Presearch data conditioning in the *Kepler* Science Operations Center pipeline

Joseph D. Twicken*[a], Hema Chandrasekaran[a], Jon M. Jenkins[a],
Jay P. Gunter[b], Forrest Girouard[b], Todd C. Klaus[b]
[a]SETI Institute/NASA Ames Research Center, MS 244-30, Moffett Field, CA, USA 94035-1000;
[b]Orbital Sciences Corporation/NASA Ames Research Center, MS 244-30, Moffett Field, CA, USA 94035-1000

ABSTRACT

We describe the Presearch Data Conditioning (PDC) software component and its context in the *Kepler* Science Operations Center (SOC) Science Processing Pipeline. The primary tasks of this component are to correct systematic and other errors, remove excess flux due to aperture crowding, and condition the raw flux light curves for over 160,000 long cadence (~thirty minute) and 512 short cadence (~one minute) stellar targets. Long cadence corrected flux light curves are subjected to a transiting planet search in a subsequent pipeline module. We discuss science algorithms for long and short cadence PDC: identification and correction of unexplained (i.e., unrelated to known anomalies) discontinuities; systematic error correction; and removal of excess flux due to aperture crowding. We discuss the propagation of uncertainties from raw to corrected flux. Finally, we present examples from *Kepler* flight data to illustrate PDC performance. Corrected flux light curves produced by PDC are exported to the Multi-mission Archive at Space Telescope [Science Institute] (MAST) and are made available to the general public in accordance with the NASA/Kepler data release policy.

Keywords: Kepler, transit photometry, light curve, systematic error correction.

1. INTRODUCTION

The *Kepler Mission* is designed to detect (habitable) Earth-size planets transiting Sun-like stars[1]. The spacecraft was launched on 6 March 2009 into an Earth-trailing heliocentric orbit with a period of 373 days. Pointing of the *Kepler* photometer is maintained to support imaging of the same star field continuously over the life of the mission (nominally 3.5 years for the primary mission). The field of view of the *Kepler* photometer is ~115 square degrees. Incident light is collected by 42 charge-coupled device (CCD) detectors (94.6 million total pixels) on the focal plane assembly. There are two readout channels (or module outputs) per CCD, for a total of 84 on the focal plane. Short exposures are integrated onboard to produce one image every 29.4 minutes for over 160,000 long cadence[2] (LC) targets, and one image every 0.98 minutes for 512 short cadence[3] (SC) targets. The spacecraft is rolled by 90 degrees on a quarterly basis so that the solar panels are continuously directed toward the Sun. Flux from any given stellar target is therefore captured by a different CCD detector from one science data acquisition season to the next.

Science data acquired in-flight are processed in the *Kepler* Science Operations Center (SOC) Science Processing Pipeline[4,5]. Pixel values are calibrated for each cadence in the Calibration (CAL) software component[6]. Raw flux light curves are extracted and target photocenters (centroids) are computed in the Photometric Analysis (PA) component[7]. Systematic and other errors are corrected in the Presearch Data Conditioning (PDC) component described in this paper. Long cadence corrected flux light curves are then subjected to a search for transiting planets. The Transiting Planet Search (TPS) pipeline module[8] returns a Threshold Crossing Event (TCE) for each target and trial transit pulse that exceeds the specified detection threshold. A transiting planet model is fitted to the corrected flux light curves in the Data Validation (DV) pipeline module[9,10] for targets with TCEs. A transit signature obtained from the fitted parameters is subsequently removed from the corrected flux time series for each candidate planet, and a search is conducted for additional candidate planets. A suite of automated tests is performed when no additional candidate planets can be identified. The purpose of the automated tests is to facilitate identification of the true candidate planets from the large number of false positive transiting planet detections (astrophysical and otherwise).

*joseph.twicken@nasa.gov; kepler.nasa.gov

The first task of PDC is to correct systematic and other errors in the raw flux light curves produced by the PA module. Flux discontinuities that cannot be attributed to known spacecraft or data anomalies are identified and corrected. Systematic error correction is then performed by removing flux signatures in the respective light curves that are correlated with ancillary engineering or pipeline data. Systematics in the flight data are attributable to a variety of sources, and are present on multiple time scales over a wide dynamic range. The second task of PDC is to remove excess flux from the light curves due to background sources within the respective target apertures (crowding). The final task of PDC is to condition the light curves for the transiting planet search. This involves identification and removal of flux outliers and filling of data gaps. Outlier identification and data gap filling are not addressed in this paper.

Corrected flux light curves are exported to the Multi-mission Archive at Space Telescope [Science Institute] (MAST) and are made available to the general public in accordance with the NASA/Kepler data release policy. Raw flux light curves, centroids, and barycentric timestamp corrections generated in PA are also exported to the MAST. Flux outliers identified and removed in PDC for the purpose of conditioning light curves for the transiting planet search in the SOC pipeline are restored in the corrected flux light curves exported to the MAST. Furthermore, filled data gaps are removed from the exported light curves.

An overview of PDC and the flow of data through the pipeline module are presented in Section 2. PDC science algorithms are described in Section 3; identification and correction of random flux discontinuities are discussed in Section 3.1; systematic error correction is discussed in Section 3.2; and removal of excess flux due to aperture crowding is discussed in Section 3.3. A summary and conclusions are presented in Section 4.

2. PRESEARCH DATA CONDITIONING OVERVIEW AND DATA FLOW

The primary tasks of PDC are to correct systematic and other errors in the raw flux light curves produced in the PA module, to remove excess flux in light curves due to crowding in the respective target apertures, and to condition light curves for the transiting planet search performed in the TPS module by identifying and removing flux outliers and filling data gaps.

Systematic errors are present in the flight science data on a range of time scales and may be traced to a variety of sources[2,4]. Targets also exhibit native variability over a range of time scales and with a wide variety of astrophysical signatures[2]. The goal of PDC is to remove systematic errors from raw flux light curves while leaving the native target variation and astrophysical signatures intact. It is difficult, if not impossible, to do this for all targets. There is an attempt in the current code release (SOC 6.1) to identify those targets for which systematic error correction performs badly; in these cases, the raw flux is passed through to the back end of PDC uncorrected.

