Transiting planet search in the *Kepler* pipeline

Jon M. Jenkins°[a], Hema Chandrasekaran[a,b], Sean D. McCauliff[c], Douglas A. Caldwell[a], Peter Tenenbaum[a], Jie Li[a], Todd C. Klaus[c], Miles T. Cote[d], Christopher Middour[c]

[a]SETI Institute/NASA Ames Research Center, M/S 244-30, Moffett Field, CA USA 94305
[b]Lawrence Livermore National Laboratory, P.O. Box 808, L-478, Livermore, CA USA 94551
[c]Orbital Sciences Corporation/NASA Ames Research Center, M/S 244-30, Moffett Field, CA USA 94305
[d]NASA Ames Research Center, M/S 244-30, Moffett Field, CA USA 94305

ABSTRACT

The *Kepler Mission* simultaneously measures the brightness of more than 160,000 stars every 29.4 minutes over a 3.5-year mission to search for transiting planets. Detecting transits is a signal-detection problem where the signal of interest is a periodic pulse train and the predominant noise source is non-white, non-stationary (1/f) type process of stellar variability. Many stars also exhibit coherent or quasi-coherent oscillations. The detection algorithm first identifies and removes strong oscillations followed by an adaptive, wavelet-based matched filter. We discuss how we obtain super-resolution detection statistics and the effectiveness of the algorithm for *Kepler* flight data.

Keywords: *Kepler Mission*, exoplanet, transit, detection algorithm

1. INTRODUCTION

The *Kepler Mission* continuously observes ∼160,000 target stars in *Kepler's* 115-square-degree field of view, seeking to discover Earth-like planets transiting Sun-like stars by detecting photometric signatures of transits.[1,2] The CCDs are read out every 6.52 s and co-added for 29.4 minutes*, after which the pixels of interest for each target star are compressed and stored on board the spacecraft's Solid State Recorder. Approximately 1,500 samples are stored on board for each pixel of interest per month for a total of $\sim 6 \times 10^6$ pixels. These data are downlinked at monthly contacts with the Deep Space Network and travel through the Ground System to the *Kepler* Science Operations Center located at NASA Ames Research Center where they are processed to derive brightness measurements of each star for each time step and to search for planetary transit signatures. *Kepler's* mission will last at least 3.5 years† to permit detection of at least three transits of Earth-size planets orbiting Sun-like stars in the habitable zone, that range of distances at which liquid water could pool on the surface.

The *Kepler* Science Processing Pipeline (hereafter referred to as the Pipeline) is designed to process science data collected from the *Kepler* Photometer to furnish calibrated pixels, raw and systematic error corrected flux time series, and centroid time series.[3–5] Figure 1 shows the processing sequence. Raw pixels are calibrated by the Calibration (CAL) component to remove on-chip artifacts such as shutterless readout smear and perform standard astronomical corrections such as bias and dark current removal.[6] Photometry is extracted from optimal apertures placed about each target star image by the Photometric Analysis (PA) component,[7] which estimates and removes background flux and identifies and removes cosmic ray hits. PA also centroids each target star image to provide measurements of the locations of each target star on each frame, enabling reconstruction of the photometer pointing and elimination of false positives. The component Pre-search Data Conditioning (PDC) removes signatures in the light curves correlated with instrumental variables such

°Further author information: Send correspondence to J.M.J.: E-mail: Jon.Jenkins@nasa.gov
©2010 Society of Photo-Optical Instrumentation Engineers. One print or electronic copy may be made for personal use only. Systematic reproduction and distribution, duplication of any material in this paper for a fee or for commercial purposes, or modification of the content of the paper are prohibited.

*Each 29.4-min data integration interval is called a Long Cadence (LC); the data is called LC data to distinguish it from data collected at ∼one-minute intervals, called Short Cadence data.

†The principal limit to *Kepler's* mission lifetime is the supply of hydrazine fuel used to de-spin the reaction wheels every three days, of which there is sufficient supply to last for six or more years.

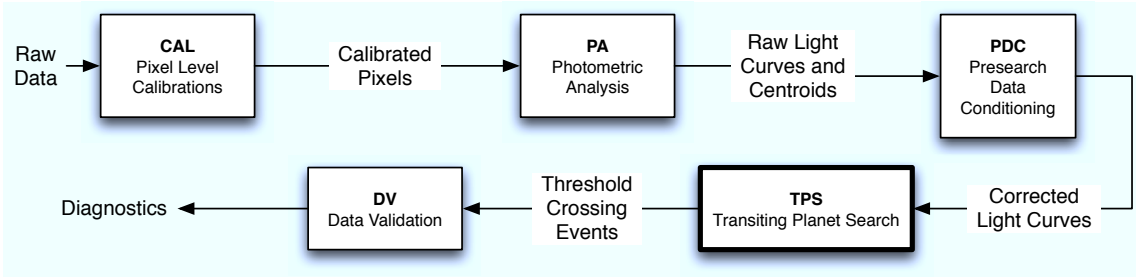

Figure 1. Data Flow diagram for the SOC Science Processing Pipeline. Several processing steps must be completed before TPS, which lies at the heart of the Pipeline, can perform its search for signatures of transiting planets. First, CAL calibrates the raw pixels downlinked from the spacecraft to remove on-chip artifacts and to place the measurements on a linear scale with estimated uncertainties. Second, PA identifies and removes cosmic rays from the pixel time series, estimates and subtracts the background flux and then sums the resulting pixel values over the photometric aperture underneath each target star image. Third, PDC identifies and removes signatures of systematic effects in the photometric time series, such as changes in pointing or focus, and fills any gaps to condition the time series for TPS. TPS then searches the corrected flux time series for signatures of periodic pulse trains indicative of transiting planets. Threshold-crossing events flagged by TPS are examined in detail by DV to establish or break confidence in the transit-like features as planetary signatures.

as pointing offsets and focus changes and removes outliers and step discontinuities due to pixel sensitivity drops.[8] At this point Transiting Planet Search (TPS) applies a wavelet-based, adaptive matched filter to identify transit-like features with durations in the range of 1 to 16 hours.[9] Light curves whose maximum folded detection statistic exceeds 7.1σ are designated Threshold Crossing Events (TCEs) and subjected to a suite of diagnostic tests in Data Validation (DV) to fit a planetary model to the data and to establish or break confidence in the planetary nature of the transit-like events.[10,11]

Monitoring the photometric precision obtained by *Kepler* is a high priority and is obtained on a monthly basis as a by-product of the noise characterization performed by TPS. The photometric precision metric is called Combined Differential Photometric Precision (CDPP) and is defined as the root mean square (RMS) photometric noise on transit timescales.[2] Each month CDPP, along with a suite of performance metrics developed during processing as the data proceed through the pipeline, is monitored and reported by the Photometric Performance Assessment (PPA) component.[12] The SOC is required to monitor CDPP for transit durations of 3, 6, and 12 hours. The typical duration of transit varies from a few hours for close-in planets to 16 hours for a Mars-size orbit.[2] Thus, TPS contributes in two primary ways: 1) it produces 3-, 6- and 12-hour CDPP estimates for each star each month, and 2) it searches for periodic transit pulse sequences.

