press

Bryson, S. T., et al. 2010, ApJ, this issue
Caldwell, D. A., et al. 2010, ApJ, this issue
Gautier, N. T., et al. 2010, ApJ, this issue
Jenkins, J. M., D. A. Caldwell, & W. J. Borucki 2002, ApJ, 564, 495
Jenkins, J. M. 2002, ApJ, 575, 493
Jenkins, J. M., et al. 2010, ApJ, this issue
Haas, M., et al. 2010, ApJ, this issue

Koch, D. G., et al. 2010, ApJ, this issue
Monet, D, et al. 2010, ApJ, this issue
Van Cleve, J., D. A. Caldwell, 2009, Kepler Instrument Handbook, KSCI 19033-001, (Moffett Field, CA: NASA Ames Research Center)
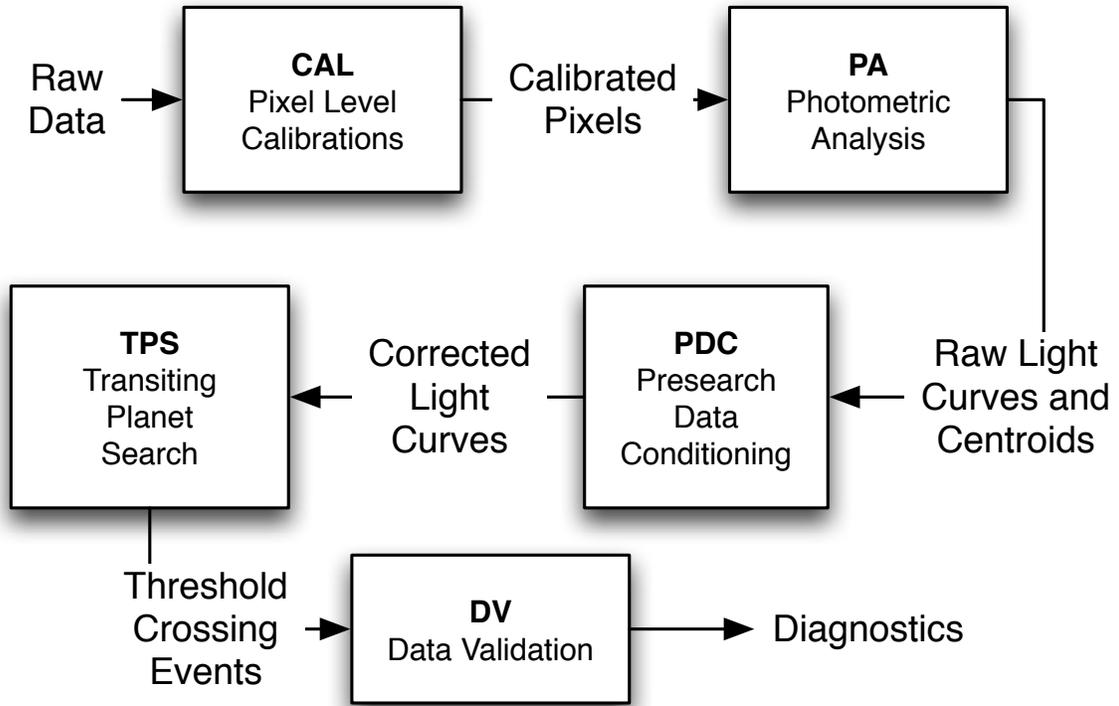
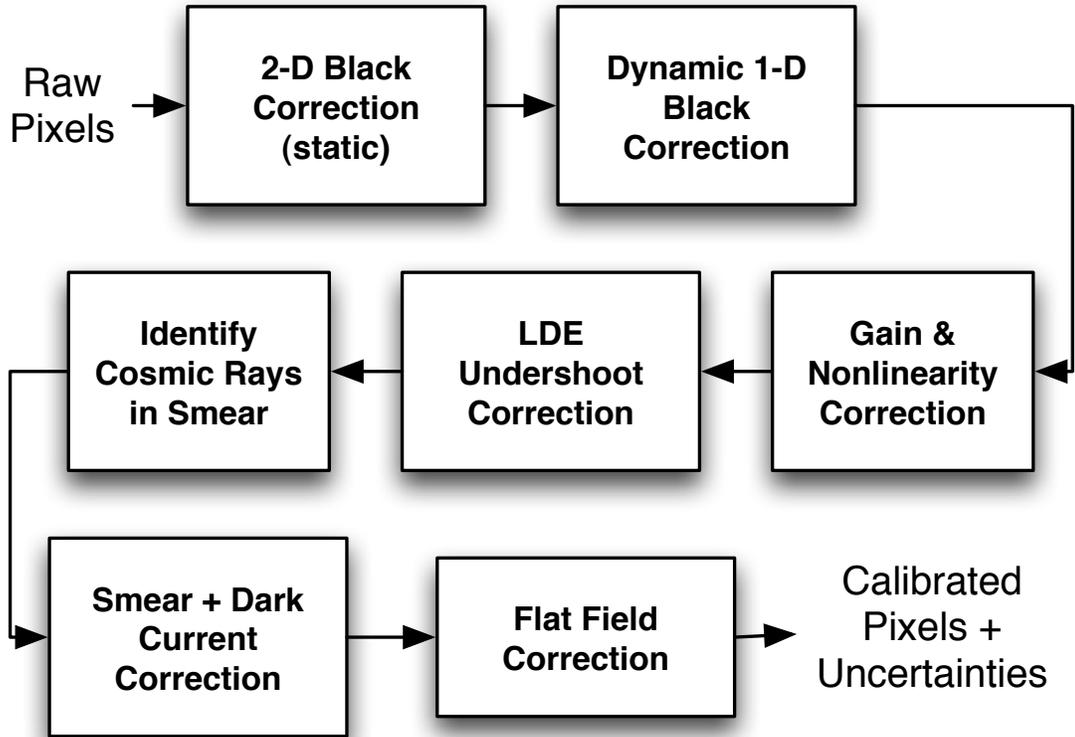FIG. 1.— Data flow diagram for the SOC Science Pipeline.