The standard PDC unit of work[11,12] for LC and SC science data processing is a single module output for a duration of one quarter (LC) or one month (SC). Raw flux light curves for all targets on a module output are provided as input through the module interface[11,12] and corrected flux light curves are passed back through the module interface both with and without fitted harmonic content. Indices of flux outliers are produced as an output of PDC along with the associated outlier values and uncertainties. Indices of filled data gaps are also returned by PDC. Unless a raw flux light curve is pathological, all data gaps should be filled by PDC. Corrected flux light curves (with harmonic content) for LC and SC targets are exported periodically to the MAST. Outlier values and uncertainties are restored in the export files and the filled data gaps are removed.

Data flow in the PDC pipeline module is shown in Figure 1. Data generally flows from left to right and top to bottom in the figure. Functional blocks in the diagram will be referenced with bold type where they are discussed in the text, and section numbers will be included for the algorithms that are described in detail later in this paper.

Synchronize Ancillary Data. Ancillary engineering data, ancillary pipeline data (i.e., ancillary data produced in another pipeline module), and temporal motion polynomial sequences[7] (produced for the given module output in PA) are synchronized to mid-cadence timestamps in the unit of work to support systematic error correction. Signatures that are correlated with synchronized ancillary data are removed from raw flux light curves to correct the systematic errors. It is not necessary that ancillary engineering data, ancillary pipeline data, and motion polynomials all be present in the PDC input structure. Systematic error correction is performed given whatever ancillary data are available. The synchronization process involves binning to cadence, sub-cadence, or super-cadence intervals followed by digital resampling (decimation or interpolation) where necessary. Gaps may be filled in the synchronization process.

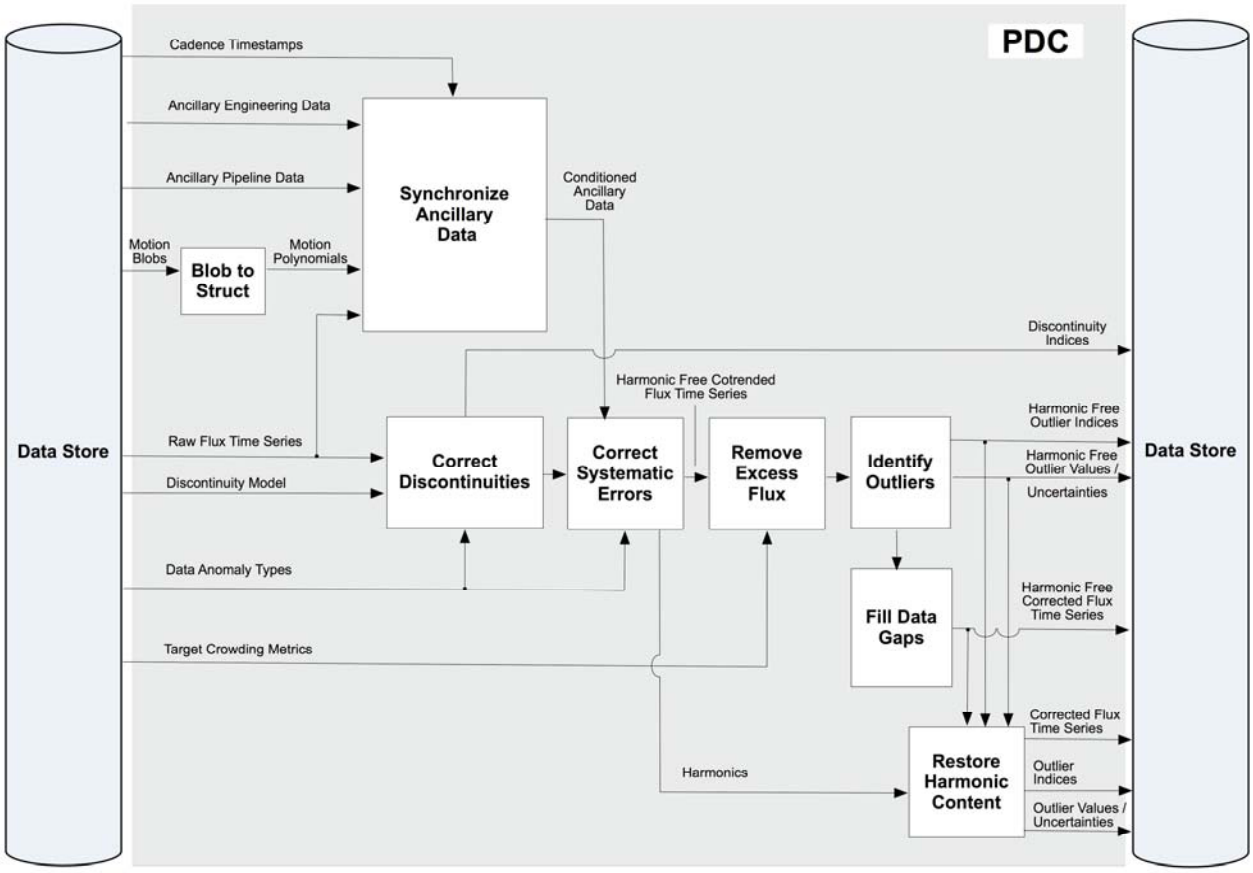

Figure 1: Data flow diagram for the Presearch Data Conditioning (PDC) pipeline module. Inputs are shown at the left and outputs are shown at the right. Inputs are obtained from the Data Store and outputs are written to the Data Store.