The *Kepler* spacecraft rotates by 90° approximately every 93 days to keep its solar arrays directed towards the Sun.[5] The first 10 days of science data obtained as the last activity during commissioning is referred to as Q0. There were $\sim$34 days of observations during Q1 following Q0 at the same orientation. Subsequent quarters are referred to as Q2, Q3, etc., and these each contain $\sim$90 days of observations. Transit searches are performed nominally every three months after each quarterly data set has been downlinked to the ground in as complete a state as possible and processed from CAL through PDC. As illustrated in Figure 2, there are three major subcomponents in TPS needed to facilitate the full transit search. Since each target star falls on a different CCD in each quarter, TPS needs to combine the quarterly segments together in such a way as to minimize the edge effects and maximize the uniformity of the apparent depths of planetary transit signatures across the entire data set. The first component of TPS "stitches" the quarterly segments of each flux time series together before presenting it to the transit detection component. The second component characterizes the observation noise as a function of time from a transit's point of view and correlates a transit pulse with the time series to estimate the likelihood that a transit is occurring at each point in time. These tasks are accomplished by a wavelet-based, adaptive matched filter as per Ref. 9. The third and final component of TPS uses the noise characterization and correlation time series to search for periodic transit pulse sequences by folding the data over trial orbital periods spanning the range from one day to the length of the current data set.

A number of issues identified since science operations commenced on May 12, 2009 have required significant modifications to the Science Pipeline and to TPS. Many target stars exhibit coherent or quasi-coherent oscillations. The wavelet-based detector was designed to deal with solar-like stellar variability for which any such oscillations occur on timescales much shorter than the LC observation interval of 29.4 minutes, and while it works well for broad-band, non-white noise processes, it is not optimal for coherent background noise that is concentrated in the frequency domain. To mitigate this

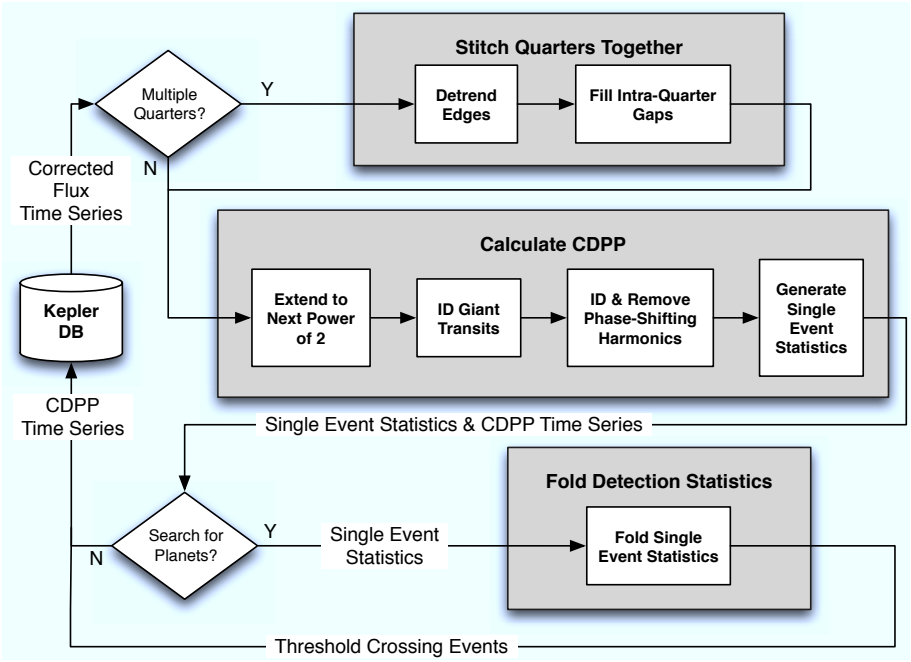

Figure 2. Block diagram for TPS. When TPS is run in transit search mode, the data from the beginning of the mission to the most recent data are first stitched together at the boundaries between quarterly segments. Single-event statistics are generated over a range of transit durations from 1 to 12 hours which are then folded at trial periods from one day to the length of the observations. A by-product of the generation of the single-event statistics is a measure of the photometric precision for each star, called CDPP. Stars for which the multiple-event statistics exceed 7.1σ are designated Threshold Crossing Events (TCEs) and persisted to the *Kepler Data Base*, along with the CDPP time series and other information, such as the epoch and period of the most likely transit pulse train. For monthly data sets, TPS measures the photometric precision achieved for as many as 169,000 target stars, in which case the first and third subcomponents are skipped.

phenomena, TPS has been modified to include the ability to identify and remove phase-shifting, harmonic components. This code is also used by PDC to condition the flux time series prior to identifying and removing instrumental signatures. The algorithm is based on that of Ref. 13. This step is performed after the quarterly segments have been conditioned, just prior to the noise characterization.

This article is organized as follows: Stitching multiple quarters of data together is presented in Sec. 2. Sec. 3 discusses the identification and removal of harmonic processes and the identification of deep transits and eclipses, and summarizes the wavelet-based, adaptive matched filter. Sec. 4 describes the software used to fold the single-event statistics to search for periodic transit signatures. The conclusion and summary of future work is given in Sec. 5.

2. STITCHING MULTIPLE QUARTERS

This subcomponent of TPS is under development, as up to now we have only exercised TPS on individual quarterly light curves. This feature will be completed as part of the next SOC development cycle and is scheduled for release early in 2011 as part of the SOC 7.0 release. Therefore, the details of implementation for this subcomponent are subject to change as we implement and test it.