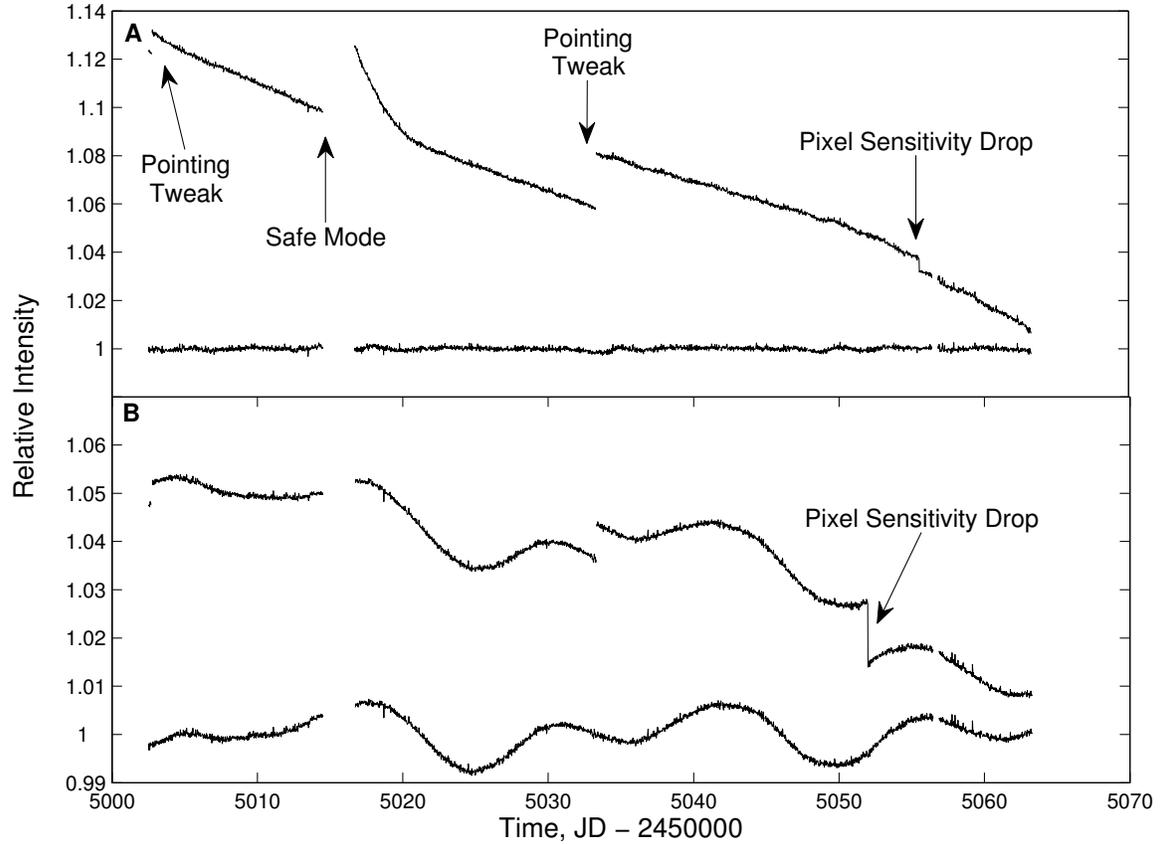Fig. 2.— Data flow diagram for the Calibration Pipeline module.

FIG. 3.— Raw and systematic error-corrected light curves for two different stars. The raw light curves exhibit step discontinuities due to pointing offsets, thermal transients due to safe modes and pixel sensitivity changes, as well as pointing errors and focus changes, as indicated in the figure. Panel A shows the raw flux time series (top curve, offset) and the PDC flux time series (bottom curve) of a $Kp = 15.6$ dwarf star. Panel B shows the raw and PDC flux time series for a $Kp = 14.9$ dwarf star that displays less sensitivity to the thermal transients. Both stars' corrected flux time series are significantly improved by the detrending afforded by PDC while retaining intrinsic stellar variability on timescales of several weeks.