Correct Discontinuities (3.1). Random flux discontinuities have been observed since the first flight data were acquired[4]. The random discontinuities are differentiated from discontinuities introduced into many target light curves by spacecraft activities and anomalies (monthly downlinks, safe modes) and commanded attitude adjustments. It is likely that the random flux discontinuities are caused by impacts of cosmic rays or other energetic particles on CCD pixels. Prior to correction of systematic errors in PDC, an algorithm is employed to identify and correct random discontinuities in the raw flux light curves. A discontinuity template is correlated against the numerical second derivative of raw flux for each target and a threshold is applied to identify significant events. In addition to cadence indices of detected discontinuities, step sizes are also estimated. Discontinuities are subsequently corrected for each target by adjusting the flux values following each identified discontinuity by the associated step size. The process of discontinuity identification and correction is iterated for each target to allow for correction of multiple cadence discontinuities.

Correct Systematic Errors (3.2). Systematic errors are corrected by a process referred to as *cotrending*. A design matrix is created by separately filtering each of the synchronized ancillary data time series into selectable lowpass, midpass, and highpass components. Each raw flux time series is then projected into the column space of the design matrix in a least squares sense, and the residual (with mean level restored) between the raw flux and least squares fit determines the systematic error corrected flux for each target. This process essentially removes flux signatures that are correlated with the ancillary data on the specified time scales. A Singular Value Decomposition (SVD) is utilized to perform the projection in a computationally efficient and numerically stable manner. Uncertainties in corrected flux values are propagated from uncertainties in the raw flux values in accordance with standard methods. Memory limitations prevent the creation of full covariance matrices for each corrected flux time series, however.

An attempt is made to identify variable targets and to fit the light curve for each such target with a superposition of phase shifting harmonics. Reliable identification of variable targets is difficult, however, in the presence of large data anomalies. Harmonic content is subsequently removed from light curves of the variable targets for which harmonic fitting was successful. All of the harmonic free light curves are again subjected to the cotrending process. A decision is made whether to use the standard or harmonic free cotrending result for each of the targets initially identified as variable. Systematic error correction performance is finally evaluated for all targets in the unit of work and error-corrected flux is replaced with raw flux for each of the targets determined to have been badly corrected.

Remove Excess Flux (3.3). Following the correction of systematic errors, excess flux due to crowding in the optimal aperture[7] is removed from the light curve for each target. The amount of excess flux is determined from the crowding metric that is provided to PDC for each target. The crowding metric is defined as the fraction of flux in the optimal aperture due to the target itself. The metric is computed in the Target and Aperture Definitions (TAD) component[13] of the SOC pipeline. A single value is provided for each target and target table even though crowding may vary with long-term motion of targets and background sources primarily due to Differential Velocity Aberration (DVA).

Identify Outliers. Outliers are identified in each flux time series based on robust estimates of the mean and standard deviation in a sliding scan window. The window size and outlier detection threshold are PDC module parameters[11] and are separately tuned for long and short cadence units of work. Flux values marked as outliers are gapped and later filled along with other data gaps. The purpose of outlier identification and removal is to prevent the triggering of false TCEs in TPS. Outlier values and uncertainties are restored in the corrected flux light curves exported to the MAST.

Fill Data Gaps. An attempt is made to fill all data gaps in PDC. The transiting planet search requires that samples be available for all cadences. Gap filling for each target proceeds in two steps: first, “short” data gaps are filled, and then any remaining “long” data gaps are filled. Short and long refer not to the type of cadence data being processed, but to the length of the gaps to be filled. The boundary between short and long data gaps is determined by the gap filling module parameter set.

Short data gaps are sequentially filled with available flux samples to the left and right of the respective gaps. An autoregressive algorithm is employed to estimate sample values in the gaps with a linear prediction based on the flux correlation in the neighborhood of the gap. Uncertainties in short gap-filled samples are produced from uncertainties in the samples used to fill them. Long data gaps are filled in a process that involves folding and tapering blocks of available samples from the left and right of the respective gaps. Wavelet domain coefficients are then adjusted to ensure statistical continuity across the filled gaps. There is no attempt to estimate uncertainties for long data gap filled samples. Gap-filled values are not included in the corrected flux light curves exported to the MAST.

Restore Harmonic Content. Harmonic content identified and removed for harmonically variable targets is restored to the corrected flux light curves before PDC runs to completion. Harmonic content is also restored to the outlier values identified earlier. PDC therefore produces two corrected light curves and two sets of outliers for each target, one based on the standard flux time series for the given target, and one based on the harmonic free flux time series. For targets without fitted harmonic content, the standard and harmonic free results are identical.

Kepler is first and foremost a transit photometry mission. Every effort is made to preserve transits in PDC and to prevent them from compromising performance of the science algorithms. When discontinuities are identified and corrected, an attempt is made to ensure that large transits (and other astrophysical events such as binary eclipses and flares) do not trigger the detector. In the correction of systematic errors, an attempt is made to prevent large transits and astrophysical events from corrupting the least squares fitting process. These events are restored after fitting is performed.

Large transits and astrophysical events are also masked prior to performing outlier identification. There are two reasons for doing so. First, transits are masked to prevent them from corrupting the estimates of mean and standard deviation that are utilized in setting the robust outlier detection threshold. Second, transits are masked in order to prevent them from being identified as outliers. An attempt is also made in the gap filling process to prevent transits and other astrophysical events in available science data samples from being used to fill both short and long data gaps.

3. PRESEARCH DATA CONDITIONING SCIENCE ALGORITHMS

PDC science algorithms are discussed in this section with flow charts and illustrations where appropriate.

3.1 Discontinuity identification and correction

Random flux discontinuities have been observed for some targets since the first flight data were acquired[4]. The random discontinuities are differentiated from discontinuities introduced into many of the target light curves as a result of spacecraft activities and anomalies (monthly downlinks, safe modes) and commanded attitude adjustments. Random discontinuities are most often attributed to abrupt decreases in sensitivity, perhaps due to impacts of cosmic rays or other energetic particles on CCD pixels. They are sometimes followed by a partial exponential rebound. There is no attempt in the current PDC release (6.1) to model and correct exponential rebounds.