The first step in this process is to detrend the edges of each quarterly segment in order to "stitch" together the segments into one continuous time series. The median of each segment is first subtracted from the light curve and divided into it to obtain a time series that represents the time evolution of fractional changes in the brightness of the target star. Next, a line is robustly fitted to the first and last day of each quarter. The slopes and values of the fitted lines at the ends of the segment then completely determine the coefficients of a cubic polynomial that is then subtracted from the segment. Depending on the details of the stellar variability exhibited in the light curve, the cubic polynomial may introduce large excursions from the mean flux level (now zero). To identify if this is the case, the residual is tested against two statistical

criteria used to determine if the residual is well-modeled as a zero-mean stochastic process. The number of positive data points is compared to the number of negative data points, and then the area under the positive points is compared to the negative area under the negative points (essentially, the sum of the positive points compared to the sum of the absolute value of the negative points). If the negative and positive metrics of these tests are within a specified tolerance of each other (typically taken to be 20%), then TPS proceeds to the next step. If these criteria are not satisfied, then TPS robustly fits constrained polynomials to the residual whose value and slope is zero at the end points of the segment starting with a quartic polynomial, and retests the residual against these criteria. The order of the polynomial is increased until either the criteria are met or a specified maximum polynomial order (10) is reached.

All polynomials $p(x)$ whose values and slopes are zero at the endpoints 0 and 1 have the form

$$p(x) = x^2(x-1)^2 q(x), \tag{1}$$

where x is the time t normalized as $x = (t-t_1)/(t_N-t_1)$ by the first and last time tags t_1 and t_N, and $q(x) = c_0 + c_1 x^2 + \ldots + c_{M-1} x^{M-1}$ is a normal Mth-order polynomial. The design matrix for solving for such a constrained polynomial is the standard one whose rows are multiplied by the constraint polynomial terms evaluated at each time tag, $x^2(x-1)^2$. That is, the elements of design matrix A are defined by

$$A_{i,j} = x_i^2 (x_i-1)^2 x_i^j, \tag{2}$$

where $i = 1,\ldots,N$ and $j = 0,\ldots,M-1$.

The next step in stitching the quarterly segments together is to fill the gaps between them. Gaps within quarters are filled by PDC prior to TPS. Short data gaps of a few days or less are filled by an autocorrelation approach using auto-regressive stochastic modeling as per Ref. 9. Filling longer data gaps, as is necessary for target stars that are not observed each and every quarter, will require other methods. Currently, the long data gap fill method employed by the Pipeline reflects and tapers data on either segment of a gap across the gap and then performs a wavelet analysis to adjust the fill data to make the amplitude of the stochastic variations consistent with those of data adjacent to the gap. For light curves with entire quarterly segments missing we will likely modify the reflection approach to deal with cases where the gap may be longer than the available data on one or the other side of the gap, as will happen for targets observed in Q1 but not Q2, and then observed in Q3. The gap filling is necessary to allow the next subcomponent of TPS to operate: applying a wavelet-based, matched filter to either calculate CDPP on a monthly or quarterly basis, or to furnish the single-event statistics for the full transit search.

3. CALCULATING CDPP

This subcomponent of TPS performs the noise characterization central to the task of detecting transiting extrasolar planets. Prior to applying the wavelet-based, adaptive matched filter it is necessary to extend the flux time series to the next power of 2 as this detection scheme invokes Fast Fourier Transforms (FFTs). It is also necessary to screen out coherent and quasi-coherent harmonic signals in the flux time series, as well as deep transits and eclipses, for which the original wavelet-based approach is not well suited.

The first step is to identify strong transit signatures, then identify and remove deep transits, and finally, the time series can be extended to the next power of 2 via methods in Ref. 9.

3.1 Identifying Giant Transiting Planets and Eclipses

The transit detection scheme baselined for the Pipeline is designed to search for weak transit signatures buried in solar-like variability and observation noise. The noise characterization will sense the presence of strong transit signatures from giant planets or eclipsing binaries and tends to "annihilate" them as part of the pre-whitening step in the detection process. To identify such signatures, TPS applies Akaike's Information Criterion[14] to fit a polynomial to each light curve and identifies clusters of negative residuals that are many median absolute deviations from the fit. Such points are removed and the process is repeated until no additional sets of consecutive, highly negative points are identified. The residuals are subjected to a search for harmonic signatures.

3.2 Identifying and Removing Phase-Shifting Harmonics

Once the deep transits and eclipses have been identified, the cadences containing such events are temporarily filled using the autocorrelation-based short data gap fill algorithm. The time series is extended to the next power of 2 using the approach of Ref. 9 and a Hanning window-weighted periodogram is formed. The background Power Spectral Density (PSD) of any broadband, non-white noise process in the data is estimated in a two-step process. First, a median filter is applied to the periodogram and then the result smoothed with a moving average window. The median filter ignores isolated peaks in the periodogram. Next this background PSD estimate is divided pointwise into the periodogram. The whitened PSD is then examined for statistically significant peaks, and the frequency bins of such peaks are fed to a nonlinear least squares fitter as the seed values for a fit in the time domain to phase-shifting harmonic signals. These are sinusoids in time that allow for the center frequency to shift linearly in time. For complex harmonic signals, this process can take a while. Figure 3 shows two examples of light curves with strong, coherent harmonic features as they are fitted and removed with this approach. The resulting harmonic-cleaned flux time series is then ready for the wavelet-based matched filter.

3.3 A Wavelet-Based Matched Filter

The optimal detector for a deterministic signal in colored Gaussian noise is a pre-whitening filter followed by a simple matched filter.[15] In TPS we implement the wavelet-based matched filter as per Ref. 9 using Debauchies' 12-tap wavelets.[16] The wavelet-based matched filter uses an octave-band filter bank to separate the input flux time series into different band passes to estimate the PSD of the background noise process as a function of time. This scheme is analogous to a graphic equalizer for an audio system. TPS constantly measures the "loudness" of the signal in each bandpass and then dials the gain for that channel so that the resulting noise power is flat across the entire spectrum. Flattening the power spectrum transforms the detection problem for colored noise into a simple one for white Gaussian noise (WGN), but also distorts transit waveforms in the flux time series. TPS correlates the trial transit pulse with the input flux time series in the whitened domain, accounting for the distortion resulting from the pre-whitening process. This is analogous to visiting a funhouse "hall of mirrors" with a friend of yours and seeking to identify your friend's face by looking in the mirrors. By examining the way that your own face is distorted in each mirror, you can predict what your friend's face will look like in each particular mirror, given that you know what your friend's face looks like without distortion. Let's briefly review the wavelet-based matched filter.

Let $x(n)$ be a flux time series. Then we define the over-complete wavelet transform (OWT) of $x(n)$ as

$$\mathbb{W}\{x(n)\} = \{x_1(n),\ x_2(n),\ \ldots,\ x_M(n)\}, \tag{3}$$

where

$$x_i(n) = h_i(n) * x(n),\ i = 1,\ 2,\ \ldots,\ M, \tag{4}$$

and '$*$' denotes convolution, and $h_i(n)$ for $i = 1, \ldots, M$ are the impulse responses of the filters in the filter bank implementation of the wavelet expansion with corresponding frequency responses $H_i(\omega)$ for $i = 1, \ldots, M$.

Figure 4 is a signal flow graph illustrating the process. The filter, H_1, is a high-pass filter that passes frequency content from half the Nyquist frequency, $f_{Nyquist}$, to the Nyquist frequency ($[f_{Nyquist}/2, f_{Nyquist}]$). The next filter, H_2, passes frequency content in the interval $[f_{Nyquist}/4, f_{Nyquist}/2]$, as illustrated in Figure 5. Each successive filter passes frequency content in a lower bandpass until the final filter, H_M, the lowest bandpass, which passes DC content as well. The number of filters is dictated by the number of observations and the length of the mother wavelet filter chosen to implement the filterbank. In this wavelet filter bank there is no decimation of the outputs so that there are M times as many points in the wavelet expansion of a flux time series, $\{x_i(n)\}$, $i = 1, \ldots, M$, as there were in the original flux time series $x(n)$. This representation has the advantage of being shift invariant, so that we need only compute the wavelet expansion of a trial transit pulse, $s(n)$, once. The noise in each channel of the filter bank is assumed to be white and Gaussian and its power is estimated as a function of time by a moving variance estimator (essentially a moving average of the squares of the data points) with an analysis window chosen to be significantly longer than the duration of the trial transit pulse.

The detection statistic is computed by multiplying the whitened wavelet coefficients of the data by the whitened wavelet coefficients of the transit pulse:

$$T = \frac{\tilde{\boldsymbol{x}} \cdot \tilde{\boldsymbol{s}}}{\sqrt{\tilde{\boldsymbol{s}} \cdot \tilde{\boldsymbol{s}}}} = \frac{\sum_{i=1}^{M} 2^{-\min(i,M-1)} \sum_{n=1}^{N} [x_i(n)/\hat{\sigma}_i(n)]\,[s_i(n)/\hat{\sigma}_i(n)]}{\sqrt{\sum_{i=1}^{M} 2^{-\min(i,M-1)} \sum_{n=1}^{N} s_i^2(n)/\hat{\sigma}_i^2(n)}}, \tag{5}$$

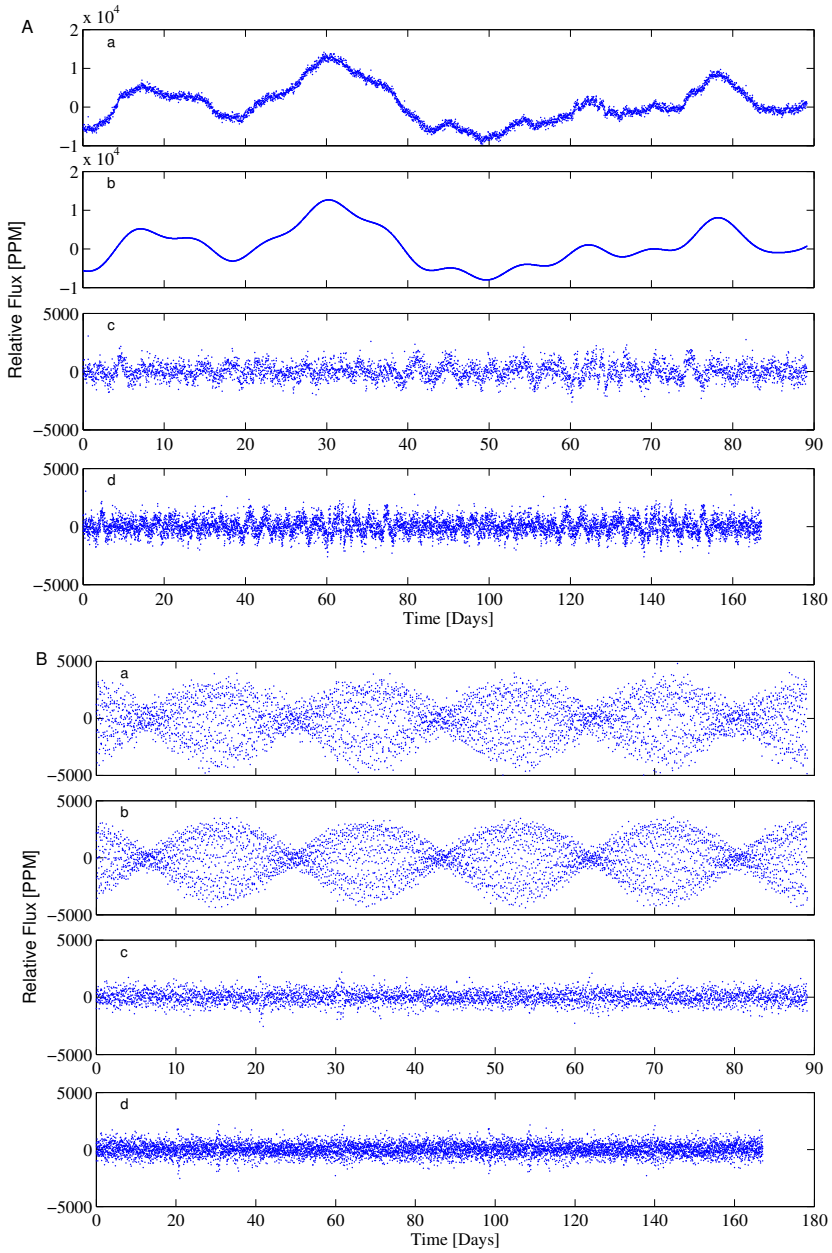

Figure 3. Harmonic removal and extension of two flux time series. A: a star with low-freqency oscillations; B: a star with high-frequency oscillations and amplitude modulation. a: original flux time series; b: detected harmonic signature; c: flux time series with harmonics subtracted; d: harmonic-free time series extended to 8192 samples.