A flow chart describing the discontinuity identification algorithm is shown in Figure 2. The process is performed independently on all raw flux light curves in a given unit of work. An attempt is first made to replace giant transits and other astrophysical events. Savitzky-Golay filtering is performed on the raw flux time series to compute the numerical derivatives for orders zero, one, and two. A sliding discontinuity template (which is provided as a parameter through the module interface) is then correlated against the filtered second derivative of the time series. Statistics are computed for the correlation time series and a threshold is applied to identify discontinuity candidates.

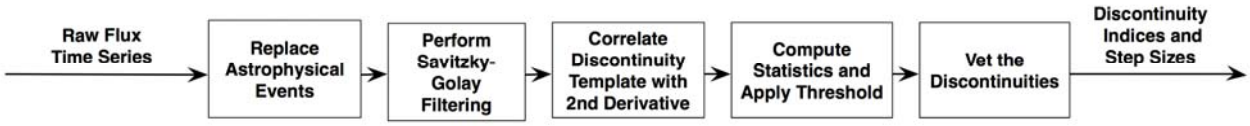

Figure 2: Flow chart for the discontinuity identification algorithm. The process is performed independently for each target in the unit of work.

Candidates are vetted before results are returned. Discontinuities that coincide with known spacecraft and data anomalies are excluded, as these are addressed as part of the systematic error correction process. Discontinuities that may have been identified as artifacts of interpolated data gaps or masked astrophysical events are also discarded. Discontinuities that fail a gradient test and are apparently due to single cadence outliers are discarded as well. Cadence indices and discontinuity step sizes are returned for each target for all discontinuity candidates that survive the vetting process.

Correction of discontinuities is straightforward. For any given target, the portion of the flux time series following each identified discontinuity is adjusted by the estimated discontinuity step size until all discontinuities have been corrected. The process of discontinuity identification and correction is repeated until no additional discontinuities are found or an iteration limit is reached. The iterative process allows multiple cadence discontinuities to be identified and corrected. If discontinuities are still identified for a given target after the iteration limit has been reached, the process is deemed to have failed for that target and the initial flux values are restored without correction of any discontinuities.

A discontinuity detection example for a long cadence Q3 target[14] on module output 7.3 is shown in Figure 3. The raw flux time series is plotted versus cadence in the upper panel. There are two flux discontinuities present that are not the result of known spacecraft activities or anomalies. The first of these occurs just prior to the downlink following the first month of Q3, and the second occurs nearly a week before the downlink following the second month of Q3. The associated discontinuity detection statistics are shown in the lower figure. Both events clearly exceed the specified 5σ detection threshold. Detection statistics for cadence indices returned by the detector are circled in the lower figure. Multiple random discontinuities are not common for targets in a single unit of work. The discontinuity following the second downlink of the quarter (near cadence 3000) and the associated thermal transient are addressed later in PDC when systematic error correction is performed.

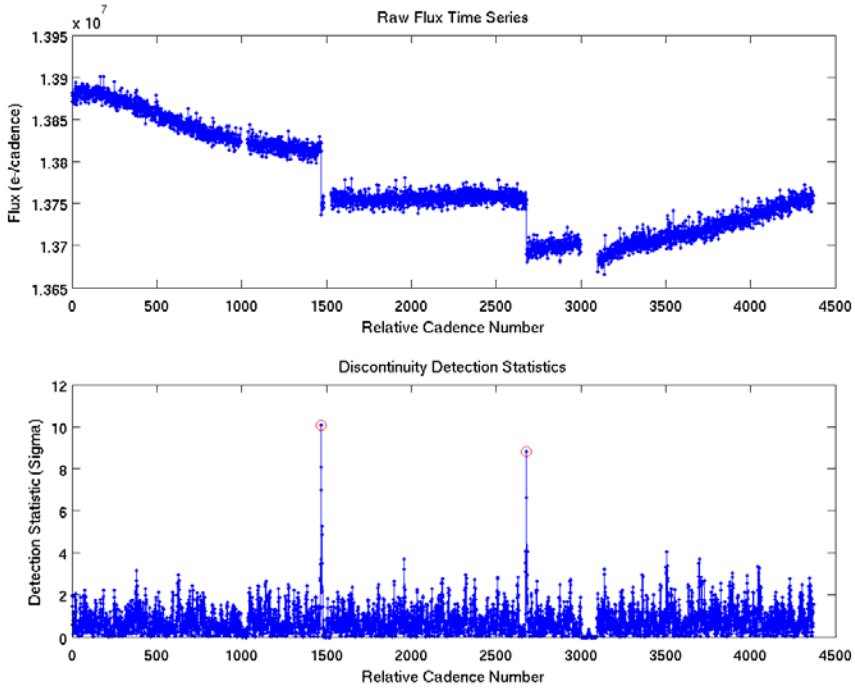

Figure 3: A raw flux time series with two random discontinuities is shown in the upper panel. The detection statistics formulated by correlating the second derivative of the time series with the discontinuity template are plotted in the lower panel. The detection statistics are circled for the cadence indices returned by the detector.

3.2 Systematic error correction

Systematic errors are introduced into long and short cadence light curves by a variety of sources over a broad spectrum of dynamic ranges and time scales. In the first year of science data collection it has become apparent that the systematic errors are caused primarily by target motion at the pixel or sub-pixel level. Target motion in turn produces changes in target flux levels. The motion polynomials[7] produced in PA by fitting the centroids of selected targets for each cadence as a function of the celestial target coordinates are therefore well suited for removing systematic effects on a module output basis.

The dominant long-term systematic effect is DVA, which causes targets to trace small ellipses on the respective CCD detectors over the period of the heliocentric orbit of the photometer[2]. The maximum motion due to DVA is 0.6 pixels per observing quarter[2]. Other significant systematic errors[2,4] in flight science data have resulted from variable (eclipsing binary) Fine Guidance Sensor (FGS) reference targets, short-period (~3 hours) reaction wheel heater cycling, long duration (~4-5 days) thermal transients following safe modes and monthly downlinks, and commanded photometer attitude adjustments. Early in the mission it was necessary to perform multiple attitude adjustments in a single quarter (Q2) to accommodate drift in photometer pointing[14].