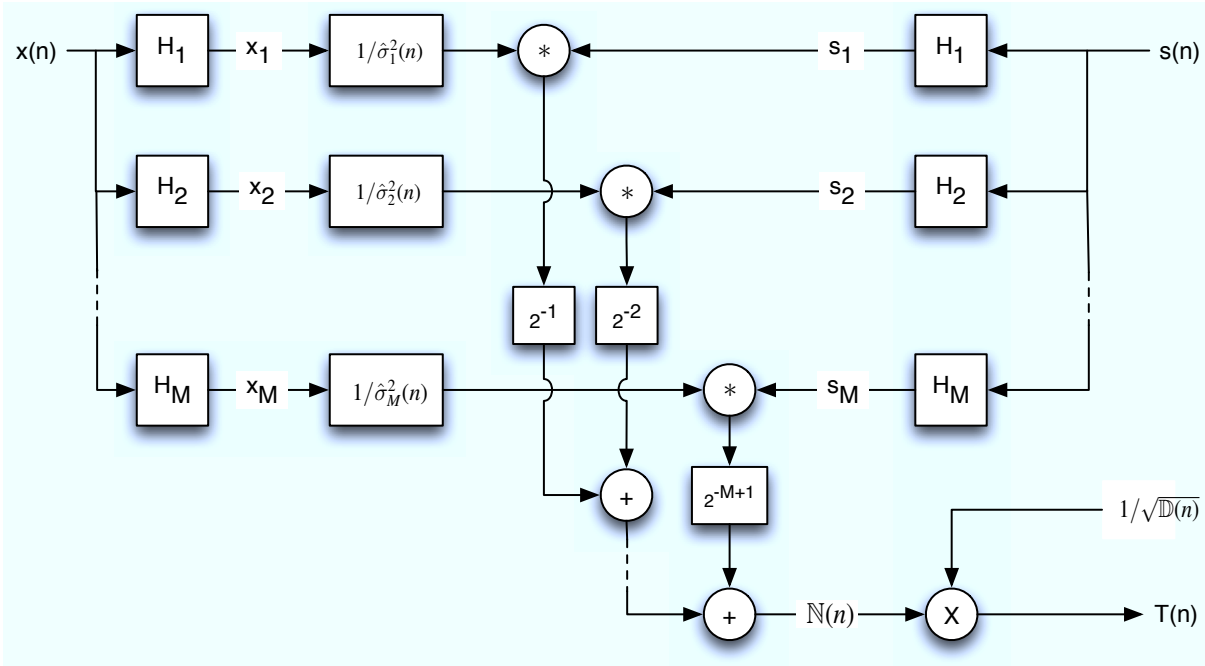

Figure 4. Signal flow diagram for TPS. The wavelet-based matched filter is implemented as a filter bank with bandpass filters $\mathrm{H}_1, \ldots, \mathrm{H}_M$ progressing from high frequencies to low frequencies. The flux time series, $x(n)$, is expanded into M time series $x_i(n)$, for $i = 1, \ldots, M$. Noise power, $\sigma_i^2(n)$, $i = 1, \ldots, M$ is estimated for each bandpass and then divided into the channel time series, $x_i(n)$, in order to whiten the flux time series in the wavelet domain. The trial transit pulse is processed through a copy of the filter bank and convolved with the doubly pre-whitened flux time series in each bandpass. Parseval's theorem for undecimated, wavelet representations allows us to combine the results for each bandpass together to form the numerator term, $\mathbb{N}(n)$, of Eq. 7. A similar filterbank arrangement is used to furnish $\mathbb{D}(n)$ from Eq. 7 by replacing the flux time series $x(n)$ in this flow diagram with the trial transit pulse $s(n)$, and by using the same bandpass noise estimates to inform the pre-whitening. The single-event detection statistic, $T(n)$, is obtained by dividing the correlation term, $\mathbb{N}(n)$ by the square root of the denominator term, $\mathbb{D}(n)$.

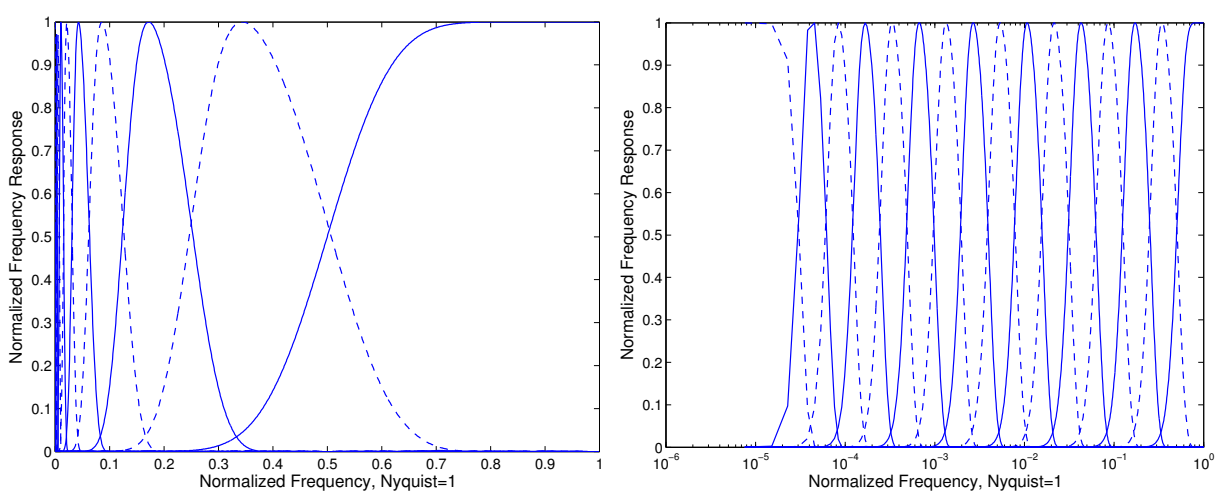

Figure 5. Frequency responses of the filters in the octave-band filterbank for a wavelet expansion corresponding to the signal flow graph in Figure 4 using Debauchies' 12-tap filter. Left: frequency responses on a linear frequency scale. Right: frequency response on a logarithmic frequency scale, illustrating the "constant-Q" property of an octave-band wavelet analysis.

where the time-varying channel variance estimates are given by

$$\hat{\sigma}_i^2(n) = \frac{1}{2^i K + 1} \sum_{k=n-2^{i-1}K}^{n+2^{i-1}K} x_i^2(k),\ i = 1,\ \ldots,\ M, \tag{6}$$

where each component $x_i(n)$ is periodically extended in the usual fashion and $2K+1$ is the length of the variance estimation window for the shortest time scale. In TPS, K is a parameter set to typically 50 times the trial transit duration.

To compute the detection statistic, $T(n)$, for a given transit pulse centered at all possible time steps, we simply "doubly whiten" $\mathbb{W}\{x(n)\}$ (i. e., divide $x_i(n)$ point-wise by $\hat{\sigma}_i^2(n)$, for $i = 1, \ldots, M$), correlate the results with $\mathbb{W}\{s(n)\}$, and apply the dot product relation, performing the analogous operations for the denominator, noting that $\hat{\sigma}_i^{-2}(n)$ is itself a time series:

$$T(n) = \frac{\mathbb{N}(n)}{\sqrt{\mathbb{D}(n)}} = \frac{\sum_{i=1}^{M} 2^{-\min(i,M-1)}\ [x_i(n)/\hat{\sigma}_i^2(n)] * s_i(-n)}{\sqrt{\sum_{i=1}^{M} 2^{-\min(i,M-1)}\ \hat{\sigma}_i^{-2}(n) * s_i^2(-n)}}. \tag{7}$$

Note that the "–" in $s_i(-n)$ indicates time reversal. The numerator term, $\mathbb{N}(n)$, is essentially the correlation of the reference transit pulse with the data. If the data were WGN then the result could be obtained by simply convolving the transit pulse with the flux time series. The expected value of Eq. 7 under that alternative hypothesis for which $x_i(n) = s_i(n)$ is $\sqrt{\sum_{i=1}^{M} 2^{-\min(i,M-1)}\ \hat{\sigma}_i^{-2}(n) * s_i^2(-n)}$. Thus, $\sqrt{\mathbb{D}(n)}$ is the expected signal-to-noise ratio (SNR) of the reference transit in the data as a function of time. The CDPP estimate is obtained as

$$CDPP(n) = 1 \times 10^6 / \sqrt{\mathbb{D}(n)}, \tag{8}$$

in units of parts per million.

For stars with identified giant planet transits or eclipses, an alternate route is taken to estimate the correlation and expected SNR. The data located in transit are removed and filled by a simple linear interpolation. The resulting time series is then high-pass filtered to remove trends on timescales >3 days and then a simple matched filter is convolved with the resulting time series. A moving variance supplies the information necessary to inform the expected SNR. Figure 6 illustrates the process of estimating CDPP for a star exhibiting strong transit-like features. Once the time-varying power spectral analysis performed by TPS, we can search for periodic transit pulses.

4. FOLDING DETECTION STATISTICS

The third and final stage of TPS is to fold the single-event detection statistics developed in Sec. 3 over the range of potential orbital periods. Applying a matched filter for a deterministic signal with unknown parameters is equivalent to performing a linear least-squares fit at each trial point in parameter space, which for transit sequences is the triple composed of the epoch (or time to first transit), orbital period, and transit duration, $\{t_0, T_p, D\}$. Clearly, we can't test for all possible points so we must lay down a grid in parameter space that balances the need to preserve sensitivity with the need for speed.

As given in Ref. 17, one measure of sensitivity is the correlation coefficient between the model planetary signatures of neighboring points in parameter space. If we specify the minimum correlation coefficient, ρ, required between neighboring models, then we can derive the step sizes in period, epoch, and duration. For the case of simple rectangular pulse trains, a real transit will have a correlation coefficient with the best-matched model of no worse than $\rho + (1-\rho)/2$. The correlation coefficient as a function of the change in epoch, Δt_0, is given by $c(\Delta t_0) = (D - \Delta t_0)/D = 1 - \Delta t_0/D$, where D is the trial transit duration. Similarly, for a change in transit duration we have $c(\Delta D) = (D - \Delta D)/D = 1 - \Delta D/D$, so that $\Delta D = (1-\rho)D$. So for a given minimum correlation coefficient, ρ, we have $\Delta t_0 = (1-\rho)D$. The step size in orbital period, ΔT_p, is strongly influenced by the number of transits expected in the data set at the trial period itself. In this case, $c \approx 1 - N\Delta T_p/4D$, where N is the number of expected transits, or the ratio of the length of the data set to the trial period, so that

$$\Delta T_p = 4(1-\rho)D/N = 4\Delta t_0/N. \tag{9}$$

The default choice for TPS is $\rho = 0.9$ for orbital period and epoch. Trial transit duration is specified by a discrete list furnished to TPS, and we have accepted $\rho = 0.5$ for the transit duration minimum correlation coefficient, although this will be tightened up as we proceed to search for transiting planets over multiple quarters. Starting with the minimum trial

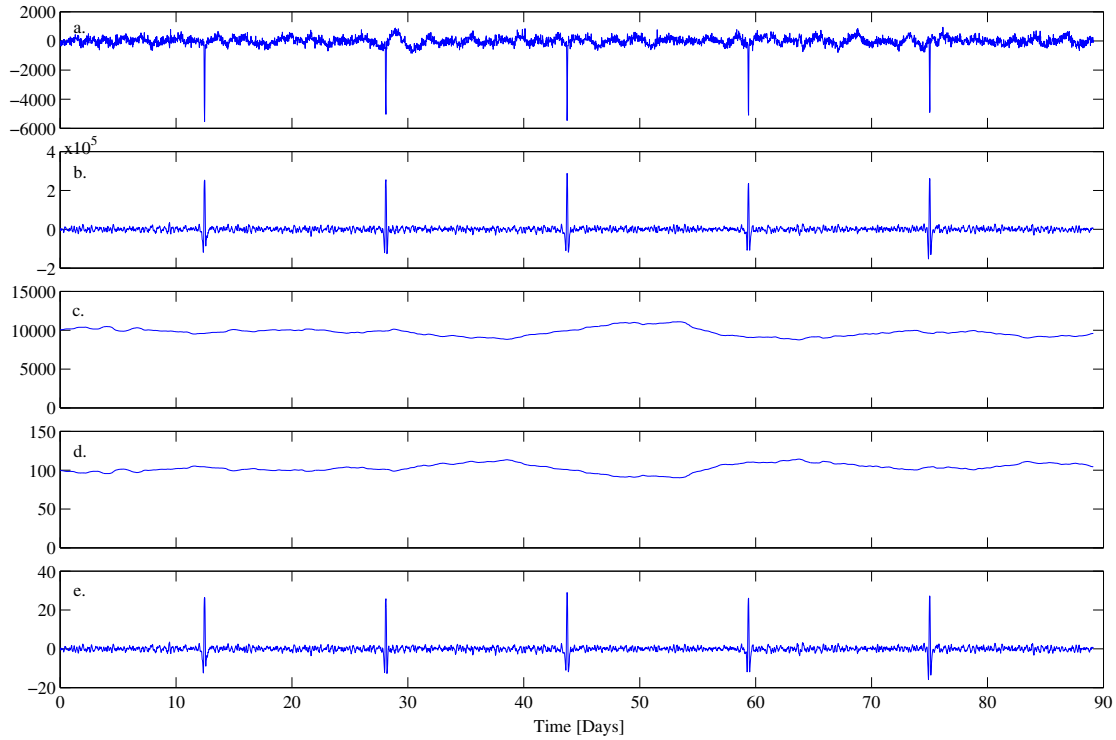

Figure 6. Calculation of CDPP for one target star. a: Normalized target flux in parts per million (ppm). b: Correlation time series $\mathbb{N}(n)$ from Eq. 7. c: Normalization time series $\mathbb{D}(n)$ from Eq. 7. d: 3-hr CDPP time series. e: Single-event statistic time series, $T(n)$. In all cases, the trial transit pulse, $s(n)$, is a square pulse of unit depth and 3-hour duration.

orbital period (usually 1 day), TPS applies Eq. 9 to determine the next trial orbital period, continuing until the maximum trial orbital period, the length of the time series, is reached. To form a multiple-event statistic for given point $\{t_0, T_p, D\}$, TPS computes the correlation and SNR time series, $\mathbb{N}(n)$ and $\mathbb{D}(n)$, and then loops over the trial orbital periods, folding these time series at each orbital period (rounded to the nearest number of samples) and summing the numerator and denominator terms falling in the each epoch bin. TPS identifies the maximum multiple-event statistic and its corresponding epoch. TPS also identifies and returns the maximum single-event statistic for each trial transit duration.[18] Figure 8 illustrates this process for the flux time series appearing in Figure 6.