The ability to correct systematic errors in PDC directly impacts performance of the transiting planet search in TPS and hence, the detectability of the very planets that the *Kepler Mission* was designed to discover. Systematic errors must be corrected so that they do not trigger massive numbers of TCEs in the transiting planet search, and so that they do not prevent detection of Earth-size planets transiting pipeline targets. The scale of the systematic errors in the light curves may be multiple orders of magnitude larger than the transit signatures of such planets.

Systematic error correction is performed in PDC by identifying signatures in the raw flux light curves that are correlated with ancillary engineering and pipeline data and temporal motion polynomial sequences. Ancillary data is first synchronized to mid-cadence timestamps of the science data. A least squares fitting algorithm is employed, utilizing the SVD of the ancillary design matrix. The projection of raw flux for each target into the column space of the design matrix

is based on the rank of the design matrix and is performed in a computationally efficient and numerically stable manner. Uncertainties are propagated for each target given the linear transformation of raw to cotrended flux, although the full covariance matrix for the cotrended flux cannot be computed due to memory constraints.

A flow chart describing the systematic error correction algorithm is shown in Figure 4. The synchronized time series (to the cadence timestamps) for the respective ancillary engineering and pipeline data mnemonics and temporal motion polynomial coefficient sequences are packed into the columns of a design matrix. The length of each time series (and hence the number of rows in the design matrix) is equal to the number of cadences in the unit of work. The mean value is subtracted from each synchronized time series, and each is then divided by its maximum absolute value for the purpose of numerical conditioning. A constant column (containing all ones) is included in the matrix. Gapped values are interpolated so that the columns may be subsequently filtered into bandpass components. There cannot be any gaps in the synchronized data, however, on cadences with valid science data.

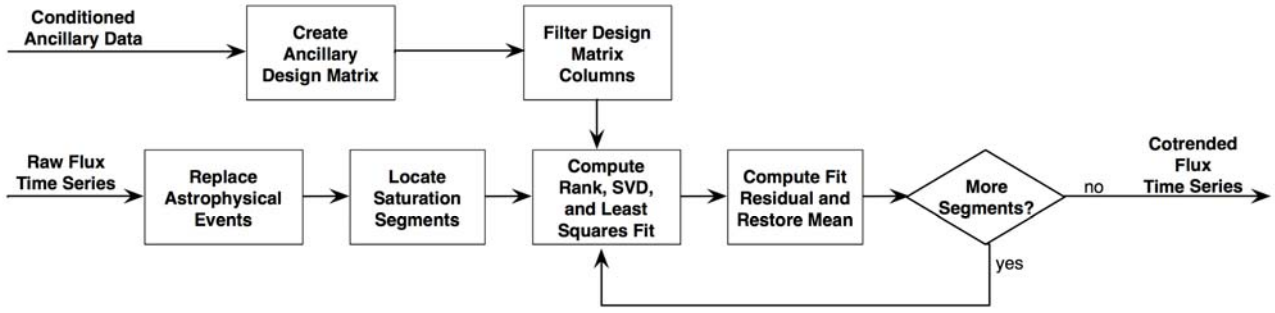

Figure 4: Flow chart for the systematic error correction algorithm. The design matrix is generated and filtered once for all targets in the unit of work. The Singular Value Decomposition (SVD) and least squares projection are also performed once for all targets (as long as they have matching data gaps). Targets with saturation segments, however, must be corrected segment by segment. Saturated segments are demarcated by abrupt changes in curvature as pixels in the associated optimal aperture enter or exit saturation.

The columns of the design matrix (with the exception of the constant column) are then filtered into selectable bandpass (lowpass, midpass, and highpass) components. This permits correction of systematic errors by separately identifying signatures in the target light curves that are correlated with the ancillary data on multiple time scales. The bandpass components are obtained in a cascade of Savitzky-Golay filters. Flux for each target is first filtered into lowpass and highpass components, and the initial highpass component is optionally filtered again into midpass and highpass components. The filter orders and durations (which determine the frequency cutoffs) are tuned separately for long and short cadence science data processing. Generation of the design matrix and filtering of the columns is performed once per PDC unit of work.

Astrophysical events (such as large transits, eclipses, flares, and microlensing events) are identified in the raw flux light curves prior to performing the cotrend fit, and are replaced temporarily with values interpolated across the cadences of the respective events. Random noise is added to the interpolated values based on the statistics of the light curves excluding the astrophysical events. The purpose of this is to prevent astrophysical events from perturbing the fit of the synchronized ancillary data to the raw flux light curves. The intent of PDC is to fit (and remove) systematic error signatures in the science data; it is not the intent to fit the astrophysical events.

Large systematic effects in the light curves due to thermal transients (following safe modes and monthly downlinks) and photometer attitude adjustments may be inadvertently misidentified as astrophysical events. If they are subsequently masked from the least squares fitting, then they cannot be corrected. To resolve this problem, the astrophysical events are vetted against the known spacecraft and data anomalies that are provided as input to all pipeline modules. If an identified event occurs on or near a known anomaly (monthly downlink, safe mode, or commanded attitude adjustment), the event is not replaced prior to cotrending. This unfortunately represents an engineering tradeoff. True astrophysical events that occur on or near major spacecraft and data anomalies are compromised so that those same anomalies may be corrected in the flux for most targets.

After astrophysical events have been identified and replaced temporarily for all targets, any saturated time series segments are located for each target that is sufficiently bright to saturate the CCD detectors. Saturated segments are demarcated by changes in the curvature of the respective flux time series (with astrophysical events removed). A Savitzky-Golay filter is utilized to compute the numerical second derivatives and a threshold is applied for the bright targets based on the statistics of the second derivative time series. If breakpoints are identified in the flux for saturated targets, those targets are separately cotrended segment by segment after processing has completed for all of the targets without saturation breakpoints.

Cotrending is performed with a linear least squares fit of the filtered ancillary design matrix columns to the raw flux time series for each target. Gaps are first squeezed from the raw flux light curves and the associated rows of the design matrix. The data gaps must match for targets that are cotrended at once. If that is not true for all targets in the unit of work, then cotrending must be performed separately on subsets of the targets that do have matching data gaps. Without any loss of generality, given the design matrix A and a raw flux time series f_{RAW}, we seek to find the least squares solution to the set of linear equations:

$$Ax = f_{RAW} \tag{1}$$

Let the dimension of A be mxn and the reduced SVD of A be denoted by

$$A = USV' \tag{2}$$

where U has dimension mxn, S is a diagonal matrix of singular values with dimension nxn, and V also has dimension nxn if $m > n$, as is generally the case in PDC. It may then be shown with matrix algebraic manipulation that the least squares solution to (1) is given by

$$x = VS^{-1}U' f_{RAW} \tag{3}$$

In PDC, we are more interested in the fit to the raw flux, however, than we are in the actual fit coefficients x. Given equations (2) and (3), we may compute the least squares fit of the filtered ancillary data to the raw flux light curve by

$$f_{FIT} = Ax = USV'(VS^{-1}U' f_{RAW}) = UU' f_{RAW} \tag{4}$$

The projection of the raw flux into the column space of the design matrix depends only on the unitary matrix U. If the design matrix A is not full rank (i.e., the columns of the design matrix are not independent), then we seek to limit the dimension of the least squares fit. If the rank of the design matrix A is denoted by r, and U_r denotes the first r columns of U, then in PDC the least squares fit is performed as follows:

$$f_{FIT} = U_r U_r' f_{RAW} \tag{5}$$

The residual between the raw flux (with astrophysical events restored) and the fitted flux determines the cotrended flux f_{COT}. The mean raw flux level μ_{RAW} is also included as follows:

$$f_{COT} = f_{RAW} - f_{FIT} + \mu_{RAW} = (I - U_r U_r') f_{RAW} + \mu_{RAW} \tag{6}$$

Propagation of uncertainties from raw to cotrended flux is straightforward in principle. Memory constraints prevent computation of the full covariance matrix for the cotrended flux, which has dimension mxm where m is the number of cadences in the unit of work. If C_{RAW} and C_{COT} denote the covariance matrices for temporal samples of the raw and cotrended flux time series for a given target, the uncertainties may be propagated (disregarding the uncertainty in the mean level which may be considered to be negligible) by

$$C_{COT} = T_{COT} C_{RAW} T_{COT}' \tag{7}$$

where the transformation T_{COT} is defined by

$$T_{COT} = I - U_r U_r' \tag{8}$$

Due to the aforementioned memory constraint, only the diagonal elements of the covariance matrix C_{COT} are computed in PDC from the diagonal covariance matrix C_{RAW}. The uncertainties in the cotrended flux are given by the square root of

the respective diagonal elements of the covariance matrix C_{COT}. Diagonal elements of C_{RAW} are squares of the uncertainties in the raw flux time series produced by PA.

It has been stated earlier that the intent of cotrending is to fit and remove systematic signatures in the data and not to fit the astrophysical events (such as transits, eclipses, and flares). To that end, an attempt is made to mask such events from the least squares fitting process. The situation becomes complicated, however, when native variability of the targets is considered. The least squares combination of filtered ancillary data may corrupt variability that is inherent in the stellar targets, and in some cases it has been observed to remove the variability completely.

After all targets have been cotrended as described above, an attempt is made to identify variable targets in the unit of work. Those targets for which the center-to-peak flux variation is observed to exceed a specified threshold are flagged. The typical variability threshold in the pipeline is set to 0.5% of the median target flux level. Coarse detrending (accounting for known spacecraft and data anomalies, thermal transient characteristics, and DVA) is performed on the raw flux light curves for the variable targets and an attempt is made to fit the detrended flux for each with a superposition of phase shifting harmonics[2,4,8]. The phase shifting harmonics differ from a conventional Fourier representation in that the frequencies are permitted to vary linearly with time. Fitted harmonic content is removed from the raw flux for each of the variable targets and saved for restoration later in PDC. The residual flux light curves with harmonic content removed are then subjected to the cotrending process as before to remove systematic effects. The harmonic content is identically zero in cases where phase shifting harmonics cannot be fitted to the variable light curves.

Reliable identification of variable targets is difficult in the presence of large data anomalies. Once the apparently variable targets have been corrected, a decision is made for each regarding whether to utilize the standard or harmonic free cotrending result going forward. Performance is assessed in each case based on a robust estimate of the ratio of the power at short time scales (defined by module parameter) in the cotrended result to the power at the same time scales in the raw flux. If a target does not appear to be variable (excluding large transits or other astrophysical events) after cotrending without removal of harmonic content and if the standard systematic error correction performance appears to be good, then the standard result is retained. Otherwise, the cotrending result is retained for the residual flux after removal of harmonics. The harmonic content is restored later in PDC.

The final step in the systematic error correction process is to identify targets for which cotrending has not performed to an acceptable degree. Raw flux (after correction of random flux discontinuities) is substituted for cotrended flux for such targets based on a comparison of the performance metric discussed above with a specified performance limit. For these targets, systematic effects are not addressed in PDC because it is not possible to do so without corrupting the inherent character of the light curves. Targets in this category are typically variable stars for which the light curves are not harmonic, or for which the phase shifting harmonics cannot adequately represent the stellar variation.

Figure 5 illustrates systematic error correction performance for a quiet target on module output 7.3 in Q2. This quarter featured a series of significant spacecraft and data anomalies[14]. The raw flux light curve is shown in the upper panel. The least squares fit of the filtered ancillary data to the raw flux is shown in the middle panel. The difference between raw and fitted flux is virtually indistinguishable on the scale shown in the figure. All of the information regarding anomaly-induced flux discontinuities and recovery transients is captured by the ancillary data and motion polynomial sequences utilized for correction of the systematic errors. The fit residual is shown in the lower panel with the mean flux level restored. The scale of the raw flux artifacts for this target due to the various spacecraft and data anomalies is multiple orders of magnitude larger than the transit depth of an Earth-size planet orbiting a Sun-like star.