To preserve sensitivity to short duration transits and small orbital periods, TPS supports a super-resolution search with respect to epoch and orbital period. This is accomplished by shifting the trial transit pulse by a fraction of a transit duration, generating the single-event statistic time series components for this shifted transit, then interleaving the results with the original transit pulse's single-event statistics. For example, a three-hour transit pulse lasts six LC samples: $[0,-1,-1,-1,-1,-1,-1,0]$. Shifting this transit by 10 minutes or one third of an LC earlier we obtain the sequence $\left[-\frac{1}{3},-1,-1,-1,-1,-1,-\frac{2}{3},0\right]$ with corresponding single-event detection statistics $\mathbb{N}_{+1/3}(n)$ and $\mathbb{D}_{+1/3}(n)$. Shifting the original transit pulse by 10 minutes later, we obtain the sequence $\left[0,-\frac{2}{3},-1,-1,-1,-1,-1,-\frac{1}{3}\right]$ with corresponding single-event detection statistics $\mathbb{N}_{-1/3}(n)$ and $\mathbb{D}_{-1/3}(n)$. We combine the results from all three analyses schematically as

$$\mathbb{N}(n) = \{\ldots, \mathbb{N}_{+1/3}(k), \mathbb{N}_0(k), \mathbb{N}_{-1/3}(k), \mathbb{N}_{+1/3}(k+1), \mathbb{N}_0(k+1), \mathbb{N}_{-1/3}(k+1), \ldots\}, \tag{10}$$

where we've denoted the original time series as $\mathbb{N}_0(n)$. A similar expression applies for the super-resolution denominator term, $\mathbb{D}(n)$. The folding proceeds exactly as before, except that now a sample is 9.8 minutes rather than 29.4 minutes.

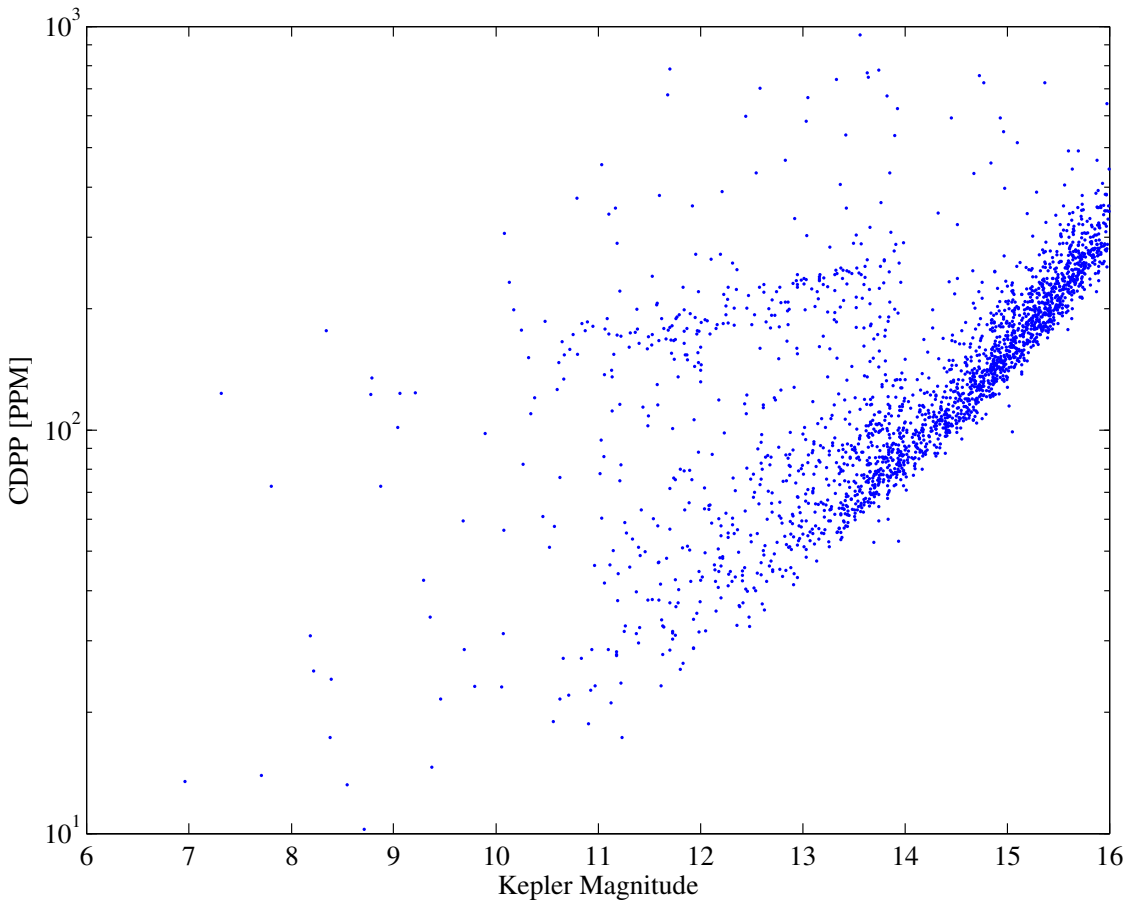

Figure 7. Three-hour CDPP as a function of *Kepler* magnitude for 2,286 stars on Module 7, Output 3, for one representative quarter.

5. CONCLUSIONS

Five planets have been discovered and announced by the *Kepler* team as of January 2010.[1] Several hundred potential planets are being vetted and followed by the Followup Observing Program. TPS has been quite productive in identifying Threshold Crossing Events in individual quarters and soon will be capable of detecting planetary signatures across the complete data set. TPS does trigger TCEs for a significant number of non-transit or eclipse events due to pixel sensitivity dropouts, flare events, and other isolated and cluster outliers. Near-term development includes better identification of step discontinuities due to pixel sensitivity dropouts in systematic error corrections made by PDC and also increasing the robustness of TPS to such events. These steps should reduce the number of TCEs that are analyzed by the DV component while preserving TPS's sensitivity to transit signatures.