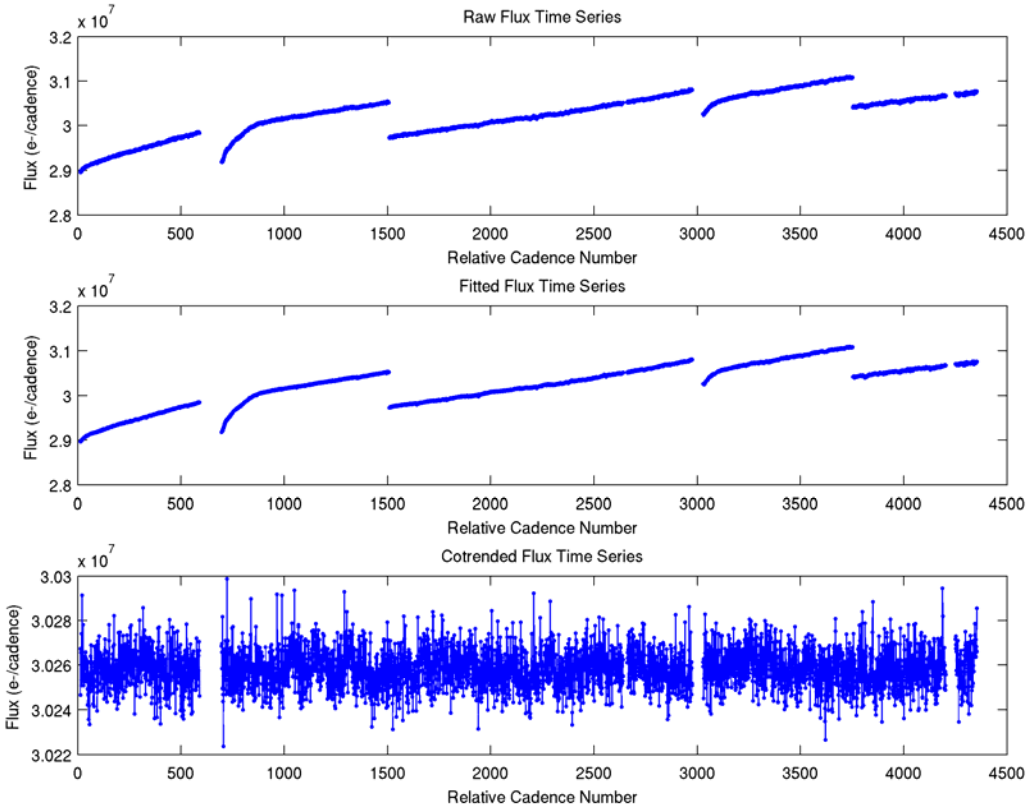

Figure 5: The raw flux time series for a quiet long cadence Q2 target is shown in the upper panel. A number of significant spacecraft and data anomalies are clearly visible in the light curve. These include thermal transient at the start of the quarter, safe mode recovery transient near cadence 700, attitude adjustment near cadence 1500, Earth point recovery transient near cadence 3000, and attitude adjustment near cadence 3750. There is also a loss of fine point data gap near the end of the quarter. The least squares fit of the synchronized ancillary data to the raw flux is shown in the middle panel. The fit residual with mean level restored is shown in the lower panel.

Systematic error correction performance for a Q1 target[14] on module output 2.1 is shown in Figure 6. This module output has been observed to be particularly sensitive to focus changes in the photometer. A 200-cadence segment of the raw flux light curve is shown in the upper panel. The least squares fit of the filtered ancillary data to the raw flux is shown in the middle panel. The oscillations in the raw flux are caused by the cycling of a reaction wheel heater outside the photometer. The mechanism by which heat generated outside the photometer causes the focus to change is not yet understood. The cotrending process nevertheless produces a fit that essentially tracks the flux oscillations, and the oscillations are significantly reduced in the residual flux shown in the lower panel.

It should be noted that the peak-to-peak variation in flux oscillations for this target are on the order of 0.1% (before removal of excess flux due to aperture crowding), which is approximately equivalent to the transit depth of a Neptune-size planet orbiting a Sun-like star, and ten times the transit depth of an Earth-size planet orbiting a Sun-like star. Hence, even small-scale systematic effects in the flight data dwarf the transit signatures that the *Kepler Mission* has been designed to detect.

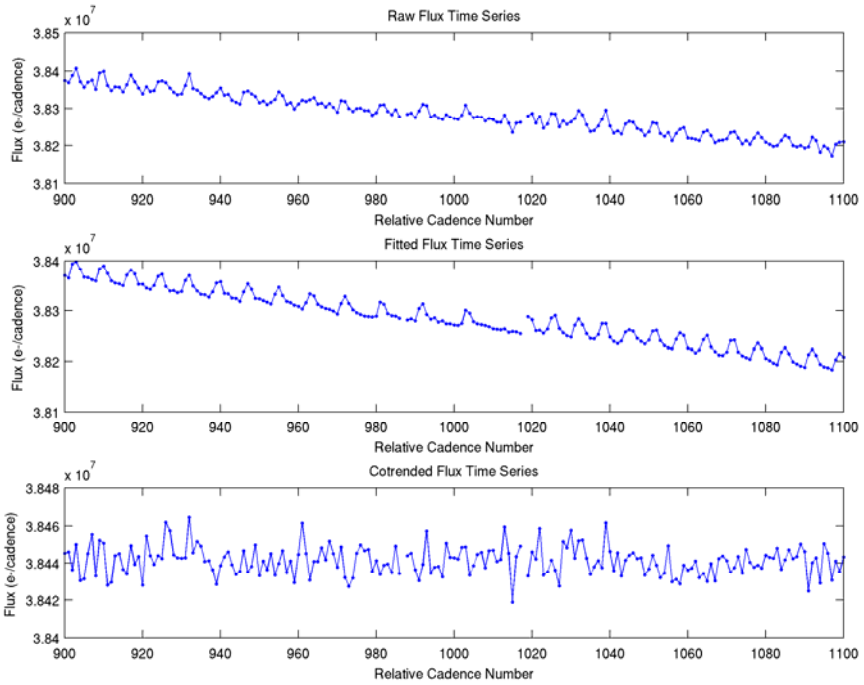

Figure 6: 200-cadence segment of a raw flux time for a quiet long cadence target in Q1 is shown in the upper panel. Oscillation in the light curve is due to cycling of a reaction wheel heater outside the photometer. The least squares fit of the synchronized ancillary data to the raw flux is shown in the middle panel. The corresponding segment of the fit residual with mean level restored is shown in the lower panel.