ACKNOWLEDGMENTS

The authors would like to thank David Pletcher for his leadership in the SOC, and Bill Borucki and David Koch for their leadership of the *Kepler Mission*. We also thank Sue Blumenberg for her careful reading and editing of this manuscript.

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

REFERENCES

[1] Borucki, W., *et al.*, "*Kepler* planet-detection mission: introduction and first results," *Science* **327**, 977–980 (2010).

[2] Koch, D. G., *et al.*, "*Kepler* mission design, realized photometric performance, and early science," *ApJL* **713(2)**, L79–L86 (2010).

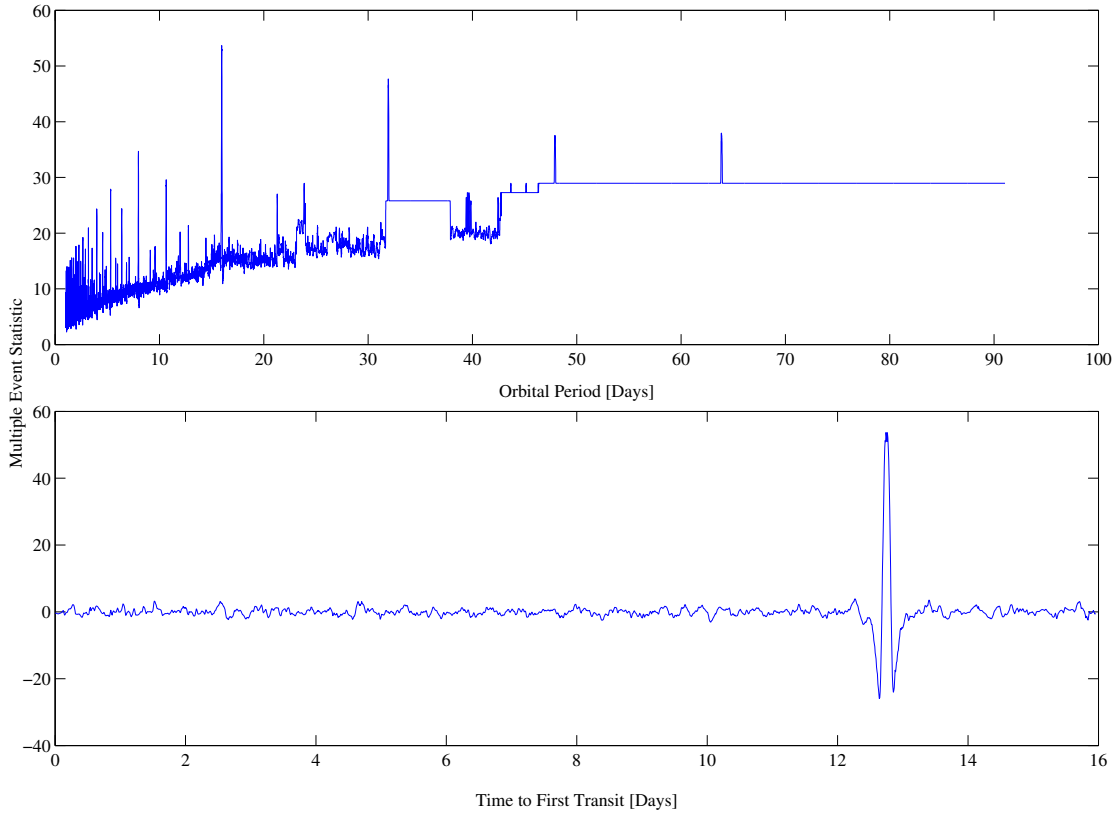

Figure 8. Multiple-event statistics determined by folding the single-event statistics distribution. Top: maximum multiple-event statistic as a function of fold interval (orbital period), showing a peak at 15.97 days, corresponding to the orbital period of the transiting object in the data of Figure 4. Bottom: multiple-event statistic for 15.97 day period as a function of lag time, showing a peak at 12.74 days, corresponding to the mid-time of the first transit shown in Figure 4.

[3] Jenkins, J. M., *et al.*, "Overview of the *Kepler* science processing pipeline," *ApJL* **713(2)**, L87–L91 (2010).

[4] Middour, C., *et al.*, "*Kepler* Science Operations Center architecture," *Proc. SPIE* **7740**, in press (2010).

[5] Haas, M., *et al.*, "*Kepler* science operations," *ApJL* **713(2)**, L115–L119 (2010).

[6] Quintana, E. M., *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] Twicken, J. D., *et al.*, "Presearch data conditioning in the *Kepler* Science Operations Center pipeline," *Proc. SPIE* **7740**, in press (2010).

[9] Jenkins, J. M., "The impact of solar-like variability on the detectability of transiting terrestrial planets," *ApJ* **575(1)**, 493–505 (2002).

[10] Tenenbaum, P., *et al.*, "An algorithm for fitting of planet models to *Kepler* light curves," *Proc. SPIE* **7740**, in press (2010).

[11] Wu, H., *et al.*, "Data validation in the *Kepler* Science Operations Center pipeline," *Proc. SPIE* **7740**, in press (2010).

[12] Li, J., *et al.*, "Photometer performance assessment in *Kepler* science data processing," *Proc. SPIE* **7740**, in press (2010).

[13] Jenkins, J. M., and Doyle, L. R., "Detecting reflected light from close-in extrasolar giant planets with the *Kepler* photometer," *ApJ* **595**, 429–445 (2003).

[14] Akaike, H., "A new look at the statistical model identification," *IEEE Trans. on Auto. Control* **19(6)**, 716–723 (1974).

[15] Kay, S., "Adaptive detection for unknown noise power spectral densities," *IEEE Trans. on Sig. Proc.* **47(1)**, 10–21(1999).

[16] Debauchies, I., "Orthonormal bases of compactly supported wavelets," *Comm. on Pure & Appl. Math.* **41**, 909–996 (1988).

[17] Jenkins, J. M., Doyle, L. R., and Cullers, K, "A matched-filter method for ground-based sub-noise detection of extrasolar planets in eclipsing binaries: Application to CM Draconis," *Icarus* **119**, 244–260 (1996).

[18] McCauliff, S. D., *et al.*, "The *Kepler* DB: a database management system for arrays, sparse arrays, and binary objects," *Proc. SPIE* **7740**, in press (2010).