3.3 Excess flux removal

Optimal apertures[7] may include flux from sources other than the targets with which they are associated. It is necessary to estimate and remove excess flux in order that the relative transit depths in the corrected flux light curves produced by PDC faithfully represent the true transit depths of the target system. Otherwise, transits will be systematically diluted and planet radii will be systematically underestimated in the pipeline and by consumers of pipeline data.

The so-called *crowding metric* is computed in the TAD module[13] for each target and is defined to be the fraction of flux in the optimal aperture due to the target itself. The crowding metric is a scalar value representing the average aperture crowding over the effective date range of a given target table. Crowding may be dynamic, however, changing as background sources of light enter and exit the optimal aperture due to DVA. To that extent, computation of the crowding metric and removal of excess flux in PDC are only approximate and may need to be revisited in the future.

For each target, a constant excess flux level is estimated over the duration of the unit of work based on the crowding metric for the given target and the median value of the cotrended flux for that target. The excess flux level is then subtracted from the cotrended flux value for every cadence with valid science data. Let the crowding metric for a given target be denoted by α and the median value of the cotrended flux time series be denoted by m_{COT}. The constant excess flux level f_{XS} due to crowding in the optimal aperture is then estimated by

$$f_{XS} = (1-\alpha)m_{COT} \tag{9}$$

The crowding-corrected flux time series f_{COR} is finally determined by subtracting the excess flux level from the cotrended flux time series as follows:

$$f_{COR} = f_{COT} - f_{XS} = f_{COT} - (1-\alpha)m_{COT} \tag{10}$$

Uncertainties are not propagated for the excess flux correction. Uncertainties for the crowding-corrected flux values are set equal to those of the cotrended flux. The uncertainty in the median flux estimate over all cadences is assumed to be negligible in comparison with the uncertainty in the systematically error-corrected flux value for any given cadence.

4. SUMMARY AND CONCLUSIONS

The primary tasks of the PDC module of the *Kepler* SOC Science Processing Pipeline are to correct systematic and other errors in the raw flux light curves, remove excess flux in the light curves due to crowding of the respective target apertures, and condition light curves for the transiting planet search by identifying and removing flux outliers and filling data gaps. We first presented an overview of the PDC module. We have then shown how random flux discontinuities are identified and corrected. We have discussed the correction of systematic errors and shown examples of the correction of both large and small scale systematic effects in the flight data. We have described removal of excess flux from light curves due to background sources in the respective target apertures. We have also described the propagation of uncertainties from raw to corrected flux light curves in PDC. Corrected flux light curves produced in PDC are exported to the MAST and made available to the general public in accordance with the NASA/Kepler data release policy. It has proven very difficult to perform the PDC error corrections properly for all targets while still preserving their native variability. The shortcomings of the current PDC release (SOC 6.1) are recognized, and it remains a work in progress.

ACKNOWLEDGMENTS

The authors would like to thank Greg Orzech, Jeffrey Van Cleve, Michael Fanelli and Bill Wohler for reviewing this paper.

Funding for the *Kepler Mission* has been provided by the NASA's Science Mission Directorate.

REFERENCES

[1] Koch, D.G., et al., "*Kepler Mission* Design, Realized Photometric Performance, and Early Science," Astrophysical Journal Letters, 713 (2), L79-L86 (2010).
[2] Jenkins, J.M., et al., "Initial Characteristics of *Kepler* Long Cadence Data for Detecting Transiting Planets," Astrophysical Journal Letters, 713 (2), L120-L125 (2010).
[3] Gilliland, R.L., et al., "Initial Characteristics of *Kepler* Short Cadence Data," Astrophysical Journal Letters, 713 (2), L160-L163 (2010).
[4] Jenkins, J.M., et al., "Overview of the *Kepler* Science Processing Pipeline," Astrophysical Journal Letters, 713 (2), L87-L91 (2010).
[5] Middour, C., et al., "*Kepler* Science Operations Center architecture," Proc. SPIE 7740, in press (2010).
[6] Quintana, E.V., et al., "Pixel-level calibration in the *Kepler* Science Operations Center pipeline," Proc. SPIE 7740, in press (2010).
[7] Twicken, J.D., et al., "Photometric analysis in the *Kepler* Science Operations Center pipeline," Proc. SPIE 7740, in press (2010).
[8] Jenkins, J.M., et al., "Transiting planet search in the *Kepler* pipeline," Proc. SPIE 7740, in press (2010).
[9] Wu, H., et al., "Data validation in the *Kepler* Science Operations Center pipeline," Proc. SPIE 7740, in press (2010).
[10] Tenenbaum, P., et al., "An algorithm for fitting of planet models to *Kepler* light curves," Proc. SPIE 7740, in press (2010).
[11] Klaus, T.C., et al., "The *Kepler* Science Operations Center pipeline framework," Proc. SPIE 7740, in press (2010).
[12] Klaus, T.C., et al., "The *Kepler* Science Operations Center pipeline framework extensions," Proc. SPIE 7740, in press (2010).
[13] Bryson, S.T., et al., "Selecting pixels for *Kepler* downlink," Proc. SPIE 7740, in press (2010).
[14] Haas, M.R., et al., "*Kepler* Science Operations," Astrophysical Journal Letters, 713 (2), L115-L119 (2010).